Applied Regression Analysis: A Research Tool, Second Edition

John O. Rawlings
Sastry G. Pantula
David A. Dickey

Springer

Springer Texts in Statistics

Advisors:
George Casella Stephen Fienberg Ingram Olkin

Springer
New York
Berlin
Heidelberg
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Springer Texts in Statistics

Alfred: Elements of Statistics for the Life and Social Sciences
Berger: An Introduction to Probability and Stochastic Processes
Bilodeau and Brenner: Theory of Multivariate Statistics
Blom: Probability and Statistics: Theory and Applications
Brockwell and Davis: An Introduction to Times Series and Forecasting
Chow and Teicher: Probability Theory: Independence, Interchangeability, Martingales, Third Edition
Christensen: Plane Answers to Complex Questions: The Theory of Linear Models, Second Edition
Christensen: Linear Models for Multivariate, Time Series, and Spatial Data
Christensen: Log-Linear Models and Logistic Regression, Second Edition
Creighton: A First Course in Probability Models and Statistical Inference
Dean and Voss: Design and Analysis of Experiments
du Toit, Steyn, and Stumpf: Graphical Exploratory Data Analysis
Durrett: Essentials of Stochastic Processes
Edwards: Introduction to Graphical Modelling, Second Edition
Finkelstein and Levin: Statistics for Lawyers
Flury: A First Course in Multivariate Statistics
Jobson: Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design
Jobson: Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods
Kalbfleisch: Probability and Statistical Inference, Volume I: Probability, Second Edition
Kalbfleisch: Probability and Statistical Inference, Volume II: Statistical Inference, Second Edition
Karr: Probability
Keyfitz: Applied Mathematical Demography, Second Edition
Kiefer: Introduction to Statistical Inference
Kokoska and Nevison: Statistical Tables and Formulae
Kulkarni: Modeling, Analysis, Design, and Control of Stochastic Systems
Lehmann: Elements of Large-Sample Theory
Lehmann: Testing Statistical Hypotheses, Second Edition
Lehmann and Casella: Theory of Point Estimation, Second Edition
Lindman: Analysis of Variance in Experimental Design
Lindsey: Applying Generalized Linear Models
Madansky: Prescriptions for Working Statisticians
McPherson: Applying and Interpreting Statistics: A Comprehensive Guide, Second Edition
Mueller: Basic Principles of Structural Equation Modeling: An Introduction to LISREL and EQS

(continued after index)

John O. Rawlings
Sastry G. Pantula
David A. Dickey

Applied Regression Analysis

A Research Tool

Second Edition

With 78 Figures

Springer

John O. Rawlings
Sastry G. Pantula
David A. Dickey
Department of Statistics
North Carolina State University
Raleigh, NC 27695
USA

Library of Congress Cataloging-in-Publication Data
Rawlings, John O., 1932–
Applied regression analysis: a research tool. — 2nd ed. / John O. Rawlings, Sastry G. Pentula, David A. Dickey.
p. cm. — (Springer texts in statistics)
Includes bibliographical references and indexes.
ISBN 0-387-98454-2 (hardcover: alk. paper)
1. regression analysis. I. Pentula, Sastry G. II. Dickey, David A. III. Title. IV. Series.
QA278.2.R38 1998
519.5′36—dc21 97-48858

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

ISBN 0-387-98454-2 Springer-Verlag New York Berlin Heidelberg SPIN 10660129

To

Our Families

PREFACE

This text is a new and improved edition of Rawlings (1988). It is the outgrowth of several years of teaching an applied regression course to graduate students in the sciences. Most of the students in these classes had taken a two-semester introduction to statistical methods that included experimental design and multiple regression at the level provided in texts such as Steel, Torrie, and Dickey (1997) and Snedecor and Cochran (1989). For most, the multiple regression had been presented in matrix notation.

The basic purpose of the course and this text is to develop an understanding of least squares and related statistical methods without becoming excessively mathematical. The emphasis is on regression concepts, rather than on mathematical proofs. Proofs are given only to develop facility with matrix algebra and comprehension of mathematical relationships. Good students, even though they may not have strong mathematical backgrounds, quickly grasp the essential concepts and appreciate the enhanced understanding. The learning process is reinforced with continuous use of numerical examples throughout the text and with several case studies. Some numerical and mathematical exercises are included to whet the appetite of graduate students.

The first four chapters of the book provide a review of simple regression in algebraic notation (Chapter 1), an introduction to key matrix operations and the geometry of vectors (Chapter 2), and a review of ordinary least squares in matrix notation (Chapters 3 and 4). Chapter 4 also provides a foundation for the testing of hypotheses and the properties of sums of squares used in analysis of variance. Chapter 5 is a case study giving a complete multiple regression analysis using the methods reviewed in the

first four chapters. Then Chapter 6 gives a brief geometric interpretation of least squares illustrating the relationships among the data vectors, the link between the analysis of variance and the lengths of the vectors, and the role of degrees of freedom. Chapter 7 discusses the methods and criteria for determining which independent variables should be included in the models. The next two chapters include special classes of multiple regression models. Chapter 8 introduces polynomial and trigonometric regression models. This chapter also discusses response curve models that are linear in the parameters. Class variables and the analysis of variance of designed experiments (models of less than full rank) are introduced in Chapter 9.

Chapters 10 through 14 address some of the problems that might be encountered in regression. A general introduction to the various kinds of problems is given in Chapter 10. This is followed by discussions of regression diagnostic techniques (Chapter 11), and scaling or transforming variables to rectify some of the problems (Chapter 12). Analysis of the correlational structure of the data and biased regression are discussed as techniques for dealing with the collinearity problem common in observational data (Chapter 13). Chapter 14 is a case study illustrating the analysis of data in the presence of collinearity.

Models that are nonlinear in the parameters are presented in Chapter 15. Chapter 16 is another case study using polynomial response models, nonlinear modeling, transformations to linearize, and analysis of residuals. Chapter 17 addresses the analysis of unbalanced data. Chapter 18 (new to this edition) introduces linear models that have more than one random effect. The ordinary least squares approach to such models is given. This is followed by the definition of the variance–covariance matrix for such models and a brief introduction to mixed effects and random coefficient models. The use of iterative maximum likelihood estimation of both the variance components and the fixed effects is discussed. The final chapter, Chapter 19, is a case study of the analysis of unbalanced data.

We are grateful for the assistance of many in the development of this book. Of particular importance have been the dedicated editing of the earlier edition by Gwen Briggs, daughter of John Rawlings, and her many suggestions for improvement. It is uncertain when the book would have been finished without her support. A special thanks goes to our former student, Virginia Lesser, for her many contributions in reading parts of the manuscript, in data analysis, and in the enlistment of many data sets from her graduate student friends in the biological sciences. We are indebted to our friends, both faculty and students, at North Carolina State University for bringing us many interesting consulting problems over the years that have stimulated the teaching of this material. We are particularly indebted to those (acknowledged in the text) who have generously allowed the use of their data. In this regard, Rick Linthurst warrants special mention for his stimulating discussions as well as the use of his data. We acknowledge the encouragement and valuable discussions of colleagues in the Department

of Statistics at NCSU, and we thank Matthew Sommerville for checking answers to the exercises. We wish to thank Sharon Sullivan and Dawn Haines for their help with LaTeX. Finally, we want to express appreciation for the critical reviews and many suggestions provided for the first edition by the Wadsworth Brooks/Cole reviewers: Mark Conaway, University of Iowa; Franklin Graybill, Colorado State University; Jason Hsu, Ohio State University; Kenneth Koehler, Iowa State University; B. Lindsay, The Pennsylvania State University; Michael Meridith, Cornell University; M. B. Rajarshi, University of Poona (India); Muni Srivastava, University of Toronto; and Patricia Wahl, University of Washington; and for the second edition by the Springer-Verlag reviewers.

Acknowledgment is given for the use of material in the appendix tables. Appendix Table A.7 is reproduced in part from Tables 4 and 6 of Durbin and Watson (1951) with permission of the Biometrika Trustees. Appendix Table A.8 is reproduced with permission from Shapiro and Francia (1972), *Journal of the American Statistical Association*. The remaining appendix tables have been computer generated by one of the authors. We gratefully acknowledge permission of other authors and publishers for use of material from their publications as noted in the text.

Note to the Reader

Most research is aimed at quantifing relationships among variables that either measure the end result of some process or are likely to affect the process. The process in question may be any biological, chemical, or physical process of interest to the scientist. The quantification of the process may be as simple as determining the degree of association between two variables or as complicated as estimating the many parameters of a very detailed nonlinear mathematical model of the system.

Regardless of the degree of sophistication of the model, the most commonly used statistical method for estimating the parameters of interest is the method of **least squares**. The criterion applied in least squares estimation is simple and has great intuitive appeal. The researcher chooses the model that is believed to be most appropriate for the project at hand. The parameters for the model are then estimated such that the predictions from the model and the observed data are in as good agreement as possible as measured by the **least squares criterion**, minimization of the sum of squared differences between the predicted and the observed points.

Least squares estimation is a powerful research tool. Few assumptions are required and the estimators obtained have several desirable properties. Inference from research data to the true behavior of a process, however, can be a difficult and dangerous step due to unrecognized inadequacies in the data, misspecification of the model, or inappropriate inferences of

causality. As with any research tool it is important that the least squares method be thoroughly understood in order to eliminate as much misuse or misinterpretation of the results as possible. There is a distinct difference between understanding and pure memorization. Memorization can make a good technician, but it takes understanding to produce a master. A discussion of the **geometric interpretation** of least squares is given to enhance your understanding. You may find your first exposure to the geometry of least squares somewhat traumatic but the visual perception of least squares is worth the effort. We encourage you to tackle the topic in the spirit in which it is included.

The general topic of least squares has been broadened to include statistical techniques associated with **model development and testing**. The backbone of least squares is the classical multiple regression analysis using the linear model to relate several independent variables to a response or dependent variable. Initially, this classical model is assumed to be appropriate. Then methods for detecting inadequacies in this model and possible remedies are discussed.

The connection between the analysis of variance for designed experiments and multiple regression is developed to build the foundation for the analysis of **unbalanced data**. (This also emphasizes the generality of the least squares method.) Interpretation of unbalanced data is difficult. It is important that the application of least squares to the analysis of such data be understood if the results from computer programs designed for the analysis of unbalanced data are to be used correctly.

The objective of a research project determines the amount of effort to be devoted to the development of realistic models. If the intent is one of prediction only, the degree to which the model might be considered realistic is immaterial. The only requirement is that the predictions be adequately precise in the region of interest. On the other hand, realism is of primary importance if the goal is a thorough understanding of the system. The simple linear additive model can seldom be regarded as a realistic model. It is at best an approximation of the true model. Almost without exception, models developed from the basic principles of a process will be nonlinear in the parameters. The least squares estimation principle is still applicable but the mathematical methods become much more difficult. You are introduced to **nonlinear least squares regression methods** and some of the more common nonlinear models.

Least squares estimation is controlled by the correlational structure observed among the independent and dependent variables in the data set. Observational data, data collected by observing the state of nature according to some sampling plan, will frequently cause special problems for least squares estimation because of strong correlations or, more generally, near-linear dependencies among the independent variables. The seriousness of the problems will depend on the use to be made of the analyses. Understanding the correlational structure of the data is most helpful in in-

terpreting regression results and deciding what inferences might be made. Principal component analysis is introduced as an aid in characterizing the correlational structure of the data. A graphical procedure, Gabriel's biplot, is introduced to help visualize the correlational structure. Principal component analysis also serves as an introduction to **biased regression methods**. Biased regression methods are designed to alleviate the deleterious effects of near-linear dependencies (among the independent variables) on ordinary least squares estimation.

Least squares estimation is a powerful research tool and, with modern low cost computers, is readily available. This ease of access, however, also facilitates misuse. Proper use of least squares requires an understanding of the basic method and assumptions on which it is built, and an awareness of the possible problems and their remedies. In some cases, alternative methods to least squares estimation might be more appropriate. It is the intent of this text to convey the basic understanding that will allow you to use least squares as an effective research tool.

The data sets used in this text are available on the internet at

http://www.stat.ncsu.edu/publications/rawlings/applied_least_squares

or through a link at the Springer-Verlag page. The "readme" file explains the contents of each data set.

Raleigh, North Carolina
March 4, 1998

John O. Rawlings
Sastry G. Pantula
David A. Dickey

CONTENTS

1

REVIEW OF SIMPLE REGRESSION

This chapter reviews the elementary regression results for a linear model in one variable. The primary purpose is to establish a common notation and to point out the need for matrix notation. A light reading should suffice for most students.

Modeling refers to the development of mathematical expressions that describe in some sense the behavior of a random variable of interest. This variable may be the price of wheat in the world market, the number of deaths from lung cancer, the rate of growth of a particular type of tumor, or the tensile strength of metal wire. In all cases, this variable is called the **dependent variable** and denoted with Y. A subscript on Y identifies the particular unit from which the observation was taken, the time at which the price was recorded, the county in which the deaths were recorded, the experimental unit on which the tumor growth was recorded, and so forth. Most commonly the modeling is aimed at describing how the **mean** of the dependent variable $\mathcal{E}(Y)$ changes with changing conditions; the variance of the dependent variable is assumed to be unaffected by the changing conditions.

Other variables which are thought to provide information on the behavior of the dependent variable are incorporated into the model as predictor or explanatory variables. These variables are called the **independent variables** and are denoted by X with subscripts as needed to identify different independent variables. Additional subscripts denote the observational unit from which the data were taken. The Xs are assumed to be known con-

stants. In addition to the Xs, all models involve unknown constants, called **parameters**, which control the behavior of the model. These parameters are denoted by Greek letters and are to be estimated from the data.

The mathematical complexity of the model and the degree to which it is a realistic model depend on how much is known about the process being studied and on the purpose of the modeling exercise. In preliminary studies of a process or in cases where prediction is the primary objective, the models usually fall into the class of models that are **linear in the parameters**. That is, the parameters enter the model as simple coefficients on the independent variables or functions of the independent variables. Such models are referred to loosely as **linear models**. The more realistic models, on the other hand, are often **nonlinear in the parameters**. Most growth models, for example, are nonlinear models. Nonlinear models fall into two categories: **intrinsically linear models**, which can be linearized by an appropriate transformation on the dependent variable, and those that cannot be so transformed. Most of the discussion is devoted to the linear class of models and to those nonlinear models that are intrinsically linear. Nonlinear models are discussed in Section 12.2 and Chapter 15.

1.1 The Linear Model and Assumptions

Model

The simplest linear model involves only one independent variable and states that the true mean of the dependent variable changes at a constant rate as the value of the independent variable increases or decreases. Thus, the functional relationship between the true mean of Y_i, denoted by $\mathcal{E}(Y_i)$, and X_i is the equation of a straight line:

$$\mathcal{E}(Y_i) = \beta_0 + \beta_1 X_i. \tag{1.1}$$

β_0 is the intercept, the value of $\mathcal{E}(Y_i)$ when $X = 0$, and β_1 is the slope of the line, the rate of change in $\mathcal{E}(Y_i)$ per unit change in X.

Assumptions

The observations on the dependent variable Y_i are assumed to be random observations from populations of random variables with the mean of each population given by $\mathcal{E}(Y_i)$. The deviation of an observation Y_i from its population mean $\mathcal{E}(Y_i)$ is taken into account by adding a random error ϵ_i to give the statistical model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i. \tag{1.2}$$

The subscript i indicates the particular observational unit, $i = 1, 2, \ldots, n$. The X_i are the n observations on the independent variable and are assumed to be measured without error. That is, the observed values of X are assumed to be a set of known constants. The Y_i and X_i are paired observations; both are measured on every observational unit.

The random errors ϵ_i have zero mean and are assumed to have common variance σ^2 and to be pairwise independent. Since the only random element in the model is ϵ_i, these assumptions imply that the Y_i also have common variance σ^2 and are pairwise independent. For purposes of making tests of significance, the random errors are assumed to be normally distributed, which implies that the Y_i are also normally distributed. The random error assumptions are frequently stated as

$$\epsilon_i \sim \ NID(0, \sigma^2), \tag{1.3}$$

where NID stands for "normally and independently distributed." The quantities in parentheses denote the mean and the variance, respectively, of the normal distribution.

1.2 Least Squares Estimation

The simple linear model has two parameters β_0 and β_1, which are to be estimated from the data. If there were no random error in Y_i, any two data points could be used to solve explicitly for the values of the parameters. The random variation in Y, however, causes each pair of observed data points to give different results. (All estimates would be identical only if the observed data fell exactly on the straight line.) A method is needed that will combine all the information to give one solution which is "best" by some criterion.

Least Squares Criterion

The **least squares estimation procedure** uses the criterion that the solution must give the smallest possible sum of squared deviations of the observed Y_i from the estimates of their true means provided by the solution. Let $\widehat{\beta}_0$ and $\widehat{\beta}_1$ be numerical estimates of the parameters β_0 and β_1, respectively, and let

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i \tag{1.4}$$

be the estimated mean of Y for each X_i, $i = 1, \ldots, n$. Note that $\widehat{Y}_i$ is obtained by substituting the estimates for the parameters in the functional form of the model relating $\mathcal{E}(Y_i)$ to X_i, equation 1.1. The least squares principle chooses $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimize the sum of squares of the residuals, SS(Res):

$$\begin{aligned} SS(Res) &= \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 \\ &= \sum e_i^2, \end{aligned} \tag{1.5}$$

where $e_i = (Y_i - \widehat{Y}_i)$ is the observed residual for the ith observation. The summation indicated by $\sum$ is over all observations in the data set as indi-

cated by the index of summation, $i = 1$ to n. (The index of summation is omitted when the limits of summation are clear from the context.)

The estimators for β_0 and β_1 are obtained by using calculus to find the values that minimize SS(Res). The derivatives of SS(Res) with respect to $\widehat{\beta}_0$ and $\widehat{\beta}_1$ in turn are set equal to zero. This gives two equations in two unknowns called the **normal equations**:

$$\begin{aligned} n(\widehat{\beta}_0) + (\sum X_i)\widehat{\beta}_1 &= \sum Y_i \\ (\sum X_i)\widehat{\beta}_0 + (\sum X_i^2)\widehat{\beta}_1 &= \sum X_i Y_i. \end{aligned} \tag{1.6}$$

Solving the normal equations simultaneously for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ gives the estimates of β_1 and β_0 as

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2} \\ \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X}. \end{aligned} \tag{1.7}$$

Note that $x_i = (X_i - \overline{X})$ and $y_i = (Y_i - \overline{Y})$ denote observations expressed as deviations from their sample means $\overline{X}$ and $\overline{Y}$, respectively. The more convenient forms for hand computation of sums of squares and sums of products are

$$\begin{aligned} \sum x_i^2 &= \sum X_i^2 - \frac{(\sum X_i)^2}{n} \\ \sum x_i y_i &= \sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}. \end{aligned} \tag{1.8}$$

Thus, the computational formula for the slope is

$$\widehat{\beta}_1 = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}. \tag{1.9}$$

These estimates of the parameters give the regression equation

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i. \tag{1.10}$$

Example 1.1

The computations for the linear regression analysis are illustrated using treatment mean data from a study conducted by Dr. A. S. Heagle at North Carolina State University on effects of ozone pollution on soybean yield (Table 1.1). Four dose levels of ozone and the resulting mean seed yield of soybeans are given. The dose of ozone is the average concentration (parts per million, ppm) during the growing season. Yield is reported in grams per plant.

TABLE 1.1. *Mean yields of soybean plants (gms per plant) obtained in response to the indicated levels of ozone exposure over the growing season. (Data courtesy of Dr. A. S. Heagle, USDA and North Carolina State University.)*

X		Y	
Ozone (ppm)		*Yield (gm/plt)*	
	.02		242
	.07		237
	.11		231
	.15		201
$\sum X_i =$	.35	$\sum Y_i =$	911
$\overline{X} =$	.0875	$\overline{Y} =$	227.75
$\sum X_i^2 =$	.0399	$\sum Y_i^2 =$	$208,495$
	$\sum X_i Y_i = 76.99$		

Assuming a linear relationship between yield and ozone dose, the simple linear model, described by equation 1.2, is appropriate. The estimates of β_0 and β_1 obtained from equations 1.7 and 1.9 are

$$\begin{aligned} \widehat{\beta}_1 &= \frac{76.99 - \frac{(.35)(911)}{4}}{.0399 - \frac{(.35)^2}{4}} = -293.531 \\ \widehat{\beta}_0 &= 227.75 - (-293.531)(.0875) = 253.434. \end{aligned} \tag{1.11}$$

The least squares regression equation characterizing the effects of ozone on the mean yield of soybeans in this study, assuming the linear model is correct, is

$$\widehat{Y}_i = 253.434 - 293.531 X_i. \tag{1.12}$$

The interpretation of $\widehat{\beta}_1 = -294$ is that the mean yield is expected to decrease, since the slope is negative, by 294 grams per plant with each 1 ppm increase in ozone, or 2.94 grams with each .01 ppm increase in ozone. The observed range of ozone levels in the experiment was .02 ppm to .15 ppm. Therefore, it would be an unreasonable extrapolation to expect this rate of decrease in yield to continue if ozone levels were to increase, for example, to as much as 1 ppm. It is safe to use the results of regression only within the range of values of the independent variable. The intercept, $\widehat{\beta}_0 = 253$ grams, is the value of Y where the regression line crosses the Y-axis. In this case, since the lowest dose is .02 ppm, it would be an extrapolation to interpret $\widehat{\beta}_0$ as the estimate of the mean yield when there is no ozone. ■

TABLE 1.2. *Observed values, estimated values, and residuals for the linear regression of soybean yield on ozone dosage.*

Y_i	$\widehat{Y}_i$	e_i	e_i^2
242	247.563	−5.563	30.947
237	232.887	4.113	16.917
231	221.146	9.854	97.101
201	209.404	−8.404	70.627
		$\sum e_i = 0.0$	$\sum e_i^2 = 215.592$

1.3 Predicted Values and Residuals

The regression equation from Example 1.1 can be evaluated to obtain estimates of the mean of the dependent variable Y at chosen levels of the independent variable. Of course, the validity of such estimates is dependent on the assumed model being correct, or at least a good approximation to the correct model within the limits of the pollution doses observed in the study.

Estimates and Predictions

Each quantity computed from the fitted regression line $\widehat{Y}_i$ is used as both (1) the **estimate** of the population mean of Y for that particular value of X and (2) the **prediction** of the value of Y one might obtain on some future observation at that level of X. Hence, the $\widehat{Y}_i$ are referred to both as **estimates** and as **predicted values.** On occasion we write $\widehat{\boldsymbol{Y}}_{pred_i}$ to clearly imply the second interpretation.

Residuals

If the observed values Y_i in the data set are compared with their corresponding values $\widehat{Y}_i$ computed from the regression equation, a measure of the degree of agreement between the model and the data is obtained. Remember that the least squares principle makes this agreement as "good as possible" in the least squares sense. The residuals

$$e_i = Y_i - \widehat{Y}_i \tag{1.13}$$

measure the discrepancy between the data and the fitted model. The results for Example 1.1 are shown in Table 1.2. Notice that the residuals sum to zero, as they always will when the model includes the constant term β_0. The least squares estimation procedure has minimized the sum of squares of the e_i. That is, there is no other choice of values for the two parameters β_0 and β_1 that will provide a smaller $\sum e_i^2$.

Example 1.2

A plot of the regression equation and the data from Example 1.1 (Figure 1.1) provides a visual check on the arithmetic and the adequacy with which the equation characterizes the data. The regression line crosses the Y-axis at the value of $\widehat{\beta}_0 = 253.4$. The negative sign on $\widehat{\beta}_1$ is reflected in

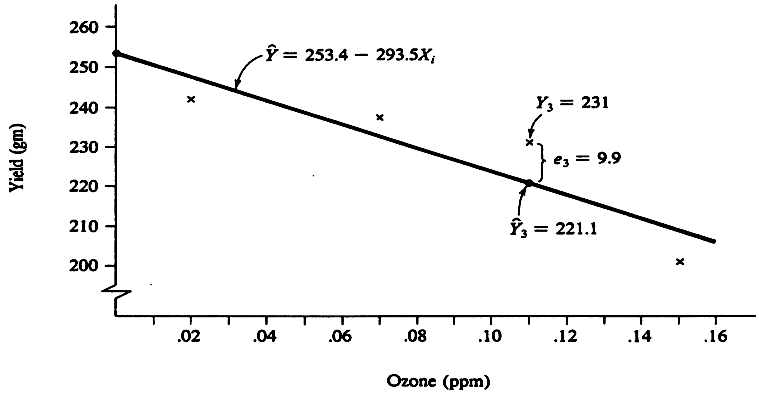

FIGURE 1.1. *Regression of soybean yield on ozone level.*

the negative slope. Inspection of the plot shows that the regression line decreases to approximately $Y = 223$ when $X = .1$. This is a decrease of 30 grams of yield over a .1 ppm increase in ozone, or a **rate** of change of -300 grams in Y for each unit increase in X. This is reasonably close to the computed value of -293.5 grams per ppm. Figure 1.1 shows that the regression line "passes through" the data as well as could be expected from a straight-line relationship. The pattern of the deviations from the regression line, however, suggests that the linear model may not adequately represent the relationship. ■

1.4 Analysis of Variation in the Dependent Variable

The residuals are defined in equation 1.13 as the deviations of the observed values from the estimated values provided by the regression equation. Alternatively, each observed value of the dependent variable Y_i can be written as the sum of the estimated population mean of Y for the given value of X and the corresponding residual:

$$Y_i = \widehat{Y}_i + e_i. \tag{1.14}$$

$\widehat{Y}$ is the part of the observation Y_i "accounted for" by the model, whereas e_i reflects the "unaccounted for" part.

SS(Model) and SS(RES)

The **total uncorrected sum of squares** of Y_i, $\text{SS(Total}_{\text{uncorr}}) = \sum Y_i^2$, can be similarly partitioned. Substitute $\widehat{Y}_i + e_i$ for each Y_i and

expand the square. Thus,

$$\begin{aligned}\sum Y_i^2 &= \sum(\widehat{Y}_i + e_i)^2 \\ &= \sum \widehat{Y}_i^2 + \sum e_i^2 \\ &= \text{SS(Model)} + \text{SS(Res)}. \end{aligned} \tag{1.15}$$

(The cross-product term $\sum \widehat{Y}_i e_i$ is zero, as can readily be shown with the matrix notation of Chapter 3. Also see Exercise 1.22.) The term SS(Model) is the sum of squares "accounted for" by the model; SS(Res) is the "unaccounted for" part of the sum of squares. The forms $\text{SS(Model)} = \sum \widehat{Y}_i^2$ and $\text{SS(Res)} = \sum e_i^2$ show the origins of these sums of squares. The more convenient computational forms are

$$\begin{aligned}\text{SS(Model)} &= n\overline{Y}^2 + \widehat{\beta}_1^2 \sum (X_i - \overline{X})^2 \\ \text{SS(Res)} &= \text{SS(Total}_{\text{uncorr}}) - \text{SS(Model)}. \end{aligned} \tag{1.16}$$

The partitioning of the total uncorrected sum of squares can be reexpressed in terms of the **corrected sum of squares** by subtracting the sum of squares due to correction for the mean, the correction factor $n\overline{Y}^2$, from each side of equation 1.15:

$$\text{SS(Total}_{\text{uncorr}}) - n\overline{Y}^2 = [\text{SS(Model)} - n\overline{Y}^2] + \text{SS(Res)}$$

or, using equation 1.16,

$$\begin{aligned}\sum y_i^2 &= \widehat{\beta}_1^2 \sum (X_i - \overline{X})^2 + \sum e_i^2 \\ &= \text{SS(Regr)} + \text{SS(Res)}. \end{aligned} \tag{1.17}$$

Notice that lower case y is the deviation of Y from $\overline{Y}$ so that $\sum y_i^2$ is the **corrected total sum of squares**. Henceforth, SS(Total) is used to denote the corrected sum of squares of the dependent variable. Also notice that SS(Model) denotes the sum of squares attributable to the entire model, whereas SS(Regr) denotes only that part of SS(Model) that exceeds the correction factor. The correction factor is the sum of squares for a model that contains *only* the constant term β_0. Such a model postulates that the mean of Y is a constant, or is unaffected by changes in X. Thus, SS(Regr) measures the *additional* information provided by the independent variable.

Degrees of Freedom

The degrees of freedom associated with each sum of squares is determined by the sample size n and the number of parameters p' in the model. [We use p' to denote the number of parameters in the model and p (without the prime) to denote the number of independent variables; $p' = p + 1$ when the model includes an intercept as in equation 1.2.] The degrees of freedom associated with SS(Model) is $p' = 2$; the degrees of freedom associated with SS(Regr) is always 1 less to account for subtraction of the correction factor,

TABLE 1.3. *Partitions of the degrees of freedom and sums of squares for yield of soybeans exposed to ozone (courtesy of Dr. A. S. Heagle, N.C. State University).*

Source of Variation	*Degrees of Freedom*	*Sum of Squares*		*Mean Square*
Total $_{\text{uncorr}}$	$n = 4$	$\sum Y_i^2 =$	208,495.00	
Corr. factor	1	$n\overline{Y}^2 =$	207,480.25	
$\text{Total}_{\text{corr}}$	$n - 1 = 3$	$\sum y_i^2 =$	1,014.75	
Due to model	$p' = 2$	$\sum \widehat{Y}_i^2 =$	208,279.39	
Corr. factor	1		207,480.25	
Due to regr.	$p' - 1 = 1$	$\sum \widehat{Y}_i^2 - n\overline{Y}^2 =$	799.14	799.14
Residual	$n - p' = 2$	$\sum e_i^2 =$	215.61	107.81

TABLE 1.4. *Analysis of variance of yield of soybeans exposed to ozone pollution (courtesy of Dr. A. S. Heagle, N.C. State University).*

Source	*d.f.*	*SS*	*MS*
Total	3	1014.75	
Due to regr.	1	799.14	799.14
Residual	2	215.61	107.81

which has 1 degree of freedom. SS(Res) will contain the $(n - p')$ degrees of freedom not accounted for by SS(Model). The mean squares are found by dividing each sum of squares by its degrees of freedom.

Example 1.3

The partitions of the degrees of freedom and sums of squares for the ozone data from Example 1.1 are given in Table 1.3. The definitional formulae for the sums of squares are included. An abbreviated form of Table 1.3, omitting the total uncorrected sum of squares, the correction factor, and SS(Model), is usually presented as the analysis of variance table (Table 1.4). ■

Coefficient of Determination

One measure of the contribution of the independent variable(s) in the model is the **coefficient of determination**, denoted by R^2:

$$R^2 = \frac{\text{SS(Regr)}}{\sum y_i^2}. \tag{1.18}$$

This is the proportion of the (corrected) sum of squares of Y attributable to the information obtained from the independent variable(s). The coefficient of determination ranges from zero to one and is the square of the product

moment correlation between Y_i and $\widehat{Y}_i$. If there is only one independent variable, it is also the square of the correlation coefficient between Y_i and X_i.

Example 1.4

The coefficient of determination for the ozone data from Example 1.1 is

$$R^2 = \frac{799.14}{1,014.75} = .7875.$$

The interpretation of R^2 is that 79% of the variation in the dependent variable, yield of soybeans, is "explained" by its linear relationship with the independent variable, ozone level. Caution must be exercised in the interpretation given to the phrase "explained by X." In this example, the data are from a controlled experiment where the level of ozone was being controlled in a properly replicated and randomized experiment. It is therefore reasonable to infer that any significant association of the variation in yield with variation in the level of ozone reflects a causal effect of the pollutant. If the data had been observational data, random observations on nature as it existed at some point in time and space, there would be no basis for inferring causality. Model-fitting can only reflect associations in the data. With observational data there are many reasons for associations among variables, only one of which is causality. ■

Expected Mean Squares

If the model is correct, the residual mean square is an unbiased estimate of σ^2, the variance among the random errors. The regression mean square is an unbiased estimate of $\sigma^2 + \beta_1^2(\sum x_i^2)$, where $\sum x_i^2 = \sum(X_i - \overline{X})^2$. These are referred to as the **mean square expectations** and are denoted by $\mathcal{E}$[MS(Res)] and $\mathcal{E}$[MS(Regr)]. Notice that MS(Regr) is estimating the same quantity as MS(Res) *plus* a positive quantity that depends on the magnitude of β_1 and $\sum x_i^2$. Thus, any linear relationship between Y and X, where $\beta_1 \neq 0$, will on the average make MS(Regr) larger than MS(Res). Comparison of MS(Regr) to MS(Res) provides the basis for judging the importance of the relationship.

Example 1.5

The estimate of σ^2 is denoted by s^2. For the data of Example 1.1, MS(Res) = s^2 = 107.81 (Table 1.4). MS(Regr) = 799.14 is much larger than s^2, which suggests that β_1 is not zero. Testing of the null hypothesis that $\beta_1 = 0$ is discussed in Section 1.6. ■

1.5 Precision of Estimates

Any quantity computed from random variables is itself a random variable. Thus, $\overline{Y}$, $\widehat{Y}$, e, $\widehat{\beta}_0$, and $\widehat{\beta}_1$ are random variables computed from the Y_i. Measures of precision, variances or standard errors of the estimates, provide a basis for judging the reliability of the estimates.

Variance of a Linear Function

The computed regression coefficients, the $\widehat{Y}_i$, and the residuals are all **linear functions** of the Y_i. Their variances can be determined using the basic definition of the variance of a linear function. Let $U = \sum a_i Y_i$ be an arbitrary linear function of the random variables Y_i, where the a_i are constants. The general formula for the variance of U is

$$\text{Var}(U) = \sum a_i^2 \text{Var}(Y_i) + \sum \sum_{i \neq j} a_i a_j \text{Cov}(Y_i, Y_j), \tag{1.19}$$

where the double summation is over all $n(n-1)$ possible pairs of terms where i and j are not equal. $\text{Cov}(\cdot, \cdot)$ denotes the covariance between the two variables indicated in the parentheses. (Covariance measures the tendency of two variables to increase or decrease together.) When the random variables are independent, as is assumed in the usual regression model, all of the covariances are zero and the double summation term disappears. If, in addition, the variances of the random variables are equal, again as in the usual regression model where $\text{Var}(Y_i) = \sigma^2$ for all i, the variance of the linear function reduces to

$$\text{Var}(U) = (\sum a_i^2)\sigma^2. \tag{1.20}$$

Variances of linear functions play an extremely important role in every aspect of statistics. Understanding the derivation of variances of linear functions will prove valuable; for this reason, we now give several examples.

Example 1.6

The variance of the sample mean of n observations is derived. The coefficient a_i on each Y_i in the sample mean is $1/n$. If the Y_i have common variance σ^2 and zero covariances (for example, if they are independent), equation 1.20 applies. The sum of squares of the coefficients is

$$\sum a_i^2 = n\left(\frac{1}{n}\right)^2 = \frac{1}{n}$$

and the variance of the mean becomes

$$\text{Var}(\overline{Y}) = \frac{\sigma^2}{n}, \tag{1.21}$$

which is the well-known result for the variance of the sample mean. ■

Example 1.7

In this example, the variance is derived for a linear contrast of three treatment means,

$$C = \overline{Y}_1 + \overline{Y}_2 - 2\overline{Y}_3. \tag{1.22}$$

If each mean is the average of n independent observations from the same population, the variance of each sample mean is equal to $\text{Var}(\overline{Y}_i) = \sigma^2/n$ and all covariances are zero. The coefficients on the $\overline{Y}_i$ are 1, 1, and -2. Thus,

$$\begin{aligned} \text{Var}(C) &= (1)^2\text{Var}(\overline{Y}_1) + (1)^2\text{Var}(\overline{Y}_2) + (-2)^2\text{Var}(\overline{Y}_3) \\ &= (1+1+4)\left(\frac{\sigma^2}{n}\right) = 6\left(\frac{\sigma^2}{n}\right). \end{aligned} \tag{1.23}$$

■

Variance of $\widehat{\beta}_1$

We now turn to deriving the variances of $\widehat{\beta}_1$, $\widehat{\beta}_0$, and $\widehat{Y}_i$. To determine the variance of $\widehat{\beta}_1$ express

$$\widehat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \tag{1.24}$$

as

$$\widehat{\beta}_1 = \left(\frac{x_1}{\sum x_i^2}\right) Y_1 + \left(\frac{x_2}{\sum x_i^2}\right) Y_2 + \cdots + \left(\frac{x_n}{\sum x_i^2}\right) Y_n. \tag{1.25}$$

(See Exercise 1.16 for justification for replacing y_i with Y_i.) The coefficient on each Y_i is $x_i/\sum x_j^2$, which is a constant in the regression model. The Y_i are assumed to be independent and to have common variance σ^2. Thus, the variance of $\widehat{\beta}_1$ is

$$\begin{aligned} \text{Var}(\widehat{\beta}_1) &= \left(\frac{x_1}{\sum x_i^2}\right)^2 \sigma^2 + \left(\frac{x_2}{\sum x_i^2}\right)^2 \sigma^2 + \cdots + \left(\frac{x_n}{\sum x_i^2}\right)^2 \sigma^2 \\ &= \frac{\sum x_i^2}{(\sum x_i^2)^2}\sigma^2 = \frac{\sigma^2}{\sum x_i^2}. \end{aligned} \tag{1.26}$$

Variance of $\widehat{\beta}_0$

Determining the variance of the intercept

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X} \tag{1.27}$$

is a little more involved. The random variables in this linear function are $\overline{Y}$ and $\widehat{\beta}_1$; the coefficients are 1 and $(-\overline{X})$. Equation 1.19 can be used to obtain the variance of $\widehat{\beta}_0$:

$$\text{Var}(\widehat{\beta}_0) = \text{Var}(\overline{Y}) + (-\overline{X})^2\text{Var}(\widehat{\beta}_1) + 2(-\overline{X})\text{Cov}(\overline{Y},\ \widehat{\beta}_1). \tag{1.28}$$

It has been shown that the $\text{Var}(\overline{Y}) = \sigma^2/n$ and $\text{Var}(\widehat{\beta}_1) = \sigma^2/\sum x_i^2$, but $\text{Cov}(\overline{Y}, \widehat{\beta}_1)$ remains to be determined.

Covariances of Linear Functions

The covariance between two linear functions is only slightly more complicated than the variance of a single linear function. Let U be the linear function defined earlier with a_i as coefficients and let W be a second linear function of the same random variables using d_i as coefficients:

$$U = \sum a_i Y_i \qquad \text{and} \qquad W = \sum d_i Y_i.$$

The covariance of U and W is given by

$$\text{Cov}(U, W) = \sum a_i d_i \text{Var}(Y_i) + \sum \sum_{i \neq j} a_i d_j \text{Cov}(Y_i, Y_j), \tag{1.29}$$

where the double summation is again over all $n(n-1)$ possible combinations of different values of the subscripts. If the Y_i are independent, the covariances are zero and equation 1.29 reduces to

$$\text{Cov}(U, W) = \sum a_i d_i Var(Y_i). \tag{1.30}$$

Note that products of the corresponding coefficients are being used, whereas the squares of the coefficients were used in obtaining the variance of a linear function.

Variance of $\widehat{\boldsymbol{\beta}}_0$ (cont.)

Returning to the derivation of $\text{Var}(\widehat{\beta}_0)$, where U and W are $\overline{Y}$ and $\widehat{\beta}_1$, we note that the corresponding coefficients for each Y_i are $1/n$ and $x_i/\sum x_j^2$, respectively. Thus, the covariance between $\overline{Y}$ and $\widehat{\beta}_1$ is

$$\begin{aligned} \text{Cov}(\overline{Y}, \widehat{\beta}_1) &= \sum \left(\frac{1}{n}\right)\left(\frac{x_i}{\sum x_j^2}\right) \text{Var}(Y_i) \\ &= \left(\frac{1}{n}\right)\left(\frac{\sum x_i}{\sum x_j^2}\right)\sigma^2 = 0, \end{aligned} \tag{1.31}$$

since $\sum x_i = 0$. Thus, the variance of $\widehat{\beta}_0$ reduces to

$$\begin{aligned} \text{Var}(\widehat{\beta}_0) &= \text{Var}(\overline{Y}) + (\overline{X})^2 \text{Var}(\widehat{\beta}_1) \\ &= \frac{\sigma^2}{n} + \overline{X}^2 \frac{\sigma^2}{\sum x_i^2} \\ &= \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum x_i^2}\right)\sigma^2. \end{aligned} \tag{1.32}$$

Variance of $\widehat{\boldsymbol{Y}}_i$

Recall that $\widehat{\beta}_0$ is the estimated mean of Y when $X = 0$, and thus $\text{Var}(\widehat{\beta}_0)$ can be thought of as the $\text{Var}(\widehat{Y})$ for $X = 0$. The formula for $\text{Var}(\widehat{\beta}_0)$ can be used to obtain the variance of $\widehat{Y}_i$ for any given value of X_i by replacing $\overline{X}$ with $(X_i - \overline{X})$. Since

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i = \overline{Y} + \widehat{\beta}_1 (X_i - \overline{X}), \tag{1.33}$$

we have

$$\text{Var}(\widehat{Y}_i) = \left[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum x_j^2}\right]\sigma^2. \tag{1.34}$$

The variance of the fitted value attains its minimum of σ^2/n when the regression equation is being evaluated at $X_i = \overline{X}$, and increases as the value of X at which the equation is being evaluated moves away from $\overline{X}$. Equation 1.34 gives the appropriate variance when $\widehat{Y}_i$ is being used as the **estimate** of the true mean $\beta_0 + \beta_i X_i$ of Y at the specific value X_i of X.

Variance of Predictions

Consider the problem of predicting some future observation $Y_0 = \beta_0 + \beta_1 X_0 + \epsilon_0$, at a specific value X_0 of X, where ϵ_0 is assumed to be $N(0, \sigma^2)$, independent of the current observations. Recall that $\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 X_0$ is used as an estimate of the mean $\beta_0 + \beta_1 X_0$ of Y_0. Since the best prediction for ϵ_0 is its mean zero, $\widehat{Y}_0$ is also used as the predictor of Y_0. The variance for prediction must take into account the fact that the quantity being predicted is itself a random variable. The success of the prediction will depend on how small the difference is between $\widehat{Y}_0$ and the future observation Y_0. The difference $Y_0 - \widehat{Y}_0$ is called the **prediction error**. The average squared difference between $\widehat{Y}_0$ and Y_0, $\mathcal{E}(\widehat{Y}_0 - Y_0)^2$, is called the **mean squared error of prediction**. If the model is correct and prediction is for an individual in the same population from which the data were obtained, so that $\mathcal{E}(\widehat{Y}_0 - Y_0) = 0$, the mean squared error is also the variance of prediction. Assuming this to be the case, the **variance for prediction** $\text{Var}(\widehat{Y}_{pred_0})$ is the variance of the difference between $\widehat{Y}_0$ and the future observation Y_0:

$$\begin{aligned}
\text{Var}(\widehat{Y}_{pred_0}) &= \text{Var}(\widehat{Y}_0 - Y_0) \\
&= \text{Var}(\widehat{Y}_0) + \sigma^2 \\
&= \left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum x_i^2}\right]\sigma^2.
\end{aligned} \tag{1.35}$$

Comparing equation 1.35 with equation 1.34, where X_0 is a particular X_i, we observe that the variance for prediction is the variance for estimation of the mean *plus* the variance of the quantity being predicted.

The derived variances are the *true* variances; they depend on knowledge of σ^2. $\text{Var}(\cdot)$ and σ^2 are used to designate true variances. *Estimated* variances are obtained by replacing σ^2 in the variance equations with an estimate of σ^2. The residual mean square from the analysis provides an estimate of σ^2 if the correct model has been fitted. As shown later, estimates of σ^2 that are not dependent on the correct regression model being used are available in some cases. The estimated variances obtained by substituting s^2 for σ^2 are denoted by $s^2(\cdot)$, with the quantity in parentheses designating the random variable to which the variance applies.

TABLE 1.5. *Summary of important formulae in simple regression.*

Formula	*Estimate of (or formula for)*
$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$	β_0
$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$	$\mathcal{E}(Y_i)$
$e_i = Y_i - \widehat{Y}_i$	ϵ_i
$\text{SS}(\text{Total}_{\text{uncorr}}) = \sum Y_i^2$	Total uncorrected sum of squares
$\text{SS}(\text{Total}) = \sum Y_i^2 - (\sum Y_i)^2/n$	Total corrected sum of squares
$\text{SS}(\text{Model}) = n\overline{Y}^2 + \widehat{\beta}_1^2(\sum x_i^2)$	Sum of squares due to model
$\text{SS}(\text{Regr}) = \widehat{\beta}_1^2(\sum x_i^2)$	Sum of squares due to X
$\text{SS}(\text{Res}) = \text{SS}(\text{Total}) - \text{SS}(\text{Regr})$	Residual sum of squares
$R^2 = \text{SS}(\text{Regr})/\text{SS}(\text{Total})$	Coefficient of determination
$s^2(\widehat{\beta}_1) = s^2/\sum x_i^2$	Variance of $\widehat{\beta}_1$
$s^2(\widehat{\beta}_0) = \left[\frac{1}{n} + \overline{X}^2/\sum x_i^2\right] s^2$	Variance of $\widehat{\beta}_0$
$s^2(\widehat{Y}_i) = \left[\frac{\frac{1}{n} + (X_i - \overline{X})^2}{\sum x_i^2}\right] s^2$	Variance of estimated mean at X_i
$s^2(\widehat{Y}_{pred_0}) = \left[\frac{1 + \frac{1}{n} + (X_0 - \overline{X})^2}{\sum x_i^2}\right] s^2$	Variance of prediction at X_0

Table 1.5 provides a summary to this point of the important formulae in linear regression with one independent variable.

Example 1.8

For the ozone data from Example 1.1, $s^2 = 107.81$, $n = 4$, and $\sum x_i^2 = [.0399 - (.35)^2/4] = .009275$. Thus, the estimated variances for the linear functions are:

$$\begin{aligned} s^2(\widehat{\beta}_1) &= \frac{s^2}{\sum x_i^2} = \frac{107.81}{.009275} = 11,623.281 \\ s^2(\widehat{\beta}_0) &= \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum x_i^2}\right) s^2 \\ &= \left[\frac{1}{4} + \frac{(.0875)^2}{.009275}\right] (107.81) = 115.942 \\ s^2(\widehat{Y}_1) &= \left(\frac{1}{n} + \frac{(X_1 - \overline{X})^2}{\sum x_i^2}\right) s^2 \\ &= \left[\frac{1}{4} + \frac{(.02 - .0875)^2}{.009275}\right] (107.81) = 79.91. \end{aligned}$$

Making appropriate changes in the values of X_i gives the variances of the remaining $\widehat{Y}_i$:

$$\begin{aligned} s^2(\widehat{Y}_2) &= 30.51, \\ s^2(\widehat{Y}_3) &= 32.84, \text{ and} \\ s^2(\widehat{Y}_4) &= 72.35. \end{aligned}$$

Note that $\widehat{Y}_1$ may also be used to predict the yield Y_0 of a future observation at the ozone level $X_0 = X_1 = .02$. The variance for prediction of Y_0 would be $\text{Var}(\widehat{Y}_1)$ increased by the amount σ^2. Thus, an estimated variance of prediction for Y_0 is $s^2(\widehat{Y}_1) + s^2 = 187.72$. Similarly, the estimated variance for predictions of future yields at ozone levels 0.07, 0.11, and 0.15 are 138.32, 140.65, and 180.16, respectively. ■

1.6 Tests of Significance and Confidence Intervals

Tests of Significance

The most common hypothesis of interest in simple linear regression is the hypothesis that the true value of the linear regression coefficient, the slope, is zero. This says that the dependent variable Y shows neither a linear increase nor decrease as the independent variable changes. In some cases, the nature of the problem will suggest other values for the null hypothesis. The computed regression coefficients, being random variables, will never

exactly equal the hypothesized value even when the hypothesis is true. The role of the test of significance is to protect against being misled by the random variation in the estimates. Is the difference between the observed value of the parameter $\widehat{\beta}_1$ and the hypothesized value of the parameter greater than can be reasonably attributed to random variation? If so, the null hypothesis is rejected.

To accommodate the more general case, the null hypothesis is written as $H_0 : \beta_1 = m$, where m is any constant of interest and of course can be equal to zero. The alternative hypothesis is $H_a : \beta_1 \neq m$, $H_a : \beta_1 > m$, or $H_a : \beta_1 < m$ depending on the expected behavior of β_1 if the null hypothesis is *not* true. In the first case, $H_a : \beta_1 \neq m$ is referred to as the two-tailed alternative hypothesis (interest is in detecting departures of β_1 from m in either direction) and leads to a two-tailed test of significance. The latter two alternative hypotheses, $H_a : \beta_1 > m$ and $H_a : \beta_1 < m$, are one-tailed alternatives and lead to one-tailed tests of significance.

If the random errors in the model, the ϵ_i, are normally distributed, the Y_i and any linear function of the Y_i will be normally distributed [see Searle (1971)]. Thus, $\widehat{\beta}_1$ is normally distributed with mean β_1 ($\widehat{\beta}_1$ is shown to be unbiased in Chapter 3) and variance $\text{Var}(\widehat{\beta}_1)$. If the null hypothesis that $\beta_1 = m$ is true, then $\widehat{\beta}_1 - m$ is normally distributed with mean zero. Thus,

$$t = \frac{\widehat{\beta}_1 - m}{s(\widehat{\beta}_1)} \tag{1.36}$$

is distributed as Student's t with degrees of freedom determined by the degrees of freedom in the estimate of σ^2 in the denominator. The computed t-value is compared to the appropriate critical value of Student's t, (Appendix Table A), determined by the Type I error rate α and whether the alternative hypothesis is one-tailed or two-tailed. The critical value of Student's t for the two-tailed alternative hypothesis places probability $\alpha/2$ in each tail of the distribution. The critical values for the one-tailed alternative hypotheses place probability α in only the upper or lower tail of the distribution, depending on whether the alternative is $\beta_1 > m$ or $\beta_1 < m$, respectively.

Example 1.9

The estimate of β_1 for Heagle's ozone data from Example 1.1 was $\widehat{\beta}_1 = -293.53$ with a standard error of $s(\widehat{\beta}_1) = \sqrt{11,623.281} = 107.81$. Thus, the computed t-value for the test of $H_0 : \beta_1 = 0$ is

$$t = \frac{-293.53}{107.81} = -2.72.$$

The estimate of σ^2 in this example has only two degrees of freedom. Using the two-tailed alternative hypothesis and $\alpha = .05$ gives a critical t-value of

$t_{(.025,2)} = 4.303$. Since $|t| < 4.303$, the conclusion is that the data do not provide convincing evidence that β_1 is different from zero.

In this example one might expect the increasing levels of ozone to *depress* the yield of soybeans; that is, the slope would be negative if not zero. The appropriate one-tailed alternative hypothesis would be $H_a : \beta_1 < 0$. For this one-tailed test, the critical value of t for $\alpha = .05$ is $t_{(.05,2)} = 2.920$. Although the magnitude of the computed t is close to this critical value, strict adherence to the $\alpha = .05$ size of test leads to the conclusion that there is insufficient evidence in these data to infer a real (linear) effect of ozone on soybean yield. (From a practical point of view, one would begin to suspect a real effect of ozone and seek more conclusive data.) ■

In a similar manner, t-tests of hypotheses about β_0 and any of the $\widehat{Y}_i$ can be constructed. In each case, the numerator of the t-statistic is the difference between the estimated value of the parameter and the hypothesized value, and the denominator is the standard deviation (or standard error) of the estimate. The degrees of freedom for Student's t is always the degrees of freedom associated with the estimate of σ^2.

The F-statistic can be used as an alternative to Student's t for two-tailed hypotheses about the regression coefficients. It was indicated earlier that MS(Regr) is an estimate of $\sigma^2 + \beta_1^2 \sum x_i^2$ and that MS(Res) is an estimate of σ^2. If the null hypothesis that $\beta_1 = 0$ is true, both MS(Regr) and MS(Res) are estimating σ^2. As β_1 deviates from zero, MS(Regr) will become increasingly larger (on the average) than MS(Res). Therefore, a ratio of MS(Regr) to MS(Res) appreciably larger than unity would suggest that β_1 is not zero. This ratio of MS(Regr) to MS(Res) follows the F-distribution when the assumption that the residuals are normally distributed is valid and the null hypothesis is true.

Example 1.10

For the ozone data of Example 1.1, the ratio of variances is

$$F = \frac{\text{MS(Regr)}}{\text{MS(Res)}} = \frac{799.14}{107.81} = 7.41.$$

This can be compared to the critical value of the F-distribution with 1 degree of freedom in the numerator and 2 degrees of freedom in the denominator, $F_{(.05,1,2)} = 18.51$ for $\alpha = .05$ (Appendix Table A.3), to determine whether MS(Regr) is sufficiently larger than MS(Res) to rule out chance as the explanation. Since $F = 7.41 < 18.51$, the conclusion is that the data do not provide conclusive evidence of a linear effect of ozone. The F-ratio with 1 degree of freedom in the numerator is the square of the corresponding t-statistic. Therefore, the F and the t are equivalent tests for this two-tailed alternative hypothesis. ■

Confidence Intervals

Confidence interval estimates of parameters are more informative than point estimates because they reflect the precision of the estimates. The 95% confidence interval estimate of β_1 and β_0 are, respectively,

$$\widehat{\beta}_1 \pm t_{(.025,\nu)} s(\widehat{\beta}_1) \tag{1.37}$$

and

$$\widehat{\beta}_0 \pm t_{(.025,\nu)} s(\widehat{\beta}_0), \tag{1.38}$$

where ν is the degrees of freedom associated with s^2.

Example 1.11

The 95% confidence interval estimate of β_1 for Example 1.1 is

$$-293.53 \pm (4.303)(107.81)$$

or $(-757,\ 170)$.

The confidence interval estimate indicates that the true value may fall anywhere between -757 and 170. This very wide range conveys a high degree of uncertainty (lack of confidence) in the point estimate $\widehat{\beta}_1 = -293.53$. Notice that the interval includes zero. This is consistent with the conclusions from the t-test and the F-test that $H_0 : \beta_1 = 0$ cannot be rejected.

The 95% confidence interval estimate of β_0 is

$$253.43 \pm (4.303)(10.77)$$

or (207.1, 299.8). The value of β_0 might reasonably be expected to fall anywhere between 207 and 300 based on the information provided by this study. ■

In a similar manner, interval estimates of the true mean of Y for various values of X are computed using $\widehat{Y}_i$ and their standard errors. Frequently, these confidence interval estimates of $\mathcal{E}(Y_i)$ are plotted with the regression line and the observed data. Such graphs convey an overall picture of how well the regression represents the data and the degree of confidence one might place in the results. Figure 1.2 shows the results for the ozone example. The confidence coefficient of .95 applies individually to the confidence intervals on each estimated mean. Simultaneous confidence intervals are discussed in Section 4.6.

The failure of the tests of significance to detect an effect of ozone on the yield of soybeans is, in this case, a reflection of the lack of power in this small data set. This lack of power is due primarily to the limited degrees of freedom available for estimating σ^2. In defense of the research project from which these data were borrowed, we must point out that only a portion of the data (the set of treatment means) is being used for this illustration. The complete data set from this experiment provides for an adequate estimate of error and shows that the effects of ozone are highly significant. The complete data are used at a later time.

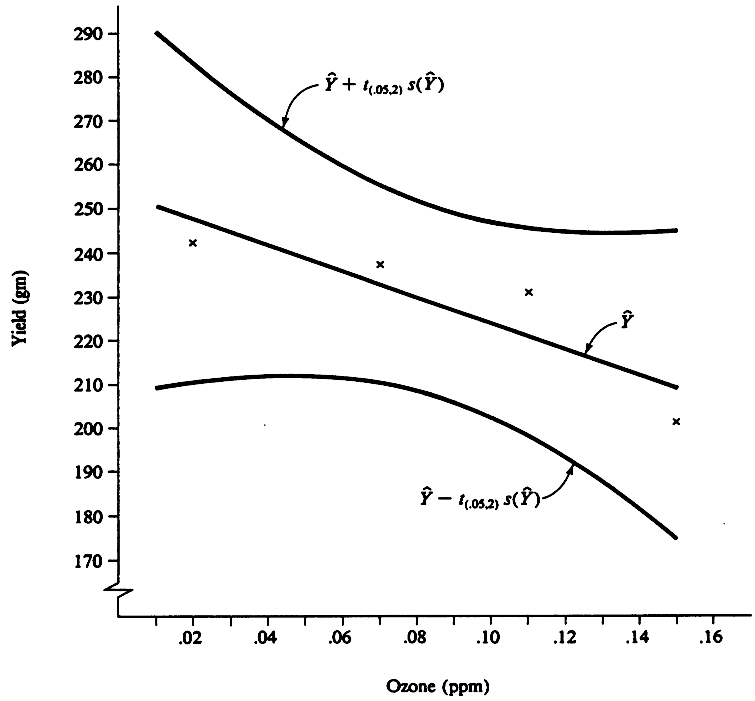

FIGURE 1.2. *The regression of soybean mean yield (grams per plant) on ozone (ppm) showing the individual confidence interval estimates of the mean response.*

1.7 Regression Through the Origin

In some situations the regression line is expected to pass through the origin. That is, the true mean of the dependent variable is expected to be zero when the value of the independent variable is zero. Many growth models, for example, would pass through the origin. The amount of chemical produced in a system requiring a catalyst would be zero when there is no catalyst present. The linear regression model is forced to pass through the origin by setting β_0 equal to zero. The linear model then becomes

$$Y_i = \beta_1 X_i + \epsilon_i. \tag{1.39}$$

There is now only one parameter to be estimated and application of the least squares principle gives

$$\beta_1(\sum X_i^2) = \sum X_i Y_i \tag{1.40}$$

as the only normal equation to be solved. The solution is

$$\widehat{\beta}_1 = \frac{\sum X_i Y_i}{\sum X_i^2}. \tag{1.41}$$

Both the numerator and denominator are now *uncorrected* sums of products and squares. The regression equation becomes

$$\widehat{Y}_i = \widehat{\beta}_1 X_i, \tag{1.42}$$

and the residuals are defined as before,

$$e_i = Y_i - \widehat{Y}_i. \tag{1.43}$$

Unlike the model with an intercept, in the no-intercept model the sum of the residuals is not necessarily zero.

The uncorrected sum of squares of Y can still be partitioned into the two parts

$$\text{SS(Model)} = \sum \widehat{Y}_i^2 \tag{1.44}$$

and

$$\text{SS(Res)} = \sum (Y_i - \widehat{Y}_i)^2 = \sum e_i^2. \tag{1.45}$$

Since only one parameter is involved in determining $\widehat{Y}_i$, SS(Model) has only 1 degree of freedom and cannot be further partitioned into the correction for the mean and SS(Regr). For the same reason, the residual sum of squares has $(n-1)$ degrees of freedom. The residual mean square is an estimate of σ^2

if the model is correct. The expectation of the MS(Model) is $\mathcal{E}$[MS(Model)] $= \sigma^2 + \beta_1^2(\sum X_i^2)$. This is the same form as $\mathcal{E}$[MS(Regr)] for a model with an intercept except here the sum of squares for X is the uncorrected sum of squares.

The variance of $\widehat{\beta}_1$ is determined using the rules for the variance of a linear function (see equations 1.25 and 1.26). The coefficients on the Y_i for the no-intercept model are $X_i / \sum X_j^2$. With the same assumptions of independence of the Y_i and common variance σ^2, the variance of $\widehat{\beta}_1$ is

$$\begin{aligned} \text{Var}(\widehat{\beta}_1) &= \left[\left(\frac{X_1}{\sum X_j^2} \right)^2 + \left(\frac{X_2}{\sum X_j^2} \right)^2 + \cdots + \left(\frac{X_n}{\sum X_j^2} \right)^2 \right] \sigma^2 \\ &= \frac{\sigma^2}{\sum X_j^2}. \end{aligned} \tag{1.46}$$

The divisor on σ^2, the *uncorrected* sum of squares for the independent variable, will always be larger (usually much larger) than the corrected sum of squares. Therefore, the estimate of $\widehat{\beta}_1$ in equation 1.41 will be much more precise than the estimate in equation 1.9 when a no-intercept model is appropriate. This results because one parameter, β_0, is assumed to be *known*.

The variance of $\widehat{Y}_i$ is most easily obtained by viewing it as a linear function of $\widehat{\beta}_1$:

$$\widehat{Y}_i = X_i \widehat{\beta}_1. \tag{1.47}$$

Thus, the variance is

$$\begin{aligned} \text{Var}(\widehat{Y}_i) &= X_i^2 \text{Var}(\widehat{\beta}_1) \\ &= \left(\frac{X_i^2}{\sum X_j^2} \right) \sigma^2. \end{aligned} \tag{1.48}$$

Estimates of the variances are obtained by substitution of s^2 for σ^2.

Example 1.12

Regression through the origin is illustrated using data on increased risk incurred by individuals exposed to a toxic agent. Such health risks are often expressed as relative risk, the ratio of the rate of incidence of the health problem for those exposed to the rate of incidence for those not exposed to the toxic agent. A relative risk of 1.0 implies no increased risk of the disease from exposure to the agent. Table 1.6 gives the relative risk to individuals exposed to differing levels of dust in their work environments. Dust exposure is measured as the average number of particles/ft^3/year scaled by dividing by 10^6. By definition, the expected relative risk is 1.0 when exposure is zero. Thus, the regression line relating relative risk to

TABLE 1.6. *Relative risk of exposure to dust for nine groups of individuals. Dust exposure is reported in particles/ft^3/year and scaled by dividing by* 10^6.

X = *Dust Exposure*	*Relative Risk*	Y = *Relative Risk* -1
75	1.10	.10
100	1.05	.05
150	.97	$-.03$
350	1.90	.90
600	1.83	.83
900	2.45	1.45
1,300	3.70	2.70
1,650	3.52	2.52
2,250	4.16	3.16
$\sum X_i = 7,375$		$\sum Y_i = 11.68$
$\sum X_i^2 = 10,805,625$		$\sum Y_i^2 = 27.2408$
	$\sum X_i Y_i = 16,904$	

exposure should have an intercept of 1.0 or, equivalently, the regression line relating $Y =$ (relative risk $- 1$) to exposure should pass through the origin. The variable Y and key summary statistics on X and Y are included in Table 1.6.

Assuming a linear relationship and zero intercept, the point estimate of the slope β_1 of the regression line is

$$\widehat{\beta}_1 = \frac{\sum X_i Y_i}{\sum X_i^2} = \frac{16,904}{10,805,625} = .00156.$$

The estimated increase in relative risk is .00156 for each increase in dust exposure of 1 million particles per cubic foot per year. The regression equation is

$$\widehat{Y}_i = .00156 X_i.$$

When $X_i = 0$, the value of $\widehat{Y}_i$ is zero and the regression equation has been forced to pass through the origin.

The regression partitions each observation Y_i into two parts; that accounted for by the regression through the origin $\widehat{Y}_i$, and the residual or deviation from the regression line e_i (Table 1.7). The sum of squares attributable to the model,

$$\text{SS(Model)} = \sum \widehat{Y}_i^2 = 26.4441,$$

and the sum of squares of the residuals,

$$\text{SS(Res)} = \sum e_i^2 = .7967,$$

TABLE 1.7. *Y_i, $\widehat{Y}_i$, and e_i from linear regression through the origin of increase in relative risk (Y = relative risk − 1) on exposure level.*

Y_i	$\widehat{Y}_i$	e_i
.10	.1173	−.0173
.05	.1564	−.1064
−.03	.2347	−.2647
.90	.5475	.3525
.83	.9386	−.1086
1.45	1.4079	.0421
2.70	2.0337	.6663
2.52	2.5812	−.0612
3.16	3.5198	−.3598
$\sum Y_i^2 = 27.2408$	$\sum \widehat{Y}_i^2 = 26.4441$	$\sum e_i^2 = .7967$

TABLE 1.8. *Summary analysis of variance for regression through the origin of increase in relative risk on level of exposure to dust particles.*

Source	*d.f.*	*SS*	*MS*	$\mathcal{E}(MS)$
$\text{Total}_{\text{uncorr}}$	n=9	27.2408		
Due to model	p=1	26.4441	26.4441	$\sigma^2 + \beta_1^2(\sum X_i^2)$
Residual	$n-p$=8	.7967	.0996	σ^2

partition the total uncorrected sum of squares,

$$\sum Y_i^2 = 27.2408.$$

In practice, the sum of squares due to the model is more easily computed as

$$\begin{aligned} \text{SS(Model)} &= \widehat{\beta}_1^2 \left(\sum X_i^2 \right) \\ &= (.00156437)^2 (10,805,625) = 26.4441. \end{aligned}$$

The residual sum of squares is computed by difference. The summary analysis of variance, including the mean square expectations, is given in Table 1.8.

When the no-intercept model is appropriate, MS(Res) is an estimate of σ^2. MS(Model) is an estimate of σ^2 plus a quantity that is positive if β_1 is not zero. The ratio of the two mean squares provides a test of significance for $H_0 : \beta_1 = 0$. This is an F-test with one and eight degrees of freedom, if the assumption of normality is valid, and is significant beyond $\alpha = .001$.

There is clear evidence that the linear regression relating increased risk to dust exposure is not zero.

The estimated variance of β_1 is

$$s^2(\widehat{\beta}_1) = \frac{s^2}{\sum X_i^2} = \frac{.09958533}{10,805,625} = 92.161 \times 10^{-10}$$

or

$$s(\widehat{\beta}_1) = 9.6 \times 10^{-5} = .000096.$$

Since each $\widehat{Y}_i$ is obtained by multiplying $\widehat{\beta}_1$ by the appropriate X_i, the estimated variance of a $\widehat{Y}_i$ is

$$\begin{aligned} s^2(\widehat{Y}_i) &= X_i^2[s^2(\widehat{\beta}_1)] \\ &= (92.161 \times 10^{-10})X_i^2 \end{aligned}$$

if $\widehat{Y}_i$ is being used as an estimate of the true mean of Y for that value of X. If $\widehat{Y}_i$ is to be used for prediction of a future observation with dust exposure X_i, then the variance for prediction is

$$\begin{aligned} s^2(\widehat{Y}_{pred_i)} &= s^2 + s^2(\widehat{Y}_i) \\ &= .09958 + (92.161 \times 10^{-10})X_i^2. \end{aligned}$$

The variances and the standard errors provide measures of precision of the estimate and are used to construct tests of hypotheses and confidence interval estimates.

The data and a plot of the fitted regression line are shown in Figure 1.3. The 95% confidence interval estimates of the mean response $\mathcal{E}(\widehat{Y}_i)$ are shown as bands on the regression line in the figure. Notice that with regression through the origin the confidence bands go to zero as the origin is approached. This is consistent with the model assumption that the mean of Y is known to be zero when $X = 0$. Although the fit appears to be reasonable, there are suggestions that the model might be improved. The three lowest exposures fall below the regression line and very near zero; these levels of exposure may not be having as much impact as linear regression through the origin would predict. In addition, the largest residual, $e_7 = .6663$, is particularly noticeable. It is nearly twice as large as the next largest residual and is the source of over half of the residual sum of squares (see Table 1.7). This large positive residual and the overall pattern of residuals suggests that a curvilinear relationship without the origin being forced to be zero would provide a better fit to the data. In practice, such alternative models would be tested before this linear no-intercept model would be adopted. We forgo testing the need for a curvilinear relationship at this time (fitting curvilinear models is discussed in Chapters 3 and 8)

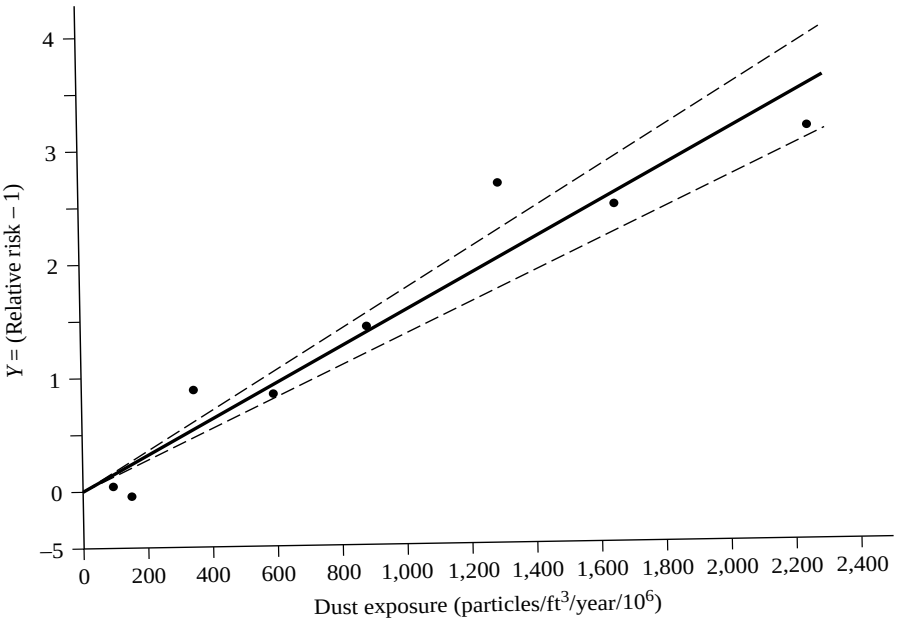

FIGURE 1.3. *Regression of increase in relative risk on exposure to dust particles with the regression forced through the origin. The bands on the regression line connect the limits of the 95% confidence interval estimates of the means.*

and continue with this example to illustrate testing the appropriateness of the no-intercept model assuming the linear relationship is appropriate.

The test of the assumption that β_0 is zero is made by temporarily adopting a model that allows a nonzero intercept. The estimate obtained for the intercept is then used to test the null hypothesis that β_0 is zero. Including an intercept in this example gives $\widehat{\beta}_0 = .0360$ with $s(\widehat{\beta}_0) = .1688$. (The residual mean square from the intercept model is $s^2 = .1131$ with seven degrees of freedom.) The t-test for the null hypothesis that $\widehat{\beta}_0$ is zero is

$$t = \frac{.0360}{.1688} = .213$$

and is not significant; $t_{(.025,7)} = 2.365$. There is no indication in these data that the no-intercept model is inappropriate. (Recall that this test has been made assuming the linear relationship is appropriate. If the model were expanded to allow a curvilinear response, the test of the null hypothesis that $\beta_0 = 0$ might become significant.) An equivalent test of the null hypothesis that $\beta_0 = 0$ can be made using the difference between the residual sums of squares from the intercept and no-intercept models. This test is discussed in Chapter 4. ■

1.8 Models with Several Independent Variables

Most models will use more than one independent variable to explain the behavior of the dependent variable. The linear additive model can be extended to include any number of independent variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_p X_{ip} + \epsilon_i. \tag{1.49}$$

The subscript notation has been extended to include a number on each X and β to identify each independent variable and its regression coefficient. There are p independent variables and, including β_0, $p' = p+1$ parameters to be estimated.

The usual least squares assumptions apply. The ϵ_i are assumed to be independent and to have common variance σ^2. For constructing tests of significance or confidence interval statements, the random errors are also assumed to be normally distributed. The independent variables are assumed to be measured without error.

The least squares method of estimation applied to this model requires that estimates of the $p+1$ parameters be found such that

$$\begin{aligned} \text{SS(Res)} &= \sum (Y_i - \widehat{Y}_i)^2 \\ &= \sum (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{i1} - \widehat{\beta}_2 X_{i2} - \cdots - \widehat{\beta}_p X_{ip})^2 \end{aligned} \tag{1.50}$$

is minimized. The $\widehat{\beta}_j,\ j = 0, 1, \ldots, p$, are the estimates of the parameters. The values of $\widehat{\beta}_j$ that minimize SS(Res) are obtained by setting the derivative of SS(Res) with respect to each $\widehat{\beta}_j$ in turn equal to zero. This gives $(p + 1)$ normal equations that must be solved simultaneously to obtain the least squares estimates of the $(p + 1)$ parameters.

It is apparent that the problem is becoming increasingly difficult as the number of independent variables increases. The algebraic notation becomes particularly cumbersome. For these reasons, matrix notation and matrix algebra are used to develop the regression results for the more complicated models. The next chapter is devoted to a brief review of the key matrix operations needed for the remainder of the text.

1.9 Violation of Assumptions

Basic Assumptions

In Section 1.1, we assumed that

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

where the random errors ϵ_i are normally distributed independent random variables with mean zero and constant variance σ^2, and the X_i are n observations on the independent variable that is measured without error. Under these assumptions, the least squares estimators of β_0 and β_1 are the best (minimum variance) among all possible unbiased estimators. Statistical inference procedures, such as hypothesis testing and confidence and prediction intervals, considered in the previous section are valid under these assumptions. Here we briefly indicate the effects of violation of assumptions on estimation and statistical inference. A more detailed discussion of problem areas in least squares and possible remedies is presented in Chapters 10 through 14.

Major problem areas in least squares analysis relate to failure of the four basic assumptions — normality, independence and constant variance of the errors, and the independent variable being measured without error. When only the assumption of normality is violated, the least squares estimators continue to have the smallest variance among all *linear* (in $\boldsymbol{Y}$) unbiased estimators. The assumption of normality is not needed for the partitioning of total variation or for estimating the variance. However, it is needed for tests of significance and construction of confidence and prediction intervals. Although normality is a reasonable assumption in many situations, it is not appropriate for count data and for some time-to-failure data that tend to have asymmetric distributions. Transformations of the dependent variable and alternative estimation procedures are used in such situations. Also, in many situations with large n, statistical inference procedures based on t- and F-statistics are approximately valid, even though the normality assumption is not valid.

Normality

Correlated Errors

When data are collected in a time sequence, the errors associated with an observation at one point in time will tend to be correlated with the errors of the immediately adjacent observations. Economic and meteorological variables measured over time and repeated measurements over time on the same experimental unit, such as in plant and animal growth studies, will usually have correlated errors. When the errors are correlated, the least squares estimators continue to be unbiased, but are no longer the best estimators. Also, in this case, the variance estimators obtained using equations 1.26 and 1.32 are seriously biased. Alternative estimation methods for correlated errors are discussed in Chapter 12.

Nonconstant Variance

In some situations, the variability in the errors increases with the independent variable or with the mean of the response variable. For example, in some yield data, the mean and the variance of the yield both increase with the amount of seeds (or fertilizer) used. Consider the model

$$\begin{aligned} Y_i &= (\beta_0 + \beta_1 X_i)u_i \\ &= \beta_0 + \beta_1 X_i + (\beta_0 + \beta_1 X_i)(u_i - 1) \\ &= \beta_0 + \beta_1 X_i + \epsilon_i, \end{aligned}$$

where the errors u_i are multiplicative and have mean one and constant variance. Then the variance of ϵ_i is proportional to $(\beta_0 + \beta_1 X_i)^2$. The effect of nonconstant (heterogeneous) variances on least squares estimators is similar to that of correlated errors. The least squares estimators are no longer efficient and the variance formulae in equations 1.26 and 1.32 are not valid. Alternative methods are discussed in Chapter 11.

Measurement Error

When the independent variable is measured with error or when the model is misspecified by omitting important independent variables, least squares estimators will be biased. In such cases, the variance estimators are also biased. Methods for detecting model misspecification and estimation in measurement error models are discussed in later chapters. Also, the effect of overly influential data points and outliers is discussed later.

1.10 Summary

This chapter has reviewed the basic elements of least squares estimation for the simple linear model containing one independent variable. The more complicated linear model with several independent variables was introduced and is pursued using matrix notation in subsequent chapters. The student should understand these concepts:

- the form and basic assumptions of the linear model;
- the least squares criterion, the estimators of the parameters obtained using this criterion, and measures of precision of the estimates;

- the use of the regression equation to obtain estimates of mean values and predictions, and appropriate measures of precision for each; and
- the partitioning of the total variability of the response variable into that explained by the regression equation and the residual or unexplained part.

1.11 Exercises

1.1. Use the least squares criterion to derive the normal equations, equation 1.6, for the simple linear model of equation 1.2.

1.2. Solve the normal equations, equation 1.6, to obtain the estimates of β_0 and β_1 given in equation 1.7.

1.3. Use the statistical model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

to show that $\epsilon_i \sim \text{NID}(0, \sigma^2)$ implies each of the following:

(a) $\mathcal{E}(Y_i) = \beta_0 + \beta_1 X_i$,

(b) $\sigma^2(Y_i) = \sigma^2$, and

(c) $\text{Cov}(Y_i, Y_{i'}) = 0$, $i \neq i'$.

For Parts (b) and (c), use the following definitions of variance and covariance.

$$\begin{aligned} \sigma^2(Y_i) &= \mathcal{E}\{[Y_i - \mathcal{E}(Y_i)]^2\} \\ \text{Cov}(Y_i,\ Y_{i'}) &= \mathcal{E}\{[Y_i - \mathcal{E}(Y_i)][Y_{i'} - \mathcal{E}(Y_{i'})]\}. \end{aligned}$$

1.4. The data in the accompanying table relate heart rate at rest Y to kilograms body weight X.

X		Y	
90		62	
86		45	
67		40	
89		55	
81		64	
75		53	
$\sum X_i =$	488	$\sum Y_i =$	319
$\sum X_i^2 =$	40,092	$\sum Y_i^2 =$	17,399
	$\sum X_i Y_i = 26,184$		

(a) Graph these data. Does it appear that there is a linear relationship between body weight and heart rate at rest?

(b) Compute $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and write the regression equation for these data. Plot the regression line on the graph from Part (a). Interpret the estimated regression coefficients.

(c) Now examine the data point (67, 40). If this data point were removed from the data set, what changes would occur in the estimates of β_0 and β_1?

(d) Obtain the point estimate of the mean of Y when $X = 88$. Obtain a 95% confidence interval estimate of the mean of Y when $X = 88$. Interpret this interval statement.

(e) Predict the heart rate for a particular subject weighing 88kg using both a point prediction and a 95% confidence interval. Compare these predictions to the estimates computed in Part (d).

(f) Without doing the computations, for which measured X would the corresponding $\widehat{Y}$ have the smallest variance? Why?

1.5. Use the data and regression equation from Exercise 1.4 and compute $\widehat{Y}_i$ for each value of X. Compute the product moment correlations between

(a) X_i and Y_i,

(b) Y_i and $\widehat{Y}_i$, and

(c) X_i and $\widehat{Y}_i$.

Compare these correlations to each other and to the coefficient of determination R^2. Can you prove algebraically the relationships you detect?

1.6. Show that

$$\text{SS(Model)} = n\overline{Y}^2 + \widehat{\beta}_1^2 \sum (X_i - \overline{X})^2 \qquad \text{(equation 1.16)}.$$

1.7. Show that

$$\sum (Y_i - \widehat{Y}_i)^2 = \sum y_i^2 - \widehat{\beta}_1^2 \sum (X_i - \overline{X})^2.$$

Note that $\sum y_i^2$ is being used to denote the *corrected* sum of squares.

1.8. Show algebraically that $\sum e_i = 0$ when the simple linear regression equation includes the constant term β_0. Show algebraically that this is not true when the simple linear regression does not include the intercept.

1.9. The following data relate biomass production of soybeans to cumulative intercepted solar radiation over an eight-week period following emergence. Biomass production is the mean dry weight in grams of independent samples of four plants. (Data courtesy of Virginia Lesser and Dr. Mike Unsworth, North Carolina State University.)

X *Solar Radiation*	Y *Plant Biomass*
29.7	16.6
68.4	49.1
120.7	121.7
217.2	219.6
313.5	375.5
419.1	570.8
535.9	648.2
641.5	755.6

(a) Compute $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for the linear regression of plant biomass on intercepted solar radiation. Write the regression equation.

(b) Place 95% confidence intervals on β_1 and β_0. Interpret the intervals.

(c) Test $H_0 : \beta_1 = 1.0$ versus $H_a : \beta_1 \neq 1.0$ using a t-test with $\alpha = .1$. Is your result for the t-test consistent with the confidence interval from Part (b)? Explain.

(d) Use a t-test to test $H_0 : \beta_0 = 0$ against $H_a : \beta_0 \neq 0$. Interpret the results. Now fit a regression with $\beta_0 = 0$. Give the analysis of variance for the regression through the origin and use an F-test to test $H_0 : \beta_0 \neq 0$. Compare the results of the t-test and the F-test. Do you adopt the model with or without the intercept?

(e) Compute $s^2(\widehat{\beta}_1)$ for the regression equation without an intercept. Compare the variances of the estimates of the slopes $\widehat{\beta}_1$ for the two models. Which model provides the greater precision for the estimate of the slope?

(f) Compute the 95% confidence interval estimates of the mean biomass production for $X = 30$ and $X = 600$ for both the intercept and the no-intercept models. Explain the differences in the intervals obtained for the two models.

1.10. A linear regression was run on a set of data using an intercept and one independent variable. You are given only the following information:

(1) $\widehat{Y}_i = 11.5 - 1.5X_i$.

(2) The t-test for $H_0 : \beta_1 = 0$ was nonsignificant at the $\alpha = .05$ level. A computed t of -4.087 was compared to $t_{(.05,2)}$ from Appendix Table A.1.

(3) The estimate of σ^2 was $s^2 = 1.75$.

(a) Complete the analysis of variance table using the given results.

(b) Compute and interpret the coefficient of determination R^2.

1.11. An experiment has yielded sample means for four treatment regimes, $\overline{Y}_1$, $\overline{Y}_2$, $\overline{Y}_3$, and $\overline{Y}_4$. The numbers of observations in the four means are $n_1 = 4$, $n_2 = 6$, $n_3 = 3$, and $n_4 = 9$. The pooled estimate of σ^2 is $s^2 = 23.5$.

(a) Compute the variance of each treatment mean.

(b) Compute the variance of the mean contrast $C = \overline{Y}_3 + \overline{Y}_4 - 2\overline{Y}_1$.

(c) Compute the variance of $(\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3)/3$.

(d) Compute the variance of $(4\overline{Y}_1 + 6\overline{Y}_2 + 3\overline{Y}_3)/13$.

1.12. Obtain the normal equations and the least squares estimates for the model

$$Y_i = \mu + \beta_1 x_i + \epsilon_i,$$

where $x_i = (X_i - \overline{X})$. Compare the results to equation 1.6. (The model expressed in this form is referred to as the "centered" model; the independent variable has been shifted to have mean zero.)

1.13. Recompute the regression equation and analysis of variance for the Heagle ozone data (Table 1.1) using the centered model,

$$Y_i = \mu + \beta_1 x_i + \epsilon_i,$$

where $x_i = (X_i - \overline{X})$. Compare the results with those in Tables 1.2 to 1.4.

1.14. Derive the normal equation for the no-intercept model, equation 1.40, and the least squares estimate of the slope, equation 1.41.

1.15. Derive the variance of $\widehat{\beta}_1$ and $\widehat{Y}_i$ for the no-intercept model.

1.16. Show that

$$\sum (X_i - \overline{X})(Y_i - \overline{Y}) = \sum (X_i - \overline{X})Y_i = \sum X_i(Y_i - \overline{Y}).$$

1.17. The variance of $\widehat{Y}_{pred_0}$ as given by equation 1.35 is for the prediction of a single future observation. Derive the variance of a prediction of the *mean* of q future observations all having the same value of X.

1.18. An experimenter wants to design an experiment for estimating the rate of change in a dependent variable Y as an independent variable X is changed. He is convinced from previous experience that the relationship is linear in the region of interest, between $X = 0$ and $X = 11$. He has enough resources to obtain 12 observations. Use $\sigma^2(\widehat{\beta}_1)$, equation 1.26, to show the researcher the best allocation of the design points (choices of X-values). Compare $\sigma^2(\widehat{\beta}_1)$ for this optimum allocation with an allocation of one observation at each interger value of X from $X = 0$ and $X = 11$.

1.19. The data in the table relate seed weight of soybeans, collected for six successive weeks following the start of the reproductive stage, to cumulative seasonal solar radiation for two levels of chronic ozone exposure. Seed weight is mean seed weight (grams per plant) from independent samples of four plants. (Data courtesy of Virginia Lesser and Dr. Mike Unsworth.)

Low Ozone		*High Ozone*	
Radiation	*Seed Weight*	*Radiation*	*Seed Weight*
118.4	.7	109.1	1.3
215.2	2.9	199.6	4.8
283.9	5.6	264.2	6.5
387.9	8.7	358.2	9.4
451.5	12.4	413.2	12.9
515.6	17.4	452.5	12.3

(a) Determine the linear regression of seed weight on radiation separately for each level of ozone. Determine the similarity of the two regressions by comparing the confidence interval estimates of the two intercepts and the two slopes and by visual inspection of plots of the data and the regressions.

(b) Regardless of your conclusion in Part (a), assume that the two regressions are the same and estimate the common regression equation.

1.20. A hotel experienced an outbreak of Pseudomona dermatis among its guests. Physicians suspected the source of infection to be the hotel whirlpool-spa. The data in the table give the number of female guests and the number infected by categories of time (minutes) spent in the whirlpool.

Time (Minutes)	*Number of Guests*	*Number Infected*
0–10	8	1
11–20	12	3
21–30	9	3
31–40	14	7
41–50	7	4
51–60	4	3
61–70	2	2

(a) Can the incidence of infection (number infected/number exposed) be characterized by a linear regression on time spent in the whirlpool? Use the midpoint of the time interval as the independent variable. Estimate the intercept and the slope, and plot the regression line and the data.

(b) Review each of the basic assumptions of least squares regression and comment on whether each is satisfied by these data.

1.21. Hospital records were examined to assess the link between smoking and duration of illness. The data reported in the table are the number of hospital days (per 1,000 person-years) for several classes of individuals, the average number of cigarettes smoked per day, and the number of hospital days for control groups of nonsmokers for each class. (The control groups consist of individuals matched as nearly as possible to the smokers for several primary health factors other than smoking.)

# Hospital Days (Smokers)	*#Cigarettes Smoked/Day*	*#Hospital Days (Nonsmokers)*
215	10	201
185	5	180
334	15	297
761	45	235
684	25	520
368	30	210
1275	50	195
3190	45	835
3520	60	435
428	20	312
575	5	590
2280	45	1131
2795	60	225

(a) Plot the logarithm of number of hospital days (for the smokers) against number of cigarettes. Do you think a linear regression will adequately represent the relationship?

(b) Plot the logarithm of number of hospital days for smokers minus the logarithm of number of hospital days for the control group against number of cigarettes. Do you think a linear regression will adequately represent the relationship? Has subtraction of the control group means reduced the dispersion?

(c) Define $Y = \ln(\#\text{ days for smokers}) - \ln(\#\text{ days for nonsmokers})$ and $X = (\#\text{cigarettes})^2$. Fit the linear regression of Y on X. Make a test of significance to determine if the intercept can be set to zero. Depending on your results, give the regression equation, the standard errors of the estimates, and the summary analysis of variance.

1.22. Use the normal equations in 1.6 to show that

(a) $\sum X_i Y_i = \sum X_i \widehat{Y}_i$.

(b) $\sum X_i e_i = 0$.

(c) $\sum \widehat{Y}_i e_i = 0$. (Hint: use Exercise 1.8).

1.23 Consider the regression through the origin model in equation 1.39. Suppose $X_i \geq 0$. Define $\tilde{\beta}_1 = \sum Y_i / \sum X_i$ and $\widehat{\beta}_1 = \sum X_i Y_i / \sum X_i^2$.

(a) Show that $\tilde{\beta}_1$ and $\widehat{\beta}_1$ are unbiased for β_1.

(b) Compare the variances of $\tilde{\beta}_1$ and $\widehat{\beta}_1$.

2
INTRODUCTION TO MATRICES

Chapter 1 reviewed simple linear regression in algebraic notation and showed that the notation for models involving several variables is very cumbersome.

This chapter introduces matrix notation and all matrix operations that are used in this text. Matrix algebra greatly simplifies the presentation of regression and is used throughout the text. Sections 2.7 and 2.8 are not used until later in the text and can be omitted for now.

Matrix algebra is extremely helpful in multiple regression for simplifying notation and algebraic manipulations. You must be familiar with the basic operations of matrices in order to understand the regression results presented. A brief introduction to the key matrix operations is given in this chapter. You are referred to matrix algebra texts, for example, Searle (1982), Searle and Hausman (1970), or Stewart (1973), for more complete presentations of matrix algebra.

2.1 Basic Definitions

Matrix

A **matrix** is a rectangular array of numbers arranged in orderly rows and columns. Matrices are denoted with boldface capital letters. The following

are examples.

$$\boldsymbol{Z} = \begin{bmatrix} 1 & 2 \\ 6 & 4 \\ 5 & 7 \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & 5 \\ 1 & 6 \\ 1 & 4 \\ 1 & 9 \\ 1 & 2 \\ 1 & 6 \end{bmatrix}$$

$$\boldsymbol{B} = \begin{bmatrix} 15 & 7 & -1 & 0 \\ 15 & 5 & -2 & 10 \end{bmatrix}.$$

Elements

The numbers that form a matrix are called the **elements** of the matrix. A general matrix could be denoted as

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

The subscripts on the elements denote the row and column, respectively, in which the element appears. For example, a_{23} is the element found in the second row and third column. The row number is always given first.

Order

The **order** of a matrix is its size given by the number of rows and columns. The first matrix given, $\boldsymbol{Z}$, is of order (3, 2). That is, $\boldsymbol{Z}$ is a 3×2 matrix, since it has three rows and two columns. Matrix $\boldsymbol{A}$ is an $m \times n$ matrix.

Rank

The **rank** of a matrix is defined as the number of linearly independent columns (or rows) in the matrix. Any subset of columns of a matrix are **linearly independent** if no column in the subset can be expressed as a linear combination of the others in the subset. The matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 0 & 6 \\ 5 & 3 & 13 \end{bmatrix}$$

contains a linear dependency among its columns. The first column multiplied by two and added to the second column produces the third column. In fact, any one of the three columns of $\boldsymbol{A}$ can be written as a linear combination of the other two columns. On the other hand, any *two* columns of $\boldsymbol{A}$ are linearly independent since one cannot be produced as a multiple of the other. Thus, the rank of the matrix $\boldsymbol{A}$, denoted by $r(\boldsymbol{A})$, is two.

Full-Rank Matrices

If there are no linear dependencies among the columns of a matrix, the matrix is said to be of **full rank**, or **nonsingular**. If a matrix is *not* of full rank it is said to be **singular**. The number of linearly independent rows of a matrix will always equal the number of linearly independent columns. The linear dependency among the rows of $\boldsymbol{A}$ is shown by 9(row1)+ 7(row2) = 6(row3). The critical matrices in regression will almost always

have fewer columns than rows and, therefore, rank is more easily visualized by inspection of the columns.

The collection of all linear combinations of columns of $\boldsymbol{A}$ is called the **column space** of $\boldsymbol{A}$ or the space spanned by the columns of $\boldsymbol{A}$. **Column Space**

2.2 Special Types of Matrices

A **vector** is a matrix having only one row or one column, and is called a row or column vector, respectively. Although vectors are often designated with boldface lowercase letters, this convention is not followed rigorously in this text. A boldface capital letter is used to designate a data vector and a boldface Greek letter is used for vectors of parameters. Thus, for example, **Vector**

$$\boldsymbol{v} = \begin{pmatrix} 3 \\ 8 \\ 2 \\ 1 \end{pmatrix} \text{ is a } 4 \times 1 \text{ column vector.}$$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 & \mu_2 & \mu_3 \end{pmatrix} \text{ is a } 1 \times 3 \text{ row vector.}$$

We usually define the vectors as column vectors but they need not be. A single number such as 4, -2.1, or 0 is called a **scalar**.

A **square matrix** has an equal number of rows and columns. **Square Matrix**

$$\boldsymbol{D} = \begin{bmatrix} 2 & 4 \\ 6 & 7 \end{bmatrix} \text{ is a } 2 \times 2 \text{ square matrix.}$$

A **diagonal matrix** is a square matrix in which all elements are zero except the elements on the main diagonal, the diagonal of elements, a_{11}, a_{22}, $\ldots$, a_{nn}, running from the upper left postion to the lower right position. **Diagonal Matrix**

$$\boldsymbol{A} = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 8 \end{bmatrix} \text{ is a } 3 \times 3 \text{ diagonal matrix.}$$

An **identity matrix** is a diagonal matrix having all the diagonal elements equal to 1; such a matrix is denoted by $\boldsymbol{I}_n$. The subscript identifies the order of the matrix and is omitted when the order is clear from the context. **Identity Matrix**

$$\boldsymbol{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ is a } 3 \times 3 \text{ identity matrix.}$$

After matrix multiplication is discussed, it can be verified that multiplying any matrix by the identity matrix will not change the original matrix.

A **symmetric matrix** is a square matrix in which element a_{ij} equals element a_{ji} for all i and j. The elements form a symmetric pattern around the diagonal of the matrix. **Symmetric Matrix**

$$\boldsymbol{A} = \begin{bmatrix} 5 & -2 & 3 \\ -2 & 4 & -1 \\ 3 & -1 & 8 \end{bmatrix} \text{ is a } 3 \times 3 \text{ symmetric matrix.}$$

Note that the first row is identical to the first column, the second row is identical to the second column, and so on.

2.3 Matrix Operations

The **transpose** of a matrix $\boldsymbol{A}$, designated $\boldsymbol{A}'$, is the matrix obtained by using the rows of $\boldsymbol{A}$ as the columns of $\boldsymbol{A}'$. If **Transpose**

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 3 & 8 \\ 4 & 1 \\ 5 & 9 \end{bmatrix},$$

the transpose of $\boldsymbol{A}$ is

$$\boldsymbol{A}' = \begin{bmatrix} 1 & 3 & 4 & 5 \\ 2 & 8 & 1 & 9 \end{bmatrix}.$$

If a matrix $\boldsymbol{A}$ has order $m \times n$, its transpose $\boldsymbol{A}'$ has order $n \times m$. A symmetric matrix is equal to its transpose: $\boldsymbol{A}' = \boldsymbol{A}$.

Addition of two matrices is defined if and only if the matrices are of the same order. Then, addition (or subtraction) consists of adding (or subtracting) the corresponding elements of the two matrices. For example, **Addition**

$$\begin{bmatrix} 1 & 2 \\ 3 & 8 \end{bmatrix} + \begin{bmatrix} 7 & -6 \\ 8 & 2 \end{bmatrix} = \begin{bmatrix} 8 & -4 \\ 11 & 10 \end{bmatrix}.$$

Addition is commutative: $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A}$.

Multiplication of two matrices is defined if and only if the number of *columns* in the first matrix equals the number of *rows* in the second matrix. If $\boldsymbol{A}$ is of order $r \times s$ and $\boldsymbol{B}$ is of order $m \times n$, the matrix product $\boldsymbol{AB}$ exists only if $s = m$. The matrix product $\boldsymbol{BA}$ exists only if $r = n$. Multiplication is most easily defined by first considering the multiplication of a row vector times a column vector. Let $\boldsymbol{a}' = (\, a_1 \quad a_2 \quad a_3 \,)$ and $\boldsymbol{b}' = (\, b_1 \quad b_2 \quad b_3 \,)$. (Notice that both $\boldsymbol{a}$ and $\boldsymbol{b}$ are defined as column vectors.) Then, the product of $\boldsymbol{a}'$ and $\boldsymbol{b}$ is **Multiplication**

$$\begin{aligned} \boldsymbol{a}'\boldsymbol{b} &= (\, a_1 \quad a_2 \quad a_3 \,) \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \\ &= a_1 b_1 + a_2 b_2 + a_3 b_3. \end{aligned} \tag{2.1}$$

The result is a scalar equal to the sum of products of the corresponding elements. Let

$$\boldsymbol{a}' = \begin{pmatrix} 3 & 6 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{b}' = \begin{pmatrix} 2 & 4 & 8 \end{pmatrix}.$$

The matrix product is

$$\boldsymbol{a}'\boldsymbol{b} = \begin{pmatrix} 3 & 6 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 4 \\ 8 \end{pmatrix} = 6 + 24 + 8 = 38.$$

Matrix multiplication is defined as a sequence of vector multiplications. Write

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \quad \text{as} \quad \boldsymbol{A} = \begin{pmatrix} \boldsymbol{a}_1' \\ \boldsymbol{a}_2' \end{pmatrix},$$

where $\boldsymbol{a}_1' = \begin{pmatrix} a_{11} & a_{12} & a_{13} \end{pmatrix}$ and $\boldsymbol{a}_2' = \begin{pmatrix} a_{21} & a_{22} & a_{23} \end{pmatrix}$ are the 1×3 row vectors in $\boldsymbol{A}$. Similarly, write

$$\boldsymbol{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} \quad \text{as } \boldsymbol{B} = \begin{pmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 \end{pmatrix},$$

where $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are the 3×1 column vectors in $\boldsymbol{B}$. Then the product of $\boldsymbol{A}$ and $\boldsymbol{B}$ is the 2×2 matrix

$$\boldsymbol{A}\boldsymbol{B} = \boldsymbol{C} = \begin{bmatrix} \boldsymbol{a}_1'\boldsymbol{b}_1 & \boldsymbol{a}_1'\boldsymbol{b}_2 \\ \boldsymbol{a}_2'\boldsymbol{b}_1 & \boldsymbol{a}_2'\boldsymbol{b}_2 \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}, \tag{2.2}$$

where

$$\begin{aligned} c_{11} &= \boldsymbol{a}_1'\boldsymbol{b}_1 = \sum_{j=1}^{3} a_{1j}b_{j1} = a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} \\ c_{12} &= \boldsymbol{a}_1'\boldsymbol{b}_2 = \sum_{j=1}^{3} a_{1j}b_{j2} = a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ c_{21} &= \boldsymbol{a}_2'\boldsymbol{b}_1 = \sum_{j=1}^{3} a_{2j}b_{j1} = a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} \\ c_{22} &= \boldsymbol{a}_2'\boldsymbol{b}_2 = \sum_{j=1}^{3} a_{2j}b_{j2} = a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32}. \end{aligned}$$

In general, element c_{ij} is obtained from the vector multiplication of the ith *row* vector from the first matrix and the jth *column* vector from the second matrix. The resulting matrix $\boldsymbol{C}$ has the number of rows equal to

the number of rows in $\boldsymbol{A}$ and number of columns equal to the number of columns in $\boldsymbol{B}$.

Example 2.1

Let

$$\boldsymbol{T} = \begin{bmatrix} 1 & 2 \\ 4 & 5 \\ 3 & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{W} = \begin{pmatrix} -1 \\ 3 \end{pmatrix}.$$

The product $\boldsymbol{WT}$ is not defined since the number of columns in $\boldsymbol{W}$ is not equal to the number of rows in $\boldsymbol{T}$. The product $\boldsymbol{TW}$, however, is defined:

$$\begin{aligned} \boldsymbol{TW} &= \begin{bmatrix} 1 & 2 \\ 4 & 5 \\ 3 & 0 \end{bmatrix} \begin{pmatrix} -1 \\ 3 \end{pmatrix} \\ &= \begin{pmatrix} (1)(-1)+(2)(3) \\ (4)(-1)+(5)(3) \\ (3)(-1)+(0)(3) \end{pmatrix} = \begin{pmatrix} 5 \\ 11 \\ -3 \end{pmatrix}. \end{aligned}$$

The resulting matrix is of order 3×1 with the elements being determined by multiplication of the corresponding row vector from $\boldsymbol{T}$ with the column vector in $\boldsymbol{W}$. ■

Matrix multiplication is not commutative; $\boldsymbol{AB}$ does not necessarily equal $\boldsymbol{BA}$ even if both products exist. As for the matrices $\boldsymbol{W}$ and $\boldsymbol{T}$ in Example 2.1, the matrices are not of the proper order for multiplication to be defined in both ways. The first step in matrix multiplication is to verify that the matrices do conform (have the proper order) for multiplication.

The transpose of a product is equal to the product *in reverse order* of the transposes of the two matrices. That is,

$$(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'. \tag{2.3}$$

The transpose of the product of $\boldsymbol{T}$ and $\boldsymbol{W}$ from Example 2.1 is

$$\begin{aligned} (\boldsymbol{TW})' = \boldsymbol{W}'\boldsymbol{T}' &= \begin{pmatrix} -1 & 3 \end{pmatrix} \begin{bmatrix} 1 & 4 & 3 \\ 2 & 5 & 0 \end{bmatrix} \\ &= \begin{pmatrix} 5 & 11 & -3 \end{pmatrix}. \end{aligned}$$

Scalar multiplication is the multiplication of a matrix by a single number. Every element in the matrix is multiplied by the scalar. Thus,

$$3 \begin{bmatrix} 2 & 1 & 7 \\ 3 & 5 & 9 \end{bmatrix} = \begin{bmatrix} 6 & 3 & 21 \\ 9 & 15 & 27 \end{bmatrix}.$$

Determinant

The **determinant** of a matrix is a scalar computed from the elements of

the matrix according to well-defined rules. Determinants are defined only for square matrices and are denoted by $|\boldsymbol{A}|$, where $\boldsymbol{A}$ is a square matrix. The determinant of a 1×1 matrix is the scalar itself. The determinant of a 2×2 matrix,

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

is defined as

$$|\boldsymbol{A}| \quad = \quad a_{11}a_{22} - a_{12}a_{21}. \tag{2.4}$$

For example, if

$$\boldsymbol{A} = \begin{bmatrix} 1 & 6 \\ -2 & 10 \end{bmatrix},$$

the determinant of $\boldsymbol{A}$ is

$$|\boldsymbol{A}| = (1)(10) - (6)(-2) = 22.$$

The determinants of higher-order matrices are obtained by expanding the determinants as linear functions of determinants of 2×2 submatrices. First, it is convenient to define the **minor** and the **cofactor** of an element in a matrix. Let $\boldsymbol{A}$ be a square matrix of order n. For any element a_{rs} in $\boldsymbol{A}$, a square matrix of order $(n-1)$ is formed by eliminating the row and column containing the element a_{rs}. Label this matrix $\boldsymbol{A}_{rs}$, with the subscripts designating the row and column eliminated from $\boldsymbol{A}$. Then $|\boldsymbol{A}_{rs}|$, the determinant of $\boldsymbol{A}_{rs}$, is called the **minor** of the element a_{rs}. The product $\theta_{rs} = (-1)^{r+s} |\boldsymbol{A}_{rs}|$ is called the **cofactor** of a_{rs}. Each element in a square matrix has its own minor and cofactor.

The determinant of a matrix of order n is expressed in terms of the elements of any row or column and their cofactors. Using row i for illustration, we can express the determinant of $\boldsymbol{A}$ as

$$|\boldsymbol{A}| \quad = \quad \sum_{j=1}^{n} a_{ij}\theta_{ij}, \tag{2.5}$$

where each θ_{ij} contains a determinant of order $(n-1)$. Thus, the determinant of order n is expanded as a function of determinants of one less order. Each of these determinants, in turn, is expanded as a linear function of determinants of order $(n-2)$. This substitution of determinants of one less order continues until $|\boldsymbol{A}|$ is expressed in terms of determinants of 2×2 submatrices of $\boldsymbol{A}$.

The first step of the expansion is illustrated for a 3×3 matrix $\boldsymbol{A}$. To compute the determinant of $\boldsymbol{A}$, choose any row or column of the matrix. For each element of the row or column chosen, compute the cofactor of the element. Then, if the ith row of $\boldsymbol{A}$ is used for the expansion,

$$|\boldsymbol{A}| = a_{i1}\theta_{i1} + a_{i2}\theta_{i2} + a_{i3}\theta_{i3}. \tag{2.6}$$

Example 2.2

For illustration, let

$$\boldsymbol{A} = \begin{bmatrix} 2 & 4 & 6 \\ 1 & 2 & 3 \\ 5 & 7 & 9 \end{bmatrix}$$

and use the first row for the expansion of $|\boldsymbol{A}|$. The cofactors of the elements in the first row are

$$\begin{aligned} \theta_{11} &= (-1)^{(1+1)} \begin{vmatrix} 2 & 3 \\ 7 & 9 \end{vmatrix} = (18 - 21) = -3, \\ \theta_{12} &= (-1)^{(1+2)} \begin{vmatrix} 1 & 3 \\ 5 & 9 \end{vmatrix} = -(9 - 15) = 6, \text{ and} \\ \theta_{13} &= (-1)^{(1+3)} \begin{vmatrix} 1 & 2 \\ 5 & 7 \end{vmatrix} = (7 - 10) = -3. \end{aligned}$$

Then, the determinant of $\boldsymbol{A}$ is

$$|\boldsymbol{A}| = 2(-3) + 4(6) + 6(-3) = 0$$

. ■

If the determinant of a matrix is zero, the matrix is **singular**, or it is not of full rank. Otherwise, the matrix is **nonsingular**. Thus, the matrix $\boldsymbol{A}$ in Example 2.2 is singular. The linear dependency is seen by noting that row 1 is equal to twice row 2. The determinants of larger matrices rapidly become difficult to compute and are obtained with the help of a computer.

Inverse of a Matrix

Division in the usual sense does not exist in matrix algebra. The concept is replaced by multiplication by the **inverse** of the matrix. The inverse of a matrix $\boldsymbol{A}$, designated by $\boldsymbol{A}^{-1}$, is defined as the matrix that gives the identity matrix when multiplied by $\boldsymbol{A}$. That is,

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}. \tag{2.7}$$

The inverse of a matrix may not exist. A matrix has a *unique inverse* if and only if the matrix is square and nonsingular. A matrix is nonsingular if and only if its determinant is not zero.

The inverse of a 2×2 matrix is easily computed. If

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

then

$$\boldsymbol{A}^{-1} = \frac{1}{|\boldsymbol{A}|} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}. \tag{2.8}$$

Note the rearrangement of the elements and the use of the determinant of $\boldsymbol{A}$ as the scalar divisor. For example, if

$$\boldsymbol{A} = \begin{bmatrix} 4 & 3 \\ 1 & 2 \end{bmatrix}, \quad \text{then} \quad \boldsymbol{A}^{-1} = \begin{bmatrix} \frac{2}{5} & -\frac{3}{5} \\ -\frac{1}{5} & \frac{4}{5} \end{bmatrix}.$$

That this is the inverse of $\boldsymbol{A}$ is verified by multiplication of $\boldsymbol{A}$ and $\boldsymbol{A}^{-1}$:

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \begin{bmatrix} 4 & 3 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} \frac{2}{5} & -\frac{3}{5} \\ -\frac{1}{5} & \frac{4}{5} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The inverse of a matrix is obtained in general by (1) replacing every element of the matrix with its cofactor, (2) transposing the resulting matrix, and (3) dividing by the determinant of the original matrix, as illustrated in the next example.

Example 2.3

Consider the following matrix,

$$\boldsymbol{B} = \begin{bmatrix} 1 & 3 & 2 \\ 4 & 5 & 6 \\ 8 & 7 & 9 \end{bmatrix}.$$

The determinant of $\boldsymbol{B}$ is

$$\begin{aligned} |\boldsymbol{B}| &= 1\begin{vmatrix} 5 & 6 \\ 7 & 9 \end{vmatrix} - 3\begin{vmatrix} 4 & 6 \\ 8 & 9 \end{vmatrix} + 2\begin{vmatrix} 4 & 5 \\ 8 & 7 \end{vmatrix} \\ &= (45-42) - 3(36-48) + 2(28-40) \\ &= 15. \end{aligned}$$

The unique inverse of $\boldsymbol{B}$ exists since $|\boldsymbol{B}| \neq 0$. The cofactors for the elements of the first row of $\boldsymbol{B}$ were used in obtaining $|\boldsymbol{B}|$: $\theta_{11} = 3,\ \theta_{12} = 12,\ \theta_{13} = -12$. The remaining cofactors are:

$$\theta_{21} = -\begin{vmatrix} 3 & 2 \\ 7 & 9 \end{vmatrix} = -13 \quad \theta_{22} = \begin{vmatrix} 1 & 2 \\ 8 & 9 \end{vmatrix} = -7 \quad \theta_{23} = -\begin{vmatrix} 1 & 3 \\ 8 & 7 \end{vmatrix} = 17$$
$$\theta_{31} = \begin{vmatrix} 3 & 2 \\ 5 & 6 \end{vmatrix} = 8 \qquad \theta_{32} = -\begin{vmatrix} 1 & 2 \\ 4 & 6 \end{vmatrix} = 2 \quad \theta_{33} = \begin{vmatrix} 1 & 3 \\ 4 & 5 \end{vmatrix} = -7.$$

Thus, the matrix of cofactors is

$$\begin{bmatrix} 3 & 12 & -12 \\ -13 & -7 & 17 \\ 8 & 2 & -7 \end{bmatrix}$$

and the inverse of $\boldsymbol{B}$ is

$$\boldsymbol{B}^{-1} = \frac{1}{15}\begin{bmatrix} 3 & -13 & 8 \\ 12 & -7 & 2 \\ -12 & 17 & -7 \end{bmatrix}.$$

Notice that the matrix of cofactors has been transposed and divided by $|\boldsymbol{B}|$ to obtain $\boldsymbol{B}^{-1}$. It is left as an exercise to verify that this is the inverse of $\boldsymbol{B}$. As with the determinants, computers are used to find the inverses of larger matrices. ■

Inverse of a Diagonal Matrix

Note that if $\boldsymbol{A}$ is a diagonal nonsingular matrix, then $\boldsymbol{A}^{-1}$ is also a diagonal matrix where the diagonal elements of $\boldsymbol{A}^{-1}$ are the reciprocals of the diagonal elements of $\boldsymbol{A}$. That is, if

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{bmatrix},$$

where $a_{ii} \neq 0$, then

$$\boldsymbol{A} = \begin{bmatrix} a_{11}^{-1} & 0 & 0 & \cdots & 0 \\ 0 & a_{22}^{-1} & 0 & \cdots & 0 \\ 0 & 0 & a_{33}^{-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn}^{-1} \end{bmatrix}.$$

Also, if $\boldsymbol{A}$ and $\boldsymbol{B}$ are two nonsingular matrices, then

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{A}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}^{-1} \end{bmatrix}.$$

2.4 Geometric Interpretations of Vectors

The elements of an $n \times 1$ vector can be thought of as the coordinates of a point in an n-dimensional coordinate system. The vector is represented in this n-space as the directional line connecting the origin of the coordinate system to the point specified by the elements. The direction of the vector is *from* the origin *to* the point; an arrowhead at the terminus indicates direction.

Vector Length

To illustrate, let $\boldsymbol{x}' = (3 \quad 2)$. This vector is of order two and is plotted in two-dimensional space as the line vector going from the origin (0, 0) to

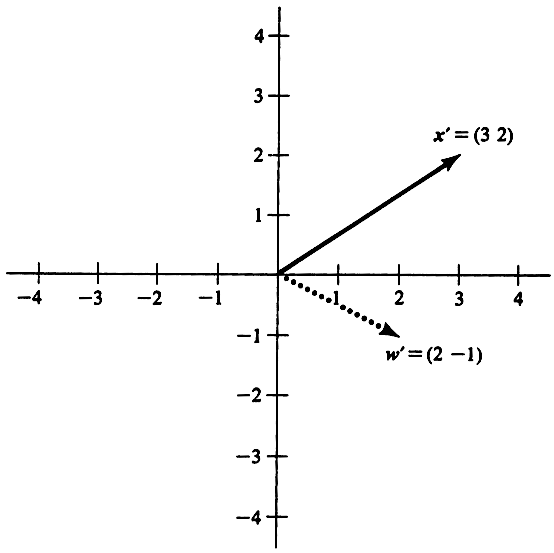

FIGURE 2.1. *The geometric representation of the vectors* $\boldsymbol{x}' = (3, 2)$ *and* $\boldsymbol{w}' = (2, -1)$ *in two-dimensional space.*

the point (3, 2) (see Figure 2.1). This can be viewed as the hypotenuse of a right triangle whose sides are of length 3 and 2, the elements of the vector $\boldsymbol{x}$. The length of $\boldsymbol{x}$ is then given by the Pythagorean theorem as the square root of the *sum of squares* of the elements of $\boldsymbol{x}$. Thus,

$$\text{length}(\boldsymbol{x}) = \sqrt{3^2 + 2^2} = \sqrt{13} = 3.61.$$

This result extends to the length of any vector regardless of its order. The sum of squares of the elements in a column vector $\boldsymbol{x}$ is given by (the matrix multiplication) $\boldsymbol{x}'\boldsymbol{x}$. Thus, the length of any vector $\boldsymbol{x}$ is

$$\text{length}(\boldsymbol{x}) = \sqrt{\boldsymbol{x}'\boldsymbol{x}}. \tag{2.9}$$

Space Defined by $\boldsymbol{x}$

Multiplication of $\boldsymbol{x}$ by a scalar defines another vector that falls precisely on the line formed by extending the vector $\boldsymbol{x}$ indefinitely in both directions. For example,

$$\boldsymbol{u}' = (-1)\boldsymbol{x}' = (\,-3 \quad -2\,)$$

falls on the extension of $\boldsymbol{x}$ in the negative direction. Any point on this indefinite extension of $\boldsymbol{x}$ in both directions can be "reached" by multiplication of $\boldsymbol{x}$ with an appropriate scalar. This set of points constitutes the **space** defined by $\boldsymbol{x}$, or the space **spanned** by $\boldsymbol{x}$. It is a one-dimensional subspace of the two-dimensional space in which the vectors are plotted. A single

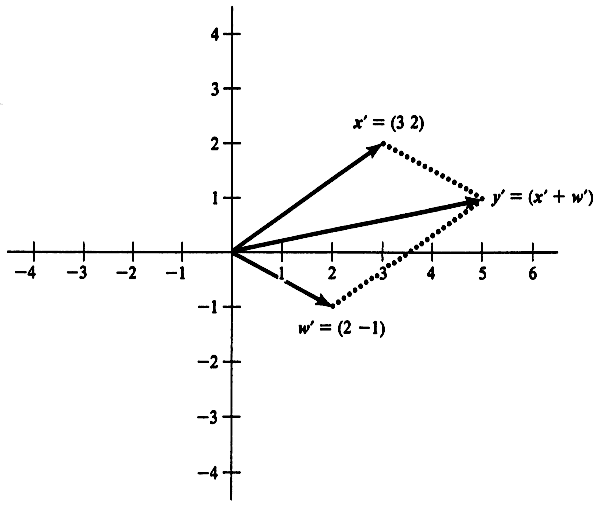

FIGURE 2.2. *Geometric representation of the sum of two vectors.*

vector of order n defines a one-dimensional subspace of the n-dimensional space in which the vector falls.

Linear Independence

The second vector $\boldsymbol{w}' = (2 \quad 1)$, shown in Figure 2.1 with a dotted line, defines another one-dimensional subspace. The two subspaces defined by $\boldsymbol{x}$ and $\boldsymbol{w}$ are disjoint subspaces (except for the common origin). The two vectors are said to be **linearly independent** since neither falls in the subspace defined by the other. This implies that one vector cannot be obtained by multiplication of the other vector by a scalar.

Two-Dimensional Subspace

If the two vectors are considered jointly, any point in the plane can be "reached" by an appropriate linear combination of the two vectors. For example, the sum of the two vectors gives the vector $\boldsymbol{y}$ (see Figure 2.2),

$$\boldsymbol{y}' = \boldsymbol{x}' + \boldsymbol{w}' = (3 \quad 2) + (2 \quad -1) = (5 \quad 1).$$

The two vectors $\boldsymbol{x}$ and $\boldsymbol{w}$ define, or span, the two-dimensional subspace represented by the plane in Figure 2.2. Any third vector of order 2 in this two-dimensional space *must* be a linear combination of $\boldsymbol{x}$ and $\boldsymbol{w}$. That is, there *must* be a linear dependency among any three vectors that fall on this plane.

Vector Addition

Geometrically, the vector $\boldsymbol{x}$ is added to $\boldsymbol{w}$ by moving $\boldsymbol{x}$, while maintaining its direction, until the base of $\boldsymbol{x}$ rests on the terminus of $\boldsymbol{w}$. The resultant vector $\boldsymbol{y}$ is the vector from the origin (0, 0) to the new terminus of $\boldsymbol{x}$. The same result is obtained by moving $\boldsymbol{w}$ along the vector $\boldsymbol{x}$. This is equivalent

to completing the parallelogram using the two original vectors as adjacent sides. The sum $\boldsymbol{y}$ is the diagonal of the parallelogram running from the origin to the opposite corner (see Figure 2.2). Subtraction of two vectors, say $\boldsymbol{w}' - \boldsymbol{x}'$, is most easily viewed as the addition of $\boldsymbol{w}'$ and $(-\boldsymbol{x}')$.

Three-Dimensional Subspace

Vectors of order 3 are considered briefly to show the more general behavior. Each vector of order 3 can be plotted in three-dimensional space; the elements of the vector define the endpoint of the vector. Each vector individually defines a one-dimensional subspace of the three-dimensional space. This subspace is formed by extending the vector indefinitely in both directions. Any *two* vectors define a two-dimensional subspace if the two vectors are **linearly independent**—that is, as long as the two vectors do not define the same subspace. The two-dimensional subspace defined by two vectors is the set of points in the *plane* defined by the origin and the endpoints of the two vectors. The two vectors defining the subspace and any linear combination of them lie in this plane.

A three-dimensional space contains an infinity of two-dimensional subspaces. These can be visualized by rotating the plane around the origin. Any third vector that does not fall in the original plane will, in conjunction with either of the first two vectors, define another plane. Any three linearly independent vectors, or any two planes, completely define, or span, the three-dimensional space. Any fourth vector in that three-dimensional subspace must be a linear function of the first three vectors. That is, any four vectors in a three-dimensional subspace *must* contain a linear dependency.

The general results are stated in the box:

1. Any vector of order n can be plotted in n-dimensional space and defines a one-dimensional subspace of the n-dimensional space.
2. Any p linearly independent vectors of order n, $p < n$, define a p-dimensional subspace.
3. Any $p + 1$ vectors in a p-dimensional subspace *must* contain a linear dependency.

Orthogonal Vectors

Two vectors $\boldsymbol{x}$ and $\boldsymbol{w}$ of the same order are **orthogonal** vectors if the vector product

$$\boldsymbol{x}'\boldsymbol{w} \quad = \quad \boldsymbol{w}'\boldsymbol{x} = 0. \tag{2.10}$$

If

$$\boldsymbol{x} = \begin{pmatrix} 1 \\ 0 \\ -1 \\ 4 \end{pmatrix} \quad \text{and} \quad \boldsymbol{w} = \begin{pmatrix} 3 \\ 4 \\ -1 \\ -1 \end{pmatrix},$$

then $\boldsymbol{x}$ and $\boldsymbol{w}$ are orthogonal because

$$\boldsymbol{x}'\boldsymbol{w} = (1)(3) + (0)(4) + (-1)(-1) + (4)(-1) = 0.$$

Geometrically, two orthogonal vectors are perpendicular to each other or they form a right angle at the origin.

Linearly Dependent Vectors

Two **linearly dependent** vectors form angles of 0 or 180 degrees at the origin. All other angles reflect vectors that are neither orthogonal nor linearly dependent. In general, the cosine of the angle α between two (column) vectors $\boldsymbol{x}$ and $\boldsymbol{w}$ is

$$\cos(\alpha) = \frac{\boldsymbol{x}'\boldsymbol{w}}{\sqrt{\boldsymbol{x}'\boldsymbol{x}}\sqrt{\boldsymbol{w}'\boldsymbol{w}}}. \tag{2.11}$$

If the elements of each vector have mean zero, the *cosine* of the angle formed by two vectors is the **product moment correlation** between the two columns of data in the vectors. Thus, orthogonality of two such vectors corresponds to a zero correlation between the elements in the two vectors. If two such vectors are linearly dependent, the correlation coefficient between the elements of the two vectors will be either $+1.0$ or -1.0 depending on whether the vectors have the same or opposite directions.

2.5 Linear Equations and Solutions

A set of r linear equations in s unknowns is represented in matrix notation as $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$, where $\boldsymbol{x}$ is a vector of the s unknowns, $\boldsymbol{A}$ is the $r \times s$ matrix of known coefficients on the s unknowns, and $\boldsymbol{y}$ is the $r \times 1$ vector of known constants on the right-hand side of the equations.

A set of equations may have (1) no solution, (2) a unique solution, or (3) an infinite number of solutions. In order to have at least one solution, the equations must be **consistent**. This means that any linear dependencies among the rows of $\boldsymbol{A}$ must also exist among the corresponding elements of $\boldsymbol{y}$ (Searle and Hausman, 1970). For example, the equations

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 3 & 3 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 10 \\ 9 \end{pmatrix}$$

are **inconsistent** since the second row of $\boldsymbol{A}$ is twice the first row but the second element of $\boldsymbol{y}$ is *not* twice the first element. Since they are *not* consistent, there is no solution to this set of equations. Note that $\boldsymbol{x}' = (\,1 \quad 1 \quad 1\,)$ satisfies the first and third equations but not the second. If the second element of $\boldsymbol{y}$ were 12 instead of 10, the equations would be consistent and the solution $\boldsymbol{x}' = (\,1 \quad 1 \quad 1\,)$ would satisfy all three equations.

Consistent Equations

One method of determining if a set of equations is consistent is to compare the rank of $\boldsymbol{A}$ to the rank of the augmented matrix $[\boldsymbol{A}\ \boldsymbol{y}]$. The equations are consistent if and only if

$$r(\boldsymbol{A}) = r([\boldsymbol{A}\ \boldsymbol{y}]). \tag{2.12}$$

Rank can be determined by using elementary (row and column) operations to reduce the elements below the diagonal to zero. Operations such as addition of two rows, interchanging rows, and obtaining a scalar multiple of a row are called elementary row operations. (In a rectangular matrix, the diagonal is defined as the elements $a_{11}, a_{22}, \ldots, a_{dd}$, where d is the lesser of the number of rows and number of columns.) The number of rows with at least one nonzero element after reduction is the rank of the matrix.

Example 2.4

Elementary operations on

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 3 & 3 \end{bmatrix}$$

give

$$\boldsymbol{A}^* = \begin{bmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & 0 & 0 \end{bmatrix}$$

so that $r(\boldsymbol{A}) = 2$. [The elementary operations to obtain $\boldsymbol{A}^*$ are (1) subtract 2 times row 1 from row 2, (2) subtract 3 times row 1 from row 3, and (3) interchange rows 2 and 3.] The same elementary operations, plus interchanging columns 3 and 4, on the augmented matrix

$$[\boldsymbol{A}\ \boldsymbol{y}] = \begin{bmatrix} 1 & 2 & 3 & 6 \\ 2 & 4 & 6 & 10 \\ 3 & 3 & 3 & 9 \end{bmatrix}$$

give

$$[\boldsymbol{A}\ \boldsymbol{y}]^* = \begin{bmatrix} 1 & 2 & 6 & 3 \\ 0 & -3 & -9 & -6 \\ 0 & 0 & -2 & 0 \end{bmatrix}.$$

Thus, $r([\boldsymbol{A}\ \boldsymbol{y}]) = 3$. Since $r([\boldsymbol{A}\ \boldsymbol{y}]) \neq r(\boldsymbol{A})$, the equations are not consistent and, therefore, they have no solution. ■

Unique Solution

Consistent equations either have a unique solution or an infinity of solutions. If $r(\boldsymbol{A})$ equals the number of unknowns, the solution is unique and is given by

1. $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y}$, when $\boldsymbol{A}$ is square; or
2. $\boldsymbol{x} = \boldsymbol{A}_1^{-1}\boldsymbol{y}$, where $\boldsymbol{A}_1$ is a full rank submatrix of $\boldsymbol{A}$, when $\boldsymbol{A}$ is rectangular.

Example 2.5

The equations $\boldsymbol{Ax} = \boldsymbol{y}$ with

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 3 & 3 \\ 5 & 7 \end{bmatrix} \quad \text{and} \quad \boldsymbol{y} = \begin{pmatrix} 6 \\ 9 \\ 21 \end{pmatrix}$$

are consistent. (Proof of consistency is left as an exercise.) The rank of $\boldsymbol{A}$ equals the number of unknowns $[r(\boldsymbol{A}) = 2]$, so that the solution is unique. Any two linearly independent equations in the system of equations can be used to obtain the solution. Using the first two rows gives the full-rank equations

$$\begin{bmatrix} 1 & 2 \\ 3 & 3 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 6 \\ 9 \end{pmatrix}$$

with the solution

$$\begin{aligned} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= \begin{bmatrix} 1 & 2 \\ 3 & 3 \end{bmatrix}^{-1} \begin{pmatrix} 6 \\ 9 \end{pmatrix} \\ &= \frac{1}{3} \begin{bmatrix} -3 & 2 \\ 3 & -1 \end{bmatrix} \begin{pmatrix} 6 \\ 9 \end{pmatrix} = \begin{pmatrix} 0 \\ 3 \end{pmatrix}. \end{aligned}$$

Notice that the solution $\boldsymbol{x}' = (\,0 \quad 3\,)$ satisfies the third equation also. ■

Infinite Solutions

When $r(\boldsymbol{A})$ in a consistent set of equations is less than the number of unknowns, there is an infinity of solutions.

Example 2.6

Consider the equations $\boldsymbol{Ax} = \boldsymbol{y}$ with

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 3 & 3 \end{bmatrix} \quad \text{and } \boldsymbol{y} = \begin{pmatrix} 6 \\ 12 \\ 9 \end{pmatrix}.$$

The rank of $\boldsymbol{A}$ is $r(\boldsymbol{A}) = 2$ and elementary operations on the augmented matrix $[\boldsymbol{A}\ \boldsymbol{y}]$ give

$$[\boldsymbol{A}\ \boldsymbol{y}]^* = \begin{bmatrix} 1 & 2 & 3 & 6 \\ 0 & -3 & -6 & -18 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Thus, $r([\boldsymbol{A}\ \boldsymbol{y}]) = 2$, which equals $r(\boldsymbol{A})$, and the equations are consistent. However, $r(\boldsymbol{A})$ is less than the number of unknowns so that there is an

infinity of solutions. This infinity of solutions comes from the fact that one element of $\boldsymbol{x}$ can be chosen arbitrarily and the remaining two chosen so as to satisfy the set of equations. For example, if x_1 is chosen to be 1, the solution is $\boldsymbol{x}' = (1 \quad 1 \quad 1)$, whereas if x_1 is chosen to be 2, the solution is $\boldsymbol{x}' = (2 \quad -1 \quad 2)$. ■

Solutions Using Generalized Inverses

A more general method of finding a solution to a set of consistent equations involves the use of **generalized inverses**. There are several definitions of generalized inverses [see Searle (1971), Searle and Hausman (1970), and Rao (1973)]. An adequate definition for our purposes is the following (Searle and Hausman, 1970).

> A generalized inverse of $\boldsymbol{A}$ is any matrix $\boldsymbol{A}^-$ that satisfies the condition $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$.

($\boldsymbol{A}^-$ is used to denote a generalized inverse.) The generalized inverse is not unique (unless $\boldsymbol{A}$ is square and of full rank, in which case $\boldsymbol{A}^- = \boldsymbol{A}^{-1}$). A generalized inverse can be used to express a solution to a set of consistent equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ as $\boldsymbol{x} = \boldsymbol{A}^-\boldsymbol{y}$. This solution is unique only when $r(\boldsymbol{A})$ equals the number of unknowns in the set of equations. (The computer is used to obtain generalized inverses when needed.)

Example 2.7

For illustration, consider the set of consistent equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ where

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 3 & 3 \\ 5 & 7 \end{bmatrix} \quad \text{and} \quad \boldsymbol{y} = \begin{pmatrix} 6 \\ 9 \\ 21 \end{pmatrix}.$$

It has been shown that $r(\boldsymbol{A}) = 2$ which equals the number of unknowns so that the solution is unique. A generalized inverse of $\boldsymbol{A}$ is

$$\boldsymbol{A}^- = \frac{1}{18}\begin{bmatrix} -10 & 16 & -4 \\ 8 & -11 & 5 \end{bmatrix}$$

and the unique solution is given by

$$\boldsymbol{x} = \boldsymbol{A}^-\boldsymbol{y} = \begin{pmatrix} 0 \\ 3 \end{pmatrix}.$$

It is left as an exercise to verify the matrix multiplication of $\boldsymbol{A}^-\boldsymbol{y}$ and that $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$. ■

Example 2.8

For another illustration, consider again the consistent equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ from Example 2.6, where

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 3 & 3 \end{bmatrix} \quad \text{and } \boldsymbol{y} = \begin{pmatrix} 6 \\ 12 \\ 9 \end{pmatrix}.$$

This system has been shown to have an infinity of solutions. A generalized inverse of $\boldsymbol{A}$ is

$$\boldsymbol{A}^{-} = \begin{bmatrix} -\frac{1}{10} & -\frac{2}{10} & \frac{4}{9} \\ 0 & 0 & \frac{1}{9} \\ \frac{1}{10} & \frac{2}{10} & -\frac{2}{9} \end{bmatrix},$$

which gives the solution

$$\boldsymbol{x} = \boldsymbol{A}^{-}\boldsymbol{y} = (\,1 \quad 1 \quad 1\,)'.$$

This happens to agree with the first solution obtained in Example 2.6. Again, it is left as an exercise to verify that $\boldsymbol{x} = \boldsymbol{A}^{-}\boldsymbol{y}$ and $\boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{A} = \boldsymbol{A}$. A different generalized inverse of $\boldsymbol{A}$ may lead to another solution of the equations. ■

2.6 Orthogonal Transformations and Projections

The linear transformation of vector $\boldsymbol{x}$ to vector $\boldsymbol{y}$, both of order n, is written as $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{A}$ is the $n \times n$ matrix of coefficients effecting the transformation. The transformation is a one-to-one transformation only if $\boldsymbol{A}$ is nonsingular. Then, the inverse transformation of $\boldsymbol{y}$ to $\boldsymbol{x}$ is $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y}$.

Orthogonal Transformations

A linear transformation is an **orthogonal** transformation if $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}$. This condition implies that the row vectors of $\boldsymbol{A}$ are orthogonal and of unit length. Orthogonal transformations maintain distances and angles between vectors. That is, the spatial relationships among the vectors are not changed with orthogonal transformations.

Example 2.9

For illustration, let $\boldsymbol{y}_1' = (\,3 \quad 10 \quad 20\,)$, $\boldsymbol{y}_2' = (\,6 \quad 14 \quad 21\,)$, and

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & 2 & -1 \end{bmatrix}.$$

Then

$$\boldsymbol{x}_1 = \boldsymbol{A}\boldsymbol{y}_1 = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & 2 & -1 \end{bmatrix} \begin{pmatrix} 3 \\ 10 \\ 20 \end{pmatrix} = \begin{pmatrix} 33 \\ 17 \\ -3 \end{pmatrix}$$

and

$$\boldsymbol{x}_2 = \boldsymbol{A}\boldsymbol{y}_2 = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & 2 & -1 \end{bmatrix} \begin{pmatrix} 6 \\ 14 \\ 21 \end{pmatrix} = \begin{pmatrix} 41 \\ 15 \\ 1 \end{pmatrix}$$

are linear transformations of $\boldsymbol{y}_1$ to $\boldsymbol{x}_1$ and $\boldsymbol{y}_2$ to $\boldsymbol{x}_2$. These are not orthogonal transformations because

$$\boldsymbol{A}\boldsymbol{A}' = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 6 \end{bmatrix} \neq \boldsymbol{I}.$$

The rows of $\boldsymbol{A}$ are mutually orthogonal (the off-diagonal elements are zero) but they do not have unit length. This can be made into an orthogonal transformation by scaling each row vector of $\boldsymbol{A}$ to have unit length by dividing each vector by its length. Thus,

$$\boldsymbol{x}_1^* = \boldsymbol{A}^*\boldsymbol{y}_1 = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \end{bmatrix} \boldsymbol{y}_1 = \begin{pmatrix} \frac{33}{\sqrt{3}} \\ \frac{17}{\sqrt{2}} \\ -\frac{3}{\sqrt{6}} \end{pmatrix}$$

and

$$\boldsymbol{x}_2^* = \boldsymbol{A}^*\boldsymbol{y}_2 = \begin{pmatrix} \frac{41}{\sqrt{3}} \\ \frac{15}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} \end{pmatrix}$$

are orthogonal transformations. It is left as an exercise to verify that the orthogonal transformation has maintained the distance between the two vectors; that is, verify that

$$(\boldsymbol{y}_1 - \boldsymbol{y}_2)'(\boldsymbol{y}_1 - \boldsymbol{y}_2) = (\boldsymbol{x}_1^* - \boldsymbol{x}_2^*)'(\boldsymbol{x}_1^* - \boldsymbol{x}_2^*) = 26.$$

[The squared distance between two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ is $(\boldsymbol{u} - \boldsymbol{v})'(\boldsymbol{u} - \boldsymbol{v})$.] ■

Projections

Projection of a vector onto a subspace is a special case of a transformation. (Projection is a key step in least squares.) The objective of a projection is to transform $\boldsymbol{y}$ in n-dimensional space to that vector $\widehat{\boldsymbol{y}}$ in a subspace such that $\widehat{\boldsymbol{y}}$ is as close to $\boldsymbol{y}$ as possible. A linear transformation of $\boldsymbol{y}$ to $\widehat{\boldsymbol{y}}$, $\widehat{\boldsymbol{y}} = \boldsymbol{P}\boldsymbol{y}$, is a **projection** if and only if $\boldsymbol{P}$ is idempotent and symmetric (Rao, 1973), in which case $\boldsymbol{P}$ is referred to as a projection matrix.

Idempotent Matrices

An **idempotent** matrix is a square matrix that remains unchanged when multiplied by itself. That is, the matrix $\boldsymbol{A}$ is idempotent if $\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$. It can be verified that the rank of an idempotent matrix is equal to the sum of the elements on the diagonal (Searle, 1982; Searle and Hausman, 1970). This sum of elements on the diagonal of a square matrix is called the **trace** of

the matrix and is denoted by $\mathrm{tr}(\boldsymbol{A})$. Symmetry is not required for a matrix to be idempotent. However, all idempotent matrices with which we are concerned are symmetric.

The subspace of a projection is defined, or spanned, by the columns or rows of the projection matrix $\boldsymbol{P}$. If $\boldsymbol{P}$ is a projection matrix, $(\boldsymbol{I}-\boldsymbol{P})$ is also a projection matrix. However, since $\boldsymbol{P}$ and $(\boldsymbol{I}-\boldsymbol{P})$ are orthogonal matrices, the projection by $(\boldsymbol{I}-\boldsymbol{P})$ is onto the subspace *orthogonal* to that defined by $\boldsymbol{P}$. The rank of a projection matrix is the dimension of the subspace onto which it projects and, since the projection matrix is idempotent, the rank is equal to its trace.

Example 2.10

The matrix

$$\boldsymbol{A} = \frac{1}{6}\begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix}$$

is idempotent since

$$\begin{aligned} \boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}^2 &= \frac{1}{6}\begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix} \frac{1}{6}\begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix} \\ &= \frac{1}{6}\begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix} = \boldsymbol{A}. \end{aligned}$$

The rank of $\boldsymbol{A}$ is given by

$$r(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A}) = \frac{1}{6}(5+2+5) = 2.$$

■

Since $\boldsymbol{A}$ is symmetric, it is also a projection matrix. Thus, the linear transformation

$$\widehat{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{y}_1 = \frac{1}{6}\begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix} \begin{pmatrix} 3 \\ 10 \\ 20 \end{pmatrix} = \begin{pmatrix} 2.5 \\ 11.0 \\ 19.5 \end{pmatrix}$$

is a projection of $\boldsymbol{y}_1 = (3 \quad 10 \quad 20)'$ onto the subspace defined by the columns of $\boldsymbol{A}$. The vector $\widehat{\boldsymbol{y}}$ is the unique vector in this subspace that is closest to $\boldsymbol{y}_1$. That is, $(\boldsymbol{y}_1 - \widehat{\boldsymbol{y}})'(\boldsymbol{y}_1 - \widehat{\boldsymbol{y}})$ is a minimum. Since $\boldsymbol{A}$ is a projection matrix, so is

$$\boldsymbol{I} - \boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \frac{1}{6}\begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix}.$$

Then,

$$\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{y}_1 = \frac{1}{6}\begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix}\begin{pmatrix} 3 \\ 10 \\ 20 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ -1 \\ \frac{1}{2} \end{pmatrix}$$

is a projection onto the subspace orthogonal to the subspace defined by $\boldsymbol{A}$. Note that $\widehat{\boldsymbol{y}}'\boldsymbol{e} = 0$ and $\widehat{\boldsymbol{y}} + \boldsymbol{e} = \boldsymbol{y}_1$. ■

2.7 Eigenvalues and Eigenvectors

Eigenvalues and eigenvectors of matrices are needed for some of the methods to be discussed, including principal component analysis, principal component regression, and assessment of the impact of collinearity (see Chapter 13). Determining the eigenvalues and eigenvectors of a matrix is a difficult computational problem and computers are used for all but the very simplest cases. However, the reader needs to develop an understanding of the eigenanalysis of a matrix.

The discussion of eigenanalysis is limited to real, symmetric, nonnegative definite matrices and, then, only key results are given. The reader is referred to other texts [such as Searle and Hausman (1970)] for more general discussions. In particular, Searle and Hausman (1970) show several important applications of eigenanalysis of asymmetric matrices. **Real matrices** do not contain any complex numbers as elements. **Symmetric, nonnegative definite matrices** are obtained from products of the type $\boldsymbol{B}'\boldsymbol{B}$ and, if used as defining matrices in quadratic forms (see Chapter 4), yield only zero or positive scalars.

Definitions

It can be shown that for a **real**, **symmetric** matrix $\boldsymbol{A}$ $(n \times n)$ there exists a set of n scalars λ_i, and n nonzero vectors $\boldsymbol{z}_i$, $i = 1, \ldots, n$, such that

$$\begin{aligned} \boldsymbol{A}\boldsymbol{z}_i &= \lambda_i \boldsymbol{z}_i, \\ \text{or} \quad \boldsymbol{A}\boldsymbol{z}_i - \lambda_i \boldsymbol{z}_i &= \boldsymbol{0}, \\ \text{or} \quad (\boldsymbol{A} - \lambda_i \boldsymbol{I})\boldsymbol{z}_i &= \boldsymbol{0}, \quad i = 1, \ldots, n. \end{aligned} \tag{2.13}$$

The λ_i are the n **eigenvalues** (characteristic roots or latent roots) of the matrix $\boldsymbol{A}$ and the $\boldsymbol{z}_i$ are the corresponding (column) **eigenvectors** (characteristic vectors or latent vectors).

Solution

There are nonzero solutions to equation 2.13 only if the matrix $(\boldsymbol{A} - \lambda_i \boldsymbol{I})$ is less than full rank—that is, only if the determinant of $(\boldsymbol{A} - \lambda_i \boldsymbol{I})$ is zero. The λ_i are obtained by solving the general determinantal equation

$$|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0. \tag{2.14}$$

Since $\boldsymbol{A}$ is of order $n \times n$, the determinant of $(\boldsymbol{A} - \lambda \boldsymbol{I})$ is an nth degree polynomial in λ. Solving this equation gives the n values of λ, which are not necessarily distinct. Each value of λ is then used in turn in Equation 2.13 to find the companion eigenvector $\boldsymbol{z}_i$.

When the eigenvalues are distinct, the vector solution to Equation 2.13 is unique except for an arbitrary scale factor and sign. By convention, each eigenvector is defined to be the solution vector scaled to have unit length; that is, $\boldsymbol{z}_i'\boldsymbol{z}_i = 1$. Furthermore, the eigenvectors are mutually orthogonal; $\boldsymbol{z}_i'\boldsymbol{z}_j = 0$ when $i \neq j$. When the eigenvalues are not distinct, there is an additional degree of arbitrariness in defining the subsets of vectors corresponding to each subset of nondistinct eigenvalues. Nevertheless, the eigenvectors for each subset can be chosen so that they are mutually orthogonal as well as orthogonal to the eigenvectors of all other eigenvalues. Thus, if $\boldsymbol{Z} = (\,\boldsymbol{z}_1 \quad \boldsymbol{z}_2 \quad \cdots \quad \boldsymbol{z}_n\,)$ is the matrix of eigenvectors, then $\boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{I}$. This implies that $\boldsymbol{Z}'$ is the inverse of $\boldsymbol{Z}$ so that $\boldsymbol{Z}\boldsymbol{Z}' = \boldsymbol{I}$ as well.

Decomposition of a Matrix

Using $\boldsymbol{Z}$ and $\boldsymbol{L}$, defined as the diagonal matrix of the λ_i, we can write the initial equations $\boldsymbol{A}\boldsymbol{z}_i = \lambda_i \boldsymbol{z}_i$ as

$$\begin{aligned} \boldsymbol{A}\boldsymbol{Z} &= \boldsymbol{Z}\boldsymbol{L}, && (2.15)\\ \text{or}\quad \boldsymbol{Z}'\boldsymbol{A}\boldsymbol{Z} &= \boldsymbol{L}, && (2.16)\\ \text{or}\quad \boldsymbol{A} &= \boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}'. && (2.17) \end{aligned}$$

Equation 2.17 shows that a real symmetric matrix $\boldsymbol{A}$ can be transformed to a diagonal matrix by pre- and postmultiplying by $\boldsymbol{Z}'$ and $\boldsymbol{Z}$, respectively. Since $\boldsymbol{L}$ is a diagonal matrix, equation 2.17 shows that $\boldsymbol{A}$ can be expressed as the sum of matrices:

$$\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}' = \sum \lambda_i (\boldsymbol{z}_i \boldsymbol{z}_i'), \qquad (2.18)$$

where the summation is over the n eigenvalues and eigenvectors. Each term is an $n \times n$ matrix of rank 1 so that the sum can be viewed as a decomposition of the matrix $\boldsymbol{A}$ into n matrices that are mutually orthogonal. Some of these may be zero matrices if the corresponding λ_i are zero. The rank of $\boldsymbol{A}$ is revealed by the number of nonzero eigenvalues λ_i.

Example 2.11

For illustration, consider the matrix

$$\boldsymbol{A} = \begin{bmatrix} 10 & 3 \\ 3 & 8 \end{bmatrix}.$$

The eigenvalues of $\boldsymbol{A}$ are found by solving the determinantal equation (equation 2.14),

$$|(\boldsymbol{A} - \lambda \boldsymbol{I})| = \left| \begin{bmatrix} 10 - \lambda & 3 \\ 3 & 8 - \lambda \end{bmatrix} \right| = 0$$

or

$$(10-\lambda)(8-\lambda)-9 \quad = \quad \lambda^2-18\lambda+71=0.$$

The solutions to this quadratic (in λ) equation are

$$\lambda_1 = 12.16228 \quad \text{and} \quad \lambda_2 = 5.83772$$

arbitrarily ordered from largest to smallest. Thus, the matrix of eigenvalues of $\boldsymbol{A}$ is

$$\boldsymbol{L} = \begin{bmatrix} 12.16228 & 0 \\ 0 & 5.83772 \end{bmatrix}.$$

The eigenvector corresponding to $\lambda_1 = 12.16228$ is obtained by solving equation 2.13 for the elements of $\boldsymbol{z}_1$:

$$(\boldsymbol{A} - 12.16228\boldsymbol{I}) \begin{pmatrix} z_{11} \\ z_{21} \end{pmatrix} = 0$$

or

$$\begin{bmatrix} -2.162276 & 3 \\ 3 & -4.162276 \end{bmatrix} \begin{pmatrix} z_{11} \\ z_{21} \end{pmatrix} = 0.$$

Arbitrarily setting $z_{11} = 1$ and solving for z_{21}, using the first equation, gives $z_{21} = .720759$. Thus, the vector $\boldsymbol{z}_1' = (\,1 \quad .720759\,)$ satisfies the first equation (and it can be verified that it also satisfies the second equation). Rescaling this vector so it has unit length by dividing by

$$\text{length}(\boldsymbol{z}_1) = \sqrt{\boldsymbol{z}_1'\boldsymbol{z}_1} = \sqrt{1.5194935} = 1.232677$$

gives the first eigenvector

$$\boldsymbol{z}_1 = (\,.81124 \quad .58471\,)'.$$

The elements of $\boldsymbol{z}_2$ are found in the same manner to be

$$\boldsymbol{z}_2 = (\,-.58471 \quad .81124\,)'.$$

Thus, the matrix of eigenvectors for $\boldsymbol{A}$ is

$$\boldsymbol{Z} = \begin{bmatrix} .81124 & -.58471 \\ .58471 & .81124 \end{bmatrix}.$$

Notice that the first *column* of $\boldsymbol{Z}$ is the first eigenvector, and the second column is the second eigenvector. ■

Example 2.12

Continuing with Example 2.11, notice that the matrix $\boldsymbol{A}$ is of rank two because both eigenvalues are nonzero. The decomposition of $\boldsymbol{A}$ into two

orthogonal matrices each of rank one, $\boldsymbol{A} = \boldsymbol{A}_1 + \boldsymbol{A}_2$, equation 2.18, is given by

$$\begin{aligned}\boldsymbol{A}_1 = \lambda_1 \boldsymbol{z}_1 \boldsymbol{z}_1' &= 12.16228 \begin{pmatrix} .81124 \\ .58471 \end{pmatrix} (.81124 \quad .58471) \\ &= \begin{bmatrix} 8.0042 & 5.7691 \\ 5.7691 & 4.1581 \end{bmatrix}\end{aligned}$$

and

$$\boldsymbol{A}_2 = \lambda_2 \boldsymbol{z}_2 \boldsymbol{z}_2' = \begin{bmatrix} 1.9958 & -2.7691 \\ -2.7691 & 3.8419 \end{bmatrix}.$$

Since the two columns of $\boldsymbol{A}_1$ are multiples of the same vector $\boldsymbol{u}_1$, they are linearly dependent and, therefore, $r(\boldsymbol{A}_1) = 1$. Similarly, $r(\boldsymbol{A}_2) = 1$. Multiplication of $\boldsymbol{A}_1$ with $\boldsymbol{A}_2$ shows that the two matrices are orthogonal to each other: $\boldsymbol{A}_1 \boldsymbol{A}_2 = \boldsymbol{0}$, where $\boldsymbol{0}$ is a 2×2 matrix of zeros. Thus, the eigenanalysis has decomposed the rank-2 matrix $\boldsymbol{A}$ into two rank-1 matrices. It is left as an exercise to verify the multiplication and that $\boldsymbol{A}_1 + \boldsymbol{A}_2 = \boldsymbol{A}$. ■

Notice that the sum of the eigenvalues in Example 2.11, $\lambda_1 + \lambda_2 = 18$, is equal to $\text{tr}(\boldsymbol{A})$. This is a general result: the sum of the eigenvalues for any square symmetric matrix is equal to the trace of the matrix. Furthermore, the trace of each of the component rank-1 matrices is equal to its eigenvalue:

$$\text{tr}(\boldsymbol{A}_1) = \lambda_1 \quad \text{and} \quad \text{tr}(\boldsymbol{A}_2) = \lambda_2.$$

Note that for $\boldsymbol{A} = \boldsymbol{B}'\boldsymbol{B}$, we have

$$\boldsymbol{z}_i' \boldsymbol{A} \boldsymbol{z}_i = \lambda_i \boldsymbol{z}_i' \boldsymbol{z}_i$$

and

$$\begin{aligned}\lambda_i &= \frac{\boldsymbol{z}_i' \boldsymbol{A} \boldsymbol{z}_i}{\boldsymbol{z}_i' \boldsymbol{z}_i} = \frac{\boldsymbol{z}_i' \boldsymbol{B}' \boldsymbol{B} \boldsymbol{z}_i}{\boldsymbol{z}_i' \boldsymbol{z}_i} \\ &= \frac{\boldsymbol{c}_i' \boldsymbol{c}_i}{\boldsymbol{z}_i' \boldsymbol{z}_i},\end{aligned}$$

where $\boldsymbol{c}_i = \boldsymbol{B}\boldsymbol{z}_i$. Therefore, if $\boldsymbol{A} = \boldsymbol{B}'\boldsymbol{B}$ for some real matrix $\boldsymbol{B}$, then the eigenvalues of $\boldsymbol{A}$ are nonnegative. Symmetric matrices with nonnegative eigenvalues are called **nonnegative definite matrices**.

2.8 Singular Value Decomposition

The eigenanalysis, Section 2.7, applies to a square symmetric matrix. In this section, the eigenanalysis is used to develop a similar decomposition,

called the **singular value decomposition**, for a rectangular matrix. The singular value decomposition is then used to give the **principal component analysis**.

Singular Value Decomposition

Let $\boldsymbol{X}$ be an $n \times p$ matrix with $n > p$. Then $\boldsymbol{X}'\boldsymbol{X}$ is a square symmetric matrix of order $p \times p$. From Section 2.7, $\boldsymbol{X}'\boldsymbol{X}$ can be expressed in terms of its eigenvalues $\boldsymbol{L}$ and eigenvectors $\boldsymbol{Z}$ as

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}'. \tag{2.19}$$

Here $\boldsymbol{L}$ is a diagonal matrix consisting of eigenvalues $\lambda_1, \ldots, \lambda_p$ of $\boldsymbol{X}'\boldsymbol{X}$. From Section 2.7, we know that $\lambda_1, \ldots, \lambda_p$ are nonnegative. Similarly, $\boldsymbol{X}\boldsymbol{X}'$ is a square symmetric matrix but of order $n \times n$. The rank of $\boldsymbol{X}\boldsymbol{X}'$ will be at most p so there will be at most p nonzero eigenvalues; they are in fact the same p eigenvalues obtained from $\boldsymbol{X}'\boldsymbol{X}$. In addition, $\boldsymbol{X}\boldsymbol{X}'$ will have at least $n - p$ eigenvalues that are zero. These $n - p$ eigenvalues and their vectors are dropped in the following. Denote with $\boldsymbol{U}$ the matrix of eigenvectors of $\boldsymbol{X}\boldsymbol{X}'$ that correspond to the p eigenvalues common to $\boldsymbol{X}'\boldsymbol{X}$. Each eigenvector $\boldsymbol{u}_i$ will be of order $n \times 1$. Then,

$$\boldsymbol{X}\boldsymbol{X}' = \boldsymbol{U}\boldsymbol{L}\boldsymbol{U}'. \tag{2.20}$$

Equations 2.19 and 2.20 jointly imply that the rectangular matrix $\boldsymbol{X}$ can be written as

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{L}^{1/2}\boldsymbol{Z}', \tag{2.21}$$

where $\boldsymbol{L}^{1/2}$ is the diagonal matrix of the positive square roots of the p eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$. Thus, $\boldsymbol{L}^{1/2}\boldsymbol{L}^{1/2} = \boldsymbol{L}$. Equation 2.21 is the **singular value decomposition** of the rectangular matrix $\boldsymbol{X}$. The elements of $\boldsymbol{L}^{1/2}$, $\lambda_i^{1/2}$ are called the **singular values** and the *column* vectors in $\boldsymbol{U}$ and $\boldsymbol{Z}$ are the left and right singular vectors, respectively.

Since $\boldsymbol{L}^{1/2}$ is a diagonal matrix, the singular value decomposition expresses $\boldsymbol{X}$ as a sum of p rank-1 matrices,

$$\boldsymbol{X} = \sum \lambda_i^{1/2} \boldsymbol{u}_i \boldsymbol{z}_i', \tag{2.22}$$

where summation is over $i = 1, \ldots, p$. Furthermore, if the eigenvalues have been ranked from largest to smallest, the first of these matrices is the "best" rank-1 approximation to $\boldsymbol{X}$, the *sum* of the first two matrices is the "best" rank-2 approximation of $\boldsymbol{X}$, and so forth. These are "best" approximations in the least squares sense; that is, no other matrix (of the same rank) will give a better agreement with the original matrix $\boldsymbol{X}$ as measured by the sum of squared differences between the corresponding elements of $\boldsymbol{X}$ and the approximating matrix (Householder and Young, 1938). The goodness of fit of the approximation in each case is given by the ratio of the sum of the eigenvalues (squares of the singular values)

used in the approximation to the sum of *all* eigenvalues. Thus, the rank-1 approximation has a goodness of fit of $\lambda_1 / \sum \lambda_i$, the rank-2 approximation has a goodness of fit of $(\lambda_1 + \lambda_2) / \sum \lambda_i$, and so forth.

Recall that there is an arbitrariness of sign for the eigenvectors obtained from the eigenalysis of $\boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{X}\boldsymbol{X}'$. Thus, care must be exercised in choice of sign for the eigenvectors in reconstructing $\boldsymbol{X}$ or lower-order approximations of $\boldsymbol{X}$ when the left and right eigenvectors have been obtained from eigenanalyses. This is not a problem when $\boldsymbol{U}$ and $\boldsymbol{Z}$ have been obtained directly from the singular value decomposition of $\boldsymbol{X}$.

Example 2.13

Singular value decomposition is illustrated using data on average minimum daily temperature X_1, average maximum daily temperature X_2, total rainfall X_3, and total growing degree days X_4, for six locations. The data were reported by Saeed and Francis (1984) to relate environmental conditions to cultivar by environment interactions in sorghum and are used with their kind permission. Each variable has been centered to have zero mean, and standardized to have unit sum of squares. (The centering and standardization are not necessary for a singular value decomposition. The centering removes the mean effect of each variable so that the dispersion about the mean is being analyzed. The standardization puts all variables on an equal basis and is desirable in most cases, particularly when the variables have different units of measure.) The $\boldsymbol{X}$ matrix is

$$\begin{aligned}\boldsymbol{X} &= (\boldsymbol{X}_1 \quad \boldsymbol{X}_2 \quad \boldsymbol{X}_3 \quad \boldsymbol{X}_4) \\ &= \begin{bmatrix} .178146 & -.523245 & .059117 & -.060996 \\ .449895 & -.209298 & .777976 & .301186 \\ -.147952 & .300866 & -.210455 & -.053411 \\ -.057369 & .065406 & .120598 & -.057203 \\ -.782003 & -.327028 & -.210455 & -.732264 \\ .359312 & .693299 & -.536780 & .602687 \end{bmatrix}.\end{aligned}$$

The singular value decomposition of $\boldsymbol{X}$ into $\boldsymbol{U}\boldsymbol{L}^{1/2}\boldsymbol{Z}'$ gives

$$\boldsymbol{U} = \begin{bmatrix} -.113995 & .308905 & -.810678 & .260088 \\ .251977 & .707512 & .339701 & -.319261 \\ .007580 & -.303203 & .277432 & .568364 \\ -.028067 & .027767 & .326626 & .357124 \\ -.735417 & -.234888 & .065551 & -.481125 \\ .617923 & -.506093 & -.198632 & -.385189 \end{bmatrix}$$

$$\boldsymbol{L}^{1/2} = \begin{bmatrix} 1.496896 & 0 & 0 & 0 \\ 0 & 1.244892 & 0 & 0 \\ 0 & 0 & .454086 & 0 \\ 0 & 0 & 0 & .057893 \end{bmatrix}$$

$$\boldsymbol{Z} = \begin{bmatrix} .595025 & .336131 & .383204 & .621382 \\ .451776 & .540753 & .657957 & .265663 \\ .004942 & .768694 & .639051 & .026450 \\ .664695 & .060922 & .108909 & .736619 \end{bmatrix}.$$

The columns of $\boldsymbol{U}$ and $\boldsymbol{Z}$ are the left and right singular vectors, respectively. The first column of $\boldsymbol{U}$, $\boldsymbol{u}_1$, the first column of $\boldsymbol{Z}$, $\boldsymbol{z}_1$, and the first singular value, $\lambda_1 = 1.496896$, give the best rank-1 approximation of $\boldsymbol{X}$,

$$\begin{aligned} \boldsymbol{A}_1 &= \lambda_1^{1/2} \boldsymbol{u}_1 \boldsymbol{z}_1' \\ &= (1.4969) \begin{pmatrix} -.1140 \\ .2520 \\ .0076 \\ -.0281 \\ -.7354 \\ .6179 \end{pmatrix} (.5950 \quad .4518 \quad .0049 \quad .6647) \\ &= \begin{bmatrix} -.101535 & -.077091 & -.000843 & -.113423 \\ .224434 & .170403 & .001864 & .250712 \\ .006752 & .005126 & .000056 & .007542 \\ -.024999 & -.018981 & -.000208 & -.027927 \\ -.655029 & -.497335 & -.005440 & -.731725 \\ .550378 & .417877 & .004571 & .614820 \end{bmatrix}. \end{aligned}$$

The goodness of fit of $\boldsymbol{A}_1$ to $\boldsymbol{X}$ is measured by

$$\frac{\lambda_1}{\sum \lambda_i} = \frac{(1.4969)^2}{4} = .56$$

or the sum of squares of the differences between the elements of $\boldsymbol{X}$ and $\boldsymbol{A}_1$, the lack of fit, is 44% of the total sum of squares of the elements in $\boldsymbol{X}$. This is not a very good approximation.

The rank-2 approximation to $\boldsymbol{X}$ is obtained by *adding* to $\boldsymbol{A}_1$ the matrix $\boldsymbol{A}_2 = \lambda_2^{1/2} \boldsymbol{u}_2 \boldsymbol{z}_2'$. This gives

$$\boldsymbol{A}_1 + \boldsymbol{A}_2 = \begin{bmatrix} .027725 & -.285040 & .295197 & -.089995 \\ .520490 & -.305880 & .678911 & .304370 \\ -.120122 & .209236 & -.290091 & -.015453 \\ -.013380 & -.037673 & .026363 & -.025821 \\ -.753317 & -.339213 & -.230214 & -.749539 \\ .338605 & .758568 & -.479730 & .576438 \end{bmatrix},$$

which has goodness of fit

$$\frac{\lambda_1 + \lambda_2}{\sum \lambda_i} = \frac{(1.4969)^2 + (1.2449)^2}{4} = .95.$$

In terms of approximating $\boldsymbol{X}$ with the rank-2 matrix $\boldsymbol{A}_1+\boldsymbol{A}_2$, the goodness of fit of .95 means that the sum of squares of the discrepancies between $\boldsymbol{X}$ and $(\boldsymbol{A}_1+\boldsymbol{A}_2)$ is 5% of the total sum of squares of all elements in $\boldsymbol{X}$. The sum of squares of all elements in $\boldsymbol{X}$ is $\sum\lambda_i$, the sum of squares of all elements in $(\boldsymbol{A}_1+\boldsymbol{A}_2)$ is $(\lambda_1+\lambda_2)$, and the sum of squares of all elements in $[\boldsymbol{X}-(\boldsymbol{A}_1+\boldsymbol{A}_2)]$ is $(\lambda_3+\lambda_4)$. In terms of the geometry of the data vectors, the goodness of fit of .95 means that 95% of the dispersion of the "cloud" of points in the original four-dimensional space is, in reality, contained in two dimensions, or the points in four-dimensional space very nearly fall on a plane. Only 5% of the dispersion is lost if the third and fourth dimensions are ignored.

Using all four singular values and their singular vectors gives the complete decomposition of $\boldsymbol{X}$ into four orthogonal rank-1 matrices. The sum of the four matrices equals $\boldsymbol{X}$, within the limits of rounding error. The analysis has shown, by the relatively small size of the third and fourth singular values, that the last two dimensions contain little of the dispersion and can safely be ignored in interpretation of the data. ■

Principal Component Analysis

The singular value decomposition is the first step in **principal component analysis**. Using the result $\boldsymbol{X}=\boldsymbol{U}\boldsymbol{L}^{1/2}\boldsymbol{Z}'$ and the property that $\boldsymbol{Z}'\boldsymbol{Z}=\boldsymbol{I}$, one can define the $n\times p$ matrix $\boldsymbol{W}$ as

$$\boldsymbol{W} \quad = \quad \boldsymbol{X}\boldsymbol{Z}=\boldsymbol{U}\boldsymbol{L}^{1/2}. \tag{2.23}$$

The first column of $\boldsymbol{Z}$ is the first of the right singular vectors of $\boldsymbol{X}$, or the first eigenvector of $\boldsymbol{X}'\boldsymbol{X}$. Thus, the coefficients in the first eigenvector define the particular linear function of the columns of $\boldsymbol{X}$ (of the original variables) that generates the first column of $\boldsymbol{W}$. The second column of $\boldsymbol{W}$ is obtained using the second eigenvector of $\boldsymbol{X}'\boldsymbol{X}$, and so on. Notice that $\boldsymbol{W}'\boldsymbol{W}=\boldsymbol{L}$. Thus, $\boldsymbol{W}$ is an $n\times p$ matrix that, unlike $\boldsymbol{X}$, has the property that all its columns are orthogonal. ($\boldsymbol{L}$ is a diagonal matrix so that all off-diagonal elements, the sums of products between columns of $\boldsymbol{W}$, are zero.) The sum of squares of the ith column of $\boldsymbol{W}$ is λ_i, the ith diagonal element of $\boldsymbol{L}$. Thus, if $\boldsymbol{X}$ is an $n\times p$ matrix of observations on p variables, each column of $\boldsymbol{W}$ is a new variable defined as a linear transformation of the original variables. The ith new variable has sum of squares λ_i and all are pairwise orthogonal. This analysis is called the **principal component analysis** of $\boldsymbol{X}$, and the columns of $\boldsymbol{W}$ are the **principal components** (sometimes called principal component scores).

Principal component analysis is used where the columns of $\boldsymbol{X}$ correspond to the observations on different variables. The transformation is to a set of orthogonal variables such that the first principal component accounts for the largest possible amount of the total dispersion, measured by λ_1, the second principal component accounts for the largest possible amount of the remaining dispersion λ_2, and so forth. The total dispersion is given by the

sum of all eigenvalues, which is equal to the sum of squares of the original variables; $\text{tr}(\boldsymbol{X}'\boldsymbol{X}) = \text{tr}(\boldsymbol{W}'\boldsymbol{W}) = \sum \lambda_i$.

Example 2.14

For the Saeed and Francis data, Example 2.13, each column of $\boldsymbol{Z}$ contains the coefficients that define one of the principal components as a linear function of the original variables. The first vector in $\boldsymbol{Z}$,

$$\boldsymbol{z}_1 = (\ .5950 \quad .4518 \quad .0049 \quad .6647\ )',$$

has similar first, second, and fourth coefficients with the third coefficient being near zero. Thus, the first principal component is essentially an average of the three temperature variables X_1, X_2, and X_4. The second column vector in $\boldsymbol{Z}$,

$$\boldsymbol{z}_2 = (\ .3361 \quad -.5408 \quad .7687 \quad .0609\ )',$$

gives heavy positive weight to X_3, heavy negative weight to X_2, and moderate positive weight to X_1. Thus, the second principal component will be large for those observations that have high rainfall X_3, and small difference between the maximum and minimum daily temperatures X_2 and X_1.

The third and fourth principal components account for only 5% of the total dispersion. This small amount of dispersion may be due more to random "noise" than to real patterns in the data. Consequently, the interpretation of these components may not be very meaningful. The third principal component will be large when there is high rainfall and large difference between the maximum and minimum daily temperatures,

$$\boldsymbol{z}_3 = (\ -.3832 \quad .6580 \quad .6391 \quad -.1089\ )'.$$

The variable degree days X_4 has little involvement in the second and third principal components; the fourth coefficient is relatively small. The fourth principal component is determined primarily by the difference between an average minimum daily temperature and degree days,

$$\boldsymbol{z}_4 = (\ .6214 \quad .2657 \quad -.0265 \quad -.7366\ )'. \qquad \blacksquare$$

The principal component vectors are obtained either by the multiplication $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{L}^{1/2}$ or $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{Z}$. The first is easier since it is the simple scalar multiplication of each column of $\boldsymbol{U}$ with the appropriate $\lambda_i^{1/2}$.

Example 2.15

The principal component vectors for the Saeed and Francis data of Ex-

ample 2.13 are (with some rounding)

$$\boldsymbol{W} = \begin{bmatrix} -.1706 & .3846 & -.3681 & .0151 \\ .3772 & .8808 & .1543 & -.0185 \\ .0113 & -.3775 & .1260 & .0329 \\ -.0420 & .0346 & .1483 & .0207 \\ -1.1008 & -.2924 & .0298 & -.0279 \\ .9250 & -.6300 & -.0902 & -.0223 \end{bmatrix}.$$

The sum of squares of the first principal component, the first column of $\boldsymbol{W}$, is $\lambda_1 = (1.4969)^2 = 2.2407$. Similarly, the sums of squares for the second, third, and fourth principal components are

$$\begin{aligned} \lambda_2 &= (1.2449)^2 = 1.5498 \\ \lambda_3 &= (.4541)^2 = .2062 \\ \lambda_4 &= (.0579)^2 = .0034. \end{aligned}$$

These sum to 4.0, the total sum of squares of the original three variables after they were standardized. The proportion of the total sum of squares accounted for by the first principal component is $\lambda_1 / \sum \lambda_i = 2.2407/4 = .56$ or 56%. The first two principal components account for $(\lambda_1 + \lambda_2)/4 = 3.79/4 = .95$ or 95% of the total sum of squares of the four original variables.

Each of the original data vectors in $\boldsymbol{X}$ was a vector in six-dimensional space and, together, the four vectors defined a four-dimensional subspace. These vectors were not orthogonal. The four vectors in $\boldsymbol{W}$, the principal component vectors, are linear functions of the original vectors and, as such, they fall in the *same* four-dimensional subspace. The principal component vectors, however, are orthogonal and defined such that the first principal component vector has the largest possible sum of squares. This means that the direction of the first principal component axis coincides with the major axis of the elipsoid of observations, Figure 2.3. Note that the "cloud" of observations, the data points, does *not* change; only the axes are being redefined. The second principal component has the largest possible sum of squares of all vectors orthogonal to the first, and so on. The fact that the first two principal components account for 95% of the sum of squares in this example shows that very little of the dispersion among the data points occurs in the third and fourth principal component dimensions. In other words, the variability among these six locations in average minimum and average maximum temperature, total rainfall, and total growing degree days, can be adequately described by considering *only* the two dimensions (or variables) defined by the first two principal components.

The plot of the first two principal components from the Saeed and Francis data, Figure 2.3, shows that locations 5 and 6 differ from each other primarily in the first principal component. This component was noted earlier to be mainly a temperature difference; location 6 is the warmer and has

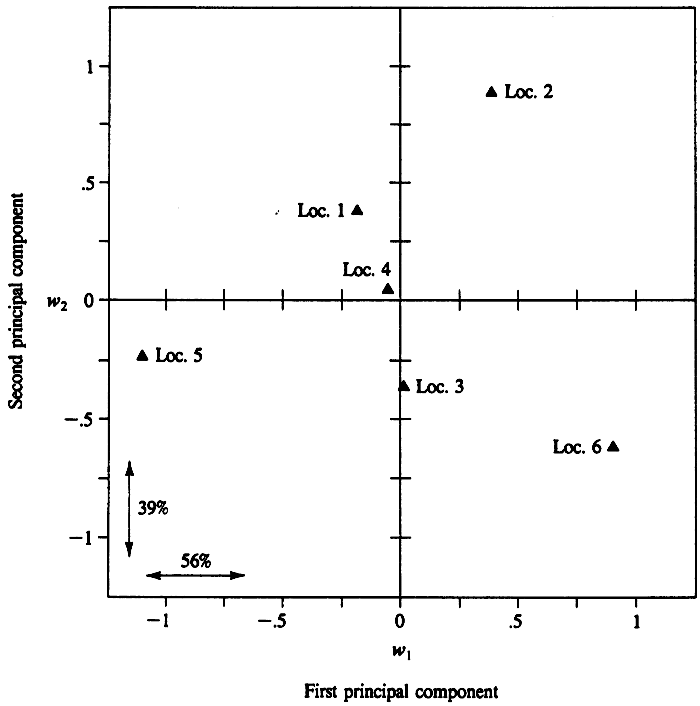

FIGURE 2.3. *The first two principal components of the Saeed and Francis (1984) data on average minimum temperature, average maximum temperature, total rainfall, and growing degree days for six locations. The first principal component primarily reflects average temperature. The second principal component is a measure of rainfall minus the spread between minimum and maximum temperature.*

the longer growing season. The other four locations differ primarily in the second principal component which reflects amount of rainfall and the difference in maximum and minimum temperature. Location 2 has the highest rainfall and tends to have a large difference in maximum and minimum daily temperature. Location 6 is also the lowest in the second principal component indicating a lower rainfall and small difference between the maximum and minimum temperature. Thus, location 6 appears to be a relatively hot, dry environment with somewhat limited diurnal temperature variation. ■

2.9 Summary

This chapter has presented the key matrix operations that are used in this text. The student must be able to use matrix notation and matrix operations. Of particular importance are

- the concepts of rank and the transpose of a matrix;
- the special types of matrices: square, symmetric, diagonal, identity, and idempotent;
- the elementary matrix operations of addition and multiplication; and
- the use of the inverse of a square symmetric matrix to solve a set of equations.

The geometry of vectors and projections is useful in understanding least squares principles. Eigenanalysis and singular value decomposition are used later in the text.

2.10 Exercises

2.1. Let

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 2 & 4 \\ -1 & 2 \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 3 & -4 \end{bmatrix},$$

$$\boldsymbol{c}' = \begin{pmatrix} 1 & 2 & 0 \end{pmatrix}, \quad \text{and} \quad d = 2, \text{ a scalar.}$$

Perform the following operations, if possible. If the operation is not possible, explain why.

(a) $\boldsymbol{c}'\boldsymbol{A}$

(b) $\boldsymbol{A}'\boldsymbol{c}$

(c) $\boldsymbol{B}' + \boldsymbol{A}$

(d) $\boldsymbol{c}'\boldsymbol{B}$

(e) $\boldsymbol{A} - d$

(f) $(d\boldsymbol{B}' + \boldsymbol{A})$.

2.2. Find the rank of each of the following matrices. Which matrices are of full rank?

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \boldsymbol{B} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{C} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix}.$$

2.3. Use $\boldsymbol{B}$ in Exercise 2.2 to compute $\boldsymbol{D} = \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'$. Determine whether $\boldsymbol{D}$ is idempotent. What is the rank of $\boldsymbol{D}$?

2.4. Find a_{ij} elements to make the following matrix symmetric. Can you choose a_{33} to make the matrix idempotent?

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & a_{13} & 4 \\ 2 & -1 & 0 & a_{24} \\ 6 & 0 & a_{33} & -2 \\ a_{41} & 8 & -2 & 3 \end{bmatrix}.$$

2.5. Verify that $\boldsymbol{A}$ and $\boldsymbol{B}$ are inverses of each other.

$$\boldsymbol{A} = \begin{bmatrix} 10 & 5 \\ 3 & 2 \end{bmatrix} \quad \boldsymbol{B} = \begin{bmatrix} \frac{2}{5} & -1 \\ -\frac{3}{5} & 2 \end{bmatrix}.$$

2.6. Find b_{41} such that $\boldsymbol{a}$ and $\boldsymbol{b}$ are orthogonal.

$$\boldsymbol{a} = \begin{pmatrix} 2 \\ 0 \\ -1 \\ 3 \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} 6 \\ -1 \\ 3 \\ b_{41} \end{pmatrix}.$$

2.7. Plot the following vectors on a two-dimensional coordinate system.

$$\boldsymbol{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \boldsymbol{v}_2 = \begin{pmatrix} 4 \\ 1 \end{pmatrix} \quad \boldsymbol{v}_3 = \begin{pmatrix} 1 \\ -4 \end{pmatrix}.$$

By inspection of the plot, which pairs of vectors appear to be orthogonal? Verify numerically that they are orthogonal *and* that all other

pairs in this set are *not* orthogonal. Explain from the geometry of the plot how you know there is a linear dependency among the three vectors.

2.8. The three vectors in Exercise 2.7 are linearly dependent. Find the linear function of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ that equals $\boldsymbol{v}_3$. Set the problem up as a system of linear equations to be solved. Let $\boldsymbol{V} = (\,\boldsymbol{v}_1 \quad \boldsymbol{v}_2\,)$, and let $\boldsymbol{x}' = (\,x_1 \quad x_2\,)$ be the vector of unknown coefficients. Then, $\boldsymbol{V}\boldsymbol{x} = \boldsymbol{v}_3$ is the system of equations to be solved for $\boldsymbol{x}$.

(a) Show that the system of equations is consistent.

(b) Show that there is a unique solution.

(c) Find the solution.

2.9. Expand the set of vectors in Exercise 2.7 to include a fourth vector, $\boldsymbol{v}_4' = (\,8 \quad 5\,)$. Reformulate Exercise 2.8 to include the fourth vector by including $\boldsymbol{v}_4$ in $\boldsymbol{V}$and an additional coefficient in $\boldsymbol{x}$. Is this system of equations consistent? Is the solution unique? Find a solution. If solutions are not unique, find another solution.

2.10. Use the determinant to determine which of the following matrices has a unique inverse.

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 4 & 10 \end{bmatrix} \quad \boldsymbol{B} = \begin{bmatrix} 4 & -1 \\ 0 & 6 \end{bmatrix} \quad \boldsymbol{C} = \begin{bmatrix} 6 & 3 \\ 4 & 2 \end{bmatrix}.$$

2.11. Given the following matrix,

$$\boldsymbol{A} = \begin{bmatrix} 3 & \sqrt{2} \\ \sqrt{2} & 2 \end{bmatrix},$$

(a) find the eigenvalues and eigenvectors of $\boldsymbol{A}$.

(b) What do your findings tell you about the rank of $\boldsymbol{A}$?

2.12. Given the following eigenvalues with their corresponding eigenvectors, and knowing that the original matrix was square and symmetric, reconstruct the original matrix.

$$\begin{aligned} \lambda_1 &= 6 \quad \boldsymbol{z}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ \lambda_2 &= 2 \quad \boldsymbol{z}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \end{aligned}$$

2.13. Find the inverse of the following matrix,

$$\boldsymbol{A} = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 10 & 2 \\ 0 & 2 & 3 \end{bmatrix}.$$

2.14. Let

$$\boldsymbol{X} = \begin{bmatrix} 1 & .2 & 0 \\ 1 & .4 & 0 \\ 1 & .6 & 0 \\ 1 & .8 & 0 \\ 1 & .2 & .1 \\ 1 & .4 & .1 \\ 1 & .6 & .1 \\ 1 & .8 & .1 \end{bmatrix} \qquad \boldsymbol{Y} = \begin{pmatrix} 242 \\ 240 \\ 236 \\ 230 \\ 239 \\ 238 \\ 231 \\ 226 \end{pmatrix}.$$

(a) Compute $\boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{X}'\boldsymbol{Y}$. Verify by separate calculations that the $(i, j) = (2, 2)$ element in $\boldsymbol{X}'\boldsymbol{X}$ is the sum of squares of column 2 in $\boldsymbol{X}$. Verify that the $(2, 3)$ element is the sum of products between columns 2 and 3 of $\boldsymbol{X}$. Identify the elements in $\boldsymbol{X}'\boldsymbol{Y}$ in terms of sums of squares or products of the columns of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

(b) Is $\boldsymbol{X}$ of full column rank? What is the rank of $\boldsymbol{X}'\boldsymbol{X}$?

(c) Obtain $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. What is the rank of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$? Verify by matrix multiplication that $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}$.

(d) Compute $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and verify by matrix multiplication that $\boldsymbol{P}$ is idempotent. Compute the trace $\mathrm{tr}(\boldsymbol{P})$. What is $r(\boldsymbol{P})$?

2.15. Use $\boldsymbol{X}$ as defined in Exercise 2.14.

(a) Find the singular value decomposition of $\boldsymbol{X}$. Explain what the singular values tell you about the rank of $\boldsymbol{X}$.

(b) Compute the rank-1 approximation of $\boldsymbol{X}$; call it $\boldsymbol{A}_1$. Use the singular values to state the "goodness of fit" of this rank-1 approximation.

(c) Use $\boldsymbol{A}_1$ to compute a rank-1 approximation of $\boldsymbol{X}'\boldsymbol{X}$; that is, compute $\boldsymbol{A}_1'\boldsymbol{A}_1$. Compare $\mathrm{tr}(\boldsymbol{A}_1'\boldsymbol{A}_1)$ with λ_1 and $\mathrm{tr}(\boldsymbol{X}'\boldsymbol{X})$.

2.16. Use $\boldsymbol{X}'\boldsymbol{X}$ as computed in Exercise 2.14.

(a) Compute the eigenanalysis of $\boldsymbol{X}'\boldsymbol{X}$. What is the relationship between the singular values of $\boldsymbol{X}$ obtained in Exercise 2.15 and the eigenvalues obtained for $\boldsymbol{X}'\boldsymbol{X}$?

(b) Use the results of the eigenanalysis to compute the rank-1 approximation of $\boldsymbol{X}'\boldsymbol{X}$. Compare this result to the approximation of $\boldsymbol{X}'\boldsymbol{X}$ obtained in Exercise 2.15.

(c) Show algebraically that they should be identical.

2.17. Verify that

$$\boldsymbol{A} = \frac{1}{15}\begin{bmatrix} 3 & -13 & 8 \\ 12 & -7 & 2 \\ -12 & 17 & -7 \end{bmatrix}$$

is the inverse of

$$\boldsymbol{B} = \begin{bmatrix} 1 & 3 & 2 \\ 4 & 5 & 6 \\ 8 & 7 & 9 \end{bmatrix}.$$

2.18. Show that the equations $\boldsymbol{Ax} = \boldsymbol{y}$ are consistent where

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 3 & 3 \\ 5 & 7 \end{bmatrix} \quad \text{and} \quad \boldsymbol{y} = \begin{pmatrix} 6 \\ 9 \\ 21 \end{pmatrix}.$$

2.19. Verify that

$$\boldsymbol{A}^{-} = \frac{1}{18}\begin{bmatrix} -10 & 16 & -4 \\ 8 & -11 & 5 \end{bmatrix}$$

is a generalized inverse of

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 3 & 3 \\ 5 & 7 \end{bmatrix}.$$

2.20. Verify that

$$\boldsymbol{A}^{-} = \begin{bmatrix} -\frac{1}{10} & -\frac{2}{10} & \frac{4}{9} \\ 0 & 0 & \frac{1}{9} \\ \frac{1}{10} & \frac{2}{10} & -\frac{2}{9} \end{bmatrix}$$

is a generalized inverse of

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 3 & 3 \end{bmatrix}.$$

2.21. Use the generalized inverse in Exercise 2.20 to obtain a solution to the equations $\boldsymbol{Ax} = \boldsymbol{y}$, where $\boldsymbol{A}$ is defined in Exercise 2.20 and $\boldsymbol{y} = (\,6 \quad 12 \quad 9\,)'$. Verify that the solution you obtained satisfies $\boldsymbol{Ax} = \boldsymbol{y}$.

2.22. The eigenanalysis of

$$\boldsymbol{A} = \begin{bmatrix} 10 & 3 \\ 3 & 8 \end{bmatrix}$$

in Section 2.7 gave

$$\boldsymbol{A}_1 = \begin{bmatrix} 8.0042 & 5.7691 \\ 5.7691 & 4.1581 \end{bmatrix} \quad \text{and} \quad \boldsymbol{A}_2 = \begin{bmatrix} 1.9958 & -2.7691 \\ -2.7691 & 3.8419 \end{bmatrix}.$$

Verify the multiplication of the eigenvectors to obtain $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$. Verify that $\boldsymbol{A}_1 + \boldsymbol{A}_2 = \boldsymbol{A}$, and that $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are orthogonal to each other.

2.23. In Section 2.6, a linear transformation of $\boldsymbol{y}_1 = (\,3 \quad 10 \quad 20\,)'$ to $\boldsymbol{x}_1 = (\,33 \quad 17 \quad -3\,)'$ and of $\boldsymbol{y}_2 = (\,6 \quad 14 \quad 21\,)'$ to $\boldsymbol{x}_2 = (\,41 \quad 15 \quad 1\,)'$ was made using the matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & 2 & -1 \end{bmatrix}.$$

The vectors of $\boldsymbol{A}$ were then standardized so that $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}$ to produce the *orthogonal* transformation of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ to

$$\boldsymbol{x}_1^* = (\,33/\sqrt{3} \quad 17/\sqrt{2} \quad -3/\sqrt{6}\,)'$$

and

$$\boldsymbol{x}_2^* = (\,41/\sqrt{3} \quad 15/\sqrt{2} \quad 1/\sqrt{6}\,)',$$

respectively. Show that the squared distance between $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ is unchanged when the orthogonal transformation is made but not when the nonorthogonal transformation is made. That is, show that

$$(\boldsymbol{y}_1 - \boldsymbol{y}_2)'(\boldsymbol{y}_1 - \boldsymbol{y}_2) = (\boldsymbol{x}_1^* - \boldsymbol{x}_2^*)'(\boldsymbol{x}_1^* - \boldsymbol{x}_2^*)$$

but that

$$(\boldsymbol{y}_1 - \boldsymbol{y}_2)'(\boldsymbol{y}_1 - \boldsymbol{y}_2) \neq (\boldsymbol{x}_1 - \boldsymbol{x}_2)'(\boldsymbol{x}_1 - \boldsymbol{x}_2).$$

2.24. (a) Let $\boldsymbol{A}$ be an $m \times n$ matrix and $\boldsymbol{B}$ be an $n \times m$ matrix. Then show that $\text{tr}(\boldsymbol{AB}) = \text{tr}(\boldsymbol{BA})$.
(b) Use (a) to show that $\text{tr}(\boldsymbol{ABC}) = \text{tr}(\boldsymbol{BCA})$, where $\boldsymbol{C}$ is an $m \times m$ matrix.

2.25. Let $\boldsymbol{a}^*$ be an $m \times 1$ vector with $\boldsymbol{a}^{*\prime}\boldsymbol{a}^* > 0$. Define $\boldsymbol{a} = \boldsymbol{a}^*/(\boldsymbol{a}^{*\prime}\boldsymbol{a}^*)^{1/2}$ and $\boldsymbol{A} = \boldsymbol{a}\boldsymbol{a}'$. Show that $\boldsymbol{A}$ is a symmetric idempotent matrix of rank 1.

2.26. Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two $m \times 1$ vectors that are orthogonal to each other. Define $\boldsymbol{A} = \boldsymbol{a}\boldsymbol{a}'$ and $\boldsymbol{B} = \boldsymbol{b}\boldsymbol{b}'$. Show that $\boldsymbol{AB} = \boldsymbol{BA} = \boldsymbol{0}$, a matrix of zeros.

2.27. **Gram–Schmidt orthogonalization**. An orthogonal basis for a space spanned by some vectors can be obtained using the Gram–Schmidt orthogonalization procedure.

(a) Consider two linearly independent vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. Define $\boldsymbol{z}_1 = \boldsymbol{v}_1$ and $\boldsymbol{z}_2 = \boldsymbol{v}_2 - \boldsymbol{v}_1 c_{2.1}$, where $c_{2.1} = (\boldsymbol{v}_1'\boldsymbol{v}_2)/(\boldsymbol{v}_1'\boldsymbol{v}_1)$. Show that $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are orthogonal. Also, show that $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ span the same space as $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$.

(b) Consider three linearly independent vectors $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, and $\boldsymbol{v}_3$. Define $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ as in (a) and $\boldsymbol{z}_3 = \boldsymbol{v}_3 - c_{3.1}\boldsymbol{z}_1 - c_{3.2}\boldsymbol{z}_2$, where $c_{3.i} = (\boldsymbol{z}_i'\boldsymbol{v}_3)/(\boldsymbol{z}_i'\boldsymbol{z}_i)$, $i = 1,\ 2$. Show that $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, and $\boldsymbol{z}_3$ are mutually orthogonal and span the same space as $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, and $\boldsymbol{v}_3$.

3

MULTIPLE REGRESSION IN MATRIX NOTATION

We have reviewed linear regression in algebraic notation and have introduced the matrix notation and operations needed to continue with the more complicated models.

This chapter presents the model, and develops the normal equations and solution to the normal equations for a general linear model involving any number of independent variables. The matrix formulation for the variances of linear functions is used to derive the measures of precision of the estimates.

Chapter 1 provided an introduction to multiple regression and suggested that a more convenient notation was needed. Chapter 2 familiarized you with matrix notation and operations with matrices. This chapter states multiple regression results in matrix notation. Developments in the chapter are for full rank models. Less than full rank models that use generalized inverses are discussed in Chapter 9.

3.1 The Model

The linear additive model for relating a dependent variable to p independent variables is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i. \tag{3.1}$$

The subscript i denotes the observational unit from which the observations on Y and the p independent variables were taken. The second subscript designates the independent variable. The sample size is denoted with $n, i = 1, \ldots, n$, and p denotes the number of independent variables. There are $(p+1)$ parameters $\beta_j, j = 0, \ldots, p$ to be estimated when the linear model includes the intercept β_0. For convenience, we use $p' = (p+1)$. In this book we assume that $n > p'$. Four matrices are needed to express the linear model in matrix notation:

Matrix Definitions

$\boldsymbol{Y}$: the $n \times 1$ column vector of observations on the dependent variable Y_i;

$\boldsymbol{X}$: the $n \times p'$ matrix consisting of a column of ones, which is labeled $\mathbf{1}$, followed by the p column vectors of the observations on the independent variables;

$\boldsymbol{\beta}$: the $p' \times 1$ vector of parameters to be estimated; and

$\boldsymbol{\epsilon}$: the $n \times 1$ vector of random errors.

With these definitions, the linear model can be written as

$$\boldsymbol{Y} \quad = \quad \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.2}$$

or

$$\underset{(n\times 1)}{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}} = \underset{(n\times p')}{\begin{bmatrix} 1 & X_{11} & X_{12} & X_{13} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & X_{23} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n3} & \cdots & X_{np} \end{bmatrix}} \underset{(p'\times 1)}{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}} + \underset{(n\times 1)}{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}.$$

The $\boldsymbol{X}$ Matrix

Each column of $\boldsymbol{X}$ contains the values for a particular independent variable. The elements of a particular row of $\boldsymbol{X}$, say row r, are the coefficients on the corresponding parameters in $\boldsymbol{\beta}$ that give $\mathcal{E}(Y_r)$. Notice that β_0 has the constant multiplier 1 for all observations; hence, the column vector $\mathbf{1}$ is the first column of $\boldsymbol{X}$. Multiplying the first row of $\boldsymbol{X}$ by $\boldsymbol{\beta}$, and adding the first element of $\boldsymbol{\epsilon}$ confirms that the model for the first observation is

$$Y_1 \quad = \quad \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_p X_{1p} + \epsilon_1.$$

The vectors $\boldsymbol{Y}$ and $\boldsymbol{\epsilon}$ are random vectors; the elements of these vectors are random variables. The matrix $\boldsymbol{X}$ is considered to be a matrix of known constants. A model for which $\boldsymbol{X}$ is of full column rank is called a **full-rank** model.

The $\boldsymbol{\beta}$ Vector

The vector $\boldsymbol{\beta}$ is a vector of unknown constants to be estimated from the data. Each element β_j is a partial regression coefficient reflecting the change in the dependent variable per unit change in the jth independent variable,

assuming all other independent variables are held constant. The definition of each partial regression coefficient is dependent on the set of independent variables in the model. Whenever clarity demands, the subscript notation on β_j is expanded to identify explicitly both the independent variable to which the coefficient applies and the other independent variables in the model. For example, $\beta_{2.13}$ would designate the partial regression coefficient for X_2 in a model that contains X_1, X_2, and X_3.

The Random Vector $\boldsymbol{\epsilon}$

It is common to assume that ϵ_i are independent and identically distributed as normal random variables with mean zero and variance σ^2. Since ϵ_i are assumed to be independent of each other, the covariance between ϵ_i and ϵ_j is zero for any $i \neq j$. The joint probability density function of $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ is

$$\prod_{i=1}^{n}[(2\pi)^{-1/2}\sigma^{-1}\, e^{-\epsilon_i^2/2\sigma^2}] \quad = \quad (2\pi)^{-n/2}\sigma^{-n}\, e^{-\sum_{i=1}^{n}\epsilon_i^2/2\sigma^2}. \quad (3.3)$$

The random vector $\boldsymbol{\epsilon}$ is a vector $(\,\epsilon_1 \quad \epsilon_2 \quad \cdots \quad \epsilon_n\,)'$ consisting of random variables ϵ_i.

The $\boldsymbol{Y}$ Vector

Since the elements of $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are assumed to be constants, the $\boldsymbol{X\beta}$ term in the model is a vector of constants. Thus, $\boldsymbol{Y}$ is a random vector that is the sum of the constant vector $\boldsymbol{X\beta}$ and the random vector $\boldsymbol{\epsilon}$. Since ϵ_i are assumed to be independent $N(0, \sigma^2)$ random variables, we have that

1. Y_i is a normal random variable with mean $\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$ and variance σ^2;
2. Y_i are independent of each other.

The covariance between Y_i and Y_j is zero for $i \neq j$. The joint probability density function of $Y_1, \ldots, Y_n$ is

$$(2\pi)^{-n/2}\sigma^{-n}\, e^{-\sum[Y_i-(\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip})]^2/2\sigma^2}. \quad (3.4)$$

Importance of Normality Assumption

The conventional tests of hypotheses and confidence interval estimates of the parameters are based on the assumption that the estimates are normally distributed. Thus, the assumption of normality of the ϵ_i is critical for these purposes. However, normality is not required for least squares estimation. Even in the absence of normality, the least squares estimates are the best linear unbiased estimates (b.l.u.e.). They are best in the sense of having minimum variance among all linear unbiased estimators. If normality does hold, the maximum likelihood estimators are derived using the criterion of finding those values of the parameters that would have maximized the probability of obtaining the particular sample, called the likelihood function. Maximizing the likelihood function in equation 3.4 with respect to $\boldsymbol{\beta} = (\,\beta_0 \quad \beta_1 \quad \cdots \quad \beta_p\,)'$ is equivalent to minimizing the sum of squares in the exponent, and hence the least squares estimates coincide

with maximum likelihood estimates. The reader is referred to statistical theory texts such as Searle (1971), Graybill (1961), and Cramér (1946) for further discussion of maximum likelihood estimation.

Example 3.1

For the ozone data used in Example 1.1 (see Table 1.1 on page 5),

$$\boldsymbol{X} = \begin{bmatrix} 1 & .02 \\ 1 & .07 \\ 1 & .11 \\ 1 & .15 \end{bmatrix} \qquad \boldsymbol{Y} = \begin{pmatrix} 242 \\ 237 \\ 231 \\ 201 \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

and $\boldsymbol{\epsilon}$ is the vector of four (unobservable) random errors. ■

3.2 The Normal Equations and Their Solution

In matrix notation, the normal equations are written as

$$\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}. \tag{3.5}$$

The normal equations are always consistent and hence will always have a solution of the form

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y}. \tag{3.6}$$

If $\boldsymbol{X}'\boldsymbol{X}$ has an inverse, then the normal equations have a unique solution given by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y}). \tag{3.7}$$

$\boldsymbol{X}'\boldsymbol{X}$

The multiplication $\boldsymbol{X}'\boldsymbol{X}$ generates a $p' \times p'$ matrix where the diagonal elements are the sums of squares of each of the independent variables and the off-diagonal elements are the sums of products between independent variables. The general form of $\boldsymbol{X}'\boldsymbol{X}$ is

$$\begin{bmatrix} n & \sum X_{i1} & \sum X_{i2} & \cdots & \sum X_{ip} \\ \sum X_{i1} & \sum X_{i1}^2 & \sum X_{i1}X_{i2} & \cdots & \sum X_{i1}X_{ip} \\ \sum X_{i2} & \sum X_{i1}X_{i2} & \sum X_{i2}^2 & \cdots & \sum X_{i2}X_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum X_{ip} & \sum X_{i1}X_{ip} & \sum X_{i2}X_{ip} & \cdots & \sum X_{ip}^2 \end{bmatrix}. \tag{3.8}$$

Summation in all cases is over $i = 1$ to n, the n observations in the data. When only one independent variable is involved, $\boldsymbol{X}'\boldsymbol{X}$ consists of only the upper-left 2×2 matrix. Inspection of the normal equations in Chapter 1, equation 1.6, reveals that the elements in this 2×2 matrix are the coefficients on $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

The elements of the matrix product $\boldsymbol{X}'\boldsymbol{Y}$ are the sums of products between each independent variable in turn and the dependent variable: **$\boldsymbol{X}'\boldsymbol{Y}$**

$$\boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} \sum Y_i \\ \sum X_{i1}Y_i \\ \sum X_{i2}Y_i \\ \vdots \\ \sum X_{ip}Y_i \end{pmatrix}. \tag{3.9}$$

The first element $\sum Y_i$ is the sum of products between the vector of ones (the first column of $\boldsymbol{X}$) and $\boldsymbol{Y}$. Again, if only one independent variable is involved, $\boldsymbol{X}'\boldsymbol{Y}$ consists of only the first two elements. The reader can verify that these are the right-hand sides of the two normal equations, equation 1.6.

A Unique Solution

The unique solution to the normal equations exists only if the inverse of $\boldsymbol{X}'\boldsymbol{X}$ exists. This, in turn, requires that the matrix $\boldsymbol{X}$ be of full column rank; that is, there can be no linear dependencies among the independent variables. The practical implication is that there can be no redundancies in the information contained in $\boldsymbol{X}$. For example, the amount of nitrogen in a diet is sometimes converted to the amount of protein by multiplication by a constant. Because the same information is reported two ways, a linear dependency occurs if both are included in $\boldsymbol{X}$. Suppose the independent variables in a genetics problem include three variables reporting the observed sample frequencies of three possible alleles (for a particular locus). These three variables, and the **1** vector, create a linear dependency since the sum of the three variables, the sum of the allelic frequencies, must be 1.0. Only two of the allelic frequencies need be reported; the third is redundant since it can be computed from the first two and the column of ones.

It is always possible to rewrite the model such that the redundancies among the independent variables are eliminated and the corresponding $\boldsymbol{X}$ matrix is of full rank. In this chapter, $\boldsymbol{X}$ is assumed to be of full column rank. The case where $\boldsymbol{X}$ is not of full rank is discussed in Chapter 9.

Example 3.2

Matrix operations using $\boldsymbol{X}$ and $\boldsymbol{Y}$ from the ozone example, Example 1.1,

give

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 4 & .3500 \\ .3500 & .0399 \end{bmatrix} \qquad \boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} 911 \\ 76.99 \end{pmatrix}$$

and

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 1.07547 & -9.43396 \\ -9.43396 & 107.81671 \end{bmatrix}.$$

The estimates of the regression coefficients are

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} 253.434 \\ -293.531 \end{pmatrix}.$$

■

3.3 The $\widehat{Y}$ and Residuals Vectors

$\widehat{\boldsymbol{Y}}$ and $\boldsymbol{P}$

The vector of estimated means of the dependent variable $\boldsymbol{Y}$ for the values of the independent variables in the data set is computed as

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}. \tag{3.10}$$

This is the simplest way to compute $\widehat{\boldsymbol{Y}}$. It is useful for later results, however, to express $\widehat{\boldsymbol{Y}}$ as a linear function of $\boldsymbol{Y}$ by substituting $[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}]$ for $\widehat{\boldsymbol{\beta}}$. Thus,

$$\begin{aligned} \widehat{\boldsymbol{Y}} &= [\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{Y} \\ &= \boldsymbol{P}\boldsymbol{Y}. \end{aligned} \tag{3.11}$$

Equation 3.11 defines the matrix $\boldsymbol{P}$, an $n \times n$ matrix determined entirely by the Xs. This matrix plays a particularly important role in regression analysis. It is a symmetric matrix ($\boldsymbol{P}' = \boldsymbol{P}$) that is also idempotent ($\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$), and is therefore a projection matrix (see Section 2.6). Equation 3.11 shows that $\widehat{\boldsymbol{Y}}$ is a linear function of $\boldsymbol{Y}$ with the coefficients given by $\boldsymbol{P}$. (For example, the first row of $\boldsymbol{P}$ contains the coefficients for the linear function of all Y_i that gives $\widehat{\boldsymbol{Y}}_1$.)

Example 3.3

For the Heagle ozone data used in Example 1.1,

$$\begin{aligned} \boldsymbol{P} &= \begin{bmatrix} 1 & .02 \\ 1 & .07 \\ 1 & .11 \\ 1 & .15 \end{bmatrix} \begin{bmatrix} 1.0755 & -9.4340 \\ -9.4340 & 107.8167 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ .02 & .07 & .11 & .15 \end{bmatrix} \\ &= \begin{bmatrix} .741240 & .377358 & .086253 & -.204852 \\ .377358 & .283019 & .207547 & .132075 \\ .086253 & .207547 & .304582 & .401617 \\ -.204852 & .132075 & .401617 & .671159 \end{bmatrix}. \end{aligned}$$

Thus, for example,

$$\widehat{Y}_1 = .741Y_1 + .377Y_2 + .086Y_3 - .205Y_4.$$

■

The residuals vector $\boldsymbol{e}$ reflects the lack of agreement between the observed $\boldsymbol{Y}$ and the estimated $\widehat{\boldsymbol{Y}}$: $\boldsymbol{e}$

$$\boldsymbol{e} \quad = \quad \boldsymbol{Y} - \widehat{\boldsymbol{Y}}. \tag{3.12}$$

As with $\widehat{\boldsymbol{Y}}, \boldsymbol{e}$ can be expressed as a linear function of $\boldsymbol{Y}$ by substituting $\boldsymbol{PY}$ for $\widehat{\boldsymbol{Y}}$:

$$\boldsymbol{e} \quad = \quad \boldsymbol{Y} - \boldsymbol{PY} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}. \tag{3.13}$$

Recall that least squares estimation minimizes the sum of squares of the residuals; $\widehat{\boldsymbol{\beta}}$ has been chosen so that $\boldsymbol{e}'\boldsymbol{e}$ is a minimum. Like $\boldsymbol{P}$, $(\boldsymbol{I} - \boldsymbol{P})$ is symmetric and idempotent.

This has partitioned $\boldsymbol{Y}$ into two parts, that accounted for by the model $\widehat{\boldsymbol{Y}} + \boldsymbol{e}$ $\widehat{\boldsymbol{Y}}$ and the residual $\boldsymbol{e}$. That the two parts are additive is evident from the fact that $\boldsymbol{e}$ was obtained by difference (equation 3.12), or can be demonstrated as follows.

$$\widehat{\boldsymbol{Y}} + \boldsymbol{e} \quad = \quad \boldsymbol{PY} + (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} = (\boldsymbol{P} + \boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} = \boldsymbol{Y}. \tag{3.14}$$

Example 3.4

Continuing with Example 3.3, we obtain

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & .02 \\ 1 & .07 \\ 1 & .11 \\ 1 & .15 \end{bmatrix} \begin{pmatrix} 253.434 \\ -293.531 \end{pmatrix} = \begin{pmatrix} 247.563 \\ 232.887 \\ 221.146 \\ 209.404 \end{pmatrix}.$$

The residuals are

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \begin{pmatrix} -5.563 \\ 4.113 \\ 9.854 \\ -8.404 \end{pmatrix}.$$

The results from the ozone example are summarized in Table 3.1. ■

TABLE 3.1. *Results for the linear regression of soybean yield on levels of ozone.*

$\boldsymbol{X}_i$	$\boldsymbol{Y}_i$	$\widehat{Y}_i$	e_i
0.02	242	247.563	−5.563
0.07	237	232.887	4.113
0.11	231	221.146	9.854
0.15	201	209.404	−8.404

3.4 Properties of Linear Functions of Random Vectors

Note that $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{Y}}$, and $\boldsymbol{e}$ are random vectors because they are functions of the random vector $\boldsymbol{Y}$. In the previous sections, these vectors are expressed as linear functions $\boldsymbol{AY}$ of $\boldsymbol{Y}$. The matrix $\boldsymbol{A}$ is

- $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ for $\widehat{\boldsymbol{\beta}}$,
- $\boldsymbol{P}$ for $\widehat{\boldsymbol{Y}}$, and
- $(\boldsymbol{I}-\boldsymbol{P})$ for $\boldsymbol{e}$.

Before studying the properties of $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{Y}}$, and $\boldsymbol{e}$, it is useful to study the general properties of linear functions of random vectors.

Random Vectors $\boldsymbol{Z}$

Let $\boldsymbol{Z} = (\, z_1 \quad \cdots \quad z_n \,)'$ be a random vector consisting of random variables $z_1, z_2, \ldots, z_n$. The mean μ_z of the random vector $\boldsymbol{Z}$ is defined as an $n \times 1$ vector with the ith coordinate given by $\mathcal{E}(z_i)$. The variance–covariance matrix $\boldsymbol{V}_z$ for $\boldsymbol{Z}$ is defined as an $n \times n$ symmetric matrix with the diagonal elements equal to the variances of the random variables (in order) and the (i,j)th off-diagonal element equal to the covariance between z_i and z_j. For example, if $\boldsymbol{Z}$ is a 3×1 vector of random variables $z_1, z_2,$ and z_3, then the mean vector of $\boldsymbol{Z}$ is the 3×1 vector

$\mathcal{E}(\boldsymbol{Z})$

$$\mathcal{E}(\boldsymbol{Z}) = \begin{pmatrix} \mathcal{E}(z_1) \\ \mathcal{E}(z_2) \\ \mathcal{E}(z_3) \end{pmatrix} = \boldsymbol{\mu}_z = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} \tag{3.15}$$

and the variance–covariance matrix is the 3×3 matrix

$\mathbf{Var}(\boldsymbol{Z})$

$$\begin{aligned} \mathbf{Var}(\boldsymbol{Z}) &= \begin{bmatrix} Var(z_1) & Cov(z_1,z_2) & Cov(z_1,z_3) \\ Cov(z_2,z_1) & Var(z_2) & Cov(z_2,z_3) \\ Cov(z_3,z_1) & Cov(z_3,z_2) & Var(z_3) \end{bmatrix} \\ &= \boldsymbol{V}_z \end{aligned} \tag{3.16}$$

$$= \begin{bmatrix} \mathcal{E}[(z_1-\mu_1)^2] & \mathcal{E}[(z_1-\mu_1)(z_2-\mu_2)] & \mathcal{E}[(z_1-\mu_1)(z_3-\mu_3)] \\ \mathcal{E}[(z_2-\mu_2)(z_1-\mu_1)] & \mathcal{E}[(z_2-\mu_2)^2] & \mathcal{E}[(z_2-\mu_2)(z_3-\mu_3)] \\ \mathcal{E}[(z_3-\mu_3)(z_1-\mu_1)] & \mathcal{E}[(z_3-\mu_3)(z_2-\mu_2)] & \mathcal{E}[(z_3-\mu_3)^2] \end{bmatrix}$$

$$= \mathcal{E}\{[\boldsymbol{Z}-\mathcal{E}(\boldsymbol{Z})][\boldsymbol{Z}-\mathcal{E}(\boldsymbol{Z})]'\}. \tag{3.17}$$

Linear Functions of $\boldsymbol{Z}$

Let $\boldsymbol{Z}$ be an $n \times 1$ random vector with mean $\boldsymbol{\mu}_z$ and variance–covariance matrix $\boldsymbol{V}_z$. Let

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{a}_1' \\ \boldsymbol{a}_2' \\ \vdots \\ \boldsymbol{a}_k' \end{pmatrix}$$

be a $k \times n$ matrix of constants. Consider the linear transformation $\boldsymbol{U} = \boldsymbol{AZ}$. That is, $\boldsymbol{U}$ is a $k \times 1$ vector given by

$\boldsymbol{U}=\boldsymbol{AZ}$

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{a}_1'\boldsymbol{Z} \\ \boldsymbol{a}_2'\boldsymbol{Z} \\ \vdots \\ \boldsymbol{a}_k'\boldsymbol{Z} \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{bmatrix}. \tag{3.18}$$

Note that

$$\begin{aligned} \mathcal{E}(u_i) &= \mathcal{E}(\boldsymbol{a}_i'\boldsymbol{Z}) \\ &= \mathcal{E}[a_{i1}z_1 + a_{i2}z_2 + \cdots + a_{in}z_n] \\ &= a_{i1}\mathcal{E}(z_1) + a_{i2}\mathcal{E}(z_2) + \cdots + a_{in}\mathcal{E}(z_n) \\ &= \boldsymbol{a}_i'\boldsymbol{\mu}_z, \end{aligned}$$

and hence

$\mathcal{E}(\boldsymbol{U})$

$$\begin{aligned} \mathcal{E}[\boldsymbol{U}] &= \begin{bmatrix} \mathcal{E}(u_1) \\ \mathcal{E}(u_2) \\ \vdots \\ \mathcal{E}(u_k) \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1'\boldsymbol{\mu}_z \\ \boldsymbol{a}_2'\boldsymbol{\mu}_z \\ \vdots \\ \boldsymbol{a}_k'\boldsymbol{\mu}_z \end{bmatrix} \\ &= \boldsymbol{A}\boldsymbol{\mu}_z. \end{aligned} \tag{3.19}$$

The $k \times k$ variance–covariance matrix for $\boldsymbol{U}$ is given by

$\mathbf{Var}(\boldsymbol{U})$

$$\begin{aligned} \mathbf{Var}(\boldsymbol{U}) &= \boldsymbol{V}_u \\ &= \mathcal{E}[\boldsymbol{U}-\mathcal{E}(\boldsymbol{U})][\boldsymbol{U}-\mathcal{E}(\boldsymbol{U})]'. \end{aligned}$$

Substitution of $\boldsymbol{AZ}$ for $\boldsymbol{U}$ and factoring gives

$$\begin{aligned}\boldsymbol{V}_u &= \mathcal{E}[\boldsymbol{AZ} - \boldsymbol{A\mu}_z][\boldsymbol{AZ} - \boldsymbol{A\mu}_z]' \\ &= \mathcal{E}\,\boldsymbol{A}[\boldsymbol{Z} - \boldsymbol{\mu}_z][\boldsymbol{Z} - \boldsymbol{\mu}_z]'\boldsymbol{A}' \\ &= \boldsymbol{A}\mathcal{E}[\boldsymbol{Z} - \boldsymbol{\mu}_z][\boldsymbol{Z} - \boldsymbol{\mu}_z]'\boldsymbol{A}' \\ &= \boldsymbol{A}[\mathbf{Var}(\boldsymbol{Z})]\boldsymbol{A}' \\ &= \boldsymbol{AV}_z\boldsymbol{A}'. \end{aligned} \tag{3.20}$$

The factoring of matrix products must be done carefully; remember that matrix multiplication is not commutative. Therefore, $\boldsymbol{A}$ is factored both to the left (from the first quantity in square brackets) and to the right (from the transpose of the second quantity in square brackets). Remember that transposing a product reverses the order of multiplication $(\boldsymbol{CD})' = \boldsymbol{D}'\boldsymbol{C}'$. Since $\boldsymbol{A}$ is a matrix of constants it can be factored outside the expectation operator. This leaves an inner matrix which by definition is $\mathbf{Var}(\boldsymbol{Z})$.

Note that, if $\mathbf{Var}(\boldsymbol{Z}) = \sigma^2\boldsymbol{I}$, then

$$\begin{aligned}\mathbf{Var}(\boldsymbol{U}) &= \boldsymbol{A}[\sigma^2\boldsymbol{I}]\boldsymbol{A}' \\ &= \boldsymbol{AA}'\sigma^2. \end{aligned} \tag{3.21}$$

The ith diagonal element of $\boldsymbol{AA}'$ is the sum of squares of the coefficients $(\boldsymbol{a}_i'\boldsymbol{a}_i)$ of the ith linear function $u_i = \boldsymbol{a}_i'\boldsymbol{Z}$. This coefficient multiplied by σ^2 gives the variance of the ith linear function. The (i,j)th off-diagonal element is the sum of products of the coefficients $(\boldsymbol{a}_i'\boldsymbol{a}_j)$ of the ith and jth linear functions and, when multiplied by σ^2, gives the covariance between two linear functions $u_i = \boldsymbol{a}_i'\boldsymbol{Z}$ and $u_j = \boldsymbol{a}_j'\boldsymbol{Z}$.

Note that if $\boldsymbol{A}$ is just a vector $\boldsymbol{a}'$, then $u = \boldsymbol{a}'\boldsymbol{Z}$ is a linear function of $\boldsymbol{Z}$. The variance of u is expressed in terms of $\mathbf{Var}(\boldsymbol{Z})$ as

$$\sigma^2(u) = \boldsymbol{a}'\mathbf{Var}(\boldsymbol{Z})\boldsymbol{a}. \tag{3.22}$$

If $\mathbf{Var}(\boldsymbol{Z}) = \boldsymbol{I}\sigma^2$, then

$$\sigma^2(u) = \boldsymbol{a}'(\boldsymbol{I}\sigma^2)\boldsymbol{a} = \boldsymbol{a}'\boldsymbol{a}\sigma^2. \tag{3.23}$$

Notice that $\boldsymbol{a}'\boldsymbol{a}$ is the sum of squares of the coefficients of the linear function $\sum a_i^2$, which is the result given in Section 1.5.

Two examples illustrate the derivation of variances of linear functions using the preceding important results.

Example 3.5

Matrix notation is used to derive the familiar expectation and variance of a sample mean. Suppose $Y_1, Y_2, \ldots, Y_n$ are independent random variables with mean μ and variance σ^2. Then, for $\boldsymbol{Y} = (\,Y_1 \quad Y_2 \quad \cdots \quad Y_n\,)'$,

$$\mathcal{E}(\boldsymbol{Y}) = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} = \boldsymbol{\mu}\mathbf{1}$$

and

$$\mathbf{Var}(\boldsymbol{Y}) = \boldsymbol{I}\sigma^2.$$

The mean of a sample of n observations, $\overline{\boldsymbol{Y}} = \sum Y_i/n$, is written in matrix notation as

$$\overline{Y} = \left(\begin{array}{cccc} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{array}\right)\boldsymbol{Y}. \tag{3.24}$$

Thus, $\overline{Y}$ is a linear function of $\boldsymbol{Y}$ with the vector of coefficients being

$$\boldsymbol{a}' = \left(\begin{array}{cccc} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{array}\right).$$

Then,

$$\mathcal{E}(\overline{Y}) = \boldsymbol{a}'\mathcal{E}(\boldsymbol{Y}) = \boldsymbol{a}'\mathbf{1}\mu = \mu \tag{3.25}$$

and

$$\begin{aligned} \mathrm{Var}(\overline{Y}) &= \boldsymbol{a}'[\mathbf{Var}(\boldsymbol{Y})]\boldsymbol{a} = \boldsymbol{a}'(\boldsymbol{I}\sigma^2)\boldsymbol{a} \\ &= \left(\begin{array}{cccc} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{array}\right)(\boldsymbol{I}\sigma^2)\begin{bmatrix} \frac{1}{n} \\ \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{bmatrix} \\ &= n\left(\frac{1}{n}\right)^2\sigma^2 = \frac{\sigma^2}{n}. \end{aligned} \tag{3.26}$$

■

Example 3.6

For the second example, consider two linear contrasts on a set of four treatment means with n observations in each mean. The random vector in this case is the vector of the four treatment means. If the means have been computed from random samples from four populations with means μ_1, μ_2, μ_3, and μ_4 and equal variance σ^2, then the variance of each sample mean will be σ^2/n (equation 3.26, and all covariances between the means will be zero. The mean of the vector of sample means $\overline{\boldsymbol{Y}} = \left(\begin{array}{cccc} \overline{Y}_1 & \overline{Y}_2 & \overline{Y}_3 & \overline{Y}_4 \end{array}\right)'$ is $\boldsymbol{\mu} = \left(\begin{array}{cccc} \mu_1 & \mu_2 & \mu_3 & \mu_4 \end{array}\right)'$. The variance–covariance matrix for the vector of means $\overline{\boldsymbol{Y}}$ is $\mathbf{Var}(\overline{\boldsymbol{Y}}) = \boldsymbol{I}(\sigma^2/n)$. Assume that the two linear contrasts of interest are

$$c_1 = \overline{Y}_1 - \overline{Y}_2 \qquad \text{and} \qquad c_2 = \overline{Y}_1 - 2\overline{Y}_2 + \overline{Y}_3.$$

Notice that $\overline{Y}_4$ is not involved in these contrasts. The contrasts can be written as

$$\boldsymbol{C} = \boldsymbol{A}\boldsymbol{Y}, \tag{3.27}$$

where

$$\boldsymbol{C} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{A} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & -2 & 1 & 0 \end{bmatrix}.$$

Then,

$$\mathcal{E}(\boldsymbol{C}) = \boldsymbol{A}\mathcal{E}(\overline{\boldsymbol{Y}}) = \boldsymbol{A}\boldsymbol{\mu} = \begin{bmatrix} \mu_1 - \mu_2 \\ \mu_1 - 2\mu_2 + \mu_3 \end{bmatrix} \tag{3.28}$$

and

$$\begin{aligned} \mathbf{Var}(\boldsymbol{C}) &= \boldsymbol{A}[\mathbf{Var}(\overline{\boldsymbol{Y}})]\boldsymbol{A}' = \boldsymbol{A}\left[\boldsymbol{I}\left(\frac{\sigma^2}{n}\right)\right]\boldsymbol{A}' \\ &= \boldsymbol{A}\boldsymbol{A}'\left(\frac{\sigma^2}{n}\right) = \begin{bmatrix} 2 & 3 \\ 3 & 6 \end{bmatrix}\frac{\sigma^2}{n}. \end{aligned} \tag{3.29}$$

Thus, the variance of c_1 is $2\sigma^2/n$, the variance of c_2 is $6\sigma^2/n$, and the covariance between the two contrasts is $3\sigma^2/n$. ■

Multivariate Normal Distribution

We now develop the multivariate normal distribution and present some properties of multivariate normal random vectors. We first define a multivariate random vector when the elements of the vector are mutually independent. We then extend the results to normal random vectors with a nonzero mean and a variance–covariance matrix that is not necessarily diagnonal. Finally, we present a result for linear functions of normal random vectors.

Normal Random Vectors

Suppose $z_1, z_2, \ldots, z_n$ are independent normal random variables with mean zero and variance σ^2. Then, the random vector $\boldsymbol{Z} = (\, z_1 \quad \cdots \quad z_n \,)'$ is said to have a multivariate normal distribution with mean $\mathbf{0} = (\, 0 \quad \cdots \quad 0 \,)'$ and variance–covariance matrix $\boldsymbol{V}_z = \boldsymbol{I}\sigma^2$. This is denoted as

$$\boldsymbol{Z} \sim N(\mathbf{0}, \boldsymbol{I}\sigma^2).$$

The probability density function of $\boldsymbol{Z}$ is given in equation (3.3) and can also be expressed as

$$(2\pi)^{-n/2}|\boldsymbol{I}\sigma^2|^{-1/2}\, e^{-\left[\boldsymbol{Z}'(\boldsymbol{I}\sigma^2)^{-1}\boldsymbol{Z}/2\right]}. \tag{3.30}$$

It is a general result that if $\boldsymbol{U}$ is any linear function $\boldsymbol{U} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{b}$, where $\boldsymbol{A}$ is a $k \times n$ matrix of constants and $\boldsymbol{b}$ is a $k \times 1$ vector of constants, then $\boldsymbol{U}$ is itself normally distributed with mean $\boldsymbol{\mu}_u = \boldsymbol{b}$ and variance–covariance matrix $\mathbf{Var}(\boldsymbol{U}) = \boldsymbol{V}_u = \boldsymbol{A}\boldsymbol{A}'\sigma^2$ (Searle, 1971). The random vector $\boldsymbol{U}$ has a multivariate normal distribution which is denoted by

$$\boldsymbol{U} \sim N(\boldsymbol{\mu}_u, \boldsymbol{V}_u). \tag{3.31}$$

If $\boldsymbol{A}$ is of rank k, then the probability density function of $\boldsymbol{U}$ is given by

$$(2\pi)^{-k/2}|\boldsymbol{V}_u|^{-1/2}\, e^{-(1/2)\left\{[\boldsymbol{U}-\boldsymbol{\mu}_u]'\boldsymbol{V}_u^{-1}[\boldsymbol{U}-\boldsymbol{\mu}_u]\right\}}. \tag{3.32}$$

The preceding result holds for vectors other than $\boldsymbol{Z}$ also. For example, if $\boldsymbol{U} \sim N(\boldsymbol{\mu}_u, \boldsymbol{V}_u)$ and if

$$\boldsymbol{Y} \quad = \quad \boldsymbol{B}\boldsymbol{U} + \boldsymbol{c}, \tag{3.33}$$

where $\boldsymbol{B}$ is a matrix of constants and $\boldsymbol{c}$ is a vector of constants, then

$$\boldsymbol{Y} \sim N(\boldsymbol{\mu}_y, \boldsymbol{V}_y), \tag{3.34}$$

where $\boldsymbol{\mu}_y = \boldsymbol{B}\boldsymbol{\mu}_u + \boldsymbol{c}$ and $\boldsymbol{V}_y = \boldsymbol{B}\boldsymbol{V}_u\boldsymbol{B}'$. In Examples 3.5 and 3.6, if the data are assumed to be from a normal population, then $\overline{\boldsymbol{Y}}$ in equation 3.24 is $N(\mu, \sigma^2/n)$ and $\boldsymbol{C}$ in equation 3.27 is

$$N\left(\begin{bmatrix}\mu_1 - \mu_2 \\ \mu_1 - 2\mu_2 + \mu_3\end{bmatrix}, \begin{bmatrix}2 & 3 \\ 3 & 6\end{bmatrix}\frac{\sigma^2}{n}\right).$$

3.5 Properties of Regression Estimates

The estimated regression coefficients $\widehat{\boldsymbol{\beta}}$, the fitted values $\widehat{\boldsymbol{Y}}$, and the residuals $\boldsymbol{e}$ are all linear functions of the original observations $\boldsymbol{Y}$. Recall that

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Since we have assumed that ϵ_i are independent random variables with mean zero and variance σ^2, we have

$$\mathcal{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$$

and

$$Var(\boldsymbol{\epsilon}) = \boldsymbol{I}\sigma^2.$$

The Y Vector

Note that

$$\begin{aligned}\mathcal{E}(\boldsymbol{Y}) &= \mathcal{E}[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}] = \mathcal{E}[\boldsymbol{X}\boldsymbol{\beta}] + \mathcal{E}[\boldsymbol{\epsilon}] \\ &= \boldsymbol{X}\boldsymbol{\beta}\end{aligned} \tag{3.35}$$

and

$$\mathbf{Var}(\boldsymbol{Y}) = \mathbf{Var}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathbf{Var}(\boldsymbol{\epsilon}) = \boldsymbol{I}\sigma^2. \tag{3.36}$$

Here, $\mathbf{Var}(\boldsymbol{Y})$ is the same as $\mathbf{Var}(\boldsymbol{\epsilon})$ since adding a constant like $\boldsymbol{X}\boldsymbol{\beta}$ to a random variable does not change the variance. When $\boldsymbol{\epsilon}$ is normally distributed, $\boldsymbol{Y}$ is also multivariate normally distributed. Thus,

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}). \tag{3.37}$$

This result is based on the assumption that the linear model used is the correct model. If important independent variables have been omitted or if the functional form of the model is not correct, $\boldsymbol{X\beta}$ will not be the expectation of $\boldsymbol{Y}$. Assuming that the model is correct, the joint probability density function of $\boldsymbol{Y}$ is given by

$$\begin{aligned}(2\pi)^{-n/2}|\boldsymbol{I}\sigma^2|^{-1/2}e^{-(1/2)\{(\boldsymbol{Y}-\boldsymbol{X\beta})'(\boldsymbol{I}\sigma^2)^{-1}(\boldsymbol{Y}-\boldsymbol{X\beta})\}}\\ = (2\pi)^{-n/2}\sigma^{-n}e^{-(1/2\sigma^2)(\boldsymbol{Y}-\boldsymbol{X\beta})'(\boldsymbol{Y}-\boldsymbol{X\beta})}.\end{aligned} \tag{3.38}$$

$\widehat{\boldsymbol{\beta}}$ Vector

Expressing $\widehat{\boldsymbol{\beta}}$ as $\widehat{\boldsymbol{\beta}} = [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{Y}$ shows that the estimates of the regression coefficients are linear funtions of the dependent variable $\boldsymbol{Y}$, with the coefficients being given by $\boldsymbol{A} = [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']$. Since the Xs are constants, the matrix $\boldsymbol{A}$ is also constant. If the model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ is correct, the expectation of $\boldsymbol{Y}$ is $\boldsymbol{X\beta}$ and the expectation of $\widehat{\boldsymbol{\beta}}$ is

$$\begin{aligned}\mathcal{E}(\widehat{\boldsymbol{\beta}}) &= [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\mathcal{E}(\boldsymbol{Y})\\ &= [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{X\beta}\\ &= [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}]\boldsymbol{\beta}\\ &= \boldsymbol{\beta}.\end{aligned} \tag{3.39}$$

This shows that $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ *if* the chosen model is correct. If the chosen model is *not* correct, say $\mathcal{E}(\boldsymbol{Y}) = \boldsymbol{X\beta} + \boldsymbol{Z\gamma}$ instead of $\boldsymbol{X\beta}$, then $[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\mathcal{E}(\boldsymbol{Y})$ does not necessarily simplify to $\boldsymbol{\beta}$.

Assuming that the model is correct,

$$\begin{aligned}\mathbf{Var}(\widehat{\boldsymbol{\beta}}) &= [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'][\mathbf{Var}(\boldsymbol{Y})][(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']'\\ &= [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{I}\sigma^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']'.\end{aligned}$$

Recalling that the transpose of a product is the product of transposes in reverse order [i.e., $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$], that $\boldsymbol{X}'\boldsymbol{X}$ is symmetric, and that the inverse of a transpose is the transpose of the inverse, we obtain

$$\begin{aligned}\mathbf{Var}(\widehat{\boldsymbol{\beta}}) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2\\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2.\end{aligned} \tag{3.40}$$

Thus, the variances and covariances of the estimated regression coefficients are given by the elements of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ multiplied by σ^2. The diagonal elements give the variances in the order in which the regression coefficients are listed in $\boldsymbol{\beta}$ and the off-diagonal elements give their covariances. When $\boldsymbol{\epsilon}$ is normally distributed, $\widehat{\boldsymbol{\beta}}$ is also multivariate normally distributed. Thus,

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2). \tag{3.41}$$

Example 3.7

In the ozone example, Example 3.3,

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 1.0755 & -9.4340 \\ -9.4340 & 107.8167 \end{bmatrix}.$$

Thus, $\text{Var}(\widehat{\beta}_0) = 1.0755\sigma^2$ and $\text{Var}(\widehat{\beta}_1) = 107.8167\sigma^2$. The covariance between $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is $\text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = -9.4340\sigma^2$. ■

Recall that the vector of estimated means $\widehat{\boldsymbol{Y}}$ is given by

$$\widehat{\boldsymbol{Y}} = [\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{Y}.$$

Therefore, using $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$, the expectation of $\widehat{\boldsymbol{Y}}$ is

$$\mathcal{E}(\widehat{\boldsymbol{Y}}) = \boldsymbol{P}\mathcal{E}(\boldsymbol{Y}) = \boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta}. \tag{3.42}$$

Thus, $\widehat{\boldsymbol{Y}}$ is an unbiased estimator of the mean of $\boldsymbol{Y}$ for the particular values of $\boldsymbol{X}$ in the data set, again *if* the model is correct. The fact that $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ can be verified using the definition of $\boldsymbol{P}$:

$$\begin{aligned} \boldsymbol{P}\boldsymbol{X} &= [\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{X} \\ &= \boldsymbol{X}[(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})] \\ &= \boldsymbol{X}. \end{aligned} \tag{3.43}$$

The variance–covariance matrix of $\widehat{\boldsymbol{Y}}$ can be derived using either the relationship $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ or $\widehat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{Y}$. Recall that $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Applying the rules for variances of linear functions to the first relationship gives

$$\begin{aligned} \mathbf{Var}(\widehat{\boldsymbol{Y}}) &= \boldsymbol{X}[\mathbf{Var}(\widehat{\boldsymbol{\beta}})]\boldsymbol{X}' \\ &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2 \\ &= \boldsymbol{P}\sigma^2. \end{aligned} \tag{3.44}$$

The derivation using the second relationship gives

$$\begin{aligned} \mathbf{Var}(\widehat{\boldsymbol{Y}}) &= \boldsymbol{P}[\mathbf{Var}(\boldsymbol{Y})]\boldsymbol{P}' \\ &= \boldsymbol{P}\boldsymbol{P}'\sigma^2 \\ &= \boldsymbol{P}\sigma^2, \end{aligned} \tag{3.45}$$

since $\boldsymbol{P}$ is symmetric and idempotent. Therefore, the matrix $\boldsymbol{P}$ multiplied by σ^2 gives the variances and covariances for all $\widehat{Y}_i$. $\boldsymbol{P}$ is a large $n \times n$ matrix and at times only a few elements are of interest. The variances of any subset of the $\widehat{Y}_i$ can be determined by using only the rows of $\boldsymbol{X}$, say $\boldsymbol{X}_r$, that correspond to the data points of interest and applying the first derivation. This gives

$$\mathbf{Var}(\widehat{\boldsymbol{Y}}_r) = \boldsymbol{X}_r[\mathbf{Var}(\widehat{\boldsymbol{\beta}})]\boldsymbol{X}_r' = \boldsymbol{X}_r(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_r'\sigma^2. \tag{3.46}$$

When $\boldsymbol{\epsilon}$ is normally distributed,

$$\widehat{\boldsymbol{Y}} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{P}\sigma^2). \tag{3.47}$$

$\boldsymbol{e}$ Vector

Recall that the vector of residuals $\boldsymbol{e}$ is given by $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$. Therefore, the expectation of $\boldsymbol{e}$ is

$$\begin{aligned} \mathcal{E}(\boldsymbol{e}) &= (\boldsymbol{I}-\boldsymbol{P})\mathcal{E}(\boldsymbol{Y}) = (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{X}\boldsymbol{\beta} \\ &= (\boldsymbol{X}-\boldsymbol{P}\boldsymbol{X})\boldsymbol{\beta} = (\boldsymbol{X}-\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{0}, \end{aligned} \tag{3.48}$$

where $\boldsymbol{0}$ is an $n \times 1$ vector of zeros. Thus, the residuals are random variables with mean zero.

The variance–covariance matrix of the residual vector $\boldsymbol{e}$ is

$$\mathbf{Var}(\boldsymbol{e}) = (\boldsymbol{I}-\boldsymbol{P})\sigma^2 \tag{3.49}$$

again using the result that $(\boldsymbol{I}-\boldsymbol{P})$ is a symmetric idempotent matrix. If the vector of regression errors $\boldsymbol{\epsilon}$ is normally distributed, then the vector of regression residuals satisfies

$$\boldsymbol{e} \sim N(\boldsymbol{0}, (\boldsymbol{I}-\boldsymbol{P})\sigma^2). \tag{3.50}$$

Prediction $\widehat{\boldsymbol{Y}}_0$

Prediction of a future random observation, $Y_0 = \boldsymbol{x}_0'\boldsymbol{\beta} + \epsilon_0$ at a given vector of independent variables $\boldsymbol{x}_0'$, is given by $\widehat{Y}_0 = \boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}$. It is easy to see that

$$\widehat{Y}_0 \sim N(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}, \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0\sigma^2). \tag{3.51}$$

This result is used to construct confidence intervals for the mean $\boldsymbol{x}_0'\boldsymbol{\beta}$.

If the future ϵ_0 is assumed to be a normal random variable with mean zero and variance σ^2, and is independent of the historic errors of $\boldsymbol{\epsilon}$, then the prediction error $Y_0 - \widehat{Y}_0 = \boldsymbol{x}_0'(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + \epsilon_0$ satisfies

$$Y_0 - \widehat{Y}_0 \sim N\left(0, [1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0]\sigma^2\right). \tag{3.52}$$

This result is used to construct a confidence interval for an individual Y_0 that we call a prediction interval for Y_0. Recall that the variance of $(Y_0 - \widehat{Y}_0)$ is denoted by $\mathrm{Var}(\widehat{Y}_{\mathrm{pred}_0})$.

Example 3.8

The matrix $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ was computed for the ozone example in Example 3.3. Thus, with some rounding of the elements in $\boldsymbol{P}$,

$$\begin{aligned} \mathbf{Var}(\widehat{\boldsymbol{Y}}) &= \boldsymbol{P}\sigma^2 \\ &= \begin{bmatrix} .741 & .377 & .086 & -.205 \\ .377 & .283 & .208 & .132 \\ .086 & .208 & .305 & .402 \\ -.205 & .132 & .402 & .671 \end{bmatrix} \sigma^2. \end{aligned}$$

The variance of the estimated mean of Y when the ozone level is .02 ppm is $\text{Var}(\widehat{Y}_1) = .741\sigma^2$. For the ozone level of .11 ppm, the variance of the estimated mean is $\text{Var}(\widehat{Y}_3) = .305\sigma^2$. The covariance between the two estimated means is $\text{Cov}(\widehat{Y}_1, \widehat{Y}_3) = .086\sigma^2$.

The variance–covariance matrix of the residuals is obtained by $\mathbf{Var}(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{P})\sigma^2$. Thus,

$$\begin{aligned} \text{Var}(e_1) &= (1 - .741)\sigma^2 = .259\sigma^2 \\ \text{Var}(e_3) &= (1 - .305)\sigma^2 = .695\sigma^2 \\ \text{Cov}(e_1, e_3) &= -\text{Cov}(\widehat{Y}_1, \widehat{Y}_3) = -.086\sigma^2. \end{aligned}$$

It is important to note that the variances of the *least squares* residuals are not equal to σ^2 and the covariances are not zero. The assumption of equal variances and zero covariances applies to the ϵ_i, not the e_i. ■

$\mathbf{Var}(\widehat{Y}_i) \leq \mathbf{Var}(Y_i)$

The variance of any particular $\widehat{Y}_i$ and the variance of the corresponding e_i will always add to σ^2 because

$$\begin{aligned} \mathbf{Var}(\boldsymbol{Y}) &= \mathbf{Var}(\widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{\epsilon}}) \\ &= \mathbf{Var}(\widehat{\boldsymbol{Y}}) + \mathbf{Var}(\widehat{\boldsymbol{\epsilon}}) + \mathbf{Cov}(\widehat{\boldsymbol{Y}}, \widehat{\boldsymbol{\epsilon}}) + \mathbf{Cov}(\widehat{\boldsymbol{\epsilon}}, \widehat{\boldsymbol{Y}}) \\ &= \boldsymbol{P}\sigma^2 + (\boldsymbol{I} - \boldsymbol{P})\sigma^2 + \boldsymbol{P}(\boldsymbol{I} - \boldsymbol{P})\sigma^2 + (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{P}\sigma^2 \\ &= \boldsymbol{P}\sigma^2 + (\boldsymbol{I} - \boldsymbol{P})\sigma^2 \\ &= \boldsymbol{I}\sigma^2. \end{aligned} \tag{3.53}$$

Since variances cannot be negative, each diagonal element of $\boldsymbol{P}$ must be between zero and one: $0 < v_{ii} < 1.0$, where v_{ii} is the ith diagonal element of $\boldsymbol{P}$. Thus, the variance of any $\widehat{Y}_i$ is always less than σ^2, the variance of the individual observations. This shows the advantage of fitting a continuous response model, assuming the model is correct, over simply using the individual observed data points as estimates of the mean of Y for the given values of the Xs. The greater precision from fitting a response model comes from the fact that each $\widehat{Y}_i$ uses information from the surrounding data points. The gain in precision can be quite striking. In Example 3.8, the precision obtained on the estimates of the means for the two intermediate levels of ozone using the linear response equation were $.283\sigma^2$ and $.305\sigma^2$. To attain the same degree of precision without using the response model would have required more than three observations at each level of ozone.

Role by $\boldsymbol{X}$s

Equation 3.53 implies that data points having low variance on $\widehat{Y}_i$ will have high variance on e_i and vice versa. Belsley, Kuh, and Welsch (1980) show that the diagonal elements of $\boldsymbol{P}$, v_{ii} can be interpreted as measures of distance of the corresponding data points from the center of the X-space (from $\overline{X}$ in the case of one independent variable). Points that are far from the center of the X-space have relatively large v_{ii} and, therefore, relatively

high variance on $\widehat{Y}_i$ and low variance on e_i. The smaller variance of the residuals for the points far from the "center of the data" indicates that the fitted regression line or response surface tends to come closer to the observed values for these points. This aspect of $\boldsymbol{P}$ is used later to detect the more influential data points.

Controlling Precision

The variances (and covariances) have been expressed as multiples of σ^2. The coefficients are determined entirely by the $\boldsymbol{X}$ matrix, a matrix of constants that depends on the model being fit and the levels of the independent variables in the study. In designed experiments, the levels of the independent variables are subject to the control of the researcher. Thus, except for the magnitude of σ^2, the precision of the experiment is under the control of the researcher and can be known before the experiment is run. The efficiencies of alternative experimental designs can be compared by computing $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\boldsymbol{P}$ for each design. The design giving the smallest variances for the quantities of interest would be preferred.

3.6 Summary of Matrix Formulae

Model:	$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
Normal equations:	$(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$
Parameter estimates:	$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$
Fitted values:	$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$
	$= \boldsymbol{P}\boldsymbol{Y}$, where $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$
Residuals:	$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$
	$= (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$
Variance of $\widehat{\boldsymbol{\beta}}$:	$\mathbf{Var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$
Variance of $\widehat{\boldsymbol{Y}}$:	$\mathbf{Var}(\widehat{\boldsymbol{Y}}) = \boldsymbol{P}\sigma^2$
Variance of $\boldsymbol{e}$:	$\mathbf{Var}(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{P})\sigma^2$

3.7 Exercises

3.1. The linear model in ordinary least squares is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \mathcal{E}$. Assume there are 30 observations and five independent variables (containing no linear dependencies). Give the order and rank of:

(a) $\boldsymbol{Y}$.

(b) $\boldsymbol{X}$ (without an intercept in the model).

(c) $\boldsymbol{X}$ (with an intercept in the model).

(d) $\boldsymbol{\beta}$ (without an intercept in the model).

(e) $\boldsymbol{\beta}$ (with an intercept in the model).

(f) $\boldsymbol{\epsilon}$.

(g) $(\boldsymbol{X}'\boldsymbol{X})$ (with an intercept in the model).

(h) $\boldsymbol{P}$ (with an intercept in the model).

3.2. For each of the following matrices, indicate whether there will be a unique solution to the normal equations. Show how you arrived at your answer.

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 3 & 8 \\ 1 & 0 & 6 \\ 1 & -1 & 2 \end{bmatrix} \quad \boldsymbol{X}_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} \quad \boldsymbol{X}_3 = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 1 & 2 \\ 1 & -3 & -6 \\ 1 & -1 & -2 \end{bmatrix}.$$

3.3. You have a data set with four independent variables and $n = 42$ observations. If the model is to include an intercept, what would be the order of $\boldsymbol{X}'\boldsymbol{X}$? Of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$? Of $\boldsymbol{X}'\boldsymbol{Y}$? Of $\boldsymbol{P}$?

3.4. A data set with one independent variable and an intercept gave the following $(\boldsymbol{X}'\boldsymbol{X})^{-1}$,

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} \frac{31}{177} & \frac{-3}{177} \\ \frac{-3}{177} & \frac{6}{177} \end{bmatrix}.$$

How many observations were there in the data set? Find $\sum X_i^2$. Find the corrected sum of squares for the independent variable.

3.5. The data in the accompanying table relate grams plant dry weight Y to percent soil organic matter X_1, and kilograms of supplemental soil nitrogen added per 1,000 square meters X_2:

	Y	X_1	X_2
	78.5	7	2.6
	74.3	1	2.9
	104.3	11	5.6
	87.6	11	3.1
	95.9	7	5.2
	109.2	11	5.5
	102.7	3	7.1
Sums:	652.5	51	32.0
Means:	93.21	7.29	4.57

(a) Define $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ for a model involving both independent variables and an intercept.

(b) Compute $\boldsymbol{X'X}$ and $\boldsymbol{X'Y}$.

(c) $(\boldsymbol{X'X})^{-1}$ for this problem is

$$(\boldsymbol{X'X})^{-1} = \begin{bmatrix} 1.7995972 & -.0685472 & -.2531648 \\ -.0685472 & .0100774 & -.0010661 \\ -.2531648 & -.0010661 & .0570789 \end{bmatrix}.$$

Verify that this is the inverse of $\boldsymbol{X'X}$. Compute $\widehat{\boldsymbol{\beta}}$ and write the regression equation. Interpret each estimated regression coefficient. What are the units of measure attached to each regression coefficient?

(d) Compute $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{e}$.

(e) The $\boldsymbol{P}$ matrix in this case is a 7×7 matrix. Illustrate the computation of $\boldsymbol{P}$ by computing v_{11}, the first diagonal element, and v_{12}, the second element in the first row. Use the preceding results and these two elements of $\boldsymbol{P}$ to give the appropriate coefficient on σ^2 for each of the following variances.

(i) $\text{Var}(\widehat{\beta}_1)$

(ii) $\text{Var}(\widehat{Y}_1)$

(iii) $\text{Var}(\widehat{Y}_{pred_1})$

(iv) $\text{Var}(e_1)$.

3.6. Use the data in Exercise 3.5. Center each independent variable by subtracting the column mean from each observation in the column. Compute $\boldsymbol{X'X}, \boldsymbol{X'Y}$, and $\widehat{\boldsymbol{\beta}}$ using the centered data. Were the computations simplified by using centered data? Show that the regression equation obtained using centered data is equivalent to that obtained with the original uncentered data. Compute $\boldsymbol{P}$ using the centered data and compare it to that obtained using the uncentered data.

3.7. The matrix $\boldsymbol{P}$ for the Heagle ozone data is given in Example 3.3. Verify that $\boldsymbol{P}$ is symmetric and idempotent. What is the linear function of Y_i that gives $\widehat{Y}_3$?

3.8. Compute $(\boldsymbol{I} - \boldsymbol{P})$ for the Heagle ozone data. Verify that $(\boldsymbol{I} - \boldsymbol{P})$ is idempotent and that $\boldsymbol{P}$ and $(\boldsymbol{I} - \boldsymbol{P})$ are orthogonal to each other. What does the orthogonality imply about the vectors $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{e}$?

3.9. This exercise uses the Lesser–Unsworth data in Exercise 1.19, in which seed weight is related to cumulative solar radiation for two levels of exposure to ozone. Assume that "low ozone" is an exposure of .025 ppm and that "high ozone" is an exposure of .07 ppm.

(a) Set up $\boldsymbol{X}$ and $\boldsymbol{\beta}$ for the regression of seed weight on cumulative solar radiation *and* ozone level. Center the independent variables and include an intercept in the model. Estimate the regression equation and interpret the result.

(b) Extend the model to include an independent variable that is the product term between centered cumulative solar radiation and centered ozone level. Estimate the regression equation for this model and interpret the result. What does the presence of the product term contribute to the regression equation?

3.10. This exercise uses the data from Exercise 1.21 (number of hospital days for smokers, number of cigarettes smoked, and number of hospital days for control groups of nonsmokers). Exercise 1.21 used the information from the nonsmoker control groups by defining the dependent variable as $Y = \ln(\text{number of hospital days for smokers/number of hospital days for nonsmokers})$. Another method of taking into account the experience of the nonsmokers is to use $X_2 = \ln(\text{number of hospital days for nonsmokers})$ as an independent variable.

(a) Set up $\boldsymbol{X}$ and $\boldsymbol{\beta}$ for the regression of $Y = \ln(\text{number of hospital days for smokers})$ on $X_1 = (\text{number cigarettes})^2$ and $X_2 = \ln(\text{number of hospital days for nonsmokers})$.

(b) Estimate the regression equation and interpret the results. What value of β_2 would correspond to using the nonsmoker experience as was done in Exercise 1.21?

3.11. The data in the table relate the annual catch of Gulf Menhaden, *Brevoortia patronus*, to fishing pressure for 1964 to 1979 (Nelson and

Ahrenholz, 1986).

Year	*Met.*	*Catch* *Ton* $\times 10^{-3}$	*Number* *Vessels*	*Pressure* *Vessel-Ton-Weeks* $\times 10^{-3}$
1964		409.4	76	282.9
1965		463.1	82	335.6
1966		359.1	80	381.3
1967		317.3	76	404.7
1968		373.5	69	382.3
1969		523.7	72	411.0
1970		548.1	73	400.0
1971		728.2	82	472.9
1972		501.7	75	447.5
1973		486.1	65	426.2
1974		578.6	71	485.5
1975		542.6	78	536.9
1976		561.2	81	575.9
1977		447.1	80	532.7
1978		820.0	80	574.3
1979		777.9	77	533.9

Run a linear regression of catch (Y) on fishing pressure (X_1) and number of vessels (X_2). Include an intercept in the model. Interpret the regression equation.

3.12. A simulation model for peak water flow from watersheds was tested by comparing measured peak flow (cfs) from 10 storms with predictions of peak flow obtained from the simulation model. Q_o and Q_p are the observed and predicted peak flows, respectively. Four independent variables were recorded:

X_1 = area of watershed (mi^2),

X_2 = average slope of watershed (in percent),

X_3 = surface absorbency index (0 = complete absorbency, 100 = no absorbency), and

X_4 = peak intensity of rainfall (in/hr) computed on half-hour time intervals.

Q_0	Q_p	X_1	X_2	X_3	X_4
28	32	.03	3.0	70	.6
112	142	.03	3.0	80	1.8
398	502	.13	6.5	65	2.0
772	790	1.00	15.0	60	.4
2,294	3,075	1.00	15.0	65	2.3
2,484	3,230	3.00	7.0	67	1.0
2,586	3,535	5.00	6.0	62	.9
3,024	4,265	7.00	6.5	56	1.1
4,179	6,529	7.00	6.5	56	1.4
710	935	7.00	6.5	56	.7

(a) Use $Y = \ln(Q_o/Q_p)$ as the dependent variable. The dependent variable will have the value zero if the observed and predicted peak flows agree. Set up the regression problem to determine whether the discrepancy Y is related to any of the four independent variables. Use an intercept in the model. Estimate the regression equation.

(b) Further consideration of the problem suggested that the discrepancy between observed and predicted peak flow Y might go to zero as the values of the four independent variables approach zero. Redefine the regression problem to eliminate the intercept (force $\beta_0 = 0$), and estimate the regression equation.

(c) Rerun the regression (without the intercept) using only X_1 and X_4; that is, omit X_2 and X_3 from the model. Do the regression coefficients for X_1 and X_4 change? Explain why they do or do not change.

(d) Describe the change in the standard errors of the estimated regression coefficients as the intercept was dropped [Part (a) versus Part (b)] and as X_2 and X_3 were dropped from the model [Part (b) versus Part (c)].

3.13. You have fit a linear model using $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{X}$ involves r independent variables. Now assume that the true model involves an *additional* s independent variables contained in $\boldsymbol{Z}$. That is, the true model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\gamma}$ is the vector of regression coefficients for the independent variables contained in $\boldsymbol{Z}$.

(a) Find $\mathcal{E}(\widehat{\boldsymbol{\beta}})$ and show that, in general, $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ is a biased estimate of $\boldsymbol{\beta}$.

(b) Under what conditions would $\widehat{\boldsymbol{\beta}}$ be unbiased?

3.14. The accompanying table shows the part of the data reported by Cameron and Pauling (1978) related to the effects of supplemental ascorbate, vitamin C, in the treatment of colon cancer. The data are taken from Andrews and Herzberg (1985) and are used with permission.

Sex	*Age*	*Days* [a]	*Control* [b]
F	76	135	18
F	58	50	30
M	49	189	65
M	69	1,267	17
F	70	155	57
F	68	534	16
M	50	502	25
F	74	126	21
M	66	90	17
F	76	365	42
F	56	911	40
M	65	743	14
F	74	366	28
M	58	156	31
F	60	99	28
M	77	20	33
M	38	274	80

[a]Days = number of days survival after date of untreatability.
[b]Control = average number of days survival of 10 control patients.

Use $Y = \ln(\text{days})$ as the dependent variable and $X_1 = \text{sex}$ (coded -1 for males and $+1$ for females), $X_2 = \text{age}$, and $X_3 = \ln(\text{control})$ in a multiple regression to determine if there is any relationship between days survival and sex and age. Define $\boldsymbol{X}$ and $\boldsymbol{\beta}$, and estimate the regression equation. Explain why X_3 is in the model if the purpose is to relate survival to X_1 and X_2.

3.15. Suppose $\boldsymbol{U} \sim N(\boldsymbol{\mu}_u, \boldsymbol{V}_u)$. Let

$$\boldsymbol{U} = \begin{pmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{pmatrix}, \quad \boldsymbol{\mu}_u = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{V}_u = \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix}.$$

Use equation 3.32 to show that $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are independent if $\boldsymbol{V}_{12} = 0$. That is, if $\boldsymbol{u}$ is multivariate normal, then $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ uncorrelated implies $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are independent. (The joint density of $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$

is the product of the marginal densities of $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, if and only if $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are independent.)

3.16. Consider the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Let

$$\boldsymbol{U} = \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \boldsymbol{e} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \\ (\boldsymbol{I} - \boldsymbol{P}) \end{bmatrix} \boldsymbol{Y}.$$

Find the distribution of $\boldsymbol{U}$. Show that $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{e}$ are independent. (Hint: Use the result in equation 3.31 and Exercise 3.15.)

4

ANALYSIS OF VARIANCE AND QUADRATIC FORMS

The previous chapter developed the regression results involving linear functions of the dependent variable, $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{Y}}$, and $\boldsymbol{e}$. All were shown to be normally distributed random variables if $\boldsymbol{Y}$ was normally distributed.

This chapter develops the distributional results for all quadratic functions of $\boldsymbol{Y}$. The distribution of quadratic forms is used to develop tests of hypotheses, confidence interval estimates, and joint confidence regions for $\boldsymbol{\beta}$.

The estimates of the regression coefficients, the estimated means, and the residuals have been presented in matrix notation; all were shown to be **linear functions** of the original observations $\boldsymbol{Y}$. In this chapter it is shown that the model, regression and residual sums of squares, and the sums of squares used for testing a linear contrast or a collection of linear hypotheses are all **quadratic forms** of Y. This means that each sum of squares can be written as $\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}$, where $\boldsymbol{A}$ is a matrix of coefficients called the **defining matrix**. $\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}$ is referred to as a **quadratic form** in $\boldsymbol{Y}$.

The aim of model fitting is to explain as much of the variation in the dependent variable as possible from information contained in the independent variables. The contributions of the independent variables to the model are measured by partitions of the total sum of squares of $\boldsymbol{Y}$ attributable to, or "explained" by, the independent variables. Each of these partitions of the sums of squares is a quadratic form in $\boldsymbol{Y}$. The degrees of freedom associated with a particular sum of squares and the orthogonality between different sums of squares are determined by the defining matrices in the

quadratic forms. The matrix form for a sum of squares makes the computations simple if one has access to a computer package for matrix algebra. Also, the expectations and variances of the sums of squares are easily determined in this form. We give a brief introduction to quadratic forms and their properties. We also discuss how the properties of quadratic forms are useful for testing linear hypotheses and for the analysis of variance of the dependent variable $\boldsymbol{Y}$.

4.1 Introduction to Quadratic Forms

Quadratic Form for One Contrast

Consider first a sum of squares with which you are familiar from your earlier statistical methods courses, the sum of squares attributable to a linear contrast. Suppose you are interested in the linear contrast

$$C_1^* = Y_1 + Y_2 - 2Y_3. \tag{4.1}$$

The sum of squares due to this contrast is

$$\mathrm{SS}(C_1^*) = \frac{(C_1^*)^2}{6}. \tag{4.2}$$

The divisor of 6 is the sum of squares of the coefficients of the contrast. This divisor has been chosen to make the coefficient of σ^2 in the expectation of the sum of squares equal to 1. If we reexpress C_1^* so that the coefficients on the Y_i include $1/\sqrt{6}$, the sum of squares due to the contrast is the square of the contrast. Thus, $C_1 = C_1^*/\sqrt{6}$ can be written in matrix notation as

$$C_1 = \boldsymbol{a}'\boldsymbol{Y} = \frac{1}{\sqrt{6}}Y_1 + \frac{1}{\sqrt{6}}Y_2 - \frac{2}{\sqrt{6}}Y_3 \tag{4.3}$$

by defining $\boldsymbol{a} = (\,1/\sqrt{6} \quad 1/\sqrt{6} \quad -2/\sqrt{6}\,)'$ and $\boldsymbol{Y} = (\,Y_1 \quad Y_2 \quad Y_3\,)'$. The sum of squares for C_1 is then

$$\begin{aligned} \mathrm{SS}(C_1) &= C_1^2 = (\boldsymbol{a}'\boldsymbol{Y})'(\boldsymbol{a}'\boldsymbol{Y}) \\ &= \boldsymbol{Y}'(\boldsymbol{a}\boldsymbol{a}')\boldsymbol{Y} \\ &= \boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}. \end{aligned} \tag{4.4}$$

Defining Matrix

Thus, $\mathrm{SS}(C_1)$ has been written as a **quadratic form** in $\boldsymbol{Y}$ where $\boldsymbol{A}$, the **defining matrix**, is the 3×3 matrix $\boldsymbol{A} = \boldsymbol{a}\boldsymbol{a}'$. The multiplication $\boldsymbol{a}\boldsymbol{a}'$ for this contrast gives

$$\boldsymbol{A} = \boldsymbol{a}\boldsymbol{a}' = \begin{bmatrix} \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \end{bmatrix} \left(\begin{array}{ccc} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \end{array} \right)$$

$$= \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & -\frac{2}{6} \\ \frac{1}{6} & \frac{1}{6} & -\frac{2}{6} \\ -\frac{2}{6} & -\frac{2}{6} & \frac{4}{6} \end{bmatrix}. \tag{4.5}$$

Completing the multiplication of the quadratic form gives

$$\begin{aligned} \boldsymbol{Y'AY} &= \begin{pmatrix} Y_1 & Y_2 & Y_3 \end{pmatrix} \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & -\frac{2}{6} \\ \frac{1}{6} & \frac{1}{6} & -\frac{2}{6} \\ -\frac{2}{6} & -\frac{2}{6} & \frac{4}{6} \end{bmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} \\ &= \frac{1}{6}[Y_1(Y_1 + Y_2 - 2Y_3) + Y_2(Y_1 + Y_2 - 2Y_3) \\ &\quad + Y_3(-2Y_1 - 2Y_2 + 4Y_3)] \\ &= \frac{1}{6}Y_1^2 + \frac{1}{6}Y_2^2 + \frac{4}{6}Y_3^2 + \frac{2}{6}Y_1Y_2 - \frac{4}{6}Y_1Y_3 - \frac{4}{6}Y_2Y_3. \end{aligned} \tag{4.6}$$

This result is verified by expanding the square of C_1^*, equation 4.1, in terms of Y_i and dividing by 6.

Comparison of the elements of $\boldsymbol{A}$, equation 4.5, with the expansion, equation 4.6, shows that the diagonal elements of the defining matrix are the coefficients on the squared terms and the *sums* of the symmetric off-diagonal elements are the coefficients on the product terms. The defining matrix for a quadratic form is always written in this symmetric form.

Consider a second linear contrast on $\boldsymbol{Y}$ that is orthogonal to C_1. Let $C_2 = (Y_1 - Y_2)/\sqrt{2} = \boldsymbol{d'Y}$ where $\boldsymbol{d} = (\,1/\sqrt{2} \quad -1/\sqrt{2} \quad 0\,)'$. The sum of squares for this contrast is

$$\mathrm{SS}(C_2) = \boldsymbol{Y'DY}, \tag{4.7}$$

where the defining matrix is

$$\boldsymbol{D} = \boldsymbol{dd'} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ -\frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{4.8}$$

Degrees of Freedom

Each of these sums of squares has 1 degree of freedom since a single linear contrast is involved in each case. The **degrees of freedom** for a quadratic form are equal to the rank of the defining matrix which, in turn, is equal to the trace of the defining matrix if the defining matrix is **idempotent**. (The defining matrix for a quadratic form does not have to be idempotent.

However, the quadratic forms with which we are concerned have idempotent defining matrices.) The defining matrices $\boldsymbol{A}$ and $\boldsymbol{D}$ in the two examples are idempotent. It is left to the reader to verify that $\boldsymbol{AA} = \boldsymbol{A}$ and $\boldsymbol{DD} = \boldsymbol{D}$ (see Exercise 2.25). $\boldsymbol{A}$ and $\boldsymbol{D}$ would not have been idempotent if, for example, the $1/\sqrt{6}$ and $1/\sqrt{2}$ had not been incorporated into the coefficient vectors. Notice that $\text{tr}(\boldsymbol{A}) = \text{tr}(\boldsymbol{D}) = 1$, the degrees of freedom for each contrast.

Quadratic Form—Joint Functions

The quadratic forms defined by $\boldsymbol{A}$ and $\boldsymbol{D}$ treated each linear function separately. That is, each quadratic form was a sum of squares with 1 degree of freedom. The two linear functions can be considered jointly by defining the coefficient matrix $\boldsymbol{K}'$ to be a 2×3 matrix containing the coefficients for both contrasts:

$$\boldsymbol{K}'\boldsymbol{Y} = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}. \tag{4.9}$$

The defining matrix for quadratic form $\boldsymbol{Y}'\boldsymbol{K}\boldsymbol{K}'\boldsymbol{Y}$ is

$$\boldsymbol{F} = \boldsymbol{K}\boldsymbol{K}' = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{bmatrix}. \tag{4.10}$$

In this example, the defining matrix $\boldsymbol{F}$ is idempotent and its trace indicates that there are 2 degrees of freedom for this sum of squares. (The quadratic form defined in this way is idempotent only because the two original contrasts were orthogonal to each other, $\boldsymbol{a}'\boldsymbol{d} = 0$. The general method of defining quadratic forms, sums of squares, for specific hypotheses is discussed in Section 4.5.1.)

Orthogonal Quadratic Forms

Two quadratic forms (of the same vector $\boldsymbol{Y}$) are orthogonal if the product of the defining matrices is $\boldsymbol{0}$. Orthogonality of the two quadratic forms in the example is verified by the multiplication of $\boldsymbol{A}$ and $\boldsymbol{D}$:

$$\boldsymbol{DA} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ -\frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & -\frac{2}{6} \\ \frac{1}{6} & \frac{1}{6} & -\frac{2}{6} \\ -\frac{2}{6} & -\frac{2}{6} & \frac{4}{6} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{4.11}$$

which equals $\boldsymbol{AD}$ since $\boldsymbol{A}$, $\boldsymbol{D}$, and $\boldsymbol{DA}$ are all symmetric. Note that $\boldsymbol{DA} = \boldsymbol{dd}'\boldsymbol{aa}'$ and will be zero if $\boldsymbol{d}'\boldsymbol{a} = 0$. Thus, the quadratic forms associated with two linear functions will be orthogonal if the two vectors of

coefficients are orthogonal—that is, if the sum of products of the coefficient vectors $\boldsymbol{d}'\boldsymbol{a}$ is zero (see Exercise 2.26). When the two linear functions are orthogonal, the sum of sums of squares (and degrees of freedom) of the two contrasts considered individually will equal the sum of squares (and degrees of freedom) of the two contrasts considered jointly. For this additivity to hold when more than two linear functions are considered, all must be pairwise orthogonal. Orthogonality of quadratic forms implies that the two pieces of information contained in the individual sums of squares are independent.

The quadratic forms of primary interest in this text are the sums of squares associated with analyses of variance, regression analyses, and tests of hypotheses. All have idempotent defining matrices.

The following facts about quadratic forms are important [see Searle (1971) for more complete discussions on quadratic forms].

1. Any sum of squares can be written as $\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}$, where $\boldsymbol{A}$ is a square symmetric nonnegative definite matrix.
2. The degrees of freedom associated with any quadratic form equal the rank of the defining matrix, which equals its trace when the matrix is idempotent.
3. Two quadratic forms are orthogonal if the product of their defining matrices is the null matrix $\mathbf{0}$.

Example 4.1

For illustration of quadratic forms, let

$$\boldsymbol{Y} = (\,3.55 \quad 3.49 \quad 3.67 \quad 2.76 \quad 1.195\,)'$$

be the vector of mean disease scores for a fungus disease on alfalfa. The five treatments were five equally spaced day/night temperature regimes under which the plants were growing at the time of inoculation with the fungus. The total uncorrected sum of squares is

$$\boldsymbol{Y}'\boldsymbol{Y} = 3.55^2 + 3.49^2 + \cdots + 1.195^2 = 47.2971.$$

The defining matrix for this quadratic form is the identity matrix of order 5. Since $\boldsymbol{I}$ is an idempotent matrix and $\mathrm{tr}(\boldsymbol{I}) = 5$, this sum of squares has 5 degrees of freedom.

The linear function of $\boldsymbol{Y}$ that gives the total disease score over all treatments is given by $\sum Y_i = \boldsymbol{a}_1'\boldsymbol{Y}$, where

$$\boldsymbol{a}_1' = (\,1 \quad 1 \quad 1 \quad 1 \quad 1\,)'.$$

The sum of squares due to correction for the mean, the correction factor, is $(\sum Y_i)^2/5 = 43.0124$. This is written as a quadratic form as

$$\boldsymbol{Y}'(\boldsymbol{J}/5)\boldsymbol{Y},$$

where $\boldsymbol{J} = \boldsymbol{a}_1\boldsymbol{a}_1'$ is a 5×5 matrix of ones. The defining matrix $\boldsymbol{J}/5$ is an idempotent matrix with $\text{tr}(\boldsymbol{J}/5) = 1$. Therefore, the sum of squares due to correction for the mean has 1 degree of freedom.

Based on orthogonal polynomial coefficients for five equally spaced treatments, the linear contrast for temperature effects is given by

$$C_2^* = \boldsymbol{a}_2^{*'}\boldsymbol{Y} = (-2 \quad -1 \quad 0 \quad 1 \quad 2)\,\boldsymbol{Y}.$$

Incorporating the divisor $\sqrt{\boldsymbol{a}_2^{*'}\boldsymbol{a}_2^*} = \sqrt{10}$ into the vector of coefficients gives

$$\boldsymbol{a}_2 = \left(-\tfrac{2}{\sqrt{10}} \quad -\tfrac{1}{\sqrt{10}} \quad 0 \quad \tfrac{1}{\sqrt{10}} \quad \tfrac{2}{\sqrt{10}}\right)'.$$

The sum of squares due to the linear regression on temperature is given by the quadratic form

$$\boldsymbol{Y}'\boldsymbol{A}_2\boldsymbol{Y} = 2.9594,$$

where

$$\boldsymbol{A}_2 = \boldsymbol{a}_2\boldsymbol{a}_2' = \begin{bmatrix} .4 & .2 & 0 & -.2 & -.4 \\ .2 & .1 & 0 & -.1 & -.2 \\ 0 & 0 & 0 & 0 & 0 \\ -.2 & -.1 & 0 & .1 & .2 \\ -.4 & -.2 & 0 & .2 & .4 \end{bmatrix}.$$

The defining matrix $\boldsymbol{A}_2$ is idempotent with $\text{tr}(\boldsymbol{A}_2) = 1$ and, therefore, the sum of squares has 1 degree of freedom.

The orthogonal polynomial coefficients for the quadratic term, including division by the square root of the sum of squares of the coefficients, is

$$\boldsymbol{a}_3 = \frac{1}{\sqrt{14}}\,(2 \quad -1 \quad -2 \quad -1 \quad 2)'.$$

The sum of squares due to quadratic regression is given by the quadratic form

$$\boldsymbol{Y}'\boldsymbol{A}_3\boldsymbol{Y} = 1.2007,$$

where

$$\boldsymbol{A}_3 = \boldsymbol{a}_3\boldsymbol{a}_3' = \begin{bmatrix} .2857 & -.1429 & -.2857 & -.1429 & .2857 \\ -.1429 & .0714 & .1429 & .0714 & -.1429 \\ -.2857 & .1429 & .2857 & .1429 & -.2857 \\ -.1429 & .0714 & .1429 & .0714 & -.1429 \\ .2857 & -.1429 & -.2857 & -.1429 & .2857 \end{bmatrix}.$$

The defining matrix $\boldsymbol{A}_3$ is idempotent and $\text{tr}(\boldsymbol{A}_3) = 1$ so that this sum of squares also has 1 degree of freedom.

It is left to the reader to verify that each of the defining matrices $\boldsymbol{J}/5$, $\boldsymbol{A}_2$, and $\boldsymbol{A}_3$ is idempotent and that they are pairwise orthogonal to each

other. Since they are orthogonal to each other, these three sums of squares represent independent pieces of information. However, they are *not* orthogonal to the uncorrected total sum of squares; the defining matrix $\boldsymbol{I}$ is not orthogonal to $\boldsymbol{J}/5$, $\boldsymbol{A}_2$, or $\boldsymbol{A}_3$. In fact, as is known from your previous experience, the sums of squares defined by $\boldsymbol{J}/5$, $\boldsymbol{A}_2$, and $\boldsymbol{A}_3$ are part of the total uncorrected sum of squares.

We could continue the partitioning of the uncorrected total sum of squares by defining two other mutually orthogonal idempotent matrices, say $\boldsymbol{A}_4$ and $\boldsymbol{A}_5$, that have rank one; are pairwise orthogonal to $\boldsymbol{J}/5$, $\boldsymbol{A}_2$, and $\boldsymbol{A}_3$; and for which the sum of all five matrices is $\boldsymbol{I}$. The sums of squares defined by these five matrices would form a complete single degree of freedom partitioning of the total uncorrected sum of squares $\boldsymbol{Y}'\boldsymbol{Y}$. ■

4.2 Analysis of Variance

The vector of observations on the dependent variable Y was partitioned in Chapter 3 into the vector of estimated means of Y, $\widehat{\boldsymbol{Y}}$and the residuals vector $\boldsymbol{e}$. That is,

$$\boldsymbol{Y} = \widehat{\boldsymbol{Y}} + \boldsymbol{e}. \tag{4.12}$$

This partitioning of $\boldsymbol{Y}$ is used to provide a similar partitioning of the total sum of squares of the dependent variable.

Partitioning of $\boldsymbol{Y}'\boldsymbol{Y}$

It has been previously noted that the product

$$\boldsymbol{Y}'\boldsymbol{Y} = \sum Y_i^2 \tag{4.13}$$

gives the total sum of squares SS(Total) of the elements in the column vector $\boldsymbol{Y}$. This is a quadratic form where the defining matrix is the identity matrix $\boldsymbol{Y}'\boldsymbol{Y} = \boldsymbol{Y}'\boldsymbol{I}\boldsymbol{Y}$. The matrix $\boldsymbol{I}$ is an idempotent matrix and its trace is equal to its order, indicating that the total (uncorrected) sum of squares has degrees of freedom equal to the number of elements in the vector. The identity matrix is the *only* full rank idempotent matrix.

Since $\boldsymbol{Y} = \widehat{\boldsymbol{Y}} + \boldsymbol{e}$,

$$\boldsymbol{Y}'\boldsymbol{Y} = (\widehat{\boldsymbol{Y}} + \boldsymbol{e})'(\widehat{\boldsymbol{Y}} + \boldsymbol{e}) = \widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{Y}}'\boldsymbol{e} + \boldsymbol{e}'\widehat{\boldsymbol{Y}} + \boldsymbol{e}'\boldsymbol{e}.$$

Substituting $\widehat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{Y}$ and $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$ gives

$$\begin{aligned}\boldsymbol{Y}'\boldsymbol{Y} &= (\boldsymbol{P}\boldsymbol{Y})'(\boldsymbol{P}\boldsymbol{Y}) + (\boldsymbol{P}\boldsymbol{Y})'[(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}] + [(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}]'(\boldsymbol{P}\boldsymbol{Y}) \\ &\quad + [(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}]'[(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}] \\ &= \boldsymbol{Y}'\boldsymbol{P}'\boldsymbol{P}\boldsymbol{Y} + \boldsymbol{Y}'\boldsymbol{P}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} + \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{P})'\boldsymbol{P}\boldsymbol{Y} \\ &\quad + \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{P})'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}. \end{aligned} \tag{4.14}$$

Both $\boldsymbol{P}$ and $(\boldsymbol{I}-\boldsymbol{P})$ are symmetric and idempotent so that $\boldsymbol{P}'\boldsymbol{P}=\boldsymbol{P}$ and $(\boldsymbol{I}-\boldsymbol{P})'(\boldsymbol{I}-\boldsymbol{P})=(\boldsymbol{I}-\boldsymbol{P})$. The two middle terms in equation 4.14 are zero since the two quadratic forms are orthogonal to each other:

$$\boldsymbol{P}'(\boldsymbol{I}-\boldsymbol{P})=\boldsymbol{P}-\boldsymbol{P}=\boldsymbol{0}.$$

Thus,

$$\boldsymbol{Y}'\boldsymbol{Y} \;=\; \boldsymbol{Y}'\boldsymbol{P}\boldsymbol{Y}+\boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}=\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}}+\boldsymbol{e}'\boldsymbol{e}. \tag{4.15}$$

The total uncorrected sum of squares has been partitioned into two quadratic forms with defining matrices $\boldsymbol{P}$ and $(\boldsymbol{I}-\boldsymbol{P})$, respectively. $\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}}$ is that part of $\boldsymbol{Y}'\boldsymbol{Y}$ that can be attributed to the model being fit and is labeled SS(Model). The second term $\boldsymbol{e}'\boldsymbol{e}$ is that part of $\boldsymbol{Y}'\boldsymbol{Y}$ not explained by the model. It is the residual sum of squares after fitting the model and is labeled SS(Res).

Degrees of Freedom

The orthogonality of the quadratic forms ensures that SS(Model) and SS(Res) are additive partitions. The degrees of freedom associated with each will depend on the rank of the defining matrices. The rank of $\boldsymbol{P}=[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']$ is determined by the rank of $\boldsymbol{X}$. For full-rank models, the rank of $\boldsymbol{X}$ is equal to the number of columns in $\boldsymbol{X}$, which is also the number of parameters in $\boldsymbol{\beta}$. Thus, the degrees of freedom for SS(Model) is p' when the model is of full rank.

The $r(\boldsymbol{P})$ is also given by $\mathrm{tr}(\boldsymbol{P})$ since $\boldsymbol{P}$ is idempotent. A result from matrix algebra states that $\mathrm{tr}(\boldsymbol{A}\boldsymbol{B})=\mathrm{tr}(\boldsymbol{B}\boldsymbol{A})$. (See Exercise 2.24.) Note the rotation of the matrices in the product. Using this property, with $\boldsymbol{A}=\boldsymbol{X}$ and $\boldsymbol{B}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ we have

$$\begin{aligned}\mathrm{tr}(\boldsymbol{P}) &= \mathrm{tr}[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']=\mathrm{tr}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}]\\ &= \mathrm{tr}(\boldsymbol{I}_{p'})=p'.\end{aligned} \tag{4.16}$$

The subscript on $\boldsymbol{I}$ indicates the order of the identity matrix. The degrees of freedom of SS(Res), $n-p'$, are obtained by noting the additivity of the two partitions or by observing that the $\mathrm{tr}(\boldsymbol{I}-\boldsymbol{P})=\mathrm{tr}(\boldsymbol{I}_n)-\mathrm{tr}(\boldsymbol{P})=(n-p')$. The order of this identity matrix is n. For each sum of squares, the corresponding **mean square** is obtained by dividing the sum of squares by its degrees of freedom.

Computational Forms

The expressions for the quadratic forms, equation 4.15, are the definitional forms; they show the nature of the sums of squares being computed. There are, however, more convenient computational forms. The computational form for SS(Model) $=\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}}$ is

$$\text{SS(Model)} \;=\; \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}, \tag{4.17}$$

and is obtained by substituting $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ for the the first $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ for the second. Thus, the sum of squares due to the model can be computed

TABLE 4.1. *Analysis of variance summary for regression analysis.*

Source of Variation	*Degrees of Freedom*	*Sum of Squares*	
		Definitional Formula	*Computational Formula*
$\text{Total}_{(\text{uncorr})}$	$r(\boldsymbol{I}) = n$	$\boldsymbol{Y}'\boldsymbol{Y}$	
Due to model	$r(\boldsymbol{P}) = p'$	$\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}} = \boldsymbol{Y}'\boldsymbol{P}\boldsymbol{Y}$	$\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}$
Residual	$r(\boldsymbol{I}-\boldsymbol{P}) = (n-p')$	$\boldsymbol{e}'\boldsymbol{e} = \boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$	$\boldsymbol{Y}'\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}$

without computing the vector of fitted values or the $n \times n$ matrix $\boldsymbol{P}$. The $\widehat{\boldsymbol{\beta}}$ vector is much smaller than $\widehat{\boldsymbol{Y}}$, and $\boldsymbol{X}'\boldsymbol{Y}$ will have already been computed. Since the two partitions are additive, the simplest computational form for SS(Res)= $\boldsymbol{e}'\boldsymbol{e}$ is by subtraction:

$$\text{SS(Res)} \quad = \quad \boldsymbol{Y}'\boldsymbol{Y} - \text{SS(Model)}. \tag{4.18}$$

The definitional and computational forms for this partitioning of the total sum of squares are summarized in Table 4.1.

Example 4.2

(*Continuation of Example* 3.8) The partitioning of the sums of squares is illustrated using the Heagle ozone example (Table 3.1, page 82). The total uncorrected sum of squares with four degrees of freedom is

$$\begin{aligned} \boldsymbol{Y}'\boldsymbol{Y} &= \begin{pmatrix} 242 & 237 & 231 & 201 \end{pmatrix} \begin{pmatrix} 242 \\ 237 \\ 231 \\ 201 \end{pmatrix} \\ &= 242^2 + 237^2 + 231^2 + 201^2 = 208,495. \end{aligned}$$

The sum of squares attributable to the model, SS(Model), can be obtained from the definitional formula, using $\widehat{\boldsymbol{Y}}$ from Table 3.1, as

$$\begin{aligned} \widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}} &= \begin{pmatrix} 247.563 & 232.887 & 221.146 & 209.404 \end{pmatrix} \begin{pmatrix} 247.563 \\ 232.887 \\ 221.146 \\ 209.404 \end{pmatrix} \\ &= 247.563^2 + 232.887^2 + 221.146^2 + 209.404^2 \\ &= 208,279.39. \end{aligned}$$

The more convenient computational formula gives

$$\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} \quad = \quad \begin{pmatrix} 253.434 & -293.531 \end{pmatrix} \begin{pmatrix} 911 \\ 76.99 \end{pmatrix} = 208,279.39.$$

(See the text following equation 3.12 for $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{X}'\boldsymbol{Y}$.)

The definitional formula for the residual sum of squares (see Table 3.1 for $\boldsymbol{e}$) gives

$$\begin{aligned}\boldsymbol{e}'\boldsymbol{e} &= \begin{pmatrix} -5.563 & 4.113 & 9.854 & -8.404 \end{pmatrix} \begin{pmatrix} -5.563 \\ 4.113 \\ 9.854 \\ -8.404 \end{pmatrix} \\ &= 215.61.\end{aligned}$$

The simpler computational formula gives

$$\begin{aligned}\text{SS(Res)} &= \boldsymbol{Y}'\boldsymbol{Y} - \text{SS(Model)} = 208,495 - 208,279.39 \\ &= 215.61.\end{aligned}$$

■

Meaning of SS(Regr)

The total *uncorrected* sum of squares has been partitioned into that due to the entire model and a residual sum of squares. Usually, however, one is interested in explaining the variation of Y about its mean, rather than about zero, and in how much the information from the independent variables contributes to this explanation. If no information is available from independent variables, the best predictor of Y is the best available estimate of the population mean. When independent variables are available, the question of interest is how much information the independent variables contribute to the prediction of Y beyond that provided by the overall mean of Y.

The measure of the additional information provided by the independent variables is the difference between SS(Model) when the independent variables are included and SS(Model) when no independent variables are included. The model with no independent variables contains only one parameter, the overall mean μ. When μ is the only parameter in the model, SS(Model) is labeled SS(μ). [SS(μ) is commonly called the **correction factor**.] The *additional* sum of squares accounted for by the independent variable(s) is called the regression sum of squares and labeled SS(Regr). Thus,

$$\text{SS(Regr)} = \text{SS(Model)} - \text{SS}(\mu), \tag{4.19}$$

where SS(Model) is understood to be the sum of squares due to the model containing the independent variables.

SS(μ)

The sum of squares due to μ alone, SS(μ), is determined using matrix notation in order to show the development of the defining matrices for the quadratic forms. The model when μ is the only parameter is still written in the form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, but now $\boldsymbol{X}$ is only a column vector of ones and

$\boldsymbol{\beta} = \mu$, a single element. The column vector of ones is labeled $\mathbf{1}$. Then,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{Y} = \left(\frac{1}{n}\right)\mathbf{1}'\boldsymbol{Y} = \overline{Y} \tag{4.20}$$

and

$$\begin{aligned} \mathrm{SS}(\mu) &= \widehat{\boldsymbol{\beta}}'(\mathbf{1}'\boldsymbol{Y}) = \left(\frac{1}{n}\right)(\mathbf{1}'\boldsymbol{Y})'(\mathbf{1}'\boldsymbol{Y}) \\ &= \boldsymbol{Y}'(\frac{1}{n}\mathbf{1}\mathbf{1}')\boldsymbol{Y}. \end{aligned} \tag{4.21}$$

Notice that $\mathbf{1}'\boldsymbol{Y} = \sum Y_i$ so that $\mathrm{SS}(\mu)$ is $(\sum Y_i)^2/n$, the familiar result for the sum of squares due to correcting for the mean. Multiplication of $\mathbf{1}\mathbf{1}'$ gives an $n \times n$ matrix of ones. Convention labels this the $\boldsymbol{J}$ matrix. Thus, the defining matrix for the quadratic form giving the correction factor is

$$\frac{1}{n}(\mathbf{1}\mathbf{1}') = \frac{1}{n}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix} = \frac{1}{n}\boldsymbol{J}. \tag{4.22}$$

The matrix $(\boldsymbol{J}/\mathrm{n})$ is idempotent with rank equal to $\mathrm{tr}(\boldsymbol{J}/\mathrm{n}) = 1$ and, hence, the correction factor has 1 degree of freedom.

Quadratic form for SS(Regr)

The *additional* sum of squares attributable to the independent variable(s) in a model is then

$$\begin{aligned} \mathrm{SS(Regr)} &= \mathrm{SS(Model)} - \mathrm{SS}(\mu) \\ &= \boldsymbol{Y}'\boldsymbol{P}\boldsymbol{Y} - \boldsymbol{Y}'(\boldsymbol{J}/n)\boldsymbol{Y} \\ &= \boldsymbol{Y}'(\boldsymbol{P} - \boldsymbol{J}/n)\boldsymbol{Y}. \end{aligned} \tag{4.23}$$

Thus, the defining matrix for SS(Regr) is $(\boldsymbol{P} - \boldsymbol{J}/n)$. The defining matrix $\boldsymbol{J}/n$ is orthogonal to $(\boldsymbol{P} - \boldsymbol{J}/n)$ and $(\boldsymbol{I} - \boldsymbol{P})$ (see exercise 4.15) so that the total sum of squares is now partitioned into three orthogonal components:

$$\begin{aligned} \boldsymbol{Y}'\boldsymbol{Y} &= \boldsymbol{Y}'(\boldsymbol{J}/n)\boldsymbol{Y} + \boldsymbol{Y}'(\boldsymbol{P} - \boldsymbol{J}/n)\boldsymbol{Y} + \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} \\ &= \mathrm{SS}(\mu) + \mathrm{SS(Regr)} + \mathrm{SS(Res)} \end{aligned} \tag{4.24}$$

with 1, $(p' - 1) = p$, and $(n - p')$ degrees of freedom, respectively. Usually $\mathrm{SS}(\mu)$ is subtracted from $\boldsymbol{Y}'\boldsymbol{Y}$ and only the *corrected* sum of squares partitioned into SS(Regr) and SS(Res) reported.

Example 4.3

For the Heagle ozone example, Example 4.2,

$$\mathrm{SS}(\mu) = \frac{(911)^2}{4} = 207,480.25$$

TABLE 4.2. *Summary analysis of variance for the regression of soybean yield on ozone exposure (Data courtesy A. S. Heagle, N. C. State University).*

Source of Variation	*d.f.*	*Sum of Squares*			*Mean Squares*
$\text{Total}_{\text{uncorr}}$	4	$\boldsymbol{Y}'\boldsymbol{Y}$	=	208,495.00	
Mean	1	$n\overline{Y}^2$	=	207,480.25	
$\text{Total}_{\text{corr}}$	3	$\boldsymbol{Y}'\boldsymbol{Y} - n\overline{Y}^2$	=	1,014.75	
Regression	1	$\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} - n\overline{Y}^2$	=	799.14	799.14
Residuals	2	$\boldsymbol{Y}'\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}$	=	215.61	107.81

so that

$$\text{SS(Regr)} = 208,279.39 - 207,480.25 = 799.14.$$

The analysis of variance for this example is summarized in Table 4.2. ■

The key points to remember are summarized in the following.

- The rank of $\boldsymbol{X}$ is equal to the number of linearly independent columns in $\boldsymbol{X}$.
- The model is a full rank model if the rank of $\boldsymbol{X}$ equals the number of columns of $\boldsymbol{X}$, $(n > p')$.
- The *unique* ordinary least squares solution exists only if the model is of full rank.
- The defining matrices for the quadratic forms in regression are all idempotent. Examples are $\boldsymbol{I}$, $\boldsymbol{P}$, $(\boldsymbol{I} - \boldsymbol{P})$, and $\boldsymbol{J}/n$.
- The defining matrices $\boldsymbol{J}/n$, $(\boldsymbol{P} - \boldsymbol{J}/n)$, and $(\boldsymbol{I} - \boldsymbol{P})$ are pairwise orthogonal to each other and sum to $\boldsymbol{I}$. Consequently, they partition the total uncorrected sum of squares into orthogonal sums of squares.
- The degrees of freedom for a quadratic form are determined by the rank of the defining matrix which, when it is idempotent, equals its trace. For a full rank model,

$$\begin{aligned} r(\boldsymbol{I}) &= n, \quad \text{the only full rank idempotent matrix} \\ r(\boldsymbol{P}) &= p' \\ r(\boldsymbol{J}/n) &= 1 \\ r(\boldsymbol{P} - \boldsymbol{J}/n) &= p \\ r(\boldsymbol{I} - \boldsymbol{P}) &= n - p'. \end{aligned}$$

4.3 Expectations of Quadratic Forms

Each of the quadratic forms computed in the analysis of variance of Y is estimating some function of the parameters of the model. The expectations of these quadratic forms must be known if proper use is to be made of the sums of squares and their mean squares. The following results are stated without proofs. The reader is referred to Searle (1971) for more complete development.

General Results

Let $\mathcal{E}(\boldsymbol{Y}) = \boldsymbol{\mu}$, a general vector of expectations, and let $\mathbf{Var}(\boldsymbol{Y}) = \boldsymbol{V}_y = \boldsymbol{V}\sigma^2$, a general variance–covariance matrix. Then the general result for the expectation of the quadratic form $\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}$ is

$$\begin{aligned} \mathcal{E}(\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}) &= \mathrm{tr}(\boldsymbol{A}\boldsymbol{V}_y) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu} \\ &= \sigma^2\mathrm{tr}(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}. \end{aligned} \tag{4.25}$$

Under ordinary least squares assumptions, $\mathcal{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ and $\mathbf{Var}(\boldsymbol{Y}) = \boldsymbol{I}\sigma^2$ and the expectation of the quadratic form becomes

$$\mathcal{E}(\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}) = \sigma^2\mathrm{tr}(\boldsymbol{A}) + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}. \tag{4.26}$$

The expectations of the quadratic forms in the analysis of variance are obtained from this general result by replacing $\boldsymbol{A}$ with the appropriate defining matrix. When $\boldsymbol{A}$ is idempotent, the coefficient on σ^2 is the degrees of freedom for the quadratic form.

$\mathcal{E}$[SS(Model)]

The expectation of SS(Model) is

$$\begin{aligned} \mathcal{E}[\mathrm{SS(Model)}] &= \mathcal{E}(\boldsymbol{Y}'\boldsymbol{P}\boldsymbol{Y}) = \sigma^2\mathrm{tr}(\boldsymbol{P}) + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta} \\ &= p'\sigma^2 + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}, \end{aligned} \tag{4.27}$$

since $\mathrm{tr}(\boldsymbol{P}) = p'$ and $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$. Notice that the second term in equation 4.27 is a quadratic form in $\boldsymbol{\beta}$, including β_0 the intercept.

$\mathcal{E}$[SS(Regr)]

The expectation for SS(Regr) is

$$\begin{aligned} \mathcal{E}[\mathrm{SS(Regr)}] &= \mathcal{E}[\boldsymbol{Y}'(\boldsymbol{P} - \boldsymbol{J}/n)\boldsymbol{Y}] \\ &= \sigma^2\mathrm{tr}(\boldsymbol{P} - \boldsymbol{J}/n) + \boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{P} - \boldsymbol{J}/n)\boldsymbol{X}\boldsymbol{\beta} \\ &= p\sigma^2 + \boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{J}/n)\boldsymbol{X}\boldsymbol{\beta}, \end{aligned} \tag{4.28}$$

since $\boldsymbol{X}'\boldsymbol{P} = \boldsymbol{X}'$. This quadratic form in $\boldsymbol{\beta}$ differs from that for $\mathcal{E}$[SS(Model)] in that $\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{J}/n)\boldsymbol{X}$ is a matrix of *corrected* sums of squares and products of the X_j. Since the first column of $\boldsymbol{X}$ is a constant, the sums of squares and products involving the first column are zero. Thus, the first row and column of $\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{J}/n)\boldsymbol{X}$ contain only zeros, which removes β_0 from the quadratic expression (see Exercise 4.16). Only the regression coefficients for the independent variables are involved in the expectation of the regression sum of squares.

$\mathcal{E}$[SS(Res)]

The expectation for SS(Res) is

$$\begin{aligned}\mathcal{E}[\text{SS(Res)}] &= \mathcal{E}[\boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}] \\ &= \sigma^2\text{tr}(\boldsymbol{I}-\boldsymbol{P}) + \boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{X}\boldsymbol{\beta} \\ &= (n-p')\sigma^2 + \boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{X}-\boldsymbol{X})\boldsymbol{\beta} \\ &= (n-p')\sigma^2. \end{aligned} \tag{4.29}$$

Expectations of Mean Squares

The coefficient on σ^2 in each expectation is the degrees of freedom for the sum of squares. After division of each expectation by the appropriate degrees of freedom to convert sums of squares to mean squares, the coefficient on σ^2 will be 1 in each case:

$$\mathcal{E}[\text{MS(Regr)}] = \sigma^2 + [\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{J}/n)\boldsymbol{X}\boldsymbol{\beta}]/p \tag{4.30}$$

$$\mathcal{E}[\text{MS(Res)}] = \sigma^2. \tag{4.31}$$

This shows that the residual mean square MS(Res) is an unbiased estimate of σ^2. The regression mean square MS(Regr) is an estimate of σ^2 plus a quadratic function of all β_j except β_0. Comparison of MS(Regr) and MS(Res), therefore, provides the basis for judging the importance of the regression coefficients or, equivalently, of the independent variables. Since the second term in $\mathcal{E}$[MS(Regr)] is a *quadratic* function of $\boldsymbol{\beta}$, which cannot be negative, any contribution from the independent variables to the predictability of Y_i makes MS(Regr) larger *in expectation* than MS(Res). The ratio of the observed MS(Regr) to the observed MS(Res) provides the test of significance of the composite hypothesis that all β_j, except β_0, are zero. Tests of significance are discussed more fully in the following sections.

The expectations assume that the model used in the analysis of variance is the correct model. This is imposed in the preceding derivations when $\boldsymbol{X}\boldsymbol{\beta}$ is substituted for $\mathcal{E}(\boldsymbol{Y})$. For example, if $\mathcal{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} \neq \boldsymbol{X}\boldsymbol{\beta}$, but we fit the model $\mathcal{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$, then

$$\begin{aligned}\mathcal{E}[\text{SS(Res)}] &= \sigma^2\text{tr}(\boldsymbol{I}-\boldsymbol{P}) + [\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{\gamma}]'(\boldsymbol{I}-\boldsymbol{P})[\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{\gamma}] \\ &= \sigma^2(n-p') + \boldsymbol{\gamma}'\boldsymbol{Z}'(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Z}\boldsymbol{\gamma} \end{aligned} \tag{4.32}$$

and

$$\mathcal{E}[\text{MS(Res)}] = \sigma^2 + \boldsymbol{\gamma}'\boldsymbol{Z}'(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Z}\boldsymbol{\gamma}/(n-p'). \tag{4.33}$$

The second term in equation 4.33 represents a quadratic function of regression coefficients of important variables that were mistakenly omitted from the model. From equation 4.33, it can be seen that MS(Res), in such cases, will be a positively biased estimate of σ^2.

Example 4.4

From Example 4.3 using the ozone data, the estimate of σ^2 obtained from MS(Res) is $s^2 = 107.81$ (Table 4.2). This is a very poor estimate of σ^2 since it has only two degrees of freedom. Nevertheless, this estimate of σ^2 is used for now. (A better estimate is obtained in Section 4.7.) ■

Estimated Variances

In Chapter 3, the variance–covariance matrices for $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{Y}}$, and $\boldsymbol{e}$ were expressed in terms of the true variance σ^2. *Estimates* of the variance–covariance matrices are obtained by substituting $s^2 = 107.81$ for σ^2 in each $\mathbf{Var}(\cdot)$ formula; $\boldsymbol{s^2}(\cdot)$ is used to denote an *estimated* variance–covariance matrix. (Note the boldface type to distinguish the matrix of estimates from individual variances.)

Example 4.5

In the ozone example, Example 4.3,

$$\begin{aligned}
\boldsymbol{s^2}(\widehat{\boldsymbol{\beta}}) &= (\boldsymbol{X'X})^{-1}s^2 \\
&= \begin{bmatrix} 1.0755 & -9.4340 \\ -9.4340 & 107.8167 \end{bmatrix} 107.81 \\
&= \begin{bmatrix} 115.94 & -1,017.0 \\ -1,017.0 & 11,623 \end{bmatrix}.
\end{aligned}$$

Thus,

$$\begin{aligned}
s^2(\widehat{\beta}_0) &= (1.0755)(107.81) = 115.94, \\
s^2(\widehat{\beta}_1) &= (107.8167)(107.81) = 11,623, \text{ and} \\
\mathrm{Cov}(\widehat{\beta}_0,\ \widehat{\beta}_1) &= (-9.4340)(107.81) = -1,017.0.
\end{aligned}$$

In each case, the first number in the product is the appropriate coefficient from the $(\boldsymbol{X'X})^{-1}$ matrix; the second number is s^2. (It is only coincidence that the lower right diagonal element of $(\boldsymbol{X'X})^{-1}$ is almost identical to s^2.) ■

The estimated variance–covariance matrices for $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{e}$ are found similarly by replacing σ^2 with s^2 in the corresponding variance–covariance matrices.

4.4 Distribution of Quadratic Forms

The probability distributions of the quadratic forms provide the basis for parametric tests of significance. It is at this point (and in making confidence interval statements about the parameters) that the normality assumption on the ϵ_i comes into play. The results are summarized assuming that normality of $\boldsymbol{\epsilon}$ and therefore normality of $\boldsymbol{Y}$ are satisfied. When normality is not satisfied, the parametric tests of significance must be regarded as approximations.

A general result from statistical theory [see, for example, Searle (1971)] states:

If $\boldsymbol{Y}$ is normally distributed, with $\mathcal{E}(\boldsymbol{Y}) = \boldsymbol{\mu}$ and $\mathbf{Var}(\boldsymbol{Y}) = \boldsymbol{V}\sigma^2$, where $\boldsymbol{V}$ is a nonsingular matrix ($\boldsymbol{\mu}$ may be $\boldsymbol{X\beta}$ and $\boldsymbol{V}$ may be $\boldsymbol{I}$), then

1. a quadratic form $\boldsymbol{Y}'(\boldsymbol{A}/\sigma^2)\boldsymbol{Y}$ is distributed as a **noncentral chi-square** with

   (a) degrees of freedom equal to the rank of $\boldsymbol{A}$, $df = r(\boldsymbol{A})$, and

   (b) noncentrality parameter $\Omega = (\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu})/2\sigma^2$,

   if $\boldsymbol{AV}$ is idempotent (if $\boldsymbol{V} = \boldsymbol{I}$, the condition reduces to $\boldsymbol{A}$ being idempotent);
2. quadratic forms $\boldsymbol{Y}'\boldsymbol{AY}$ and $\boldsymbol{Y}'\boldsymbol{BY}$ are independent of each other if $\boldsymbol{AVB} = \boldsymbol{0}$ (if $\boldsymbol{V} = \boldsymbol{I}$, the condition reduces to $\boldsymbol{AB} = \boldsymbol{0}$; that is, $\boldsymbol{A}$ and $\boldsymbol{B}$ are orthogonal to each other); and
3. a quadratic function $\boldsymbol{Y}'\boldsymbol{AY}$ is independent of a linear function $\boldsymbol{BY}$ if $\boldsymbol{BVA} = \boldsymbol{0}$. (If $\boldsymbol{V} = \boldsymbol{I}$, the condition reduces to $\boldsymbol{BA} = \boldsymbol{0}$.)

Application to Regression

In the normal multiple regression model, the following hold.

1. The sums of squares for model, mean, regression, and residuals all involve defining matrices that are idempotent. Recall that

$$\text{SS(Model)}/\sigma^2 \quad = \quad \boldsymbol{Y}'\boldsymbol{PY}/\sigma^2.$$

   Since $\boldsymbol{P}$ is idempotent, $\text{SS(Model)}/\sigma^2$ is distributed as a chi-square random variable with $r(\boldsymbol{P}) = p'$ degrees of freedom and noncentrality parameter

$$\Omega \quad = \quad \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{PX}\boldsymbol{\beta}/2\sigma^2 = \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}/2\sigma^2.$$

   Similarly:

   (a) $\text{SS}(\mu)/\sigma^2 = \boldsymbol{Y}'(\boldsymbol{J}/n)\boldsymbol{Y}/\sigma^2$ is distributed as a chi-square random variable with $r(\boldsymbol{J}/n) = 1$ degree of freedom and noncentrality parameter

$$\Omega \quad = \quad \boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{J}/n)\boldsymbol{X}\boldsymbol{\beta}/2\sigma^2 = (\mathbf{1}'\boldsymbol{X}\boldsymbol{\beta})^2/2n\sigma^2.$$

   (b) $\text{SS(Regr)}/\sigma^2 = \boldsymbol{Y}'(\boldsymbol{P}-\boldsymbol{J}/n)\boldsymbol{Y}/\sigma^2$ is distributed as a chi-square random variable with $r(\boldsymbol{P}-\boldsymbol{J}/n) = p$ (see Exercise 4.15) degrees of freedom and noncentrality parameter

$$\Omega \quad = \quad [\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{P} - \boldsymbol{J}/n)\boldsymbol{X}\boldsymbol{\beta}]/2\sigma^2 = [\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{J}/n)\boldsymbol{X}\boldsymbol{\beta}]/2\sigma^2.$$

(c) $\text{SS(Res)}/\sigma^2 = \boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}/\sigma^2$ is distributed as a chi-square random variable with $r(\boldsymbol{I}-\boldsymbol{P}) = (n-p')$ degrees of freedom and noncentrality parameter

$$\Omega \quad = \quad \boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{X}\boldsymbol{\beta}/2\sigma^2 = 0.$$

That is, $\text{SS(Res)}/\sigma^2$ has a *central* chi-square distribution with degrees of freedom $(n-p')$. (A **central** chi-square distribution has noncentrality parameter equal to zero.)

2. Since $(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{P}-\boldsymbol{J}/n) = 0$ (see Exercise 4.15), $\text{SS(Res)} = \boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$ and $\text{SS(Regr)} = \boldsymbol{Y}'(\boldsymbol{P}-\boldsymbol{J}/n)\boldsymbol{Y}$ are independent. Similarly, since $\boldsymbol{P}(\boldsymbol{I}-\boldsymbol{P}) = 0$, $\boldsymbol{J}/n(\boldsymbol{P}-\boldsymbol{J}/n) = 0$, and $\boldsymbol{J}/n(\boldsymbol{I}-\boldsymbol{P}) = 0$, we have that SS(Model) and SS(Res) are independent, $\text{SS}(\mu)$ and SS(Regr) are independent, and $\text{SS}(\mu)$ and SS(Res) are independent, respectively.

3. Since $\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{P}) = 0$, any linear function $\boldsymbol{K}'\widehat{\boldsymbol{\beta}} = \boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{Y}$ is independent of $\text{SS(Res)} = \boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$. This follows from noting that $\boldsymbol{B}(\boldsymbol{I}-\boldsymbol{P}) = \boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{P}) = \boldsymbol{0}$.

Thus, the normality assumption on $\boldsymbol{\epsilon}$ implies that the sums of squares, divided by σ^2, are chi-square random variables. The chi-square distribution and the orthogonality between the quadratic forms provide the basis for the usual tests of significance. For example, when the null hypothesis is true, the t-statistic is the ratio of a normal deviate to the square root of a scaled independent central chi-square random variable. The F-statistic is the ratio of a scaled noncentral chi-square random variable (central chi-square random variable if the null hypothesis is true) to a scaled independent central chi-square random variable. The scaling in each case is division of the chi-square random variable by its degrees of freedom.

Noncentrality Parameter and *F*-Test

The noncentrality parameter $\Omega = (\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu})/2\sigma^2$ is important for two reasons: the condition that makes the noncentrality parameter of the numerator of the F-ratio equal to zero is an explicit statement of the null hypothesis; and the power of the test to detect a false null hypothesis is determined by the magnitude of the noncentrality parameter. The noncentrality parameter of the chi-square distribution is the second term of the expectations of the quadratic forms divided by 2 (see equation 4.25). $\text{SS(Res)}/\sigma^2$ is a central chi-square since the second term was zero (equation 4.29). The noncentrality parameter for $\text{SS(Regr)}/\sigma^2$ (see equation 4.28) is

$$\Omega \quad = \quad \frac{\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{J}/n)\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2}, \tag{4.34}$$

which is a quadratic form involving all β_j except β_0. Thus, $\text{SS(Regr)}/\sigma^2$ is a *central* chi-square only if $\Omega = 0$, which requires $(\boldsymbol{I}-\boldsymbol{J}/n)\boldsymbol{X}\boldsymbol{\beta} = 0$. Since

$\boldsymbol{X}$ is assumed to be of full rank, it can be shown that $\Omega = 0$ if and only if $\beta_1 = \beta_2 = \cdots = \beta_p = 0$. Therefore, the F-ratio using

$$F = \frac{\text{MS(Regr)}}{\text{MS(Res)}}$$

is a test of the composite hypothesis that all β_j, except β_0, equal zero. This hypothesis is stated as

$$\begin{aligned} H_0 : \boldsymbol{\beta}^* &= 0 \\ H_a : \boldsymbol{\beta}^* &\neq 0, \end{aligned}$$

where $\boldsymbol{\beta}^*$ is the $p \times 1$ vector of regression coefficients excluding β_0.

An observed F-ratio, equation 4.35, sufficiently greater than 1 suggests that the noncentrality parameter is not zero. The larger the noncentrality parameter for the numerator chi-square, the larger will be the F-ratio, on the average, and the greater will be the probability of detecting a false null hypothesis. This probability, by definition, is the **power of the test**. (The power of an F-test is also increased by increasing the degrees of freedom for each chi-square, particularly the denominator chi-square.) All of the quantities except $\boldsymbol{\beta}$ in the noncentrality parameter are known *before* the experiment is run (in those cases where the Xs are subject to the control of the researcher). Therefore, the relative powers of different experimental designs can be evaluated before the final design is adopted.

Example 4.6

In the Heagle ozone example, Example 4.2,

$$F = \frac{\text{MS(Regr)}}{\text{MS(Res)}} = \frac{799.14}{107.81} = 7.41.$$

The critical value for $\alpha = .05$ with 1 and 2 degrees of freedom is $F_{(.05;1,2)} = 18.51$. The conclusion is that these data do not provide sufficient evidence to reject the null hypothesis that β_1 equals zero. Even though MS(Regr) is considerably larger than MS(Res), the difference is not sufficient to be confident that it is not due to random sampling variation from the underlying chi-square distributions. The large critical value of F, 18.51, is a direct reflection of the very limited degrees of freedom for MS(Res) and, consequently, large sampling variation in the F-distribution. A later analysis that uses a more precise estimate of σ^2 (more degrees of freedom) but the same MS(Regr) shows that β_1 clearly is not zero. ■

The key points from this section are summarized as follows.

1. The expectations of the quadratic forms are model dependent. If the incorrect model has been used, the expectations are incorrect. This is particularly critical for the

MS(Res) since it is used repeatedly as the estimate of σ^2. For this reason it is desirable to obtain an estimate of σ^2 that is not model dependent. This is discussed in Section 4.7.

2. The expectations of the mean squares provide the basis for choosing the appropriate mean squares for tests of hypotheses with the F-test; the numerator and denominator mean squares must have the same expectations if the null hypothesis is true and the expectation of the numerator mean square must be larger if the alternative hypothesis is true.
3. The assumption of a normal probability distribution for the residuals is necessary for the conventional tests of significance and confidence interval estimates of the parameters to be correct. Although tests of significance appear to be reasonably robust against nonnormality, they must be regarded as approximations when the normality assumption is not satisfied.

4.5 General Form for Hypothesis Testing

The ratio of MS(Regr) to MS(Res) provides a test of the null hypothesis that all β_j, except β_0, are simultaneously equal to zero. More flexibility is needed in constructing tests of hypotheses than is allowed by this procedure. This section presents a general method of constructing tests for any hypothesis involving linear functions of $\boldsymbol{\beta}$. The null hypothesis may involve a single linear function, a **simple hypothesis**, or it may involve several linear functions simultaneously, a **composite hypothesis**.

4.5.1 The General Linear Hypothesis

The general linear hypothesis is defined as

$$\begin{aligned} H_0 : \boldsymbol{K}'\boldsymbol{\beta} &= \boldsymbol{m} \\ H_a : \boldsymbol{K}'\boldsymbol{\beta} &\neq \boldsymbol{m}, \end{aligned} \tag{4.35}$$

where $\boldsymbol{K}'$ is a $k \times p'$ matrix of coefficients defining k linear functions of the β_j to be tested. Each row of $\boldsymbol{K}'$ contains the coefficients for one linear function; $\boldsymbol{m}$ is a $k \times 1$ vector of constants, frequently zeros. The k linear equations in H_0 must be linearly independent (but they need not be orthogonal). Linear independence implies that $\boldsymbol{K}'$ is of full rank, $r(\boldsymbol{K}) = k$, and ensures that the equations in H_0 are consistent for every choice of $\boldsymbol{m}$ (see Section

2.5). The number of linear functions in H_0 cannot exceed the number of parameters in $\boldsymbol{\beta}$; otherwise, $\boldsymbol{K}'$ would not be of rank k.

Example 4.7

Suppose $\boldsymbol{\beta}' = (\,\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3\,)$ and you wish to test the composite null hypothesis that $\beta_1 = \beta_2$, $\beta_1 + \beta_2 = 2\beta_3$, and $\beta_0 = 20$ or, equivalently,

$$\begin{array}{lrcl} H_0: & \beta_1 - \beta_2 & = & 0 \\ & \beta_1 + \beta_2 - 2\beta_3 & = & 0 \\ & \beta_0 & = & 20 \end{array} . \tag{4.36}$$

These three linear functions can be written in the form $\boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m}$ by defining

$$\boldsymbol{K}' = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 1 & -2 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{m} = \begin{pmatrix} 0 \\ 0 \\ 20 \end{pmatrix}. \tag{4.37}$$

The alternative hypothesis is $H_a : \boldsymbol{K}'\boldsymbol{\beta} \neq \boldsymbol{m}$. The null hypothesis is violated if any one or more of the equalities in H_0 is not true. ■

Estimator and Variance

The least squares estimate of $\boldsymbol{K}'\boldsymbol{\beta} - \boldsymbol{m}$ is obtained by substituting the least squares estimate $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ to obtain $\boldsymbol{K}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m}$. Under the ordinary least squares assumptions, including normality, $\boldsymbol{K}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m}$ is normally distributed with mean $\mathcal{E}(\boldsymbol{K}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m}) = \boldsymbol{K}'\boldsymbol{\beta} - \boldsymbol{m}$, which is zero if the null hypothesis is true, and variance–covariance matrix $\mathbf{Var}(\boldsymbol{K}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m}) = \boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}\sigma^2 = \boldsymbol{V}\sigma^2$, say. The variance is obtained by applying the rules for variances of linear functions (see Section 3.4).

Sum of Squares

The sum of squares for the linear hypothesis $H_0 : \boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m}$ is computed by [see Searle (1971)]

$$Q = (\boldsymbol{K}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m})'[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}]^{-1}(\boldsymbol{K}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m}). \tag{4.38}$$

This is a quadratic form in $\boldsymbol{K}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m}$ with defining matrix

$$\boldsymbol{A} = [\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}]^{-1} = \boldsymbol{V}^{-1}. \tag{4.39}$$

The defining matrix, except for division by σ^2, is the inverse of the variance–covariance matrix of the linear functions $\boldsymbol{K}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m}$. Thus, $\mathrm{tr}(\boldsymbol{A}\boldsymbol{V}) = \mathrm{tr}(\boldsymbol{I}_k) = k$ and the expectation of Q (see equation 4.25) is

$$\mathcal{E}(Q) = k\sigma^2 + (\boldsymbol{K}'\boldsymbol{\beta} - \boldsymbol{m})'[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}]^{-1}(\boldsymbol{K}'\boldsymbol{\beta} - \boldsymbol{m}). \tag{4.40}$$

With the assumption of normality, Q/σ^2 is distributed as a noncentral chi-square random variable with k degrees of freedom. This is verified by noting

that $\boldsymbol{AV} = \boldsymbol{I}_k$, which is idempotent (see Section 4.4). The degrees of freedom are determined from $\mathrm{r}(\boldsymbol{A}) = \mathrm{r}(\boldsymbol{K}) = k$. The noncentrality parameter is

$$\Omega = \frac{(\boldsymbol{K}'\boldsymbol{\beta} - \boldsymbol{m})'[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}]^{-1}(\boldsymbol{K}'\boldsymbol{\beta} - \boldsymbol{m})}{2\sigma^2},$$

which is zero when the null hypothesis is true. Thus, Q/k is an appropriate numerator mean square for an F-test of the stated hypothesis.

F-Test

The appropriate denominator of the F-test is any unbiased and independent estimate of σ^2; usually MS(Res) is used. Thus,

$$F = \frac{Q/\mathrm{r}(\boldsymbol{K})}{s^2} \tag{4.41}$$

is a proper F-test of $H_0 : \boldsymbol{K}'\boldsymbol{\beta} - \boldsymbol{m} = 0$ with numerator degrees of freedom equal to $\mathrm{r}(\boldsymbol{K})$ and denominator degrees of freedom equal to the degrees of freedom in s^2. Since $\boldsymbol{K}'\widehat{\boldsymbol{\beta}}$ is independent of SS(Res), Q is independent of MS(Res).

This general formulation provides a convenient method for testing any hypotheses of interest and is particularly useful when the computations are being done with a matrix algebra computer program. It is important to note, however, that all sums of squares for hypotheses are dependent on the particular model being used. In general, deleting an independent variable or adding an independent variable to the model will change the sum of squares for every hypothesis.

4.5.2 Special Cases of the General Form

Three special cases of the general linear hypothesis are of interest.

Case 1. A simple hypothesis.
When a **simple hypothesis** on $\boldsymbol{\beta}$ is being tested, $\boldsymbol{K}'$ is a single row vector so that $[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}]$ is a scalar. Its inverse is $1/[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}]$. The sum of squares for the hypothesis can be written as

$$Q = \frac{(\boldsymbol{K}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m})^2}{\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}} \tag{4.42}$$

and has 1 degree of freedom. The numerator of Q is the square of the linear function of $\widehat{\boldsymbol{\beta}}$ and the denominator is its variance, except for σ^2. Thus, the F-ratio is

$$F = \frac{(\boldsymbol{K}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m})^2}{[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}]s^2}. \tag{4.43}$$

The F-test of a simple hypothesis is the square of a two-tailed t-test:

$$t = \frac{\boldsymbol{K}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m}}{\{[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}]s^2\}^{1/2}}. \tag{4.44}$$

The denominator is the standard error of the linear function in the numerator.

Case 2. k specific β_j equal zero.
The null hypothesis of interest is that each of k specific regression coefficients is zero. For this case $\boldsymbol{K}'$ is a $k \times p'$ matrix consisting of zeros except for a single one in each row to identify the β_j being tested; $\boldsymbol{m} = \boldsymbol{0}$. With this $\boldsymbol{K}'$, the matrix multiplication $[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}]$ extracts from $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ the $k \times k$ submatrix consisting of the coefficients for the variances and covariances of the k $\widehat{\beta}_j$ being tested. Suppose the null hypothesis to be tested is that β_1, β_3, and β_5 are each equal to zero. The sum of squares Q has the form

$$Q = \begin{pmatrix} \widehat{\beta}_1 & \widehat{\beta}_3 & \widehat{\beta}_5 \end{pmatrix} \begin{bmatrix} c_{11} & c_{13} & c_{15} \\ c_{31} & c_{33} & c_{35} \\ c_{51} & c_{53} & c_{55} \end{bmatrix}^{-1} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_3 \\ \widehat{\beta}_5 \end{pmatrix}, \tag{4.45}$$

where c_{ij} is the element from row $(i+1)$ and column $(j+1)$ of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

The sum of squares for this hypothesis measures the contribution of this subset of k independent variables to a model that already contains the other independent variables. This sum of squares is described as the sum of squares for these k variables **adjusted for** the other independent variables in the model.

Case 3. One β_j equals zero; the partial sum of squares.
Partial Sum of Squares
The third case is a further simplification of the first two. The hypothesis is that a single β_j is zero; $H_0 : \beta_j = 0$. For this hypothesis, $\boldsymbol{K}'$ is a row vector of zeros except for a one in the column corresponding to the β_j being tested. As described in case 2, the sum of squares for this hypothesis is the contribution of X_j *adjusted for* all other independent variables in the model. This sum of squares is called the **partial sum of squares** for the jth independent variable.

The matrix multiplication $[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}]$ in Q extracts only the $(j+1)$st diagonal element c_{jj} from $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. This is the coefficient for the variance of $\widehat{\beta}_j$. The sum of squares, with one degree of freedom, is

$$Q = \frac{\widehat{\beta}_j^2}{c_{jj}}. \tag{4.46}$$

This provides an easy method of computing the *partial* sum of squares for any independent variable. For this case, the two-tailed t-test is

$$t = \frac{\widehat{\beta}_j}{(c_{jj}s^2)^{1/2}}. \tag{4.47}$$

4.5.3 A Numerical Example

Example 4.8
For illustration of the use of the general linear hypothesis, data from a

physical fitness program at N. C. State University are used. (The data were provided by A. C. Linnerud and are used with his permission.) Measurements were taken on $n = 31$ men. In addition to age and weight, oxygen uptake (Y), run time (X_1), heart rate while resting (X_2), heart rate while running (X_3), and maximum heart rate (X_4) while running 1.5 miles were recorded for each subject. The data are given in Table 4.3. The results we discuss are from the regression of oxygen uptake Y on the four variables X_1, X_2, X_3, and X_4.

The model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta} = (\,\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_4\,)'$ with the subscripts matching the identification of the independent variables given previously. The estimated regression equation is

$$\widehat{Y}_i \;=\; 84.26902 - 3.06981X_{i1} + .00799X_{i2} - .11671X_{i3} + .08518X_{i4}.$$

The analysis of variance for this model is summarized in Table 4.4. The residual mean square $s^2 = 7.4276$ is the estimate of σ^2 and has 26 degrees of freedom. The tests of hypotheses on $\boldsymbol{\beta}$ require $(\boldsymbol{X}'\boldsymbol{X})^{-1}$:

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} \;=\; \begin{bmatrix} 17.42309 & -.159620 & .007268 & -.014045 & -.077966 \\ -.159620 & .023686 & -.001697 & -.000985 & .000948 \\ .007268 & -.001697 & .000778 & -.000094 & -.000085 \\ -.014045 & -.000985 & -.000094 & .000543 & -.000356 \\ -.077966 & .000948 & -.000085 & -.000356 & .000756 \end{bmatrix}.$$

The first example tests the composite null hypothesis that the two regression coefficients β_2 and β_4 are zero, $H_0 : \beta_2 = \beta_4 = 0$. The alternative hypothesis is that either one or both are not zero. This null hypothesis is written in the general form as

$$\boldsymbol{K}'\boldsymbol{\beta} \;=\; \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Multiplication of the first row vector of $\boldsymbol{K}'$ with $\boldsymbol{\beta}$ gives $\beta_2 = 0$; the second row gives $\beta_4 = 0$.

There are two degrees of freedom associated with the sum of squares for this hypothesis, since $\mathrm{r}(\boldsymbol{K}) = 2$. The sum of squares is

$$\begin{aligned} Q &= (\boldsymbol{K}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m})'[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}]^{-1}(\boldsymbol{K}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m}) \\ &= \begin{pmatrix} .00799 \\ .08518 \end{pmatrix}' \begin{bmatrix} .0007776 & -.0000854 \\ -.0000854 & .0007560 \end{bmatrix}^{-1} \begin{pmatrix} .00799 \\ .08518 \end{pmatrix} \\ &= 10.0016. \end{aligned}$$

Notice that the product $\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}$ extracts the c_{22}, c_{24}, c_{42}, and c_{44} elements from $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. The F-test of the null hypothesis is

$$F \;=\; \frac{Q/2}{s^2} = \frac{(10.0016)/2}{7.4276} = .673.$$

TABLE 4.3. *Physical fitness measurements on 31 men involved in a physical fitness program at North Carolina State University. The variables measured were age (years), weight (kg), oxygen uptake rate (ml per kg body weight per minute), time to run 1.5 miles (minutes), heart rate while resting, heart rate while running (at the same time oxygen uptake was measured), and maximum heart rate while running. (Data courtesy A. C. Linnerud, N. C. State University.)*

				Heart Rate		
Age (yrs)	*Weight (kg)*	*O_2 Uptake (ml/kg/min)*	*Time (min)*	*Resting*	*Running*	*Maximum*
44	89.47	44.609	11.37	62	178	182
40	75.07	45.313	10.07	62	185	185
44	85.84	54.297	8.65	45	156	184
42	68.15	59.571	8.17	40	166	172
38	89.02	49.874	9.22	55	178	180
47	77.45	44.811	11.63	58	176	176
40	75.98	45.681	11.95	70	176	180
43	81.19	49.091	10.85	64	162	170
44	81.42	39.442	13.08	63	174	176
38	81.87	60.055	8.63	48	170	186
44	73.03	50.541	10.13	45	168	168
45	87.66	37.388	14.03	56	186	192
45	66.45	44.754	11.12	51	176	176
47	79.15	47.273	10.60	47	162	164
54	83.12	51.855	10.33	50	166	170
49	81.42	49.156	8.95	44	180	185
51	69.63	40.836	10.95	57	168	172
51	77.91	46.672	10.00	48	162	168
48	91.63	46.774	10.25	48	162	164
49	73.37	50.388	10.08	67	168	168
57	73.37	39.407	12.63	58	174	176
54	79.38	46.080	11.17	62	156	176
52	76.32	45.441	9.63	48	164	166
50	70.87	54.625	8.92	48	146	186
51	67.25	45.118	11.08	48	172	172
54	91.63	39.203	12.88	44	168	172
51	73.71	45.790	10.47	59	186	188
57	59.08	50.545	9.93	49	148	160
49	76.32	48.673	9.40	56	186	188
48	61.24	47.920	11.50	52	170	176
52	82.78	47.467	10.50	53	170	172

TABLE 4.4. *Summary analysis of variance for the regression of oxygen uptake on run time, heart rate while resting, heart rate while running, and maximum heart rate.*

Source	*d.f.*	*SS*	*MS*
$\text{Total}_{\text{corr}}$	30	851.3815	
Regression	4	658.2368	164.5659
Residual	26	193.1178	$7.4276 = s^2$

The computed F is much smaller than the critical value $F_{(.05,2,26)} = 3.37$ and, therefore, there is no reason to reject the null hypothesis that β_2 and β_4 are both zero.

The second hypothesis illustrates a case where $\boldsymbol{m} \neq \mathbf{0}$. Suppose prior information suggested that the intercept β_0 for a group of men of this age and weight should be 90. Then the null hypothesis of interest is $\beta_0 = 90$ and, for illustration, we construct a composite hypothesis by adding this constraint to the two conditions in the first null hypothesis. The null hypothesis is now

$$H_0 : \boldsymbol{K}'\boldsymbol{\beta} - \boldsymbol{m} = \mathbf{0},$$

where

$$\boldsymbol{K}'\boldsymbol{\beta} - \boldsymbol{m} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} - \begin{pmatrix} 90 \\ 0 \\ 0 \end{pmatrix}.$$

For this hypothesis

$$(\boldsymbol{K}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m}) = \begin{pmatrix} 84.26902 - 90 \\ .00799 \\ .08518 \end{pmatrix} = \begin{pmatrix} -5.73098 \\ .00799 \\ .08518 \end{pmatrix}$$

and

$$[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}]^{-1} = \begin{bmatrix} 17.423095 & .0072675 & -.0779657 \\ .0072675 & .0007776 & -.0000854 \\ -.0779657 & -.0000854 & .0007560 \end{bmatrix}^{-1}.$$

Notice that $(\boldsymbol{K}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m})$ causes the hypothesized $\beta_0 = 90$ to be subtracted from the estimated $\widehat{\beta}_0 = 84.26902$. The sum of squares for this composite hypothesis is

$$Q = (\boldsymbol{K}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m})'[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}]^{-1}(\boldsymbol{K}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m}) = 11.0187$$

and has 3 degrees of freedom. The computed F-statistic is

$$F = \frac{Q/3}{s^2} = \frac{11.0187/3}{7.4276} = .494,$$

which, again, is much less than the critical value of F for $\alpha = .05$ and 3 and 26 degrees of freedom, $F_{(.05,3,26)} = 2.98$. There is no reason to reject the null hypothesis that $\beta_0 = 90$ and $\beta_2 = \beta_4 = 0$. ■

4.5.4 Computing Q from Differences in Sums of Squares

Full and Reduced Models

As an alternative to the general formula for Q, equation 4.38, the sum of squares for any hypothesis can be determined from the difference between the residual sums of squares of two models. The current model, in the context of which the null hypothesis is to be tested, is called the **full model**. This model must include all parameters involved in the null hypothesis and will usually include additional parameters. The second model is obtained from the full model by assuming the null hypothesis is true and imposing its constraints on the full model. The model obtained in this way is called the **reduced model** because it will always have fewer parameters than the full model. For example, the null hypothesis $H_0 : \beta_2 = c$, where c is some known constant, gives a reduced model in which β_2 has been replaced with the constant c. Consequently, β_2 is no longer a parameter to be estimated.

Computing Q

The reduced model is a special case of the full model and, hence, its residual sum of squares must always be at least as large as the residual sum of squares for the full model. It can be shown that, for any general hypothesis, the sum of squares for the hypothesis can be computed as

$$Q = \text{SS}(\text{Res}_{\text{reduced}}) - \text{SS}(\text{Res}_{\text{full}}), \tag{4.48}$$

where "reduced" and "full" identify the two models.

Degrees of Freedom

There are $(n-p')$ degrees of freedom associated with $\text{SS}(\text{Res}_{\text{full}})$. Generating the reduced model by imposing the k linearly independent constraints of the null hypothesis on the full model reduces the number of parameters from p' to $(p'-k)$. Thus, $\text{SS}(\text{Res}_{\text{reduced}})$ has $[n-(p'-k)]$ degrees of freedom. Therefore, Q will have $[(n-p'+k)-(n-p')] = k$ degrees of freedom.

Illustration

Assume $\boldsymbol{X}$ is a full-rank matrix of order $n \times 4$ and $\boldsymbol{\beta}' = (\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3)$. Suppose the null hypothesis to be tested is

$$H_0 : \boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m},$$

where

$$\boldsymbol{K}' = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{m} = \begin{pmatrix} 0 \\ 20 \end{pmatrix}.$$

The full model was

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i.$$

If there were n observations, the residual sum of squares from this model would have $(n-4)$ degrees of freedom. The null hypothesis states that (1) $\beta_1 = \beta_2$ and (2) $\beta_0 = 20$. The reduced model is generated by imposing on the full model the conditions stated in the null hypothesis. Since the null hypothesis states that β_1 and β_2 are equal, one of these two parameters, say β_2, can be eliminated by substitution of β_1 for β_2. Similarly, β_0 is replaced with the constant 20. These substitutions give the reduced model:

$$Y_i = 20 + \beta_1 X_{i1} + \beta_1 X_{i2} + \beta_3 X_{i3} + \epsilon_i.$$

Moving the constant 20 to the left side of the equality and collecting the two terms that involve β_1 gives

$$Y_i - 20 = \beta_1 (X_{i1} + X_{i2}) + \beta_3 X_{i3} + \epsilon_i$$

or

$$Y_i^* = \beta_1 X_{i1}^* + \beta_3 X_{i3} + \epsilon_i,$$

where $Y_i^* = Y_i - 20$ and $X_{i1}^* = X_{i1} + X_{i2}$.

In matrix notation, the reduced model is

$$\boldsymbol{Y}^* = \boldsymbol{X}^* \boldsymbol{\beta}^* + \boldsymbol{\epsilon},$$

where

$$\boldsymbol{Y}^* = \begin{pmatrix} Y_1 - 20 \\ Y_2 - 20 \\ \vdots \\ Y_n - 20 \end{pmatrix}$$

$$\boldsymbol{X}^* = \begin{bmatrix} (X_{11} + X_{12}) & X_{13} \\ (X_{21} + X_{22}) & X_{23} \\ \vdots & \vdots \\ (X_{n1} + X_{n2}) & X_{n3} \end{bmatrix} = \begin{bmatrix} X_{11}^* & X_{13} \\ X_{21}^* & X_{23} \\ \vdots & \vdots \\ X_{n1}^* & X_{n3} \end{bmatrix}$$

and

$$\boldsymbol{\beta}^* = \begin{pmatrix} \beta_1 \\ \beta_3 \end{pmatrix}.$$

The rank of $\boldsymbol{X}^*$ is 2 so that $\mathrm{SS}(\mathrm{Res}_{\mathrm{reduced}})$ will have $(n-2)$ degrees of freedom. Consequently,

$$Q = \mathrm{SS}(\mathrm{Res}_{\mathrm{reduced}}) - \mathrm{SS}(\mathrm{Res}_{\mathrm{full}})$$

will have $[(n-2)-(n-4)] = 2$ degrees of freedom. Note that this agrees with $r(\boldsymbol{K}') = 2$.

The F-test of the null hypothesis is

$$F = \frac{Q/2}{s^2}$$

with 2 and ν degrees of freedom, where ν is the degrees of freedom in s^2. The denominator of F must be an unbiased estimate of σ^2 and must be statistically independent of the numerator sum of squares. This condition is satisfied if σ^2 is estimated from a model that contains at least all of the terms in the full model or is estimated from independent information such as provided by true replication (see Section 4.7).

Example 4.9

The oxygen consumption example, Example 4.8, is used to illustrate computation of Q using the difference between the residual sums of squares for full and reduced models. The reduced model for the first hypothesis tested, $H_0: \beta_2 = \beta_4 = 0$, is obtained from the full model by setting β_2 and β_4 equal to zero. This leaves a bivariate model containing only X_1 and X_3. Fitting this reduced model gives a residual sum of squares of

$$\text{SS}(\text{Res}_{\text{reduced}}) = 203.1194$$

with $[n-(p'-k)] = (31-3) = 28$ degrees of freedom. The residual sum of squares from the full model was

$$\text{SS}(\text{Res}_{\text{full}}) = 193.1178$$

with $(n-p') = 31-5 = 26$ degrees of freedom. The difference gives

$$\begin{aligned} Q &= \text{SS}(\text{Res}_{\text{reduced}}) - \text{SS}(\text{Res}_{\text{full}}) \\ &= 203.1194 - 193.1178 = 10.0016. \end{aligned}$$

with $(28-26) = 2$ degrees of freedom. This agrees, as it should, with the earlier result for Q obtained in Example 4.8.

The second hypothesis tested in the previous example included the statement that $\beta_0 = 90$ in addition to $\beta_2 = \beta_4 = 0$. The reduced model for this null hypothesis is

$$Y_i = 90 + \beta_1 X_{i1} + \beta_3 X_{i3} + \epsilon_i$$

or

$$(Y_i - 90) = \beta_1 X_{i1} + \beta_3 X_{i3} + \epsilon_i.$$

The reduced model has a new dependent variable formed by subtracting 90 from every Y_i, has only X_1 and X_3 as independent variables, and has *no intercept*. The residual sum of squares from this model is

$$\text{SS}(\text{Res}_{\text{reduced}}) = 204.1365$$

with $(31-2) = 29$ degrees of freedom. The $\text{SS}(\text{Res}_{\text{full}})$ is the same as before and the difference gives

$$Q = 204.1365 - 193.1178 = 11.0187$$

with 3 degrees of freedom. ■

Caution

The sum of squares Q for any null hypothesis can *always* be computed as a difference in *residual* sums of squares. For null hypotheses where $\boldsymbol{m} = \mathbf{0}$, the same result can be obtained, sometimes more conveniently, by taking the difference in the *model* sums of squares; that is,

$$Q = \text{SS}(\text{Model}_{\text{full}}) - \text{SS}(\text{Model}_{\text{reduced}}).$$

This follows from noting that

$$\text{SS}(\text{Model}_{\text{full}}) = \text{SS}(\text{Total}) - \text{SS}(\text{Res}_{\text{full}})$$

and

$$\text{SS}(\text{Model}_{\text{reduced}}) = \text{SS}(\text{Total}) - \text{SS}(\text{Res}_{\text{reduced}}).$$

If β_0 is in the model and *not* involved in the null hypothesis $\boldsymbol{K}'\boldsymbol{\beta} = \mathbf{0}$, the differences in regression sums of squares, $\text{SS}(\text{Regr}_{\text{full}}) - \text{SS}(\text{Regr}_{\text{reduced}})$, will also give Q. The first hypothesis in Example 4.9 involved only β_2 and β_4 and had $\boldsymbol{m} = \mathbf{0}$. The sum of squares due to regression for the reduced model was $\text{SS}(\text{Regr}_{\text{reduced}}) = 648.2622$. Comparison of this to $\text{SS}(\text{Regr}_{\text{full}}) = 658.2638$ verifies that the difference again gives $Q = 10.0016$.

The difference in regression sums of squares, however, cannot be used to compute Q in the second example where $\beta_0 = 20$ is included in the null hypothesis. In this case, SS(Total) for the reduced model is based on Y_i^* and hence it is different from SS(Total) for the full model. Consequently, it is important to develop the habit of either always using the residual sums of squares, since that procedure always gives the correct answer, or being very cautious in the use of differences in regression sums of squares to compute Q.

4.5.5 The R-Notation to Label Sums of Squares

The sum of squares for the null hypothesis that each of a subset of the partial regression coefficients is zero is dependent on both the specific subset of parameters in the null hypothesis and on the set of all parameters in the model. To clearly specify both in each case, a more convenient notation for sums of squares is needed. For this purpose, the commonly used *R-notation* is introduced.

Let $R(\beta_0\ \beta_1\ \beta_2\ \ldots\ \beta_p) = \text{SS}(\text{Model})$ denote the sum of squares due to the model containing the parameters listed in parentheses. The sum

of squares for the hypothesis that a subset of β_j is zero can be obtained by subtraction of SS(Model) for the reduced model from that for the full model. Assume the subset of β_j being tested against zero consists of the last k β_j. Then

$$\begin{aligned} \text{SS}(\text{Model}_{\text{full}}) &= R(\beta_0\ \beta_1\ \ldots\ \beta_{p-k}\ \beta_{p-k+1}\ \ldots\ \beta_p), \\ \text{SS}(\text{Model}_{\text{reduced}}) &= R(\beta_0\ \beta_1\ \ldots\ \beta_{p-k}) \end{aligned}$$

and

$$\begin{aligned} Q &= \text{SS}(\text{Model}_{\text{full}}) - \text{SS}(\text{Model}_{\text{reduced}}) \\ &= R(\beta_0\ \beta_1\ \ldots\ \beta_{p-k}\ \beta_{p-k+1}\ \ldots\ \beta_p) - R(\beta_0\ \beta_1\ \ldots\ \beta_{p-k}). \end{aligned} \tag{4.49}$$

The final R-notation expresses this difference in sums of squares as

$$R(\beta_{p-k+1}\ \beta_{p-k+2}\ \ldots\ \beta_p|\beta_0\ \beta_1\ \ldots\ \beta_{p-k}). \tag{4.50}$$

The β_j appearing before the vertical bar are those specified to be zero by the null hypothesis, whereas the β_j appearing after the bar are those for which the former are adjusted. Alternatively, the full model consists of all parameters in parentheses, whereas the reduced model contains only those parameters appearing *after* the bar. In this notation,

$$\begin{aligned} \text{SS}(\text{Regr}) &= \text{SS}(\text{Model}) - \text{SS}(\mu) \\ &= R(\beta_1\ \beta_2\ \ldots\ \beta_p|\beta_0). \end{aligned} \tag{4.51}$$

Partial Sums of Squares

To illustrate the R-notation, consider a linear model that contains three independent variables plus an intercept, given by

$$Y_i = \beta_0 X_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \tag{4.52}$$

where ϵ_i are NID(0, σ^2) and $X_{i0} = 1$. The *partial* sums of squares for this model would be

$R(\beta_1|\beta_0\ \beta_2\ \beta_3)$,

$R(\beta_2|\beta_0\ \beta_1\ \beta_3)$, and

$R(\beta_3|\beta_0\ \beta_1\ \beta_2)$.

Each is the additional sum of squares accounted for by the parameter (or its corresponding variable) appearing before the vertical bar when added to a model that already contains the parameters appearing after the bar. Each is the appropriate numerator sum of squares for testing the simple hypothesis $H_0:\ \beta_j = 0$, for $j = 1,\ 2,$ and 3, respectively.

Consider the model

$$Y_i = \beta_0 X_{i0} + \beta_3 X_{i3} + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \tag{4.53}$$

where we have changed the order of the independent variables in model 4.52. The partial sums of squares for X_3, X_1, and X_2 are

$$R(\beta_3|\beta_0\ \beta_1\ \beta_2),\ R(\beta_1|\beta_0\ \beta_3\ \beta_2),\ \text{and}\ R(\beta_2|\beta_0\ \beta_3\ \beta_1),$$

respectively, and are the same as those obtained for model 4.52. That is, the partial sums of squares for the independent variables of a given model are independent of the order in which the variables are listed in the model.

Sequential Sums of Squares

The **sequential sums of squares** measure the contributions of the variables as they are *added* to the model in a particular sequence. The sequential sum of squares for X_j is the increase in SS(Regr), or the decrease in SS(Res), when X_j is added to the existing model. This sum of squares measures the contribution of X_j adjusted *only* for those independent variables that preceded X_j in the model-building sequence.

For illustration, suppose a model is to be built by adding variables in the sequence X_0, X_1, X_2, and X_3 as in model 4.52. The first model to be fit will contain X_0 (the intercept) and X_1. SS(Regr) from this model is the sequential sum of squares for X_1. In the R-notation, this sequential sum of squares is given by $R(\beta_1|\beta_0)$. The second model to be fit will contain X_0, X_1, and X_2. The sequential sum of squares for X_2 is SS(Regr) for this model minus SS(Regr) for the first model and, in R-notation, it is given by $R(\beta_2|\beta_0\ \beta_1)$. The third model to be fit will contain the intercept and all three independent variables. The sequential sum of squares for X_3 is SS(Regr) for this three-variable model minus SS(Regr) for the preceding two-variable model. In R-notation, the sequential sum of squares for X_3 is $R(\beta_3|\beta_0\ \beta_1\ \beta_2)$. Note that because X_3 is the last variable added to the model, the sequential sum of squares for X_3 coincides with the partial sum of squares for X_3.

Consider now equation 4.53 where the model is built in the sequence X_0, X_3, X_1, and X_2. The sequential sums of squares for X_3, X_1, and X_2 are $R(\beta_3|\beta_0)$, $R(\beta_1|\beta_0\ \beta_3)$, and $R(\beta_2|\beta_0\ \beta_3\ \beta_1)$. These are different from the sequential sums of squares obtained in the model 4.52. That is, the sequential sums of squares are *dependent* on the order in which the variables are added to the model. It should be clear from the definition of the R-notation that the ordering of the parameters after the vertical bar is immaterial.

Using Sequential Sums of Squares

The *partial* sums of squares measure the contributions of the individual variables with each adjusted for *all* other independent variables in the model (see Section 4.5.2) and are appropriate for testing simple hypotheses of the form $H_0:\ \beta_j = 0$. Each sequential sum of squares is the appropriate sum of squares for testing the jth partial regression coefficient, $H_0:\ \beta_j = 0$, for a model that contains X_j and *only* those independent variables that preceded X_j in the sequence. For example, the sequential sum of squares, $R(\beta_1|\beta_0)$, for X_1 is appropriate for testing $H_0:\ \beta_1 = 0$ in the model $Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$. Note that this model assumes that β_2 and β_3 of model 4.52 are zero. The sequential sum of squares $R(\beta_2|\beta_0\ \beta_1)$ for X_2 is appropriate

for testing $H_0 : \beta_2 = 0$ in the model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$. This model assumes $\beta_3 = 0$ in model 4.52. Similarly, the sequential sum of squares $R(\beta_3|\beta_0)$ (in model 4.53) is appropriate for testing $H_0 : \beta_3 = 0$ in the model $Y_i = \beta_0 + \beta_3 X_{i3} + \epsilon_i$. This is, however, not appropriate for testing $H_0 : \beta_3 = 0$ in the model $Y_i = \beta_0 + \beta_3 X_{i3} + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$.

The *partial* sums of squares, although useful for testing simple hypotheses of the form $H_0 : \beta_j = 0$, are not useful for testing joint hypotheses of the form $H_0 : \beta_j = 0$, $\beta_k = 0$ or $H_0 : \beta_j = 0$, $\beta_k = 0$, $\beta_l = 0$. The sequential sums of squares can be combined to obtain appropriate sums of squares for testing certain joint hypotheses. For example, if we wish to test $H_0 : \beta_2 = \beta_3 = 0$ in model 4.52, we know that the appropriate numerator sum of squares is

$$\begin{aligned} R(\beta_2\ \beta_3|\beta_0\ \beta_1) &= R(\beta_0\ \beta_1\ \beta_2\ \beta_3) - R(\beta_0\ \beta_1) \\ &= [R(\beta_0\ \beta_1\ \beta_2\ \beta_3) - R(\beta_0\ \beta_1\ \beta_2)] \\ &\quad + [R(\beta_0\ \beta_1\ \beta_2) - R(\beta_0\ \beta_1)] \\ &= R(\beta_3|\beta_0\ \beta_1\ \beta_2) + R(\beta_2|\beta_0\ \beta_1) \\ &= \text{sum of the sequential sums of squares for} \\ &\quad X_2 \text{ and } X_3 \text{ in model 4.52.} \end{aligned}$$

Similarly, if we wish to test the hypothesis that $H_0 : \beta_1 = \beta_2 = 0$ in model 4.53 (or 4.52), the appropriate sum of squares is

$$\begin{aligned} R(\beta_1\ \beta_2|\beta_0\ \beta_3) &= R(\beta_0\ \beta_3\ \beta_1\ \beta_2) - R(\beta_0\ \beta_3) \\ &= [R(\beta_0\ \beta_3\ \beta_1\ \beta_2) - R(\beta_0\ \beta_3\ \beta_1)] \\ &\quad + [R(\beta_0\ \beta_3\ \beta_1) - R(\beta_0\ \beta_3)] \\ &= R(\beta_2|\beta_0\ \beta_3\ \beta_1) + R(\beta_1|\beta_0\ \beta_3) \\ &= \text{sum of the sequential sums of squares for} \\ &\quad X_2 \text{ and } X_1 \text{ in model 4.53.} \end{aligned}$$

Note that the sequential sums of squares from model 4.52 cannot be used for testing $H_0 : \beta_1 = \beta_2 = 0$. Note that in both models, equations 4.52 and 4.53,

$$\begin{aligned} \text{SS(Regr)} &= R(\beta_1\ \beta_2\ \beta_3|\beta_0) \\ &= R(\beta_1|\beta_0) + R(\beta_2|\beta_0\ \beta_1) + R(\beta_3|\beta_0\ \beta_1\ \beta_2) \\ &= R(\beta_3|\beta_0) + R(\beta_1|\beta_0\ \beta_3) + R(\beta_2|\beta_0\ \beta_1\ \beta_3). \end{aligned}$$

That is, the sequential sums of squares are an additive partition of SS(Regr) for the full model.

There are some models (for example, purely nested models and polynomial response models) where there is a logical order in which terms should be added to the model. In such cases, the sequential sums of squares provide the appropriate tests for determining which terms are to be retained

TABLE 4.5. *Regression sum of squares, sequential sums of squares, and the residual sum of squares for the oxygen uptake example.*

Sequential Sums of Squares			*d.f.*	*F*
SS(Regr) =	$R(\beta_1\ \beta_3\ \beta_2\ \beta_4\|\beta_0)$	=658.2638	4	22.16
	$R(\beta_1\|\beta_0)$	=632.9001	(1)	
	$R(\beta_3\|\beta_0\ \beta_1)$	= 15.3621	(1)	
	$R(\beta_2\|\beta_0\ \beta_1\ \beta_3)$	= .4041	(1)	
	$R(\beta_4\|\beta_0\ \beta_1\ \beta_3\ \beta_2)$	= 9.5975	(1)	
SS(Error)		=193.1178	26	

in the model. In other cases, prior knowledge of the behavior of the system will suggest a logical ordering of the variables according to their relative importance. Use of this prior information and sequential sums of squares should simplify the process of determining an appropriate model.

4.5.6 Example: Sequential and Partial Sums of Squares

Example 4.10

The oxygen uptake example, Example 4.8, is used to illustrate the *R*-notation and the sequential and partial sums of squares. The sum of squares due to regression for the full model was

$$\text{SS(Regr)} = R(\beta_1\ \beta_2\ \beta_3\ \beta_4|\beta_0) = 658.2638$$

with four degrees of freedom (Table 4.4). The *sequential* sums of squares, from fitting the model in the order X_1, X_3, X_2, and X_4 are shown in Table 4.5. Each sequential sum of squares measures the stepwise improvement in the model realized from adding one independent variable. The sequential sums of squares add to the total regression sum of squares, $\text{SS(Regr)} = R(\beta_1\ \beta_3\ \beta_2\ \beta_4|\beta_0) = 658.2638$; that is, this is an orthogonal partitioning.

The regression sum of squares, $R(\beta_1\ \beta_3\ \beta_2\ \beta_4|\beta_0)$ is used to test the composite hypothesis $H_0 : \beta_1 = \beta_3 = \beta_2 = \beta_4 = 0$. This gives $F = 22.16$ which, with 4 and 26 degrees of freedom is highly significant. That is, there is evidence to believe that the independent variables need to be included in the model to account for the variability in oxygen consumption among runners.

Adjacent sequential sums of squares at the end of the list can be added to generate the appropriate sum of squares for a composite hypothesis. For example, the sequential sums of squares $R(\beta_2|\beta_0\ \beta_1\ \beta_3)$ and $R(\beta_4|\beta_0\ \beta_1\ \beta_3\ \beta_2)$ for X_2 and X_4, respectively, in Table 4.5, can be added to give the additional sum of squares one would obtain from adding *both* X_2 and X_4 in one step to the model containing only X_1 and X_3 (and the intercept). Thus,

$$R(\beta_2|\beta_0\ \beta_1\ \beta_3) + R(\beta_4|\beta_0\ \beta_1\ \beta_3\ \beta_2) \quad = \quad .4041 + 9.5975$$

TABLE 4.6. *Cumulative sequential sums of squares, the null hypothesis being tested by each cumulative sum of squares, and the F-test of the null hypothesis for the oxygen uptake example.*

Cumulative Sequential Sums of Squares	*d.f.*	*Null Hypothesis*	*F*
$R(\beta_1\ \beta_3\ \beta_2\ \beta_4\|\beta_0)$=658.2638	4	$\beta_1 = \beta_3 = \beta_2 = \beta_4 = 0$	22.16
$R(\beta_3\ \beta_2\ \beta_4\|\beta_0\ \beta_1)$= 25.3637	3	$\beta_3 = \beta_2 = \beta_4 = 0$	.67
$R(\beta_2\ \beta_4\|\beta_0\ \beta_1\ \beta_3)$= 10.0026	2	$\beta_2 = \beta_4 = 0$	1.14
$R(\beta_4\|\beta_0\ \beta_1\ \beta_3\ \beta_2)$= 9.5975	1	$\beta_4 = 0$	1.29
SS(Error) =193.1178	26		

$$\begin{aligned} &= 10.0016 \\ &= R(\beta_2\ \beta_4|\beta_0\ \beta_1\ \beta_3) \end{aligned}$$

in the R-notation. This is the appropriate sum of squares for testing the composite hypothesis that both β_2 and β_4 are zero. This gives $F = .67$ which, with 2 and 26 degrees of freedom, does not approach significance. That is, the run time X_1 and the heart rate while running X_3 are sufficient to account for oxygen consumption differences among runners.

If this particular ordering of the variables was chosen because it was expected that X_1 (run time) likely would be the most important variable with the others being of secondary importance, it is logical to test the composite null hypothesis $H_0 : \beta_3 = \beta_2 = \beta_4 = 0$. The sum of the sequential sums of squares for X_3, X_2, and X_4 is the appropriate sum of squares and gives $R(\beta_3\ \beta_2\ \beta_4|\beta_0\ \beta_1) = 25.3637$ with 3 degrees of freedom. This gives $F = 1.14$ which, with 3 and 26 degrees of freedom, does not approach significance. This single test supports the contention that X_1 alone is sufficient to account for oxygen consumption differences among the runners. (Since the variables are not orthogonal, this does *not* rule out the possibility that a model based on the other three variables might do better.)

The cumulative sequential sums of squares (from bottom to top) and the corresponding F-statistics and null hypotheses are summarized in Table 4.6. The appropriate sum of squares to test the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$ is $R(\beta_2\ \beta_3|\beta_0\ \beta_1\ \beta_4)$. This sum of squares cannot be obtained from the sums of squares given in Tables 4.5 and 4.6. The sum of squares $R(\beta_2\ \beta_3|\beta_0\ \beta_1\ \beta_4)$ may be obtained by adding the sequential sums of squares for X_2 and X_3 from fitting the model in the order X_0, X_1, X_4, X_2, and X_3.

The *partial* sums of squares, their null hypotheses, and the F-tests are shown in Table 4.7. This is not an orthogonal partitioning; the partial sums of squares will not add to SS(Regr). Each partial sum of squares reflects the contribution of the particular variable as if it were the last to be considered for the model. Hence, it is the appropriate sum of squares

TABLE 4.7. *Partial sums of squares, the null hypothesis being tested by each, and the F-test of the null hypothesis for the oxygen uptake example.*

Partial Sum of Squares	*Null Hypothesis*	F^a
$R(\beta_1\|\beta_0,\ \beta_2,\ \beta_3,\ \beta_4) = 397.8664$	$\beta_1 = 0$	53.57
$R(\beta_3\|\beta_0,\ \beta_1,\ \beta_2,\ \beta_4) = 25.0917$	$\beta_3 = 0$	3.38
$R(\beta_2\|\beta_0,\ \beta_1,\ \beta_3,\ \beta_4) = .0822$	$\beta_2 = 0$	.01
$R(\beta_4\|\beta_0,\ \beta_1,\ \beta_2,\ \beta_3) = 9.5975$	$\beta_4 = 0$	1.29

[a]All F-tests were computed using the residual mean square from the full model.

for deciding whether the variable might be omitted. The null hypotheses in Table 4.7 reflect the adjustment of each partial regression coefficient for all other independent variables in the model.

The partial sum of squares for X_2, $R(\beta_2|\beta_0\ \beta_1\ \beta_3\ \beta_4) = .0822$ is much smaller than $s^2 = 7.4276$ and provides a clear indication that this variable does not make a significant contribution to a model that already contains X_1, X_3, and X_4. The next logical step in building the model based on tests of the partial sums of squares would be to omit X_2. Even though the tests for β_3 and β_4 are also nonsignificant, one must be cautious in omitting more than one variable at a time on the basis of the partial sums of squares. The partial sums of squares are dependent on which variables are in the model; it will almost always be the case that all partial sums of squares will change when a variable is dropped. (In this case, we know from the sequential sums of squares that all three variables can be dropped. A complete discussion on choice of variables is presented in Chapter 7.) ■

4.6 Univariate and Joint Confidence Regions

Confidence interval estimates of parameters convey more information to the reader than do simple point estimates. Univariate confidence intervals for several parameters, however, do not take into account correlations among the estimators of the parameters. Furthermore, the individual confidence coefficients do not reflect the overall degree of confidence in the joint statements. Joint confidence regions address these two points. Univariate confidence interval estimates are discussed briefly before proceeding to a discussion of joint confidence regions.

4.6.1 Univariate Confidence Intervals

Confidence Intervals for $\boldsymbol{\beta_j}$

If $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{I}\sigma^2)$, then $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{Y}}$ have multivariate normal distributions (see equation 3.37). With normality, the classical $(1 - \alpha)100\%$ confidence

interval estimate of each β_j is

$$\widehat{\beta}_j \quad \pm \quad t_{(\alpha/2,\nu)} s(\widehat{\beta}_j), \quad j = 0, \ldots, p, \tag{4.54}$$

where $t_{(\alpha/2,\nu)}$ is the value of the Student's t-distribution, with ν degrees of freedom, that puts $\alpha/2$ probability in the upper tail. [In the usual multiple regression problem, $\nu = (n-p')$.] The standard error of $\widehat{\beta}_j$ is $s(\widehat{\beta}_j) = \sqrt{c_{jj}s^2}$ where s^2 is estimated with ν degrees of freedom and c_{jj} is the $(j+1)$th diagonal element from $(\boldsymbol{X'X})^{-1}$.

Confidence Interval for $\boldsymbol{\mathcal{E}(Y_0)}$

Similarly, the $(1-\alpha)100\%$ confidence interval estimate of the mean of Y for a particular choice of values for the independent variables, say $\boldsymbol{x}_0' = (\,1 \quad X_{01} \quad \cdots \quad X_{0p}\,)$, is

$$\widehat{Y}_0 \quad \pm \quad t_{(\alpha/2,\nu)} s(\widehat{Y}_0), \tag{4.55}$$

where $\widehat{Y}_0 = \boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}$; $s(\widehat{Y}_0) = \sqrt{\boldsymbol{x}_0'(\boldsymbol{X'X})^{-1}\boldsymbol{x}_0 s^2}$, in general, or $s(\widehat{Y}_0) = \sqrt{v_{ii}s^2}$ if $\boldsymbol{x}_0'$ corresponds to the ith row of $\boldsymbol{X}$; v_{ii} is the ith diagonal element in $\boldsymbol{P}$; $t_{(\alpha/2,\nu)}$ is as defined for equation 4.54.

Prediction Interval for $\boldsymbol{Y_0}$

A $(1-\alpha)100\%$ prediction interval of $Y_0 = \boldsymbol{x}_0'\boldsymbol{\beta} + \epsilon$, for a particular choice of values of the independent variables, say $\boldsymbol{x}_0' = (\,1 \quad X_{01} \quad \cdots \quad X_{0p}\,)$ is

$$\widehat{Y}_0 \quad \pm \quad t_{(\alpha/2,\nu)} s(Y_0 - \widehat{Y}_0), \tag{4.56}$$

where $\widehat{Y}_0 = \boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}$ and $s(Y_0 - \widehat{Y}_0) = \sqrt{s^2[1 + \boldsymbol{x}_0'(\boldsymbol{X'X})^{-1}\boldsymbol{x}_0]}$.

Example 4.11

The univariate confidence intervals are illustrated with the oxygen uptake example (see Example 4.8). $s^2 = 7.4276$ was estimated with 26 degrees of freedom. The value of Student's t for $\alpha = .05$ and 26 degrees of freedom is $t_{(.025,26)} = 2.056$. The point estimates of the parameters and the estimated variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ were

$$\widehat{\boldsymbol{\beta}}' = (\,84.2690 \quad -3.0698 \quad .0080 \quad -.1167 \quad .0852\,)$$

and

$$\begin{aligned} \boldsymbol{s^2}(\widehat{\boldsymbol{\beta}}) &= (\boldsymbol{X'X})^{-1}s^2 \\ &= \begin{bmatrix} 129.4119 & -1.185591 & .053980 & -.104321 & -.579099 \\ -1.185591 & .175928 & -.012602 & -.007318 & .007043 \\ .053980 & -.012602 & .005775 & -.000694 & -.000634 \\ -.104321 & -.007318 & -.000694 & .004032 & -.002646 \\ -.579099 & .007043 & -.000634 & -.002646 & .005616 \end{bmatrix}. \end{aligned}$$

The square root of the $(j+1)$st diagonal element gives $s(\widehat{\beta}_j)$. If $\widehat{\boldsymbol{d}}$ is defined as the column vector of $s(\widehat{\beta}_j)$, the univariate 95% confidence interval

estimates can be computed as

$$\begin{aligned} \mathrm{CL}(\widehat{\boldsymbol{\beta}}) &= [\widehat{\boldsymbol{\beta}} - t_{(\alpha/2,\nu)}\widehat{\boldsymbol{d}} \quad \widehat{\boldsymbol{\beta}} + t_{(\alpha/2,\nu)}\widehat{\boldsymbol{d}}] \\ &= \begin{bmatrix} 60.880 & 107.658 \\ -3.932 & -2.207 \\ -.148 & .164 \\ -.247 & .014 \\ -.069 & .239 \end{bmatrix}, \end{aligned}$$

where the two columns give the lower and upper limits, respectively, for the β_j in the same order as listed in $\boldsymbol{\beta}$. ■

4.6.2 Simultaneous Confidence Statements

For the classical univariate confidence intervals, the confidence coefficient $(1-\alpha) = .95$ applies to each confidence statement. The level of confidence associated with the statement that all five intervals simultaneously contain their respective parameters is much lower. If the five intervals were statistically independent, which they are not, the overall or joint confidence coefficient would be only $(1-\alpha)^5 = .77$.

There are two procedures that keep the joint confidence coefficient for several simultaneous statements near a prechosen level $(1-\alpha)$. The oldest and simplest procedure, commonly called the **Bonferroni method**, constructs the individual confidence intervals as given in equations 4.54 and 4.55, but uses $\alpha^* = \alpha/k$ where k is the number of simultaneous intervals or statements. That is, in equation 4.54, $t_{(\alpha/2,\nu)}$ is replaced with $t_{(\alpha/2k,\nu)}$. This procedure ensures that the true joint confidence coefficient for the k simultaneous statements is *at least* $(1-\alpha)$.

The Bonferroni simultaneous confidence intervals for the p' parameters in $\boldsymbol{\beta}$ are given by

$$\widehat{\beta}_j \quad \pm \quad t_{(\alpha/2p',\nu)}s(\widehat{\beta}_j). \tag{4.57}$$

This method is particularly suitable for obtaining simultaneous confidence intervals for k *prespecified* (prior to analyzing the data) parameters or linear combinations of parameters. When k is small, generally speaking, the Bonferroni simultaneous confidence intervals are not very wide. However, if k is large, the Bonferroni intervals tend to be wide (conservative) and the simultaneous coverage may be much larger than the specified confidence level $(1-\alpha)$. For example, if we are interested in obtaining simultaneous confidence intervals of all pairwise differences of p parameters (e.g., treatment means), then k is $p(p+1)/2$ which is large even for moderate values of p. The Bonferroni method is not suitable for obtaining simultaneous confidence intervals for *all* linear combinations. In this case, k is infinity

and the Bonferroni intervals would be the entire space. For example, in a simple linear regression, if we wish to compute a confidence *band* on the entire regression line, then the Bonferroni simultaneous band would be the entire space.

Scheffé's Method

The second procedure applies the general approach developed by Scheffé (1953). **Scheffé's method** provides simultaneous confidence statements for *all* linear combinations of a set of parameters in a d-dimensional subspace of the p'-dimensional parameter space. The Scheffé joint confidence intervals for the p' parameters in $\boldsymbol{\beta}$ and the means of $\boldsymbol{Y}$, $\mathcal{E}(Y_i)$, are obtained from equations 4.54 and 4.55 by replacing $t_{(\alpha/2,\nu)}$ with $[p'F_{(\alpha,p',\nu)}]^{1/2}$. (If only a subset of d linearly independent parameters β_j is of interest, $t_{(\alpha/2,\nu)}$ is replaced with $[dF_{(\alpha,d,\nu)}]^{1/2}$.) That is,

$$\widehat{\beta}_j \pm (p'F_{(\alpha,p',\nu)})^{1/2} s(\widehat{\beta}_j) \tag{4.58}$$

$$\widehat{Y}_0 \pm (p'F_{(\alpha,p',\nu)})^{1/2} s(\widehat{Y}_0). \tag{4.59}$$

This method provides simultaneous statements for all linear combinations of the set of parameters. As with the Bonferroni intervals, the joint confidence coefficient for the Scheffé intervals is at least $(1-\alpha)$. That is, the confidence coefficient of $(1-\alpha)$ applies to all confidence statements on the β_j, the $\mathcal{E}(Y_i)$, plus all other linear functions of β_j of interest. Thus , equation 4.59 can be used to establish a confidence *band* on the entire regression surface by computing Scheffé confidence intervals for $\mathcal{E}(Y_0)$ for all values of the independent variables in the region of interest. The confidence band for the simple linear regression case was originally developed by Working and Hotelling (1929) and frequently carries their names.

The reader is referred to Miller (1981) for more complete presentations on Bonferroni and Scheffé methods. Since the Scheffé method provides simultaneous confidence statements on *all* linear functions of a set of parameters, the Scheffé intervals will tend to be longer than Bonferroni intervals, particularly when a small number of simultaneous statements is involved (Miller, 1981). One would choose the method that gave the shorter intervals for the particular application.

Example 4.12

The oxygen uptake model of Example 4.8 has $p' = 5$ parameters and $\nu = 26$ degrees of freedom for s^2. In order to attain an overall confidence coefficient no smaller than $(1-\alpha) = .95$ with the Bonferroni method, $\alpha^* = .05/5 = .01$ would be used, for which $t_{(.01/2,26)} = 2.779$. Using this value of t in equation 4.54 gives the Bonferroni simultaneous confidence intervals

with an *overall* confidence coefficient at least as large as $(1 - \alpha) = .95$:

$$\mathrm{CL}_{\mathbf{B}}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} 52.655 & 115.883 \\ -4.235 & -1.904 \\ -.203 & -.219 \\ -.293 & .060 \\ -.123 & .293 \end{bmatrix}.$$

The Scheffé simultaneous intervals for the $p' = 5$ parameters in $\boldsymbol{\beta}$ are obtained by using $[p' F_{(.05,5,26)}]^{1/2} = [5(2.59)]^{1/2} = 3.599$ in place of $t_{(\alpha/2,\nu)}$ in equation 4.54. The results are

$$\mathrm{CL}_{\mathbf{S}}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} 43.331 & 125.207 \\ -4.579 & -1.560 \\ -.265 & .281 \\ -.345 & .112 \\ -.184 & .355 \end{bmatrix}.$$

The Bonferroni and Scheffé simultaneous confidence intervals will always be wider than the classical univariate confidence intervals in which the confidence coefficient applies to each interval. In this example, the Scheffé intervals are wider than the Bonferroni intervals. ■

The $100(1 - \alpha)\%$ simultaneous confidence intervals for $\boldsymbol{\beta}$ obtained using either Bonferroni or Sheffé methods, provide confidence intervals for each individual parameter β_j in such a way that the p'-dimensional region formed by the intersection of the p'-simultaneous confidence intervals gives at least a $100(1-\alpha)\%$ joint confidence region for *all* parameters. The shape of this joint confidence region is rectangular or cubic. Sheffé also derives an ellipsoidal $100(1 - \alpha)\%$ joint confidence region for all parameters that is contained in the boxed region obtained by the Sheffé simultaneous confidence intervals. This distinction is illustrated after joint confidence regions are defined in the next section.

4.6.3 Joint Confidence Regions

A joint confidence region for all p' parameters in $\boldsymbol{\beta}$ is obtained from the inequality

$$(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \leq p' s^2 F_{(\alpha,p',\nu)}, \tag{4.60}$$

where $F_{(\alpha,p',\nu)}$ is the value of the F-distribution with p' and ν degrees of freedom that leaves probability α in the upper tail; ν is the degrees of freedom associated with the s^2. The left-hand side of this inequality is a quadratic form in $\boldsymbol{\beta}$, because $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{X}'\boldsymbol{X}$ are known quantities computed from the data. The right-hand side is also known from the data. Solving

this quadratic form for the boundary of the inequality establishes a p'-dimensional ellipsoid which is the $100(1-\alpha)\%$ joint confidence region for all the parameters in the model. The slope of the axes and eccentricity of the ellipsoid show the direction and strength, respectively, of correlations between the estimates of the parameters.

Interpretation

An ellipsoidal confidence region with more than two or three dimensions is difficult to interpret. Specific choices of $\boldsymbol{\beta}$ can be checked, with a computer program, to determine whether they fall inside or outside the confidence region. The multidimensional region, however, must be viewed two or at most three dimensions at a time. One approach to visualizing the joint confidence region is to evaluate the p'-dimensional joint confidence region for specific values of all but two of the parameters. Each set of specified values produces an ellipse that is a two-dimensional "slice" of the multidimensional region. To develop a picture of the entire region, two-dimensional "slices" can be plotted for several choices of values for the other parameters.

An alternative to using the p'-dimensional joint confidence region for all parameters is to construct joint confidence regions for two parameters at a time *ignoring* the other $(p'-2)$ parameters. The quadratic form for the joint confidence region for a subset of two parameters is obtained from that for all parameters, equation 4.60, by

1. replacing $(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ with the corresponding vectors involving *only* the two parameters of interest;
2. replacing $(\boldsymbol{X}'\boldsymbol{X})$ with the *inverse* of the 2×2 variance–covariance matrix for the two parameters; and
3. replacing $p's^2F_{(\alpha,p',\nu)}$ with $2F_{(\alpha,2,\nu)}$. Notice that s^2 is not in the second quantity since it has been included in the variance–covariance matrix in step 2.

Thus, if β_j and β_k are the two distinct parameters of interest, the joint confidence region is given by

$$\left[\begin{pmatrix}\widehat{\beta}_j\\ \widehat{\beta}_k\end{pmatrix}-\begin{pmatrix}\beta_j\\ \beta_k\end{pmatrix}\right]' (\boldsymbol{s^2}(\widehat{\beta}_j,\widehat{\beta}_k))^{-1}\left[\begin{pmatrix}\widehat{\beta}_j\\ \widehat{\beta}_k\end{pmatrix}-\begin{pmatrix}\beta_j\\ \beta_k\end{pmatrix}\right] \leq 2F_{(\alpha,2,\nu)}. \qquad (4.61)$$

The confidence coefficient $(1-\alpha)$ applies to the joint statement on the two parameters being considered at the time. This procedure takes into account the joint distribution of $\widehat{\beta}_j$ and $\widehat{\beta}_k$ but ignores the values of the other parameters. Since this bivariate joint confidence region ignores the joint distribution of $\widehat{\beta}_j$ and $\widehat{\beta}_k$ with the other $(p'-2)$ parameter estimates, it suffers from the same conceptual problem as the univariate confidence intervals.

Example 4.13

The oxygen uptake data, given in Example 4.8, are used to illustrate joint confidence regions, but the model is simplified to include only an intercept and two independent variables, time to run 1.5 miles (X_1) and heart rate while running (X_3). The estimate of $\boldsymbol{\beta}$, $\boldsymbol{X'X}$, and the variance–covariance matrix for $\widehat{\boldsymbol{\beta}}$ for this reduced model are

$$\begin{aligned} \widehat{\boldsymbol{\beta}}' &= (\,93.0888 \quad -3.14019 \quad -0.073510\,) \\ \boldsymbol{X'X} &= \begin{bmatrix} 31 & 328.17 & 5,259 \\ 328.17 & 3531.797 & 55,806.29 \\ 5,259 & 55,806.29 & 895,317 \end{bmatrix} \end{aligned}$$

and

$$\boldsymbol{s^2}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} 68.04308 & -.47166 & -.37028 \\ -.47166 & .13933 & -.00591 \\ -.37028 & -.00591 & .00255 \end{bmatrix}.$$

The residual mean square from this model is $s^2 = 7.25426$ with 28 degrees of freedom.

The joint confidence region for all three parameters is obtained from equation 4.60 and is a three-dimensional ellipsoid. The right-hand side of equation 4.60 is

$$p' s^2 F_{(\alpha,3,28)} = 3(7.25426)(2.95)$$

if $\alpha = .05$. This choice of α gives a confidence coefficient of .95 that applies to the joint statement involving all three parameters. The three-dimensional ellipsoid is portrayed in Figure 4.1 with three two-dimensional "slices" (solid lines) from the ellipsoid at $\beta_0 = 76.59, 93.09,$ and 109.59. These choices of β_0 correspond to $\widehat{\beta}_0$ and $\widehat{\beta}_0 \pm 2s(\widehat{\beta}_0)$. The "slices" indicate that the ellipsoid is extremely thin in one plane but only slightly elliptical in the other, much like a slightly oval pancake. This is reflecting the high correlation between $\widehat{\beta}_0$ and $\widehat{\beta}_3$ of $-.89$ and the more moderate correlations of $-.15$ and $-.31$ between $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and between $\widehat{\beta}_1$ and $\widehat{\beta}_3$, respectively.

The bivariate joint confidence region for $\widehat{\beta}_1$ and $\widehat{\beta}_3$ *ignoring* $\widehat{\beta}_0$, obtained from equation 4.61, is shown in Figure 4.1 as the ellipse drawn with the dashed line. The variance–covariance matrix to be inverted in equation 4.61 is the lower-right 2×2 matrix in $\boldsymbol{s^2}(\widehat{\boldsymbol{\beta}})$. The right-hand side of the inequality is $2F_{(\alpha,2,28)} = 2(3.34)$ if $\alpha = .05$. The confidence coefficient of .95 applies to the joint statement involving *only* β_1 and β_3. The negative slope in this ellipse reflects the moderate negative correlation between $\widehat{\beta}_1$ and $\widehat{\beta}_3$. For reference, the Bonferroni confidence intervals for β_1 and β_3, ignoring β_0, using a joint confidence coefficient of .95 are shown by the corners of the rectangle enclosing the intersection region.

The implications as to what are "acceptable" combinations of values for the parameters are very different for the two joint confidence regions. The

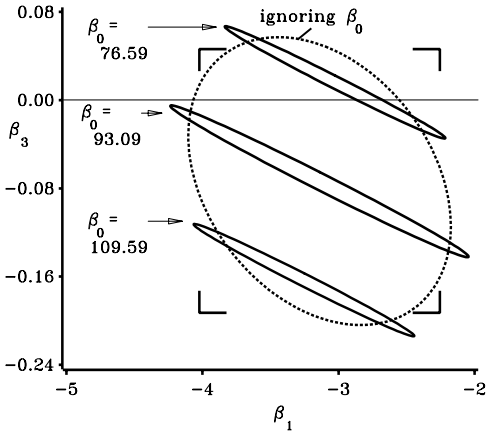

FIGURE 4.1. *Two-dimensional "slices" of the joint confidence region for the regression of oxygen uptake on time to run 1.5 miles* (X_1)*, and heart rate while running* (X_3) *(solid ellipses), and the two-dimensional joint confidence region for* β_1 *and* β_3 *ignoring* β_0 *(dashed ellipse). The intersection of the Bonferroni univariate confidence intervals is shown as the corners of the rectangle formed by the intersection .*

joint confidence region for all parameters is much more restrictive than the bivariate joint confidence region or the univariate confidence intervals would indicate. Allowable combinations of β_1 and β_3 are very dependent on choice of β_0. Clearly, univariate confidence intervals and joint confidence regions that do not involve all parameters can be misleading. ■

The idea of obtaining joint confidence regions in equation 4.60 can also be extended to obtain **joint prediction regions**. Let $\boldsymbol{X}_0 : k \times p'$ be a set of k linearly independent vectors of explanatory variables at which we wish to predict $\boldsymbol{Y}_0$. That is, we wish to simultaneously predict

$$\boldsymbol{Y}_0 = \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{\epsilon}_0, \tag{4.62}$$

where $\boldsymbol{\epsilon}_0$ is $N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_k)$ and is assumed to be independent of $\boldsymbol{Y}$. The best linear unbiased predictor of $\boldsymbol{Y}_0$ is

$$\widehat{\boldsymbol{Y}}_0 = \boldsymbol{X}_0\widehat{\boldsymbol{\beta}}, \tag{4.63}$$

where $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. Note that the prediction error vector

$$\begin{aligned} \boldsymbol{Y}_0 - \widehat{\boldsymbol{Y}}_0 &= \boldsymbol{X}_0(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + \boldsymbol{\epsilon}_0 \\ &\sim N(\boldsymbol{0}, \sigma^2[\boldsymbol{I}_k + \boldsymbol{X}_0(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_0']). \end{aligned} \tag{4.64}$$

A joint $100(1-\alpha)\%$ prediction region is obtained from the inequality

$$(\boldsymbol{Y}_0 - \widehat{\boldsymbol{Y}}_0)'[\boldsymbol{I}_k + \boldsymbol{X}_0(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_0']^{-1}(\boldsymbol{Y}_0 - \widehat{\boldsymbol{Y}}_0) \leq ks^2 F_{(\alpha,k,\nu)}, \quad (4.65)$$

where ν is the degrees of freedom associated with s^2. The Bonferroni prediction intervals are given by

$$(\boldsymbol{Y}_0 - \widehat{\boldsymbol{Y}}_0) \pm t_{(\alpha/2k,\nu)}\boldsymbol{d}\, s, \quad (4.66)$$

where $\boldsymbol{d}$ is a vector of the diagonal elements of $[\boldsymbol{I}_k + \boldsymbol{X}_0(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_0']$. The corresponding Sheffé prediction intervals are given by

$$(\boldsymbol{Y}_0 - \widehat{\boldsymbol{Y}}_0) \pm [kF_{(\alpha,k,\nu)}]^{1/2}\boldsymbol{d}\, s. \quad (4.67)$$

4.7 Estimation of Pure Error

The residual mean square has been used, until now, as the estimate of σ^2. One of the problems with this procedure is the dependence of the residual mean square on the model being fit. Any inadequacies in the model, important independent variables omitted, or an incorrect form of the model will cause the residual mean square to overestimate σ^2. An estimate of σ^2 is needed that is not as dependent on the choice of model being fit at the time.

Definition of Pure Error

The variance σ^2 is the variance of the ϵ_i about zero or, equivalently, the variance of Y_i about their true means $\mathcal{E}(Y_i)$. The concept of modeling Y_i assumes that $\mathcal{E}(Y_i)$ is determined by some unknown function of the relevant independent variables. Let $\boldsymbol{x}_i'$ be the row vector of values of all relevant independent variables for the ith observation. Then, all Y_i that have the same $\boldsymbol{x}_i'$ also will have the same true mean regardless of whether the correct model is known. Hence, σ^2 is by definition the variance among statistically independent observations that have the same $\boldsymbol{x}_i'$. Such repeated observations are called **true replicates**. The sample variance of the Y_i among true replicates provides a direct estimate of σ^2 that is independent of the choice of model. (It is, however, dependent on having identified and taken data on all relevant independent variables.) The estimate of σ^2 obtained from true replication is called **pure error**. When several sets of replicate observations are available, the best estimate of σ^2 is obtained by pooling all estimates.

True replication is almost always included in the design of controlled experiments. For example, the estimate of experimental error from the completely random design or the randomized complete block design when there is no block-by-treatment interaction is the estimate of pure error. **Observational studies**, on the other hand, seldom have true replication since they impose no control over the independent variables. Then, true replication occurs only by chance and is very unlikely if several independent

TABLE 4.8. *Replicate yield data for soybeans exposed to chronic levels of ozone and estimates of pure error. (Data courtesy A. S. Heagle, North Carolina State University.)*

	Ozone Level (ppm)			
	.02	.07	.11	.15
	238.3	235.1	236.2	178.7
	270.7	228.9	208.0	186.0
	210.0	236.2	243.5	206.9
	248.7	255.0	233.0	215.3
	242.4	228.9	233.0	219.5
$\overline{Y}_i$	242.02	236.82	230.74	201.28
s_i^2	476.61	114.83	179.99	325.86

variables are involved. In addition, apparent replicates in the observational data may not, in fact, be true replicates due to important variables having been overlooked. Pseudoreplication or **near replication** is sometimes used with observational data to estimate σ^2. These are sets of observations in which the values of the independent variables fall within a relatively narrow range.

Example 4.14

To illustrate the estimation of pure error, the ozone example used in Example 1.1 is used. The four observations used in that section were the means of five replicate experimental units *at each level of ozone* from a completely random experimental design. The full data set, the treatment means, and the estimates of pure error within each ozone level are given in Table 4.8.

Each s^2 is estimated from the variance among the five observations for each ozone level, with 4 degrees of freedom, and is an unbiased estimate of σ^2. Since each is the variation of Y_{ij} about $\overline{Y}_i$ for a given level of ozone, the estimates are in no way affected by the form of the response model that might be chosen to represent the response of yield to ozone. Figure 4.2 illustrates that the variation among the replicate observations for a given level of ozone is unaffected by the form of the regression line fit to the data. The best estimate of σ^2 is the pooled estimate

$$\begin{aligned} s^2 &= \frac{\sum(n_i-1)s_i^2}{\sum(n_i-1)} = \frac{4(476.61)+\cdots+4(325.86)}{16} \\ &= 274.32 \end{aligned}$$

with 16 degrees of freedom, where $n_i = 4,\ i$ =1, 2, 3, 4.

The analysis of variance for the completely random design is given (Table 4.9) to emphasize that s^2 is the experimental error from that analysis. The previous regression analysis (Section 1.4, Tables 1.3 and 1.4) used the

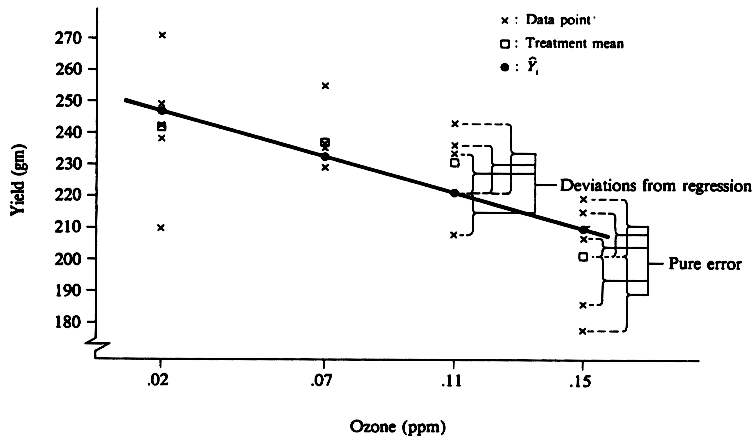

FIGURE 4.2. *Comparison of "pure error" and "deviations from regression" using the data on soybean response to ozone.*

TABLE 4.9. *The analysis of variance for the completely random experimental design for the yield response of soybean to ozone.*

Source	*d.f.*	*SS*	*MS*
$\text{Total}_{(\text{corr})}$	19	9366.61	
Treatments	3	4977.47	1659.16
Regression	1	3956.31	3956.31
Lack of Fit	2	1021.16	510.58
Pure Error	16	4389.14	274.32

treatment means (of $r = 5$ observations). Thus, the sums of squares from that analysis have to be multiplied by $r = 5$ to put them on a "per observation" basis. That analysis of variance, Table 1.4, partitioned the sum of squares among the four treatment means into 1 degree of freedom for the linear regression of Y on ozone level and 2 degrees of freedom for **lack of fit** of linear regression. The middle three lines of Table 4.9 contain the results from the original analysis multiplied by $r = 5$. The numbers differ slightly due to rounding the original means to whole numbers. ■

The expectations of the mean squares in the analysis of variance show what function of the parameters each mean square is estimating. The mean square expectations for the critical lines in Table 4.9 are

$$\begin{aligned} \mathcal{E}[\text{MS(Regr)}] &= \sigma^2 + \beta_1^2 \sum x_i^2, \\ \mathcal{E}[\text{MS(Lack of fit)}] &= \sigma^2 + (\text{Model bias})^2, \\ \mathcal{E}[\text{MS(Pure error)}] &= \sigma^2. \end{aligned} \tag{4.68}$$

Recall that $\sum x_i^2$ is used to indicate the *corrected* sum of squares of the independent variable.

The square on "model bias" emphasizes that any inadequacies in the model cause this mean square to be larger, in expectation, than σ^2. Thus, the "lack of fit" mean square is an unbiased estimate of σ^2 only if the linear model is correct. Otherwise, it is biased upwards. On the other hand, the "pure error" estimate of σ^2 obtained from the replication in the experiment is unbiased regardless of whether the assumed linear relationship is correct.

The mean square expectation of MS(Regr) is shown as if the linear model relating yield to ozone level is correct. If the model is not correct (for example, if the treatment differences are not due solely to ozone differences), the second term in $\mathcal{E}$[MS(Regr)] will include contributions from all variables that are correlated with ozone levels. This is the case even if the variables have not been identified. The advantage of controlled experiments such as this ozone study is that amount of ozone is, presumably, the only variable changing consistently over the ozone treatments. Random assignment of treatments to the experimental units should destroy any correlation between ozone level and any incidental environmenal variable. Thus, treatment differences in this controlled study can be attributed to the effects of ozone and $\mathcal{E}$[MS(Regr)] should not be biased by the effects of any uncontrolled variables. One should not overlook, however, this potential for bias in the regression sum of squares, particularly when observational data are being analyzed.

Adequacy of the Model

The independent estimate of pure error, experimental error, provides the basis for two important tests of significance. The **adequacy of the model** can be checked by testing the null hypothesis that "model bias" is zero. Any inadequacies in the linear model will make this mean square larger than

σ^2 on the average. Such inadequacies could include omitted independent variables as well as any curvilinear response to ozone.

Example 4.15

In the ozone example, Example 4.14, the test of the adequacy of the linear model is

$$F = \frac{\text{MS(Lack of fit)}}{\text{MS(Pure error)}} = \frac{510.58}{274.32} = 1.86,$$

which, if the model is correct, is distributed as F with 2 and 16 degrees of freedom. Comparison against the critical value $F_{(.05,2,16)} = 3.63$ shows this to be nonsignificant, indicating that there is no evidence in these data that the linear model is inadequate for representing the response of soybean to ozone. ■

$H_0 : \beta_1 = 0$

The second hypothesis of interest is $H_0 : \beta_1 = 0$ against the alternative hypothesis $H_a : \beta_1 \neq 0$. If the fitted model is *not* adequate, then the parameter β_1 may not have the same interpretation as when the model is adequate. Therefore, when the model is not adequate, it does not make sense to test $H_0 : \beta_1 = 0$.

Suppose that the fitted model is adequate and we are interested in testing $H_0 : \beta_1 = 0$. The ratio of regression mean square to an estimate of σ^2 provides a test of this hypothesis. The mean square expectations show that both mean squares estimate σ^2 when the null hypothesis is true and that the numerator becomes increasingly larger as β_1 deviates from zero. One estimate of σ^2 is, again, the pure error estimate or experimental error.

Example 4.16

For the ozone example, a test statistic for testing $H_0 : \beta_1 = 0$ is

$$F = \frac{\text{MS(Regr)}}{\text{MS(Pure error)}} = \frac{3,956.31}{274.32} = 14.42.$$

Comparing this to the critical value for $\alpha = .01$, $F_{(.01,1,16)} = 8.53$, indicates that the null hypothesis that $\beta_1 = 0$ should be rejected. This conclusion differs from that of the analysis in Chapter 1 because σ^2 is now estimated with many more degrees of freedom. As a result, the test has more power for detecting departures from the null hypothesis. ■

Note that, if the model is truly adequate, then the mean square for lack of fit is also an estimate of σ^2. A pooled estimate of σ^2 is given by the sum of SS(Lack of fit) and SS(Pure error) divided by the sum of the corresponding degrees of freedom.

Example 4.17

For the ozone example, consider the analysis of variance given in Ta-

TABLE 4.10. *The analysis of variance for the ozone data.*

Source	*d.f.*	*SS*	*MS*
Total $_{(\text{corr})}$	19	9,366.61	
Regression	1	3,956.31	3,956.31
Error	18	5,410.30	300.57
Lack of Fit	2	1,021.16	510.58
Pure Error	16	4,389.14	274.32

ble 4.10. Based on the pooled error, a test statistic for testing $H_0 : \beta_1 = 0$ is

$$F = \frac{\text{MS(Regression)}}{\text{MS(Error)}} = \frac{3,956.31}{300.57} = 13.16.$$

Comparing this to the critical value for $\alpha = .01$, $F_{(.01,1,18)} = 8.29$, indicates that $H_0 : \beta_1 = 0$ should be rejected. This F-statistic coincides with the F-statistic given in Chapter 1 for testing $H_0 : \beta_1 = 0$ in the model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ when all of the data in Table 4.8 (instead of only the means, Table 1.1) are used. This test statistic is more powerful than that based on the MS(Pure error). However, if the fitted model is inadequate, then MS(Error) is no longer an unbiased estimate of σ^2, whereas MS(Pure error) is even if the fitted model is not adequate.

Finally, a composite test for $H_0 : \beta_1 = 0$ and that the model is adequate is given by

$$\begin{aligned} F &= \frac{[\text{SS(Regression)} + \text{SS(Lack of fit)}]/(1+2)}{\text{MS(Pure Error)}} \\ &= \frac{(3,956.31 + 1,021.16)/3}{274.32} = \frac{1659.16}{274.32} \\ &= 6.05. \end{aligned}$$

Comparing this to the critical value for $\alpha = .01$, $F_{(.01,3,16)} = 3.24$, indicates that either the model is not adequate or β_1 is not zero. This is equivalent to testing the null hypothesis of no treatment effects in the analysis of variance which is discussed in Chapter 9. ■

In summary, multiple, statistically independent observations on the dependent variable for given values of all relevant independent variables is called true replication. True replication provides for an unbiased estimate of σ^2 that is not dependent on the model being used. The estimate of pure error provides a basis for testing the adequacy of the model. True replication should be designed into all studies where possible and the pure error estimate of σ^2, rather than a residual mean square estimate, used for tests of significance and standard errors.

4.8 Exercises

4.1. A dependent variable $\boldsymbol{Y}$ (20×1) was regressed onto 3 independent variables plus an intercept (so that $\boldsymbol{X}$ was of dimension 20×4). The following matrices were computed.

$$\boldsymbol{X'X} = \begin{bmatrix} 20 & 0 & 0 & 0 \\ 0 & 250 & 401 & 0 \\ 0 & 401 & 1,013 & 0 \\ 0 & 0 & 0 & 128 \end{bmatrix} \qquad \boldsymbol{X'Y} = \begin{pmatrix} 1,900.00 \\ 970.45 \\ 1,674.41 \\ -396.80 \end{pmatrix}$$

$$\boldsymbol{Y'Y} = 185.883.$$

(a) Compute $\widehat{\boldsymbol{\beta}}$ and write the regression equation.

(b) Compute the analysis of variance of $\boldsymbol{Y}$. Partition the sum of squares due to the model into a part due to the mean and a part due to regression on the Xs *after adjustment* for the mean. Summarize the results, including degrees of freedom and mean squares, in an analysis of variance table.

(c) Compute the estimate of σ^2 and the standard error for each regression coefficient. Compute the covariance between $\widehat{\beta}_1$ and $\widehat{\beta}_2$, $\text{Cov}(\widehat{\beta}_1,\ \widehat{\beta}_2)$. Compute the covariance between $\widehat{\beta}_1$ and $\widehat{\beta}_3$, $\text{Cov}(\widehat{\beta}_1,\ \widehat{\beta}_3)$.

(d) Drop X_3 from the model. Reconstruct $\boldsymbol{X'X}$ and $\boldsymbol{X'Y}$ for this model without X_3 and repeat Questions (a) and (b). Put X_3 back in the model but drop X_2 and repeat (a) and (b).

(i) Which of the two independent variables X_2 or X_3 made the greater contribution to Y in the presence of the remaining Xs; that is, compare $R(\beta_2|\beta_0,\ \beta_1,\ \beta_3)$ and $R(\beta_3|\beta_0,\ \beta_1,\ \beta_2)$.

(ii) Explain why $\widehat{\beta}_1$ changed in value when X_2 was dropped but not when X_3 was dropped.

(iii) Explain the differences in meaning of β_1 in the three models.

(e) From inspection of $\boldsymbol{X'X}$ how can you tell that $X_1,\ X_2$, and X_3 were expressed as deviations from their respective means? Would $(\boldsymbol{X'X})^{-1}$ have been easier or harder to obtain if the original Xs (without subtraction of their means) had been used? Explain.

4.2. A regression analysis led to the following $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$ matrix and estimate of σ^2.

$$\frac{1}{70}\begin{bmatrix} 62 & 18 & -6 & -10 & 6 \\ 18 & 26 & 24 & 12 & -10 \\ -6 & 24 & 34 & 24 & -6 \\ -10 & 12 & 24 & 26 & 18 \\ 6 & -10 & -6 & 18 & 62 \end{bmatrix}, \qquad s^2 = .06.$$

(a) How many observations were in the data set?

(b) How many linearly independent columns are in $\boldsymbol{X}$—that is, what is the rank of $\boldsymbol{X}$? How many degrees of freedom are associated with the *model* sum of squares? Assuming the model contained an intercept, how many degrees of freedom are associated with the *regression* sum of squares?

(c) Suppose $\boldsymbol{Y} = (\,82 \quad 80 \quad 75 \quad 67 \quad 55\,)'$. Compute the estimated mean $\widehat{Y}_1$ of Y corresponding to the first observation. Compute $s^2(\widehat{Y}_1)$. Find the residual e_1 for the first observation and compute its variance. For which data point will $\widehat{Y}_i$ have the smallest variance? For which data point will e_i have the largest variance?

4.3. The following $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, $\widehat{\boldsymbol{\beta}}$, and residual sum of squares were obtained from the regression of plant dry weight (grams) from $n = 7$ experimental fields on percent soil organic matter (X_1) and kilograms of supplemental nitrogen per 1000 m^2 (X_2). The regression model included an intercept.

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 1.7995972 & -.0685472 & -.2531648 \\ -.0685472 & .0100774 & -.0010661 \\ -.2531648 & -.0010661 & .0570789 \end{bmatrix}$$

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 51.5697 \\ 1.4974 \\ 6.7233 \end{pmatrix}, \qquad \mathrm{SS(Res)} = 27.5808.$$

(a) Give the regression equation and interpret each regression coefficient. Give the units of measure of each regression coefficient.

(b) How many degrees of freedom does SS(Res) have? Compute s^2, the variance of $\widehat{\beta}_1$, and the covariance of $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

(c) Determine the 95% univariate confidence interval estimates of β_1 and β_2. Compute the Bonferroni and the Scheffé confidence intervals for β_1 and β_2 using a joint confidence coefficient of .95.

(d) Suppose previous experience has led you to believe that one percentage point increase in organic matter is equivalent to .5 kilogram/1,000 m^2 of supplemental nitrogen in dry matter production. Translate this statement into a null hypothesis on the regression coefficients. Use a t-test to test this null hypothesis against the alternative hypothesis that supplemental nitrogen is *more* effective than this statement would imply.

(e) Define $\boldsymbol{K}'$ and $\boldsymbol{m}$ for the general linear hypothesis $H_0 : \boldsymbol{K}'\boldsymbol{\beta} - \boldsymbol{m} = \boldsymbol{0}$ for testing $H_0 : \ 2\beta_1 = \beta_2$. Compute Q and complete the test of significance using the F-test. What is the alternative hypothesis for this test?

(f) Give the reduced model you obtain if you impose the null hypothesis in (e) on the model. Suppose this reduced model gave a SS(Res) = 164.3325. Use this result to complete the test of the hypothesis.

4.4. The following analysis of variance summarizes the regression of Y on two independent variables plus an intercept.

Source	*d.f.*	*SS*	*MS*
Total$_{(\text{ corr})}$	26	1,211	
Regression	2	1,055	527.5
Residual	24	156	6.5
Variable	*Sequential SS*		*Partial SS*
X_1	263		223
X_2	792		792

(a) Your estimate of β_1 is $\widehat{\beta}_1 = 2.996$. A friend of yours regressed Y on X_1 and found $\widehat{\beta}_1 = 3.24$. Explain the difference in these two estimates.

(b) Label each sequential and partial sum of squares using the R-notation. Explain what $R(\beta_1|\beta_0)$ measures.

(c) Compute $R(\beta_2|\beta_0)$ and explain what it measures.

(d) What is the regression sum of squares due to X_1 after adjustment for X_2?

(e) Make a test of significance (use $\alpha = .05$) to determine if X_1 should be retained in the model with X_2.

(f) The original data contained several sets of observations having the same values of X_1 and X_2. The pooled variance from these replicate observations was $s^2 = 3.8$ with eight degrees of freedom. With this information, rewrite the analysis of variance to show the partitions of the "residual" sum of squares into "pure error" and "lack of fit." Make a test of significance to determine whether the model using X_1 and X_2 is adequate.

4.5. The accompanying table presents data on one dependent variable and five independent variables.

Y	X_1	X_2	X_3	X_4	X_5
6.68	32.6	4.78	1,092	293.09	17.1
6.31	33.4	4.62	1,279	252.18	14.0
7.13	33.2	3.72	511	109.31	12.7
5.81	31.2	3.29	518	131.63	25.7
5.68	31.0	3.25	582	124.50	24.3
7.66	31.8	7.35	509	95.19	.3
7.30	26.4	4.92	942	173.25	21.1
6.19	26.2	4.02	952	172.21	26.1
7.31	26.6	5.47	792	142.34	19.8

(a) Give the linear model in matrix form for regressing Y on the five independent variables. Completely define each matrix and give its order and rank.

(b) The following quadratic forms were computed.

$$\begin{array}{llll} \boldsymbol{Y}'\boldsymbol{P}\boldsymbol{Y} & = 404.532 & \boldsymbol{Y}'\boldsymbol{Y} & = 405.012 \\ \boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y} & = 0.480 & \boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{J}/n)\boldsymbol{Y} & = 4.078 \\ \boldsymbol{Y}'(\boldsymbol{P}-\boldsymbol{J}/n)\boldsymbol{Y} & = 3.598 & \boldsymbol{Y}'(\boldsymbol{J}'/n)\boldsymbol{Y} & = 400.934. \end{array}$$

Use a matrix algebra computer program to reproduce each of these sums of squares. Use these results to give the complete analysis of variance summary.

(c) The partial sums of squares for X_1, X_2, X_3, X_4, and X_5 are .895, .238, .270, .337, and .922, respectively. Give the R-notation that describes the partial sum of squares for X_2. Use a matrix algebra program to verify the partial sum of squares for X_2.

(d) Assume that none of the partial sums of squares for X_2, X_3, and X_4 is significant and that the partial sums of squares for X_1 and X_5 are significant (at $\alpha = .05$). Indicate whether each of the following statements is valid based on these results. If it is not a valid statement, explain why.

(i) X_1 and X_5 are important causal variables whereas X_2, X_3, and X_4 are not.

(ii) X_2, X_3, and X_4 can be dropped from the model with no meaningful loss in predictability of Y.

(iii) There is no need for all five independent variables to be retained in the model.

4.6. This exercise continues with the analysis of the peak water flow data used in Exercise 3.12. In that exercise, several regressions were run to relate $Y = \ln(Q_0/Q_p)$ to three characteristics of the watersheds and a measure of storm intensity. Y measures the discrepancy between peak water flow predicted from a simulation model (Q_p) and observed

peak water flow (Q_0). The four independent variables are described in Exercise 3.12.

(a) The first model used an intercept and all four independent variables.

(i) Compute SS(Model), SS(Regr), and SS(Res) for this model and summarize the results in the analysis of variance table. Show degrees of freedom and mean squares.

(ii) Obtain the partial sum of squares for each independent variable and the sequential sums of squares for the variables added to the model in the order X_1, X_4, X_2, X_3.

(iii) Use tests of significance ($\alpha = .05$) to determine which partial regression coefficients are different from zero. What do these tests suggest as to which variables might be dropped from the model?

(iv) Construct a test of the null hypothesis $H_0 : \beta_0 = 0$ using the general linear hypothesis. What do you conclude from this test?

(b) The second model used the four independent variables but forced the intercept to be zero.

(i) Compute SS(Model), SS(Res), and the partial and sequential sums of squares for this model. Summarize the results in the analysis of variance table.

(ii) Use the difference in SS(Res) between this model with no intercept and the previous model with an intercept to test $H_0 : \beta_0 = 0$. Compare the result with that obtained under (iv) in Part (a).

(iii) Use tests of significance to determine which partial regression coefficients in this model are different from zero. What do these tests tell you in terms of simplifying the model?

(c) The third model used the zero-intercept model and only X_1 and X_4.

(i) Use the results from this model and the zero-intercept model in Part (b) to test the composite null hypothesis that β_2 and β_3 are both zero.

(ii) Use the general linear hypothesis to construct the test of the composite null hypothesis that β_2 and β_3 in the model in Part (b) are both zero. Define $\boldsymbol{K}'$ and $\boldsymbol{m}$ for this hypothesis, compute Q, and complete the test of significance. Compare these two tests.

4.7. Use the data on annual catch of Gulf Menhaden, number of fishing vessels, and fishing effort given in Exercise 3.11.

(a) Complete the analysis of variance for the regression of catch (Y) on fishing effort (X_1) and number of vessels (X_2) with an intercept in the model. Determine the partial sums of squares for each independent variable. Estimate the standard errors for the regression coefficients and construct the Bonferroni confidence intervals for each using a joint confidence coefficient of 95%. Use the regression equation to predict the "catch" if number of vessels is limited to $X_2 = 70$ and fishing effort is restricted to $X_1 = 400$. Compute the variance of this prediction and the 95% confidence interval estimate of the prediction.

(b) Test the hypothesis that the variable "number of vessels" does not add significantly to the explanation of variation in "catch" provided by "fishing effort" alone (use $\alpha = .05$). Test the hypothesis that "fishing effort" does not add significantly to the explanation provided by "number of vessels" alone.

(c) On the basis of the tests in Part (b) would you keep both X_1 and X_2 in the model, or would you eliminate one from the model? If one should be eliminated, which would it be? Does the remaining variable make a significant contribution to explaining the variation in "catch"?

(d) Suppose consideration is being given to controlling the annual catch by limiting either the number of fishing vessels or the total fishing effort. What is your recommendation and why?

4.8. This exercise uses the data in Exercise 3.14 relating $Y = \ln(\text{days survival})$ for colon cancer patients receiving supplemental ascorbate to the variables sex (X_1), age of patient (X_2), and ln(average survival of control group) (X_3).

(a) Complete the analysis of variance for the model using all three variables plus an intercept. Compute the partial sum of squares for each independent variable using the formula $\widehat{\beta}_j^2/c_{jj}$. Demonstrate that each is the same as the sum of squares one obtains by computing Q for the general linear hypothesis that the corresponding β_j is zero. Compute the standard error for each regression coefficient and the 95% confidence interval estimates.

(b) Does information on the length of survival time of the control group (X_3) help explain the variation in Y? Support your answer with an appropriate test of significance.

(c) Test the null hypothesis that "sex of patient" has no effect on survival beyond that accounted for by "age" and survival of the control group. Interpret the results.

(d) Test the null hypothesis that "age of patient" has no effect on survival beyond that accounted for by "sex" and survival time of the control group. Intrepret the results.

(e) Test the composite hypothesis that $\beta_1 = \beta_2 = \beta_3 = 0$. From these results, what do you conclude about the effect of sex and age of patient on the mean survival time of patients in this study receiving supplemental ascorbate? With the information available in these data, what would you use as the best estimate of the mean ln(days survival)?

4.9. The Lesser–Unsworth data (Exercise 1.19) was used in Exercise 3.9 to estimate a bivariate regression equation relating seed weight to cumulative solar radiation and level of ozone pollution. This exercise continues with the analysis of that model using the centered independent variables.

(a) The more complex model used in Exercise 3.9 included the independent variables cumulative solar radiation, ozone level, and the product of cumulative solar radiation and ozone level (plus an intercept).

(i) Construct the analysis of variance for this model showing sums of squares, degrees of freedom, and mean squares. What is the estimate of σ^2?

(ii) Compute the standard errors for each regression coefficient. Use a *joint* confidence coefficient of 90% and construct the Bonferroni confidence intervals for the four regression coefficients. Use the confidence intervals to draw conclusions about which regression coefficients are clearly different from zero.

(iii) Construct a test of the null hypothesis that the regression coefficient for the product term is zero (use $\alpha = .05$). Does your conclusion from this test agree with your conclusion based on the Bonferroni confidence intervals? Explain why they need not agree.

(b) The simpler model in Exercise 3.9 did not use the product term. Construct the analysis of variance for the model using only the two independent variables cumulative solar radiation and ozone level.

(i) Use the residual sums of squares from the two analyses to test the null hypothesis that the regression coefficient on the product term is zero (use $\alpha = .05$). Does your conclusion agree with that obtained in Part (a)?

(ii) Compute the standard errors of the regression coefficients for this reduced model. Explain why they differ from those computed in Part (a).

(iii) Compute the estimated mean seed weight for the mean level of cumulative solar radiation and .025 ppm ozone. Compute

the estimated mean seed weight for the mean level of radiation and .07 ppm ozone. Use these two results to compute the estimated mean loss in seed weight if ozone changes from .025 to .07 ppm. Define a matrix of coefficients $\boldsymbol{K}'$ such that these three linear functions of $\widehat{\boldsymbol{\beta}}$ can be written as $\boldsymbol{K}'\widehat{\boldsymbol{\beta}}$. Use this matrix form to compute their variances and covariances.

(iv) Compute and plot the 90% *joint* confidence region for β_1 and β_2 *ignoring* β_0. (This joint confidence region will be an ellipse in the two dimensions β_1 and β_2.)

4.10. This is a continuation of Exercise 3.10 using the number of hospital days for smokers from Exercise 1.21. The dependent variable is $Y =$ ln(number of hospital days for smokers). The independent variables are $X_1 =$ (number of cigarettes)^2 and $X_2 =$ ln(number of hospital days for nonsmokers). Note that X_1 is the square of number of cigarettes.

(a) Plot Y against number of cigarettes and against the square of number of cigarettes. Do the plots provide any indication of why the square of number of cigarettes was chosen as the independent variable?

(b) Complete the analysis of variance for the regression of Y on X_1 and X_2. Does the information on number of hospital days for nonsmokers help explain the variation in number of hospital days for smokers? Make an appropriate test of significance to support your statement. Is Y, after adjustment for number of hospital days for nonsmokers, related to X_1? Make a test of significance to support your statement. Are you willing to conclude from these data that number of cigarettes smoked has a direct effect on the average number of hospital days?

(c) It is logical in this problem to expect the number of hospital days for smokers to approach that of nonsmokers as the number of cigarettes smoked goes to zero. This implies that the intercept in this model might be expected to be zero. One might also expect β_2 to be equal to one. (Explain why.) Set up the general linear hypothesis for testing the composite null hypothesis that $\beta_0 = 0$ and $\beta_2 = 1.0$. Complete the test of significance and state your conclusions.

(d) Construct the *reduced* model implied by the composite null hypothesis under (c). Compute the regression for this reduced model, obtain the residual sum of squares, and use the difference in residual sums of squares for the full and reduced models to test the composite null hypothesis. Do you obtain the same result as in (c)?

(e) Based on the preceding tests of significance, decide which model you feel is appropriate. State the regression equation for your adopted model. Include standard errors on the regression coefficients.

4.11. You are given the following matrices computed for a regression analysis.

$$\boldsymbol{X'X} = \begin{bmatrix} 9 & 136 & 269 & 260 \\ 136 & 2114 & 4176 & 3583 \\ 269 & 4176 & 8257 & 7104 \\ 260 & 3583 & 7104 & 12276 \end{bmatrix} \qquad \boldsymbol{X'Y} = \begin{pmatrix} 45 \\ 648 \\ 1,283 \\ 1,821 \end{pmatrix}$$

$$(\boldsymbol{X'X})^{-1} = \begin{bmatrix} 9.610932 & .0085878 & -.2791475 & -.0445217 \\ .0085878 & .5099641 & -.2588636 & .0007765 \\ -.2791475 & -.2588636 & .1395 & .0007396 \\ -.0445217 & .0007765 & .0007396 & .0003698 \end{bmatrix}$$

$$(\boldsymbol{X'X})^{-1}\boldsymbol{X'Y} = \begin{bmatrix} -1.163461 \\ .135270 \\ .019950 \\ .121954 \end{bmatrix}, \qquad \boldsymbol{Y'Y} = 285.$$

(a) Use the preceding results to complete the analysis of variance table.

(b) Give the computed regression equation and the standard errors of the regression coefficients.

(c) Compare each estimated regression coefficient to its standard error and use the t-test to test the simple hypothesis that each regression coefficient is equal to zero. State your conclusions (use $\alpha = .05$).

(d) Define the $\boldsymbol{K'}$ and $\boldsymbol{m}$ for the composite hypothesis that $\beta_0 = 0$, $\beta_1 = \beta_3$, and $\beta_2 = 0$. Give the rank of $\boldsymbol{K'}$ and the degrees of freedom associated with this test.

(e) Give the reduced model for the composite hypothesis in Part (d).

4.12. You are given the following *sequential* and *partial* sums of squares from a regression analysis.

$$\begin{aligned} R(\beta_3|\beta_0) &= 56.9669 & R(\beta_3|\beta_0\ \beta_1\ \beta_2) &= 40.2204 \\ R(\beta_1|\beta_0\ \beta_3) &= 1.0027 & R(\beta_1|\beta_0\ \beta_2\ \beta_3) &= .0359 \\ R(\beta_2|\beta_0\ \beta_1\ \beta_3) &= .0029 & R(\beta_2|\beta_0\ \beta_1\ \beta_3) &= .0029. \end{aligned}$$

Each sequential and partial sum of squares can be used for the numerator of an F-test. Clearly state the null hypothesis being tested in each case.

4.13. A regression analysis using an intercept and one independent variable gave

$$\widehat{Y}_i = 1.841246 + .10934X_{i1}.$$

The variance–covariance matrix for $\widehat{\boldsymbol{\beta}}$ was

$$\boldsymbol{s^2}(\widehat{\beta}) = \begin{bmatrix} .1240363 & -.002627 \\ -.002627 & .0000909 \end{bmatrix}.$$

(a) Compute the 95% confidence interval estimate of $\widehat{\beta}_1$. The estimate of σ^2 used to compute $s^2(\widehat{\beta}_1)$ was $s^2 = 1.6360$, the residual mean square from the model using only X_0 and X_1. The data had $n = 34$ observations.

(b) Compute $\widehat{Y}$ for $X_1 = 4$. Compute the variance of $\widehat{Y}$ if it is being used to estimate the mean of Y when $X_1 = 4$. Compute the variance of $\widehat{Y}$ if it is being used to *predict* a future observation at $X_1 = 4$.

4.14. You are given the following matrix of simple (product moment) correlations among a dependent variable Y (first variable) and three independent variables.

$$\begin{bmatrix} 1.0 & -.538 & -.543 & .974 \\ -.538 & 1.0 & .983 & -.653 \\ -.543 & .983 & 1.0 & -.656 \\ .974 & -.653 & -.656 & 1.0 \end{bmatrix}.$$

(a) From inspection of the correlation matrix, which independent variable would account for the greatest variability in Y? What proportion of the corrected sum of squares in Y would be accounted for by this variable? If Y were regressed on all three independent variables (plus an intercept), would the coefficient of determination for the multiple regression be smaller or larger than this proportion?

(b) Inspection of the three pairwise correlations among the X variables suggests that at least one of the independent variables will not be useful for the regression of Y on the Xs. Explain exactly the basis for this statement and why it has this implication.

4.15. Let $\boldsymbol{X}$ be an $n \times p'$ matrix with rank p'. Suppose the first column of $\boldsymbol{X}$ is $\mathbf{1}$, a column of 1s. Then, show that

(a) $\boldsymbol{P}\mathbf{1} = \mathbf{1}$.

(b) $(\boldsymbol{J}/n)$, $\boldsymbol{P} - \boldsymbol{J}/n$, and $(\boldsymbol{I} - \boldsymbol{P})$ are idempotent and pairwise orthogonal, where $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{J}/n$ is given in equation 4.22.

4.16. Let $\boldsymbol{X}$ be a full rank $n \times p'$ matrix given in equation 3.2. For $\boldsymbol{J}$ given in equation 4.22,

(a) show that

$$(\boldsymbol{I} - \boldsymbol{J}/n)\boldsymbol{X} = \begin{bmatrix} 0 & x_{11} & x_{12} & \cdots & x_{1p} \\ 0 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix},$$

where $x_{ij} = X_{ij} - \overline{X}_{\cdot j}$ and $\overline{X}_{\cdot j} = n^{-1}\sum_{i=1}^{n} X_{ij}$; and

(b) hence, show that $\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{J}/n)\boldsymbol{X}$ has zero in each entry of the first row and first column.

5

CASE STUDY: FIVE INDEPENDENT VARIABLES

The last two chapters completed the presentation of the basic regression results for linear models with any number of variables.

This chapter demonstrates the application of least squares regression to a problem involving five independent variables. The full model is fit and then the model is simplified to a two-variable model that conveys most of the information on Y.

The basic steps in ordinary regression analysis have now been covered. This chapter illustrates the application of these methods. Computations and interpretations of the regression results are emphasized.

5.1 Spartina Biomass Production in the Cape Fear Estuary

The data considered are part of a larger study conducted by Dr. Rick Linthurst (1979) at North Carolina State University as his Ph.D. thesis research. The purpose of his research was to identify the important soil characteristics influencing aerial biomass production of the marsh grass *Spartina alterniflora* in the Cape Fear Estuary of North Carolina.

One phase of Linthurst's research consisted of sampling three types of **Design**

Spartina vegetation (revegetated "dead" areas, "short" Spartina areas, and "tall" Spartina areas) in each of three locations (Oak Island, Smith Island, and Snows Marsh). Samples of the soil substrate from 5 random sites within each location–vegetation type (giving 45 total samples) were analyzed for 14 soil physicochemical characteristics each month for several months. In addition, above-ground biomass at each sample site was measured each month. The data used in this case study involve only the September sampling and these five substrate measurements:

$X_1 =$ salinity $^{\circ}/_{\infty}(SALINITY)$

$X_2 =$ acidity as measured in water (pH)

$X_3 =$ potassium ppm (K)

$X_4 =$ sodium ppm (Na)

$X_5 =$ zinc ppm (Zn).

The dependent variable Y is aerial biomass gm^{-2}. The data from the September sampling for these six variables are given in Table 5.1. The objective of this phase of the Linthurst research was to identify the substrate variables showing the stronger relationships to biomass. These variables would then be used in controlled studies to investigate causal relationships. The purpose of this case study is to use multiple linear regression to relate *total* variability in Spartina biomass production to total variability in the five substrate variables. For this analysis, total variation among vegetation types, locations, and samples within vegetation types and locations is being used. It is left as an exercise for the student to study separately the relationships shown by the variation *among* vegetation types and locations (using the location–vegetation type means) and the relationships shown by the variation among samples within location–vegetation type combinations.

Objective

5.2 Regression Analysis for the Full Model

The initial model assumes that *BIOMASS*, Y, can be adequately characterized by linear relationships with the five independent variables plus an intercept. Thus, the linear model

Model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{5.1}$$

is completely specified by defining $\boldsymbol{Y}$, $\boldsymbol{X}$, and $\boldsymbol{\beta}$ and stating the appropriate assumptions about distribution of the random errors $\boldsymbol{\epsilon}$. $\boldsymbol{Y}$ is the vector of *BIOMASS* measurements

$$\boldsymbol{Y} = (\,676 \quad 516 \quad \cdots \quad 1,560\,)'\,.$$

TABLE 5.1. *Aerial biomass (BIO) and five physicochemical properties of the substrate (salinity (SAL), pH, K, Na, and Zn) in the Cape Fear Estuary of North Carolina. (Data used with permission of Dr. R. A. Linthurst.)*

Obs.	*Loc.*	*Type*	*BIO*	*SAL*	*pH*	*K*	*Na*	*Zn*
1	OI	DVEG	676	33	5.00	1,441.67	35,185.5	16.4524
2	OI	DVEG	516	35	4.75	1,299.19	28,170.4	13.9852
3	OI	DVEG	1,052	32	4.20	1,154.27	26,455.0	15.3276
4	OI	DVEG	868	30	4.40	1,045.15	25,072.9	17.3128
5	OI	DVEG	1,008	33	5.55	521.62	31,664.2	22.3312
6	OI	SHRT	436	33	5.05	1,273.02	25,491.7	12.2778
7	OI	SHRT	544	36	4.25	1,346.35	20,877.3	17.8225
8	OI	SHRT	680	30	4.45	1,253.88	25,621.3	14.3516
9	OI	SHRT	640	38	4.75	1,242.65	27,587.3	13.6826
10	OI	SHRT	492	30	4.60	1,281.95	26,511.7	11.7566
11	OI	TALL	984	30	4.10	553.69	7,886.5	9.8820
12	OI	TALL	1,400	37	3.45	494.74	14,596.0	16.6752
13	OI	TALL	1,276	33	3.45	525.97	9,826.8	12.3730
14	OI	TALL	1,736	36	4.10	571.14	11,978.4	9.4058
15	OI	TALL	1,004	30	3.50	408.64	10,368.6	14.9302
16	SI	DVEG	396	30	3.25	646.65	17,307.4	31.2865
17	SI	DVEG	352	27	3.35	514.03	12,822.0	30.1652
18	SI	DVEG	328	29	3.20	350.73	8,582.6	28.5901
19	SI	DVEG	392	34	3.35	496.29	12,369.5	19.8795
20	SI	DVEG	236	36	3.30	580.92	14,731.9	18.5056
21	SI	SHRT	392	30	3.25	535.82	15,060.6	22.1344
22	SI	SHRT	268	28	3.25	490.34	11,056.3	28.6101
23	SI	SHRT	252	31	3.20	552.39	8,118.9	23.1908
24	SI	SHRT	236	31	3.20	661.32	13,009.5	24.6917
25	SI	SHRT	340	35	3.35	672.15	15,003.7	22.6758
26	SI	TALL	2,436	29	7.10	528.65	10,225.0	0.3729
27	SI	TALL	2,216	35	7.35	563.13	8,024.2	0.2703
28	SI	TALL	2,096	35	7.45	497.96	10,393.0	0.3205
29	SI	TALL	1,660	30	7.45	458.38	8,711.6	0.2648
30	SI	TALL	2,272	30	7.40	498.25	10,239.6	0.2105
31	SM	DVEG	824	26	4.85	936.26	20,436.0	18.9875
32	SM	DVEG	1,196	29	4.60	894.79	12,519.9	20.9687
33	SM	DVEG	1,960	25	5.20	941.36	18,979.0	23.9841
34	SM	DVEG	2,080	26	4.75	1,038.79	22,986.1	19.9727
35	SM	DVEG	1,764	26	5.20	898.05	11,704.5	21.3864
36	SM	SHRT	412	25	4.55	989.87	17,721.0	23.7063
37	SM	SHRT	416	26	3.95	951.28	16,485.2	30.5589
38	SM	SHRT	504	26	3.70	939.83	17,101.3	26.8415
39	SM	SHRT	492	27	3.75	925.42	17,849.0	27.7292
40	SM	SHRT	636	27	4.15	954.11	16,949.6	21.5699
41	SM	TALL	1,756	24	5.60	720.72	11,344.6	19.6531
42	SM	TALL	1,232	27	5.35	782.09	14,752.4	20.3295
43	SM	TALL	1,400	26	5.50	773.30	13,649.8	19.5880
44	SM	TALL	1,620	28	5.50	829.26	14,533.0	20.1328
45	SM	TALL	1,560	28	5.40	856.96	16,892.2	19.2420

$\boldsymbol{X}$ (45×6) consists of the column vector $\mathbf{1}$, the 45×1 column vector of ones, and the five column vectors of data for the substrate variables $\boldsymbol{X}_1 = SALINITY$, $\boldsymbol{X}_2 = pH$, $\boldsymbol{X}_3 = K$, $\boldsymbol{X}_4 = Na$, and $\boldsymbol{X}_5 = Zn$:

$$\begin{aligned} \boldsymbol{X} &= [\,\mathbf{1} \quad \boldsymbol{X}_1 \quad \boldsymbol{X}_2 \quad \boldsymbol{X}_3 \quad \boldsymbol{X}_4 \quad \boldsymbol{X}_5\,] \\ &= \begin{bmatrix} 1 & 33 & 5.00 & 1,441.67 & 35,184.5 & 16.4524 \\ 1 & 35 & 4.75 & 1,299.19 & 28,170.4 & 13.9852 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 28 & 5.40 & 856.96 & 16,892.2 & 19.2420 \end{bmatrix}. \end{aligned} \tag{5.2}$$

The vector of parameters is

$$\boldsymbol{\beta} = (\,\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_4 \quad \beta_5\,)'. \tag{5.3}$$

The random errors $\boldsymbol{\epsilon}$ are assumed to be normally distributed, $\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \boldsymbol{I}\sigma^2\right)$. The assumption that the variance–covariance matrix for $\boldsymbol{\epsilon}$ is $\boldsymbol{I}\sigma^2$ contains the two assumptions of independence of the errors and common variance.

5.2.1 The Correlation Matrix

A useful starting point in any multiple regression analysis is to compute the matrix of correlations among all variables including the dependent variable. This provides a "first look" at the simple linear relationships among the variables. The correlation matrix is obtained by

$$\widehat{\boldsymbol{\rho}} = \boldsymbol{S}\left[\boldsymbol{W}'(\boldsymbol{I} - \boldsymbol{J}/n)\boldsymbol{W}\right]\boldsymbol{S}, \tag{5.4}$$

where $n = 45$, $\boldsymbol{I}$ is an identity matrix (45×45), $\boldsymbol{J}$ is a (45×45) matrix of ones, $\boldsymbol{W}$ is the (45×6) matrix of $BIOMASS$ (Y) and the five independent variables, and $\boldsymbol{S}$ is a diagonal matrix of the reciprocals of the square roots of the corrected sums of squares of each variable. The corrected sums of squares are given by the diagonal elements of $\boldsymbol{W}'(\boldsymbol{I} - \boldsymbol{J}/n)\boldsymbol{W}$. For the Linthurst data,

$$\widehat{\boldsymbol{\rho}} = \begin{array}{c} \begin{matrix} \boldsymbol{Y} & SAL & pH & K & Na & Zn \end{matrix} \\ \begin{bmatrix} 1 & -.103 & .774 & -.205 & -.272 & -.624 \\ -.103 & 1 & -.051 & -.021 & .162 & -.421 \\ .774 & -.051 & 1 & .019 & -.038 & -.722 \\ -.205 & -.021 & .019 & 1 & .792 & .074 \\ -.272 & .162 & -.038 & .792 & 1 & .117 \\ -.624 & -.421 & -.722 & .074 & .117 & 1 \end{bmatrix} \end{array}.$$

The first row of $\widehat{\boldsymbol{\rho}}$ contains the simple correlations of the dependent variable with each of the independent variables. The two variables pH and Zn have reasonably high correlations with $BIOMASS$. They would "account for" 60% $(r^2 = .774^2)$ and 39%, respectively, of the variation in $BIOMASS$ if

TABLE 5.2. *Results of the regression of BIOMASS on the five independent variables SALINITY, pH, K, Na, and Zn (Linthurst September data).*

Variable	$\widehat{\beta}_j$	$s(\widehat{\beta}_j)$	t	*Partial SS*
SAL	−30.285	24.031	−1.26	251,921
pH	305.525	87.879	3.48	1,917,306
K	−.2851	.3484	−.82	106,211
Na	−.0087	.0159	−.54	47,011
Zn	−20.676	15.055	−1.37	299,209

Analysis of variance for BIOMASS

Source	*d.f.*	*Sum of Squares*	*Mean Square*	
Total	44	19,170,963		
Regression	5	12,984,700	2,596,940	$F = 16.37$
Residual	39	6,186,263	158,622	

used separately as the only independent variable in the regressions. *Na* and *K* are about equally correlated with *BIOMASS* but at a much lower level than *pH* and *Zn*. There appears to be almost no correlation between *SALINITY* and *BIOMASS*.

There are two high correlations among the independent variables, *K* and *Na* with $r = .79$ and *pH* and *Zn* at $r = -.72$. The impact of these correlations on the regression results is noted as the analysis proceeds. With the exception of a moderate correlation between *SALINITY* and *Zn*, all other correlations are quite small.

5.2.2 Multiple Regression Results: Full Model

Summary of Results

The results of the multiple regression analysis using all five independent variables are summarized in Table 5.2. There is a strong relationship between *BIOMASS* and the independent variables. The coefficient of determination R^2 is .677. (See Table 1.5, page 15, for the definition of coefficient of determination.) Thus, 68% of the sums of squares in *BIOMASS* can be associated with the variation in these five independent variables. The test of the composite hypothesis that all five regression coefficients are zero is highly significant; $F = 16.37$ compared to $F_{(.01,5,39)} = 3.53$.

Computations

The computations for this analysis were done using a matrix algebra computer program [SAS/IML (SAS Institute Inc., 1989d)] operating on the $\boldsymbol{X}$ and $\boldsymbol{Y}$ matrices only. The steps in the language of SAS/IML and an explanation of each step is given in Table 5.3. The simplicity of matrix arithmetic can be appreciated only if one attempts to do the analysis with, say, a hand calculator.

TABLE 5.3. *The matrix algebra steps for the regression analysis as written for SAS/IML,*[a] *an interactive matrix programming language. It is assumed that* $\boldsymbol{Y}$ *and* $\boldsymbol{X}$ *have been properly defined in the matrix program and that* $\boldsymbol{X}$ *is of full rank.*

SAS/IML Program Step[a]	*Matrix Being Computed*
INVX=INV(X`*X);	$(\boldsymbol{X}'\boldsymbol{X})^{-1}$
$\boldsymbol{X}$` indicates transpose of $\boldsymbol{X}$	
B=INVX*X`*Y;	$\widehat{\boldsymbol{\beta}}$
CF=SUM(Y)##2/NROW(X);	$\boldsymbol{Y}'(\boldsymbol{J}/n)\boldsymbol{Y} = (\sum Y)^2/n$
The "##2" squares SUM(Y)	
SST=Y`*Y-CF;	$\boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{J}/n)\boldsymbol{Y}$
Corrected sum of squares for *BIOMASS*	
SSR=B`*X`*Y-CF;	$\boldsymbol{Y}'(\boldsymbol{P} - \boldsymbol{J}/n)\boldsymbol{Y} = \text{SS(Regr)}$
Notice that $\boldsymbol{P}$ need not be computed	
SSE=SST-SSR;	$\boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} = \text{SS(Res)}$
S2=SSE/(NROW(X)-NCOL(X));	s^2
The estimate of σ^2 with degrees of freedom $= n - r(\boldsymbol{X})$	
SEB=SQRT(VECDIAG(INVX)*S2);	Standard errors of $\widehat{\boldsymbol{\beta}}$
"VECDIAG" creates a vector from diagonal elements	
T=B/SEB;	t for $H_0 : \beta_j = 0$
"/" indicates elementwise division of B by SEB	
PART=B##2/VECDIAG(INVX);	Partial sums of squares
YHAT=X*B;	$\widehat{\boldsymbol{Y}}$, estimated means for $\boldsymbol{Y}$
E=Y-YHAT;	$\boldsymbol{e}$, estimated residuals

[a]Program steps for SAS/IML (1985a), an interactive matrix language program developed by SAS Institute, Inc., Cary, North Carolina.

Obtaining $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is the most difficult and requires the use of a computer for all but the simplest problems. Most of the other computations are relatively easy. Notice that the large 45×45 $\boldsymbol{P}$ matrix is not computed and generally is not needed in its entirety. The $\widehat{\boldsymbol{Y}}$ vector is more easily computed as $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, rather than as $\widehat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{Y}$. The only need for $\boldsymbol{P}$ is for $\mathbf{Var}(\widehat{\boldsymbol{Y}}) = \boldsymbol{P}\sigma^2$ and $\mathbf{Var}(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{P})\sigma^2$. Even then, the variance of an individual $\widehat{Y}_i$ or e_i of interest can be computed using only the ith row of $\boldsymbol{X}$, rather than the entire $\boldsymbol{X}$ matrix.

Residual Mean Square

The residual mean square, $s^2 = 158,622$ with 39 degrees of freedom, is an unbiased estimate of σ^2 if this five-variable model is the correct model. Of course, this is almost certainly not the correct model because (1) important variables may have been excluded, or (2) the mathematical form of the model may not be correct. (Including *unimportant* variables will not generally bias the estimate of σ^2.) Therefore, s^2 must be regarded as the tentative "best" estimate of σ^2 and is used for tests of significance and for computing the standard errors of the estimates.

Inconsistencies in the Results

The regression of *BIOMASS* on these five independent variables is highly significant. Yet, only one partial regression coefficient $\widehat{\beta}_2$ for *pH* is significantly different from zero, with $t = 3.48$. Recall that the simple correlation between *BIOMASS* and *pH* showed that *pH* alone would account for 60%, or 11.5 million, of the total corrected sum of squares for *BIOMASS*. When *pH* is used in a model with the other four variables, however, its partial sum of squares, 1,917,306, is only 10% of the total sum of squares and less than 15% of the regression sum of squares for all five variables. On the other hand, the partial sum of squares for *SALINITY* is larger than the simple correlation between *BIOMASS* and *SALINITY* would suggest.

These apparent inconsistencies are typical of regression results when the independent variables are not orthogonal. They are not inconsistencies if the meaning of the word "partial" in partial regression coefficients and partial sums of squares is remembered. "Partial" indicates that the regression coefficient or the sum of squares is the contribution of that particular independent variable *after* taking into account the effects of all other independent variables. Only when an independent variable is orthogonal to *all* other independent variables are its simple and partial regression coefficients and its simple and partial sums of squares equal.

5.3 Simplifying the Model

Removing Variables

The t-tests of the partial regression coefficients $H_0 : \beta_j = 0$ would seem to suggest that four of the five independent variables are unimportant and could be dropped from the model. The dependence of the partial regression coefficients and sums of squares on the other variables in the model, however, means that one must be cautious in removing more than one vari-

able at a time from the regression model. Removing one variable from the model will cause the regression coefficients and the partial sums of squares for the remaining variables to change (unless they are orthogonal to the variable dropped). These results do indicate that not all five independent variables are needed in the model. It would appear that any one of the four, *SALINITY*, *pH*, *Na*, or *K*, could be dropped without causing a significant decline in predictability of *BIOMASS*. It is not clear at this stage of the analysis, however, that more than one can be dropped.

There are several approaches for deciding which variables to include in the final model. These are studied in Chapter 7. For this example, one variable at a time is eliminated— the one whose elimination will cause the smallest increase in the residual sum of squares. The process will stop when the partial sums of squares for all variables remaining in the model are significant ($\alpha = .05$). As discussed in Chapter 7, data-driven variable selection and multiple testing to arrive at the final model alter the true significance levels; probability levels and confidence intervals should be used with caution.

A 4-Variable Model

The variable *Na* has the smallest partial sum of squares in the five-variable model. This means that *Na* is the least important of the five variables in accounting for the variability in *BIOMASS after* the contributions of the other four variables have been taken into account. As a result, *Na* is the logical variable to eliminate first. And, since the partial sum of squares for *Na*, $R(\beta_4 \mid \beta_1\ \beta_2\ \beta_3\ \beta_5\ \beta_0) = 47,011$ is not significant, there is no reason $X_4 = Na$ should not be eliminated.

Dropping *Na* means that $\boldsymbol{X}$ must be redefined to be the 45×5 matrix consisting of $\mathbf{1}$, $\boldsymbol{X}_1 = SALINITY$, $\boldsymbol{X}_2 = pH$, $\boldsymbol{X}_3 = K$, and $\boldsymbol{X}_5 = Zn$; the column vector of *Na* observations $\boldsymbol{X}_4$ is removed from $\boldsymbol{X}$. Similarly, $\boldsymbol{\beta}$ must be redefined by removing β_4. The regression analysis using these four variables (Table 5.4) shows the decrease in the regression sum of squares, now with four degrees of freedom, and the increase in the residual sum of squares to be exactly equal to the partial sum of squares for *Na* in the previous stage. This demonstrates the meaning of "partial sum of squares." In the absence of independent information on σ^2, the residual mean square from this reduced model is now used (tentatively) as the estimate of σ^2, $s^2 = 155,832$. (Notice that the increase in the residual sum of squares does not necessarily imply an increase in the residual mean square.)

A 3-Variable Model

The partial sums of squares at the four-variable stage (Table 5.4) show *SALINITY* and *Zn* to be equally unimportant to the model; the partial sum of squares for *Zn* is slightly smaller and both are nonsignificant. The next step in the search for the final model is to eliminate one of these two variables. Again, it is not safe to assume that both variables can be dropped since they are not orthogonal.

Since *Zn* has the slightly smaller partial sum of squares, *Zn* will be eliminated and *pH*, *SALINITY*, and *K* retained as the three-variable model. One could have used the much higher simple correlation between *Zn* and

TABLE 5.4. *Results of the regression of BIOMASS on the four independent variables SALINITY, pH, K, and Zn (Linthurst data).*

Variable	$\widehat{\beta}_j$	$\mathrm{s}(\widehat{\beta}_j)$	t	*Partial SS*
Sal	−35.94	21.48	−1.67	436,496
pH	293.9	84.5	3.48	1,885,805
K	−0.439	0.202	−2.17	732,606
Zn	−23.45	14.04	−1.67	434,796

Analysis of variance

Source	*d.f.*	*Sum of Squares*	*Mean Square*
Total	44	19,170,963	
Regression	4	12,937,689	3,234,422
Residual	40	6,233,274	155,832

TABLE 5.5. *Results of the regression of BIOMASS on the three independent variables SALINITY, pH, and K (Linthurst data).*

Variable	$\widehat{\beta}_j$	$\mathrm{s}(\widehat{\beta}_j)$	t	*Partial SS*
SAL	−12.06	16.37	−.74	88,239
pH	410.21	48.83	8.40	11,478,835
K	−.490	.204	−2.40	935,178

Analysis of variance

Source	*d.f.*	*Sum of Squares*	*Mean Square*
Total	44	19,170,963	
Regression	3	12,502,893	4,167,631
Residual	41	6,668,070	162,636

BIOMASS, $r = -.62$ versus $r = -.10$, to argue that *SALINITY* is the variable to eliminate at this stage. This is a somewhat arbitrary choice with the information at hand, and illustrates one of the problems of this sequential method of searching for the appropriate model. There is no assurance that choosing to eliminate *Zn* first will lead to the best model by whatever criterion is used to measure "goodness" of the model.

Again, $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are redefined, so that *Zn* is eliminated, and the computations repeated. This analysis gives the results in Table 5.5. The partial sum of squares for *pH* increases dramatically when *Zn* is dropped from the model, from 1.9 million to 11.5 million. This is due to the strong correlation between *pH* and *Zn* ($r = -.72$). When two independent variables are highly correlated, either positively or negatively, much of the predictive information contained in either can be usurped by the other. Thus, a

TABLE 5.6. *Results of the regression of BIOMASS on the two independent variables pH, and K (Linthurst data).*

Variable	$\widehat{\beta}_j$	$\mathrm{s}(\widehat{\beta}_j)$	t	*Partial SS*
pH	412.04	48.50	8.50	11,611,782
K	−0.487	0.203	−2.40	924,266

Analysis of variance

Source	*d.f.*	*Sum of Squares*	*Mean Square*
Total	44	19,170,963	
Regression	2	12,414,654	6,207,327
Residual	42	6,756,309	160,865

very important variable may appear as insignificant if the model contains a correlated variable and, conversely, an otherwise unimportant variable may take on false significance.

A 2-Variable Model

The contribution of *SALINITY* in the three-variable model is even smaller than it was before *Zn* was dropped and is far from being significant. The next step is to drop *SALINITY* from the model. In this particular example, one would not have been misled by eliminating both *SALINITY* and *Zn* at the previous step. This is not true in general.

The two-variable model containing pH and K gives the results in Table 5.6. Since the partial sums of squares for both pH and K are significant, the simplification of the model will stop with this two-variable model. The degree to which the linear model consisting of the two variables pH and K accounts for the variability in *BIOMASS* is $R^2 = .65$, only slightly smaller than the $R^2 = .68$ obtained with the original five-variable model.

5.4 Results of the Final Model

The Equation

This particular method of searching for an appropriate model led to the two-variable model consisting of pH and K. The regression equation is

$$\widehat{Y}_i = -507.0 + 412.0X_{i2} - 0.4871X_{i3} \tag{5.5}$$

or, expressed in terms of the centered variables,

$$\widehat{Y}_i = 1000.8 + 412.0(X_{i2} - 4.60) - .4871(X_{i3} - 797.62),$$

where $X_2 = pH$ and $X_3 = K$. This equation accounts for 65% of the variation in the observed values of aerial *BIOMASS*. That is, the predicted values computed from $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ account for 65% of the variation of *BIOMASS* or, conversely, the sum of squares of the residuals $\boldsymbol{e}'\boldsymbol{e}$ is 35% of the original

corrected sum of squares of *BIOMASS*. The square root of R^2 is the simple correlation between *BIOMASS* and $\widehat{\boldsymbol{Y}}$:

$$r(\boldsymbol{Y},\ \widehat{\boldsymbol{Y}}) \quad = \quad \sqrt{.65} = .80.$$

The estimate of σ^2 from this final model is $s^2 = 160,865$ with $(n - p') = 42$ degrees of freedom. The variance–covariance matrix for the regression coefficients is

$s^2(\widehat{\boldsymbol{\beta}})$

$$\begin{aligned}
\boldsymbol{s^2}(\widehat{\boldsymbol{\beta}}) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}s^2 \\
&= \begin{bmatrix} .4865711 & -.0663498 & -.0001993 \\ -.0663498 & .0146211 & -.0000012 \\ -.0001993 & -.0000012 & .00000026 \end{bmatrix} (160,865) \\
&= \begin{bmatrix} 78,272 & -10,673 & -32.0656 \\ -10,673 & 2,352.0 & -0.18950 \\ -32.0656 & -0.18950 & 0.04129 \end{bmatrix}.
\end{aligned}$$

The square roots of the diagonal elements give the standard errors of the estimated regression coefficients in the order in which they are listed in $\boldsymbol{\beta}$. In this model,

$$\boldsymbol{\beta} = (\,\beta_0 \quad \beta_2 \quad \beta_3\,)'.$$

Thus, the standard errors of the estimated regression coefficients are

$$\begin{aligned}
s(\widehat{\beta}_0) &= \sqrt{78,272} = 280 \\
s(\widehat{\beta}_2) &= \sqrt{2,352.0} = 48.5 \qquad (5.6) \\
s(\widehat{\beta}_3) &= \sqrt{.04129} = .2032.
\end{aligned}$$

The regression coefficients for pH and K are significantly different from zero as shown by the t-test (Table 5.6). The critical value of Student's t is $t_{(.05/2,42)} = 2.018$. (The intercept $\widehat{\beta}_0 = -507.0$ is not significantly different from zero, $t = -1.81$, and if one had reason to believe that β_0 should be zero the intercept could be dropped from the model.)

Univariate Confidence Intervals

The *univariate 95% confidence interval estimates* of the regression coefficients (Section 4.6.1),

$$\widehat{\beta}_j \quad \pm \quad t_{(.05/2,42)}s(\widehat{\beta}_j)$$

are

$$\begin{aligned}
-1,072 < \ \beta_0 \ &< 58 \\
314 < \ \beta_2 \ &< 510 \\
-.898 < \ \beta_3 \ &< -.077.
\end{aligned}$$

The value of Student's t for these intervals is $t_{(.05/2,42)} = 2.018$. The confidence coefficient of .95 applies to each interval statement.

Bonferroni Confidence Intervals

The **Bonferroni confidence intervals** (Section 4.6.2), using a joint confidence coefficient of .95, are

$$\begin{aligned} -1,206 < \ & \beta_0 \ < 192 \\ 291 < \ & \beta_2 \ < 533 \\ -.995 < \ & \beta_3 \ < .021. \end{aligned}$$

The joint confidence of $1 - \alpha$ is obtained by using the value of Student's t for $\alpha^* = \alpha/2p' : \ t_{(.05/(2\times3),42)} = 2.50$.

The Bonferroni intervals are necessarily wider than the univariate confidence intervals to allow for the fact that the confidence coefficient of .95 applies to the statement that all three intervals contain their true regression coefficients. In this example, the Bonferroni interval for β_3 overlaps zero whereas the univariate 95% confidence interval did not.

Joint Confidence Region

The *95% joint confidence region* for the three regression coefficients is determined from the quadratic inequality shown in equation 4.60 (Section 4.6.3). This three-dimensional 95% confidence ellipsoid is shown in Figure 5.1 for the Linthurst data. The outer box in Figure 5.1 is the Scheffé 95% confidence region. The inner box in the figure is the Bonferroni confidence region.

The ellipsoid in Figure 5.1 has been constructed using 19 cross-sectional planes in each of the three dimensions. The cross-sectional slices were chosen equally spaced and such that the most extreme in each direction coincided with a side of the Bonferroni box. These extreme slices and areas of the ellipsoid that extend beyond have been darkened to clearly show the portions of the joint confidence ellipsoid that extend beyond the Bonferroni box. Although the ellipsoid extends beyond the Bonferroni box in several areas, it is clear that the ellipsoid takes less volume of the parameter space to ensure 95% confidence in this example.

The sides of the Scheffé box (Figure 5.1) are tangent to the confidence ellipsoid and, consequently, the Scheffé box completely contains the ellipsoid. It can be shown in this particular example that the volume of the Bonferroni box is approximately 63% of the volume of the Scheffé box.

To more clearly show the shape of the joint confidence ellipsoid, the slices created by two sides of the Bonferroni box and the midplane in one dimension have been projected onto the floor in Figure 5.2. The slices show that the ellipsoid is very flattened in one dimension and clearly illustrate the strong interdependence among the regression coefficients as to what constitutes "acceptable" values of the parameters. Also inscribed on the floor is the two-dimensional 95% confidence ellipse calculated from the 2×2 variance–covariance matrix of $\widehat{\beta}_2$ and $\widehat{\beta}_3$ ignoring $\widehat{\beta}_0$. This shows that the two-dimensional confidence ellipse is *not* a projection of the three-dimensional confidence ellipsoid.

The general shape of the confidence region can be seen from the three-dimensional figure. However, it is very difficult to read the parameter values

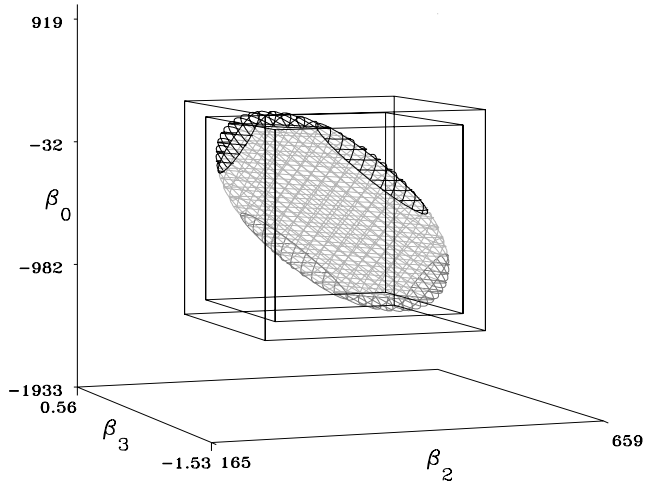

FIGURE 5.1. *Three-dimensional 95% joint confidence region (ellipsoid) for β_0, β_2, and β_3. The intersection of the Bonferroni confidence intervals (inner box) and the intersection of the Scheffé confidence intervals (outer box).*

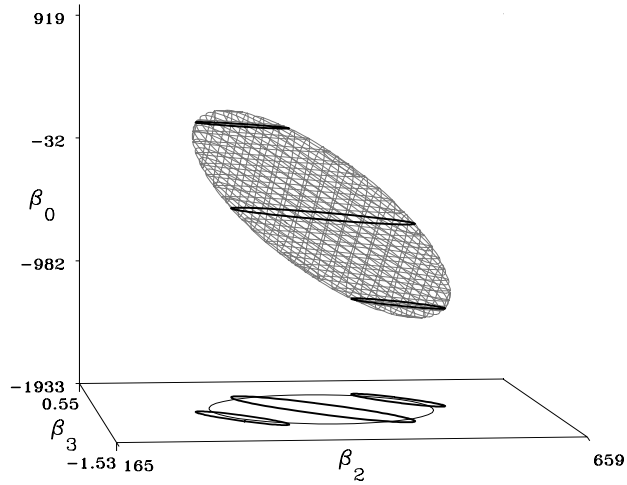

FIGURE 5.2. *Three-dimensional 95% joint confidence region for β_0, β_2, and β_3 showing projections of three 2-dimensional slices, corresponding to three values of β_0, onto the floor. The three values of β_0 chosen to define the slices were the midpoint and the limits of the 95% Bonferroni confidence interval for β_0.*

corresponding to any particular point in the figure. Furthermore, the joint confidence ellipsoid for more than three parameters cannot be pictured.

A more useful presentation of the joint confidence region is obtained by plotting two-dimensional "slices" through the ellipsoid for pairs of parameters of particular interest. This is done by evaluating the joint confidence equation at specific values of the other parameters. Three such two-dimensional ellipses for β_2 and β_3 are those shown in Figure 5.2. These slices help picture the three-dimensional ellipsoid but they are *not* to be interpreted individually as joint confidence regions for β_2 and β_3.

Alternatively, one can determine the two-dimensional 95% joint confidence region for β_2 and β_3 ignoring β_0. This region is also shown in Figure 5.2 as the larger ellipse on the floor of the figure. In this case, $\widehat{\beta}_2$ and $\widehat{\beta}_3$ are only slightly negatively correlated so that the two-dimensional joint confidence region is only slightly elliptical. The very elliptical slices from the original joint confidence region show that the choice of β_2 and β_3 for a given value of β_0 are more restricted than the two-dimensional joint confidence region would lead one to believe. This illustrates the information obscured by confidence intervals or regions that do not take into account the joint distribution of the full set of parameter estimates.

Two-dimensional slices through the joint confidence region in another direction, for given values of β_2, and the two-dimensional confidence region for β_0 and β_3 ignoring β_2 are shown in Figure 5.3. The strong negative correlation between $\widehat{\beta}_0$ and $\widehat{\beta}_3$ is evident in the two-dimensional joint confidence region and the slices from the three-dimensional region. Again, it is clear that reasonable combinations of β_0 and β_3 are dependent on the assumed value of β_2, a result that is not evident from the two-dimensional joint confidence region ignoring β_2.

$\widehat{Y}_1$ and $s^2(\widehat{Y}_1)$

$\widehat{\boldsymbol{Y}}$ and $\boldsymbol{e}$ for this example are not given. They are easily computed as shown in Table 5.2. Likewise, $\boldsymbol{s^2}(\widehat{\boldsymbol{Y}}) = \boldsymbol{P}s^2$ and $\boldsymbol{s^2}(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{P})s^2$ are not given; each is a 45×45 matrix. Computation of $\widehat{Y}_i$ and its variance is illustrated using the first data point. Each $\widehat{Y}_i$ is computed using the corresponding row vector from $\boldsymbol{X}$, which is designated $\boldsymbol{x}_i'$. For the first observation,

$$\boldsymbol{x}_1' = (1 \quad 5.00 \quad 1,441.67).$$

Thus,

$$\begin{aligned}\widehat{Y}_1 &= \boldsymbol{x}_1'\widehat{\boldsymbol{\beta}} \\ &= (1 \quad 5.00 \quad 1,441.67)\begin{pmatrix} -506.9774 \\ 412.0392 \\ -.4871 \end{pmatrix} = 850.99.\end{aligned}$$

The variance of $\widehat{Y}_1$, used as an estimate of the mean aerial *BIOMASS* at this specific level of *pH* (X_2) and *K* (X_3), is $s^2(\widehat{Y}_1) = v_{11}s^2$, where v_{11} is

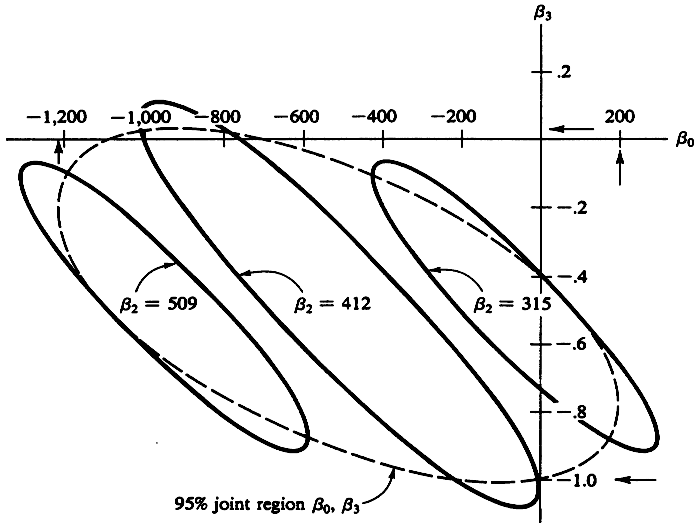

FIGURE 5.3. *Two-dimensional slices of the joint confidence region for three values of β_2 and the joint confidence region for β_0 and β_3 ignoring β_2 (shown in dashed line). The arrows indicate the limits of the intersection of the Bonferroni confidence intervals for β_0 and β_3.*

the first diagonal element from $\boldsymbol{P}$. The ith diagonal element of $\boldsymbol{P}$ can be obtained individually as $v_{ii} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$. Or, the variance for any one $\widehat{Y}_i$ is obtained as the variance of a linear function of $\widehat{\boldsymbol{\beta}}$. Thus,

$$\begin{aligned} s^2(\widehat{Y}_1) &= \boldsymbol{x}_1'[\boldsymbol{s^2}(\widehat{\boldsymbol{\beta}})]\boldsymbol{x}_1 \\ &= \begin{pmatrix} 1 & 5.00 & 1,441.67 \end{pmatrix} \begin{bmatrix} 78,272 & -10,673 & -32.0656 \\ -10,673 & 2,352.0 & -.18950 \\ -32.0656 & -.18950 & .04129 \end{bmatrix} \begin{pmatrix} 1 \\ 5.00 \\ 1,441.67 \end{pmatrix} \\ &= 20,978.78. \end{aligned}$$

Its standard error is

$$s(\widehat{Y}_1) = \sqrt{20,978.78} = 144.8.$$

If $\widehat{Y}_1$ is used as a prediction of a future observation Y_0 at the specified level $\boldsymbol{x}_1'$, then the variance of the prediction error is the variance of $\widehat{Y}_1$ increased by $s^2 = 160,865$. This accounts for the variability of the random variable being predicted. This gives

$$\begin{aligned} s^2(\widehat{Y}_{pred_1}) &= s^2(Y_0 - \widehat{Y}_1) \\ &= 20,979 + 160,865 = 181,843 \end{aligned}$$

or the standard error of prediction is

$$s(\widehat{Y}_{pred_1}) \quad = \quad \sqrt{181,843} = 426.4.$$

e_1 and $s^2(e_1)$

The residual for the first observation is

$$e_1 \quad = \quad Y_1 - \widehat{Y}_1 = 676 - 850.99 = -174.99.$$

The estimated variance of e_1 is

$$s^2(e_1) \quad = \quad (1 - v_{11})s^2.$$

Since $s^2(\widehat{Y}_1) = v_{11}s^2$ has already been computed, $s^2(e_1)$ is easily obtained as

$$\begin{aligned} s^2(e_1) &= s^2 - s^2(\widehat{Y}_1) \\ &= 160,865 - 20,979 = 139,886. \end{aligned}$$

The standard error is

$$s(e_1) \quad = \quad \sqrt{139,886} = 374.0.$$

These variances are used to compute confidence interval estimates for each of the corresponding parameters. Student's t has 42 degrees of freedom, the degrees of freedom in the estimate of σ^2. For illustration, the 95% confidence interval estimate of the mean *BIOMASS* production when $pH = 5.00$ and $K = 1,441.67$ ppm, $\mathcal{E}(Y_1)$, is

Confidence Intervals on $\mathcal{E}(\widehat{Y}_i)$

$$\widehat{Y}_1 \quad \pm \quad t_{(.05/2,42)}s(\widehat{Y}_1)$$

or

$$850.99 \quad \pm \quad (2.018)(144.8),$$

which becomes

$$558.7 \quad < \mathcal{E}(Y_1) < \quad 1,143.3.$$

These results indicate that, with 95% confidence, the true mean *BIOMASS* for $pH = 5.00$ and $K = 1,441.67$ is between 559 and 1,143 gm^{-2}.

Prediction Intervals for Y_0

If we wish to predict the *BIOMASS* production Y_0 at $\boldsymbol{x}_0 = \boldsymbol{x}_1$ ($pH = 5.00$ and $K = 1,441.67$), then a 95% prediction interval for Y_0 is given by

$$\widehat{Y}_1 \quad \pm \quad t_{(.025,42)}s(Y_0 - \widehat{Y}_1),$$

which gives

$$-9.60 < Y_0 < 1,711.5.$$

Since *BIOMASS* cannot be negative, this is usually reported as

$$0 < Y_0 < 1,711.5.$$

This example stops at this point. A complete analysis includes plots of regression results to verify that the regression equation gives a reasonable characterization of the observed data and that the residuals are behaving as they should. Such an extended analysis, however, would get into topics that are discussed in Chapters 7 and 9.

5.5 General Comments

The original objective of the Linthurst research was to identify important soil variables that were influencing the amount of *BIOMASS* production in the marshes. The wording of this objective implies that the desire is to establish causal links.

Cannot Infer Causality

Observational data *cannot* be used to establish causal relationships. Any analysis of observational data must build on the observed relationships, or the correlational structure, in the sample data. There are many reasons why correlations might exist in any set of data, only one of which is a causal pathway involving the variables. Some of the correlations observed will be fortuitous, accidents of the sampling of random variables. This is particularly likely if small numbers of observations are taken or if the sample points are not random. Some of the correlations will result from accidents of nature or from the variables being causally related to other unmeasured variables which, in turn, are causally related to the dependent variable. Even if the linkage between an independent and dependent variable is causal in origin, the direction of the causal pathway cannot be established from the observational data alone. The only way causality can be established is in controlled experiments where the causal variable is changed and the impact on the response variable observed.

Thus, it is incorrect in this case study to conclude that *pH* and *K* are important causal variables in *BIOMASS* production. The least squares analysis has established only that variation in *BIOMASS* is associated with variation in *pH* and *K*. The *reason* for the association is not established. Furthermore, there is no assurance that this analysis has identified all of the variables which show significant association with *BIOMASS*. The reasonably high correlation between *pH* and *Zn*, for example, has caused the regression analysis to eliminate *Zn* from the model; the partial sum of squares for *Zn* is nonsignificant after adjustment for *pH*. This sequential method of building the model may have eliminated an important causal variable.

Another common purpose of least squares is to develop prediction equations for the behavior of the dependent variable. Observational data are

Interpreting the Regression Equation

frequently the source of information for this purpose. Even here, care must be used in interpreting the results. The results from this case study predict that, on the average, *BIOMASS* production changes by 412 gm^{-2} for each unit change in *pH* and $-.5$ gm^{-2} for each ppm change in *K*. This prediction is appropriate for the population being sampled by this set of data, the marshes in the Cape Fear Estuary of North Carolina. It is not appropriate if the population has been changed by some event nor is it appropriate for points outside the population represented by the sample.

The regression coefficient for *pH* gives the expected change in *BIOMASS* per unit change in *pH*. This statement treats the other variables in the system two different ways, depending on whether they are included in the prediction equation. The predicted change in *BIOMASS* per unit change in *pH* ignores all variables *not included* in the final prediction equation. This means that any change in *pH*, for which a prediction is being made, will be accompanied by simultaneous changes in these ignored variables. The nature of these changes will be controlled by the correlational structure of the data. For example, *Zn* would be expected to decrease on the average as *pH* is increased due to the negative correlation between the two variables. Thus, this predicted change in *BIOMASS* is really associated with the *simultaneous* increase in *pH and* decrease in *Zn* . It is incorrect to think the prediction is for a situation where, somehow, *Zn* is not allowed to change.

On the other hand, the predicted change of 412 gm^{-2} *BIOMASS* associated with a unit change in *pH* assumes that the other variables included in the prediction equation, in this case *K*, are being held constant. Again, this is unrealistic when the variables in the regression equation are correlated.

The appropriate view of the regression equation obtained from observational data is as a description of the response surface of the dependent variable, where the independent variables in the equation are serving as surrogates for the many variables that have been omitted from the equation. The partial regression coefficients are the slopes of the response surface in the directions represented by the corresponding independent variables. Any attempt to ascribe these slopes, or changes, to the particular independent variables in the model implicitly assumes a causal relationship of the independent variable to the dependent variable and that all other variables in the system, for which the variables in the equation serve as surrogates, are unimportant in the process.

The response surface equation obtained from observational data can serve as a useful prediction equation as long as care is taken to ensure that the points for which predictions are to be made are valid points in the sampled population. This requires that the values of the independent variables for the prediction points must be in the sample space. It is easy, for example, when one variable at a time is being changed, to create prediction points that are outside the sample space. Predictions for these points can be very much in error.

5.6 Exercises

The data in the accompanying table are simulated data on peak rate of flow Q (cfs) of water from six watersheds following storm episodes. The storm episodes have been chosen from a larger data set to give a range of storm intensities. The independent variables are

$X_1 =$ Area of watershed (mi^2)

$X_2 =$ Area impervious to water (mi^2)

$X_3 =$ Average slope of watershed (percent)

$X_4 =$ Longest stream flow in watershed (thousands of feet)

$X_5 =$ Surface absorbency index, 0 = complete absorbency, 100 = no absorbency

$X_6 =$ Estimated soil storage capacity (inches of water)

$X_7 =$ Infiltration rate of water into soil (inches/hour)

$X_8 =$ Rainfall (inches)

$X_9 =$ Time period during which rainfall exceeded $\frac{1}{4}$ inch/hr.

Computations with this set of data will require a computer.

5.1. Compute the correlation matrix for all variables including the dependent variable Q. By inspection of the correlations determine which variables are most likely to contribute significantly to variation in Q. If you could use only one independent variable in your model, which would it be?

5.2. Compute the correlation matrix using $LQ = \log(Q)$ and the logarithms of all independent variables. How does this change the correlations and your conclusions about which variables are most likely to contribute significantly to variation in LQ?

5.3. Use $LQ = \ln(Q)$ as the dependent variable and the logarithm of all nine independent variables plus an intercept as the "full" model. Compute the least squares regression equation and test the composite null hypothesis that all partial regression coefficients for the independent variables are zero. Compare the estimated partial regression coefficients to their standard errors. Which partial regression coefficients are significantly different from zero? Which independent variable would you eliminate first to simplify the model?

5.4. Eliminate the least important variable from the model in Exercise 5.3 and recompute the regression. Are all partial sums of squares for the remaining variables significant ($\alpha = .05$)? If not, continue to eliminate the least important independent variable at each stage and recompute

Peak flow data from six watersheds.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Q
.03	.006	3.0	1	70	1.5	.25	1.75	2.0	46
.03	.006	3.0	1	70	1.5	.25	2.25	3.7	28
.03	.006	3.0	1	70	1.5	.25	4.00	4.2	54
.03	.021	3.0	1	80	1.0	.25	1.60	1.5	70
.03	.021	3.0	1	80	1.0	.25	3.10	4.0	47
.03	.021	3.0	1	80	1.0	.25	3.60	2.4	112
.13	.005	6.5	2	65	2.0	.35	1.25	.7	398
.13	.005	6.5	2	65	2.0	.35	2.30	3.5	98
.13	.005	6.5	2	65	2.0	.35	4.25	4.0	191
.13	.008	6.5	2	68	.5	.15	1.45	2.0	171
.13	.008	6.5	2	68	.5	.15	2.60	4.0	150
.13	.008	6.5	2	68	.5	.15	3.90	3.0	331
1.00	.023	15.0	10	60	1.0	.20	.75	1.0	772
1.00	.023	15.0	10	60	1.0	.20	1.75	1.5	1,268
1.00	.023	15.0	10	60	1.0	.20	3.25	4.0	849
1.00	.023	15.0	10	65	2.0	.20	1.80	1.0	2,294
1.00	.023	15.0	10	65	2.0	.20	3.10	2.0	1,984
1.00	.023	15.0	10	65	2.0	.20	4.75	6.0	900
3.00	.039	7.0	15	67	.5	.50	1.75	2.0	2,181
3.00	.039	7.0	15	67	.5	.50	3.25	4.0	2,484
3.00	.039	7.0	15	67	.5	.50	5.00	6.5	2,450
5.00	.109	6.0	15	62	1.5	.60	1.50	1.5	1,794
5.00	.109	6.0	15	62	1.5	.60	2.75	3.0	2,067
5.00	.109	6.0	15	62	1.5	.60	4.20	5.0	2,586
7.00	.055	6.5	19	56	2.0	.50	1.80	2.0	2,410
7.00	.055	6.5	19	56	2.0	.50	3.25	4.0	1,808
7.00	.055	6.5	19	56	2.0	.50	5.25	6.0	3,024
7.00	.063	6.5	19	56	1.0	.50	1.25	2.0	710
7.00	.063	6.5	19	56	1.0	.50	2.90	3.4	3,181
7.00	.063	6.5	19	56	1.0	.50	4.76	5.0	4,279

the regression. Stop when all independent variables in the model are significant (use $\alpha = .05$). What do the results indicate about the need for the intercept? Does it make sense to have $\beta_0 = 0$ in this exercise? Summarize the results of your final model in an analysis of variance table. Discuss in words your conclusions about what factors are important in peak flow rates.

5.5. Determine the 95% univariate confidence interval estimates of the regression coefficients for your final model. Determine the 95% Bonferroni confidence interval estimates. Determine also the 95% Scheffé confidence interval estimates.

5.6. Construct the 95% joint confidence region for the partial regression coefficients for X_8 and X_9 *ignoring* the parameters for the other variables in your final model in Exercise 5.4.

6
GEOMETRY OF LEAST SQUARES

Matrix notation has been used to present least squares regression and the application of least squares has been demonstrated. This chapter presents the geometry of least squares. The data vectors are represented by vectors plotted in n-space and the basic concepts of least squares are illustrated using relationships among the vectors. The intent of this chapter is to give insight into the basic principles of least squares. This chapter is not essential for an understanding of the remaining topics.

All concepts of ordinary least squares can be visualized by applying a few principles of geometry. Many find the geometric interpretation more helpful than the cumbersome algebraic equations in understanding the concepts of least squares. Partial regression coefficients, sums of squares, degrees of freedom, and most of the properties and problems of ordinary least squares have direct *visual* analogues in the geometry of vectors.

This chapter is presented solely to enhance your understanding. Although the first exposure to the geometric interpretation may seem somewhat confusing, the geometry usually enhances understanding of the least squares concepts. You are encouraged to study this chapter in the spirit in which it is presented. It is not an essential chapter for the use and understanding of regression. Review of Section 2.4 before reading this chapter may prove helpful.

6.1 Linear Model and Solution

X-Space

In the geometric interpretation of least squares, $\boldsymbol{X}$ is viewed as a collection of p' column vectors. It is assumed for this discussion that the column vectors of $\boldsymbol{X}$ are linearly independent (any linear dependencies that might have existed in $\boldsymbol{X}$ have been eliminated). Each column vector of $\boldsymbol{X}$ can be plotted as a vector in n-dimensional space (see Section 2.4). That is, the n elements in each column vector provide the coordinates for identifying the endpoint of the vector plotted in n-space. The p' vectors *jointly* define a p'-dimensional subspace of the n-dimensional space in which they are plotted ($p' < n$). This p'-dimensional subspace consists of the set of points that can be reached by *linear* functions of the p' vectors of $\boldsymbol{X}$. This subspace is called the **$\boldsymbol{X}$-space**. (When the vectors of $\boldsymbol{X}$ are not linearly independent, the dimensionality of the X-space is determined by the rank of $\boldsymbol{X}$.)

$\mathcal{E}(\boldsymbol{Y})$ Vector

The $\boldsymbol{Y}$ vector is also a vector in n-dimensional space. Its expectation

$$\mathcal{E}(\boldsymbol{Y}) \quad = \quad \boldsymbol{X}\boldsymbol{\beta} = \beta_0\mathbf{1} + \beta_1\boldsymbol{X}_1 + \cdots + \beta_p\boldsymbol{X}_p \tag{6.1}$$

is a linear function of the column vectors of $\boldsymbol{X}$ with the elements of $\boldsymbol{\beta}$ being the coefficients. Thus, the linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{6.2}$$

says that the mean vector $\mathcal{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ falls *exactly* in the X-space. The specific point at which $\mathcal{E}(\boldsymbol{Y})$ falls is determined by the true, and unknown, partial regression coefficients in $\boldsymbol{\beta}$.

$\boldsymbol{Y}$ Vector

The vector of observations on the dependent variable $\boldsymbol{Y}$will fall somewhere in n-dimensional space around its mean $\mathcal{E}(\boldsymbol{Y})$, with its exact position being determined by the random elements in $\boldsymbol{\epsilon}$. The model (equation 6.2) states that $\boldsymbol{Y}$ is the sum of the two vectors $\mathcal{E}(\boldsymbol{Y})$ and $\boldsymbol{\epsilon}$. Although $\mathcal{E}(\boldsymbol{Y})$ is in the X-space, $\boldsymbol{\epsilon}$ and, consequently, $\boldsymbol{Y}$ are random vectors in n-dimensional space. Neither $\boldsymbol{\epsilon}$ nor $\boldsymbol{Y}$ will fall in the X-space (unless an extremely unlikely sample has been drawn).

Example 6.1

To illustrate these relationships, we must limit ourselves to three-dimensional space. The concepts illustrated in two and three dimensions extend to n-dimensional geometry. Assume that $\boldsymbol{X}$ consists of two vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, each of order 3, so that they can be plotted in three-dimensional space (Figure 6.1). The plane in Figure 6.1 represents the two-dimensional subspace defined by $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. The vector $\mathcal{E}(\boldsymbol{Y})$ lies in this plane and represents the true mean vector of $\boldsymbol{Y}$, as the linear function of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ expressed in the model. The dashed lines in Figure 6.1 show the addition of the vectors $\beta_1\boldsymbol{X}_1$ and $\beta_2\boldsymbol{X}_2$ to give the vector $\mathcal{E}(\boldsymbol{Y})$. This, of course, assumes that the model is correct. In practice, $\mathcal{E}(\boldsymbol{Y})$ is not known because $\boldsymbol{\beta}$ is not known. One of the purposes of the regression analysis is to find "best" estimates of β_1 and β_2. ■

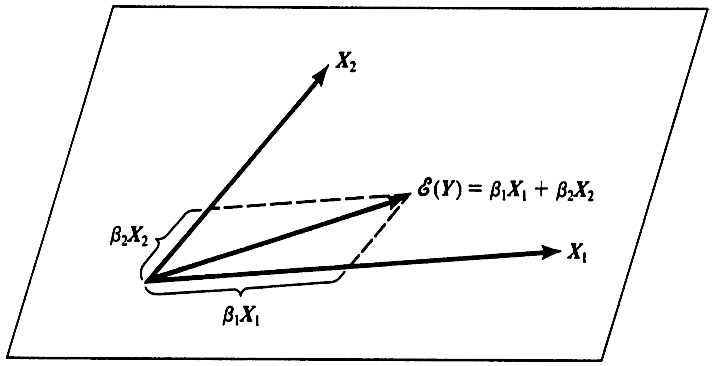

FIGURE 6.1. *The geometric interpretation of $\mathcal{E}(\boldsymbol{Y})$ as a linear function of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. The plane represents the space defined by the two independent vectors. The vector $\mathcal{E}(\boldsymbol{Y})$ is shown as the sum of $\beta_1\boldsymbol{X}_1$ and $\beta_2\boldsymbol{X}_2$.*

The Partial Regression Coefficients

The position of $\mathcal{E}(\boldsymbol{Y})$ in Figure 6.1 represents a case where both β_1 and β_2 are positive; the vectors to be added to give $\mathcal{E}(\boldsymbol{Y})$, $\beta_1\boldsymbol{X}_1$ and $\beta_2\boldsymbol{X}_2$, have the same direction as the original vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. When $\mathcal{E}(\boldsymbol{Y})$ falls outside the angle formed by $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, one or both of the regression coefficients must be negative. Multiplication of a vector by a negative coefficient reverses the direction of the vector. For example, $-.1\boldsymbol{X}_1$ defines a vector that is $\frac{1}{10}$ the length of $\boldsymbol{X}_1$ and has opposite direction to $\boldsymbol{X}_1$. Figure 6.2 partitions the two-dimensional X-space according to the signs β_1 and β_2 take when $\mathcal{E}(\boldsymbol{Y})$ falls in the particular region. Figure 6.3 uses the same X-space and $\mathcal{E}(\boldsymbol{Y})$ as Figure 6.1 but includes $\boldsymbol{Y}$, at some distance from $\mathcal{E}(\boldsymbol{Y})$ and *not* in the X-space (because of $\boldsymbol{\epsilon}$), and $\widehat{\boldsymbol{Y}}$. Since $\widehat{\boldsymbol{Y}}$ is a linear function of the columns of $\boldsymbol{X}$, $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, it must fall in the X-space. The *estimated* regression coefficients $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are shown as the multiples of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ that give $\widehat{\boldsymbol{Y}}$ when summed. The estimated regression coefficients serve the same role in determining $\widehat{\boldsymbol{Y}}$ that the true regression coefficients β_1 and β_2 do in determining $\mathcal{E}(\boldsymbol{Y})$. Of course, $\widehat{\boldsymbol{Y}}$ will almost certainly never coincide with $\mathcal{E}(\boldsymbol{Y})$. Figure 6.3 is drawn so that both $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are positive. The signs of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are determined by the region of the X-space in which $\widehat{\boldsymbol{Y}}$ falls, as illustrated in Figure 6.2 for β_1 and β_2.

The e Vector

The short vector connecting $\widehat{\boldsymbol{Y}}$ to $\boldsymbol{Y}$ in Figure 6.3 is the vector of residuals $\boldsymbol{e}$. The least squares principle requires that $\widehat{\boldsymbol{\beta}}$, and hence $\widehat{\boldsymbol{Y}}$, be chosen such that $\sum(Y_i - \widehat{Y}_i)^2 = \boldsymbol{e}'\boldsymbol{e}$ is minimized. But $\boldsymbol{e}'\boldsymbol{e}$ is the squared length of $\boldsymbol{e}$. Geometrically, it is the squared distance from the end of the $\boldsymbol{Y}$ vector to the end of the $\widehat{\boldsymbol{Y}}$ vector. Thus, $\widehat{\boldsymbol{Y}}$ must be that unique vector in the X-space that is closest to $\boldsymbol{Y}$ in n-space. The closest point on the plane to $\boldsymbol{Y}$ (in Figure 6.3) is the point that would be reached with a perpendicular

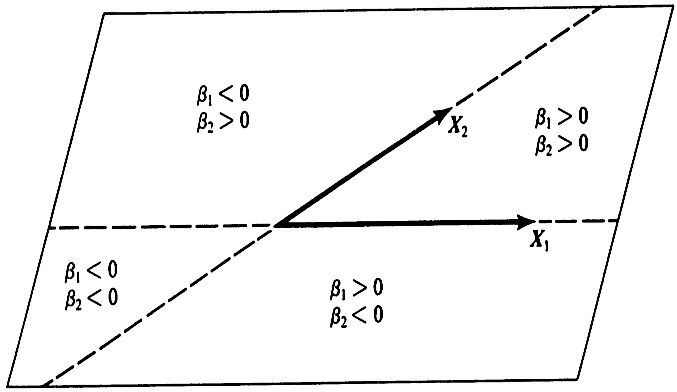

FIGURE 6.2. *Partitions of the two-dimensional X-space according to the signs* β_1 *and* β_2 *take when* $\mathcal{E}(\boldsymbol{Y})$ *falls in the indicated region.*

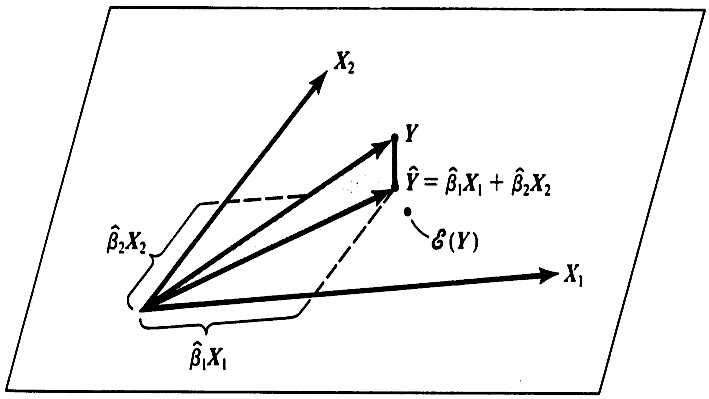

FIGURE 6.3. *The geometric relationship of* $\boldsymbol{Y}$ *and* $\widehat{\boldsymbol{Y}}$ *to the X-space.* $\boldsymbol{Y}$ *is not in the plane defined by* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$*. The perpendicular projection from* $\boldsymbol{Y}$ *to the plane defines the vector* $\widehat{\boldsymbol{Y}}$*, which is in the plane. The estimated regression coefficients are the proportions of* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$ *that, when added, give* $\widehat{\boldsymbol{Y}}$*. The short vector connecting* $\widehat{\boldsymbol{Y}}$ *to* $\boldsymbol{Y}$ *is the vector of residuals* $\boldsymbol{e}$*.*

projection from $\boldsymbol{Y}$ to the plane. That is, $\boldsymbol{e}$ must be perpendicular to the X-space. $\widehat{\boldsymbol{Y}}$ is shown as the shadow on the plane cast by $\boldsymbol{Y}$ with a light directly "overhead."

Visualize the floor of a room being the plane defined by the X-space. Let one corner of the room at the floor be the origin of the three-dimensional coordinate system, the line running along one baseboard be the $\boldsymbol{X}_1$ vector, and the line running along the adjoining baseboard be the $\boldsymbol{X}_2$ vector. Thus, the floor of the room is the X-space. Let the $\boldsymbol{Y}$ vector run from the origin to some point in the ceiling. It is obvious that the point on the floor *closest* to this point in the ceiling is the point directly beneath. That is, the "projection" of $\boldsymbol{Y}$ onto the X-space must be a perpendicular projection onto the floor. A line from the end of $\boldsymbol{Y}$ to $\widehat{\boldsymbol{Y}}$ must form a right angle with the floor. This "vertical" line from $\boldsymbol{Y}$ to $\widehat{\boldsymbol{Y}}$ is the vector of observed residuals $\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ (plotted at $\widehat{\boldsymbol{Y}}$ instead of at the origin). The two vectors $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{e}$ clearly add to $\boldsymbol{Y}$.

Common sense told us that $\boldsymbol{e}$ must be perpendicular to the plane for $\widehat{\boldsymbol{Y}}$ to be the closest possible vector to $\boldsymbol{Y}$. The least squares procedure requires this to be the case. Note that,

$$\begin{aligned} \boldsymbol{X}'\boldsymbol{e} &= \boldsymbol{X}'(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}) = \boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ &= \boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ &= \boldsymbol{0}, \end{aligned} \tag{6.3}$$

since we know that from the normal equations

$$\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}.$$

The statement $\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0}$ shows that $\boldsymbol{e}$ must be orthogonal (or perpendicular) to each of the column vectors in $\boldsymbol{X}$. (The sum of products of the elements of $\boldsymbol{e}$ with those of each vector in $\boldsymbol{X}$ is zero.) Hence, $\boldsymbol{e}$ must be perpendicular to any linear function of these vectors in order for the result to be a least squares result.

$\boldsymbol{P}$ Matrix

$\widehat{\boldsymbol{Y}}$ may also be written as $\widehat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{Y}$. The matrix $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the matrix that *projects* $\boldsymbol{Y}$ onto the p'-dimensional subspace defined by the columns of $\boldsymbol{X}$. In other words, premultiplying $\boldsymbol{Y}$ by $\boldsymbol{P}$ gives $\widehat{\boldsymbol{Y}}$ such that the vector $\boldsymbol{e}$ is perpendicular to the X-space and as short as possible. $\boldsymbol{P}$ is called a **projection matrix**; hence its label $\boldsymbol{P}$.

Example 6.2

Consider the model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{X} = (1 \quad 1)'$ and $\boldsymbol{\beta}$ is a scalar. In this case, the X-space is one-dimensional and given by the straight line $Z_2 = Z_1$, where Z_1 and Z_2 represent the coordinates of a two-dimensional plane. The $\mathcal{E}(\boldsymbol{Y})$ vector is given by $(\beta \quad \beta)'$, which is a point on the straight line $Z_2 = Z_1$. Suppose we

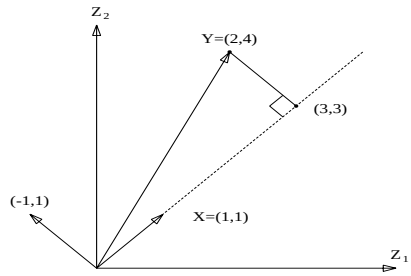

FIGURE 6.4. *Geometric interpretation of the regression in Example 6.2.*

observe $\boldsymbol{Y}$ to be $\boldsymbol{Y} = (\,2 \quad 4\,)'$. This is a vector in the two-dimensional plane (Figure 6.4). Since $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = (2)^{-1}6 = 3$, we have $\widehat{\boldsymbol{Y}} = (\,3 \quad 3\,)'$. Note that $\widehat{\boldsymbol{Y}}$ is a point (vector) in the X-space that is the closest to the observed vector $\boldsymbol{Y}$. The line that connects $\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}$ is perpendicular to (orthogonal to) the straight line given by $Z_2 - Z_1 = 0$ which is the X-space. The residual vector is given by $\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\,-1 \quad 1\,)'$. It is easy to verify that $\boldsymbol{X}'\boldsymbol{e} = (\,1 \quad 1\,)(\,-1 \quad 1\,)' = 0$. ■

The results of this section are summarized as follows.

1. $\boldsymbol{Y}$ is a vector in n-space.

2. Each column vector of $\boldsymbol{X}$ is a vector in n-space.

3. The p' linearly independent vectors of $\boldsymbol{X}$ define a p'-dimensional subspace.

4. The linear model specifies that $\mathcal{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ is in the X-space; the vector $\boldsymbol{Y}$ is (almost certainly) not in the X-space.

5. The least squares solution $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{P}\boldsymbol{Y}$ is that point in the X-space that is closest to $\boldsymbol{Y}$.

6. The residuals vector $\boldsymbol{e}$ is orthogonal to the X-space.

7. The right triangle formed by $\boldsymbol{Y}$, $\widehat{\boldsymbol{Y}}$, and $\boldsymbol{e}$ expresses $\boldsymbol{Y}$ as the sum of the other two vectors, $\boldsymbol{Y} = \widehat{\boldsymbol{Y}} + \boldsymbol{e}$.

6.2 Sums of Squares and Degrees of Freedom

Length of $\boldsymbol{Y}$

The Pythagorean theorem in two-dimensional space states that the length of the hypotenuse of a right triangle is the square root of the sum of the squares of the sides of the triangle. In Section 2.4 it was explained that this extends into n dimensions—the length of any vector is the square root of the sum of the squares of *all* its elements. Thus, $\boldsymbol{Y}'\boldsymbol{Y}$, the uncorrected sum of squares of the dependent variable, is the squared length of the vector $\boldsymbol{Y}$.

Partitioning the Total Sum of Squares

The vectors $\boldsymbol{Y}$, $\widehat{\boldsymbol{Y}}$, and $\boldsymbol{e}$ form a right triangle with $\boldsymbol{Y}$ being the hypotenuse (Figure 6.3). One side of the triangle $\widehat{\boldsymbol{Y}}$ lies in the X-space; the other side $\boldsymbol{e}$ is perpendicular to the X-space. The Pythagorean theorem can be used to express the length of $\boldsymbol{Y}$ in terms of the lengths of $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{e}$:

$$\text{length}(\boldsymbol{Y}) = \sqrt{[\text{length}(\widehat{\boldsymbol{Y}})]^2 + [\text{length}(\boldsymbol{e})]^2}.$$

Squaring both sides yields

$$\boldsymbol{Y}'\boldsymbol{Y} = \widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}} + \boldsymbol{e}'\boldsymbol{e}. \tag{6.4}$$

Thus, the partitioning of the total sum of squares of $\boldsymbol{Y}'\boldsymbol{Y}$ into SS(Model) = $\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}}$ and SS(Res) = $\boldsymbol{e}'\boldsymbol{e}$ corresponds to expressing the squared length of the vector $\boldsymbol{Y}$ in terms of the squared lengths of the sides of the right triangle.

Example 6.3

In Example 6.2, note that

$$\begin{aligned}
\boldsymbol{Y}'\boldsymbol{Y} &= \begin{pmatrix} 2 & 4 \end{pmatrix}\begin{pmatrix} 2 \\ 4 \end{pmatrix} = 20 \\
\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}} &= \begin{pmatrix} 3 & 3 \end{pmatrix}\begin{pmatrix} 3 \\ 3 \end{pmatrix} = 18 \\
\boldsymbol{e}'\boldsymbol{e} &= \begin{pmatrix} -1 & 1 \end{pmatrix}\begin{pmatrix} -1 \\ 1 \end{pmatrix} = 2
\end{aligned}$$

and hence equation 6.4 is satisfied. ■

SS(Res)

The "room" analogy given in Figure 6.3 can be used to show another property of least squares regression. The regression of $\boldsymbol{Y}$ on one independent variable, say $\boldsymbol{X}_1$, cannot give a smaller residual sum of squares $\boldsymbol{e}'\boldsymbol{e}$ than the regression on $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ jointly. The X-space defined by $\boldsymbol{X}_1$

alone is the set of points along the baseboard representing $\boldsymbol{X}_1$. Therefore, the projection of $\boldsymbol{Y}$ onto the space defined only by $\boldsymbol{X}_1$ (as if $\boldsymbol{X}_1$ were the only variable in the regression) must be to a point along this baseboard. The subspace defined by $\boldsymbol{X}_1$ alone is part of the subspace defined jointly by $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Therefore, no point along this baseboard can be closer to the end of the $\boldsymbol{Y}$ vector than the closest point on the entire floor (the X-space defined by $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ jointly). The two vectors of residuals, that from the regression of $\boldsymbol{Y}$ on $\boldsymbol{X}_1$ alone and that from the regression of $\boldsymbol{Y}$ on $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ jointly, would be the same length only if the projection onto the floor happened to fall exactly at the baseboard. In this case, $\widehat{\beta}_2$ must be zero. This illustrates a general result that the residual sum of squares from the regression of $\boldsymbol{Y}$ on a subset of independent variables *cannot* be smaller than the residual sum of squares from the regression on the full set of independent variables.

Degrees of Freedom

The "degrees of freedom" associated with each sum of squares is the number of dimensions in which that vector is "free to move." $\boldsymbol{Y}$ is free to fall anywhere in n-dimensional space and, hence, has n degrees of freedom. $\widehat{\boldsymbol{Y}}$, on the other hand, must fall in the X-space and, hence, has degrees of freedom equal to the dimension of the X-space—two in Figure 6.3 or p' in general. The residual vector $\boldsymbol{e}$ can fall anywhere in the subspace of the n-dimensional space that is *orthogonal* to the X-space. This subspace has dimensionality $(n - p')$ and, hence, $\boldsymbol{e}$ has $(n - p')$ degrees of freedom. In Figure 6.3, $\boldsymbol{e}$ has $(3 - 2) = 1$ degree of freedom. In general, the degrees of freedom associated with $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{e}$ will be $r(\boldsymbol{X})$ and $[n - r(\boldsymbol{X})]$, respectively.

Figures 6.1 through 6.3 have been described as if all vectors were of order 3 so that they could be fully represented in the three-dimensional figures. This is being more restrictive than needed. Three vectors of *any* order define a three-dimensional subspace and, if one forgoes plotting the individual vectors in n-space, the relationships among the three vectors can be illustrated in three dimensions as in Figures 6.1 through 6.3.

Example 6.4

This example uses the data from Exercise 1.4, which relate heart rate at rest to kilograms of body weight. The model to be fit includes an intercept so that the two vectors defining the X-space are $\mathbf{1}$, the vector of ones, and $\boldsymbol{X}_1$, the vector of body weights. The $\boldsymbol{Y}$ and $\boldsymbol{X}_1$ vectors in the original data are an order of magnitude longer than $\mathbf{1}$, so that both $\boldsymbol{Y}$ and $\boldsymbol{X}_1$ have been scaled by $\frac{1}{20}$ for purposes of this illustration. The rescaled data are

$$\begin{aligned} \boldsymbol{X}_1' &= (\,4.50 \quad 4.30 \quad 3.35 \quad 4.45 \quad 4.05 \quad 3.75\,) \\ \boldsymbol{Y}' &= (\,3.10 \quad 2.25 \quad 2.00 \quad 2.75 \quad 3.20 \quad 2.65\,). \end{aligned}$$

The X-space is defined by $\mathbf{1}$ and $\boldsymbol{X}_1$. The lengths of the vectors are

$$\begin{aligned} \text{length}(\mathbf{1}) &= \sqrt{\mathbf{1}'\mathbf{1}} = \sqrt{6} = 2.45 \\ \text{length}(\boldsymbol{X}_1) &= \sqrt{\boldsymbol{X}_1'\boldsymbol{X}_1} = \sqrt{100.23} = 10.01 \end{aligned}$$

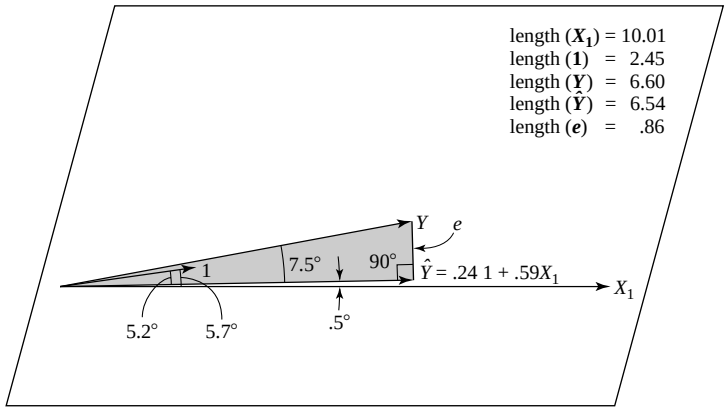

FIGURE 6.5. *Geometric interpretation of the regression of heart rate at rest* (Y) *on kilograms body weight* (X_1). *The plane in the figure is the* X*-space defined by* $\mathbf{1}$ *and* $\boldsymbol{X}_1$. *The data are from Exercise 1.4 with both* $\boldsymbol{X}_1$ *and* $\boldsymbol{Y}$ *scaled by* $\frac{1}{20}$. *Angles between vectors are shown in degrees.* $\boldsymbol{Y}$ *protrudes away from the plane at an angle of* $7.5°$. *Perpendicular projection of* $\boldsymbol{Y}$ *onto the plane defines* $\widehat{\boldsymbol{Y}}$ *which forms an angle of* $5.2°$ *with* $\mathbf{1}$ *and* $.5°$ *with* $\boldsymbol{X}_1$.

and the angle between the two vectors $\theta(\mathbf{1},\ \boldsymbol{X}_1)$ is

$$\begin{aligned}\theta(\mathbf{1}, \boldsymbol{X}_1) &= \arccos\left(\frac{\mathbf{1}'\boldsymbol{X}_1}{\sqrt{\mathbf{1}'\mathbf{1}}\sqrt{\boldsymbol{X}_1'\boldsymbol{X}_1}}\right) \\ &= \arccos\left(\frac{24.4}{\sqrt{6}\sqrt{100.23}}\right) = 5.7°.\end{aligned}$$

The vectors $\mathbf{1}$ and $\boldsymbol{X}_1$ are plotted in Figure 6.5 using their relative lengths and the angle between them. The X-space defined by $\mathbf{1}$ and $\boldsymbol{X}_1$ is the plane represented by the parallelogram.

The $\boldsymbol{Y}$ vector is drawn as protruding above the surface of the plane at an angle of $\theta(\boldsymbol{Y},\ \widehat{\boldsymbol{Y}}) = 7.5°$, the angle between $\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}$. [All angles are computed as illustrated for $\theta(\mathbf{1},\ \boldsymbol{X}_1)$.] The length of $\boldsymbol{Y}$ is

$$\text{length}(\boldsymbol{Y}) = \sqrt{\boldsymbol{Y}'\boldsymbol{Y}} = \sqrt{43.4975} = 6.60.$$

This is the square root of the uncorrected sum of squares of Y which, since $\boldsymbol{Y}$ can fall anywhere in six-dimensional space, has six degrees of freedom. The projection of $\boldsymbol{Y}$ onto the plane defines $\widehat{\boldsymbol{Y}}$ as the sum

$$\widehat{\boldsymbol{Y}} = (.24)\mathbf{1} + (.59)\boldsymbol{X}_1.$$

The angles between $\widehat{\boldsymbol{Y}}$ and the two X-vectors are

$$\theta(\widehat{\boldsymbol{Y}},\ \mathbf{1}) = 5.2°$$

and

$$\theta(\widehat{\boldsymbol{Y}},\ \boldsymbol{X}_1) = .5^\circ.$$

The length of $\widehat{\boldsymbol{Y}}$ is the square root of SS(Model):

$$\text{length}(\widehat{\boldsymbol{Y}}) = \sqrt{\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}}} = \sqrt{42.7552} = 6.539.$$

Since $\widehat{\boldsymbol{Y}}$ must fall in the two-dimensional X-space, SS(Model) has two degrees of freedom. The residuals vector $\boldsymbol{e}$ connecting $\widehat{\boldsymbol{Y}}$ to $\boldsymbol{Y}$ is perpendicular to the plane and its length is the square root of SS(Res):

$$\text{length}(\boldsymbol{e}) = \sqrt{\boldsymbol{e}'\boldsymbol{e}} = \sqrt{.7423} = .862.$$

Since $\boldsymbol{e}$ must be orthogonal to the X-space, SS(Res) has four degrees of freedom. Thus, the squared lengths of $\boldsymbol{Y}$, $\widehat{\boldsymbol{Y}}$, and $\boldsymbol{e}$ and the dimensions in which each is free to move reflect the analysis of variance of the regression results.

In this example, $\widehat{\boldsymbol{Y}}$ falls very close to $\boldsymbol{X}_1$; the angle between the two vectors is only .5°. This suggests that very nearly the same predictability of $\boldsymbol{Y}$ would be obtained from the regression of $\boldsymbol{Y}$ on $\boldsymbol{X}_1$ alone—that is, if the model forced the regression line to pass through the origin. If the no-intercept model is adopted, the X-space becomes the one-dimensional space defined by $\boldsymbol{X}_1$. The projection of $\boldsymbol{Y}$ onto this X-space gives $\widehat{\boldsymbol{Y}} = .65\boldsymbol{X}_1$. That is, $\widehat{\boldsymbol{Y}}$ falls *on* $\boldsymbol{X}_1$. The length of $\widehat{\boldsymbol{Y}}$ is

$$\text{length}(\widehat{\boldsymbol{Y}}) = \sqrt{42.7518} = 6.538,$$

which is trivially shorter than that obtained with the intercept model, $\sqrt{42.7518}$ versus $\sqrt{42.7552}$. The residuals vector is, correspondingly, only slightly longer:

$$\text{length}(\boldsymbol{e}) = \sqrt{.7457} = .864.$$

■

6.3 Reparameterization

Consider the model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{6.5}$$

Let $\boldsymbol{C}$ be a $p' \times p'$ nonsingular matrix. Then, we can rewrite the model shown in equation 6.5 also as

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{C}\boldsymbol{C}^{-1}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \boldsymbol{W}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \end{aligned} \tag{6.6}$$

where $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{C}$ and $\boldsymbol{\alpha} = \boldsymbol{C}^{-1}\boldsymbol{\beta}$. Here the model in equation 6.6 is a *reparameterization* of the model in equation 6.5. Note that since $\boldsymbol{C}$ is nonsingular, the W-space, the p'-dimensional subspace spanned by the p' columns of $\boldsymbol{W}$, is the same as the X-space. Recall that $\widehat{\boldsymbol{Y}}$ is the point in the X-space that is closest to $\boldsymbol{Y}$ and is given by $\widehat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{Y} = \boldsymbol{P}_X\boldsymbol{Y}$, where the X subscript identifies the projection matrix based on $\boldsymbol{X}$, $\boldsymbol{P}_X = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Since the W-space is the same as the X-space, we have $\widehat{\boldsymbol{Y}} = \boldsymbol{P}_W\boldsymbol{Y}$ and $\boldsymbol{P}_W = \boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}' = \boldsymbol{P}_X$. See Exercise 6.9.

In Chapter 8, we consider orthogonal polynomial models that are reparameterizations of polynomial models. We show that they are also reparameterizations of analysis of variance models. Also, the models where the input variables are centered are reparameterizations of corresponding uncentered models.

Example 6.5

Consider the model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Then, the X-space consists of all points of the form $(\, z_1 \quad z_2 \quad 0 \,)'$. In terms of the "room" analogy considered in Figure 6.3, the X-space consists of the floor. Suppose we observe the $\boldsymbol{Y}$ vector to be $\boldsymbol{Y} = (\, 2 \quad 4 \quad 3 \,)'$. Then,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = (\, 2 \quad 4 \,)'$$

and

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\, 2 \quad 4 \quad 0 \,)'.$$

Figure 6.6 shows the vector $\boldsymbol{Y} = (\, 2 \quad 4 \quad 3 \,)'$ and its projection $\widehat{\boldsymbol{Y}} = (\, 2 \quad 4 \quad 0 \,)'$ in a plane that forms the "floor" of the plot. We can think of this "floor" as the plane spanned by the vectors $\boldsymbol{X}_1 = (\, 1 \quad 0 \quad 0 \,)'$ and $\boldsymbol{X}_2 = (\, 0 \quad 1 \quad 0 \,)'$. Around the origin, on the floor of the plot, we have placed for reference circles of radii 1 and 4. The vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, each of unit length, are shown with the end of each vector touching the unit circle. The vectors are also extended to $2\boldsymbol{X}_1 = (\, 2 \quad 0 \quad 0 \,)'$ and $4\boldsymbol{X}_2 = (\, 0 \quad 4 \quad 0 \,)'$. Their sum $2\boldsymbol{X}_1 + 4\boldsymbol{X}_2 = (\, 2 \quad 4 \quad 0 \,)'$ is shown as $\widehat{\boldsymbol{Y}}$, the projection of $\boldsymbol{Y}$ onto the two-dimensional X-space.

The plane represented by the "floor" of the plot is also spanned by the two vectors $\boldsymbol{W}_1 = (\, 1 \quad 1 \quad 0 \,)'$ and $\boldsymbol{W}_2 = (\, 1 \quad 2 \quad 0 \,)'$. Thus, the floor of the plot is the set of all linear combinations of $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ (as well as all linear combinations of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$). Note that the linear combination

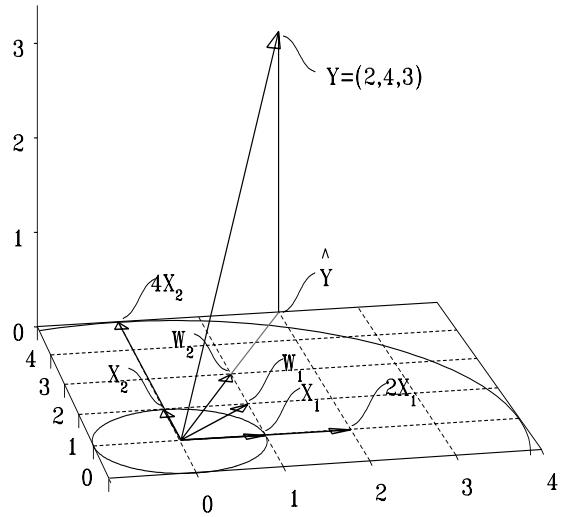

FIGURE 6.6. *Projection of* $\boldsymbol{Y}$ *onto the two-dimensional space spanned by* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$. $\widehat{\boldsymbol{Y}}$ *is equal to the sum of* $2\boldsymbol{X}_1$ *and* $4\boldsymbol{X}_2$. *Any two other vectors in the floor of the plot, say* $\boldsymbol{W}_1$ *and* $\boldsymbol{W}_2$, *will be linear combinations of* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$ *and will define the same space.* $\widehat{\boldsymbol{Y}}$ *is also obtained as a linear function of* $\boldsymbol{W}_1$ *and* $\boldsymbol{W}_2$; $\widehat{\boldsymbol{Y}} = 0\boldsymbol{W}_1 + 2\boldsymbol{W}_2$.

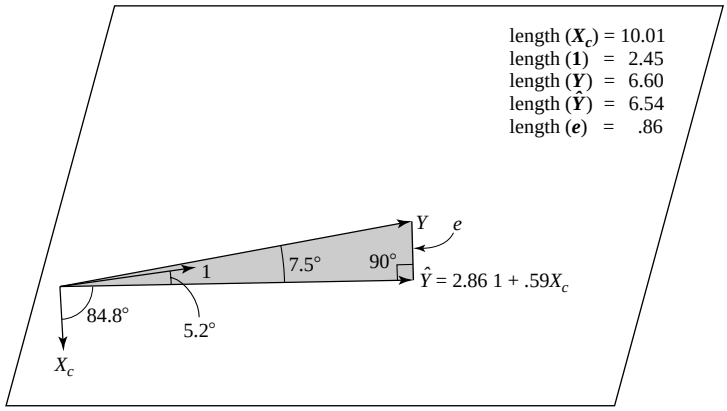

FIGURE 6.7. *Geometric interpretation of the regression of heart rate at rest (Y) on kilograms body weight using the centered variable (X_c). The plane in the figure is defined by* $\mathbf{1}$ *and* $\boldsymbol{X}_c$ *and is identical to the plane defined by* $\mathbf{1}$ *and* $\boldsymbol{X}_1$ *in Figure 6.5. All vectors are the same as in Figure 6.5 except* $\boldsymbol{X}_c$ *replaces* $\boldsymbol{X}_1$.

$0\boldsymbol{W}_1 + 2\boldsymbol{W}_2$ also gives $\widehat{\boldsymbol{Y}}$; the vector $\boldsymbol{W}_2$ is extended to $\widehat{Y}$ to illustrate this. Mathematically, all points on the floor of the plot are of the form

$$(\,a \quad b \quad 0\,)' = a\boldsymbol{X}_1 + b\boldsymbol{X}_2 = (2a - b)\boldsymbol{W}_1 + (b - a)\boldsymbol{W}_2$$

showing explicitly how any point $(\,a \quad b \quad 0\,)$ in the floor can be expressed as a linear combination of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ or of $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$. In this example $a = 2$ and $b = 4$. ■

Centered Independent Variables

It is common in least squares regression to express the model in terms of *centered* independent variables. That is, each independent variable is coded to have zero mean by subtracting the mean of the variable from each observation. The only effect, geometrically, of centering the independent variable is to shift the position, in the original X-space, of the vector representing the independent variable so that it is orthogonal to the vector $\mathbf{1}$. In general, when more than one independent variable is involved, each centered variable will be orthogonal to $\mathbf{1}$. The centering will change the angles between the vectors of the independent variables but the X-space remains as defined by the original variables. That is, the model with the centered independent variable is a reparameterization of the original model. See Exercise 6.11.

Example 6.6

The geometric interpretation of the effect of centering the independent variable is illustrated in Figure 6.7 for the heart rate/body weight data from Example 6.2. Let $\boldsymbol{X}_c$ be the centered vector. $\boldsymbol{X}_c$ is obtained by the

subtraction

$$\boldsymbol{X}_c = \boldsymbol{X}_1 - (4.0667)\,\mathbf{1},$$

where 4.0667 is the mean of the elements in $\boldsymbol{X}_1$. Since $\boldsymbol{X}_c$ is a linear function of $\mathbf{1}$ and $\boldsymbol{X}_1$, it is by definition in the space defined by $\mathbf{1}$ and $\boldsymbol{X}_1$. Thus, the X-space defined by $\mathbf{1}$ and $\boldsymbol{X}_c$ in Figure 6.7 is identical to the X-space defined by $\mathbf{1}$ and $\boldsymbol{X}_1$ in Figure 6.5. Centering the independent variable does not alter the definition of the X-space. The centered vector $\boldsymbol{X}_c$ is orthogonal to $\mathbf{1}$, because $\mathbf{1}'\boldsymbol{X}_c = 0$, and has length 1.002. $\boldsymbol{Y}$ is the same as in Figure 6.5 and, because the X-space is the same, the projection of $\boldsymbol{Y}$ onto the X-space must give the same $\widehat{\boldsymbol{Y}}$. The regression equation, however, is now expressed in terms of a linear function of $\mathbf{1}$ and $\boldsymbol{X}_c$ rather than in terms of $\mathbf{1}$ and $\boldsymbol{X}_1$. ■

6.4 Sequential Regressions

Correction for the Mean

Equation 6.4 gave the partitioning of the total *uncorrected* sum of squares for Y. Interest is usually in partitioning the total *corrected* sum of squares. The partitioning of the corrected sum of squares is obtained by subtracting the sum of squares attributable to the mean, or the correction factor, from both $\boldsymbol{Y}'\boldsymbol{Y}$ and SS(Model):

$$\begin{aligned} \boldsymbol{Y}'\boldsymbol{Y} - \mathrm{SS}(\mu) &= [\mathrm{SS}(\mathrm{Model}) - \mathrm{SS}(\mu)] + \boldsymbol{e}'\boldsymbol{e} \\ &= \mathrm{SS}(\mathrm{Regr}) + \boldsymbol{e}'\boldsymbol{e}. \end{aligned} \tag{6.7}$$

The correction for the mean SS(μ) is the sum of squares attributable to a model that contains only the constant term β_0. Geometrically, this is equivalent to projecting $\boldsymbol{Y}$ onto the one-dimensional space defined by $\mathbf{1}$. The least squares estimate of β_0 is $\overline{Y}$, and the residuals vector from this projection is the vector of deviations of Y_i from $\overline{Y}$, $y_i = Y_i - \overline{Y}$. The squared length of this residuals vector is the *corrected* sum of squares for Y. Since the space defined by $\mathbf{1}$ is a one-dimensional space, this residuals vector lies in $(n-1)$-dimensional space and has $(n-1)$ degrees of freedom.

SS(Regr) and the partial regression coefficients are the results obtained when this residuals vector is, in turn, projected onto the p-dimensional subspace ($p = p' - 1$) defined by the independent variables where each independent variable has also been "corrected for" its mean. Thus, obtaining SS(Regr) can be viewed as a two-stage process. First, $\boldsymbol{Y}$ and the independent variables are each projected onto the space defined by $\mathbf{1}$. Then, the *residuals* vector for $\boldsymbol{Y}$ is projected onto the space defined by the *residuals* vectors for the independent variables. The squared length of $\widehat{\boldsymbol{Y}}$ for this second projection is SS(Regr).

Sequential Sums of Squares

The **sequential sum of squares** for an independent variable is an extension of this process. Now, however, $\boldsymbol{Y}$ and the independent variable of current interest are first projected onto the space defined by *all* independent variables that precede the current X in the model, not just $\mathbf{1}$. Then, the *residuals* vector for $\boldsymbol{Y}$ (call it $\boldsymbol{e}_y$) is projected onto the space defined by the *residuals* vector for the current X (call it $\boldsymbol{e}_x$). The sequential sum of squares for the current independent variable is the squared length of $\widehat{\boldsymbol{Y}}$ for this projection of $\boldsymbol{e}_y$ onto $\boldsymbol{e}_x$. Note that both the dependent variable and the current independent variable have been "adjusted" for all preceding independent variables. At each step in the sequential analysis, the new X-space is a one-dimensional space and, therefore, the sequential sum of squares at each stage has one degree of freedom.

Since the residuals vector in least squares is always orthogonal to the X-space onto which $\boldsymbol{Y}$ is projected, $\boldsymbol{e}_y$ and $\boldsymbol{e}_x$ are both orthogonal to *all* independent variables previously included in the model. Because of this orthogonality to the previous X-space, the sequential sums of squares and degrees of freedom are additive. That is, the sum of the sequential sums of squares and the sum of the degrees of freedom for each step are equal to what would have been obtained if a single model containing all independent variables had been used.

6.5 The Collinearity Problem

Definition of Collinearity

The partial regression coefficient and partial sum of squares for any independent variable are, in general, dependent on which other independent variables are in the model. In the case study in Chapter 5, it was observed that the changes in regression coefficients and sums of squares as other variables were added to or removed from the model could be large. This dependence of the regression results for each variable on what other variables are in the model derives from the independent variables *not* being mutually orthogonal. Lack of orthogonality of the independent variables is to be expected in observational studies, those in which the researcher is restricted to making observations on nature as it exists. In such studies, the researcher

> ... cannot impose on a subject, or withhold from the subject, a procedure or treatment whose effects he desires to discover, or cannot assign subjects at random to different procedures. (Cochran, 1983).

On the other hand, controlled experiments are usually designed to avoid the collinearity problems.

The extreme case of nonorthogonality, where two or more independent variables are very nearly linearly dependent, creates severe problems in least

squares regression. This is referred to as the **collinearity problem**. The regression coefficients become extremely unstable; they are very sensitive to small random errors in $\boldsymbol{Y}$ and may fluctuate wildly as independent variables are added to or removed from the model. The instability in the regression results is reflected in very large standard errors for the partial regression coefficients. Frequently, none of the individual partial regression coefficients will be significantly different from zero even though their combined effect is highly significant.

Geometry of Collinearity

The impact of collinearity is illustrated geometrically in Figure 6.8. Consider the model and a reparameterization of the model given by

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{bmatrix} X_{11} & X_{21} \\ X_{12} & X_{22} \\ X_{13} & X_{23} \end{bmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}$$
$$= \begin{bmatrix} W_{11} & W_{21} \\ W_{12} & W_{22} \\ W_{13} & W_{23} \end{bmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}.$$

Suppose that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are orthogonal to each other, whereas $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are not orthogonal. $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ represent two vectors that show some degree of collinearity. The X-space and W-space are the same since one is a reparameterization of the other. This two-dimensional space is shown as the "floor" in the three-dimensional figure, panel (a), and as the plane in panels (b) and (c) of Figure 6.8. The central 95% of the population of all possible $\boldsymbol{Y}$-vectors is represented in the three-dimensional figure, panel (a), as the shaded sphere.

Recall that $\widehat{\boldsymbol{Y}}$ is the projection of $\boldsymbol{Y}$ onto the "floor" (= X-space = W-space). The circular area on the "floor" encloses the collection of all projections $\widehat{\boldsymbol{Y}}$ of the points $\boldsymbol{Y}$ in the sphere. Two possible projections $\widehat{\boldsymbol{Y}}_1$ and $\widehat{\boldsymbol{Y}}_2$ (on opposing edges of the circle), representing two independent $\boldsymbol{Y}$, are used to illustrate the relative sensitivity of the partial regression coefficients to variation in $\boldsymbol{Y}$ in the collinear case compared to the orthogonal case. It is assumed that the linear model $\mathcal{E}(\boldsymbol{Y})$ is known and that the input variables are fixed and measured without error. Thus, only the effect of variation in $\boldsymbol{Y}$, different samples of $\boldsymbol{\epsilon}$, is being illustrated by the difference between $\widehat{\boldsymbol{Y}}_1$ and $\widehat{\boldsymbol{Y}}_2$ in Figure 6.8.

The partial regression coefficients are the multipliers that get attached to each of the vectors so that the vector addition gives $\widehat{\boldsymbol{Y}}$. The vector addition is illustrated in Figure 6.8 by completion of the parallelogram for each $\widehat{\boldsymbol{Y}}$. This is most easily seen in panel (b) for the orthogonal vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ and in panel (c) for the nonorthogonal vectors $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$. The point to note is that the change in γ_1 and γ_2, the partial regression coefficients for the nonorthogonal system [panel (c)], as one shifts from $\widehat{\boldsymbol{Y}}_1$ to $\widehat{\boldsymbol{Y}}_2$ is much greater than the corresponding change in β_1 and β_2, the partial regression coefficients for the orthogonal system [panel (b)]. This illustrates

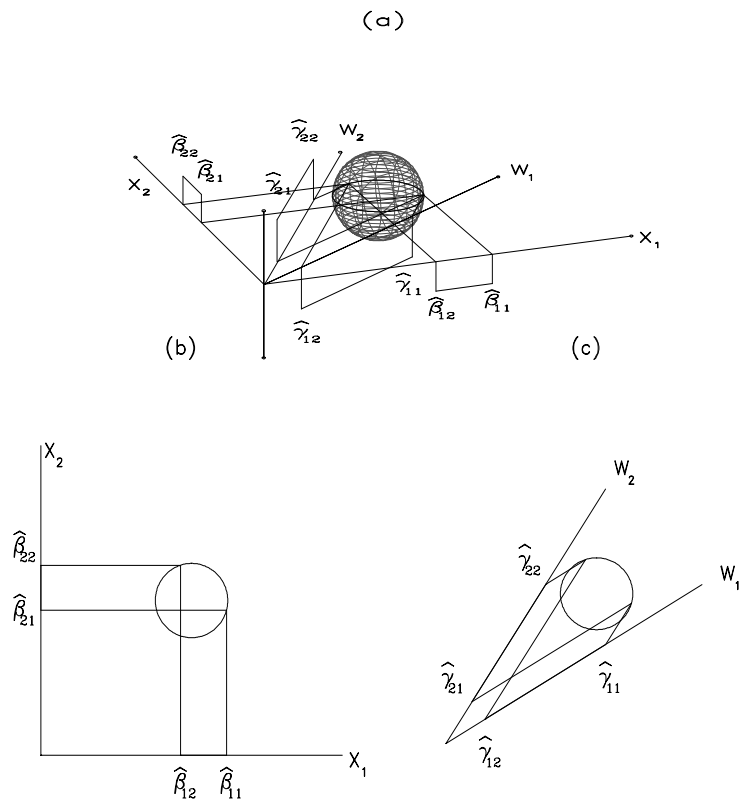

FIGURE 6.8. *Illustration of the effect of collinearity on the stability of the partial regression coefficients. The points in the shaded sphere centered on the plane [panel (a)] represent 95% of a population of three-dimensional vectors* $\boldsymbol{Y}$. $\mathcal{E}(\boldsymbol{Y})$ *is at the center of the sphere and at the center of the circle of projections of all* $\boldsymbol{Y}$ *onto the two-dimensional plane spanned by either the two orthogonal vectors* $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ *or the two nonorthogonal (somewhat collinear) vectors* $\boldsymbol{W}_1$ *and* $\boldsymbol{W}_2$. *Points shown on opposite sides of the circle represent* $\widehat{Y}_1$ *and* $\widehat{Y}_2$, *projections from two independent* $\boldsymbol{Y}$. *The parallelograms connecting each* $\widehat{\boldsymbol{Y}}$ *to the two sets of vectors show the relative magnitudes of the partial regression coefficients for the pair of orthogonal vectors [panel (b)] and the pair of nonorthogonal vectors [panel (c)].*

the greater sensitivity of the partial regression coefficients in the presence of collinearity; comparable changes in $\boldsymbol{Y}$ cause larger changes in the partial regression coefficients when the vectors are not orthogonal. As the two vectors approach collinearity (the angle between the vectors approaches 0° or 180°), the sensitivity of the regression coefficients to random changes in $\boldsymbol{Y}$ increases dramatically. In the limit, when the angle is 0° or 180°, the two vectors are linearly dependent and no longer define a two-dimensional subspace. In such cases, it is not possible to estimate β_1 and β_2 separately; only the joint effect of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ on $\boldsymbol{Y}$ is estimable.

Variation in the X-Vectors

Figure 6.8 illustrates the relative impact of variation in $\boldsymbol{\epsilon}$ on the partial regression coefficients in the orthogonal and nonorthogonal cases. In most cases, and particularly when the data are observational, the X-vectors are also subject to random variation in the population being sampled. Consequently, even if the independent variables are measured without error, repeated samples of the population will yield different X-vectors. Measurement error on the independent variables adds another component of variation to the X-vectors. Geometrically, this means that the X-space defined by the observed Xs, the plane in Figure 6.8, will vary from sample to sample; the amount of variation in the plane will depend on the amount of sampling variation and measurement error in the independent variables.

The impact of sampling variation and measurement error in the independent variables is magnified with increasing collinearity of the X-vectors. Imagine balancing a cardboard (the plane) on two pencils (the vectors). If the pencils are at right angles, the plane is relatively insensitive to small movements in the tips of the pencils. On the other hand, if the pencils form a very small angle with each other (the vectors are nearly collinear), the plane becomes very unstable and its orientation changes drastically as the pencils are shifted even slightly. In the limit as the angle goes to 0° (the two vectors are linearly dependent), the pencils merge into one and in one direction all support for the plane disappears.

In summary, collinearity causes the partial regression coefficients to be sensitive to small changes in $\boldsymbol{Y}$; the solution to the normal equations becomes unstable. In addition, sampling variation and measurement error in the independent variables causes the X-space to be poorly defined, which magnifies the sensitivity of the partial regression coefficients to collinearity. The instability in the least squares solution due to variation in $\boldsymbol{Y}$ is reflected in larger standard errors on the partial regression coefficients. The instability due to sampling variation in the independent variables, however, is ignored in the usual regression analysis because the independent variables are assumed to be fixed constants.

6.6 Summary

The following regression results are obtained from the geometric interpretation of least squares.

1. The data vectors $\boldsymbol{Y}$ and $\boldsymbol{X}_j$ are vectors in n-dimensional space.
2. The linear model states that the true mean of $\boldsymbol{Y}$, $\mathcal{E}(\boldsymbol{Y})$, is in the X-space, a p'-dimensional subspace of the n-dimensional space.
3. $\widehat{\boldsymbol{Y}}$ is the point in the X-space closest to $\boldsymbol{Y}$; $\boldsymbol{e}$ is orthogonal to the X-space.
4. The partial regression coefficients multiplied by their respective X-vectors define the set of vectors that must be added to "reach" $\widehat{\boldsymbol{Y}}$.
5. The vectors $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{e}$ are the two sides of a right triangle whose hypotenuse is $\boldsymbol{Y}$. Thus, $\boldsymbol{Y} = \widehat{\boldsymbol{Y}} + \boldsymbol{e}$.
6. The squared lengths of the sides of the right triangle give the partitioning of the sums of squares of $\boldsymbol{Y}$: $\boldsymbol{Y}'\boldsymbol{Y} = \widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}} + \boldsymbol{e}'\boldsymbol{e}$.
7. The correlation structure among the Xs influences the regression results. In general, $\widehat{\beta}_1 \neq \widehat{\beta}_{1.2}$. However if $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are orthogonal, then $\widehat{\beta}_1 = \widehat{\beta}_{1.2}$.
8. Regression of $\boldsymbol{Y}$ on one independent variable, say $\boldsymbol{X}_1$, cannot give smaller $\boldsymbol{e}'\boldsymbol{e}$ than regression on $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ jointly. More generally, regression on a subset of independent variables cannot give a better fit (smaller $\boldsymbol{e}'\boldsymbol{e}$) than regression on all variables.
9. If $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are nearly collinear, small variations in $\boldsymbol{Y}$ cause large shifts in the partial regression coefficients. The regression results become unstable.

6.7 Exercises

6.1. Use Figure 6.3 as plotted to approximate the values of $\widehat{\beta}_1$and $\widehat{\beta}_2$. Where would $\widehat{\boldsymbol{Y}}$ have to have fallen for $\widehat{\beta}_1$ to be negative? For $\widehat{\beta}_2$ to be negative? For both to be negative?

6.2. Construct a figure similar to Figure 6.3 except draw the projection of $\boldsymbol{Y}$ onto the space defined by $\boldsymbol{X}_1$. Similarly, draw the projection of $\boldsymbol{Y}$ onto the space defined by $\boldsymbol{X}_2$.

(a) Approximate the values of the simple regression coefficients in each case and compare them to the partial regression coefficients in Figure 6.3.

(b) Identify the residuals vector in both cases and in Figure 6.3.

(c) Convince yourself that the shortest residuals vector is the one in Figure 6.3.

6.3. Construct a diagram similar to Figure 6.3 except make $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ orthogonal to each other. Convince yourself that, when the independent variables are orthogonal, the simple regression coefficients from the projection of $\boldsymbol{Y}$ onto $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ separately equal the partial regression coefficients from the projection of $\boldsymbol{Y}$ onto the space defined by $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ jointly.

6.4. Assume we have two vectors of order 10, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Jointly these two vectors define a plane, a 2-dimensional subspace of the original 10-dimensional space. Let $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ be an arbitrary coordinate system for this 2-dimensional subspace. Represent the vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ *in this plane* by the coordinates of the two vectors $\boldsymbol{Z}_1 = (\,5 \quad 2\,)'$ and $\boldsymbol{Z}_2 = (\,0 \quad -4\,)'$. Suppose the projection of $\boldsymbol{Y}$ onto this plane plots at $(\,-1 \quad 3\,)'$ in this coordinate system.

(a) Use your figure to approximate the regression coefficients for the regression of $\boldsymbol{Y}$ on $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

(b) From your figure compute the sum of squares due to the regression of $\boldsymbol{Y}$ on $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ jointly. How many degrees of freedom does this sum of squares have?

(c) Do you have enough information to compute the residual sum of squares? How many degrees of freedom would the residual sum of squares have?

(d) Suppose someone told you that the original vector $\boldsymbol{Y}$ had length 3. Would there be any reason to doubt their statement?

6.5. Plot the two vectors $\boldsymbol{X}_1 = (\,5 \quad 0\,)'$ and $\boldsymbol{X}_2 = (\,-4 \quad .25\,)'$. Suppose two different samples of $\boldsymbol{Y}$ give projections onto this X-space at $\widehat{\boldsymbol{Y}}_1 = (\,4 \quad .5\,)'$ and $\widehat{\boldsymbol{Y}}_2 = (\,4 \quad -.5\,)'$. Approximate from the graph the partial regression coefficients for the two cases. Note the shift in the partial regression coefficients for the two cases. Compare this shift to what would have been realized if $\boldsymbol{X}_2 = (\,0 \quad 4\,)'$, orthogonal to $\boldsymbol{X}_1$.

6.6. Data from Exercise 1.9 relating plant biomass $\boldsymbol{Y}$ to total accumulated solar radiation $\boldsymbol{X}$ was used to fit a no-intercept model. $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{e}$ were determined from the regression equation. The matrix $\boldsymbol{W}$ (8×4) was defined as

$$\boldsymbol{W} = [\boldsymbol{X}\ \boldsymbol{Y}\ \widehat{\boldsymbol{Y}}\ \boldsymbol{e}]$$

and the following matrix of sums of squares and products was computed.

$$\boldsymbol{W}'\boldsymbol{W} = \begin{bmatrix} 1,039,943.1 & 1,255,267.1 & 1,255,267.1 & 0 \\ 1,255,267.1 & 1,523,628.9 & 1,515,174.7 & 8,454.2 \\ 1,255,267.1 & 1,515,174.7 & 1,515,174.7 & 0 \\ 0 & 8,454.2 & 0 & 8,454.2 \end{bmatrix}.$$

(a) Determine the length of each (column) vector in $\boldsymbol{W}$.

(b) Compute the angles between all pairs of vectors.

(c) Use the lengths of the vectors and the angles between the vectors to show graphically the regression results. What is the dimension of the X-space? Why is the angle between $\boldsymbol{X}$ and $\widehat{\boldsymbol{Y}}$ zero? Estimate the regression coefficient from the figure you construct.

6.7. This exercise uses the data given in Exercise 1.19 relating seed weight of soybeans $\boldsymbol{Y}$ to cumulative seasonal solar radiation $\boldsymbol{X}$ for two levels of ozone exposure. For simplicity in plotting, rescale $\boldsymbol{X}$ by dividing by 2 and $\boldsymbol{Y}$ by dividing by 100 for this exercise.

(a) Use the "Low Ozone" data to compute the linear regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$ (with an intercept). Compute $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{e}$, the lengths of all vectors, and the angle between each pair of vectors. Use the vector lengths and angles to display graphically the regression results (similar to Figure 6.5). Use your figure to "estimate" the regression coefficients. From the relative positions of the vectors, what is your judgment as to whether the intercept is needed in the model?

(b) Repeat Part (a) using the "High Ozone" data.

(c) Compare the graphical representations of the two regressions. What is your judgment as to whether the regressions are homogeneous—that is, are the same basic relationships—within the limits of random error, illustrated in both figures.

6.8. The angle θ between the intercept vector $\mathbf{1}$ and an independent variable vector $\boldsymbol{X}$ depends on the coefficient of variation of the independent variable. Use the relationship

$$\cos(\theta) = \frac{\mathbf{1}'\boldsymbol{X}}{\sqrt{\mathbf{1}'\mathbf{1}}\sqrt{\boldsymbol{X}'\boldsymbol{X}}}$$

to show the relationship to the coefficient of variation. What does this relationship imply about the effect on the angle of scaling the $\boldsymbol{X}$ by a constant? What does it imply about the effect of adding a constant to or subtracting a constant from $\boldsymbol{X}$?

6.9. Consider the reparameterization

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$$

of the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{C}$ and $\boldsymbol{C}$ is nonsingular.

(a) Show that W-space is the same as the X-space.

(b) Show that $\boldsymbol{P}_W = \boldsymbol{P}_X$.

(c) Express $\widehat{\boldsymbol{\alpha}}$ as a function of $\widehat{\boldsymbol{\beta}}$.

6.10 Verify the results of Exercise 6.9 for the data in Example 6.5.

6.11 Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

(a) Let $X_{ci} = X_i - \overline{X}$ denote the centered input variable. Show that

$$Y_i = \alpha_0 + \alpha_1 X_{ci} + \epsilon_i$$

is a reparameterization of the preceding model.

(b) Express α_0 and α_1 in terms of β_0 and β_1 and vice versa.

(c) Are there any advantages in using the centered model?

7

MODEL DEVELOPMENT: VARIABLE SELECTION

The discussion of least squares regression thus far has presumed that the model was known with respect to which variables were to be included and the form these variables should take.

This chapter discusses methods of deciding which variables should be included in the model. It is still assumed that the variables are in the appropriate form. The effect of variable selection on least squares, the use of automated methods of selecting variables, and criteria for choice of subset model are discussed.

The previous chapters dealt with computation and interpretation of least squares regression. With the exception of the case study in Chapter 5, it has been assumed that the independent variables to be used in the model, and the form in which they would be expressed, were known. The properties of the least squares estimators were based on the assumption that the model was correct.

Most regression problems, however, require decisions on which variables to include in the model, the form the variables should take (for example, X, X^2, $1/X$, etc.), and the functional form of the model. This chapter discusses the choice of variables to include in the model. It is assumed that there is a set of t candidate variables, which presumably includes all relevant variables, from which a subset of r variables is to be chosen for the

regression equation. The candidate variables may include different forms of the same basic variable, such as X and X^2, and the selection process may include constraints on which variables are to be included. For example, X may be forced into the model if X^2 is in the selected subset; this a common constraint in building polynomial models (see Chapter 8).

These distinct problem areas are related to this general topic:

1. the theoretical effects of variable selection on the least squares regression results;
2. the computational methods for finding the "best" subset of variables for each subset size; and
3. the choice of subset size (for the final model), or the "stopping rule."

An excellent review of these topics is provided by Hocking (1976). This chapter gives some of the key results on the effects of variable selection, discusses the conceptual operation of automated variable selection procedures (without getting involved in the computational algorithms), and presents several of the commonly used criteria for choice of subset size.

7.1 Uses of the Regression Equation

The purpose of the least squares analysis—how the regression equation is to be used—will influence the manner in which the model is constructed. Hocking (1976) relates these potential uses of regression equations given by Mallows (1973b):

1. providing a good description of the behavior of the response variable;
2. prediction of future responses and estimation of mean responses;
3. extrapolation, or prediction of responses outside the range of the data;
4. estimation of parameters;
5. control of a process by varying levels of input; and
6. developing realistic models of the process.

Each objective has different implications on how much emphasis is placed on eliminating variables from the model, on how important it is that the retained variables be causally related to the response variable, and on the amount of effort devoted to making the model realistic. The concern in this chapter is the selection of variables. Decisions on causality and realism must depend on information from outside the specific data set—for example, on details of how the data were obtained (the experimental design), and on fundamental knowledge of how the particular system operates.

Describing Behavior of Y

When the object is **simple description** of the behavior of the response variable in a particular data set, there is little reason to be concerned about elimination of variables from the model, about causal relationships, or about the realism of the model. The best description of the response variable, in terms of minimum residual sum of squares, will be provided by the full model, and it is unimportant whether the variables are causally related or the model is realistic.

Why Eliminate Variables?

Elimination of variables becomes more important for the other purposes of least squares regression. Regression equations with fewer variables have the appeal of simplicity, as well as an economic advantage in terms of obtaining the necessary information to use the equations. In addition, there is a theoretical advantage of eliminating irrelevant variables and, in some cases, even variables that contain some predictive information about the response variable; this is discussed in Section 7.2. The motivation to eliminate variables is tempered by the biases and loss of predictability that are introduced when relevant variables are eliminated. The objective is to reach a compromise where the final equation satisfies the purpose of the study.

Prediction and Estimation

Of the uses of regression, **prediction** and **estimation** of mean responses are the most tolerant toward eliminating variables. At the same time, it is relatively unimportant whether the variables are causally related or the model is realistic. It is tacitly assumed that prediction and estimation are to be within the X-space of the data and that the system continues to operate as it did when the data were collected. Thus, any variables that contain predictive information on the dependent variable, and for which information can be obtained at a reasonable cost, are useful variables. Of course, more faith could be placed in predictions and estimates based on established causal relationships, because of the protection such models provide against inadvertent extrapolations and unrecognized changes in the correlational structure of the system.

Extrapolation

Extrapolation requires more care in choice of variables. There should be more concern that all relevant variables are retained so that the behavior of the system is described as fully as possible. Extrapolations (beyond the X-space of the data) are always dangerous but can become disastrous if the equation is not a reasonably correct representation of the true model. Any extrapolation carries with it the assumption that the correlational structure observed in the sample continues outside the sample space. Validation and continual updating are essential for equations that are intended to be used for extrapolations (such as forecasts).

Estimation of Parameters

One should also be conservative in eliminating variables when **estimation of parameters** is the objective. This is to avoid the bias introduced when a relevant variable is dropped (see Section 7.2). There is an advantage in terms of reduced variance of the estimates if variables truly unrelated to the dependent variable are dropped.

Control of a System

Control of a system also implies that good estimates of the parameters are needed, but it further implies that the independent variables must have

a causal effect on the response variable. Otherwise, one cannot intervene in a system and effect a change by altering the value of independent variables.

Developing Realistic Models

The objective of basic research is often related to **building realistic models**, usually the most preliminary stages of model building. Understanding the process is the ultimate goal. Whether explicitly stated or not, there will be the desire to identify the variables that are important, through some causal link, in the expression of the dependent variable. For this purpose, variable selection procedures based on the observed correlational structure in a particular set of data become relatively unimportant. At best, they can serve as tools for identifying classes of variables that warrant further study of the causal relationships, usually in controlled experiments. As the objective of the research becomes more oriented toward understanding the process, there will be increasing emphasis on developing models whose functional forms realistically reflect the behavior of the system.

Approach Depends on Purpose

The purpose of introducing these differing objectives is to emphasize that the approach to the selection of variables will depend on the objectives of the analysis. Furthermore, how far a researcher can move in the direction of establishing the importance of variables or causality depends on the source and nature of the data. Least squares regression results reflect only the correlational structure of the data being analyzed. Of itself, least squares analysis cannot establish causal relationships. Causality can be established only from controlled experiments in which the value of the suspected causal variable is changed and the impact on the dependent variable measured. The results from any variable selection procedure, and particularly those that are automated, need to be studied carefully to make sure the models suggested are consistent with the state of knowledge of the process being modeled. ***No variable selection procedure can substitute for the insight of the researcher.***

7.2 Effects of Variable Selection on Least Squares

The effects of variable selection on the least squares results are explicitly developed only for the case where selection is *not* based on information from the current data. This often is not the case, as in the variable selection techniques discussed in this chapter, but the theoretical results for this situation provide motivation for variable selection.

Theoretical Effects of Eliminating Variables

Assume that the correct model involves t independent variables but that a subset of p variables (chosen randomly or on the basis of external information) is used in the regression equation. Let $\boldsymbol{X}_p$ and $\boldsymbol{\beta}_p$ denote submatrices of $\boldsymbol{X}$ and $\boldsymbol{\beta}$ that relate to the p selected variables. $\widehat{\boldsymbol{\beta}}_p$ denotes the least squares estimate of $\boldsymbol{\beta}_p$ obtained from the p-variate subset model. Similarly, $\widehat{Y}_{pi}$, $\widehat{Y}_{\text{pred}_{pi}}$, and $\text{MS}(\text{Res}_p)$ denote the estimated mean for the ith

observation, the prediction for the ith observation, and the mean squared residual, respectively, obtained from the p-variate subset model. Hocking (1976) summarizes the following properties.

1. $\text{MS}(\text{Res}_p)$ is a *positively* biased estimate of σ^2 unless the true regression coefficients for all deleted variables are zero. (See Exercise 7.13.)

2. $\widehat{\boldsymbol{\beta}}_p$ is a biased estimate of $\boldsymbol{\beta}_p$ and $\widehat{Y}_{p_i}$ is a biased estimate of $\mathcal{E}(Y_i)$ unless the true regression coefficient for each deleted variable is zero or, in the case of $\widehat{\boldsymbol{\beta}}_p$, each deleted variable is orthogonal to the p retained variables. (See Exercise 7.13.)

3. $\widehat{\boldsymbol{\beta}}_p$, $\widehat{Y}_{pi}$, and $\widehat{Y}_{\text{pred}_{pi}}$ are generally *less* variable than the corresponding statistics obtained from the t-variate model. (See Exercise 7.13.)

4. There are conditions under which the mean squared errors (variance plus squared bias) of $\widehat{\boldsymbol{\beta}}_p$, $\widehat{Y}_{pi}$, and $\widehat{Y}_{\text{pred}_{pi}}$ are smaller than the variances of the estimates obtained from the t-variate model.

Thus, a bias penalty is paid whenever relevant variables, those with $\beta_j \neq 0$, are omitted from the model (Statements 1 and 2). On the other hand, there is an advantage in terms of decreased variance for both estimation and prediction if variables are deleted from the model (Statement 3). Furthermore, there may be cases in which there is a gain in terms of mean squared error of estimation and prediction from omitting variables whose true regression coefficients are *not* zero (Statement 4).

Sample-Based Selection of Variables

These results provide motivation for selecting subsets of variables, but they do not apply directly to the usual case where variable selection is based on analyses of the current data. The general nature of these effects may be expected to persist, but selection of variables based on their performance in the sample data introduces another class of biases that confound these results. The process of searching through a large number of potential subset models for the one that best fits the data capitalizes on the random variation in the sample to "overfit" the data. That is to say, the chosen subset model can be expected to show a higher degree of agreement with the sample data than the true equation would show with the population data. Another problem of sample-based selection is that relative importance of variables *as manifested in the sample* will not necessarily reflect relative importance in the population. The best subset in the sample, by whatever criterion, need not be the best subset in the population. Important variables in the population may appear unimportant in the sample and consequently be omitted from the model, and vice versa.

Bias in Residual Mean Squared Error

Simulation studies of the effects of subset selection (Berk, 1978) gave sample mean squared errors that were biased downward as much as 25% below the population residual variance when the sample size was less than

50. The bias decreased, as sample size increased, to 2 or 3% when there were several hundred observations in the sample. The percentage bias tended to be largest when the number of variables in the subset was relatively small, $\frac{1}{5}$ to $\frac{1}{2}$ of the number of variables in the full model. This bias in the residual mean squared error translated into bias in the F-ratios for "testing" the inclusion of a variable. The bias in F tended to be largest (positive) for inclusion of the first or second predictor, dropped to near zero before half the variables were added, and became a negative bias as more variables were added.

7.3 All Possible Regressions

Nonorthogonality Among the Independent Variables

When the independent variables in the data set are orthogonal, as they might be in a designed experiment, the least squares results for each variable remain the same regardless of which other variables are in the model. In these cases, the results from a single least squares analysis can be used to choose those independent variables to keep in the model. Usually, however, the independent variables will not be orthogonal. Nonorthogonality is to be expected with observational data and will frequently occur in designed experiments due to unforeseen mishaps. Lack of orthogonality among the independent variables causes the least squares results for each independent variable to be dependent on which other variables are in the model. The full subscript notation for the partial regression coefficients and the R-notation for sums of squares explicitly identify the variables in the model for this reason.

Computing All Possible Regressions

Conceptually, the only way of ensuring that the best model for each subset size has been found is to compute all possible subset regressions. This is feasible when the total number of variables is relatively small, but rapidly becomes a major computing problem even for moderate numbers of independent variables. For example, if there are 10 independent variables from which to choose, there are $2^{10} - 1 = 1,023$ possible models to be evaluated. Much effort has been devoted to finding computing algorithms that capitalize on the computations already done for previous subsets in order to reduce the total amount of computing for all possible subsets [e.g., Furnival (1971)]. Furnival (1971) also pointed out that much less computing is required if one is satisfied with obtaining only the residual sum of squares from each subset model.

Finding Best Subsets

More recently, attention has focused on identifying the best subsets within each subset size without computing all possible subsets. These methods utilize the basic least squares property that the residual sums of squares cannot decrease when a variable is dropped from a model. Thus, comparison of residual sums of squares from different subset models is used to eliminate the need to compute other subsets. For example, if a two-variable subset

has already been found that gives a residual sum of squares less than some three-variable model, then none of the two-variable subsets of the three-variable model need be computed; they will all give residual sums of squares larger than that from the three-variable model and, hence, larger than for the two-variable model already found. The **leaps-and-bounds algorithm** of Furnival and Wilson (1974) combines comparisons of residual sums of squares for different subset models with clever control over the sequence in which subset regressions are computed. This algorithm guarantees finding the best m subset regressions within each subset size with considerably less computing than is required for all possible subsets. The RSQUARE method in PROC REG (SAS Institute Inc., 1989b) uses the leaps-and-bounds algorithm. These computing advances have made all possible regressions a viable option in most cases.

Example 7.1

The Linthurst data used in the case study in Chapter 5 are used to illustrate the model selection methods of this chapter. First, the regressions for all possible models are computed to find the "best" model for this data set and to serve as references for the stepwise methods to follow. The five independent variables used in Chapter 5 are also used here as potential variables for the model. Thus, there are $2^5 - 1 = 31$ possible regression models: 5 one-variable, 10 two-variable, 10 three-variable, 5 four-variable, and 1 five-variable model.

The RSQUARE method in PROC REG (SAS Institute, Inc., 1989b) was used to compute all possible regressions. In Table 7.1,the subset models are ranked within each subset size (p') from the best to the worst fitting model. (Table 7.1 includes the results from six criteria discussed later. For the present discussion, only the coefficient of determination R^2 is used.) The full model $p' = 6$ accounts for $100R^2 = 67.7\%$ of the variation in the dependent variable BIOMASS. No subset of the independent variables can give a larger R^2.

Of the univariate subsets, the best, pH, accounted for 59.9% of the variation in BIOMASS, 8% below the maximum. The second best univariate subset Zn accounted for only 39% of the variation in Y. The best two-variable model pH and Na accounted for 65.8%, only 2% below the maximum. The second best two-variable subset pH and K is very nearly as good, with $100R^2 = 64.8\%$. Note that the second best single variable is not contained in either of the two best two-variable subsets.

There are three 3-variable models that are equally effective for all practical purposes, with $100R^2$ ranging from 65.9% to 66.3%. All three of these subsets include pH and Na. Thus, it makes little difference which of the three variables SAL, Zn, or K is added to the best 2-variable subset. The two best 4-variable subsets are also equally effective; the best in this subset does not include the best 2-variable or 3-variable subsets.

TABLE 7.1. *Summary statistics R^2, MS(Res), R^2_{adj}, and C_p from all possible regressions for Linthurst data using the five independent variables SALINITY, pH, K, Na, and Zn. All models included an intercept. (Data used with permission.)*

p'	*Variables*	R^2	*MS(Res)*	R^2_{adj}	C_p	*AIC*	*SBC*
2	*pH*	.599	178618	.590	7.4	546.1	549.8
	Zn	.390	272011	.376	32.7	565.1	568.7
	Na	.074	412835	.053	70.9	583.8	587.5
	K	.042	427165	.020	74.8	585.4	589.0
	SAL	.011	441091	−.012	78.6	586.8	590.4
3	*pH, Na*	.658	155909	.642	2.3	541.0	546.4
	pH, K	.648	160865	.631	3.6	542.2	547.8
	pH, Zn	.608	178801	.590	8.3	547.1	552.5
	SAL, pH	.603	181030	.585	8.9	547.7	553.1
	SAL, Zn	.553	204209	.531	15.1	553.1	558.5
	Na, Zn	.430	260164	.403	29.9	564.0	569.4
	K, Zn	.415	266932	.387	31.7	565.2	570.6
	SAL, Na	.078	421031	.034	72.5	585.7	591.1
	K, Na	.074	422520	.030	72.9	585.8	591.2
	SAL, K	.053	432069	.008	75.4	586.8	592.3
4	*pH, Na, Zn*	.663	157833	.638	3.8	542.4	549.7
	pH, K, Na	.660	158811	.636	4.1	542.7	549.9
	SAL, pH, Na	.659	159424	.634	4.2	542.9	550.1
	SAL, pH, K	.652	162636	.627	5.0	543.8	551.0
	pH, K, Zn	.652	162677	.627	5.1	543.8	551.0
	SAL, pH, Zn	.637	169900	.610	6.9	545.7	553.0
	SAL, K, Zn	.577	198026	.546	14.2	552.6	559.9
	SAL, Na, Zn	.564	203666	.533	15.6	553.9	561.1
	K, Na, Zn	.430	266509	.388	31.9	566.0	573.2
	SAL, K, Na	.078	431296	.010	74.5	587.7	594.9
5	*SAL, pH, K, Zn*	.675	155832	.642	4.3	542.7	551.8
	SAL, pH, Na, Zn	.672	157312	.639	4.7	543.2	552.2
	pH, K, Na, Zn	.664	160955	.631	5.6	544.2	553.2
	SAL, pH, K, Na	.662	162137	.628	5.9	544.5	553.6
	SAL, K, Na, Zn	.577	202589	.535	16.1	554.6	563.6
6	*SAL, pH, K, Na, Zn*	.677	158622	.636	6	544.4	555.2

A key point to note from the all-possible-regressions analysis is that more than one model is in contention for nearly every subset size. With only minor differences in R^2 for the best two or three subsets in each case, it is very likely that other considerations, such as behavior of the residuals, cost of obtaining information, or prior knowledge on the importance of the variables, could shift the final choice of model away from the "best" subset.

For this example, adding a second independent variable to the model increased R^2 by 6%. However, the third, fourth, and fifth variables increased R^2 by only .4%, 1.2%, and .2%, respectively. The improvement obtained from the second variable would appear worthwhile, but the value of adding the third, fourth, and fifth variables is questionable. Further discussion of choice of subset size is delayed until the different criteria for the choice of subset size have been discussed. ■

7.4 Stepwise Regression Methods

Alternative variable selection methods have been developed that identify good (although not necessarily the best) subset models, with considerably less computing than is required for all possible regressions. These methods are referred to as **stepwise regression methods**. The subset models are identified sequentially by adding or deleting, depending on the method, the one variable that has the greatest impact on the residual sum of squares. These stepwise methods are not guaranteed to find the "best" subset for each subset size, and the results produced by different methods may not agree with each other.

Forward Selection

Forward stepwise selection of variables chooses the subset models by adding one variable at a time to the previously chosen subset. Forward selection starts by choosing as the one-variable subset the independent variable that accounts for the largest amount of variation in the dependent variable. This will be the variable having the highest simple correlation with Y. At each successive step, the variable in the subset of variables *not* already in the model that causes the largest decrease in the residual sum of squares is added to the subset. Without a termination rule, forward selection continues until all variables are in the model.

Backward Elimination

Backward elimination of variables chooses the subset models by starting with the full model and then eliminating at each step the one variable whose deletion will cause the residual sum of squares to increase the least. This will be the variable in the current subset model that has the smallest partial sum of squares. Without a termination rule, backward elimination continues until the subset model contains only one variable.

Stepwise Selection

Neither forward selection nor backward elimination takes into account the effect that the addition or deletion of a variable can have on the con-

tributions of other variables to the model. A variable added early to the model in forward selection can become unimportant after other variables are added, or variables previously dropped in backward elimination can become important after other variables are dropped from the model. The variable selection method commonly labeled **stepwise regression** is a forward selection process that rechecks at each step the importance of all previously included variables. If the partial sums of squares for any previously included variables do not meet a minimum criterion to stay in the model, the selection procedure changes to backward elimination and variables are dropped one at a time until all remaining variables meet the minimum criterion. Then, forward selection resumes.

Stepwise selection of variables requires more computing than forward or backward selection but has an advantage in terms of the number of potential subset models checked before the model for each subset size is decided. It is reasonable to expect stepwise selection to have a greater chance of choosing the best subsets in the sample data, but selection of the best subset for each subset size is not guaranteed.

Stopping Rules

The computer programs for the stepwise selection methods generally include criteria for terminating the selection process. In forward selection, the common criterion is the ratio of the reduction in residual sum of squares caused by the next candidate variable to be considered to the residual mean square from the model including that variable. This criterion can be expressed in terms of a critical "F-to-enter" or in terms of a critical "significance level to enter" (SLE), where F is the "F-test" of the partial sum of squares of the variable being considered. The forward selection terminates when no variable outside the model meets the criterion to enter. This "F-test," and the ones to follow, should be viewed only as *stopping rules* rather than as classical tests of significance. The use of the data to select the most favorable variables creates biases that invalidate these ratios as tests of significance (Berk, 1978).

The stopping rule for backward elimination is the "F-test" of the smallest partial sum of squares of the variables remaining in the model. Again, this criterion can be stated in terms of an "F-to-stay" or as a "significance level to stay" (SLS). Backward elimination terminates when all variables remaining in the model meet the criterion to stay.

The stopping rule for stepwise selection of variables uses both the forward and backward elimination criteria. The variable selection process terminates when all variables in the model meet the criterion to stay *and* no variables outside the model meet the criterion to enter (except, perhaps, for the variable that was just eliminated). The criterion for a variable to enter the model need not be the same as the criterion for the variable to stay. There is some advantage in using a more relaxed criterion for entry to force the selection process to consider a larger number of subsets of variables.

Example 7.2

(*Continuation of Example* 7.1) The **FORWARD**, **BACKWARD**, and **STEPWISE** methods of variable selection in PROC REG (SAS Institute, Inc., 1989b) are illustrated with the Linthurst data. In this program, the termination rules are expressed in terms of significance level to enter, and significance level to stay. For this example, the criteria were set at SLE = .50 in forward selection, SLS = .10 in backward elimination, and SLE = .50 and SLS = .15 in stepwise selection. The values were chosen for forward and backward selection to allow the procedures to continue through most of the subset sizes. One can then tell by inspection of the results where the selection would have terminated with more stringent criteria.

The results from the *forward selection* method applied to the Linthurst data are summarized in Table 7.2. The F-ratio given is the ratio of the partial sum of squares for the variable to the mean square residual for the model containing all previously admitted variables plus the one being considered.

The best single variable is pH which gives $(100)R^2 = 59.9\%$ (see Table 7.1) and $F = 64.3$. The corresponding significance level is far beyond the significance level needed to enter, SLE = .50. The second step of the forward selection computes the partial sums of squares for each of the remaining variables, *SALINITY*, K, *Na*, and Zn, in a model that contains pH plus that particular variable. The partial sum of squares for *Na* is the largest and gives $F = 7.26$, or Prob $> F = .0101$, which satisfies the criterion for entry. Thus, *Na* is added to the model and the selection process goes to Step 3. At the third step, the partial sum of squares for Zn is the largest and Prob $> F = .4888$ just meets the criterion for entry. $SALINITY$ meets the criterion for entry at the fourth step, and K at the fifth step.

In this case, the choice of SLE = .50 allowed all variables to be included in the model. The selection would have stopped at the two-variable model with pH and *Na* had SLE been chosen anywhere between .4888 and .0101. Any choice of SLE less than .0101 would have stopped the selection process with the one-variable model.

Forward selection chose the best subset models for $p = 1$, 2, and 3, but the second best model for $p = 4$ (see Table 7.1). This illustrates the fact that the stepwise methods are not guaranteed to find the best subset model for each subset size. In addition, the stepwise methods do not alert the user to the fact that other subsets at each stage may be as good. For example, one is not aware from the forward selection results that two other three-variable subsets [(pH, K, *Na*) and (SAL, pH, *Na*)] are essentially equivalent to the one chosen.

The stepwise regression results using *backward elimination* are summarized in Table 7.3. Starting with the full model, the procedure eliminates the variable with the smallest partial sum of squares if its sum of squares does not meet the criterion to stay in the model. In this example, the significance level to stay is set at SLS = .10. *Na* has the smallest partial sum of

TABLE 7.2. *Summary statistics for forward selection of variables for the Linthurst data using significance level for variable to enter the model of SLE = .50.*

Step	*Variable*	*Partial SS*	*MS(Res)*	R^2	F^a	$Prob > F^b$
1. Determine best single variable and test for entry:						
	Sal	204048	441091	.0106	.46	.5001
	pH	11490388	178618	.5994	64.33	.0001
	K	802872	427165	.0419	1.88	.1775
	Na	1419069	412834	.0740	3.44	.0706
	Zn	7474474	272011	.3899	27.48	.0001
	Best 1-variable model: pH				$C_p = 7.42$	
2. Determine best second variable and test for entry:						
	Sal	77327	181030	.6034	.43	.5170
	K	924266	160865	.6476	5.75	.0211
	Na	1132401	155909	.6584	7.26	.0101
	Zn	170933	178801	.6083	.96	.3338
	Best 2-variable model: pH Na				$C_p = 2.28$	
3. Determine best third variable and test for entry:						
	Sal	11778	159424	.6590	.07	.7871
	K	36938	158804	.6604	.23	.6322
	Zn	77026	157833	.6625	.49	.4888
	Best 3-variable model: pH Na Zn				$C_p = 3.80$	
4. Determine best fourth variable and test for entry:						
	SAL	178674	157312	.6718	1.136	.2929
	K	32964	160955	.6642	.205	.6533
	Best 4-variable model: pH Na Zn SAL				$Cp = 4.67$	
5. Test last variable for entry:						
	K	106211	158622	.6773	.670	.4182
	Last variable is added with SLE = .50				$C_p = 6.00$	

[a] F-test of partial sum of squares.
[b] $Prob > F$ assuming the ratio is a valid F-statistic.

TABLE 7.3. *Summary statistics for the backward elimination of variables for the Linthurst data using significance level for staying of SLS = .10. All models included an intercept.*

Step		*Variable*	*Partial SS*	$R^{2\,a}$	F^b	$Prob > F^c$
0	*Model* :	All variables; $R^2 = .6773$, $C_p = 6$, $s^2 = 158,616$ with 39 d.f.				
		SAL	251,921	.6642	1.59	.2151
		pH	1,917,306	.5773	12.09	.0013
		K	106,211	.6718	.67	.4182
		Na	46,011	.6749	.30	.5893
		Zn	299,209	.6617	1.89	.1775
1	*Model* :	Na removed; $R^2 = .6749$, $C_p = 4.30$, $s^2 = 155,824$ with 40 d.f.				
		Sal	436,496	.6521	2.80	.1020
		pH	1,885,805	.5765	12.10	.0012
		K	732,606	.6366	4.70	.0361
		Zn	434,796	.6522	2.79	.1027
2	*Model* :	Zn removed; $R^2 = .6522$, $C_p = 5.04$, $s^2 = 162,636$ with 41 d.f.				
		Sal	88,239	.6476	.54	.4656
		pH	11,478,835	.0534	70.58	.0001
		K	935,178	.6034	5.75	.0211
3	*Model* :	Sal removed; $R^2 = .6476$, $C_p = 3.59$, $s^2 = 160,865$ with 42 d.f.				
		pH	11,611,782	.0419	72.18	.0001
		K	924,266	.5994	5.75	.0211
STOP. Prob> F for each remaining variable exceeds SLS = .10. Final model contains pH, K and an intercept.						

[a] R^2 is for the model with the indicated variable removed.
[b] F-ratio for the partial sum of squares for the indicated variable.
[c] Probability of a larger F assuming it is a valid F-statistic.

squares and is eliminated from the model since Prob > F = .5893 is larger than SLS = .10. This leaves (SAL, pH, K, Zn) as the chosen four-variable subset. Of these four variables, Zn has the smallest partial sum of squares (by a very small margin over $SALINITY$) and Prob > F = .1027, slightly larger than the criterion to stay SLS = .10. Therefore, Zn is dropped from the model leaving (SAL, pH, K) as the chosen three-variable model. At the next step, SAL is dropped, giving (pH, K) as the chosen two-variable model. Both pH and K meet the criterion to stay (Prob > F is less than SLS), and the backward selection process stops with that model.

Backward elimination identifies the best four-variable subset whereas forward selection did not. On the other hand, backward elimination chose the fourth best three-variable subset and the second best two-variable subset, whereas forward selection identified the best subset at these stages. If SLS had been set low enough, say at .02, backward elimination would have gone one step further and correctly identified pH as the best one-variable subset.

The *stepwise* method of stepwise variable selection applied to the Linthurst data starts the same as forward selection (Table 7.2). After the second step when pH and Na are both in the model, the stepwise method rechecks the contribution of each variable to determine if each should stay in the model. The partial sums of squares are

$$\begin{aligned} R(\beta_{pH}|\beta_{Na}) &= 11,203,720 \\ R(\beta_{Na}|\beta_{pH}) &= 1,132,401. \end{aligned}$$

The mean square residual for this model is MS(Res) = 155,909 with 42 degrees of freedom. Both give large F-ratios with Prob > F much smaller than SLS = .15 so that both pH and Na are retained.

The forward selection phase of stepwise selection resumes with the choice of Zn as the third variable to be added (Table 7.2). Again, the contribution of each variable in the model is rechecked to determine if each should stay. The partial sums of squares are

$$\begin{aligned} R(\beta_{pH}|\beta_{Na}\ \beta_{Zn}) &= 4,455,726, \\ R(\beta_{Na}|\beta_{pH}\ \beta_{Zn}) &= 1,038,493, \quad \text{and} \\ R(\beta_{Zn}|\beta_{pH}\ \beta_{Na}) &= 77,026. \end{aligned}$$

The mean square residual for this model is MS(Res) = 157,833 with 41 degrees of freedom. Both pH and Na meet the criterion to stay, but the F-ratio for Zn is less than 1.0 with Prob > F = .4888, which does not meet the criterion of SLS = .15. Therefore, Zn, which has just been added, is immediately dropped from the model.

The stepwise procedure then checks to see if any variables other than Zn meet the criterion to enter the model. The two remaining variables to be checked are $SALINITY$ and K. The partial sum of squares for each, adjusted for pH and Na, is given in Step 3 of the forward selection,

Table 7.2. The Prob $>$ F for both variables is larger than SLE = .50. Therefore, no other variables meet the criterion to enter the model and all variables in the model meet the criterion to stay so the selection terminates with the two-variable subset (pH, Na).

In general, the rechecking of previous decisions in stepwise selection should improve the chances of identifying the best subsets at each subset size. In this particular example, the choice of SLS = .15 caused the stepwise selection to terminate early. If SLS had been chosen equal to SLE = .50, stepwise regression would have followed the same path as forward selection until the fifth variable K had been added to the model. Then, rechecking the variables in the model would have caused Na to be dropped from the model leaving (SAL, pH, K, Zn) as the selected four-variable subset. This is the best four-variable subset (Table 7.1), which forward selection failed to identify. ■

Warnings on Using Stepwise Methods

Even in the small example just discussed, there are several close contenders within most subset sizes as shown by all possible regressions (Table 7.1). Each stepwise regression method reveals only one subset at each step and, if the stopping criteria are set to select a "best" subset size, only part of the subset models are identified. (Choice of criteria for this purpose are discussed in Section 7.5.) In general, it is not recommended that the automated stepwise regression methods be used blindly to identify a "best" model. It is imperative that any model obtained in this manner be thoroughly checked for any inadequacies (see Chapter 10) and validated against an independent data set before being adopted (see Section 7.6).

If stepwise variable selection methods are to be used, they are best used as screening tools to identify contender models. For this purpose, forward selection and backward elimination methods alone provide very narrow views of the possible models. Stepwise selection would be somewhat better. An even better option would be the joint use of all three methods. If forward selection and backward elimination identify the same subsets, then it is known that they will have identified the best subset in each subset size (Berk, 1978). One still would not have information on close contenders within each subset size. For screening purposes, the choice of the termination criteria should be such as to provide the greatest exposure to alternative models. For forward selection, this means that SLE should be large, say SLE = .5 or larger. For backward elimination, SLS should be small. For the stepwise method of selection, SLE should be large but the choice of SLS is not so easily specified. It may be worthwhile to try more than one choice of each.

For the purpose of identifying several contender models, one should not overlook the possible use of a program that utilizes the "leaps-and-bounds" algorithm, such as the RSQUARE option in PROC REG (SAS Institute, Inc., 1989b). This algorithm guarantees that the best m subset models

within each subset size will be identified. Changing m from 1 to 10 approximately doubles the computing time (Furnival and Wilson, 1974). Although the computing cost will be higher than for any of the stepwise methods, the cost may not be excessive and considerably more information is obtained.

7.5 Criteria for Choice of Subset Size

Many criteria for choice of subset size have been proposed. These criteria are based on the principle of parsimony which suggests selecting a model with small residual sum of squares with as few parameters as possible. Hocking (1976) reviews 8 stopping rules, Bendel and Afifi (1977) compare 8 (not all the same as Hocking's) in forward selection, and the RSQUARE method in PROC REG (SAS Institute, Inc., 1989b) provides the option of computing 12. Most of the criteria are monotone functions of the residual sum of squares for a given subset size and, consequently, give identical rankings of the subset models within each subset size. However, the choice of criteria may lead to different choices of subset size, and they may give different impressions of the magnitude of the differences among subset models. The latter may be particularly relevant when the purpose is to identify several competing models for further study.

Six commonly used criteria are discussed briefly. In addition, the choice of F-to-enter and F-to-stay, or the corresponding "significance levels" SLE and SLS are reviewed. The six commonly used criteria to be discussed are

- coefficient of determination R^2,
- residual mean square MS(Res),
- adjusted coefficient of determinationR^2_{adj},
- Mallows' C_p statistic, and
- two information criteria AIC and SBC.

The values for these criteria are given in Table 7.1 for all possible subsets from the Linthurst data.

7.5.1 Coefficient of Determination

Behavior of R^2

The **coefficient of determination** R^2 is the proportion of the total (corrected) sum of squares of the dependent variable "explained" by the independent variables in the model:

$$R^2 = \frac{\text{SS(Regr)}}{\text{SS(Total)}}. \tag{7.1}$$

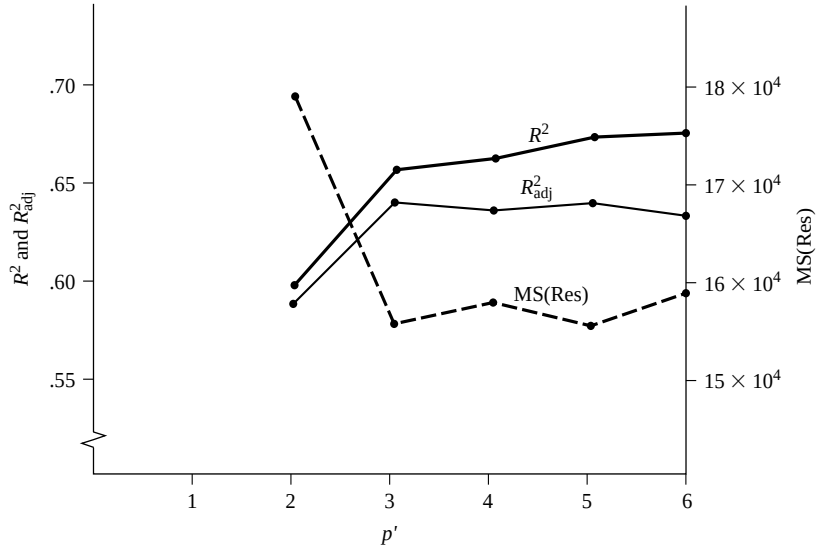

FIGURE 7.1. *R^2, R^2_{adj}, and MS(Res) plotted against p' for the best model from each subset size for the Linthurst data.*

The objective is to select a model that accounts for as much of the variation in Y as is practical. Since R^2 cannot decrease as independent variables are added to the model, the model that gives the maximum R^2 will necessarily be the model that contains all independent variables. The typical plot of R^2 against the number of variables in the model starts as a steeply upward sloping curve, then levels off near the maximum R^2 once the more important variables have been included. Thus, the use of the R^2 criterion for model building requires a judgment as to whether the increase in R^2 from additional variables justifies the increased complexity of the model. The subset size is chosen near the bend where the curve tends to flatten.

Example 7.3

(*Continuation of Example* 7.1) The best one-variable subset accounted for $100R^2 = 59.9\%$ of the variation in *BIOMASS*, the best two-variable subset accounted for $100R^2 = 65.8\%$, and the best three-variable subset accounted for $100R^2 = 66.3\%$ (see Figure 7.1 on page 221). The increase in R^2 from two to three variables was small and R^2 is close to the maximum of $100R^2 = 67.7\%$. Thus, the R^2 criterion leads to the choice of the two-variable subset containing *pH* and *Na* as the "best." ■

7.5.2 Residual Mean Square

Expected Behavior of MS(Res)

The **residual mean square** MS(Res) is an estimate of σ^2 if the model contains all relevant independent variables. If relevant independent variables have been omitted, the residual mean square is biased upward. Including an unimportant independent variable will have little impact on the residual mean square. Thus, the expected behavior of the residual mean square, as variables are added to the model, is for it to decrease toward σ^2 as important independent variables are added to the model and to fluctuate around σ^2 once all relevant variables have been included.

Behavior with Variable Selection

The previous paragraph describes the expected behavior of MS(Res) when the selection of variables is not based on sample data. Berk (1978) demonstrated with simulation that selection of variables based on the sample data causes MS(Res) to be biased downward. In his studies, the bias was as much as 25% when sample sizes were less than 50. The bias tended to reach its peak in the early stages of forward selection, when one-third to one-half of the total number of variables had been admitted to the model. In backward elimination, the bias tended to peak when slightly more than half of the variables had been eliminated. These results suggest that the pattern of MS(Res) as variables are added in a variable selection procedure will be to drop slightly below σ^2 in the intermediate stages of the selection and then return to near σ^2 as the full model is approached. It is unlikely that a bias of this magnitude would be detectable in plots of MS(Res) against number of variables, particularly in small samples where the bias is most serious.

The pattern of the residual mean squares, as variables are added to the model, is used to judge when the residual mean square is estimating σ^2 and, by inference, when the model contains the important independent variables. In larger regression problems, with many independent variables and several times as many observations, a plot of the residual mean square against the number of parameters in the model will show when the plateau has been reached. The plateau may not be clearly defined in smaller problems.

Example 7.4

For the Linthurst data (Example 7.1), MS(Res) drops from MS(Res) = 178,618 for the best one-variable subset to MS(Res) = 155,909 for the best two-variable subset, and then changes little beyond that point (see Table 7.1 and Figure 7.1). The two-variable subset would be chosen by this criterion. ■

7.5.3 Adjusted Coefficient of Determination

Behavior of R^2_{adj}

The **adjusted** R^2, which is labeled as R^2_{adj}, is a rescaling of R^2 by degrees of freedom so that it involves a ratio of mean squares rather than sums of

squares:

$$\begin{aligned} R^2_{adj} &= 1 - \frac{\text{MS(Res)}}{\text{MS(Total)}} \\ &= 1 - \frac{(1-R^2)(n-1)}{(n-p')}. \end{aligned} \tag{7.2}$$

This expression removes the impact of degrees of freedom and gives a quantity that is more comparable than R^2 over models involving different numbers of parameters. Unlike R^2, R^2_{adj} need not always increase as variables are added to the model. The value of R^2_{adj} will tend to stabilize around some upper limit as variables are added. The simplest model with R^2_{adj} near this upper limit is chosen as the "best" model. R^2_{adj} is closely related to MS(Res) (see equation 7.2) and will lead to the same conclusions.

Example 7.5

For the Linthurst data, the maximum R^2_{adj} for the one-variable subset is $R^2_{adj} = .590$ (see Table 7.1 and Figure 7.1). This increases to .642 for the two-variable subset, and then shows no further increase; $R^2_{adj} = .638$, .642, and .636 for $p' = 4$, 5, and 6, respectively. ■

7.5.4 Mallows' C_p Statistic

Behavior of C_p

The C_p statistic is an estimate of the standardized total mean squared error of estimation for the current set of data (Hocking, 1976). The C_p statistic and the C_p plot were initially described by Mallows [see Mallows (1973a) for earlier references]. The C_p statistic is computed as

$$C_p = \frac{\text{SS(Res)}_p}{s^2} + 2p' - n, \tag{7.3}$$

where SS(Res)_p is the residual sum of squares from the p-variable subset model being considered and s^2 is an estimate of σ^2, either from independent information or, more commonly, from the model containing all independent variables. When the model is correct, the residual sum of squares is an unbiased estimate of $(n-p')\sigma^2$; in this case, C_p is (approximately) equal to p'. When important independent variables have been omitted from the model, the residual sum of squares is an estimate of $(n-p')\sigma^2$ *plus* a positive quantity reflecting the contribution of the omitted variables; in this case, C_p is expected to be greater than p'.

C_p Plot

The C_p plot presents C_p as a function of p' for the better subset models and provides a convenient method of selecting the subset size and judging the competitor subsets. The usual pattern is for the minimum C_p statistic for each subset size $C_{p\,\min}$ to be much larger than p' when p' is small, to

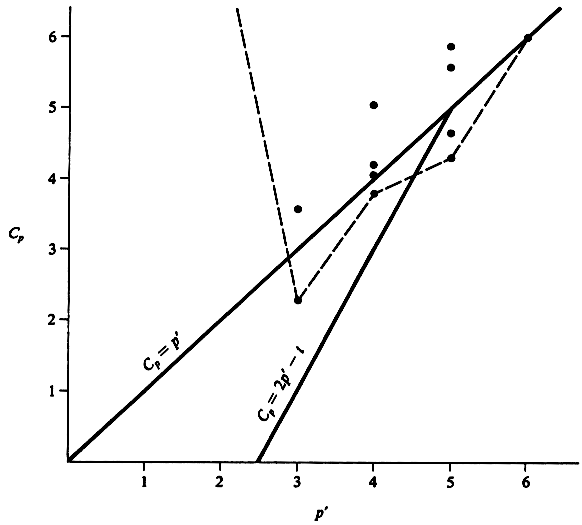

FIGURE 7.2. *The C_p plot of the Linthurst data. The dashed line connects $C_{p\,min}$ for each subset size. The two solid lines are the reference lines for subset selection according to Hocking's criteria.*

decrease toward p' as the important variables are added to the model, and then to fall below or fluctuate around p'. When the residual mean square from the full model has been used as s^2, C_p will equal p' for the full model. A value of C_p near p' indicates little bias in MS(Res) as an estimate of σ^2. (This interpretation assumes that s^2 in the denominator of C_p is an unbiased estimate of σ^2. If s^2 has been obtained from the full model, s^2 is an unbiased estimate of σ^2 only if the full model contains all relevant variables.)

Cp Criterion

Different criteria have been advanced for the use of C_p. Mallows (1973a) suggested that all subset models with small C_p and with C_p close to p' be considered for further study. Hocking (1976) defined two criteria depending on whether the model is intended primarily for prediction or for parameter estimation. He used the criterion $C_p \leq p'$ for prediction. For parameter estimation, Hocking argued that fewer variables should be eliminated from the model, to avoid excessive bias in the estimates, and provided the selection criterion $C_p \leq 2p' - t$, where t is the number of variables in the full model.

Example 7.6

The C_p plot for the Linthurst example is given in Figure 7.2. Only the smaller C_p statistics, the dots, are shown for each value of p', with the $C_{p\,min}$ values connected by the dashed line. The figure includes two reference lines corresponding to Hocking's two criteria $C_p = p'$ and $C_p = 2p' - t$. The C_p

statistics for all subsets are given in Table 7.1. For the 1-variable subsets, $C_{p\,\min} = 7.42$, well above $p' = 2$. For the 2-variable subsets, $C_{p\,\min} = 2.28$, just below $p' = 3$. The next best 2-variable subset has $C_p = 3.59$, somewhat above $p' = 3$. Three 3-variable subsets give C_p close to $p' = 4$ with $C_{p\,\min} = 3.80$. The C_p statistics for the 4-variable subsets identify two subsets with $C_p \leq p'$. Two other subsets have C_p slightly greater than p'.

Mallows' C_p criterion (which requires C_p small and near p') identifies the two-variable subsets (pH, Na) and (pH, K), and possibly the three-variable subsets (pH, Na, Zn), (pH, K, Na), and ($SALINITY$, pH, Na). Preference would be given to (pH, Na) if this model appears to be adequate when subjected to further study. Hocking's criterion for selection of the best subset model *for prediction* leads to the two-variable model (pH, Na); $C_p = 2.28$ is less than $p' = 3$. The more restrictive criterion for subset selection *for parameter estimation* leads to the best four-variable subset ($SALINITY$, pH, K, Zn); $C_p = 4.30$ is less than $2p' - t = 5$. ■

7.5.5 Information Criteria: AIC and SBC

Behavior of AIC and SBC

The Akaike (1969) Information Criterion (AIC) is computed as

$$\text{AIC}(p') = n \ln(\text{SS(Res)}_p) + 2p' - n \ln(n). \tag{7.4}$$

(Note that all logarithmic functions in this text use base e.) Since SS(Res)_p decreases as the number of independent variables increases, the first term in AIC decreases with p'. However, the second term in AIC increases with p' and serves as a penalty for increasing the number of parameters in the model. Thus, it trades off precision of fit against the number of parameters used to obtain that fit. A graph of $\text{AIC}(p')$ against p' will, in general, show a minimum value, and the appropriate value of the subset size is determined by the value of p' at which $\text{AIC}(p')$ attains its minimum value.

The AIC criterion is widely used, although it is known that the criterion tends to select models with larger subset sizes than the true model. [See Judge, Griffiths, Hill, and Lee (1980).] Because of this tendency to select models with larger number of independent variables, a number of alternative criteria have been developed. One such criterion is Schwarz (1978) Bayesian Criterion (SBC) given by

$$\text{SBC}(p') = n \ln(\text{SS(Res)}_p) + [\ln(n)]p' - n \ln(n). \tag{7.5}$$

Note that SBC uses the multiplier ln(n), (instead of 2 in AIC) for the number of parameters p' included in the model. Thus, it more heavily penalizes models with a larger number of independent variables than does AIC. The appropriate value of the subset size is determined by the value of p' at which $\text{SBC}(p')$ attains its minimum value.

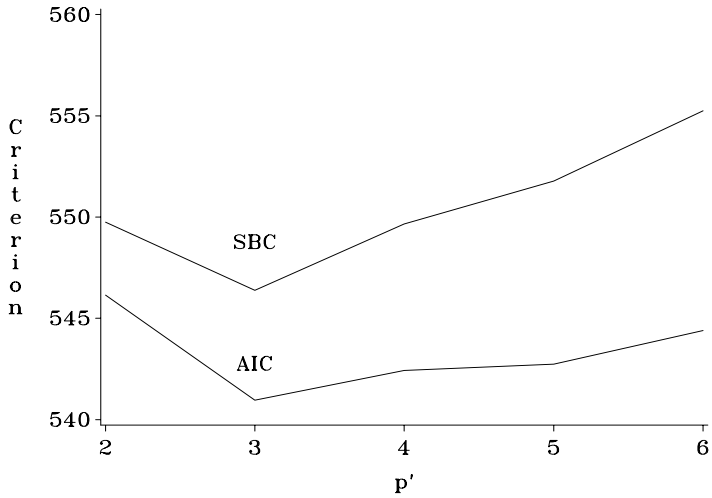

FIGURE 7.3. *Minimum AIC and SBC values plotted against p' for each subset size for the analysis of the Linthurst data.*

Example 7.7

The values of AIC and SBC for the regression analysis of the Linthurst data are given in the last two columns of Table 7.1 and the minimum values for each subset size are plotted in Figure 7.3. The minimum value for both criteria occurs at $p' = 3$ and for the model containing pH and Na as the independent variables. It should be noted that the AIC and SBC values for the two-variable containing pH and K are only slightly larger than the minimum values. ■

7.5.6 "Significance Levels" for Choice of Subset Size

Use of "Significance Levels"

F-to-enter and F-to-stay, or the equivalent "significance levels," in the stepwise variable selection methods serve as subset size selection criteria when they are chosen so as to terminate the selection process before all subset sizes have been considered. Bendel and Afifi (1977) compared several stopping rules for forward selection and showed that the sequential F-test based on a constant "significance level" compared very favorably. The optimum "significance level to enter" varied between SLE = .15 and .25. Although not the best of the criteria they studied, the sequential F-test with SLE = .15 allowed one to do "almost best" when $n - p \leq 20$. When $n - p \geq 40$, the C_p statistic was preferred over the sequential F-test but by a very slight margin if SLE = .20 were used.

This is similar to the conclusion reached by Kennedy and Bancroft (1971) for the sequential F-test but where the order of importance of the variables was known a priori. They concluded that the significance level should be .25 for forward selection and .10 for backward elimination. Bendel and Afifi did not speculate on the choice of "significance level to stay" in backward elimination. For stepwise selection, they recommended the same levels of SLE as for forward selection and half that level for SLS.

Example 7.8

For the Linthurst data of Example 7.1, the Bendel and Afifi level of SLE = .15 would have terminated forward selection with the two-variable subset (pH, Na) (see Table 7.2). The Kennedy and Bancroft suggestion of using SLS = .10 for backward elimination gives the results shown in Table 7.3 terminating with the two-variable subset (pH, K). In this case, the backward elimination barely continued beyond the second step where the least significant of the four variables had Prob $> F = .1027$. The recommended significance levels of SLE = 2(SLS) = .15 for the stepwise selection method terminates at the same point as forward selection. ■

Conclusions

In summary of the choice of subset size, some of the other conclusions of Bendel and Afifi (1977) regarding stopping rules are of importance. First, the use of all independent variables is a very poor rule unless $n - p'$ is very large. For their studies, the use of all variables was always inferior to the best stopping rule. This is consistent with the theoretical results (Section 7.2) that showed larger variances for $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{Y}}$, and $\widehat{\boldsymbol{Y}}_{\text{pred}}$ for the full models. Second, most of the stopping rules do poorly if $n - p' \leq 10$. The C_p statistic does poorly when $n - p' \leq 10$ (but is recommended for $n - p' \geq 40$). Third, the lack-of-fit test of the $(t - p')$ variables that have been dropped (an intuitively logical procedure but not discussed in this text) is generally very poor as a stopping rule regardless of the significance level used. Finally, an unbiased version of the coefficient of determination generally did poorly unless $n - p'$ was large. This suggests that R^2, and perhaps R^2_{adj} and MS(Res), may not serve as good stopping rules for subset size selection.

Mallows' C_p statistic and significance levels appear to be the most favored criteria for subset size selection. The C_p statistic was not the optimum choice of Bendel and Afifi in the intermediate-sized data sets and it did poorly for very small samples. Significance level as a criterion did slightly better than C_p in the intermediate-sized studies. The poor performance of C_p in the small samples should not be taken as an indictment. First, none of the criteria did well in such studies and, second, no variable selection routine or model building exercise should be taken seriously when the sample sizes are as small as $n - p' \leq 10$.

7.6 Model Validation

Importance of Validation

Validation of a fitted regression equation is the demonstration or confirmation that the model is sound and effective for the purpose for which it was intended. This is *not* equivalent to demonstrating that the fitted equation agrees well with the data from which it was computed. Validation of the model requires assessing the effectiveness of the fitted equation against an *independent* set of data, and is essential if confidence in the model is to be expected.

Results from the regression analysis—R^2, MS(Res), and so forth—do not necessarily reflect the degree of agreement one might obtain from future applications of the equation. The model-building exercise has searched through many possible combinations of variables and mathematical forms for the model. In addition, least squares estimation has given the best possible agreement of the chosen model with the observed data. As a result, the fitted equation is expected to fit the data from which it was computed better than it will an independent set of data. In fact, the fitted equation quite likely will fit the sample data better than the *true* model would if it were known.

A fitted model should be validated for the specific objective for which it was planned. An equation that is good for predicting Y_i in a given region of the X-space might be a poor predictor in another region of the X-space, or for estimation of a mean change in Y for a given change in X even in the same region of the X-space. These criteria are of interest:

1. Does the fitted regression equation provide unbiased predictions of the quantities of interest?

2. Is the precision of the prediction good enough (the variance small enough) to accomplish the objective of the study?

Mean Squared Error of Prediction

Both quantities, bias and variance, are sometimes incorporated into a single measure called the **mean squared error of prediction** (MSEP). Mean square error of prediction is defined as the average squared difference between independent observations and predictions from the fitted equation for the corresponding values of the independent variables. The mean squared error of prediction incorporates both the variance of prediction and the square of the bias of the prediction.

Example 7.9

For illustration, suppose a model has been developed to predict maximum rate of runoff from watersheds following rain storms. The independent variables are rate of rainfall (inches per hour), acreage of watershed, average slope of land in the watershed, soil moisture levels, soil type, amount of urban development, and amount and type of vegetative cover. The dependent variable is maximum rate of runoff (ft^3 sec^{-1}), or peak flow. Assume the immediate interest in the model is prediction of peak flow for a particular

TABLE 7.4. *Observed rate of runoff, predicted rate of runoff, and prediction error for validation of water runoff model. Results are listed in increasing order of runoff (ft^3 sec^{-1}).*

	Predicted P	*Observed* Y	*Prediction Error* $\delta = P - Y$
	2,320	2,380	−60
	3,300	3,190	110
	3,290	3,270	20
	3,460	3,530	−70
	3,770	3,980	−210
	4,210	4,390	−180
	5,470	5,400	70
	5,510	5,770	−260
	6,120	6,890	−770
	6,780	8,320	−1,540
Mean	4,423	4,712	−289

watershed. The model is to be validated for this watershed by comparing observed rates of peak flow with the model predictions for 10 episodes of rain. The observed peak flow, the predicted peak flow, and the error of prediction are given in Table 7.4 for each episode. The average prediction bias is $\overline{\delta} = -289$ ft^3 sec^{-1}; the peak flow in these data is underestimated by approximately 6%. The variance of the prediction error is $s^2(\delta) = 255,477$, or $s(\delta) = 505$ ft^3 sec^{-1}. The standard error of the estimated mean bias is $s(\overline{\delta}) = 505/\sqrt{10} = 160$. A t-test of the hypothesis that the bias is zero gives $t = -1.81$, which, with 9 degrees of freedom and $\alpha = .05$, is not significant.

The mean squared error of prediction is

$$\text{MSEP} = \frac{\boldsymbol{\delta}'\boldsymbol{\delta}}{n} = 313,450$$

or

$$\begin{aligned}\text{MSEP} &= \frac{(n-1)s^2(\delta)}{n} + (\overline{\delta})^2 \\ &= \frac{9(255,477)}{10} + (-289)^2 = 313,450.\end{aligned}$$

The bias term contributes 27% of MSEP. The square root of MSEP gives 560 ft^3 sec^{-1}, an approximate 12% error in prediction.

Even though the average bias is not significantly different from zero, the very large prediction error on the largest peak flow (Table 7.4) suggests that the regression equation is not adequate for heavy rainfalls. Review of the data from which the equation was developed shows very few episodes

of rainfall as heavy as the last in the validation data set. If the last rainfall episode is omitted from the computations, the average bias drops to $\overline{\delta} = -150$ ft^3 sec^{-1} with a standard deviation of $s(\delta) = 265$, or a standard error of the mean of $s(\overline{\delta}) = 88.2$. Again, the average bias is not significantly different from zero using these nine episodes. However, the error of prediction on the largest rainfall differs from zero by -1540/265 = -5.8 standard deviations. This is a clear indication that the regression equation is seriously biased for the more intense rainfalls and must be modified before it can be used with confidence. ■

Choosing the Data Set for Validation

In Example 7.9, the peak flow model was being validated for a particular watershed. If the intended use of the model had been prediction of peak flow from several watersheds over a large geographical area, this sample of data would have been inadequate for validation of the model. Validation on one watershed would not have provided assurance that the equation would function well over a wide range of watersheds. The data to be used for validation of a model must represent the population for which the predictions are to be made.

Splitting the Data Set

It often is impractical to obtain an adequate independent data set with which to validate a model. If the existing data set is sufficiently large, an alternative is to use those data for both estimation and validation. One approach is to divide the data set into two representative halves; one half is then used to develop the regression model and the other half is used for validation of the model. Snee (1977) suggests that the total sample size should be greater than $2p' + 25$ before splitting the sample is considered. Of course, one could reverse the roles of the two data sets and have double estimation and validation. Presumably, after the validation, and assuming satisfactory results, one would prefer to combine the information from the two halves to obtain one model which would be better than either alone.

Estimating MSEP

Methods have been devised for estimating the mean squared error of prediction MSEP when it is not practical to obtain new independent data. The C_p statistic can be considered an estimator of MSEP. Weisberg (1981) presents a method of allocating C_p to the individual observations which facilitates detecting inadequacies in the model. Another approach is to measure the discrepancy between each observation and its prediction but where that observation was not used in the development of the prediction equation. The sum of squares of these discrepancies is the *PRESS* statistic given by Allen (1971b). Let $\widehat{Y}_{\text{pred}_{i(i)}}$ be the prediction of observation i, where the (i) indicates that the ith observation was not used in the development of the regression equation. Then,

$$PRESS = \sum_{i=1}^{n} (Y_i - \widehat{Y}_{\text{pred}_{i(i)}})^2. \tag{7.6}$$

The individual discrepancies are of particular interest for model validation. Unusually large discrepancies or patterns to the discrepancies can indicate inadequacies in the model. Bunke and Droge (1984) derive a best unbiased estimator and a minimum mean squared error estimator of MSEP where there is replication and all variables are assumed to have a multivariate normal distribution.

Validation of the model based on an independent sampling of the population is to be preferred to the use of estimates of mean squared error of prediction based on the original sample data. Part of the error of prediction may arise because the original data do not adequately represent the original population. Or, the population may have changed in some respects since the original sample was taken. Estimates of MSEP computed from the original data cannot detect these sources of inadequacies in the model.

7.7 Exercises

7.1. Show algebraically the relationship between R^2 and MS(Res).

7.2. Show algebraically the relationship between R^2 and C_p, and between MS(Res) and C_p.

7.3. Substitute expectations in the numerator and denominator of the C_p statistic and show that C_p is approximately an estimate of p' when the model is correct. (This is approximate because the ratio of expectations is not the same as the expectation of the ratio.)

7.4. Use the relationship between R^2 and MS(Res), Exercise 7.1, to show equality between the two forms of R^2_{adj} in equation 7.2.

7.5. The following approach was used to determine the effect of acid rain on agricultural production. U.S. Department of Agriculture statistics on crop production, fertilizer practices, insect control, fuel costs, land costs, equipment costs, labor costs, and so forth for each county in the geographical area of interest were paired with county-level estimates of average pH of rainfall for the year. A multiple regression analysis was run in which "production ($)" was used as the dependent variable and all input costs plus pH of rainfall were used as independent variables. A stepwise regression analysis was used with pH forced to be in all regressions. The partial regression coefficient on pH from the model chosen by stepwise regression was taken as the measure of the impact of acid rain on crop production.

(a) Discuss the validity of these data for establishing a causal relationship between acid rain and crop production.

(b) Suppose a causal effect of acid rain on crop production had already been established from other research. Discuss the use of the partial regression coefficient for pH from these data to predict the change in crop production that would result if rain acidity were to be decreased. Do you see any reason the prediction might not be valid?

(c) Suppose the regression coefficient for pH were significantly negative (higher pH predicts lower crop production). Do you see any problem with inferring that stricter government air pollution standards on industry would result in an increase in crop production?

(d) Do you see any potential for bias in the estimate of the partial regression coefficient for pH resulting from the omission of other variables?

7.6. The final model in the Linthurst example in this chapter used pH and Na content of the marsh substrate as the independent variables for predicting biomass (in the forward selection and stepwise methods). The regression equation was

$$\widehat{Y}_i = -476 + 407X_{pH} - .0233X_{Na}.$$

What inference are you willing to make about the relative importance of pH and Na versus $SALINITY$, K, and Zn as biologically important variables in determining biomass? When all five variables were in the model, the partial regression coefficient for pH was a nonsignificant $-.009(\pm.016)$. Does this result modify your inference?

Exercises 7.7 through 7.12 use the simulated data on peak flow of water used in the exercises in Chapter 5. Use $LQ = \ln(Q)$ as the dependent variable with the logarithms of the nine independent variables.

7.7. Determine the total number of possible models when there are nine independent variables, as in the peak water flow problem. Your computing resources may not permit computing all possible regressions. Use a program such as the METHOD = RSQUARE option in PROC REG (SAS Institute, Inc., 1989b) to find the $n = 6$ "best" subsets in each stage. This will require using the SELECT = n option. Plot the behavior of the C_p statistic and determine the "best" model.

7.8. Use a forward selection variable selection method to search for an acceptable model for the peak flow data. Use SLE = .50 for entry of a variable into the model. What subset would you select if you used SLE = .15? Compute and plot the C_p statistic for the models from SLE = .50. What subset model do you select for prediction using C_p?

7.9. Repeat Exercise 7.8 using backward elimination. Use SLS = .10 for elimination of a variable. What subset model is selected? Compute and plot the C_p statistic for the models used and decide on the "best" model. Does backward elimination give the same model as forward selection in Exercise 7.8?

7.10. Repeat Exercise 7.8 using the stepwise method of variable selection. Use SLE = .50 and SLS = .20 for elimination of a variable from the model. What subset model is selected? Plot the C_p statistic for the models used to decide which model to adopt. Do you arrive at the same model as with forward selection? As with backward elimination?

7.11. Give a complete summary of the results for the model you adopted from the backward elimination method in Exercise 7.9. Give the analysis of variance, the partial regression coefficients, their standard errors, and R^2.

7.12. Your analysis of the peak flow data has been done on $\ln(Q)$. Reexpress your final model on the original scale (by taking the antilogarithm of your equation). Does this equation make sense; that is, are the variables the ones you would expect to be important and do they enter the equation the way common sense would suggest? Are there omitted variables you would have thought important?

7.13. Consider the model

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon},$$

where $\boldsymbol{X}_1 : n \times p'$, and $\boldsymbol{X}_2 : n \times (t - p')$ and $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$. Suppose we estimate $\boldsymbol{\beta}_1$ and σ^2 using the subset model

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}.$$

That is,

$$\widehat{\beta}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{Y}$$

and

$$\widehat{\sigma}^2 = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{P}_{X_1})\boldsymbol{Y}/(n - p').$$

(a) Show that $\mathcal{E}(\widehat{\boldsymbol{\beta}}_1) = \boldsymbol{\beta}_1 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{\beta}_2$. Under what conditions is $\widehat{\boldsymbol{\beta}}_1$ unbiased for $\boldsymbol{\beta}_1$?

(b) Using the result for quadratic forms $\mathcal{E}[\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}] = \operatorname{tr}(\boldsymbol{A}\,\mathbf{Var}(\boldsymbol{Y})) + \mathcal{E}(\boldsymbol{Y}')\boldsymbol{A}\,\mathcal{E}(\boldsymbol{Y})$, show that

$$\begin{aligned}\mathcal{E}[\widehat{\sigma}^2] &= \sigma^2 + \boldsymbol{\beta}_2'\boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{P}_{X_1})\boldsymbol{X}_2\boldsymbol{\beta}_2/(n - p') \\ &\geq \sigma^2.\end{aligned}$$

Under what conditions is $\widehat{\sigma}^2$ unbiased for σ^2?

(c) Let $\boldsymbol{X} = (\,\boldsymbol{X}_1 \quad \boldsymbol{X}_2\,)$ be of full column rank. Show that

$$\begin{aligned}(\boldsymbol{X}'\boldsymbol{X})^{-1} &= \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix}^{-1} \\ &= \begin{bmatrix} (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} + \boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}' & -\boldsymbol{A}\boldsymbol{Q}^{-1} \\ -\boldsymbol{Q}^{-1}\boldsymbol{A}' & \boldsymbol{Q}^{-1} \end{bmatrix},\end{aligned}$$

where $\boldsymbol{A} = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2$ and $\boldsymbol{Q} = \boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{P}_{X_1})\boldsymbol{X}_2$.

(d) Using (c), show that the least squares estimators of the elements in β_1, based on the subset model in (a), have smaller variance than the corresponding estimators of the full model.

8
POLYNOMIAL REGRESSION

To this point we have assumed that the relationship between the dependent variable Y and any independent variable X can be represented with a straight line. This clearly is inadequate in many cases. This chapter introduces the extensively used polynomial and trigonometric regression response models to characterize curvilinear relationships. Such models are linear in the parameters and linear least squares is appropriate for estimation of the parameters. Models that are nonlinear in the parameters are introduced in Chapter 15.

Most models previously considered have (1) specified a linear relationship between the dependent variable and each independent variable and (2) have been linear in the parameters. The linear relationship results from each independent variable appearing only to the first degree and in only one term of the model; no terms are included that contain powers or products of independent variables. This restriction forces the rate of change in the mean of the dependent variable with respect to an independent variable to be constant over all values of that and every other independent variable in the model. Linearity in the parameters means that each (additive) term in the model contains only one parameter *and* only as a multiplicative constant on the independent variable. This restriction excludes many useful mathematical forms including nearly all models developed from principles of behavior of the system. These simple models are very restrictive and should be viewed as first-order approximations to true relationships.

In this chapter, the class of models is extended to allow greater flexibility and realism by introducing the higher-degree polynomial models and trigonometric models. These models still are to be regarded as approximations to the true models for most situations. Even more realistic models that are nonlinear in the parameters are introduced in Chapter 15. Although this chapter does not dwell on the behavior of the residuals, it is important that the assumptions of least squares be continually checked. Growth data, for example, often will not satisfy the homogeneous variance assumption, and will contain correlated errors if the data are collected as repeated measurements over time on the same experimental units. For discussion on experimental designs for fitting response surfaces and for estimating the values (settings) of the independent variables that optimize the response, the reader is referred to design texts such as Box, Hunter, and Hunter (1978).

8.1 Polynomials in One Variable

An assumed linear relationship between a dependent (response) variable and an independent (input) variable implies a constant rate of change and may not represent the true relationship adequately. For example, the concentration of a drug in the blood stream may not be linear over time. Many economic time series such as the inflation index and the gross domestic product exhibit trends over time that may not be linear. Although the time to bake a cake may decrease as the temperature of the oven increases, it may not decrease linearly. In all of these examples, the rate of change in the mean of the dependent variable (Y) is not constant with respect to the independent variable (X).

Quadratic Model

The simplest extension of the straight-line model involving one independent variable is the second-order polynomial (quadratic) model,

$$\mathcal{E}(Y) = \beta_0 + \beta_1 X + \beta_2 X^2. \tag{8.1}$$

The quadratic model includes the term X^2 in addition to X. Note that this model is a special case of the multiple regression model where $X_1 = X$ and $X_2 = X^2$. Hence, the estimation methods considered in Chapter 4 are appropriate. Higher-order polynomials of the form

Polynomial Model

$$\mathcal{E}(Y) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_p X^p \tag{8.2}$$

allow increasing flexibility of the response relationship and are also special cases of the multiple regression models where $X_i = X^i, i = 1, \ldots, p$. The model in equation 8.2 is called a pth degree polynomial model.

An important aspect of the polynomial model that distinguishes it from other multiple regression models is that the mean of the dependent variable

TABLE 8.1. *Algae density measures over time.*

Day	*Rep* 1	*Rep* 2	*Day*	*Rep* 1	*Rep* 2
1	.530	.184	8	4.059	3.892
2	1.183	.664	9	4.349	4.367
3	1.603	1.553	10	4.699	4.551
4	1.994	1.910	11	4.983	4.656
5	2.708	2.585	12	5.100	4.754
6	3.006	3.009	13	5.288	4.842
7	3.867	3.403	14	5.374	4.969

is a function of a single independent variable. Even though the independent variables in a general multiple regression model may be related to each other, typically they are not assumed to be functions of one another. The fact that the "independent" variables in a simple polynomial model are functions of a single independent variable affects the interpretation of the parameters. Consider, for example, the model

$$\mathcal{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \tag{8.3}$$

In this model, β_1 is interpreted as the change in the mean of the dependent variable per unit change in X_1 at any fixed value of X_2. (Likewise, β_2 is the change in the mean of the dependent variable per unit change in X_2 at any fixed value of X_1.) However, if $X_2 = X_1^2$, then changing X_1 by a unit will also change the value of X_2 . In the second-degree model, equation 8.1, the rate of change in the mean of the dependent variable as a function of X is called the *slope* at X or the *derivative* at X. From calculus, the derivative for equation 8.1 with respect to X is

$$\frac{d\mathcal{E}(Y)}{dX} = \beta_1 + 2\beta_2 X. \tag{8.4}$$

That is, the slope of $\mathcal{E}(Y)$ depends on the value of the independent variable. The parameter β_1 is the slope only at $X = 0$. The parameter β_2 is half the velocity of change in $\mathcal{E}(Y)$ or, equivalently, it is half the rate of change in the slope of $\mathcal{E}(Y)$.

Note that any polynomial model in one variable can be represented by a curvilinear plot on a two-dimensional graph, rather than a surface in higher-dimensional space, since the dependent variable is considered as a function of a single independent variable.

Example 8.1

The data in Table 8.1 are from a growth experiment with blue-green algae *Spirulina platensis* conducted by Linda Shurtleff, North Carolina State University (data used with permission). The complete data are presented in

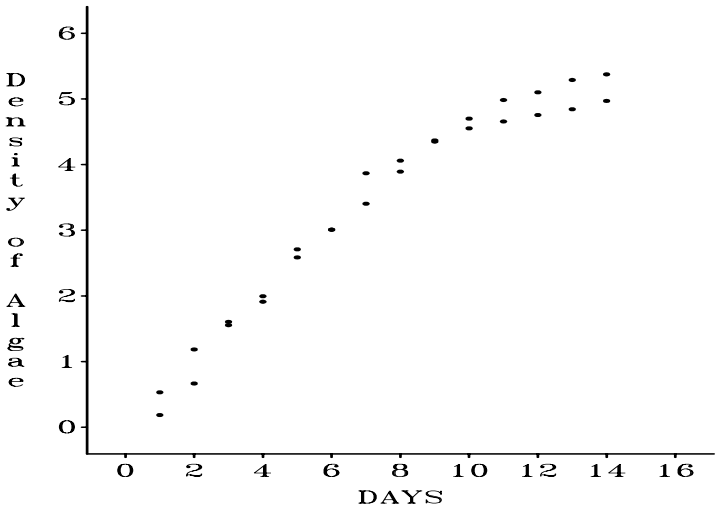

FIGURE 8.1. *Algae density versus days of study.*

Exercise 8.8. The data in Table 8.1 are for the treatment where CO_2 is bubbled through the culture. There were two replicates for this treatment, each consisting of 14 independent solutions. The 14 solutions in each replicate were randomly assigned for measurement to one of each of 14 successive days of study. The dependent variable reported is a log-scale measurement of the increased absorbance of light by the solution. This is interpreted as a measure of algae density. The plot of the algae density measurement versus days (Figure 8.1) clearly shows a curvilinear relationship. ■

Fitting Polynomials

Since polynomial response models are a special subset of multiple regression models, fitting polynomial models with least squares does not introduce any new conceptual problems. As long as the usual assumptions on the errors are appropriate, ordinary least squares can be used. The higher-degree terms are included in the model by augmenting $\boldsymbol{X}$ with columns of new variables defined as the appropriate powers of the independent variables. Testing procedures discussed for the multiple regression model are also appropriate for testing relevant hypotheses.

Example 8.2

Consider the data for the first replicate given in Example 8.1. We consider a cubic polynomial model given by

$$Y_{i1} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \epsilon_{i1}, \tag{8.5}$$

where $X_i = i$ represents the day and Y_{i1} represents the response variable for the first replicate on day i. Note that the model in equation 8.5 can be expressed as a multiple regression model given by

$$\begin{pmatrix} .530 \\ 1.183 \\ 1.603 \\ \vdots \\ 5.374 \end{pmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 14 & 196 & 2,744 \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_{14} \end{pmatrix} \tag{8.6}$$

or

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

The ordinary least squares fit of the model is given by

$$\begin{array}{rcl} \widehat{Y}_i & = & .00948 + .53074X_i + 0.00595X_i^2 - .00119X_i^3 \\ & & (.16761)\ \ (.09343) \qquad (.01422) \qquad (.00062), \end{array} \tag{8.7}$$

where the standard errors of the estimates are given in parentheses.

Assuming that a cubic model is adequate, we can test the hypotheses $H_0 : \beta_3 = 0$ and $H_0 : \beta_2 = \beta_3 = 0$. Given a cubic polynomial model, $H_0 : \beta_3 = 0$ tests the hypothesis that a quadratic polynomial model is adequate whereas $H_0 : \beta_2 = \beta_3 = 0$ tests the hypothesis that a linear trend model is adequate. The t-statistic for testing $\beta_3 = 0$ is

$$t = \frac{-.00119}{.00062} = -1.91.$$

Comparing $|t| = 1.91$ with $t_{(.025;10)} = 2.228$, we fail to reject $H_0 : \beta_3 = 0$.

To test $H_0 : \beta_2 = \beta_3 = 0$, we fit the reduced model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

and compute the F-statistic

$$\begin{array}{rcl} F & = & \dfrac{[\mathrm{SS}(\mathrm{Res}_{\mathrm{reduced}}) - \mathrm{SS}(\mathrm{Res}_{\mathrm{full}})]/2}{\mathrm{SS}(\mathrm{Res}_{\mathrm{full}})/10} \\ & = & \dfrac{[1.45812 - .13658]/2}{.01366} \\ & = & 48.37. \end{array}$$

Since $F_{(.05;2,10)} = 4.10$, we reject $H_0 : \beta_2 = \beta_3 = 0$. That is, we conclude that a linear trend model is not adequate. ■

It is interesting to note that the t-statistic for testing $H_0 : \beta_2 = 0$ in Example 8.2 is $t = .418$ and we would fail to reject $H_0 : \beta_2 = 0$. That is, we

fail to reject the *individual* null hypotheses $H_0 : \beta_2 = 0$ *and* $H_0 : \beta_3 = 0$, but we reject the *joint* null hypothesis $H_0 : \beta_2 = \beta_3 = 0$. This is due to the fact that $X_3 = X^3$ is highly correlated with the linear and quadratic variables. When the columns of an $\boldsymbol{X}$ matrix are nearly linearly dependent on each other, the matrix $\boldsymbol{X'X}$ is nearly singular and, hence, the matrix $\mathbf{Var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X'X})^{-1}\sigma^2$ tends to have large elements. That is, the standard errors of the least squares estimators will be large and the corresponding t-statistics will be small. This problem is known as the multicollinearity problem. This and other related problems are discussed in Chapter 10.

Testing Model Adequacy

Since polynomial models are special cases of multiple linear regression, diagnostics based on the residuals can be used to check the adequacy of the model. Another approach is to fit a higher-order polynomial that is deemed adequate and use statistical tests to obtain a low-order polynomial that is adequate. For example, in Example 8.2, we assume that a cubic polynomial model is adequate and test sequentially whether a quadratic polynomial model or a linear trend model is adequate. When one measurement is observed at each of k distinct values of the input variable, then it is possible to fit a polynomial of degree $(k-1)$. However, in this case, the $(k-1)$th degree polynomial will fit the k observations perfectly and the residual sum of squares will be zero. Therefore, in testing the adequacy of a polynomial model, it is important to choose a high, but not too high, order polynomial model.

Lack of Fit

When replicate measurements are observed at at least one of the values of the independent variable, an alternative test for the adequacy of the model can be used. Suppose we have n_i replicate measuements at X_i, for $i = 1, \ldots, k$. Assume that the X_is are distinct, $n_i \geq 1$, and at least one of the n_i is strictly greater than 1. In this case, we can fit a $(k-1)$th degree polynomial and the error sum of squares will have $\sum n_i - k$ degrees of freedom. Using the $(k-1)$th degree polynomial as the full model, we can test the adequacy of a low-order polynomial model. Let Y_{ij} denote the jth replicate value of the response variable at the ith value (X_i) of the independent variable. We wish to test the adequacy of the degree $q(< k)$ polynomial model

$$Y_{ij} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_q X_i^q + \epsilon_{ij}. \tag{8.8}$$

We first fit the full model

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_q X_i^q \\ &\quad + \beta_{q+1} X_i^{q+1} + \cdots + \beta_{k-1} X_i^{k-1} + \epsilon_{ij} \end{aligned} \tag{8.9}$$

and then fit the reduced model in equation 8.8. We use the F-statistic for testing $H_0 : \beta_{q+1} = \cdots = \beta_{k-1} = 0$ to test the adequacy of the model in equation 8.8. That is, we use the F-statistic

$$F = \frac{[\mathrm{SS}(\mathrm{Res}_{\mathrm{reduced}}) - \mathrm{SS}(\mathrm{Res}_{\mathrm{full}})]/(k-1-q)}{\mathrm{SS}(\mathrm{Res}_{\mathrm{full}})/(\sum n_i - k)} \tag{8.10}$$

$$= \frac{[\text{Lack-of-Fit Sum of Squares}]/(k-1-q)}{[\text{Pure Error Sum of Squares}]/(\sum n_i - k)}. \tag{8.11}$$

We show in Chapter 9 that

$$\begin{aligned} \text{Pure Error Sum of Squares} &= \text{SS}(\text{Res}_{\text{full}}) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2. \end{aligned}$$

The adequacy of the model in equation 8.8 is rejected if F is larger than $F_{(\alpha;k-1-q,\sum n_i-k)}$.

Example 8.3

In Example 8.1, we have two replicates each day. That is, we have $k = 14$ and $n_i = 2$ for $i = 1, \ldots, 14$. To test the adequacy of a quadratic polynomial model, we fit the model

$$Y_{ij} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_{ij} \tag{8.12}$$

to obtain

$$\text{SS}(\text{Res}_{\text{reduced}}) = .7984.$$

The pure-error sum of squares is

$$\begin{aligned} \text{Pure-Error Sum of Squares} &= \sum_{i=1}^{14}\sum_{j=1}^{2}(Y_{ij} - \overline{Y}_{i.})^2 \\ &= .6344 \end{aligned}$$

and hence the lack-of-fit sum of squares is

$$\begin{aligned} \text{Lack-of-Fit Sum of Squares} &= .7984 - .6344 \\ &= .1640. \end{aligned}$$

The value of the F-statistic for testing the adequacy of the quadratic polynomial model (equation 8.12) is

$$F = \frac{.1640/(14-1-2)}{.6344/(28-14)} = .329.$$

Since $F_{(.05;11,14)} = 2.57$, we fail to reject the null hypothesis that the quadratic model (equation 8.12) is adequate. Also, from Figure 8.2, we observe that the quadratic polynomial model fits the data reasonably well. Figure 8.2 also shows the fit from the full model (a 13th degree polynomial). Even though the full model has smaller residual sum of squares, we observe that the fitted curve has a considerable number of wild oscillations. These

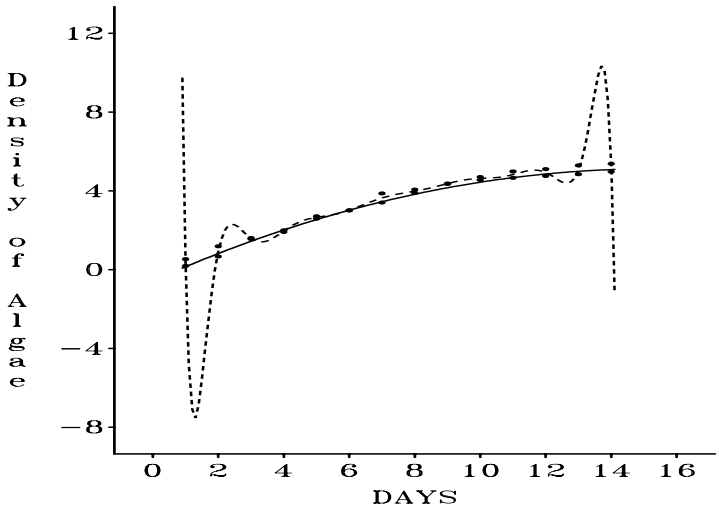

FIGURE 8.2. *Algae density data with the fitted quadratic model (solid line) and fitted 13th degree polynomial model.*

fits also indicate that care must be used when interpolating or extrapolating based on high-order polynomial models. Issues related to extrapolation are discussed further in Section 8.3.2. ■

Orthogonal Polynomials

In Example 8.2, we have observed that the "natural" polynomials X_i, X_i^2, and X_i^3 are nearly linearly dependent on each other. Such relationships among the columns of the $\boldsymbol{X}$ matrix lead to multicollinearity problems. The collinearity problems and diagnostics are discussed in Sections 10.3 and 11.3. When columns are not orthogonal to each other, the sequential and partial sums of squares of the coefficients will be different. On the other hand, if the columns are orthogonal, the sequential sums of squares equal the partial sums of squares.

Consider the cubic polynomial model in equation 8.5 given by

$$Y_{i1} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \epsilon_{i1}, \ i = 1, \ldots, 14, \tag{8.13}$$

where $X_i = i$. In this case, the sequential sums of squares $R(\beta_1|\beta_0)$ and $R(\beta_2|\beta_1\ \beta_0)$ based on the "natural" polynomials are different from the partial sums of squares $R(\beta_1|\beta_0\ \beta_2\ \beta_3)$ and $R(\beta_2|\beta_0\ \beta_1\ \beta_3)$. Define a set of orthogonal polynomials

$$\begin{aligned} O_{0i} &= 1, \\ O_{1i} &= 2X_i - 15, \end{aligned}$$

$$\begin{aligned} O_{2i} &= .5X_i^2 - 7.5X_i + 20, \quad \text{and} \\ O_{3i} &= \frac{5}{3}X_i^3 - 37.5X_i^2 + \frac{698.5}{3}X_i - 340. \end{aligned} \tag{8.14}$$

Note that O_{1i}, O_{2i}, and O_{3i} are linear combinations of the "natural" polynomials X_i, X_i^2, and X_i^3. Arranging the values of the orthogonal polynomials ($i = 1, \ldots, 14$) from equation 8.14 in a (14×4) matrix gives

$$\begin{aligned} \boldsymbol{O} &= (\boldsymbol{O}_0 \quad \boldsymbol{O}_1 \quad \boldsymbol{O}_2 \quad \boldsymbol{O}_3) \\ &= \begin{bmatrix} 1 & -13 & 13 & -143 \\ 1 & -11 & 7 & -11 \\ 1 & -9 & 2 & 66 \\ 1 & -7 & -2 & 98 \\ 1 & -5 & -5 & 95 \\ 1 & -3 & -7 & 67 \\ 1 & -1 & -8 & 24 \\ 1 & 1 & -8 & -24 \\ 1 & 3 & -7 & -67 \\ 1 & 5 & -5 & -95 \\ 1 & 7 & -2 & -98 \\ 1 & 9 & 2 & -66 \\ 1 & 11 & 7 & 11 \\ 1 & 13 & 13 & 143 \end{bmatrix}. \end{aligned} \tag{8.15}$$

Note that the columns O_0, O_1, O_2, and O_3 in the matrix $\boldsymbol{O}$ are mutually orthogonal. When the values of X_i are equally spaced, orthogonal polynomials may be obtained from tables given in Steel, Torrie, and Dickey (1997). Regardless of whether X_is are equally spaced, the orthogonal polynomials can be obtained using the Gram–Schmidt orthogonalization procedure (see Exercise 2.27) or by a computing program such as the ORPOL function in PROC IML of SAS (SAS Institute Inc., 1989d).

Given X_i, X_i^2, and X_i^3, we can obtain O_{1i}, O_{2i}, and O_{3i} as linear functions of X_i, X_i^2, and X_i^3 (equation 8.14). Also, given O_{1i}, O_{2i}, and O_{3i}, we can get back to X_i, X_i^2, and X_i^3, using

$$\begin{aligned} X_i &= 7.5 + .5\,O_{1i}, \\ X_i^2 &= 72.5 + 7.5\,O_{1i} + 2\,O_{2i}, \quad \text{and} \\ X_i^3 &= 787.5 + 98.9\,O_{1i} + 45\,O_{2i} + .6\,O_{3i}. \end{aligned} \tag{8.16}$$

Note that, from equations 8.13 and 8.16, we get

$$\begin{aligned} Y_{i1} &= \beta_0 + \beta_1(7.5 + .5\,O_{1i}) + \beta_2(72.5 + 7.5\,O_{1i} + 2\,O_{2i}) \\ &\quad + \beta_3(787.5 + 98.9\,O_{1i} + 45\,O_{2i} + .6\,O_{3i}) + \epsilon_{i1} \\ &= \gamma_0 + \gamma_1\,O_{1i} + \gamma_2\,O_{2i} + \gamma_3\,O_{3i} + \epsilon_{i1}, \end{aligned} \tag{8.17}$$

where

$$\begin{aligned}
\gamma_0 &= \beta_0 + 7.5\,\beta_1 + 72.5\,\beta_2 + 787.5\,\beta_3, \\
\gamma_1 &= .5\,\beta_1 + 7.5\,\beta_2 + 98.9\,\beta_3, \\
\gamma_2 &= 2\,\beta_2 + 45\,\beta_3, \quad \text{and} \\
\gamma_3 &= .6\,\beta_3.
\end{aligned} \tag{8.18}$$

That is, the model in equation 8.17 is a reparameterization of the model in equation 8.13. Similarly, using equations 8.14 and 8.17, or by solving equation 8.18 for the βs, we get

$$\begin{aligned}
\beta_0 &= \gamma_0 - 15\,\gamma_1 + 20\,\gamma_2 - 340\,\gamma_3, \\
\beta_1 &= 2\,\gamma_1 - 7.5\,\gamma_2 + \frac{698.5}{3}\,\gamma_3, \\
\beta_2 &= .5\,\gamma_2 - 37.5\,\gamma_3, \quad \text{and} \\
\beta_3 &= \frac{5}{3}\,\gamma_3.
\end{aligned} \tag{8.19}$$

That is, the model using Xs, equation 8.13, is equivalent to the model using the orthogonal polynomials, equation 8.17. One of the advantages of working with orthogonal polynomials is that the columns corresponding to O_{1i}, O_{2i}, and O_{3i} are mutually orthogonal and hence avoid numerical problems associated with the near-singularity. Also, the sequential and partial sums of squares coincide for the model in equation 8.17. Note also that $\beta_3 = 0$ if and only if $\gamma_3 = 0$ and $\beta_2 = \beta_3 = 0$ if and only if $\gamma_2 = \gamma_3 = 0$. Hence, testing $H_0 : \beta_3 = 0$ and $H_0 : \beta_2 = \beta_3 = 0$ in equation 8.13 is equivalent to testing $H_0 : \gamma_3 = 0$ and $H_0 : \gamma_2 = \gamma_3 = 0$, respectively, in equation 8.17.

Example 8.4

For the data in Example 8.2, we get

$$\begin{aligned}
\widehat{Y}_{i1} &= 3.48164 + .19198\,O_{1i} - .04179\,O_{2i} - .00072\,O_{3i} \\
&\quad\ (.03123) \quad (.00387) \quad\ (.00433) \quad\ (.00037).
\end{aligned}$$

Note that the t-statistic for testing $H_0 : \gamma_3 = 0$ in equation 8.17 is

$$t = \frac{-.00072}{.00037} = -1.91.$$

This is the same as the t-statistic for testing $H_0 : \beta_3 = 0$ in equation 8.13 (Example 8.2). Similarly, the F-statistic for testing $H_0 : \gamma_2 = \gamma_3 = 0$ in equation 8.13 is the same as the F-statistic we have computed for testing $H_0 : \beta_2 = \beta_3 = 0$ in Example 8.2. However, the t-statistic (-9.649) for testing $H_0 : \gamma_2 = 0$ is *not* the same as the t-statistic (.418) for testing $H_0 : \beta_2 = 0$. Using equation 8.18, a test of $H_0 : \gamma_2 = 0$ would be the same as a test of $H_0 : 2\beta_2 + 45\beta_3 = 0$. ■

TABLE 8.2. *Quarterly U. S. beer production from the first quarter of 1975 to the fourth quarter of 1982 (millions of barrels).*

	Quarter			
Year	*I*	*II*	*III*	*IV*
1975	36.14	44.60	44.15	35.72
1976	36.19	44.63	46.95	36.90
1977	39.66	49.72	44.49	36.54
1978	41.44	49.07	48.98	39.59
1979	44.29	50.09	48.42	41.39
1980	46.11	53.44	53.00	42.52
1981	44.61	55.18	52.24	41.66
1982	47.84	54.27	52.31	42.03

8.2 Trigonometric Regression Models

Measurements on a response variable (Y_t) collected over time (t), as in Example 8.3, are called time series data. Although not present in Example 8.3, such data often display periodic behavior that repeats itself every s time periods. For example, the average monthly temperatures in Raleigh may exhibit a periodic behavior that is expected to repeat itself over the years. That is, the average temperature value for January in one year is expected to be similar to January values in other years, the February value in one year is expected to be similar to February values in other years, and so forth for each month. Economic time series often exhibit periodic behavior that reflects business cycles. For example, total monthly sales of greeting cards is expected to be periodic over the years as are total monthly retail sales and housing starts. Trigonometric functions such as $\sin(\omega t)$ and $\cos(\omega t)$ are periodic over time with a **period** of $2\pi/\omega$. That is, $\sin(\omega t)$ is the same as $\sin[\omega(t + (2\pi/\omega)j)]$ for $j = 1, 2, \ldots$. Hence, time series with periodic behavior may be modeled parsimoniously using **trigonometric regression models**.

Consider, for example, quarterly U. S. beer production from the first quarter of 1975 to the fourth quarter of 1982 (Table 8.2 and Figure 8.3). We see that the behavior of the production is periodic and it is repeated over the years. Production tends to be highest in the second quarter and lowest in the fourth quarter of each year. A trigonometric regression model that may be appropriate for these data is given by

$$Y_t = \beta_0 + \beta_1 \cos(2\pi t/4) + \beta_2 \sin(2\pi t/4) + \beta_3 \cos(\pi t) + \epsilon_t. \quad (8.20)$$

The cosine and sine terms appear in pairs. The term $\sin(\pi t)$ is not included since it is identically zero in this case. The intercept column may also be

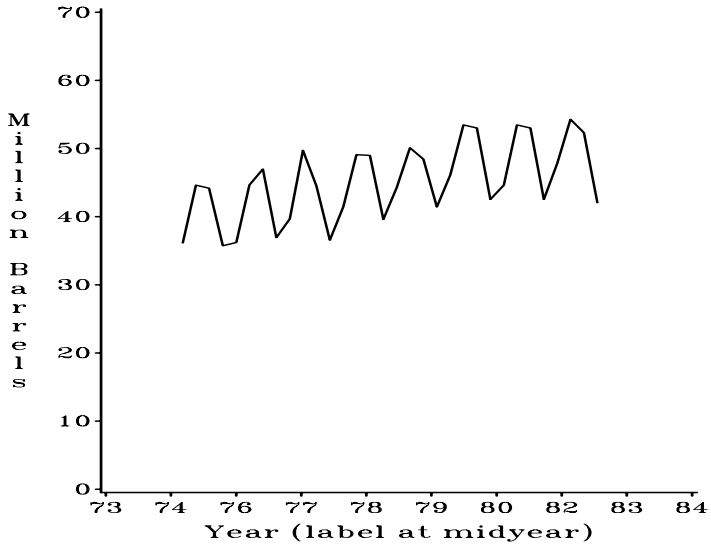

FIGURE 8.3. *Quarterly U.S. beer production versus time.*

thought of as the $\cos(0t)$ term. Note that

$$\cos(2\pi t/4) \quad = \quad \cos(2\pi(t+4j)/4) = \cos(2\pi t/4 + 2\pi j)$$

and

$$\sin(2\pi t/4) \quad = \quad \sin(2\pi(t+4j)/4) = \sin(2\pi t/4 + 2\pi j)$$

for any integer j. That is, $\cos(2\pi t/4)$ and $\sin(2\pi t/4)$ are periodic with a period of 4. They take the same value every 4 quarters. On the other hand,

$$\cos(\pi t) = \cos(\pi(t+2j)) = \cos(2\pi t/2 + 2\pi j)$$

for any integer j and, hence, it has a period of 2. That is, it takes the same value, 1 or -1, every 2 quarters. Note that this model (equation 8.20) is a special case of the multiple regression model with $X_{t1} = \cos(2\pi t/4)$, $X_{t2} = \sin(2\pi t/4)$, and $X_{t3} = \cos(\pi t)$.

The X-matrix for this model (equation 8.20) is given by

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 1 & -1 \\ 1 & -1 & 0 & 1 \\ 1 & 0 & -1 & -1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & -1 \\ 1 & -1 & 0 & 1 \\ 1 & 0 & -1 & -1 \\ 1 & 1 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & -1 \\ 1 & -1 & 0 & 1 \\ 1 & 0 & -1 & -1 \\ 1 & 1 & 0 & 1 \end{bmatrix}. \tag{8.21}$$

Note that the columns of $\boldsymbol{X}$ in equation 8.21 are mutually orthogonal and the $\boldsymbol{X}'\boldsymbol{X}$ matrix is given by

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 32 & 0 & 0 & 0 \\ 0 & 16 & 0 & 0 \\ 0 & 0 & 16 & 0 \\ 0 & 0 & 0 & 32 \end{bmatrix}.$$

In addition to the periodic behavior, Figure 8.3 shows an increasing trend in beer production over time. A more appropriate model would account for a time trend by including the term δt in the trigonometric model, equation 8.20, where δ is the linear regression coefficient for the average change in beer production per year. In this case, $\boldsymbol{X}'\boldsymbol{X}$ is no longer a diagonal matrix.

For monthly data like the average temperatures or average river flow measures that exhibit periodic behavior every 12 months, a model of the form

$$Y_t = a_0 + \sum_{j=1}^{5} [a_j \cos(2\pi jt/12) + b_j \sin(2\pi jt/12)] + a_6 \cos(\pi t) + \epsilon_t \tag{8.22}$$

may be appropriate. The trigonometric functions

$$\cos(2\pi jt/12) \qquad \text{and} \qquad \sin(2\pi jt/12),\ j = 1, \ldots, 6,$$

are periodic with a period of $12/j$. That is, they have the same value every $12/j$ months. As in the beer production example, the cosine and sine terms appear as pairs at each frequency. An interpretation of the coefficients a_j and b_j in terms of the phase angle of the trend and the period is given in Anderson (1971).

The trigonometric regression model in equation 8.22 is also a special case of the multiple linear regression model. Suppose we have data on the average monthly temperatures for the period January 1987 through December 1996. Then the $\boldsymbol{X}'\boldsymbol{X}$ matrix for the model in equation 8.22 is a 12×12 diagonal matrix with diagonal elements 120, 60, 60,..., 60, 120. That is, the columns of the X-matrix are mutually orthogonal. This orthogonality stems from the fact that the data cover complete cycles of the anticipated periodicity. If our data had included the averages for January 1997 through May 1997, a partial cycle, the columns of the X-matrix would no longer be orthogonal. Orthogonality of the columns of the X-matrix makes it simple to obtain the least squares estimators of the parameters. For this model (equation 8.22), with 10 years of data, the least squares estimators of a_j and b_j are given by

$$\begin{aligned}
\widehat{a}_0 &= \frac{1}{120}\sum_{t=1}^{120} Y_t = \overline{Y}, \\
\widehat{a}_j &= \frac{1}{60}\sum_{t=1}^{120} \cos(2\pi jt/12) Y_t, \ j = 1, \ldots, 5, \\
\widehat{b}_j &= \frac{1}{60}\sum_{t=1}^{120} \sin(2\pi jt/12) Y_t, \ j = 1, \ldots, 5, \quad \text{and} \\
\widehat{a}_6 &= \frac{1}{120}\sum_{t=1}^{120} \cos(\pi t) Y_t.
\end{aligned} \tag{8.23}$$

The residual mean square error for this model (equation 8.22) is

$$\widehat{\sigma}^2 = \frac{\sum_{t=1}^{120} Y_t^2 - 120\widehat{a}_0^2 - 60\left[\sum_{j=1}^{5}(\widehat{a}_j^2 + \widehat{b}_j^2)\right] - 120\widehat{a}_6^2}{120 - 12}, \tag{8.24}$$

where $\sum Y_t^2 - 120\widehat{a}_0^2 = \sum Y_t^2 - n\overline{Y}^2$ is seen to be the corrected total sum of squares.

As in the case of multiple regression models, t- and F-statistics can be used to test hypotheses regarding the significance of certain parameters. For example, to test the hypothesis $H_0 : a_6 = 0$, we use the t-statistic

$$t = \frac{\widehat{a}_6}{\sqrt{\widehat{\sigma}^2/120}} \tag{8.25}$$

and reject $H_0 : a_6 = 0$ if $|t| > t_{(\alpha/2;108)}$. Similarly, to test the null hypothesis $H_0 : a_5 = b_5 = 0$ (that is, no periodic component of period 12/5 months), we use the F-statistic

$$F = \frac{60\left[\widehat{a}_5^2 + \widehat{b}_5^2\right]/2}{\widehat{\sigma}^2} \tag{8.26}$$

and reject $H_0 : a_5 = b_5 = 0$ if $F > F_{(\alpha;2,108)}$. In trigonometric regression models, it is appropriate to test $a_j = b_j = 0$ simultaneously, since, as a pair, they correspond to a periodic component of period $12/j$ months.

The assumption that the errors ϵ_t in equation 8.22 are independent over time may not be realistic for time series data. For example, the temperatures in different months may be correlated with each other. If the errors are correlated, the ordinary least squares estimators may not be efficient. Also, the standard errors and the test statistics constructed under the assumption of independent errors may not be valid when the errors are correlated. We discuss in Chapter 10 appropriate methods when the assumptions are violated.

8.3 Response Curve Modeling

8.3.1 Considerations in Specifying the Functional Form

Regression to Summarize Data

The degree of realism that needs to be incorporated into a model will depend on the purpose of the regression analysis. The least demanding purpose is the simple use of a regression model to *summarize* the observed relationships in a particular set of data. There is no interest in the functional form of the model per se or in predictions to other sets of data or situations. The most demanding is the more esoteric development of mathematical models to describe the physical, chemical, and biological processes in the system. The goal of the latter is to make the model as realistic as the state of knowledge will permit.

The use of regression models simply to summarize observed relationships places no priority on realism because no inference, even to other samples, is intended. The overriding concern is that the model adequately portray the observed relationships. In practice, however, readers will often attach a predictive inference to the presentation of regression results, even if the intent of the author is simply to summarize the data.

Regression for Prediction

When the regression equation is to be used for prediction, it is beneficial to incorporate into the model prior information on the behavior of the system. This serves certain goals. First, other things being equal, the more realistic model would be expected to provide better predictions for unobserved points in the X-space, either interpolations or extrapolations. Although extrapolations are always dangerous and are to be avoided, it is not always easy, particularly with observational data, to identify points outside the sample space. Realistic models will tend to provide more protection against large errors in unintentional extrapolations than purely approximating models. Second, incorporating current beliefs about the behavior of the system into the model provides an opportunity to test and update these theories.

Use of Prior Information

The prior information used in the model may be nothing more than recognizing the general shape the response curve should take. For example, it may be that the response variable should not take negative values, or the response should approach an asymptote for high or low values of an independent variable. Recognizing such constraints on the behavior of the system will often lead to the use of nonlinear models. In some cases, these (presumably) more realistic models will also be simpler models in terms of the number of parameters to be estimated. A response with a plateau, for example, may require several terms of a polynomial model to fit the plateau, but might be characterized very well with a two-parameter exponential model. Polynomial models should not a priori be considered the simpler and nonlinear models the more complex. Models that are nonlinear in the parameters are discussed in Chapter 15.

At the other extreme, prior information on the behavior of a system may include minute details on the physical and chemical interactions in each of several different components of the system and on how these components interact to produce the final product. Such models can become extremely complex and most likely cannot be written as a single functional relationship between $\mathcal{E}(Y)$ and the independent variables. Numerical integration may be required to evaluate and combine the effects of the different components. The detailed crop growth models that predict crop yields based on daily, or even hourly, data on the environmental and cultural conditions during the growing season are examples of such models. The development of such models is not pursued in this text. They are mentioned here as an indication of the natural progression of the use of prior information in model building.

8.3.2 Polynomial Response Models

The models previously considered have been first-degree polynomial models, models in which each term contains only one independent variable to the first power. The first-degree polynomial model in two variables is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i. \tag{8.27}$$

A second-degree polynomial model includes terms, in addition to the first-degree terms, that contain squares or products of the independent variables. The full second-degree polynomial model in two variables is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{11} X_{i1}^2 + \beta_{12} X_{i1} X_{i2} + \beta_{22} X_{i2}^2 + \epsilon_i. \tag{8.28}$$

Degree of a Polynomial

The **degree** (or **order**) of an individual term in a polynomial is defined as the *sum* of the powers of the independent variables in the term. The degree of the entire polynomial is defined as the degree of the highest-degree term. All polynomial models, regardless of their degree, are linear in the parameters. For the higher-degree polynomial models, the subscript

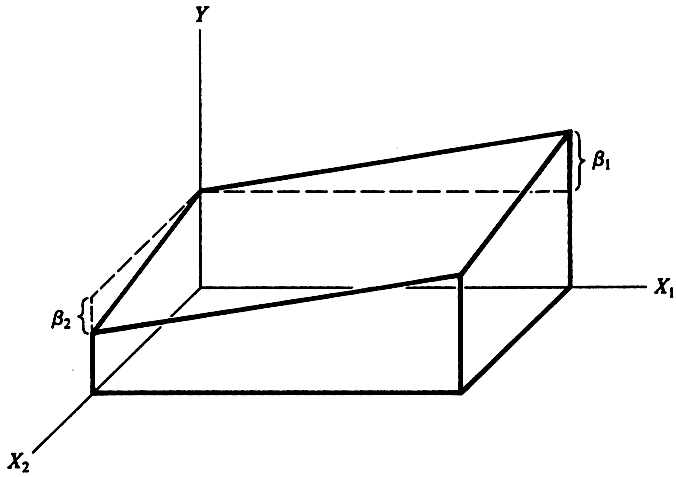

FIGURE 8.4. *A first-degree bivariate polynomial response surface.*

notation on the βs is expanded to reflect the degree of the polynomial term. In general, the number of 1s and the number of 2s in the subscript identify the powers of X_1 and X_2, respectively, in the polynomial term. For example, the two 1s identify β_{11} as the regression coefficient for the second-degree term in X_1.

The higher-degree polynomial models provide greatly increased flexibility in the response surface. Although it is unlikely that any complex process will be truly polynomial in form, the flexibility of the higher-degree polynomials allows any true model to be approximated to any desired degree of precision.

First-Degree Polynomial

The increased flexibility of the higher-degree polynomial models is illustrated with a sequence of polynomial models containing two independent variables. The first-degree polynomial model, equation 8.1, uses a plane to represent $\mathcal{E}(Y_i)$. This surface is a "table top" tilted to give the slopes $\widehat{\beta}_1$ in the X_1 direction and $\widehat{\beta}_2$ in the X_2 direction (Figure 8.4).

The properties of any response equation can be determined by observing how $\mathcal{E}(Y)$ changes as the values of the independent variables change. For the first-degree polynomial, equation 8.27, the rate of change in $\mathcal{E}(Y)$ as X_1 is changed is the constant β_1, regardless of the values of X_1 and X_2. Similarly, the rate of change in $\mathcal{E}(Y)$ as X_2 changes is determined solely by β_2. The changes in $\mathcal{E}(Y)$ as the independent variables change are given by the partial derivatives of $\mathcal{E}(Y)$ with respect to each of the independent variables. For the first-degree polynomial, the partial derivatives are the

constants β_1 and β_2:

$$\begin{aligned} \frac{\partial \mathcal{E}(Y)}{\partial X_1} &= \beta_1, \quad \text{and} \\ \frac{\partial \mathcal{E}(Y)}{\partial X_2} &= \beta_2. \end{aligned} \tag{8.29}$$

The partial derivative with respect to X_j gives the slope of the surface, or the rate of change in $\mathcal{E}(Y)$, in the X_j direction.

Second-Degree Polynomial

The polynomial model is expanded to allow the rate of change in $\mathcal{E}(Y)$ with respect to one independent variable to be dependent on the value of that variable by including a term that contains the square of the variable. For example, adding a second-degree term in X_1 to equation 8.27 gives

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_{11} X_{i1}^2 + \beta_2 X_{i2} + \epsilon_i. \tag{8.30}$$

The partial derivatives for this model are

$$\begin{aligned} \frac{\partial \mathcal{E}(Y)}{\partial X_1} &= \beta_1 + 2\beta_{11} X_{i1} \\ \frac{\partial \mathcal{E}(Y)}{\partial X_2} &= \beta_2. \end{aligned} \tag{8.31}$$

Now the rate of change in $\mathcal{E}(Y)$ with respect to X_1 is a linear function of X_1, increasing or decreasing according to the sign of β_{11}. The rate of change in $\mathcal{E}(Y)$ with respect to X_2 remains a constant β_2. Notice that the meaning of β_1 is not the same in equation 8.30 as it was in the first-degree polynomial, equation 8.27. Here β_1 is the slope of the surface in the X_1 direction *only* where $X_1 = 0$. The nature of this response surface is illustrated in Figure 8.5.

Interaction Term

The rate of change in $\mathcal{E}(Y)$ with respect to one independent variable can be made dependent on another independent variable by including the product of the two variables as a term in the model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{12} X_{i1} X_{i2} + \epsilon_i. \tag{8.32}$$

The product term $\beta_{12} X_{i1} X_{i2}$ is referred to as an **interaction term**. It allows one independent variable to influence the impact of another. The derivatives are

$$\begin{aligned} \frac{\partial \mathcal{E}(Y)}{\partial X_1} &= \beta_1 + \beta_{12} X_{i2}, \quad \text{and} \\ \frac{\partial \mathcal{E}(Y)}{\partial X_2} &= \beta_2 + \beta_{12} X_{i1}. \end{aligned} \tag{8.33}$$

The rate of change in $\mathcal{E}(Y)$ with respect to X_1 is now dependent on X_2 but not on X_1, and vice versa. Notice the symmetry of the interaction effect;

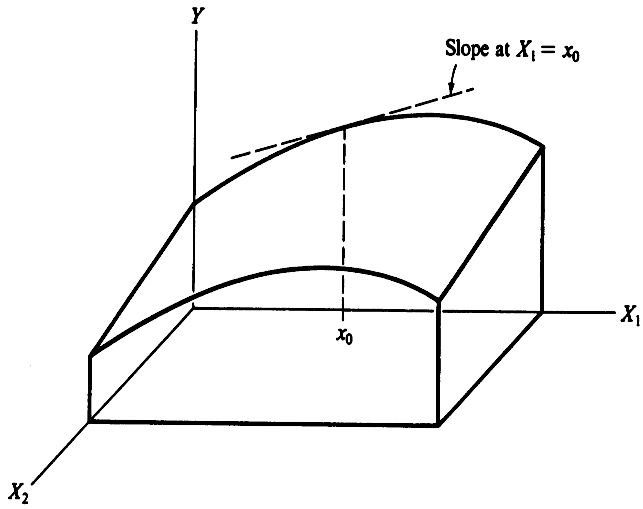

FIGURE 8.5. *A polynomial response surface that is of second degree in X_1 and first degree in X_2.*

both partial derivatives are influenced in the same manner by changes in the other variable. This particular type of interaction term is referred to as the **linear-by-linear interaction**, because the linear slope in one variable is changed linearly (at a constant rate) by changes in the other variable and vice versa. This response function gives a "twisted plane" where the response in $\mathcal{E}(Y)$ to changes in either variable is always linear but the slope is dependent on the value of the other variable. This linear-by-linear interaction is illustrated in Figure 8.6 with the three-dimensional figure in part (a) and a two-dimensional representation showing the relationship between Y and X_1 for given values of X_2. The interaction is shown by the failure of the three lines in (b) to be parallel.

Full Second-Degree Bivariate Model

The full second-degree bivariate model includes all possible second-degree terms as shown in equation 8.28. The derivatives with respect to each independent variable are now functions of both independent variables:

$$\begin{aligned} \frac{\partial \mathcal{E}(Y)}{\partial X_1} &= \beta_1 + 2\beta_{11}X_{i1} + \beta_{12}X_{i2}, \quad \text{and} \\ \frac{\partial \mathcal{E}(Y)}{\partial X_2} &= \beta_2 + 2\beta_{22}X_{i2} + \beta_{12}X_{i1}. \end{aligned} \tag{8.34}$$

The squared terms allow for a curved response in each variable. The product term allows for the surface to be "twisted" (Figure 8.7). β_1 and β_2 are the slopes of the response surface in the X_1 and X_2 directions, respectively, only at the point $X_1 = 0$ and $X_2 = 0$. A quadratic response surface will have a maximum, a minimum, or a saddle point, depending on the coefficients in

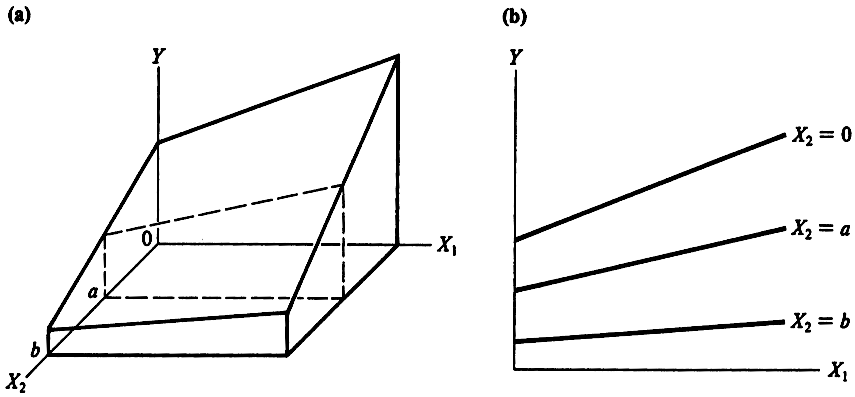

FIGURE 8.6. *Bivariate response surface (a) with interaction and (b) a two-dimensional representation of the surface.*

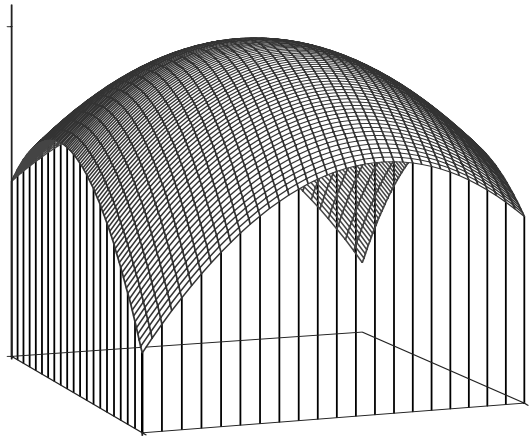

FIGURE 8.7. *A bivariate quadratic response surface with a maximum.*

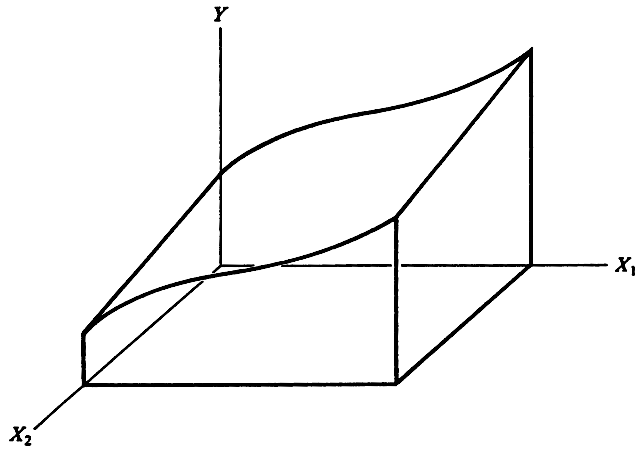

FIGURE 8.8. *A polynomial response surface with a third-degree term in* X_1.

the regression equation. The reader is referred to Box and Draper (1987) for a discussion of the analyis of the properties of quadratic response surfaces. The computer program PROC RSREG (SAS Institute Inc., 1989b) fits a full quadratic model to a set of data and provides an analysis of the properties of the response surface.

Third-Degree Polynomial

The flexibility of the polynomial models is demonstrated by showing the effects of a third-degree term for one of the variables. For example, consider the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{11} X_{i1}^2 + \beta_{111} X_{i1}^3 + \epsilon_i. \tag{8.35}$$

The partial derivative with respect to X_1 is now a quadratic function of X_1:

$$\frac{\partial \mathcal{E}(Y)}{\partial X_1} = \beta_1 + 2\beta_{11} X_{i1} + 3\beta_{111} X_{i1}^2. \tag{8.36}$$

The derivative with respect to X_2 is still β_2. An example of this response surface is shown in Figure 8.8. The full third-degree model in two variables would include all combinations of X_1 and X_2 with sums of the exponents equal to 3 or less.

Flexibility of Polynomials

Increasingly higher-degree terms can be added to the polynomial response model to give an arbitrary degree of flexibility. Any continuous response function can be approximated to any level of precision desired by a polynomial of appropriate degree. Thus, an excellent fit of a polynomial model (or, for that matter, any model) *cannot* be interpreted as an indication that it is in fact the true model. Due to this extreme flexibility, some caution is needed in the use of polynomial models; it is easy to "overfit"

a set of data with polynomial models. Nevertheless, polynomial response models have proven to be extremely useful for summarizing relationships.

Presenting the Response Surface

Polynomial models can be extended to include any number of independent variables. Presenting a multivariate response surface so it can be visualized, however, becomes increasingly difficult. Key features of the response surface (maxima, minima, inflection points) can be determined with the help of calculus. Two- or three-dimensional plots of "slices" of the multivariate surface can be obtained by evaluating the response surface equation at specific values for all independent variables other than the ones of interest.

Caution with Extrapolations

Extrapolation is particularly dangerous when higher-degree polynomial models are being used. The highest degree term in each independent variable eventually dominates the response in that dimension and the surface will "shoot off" in either the positive or negative direction, depending on the sign of the regression coefficient on that term. Thus, minor extrapolations can have serious errors. See Figure 8.2 for an example.

Fitting Polynomials

Fitting polynomial response models with least squares introduces no new conceptual problems. The model is still linear in the parameters and, as long as the usual assumptions on $\boldsymbol{\epsilon}$ are appropriate, ordinary least squares can be used. The higher-degree terms are included in the model by augmenting $\boldsymbol{X}$ with columns of new variables defined as the appropriate powers and products of the independent variables and by augmenting $\boldsymbol{\beta}$ with the respective parameters. The computational problems associated with collinearity are aggravated by the presence of the higher-degree terms because X, X^2, X^3, and so on are often highly collinear. To help alleviate this problem, orthogonal polynomials as discussed in Section 8.1 can be used (Steel, Torrie, and Dickey, 1997) or each independent variable can be centered *before* the higher-degree terms are included in $\boldsymbol{X}$. For example, the quadratic model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{11} X_{i1}^2 + \beta_{22} X_{i2}^2 + \beta_{12} X_{i1} X_{i2} + \epsilon_i \quad (8.37)$$

becomes

$$\begin{aligned} Y_i &= \gamma_0 + \gamma_1 (X_{i1} - \overline{X}_{.1}) + \gamma_2 (X_{i2} - \overline{X}_{.2}) + \gamma_{11} (X_{i1} - \overline{X}_{.1})^2 \\ &\quad + \gamma_{22} (X_{i2} - \overline{X}_{.2})^2 + \gamma_{12} (X_{i1} - \overline{X}_{.1})(X_{i2} - \overline{X}_{.2}) + \epsilon_i. \quad (8.38) \end{aligned}$$

Centering the independent variables changes the definition of the regression coefficients for all but the highest-degree terms. For example, γ_1 and γ_2 are the rates of change in $\mathcal{E}(Y)$ in the X_1 and X_2 directions, respectively, at $X_1 = \overline{X}_{.1}$ and $X_2 = \overline{X}_{.2}$, whereas β_1 and β_2 are the rates of change at $X_1 = X_2 = 0$. The relationship between the two sets of regression coefficients is obtained by expanding the square and product terms in the centered model, equation 8.38, and comparing the coefficients for similar polynomial terms with those in the original model, equation 8.37. Thus,

$$\beta_0 = \gamma_0 - \gamma_1 \overline{X}_{.1} - \gamma_2 \overline{X}_{.2} + \gamma_{11} \overline{X}_{.1}^2 + \gamma_{22} \overline{X}_{.2}^2 + \gamma_{12} \overline{X}_{.1} \overline{X}_{.2},$$

$$\begin{aligned} \beta_1 &= \gamma_1 - 2\gamma_{11}\overline{X}_{.1} - \gamma_{12}\overline{X}_{.2}, \\ \beta_2 &= \gamma_2 - 2\gamma_{22}\overline{X}_{.2} - \gamma_{12}\overline{X}_{.1}, \\ \beta_{11} &= \gamma_{11}, \qquad \beta_{22} = \gamma_{22}, \qquad \text{and} \qquad \beta_{12} = \gamma_{12}. \end{aligned} \tag{8.39}$$

When the sample X-space does not include the origin, the parameters for the centered model are more meaningful because they relate more directly to the behavior of the surface in the region of interest.

Building the Model

The polynomial model is built sequentially, starting either with a first-degree polynomial and adding progressively higher-order terms as needed, or with a high-degree polynomial and eliminating the unneeded higher-degree terms. The lowest-degree polynomial that accomplishes the degree of approximation needed or warranted by the data is adopted. The error term for the tests of significance at each stage must be an appropriate independent estimate of error, preferably estimated from true replication if available. Otherwise, the residual mean square from a model that contains at least all the terms in the more complex model being considered is used as the estimate of error.

Retaining Lower-Order Terms

It is common practice to retain in the model all lower-degree terms, regardless of their significance, that are contained in, or are subsets of, any significant term. For example, if a second-degree term is significant, the first-degree term in the same variable would be retained even if its partial regression coefficient is not significantly different from zero. If the $X_1^2X_2^2$ term is significant, the $X_1, X_2, X_1^2X_2, X_1X_2^2$, and X_1X_2 terms would be retained even if nonsignificant.

The argument for retaining lower-order terms even if not significant is based on these points. First, the meanings and values of the regression coefficients on all except the highest-degree terms change with a simple shift in origin of the independent variables. Recall that reexpressing the independent variables as deviations from their means in a quadratic model changed the meaning of the coefficient for each first-degree term. Thus, the significance or nonsignificance of a lower-order term will depend on the choice of origin for the independent variable during the analysis. A lower-order term that might have been eliminated from a regression equation because it was nonsignificant could "reappear," as a function of the higher-order regression coefficients, when the regression equation was reexpressed with different origins for the independent variables.

Second, eliminating lower-order terms from a polynomial tends to give biased interpretations of the nature of the response surface when the resulting regression equation is studied. For example, eliminating the first-degree term from a second-degree polynomial forces the critical point (maximum, minimum, or saddle point) of the fitted response surface to occur precisely at $X = 0$. (The critical point on a quadratic response surface is found by setting the partial derivatives equal to zero and solving for the values of the independent variable.) For the second-degree polynomial in one variable, the critical point is $X = -\beta_1/(2\beta_{11})$, which is forced to be zero if the first-

degree term has been dropped from the model ($\beta_1 = 0$). Even though β_1 may not be significantly different from zero, it would be more informative to investigate the nature of the response surface before such constraints are imposed. The position of the critical point could then be estimated with its standard error and appropriate inferences made.

These arguments for retaining all lower-degree polynomial terms apply when the polynomial model is being used as an approximation of some unknown model. They are not meant to apply to the case where there is a meaningful basis for a model that contains a higher-order term but not the lower-order terms. The development of a prediction equation for the volume of timber from information on diameter and height of the trees provides an illustration. Geometry would suggest that volume should be nearly proportional to the product of (diameter)^2 and height. Consequently, a model *without* the lower-order terms, diameter and diameter $\times$ height, would be realistic and appropriate.

Example 8.5

A study of the effects of salinity, temperature, and dissolved oxygen on the resistance of young coho salmon to pentachlorophenate is used to illustrate the use of polynomial models [Alderdice (1963) used with permission]. The study used a 3-factor composite design in two stages to estimate the response surface for median survival time (Y) following exposure to 3 mg/l of sodium pentachlorophenate. The treatment variables were water salinity, temperature, and dissolved oxygen content. The first 15 trials (2 replicates) used a 2^3 design of the 3 factors plus the six axial points and the center point (Table 8.3). The last 10 trials were a second-stage study to improve the definition of the center of the response surface. The basic levels of the 3 factors were 9, 5, and 1% salinity; 13, 10, and 7°C temperature; and 7.5, 5.5, and 3.5 mg/1 dissolved oxygen. The independent variables were coded as follows.

$$\begin{aligned} X_1 &= (\text{salinity} - 5\%)/4, \\ X_2 &= (\text{temperature} - 10^\circ C)/3, \text{ and} \\ X_3 &= (\text{dissolved oxygen} - 5.5 mg/l)/2. \end{aligned}$$

The dependent variable, median lethal time, was computed on samples of 10 individuals per experimental unit. The treatment combinations and the observed responses are given in Table 8.3.

It was verified by Alderdice (1963), using the first 15 trials for which there was replication, that a quadratic polynomial response model in the three independent variables was adequate for characterizing the response surface. The replication provided an unbiased estimate of experimental error, which was used to test the lack of fit of the quadratic polynomial. Alderdice then fit the full quadratic or second-degree polynomial model to all the data and presented interpretations of the trivariate response surface equation.

TABLE 8.3. *Treatment combinations of salinity (X_1), temperature (X_2), and dissolved oxygen (X_3), and median lethal time for exposure to 3 mg/l of sodium pentachlorophenate. [Data from Alderdice (1963), and used with permission.]*

	Salinity	*Temperature*	*Oxygen*	*Median Lethal Time*	
Trial	X_1	X_2	X_3	*Rep 1*	*Rep 2*
1	−1	−1	−1	53	50
2	−1	−1	1	54	42
3	−1	1	−1	40	31
4	−1	1	1	37	28
5	1	−1	−1	84	57
6	1	−1	1	76	78
7	1	1	−1	40	49
8	1	1	1	50	54
9	0	0	0	50	50
10	1.215	0	0	61	76
11	−1.215	0	0	54	45
12	0	1.215	0	39	33
13	0	−1.215	0	67	54
14	0	0	1.215	44	45
15	0	0	−1.215	61	38
16	−1.2500	−1.8867	−.6350	46	
17	.8600	−2.2200	−.4250	66	
18	1.0000	−2.2400	−.3100	68	
19	2.1165	−2.4167	−.1450	75	
20	2.5825	−2.4900	−.0800	75	
21	3.2475	−2.6667	.0800	68	
22	1.1760	−1.3333	0	78	
23	1.4700	−1.6667	0	93	
24	1.7640	−2.0000	0	96	
25	2.0580	−2.3333	0	66	

TABLE 8.4. *Partial regression coefficients for the full second-degree polynomial model in three variables for the Alderdice (1963) data.*

Term	$\widehat{\beta}_j$	$s(\widehat{\beta}_j)$	*Student's* t^a
X_1	9.127	1.772	5.151
X_2	−9.852	1.855	−5.312
X_3	.263	1.862	.141
X_1^2	−1.260	1.464	−.861
X_2^2	−6.498	2.014	−3.225
X_3^2	−2.985	2.952	−1.011
X_1X_2	−.934	1.510	−.618
X_1X_3	2.242	2.150	1.042
X_2X_3	−.139	2.138	−.065

[a]The estimate of σ^2 from this model was $s^2 = 76.533$ with 28 degrees of freedom.

Quadratic Model

For this example, the full set of data is used to develop the simplest polynomial response surface model that adequately represents the data. Since the full quadratic model appears to be more than adequate, that model is used as the starting point and higher-degree terms are eliminated if nonsignificant. In addition to the polynomial terms, the model must include a class variable "REP" to account for the differences between the two replications in the first stage and between the first and second stages. Thus, the full quadratic model is

$$\begin{aligned} Y_{ij} &= \mu + \rho_i + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_{11} X_{ij1}^2 + \beta_{22} X_{ij2}^2 + \beta_{33} X_{ij3}^2 \\ &\quad + \beta_{12} X_{ij1} X_{ij2} + \beta_{13} X_{ij1} X_{ij3} + \beta_{23} X_{ij2} X_{ij3} + \epsilon_{ij}, \end{aligned} \tag{8.40}$$

where ρ_i is the effect of the ith "rep," $i = 1, 2$ labels the two replications in stage one, $i = 3$ labels the trials in the second stage, and j designates the observation within the replication. This model allows each rep to have its own level of performance but requires the shape of the response surface to be the same over replications. The presence of the replication effects creates a singularity in $\boldsymbol{X}$ and methods of handling this are discussed in Chapter 9. For this example, we avoid the singularity by letting $\mu_i = \mu + \rho_i, i = 1, 2, 3$. Thus, $\boldsymbol{X}$ for the full-rank model consists of three columns of indicator variables, 0 or 1, identifying to which of the three replications the observation belongs, followed by nine columns of X_1, X_2, X_3, and their squares and products. The partial regression coefficients, their standard errors, and the t-statistics for this full model are given in Table 8.4.

Several of the partial regression coefficients do not approach significance, $t_{(.05/2,28)} = 2.048$; at least some terms can be eliminated from the model. It is not a safe practice, however, to delete all nonsignificant terms in one step unless the columns of the $\boldsymbol{X}$ matrix are orthogonal. The common

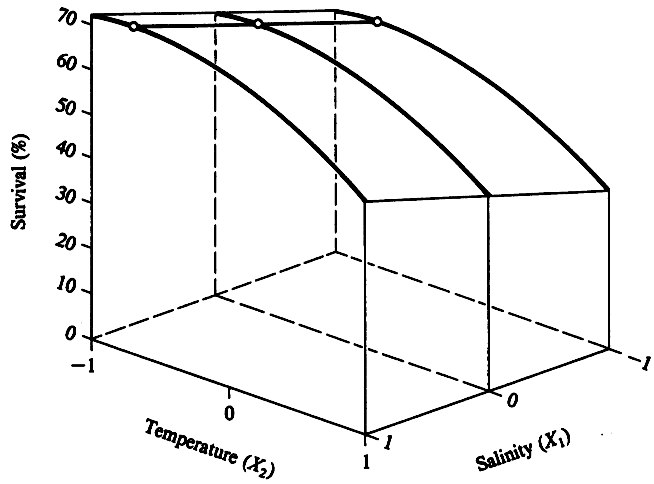

FIGURE 8.9. *Bivariate response surface relating survival time of coho salmon exposed to 3 mg/l of sodium pentachlorophenate to water temperature and water salinity. There was no significant effect of dissolved oxygen (X_3). [Data from Alderdice (1963); used with permission].*

practice with polynomial models is to eliminate the least important of the highest-degree terms at each step. In this example, the X_2X_3 term would be dropped first. Notice that X_3 is retained in the model at this stage, even though it has the smallest t-value, because there are higher-order terms in the model that contain X_3.

Final Model

The subsequent steps consist of dropping $X_1X_2, X_1X_3, X_1^2, X_3^2$, and, finally, X_3 in turn. The final polynomial model is

$$Y_{ij} = \mu_i + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_{22} X_{ij2}^2 + \epsilon_{ij}. \tag{8.41}$$

The residual mean square for this model is 69.09 with 34 degrees of freedom. (The estimate of experimental error from the replicated data is 62.84 with 14 degrees of freedom.) The regression equation, using the weighted average, 59.92, of the estimates of μ_i is

$$\begin{aligned} \widehat{Y} &= 59.92 + 9.21X_1 - 9.82X_2 - 6.896X_2^2 \\ &\quad (2.85) \quad (1.72) \quad (1.76) \quad (1.56). \end{aligned} \tag{8.42}$$

The standard errors of the estimates are shown in parentheses. Thus, within the limits of the observed values of the independent variables, survival time of coho salmon with exposure to sodium pentachlorophenate is well represented by a linear response to salinity, and a quadratic response to temperature (Figure 8.9, page 261). There is no significant effect of dissolved oxygen on survival time and there appear to be no interactions among the

three environmental factors. The linear effect of salinity is to increase survival time 9.2 minutes per coded unit of salinity, or $9.2/4 = 2.3$ minutes per percent increase in salinity. The quadratic response to temperature has a maximum at $X_2 = -\widehat{\beta}_2/(2\widehat{\beta}_{22}) = -.71$, which is 7.9°C on the original temperature scale. (The variance of the estimated maximum point is obtained by using the linear approximation of the ratio of two random variables. This is discussed in Chapter 15, for the more general case of any nonlinear function with nonlinear models.)

The maximum survival times with respect to temperature for given values of salinity are shown with the line on the surface connecting the open circles at $X_2 = -.71$. The investigated region appears to contain the maximum with respect to temperature, but the results suggest even higher salinities will produce greater survival. The linear response to salinity cannot continue without limit. Using the original *full* quadratic model to investigate the critical points on the response surface, Alderdice (1963) found a maximum at $X_1 = 3.2$ (salinity = 17.8%), $X_2 = -1.7$ (temperature = $4.9°C$), and $X_3 = 1.1$ (dissolved oxygen = 7.7 mg/l). These critical points are near the limits of the sample X-space and should be used with caution. Tests of significance indicate that the data are not adequate to support a statement on curvature with respect to salinity or on even a linear response with respect to dissolved oxygen. ■

8.4 Exercises

8.1. The critical point (maximum or minimum) on a quadratic response curve is that point where the tangent to the curve has slope zero. Plot the equation

$$Y = 10 + 2.5X - .5X^2$$

and find the value of X where the tangent to the curve has slope zero. Is the point on the response curve a maximum or a minimum? The derivative of Y with respect to X is $dY/dX = 2.5 - 1.0X$. Solve for the value of X that makes the derivative equal to zero. How does this point relate to the value of X where the tangent was zero?

8.2. Change the quadratic equation in Exercise 8.1 to

$$Y = 10 + 2.5X + .5X^2.$$

Again, plot the equation and find the value of X where the tangent to the curve has slope zero. Is this point a maximum or minimum? What characteristic in the quadratic equation determines whether the critical point is a maximum or a minimum?

8.3. The critical point on a bivariate quadratic response surface is a maximum, minimum, or saddle point. Plot the bivariate polynomial

$$Y = 10 - X_1 + 4X_2 + .25X_1^2 - .5X_2^2$$

over the region $0 < X_1 < 5$ and $2 < X_2 < 6$. Visually locate the critical point where the slopes of the tangent lines in the X_1 direction *and* the X_2 direction are zero. Is this point a maximum, a minimum, or a saddle point? Now use the partial derivatives to find this critical point.

8.4 Assume you have fit the following cubic polynomial to a set of growth data where X ranged from 6 to 20.

$$Y = 50 - 20X + 2.5X^2 - .0667X^3.$$

Plot the response equation over the interval of the data. Does it appear to have a reasonable "growth" form? Demonstrate the sensitivity of the polynomial model to extrapolation by plotting the equation over the interval $X = 0$ to $X = 30$.

8.5 You have obtained the regression equation $Y = 40 - .5X^2$ over the interval $-5 < X < 5$, where $X =$ (temperature in °F-95). Assume the partial regression coefficient for the linear term was not significant and was dropped from the model. Reexpress the regression equation in degrees centigrade, °C $= 5$(°F$-32)/9$. Find the conversion of $X =$ (°F-95) to °C and convert the regression equation. What is the linear regression coefficient in the converted equation? What do you conclude about this linear regression coefficient being different from zero *if* the coefficient on X^2, the .5, in the original equation is significantly different from zero?

8.6 The first four columns of the following data give the average precipitation (inches averaged over 30 years) in April and May for five western U. S. cities and five eastern U. S. cities. (Source: 1993 Almanac and Book of Facts. Pharos Books, Scripps Howard Company, New York.) The last three columns include numbers we use later in

the exercise.

Coast	*City*	*April*	*May*	S_E	X_E	X_W
East	Albany, N. Y.	2.9	3.3	1	2.9	0
East	Washington, D.C.	3.1	3.6	1	3.1	0
East	Jacksonville, Fla.	3.3	4.9	1	3.3	0
East	Raleigh, N.C.	2.9	3.7	1	2.9	0
East	Burlington, Vt.	2.8	3.0	1	2.8	0
West	Los Angeles, Ca.	1.2	.2	0	0	1.2
West	Seattle, Wash.	2.4	1.6	0	0	2.4
West	Portland, Ore.	2.3	2.1	0	0	2.3
West	San Diego, Ca.	2.6	1.5	0	0	2.6
West	Fresno, Ca.	1.2	.3	0	0	1.2

(a) Plot May precipitation versus April using E and W as plot symbols to represent the coasts. What do you conclude from the plot? Is it appropriate to fit a single straight line for both coasts?

(b) Regress the May precipitation on the April precipitation for each region. Add together the error sums of squares and refer to this as the full model residual sum of squares where the full model allows two different slopes and two different intercepts. Compute the difference in the two slopes and in the two intercepts.

(c) Now, regress the May precipitation on the April precipitation using all $n = 10$ points. The error sum of squares here is the reduced model residual sum of squares. The reduced model forces the same intercept and slope for the two groups. Compare the full to the reduced model using an F-test. What degrees of freedom did you use?

(d) Run a multiple regression of May precipitation on columns S_E, X_E, and X_W. What do the coefficients on X_E and X_W represent? Have you seen these numbers before? How about the error sum of squares and the coefficient on S_E? Write out the $\boldsymbol{X}$ matrix for this regression. What would happen to the rank of $\boldsymbol{X}$ if we appended the column of 10 April precipitation numbers to it?

(e) Finally run a multiple regression of May precipitation on April precipitation, S_E, and X_E. Write out the $\boldsymbol{X}$ matrix for this regression. Compute the F-test for the hypothesis that S_E and X_E can be omitted from this model. Have you seen this test before? The coefficient on X_E in this regression estimates the difference of the two slopes in (b) and thus can be used to test the hypothesis of parallel lines. Test the hypothesis that the lines have equal slopes. Omission of S_E from this model produces two lines emanating from the same origin. Test the hypothesis

that both lines have the same intercept (with possibly different slopes).

8.7 You are given the accompanying response data on concentration of a chemical as a function of time. The six sets of observations Y_1 to Y_6 represent different environmental conditions.

Time (h)	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
6	.38	.20	.34	.43	.10	.26
12	.74	.34	.69	.82	.16	.48
24	.84	.51	.74	.87	.18	.51
48	.70	.41	.62	.69	.19	.44
72	.43	.29	.43	.60	.15	.33

(a) Use cubic polynomial models to relate Y = concentration to X = time, where each environment is allowed to have its own intercept and response curve. Is the cubic term significant for any of the environments? [For the purposes of testing homogeneity in Part (c), retain the minimum-degree polynomial model that describes all responses.]

(b) Your knowledge of the process tells you that Y must be zero when $X = 0$. Test the composite null hypothesis that the six intercepts are zero using the model in Part (a) as the full model. What model do you adopt based on this test?

(c) Use the model determined from the test in Part (b) and test the homogeneity of the six response curves. State the conclusion of the test and give the model you have adopted at this stage.

8.8 The data in the table are from a growth experiment with blue-green algae *Spirulina platensis* conducted by Linda Shurtleff, North Carolina State University (data used with permission). These treatments were determined by the amount of "aeration" of the cultures:

1. no shaking and no CO_2 aeration;
2. CO_2 bubbled through the culture;
3. continuous shaking of the culture but no CO_2; and
4. CO_2 bubbled through the culture and continuous shaking of the culture.

There were two replicates for each treatment, each consisting of 14 independent solutions. The 14 solutions in each replicate and treatment were randomly assigned for measurement to 1 of each of the 14 days of study. The dependent variable reported is a log-scale measurement of the increased absorbance of light by the solution, which is interpreted as a measure of algae density. The readings for $DAYS = 0$ are a constant zero and are to be omitted from the analyses.

Growth experiment with blue-green algae.

	Treatment			
Time	*Control*		CO_2	
(days)	*Rep 1*	*Rep 2*	*Rep 1*	*Rep 2*
0	0	0	0	0
1	.220	.482	.530	.184
2	.555	.801	1.183	.664
3	1.246	1.483	1.603	1.553
4	1.456	1.717	1.994	1.910
5	1.878	2.128	2.708	2.585
6	2.153	2.194	3.006	3.009
7	2.245	2.639	3.867	3.403
8	2.542	2.960	4.059	3.892
9	2.748	3.203	4.349	4.367
10	2.937	3.390	4.699	4.551
11	3.132	3.626	4.983	4.656
12	3.283	4.003	5.100	4.754
13	3.397	4.167	5.288	4.842
14	3.456	4.243	5.374	4.969

	Treatment			
Time	*Shaking*		CO_2 + *Shaking*	
(days)	*Rep 1*	*Rep 2*	*Rep 1*	*Rep 2*
0	0	0	0	0
1	.536	.531	.740	.638
2	.974	.926	1.251	1.143
3	1.707	1.758	2.432	2.058
4	2.032	2.021	3.054	2.451
5	2.395	2.374	3.545	2.836
6	2.706	2.933	4.213	3.296
7	3.009	3.094	4.570	3.594
8	3.268	3.402	4.833	3.790
9	3.485	3.564	5.074	3.898
10	3.620	3.695	5.268	4.028
11	3.873	3.852	5.391	4.150
12	4.042	3.960	5.427	4.253
13	4.149	4.054	5.549	4.314
14	4.149	4.168	5.594	4.446

(a) Use quadratic polynomials to represent the response over time. Fit a model that allows each treatment to have its own intercept and quadratic response. Then fit a model that allows each treatment to have its own intercept but forces all to have the same quadratic response. Use the results to test the homogeneity of the responses for the four treatments. (*Note*: Use the residual mean square from the analysis of variance as your estimate of σ^2.) Use the quadratic model you have adopted at this point and define a reduced model that will test the null hypothesis that all intercepts are zero. Complete the test and state your conclusions.

(b) The test of zero intercepts in Part (a) used quadratic polynomials. Repeat the test of zero intercepts using cubic polynomials for each treatment. Summarize the results.

8.9 Assigning a visual volume score to vegetation is a nondestructive method of obtaining measures of biomass. The volume score is the volume of space occupied by the plant computed according to an extensive set of rules involving different geometric shapes. The accompanying data on volume scores and biomass dry weights for grasses were obtained for the purpose of developing a prediction equation for dry weight biomass based on the nondestructive volume score. (Data were provided by Steve Byrne, North Carolina State University, and are used with permission.)

Volume	*Dry Wt.*	*Volume*	*Dry Wt.*
5	0.8	1,753	3.4
1,201	2.2	70,300	107.6
108,936	87.5	62,000	42.3
105,000	94.4	369	1.0
1,060	4.2	4,100	6.9
1,036	0.5	177,500	205.5
33,907	67.7	91,000	120.9
48,500	72.4	2,025	5.5
314	0.6	80	1.3
1,400	3.9	54,800	110.3
46,200	87.7	51,000	26.0
76,800	86.8	55	3.4
24,000	57.6	1,605	3.4
1,575	0.5	15,262	32.1
9,788	20.7	1,362	1.5
5,650	15.1	57,176	85.1
17,731	26.5	25,000	50.5
38,059	9.3		

Use a polynomial response model to develop a prediction equation for $Y = (\text{dry weight})^{1/2}$ on $X = \ln(\text{volume} + 1)$. What degree polynomial do you need? Would it make sense in this case to force the origin to be zero? Will your fit to the data still be satisfactory if you do?

9

CLASS VARIABLES IN REGRESSION

In all previous discussions, the independent variables were continuous or quantitative variables. There are many situations in which this is too restrictive.

This chapter introduces the use of categorical or class variables in regression models. The use of class variables broadens the scope of regression to include the classical analysis of variance models and models containing both continuous and class variables, such as analysis of covariance models and models to test homogeneity of regressions over groups.

To this point, only quantitative variables have been used as independent variables in the models. This chapter extends the models to include qualitative (or categorical) variables. Quantitative variables are the result of some measurement such as length, weight, temperature, area, or volume. There is always a logical ordering attached to the measurements of such variables. Qualitative variables, on the other hand, identify the state, category, or class to which the observation belongs, such as hair color, sex, breed, or country of origin. There may or may not be a logical ordering to the classes. Such variables are called **class variables**.

Class variables greatly increase the flexibility of regression models. This chapter shows how class variables are included in regression models with the use of **indicator variables** or **dummy variables**. The classical anal-

yses of variance for the standard experimental designs are then shown to be special cases of ordinary least squares regression using class variables. This forms the basis for the more general linear model analysis of unbalanced data where conventional analyses of variance are no longer valid (Chapter 17). Then class variables and continuous variables are used jointly to discuss the test of homogeneity of regressions (Section 9.6) and the analysis of covariance (Section 9.7).

Some of the material in the analysis of variance sections of this chapter (Sections 9.2 through 9.5) is not used again until Chapter 17. This material is placed here, rather than immediately preceding Chapter 17, in order to provide the reader with an early appreciation of the generality of regression analyses, and to provide the tools for tests of homogeneity that are used from time to time throughout the text.

9.1 Description of Class Variables

Class Variables

A **class variable** identifies, by an appropriate code, the distinct classes or *levels* of the variable. For example, a code that identifies the different genetic lines, or cultivars, in a field experiment is a class variable. The classes or levels of the variable are the code names or numbers that have been assigned to represent the cultivars. The variation in the dependent variable attributable to this class variable is the total variation among the cultivar classes. It usually does not make sense to think of a continuous response curve relating a dependent variable to a class variable. There frequently is no logical ordering of the class variable or, if there is a logical ordering, the relative spacing of the classes on a quantitative scale is often not well defined.

Quantitative Variables as Class Variables

There are situations in which a quantitative variable is treated (temporarily) as a class variable. That is, the quantitative information contained in the variable is ignored and only the distinct categories or classes are considered. For example, assume the treatments in an experiment are the amounts of fertilizer applied to each experimental unit. The independent variable "amount of fertilizer" is, of course, quantitative. However, as part of the total analysis of the effects of the fertilizer, the total variation among the treatment categories is of interest. The sum of squares "among levels of fertilizer" is the treatment sum of squares and is obtained by using the variable "amount of fertilizer" as a class variable. For this purpose, the quantitative information contained in the variable "amount of fertilizer" is ignored; the variable is used only to identify the grouping or class identification of the observations. Subsequent analyses to determine the nature of the response curve would use the quantitative information in the variable.

The completely random and the randomized complete block experimental designs are used to illustrate the use of class variables in the least squares

regression model. Then, a class variable is introduced to test homogeneity of regression coefficients (for a continuous variable) over the levels of the class variable. Finally, continuous and class variables are combined to give the analysis of covariance in the regression context.

9.2 The Model for One-Way Structured Data

The model for one-way structured data, of which the completely random design (CRD) is the most common example, can be written either as

$$\begin{aligned} Y_{ij} &= \mu_i + \epsilon_{ij} \qquad \text{or} \\ Y_{ij} &= \mu + \tau_i + \epsilon_{ij}, \end{aligned} \tag{9.1}$$

where $\mu_i = \mu + \tau_i$ is the mean of the ith group or treatment and ϵ_{ij} is the random error associated with the jth observation in the ith group, $j = 1, \ldots, r$. The group mean μ_i in the first form is expressed in the second form in terms of an overall constant μ and the effect of the ith group or treatment τ_i, $i = 1, \ldots, t$. The first form is called the **means model**; the second is the classical **effects model** (equation 9.1).

The model assumes that the members of each group are randomly selected from the population of individuals in that group or, in the case of the completely random experimental design, that each treatment has been randomly assigned to r experimental units. (The number of observations in each group or treatment need not be constant but is assumed to be constant for this discussion.)

Class Variable Defined

The data set consists of two columns of information, one containing the response for the dependent variable Y_{ij} and one designating the group or treatment from which the observation came. The code used to designate the group is the class variable. In the case of the CRD, the class variable is the treatment code. For convenience, the class variable is called *treatment* and $i = 1, 2, \ldots, t$ designates the *level* of the class variable.

Model in Matrix Notation

It is easier to see the transition of this model to matrix form if the observations are listed:

$$\begin{aligned} Y_{11} &= \mu + \tau_1 + \epsilon_{11} \\ Y_{12} &= \mu + \tau_1 + \epsilon_{12} \\ &\vdots \\ Y_{1r} &= \mu + \tau_1 + \epsilon_{1r} \\ Y_{21} &= \mu + \tau_2 + \epsilon_{21} \\ &\vdots \\ Y_{2r} &= \mu + \tau_2 + \epsilon_{2r} \end{aligned} \tag{9.2}$$

$$\begin{aligned} &\vdots \\ Y_{tr} &= \mu + \tau_t + \epsilon_{tr}. \end{aligned}$$

The observations here are ordered so that the first r observations are from the first treatment, the second r observations are from the second treatment, and so forth. The total number of observations is $n = rt$ so that the vector of observations on the dependent variable $\boldsymbol{Y}$ is of order $n \times 1$. The total number of parameters is $t + 1$: μ and t τs. The vector of parameters is written

$$\boldsymbol{\beta}' = (\,\mu \quad \tau_1 \quad \tau_2 \quad \cdots \quad \tau_t\,). \tag{9.3}$$

Dummy Variables

In order to express the algebraic model (equation 9.1) in matrix form, we must define $\boldsymbol{X}$ such that the product $\boldsymbol{X\beta}$ associates μ with every observation but each τ_i with only the observations from the ith group. Including μ with every observation is the same as including the common intercept in the usual regression equation. Therefore, the first column of $\boldsymbol{X}$ is $\mathbf{1}$, a column of ones. The remaining columns of $\boldsymbol{X}$ assign the treatment effects to the appropriate observations. This is done by defining a series of **indicator variables** or **dummy variables**, variables that take only the values zero or one. A dummy variable is defined for each level of the class variable. The ith dummy variable is an $n \times 1$ column vector with ones in the rows corresponding to the observations receiving the ith treatment and zeros elsewhere. Thus, $\boldsymbol{X}$ is of order $n \times (t + 1)$.

Example 9.1

To illustrate the pattern, assume there are 4 treatments ($t = 4$) with 2 replications per treatment ($r = 2$). Then $\boldsymbol{Y}$ is an 8×1 vector, $\boldsymbol{X}$ is an 8×5 matrix, and $\boldsymbol{\beta}$ is 5×1:

$$\boldsymbol{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \\ Y_{41} \\ Y_{42} \end{pmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{pmatrix}. \tag{9.4}$$

The second column of $\boldsymbol{X}$ is the dummy variable identifying the observations from treatment 1, the third column identifies the observations from treatment 2, and so on. For this reason, the dummy variables are sometimes called **indicator variables** and $\boldsymbol{X}$ the **indicator matrix**. The reader should verify that multiplication of $\boldsymbol{X}$ by $\boldsymbol{\beta}$ generates the same pattern of model effects shown in equation 9.2. ■

$\boldsymbol{X}$ is Singular

With these definitions of $\boldsymbol{Y}$, $\boldsymbol{X}$, and $\boldsymbol{\beta}$, the model for the completely

random design can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{9.5}$$

which is the usual matrix form of the least squares model. The difference now is that $\boldsymbol{X}$ is not a full-rank matrix; $r(\boldsymbol{X})$ is less than the number of columns of $\boldsymbol{X}$. The singularity in $\boldsymbol{X}$ is evident from the fact that the sum of the last four columns is equal to the first column. This singularity indicates that the model as defined has too many parameters; it is overparameterized.

Since $\boldsymbol{X}$ is not of full rank, the unique $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ does not exist. Therefore, there is no unique solution to the normal equations as there is with the full-rank models. The absence of a unique solution indicates that at least some of the parameters in the model cannot be estimated; they are said to be **nonestimable**. (Estimability is discussed more fully later.)

SS(Regr)

Recall that the degrees of freedom associated with the model sum of squares is determined by the rank of $\boldsymbol{X}$. In full-rank models, $r(\boldsymbol{X})$ always equals the number of columns of $\boldsymbol{X}$. Here, however, there is one linear dependency among the columns of $\boldsymbol{X}$, so the rank of $\boldsymbol{X}$ is t rather than $t+1$. There will be only t degrees of freedom associated with SS(Model). Adjusting the sum of squares for μ uses 1 degree of freedom, leaving $(t-1)$ degrees of freedom for SS(Regr). This SS(Regr) is the partial sum of squares for the t dummy variables defined from the class variable. For convenience, we refer to SS(Regr) more simply as the sum of squares for the class variable. This sum of squares, with $(t-1)$ degrees of freedom, is the treatment sum of squares in the analysis of variance for the completely random experimental design.

Approaches When $\boldsymbol{X}$ Is Singular

Approaches to handling linear models that are not of full rank include:

1. redefine, or reparameterize, the model so that it is a full-rank model; or
2. use one of the nonunique solutions to the normal equations to obtain the regression results.

Reparameterization of the model was the standard approach before computers and is still used in many instances. Understanding reparameterization is helpful in understanding the results of the second approach, which is used in most computer programs for the analysis of general linear models.

9.3 Reparameterizing to Remove Singularities

Purpose

The purpose of reparameterization is to redefine the model so that it is of full rank. This is accomplished by imposing linear constraints on the parameters so as to reduce the number of unspecified parameters to equal the rank of $\boldsymbol{X}$. Then, with $\boldsymbol{X}^*$ of full rank, ordinary least squares can be

used to obtain a solution. If there is one singularity in $\boldsymbol{X}$, one constraint must be imposed, or the number of parameters must be reduced by 1. Two singularities require the number of parameters to be reduced by 2, and so on. There are several alternative reparameterizations for each case. Three common ones are illustrated, each of which gives a full-rank model.

Notation

Each reparameterization carries with it a redefinition of the parameters remaining in the model and corresponding modifications in $\boldsymbol{X}$. To distinguish the reparameterized model from the original model, an asterisk is appended to $\boldsymbol{\beta}$ and $\boldsymbol{X}$, and to the individual parameters when the same symbols are used for both sets. Thus, the reparameterized models are written as $\boldsymbol{Y} = \boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$ with $\boldsymbol{X}^*$ and $\boldsymbol{\beta}^*$ appropriately defined.

9.3.1 Reparameterizing with the Means Model

Defining the Model

The means model, letting $\mu_i = \mu + \tau_i$, is presented here as a reparameterization of the classical effects model. The $(t+1)$ parameters in the effects model are replaced with the t parameters μ_i. The model becomes

$$Y_{ij} = \mu_i + \epsilon_{ij}. \tag{9.6}$$

(This redefinition of the model is equivalent to imposing the constraint that $\mu = 0$ in the original model, leaving τ_1 to τ_t to be estimated. Because of the obvious link of the new parameters to the group means, the usual notation for a population mean μ is used in place of τ.)

Although the means model is used here as a reparameterization of the effects model, it is a valid model in its own right and is often proposed as the more direct approach to the analysis of data (Hocking, 1985). The essential difference between the two models is that the algebraic form of the classical effects model conveys the structure of the data, which in turn generates logical hypotheses and sums of squares in the analysis. The means model, on the other hand, conveys the structure of the data in constraints imposed on the μ_i and in hypotheses specified by the analyst. This text emphasizes the use of the classical effects model. The reader is referred to Hocking (1985) for discussions on the use of the means model.

Defining the Matrices

The reparameterized model is written as

$$\boldsymbol{Y} = \boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon},$$

where (for the case $t = 4$)

$$\boldsymbol{X}^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad \boldsymbol{\beta}^* = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix}. \tag{9.7}$$

The columns of $\boldsymbol{X}^*$ are the dummy variables defined for the original matrix, equation 9.4. Since the first column $\mathbf{1}$ of $\boldsymbol{X}$ is the sum of the columns of $\boldsymbol{X}^*$, the space spanned by the columns of $\boldsymbol{X}$ is the same as that spanned by the columns of $\boldsymbol{X}^*$. Thus, the model in equation 9.7 is a reparameterization of the model given by equation 9.2. For the general case, $\boldsymbol{X}^*$ will be a matrix of order $(n \times t)$, where $n = rt$ is the total number of observations. In this form, $\boldsymbol{X}^*$ is of full rank and ordinary least squares regression can be used to estimate the parameters $\boldsymbol{\beta}^*$.

Solution

The form of $\boldsymbol{X}^*$ in this reparameterization makes the least squares arithmetic particularly simple. $\boldsymbol{X}^{*'}\boldsymbol{X}^*$ is a diagonal matrix of order $(t \times t)$ with the diagonal elements being the number of replications r of each treatment. Thus, $(\boldsymbol{X}^{*'}\boldsymbol{X}^*)^{-1}$ is diagonal with diagonal elements $1/r$. $\boldsymbol{X}^{*'}\boldsymbol{Y}$ is the vector of treatment sums. The least squares solution is

$$\widehat{\boldsymbol{\beta}}^{*'} = (\,\overline{Y}_{1.} \quad \overline{Y}_{2.} \quad \cdots \quad \overline{Y}_{t.}\,), \tag{9.8}$$

which is the vector of treatment means. (A dot in a subscript indicates that the observations have been summed over that subscript; thus, $Y_{i.}$ is the ith treatment sum and $\overline{Y}_{i.}$ is the ith treatment mean.)

Meaning of $\widehat{\boldsymbol{\beta}}^*$

Since this is the least squares solution to a full-rank model, $\widehat{\boldsymbol{\beta}}^*$ is the best linear unbiased estimator of $\boldsymbol{\beta}^*$, but not of $\boldsymbol{\beta}$. (The parameters $\boldsymbol{\beta}$ in the original model are not estimable.) It is helpful in understanding the results of the reparameterized model to know what function of the original parameters is being estimated by $\widehat{\boldsymbol{\beta}}^*$. This is determined by finding the expectation of $\widehat{\boldsymbol{\beta}}^*$ in terms of the expectation of $\boldsymbol{Y}$ from the original model, $\mathcal{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$:

$$\begin{aligned} \mathcal{E}(\widehat{\boldsymbol{\beta}}^*) &= (\boldsymbol{X}^{*'}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*'}\mathcal{E}(\boldsymbol{Y}) \\ &= [(\boldsymbol{X}^{*'}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*'}\boldsymbol{X}]\boldsymbol{\beta}. \end{aligned} \tag{9.9}$$

Notice that the last $\boldsymbol{X}$ is the original matrix. Evaluating this expectation for the current reparameterization (again using $t = 4$) gives

$$\mathcal{E}(\widehat{\boldsymbol{\beta}}^*) = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{pmatrix} = \begin{pmatrix} \mu + \tau_1 \\ \mu + \tau_2 \\ \mu + \tau_3 \\ \mu + \tau_4 \end{pmatrix}. \tag{9.10}$$

Thus, each element of $\widehat{\boldsymbol{\beta}}^*$, $\widehat{\mu}_i = \overline{Y}_{i.}$, is an estimate of $\mu + \tau_i$. This is the expectation of the ith group mean under the original model.

Estimable Functions of $\boldsymbol{\beta}$

Unbiased estimators of other estimable functions of the original parameters are obtained by using appropriate linear functions of $\widehat{\boldsymbol{\beta}}^*$. For example, $(\tau_1 - \tau_2)$ is estimated unbiasedly by $\widehat{\mu}_1 - \widehat{\mu}_2 = \overline{Y}_{1.} - \overline{Y}_{2.}$. Notice, however,

that there is no linear function of $\widehat{\boldsymbol{\beta}}^*$ that provides an unbiased estimator of μ, or of one of the τ_i. These are *nonestimable functions* of the original parameters, and no reparameterization of the model will provide unbiased estimators of such nonestimable quantities. (In a general linear model, a linear combination $\boldsymbol{\lambda}'\boldsymbol{\beta}$ of parameters is said to be *estimable* if there is a linear function $\boldsymbol{a}'\boldsymbol{Y}$ that is unbiased for $\boldsymbol{\lambda}'\boldsymbol{\beta}$. If no such linear combination exists, then it is said to be *nonestimable*.)

SS(Model) and SS(Res)

The sum of squares due to this model is the uncorrected treatment sum of squares

$$\begin{aligned} \text{SS(Model)} &= \widehat{\boldsymbol{\beta}}^{*'}\boldsymbol{X}^{*'}\boldsymbol{Y} \\ &= \left[\sum_{i=1}^{t}(Y_{i.}^2)\right]/r \end{aligned} \tag{9.11}$$

because the elements of $\widehat{\boldsymbol{\beta}}^*$ are the treatment means and the elements of $\boldsymbol{X}^{*'}\boldsymbol{Y}$ are the treatment sums. The residual sum of squares is the pooled sum of squares from among the replicate observations within each group

$$\begin{aligned} \text{SS(Res)} &= \boldsymbol{Y}'\boldsymbol{Y} - \text{SS(Model)} \\ &= \sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}^2 - \frac{\sum_{i=1}^{t}(Y_{i.})^2}{r} \\ &= \sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij} - \overline{Y}_{i.})^2 \end{aligned} \tag{9.12}$$

and has $(n-t)$ degrees of freedom.

Treatment Sum of Squares

SS(Model) measures the squared deviations of the treatment means from zero. Comparisons among the treatment means are of greater interest. Sums of squares for these comparisons are generated using the general linear hypothesis (discussed in Section 4.5). For example, the sum of squares for the null hypothesis that all μ_i are equal is obtained by constructing a $\boldsymbol{K}'$ matrix of rank $(t-1)$ to account for all differences among the t treatment parameters. One such $\boldsymbol{K}'$ (for $t=4$) is

$$\boldsymbol{K}' = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}. \tag{9.13}$$

This matrix defines the three nonorthogonal but linearly independent contrasts of treatment 1 versus treatment 2, treatment 2 versus treatment 3, and treatment 3 versus treatment 4. (A linear combination $\sum a_i\mu_i$ is said to be a *contrast* of the treatment means if $\sum a_i = 0$). Any set of three linearly independent contrasts would produce the sum of squares for the hypothesis that the $t=4$ μ_i are equal. The sum of squares for this hypothesis is the treatment sum of squares for the $t=4$ treatments. In general, the

TABLE 9.1. *Relationship between the conventional analysis of variance and ordinary least squares regression computations for the completely random experimental design.*

Source of Variation	*d.f.*	*Traditional AOV SS*	*Regression SS*
$\text{Total}_{\text{uncorr}}$	rt	$\sum\sum Y_{ij}^2$	$\boldsymbol{Y}'\boldsymbol{Y}$
Model	t	$\sum(Y_{i.})^2/r$	$\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}$
C.F.	1	$n\overline{Y}^2$	$n\overline{Y}^2$
Treatments	$t-1$	$\sum(Y_{i.})^2/r - n\overline{Y}^2$	$\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} - n\overline{Y}^2$
Residual	$t(r-1)$	$\sum\sum Y_{ij}^2 - \sum(Y_{i.})^2/r$	$\boldsymbol{Y}'\boldsymbol{Y} - \text{SS(Model)}$

treatment sum of squares can be obtained by defining a matrix of contrasts $\boldsymbol{K}'$ with $r(\boldsymbol{K}') = (t-1)$.

Alternatively, the treatment sum of squares can be obtained by using the difference in sums of squares between full and reduced models. The reduced model for the null hypothesis that all μ_i are equal contains only one parameter, a constant mean μ. The sum of squares for such a model is $\text{SS}(\mu) = n\overline{Y}_{..}^2$, or the sum of squares due to correction for the mean, commonly called the **correction factor** (C.F.). Thus, the treatment sum of squares for the completely random experimental design can be obtained as $\text{SS(Model)} - \text{SS}(\mu)$. The relationship between the conventional analysis of variance and the regression analysis for the completely random design is summarized in Table 9.1.

9.3.2 Reparameterization Motivated by $\sum \tau_i = 0$

Redefining the Model

The original model defined the τ_i as deviations from μ. If μ is thought of as the overall true mean $\overline{\mu}_{.}$ of the t treatments and τ_i as $\mu_i - \overline{\mu}_{.}$, it is reasonable to impose the condition that the sum of the treatment deviations about the true mean is zero; that is, $\sum \tau_i = 0$. This implies that one τ_i can be expressed as the negative of the sum of the other τ_i. The number of parameters to be estimated is thus reduced by 1.

The constraint $\sum \tau_i = 0$ is used to express the last treatment effect τ_t in terms of the first $(t-1)$ treatment effects. Thus,

$$\tau_t = -(\tau_1 + \tau_2 + \cdots + \tau_{t-1})$$

is substituted for τ_t everywhere in the original model. In the example with $t = 4$, $\tau_4 = -(\tau_1 + \tau_2 + \tau_3)$ so that the model for each observation in the fourth group changes from

$$Y_{4j} = \mu + \tau_4 + \epsilon_{4j}$$

to

$$Y_{4j} = \mu + (-\tau_1 - \tau_2 - \tau_3) + \epsilon_{4j}.$$

This substitution eliminates τ_4, reducing the number of parameters from 5 to 4 or, in general, from $(t+1)$ to t. The vector of redefined parameters is

$$\boldsymbol{\beta}^{*'} = (\,\mu^* \quad \tau_1^* \quad \tau_2^* \quad \tau_3^*\,). \tag{9.14}$$

The design matrix $\boldsymbol{X}^*$ for this reparameterization is obtained from the original $\boldsymbol{X}$ as follows assuming $t = 4$. The dummy variable for treatment 4, the last column of $\boldsymbol{X}$, equation 9.4, identifies the observations that contain τ_4 in the model. For each such observation, the substitution of $-(\tau_1 + \tau_2 + \tau_3)$ for τ_4 is accomplished by replacing the "0" coefficients on τ_1, τ_2, and τ_3 with "-1," and dropping the dummy variable for τ_4. Thus, $\boldsymbol{X}^*$ for this reparameterization is

$\boldsymbol{X}^*$

$$\boldsymbol{X}^* = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix}. \tag{9.15}$$

It is not difficult to show that the space spanned by the columns of $\boldsymbol{X}^*$ in equation 9.15 is the same as that spanned by $\boldsymbol{X}$ in equation 9.4. See Exercise 9.7.

Again, the reparameterized model is of full rank and ordinary least squares gives an unbiased estimate of the new parameters defined in $\boldsymbol{\beta}^*$. The expectation of $\widehat{\boldsymbol{\beta}}^*$ in terms of the parameters in the original model and the means model is found from equation 9.9 using $\boldsymbol{X}^*$ from the current reparameterization. This gives

$\widehat{\boldsymbol{\beta}}^*$

$$\begin{aligned} \mathcal{E}(\widehat{\boldsymbol{\beta}}^*) &= \begin{bmatrix} 1 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ 0 & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ 0 & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \end{bmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{pmatrix} \\ &= \begin{pmatrix} \mu + \overline{\tau} \\ \tau_1 - \overline{\tau} \\ \tau_2 - \overline{\tau} \\ \tau_3 - \overline{\tau} \end{pmatrix} = \begin{pmatrix} \overline{\mu}_{\cdot} \\ \mu_1 - \overline{\mu}_{\cdot} \\ \mu_2 - \overline{\mu}_{\cdot} \\ \mu_3 - \overline{\mu}_{\cdot} \end{pmatrix}, \end{aligned} \tag{9.16}$$

where $\overline{\tau}$ is the average of the four τ_i. Note that the expectation of $\widehat{\boldsymbol{\beta}}^*$ is expressed in terms of the parameters of the original model with no constraints. The constraint $\sum \tau_i = 0$ was used only to generate a full rank

reparameterization of the original model. Thus, $\widehat{\mu}^*$ is an estimator of $\mu+\overline{\tau}$, $\widehat{\tau}_1^*$ is an estimator of $(\tau_1-\overline{\tau})$, and so forth. Note that an unbiased estimator of $\tau_4-\overline{\tau}$ is given by

$$\widehat{\tau}_4^* = -(\widehat{\tau}_1^* + \widehat{\tau}_2^* + \widehat{\tau}_3^*). \tag{9.17}$$

Functions of the $\boldsymbol{\beta_i}$

Other estimable functions of the original parameters are obtained from appropriate linear functions of $\boldsymbol{\beta}^*$. For example, the least squares estimator of the ith treatment mean $(\mu+\tau_i)$ is given by $(\widehat{\mu}^*+\widehat{\tau}_i^*)$. The estimator of the difference between two treatment effects, say $(\tau_2-\tau_3)$, is given by $(\widehat{\tau}_2^*-\widehat{\tau}_3^*)$.

Treatment Sum of Squares

The analysis of variance for the completely random design is obtained from this reparameterization in much the same way as with the means reparameterization. The sum of squares for treatments is obtained as the sum of squares for the null hypothesis

$$H_o : \tau_i^* = 0, \qquad \text{for } i = 1, 2, 3$$

or as

$$\text{SS(Model)} - \text{SS}(\mu).$$

In terms of the original parameters, this null hypothesis is satisfied only if all τ_i are equal.

9.3.3 Reparameterization Motivated by $\tau_t = 0$

Redefining the Model

Another method of reducing the number of parameters in an overparameterized model is to arbitrarily set the required number of nonestimable parameters equal to zero. In the model for the completely random experimental design, one constraint is needed so that one parameter—usually the last τ_i—is set equal to zero. In the example with four treatments, setting $\tau_4 = 0$ gives

$$\boldsymbol{\beta}^{*'} = (\,\mu^* \quad \tau_1^* \quad \tau_2^* \quad \tau_3^*\,)$$

and an $\boldsymbol{X}^*$ that contains only the first four columns of the original $\boldsymbol{X}$. Since the last column of $\boldsymbol{X}$ is the difference between the first column and the sum of the last three columns of $\boldsymbol{X}^*$, the space spanned by columns of $\boldsymbol{X}^*$ is the same as that spanned by the columns of $\boldsymbol{X}$. As with the other reparameterizations, this model is of full rank and ordinary least squares can be used to obtain the solution $\widehat{\boldsymbol{\beta}}^*$.

$\widehat{\boldsymbol{\beta}}^*$

The expectation of $\widehat{\boldsymbol{\beta}}^*$ in terms of the parameters in the original model (from equation 9.9) using the current $\boldsymbol{X}^*$ is

$$\mathcal{E}(\widehat{\boldsymbol{\beta}}^*) = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{pmatrix}$$

$$= \begin{pmatrix} \mu + \tau_4 \\ \tau_1 - \tau_4 \\ \tau_2 - \tau_4 \\ \tau_3 - \tau_4 \end{pmatrix} = \begin{pmatrix} \mu_4 \\ \mu_1 - \mu_4 \\ \mu_2 - \mu_4 \\ \mu_3 - \mu_4 \end{pmatrix}. \tag{9.18}$$

With this parameterization, $\widehat{\mu}^*$ is an estimator of the mean of the fourth treatment $\mu + \tau_4$, and each $\widehat{\tau}_i^*$ estimates the difference between the true means of the ith treatment and the fourth treatment. Hence, this reparameterization is also called the *reference cell model*. The ith treatment mean $\mu + \tau_i$ is estimated by $\widehat{\mu}^* + \widehat{\tau}_i^*$. The difference between two means $(\tau_i - \tau_{i'})$ is estimated by $(\widehat{\tau}_i^* - \widehat{\tau}_{i'}^*)$.

Treatment Sum of Squares

The treatment sum of squares for this parameterization is given as the sum of squares for the composite null hypothesis

$$H_0 : \tau_i^* = 0 \qquad \text{for } i = 1, 2, 3$$

or as

$$\text{SS(Model)} - \text{SS}(\mu).$$

In terms of the original parameters, this hypothesis implies that the first three τ_i are each equal to τ_4 (equation 9.18), or that $\tau_1 = \tau_2 = \tau_3 = \tau_4$.

Estimable Functions

Each of the three reparameterizations introduced in this section has provided estimates of the meaningful functions of the original parameters, the true means of the treatments, and all contrasts among the true treatment means. These are estimable functions of the original parameters. As a general result, if a function of the original parameters is estimable, it can be estimated from $\widehat{\beta}^*$ obtained from any reparameterization. Furthermore, the same numerical estimate for any estimable function of the original parameters will be obtained from every reparameterization. Estimability is discussed more fully in Chapter 17 and the reader is referred to Searle (1971) for the theoretical developments.

9.3.4 Reparameterization: A Numerical Example

Example 9.2

A small numerical example illustrates the three reparameterizations. An artificial data set was generated to simulate an experiment with $t = 4$ and $r = 2$. The conventional one-way model was used with the parameters chosen to be $\mu = 12$, $\tau_1 = -3$, $\tau_2 = 0$, $\tau_3 = 2$, and $\tau_4 = 4$. A random observation from a normal distribution with mean zero and unit variance was added to each expectation to simulate random error. (The τ_i are chosen so they do not add to zero for this illustration.) The vector of observations

TABLE 9.2. *Estimates obtained from simulated data for three reparameterizations of the one-way model, $t = 4$ and $r = 2$. Expectations of the estimators are in terms of the parameters of the original singular model.*

Means Model		*Reparameterization:* $\sum \tau_i = 0$		$\tau_4 = 0^a$	
$\widehat{\boldsymbol{\beta}}^*$	$\mathcal{E}(\widehat{\boldsymbol{\beta}}^*)$	$\widehat{\boldsymbol{\beta}}^*$	$\mathcal{E}(\widehat{\boldsymbol{\beta}}^*)$	$\widehat{\boldsymbol{\beta}}^*$	$\mathcal{E}(\widehat{\boldsymbol{\beta}}^*)$
8.830	$\mu + \tau_1$	12.731	$\mu + \overline{\tau}$	16.680	$\mu + \tau_4$
11.925	$\mu + \tau_2$	−3.901	$\tau_1 - \overline{\tau}$	−7.850	$\tau_1 - \tau_4$
13.490	$\mu + \tau_3$	−.806	$\tau_2 - \overline{\tau}$	−4.755	$\tau_2 - \tau_4$
16.680	$\mu + \tau_4$	.759	$\tau_3 - \overline{\tau}$	−3.190	$\tau_3 - \tau_4$

[a]The solution obtained from the general linear models solution in PROC GLM corresponds to that for $\tau_4 = 0$.

generated in this manner was

$$\boldsymbol{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \\ Y_{41} \\ Y_{42} \end{bmatrix} = \begin{bmatrix} \mu + \tau_1 + \epsilon_{11} \\ \mu + \tau_1 + \epsilon_{12} \\ \mu + \tau_2 + \epsilon_{21} \\ \mu + \tau_2 + \epsilon_{22} \\ \mu + \tau_3 + \epsilon_{31} \\ \mu + \tau_3 + \epsilon_{32} \\ \mu + \tau_4 + \epsilon_{41} \\ \mu + \tau_4 + \epsilon_{42} \end{bmatrix} = \begin{bmatrix} 8.90 \\ 8.76 \\ 11.78 \\ 12.07 \\ 14.50 \\ 12.48 \\ 16.79 \\ 16.57 \end{bmatrix}. \tag{9.19}$$

The parameter estimates from these data for each of the three reparameterizations and their expectations in terms of the original parameters are shown in Table 9.2. Most notable are the numerical differences in $\widehat{\boldsymbol{\beta}}^*$ for the different parameterizations. All convey the same information but in very different packages. The results from the means model are the most directly useful; each regression coefficient estimates the corresponding group mean. Contrasts among the τ_i are estimated by the same contrasts among the estimated regression coefficients. For example,

$$\widehat{\mu}_1^* - \widehat{\mu}_2^* = 8.8300 - 11.9250 = -3.0950$$

is an estimate of $(\tau_1 - \tau_2)$, which is known to be -3 from the simulation model.

The reparameterization motivated by the "sum" constraint gives $\widehat{\mu}^* = 12.73125$, which is an estimate of the overall mean plus the average of the treatment effects. [From the simulation model, $(\mu + \overline{\tau})$ is known to be 12.75.] Each of the other computed regression coefficients is estimating the deviation of a τ_i from $\overline{\tau}$. The estimate of $(\tau_4 - \overline{\tau})$ is obtained from equation 9.17. This gives

$$\widehat{\tau}_4^* = -(-3.90125 - .80625 + .75875) = 3.94875.$$

The sum of the first two estimates,

$$\widehat{\mu}^* + \widehat{\tau}_1^* = 12.73125 + (-3.90125) = 8.8300,$$

is an estimate of $(\mu + \tau_1)$. This estimate is identical to that obtained for $(\mu + \tau_1)$ from the means model. Similarly, the estimate of $(\tau_1 - \tau_2)$,

$$\widehat{\tau}_1^* - \widehat{\tau}_2^* = -3.90125 - (-.80625) = -3.095,$$

is the same as that obtained from the means model.

The third reparameterization motivated by $\tau_4 = 0$ gives $\widehat{\mu}^* = 16.6800$, which is an estimate of $(\mu + \tau_4)$, the true mean of the fourth group. The sum of the first two regression coefficients again estimates $(\mu + \tau_1)$ as

$$\widehat{\mu}^* + \widehat{\tau}_1^* = 16.6800 + (-7.8500) = 8.8300.$$

Each $\widehat{\tau}_1^*$ in this reparameterization estimates the difference in effects between the ith group and the fourth group. The numerical values obtained for these estimates are identical to those obtained from the other models. ■

Unique Results from Reparameterizations

The results from these three reparameterizations illustrate general results. Least squares estimates of $\boldsymbol{\beta}^*$ obtained from different reparameterizations estimate different functions of the original parameters. The relationship of the redefined parameters to those in the original model must be known in order to properly interpret these estimates. Even though the solution appears to change with the different reparameterizations, all give identical numerical estimates of every estimable function of the original parameters. This includes $\widehat{\boldsymbol{Y}} = \boldsymbol{X}^*\widehat{\boldsymbol{\beta}}^*$ and $\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$. Furthermore, sums of squares associated with any estimable contrast on $\boldsymbol{\beta}$ are identical, which implies that all parameterizations give the same analysis of variance. In Example 9.2, all models gave

$$\text{SS(Regr)} = 64.076238 \qquad \text{and} \qquad \text{SS(Res)} = 2.116250.$$

9.4 Generalized Inverse Approach

When $\boldsymbol{X}$ is not of full rank there is no unique solution to the normal equations $(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$. A general approach to models of less than full rank is to use one of the nonunique solutions to the normal equations. This is accomplished by using a **generalized inverse** of $\boldsymbol{X}'\boldsymbol{X}$. (The generalized inverse of a matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^-$.) There are many different kinds of generalized inverses which, to some extent, have different properties. The reader is referred to Searle (1971) for complete discussions on generalized

inverses. It is sufficient for now to know that a generalized inverse provides one of the infinity of solutions that satisfies the normal equations. Such a solution is denoted with $\boldsymbol{\beta}^0$ to emphasize the fact that it is not a unique solution. $\widehat{\boldsymbol{\beta}}$ is reserved as the label for the unique least squares solution when it exists. Thus,

$$\boldsymbol{\beta}^0 = (\boldsymbol{X}'\boldsymbol{X})^- \boldsymbol{X}'\boldsymbol{Y}. \tag{9.20}$$

Computers are used to obtain the generalized inverse solutions.

Estimable Functions

Since $\boldsymbol{\beta}^0$ is not unique, its elements per se are meaningless. Another generalized inverse would give another set of numbers from the same data. However, many of the regression results obtained from using a nonunique solution are unique; the same numerical results are obtained regardless of which solution is used. It was observed in Section 9.3 that all reparameterizations gave identical estimates of estimable functions of the parameters. This important result applies to all generalized inverse solutions to the normal equations. Any estimable function of the original parameters is *uniquely* estimated by the same linear function of one of the nonunique solutions $\boldsymbol{\beta}^0$. That is, if $\boldsymbol{K}'\boldsymbol{\beta}$ is estimable, then $\boldsymbol{K}'\boldsymbol{\beta}^0$ is the least squares estimate of $\boldsymbol{K}'\boldsymbol{\beta}$ *and* the estimate is unique with respect to choice of solution. Such estimates of estimable linear functions of the original parameters have all the desirable properties of least squares estimators.

Unique Results

Results concerning other unique quantities follow from this statement. For example, $\boldsymbol{X}\boldsymbol{\beta}$ is an estimable function of $\boldsymbol{\beta}$ and, hence, $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{\beta}^0$ is the unique unbiased estimate of $\boldsymbol{X}\boldsymbol{\beta}$. Then, $\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ must be unique. Since SS(Model) $= \widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}}$ and SS(Res) $= \boldsymbol{e}'\boldsymbol{e}$, these sums of squares are also unique with respect to choice of solution. The uniqueness extends to the partitions of the sums of squares, as long as the sums of squares relate to hypotheses that are estimable linear functions of the parameters.

Thus, the generalized inverse approach to models of less than full rank provides all the results of interest. The only quantities not estimated uniquely are those quantities for which the data contain no information—the nonestimable functions of $\boldsymbol{\beta}$.

PROC GLM

The generalized inverse approach is used for the least squares analysis of models of less than full rank by many computer programs, including PROC GLM (SAS Institute Inc., 1989b). In their procedure, any variable in the model that is to be regarded as a class variable must be identified in a CLASS statement in the program. Each class variable will generate one or more singularities that make the model less than full rank. (Singularities can also result from linear dependencies among continuous variables, but this chapter is concerned with the use of class variables in regression models.) Since the estimates of the regression coefficients in the singular model are not unique, PROC GLM does not print the solution $\boldsymbol{\beta}^0$ unless it is specifically requested. The unique results from the analysis are obtained by requesting estimation of specific estimable functions and tests of

testable hypotheses. (A testable hypothesis is one in which the linear functions of parameters in the null hypothesis are estimable functions.)

When a class variable is specified, PROC GLM creates $\boldsymbol{\beta}$ and the set of dummy variables for the $\boldsymbol{X}$ matrix as was done in Section 9.2. No reparameterization is done so that $\boldsymbol{X}$ remains singular. The particular generalized inverse used by PROC GLM gives the same solution as that obtained with reparameterization using the constraint $\tau_t = 0$. The solution vector in PROC GLM contains an estimate for every parameter including τ_t^*. But, because each $\widehat{\tau}_i^*$ is estimating $\tau_i - \tau_t$, the numerical value of $\widehat{\tau}_t^*$ is always zero. Thus, the PROC GLM solution for the simulated data from the completely random design is the same as that given in the last column of Table 9.2, except the vector of estimates includes $\widehat{\tau}_4^*$ in the fifth position. The estimates obtained for all estimable functions and sums of squares are identical to those obtained from the reparameterizations.

9.5 The Model for Two-Way Classified Data

Defining the Model

The conventional model for two-way classified data, of which the randomized complete block design (RCB) is the most common example, is

$$Y_{ij} = \mu + \gamma_i + \tau_j + \epsilon_{ij}, \tag{9.21}$$

where μ is an overall mean, γ_i is the effect of the ith block, τ_j is the effect of the jth treatment, and ϵ_{ij} is the random error. In this model there are two class variables—"block" and "treatment"—which identify the particular block and treatment associated with the ijth experimental unit. There are b levels ($i = 1, \ldots, b$) of the block class variable and t levels ($j = 1, \ldots, t$) of the treatment class variable.

$\boldsymbol{X}$ and $\boldsymbol{\beta}$

Defining the $\boldsymbol{X}$ matrix for this model requires b dummy variables for blocks and t dummy variables for treatments. The vector of observations is assumed to be ordered with all of the treatments occurring in order for the first block followed by the treatments in order for the second block, and so forth. The parameter vector $\boldsymbol{\beta}$ is defined with the block effects γ_i occurring before the treatment effects τ_j. For illustration, assume that $b = 2$ and $t = 4$ for a total of $bt = 8$ observations. Then,

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}, \qquad \text{and} \qquad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \gamma_1 \\ \gamma_2 \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{bmatrix}. \tag{9.22}$$

The second and third columns of $\boldsymbol{X}$ are the dummy variables for blocks; the last four columns are the dummy variables for treatments.

Degrees of Freedom

There are two linear dependencies in $\boldsymbol{X}$. The sum of the block dummy variables (columns 2 and 3) and the sum of the treatment dummy variables (the last four columns) both equal column 1. Thus, the rank of $\boldsymbol{X}$ is $r(\boldsymbol{X}) = 7-2 = 5$, which is the degrees of freedom for SS(Model). In the conventional RCB analysis of variance these degrees of freedom are partitioned into 1 for the correction factor, $(b-1) = 1$ for SS(Blocks), and $(t-1) = 3$ for SS(Treatments).

Reparameterizing Using $\gamma_2 = \tau_4 = 0$

Reparameterizing this model to make it full rank requires two constraints. The effective number of parameters must be reduced to 5, the rank of $\boldsymbol{X}$. The simplest constraints to obtain a full rank reparameterization would be to use $\gamma_2 = 0$ and $\tau_4 = 0$. These constraints have the effect of eliminating γ_2 and τ_4 from $\boldsymbol{\beta}$ and columns 3 and 7 from $\boldsymbol{X}$. Thus, $\boldsymbol{X}^*$ would be an 8×5 matrix consisting of columns 1, 2, 4, 5, and 6 from $\boldsymbol{X}$ and $\boldsymbol{\beta}^*$ would be

$$\boldsymbol{\beta}^{*'} = (\,\mu^* \quad \gamma_1^* \quad \tau_1^* \quad \tau_2^* \quad \tau_3^*\,)\,. \tag{9.23}$$

Reparameterizing Using Sum Constraints

The constraints requiring the sum of the effects to be zero would be $\sum \gamma_i = 0$ and $\sum \tau_j = 0$. These constraints are imposed by substituting $-\gamma_1$ for γ_2 and $-(\tau_1 + \tau_2 + \tau_3)$ for τ_4 in the original model. This reduces the number of parameters by two and gives

$$\boldsymbol{X}^* = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 \end{bmatrix}. \tag{9.24}$$

Either of these reparameterizations will generate the conventional analysis of variance of two-way classified data when the least squares regression concepts are applied. The full model consists of μ^*, the γ_i^*, and the τ_j^*. The residual mean square from this model estimates σ^2. The general linear hypothesis can be used to generate the sum of squares for testing the null hypothesis that γ_1^* is zero. In the more general case, this would be a composite hypothesis that all γ_i^* are zero. The sum of squares Q, generated for this hypothesis, will have 1 degree of freedom [or, in general, $(b-1)$ degrees of freedom] and is algebraically identical to SS(Blocks) in the conventional analysis of variance. Similarly, the sum of squares associated with the composite hypothesis that all τ_j^* are zero is identical to SS(Treatments) in the conventional analysis of variance. These sums of squares can also be computed from the procedure based on $[\mathrm{SS}(\mathrm{Res}_{\mathrm{reduced}}) - \mathrm{SS}(\mathrm{Res}_{\mathrm{full}})]$.

Reparameterizing Using the Means Model

The model could also be made full rank by using the means model reparameterization. Each cell of the two-way table would be assigned its own mean. Thus,

$$\boldsymbol{Y}_{ij} = \mu_{ij} + \epsilon_{ij}, \tag{9.25}$$

where $\mu_{ij} = \mu + \gamma_i + \tau_j$ in terms of the parameters of the original model. This model is different from the original, however. The original model specified a column (or treatment) effect and a row (or block) effect that added to give the "cell" effect; the same column effect was imposed on all rows and the same row effects applied to all columns. Deviations from the sum of the block and treatment effects were assumed to be random error. The means model as given, on the other hand, imposes no restrictions on the relationships among the μ_{ij}. The means model is made analogous to the classical RCB effects model by imposing constraints on the μ_{ij} so as to satisfy the conditions of no interaction in every 2×2 subtable of the $b \times t$ table of μ_{ij}. The reader is referred to Hocking (1985) for complete discussions on analyses using means models.

Generalized Inverse Approach

The generalized inverse approach also can be used for two-way classified data. The two class variables would be used to generate the singular $\boldsymbol{X}$ (equation 9.22) and a generalized inverse would be used to obtain a (nonunique) solution. SS(Res) from that analysis would be the interaction sum of squares for the two-way table, which in the RCB design is the estimate of experimental error. Appropriate hypotheses on the subsets of parameters generate the usual analysis of variance for two-way data.

Two-Way Model with Interaction Effects

A more general model for two-way classified data includes interaction effects in the model. Suppose the γ_i and τ_j are the effects of two treatment factors, A and B, with a levels of factor A and b levels of factor B. Let the interaction effects between the two factors be represented by $(\gamma\tau)_{ij}$ and assume there are r observations in each cell, $k = 1, \ldots, r$. The linear model is

$$Y_{ijk} = \mu + \gamma_i + \tau_j + (\gamma\tau)_{ij} + \epsilon_{ijk}, \tag{9.26}$$

where $i = 1, \ldots, a$ and $j = 1, \ldots, b$. In matrix notation, $\boldsymbol{\beta}$ contains $(1 + a + b + ab) = (a + 1)(b + 1)$ parameters and $\boldsymbol{X}$ contains an equal number of columns. The number of rows of $\boldsymbol{X}$ will equal the number of observations, $n = abr$. The r observations from the same treatment combination have the same expectation (equation 9.26), so that there will be ab distinct rows in $\boldsymbol{X}$ with r repeats of each.

Example 9.3

For illustration, assume $a = 2$ and $b = 4$. Then $\boldsymbol{X}$ contains 15 columns

and 8 distinct rows. Each of the 8 rows will be repeated r times. Then,

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad (9.27)$$

where only the 8 distinct rows of $\boldsymbol{X}$ are shown.

The first 7 columns of $\boldsymbol{X}$ are as defined in equation 9.22. The last 8 columns are the dummy variables for the interaction effects. The dummy variable for $(\gamma\tau)_{ij}$ takes the value 1 if the observation is from the ijth treatment combination, and 0 otherwise. The dummy variable for $(\gamma\tau)_{ij}$ can also be obtained as the element-by-element product of the dummy variables for the corresponding γ_i and τ_j effects. (This is a general result that extends to higher-order interaction effects.) Although $\boldsymbol{X}$ contains 15 columns, its rank is only 8. (The rank of $\boldsymbol{X}$ cannot be greater than the number of linearly independent rows.) Thus, there must be 7 linear dependencies among the columns of $\boldsymbol{X}$. These dependencies would have to be identified if the model were to be reparameterized. Note that each of the first 7 columns can be obtained as a linear combination of the last 8 columns. The generalized inverse approach, however, uses $\boldsymbol{X}$ as defined. ■

Computing Load

The size of $\boldsymbol{X}$ increases very rapidly as additional factors and particularly their interactions are added to the model. The number of columns of $\boldsymbol{X}$ required for each set of interaction effects is the product of the number of levels of all the factors in the interaction. The total number of parameters in a model with class variables and their interactions is the product of the number of levels plus 1 of all class variables in the model; for example, $(2 + 1)(4 + 1) = 15$ in Example 9.3. It is not uncommon for the full $\boldsymbol{X}$ matrix of a reasonably sized experiment to have more than 100 columns. The computational load of finding the generalized inverse and operating on this very large $\boldsymbol{X}$ matrix would be exorbitant without modern computers.

On the other hand, the conventional analysis of variance formulas, which result from the least squares analysis of balanced data, are computationally very efficient. Very large models can be easily analyzed. The more general approach has been introduced to demonstrate the link between least squares regression analysis and the conventional analyses of variance, and to set the stage for the analysis of unbalanced data (Chapter 17).

9.6 Class Variables To Test Homogeneity of Regressions

Consider the situation where two or more subsets of data are available, each of which provides information on the dependent variable of interest and the potential predictor variables. The subsets of data originate from different levels of one or more class variables. For example, data relating yield in corn to levels of nitrogen and phosphorous fertilization may be available for several corn hybrids grown in several environments. Yield is the dependent variable, amount of nitrogen fertilizer and amount of phosphorous fertilizer are independent variables, and "hybrid" and "environment" are two class variables.

The objective is to model the response of yield to changing rates of nitrogen and phosphorous fertilization. The question is whether a single regression equation will adequately describe the relationship for all hybrids and environments or will different regressions be required for each hybrid–environment combination. The most complete description of the response (the best fit to the data) would be obtained by allowing each combination to have its own regression equation. This would be inefficient, however, if the responses were similar over all groups; the researcher would be estimating more parameters than necessary. On the other hand, a single regression equation to represent the response for all groups will not characterize any one group as well and could be very misleading if the relationships differed among groups. The simplicity of the single regression equation is to be preferred if it can be justified. Intermediate models may allow a common regression for some independent variables but require others to have different regression coefficients for different subsets of data.

Illustration

The decision to use a regression coefficient for each subset or a common regression coefficient for all subsets is based on the test of homogeneity of regression coefficients over levels of the class variable. The test of homogeneity is illustrated assuming a linear relationship between a dependent variable and an independent variable. The general method extends to any number of independent variables and any functional relationship.

Defining the Model

Suppose the data consist of t groups with n_i observations in each group. There will be $\sum n_i = n$ data points, each consisting of an observation on the Y, X, and the class variable identifying the group from which the observations came. The most general model for this situation allows each group to have its own intercept and slope coefficient. The separate models can be written as

$$\begin{aligned} \text{Group 1}: \; Y_{1j} &= \beta_{10} + \beta_{11} X_{1j} + \epsilon_{1j} \\ \text{Group 2}: \; Y_{2j} &= \beta_{20} + \beta_{21} X_{2j} + \epsilon_{2j} \\ &\vdots \end{aligned} \tag{9.28}$$

$$\text{Group } t: \; Y_{tj} \;=\; \beta_{t0} + \beta_{t1} X_{tj} + \epsilon_{tj}.$$

If the subscript i designates the group code, or the level of the class variable, the models can be written as

$$Y_{ij} = \beta_{i0} + \beta_{i1} X_{ij} + \epsilon_{ij}, \tag{9.29}$$

where $i = 1, \ldots, t$ and $j = 1, \ldots, n_i$. This model contains $2t$ parameters: t β_0-parameters and t β_1-parameters. The random errors ϵ_{ij} for all groups are assumed to be normally and independently distributed with zero mean and common variance σ^2.

Model in Matrix Notation

The model encompassing all t groups is written in matrix notation by using t dummy variables to identify the levels of the class variable "group." Let

$$\begin{aligned}
W_{1_{ij}} &= \begin{cases} 1 & \text{if the observation is from group 1} \\ 0 & \text{otherwise} \end{cases} \\
W_{2_{ij}} &= \begin{cases} 1 & \text{if the observation is from group 2} \\ 0 & \text{otherwise} \end{cases} \\
&\vdots \\
W_{t_{ij}} &= \begin{cases} 1 & \text{if the observation is from group } t \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

Then

$$\begin{aligned}
Y_{ij} &= W_{1_{ij}}(\beta_{10} + \beta_{11} X_{1j}) + W_{2_{ij}}(\beta_{20} + \beta_{21} X_{2j}) \\
&\quad + \cdots + W_{t_{ij}}(\beta_{t0} + \beta_{t1} X_{tj}) + \epsilon_{ij} \\
&= \beta_{10} W_{1_{ij}} + \beta_{11}(W_{1_{ij}} X_{1j}) + \beta_{20} W_{2_{ij}} + \beta_{21}(W_{2_{ij}} X_{2j}) \\
&\quad + \cdots + \beta_{t0} W_{t_{ij}} + \beta_{t1}(W_{t_{ij}} X_{tj}) + \epsilon_{ij}
\end{aligned} \tag{9.30}$$

or

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{9.31}$$

where

$$\boldsymbol{X} = \begin{bmatrix}
1 & X_{11} & 0 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
1 & X_{1n_1} & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 1 & X_{21} & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 1 & X_{2n_2} & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & 1 & X_{t1} \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & 1 & X_{tn_t}
\end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix}
\beta_{10} \\ \beta_{11} \\ \beta_{20} \\ \beta_{21} \\ \vdots \\ \beta_{t0} \\ \beta_{t1}
\end{bmatrix}.$$

The odd-numbered columns of $\boldsymbol{X}$ are the dummy variables and provide for the $t\beta_0$s in the model. The even-numbered columns are the elementwise products of the dummy variables and the independent variable. These bring in the level of the X variable times the appropriate β_{i1} only when the observations are from the ith group. We assume that $\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i.})^2 > 0$, for $i = 1, \ldots, t$. That is, within each group, the X variable takes at least two distinct values. This is a full-rank model; $r(\boldsymbol{X}) = 2t$ and there are $2t$ parameters to be estimated.

The two columns associated with any particular group are orthogonal to all other columns. Therefore, the results of the least squares regression using this large model to encompass all groups are identical to the results that would be obtained if each group were analyzed separately. The SS(Model) will have $2t$ degrees of freedom and will be the sum of the SS(Model) quantities from the separate analyses. The residual mean square from this full analysis will be identical to the pooled residual mean squares from the separate analyses. The pooled residual mean square is the best estimate of σ^2 unless a pure error estimate is available.

Testing Homogeneity of Slopes

There are several tests of homogeneity of interest. The test of homogeneity of slopes of regression lines is most common in the context of allowing the intercepts to be different. Thus, the different groups are allowed to have different mean levels of Y but are required to have the same response to changes in the independent variable. The null hypothesis is the composite hypothesis

$$H_0 : \beta_{11} = \beta_{21} = \cdots = \beta_{t1}. \tag{9.32}$$

The difference in SS(Res) for full and reduced models is used to test this hypothesis of common β_1. The reduced model is obtained from equation 9.30 by replacing the t different slopes β_{i1} with a common slope β_1:

$$Y_{ij} = \beta_{10}W_{1ij} + \beta_{20}W_{2ij} + \cdots + \beta_{t0}W_{tij} + \beta_1 X_{ij} + \epsilon_{ij}. \tag{9.33}$$

The independent variable is no longer multiplied by the dummy variables W_i. The $\boldsymbol{X}$ matrix for the reduced model consists of t columns for the dummy variables plus one column of the observations on the independent variable; the X_{ij} are no longer separated by groups. The rank of $\boldsymbol{X}$ in the reduced model is $t + 1$, t degrees of freedom for estimating the t intercepts and 1 degree of freedom for estimating the common slope.

The difference between the residual sum of squares for the full model and the residual sum of squares for the reduced model,

$$Q = \mathrm{SS}(\mathrm{Res}_{\mathrm{reduced}}) - \mathrm{SS}(\mathrm{Res}_{\mathrm{full}}) \tag{9.34}$$

has $(t - 1)$ degrees of freedom, $(\sum n_i - t - 1) - (\sum n_i - 2t)$. This is the appropriate sum of squares for testing the composite null hypothesis given in equation 9.32. The test statistic is an F-ratio with $Q/(t - 1)$ as the

numerator and the residual mean square from the full model as the denominator. A nonsignificant F-ratio leads to the conclusion that the regressions of Y on X for the several groups are adequately represented by a series of parallel lines. The differences in the "heights" of the lines reflect differences of the intercepts among the groups.

Testing Homogeneity of Intercepts

The same general procedure can be used to test other hypotheses. The composite null hypothesis of common intercepts β_{i0} in the presence of heterogeneous slopes is not a meaningful hypothesis unless there is some logic in expecting the regressions for all groups to converge to a common value of Y at $X = 0$. (The intercept is usually defined as the value of Y at $X = 0$ or, if the Xs are centered, the value of Y at $X = \overline{X}$. The origin of the independent variable can be shifted by adding a constant to or subtracting a constant from each value of X so that it is possible to test convergence of the regression lines at any chosen value of X.) It is quite common, however, to test homogeneity of intercepts after having decided that the groups have common slope. For this test, the reduced model with t β_{i0}-parameters and common β_1 (equation 9.33) becomes the full model. The new reduced model for $H_0 : \beta_{10} = \beta_{20} = \cdots = \beta_{t0}$ is the simple regression model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij}. \tag{9.35}$$

The $\boldsymbol{X}$ matrix for this reduced model has only two columns, the column of ones for the intercept and the column of X_{ij}. The difference in residual sums of squares for this model and the full model will have $t-1$ degrees of freedom and is appropriate for testing the null hypothesis of equal intercepts in the presence of equal slopes.

A numerical example showing the tests of homogeneity of regression coefficients is presented in Section 9.8.

Testing Equality of Variances

In the model in equation 9.29, we have *assumed* that the variance of ϵ_{ij} is the same for all t groups. Bartlett (1937) proposed a general test for testing the equality of variances of t normal populations. Let $s_1^2, \ldots, s_t^2$ be the sample variances with $\nu_1, \ldots, \nu_t$ degrees of freedom, respectively, from t normal populations. Bartlett's test statistic is given by

$$B = \frac{1}{C}\left[\nu \log(\text{MSE}) - \sum_{i=1}^{t} \nu_i \log(s_i^2)\right], \tag{9.36}$$

where

$$C = 1 + \frac{1}{3(t-1)}\left[\sum_{i=1}^{t} {\nu_i}^{-1} - \nu^{-1}\right], \quad \text{and}$$

$$\text{MSE} = \frac{1}{\nu}\sum_{i=1}^{t} \nu_i\, s_i^2$$

and $\nu = \sum \nu_i$. In the model in equation 9.29, s_i^2 represents the residual mean square error from the simple linear regression for the ith group so

TABLE 9.3. *Pre-test and post-test scores from the listening–reading skills study at the Governor Morehead School. The test scores came from the Gilmore Oral Reading Test. (Used with permission of Dr. Larry Nelson.)*

Treatments	*Pre-Test Score (X)*	*Post-Test Score (Y)*
$T1$	89	87
	82	86
	88	94
	94	96
$T2$	89	84
	90	94
	91	97
	92	93
$T3$	89	96
	99	97
	84	100
	87	98

that $\nu_i = n_i - 2$. MSE is the residual mean square error from the full model with $\nu = \sum_{i=1}^{t}(n_i - 2)$ degrees of freedom. We reject the null hypothesis that the variances of ϵ_{ij} are equal among groups if the test statistic B is larger than $\chi^2_{(t-1;\alpha)}$.

Example 9.4

A study was conducted at the Governor Morehead School in Raleigh, North Carolina to evaluate some techniques intended to improve "listening–reading" skills of subjects who were visually impaired. The listening–reading treatments were: (1) instruction in listening techniques plus practice listening to selected readings; (2) the same as (1) but with copies of the selected readings in Braille; and (3) the same as (1) but with copies of selected readings in ink print. The number of individuals per group was four. The response data are measures of reading accuracy as measured by the Gilmore Oral Reading Test. Both pre- and post-test data were taken. The pre-test scores are intended to serve as a covariable to adjust for differences in the abilities of the subjects before the study. The data are summarized in Table 9.3.

The ultimate intent of the study was to test for differences among treatments as measured by the post-test scores *after* taking into account differences in ability levels of the individuals as measured by the pre-test scores. However, we use this study to illustrate the test of homogeneity of regressions over the three treatment groups. First, we test the homogeneity of the slope coefficients from the regression of post-test scores on pre-test scores. We fit the full model in equation 9.29 allowing each treatment group to

have its own slope and intercept. The residual sum of squares from this model is observed to be SS(Res)= 86.0842 with 6 degrees of freedom. To test the hypothesis in equation 9.32 that the slopes are equal, we fit the reduced model in equation 9.33 and compute

$$\begin{aligned} F &= \frac{[\text{SS}(\text{Res}_{\text{reduced}}) - \text{SS}(\text{Res}_{\text{full}})]/(8-6)}{\text{SS}(\text{Res}_{\text{full}})/6} \\ &= \frac{(164.2775 - 86.0842)/2}{86.0842/6} \\ &= 2.73. \end{aligned}$$

Comparing this value to $F_{(.05;2,6)} = 14.54$, we fail to reject the null hypothesis of common slopes among the three treatment groups. Now assuming that the model in equation 9.33 is the full model, we test the hypothesis that the three intercepts are equal. The F-statistic is given by

$$F = \frac{(269.9488 - 164.2775)/2}{164.2775/8} = 2.57,$$

where 269.9488 is the residual sum of squares of the reduced model given in equation 9.35. Comparing $F = 2.57$ with $F_{(.05;2,8)} = 11.044$, we fail to reject the null hypothesis that the intercepts are the same for all three treatment groups, assuming that they have common slopes.

A joint test of the hypothesis that the intercepts and the slopes are constant among the groups is given by

$$F = \frac{(269.9488 - 86.0842)/4}{86.0842/6} = 3.20.$$

Comparing this value with $F_{(.05;4,6)} = 12.04$, we fail to reject the null hypothesis that a single line is adequate for all three treatment groups. In fact, in this particular example, it is observed that neither the treatment nor the pre-test score have a significant effect on the post-treatment score. Given the small number of degrees of freedom for error, the test statistics may not be powerful enough to detect differences among the treatment groups and the significance of the pre-test score. ■

Example 9.5

The tests of significance in Example 9.4 assume that the variance of the errors in the model is the same for all three groups. Estimating the simple linear regression for the three groups separately, we obtain the residual mean squares $s_1^2 = 15.63$, $s_2^2 = 24.5$, and $s_3^2 = 2.91$ each with two degrees of freedom. Bartlett's test statistic in equation 9.36 is 1.596, which is not significant since $\chi^2_{(.05;2)} = 10.06$. Therefore, there is not enough evidence to conclude that the variances are different among the three groups.

These examples provide a good illustration of the importance of sample size in experimentation. The lack of significance of the tests in Example

9.4, and even more so in the test of variances in Example 9.5, is as likely to be due to lack of power of the tests (due to small sample size) as to the absence of true differences. In particular, an estimate of variance with only two degrees of freedom is essentially meaningless. ■

9.7 Analysis of Covariance

Covariance to Improve Precision

The classical purpose of the analysis of covariance is to improve the precision of the experiment by statistical control of variation among experimental units. A useful covariate identifies variation among the experimental units that is also associated with variation in the dependent variable. For example, variation in density of plants in the experimental units causes variation in yield of most plant species, or variation in age or body weight of animals often causes variation in rate of gain in feeding trials. The covariance analysis removes this source of variation from experimental error and adjusts the treatment means for differences attributable to the covariate. For this purpose, the covariate should not be affected by the treatments. Otherwise, adjustment for the covariate will bias the estimates of treatment effects and possibly lead to incorrect inferences.

As an illustration, consider a study to measure the effects of nutrient levels on the growth rate of a species of bacteria. It is well known that temperature has an effect on growth rate. Therefore, any differences in temperature of the experimental units can be expected to cause differences in growth rates even if the experimental units receive the same nutrient treatment. Such differences will inflate experimental error and, to the extent the nutrient groups differ in mean temperature, cause biases in the observed treatment effects. Suppose the available resources do not permit sufficient control of temperature to rule out these effects. Covariance analysis, with the measured temperature of each experimental unit as the covariate, could be used to adjust the observed growth rates to a common temperature.

Covariance to Interpret Treatment Effects

A second use of the analysis of covariance is as an aid in the interpretation of treatment effects on a primary response variable. In this case, the covariate is another response variable that may be involved in the response of the primary response variable. The questions to be addressed by the covariance analysis are whether the treatment effects on the primary response variable are essentially independent of those on the secondary variable (the covariate) and, if not, how much of the effect on the primary response variable might be attributed to the indirect effects of the treatments on the covariate. For this purpose, it is quite likely that the covariate will be affected by the treatments. (In cases such as this, a multivariate analysis of variance of the two response variables would be a more appropriate analysis.)

Analysis of covariance is a special case of regression analysis where both continuous and class variables are used. The class variables take into account the experimental design features as discussed earlier in this chapter. The covariate will (almost) always be a continuous variable for which the experimental results are to be "adjusted."

Two-Way Model with Covariate

The usual linear model for the analysis of covariance for a randomized complete block design is

$$\begin{aligned} Y_{ij} &= \mu + \tau_i + \gamma_j + \beta(X_{ij} - \overline{X}_{..}) + \epsilon_{ij} \\ &\quad \text{for } i = 1, \ldots, a \text{ treatments and } j = 1, \ldots, b \text{ blocks,} \end{aligned} \tag{9.37}$$

where the term $\beta(X_{ij} - \overline{X}_{..})$ has been added to the RCB model, equation 9.21, to incorporate the effect of the covariate X_{ij} on the dependent variable. The covariate is expressed in terms of the deviations about its sample mean $\overline{X}_{..}$. This emphasizes that it is the variation in the covariate that is of interest, and simplifies the subsequent adjustment of the treatment means. Equation 9.37 is the simplest form in which a covariate effect can be included in a model—one covariate acting in a linear manner. The covariate model can be extended to include more than one covariate and more complicated relationships.

Model in Matrix Notation

The covariance model is written in matrix form by augmenting the design matrix $\boldsymbol{X}$ and parameter vector $\boldsymbol{\beta}$ for the appropriate experimental design. $\boldsymbol{X}$ is expanded to include a column vector of $(X_{ij} - \overline{X}_{..})$. $\boldsymbol{\beta}$ is expanded to include the regression coefficient for the covariate β. The ordering of the observations for the covariate must be identical to the ordering of observations in $\boldsymbol{Y}$. The numerical example in Section 9.8 illustrates $\boldsymbol{X}$ and $\boldsymbol{\beta}$.

Quantities of Interest

The covariance model is of less than full rank, because the design matrix to which the covariate vector was appended is singular. None of the singularities, however, involves the covariate vector. Reparameterization or the generalized inverse approach is used to obtain the relevant sums of squares and to estimate the estimable functions of the parameters. The quantities of primary interest are:

1. partial sums of squares attributable to the covariate and to differences among the treatments,
2. estimate of experimental error after removal of the variation attributable to the covariate, and
3. estimated treatment means and mean contrasts after adjustment to a common level of the covariate.

The covariance analysis is first discussed as if the purpose of the analysis were to increase precision of the experiment. Then, the key changes in interpretation are noted for the case when covariance analysis is being used to help interpret the treatment effects.

TABLE 9.4. *Partial sums of squares and mean squares from the analysis of covariance for a randomized complete block design with b blocks and t treatments.*

Source	*d.f.*	*Partial SS*[a]	*MS*
Total	$bt-1$	$\boldsymbol{Y}'\boldsymbol{Y} - C.F.$	
Blocks	$b-1$	$R(\gamma'\|\boldsymbol{\tau}'\ \beta\ \mu)$	
Treatments	$t-1$	$R(\boldsymbol{\tau}'\|\boldsymbol{\gamma}'\ \beta\ \mu)$	
Covariate	1	$R(\beta\|\boldsymbol{\gamma}'\ \boldsymbol{\tau}'\ \mu)$	
Residual	$(b-1)(t-1)-1$	$\boldsymbol{Y}'\boldsymbol{Y} - R(\boldsymbol{\gamma}'\ \boldsymbol{\tau}'\ \beta\ \mu)$	s^2

[a] γ' and τ' designate the row vectors of effects for the class variables "blocks" and "treatments," respectively.

Analysis of Variance

The partial sums of squares for the class variables, "blocks" and "treatments" in the RCB, and the covariate are shown in Table 9.4. These are *not* additive partitions of the total sum of squares even when the data are balanced. The covariate destroys the orthogonality that might have been present in the basic experimental design. The error variance is estimated from the residual mean square, the "block by treatment" interaction mean square after adjustment for the covariate. The degrees of freedom for residual reflect the loss of one degree of freedom for estimating β for the covariate.

This model and analysis assume that the basic datum is one observation on the ijth experimental unit, so that the residual mean square from the regression analysis is also the error variance. If the data involve multiple samples from each experimental unit, the residual mean square in Table 9.4 will contain both experimental error and sampling error.

Covariance with Sampling

A simple way to approach analysis of covariance in the presence of sampling is to do the analysis of covariance based on the experimental unit means. The errors associated with the experimental unit means are independent and identically distributed with constant variance. Another procedure would be to use a more general model that recognizes the correlated error structure introduced by the multiple sampling on the same experimental unit. (See Chapter 18 for mixed models.)

Testing the Effect of the Covariate

The presence of the covariate reduces the residual sum of squares by the amount $R(\beta|\boldsymbol{\gamma}'\,\boldsymbol{\tau}'\,\mu)$, the partial sum of squares attributable to the covariate. This reflects the direct impact of the covariate on the magnitude of σ^2 and, hence, on the precision of the experiment. The null hypothesis that the covariate has no effect, $H_0 : \beta = 0$, is tested with

$$F = \frac{R(\beta|\boldsymbol{\gamma}'\,\boldsymbol{\tau}'\,\mu)}{s^2}, \tag{9.38}$$

which has 1 and $[(b-1)(t-1)-1]$ degrees of freedom. If F is not significant at the chosen α, it is concluded that the covariate is not important in controlling precision and the covariance analysis is abandoned. Interpretations

are based on the conventional analysis of variance. If the null hypothesis is rejected, it is concluded that the covariate is effective in increasing precision and the covariance analysis is continued to obtain estimates of treatment means and contrasts adjusted for the effects of the covariate. The residual mean square is the estimate of σ^2 for all subsequent computations.

Testing Treatment Effects

The appropriate sum of squares for testing the composite null hypothesis that all effects for a class variable are zero is the partial sum of squares for that class variable $R(\boldsymbol{\tau}'|\boldsymbol{\gamma}'\,\beta\,\mu)$ or $R(\boldsymbol{\gamma}'|\boldsymbol{\tau}'\,\beta\,\mu)$. As always, these sums of squares can be computed either by defining an appropriate $\boldsymbol{K}'$ for the general linear hypothesis or by the difference between residual sums of squares for full and reduced models. The partial sum of squares for a class variable adjusted for the covariate measures the variability among the levels of the class variable as if all observations had occurred at the mean level of the covariate. The null hypothesis that all treatment effects are zero is tested by

$$F = \frac{R(\boldsymbol{\tau}'|\boldsymbol{\gamma}'\ \beta\ \mu)/(t-1)}{s^2}. \tag{9.39}$$

Unadjusted Treatment Means

The conventional, unadjusted treatment means are computed as simple averages of the observations in each treatment. The vector of unadjusted treatment means can be written as

$$\overline{\boldsymbol{Y}} = \boldsymbol{T}'\boldsymbol{Y}, \tag{9.40}$$

where $\boldsymbol{T}$ is defined as the matrix of the t treatment dummy variables with each divided by the number of observations in the treatment. Thus, $\boldsymbol{T}$ is

$$\boldsymbol{T} = \frac{1}{b}\begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & & 0 \\ & & \ddots & \\ 0 & 0 & & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \tag{9.41}$$

when there are b observations per treatment. The expectation of $\overline{\boldsymbol{Y}}$ is

$$\mathcal{E}(\overline{\boldsymbol{Y}}) = \boldsymbol{T}'\boldsymbol{X}\boldsymbol{\beta}. \tag{9.42}$$

If the model includes a covariate, the expectation of the ith mean contains the term $\beta(\overline{X}_{i.} - \overline{X}_{..})$ in addition to the appropriate linear function of the

other model effects. Because of this term, comparisons among the treatment means include differences due to the covariate unless $\beta = 0$ or $\overline{X}_{i.}$ is the same for all treatments being compared.

Adjusted Treatment Means

The **adjusted treatment means** are designed to remove this confounding. Adjustment is accomplished either by estimating directly from $\boldsymbol{\beta}^0$ the linear function of the parameters of interest, or by subtracting an estimate of the bias term from each unadjusted treatment mean. The linear functions of the parameters that need to be estimated are appropriately defined by equation 9.42 if $\boldsymbol{X}$ is redefined by replacing the column of covariate values with a column of zeros. If this redefined $\boldsymbol{X}$ is labeled $\boldsymbol{X}_c$, the linear functions to be estimated by the adjusted treatment means are

$$\mathcal{E}(\overline{\boldsymbol{Y}}_{\text{adj}}) = \boldsymbol{T}'\boldsymbol{X}_c\boldsymbol{\beta}, \tag{9.43}$$

where $\overline{\boldsymbol{Y}}_{\text{adj}}$ denotes the vector of adjusted treatment means. The least squares estimate of the *adjusted* treatment means is given by the same linear function of the least squares solution $\boldsymbol{\beta}^0$,

$$\overline{\boldsymbol{Y}}_{\text{adj}} = \boldsymbol{T}'\boldsymbol{X}_c\boldsymbol{\beta}^0. \tag{9.44}$$

The adjusted treatment means are estimates of the treatment means for the case where all treatments have the mean level of the covariate, $\overline{X}_{i.} = \overline{X}_{..}$ for all i. The adjustment can be made to any level of the covariate, say C, by defining $\boldsymbol{X}_c$ to be the matrix with the column vector of covariate values replaced with $(C - \overline{X}_{..})$ rather than with zeros.

Alternatively, each adjusted treatment mean can be obtained by removing the bias $\beta(\overline{X}_{i.} - \overline{X}_{..})$ from the corresponding unadjusted treatment mean. This leads to the more traditional method of computing the adjusted treatment means:

$$\overline{Y}_{\text{adj}_{i.}} = \overline{Y}_{i.} - \widehat{\beta}(\overline{X}_{i.} - \overline{X}_{..}). \tag{9.45}$$

The covariance adjustment is illustrated in Figure 9.1. The diagonal line passing through the point $(\overline{X}_{..}, \overline{Y}_{..})$ is the regression line with slope $\widehat{\beta}$ relating the dependent variable to the covariate. The original observations are represented with $\times$s. The adjustment can be viewed as moving each observation along a path parallel to the fitted regression line from the observed value of the covariate $X = X_{ij}$ to the common value $X = \overline{X}_{..}$. The dots on the vertical line at $X = \overline{X}_{..}$ represent the adjusted observations. The amount each Y_{ij} is adjusted during this shift is determined by the slope of the regression line and the change in X,

$$Y_{\text{adj}_{ij}} = Y_{ij} - \widehat{\beta}(X_{ij} - \overline{X}_{..}).$$

Averaging the adjusted observations within each treatment gives the adjusted treatment means, equation 9.45.

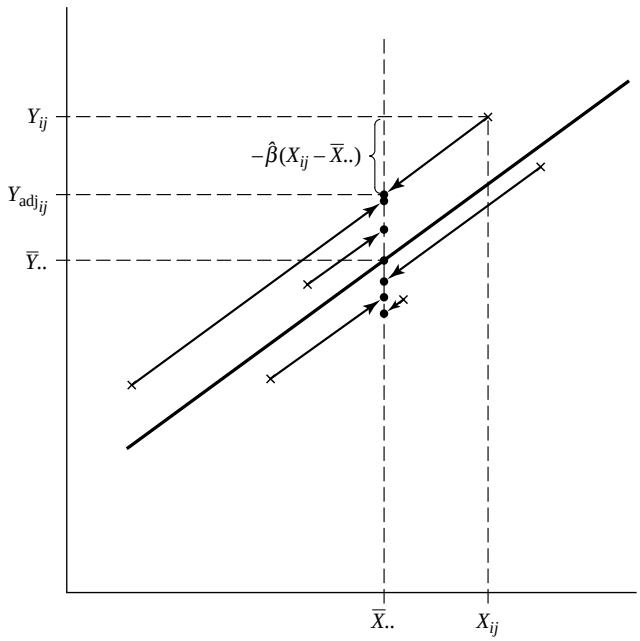

FIGURE 9.1. *Illustration of the adjustment of the response variable Y for differences in the covariate X.*

Variances of Adjusted Treatment Means

The variance–covariance matrix of the adjusted treatment means follows directly from the matrix equation for the variance of a linear function. Thus,

$$\mathbf{Var}(\overline{\boldsymbol{Y}}_{\text{adj}}) = (\boldsymbol{T}'\boldsymbol{X}_c)(\boldsymbol{X}'\boldsymbol{X})^{-}(\boldsymbol{T}'\boldsymbol{X}_c)'\sigma^2. \tag{9.46}$$

The variances of the adjusted treatment means, the diagonal elements of equation 9.46, simplify to the classical formula for the variance:

$$\sigma^2(\overline{Y}_{\text{adj}_{i.}}) = \left[\frac{1}{n} + \frac{(\overline{X}_{i.} - \overline{X}_{..})^2}{E_{xx}}\right]\sigma^2, \tag{9.47}$$

where E_{xx} is the residual sum of squares from the RCB analysis of variance *of the covariate*. That is, $E_{xx} = \sum_{i=1}^{a}\sum_{j=1}^{b}[X_{ij} - \overline{X}_{i.} - \overline{X}_{.j} + \overline{X}_{..}]^2$.

Covariance to Help Interpret Treatment Effects

When the covariance analysis is being used to aid interpretation of the treatment effects, the primary interest is in comparison of the treatment means and sums of squares before and after adjustment for the covariate. The adjustment of the means and sums of squares is *not* viewed as a method of obtaining *unbiased* estimates of treatment effects. Rather, the changes in the means and sums of squares provide some indication of the proportion of the treatment effects that can be viewed as direct effects on Y versus possible indirect effects on Y through X, or through some other variable that in turn affects both X and Y. For example, highly significant treatment effects that remain about the same after adjustment for X would suggest that most of the treatment effects on Y are essentially independent of any treatment effects on X. On the other hand, dramatic changes in the treatment effects with adjustment would suggest that X and Y are closely linked in the system being studied so that the responses of both variables to the treatments are highly correlated.

The test of the null hypothesis $H_0 : \beta = 0$ is a test of the hypothesis that the correlation between the residuals for X and the residuals for Y is zero, after both have been adjusted for block and treatment effects. If the covariate was chosen because it was expected to have a direct impact on Y, then β would be expected to be nonzero and this test would serve only as a confirmation of some link between the two variables. A nonsignificant test would suggest that the link between the two variables is very weak, or the power of the test is not adequate to detect the link. In either case, any effort devoted to interpretation of the adjusted treatment means and sums of squares would not be very productive.

9.8 Numerical Examples

Two examples are used. The first example combines several concepts covered in this chapter:

1. analysis of variance as a regression problem including reparameterization;

2. use of dummy variables to test homogeneity of regressions; and

3. analysis of covariance to aid in the interpretation of treatment effects.

The covariable in the first example can be viewed as another response variable and is expected to be affected by the treatments. A multivariate analysis of variance of the two response variables would be a more appropriate analysis.

The second example illustrates the more classical use of covariance and uses a generalized inverse solution to the normal equations.

Example 9.6

The purpose of this study was to compare ascorbic acid content in cabbage from two genetic lines (cultivars) planted on three different dates (Table 9.5). The experimental design was a completely random design with $r = 10$ experimental units for each combination of planting date and genetic line, for a total of 60 observations. It was anticipated that ascorbic acid content might be dependent on the size of the cabbage head; hence, head weight was recorded for possible use as a covariate. (The data are from the files of the late Dr. Gertrude M. Cox.)

Ascorbic acid content is the dependent variable of interest and head weight is used as a covariate. The variables "date" and "line" are treated as class variables. The first analysis is the conventional analysis of variance for the factorial experiment. Then, in anticipation of the analysis of covariance, the homogeneity of regression coefficients, relating ascorbic acid content to head size, over the six date-line treatment combinations is tested. Finally, the analysis of covariance is run.

The purpose of the covariance analysis in this example is as an aid in interpreting the effects of planting date and genetic line on ascorbic acid content, rather than for control of random variation among the experimental units. It is expected that the covariable head weight will be affected by the date and line treatment factors. Hence, adjustment of ascorbic acid content to a common head weight would redefine treatment effects. When the response variable and the covariate are affected by the treatment, a multivariate approach that studies the treatment effects is preferred.

9.8.1 Analysis of Variance

The conventional model for a factorial set of treatments in a completely random design is

$$Y_{ijk} = \mu + \gamma_i + \tau_j + (\gamma\tau)_{ij} + \epsilon_{ijk}, \tag{9.48}$$

TABLE 9.5. *Head weight and ascorbic acid content for two cabbage varieties on three planting dates.*

	Planting Date					
	16		20		21	
Line Number	*Head Wt.*	*Ascorbic Content*	*Head Wt.*	*Ascorbic Content*	*Head Wt.*	*Ascorbic Content*
39	2.5	51	3.0	65	2.2	54
	2.2	55	2.8	52	1.8	59
	3.1	45	2.8	41	1.6	66
	4.3	42	2.7	51	2.1	54
	2.5	53	2.6	41	3.3	45
	4.3	50	2.8	45	3.8	49
	3.8	50	2.6	51	3.2	49
	4.3	52	2.6	45	3.6	55
	1.7	56	2.6	61	4.2	49
	3.1	49	3.5	42	1.6	68
52	2.0	58	4.0	52	1.5	78
	2.4	55	2.8	70	1.4	75
	1.9	67	3.1	57	1.7	70
	2.8	61	4.2	58	1.3	84
	1.7	67	3.7	47	1.7	71
	3.2	68	3.0	56	1.6	72
	2.0	58	2.2	72	1.4	62
	2.2	63	2.3	63	1.0	68
	2.2	56	3.8	54	1.5	66
	2.2	72	2.0	60	1.6	72

where γ_i are the "date" effects ($i = 1, 2, 3$), τ_j are the "line" effects ($j = 1, 2$), and $(\gamma\tau)_{ij}$ are the "date by line" interaction effects. This model contains 12 parameters to define only 6 group means. Thus, there are 6 linear dependencies in the model and a full rank reparameterization requires 6 constraints. There must be 1 constraint on the γ_i, 1 on the τ_j, and 4 on the $(\gamma\tau)_{ij}$.

For this illustration, the means model is used as the reparameterized model and then the general linear hypothesis is used to partition the variation among the six treatments into "date," "line," and "date by line" sums of squares. Thus, the (full-rank) model for the analysis of variance is

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \tag{9.49}$$

where μ_{ij} is the true mean of the ijth date-line group. In this model $\boldsymbol{X}$ is of order (60×6) where each column is a dummy variable showing the incidence of the observations for one of the date-line groups. That is, the ijth dummy variable takes the value one if the observation is from the ijth date-line group; otherwise the dummy variable takes the value zero. It is assumed that the elements of $\boldsymbol{\beta}^*$ are in the order

$$\boldsymbol{\beta}^{*\prime} = (\,\mu_{11} \quad \mu_{12} \quad \mu_{21} \quad \mu_{22} \quad \mu_{31} \quad \mu_{32}\,).$$

The least squares analysis using this model gives SS(Model) = 205,041.9 with 6 degrees of freedom and SS(Residual) = 2,491.1 with 54 degrees of freedom. The least squares estimates of μ_{ij} are the group means:

$$\widehat{\boldsymbol{\beta}}^{*\prime} = (\,50.3 \quad 62.5 \quad 49.4 \quad 58.9 \quad 54.8 \quad 71.8\,).$$

Each $\widehat{\mu}_{ij}$ is estimating $\mu + \gamma_i + \tau_j + (\gamma\tau)_{ij}$, the mean of the treatment group in terms of the original parameters. These are the estimated group means for ascorbic acid ignoring any differences in head weight since the model does not include the covariate.

The partitions of SS(Model) are obtained by appropriate definition of $\boldsymbol{K}'$ for general linear hypotheses on the μ_{ij}. For this purpose, it is helpful to view the μ_{ij} as a 3×2 "date by line" table of means. The marginal means for this table $\overline{\mu}_{i.}$ and $\overline{\mu}_{.j}$ represent the "date" means and the "line" means, respectively. For each sum of squares to be computed, the appropriate null hypothesis is stated in terms of the μ_{ij}, the appropriate $\boldsymbol{K}'$ is defined for the null hypothesis, and the sum of squares Q computed using the general linear hypothesis, equation 4.38, is given. In all hypotheses $\boldsymbol{m} = \boldsymbol{0}$ and Q is computed as

$$Q = (\boldsymbol{K}'\widehat{\boldsymbol{\beta}}^*)'[\boldsymbol{K}'(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\boldsymbol{K}]^{-1}(\boldsymbol{K}'\widehat{\boldsymbol{\beta}}^*).$$

1. *Correction factor:* The sum of squares due to the correction for the mean, the correction factor, measures the deviation of the overall

mean $\overline{\mu_{..}}$ from zero. The overall mean is zero only if the sum of the μ_{ij} is zero. Therefore,

$$\begin{aligned} H_0: &\quad \overline{\mu}_{..} = 0 \text{ or } \sum\sum \mu_{ij} = 0, \\ \boldsymbol{K}_1' &= (1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1), \\ r(\boldsymbol{K}_1) &= 1 \text{ and} \\ Q_1 &= 201,492.1 \text{ with 1 degree of freedom.} \end{aligned} \tag{9.50}$$

2. *Sum of squares for "dates":* The hypothesis of no date effects is equivalent to the hypothesis that the three marginal means $\overline{\mu}_{i.}$ are equal. The equality of the three means can be expressed in terms of two linearly independent differences being zero:

$$H_0: \overline{\mu}_{1.} = \overline{\mu}_{2.} = \overline{\mu}_{3.}$$

or

$$\begin{aligned} H_0: &\quad (\mu_{11} + \mu_{12}) - (\mu_{21} + \mu_{22}) = 0 \text{ and} \\ &\quad (\mu_{11} + \mu_{12}) + (\mu_{21} + \mu_{22}) - 2(\mu_{31} + \mu_{32}) = 0, \\ \boldsymbol{K}_2' &= \begin{bmatrix} 1 & 1 & -1 & -1 & 0 & 0 \\ 1 & 1 & 1 & 1 & -2 & -2 \end{bmatrix}, \\ r(\boldsymbol{K}_2) &= 2, \text{ and} \\ Q_2 &= 909.3 \text{ with 2 degrees of freedom.} \end{aligned} \tag{9.51}$$

3. *Sum of squares for "lines":* The hypothesis of no "line" effects is equivalent to the hypothesis that the two marginal means for "lines" $\overline{\mu}_{.j}$ are equal or that the difference is zero:

$$H_0: \quad \overline{\mu}_{.1} - \overline{\mu}_{.2}$$

or

$$\begin{aligned} H_0: &\quad \mu_{11} + \mu_{21} + \mu_{31} - \mu_{12} - \mu_{22} - \mu_{32} = 0, \\ \boldsymbol{K}_3' &= (1 \quad -1 \quad 1 \quad -1 \quad 1 \quad -1), \\ r(\boldsymbol{K}_3) &= 1, \text{ and} \\ Q_3 &= 2,496.15 \text{ with 1 degree of freedom.} \end{aligned} \tag{9.52}$$

4. *Sum of squares for "dates by lines":* The null hypothesis of no interaction effects between "dates" and "lines" is equivalent to the hypothesis that the difference between lines is the same for all dates, or that the differences among dates are the same for all lines. The former is easier to visualize because there are only two lines and one difference

TABLE 9.6. *Factorial analysis of variance of ascorbic acid content of cabbage.*

Source	*d.f.*	*Sum of Squares*	*Mean Square*
Total_{uncorr}	60	207,533.0	
Model	6	205,041.9	
C.F.	1	201,492.1	
Dates	2	909.3	454.7
Lines	1	2,496.2	2,496.2
Dates × Lines	2	144.3	72.2
Residual	54	2,491.1	46.1

between lines for each date. There are three such differences which, again, require two linearly independent statements:

$$H_0: \qquad \mu_{11} - \mu_{12} = \mu_{21} - \mu_{22} = \mu_{31} - \mu_{32}$$

or

$$\begin{aligned} H_0: \quad & (\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22}) = 0 \text{ and} \\ & (\mu_{11} - \mu_{12}) + (\mu_{21} - \mu_{22}) - 2(\mu_{31} - \mu_{32}) = 0, \\ \boldsymbol{K}_4' &= \begin{bmatrix} 1 & -1 & -1 & 1 & 0 & 0 \\ 1 & -1 & 1 & -1 & -2 & 2 \end{bmatrix}, \\ r(\boldsymbol{K}_4) &= 2, \text{ and} \\ Q_4 &= 144.25 \text{ with 2 degrees of freedom.} \end{aligned} \tag{9.53}$$

The $\boldsymbol{K}'$ matrix appropriate for the hypothesis of no interaction is the more difficult matrix to define. The statements were generated using the fact that interaction measures the failure of the simple effects to be consistent over all levels of the other factor. It should be observed, however, that $\boldsymbol{K}_4'$ is easily generated as the elementwise product of each row vector in $\boldsymbol{K}_2'$ with the row vector in $\boldsymbol{K}_3'$. Interaction contrasts can always be generated in this manner.

This analysis of variance is summarized in Table 9.6. The results are identical to those from the conventional analysis of variance for a two-factor factorial in a completely random experimental design. The residual mean square serves as the denominator for F-tests of the treatment effects (if treatment effects are fixed effects). There are significant differences among the planting dates and between the two genetic lines for ascorbic acid content. The interaction between dates and lines is not significant, indicating that the difference between the lines is reasonably constant over all planting dates.

9.8.2 Test of Homogeneity of Regression Coefficients

The analysis of covariance assumes that all treatments have the same relationship between the dependent variable and the covariate. In preparation for the covariance analysis of the cabbage data (Section 9.8.3), this section gives the test of homogeneity of the regression coefficients.

The full model for the test of homogeneity allows each treatment group to have its own regression coefficient relating ascorbic acid content to head size. The means model used in the analysis of variance (equation 9.49 is expanded to give

$$Y_{ijk} = \mu_{ij} + \beta_{ij}(X_{ijk} - \overline{X}_{...}) + \epsilon_{ijk}, \tag{9.54}$$

where the ij subscripts on β allow for a different regression coefficient for each of the six treatment groups. There are now 12 parameters and $\boldsymbol{X}^*$ must be of order (60×12). Each of the additional six columns in $\boldsymbol{X}^*$ consists of the covariate values for one of the treatment groups. The elements in the column for the ijth group take the values $(X_{ijk} - \overline{X}_{...})$ if the observation is from that group and zero otherwise. These six columns can be generated by elementwise multiplication of the dummy variable for each treatment by the original vector of $(X_{ijk} - \overline{X}_{...})$. The $\boldsymbol{X}^*$ matrix has the form

$$\boldsymbol{X}^* = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{x}_{11} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{x}_{12} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{x}_{21} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{x}_{22} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{x}_{31} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{x}_{32} \end{bmatrix},$$

where each *symbol* in $\boldsymbol{X}^*$ is a column vector of order 10×1; $\boldsymbol{x}_{ij}$ is the 10×1 column vector of the deviations of head weight from the overall mean head weight for the ijth treatment group. The least squares analysis using this model gives $\text{SS}(\text{Res}_{\text{full}}) = 1847.2$ with $60 - 12 = 48$ degrees of freedom.

The reduced model for the null hypothesis of homogeneity of regression coefficients, $H_0 : \beta_{ij} = \beta$ for all ij combinations, is

$$Y_{ijk} = \mu_{ij} + \beta(X_{ijk} - \overline{X}_{...}) + \epsilon_{ijk}. \tag{9.55}$$

There are seven parameters in this reduced model—the six μ_{ij} plus the common β. (This is the covariance model that is used in the next section.) The least squares analysis of this reduced model gives $\text{SS}(\text{Res}_{\text{reduced}}) = 1,975.1$ with 53 degrees of freedom.

The difference in residual sums of squares for the full and reduced models is:

$$\begin{aligned} Q &= \text{SS}(\text{Res}_{\text{reduced}}) - \text{SS}(\text{Res}_{\text{full}}) \\ &= 1,975.1 - 1,847.2 = 127.9 \end{aligned}$$

with $53 - 48 = 5$ degrees of freedom. This is the appropriate numerator sum of squares for the F-test of the null hypothesis. The appropriate denominator for the F-test is the residual mean square from the full model,

$$s^2 = \frac{1,847.24}{48} = 38.48.$$

Thus,

$$F = \frac{127.9/5}{38.48} = .66$$

which is nonsignificant. A common regression coefficient for all treatments is sufficient for describing the relationship between ascorbic acid content and the head weight of cabbage in these data.

If the regression coefficients are heterogeneous, the covariance analysis for whatever purpose must be used with caution. The meaning of "adjusted treatment means" is not clear when the responses to the covariate differ. The choice of the common level of the covariate to which adjustment is made becomes critical. The treatment differences and even the ranking of the treatments can depend on this choice.

9.8.3 Analysis of Covariance

The analysis of covariance is used on the ascorbic acid content of cabbage as an aid in interpreting the treatment effects. The differences among adjusted treatment means are *not* to be interpreted as treatment effects. The changes in the sums of squares and treatment means as they are adjusted provide insight into the degree of relationship between the treatment effects on the two response variables, ascorbic acid content and head weight.

The model for the analysis of covariance, using the means parameterization and a common regression of ascorbic acid on head size for all groups, was given as the reduced model in the test of homogeneity, equation 9.55. The least squares analysis of this model gives the analysis of covariance. The $\boldsymbol{X}^*$ matrix from the analysis of variance is augmented with the column of observations on the covariate, expressed as deviations from the mean of the covariate. The vector of parameters is expanded to include β, the regression coefficient for the covariate.

Least squares analysis for this model gives SS(Model) = 205,557.9 with 7 degrees of freedom and SS(Residual) = 1,975.1 with 53 degrees of freedom. The decrease in the residual sum of squares from the analysis of variance model to the covariance model is due to the linear regression on the covariate. This difference in SS(Res) for the two models is the partial sum of squares for β, $R(\beta|\boldsymbol{\mu}') = 2,491.1 - 1,975.1 = 516.0$ with 1 degree of freedom, and is the appropriate numerator sum of squares for the F-test of the null hypothesis $H_0 : \beta = 0$. The denominator is the residual mean square from the covariance model, $s^2 = 1,975.1/53 = 37.3$.

TABLE 9.7. *Partial sums of squares for the analysis of covariance of ascorbic acid content for the cabbage data. The covariate is head weight.*

Source	*d.f.*	*Sum of Squares*	*Mean Square*
Total_{uncorr}	60	207,533.0	
Model	7	205,557.9	
C.F.	1	201,492.1	
Dates	2	239.8	119.9
Lines	1	1,237.3	1,237.3
Dates × Lines	2	30.7	15.4
Covariate	1	516.0	516.0
Residual	53	1,975.1	37.3

The F-test of $H_0 : \beta = 0$ is

$$F = \frac{516.0}{37.3} = 13.8$$

with 1 and 53 degrees of freedom, which is significant beyond $\alpha = .001$. This confirms that there is a significant correlation between the variation in ascorbic acid content and head size after both have been adjusted for other effects in the model. This can be interpreted as a test of the hypothesis that the correlation between the random plot-to-plot errors of the two traits is zero.

General linear hypotheses are used to compute the partial sum of squares attributable to each of the original class variables. These sums of squares will differ from the analysis of variance sums of squares because they will now be adjusted for the covariate. The $\boldsymbol{K}'$ matrices defined in the analysis of variance, equations 9.50 through 9.53, need to be augmented on the right with a column of zeros as coefficients for β so that $\boldsymbol{K}'$ and $\widehat{\boldsymbol{\beta}}^*$ conform for multiplication. These sums of squares are no longer additive partitions of the model sum of squares because the adjustment for the covariate has destroyed the orthogonality. An additional $\boldsymbol{K}'$ could be defined for the hypothesis that $\beta = 0$, but the appropriate F-test based on the difference in residual sums of squares has already been performed in the previous paragraph. The analysis of variance summary for the covariance model is given in Table 9.7.

A comparison of Tables 9.6 and 9.7 shows major decreases in the sums of squares for "dates" and "lines" after adjustment for differences in head weight. The test for "date by line" effects is nonsignificant both before and after adjustment. The sum of squares for "dates" was reduced from a highly significant 909 to a just-significant 240 ($\alpha = .05$). The sum of squares for "lines" was reduced by half but is still highly significant. These results suggest that a significant part of the variation in ascorbic acid content among dates of planting and between lines is associated with variation in

TABLE 9.8. *Adjustment of treatment means for ascorbic acid content in cabbage for differences in the covariable head weight.*

Group	*Mean Head Weight*	*Mean Ascorbic Acid (Unadjusted)*	*Adjustment* $-\widehat{\beta}(\overline{X}_{ij.} - \overline{X}_{...})$	*Mean Ascorbic Acid (Adjusted)*
11	3.18	50.3	2.64	52.94 (2.06)[a]
12	2.26	62.5	−1.50	61.00 (1.97)
21	2.80	49.4	.93	50.33 (1.95)
22	3.11	58.9	2.33	61.23 (2.03)
31	2.74	54.8	.66	55.46 (1.94)
32	1.47	71.8	−5.06	66.74 (2.36)
Mean	2.593	57.95	.00	57.95

[a]Standard errors of adjusted treatment means are shown in parentheses. The standard error on each unadjusted treatment mean is 2.15.

head size. However, not all of the variation in ascorbic acid content can be explained by variation in head size.

The estimate of the parameters is:

$$\widehat{\boldsymbol{\beta}}^{*\prime} = (\,52.94 \quad 61.00 \quad 50.33 \quad 61.23 \quad 55.46 \quad 66.74 \quad -4.503\,).$$

The $\widehat{\mu}_{ij}$ from the means reparameterization are estimates of the treatment means for ascorbic acid content, which are now adjusted for differences in head weight. (The estimate of the parameters contains the adjusted treatment means only because the means reparameterization was used and the covariate was centered. Otherwise, linear functions of the parameter estimates would have to be used to compute the adjusted means.) The estimate of the regression coefficient for the covariate is $\widehat{\beta} = -4.50265$. Each increase of 1 unit in head weight is associated with a *decrease* in ascorbic acid content of 4.5 units on the average.

The adjustments to mean ascorbic acid content for differences in mean head weight are shown in Table 9.8. The biggest adjustment is for the third planting date for line 2, which had a very small head weight and high ascorbic acid content. Adjustment for head size reduced the average difference in ascorbic acid content between the two lines from about 12 units to 10 units. The first two planting dates differ very little for either line, but the third planting date gives appreciably higher ascorbic acid content even after adjustment for smaller head size on that planting date.

The analysis shows that there is considerable genetic and environmental correlation between ascorbic acid content and head size in cabbage. Some of the higher ascorbic acid content in line 2 on the third planting date may be attributable to the smaller head size produced by that treatment

TABLE 9.9. *Average dry forage yields (lbs/A) from a study of sources and rates of phosphorus fertilization. The experimental design was a randomized complete block design with seven sources of phosphorus, each applied at two rates (lbs/A). The phosphorus content of the soil (ppm of P_2O_5) at the beginning of the study was recorded for use as a possible covariate. (Data are from the files of the late Dr. Gertrude M. Cox.)*

Treatment		*Block I*		*Block II*		*Block III*	
Source	*Rate*	*Phos.*	*Forage*	*Phos.*	*Forage*	*Phos.*	*Forage*
SUPER	40	32.0	2,475	43.2	3,400	51.2	3,436
SUPER	80	44.8	3,926	56.0	4,145	75.2	3,706
TSUPER	40	43.2	2,937	52.8	2,826	27.2	3,288
TSUPER	80	41.6	3,979	64.0	4,065	36.8	4,344
BSLAG	40	49.6	3,411	62.4	3,418	46.4	2,915
BSLAG	80	51.2	4,420	62.4	4,141	48.0	4,297
FROCK	40	48.0	3,122	75.2	3,372	22.4	1,576
FROCK	80	48.0	4,420	76.8	3,926	24.0	1,666
RROCK	40	54.4	2,334	60.8	2,530	49.6	1,275
RROCK	80	60.8	3,197	59.2	3,444	46.4	2,414
COLOID	40	72.0	3,045	59.2	2,206	19.2	540
COLOID	80	76.8	3,333	32.0	410	70.4	4,294
CAMETA	40	64.0	3,594	62.4	3,787	44.8	3,312
CAMETA	80	62.4	3,611	76.8	4,211	48.0	4,379

combination. This does not mean, however, that this adjusted mean is a better estimate of the ascorbic acid content of line 2 when planted late. The smaller head size may be an innate trait of line 2 when grown under the environmental conditions of the late planting. If so, the adjustment to a common head size underestimates the ascorbic acid content for line 2 grown under those conditions. ■

The next example illustrates the classical use of covariance to control experimental error.

Example 9.7

The data for the example are from a study to compare seven sources of phosphorus each applied at two rates (40 and 80 lbs/A). The experimental design is a randomized complete block experimental design with $b = 3$ blocks. The dependent variable is 3-year dry weight forage production (lbs/A). The covariate is soil phosphorus content (ppm P_2O_5) measured at the beginning of the study. The data are given in Table 9.9. (The data are from the files of the late Dr. Gertrude M. Cox.)

The linear model for a factorial set of treatments in a randomized complete block design is

$$Y_{ijk} = \mu + \rho_i + \gamma_j + \tau_k + (\gamma\tau)_{jk} + \epsilon_{ijk}, \tag{9.56}$$

where

$$\begin{aligned} \rho_i &= \text{effect of } i\text{th block } (i = 1, 2, 3) \\ \gamma_j &= \text{effect of } j\text{th source of phosphorus } (j = 1, \ldots, 7) \\ \tau_k &= \text{effect of } k\text{th rate of application } (k = 1, 2) \\ (\gamma\tau)_{jk} &= \text{interaction effect of } j\text{th source and } k\text{th rate.} \end{aligned}$$

The covariate is included in the model by adding the term $\beta(X_{ijk} - \overline{X}_{...})$ to equation 9.56. In this example, the covariate was measured before the treatments were applied to the experimental units, so there is no chance the covariate could have been affected by the treatments.

The analysis of variance model contains 27 parameters but the rank of $\boldsymbol{X}$ is $r(\boldsymbol{X}) = 17$; reparameterization would therefore require 10 constraints. Analysis of these data uses the generalized inverse approach, rather than reparameterization, to obtain the solution to the normal equations. PROC ANOVA and PROC GLM, the general linear models procedure, (SAS Institute Inc., 1989a, 1989b) are used for the analyses.

The analysis of variance is obtained from PROC ANOVA using the statements:

```
PROC ANOVA; CLASS BLOCK SOURCE RATE; MODEL
FORAGE = BLOCK SOURCE RATE SOURCE*RATE;
```

The CLASS statement identifies the variables that are to be regarded as class variables. Whenever a class variable is encountered in the MODEL statement, the program constructs a dummy variable for each level of the class variable. Thus, $\boldsymbol{X}$ will contain 3 dummy variables for BLOCK, 7 dummy variables for SOURCE, and 2 dummy variables for RATE. An interaction between two (or more) class variables in the MODEL statement instructs the program to construct a dummy variable for each unique joint level of the two factors; there will be 14 dummy variables for SOURCE*RATE.

The summary of the analysis of variance for the experiment is given in Table 9.10. There are significant differences among the sources of phosphorus ($\alpha = .05$) and highly significant differences between the rates of application ($\alpha = .01$). Block effects and source-by-rate interaction effects are not significant. The residual mean square is $s^2 = 735,933$ and the coefficient of variation is 26.7%.

The purpose of the covariance analysis is to use the information on soil phosphorus content to "standardize" the experimental results to a common level of soil phosphorus and, thereby, improve the precision of the comparisons. The analysis of covariance is obtained from PROC GLM (PROC

TABLE 9.10. *Analysis of variance of dry forage from the phosphorus fertilization data.*

Source	*d.f.*	*Sum of Squares*	*Mean Square*	*F*	*Prob > F*
Corrected total	41	41,719,241			
BLOCK	2	1,520,897	760,449	1.03	.3700
SOURCE	6	13,312,957	2,218,826	3.01	.0226
RATE	1	7,315,853	7,315,853	9.94	.0040
SOURCE∗RATE	6	435,267	72,544	.10	.9959
Error	26	19,134,266	735,933		

ANOVA cannot handle a continuous variable) by expanding the model statement to include the covariate PHOSDEV as follows.

```
MODEL FORAGE=BLOCK SOURCE RATE SOURCE*RATE
PHOSDEV/SOLUTION;
```

The variable PHOSDEV has been previously defined in the program as the centered covariate. The "/SOLUTION" portion of the statement requests PROC GLM to print a solution to the normal equations.

The analysis of covariance is summarized in Table 9.11. The lower two sections of Table 9.11 present the sequential sums of squares (TYPE I in SAS) and the partial sums of squares (TYPE III in SAS). Since the covariate was placed last in the model statement and the experimental design was balanced, the first four lines of the sequential sums of squares reproduce the analysis of variance sums of squares (Table 9.10).

The first question to ask of the analysis is whether the covariate has improved the precision of the comparisons. The residual mean square after adjustment for the covariate is $s^2 = 384,776$. This is a reduction of 48% from $s^2 = 735,933$ in the analysis of variance (Table 9.10). The coefficient of variation has been reduced from 26.7% to 19.3%. The reduction in the residual sum of squares is the partial sum of squares for the covariate and provides a test of the hypothesis $H_0 : \beta = 0$, where β is the regression coefficient on PHOSDEV. This test gives $F = 24.73$ with 1 and 25 degrees of freedom, which is significant beyond $\alpha = .0001$. [$\widehat{\beta} = 39.7801$ with $s(\widehat{\beta}) = 7.9996$]. The use of the covariate, initial soil phosphorus content, has greatly improved the precision of the experiment.

Adjustment of the treatment effects for differences in the covariate changed the treatment sums of squares (compare the sequential and partial sums of squares in Table 9.11) but did not change any of the conclusions from the F-tests of the treatment effects. Sources of phosphorus and rates of application remain significant, both beyond $\alpha = .01$, and the source-by-rate interaction remains nonsignificant. The absence of any interaction between sources and rates of fertilization means that differences in forage produc-

TABLE 9.11. *Covariance analysis for dry forage yield from a randomized complete block design with seven sources of phosphorus applied at two rates. The covariate is amount of soil phosphorus in the plot at the beginning of the three-year study.*

Source	*d.f.*	*Sum of Squares*	*Mean Square*	*F*	*Prob > F*
Model	16	32,099,838	2,006,240	5.21	.0001
Error	25	9,619,403	384,776		
Corrected Total	41	41,719,241			

Sequential Sums of Squares:

Source	*d.f.*	*SS*	*MS*	*F*	*Prob > F*
BLOCK	2	1,520,897	760,449	1.98	.1596
SOURCE	6	13,312,957	2,218,826	5.77	.0007
RATE	1	7,315,853	7,315,853	19.01	.0002
SOURCE*RATE	6	435,267	72,544	.19	.9773
PHOSDEV	1	9,514,863	9,514,863	24.73	.0001

Partial Sums of Squares:

Source	*d.f.*	*SS*	*MS*	*F*	*Prob > F*
BLOCK	2	1,173,100	586,550	1.52	.2373
SOURCE	6	15,417,193	2,569,532	6.68	.0003
RATE	1	3,623,100	3,623,100	9.42	.0051
SOURCE*RATE	6	999,892	166649	.43	.8497
PHOSDEV	1	9,514,863	9,514,863	24.73	.0001

tion among the 14 phosphorus fertilization treatments can be summarized in the marginal means for the two treatment factors, sources and rates. However, both sets of means need to be adjusted to remove biases due to differences in initial levels of soil phosphorus.

The adjusted SOURCE marginal means are obtained as

$$\overline{Y}_{\text{adj}_{.j.}} = \overline{Y}_{.j.} - \widehat{\beta}(\overline{X}_{.j.} - \overline{X}_{...}),$$

where $\overline{X}_{.j.}$ is the marginal mean for the covariate for those experimental plots receiving the jth source of phosphorus, and $\overline{X}_{...}$ is the overall mean for the covariate, $\widehat{\beta} = 39.7801$. This adjusts the SOURCE means to the common level of initial soil phosphorus $\overline{X}_{...} = 52.4$ ppm. Similarly, the adjusted RATE marginal means are obtained as

$$\overline{Y}_{\text{adj}_{..k}} = \overline{Y}_{..k} - \widehat{\beta}(\overline{X}_{..k} - \overline{X}_{...}).$$

The unadjusted marginal means and the steps in the adjustment to obtain the adjusted means are shown in Table 9.12. The standard errors of the adjusted treatment means are also shown. The standard errors on the unadjusted treatment means were $s(\overline{Y}_{.j.}) = 350.2$ and $s(\overline{Y}_{..k}) = 229.3$. The differences between standard errors for the unadjusted and adjusted means show a marked increase in precision from the use of the covariate.

PROC GLM computes the adjusted means as linear functions of the solution $\boldsymbol{\beta}^0$. The appropriate linear functions to be estimated for each mean are determined by the expectations of means in balanced data with the covariate set equal to $\overline{X}_{...}$. For example, the expectation of the marginal mean for the first source, BSLAG, is

$$\mathcal{E}(\overline{Y}_{.1.}) = \mu + \frac{\rho_1 + \rho_2 + \rho_3}{3} + \gamma_1 + \frac{\tau_1 + \tau_2}{2} + \frac{(\gamma\tau)_{11} + (\gamma\tau)_{12}}{2}.$$

The expectation contains, in addition to $\mu + \gamma_1$, the average of the block effects ρ_i, the average of the rate effects τ_k, and the average of the interaction effects in which source 1 is involved. The covariate is not involved in this expectation because adjusting to the mean level of the covariate is equivalent to adjusting to PHOSDEV = 0 when the centered covariate is used. This is the particular linear function of $\boldsymbol{\beta}$ that is to be estimated as the marginal FORAGE mean for SOURCE = BSLAG. The estimate is obtained by computing the *same* linear function of $\boldsymbol{\beta}^0$. The adjusted means are obtained from PROC GLM with the statement

LSMEANS SOURCE RATE/STDERR;

The "/**STDERR**" asks for the standard errors on the adjusted means to be printed.

Interpretations of the treatment effects are based on the adjusted treatment means. In this example, adjustment for differences in the covariate

TABLE 9.12. *Unadjusted and adjusted treatment means for "Source" and "Rate" of phosphorus fertilization. There was no "Rate by Source" interaction so that the experimental results are summarized in terms of the marginal means.*

Treatment	*Forage Mean (Unadj)*[a]	*Phosphorus Mean Deviation*[b]	*Covariance Adjustment*[c]	*Forage Mean (Adj.)*	*Std. Error*
SOURCE means:					
BSLAG	3,767.0	0.914	−36.4	3,730.6	253.3
CAMETA	3,815.7	7.314	−291.0	3,524.7	259.9
COLOID	2,304.7	2.514	−100.1	2,204.6	254.0
FROCK	3,013.7	−3.352	133.3	3,147.0	254.7
RROCK	2,532.3	2.781	−110.6	2,421.7	254.2
SUPER	3,514.7	−2.019	80.3	3,595.0	253.8
TSUPER	3,573.2	−8.152	324.3	3,897.5	261.5
RATE means:					
40	2,800.0	−2.895	115.2	2,915.1	137.3
80	3,634.7	2.895	−115.2	3,519.5	137.3

[a] The standard errors for the unadjusted treatment means are $s(\overline{Y}_{.j.}) = 350.2$ for the SOURCE means and $s(\overline{Y}_{..k}) = 229.3$ for the RATE means.

[b] "Phosphorus mean deviation" is $(\overline{X}_{.j.} - \overline{X}_{...})$ for SOURCE means and $(\overline{X}_{..k} - \overline{X}...)$ for RATE means.

[c] "Covariance adjustment" is $-\widehat{\beta}$(Phosphorus mean deviation) where $\widehat{\beta} = 39.7801$.

changed the ranking of the four best sources of phosphorus, which did not differ significantly, and decreased the difference between the two rates of application. The adjusted means suggest an average rate of change in forage of 15lbs/A for each lb/A of phosphorus compared to 21lbs/A suggested by the unadjusted means. ■

9.9 Exercises

9.1. Use matrix multiplication to verify that the linear model in equation 9.5, where $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are as defined in equation 9.4, generates the combinations of effects shown in equation 9.2.

9.2. Determine the number of rows and columns in $\boldsymbol{X}$ before reparameterization for one-way structured data with t groups (or treatments) and n observations in each group. How does the order of $\boldsymbol{X}$ change if there are n_i observations in each group?

9.3 Suppose you have one-way structured data with $t = 3$ groups. Define the linear model such that μ is the mean of the first group and the second and third groups are measured as deviations from the first. Is $\boldsymbol{X}$ for this model of full rank? Does this form of the model relate to any of the three reparameterizations?

9.4. The accompanying table gives survival data for tropical corn borer under field conditions in Thailand (1974). Researchers inoculated 30 experimental plots with egg masses of the corn borer on the same date by placing egg masses on each corn plant in the plot. After each of 3, 6, 9, 12, and 21 days, the plants in 6 random plots were dissected and the surviving larvae were counted. This gives a completely random experimental design with the treatments being "days after inoculation." (Data are used with permission of Dr. L. A. Nelson, North Carolina State University.)

Days After Inoculation	*Numbers of Larvae Surviving in 6 Plots*					
3	17	22	26	20	11	14
6	37	26	24	11	11	16
9	8	5	12	3	5	4
12	14	8	4	6	3	3
21	10	13	5	7	3	4

(a) Do the classical analysis of variance by hand for the completely random design. Include in your analysis a partitioning of the

sum of squares for treatments to show the linear regression on "number of days" and deviations from linearity.

(b) Regard "days after inoculation" as a class variable. Define $\boldsymbol{Y}$, $\boldsymbol{X}$, and $\boldsymbol{\beta}$ so that the model for the completely random design $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$ can be represented in matrix form. Show enough of each matrix to make evident the order in which the observations are listed. Identify the singularity that makes $\boldsymbol{X}$ not of full rank.

(c) Show the form of $\boldsymbol{X}$ and $\boldsymbol{\beta}$ for each of the three reparameterizations—the means model, the $\sum \tau_i = 0$ constraint, and the $\tau_5 = 0$ constraint.

(d) Choose one of the reparameterizations to compute $R(\boldsymbol{\tau}'|\mu)$ and SS(Res). Summarize the results in an analysis of variance table and compare with the analysis of variance obtained under (a).

(e) Use SAS PROC GLM, or a similar program for the analysis of less than full-rank models, to compute the analysis of variance. Ask for the solution to the normal equations so that "estimates" of $\boldsymbol{\beta}$ are obtained. Compare these sums of squares and estimates of $\boldsymbol{\beta}$ with the results from your reparameterization in Part (d). Show that the unbiased estimates of $\mu + \tau_1$ and $\tau_1 - \tau_2$ are the same from both analyses.

(f) Now regard X as a quantitative variable and redefine $\boldsymbol{X}$ and $\boldsymbol{\beta}$ so that $\boldsymbol{Y} = \boldsymbol{X\beta} + \epsilon$ expresses Y as a linear function of "number of days." Compute SS(Regr) and compare the result with that under Part (a). Test the null hypothesis that the linear regression coefficient is zero. Test the null hypothesis that the linear function adequately represents the relationship.

(g) Do you believe the assumptions for least squares are valid in this example? Justify.

9.5. Use $\boldsymbol{X}$ and $\boldsymbol{\beta}$ as defined for the completely random design, equation 9.4. Define $\boldsymbol{K}'$ for the null hypothesis $H_0 : \tau_1 = \tau_2$. Define $\boldsymbol{K}'$ for the null hypothesis $H_0 : \tau_3 = \tau_4$. Define $\boldsymbol{K}'$ for the *composite* null hypothesis $H_0 : \tau_1 = \tau_2$ and $\tau_3 = \tau_4$ and $\tau_1 + \tau_2 = \tau_3 + \tau_4$. Is each of these hypotheses testable? How does the sum of squares generated by the composite hypothesis relate to the analysis of variance?

9.6. Show that the means model reparameterization for the completely random design is equivalent to imposing the constraint that $\mu = 0$.

9.7. Express the columns of $\boldsymbol{X}$ in equation 9.4 as linear combinations of columns of $\boldsymbol{X}^*$ in equation 9.15. Also, express the columns of $\boldsymbol{X}^*$ in equation 9.15 as linear combinations of columns of $\boldsymbol{X}$. Thus, the space spanned by columns of $\boldsymbol{X}$ is the same as that spanned by columns of $\boldsymbol{X}^*$.

9.8. Use the means model reparameterization on a randomized complete block design with $b = 2$ and $t = 4$. As discussed in the text, this reparameterization leaves zero degrees of freedom for the estimate of error. However, experimental error can be estimated as the block-by-treatment interaction sum of squares. Define $\boldsymbol{K}'$ for the means reparameterization so that the sum of squares obtained from Q is the error sum of squares.

9.9. Show $\boldsymbol{X}^*$ and $\boldsymbol{\beta}^*$ for the model for the randomized complete block design (equation 9.21) with $b = 2$ and $t = 4$ using the constraint $\gamma_2 = 0$ and $\mu_j = \mu + \tau_j$. Determine the expectation of $\widehat{\boldsymbol{\beta}}^*$ in terms of the original parameters.

9.10. Use matrix multiplication of $\boldsymbol{X}$ and $\boldsymbol{\beta}$ in equation 9.22 to verify that the linear model in equation 9.21 is obtained.

9.11. Determine the general result for the number of columns in $\boldsymbol{X}$ for two-way classified data when there are b levels of one factor and t levels of the other factor if the model does not contain interaction effects. How many additional columns are needed if the model does contain interaction effects?

9.12. A randomized complete block experimental design was used to determine the joint effects of temperature and concentration of herbicide on absorption of 2 herbicides on a commercial charcoal material. There were 2 blocks and a total of 20 treatment combinations—2 temperatures by 5 concentrations by 2 herbicides. (The data are used with permission of Dr. J. B. Weber, North Carolina State University.)

	Temp.		*Concentration* $\times 10^5$				
Block	$^\circ C$	*Herb.*	20	40	60	80	100
1	10	*A*	.280	.380	.444	.480	.510
		B	.353	.485	.530	.564	.620
	55	*A*	.266	.332	.400	.436	.450
		B	.352	.474	.556	.590	.625
2	10	*A*	.278	.392	.440	.470	.500
		B	.360	.484	.530	.566	.611
	55	*A*	.258	.334	.390	.436	.446
		B	.358	.490	.560	.570	.600

The usual linear model for a randomized complete block experiment, $Y_{ij} = \mu + \gamma_i + \tau_j + \epsilon_{ij}$, where γ_i is the effect of the ith block and τ_j is the effect of the jth treatment, can be expanded to include the main and interaction effects of the three factors:

$$\begin{aligned} Y_{ijkl} &= \mu + \gamma_i + T_j + H_k + C_l + (TH)_{jk} + (TC)_{jl} \\ &\quad + (HC)_{kl} + (THC)_{jkl} + \epsilon_{ijkl}, \end{aligned}$$

where T_j, H_k, and C_l refer to the effects of temperature, herbicide, and concentration, respectively. The combinations of letters refer to the corresponding interaction effects.

(a) Show the form of $\boldsymbol{X}$ and $\boldsymbol{\beta}$ for the usual RCB model, the model containing γ_i and τ_j. Assume the data in $\boldsymbol{Y}$ are listed in the order that would be obtained if successive rows of data in the table were appended into one vector. What is the order of $\boldsymbol{X}$ and how many singularities does it have? Use $\gamma_2 = 0$ and $\tau_{20} = 0$ to reparameterize the model and compute the sums of squares for blocks and treatments.

(b) Define $\boldsymbol{K}'$ for the singular model in Part (a) for the composite null hypothesis that there is no temperature effect at any of the combinations of herbicide and concentration. (*Note*: τ_1 is the effect for the treatment having temperature 10°, herbicide A, and concentration 20×10^{-5}. τ_{11} is the effect for the similar treatment except with 55° temperature. The null hypothesis states that these two effects must be equal, or their difference must be zero, and similarly for all other combinations of herbicide and concentration.) How many degrees of freedom does this sum of squares have? Relate these degrees of freedom to degrees of freedom in the conventional factorial analysis of variance. Define $\boldsymbol{K}'$ for the null hypothesis that the *average* effect of temperature is zero. How many degrees of freedom does this sum of squares have and how does it relate to the analysis of variance?

(c) Show the form of $\boldsymbol{X}$ and $\boldsymbol{\beta}$ if the factorial model with *only* the main effects T_j, H_k, and C_l is used. How many singularities does this $\boldsymbol{X}$ matrix contain? Show the form of $\boldsymbol{X}^*$ if the "sum" constraints are used. Use this reparameterized form to compute the sums of squares due to temperature, due to herbicides, and due to concentration.

(d) Demonstrate how $\boldsymbol{X}$ in Part (c) is augmented to include the $(TH)_{jk}$ effects. How many columns are added to $\boldsymbol{X}$? How many additional singularities does this introduce? How many columns would be added to $\boldsymbol{X}$ to accommodate the $(TC)_{jl}$ effects? The $(HC)_{kl}$ effects? The $(THC)_{jkl}$ effects? How many singularities does each introduce?

(e) Use PROC ANOVA in SAS, or a similar computer package, to compute the full factorial analysis of variance. Regard blocks, temperature, herbicide, and concentration as class variables.

9.13. The effect of supplemental ascorbate, vitamin C, on survival time of terminal cancer patients was studied. [Data are from Cameron and Pauling (1978) as reported in Andrews and Herzberg (1985).] The

Effect of supplemental ascorbate on survival time of cancer patients.

Stomach Cancer			*Bronchus Cancer*			*Colon Cancer*		
Age	*Days*	*Cont.*	*Age*	*Days*	*Cont.*	*Age*	*Days*	*Cont.*
Females:			**F**emales:			**F**emales:		
61	124	38	48	87	13	76	135	18
62	19	36	64	115	49	58	50	30
66	45	12	**M**ales:			70	155	57
69	876	19	74	74	33	68	534	16
59	359	55	74	423	18	74	126	21
Males:			66	16	20	76	365	42
69	12	18	52	450	58	56	911	40
63	257	64	70	50	38	74	366	28
79	23	20	77	50	24	60	99	28
76	128	13	71	113	18	**M**ales:		
54	46	51	70	857	18	49	189	65
62	90	10	39	38	34	69	1,267	17
46	123	52	70	156	20	50	502	25
57	310	28	70	27	27	66	90	17
			55	218	32	65	743	14
			74	138	27	58	156	31
			69	39	39	77	20	33
			73	231	65	38	274	80

survival time (Days) of each treated patient was compared to the mean survival time of a control group (Cont.) of 10 similar patients. Age of patient was also recorded. For this exercise, the results are used from three cancer types—stomach, bronchus, and colon. There were 13, 17, and 17 patients in the three groups, respectively. For this question use the logarithm of the ratio of days survival of the treated patient to the mean days survival of his or her control group as the dependent variable.

(a) Use the means model reparameterization to compute the analysis of variance for ln(survival ratio). Determine $\boldsymbol{X}^{*\prime}\boldsymbol{X}^{*}$, $\boldsymbol{X}^{*\prime}\boldsymbol{Y}$, $\widehat{\boldsymbol{\beta}}^{*}$, SS(Model), SS(Res), and s^2. What is the least squares estimate of the mean ln(survival ratio) for each cancer group and what is the standard error of each mean? Two different kinds of hypotheses are of interest: does the treatment increase survival time; that is, is ln(survival ratio) significantly greater than zero for each type cancer; and are there significant differences among the cancer types in the effect of the treatment? Use a t-test to

test the null hypothesis that the true mean ln(survival ratio) for each group is zero. Use an F-test to test the significance of differences among cancer types.

(b) The ages of the patients in the study varied from 38 to 79; the mean age was 64.3191 years. Augment the $\boldsymbol{X}^*$ matrix in Part (a) with the vector of centered ages. Compute the residual sum of squares and the estimate of σ^2 for this model. Compute the standard error of each estimated regression coefficient. Use a t-test to test the null hypothesis that the partial regression coefficient for the regression of ln(survival ratio) on age is zero. Use the difference in residual sums of squares between this model and the previous model to test the same null hypothesis. How are these two tests related? What is your conclusion about the importance of adjusting for age differences?

(c) Since the means model was used in Part (b) and ages were expressed as deviations from the mean age, the first three regression coefficients in $\widehat{\boldsymbol{\beta}}$ are the estimates of the cancer group means *adjusted* to the mean age of 64.3191. Construct $\boldsymbol{K}'$ for the hypothesis that the true means, adjusted for age differences, of the stomach and bronchus cancer groups, the first and second groups, are the same as for colon cancer, the third group. Complete the test and state your conclusion.

(d) Describe how $\boldsymbol{X}_c^*$ would be defined to adjust all observations to age 60 for all patients. Show the form of $\boldsymbol{T}$ for averaging the adjusted observations to obtain the adjusted group means. The adjusted group means are obtained as $\boldsymbol{T}'\boldsymbol{X}_c^*\widehat{\boldsymbol{\beta}}^*$, equation 9.44. Compute $\boldsymbol{T}'\boldsymbol{X}_c^*$ and $s^2(\overline{\boldsymbol{Y}}_{adj})$ for this example.

(e) Even though the average regression on age did not appear important, it was decided that each cancer group should be allowed to have its own regression on age to verify that age was not important in any of the three groups. Illustrate how $\boldsymbol{X}^*$ would be expanded to accomodate this model and complete the test of the null hypothesis that the regressions on age are the same for all three cancer groups. State your conclusion.

9.14. The means reparameterization was used on the cabbage data example (Example 9.5) in the text. Define $\boldsymbol{\beta}^*$ and $\boldsymbol{X}^*$ for this model (equation 9.48 using the reparameterization constraints $\gamma_3 = \tau_2 = (\gamma\tau)_{31} = (\gamma\tau)_{32} = (\gamma\tau)_{12} = (\gamma\tau)_{22} = 0$. Define $\boldsymbol{K}'$ for the reparameterized model so as to obtain the sum of squares for "dates-by-lines" interaction.

9.15. Equation 9.55 defines the reduced model for $H_0 : \beta_{ij} = \beta$ for all ij. Define the reduced model for the test of homogeneity of regressions

within lines:

$$H_0 : \beta_{11} = \beta_{21} = \beta_{31} \text{ and } \beta_{12} = \beta_{22} = \beta_{32}.$$

Find SS(Res) for this reduced model and complete the test of homogeneity.

9.16. The means model was used in the cabbage data example (equation 9.49) and $\boldsymbol{K}'$ was defined to partition the sums of squares. Develop a reduced model that reflects $H_0 : \overline{\mu}_{1.} = \overline{\mu}_{2.} = \overline{\mu}_{3.}$ Use the full and reduced models to obtain the sum of squares for this hypothesis and verify that this is equivalent to that using $\boldsymbol{K}'$ (equation 9.51) in the text.

9.17. The covariance analysis of the phosphorus study in Section 9.8.3 assumed a common regression of forage yield on soil phosphorus. Use a general linear analysis program (such as PROC GLM in SAS) to test the homogeneity of regressions over the 14 treatment groups.

9.18. The Linthurst data used in Chapters 5 and 7 came from nine sites classified according to location (LOC) and type of vegetation ($TYPE$). (The data are given in Table 5.1.) Do the analysis of variance on *BIOMASS* partitioning the sum of squares into that due to LOC, $TYPE$, and LOC-by-$TYPE$ interaction. The regression models in Chapter 7 indicated that pH and Na were important variables in accounting for the variation in *BIOMASS*. Add these two variables to your analysis of variance model as covariates (center each) and compute the analysis of covariance. Obtain the adjusted LOC, $TYPE$, and LOC-by-$TYPE$ treatment means. Interpret the results of the covariance analysis. For what purpose is the analysis of covariance being used in this case?

9.19. Consider the analysis of covariance model given by

$$Y_{ij} = \mu + \tau_i + \beta(X_{ij} - \overline{X}_{..}) + \epsilon_{ij};\ i = 1, \ldots, a;\ j = 1, \ldots, r,$$

where ϵ_{ij}s are independent normal random variables with mean zero and variance σ^2.

(a) Show that all of the following models are reparameterizations of the prededing model.

(i) $Y_{ij} = \mu_i + \beta(X_{ij} - \overline{X}_{..}) + \epsilon_{ij}$.

(ii) $Y_{ij} = \mu_i^* + \beta(X_{ij} - \overline{X}_{i.}) + \epsilon_{ij}$.

(iii) $Y_{ij} = \mu^* + \tau_i^* + \beta X_{ij} + \epsilon_{ij}$.

Interpret the parameters μ_i and μ_i^*.

(b) Use the reparameterization (ii) in (a) to derive

$$\text{SSE}_{\text{full}} = \sum_{i=1}^{a}\sum_{j=1}^{r}(Y_{ij} - \overline{Y}_{i.})^2 - \widehat{\beta}^2 \sum_{i=1}^{a}\sum_{j=1}^{r}(X_{ij} - \overline{X}_{i.})^2$$

and

$$R(\beta|\mu_1^*, \ldots, \mu_a^*) = \widehat{\beta}^2 \sum_{i=1}^{a}\sum_{j=1}^{r}(X_{ij} - \overline{X}_{i.})^2,$$

where

$$\widehat{\beta} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{r}(X_{ij} - \overline{X}_{i.})Y_{ij}}{\sum_{i=1}^{a}\sum_{j=1}^{r}(X_{ij} - \overline{X}_{i.})^2}.$$

(c) To test the hypothesis that there is "no treatment effect," consider the reduced model

$$Y_{ij} = \overline{\mu} + \beta(X_{ij} - \overline{X}_{..}) + \epsilon_{ij}.$$

Show that

$$\text{SSE(Reduced)} = \sum_{i=1}^{a}\sum_{j=1}^{r}(Y_{ij} - \overline{Y}_{..})^2 - \widetilde{\beta}^2 \sum_{i=1}^{a}\sum_{j=1}^{r}(X_{ij} - \overline{X}_{..})^2,$$

where

$$\widetilde{\beta} = \sum_{i=1}^{a}\sum_{j=1}^{r}(X_{ij} - \overline{X}_{..})Y_{ij} \Bigg/ \sum_{i=1}^{a}\sum_{j=1}^{r}(X_{ij} - \overline{X}_{..})^2 \ .$$

[Note that we can now obtain $R(\tau|\beta)$ as $\text{SS}(\text{Res}_{\text{reduced}}) - \text{SS}(\text{Res}_{\text{full}})$.]

10

PROBLEM AREAS IN LEAST SQUARES

All discussions to this point have assumed that the least squares assumptions of normality, common variance, and independence are valid, and that the data are correct and representative of the intended populations.

In reality, the least squares assumptions hold only approximately and one can expect the data to contain either errors or observations that are somewhat unusual compared to the rest of the data. This chapter presents a synopsis of the problem areas that commonly arise in least squares analysis.

The least squares regression method discussed in the previous chapters was based on the assumptions that the errors are additive (to the fixed-effects part of the model) and are normally distributed independent random variables with common variance σ^2. Least squares estimation based on these assumptions is referred to as **ordinary least squares**. When the assumptions of independence and common variance hold, least squares estimators have the desirable property of being the best (minimum variance) among all possible linear unbiased estimators. When the normality assumption is satisfied, the least squares estimators are also maximum likelihood estimators.

Three of the major problem areas in least squares analysis relate to failures of the basic assumptions: normality, common variance, and indepen-

dence of the errors. Other problem areas are overly influential data points, outliers, inadequate specification of the functional form of the model, near-linear dependencies among the independent variables (collinearity), and independent variables being subject to error. This chapter is a synopsis of these problem areas with brief discussions on how they might be detected, their impact on least squares, and what might be done to remedy or at least reduce the problem. Subsequent chapters discuss in greater detail techniques for detecting the problems, transformations of variables as a means of alleviating some of the problems, and analysis of the correlational structure of the data to understand the nature of the collinearity problem. This process of checking the validity of the assumptions, the behavior of the data, and the adequacy of the model is an important step in every regression analysis. It should not, however, be regarded as a substitute for a proper validation of the regression equation against an independent set of data.

The emphasis here is on making the user aware of problem areas in the data or the model and insofar as possible removing the problems. An alternative to least squares regression when the assumptions are not satisfied is **robust regression**. Robust regression refers to a general class of statistical procedures designed to reduce the sensitivity of the estimates to failures in the assumptions of the parametric model. For example, the least squares approach is known to be sensitive to gross errors, or outliers, in the data because the solution minimizes the *squared* deviations. A robust regression procedure would reduce the impact of such errors by reducing the weight given to large residuals. This can be done by minimizing the sum of absolute residuals, for example, rather than the sum of squared residuals. In the general sense, procedures for detecting outliers and influential observations can be considered part of robust regression. Except for this connection, robust regression is not discussed in this text. The reader is referred to Huber (1981) and Hampel, Ronchetti, Rousseeuw, and Stahel(1986) for discussions on robust statistics.

10.1 Nonnormality

Importance of Normality

The assumption that the residuals ϵ are normally distributed is not necessary for estimation of the regression parameters and partitioning of the total variation. Normality is needed only for tests of significance and construction of confidence interval estimates of the parameters. The t-test, F-test, and chi-square test require the underlying random variables to be normally distributed. Likewise, the conventional confidence interval estimates depend on the normal distribution, either directly or through Student's t-distribution.

"Nonnormal" Data

Experience has shown that normality is a reasonable assumption in many cases. However, in some situations it is not appropriate to assume normality. Count data will frequently behave more like Poisson-distributed random variables. The proportion of subjects that show a response to the agent in toxicity studies is a binomially distributed random variable if the responses are independent. Time to failure in reliability studies and time to death in toxicity studies will tend to have asymmetric distributions and, hence, not be normally distributed.

The impact of nonnormality on least squares depends on the degree of departure from normality and the specific application. Nonnormality does not affect the estimation of the parameters; the least squares estimates are still the best linear unbiased estimates if the other assumptions are met. The tests of significance and confidence intervals, however, are affected by nonnormality. In general, the probability levels associated with the tests of significance or the confidence coefficients will not be correct. The F-test is generally regarded as being reasonably robust against nonnormality.

Effect on Confidence Intervals

Confidence interval estimates can be more seriously affected by nonnormality, particularly when the underlying distribution is highly skewed or has fixed boundaries. The two-tailed symmetric confidence interval estimates based on normality will not, in fact, be allocating equal probability to each tail if the distribution is asymmetric and may even violate natural boundaries for the parameter. The confidence interval estimate for proportion of affected individuals in a toxicity study, for example, may be less than zero or greater than one if the estimates ignore the nonnormality in the problem.

Detecting Nonnormality

Plots of the observed residuals $\boldsymbol{e}$ and skewness and kurtosis coefficients are helpful in detecting nonnormality. The **skewness coefficient** measures the asymmetry of the distribution whereas **kurtosis** measures the tendency of the distribution to be too flat or too peaked. The skewness coefficient for the normal distribution is 0; the kurtosis coefficient is 3.0. Some statistical computing packages provide these coefficients in the univariate statistics analysis. (Often, the kurtosis coefficient is expressed as a deviation from the value for the normal distribution.) When the sample size is sufficiently large, a frequency distribution of the residuals can be used to judge symmetry and kurtosis. A full-normal or half-normal plot, which gives a straight line under normality, is probably easier to use. These plots compare the ordered residuals from the data to the expected values of ordered observations from a normal distribution (with mean zero and unit variance). The full-normal plot uses the signed residuals; the half-normal plot uses the absolute values of the residuals. Different shapes of the normal plots reveal different kinds of departure from normality. More details on these plots are given in Section 11.1.

Improving Normality

Transformation of the dependent variable to a form that is more nearly normally distributed is the usual recourse to nonnormality. Statistical theory says that such a transformation exists if the distribution of the original

dependent variable is known. Many of the common transformations (such as the arcsin, the square root, the logarithmic, and the logistic transformations) were developed for situations in which the random variables were expected a priori to have specific nonnormal distributions.

In many cases, the sample data provide the only information available for determining the appropriate normalizing transformation. The plots of the residuals may suggest transformations, or several transformations might be tried and the one adopted that most nearly satisfies the normality criteria. Alternatively, an empirical method of estimating the appropriate power transformation might be used (Box and Cox, 1964). Chapter 12 is devoted to transformations of variables.

10.2 Heterogeneous Variances

Importance of Homogeneous Variance

The assumption of common variance plays a key role in ordinary least squares. The assumption implies that every observation on the dependent variable contains the same amount of information. Consequently, all observations in ordinary least squares receive the same weight. On the other hand, heterogeneous variances imply that some observations contain more information than others. Rational use of the data would require that more weight be given to those that contain the most information.

The minimum variance property of ordinary least squares estimators is directly dependent on this assumption. Equal weighting, as in ordinary least squares, does not give the minimum variance estimates of the parameters if the variances are not equal. Therefore, the direct impact of heterogeneous variances in ordinary least squares is a loss of precision in the estimates compared to the precision that would have been realized if the heterogeneous variances had been taken into account.

Data Having Heterogeneous Variances

Heterogeneous variance, as with nonnormality, is expected a priori with certain kinds of data. The same situations that give nonnormal distributions will usually give heterogeneous variances since the variance in most nonnormal distributions is related to the mean of the distribution. Even in situations where the underlying distributions are normal within groups, the variances of the underlying distributions may change from group to group. Most commonly, larger variances will be associated with groups having the larger means. Various plots of the residuals are useful for revealing heterogeneous variances.

Decreasing Heterogeneity

Two approaches to handling heterogeneous variances are transformation of the dependent variable and use of weighted least squares; the former is probably the more common. The transformation is chosen to make the variance homogeneous (or more nearly so) on the transformed scale. Prior information on the probability distribution of the dependent variable or

empirical information on the relationship of the variance to the mean may suggest a transformation. For example, the arcsin transformation is designed to stabilize the variance when the dependent variable is binomially distributed. Weighted least squares uses the original metric of the dependent variable but gives each observation weight according to the relative amount of information it contains. Weighted least squares is discussed in Section 12.5.1.

10.3 Correlated Errors

Correlations among the residuals may arise from many sources. It is common for data collected in a time sequence to have correlated errors; the error associated with an observation at one point in time will tend to be correlated with the errors of the immediately adjacent observations. Almost any physical process that is continuous over time will show serial correlations. Hourly measurements on the pollutant emissions from a coal smokestack, for example, have very high serial correlations. Biological studies in which repeated measurements are made over time on the same individuals, such as plant and animal growth studies or clinical trials, will usually have correlated errors.

Many of the experimental designs, including the randomized complete block design and the split-plot design, allow us to capitalize on the correlated errors among the observations within a block or within a whole plot to improve the precision of certain comparisons. The observations among samples within experimental units will have correlated errors, and the conventional analyses take these correlations into account. In some cases, however, correlations may be introduced inadvertently by the way the experiment is managed. For example, the grouping of experimental units for convenience in exposing them to a treatment, applying nutrient solution, taking measurements, and so forth, will tend to introduce positively correlated errors among the observations within the groups. These correlations are frequently overlooked and are not taken into account in the conventional analyses.

Impact of Correlated Errors

The impact of correlated errors on the ordinary least squares results is loss in precision in the estimates, similar to the effect of heterogeneous variances. Correlated errors that are not recognized appropriately in the analysis will seriously bias the estimates of variances with the direction and magnitude of the bias depending on the nature of the correlations. This, in turn, causes all measures of precision of the estimates to be biased and invalidates tests of significance.

Detecting Correlated Errors

The nature of the data frequently suggest the presence of correlated errors. Any data set collected in a time sequence should be considered suspect and treated as time series data unless the correlation can be shown to be

negligible. There are many texts devoted to the analysis of time series data (Fuller, 1996; Bloomfield, 1976). A clear understanding of the design and conduct of the experiment will reveal many potential sources of correlated errors. The more troublesome to detect are the inadvertent correlated errors arising from inadequate randomization of the experiment or failure to adhere to the randomization plan. In such cases, inordinately small error variances may provide the clue. In other cases, plotting of the residuals according to the order in which the data were collected or the grouping used in the laboratory may reveal patterns of residuals that suggest correlated errors.

Handling Correlated Errors

The remedy to the correlated errors problem is to utilize a model that takes into account the correlation structure in the data. Various time series models and analyses have been constructed to accomodate specific correlated error structures. **Generalized least squares** is a general approach to the analysis of data having correlated errors. This is an extension of weighted least squares where the entire variance–covariance matrix of the residuals is used. The difficulty with generalized least squares is that the covariances are usually not known and must be estimated from the data. This is a difficult estimation problem, unless the correlation structure is simple, and poor estimation of the correlation matrix can cause a loss in precision, rather than a gain, compared to ordinary least squares. Generalized least squares is discussed in Section 12.5.2.

10.4 Influential Data Points and Outliers

Influential Data Points

The method of ordinary least squares gives equal weight to every observation. However, every observation does not have equal impact on the various least squares results. For example, the slope in a simple linear regression problem is influenced most by the observations having values of the independent variable farthest from the mean. A single point far removed from the other data points can have almost as much influence on the regression results as all other points combined. Such observations are called **influential points** or **high leverage points**.

The potential influence of a data point on the least squares results is determined by its position in the X-space relative to the other points. In general, the more "distant" the point is from the center of the data points in the X-space, the greater is its potential for influencing the regression results.

Outliers

The term *outlier* refers to an observation which in some sense is inconsistent with the rest of the observations in the data set. An observation can be an outlier due to the dependent variable or any one or more of the independent variables having values outside expected limits. In this book the term *outlier* is restricted to a data point for which the value of the dependent

variable is inconsistent with the rest of the sample. The phrase *outlier in the residuals* refers to a data point for which the observed residual is larger than might reasonably be expected from random variation alone. The term *potentially influential observation* is used to refer to an observation that is an outlier in one or more of the independent variables. The context of the usage makes clear whether *outlier* refers to the value of the dependent variable or of the residual.

Origin of Outliers and Influential Points

A data point may be an outlier or a potentially influential point because of errors in the conduct of the study (machine malfunction; recording, coding, or data entry errors; failure to follow the experimental protocol) or because the data point is from a different population. The latter could result, for example, from management changes that take the system out of the realm of interest or the occurrence of atypical environmental conditions. A valid data point may appear to be an outlier, have an outlier in the residual, because the model being used is not adequately representing the process. On the other hand, a data point that is truly an outlier may not have an outlier residual, and almost certainly will not if it happens also to be an influential point. The influential data points tend to force the regression so that such points have small residuals.

Handling Outliers and Influential Points

Influential points and outliers need to be identified. Little confidence can be placed in regression results that have been dominated by a few observations, regardless of the total size of the study. The first concern should be to verify that these data points are correct. Clearly identifiable errors should be corrected if possible or else eliminated from the data set. Data points that are not clearly identified as errors or that are found to be correct should be studied carefully for the information they might contain about the system being studied. Do they reflect inadequacies in the model or inadequacies in the design of the study? Outliers and overly influential data points should not be discarded indiscriminately. The outlier might be the most informative observation in the study.

Detection of Influential Points

Detection of the potentially more influential points is by inspection of the diagonal elements of $\boldsymbol{P}$ the projection matrix. The diagonal elements of $\boldsymbol{P}$ are measures of the Euclidean distances between the corresponding sample points and the centroid of the sample X-space. Whether a potentially influential point has, in fact, been influential is determined by measuring directly the impact of each data point on various regression results. Appropriate influence statistics are discussed in Section 11.2.

Detection of Outliers

Outliers are detected by analysis of the observed residuals and related statistics. It is usually recommended that the residuals first be standardized to have a common variance. Some suggest the use of recursive residuals (Hedayat and Robson, 1970). A residual that is several standard deviations from zero identifies a data point that needs careful review. Plots of residuals for detecting nonnormality and heterogeneous variances are also effective in identifying outliers. The detection of outliers is discussed in Section 11.1.

10.5 Model Inadequacies

Missing Independent Variables

The ordinary least squares estimators are unbiased if the model is correct. They will not be unbiased if the model is incorrect in any of several different ways. If, for example, an important independent variable has been omitted from the model, the residual mean square is a (positively) biased estimate of σ^2 and the regression coefficients for all independent variables are biased (unless the omitted variable is orthogonal to all variables in the model). The common linear model that uses only the first power of the independent variables assumes that the relationship of Y to each of the independent variables is linear and that the effect of each independent variable is independent of the other variables. Omitting any important higher-order polynomial terms, including product terms, has the same effect as omitting an independent variable.

Approximating the "True" Model

One does not expect a complicated physical, chemical, or biological process to be linear in the parameters. In this sense, the ordinary linear least squares model (including higher-degree polynomial terms) must be considered an approximation of the true process. The rationale for using a linear model, in cases where the true relationship is almost certainly nonlinear, is that any nonlinear function can be approximated to any degree of accuracy desired with an appropiate number of terms of a linear function. Thus, the linear model is used to provide what is believed to be a satisfactory approximation in some limited region of interest. To the extent that the approximation is not adequate, the least squares results will contain biases similar to those created by omitting a variable.

Detecting Inadequacies

Detection of model inadequacies will depend on the nature of the problem and the amount of information available on the system. Bias in the residual mean square and, hence, indication of an omitted term, can be detected if an independent estimate of σ^2 is available as would be the case in most designed experiments. In other cases, previous experience might provide some idea of the size of σ^2 from which a judgment can be made as to the presence of bias in the residual mean square. Overlooked higher-order polynomial terms are usually easily detected by appropriate residuals plots. Independent variables that are missing altogether are more difficult to detect. Unusual patterns of behavior in the residuals may provide clues.

Nonlinear Models

More realistic nonlinear models might be formulated as alternatives to the linear approximations. Some nonlinear models will be such that they can be linearized by an appropriate transformation of the dependent variable. These are called **intrinsically linear models**. Ordinary least squares can be used on linearized models if the assumptions on the errors are satisfied after the transformation is made. The intrinsically nonlinear models require the use of **nonlinear least squares** for the estimation of the parameters. The nonlinear form of even the intrinsically linear models might be preferred if it is believed the least squares assumptions are more nearly

satisfied in that form. Nonlinear models and nonlinear least squares are discussed in Chapter 15.

10.6 The Collinearity Problem

Near-Singularities in $\boldsymbol{X}$

Singularity of $\boldsymbol{X}$ results when some linear function of the columns of $\boldsymbol{X}$ is exactly equal to the zero vector. Such cases become obvious when the least squares analysis is attempted because the unique $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ does not exist. A more troublesome situation arises when the matrix is only close to being singular; a linear function of the vectors is nearly zero. Redundant independent variables—the same information expressed in different forms—will cause $\boldsymbol{X}$ to be nearly singular. Interdependent variables that are closely linked in the system being studied can cause near-singularities in $\boldsymbol{X}$.

A unique solution to the normal equations exists in these nearly singular cases but the solution is very unstable. Small changes (random noise) in the variables Y or X can cause drastic changes in the estimates of the regression coefficients. The variances of the regression coefficients, for the independent variables involved in the near-singularity, become very large. In effect, the variables involved in the near-singularity can serve as surrogates for each other so that widely different combinations of the independent variables can be used to give nearly the same value of Y. The difficulties that arise from $\boldsymbol{X}$ being nearly singular are referred to collectively as the **collinearity problem**. The collinearity problem was defined geometrically in Section 6.5.

Effects of Collinearity

The impact of collinearity on least squares is very serious if primary interest is in the regression coefficients per se or if the purpose is to identify "important" variables in the process. The estimates of the regression coefficients can differ greatly from the parameters they are estimating, even to the point of having incorrect sign. The collinearity will allow "important" variables to be replaced in the model with incidental variables that are involved in the near-singularity. Hence, the regression analysis provides little indication of the relative importance of the independent variables.

The use of the regression equation for prediction is not seriously affected by collinearity as long as the correlational structure observed in the sample persists in the prediction population and prediction is carefully restricted to the sample X-space. However, prediction to a system where the observed correlation structure is not maintained or for points outside the sample space can be very misleading. The sample X-space in the presence of near-collinearities becomes very narrow in certain dimensions so that it is easy to choose prediction points that fall outside the sample space and, at the same time, difficult to detect when this has been done. Points well within the limits of each independent variable may be far outside the sample space.

Detecting Collinearity

Most regression computer programs are not designed to warn the user au-

tomatically of the presence of near-collinearities. Certain clues are present, however: unreasonable values for regression coefficients, large standard errors, nonsignificant partial regression coefficients when the model provides a reasonable fit, and known important variables appearing as unimportant (or with an opposite sign from what the theory would suggest) in the regression results. High correlations between independent variables will identify near-collinearities involving two variables but may miss those involving more than two variables. A more direct approach to detecting the presence of collinearity is with a singular value decomposition of $\boldsymbol{X}$ or an eigenanalysis of $\boldsymbol{X}'\boldsymbol{X}$. These were discussed in Sections 2.7 and 2.8. Their use and other collinearity diagnostics are discussed in Section 11.3.

Handling Collinearity

The remedies for the collinearity problem depend on the objective of the model-fitting exercise. If the objective is prediction, collinearity causes no serious problem within the sample X-space. The limitations discussed previously must be understood, however. When primary interest is in estimation of the regression coefficients, one of the biased regression methods may be useful (Chapter 13). A better solution, when possible, is to obtain new data or additional data such that the sample X-space is expanded to remove the near-singularity. It is not likely that this will be possible when the near-singularity is the result of internal constraints of the system being studied. When the primary interest of the research is to identify the "important" variables in a system or to model the system, the regression results in the presence of severe collinearity will not be very helpful and can be misleading. It is more productive for this purpose to concentrate on understanding the correlational structure of the variables and how the dependent variable fits into this structure. Principal component analysis, Gabriel's (1971) biplot, and principal component regression can be helpful in understanding this structure. These topics are discussed in Chapter 13.

10.7 Errors in the Independent Variables

The original model assumed that the independent variables were measured without error; they were considered to be constants in the regression model. With the errors-in-variables model, the true values of the independent variables are masked by measurement errors. Thus, the observed X_i is

$$X_i = Z_i + U_i, \tag{10.1}$$

where Z_i is the unobserved true value of X_i and U_i is the measurement error. The error U_i is assumed to have zero mean and variance σ_U^2. For example, in an experiment to study the effect of temperature in an oven on baking time, the observed temperature may be different from the actual temperature in the oven. Fuller (1987) gives several examples where the independent variable is measured with error. In an experiment to study

the relationship between dry weight of the plant and available nitrogen in the leaves, the independent variable is measured with error. Typically, the true nitrogen content (Z_i) in the leaves is unknown and is estimated (X_i) from a small sample of leaves. See also Carroll, Ruppert, and Stefanski (1995) for some examples.

The regression model assumes that Y_i is a function of the true value Z_i:

$$Y_i = \mu + \beta Z_i + v_i, \tag{10.2}$$

where v_i are assumed to be independent mean zero and variance σ_v^2 random variables, and are independent of Z_i and U_i. However, we estimate the parameters μ and β using the model

$$Y_i = \mu + \beta X_i + \epsilon_i. \tag{10.3}$$

Bias in $\widehat{\beta}$

The ordinary least squares estimator of β, based on the model in equation 10.3, is

$$\begin{aligned} \widehat{\beta} &= \sum x_i Y_i / \sum x_i^2 & (10.4)\\ &= \beta \left[\frac{(\sum z_i^2 + \sum z_i u_i)}{(\sum z_i^2 + 2\sum z_i u_i + \sum u_i^2)} \right] + \frac{\sum x_i v_i}{\sum x_i^2}, & (10.5) \end{aligned}$$

where x_i, z_i, and u_i represent the centered values of X_i, Z_i, and U_i, respectively. Note that, if there is no measurement error ($U_i = 0$), the first term reduces to β and, since the second term has zero expectation, $\widehat{\beta}$ is unbiased for β. However, if measurement error is present, then the first term shows the bias in β. If we assume that the Z_i are independently and identically distributed $N(0, \sigma_z^2)$, the U_i are independent and identically distributed $N(0, \sigma_U^2)$ and that the $\{Z_i\}$ and $\{U_i\}$ are independent, then Fuller (1987) shows that

$$\mathcal{E}(\widehat{\beta}) = \beta \left[\frac{\sigma_Z^2}{\sigma_Z^2 + \sigma_U^2} \right] = \beta \left[\frac{1}{1 + \sigma_U^2/\sigma_Z^2} \right]. \tag{10.6}$$

Also, if the true independent variable values Z_i are considered fixed, then the expectation of $\widehat{\beta}$ is

$$\mathcal{E}(\widehat{\beta}) \doteq \beta \left[\frac{1}{1 + n\sigma_U^2 / \sum z_i^2} \right]. \tag{10.7}$$

The denominators of equations 10.6 and 10.7 are always greater than one whenever there is measurement error, $\sigma_U^2 > 0$. Thus, $\widehat{\beta}$ is biased toward zero. The bias is small if σ_U^2 is small relative to σ_z^2 or $\sum z_i^2/n$. That is, the bias is small if the measurement errors in the independent variable are small relative to the variation in the true values of the independent variable.

Proposals for Estimating β

There have been numerous proposals for estimating β under these conditions. Some of the procedures assume that additional information is available.

(a) **Known Reliability Ratio: $\sigma_Z^2/(\sigma_Z^2+\sigma_U^2)$.** If we assume that $Z_i \sim \text{NID}(0,\sigma_Z^2), U_i \sim \text{NID}(0,\sigma_U^2)$, and that $\{Z_i\}$ and $\{U_i\}$ are independent, then

$$\widetilde{\beta}_R = \left[\frac{\sigma_Z^2+\sigma_U^2}{\sigma_Z^2}\right]\widehat{\beta} \tag{10.8}$$

is an unbiased estimator of β. Fuller (1987) gives examples from psychology, sociology, and survey sampling where the *reliability ratio* $\sigma_Z^2/(\sigma_Z^2+\sigma_U^2)$ is known.

(b) **Known Measurement Error Variance: σ_U^2.** In some situations, the scientist may have a good knowledge of the measurement error variance σ_U^2. For example, it may be possible to obtain a large number of repeated measurements to determine σ_U^2. Madansky (1959) and Fuller (1987) consider the estimator

$$\widetilde{\beta}_U = \frac{\sum x_i y_i}{\sum x_i^2 - (n-1)\sigma_U^2} \tag{10.9}$$

which adjusts the denominator for the measurement error variance.

(c) **Known Ratio of Error Variances: $\delta = \sigma_v^2/\sigma_U^2$.** Under the normality assumptions on Z_i, U_i, and v_i, Fuller (1987) shows that the maximum likelihood estimator of β is

$$\widetilde{\beta}_\delta = \frac{\sum y_i^2 - \delta\sum x_i^2 + [(\sum y_i^2 - \delta\sum x_i^2)^2 + 4\delta(\sum x_i y_i]^{1/2}}{2\sum x_i y_i}, \tag{10.10}$$

where $\delta = \sigma_v^2/\sigma_U^2$. It can be shown that $\widetilde{\beta}_\delta$ is also the "orthogonal regression" estimator of β obtained by minimizing the distance

$$\sum(Y_i - \mu - \beta Z_i)^2 + \delta\sum(X_i - Z_i)^2 \tag{10.11}$$

with respect to μ, β, $Z_1,\ldots,Z_n$. When $\delta = 1$, equation 10.11 is the sum of the Euclidean distances between the observed vector $(\,Y_i \quad X_i\,)$ and the point $(\,\mu+\beta Z_i \quad Z_i\,)$ on the line that generated it. [Carroll, Ruppert, and Stefanski (1995) prefer to restrict the use of the term "orthogonal regression" to the case where $\delta = 1$.]

Comparison of the Estimators

Riggs, Guarnieri, and Addelman (1978) used computer simulation to study the behavior of a large number of published estimators and several additional ones they developed. Fuller (1987) and Carroll, Ruppert, and Stefanski (1995) also discuss the behavior of these estimators. To summarize:

(i) $\widetilde{\beta}_U$ behaves erratically whenever measurement error variances are large;

(ii) $\widetilde{\beta}_R$ is unbiased. However, when σ_Z^2 and σ_U^2 are replaced by their estimates, the sampling distribution of $\widetilde{\beta}_R$ is highly skewed;

(iii) $\widetilde{\beta}_\delta$ tends to give highly unreliable estimates when σ_U^2 is large and n is small.

The reader is referred to the original references for more discussion of the problems and the summary of their comparisons.

Instrumental Variables

Several alternative approaches for estimating are also available. We discuss two such approaches. One approach to the errors-in-variables problem is to use information from other variables that are correlated with Z_i, but not with U_i, to obtain consistent estimators of β. Such variables are called **instrumental variables** . (A **consistent** estimator converges to the true value as the sample size gets large.) For example, the true nitrogen in the leaves Z_i may be correlated with the amount of nitrogen fertilizer W_i applied to the experimental plot. In this case, it may be reasonable to assume that W_i is not correlated with the measurement error U_i. An instrumental variable estimator of β is given by

$$\widetilde{\beta}_W = \frac{\sum w_i Y_i}{\sum w_i X_i}, \tag{10.12}$$

where w_i is the centered value of W_i. The reader is referred to Feldstein (1974), Carter and Fuller (1980), and Fuller (1987) for more discussion on the use of instrumental variables.

SIMEX Estimator

We have seen that, in the errors-in-variables model, the ordinary least squares estimator $\widehat{\beta}$ of β is biased and its expectation is given by $\beta\sigma_Z^2/(\sigma_Z^2 + \sigma_U^2)$. The effect of measurement error on the ordinary least squares estimator can also be determined experimentally via simulations. The **Simulation Extrapolation (SIMEX)** method of Cook and Stefanski (1995) determines this effect using simulations at various known levels of the measurement error and extrapolates the results to the no-measurement error case to obtain the SIMEX estimator of β.

Assume that σ_U^2 is known. Consider m data sets with independent variables $X_i^{(\lambda)} = X_i + \lambda^{1/2} U_i^*$, $i = 1, \ldots, n$, where $U_i^* \sim \text{NID}(0, \sigma_U^2)$, and λ takes known values $0 = \lambda_1 < \lambda_2 < \cdots < \lambda_m$. Note that the measurement error variance in $X_i^{(\lambda)}$ is $(1+\lambda)\sigma_U^2$ and we are considering $(m-1)$ new data sets with increasing measurement errors. The least squares estimate $\widehat{\beta}_\lambda$ from the regression of Y_i on $X_i^{(\lambda)}$ consistently estimates $\beta_\lambda = \beta\sigma_Z^2/[\sigma_Z^2 + (1+\lambda)\sigma_U^2]$. That is, as n tends to infinity, the estimator $\widehat{\beta}_\lambda$ converges to β_λ. Note that, at $\lambda = -1$, β_λ estimates β consistently. The SIMEX method uses $\widehat{\beta}_{\lambda_1}, \ldots, \widehat{\beta}_{\lambda_m}$ to fit a model for $\widehat{\beta}_\lambda$ as a function of λ and uses this function to extrapolate back to the no-measurement error case, $\lambda = -1$. This extrapolated value is called the SIMEX estimate of β. The process is described schematically in Figure 10.1.

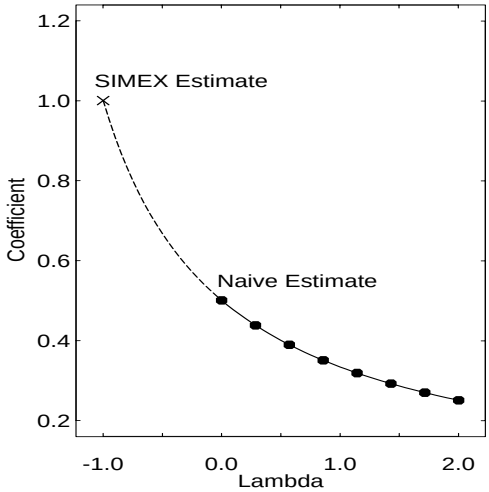

FIGURE 10.1. *A generic plot of the effect of measurement error of size* $(1+\lambda)\sigma_U^2$ *on the slope estimates. The ordinary least squares estimate occurs at* $\lambda = 0$ *and the SIMEX estimate is an extrapolation to* $\lambda = -1$.

Cook and Stefanski (1995) recommend generating several data sets at each value of λ and use the average of the estimates of β to obtain $\widehat{\beta}_\lambda$. For example, at $\lambda = .5$, generate 10 sets of $X_i^{(.5)}$; compute $\widehat{\beta}_{.5}$ for each of the 10 data sets and compute the average of these 10 estimates to get $\overline{\widehat{\beta}}_{.5}$. Similarly, obtain $\overline{\widehat{\beta}}_\lambda$ for several values of λ. Use these $\overline{\widehat{\beta}}_\lambda$s to obtain the SIMEX estimate of β. See Carroll, Ruppert, and Stefanski (1995) for properties and extensions of the SIMEX estimates.

There are serious problems associated with estimation of other parameters and variances in the errors-in-variables model. The reader is referred to Fuller (1987) and Carroll, Ruppert, and Stefanski (1995) for more complete discussions. These authors also considered extensions to multiple and nonlinear regression models with measurement errors in the independent variables.

Control with Design

The errors-in-variables issue greatly complicates the regression problem. There appears to be no one solution that does well in all situations and it is best to avoid the problem whenever possible. The bias from ordinary least squares is dependent on the ratio of σ_U^2 to σ_Z^2 or to $\sum z_i^2/n$. Thus, the problem can be minimized by designing the research so that the dispersion in X is large relative to any measurement errors. In such cases, ordinary least squares should be satisfactory.

10.8 Summary

This chapter is a synopsis of the common problems in least squares regression emphasizing their importance and encouraging the user to be critical of his or her own results. Because least squares is a powerful and widely used tool, it is important that the user be aware of its pitfalls. Some of the diagnostic techniques (such as the analysis of residuals) are useful for detection of several different problems. Similarly, some of the remedial methods (such as transformations) attack more than one kind of problem. The following three chapters are devoted to discussions of the tools for detecting the problems and some of the remedies.

10.9 Exercises

10.1. Several levels of a drug were used to assess its toxic effects on a particular animal species. Twenty-four animals were used and each was administered a particular dose of the drug. After a fixed time interval, each animal was scored as 0 if it showed no ill effects and as 1 if a toxic effect was observed. That is, the dependent variable takes the value of 0 or 1 depending on the absence or presence of a toxic reaction.

(a) Which assumptions of ordinary least squares would you expect not to be satisfied with this dependent variable?

(b) The dependent variable was used in a linear regression on dose. The resulting regression equation was $\widehat{Y} = -.214 + .159X$. Plot this regression line for $X = 1$ to $X = 8$. Superimpose on the plot what you might expect the observed data to look like if 24 approximately equally spaced dose levels were used. What problems do you see now?

(c) The researcher anticipated using $\widehat{Y}$ to estimate the proportion of affected individuals at the given dose. What is the estimated proportion of individuals that will be affected by a dose of $X = 2$ units? Use the conventional method to compute the 95% confidence interval estimate of the mean at $X = 2$ if $s^2 = .1284$ with 22 degrees of freedom, $\overline{X} = 4.5$, and $\sum(X_i - \overline{X})^2 = 126$. Comment on the nature of this interval estimate.

(d) Suppose each observation consisted of the proportion of mosquitoes in a cage of 50 that showed response to the drug rather than the response of a single animal. Would this have helped satisfy some of the least squares assumptions? Which?

10.2. Identify an independent variable in your area of research that you would not expect to be normally distributed. How is this variable usually handled in the analysis of experimental results?

10.3. Suppose there are three independent observations that are to be averaged. The known variances of the three observations are 4, 9, and 16. Two different averages are proposed, the simple arithmetic average and the weighted average where each observation is weighted by the reciprocal of its variance. Use variances of linear functions, equation 3.21 and following, to demonstrate that the weighted average has the smaller variance. Can you find any other weighting that will give an even smaller variance?

10.4. Find a data set from your area of research in which you do *not* expect the variances to be homogeneous. Explain how you expect the variances to behave. How are these data usually handled in analysis?

10.5. A plant physiologist was studying the relationship between intercepted solar radiation and plant biomass produced over the growing season. Several experimental plots under different growing conditions were monitored for radiation. Several times during the growing season biomass samples were taken from the plots to measure growth. The resulting data for each experimental plot showed cumulative solar radiation and biomass for the several times the biomass was measured during the season. Would you expect the dependent variable, biomass, to have constant variance over the growing season? Would you expect the several measurements of biomass on each plot to be statistically independent? Would you expect the measurements from different random experimental units to be statistically independent?

10.6. The relatively greater influence of observations farther from the center of the X-space can be illustrated using simple linear regression. Express the slope of the regression line as $\widehat{\beta}_1 = \sum(X_i - \overline{X})Y_i / \sum(X_i - \overline{X})^2$. In this form it is clear that a perturbation of the amount δ on any $Y_{i'}$ changes $\widehat{\beta}_1$ by the amount $\delta(X_{i'} - \overline{X}) / \sum(X_i - \overline{X})^2$. (Substitute $Y_{i'} + \delta$ for $Y_{i'}$ to get a new $\widehat{\beta}_1$ and subtract out the original $\widehat{\beta}_1$.) Assume a perturbation of $\delta = 1$ on each Y_i in turn. Compute the amount $\widehat{\beta}_1$ would change if the values of X are 0, 1, 2, and 9. Compute $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ for this example. Which observation has the largest diagonal element of $\boldsymbol{P}$?

10.7. Find an example in your field for which you might expect collinearity to be a problem. Explain why you expect there to be collinearity.

11
REGRESSION DIAGNOSTICS

Chapter 10 summarized the problems that are encountered in least squares regression and the impact of these problems on the least squares results.

This chapter presents methods for detecting problem areas. Included are graphical methods for detecting failures in the assumptions, unusual observations, and inadequacies in the model, statistics to flag observations that are dominating the regression, and methods of detecting situations in which strong relationships among the independent variables are affecting the results.

Regression diagnostics refers to the general class of techniques for detecting problems in regression—problems with either the model or the data set. This is an active field of research with many recent publications. It is not clear which of the proposed techniques will eventually prove most useful. Some of the simpler techniques that appear to be gaining favor are presented in this chapter. Belsley, Kuh, and Welsch (1980) and Cook and Weisberg (1982) are recommended for a more thorough coverage of the theory and methods of diagnostic techniques.

11.1 Residuals Analysis

Characteristics of the Residuals

Analysis of the regression residuals, or some transformation of the residuals, is very useful for detecting inadequacies in the model or problems in the data. The true errors in the regression model are assumed to be normally and independently distributed random variables with zero mean and common variance $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{I}\sigma^2)$. The observed residuals, however, are not independent and do not have common variance, even when the $\boldsymbol{I}\sigma^2$ assumption is valid. Under the usual least squares assumptions, $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$ has a multivariate normal distribution with $\mathcal{E}(\boldsymbol{e}) = \mathbf{0}$ and $\mathbf{Var}(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{P})\sigma^2$. The diagonal elements of $\mathbf{Var}(\boldsymbol{e})$ are not equal, so the observed residuals do not have common variance; the off-diagonal elements are not zero, so they are not independent.

Standardized Residuals

The heterogeneous variances in the observed residuals are easily corrected by standardizing each residual. The variances of the residuals are estimated by the diagonal elements of $(\boldsymbol{I} - \boldsymbol{P})s^2$. Dividing each residual by its standard deviation gives a **standardized residual**, denoted with r_i,

$$r_i = \frac{\boldsymbol{e}_i}{s\sqrt{(1 - v_{ii})}}, \tag{11.1}$$

where v_{ii} is the ith diagonal element of $\boldsymbol{P}$. All standardized residuals (with σ in place of s in the denominator) have unit variance. The standardized residuals behave much like a Student's t random variable except for the fact that the numerator and denominator of r_i are not independent.

Studentized Residuals

Belsley, Kuh, and Welsch (1980) suggest standardizing each residual with an estimate of its standard deviation that is independent of the residual. This is accomplished by using, as the estimate of σ^2 for the ith residual, the residual mean square from an analysis where that observation has been omitted. This variance is labeled $s^2_{(i)}$, where the subscript in parentheses indicates that the ith observation has been omitted for the estimate of σ^2. The result is the **Studentized residual**, denoted r_i^*,

$$r_i^* = \frac{e_i}{s_{(i)}\sqrt{1 - v_{ii}}}. \tag{11.2}$$

Each Studentized residual is distributed as Student's t with $(n - p' - 1)$ degrees of freedom when normality of $\boldsymbol{\epsilon}$ holds. As with e_i and r_i, the r_i^* are not independent of each other. Belsley, Kuh, and Welsch show that the $s_{(i)}$ and Studentized residuals can be obtained from the ordinary residuals without rerunning the regression with the observation omitted.

Notation

The standardized residuals r_i are called Studentized residuals in many references [e.g., Cook and Weisberg (1982); Pierce and Gray (1982); Cook and Prescott (1981); and SAS Institute, Inc. (1989b)]. Cook and Weisberg refer to r_i as the Studentized residual with *internal* Studentization in contrast to *external* Studentization for r_i^*. The r_i^* are called the cross-validatory or jackknife residuals by Atkinson (1983) and RSTUDENT by

Belsley, Kuh, and Welsch (1980) and SAS Institute, Inc. (1989b). The terms *standardized* and *Studentized* are used in this text as labels to distinguish between r_i and r_i^*.

Using Residuals

The observed residuals and the scaled versions of the observed residuals have been used extensively to study validity of the regression model and its assumptions. The heterogeneous variances of the observed residuals and the lack of independence among all three types of residuals complicate interpretation of their behavior. In addition, there is a tendency for inadequacies in the data to be spread over several residuals. For example, an outlier will have the effect of inflating residuals on several other observations and may itself have a relatively small residual. Furthermore, the residuals from least squares regression will tend to be "supernormal." That is, when the normality assumption is *not* met, the observed residuals from a least squares analysis will fit the normal distribution more closely than would the original ϵ_i (Huang and Bolch, 1974; Quesenberry and Quesenberry, 1982; Cook and Weisberg, 1982). As a result, there will be a tendency for failures in the model to go undetected when residuals are used for judging goodness-of-fit of the model.

In spite of the problems associated with their use, the observed, standardized, and Studentized residuals have proven useful for detecting model inadequacies and outliers. For most cases, the three types of residuals give very similar patterns and lead to similar conclusions. The heterogeneous variances of e_i can confound the comparisons somewhat, and for that reason use of one of the standardized residuals r_i or r_i^* is to be preferred if they are readily available. The primary advantage of the Studentized residuals over the standardized residuals is their closer connection to the t-distribution. This allows the use of Student's t as a convenient criterion for judging whether the residuals are inordinately large.

Exact tests of the behavior of the observed residuals are not available; approximations and subjective judgments must be used. The use of the standardized or Studentized residuals as a check for an outlier is a multiple testing procedure, since the residual to be tested will be the largest out of the sample of n, and appropriate allowances on α must be made. The first-order Bonferroni bound on the probability would suggest using the critical value of t for $\alpha = \alpha^*/n$, as was done for the Bonferroni confidence intervals in Chapter 4. (α^* is the desired overall significance level.) Cook and Prescott (1981), in a study assessing the accuracy of the Bonferroni significance levels for detecting outliers in linear models, conclude that the bounds can be expected to be reasonably accurate if the correlations among the residuals are not excessively large. Cook and Weisberg (1982) suggest using $\alpha = v_{ii}\alpha^*/p'$ for testing the ith Studentized residual. This choice of α maintains the overall significance level but gives greater power to cases with large v_{ii}.

Recursive Residuals

Another class of residuals, **recursive residuals**, are constructed so that they are independent and identically distributed when the model is correct

and are recommended by some for residuals analysis (Hedayat and Robson, 1970; Brown, Durbin, and Evans, 1975; Galpin and Hawkins, l984; Quesenberry, 1986). Recursive residuals are computed from a sequence of regressions starting with a base of p' observations (p' = number of parameters to be estimated) and adding one observation at each step. The regression equation computed at each step is used to compute the residual for the next observation to be added. This sequence continues until the last residual has been computed. There will be $(n - p')$ recursive residuals; the residuals from the first p' observations will be zero.

Computation of Recursive Residuals

Assume a particular ordering of the data has been adopted for the purpose of computing the recursive residuals. Let y_r and x_r' be the rth rows from $\boldsymbol{Y}$ and $\boldsymbol{X}$, respectively. Let $\boldsymbol{X}_r$ be the first r rows of $\boldsymbol{X}$ and $\widehat{\boldsymbol{\beta}}_r$ be the least squares solution using the first r observations in the chosen ordering. Then the *recursive residual* is defined as

$$w_r = \frac{y_r - \boldsymbol{x}_r'\widehat{\boldsymbol{\beta}}_{r-1}}{[1 - \boldsymbol{x}_r'(\boldsymbol{X}_{r-1}'\boldsymbol{X}_{r-1})^{-1}\boldsymbol{x}_r]^{1/2}} \tag{11.3}$$

for $r = p' + 1, \ldots, n$. The original proposal defined the recursive residuals for time sequence data. Galpin and Hawkins (1984) contend, however, that they are useful for all data sets, but particularly so when there are natural orderings to the data.

Characteristics of Recursive Residuals

Recursive residuals are independent and have common variance σ^2. Each is explicitly associated with a particular observation and, consequently, recursive residuals seem to avoid some of the "spreading" of model defects that occurs with ordinary residuals. Since the recursive residuals are independently and identically distributed, exact tests for normality and outliers can be used. The major criticisms of recursive residuals are the greater computational effort required, no residuals are associated with the first p' observations used as the base, and the residuals are not unique since the data can be ordered in different ways. Appropriate computer programs can remove the first problem. The last two are partially overcome by computing recursive residuals for different orderings of the data.

Graphical Techniques

Graphical techniques are very effective for detecting abnormal behavior of residuals. If the model is correct and the assumptions are satisfied, the residuals should appear in any plot as random variation about zero. Any convincing pattern to the residuals would suggest some inadequacy in the model or the assumptions. To emphasize the importance of plotting, Anscombe (1973) presents four (artificial) data sets that give identical least squares regression results [same $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{Y}}$, SS(Total), SS(Regression), SS(Residual), and R^2], but are strikingly different when plotted. The fitted model appears equally good in all cases if one looks only at the quantitative results. The plots of Y versus X, however, show obvious differences [Figure 11.1; adapted from Anscombe (1973)].

Anscombe Plots

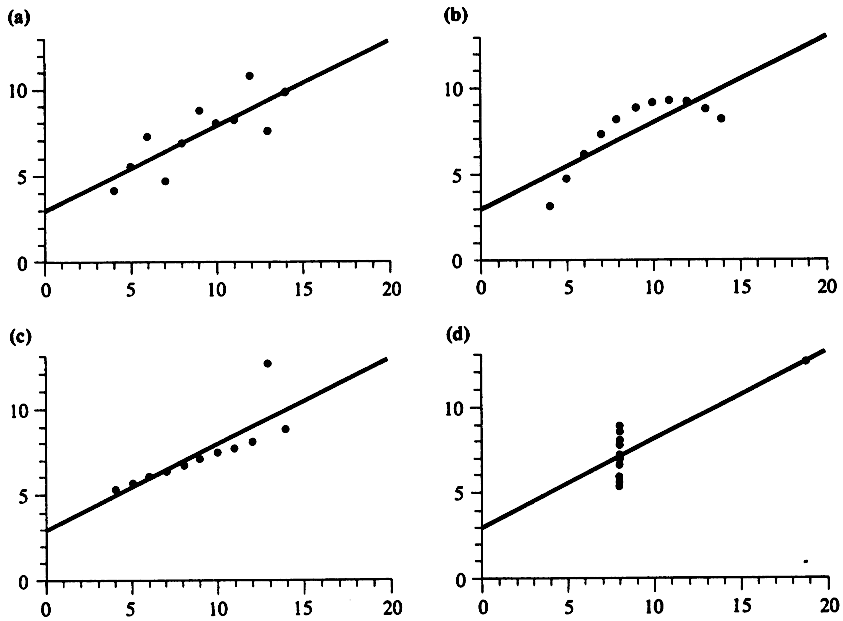

FIGURE 11.1. *Four data sets that give the same quantitative results for the linear regression of Y on X. [Adapted from Anscombe (1973).]*

The first data set, Figure 11.1(a), shows a typical linear relationship between Y and X with apparent random scatter of the data points above and below the regression line. This is the expected pattern if the model is adequate and the ordinary least squares assumptions hold.

The data in Figure 11.1(b) show a distinct quadratic relationship and a very patterned set of residuals. It is clear from the plot that the linear model is inadequate and that the fit would be almost perfect if the model were expanded to include a quadratic term.

Figure 11.1(c) illustrates a case where there is a strict linear relationship between Y and X except for one aberrant data point. Removal of this one point would cause the residual sum of squares to go to zero. The residuals pattern is a clear indication of a problem with the data or the model. If this is a valid data point, the model must be inadequate. It may be that an important independent variable has been omitted.

The data in Figure 11.1(d) represent a case where the entire regression relationship is determined by one observation. This observation is a particularly influential point because it is so far removed (on the X-scale) from the other data points. Even if this is a valid data point, one could place little faith in estimates of regression parameters so heavily dependent on a single observation.

The Anscombe plots emphasize the power of simple graphical techniques for detecting inadequacies in the model. There are several informative plots one might use. No single plot can be expected to detect all types of problems. The following plots are presented as if the ordinary residuals e_i are being used. In all cases, the standardized, Studentized, or recursive residuals could be used.

11.1.1 Plot of $\boldsymbol{e}$ *Versus* $\widehat{\boldsymbol{Y}}$

Expected Behavior

The plot of the residuals against the fitted values of the dependent variable is particularly useful. A random scattering of the points above and below the line $e = 0$ with nearly all the data points being within the band defined by $e = \pm 2s$ (Figure 11.2) is expected if the assumptions are satisfied. ($\widehat{\boldsymbol{Y}}$is used rather than $\boldsymbol{Y}$ because $\boldsymbol{e}$ is orthogonal to $\widehat{\boldsymbol{Y}}$ but not to $\boldsymbol{Y}$. A plot of $\boldsymbol{e}$ versus $\boldsymbol{Y}$ will show a pattern due to this lack of orthogonality.)

Any pattern in the *magnitude* of the dispersion about zero associated with changing $\widehat{Y}_i$ suggests heterogeneous variances of ϵ_i. The fan-shaped pattern in Figure 11.3 is the typical pattern when the variance increases with the mean of the dependent variable. This is the pattern to be expected if the dependent variable has a Poisson or a log-normal distribution, for example, or if the errors are multiplicative rather than additive. Binomially distributed data would show greater dispersion when the proportion of "successes" is in the intermediate range.

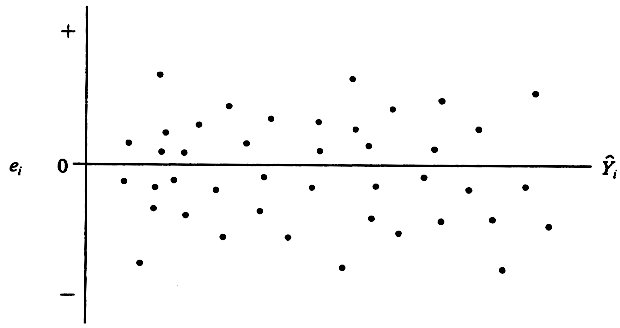

FIGURE 11.2. *Typical pattern expected for a plot of* $\mathbf{e}$ *versus* $\widehat{\boldsymbol{Y}}$ *when assumptions are met.*

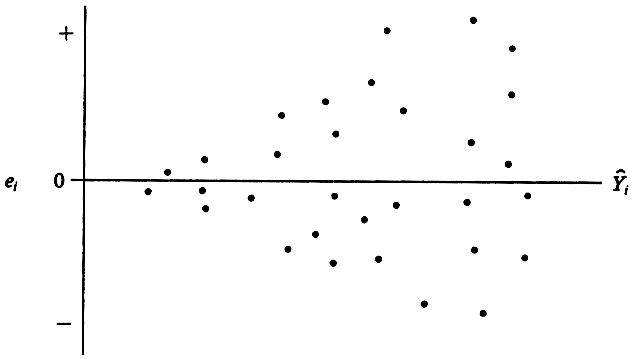

FIGURE 11.3. *Plot of* $\mathbf{e}$ *versus* $\widehat{\boldsymbol{Y}}$ *showing increasing dispersion (larger variance) with larger* $\widehat{\boldsymbol{Y}}$.

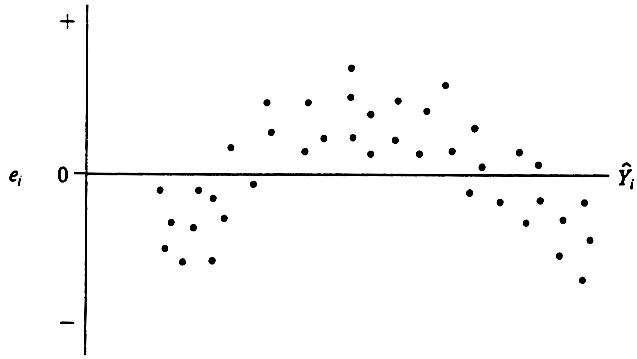

FIGURE 11.4. *An asymmetric (curved) pattern of residuals plotted against* $\widehat{\boldsymbol{Y}}$ *suggests that the model is missing an important independent variable, perhaps a quadratic term.*

Detecting Model Inadequacies

Any asymmetry of the distribution of the residuals about zero suggests a problem with the model or the basic assumptions. A majority of relatively small negative residuals and fewer but larger positive residuals would suggest a positively skewed distribution of residuals rather than the assumed symmetric normal distribution. (A skewed distribution would be more evident in either a frequency plot or a normal plot of the residuals.) A preponderance of negative residuals for some regions of $\widehat{\boldsymbol{Y}}$ and positive residuals in other regions, such as the curved pattern of residuals in Figure 11.4, suggests a systematic error in the data or an important variable missing from the model. The obvious candidate in this illustration would be the square of one of the present independent variables. A missing independent variable can cause unusual patterns of residuals depending on the scatter of the data with respect to that variable.

Outlier Residuals

An outlier residual would appear in any of the plots of $\boldsymbol{e}$ as a point well outside the band containing most of the residuals. However, an outlier in $\boldsymbol{Y}$ will not necessarily have an outlier residual.

Example 11.1

The Lesser–Unsworth data in Exercise 1.19 related seed weight of soybeans to cumulative solar radiation for plants exposed to two different levels of ozone. The Studentized residuals from the regression of $Y_i = (\text{seed weight})^{1/2}$ on solar radiation and ozone level are plotted against $\widehat{Y}_i$ in Figure 10.5. The residuals for the low and high levels of ozone are shown as dots and ×s, respectively. One observation from the high ozone treatment seems to stand out from the others. Is this residual the result of an error in the data, an incorrect model, or simply random variation in the data?

The value of this Studentized residual is $r_i^* = 2.8369$. This is distributed as Student's t with $(n - p' - 1) = 8$ degrees of freedom. The probability of

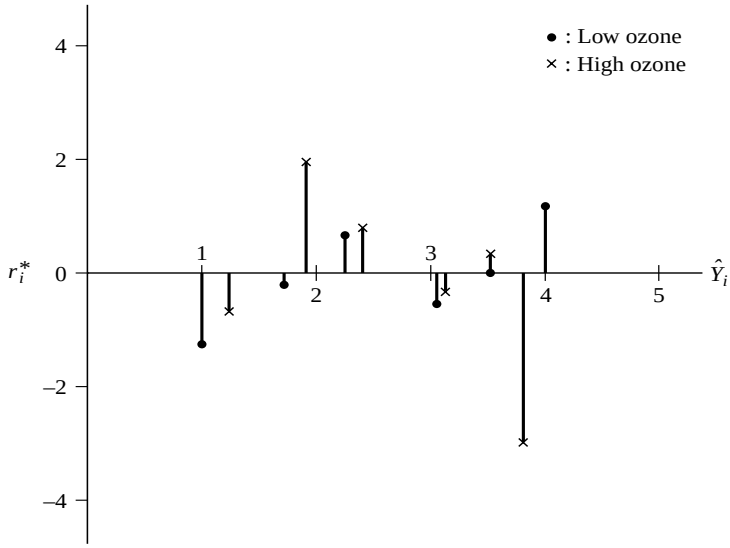

FIGURE 11.5. *Plot of r_i^* versus $\widehat{Y}_i$ for the Lesser–Unsworth data (Exercise 1.19) relating seed weight of soybeans to cumulative solar radiation for two levels of ozone exposure. The model included linear regression of (seed weight)$^{1/2}$ on ozone level and solar radiation.*

$|t| > 2.8369$ is slightly less than .02. Allowing for the fact that this is the most extreme residual out of a sample of 12, this does not appear to be unusually large. Overall, the remaining residuals tend to show an upward trend suggesting that this observation is pulling the regression line down. Inspection of the residuals by treatment, however, shows that the high ozone treatment, the ×s, have a slight downward slope. Perhaps the large residual results from an incorrect model that forces both ozone treatments to have a common regression on solar radiation. ■

Example 11.2

The standardized residuals from the regression of oxygen uptake on time to run a fixed course, resting heart rate, heart rate while running, and maximum heart rate while running, Table 4.3, are plotted against $\widehat{Y}_i$ in Figure 11.6. Although the pattern is not definitive, there is some semblance of the fan-shaped pattern of residuals suggesting heterogeneous variance. The larger dispersion for the higher levels of oxygen consumption could also result from the model being inadequate in this region. Perhaps the faster runners, who tended to use more oxygen, differed in ways not measured by the four variables. ■

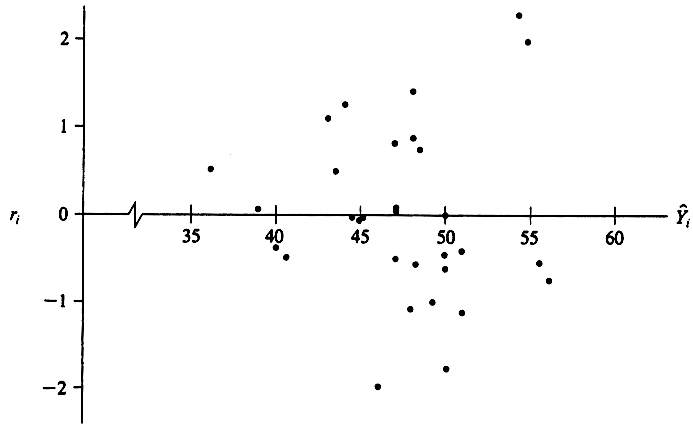

FIGURE 11.6. *Plot of r_i versus $\widehat{Y}_i$ for the regression of oxygen uptake on time, resting heart rate, running heart rate, and maximum heart rate. The original data are given in Table 4.3.*

11.1.2 Plots of $\boldsymbol{e}$ *Versus* $\boldsymbol{X}_i$

Interpretation

Plots of the residuals against the independent variables have interpretations similar to plots against $\widehat{\boldsymbol{Y}}$. Differences in magnitude of dispersion about zero suggest heterogeneous variances. A missing higher-degree polynomial term for the independent variable should be evident in these plots. However, inadequacies in the model associated with one variable, such as a missing higher-degree polynomial term, can be obscured by the effects and distribution of other independent variables. The partial regression leverage plots (discussed in Section 11.1.6) may be more revealing when several independent variables are involved.

Outliers and Influential Points

Outlier residuals will be evident. Observations that appear as isolated points at the extremes of the X_i scale are potentially influential because of their extreme values for that particular independent variable. Such points will tend to have small residuals because of their high leverage. However, data points can be far outside the sample X-space without being outside the limits of any one independent variable by having unlikely combinations of values for two or more variables. Such points are potentially influential but will not be easily detected by any univariate plots.

Example 11.3

(*Continuation of Example* 11.1) The plot of the Studentized residuals against radiation from the regression of seed weight on ozone exposure and cumulative solar radiation (Lesser–Unsworth data) is given in Figure 11.7. [Seed weight is being used as the dependent variable rather than (seed weight)$^{1/2}$ as in Figure 11.5.] One residual (not the same as in Figure 11.5)

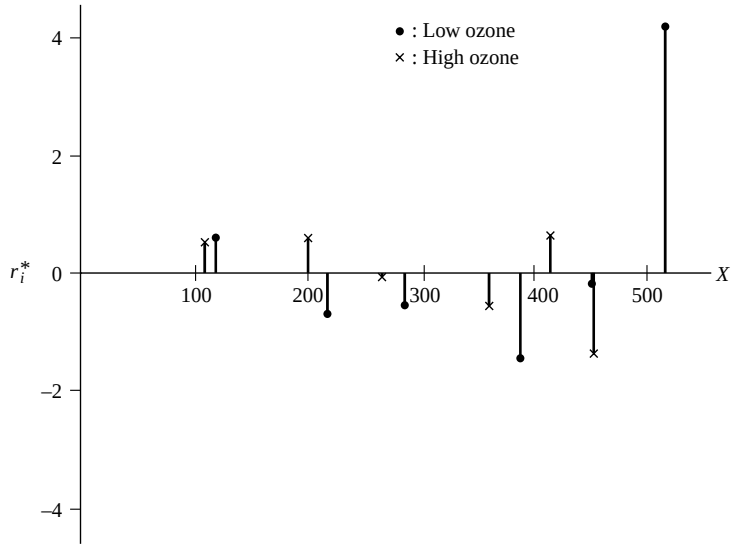

FIGURE 11.7. *Plot of the Studentized residual versus radiation (X) for the Lesser–Unsworth data. The residuals are from the regression of seed weight on ozone level and cumulative solar radiation.*

stands out as a possible outlier. In this case, $r_6^* = 4.1565$ and is very close to being significant, $\alpha^* = .05$. It is evident from the general negative slope of the other residuals that this point has had a major effect on the regression coefficient. ■

11.1.3 Plots of $\boldsymbol{e}$ *Versus Time*

Data collected over time on individual observational units will often have serially correlated residuals. That is, the residual at one point in time depends to some degree on the previous residuals. Classical time series data, such as the data generated by the continuous monitoring of some process, are readily recognized as such and are expected to have correlated residuals. Time series models and analyses take into account these serial correlations and should be used in such cases (Fuller, 1996; Bloomfield, 1976).

Causes of Correlated Residuals

There are many opportunities, however, for time effects to creep into data that normally may not be thought of as time series data. For example, resource limitations may force the researcher to run the experiment over some period of time to obtain even one observation on each treatment. This is common in industrial experiments where an entire production process may be utilized to produce an observation. The time of day or time of

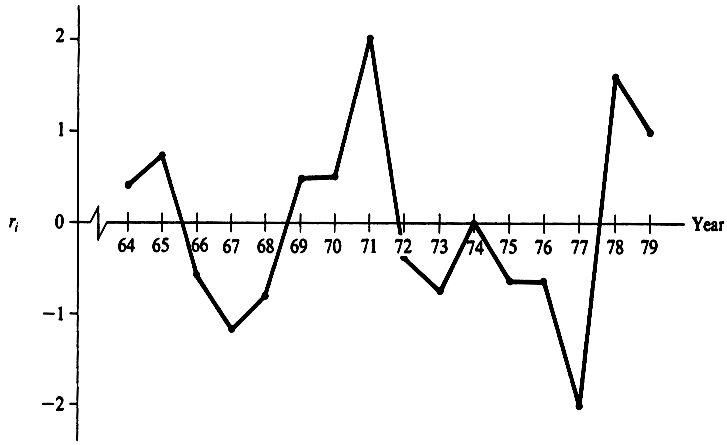

FIGURE 11.8. *Plot of r_i versus year of catch for the regression of yearly Menhaden catch on year. [Data are from Nelson and Ahrenholz (1986).]*

week can have effects on the experimental results even though the process is thought to be well controlled.

Even in biological experiments, where it is usual for all experimental units to be under observation at the same time, some phases of the study may require extended periods of time to complete. For example, autopsies on test animals to determine the incidence of precancerous cell changes may require several days. The simple recording of data in a field experiment may take several days. All such situations provide the opportunity for "time" to have an impact on the differences among the experimental observations.

A plot of the residuals against time may reveal effects not previously thought to be important and, consequently, not taken into account in the design of the study. Serial correlations will appear as a tendency of neighboring residuals to be similar in value.

Example 11.4

The standardized residuals from a regression adjusting yearly Menhaden catch from 1964 to 1979 for a linear time trend are shown in Figure 11.8. [Data are taken from Nelson and Ahrenholz (1986) and are given in Exercise 3.11.] The serial correlation is relatively weak in the case; the lag-one serial correlation is .114. (The lag-one serial correlation is the correlation between residuals one time unit apart.) Even though the serial correlation is weak, the residuals show the typical pattern of the positive and negative residuals occurring in runs. ■

Changes in the production process, drifting of monitoring equipment, time-of-day effects, time-of-week effects, and so forth, will show up as shifts

in the residuals plot. "Time" in this context can be the sequence in which the treatments are imposed, in which measurements are taken, or in which experimental units are tended. Alternatively, "time" could represent the spatial relationship of the experimental units during the course of the trial. In this case, plots of $\boldsymbol{e}$ versus "time" might detect environmental gradients within the space of the experiment.

Runs Test

The **runs test** is frequently used to detect serial correlations. The test consists of counting the number of runs, or sequences of positive and negative residuals, and comparing this result to the expected number of runs under the null hypothesis of independence. (The lack of statistical independence among the observed residuals will confound the runs test to some degree. This effect can probably be ignored as long as a reasonable proportion of the total degrees of freedom are devoted to the residual sum of squares.)

Example 11.5

(Continuation of Example 11.4*)* The data of annual catch of Menhaden for 1964 to 1979 show the following sequence of positive and negative residuals when regressed against time (see Figure 11.8):

$$+ \; + \; - \; - \; - \; + \; + \; + \; - \; - \; - \; - \; - \; - \; + \; + \; .$$

There are $u = 5$ runs in a sample consisting of $n_1 = 7$ positives and $n_2 = 9$ negatives. The cumulative probabilities for number of runs u in sample sizes of (n_1, n_2) are given by Swed and Eisenhart (1943) for $n_1 + n_2 \leq 20$. In this example with $(n_1, n_2) = (7, 9)$, the probability of $u \leq 5$ is .035, indicating significant departure from independence. Appendix Tables A.9 and A.10 give the critical number of runs to attain 5% and 1% significance levels for the runs test for $n_1 + n_2 \leq 20$. These were generated using the Swed and Eisenhart formulae. The critical 5% value for this example is $u \leq 5$. (It would take $u \leq 3$ to be significant at the 1% level.) The low number of runs in this example suggests the presence of a positive serial correlation. ■

If n_1 and n_2 are greater than 10, a normal approximation for the distribution of runs can be used, where

$$\mu = \frac{2n_1 n_2}{n_1 + n_2} + 1 \tag{11.4}$$

and

$$\sigma^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}. \tag{11.5}$$

Then

$$z = \frac{u - \mu + \frac{1}{2}}{\sigma} \tag{11.6}$$

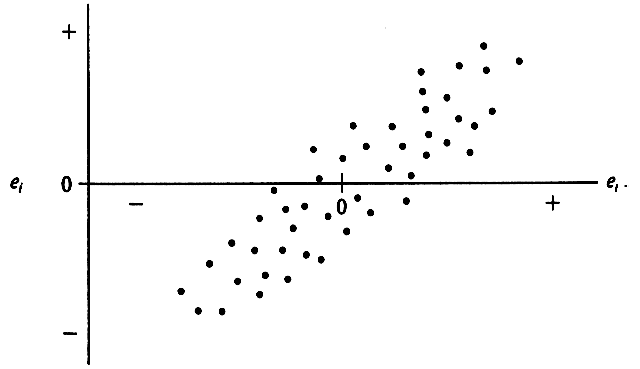

FIGURE 11.9. *Typical plot of e_i versus e_{i-1} showing a positive serial correlation among successive residuals.*

is the standardized normal deviate, where the $\frac{1}{2}$ is the correction for continuity.

Example 11.6

(Continuation of Example 11.5*)* Applying the normal approximation to the Menhaden catch data of Example 10.5, even though n_1 and n_2 are less than 10, gives $\mu = 8.875$ and $\sigma^2 = 3.60944$, which yields $z = -1.776$. The probability of z being less than -1.776 is .0384, very close to the probability of .035 taken from Swed and Eisenhart. ■

11.1.4 Plots of e_i Versus e_{i-1}

A serial correlation in time series data is more clearly revealed with a plot of each residual against the immediately preceding residual. A positive serial correlation would produce a scatter of points with a clear positive slope as in Figure 11.9.

Example 11.7

The plot of r_i versus r_{i-1} for the Menhaden data is shown in Figure 11.10. The extreme point in the upper left-hand quadrant is the plot of the second largest positive residual (1978) against the largest negative residual (1977). This sudden shift in catch from 1977 to 1978 is largely responsible for the serial correlation being as small as it is. Even so, the positive serial correlation is evident. ■

Durbin–Watson Test

The presence of a serial correlation in the residuals is also detected by the Durbin–Watson test for independence (Durbin and Watson, 1951). The

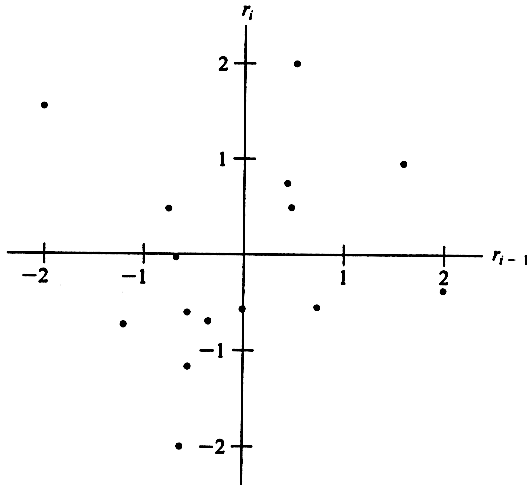

FIGURE 11.10. *Plot of r_i versus r_{i-1} for the Menhaden catch data. The residuals are from the regression of annual catch on year of catch.*

Durbin–Watson test statistic is

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \approx 2(1 - \widehat{\rho}), \tag{11.7}$$

where $\widehat{\rho}$ is the sample correlation between e_i and e_{i-1}. The Durbin–Watson statistic d gets smaller as the serial correlation increases. The one-tailed Durbin–Watson test of the null hypothesis of independence $H_0 : \rho = 0$, against the alternative hypothesis $H_a : \rho > 0$, uses two critical values d_U and d_L which depend on n, p, and the choice of α. Critical values for the Durbin–Watson test statistic are given in Appendix Table A.7. The test procedure rejects the null hypothesis if $d < d_L$, does not reject the null hypothesis if $d > d_U$, and is inconclusive if $d_L < d < d_U$. Tests of significance for the alternative hypothesis $H_a : \rho < 0$ use the same critical values d_U and d_L, but the test statistic is first subtracted from 4.

Some statistical computing packages routinely provide the Durbin–Watson test for serial correlation of the residuals. In PROC GLM (SAS Institute, Inc., 1989b), for example, the Durbin–Watson statistic is reported as part of the standard results whenever the residuals are requested, even though the data may not be time series data. The statistic is computed on the residuals in the order in which the data are listed in the data set. Care must be taken to ensure that the test is appropriate and that the ordering of the data is meaningful before the Durbin–Watson test is used. Also, note that the Durbin–Watson test is computed for the unstandardized residuals e_i.

11.1.5 Normal Probability Plots

The normal probability plot is designed to detect nonnormality. It is the plot of the ordered residuals against the normal order statistics for the appropriate sample size. The normal order statistics are the expected values of ordered observations from the normal distribution with zero mean and unit variance.

Normal Order Statistics

Let $z_1, z_2, \ldots, z_n$ be the observations from a random sample of size n. The n observations ordered (and relabeled) so that $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(n)}$ give the sample order statistics. The average for each $z_{(i)}$ over repeated samplings gives the ith order statistic for the probability distribution being sampled. These are the normal order statistics if the probability distribution being sampled is the normal distribution with zero mean and unit variance. For example, the normal order statistics for a sample of size five are -1.163, $-.495$, .0, .495, and 1.163. The expected value of the smallest observation in a sample of size five from an $N(0,1)$ distribution is -1.163, the second smallest has expectation $-.495$, and so forth.

The normal order statistics were tabled for sample sizes to $n = 204$ by Pearson and Hartley (1966), *Biometrika Tables for Statisticians*, and have been reproduced in many references [e.g., Weisberg (1985), Table D, or Rohlf and Sokal (1981), Table 27]. In some references the indexing of the normal order statistics is in the reverse order so that the first order statistic refers to the largest. The order statistics are easily approximated by any computer program that provides the inverse function of the cumulative normal distribution. Thus, $z_{(i)} \approx \Phi^{-1}(p)$, where p is chosen as a function of the ranks of the residuals. Several choices of p have been suggested. Blom's (1958) suggestion of using

$$p = \frac{R_i - \frac{3}{8}}{n + \frac{1}{4}}, \tag{11.8}$$

where R_i is the rank and n is the sample size, provides an excellent approximation if $n \geq 5$. Plotting the ordered observed residuals against their normal order statistics provides the normal plot.

Expected Behavior

The expected result from a normal plot when the residuals are a sample from a normal distribution is a straight line passing through zero with the slope of the line determined by the standard deviation of the residuals. There will be random deviations from a straight line due to sampling variation of the sample order statistics. Some practice is needed to develop judgment for the amount of departure one should allow before concluding that nonnormality is a problem. Daniel and Wood (1980) give illustrations of the amount of variation in normal probability plots of samples from normal distributions. The normal probability plots for small samples will not be very informative, because of sampling variation, unless departures from normality are large.

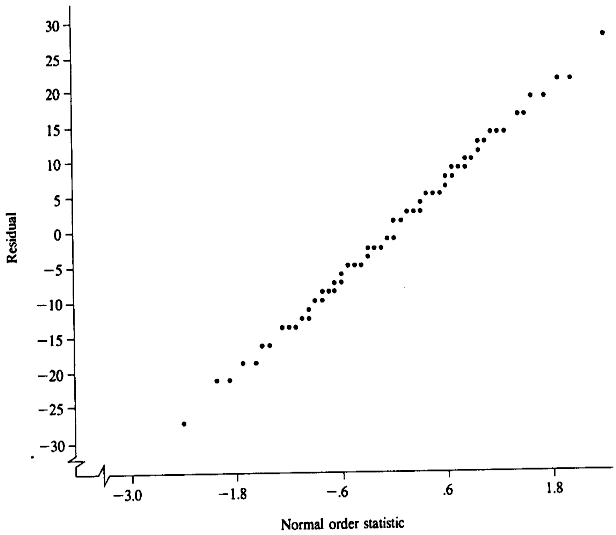

FIGURE 11.11. *Normal plot of residuals from the analysis of variance of final plant heights in a study of blue mold infection on tobacco. (Data courtesy of M. Moss and C. C. Main, North Carolina State University.)*

Example 11.8

Figure 11.11 shows a well-behaved normal plot of the residuals from an analysis of variance of final plant heights in a study of blue mold infection in tobacco. (Data provided courtesy of M. Moss and C. C. Main, North Carolina State University.) There are a total of 80 observations and the residual sum of squares has 36 degrees of freedom. The amount of dependence among the residuals will be related to the proportion of degrees of freedom used by the model, $\frac{44}{80}$ in this case. This relatively high degree of dependence among the residuals and the "supernormal" tendencies of least squares residuals mentioned earlier may be contributing to the very normal-appearing behavior of this plot. ■

Interpretation of Normal Plots

The pattern of the departure from the expected straight line suggests the nature of the nonnormality. A skewed distribution will show a curved normal plot with the direction of the curve determined by the direction of the skewness. An S-shaped curve suggests heavy-tailed or light-tailed distributions (Figure 11.12), depending on the direction of the S. (Heavy-tailed distributions have a relatively higher frequency of extreme observations than the normal distribution; light-tailed distributions have relatively fewer.) Other model defects can mimic the effects of nonnormality. For example, heterogeneous variances or outlier residuals will give the appearance of a heavy-tailed distribution. The ordinary least squares residuals are con-

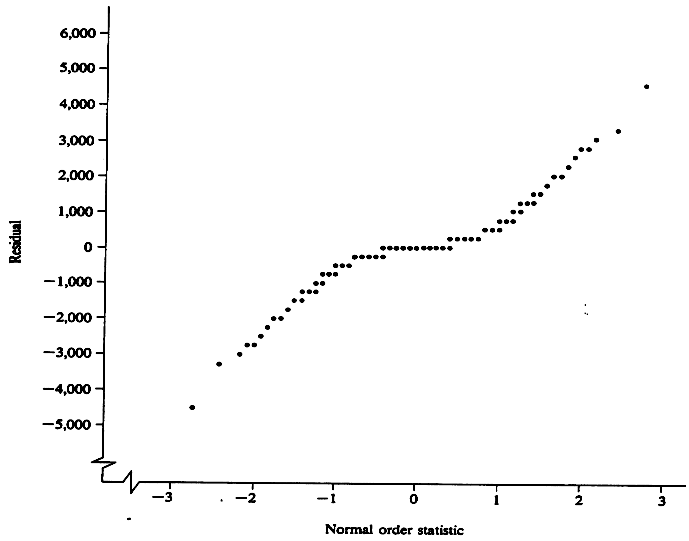

FIGURE 11.12. *A normal probability plot with a pattern typical of a heavy-tailed distribution. In this case, the S-shape resulted from heterogeneous variances in the data.*

strained to have zero mean if the model includes the intercept term, and the plot of the residuals should pass through the origin. (The recursive residuals, on the other hand, are not so constrained and, thus, the normal plot of recursive residuals need not pass through the origin even if the model is correct.) Failure to pass through the origin can be interpreted as an indication of an outlier in the base set of observations or as a model misfit such as an omitted variable (Galpin and Hawkins, 1984).

Tests for Nonnormality

There are many tests for nonnormality under independence. However, these tests must be used with caution when applied to regression residuals, since the residuals are not independent. The limiting distributions of the test statistics show that they are appropriate for regression residuals if the sample size is infinite (Pierce and Kopecky, 1979). For finite samples, however, all are approximations and the question becomes one of how large the sample must be for the approximation to be satisfactory. The required sample size will depend on the number of parameters p', and the nature of $\boldsymbol{P}$, which is determined by the configuration of the Xs (Cook and Weisberg, 1982). Simulation studies have suggested that the approximation is adequate, insofar as size of the test is concerned, for samples as small as $n = 20$ when there are four or six independent variables (White and MacDonald, 1980; Pierce and Gray, 1982), or with $n = 40$ when there are eight independent variables (Pierce and Gray, 1982). However, caution must be used; Weisberg (1980) gives an example using an experimental design ma-

trix with $n = 20$ where the observed size of the test is near $\alpha = .30$, rather than the nominal $\alpha = .10$ level.

Shapiro–Francia Statistic

It appears that many tests for normality applied to regression residuals will provide acceptable approximations if the sample size is reasonable, say $n > 40$ or $n > 80$ if p' is large. The size and power of the tests in small samples make them of questionable value. The Shapiro–Francia (1972) W' test statistic for normality, a modification of the Shapiro–Wilk (1965) W, provides a direct quantitative measure of the degree of agreement between the normal plot and the expected straight line. The Shapiro–Francia statistic is the squared correlation between the observed ordered residuals and the normal order statistics. Let $\boldsymbol{u}$ be the vector of *centered* observed ordered residuals (the e_i, r_i, or r_i^*) and let z be the vector of normal order statistics. Then

$$W' = \frac{(\boldsymbol{u}'\boldsymbol{z})^2}{(\boldsymbol{u}'\boldsymbol{u})(\boldsymbol{z}'\boldsymbol{z})}. \tag{11.9}$$

The observed residuals are expressed as deviations from their mean. The e_i will have zero mean if the model includes an intercept, but this does not apply to r_i or r_i^*. The null hypothesis of normality is rejected for sufficiently small values of W'. Critical values for W' are tabulated by Shapiro and Francia (1972) for $n = 35, 50, 51(2)99$, and are reproduced in Appendix Table A.8. For $n < 50$, the percentage points provided by Shapiro and Wilk (1965) for W are good approximations of those for W' (Weisberg, 1974).

Additional Tests

Other test statistics are frequently used as tests for nonnormality. For example, PROC UNIVARIATE (SAS Institute, Inc., 1990) uses the Shapiro–Wilk W statistic if $n < 2000$ and the Kolomogorov D statistic if $n > 2000$. PROC UNIVARIATE also reports skewness and kurtosis coefficients for the sample; these are sometimes used for testing normality.

11.1.6 Partial Regression Leverage Plots

When several independent variables are involved, the relationship of the residuals to one independent variable can be obscured by effects of other variables. **Partial regression leverage plots** are an attempt to remove the confounding effects of the other variables. Let $\boldsymbol{e}(j)$ denote the residuals from the regression of the dependent variable on all independent variables *except* the jth. Similarly, let $\boldsymbol{u}(j)$ denote the residuals from the regression of the jth independent variable on all other independent variables. The plot of $\boldsymbol{e}(j)$ versus $\boldsymbol{u}(j)$ is the partial regression leverage plot for the jth variable. Note that both $\boldsymbol{e}(j)$ and $\boldsymbol{u}(j)$ have been adjusted for all other independent variables in the model.

Interpretation

This plot reflects what the least squares regression is "seeing" when the jth variable is being added last to the model. The slope of the linear regression line in the partial regression leverage plot is the partial regression

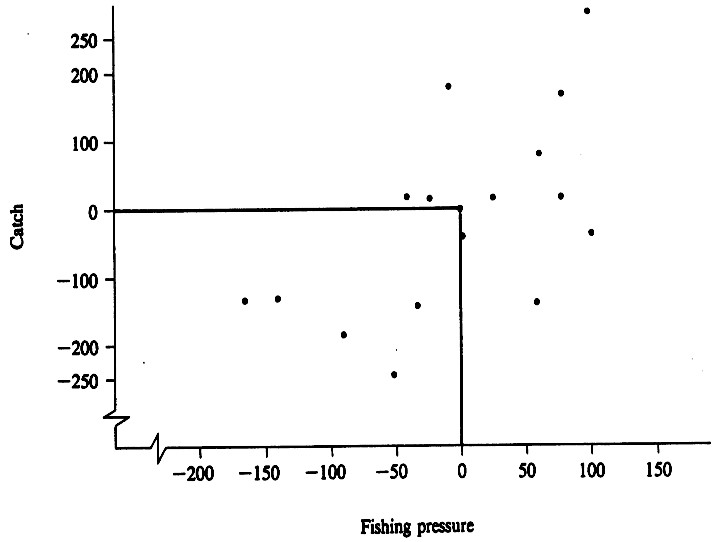

FIGURE 11.13. *Partial regression leverage plot for catch versus fishing pressure from the regression of yearly catch of Menhaden on number of vessels and fishing effort. The model included an intercept. [Data from Nelson and Ahrenholz (1986).]*

coefficient for that independent variable in the full model. The deviations from the linear regression line correspond to the residuals $\boldsymbol{e}$ from the full model.

Any curvilinear relationships not already taken into account in the model should be evident from the partial regression leverage plots. The plot is useful for detecting outliers and high-leverage points and for showing how several leverage points might be interacting to influence the partial regression coefficients.

Example 11.9

The partial regression leverage plot of catch versus fishing pressure for the Menhaden yearly catch data of Example 11.4 is given in Figure 11.13. In this case, yearly catch was regressed on number of vessels and fishing effort; the model also included an intercept. Thus, the partial residuals for catch and fishing pressure are adjusted for the intercept and number of vessels. The figure shows a clear linear relationship between catch and pressure and there may be some suggestion of a slight curvilinear relationship. None of the points appears to be an obvious outlier. The two leftmost points and the uppermost point appear to be influential points in terms of the possible curvilinear relationship. ■

11.2 Influence Statistics

Potentially Influential Points

Potentially influential points or points with high leverage are the data points that are on the fringes of the cloud of sample points in X-space. The ith diagonal element v_{ii} of the projection matrix $\boldsymbol{P}$ (called the Hat matrix in some references) can be related to the distance of the ith data point from the centroid of the X-space. This distance measure takes into account the overall shape of the cloud of sample points. For example, a data point at the side of an elliptical cloud of data points will have a larger value of v_{ii} than another data point falling at a similar distance from the centroid but along the major axis of the elliptical cloud. The ith diagonal element of $\boldsymbol{P}$ is given by

$$v_{ii} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i, \tag{11.10}$$

where $\boldsymbol{x}_i'$ is the ith row of $\boldsymbol{X}$. The limits on v_{ii} are $1/n \leq v_{ii} \leq 1/c$, where c is the number of rows of $\boldsymbol{X}$ that have the same values as the ith row. The lower bound $1/n$ is attained only if every element in $\boldsymbol{x}_i$ is equal to the mean for that independent variable—in other words, only if the data point falls on the centroid. The larger values reflect data points that are farther from the centroid. The upper limit of 1 (when $c = 1$) implies that the leverage for the data point is so high as to force the regression line to pass exactly through that point. The variance of $\widehat{Y}$ for such a point is σ^2 and the variance of the residual is zero. The average value of v_{ii} is p'/n. [There are n v_{ii}-elements and the sum $\text{tr}(\boldsymbol{P})$ is p'.] Belsley, Kuh, and Welsch (1980) suggest using $v_{ii} > 2p'/n$ to identify potentially influential points or leverage points.

Identifying the Influential Points

The diagonal elements of $\boldsymbol{P}$ only identify data points that are far from the centroid of the sample X-space. Such points are potentially but not necessarily influential in determining the results of the regression. The general procedure for assessing the influence of a point in a regression analysis is to determine the changes that occur when that observation is omitted. Several measures of influence have been developed using this concept. They differ in the particular regression result on which the effect of the deletion is measured, and the standardization used to make them comparable over observations. All influence statistics can be computed from the results of the single regression using all data.

Influence Measures

Some influence measures are discussed, each of which measures the effect of deleting the ith observation:

1. Cook's D_i, which measures the effect on $\widehat{\boldsymbol{\beta}}$;
2. DFFITS$_i$, which measures the effect on $\widehat{Y}_i$;
3. DFBETAS$_{j(i)}$, which measures the effect on $\widehat{\beta}_j$; and
4. COVRATIO$_i$, which measures the effect on the variance–covariance matrix of the parameter estimates.

The first three of these, Cook's D, DFFITS, and DFBETAS, can be thought of as special cases of a general approach for measuring the impact of deleting the ith observation on any set of k linearly independent functions of $\widehat{\boldsymbol{\beta}}$ (Cook and Weisberg, 1982). Let $\boldsymbol{U} = \boldsymbol{K}'\boldsymbol{\beta}$ be a set of linear functions of $\boldsymbol{\beta}$ of interest. Then the change in the estimate of $\boldsymbol{U}$ when the ith observation is dropped is given by $\boldsymbol{K}'(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})$, where $\widehat{\boldsymbol{\beta}}_{(i)}$ is the vector of regression coefficients estimated with the ith observation omitted. This change can be written in a quadratic form similar to the quadratic form for the general linear hypothesis in Chapter 4:

$$\frac{[\boldsymbol{K}'(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})]'[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}]^{-1}[\boldsymbol{K}'(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})]}{r(\boldsymbol{K})\widehat{\sigma}^2}. \tag{11.11}$$

If $\boldsymbol{K}'$ is chosen as $\boldsymbol{I}_{p'}$ and s^2 is used for σ^2, Cook's D results. If $\boldsymbol{K}'$ is chosen as $\boldsymbol{x}_i'$, the ith row of $\boldsymbol{X}$, and $s^2_{(i)}$ is used for σ^2, the result is $(\text{DFFITS}_i)^2$. Choosing $\boldsymbol{K}' = (\,0 \quad \ldots \quad 0 \quad 1 \quad 0 \quad \ldots \quad 0\,)$, where the 1 occurs in the $(j+1)$st position, and using $s^2_{(i)}$ for σ^2 gives $(\text{DFBETAS}_{j(i)})^2$.

11.2.1 Cook's D

Computation

Cook's D (Cook, 1977; Cook and Weisberg, 1982) is designed to measure the shift in $\widehat{\boldsymbol{\beta}}$ when a particular observation is omitted. It is a combined measure of the impact of that observation on all regression coefficients. Cook's D is defined as

$$D_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X})(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p's^2}. \tag{11.12}$$

Computationally, D_i is more easily obtained as

$$D_i = \frac{r_i^2}{p'}\left(\frac{v_{ii}}{1 - v_{ii}}\right), \tag{11.13}$$

where r_i is the standardized residual and v_{ii} is the ith diagonal element of $\boldsymbol{P}$ computed from the full regression. Notice that D_i is large if the standardized residual is large and if the data point is far from the centroid of the X-space—that is, if v_{ii} is large.

Interpretation

Cook's D measures the distance from $\widehat{\boldsymbol{\beta}}$ to $\widehat{\boldsymbol{\beta}}_{(i)}$ in terms of the joint confidence ellipsoids about $\widehat{\boldsymbol{\beta}}$. Thus, if D_i is equal to $F_{(\alpha,p',n-p')}$, the $\widehat{\boldsymbol{\beta}}_{(i)}$ vector is on the $100(1-\alpha)\%$ confidence ellipsoid of $\boldsymbol{\beta}$ computed from $\widehat{\boldsymbol{\beta}}$. This should not be treated as a test of significance. A shift in $\widehat{\boldsymbol{\beta}}$ to the ellipsoid corresponding to $\alpha = .50$ from omitting a single data point would be considered a major shift. For reference, the 50th percentile for the F-distribution is 1.0 when the numerator and denominator degrees of freedom are equal and is always less than 1 if the denominator degrees of freedom

is the larger. The 50th percentile does not get smaller than .8 unless the numerator degrees of freedom is only 1 or 2. Thus, Cook's D_i in the vicinity of .8 to 1.0 would indicate a shift to near the 50th percentile in most situations.

Cook's D can also be written in the form

$$D_i = \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})'(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{p's^2}, \tag{11.14}$$

where $\widehat{\boldsymbol{Y}}_{(i)} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{(i)}$. In this form, Cook's D can be interpreted as the Euclidean distance between $\widehat{\boldsymbol{Y}}_{(i)}$ and $\widehat{\boldsymbol{Y}}$ and, hence, measures the shift in $\widehat{\boldsymbol{Y}}$ caused by deleting the ith observation.

11.2.2 DFFITS

Computation

Equation 11.13 showed that Cook's D provides a measure of the shift in $\widehat{\boldsymbol{Y}}$ when the ith observation is not used in the estimation of $\boldsymbol{\beta}$. A closely related measure is provided by DFFITS (Belsley, Kuh, and Welsch, 1980) defined as

$$\begin{aligned} \text{DFFITS}_i &= \frac{\widehat{Y}_i - \widehat{Y}_{i(i)}}{s_{(i)}\sqrt{v_{ii}}} \\ &= \left(\frac{v_{ii}}{1 - v_{ii}}\right) \frac{e_i}{s_{(i)}(1 - v_{(ii)})^{1/2}}, \end{aligned} \tag{11.15}$$

where $\widehat{Y}_{i(i)}$ is the estimated mean for the ith observation but where the ith observation was not used in estimating $\boldsymbol{\beta}$. Notice that σ has been estimated with $s_{(i)}$, the estimate of σ obtained without the ith observation. $s_{(i)}$ is obtained without redoing the regression by using the relationship

$$(n - p' - 1)s^2_{(i)} = (n - p')s^2 - \frac{e_i^2}{1 - v_{ii}}. \tag{11.16}$$

Interpretation

The relationship of DFFITS to Cook's D is

$$D_i = (\text{DFFITS}_i)^2 \left(\frac{s^2_{(i)}}{p's^2}\right). \tag{11.17}$$

Belsley, Kuh, and Welsch (1980) suggest that DFFITS larger in absolute value than $2\sqrt{p'/n}$ be used to flag influential observations. Ignoring the difference between s^2 and $s^2_{(i)}$, this cutoff number for DFFITS suggests a cutoff of $4/n$ for Cook's D.

A modified version of Cook's D suggested by Atkinson (1983) is even more closely related to DFFITS:

$$C_i = |r_i^*| \left[\left(\frac{n - p'}{p'}\right)\left(\frac{v_{ii}}{1 - v_{ii}}\right)\right]^{1/2}$$

$$= \left(\frac{n-p'}{p'}\right)^{1/2} |\text{DFFITS}_i|. \tag{11.18}$$

The cutoff point for DFFITS for flagging large values translates into a cutoff for C_i of $2[(n-p')/n]^{1/2}$. Atkinson recommends that signed values of C_i be plotted in any of the ways customary for residuals. (This recommendation can be extended to any of the measures of influence.) Very nearly identical interpretations are obtained from DFFITS_i, Cook's D_i, and Atkinson's C_i if these reference numbers are used. There is no need to use more than one.

11.2.3 DFBETAS

Cook's D_i reveals the impact of the ith observation on the entire vector of the estimated regression coefficients. The influential observations for the individual regression coefficients are identified by $\text{DFBETAS}_{j(i)}$, $j = 0, 1, 2, \ldots, p$ (Belsley, Kuh, and Welsch, 1980), where each $\text{DFBETAS}_{j(i)}$ is the standardized change in $\widehat{\beta}_j$ when the ith observation is deleted from the analysis. Thus,

$$\text{DFBETAS}_{j(i)} = \frac{\widehat{\beta}_j - \widehat{\beta}_{j(i)}}{s_i\sqrt{c_{jj}}}, \tag{11.19}$$

where c_{jj} is the $(j+1)$st diagonal element from $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Although the formula is not quite as simple as for DFFITS_i, $\text{DFBETAS}_{j(i)}$ can also be computed from the results of the original regression. The reader is referred to Belsley, Kuh, and Welsch (1980) for details.

Interpretation

$\text{DFBETAS}_{j(i)}$ measures the change in $\widehat{\beta}_j$ in multiples of its standard error. Although this looks like a t-statistic, it should not be interpreted as a test of significance. Values of $\text{DFBETAS}_{j(i)}$ greater than 2 would certainly indicate a major, but very unlikely, impact from a single point. The cutoff point of $2/\sqrt{n}$ is suggested by Belsley, Kuh, and Welsch as the point that will tend to highlight the same proportion of influential points across data sets.

11.2.4 COVRATIO

The impact of the ith observation on the variance–covariance matrix of the estimated regression coefficients is measured by the ratio of the determinants of the two variance–covariance matrices. Belsley, Kuh, and Welsch (1980) formulate this as

$$\begin{aligned}\text{COVRATIO} &= \frac{det(s^2_{(i)}[\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)}]^{-1})}{det(s^2[\boldsymbol{X}'\boldsymbol{X}]^{-1})} \\ &= \left[\left(\frac{n-p'-1}{n-p'} + \frac{r_i^{*2}}{n-p'}\right)^p (1-v_{ii})\right]^{-1}. \end{aligned} \tag{11.20}$$

Interpretation

The determinant of a variance–covariance matrix is a generalized measure of variance. Thus, COVRATIO reflects the impact of the ith observation on the precision of the estimates of the regression coefficients. Values near 1 indicate the ith observation has little effect on the precision of the estimates. A COVRATIO greater than 1 indicates that the presence of the ith observation increases the precision of the estimates; a ratio less than 1 indicates that the presence of the observation impairs the precision of the estimates. Belsley, Kuh, and Welsch (1980) suggest that observations with values of COVRATIO outside the limits $1 \pm 3(p'/n)$ be considered influential in the sense of having an inordinate effect on either increasing or decreasing the precision of the estimates.

Using the Influence Statistics

The influence statistics are to be used as diagnostic tools for identifying the observations having the greatest impact on the regression results. Although some of the influence measures resemble test statistics, they are not to be interpreted as tests of significance for influential observations. The large number of influence statistics that can be generated can cause confusion. One should concentrate on that diagnostic tool that measures the impact on the quantity of primary interest. The first two statistics, Cook's D_i and DFFITS_i, are very similar and provide "overall" measures of the influence of each observation. One of these will be of primary interest in most problems. In those cases where interest is in the estimation of particular regression parameters, $\text{DFBETAS}_{j(i)}$ for those j of interest will be most helpful.

Example 11.10

(Continuation of Example 11.3*)* The Studentized residuals and DFFITS_i for the Lesser–Unsworth example are plotted against observation number in Figure 11.14. The close relationship between DFFITS_i and r_i^* is evident; DFFITS_i is the product of r_i^* and $(v_{ii}/(1 - v_{ii}))$ (equation 11.15). The latter is a measure of the potential leverage of the observation which, in this example, varies from .48 for observation 9 to .80 for observation 1. The suggested cutoff value for DFFITS is $2\sqrt{p'/n} = 2\sqrt{3/12} = 1$. Only DFFITS_6 exceeds this value and the residual for this observation certainly appears to be an outlier, $r_6^* = 4.16$. The closely related Cook's D_i are plotted against observation number in Figure 11.15.

The most influential point on $\widehat{\boldsymbol{\beta}}$ is observation 6 with $D_6 = 1.06$. Thus, deleting observation 6 from the analysis causes $\widehat{\boldsymbol{\beta}}$ to shift to beyond the .50 confidence ellipsoid of $\boldsymbol{\beta}$, $F_{(.50;3,9)} = .852$. (In fact, the shift in this case is to the edge of the 58.7% ellipsoid.) The cutoff point translated from DFFITS to Cook's D is $4/n = .33$; only observation 6 exceeds this number.

The impact of each observation on the estimate of β_0 and β_1, where β_1 is the regression of seed weight on total solar radiation, is shown in the plots of DFBETAS_0 and DFBETAS_1 in Figure 11.16. The suggested cutoff point for DFBETAS_j is $2/\sqrt{n} = .58$ in this example. None of the observations exceed this cutoff point for DFBETAS_0 and only observation 6 exceeds

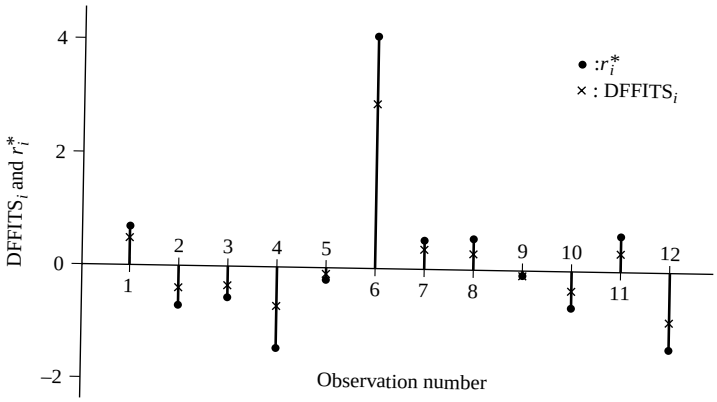

FIGURE 11.14. *Studentized residuals and DFFITS$_i$ plotted against observation number from the regression of seed weight on ozone level and cumulative solar radiation using the Lesser–Unsworth data.*

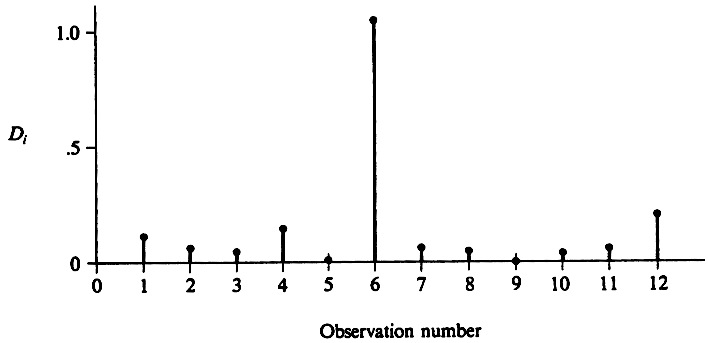

FIGURE 11.15. *Cook's D_i plotted against observation number from the regression of seed weight on ozone level and cumulative solar radiation using the Lesser–Unsworth data.*

this cutoff for DFBETAS_1. This illustrates a case where an observation has major impact on the regression [D_6, DFFITS_6, and $\text{DFBETAS}_{1(6)}$ are large] but has very little effect on the estimation of one parameter, in this case β_0.

The suggested cutoff values for COVRATIO_i in the Lesser–Unsworth example are $1 \pm 3p'/n = (0.25, 1.75)$. Observations 1, 5, and 7 exceed the upper cutoff point (values not shown) indicating that the presence of these three observations has the greatest impact on *increasing* the precision of the parameter estimates. $\text{COVRATIO}_6 = .07$ is the only one that falls below the lower limit. This indicates that the presence of observation 6 greatly *decreases* the precision of the estimates; the large residual from this observation will cause s^2 to be much larger than $s^2_{(6)}$.

The influence diagnostics on the Lesser–Unsworth example flag observation 6 as a serious problem in this analysis. This can be due to observation 6 being in error in some sense or the model not adequately representing the relationship between seed weight, solar radiation, and ozone exposure. The seed weight and radiation values for observation 6 were both the largest in the sample. There is no obvious error in either. The most logical explanation of the impact of this observation is that the linear model does not adequately represent the relationship for these extreme values. ■

11.2.5 Summary of Influence Measures

The following summarizes the influence measures.

Influence Measure	*Formula*	*Observation i May Be Influential If:*
Cook's D_i	$\frac{(\widehat{\boldsymbol{\beta}}_{(i)}-\widehat{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X})(\widehat{\boldsymbol{\beta}}_{(i)}-\widehat{\boldsymbol{\beta}})}{p's^2}$	$D_i > F_{(.5,p',n-p')}$
DFFITS_i	$\frac{\widehat{Y}_i-\widehat{Y}_{i(i)}}{s_{(i)}\sqrt{v_{ii}}}$	$\lvert\text{DFFITS}_i\rvert > 2\sqrt{p'/n}$
Atkinson's C_i	$\left(\frac{n-p'}{p'}\right)^{1/2}\lvert\text{DFFITS}_i\rvert$	$\lvert C_i\rvert > 2[(n-p')/n]^{1/2}$
$\text{DFBETAS}_{j(i)}$	$\frac{\widehat{\beta}_j-\widehat{\beta}_{j(i)}}{s_i\sqrt{c_{jj}}}$	$\lvert\text{DFBETAS}_{j(i)}\rvert > 2/\sqrt{n}$
COVRATIO_i	$\frac{\det(s^2_{(i)}[\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)}]^{-1})}{\det(s^2[\boldsymbol{X}'\boldsymbol{X}]^{-1})}$	$\text{COVRATIO}\begin{cases} < 1-3p'/n \\ > 1+3p'/n \end{cases}$

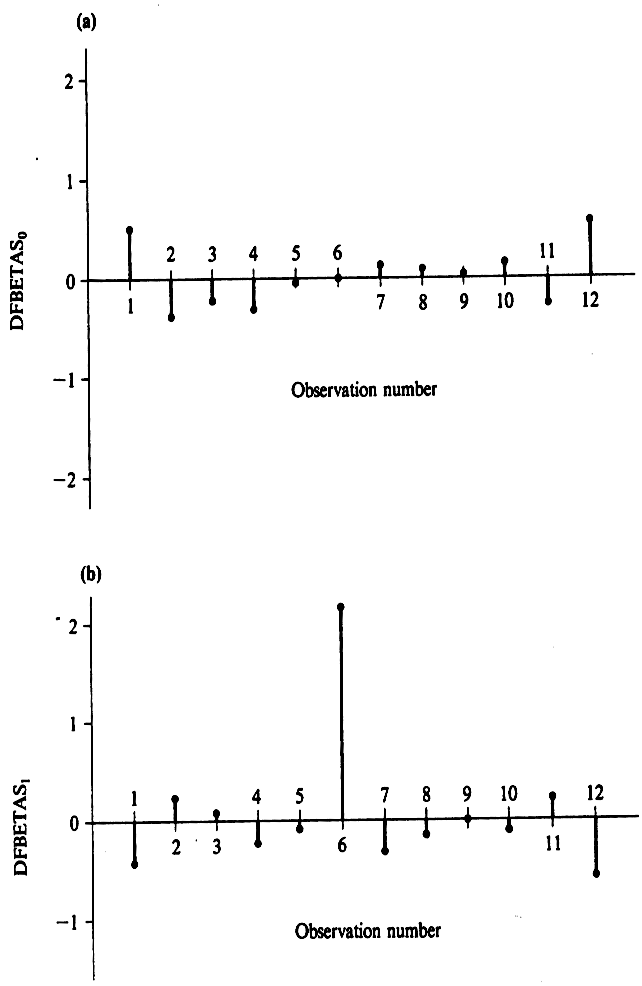

FIGURE 11.16. *$DFBETAS_{0(i)}$ and $DFBETAS_{1(i)}$ plotted against observation number from the regression of seed weight on ozone level and cumulative solar radiation using the Lesser–Unsworth data.*

11.3 Collinearity Diagnostics

Effects of Collinearity

The collinearity problem in regression refers to the set of problems created when there are near-singularities among the columns of the $\boldsymbol{X}$ matrix; certain linear combinations of the columns of $\boldsymbol{X}$ are nearly zero. This implies that there are (near) redundancies among the independent variables; essentially the same information is being provided in more than one way. Geometrically, collinearity results when at least one dimension of the X-space is very poorly defined in the sense that there is almost no dispersion among the data points in that dimension.

Limited dispersion in an independent variable results in a very poor (high variance) estimate of the regression coefficient for that variable. This can be viewed as a result of the near-collinearity between the variable and the column of ones (for the intercept) in $\boldsymbol{X}$. (A variable that has very little dispersion relative to its mean is very nearly a multiple of the vector of ones.) This is an example of collinearity that is easy to detect by simple inspection of the amount of dispersion in the individual independent variables. The more usual, and more difficult to detect, collinearity problem arises when the near-singularity involves several independent variables. The dimension of the X-space in which there is very little dispersion is some linear combination of the independent variables, and may not be detectable from inspection of the dispersion of the individual independent variables.

The result of collinearity involving several variables is high variance for the regression coefficients of *all* variables involved in the near-singularity. In addition, and perhaps more importantly, it becomes virtually impossible to separate the influences of the independent variables and very easy to pick points for prediction that are (unknowingly) outside the sample X-space, representing extrapolations.

Detecting Collinearity with Eigenanalysis

The presence of collinearity is detected with the singular value decomposition of $\boldsymbol{X}$ or the eigenanalysis of $\boldsymbol{X}'\boldsymbol{X}$ (Sections 2.7 and 2.8). The eigenvalues λ_i provide measures of the amount of dispersion in the dimensions corresponding to the principal component axes of the X-space. The elements in the eigenvectors are the coefficients (for the independent variables) defining the principal component axes. All principal components are pairwise orthogonal.

The first principal component axis is defined so as to identify the direction through the X-space that has the maximum dispersion. The second principal component axis identifies the dimension orthogonal to the first that has the second most variation and so forth until the last principal component axis identifies the dimension with the least dispersion. The relative sizes of the eigenvalues reveal the relative amounts of dispersion in the different dimensions of the X-space, and the eigenvectors identify the linear combinations of the independent variables that define those dimensions. The smaller eigenvalues, and their eigenvectors, are of particular interest for the collinearity diagnostics.

Standardizing $\boldsymbol{X}$

The eigenanalysis for purposes of detecting collinearity typically is done on $\boldsymbol{X}'\boldsymbol{X}$ after $\boldsymbol{X}$ has been scaled so that the length of each vector, the sum of squares of each column, is one. Thus, $\text{tr}(\boldsymbol{X}'\boldsymbol{X}) = p'$. This standardization is necessary to prevent the eigenanalysis from being dominated by one or two of the independent variables. The sum of the eigenvalues equals the trace of the matrix being analyzed, $\sum \lambda_k = \text{tr}(\boldsymbol{X}'\boldsymbol{X})$, which is the sum of the sum of squares of the independent variables including X_0. The independent variables in their original units of measure would contribute unequally to this total sum of squares and, hence, to the eigenvalues. A simple change of scale of a variable, such as from inches to centimeters, would change the contribution of the variable to the principal components if the vectors were not rescaled to have equal length.

The standardization of $\boldsymbol{X}$ is accomplished by dividing the elements of each column vector by the square root of the sum of squares of the elements. In matrix form, define a diagonal $(p' \times p')$ matrix $\boldsymbol{D}$, which consists of square roots of the diagonal elements of $\boldsymbol{X}'\boldsymbol{X}$. The standardized $\boldsymbol{X}$ matrix $\boldsymbol{Z}$ is given by

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{D}^{-1}. \tag{11.21}$$

The eigenanalysis is done on $\boldsymbol{Z}'\boldsymbol{Z}$.

Centering the Independent Variables

Some authors argue that the independent variables should first be centered by subtracting the mean of each independent variable. This centering makes all independent variables orthogonal to the intercept column and, hence, removes any collinearity that involves the intercept. Marquardt (1980) calls this the "nonessential collinearity." Any independent variable that has a very small coefficient of variation, small dispersion relative to its mean, will be highly collinear with the intercept and yet, when centered, be orthogonal to the intercept. Belsley, Kuh, and Welsch (1980) and Belsley (1984) argue that this correction for the mean is part of the multiple regression arithmetic and should be taken into account when assessing the collinearity problem. For further discussion on this topic, the reader is referred to Belsley (1984) and the discussions following his article by Cook (1984), Gunst (1984), Snee and Marquardt (1984), and Wood (1984).

"Nonessential" Collinearity

The seriousness of collinearity and whether it is "nonessential" collinearity depends on the specific objectives of the regression. Even under severe collinearity, certain linear functions of the parameters may be estimated with adequate precision. For example, the estimate of the *change* in Y between two points in X may be very precisely estimated even though the estimates of some of the parameters are highly variable. If these linear functions also happen to be the quantities of primary interest, the collinearity might be termed "nonessential." However, *any collinearity, including collinearity with the intercept, that destroys the stability of the quantities of interest cannot be so termed.*

This discussion of collinearity diagnostics assumes that the noncentered independent variables (scaled to have unit vector length) are being used.

The diagnostics from the centered data can be used when they are more relevant for the problem. In any specific case, it is best to look at the seriousness of the collinearity in terms of the objectives of the study.

11.3.1 Condition Number and Condition Index

Condition Number

The condition number$K(\boldsymbol{X})$ of a matrix $\boldsymbol{X}$ is defined as the ratio of the largest singular value to the smallest singular value (Belsley, Kuh, and Welsch, 1980),

$$K(\boldsymbol{X}) = \left[\frac{\lambda_{max}}{\lambda_{min}}\right]^{1/2}. \tag{11.22}$$

The condition number provides a measure of the sensitivity of the solution to the normal equations to small changes in X or Y. A large condition number indicates that a near-singularity is causing the matrix to be poorly conditioned. For reference, the condition number of a matrix is 1 when all the columns are pairwise orthogonal and scaled to have unit length; all λ_k are equal to 1.

Condition Index

The condition number concept is extended to provide the **condition index** for each (principal component) dimension of the X-space. The condition index δ_k for the kth principal component dimension of the X-space is

$$\delta_k = \left[\frac{\lambda_{max}}{\lambda_k}\right]^{1/2}. \tag{11.23}$$

The largest condition index is also the condition number $K(\boldsymbol{X})$ of the matrix. Thus, condition indices identify the dimensions of the X-space where dispersion is limited enough to cause problems with the least squares solution.

Interpretation

Belsley, Kuh, and Welsch (1980) suggest that condition indices around 10 indicate weak dependencies that may be starting to affect the regression estimates. Condition indices of 30 to 100 indicate moderate to strong dependencies and indices larger than 100 indicate serious collinearity problems. The number of condition indices in the critical range indicates the number of near-dependencies contributing to the collinearity problem.

Another measure of collinearity involves the ratios of the squares of the eigenvalues. Thisted (1980) suggested

$$mci = \sum_{j=1}^{p'} \left(\frac{\lambda_{p'}}{\lambda_j}\right)^2 \tag{11.24}$$

as a multicollinearity index, where $\lambda_{p'}$ is the smallest eigenvalue of $\boldsymbol{X}'\boldsymbol{X}$. Values of *mci* near 1.0 indicate high collinearity; values greater than 2.0 indicate little or no collinearity.

TABLE 11.1. *The singular values and the condition indices for the numerical example.*

Principal Component	*Singular Values*	*Condition Index*
1	1.7024	1.00
2	1.0033	1.70
3	0.3083	5.52
4	0.0062	273.60

Example 11.11

A small numerical example is used to illustrate the measures of collinearity. An $\boldsymbol{X}$ matrix 20×4 consists of the intercept column, and three independent variables constructed in the following way.

X_1 is the sequence of numbers 20 to 29 and repeated.

X_2 is X_1 minus 25 with the first and eleventh observations changed to -4 (from -5) to avoid a complete linear dependency.

X_3 is a periodic sequence running 5, 4, 3, 2, 1, 2, 3, 4, 5, 6, and repeated. X_3 is designed to be nearly orthogonal to the variation in X_1 and X_2.

The singular values and the condition indices using the noncentered, unit-length vectors for this $\boldsymbol{X}$ matrix are given in Table 11.1. The largest condition index $\delta_4 = 1.702410/0.006223 = 273.6$ indicates a severe collinearity problem. This is the condition number $K(\boldsymbol{X})$ of $\boldsymbol{X}$ as scaled. The condition indices for the other dimensions do not indicate any collinearity problem; they are well below the value of 10 suggested by Belsley, Kuh, and Welsch as the point at which collinearity may be severe enough to begin having an effect. The multicollinearity index of Thisted, *mci*, is very close to 1 (*mci* = 1.061),which indicates severe collinearity. ■

11.3.2 Variance Inflation Factor

Definition

Another common measure of collinearity is the **variance inflation factor** for the jth regression coefficient VIF_j. The variance inflation factors are computed from the correlation matrix $\widehat{\boldsymbol{\rho}}$ of the independent variables. Thus, the independent variables are centered and standardized to unit length. The diagonal elements $\widehat{\boldsymbol{\rho}}^{-1}$, the inverse of $\widehat{\boldsymbol{\rho}}$, are the variance inflation factors. The link between VIF_j and collinearity (of the standardized and centered variables) is through the relationship

$$VIF_j = \frac{1}{1 - R_j^2}, \tag{11.25}$$

where R_j^2 is the coefficient of determination from the regression of X_j on the other independent variables. If there is a near-singularity involving X_j and the other independent variables, R_j^2 will be near 1.0 and VIF_j will be large. If X_j is orthogonal to the other independent variables, R_j^2 will be 0 and VIF_j will be 1.0.

The term *variance inflation factor* comes from the fact that the variance of the jth regression coefficient can be shown to be directly proportional to VIF_j (Theil, 1971; Berk, 1977):

$$s^2(\widehat{\beta}_j) = \frac{\sigma^2}{\boldsymbol{x}_j'\boldsymbol{x}_j}(VIF_j), \tag{11.26}$$

where $\boldsymbol{x}_j$ is the jth *column* of the centered $\boldsymbol{X}$ matrix.

Interpretation

The variance inflation factors are simple diagnostics for detecting overall collinearity problems that do not involve the intercept. They will not detect multiple near-singularities nor identify the source of the singularities. The maximum variance inflation factor has been shown to be a lower bound on the condition number (Berk, 1977). Snee and Marquardt (1984) suggest that there is no practical difference between Marquardt's (1970) guideline for serious collinearity $VIF > 10$, and Belsley, Kuh, and Welsch's (1980) condition number of 30.

Example 11.12

The variance inflation factors computed from the correlation matrix of the independent variables for Example 11.11 are

$$\begin{aligned} VIF_1 &= 169.4, \\ VIF_2 &= 175.7, \text{ and} \\ VIF_3 &= 1.7. \end{aligned}$$

The variance inflation factors indicate that the estimates of β_1 and β_2 would be seriously affected by the very near-singularity in $\boldsymbol{X}$. In this case, the near-singularity is known to be due to the near-redundancy between X_1 and X_2. Notice that the variance inflation factor of $\widehat{\beta}_3$ is near 1, the expected result if all variables are orthogonal. The variance inflation factors are computed on the centered and scaled data, and as a result are orthogonal to the intercept column. Thus, the variance inflation factors indicate a collinearity problem in this example that does not involve the intercept. ■

11.3.3 Variance Decomposition Proportions

The variance of each estimated regression coefficient can be expressed as a function of the eigenvalues λ_k of $\boldsymbol{X}'\boldsymbol{X}$ and the elements of the eigenvectors.

Let u_{jk} be the jth element of the kth eigenvector. Then,

$$\text{Var}(\widehat{\beta}_j) = \sigma^2 \sum_k \left(\frac{u_{jk}^2}{\lambda_k} \right). \tag{11.27}$$

The summation is over the $k = 1, \ldots, p'$ principal component dimensions. Thus, the variance of each regression coefficient can be decomposed into the contributions from each of the principal components. The size of each contribution (for the variance of the jth regression coefficient) is determined by the square of the ratio of the jth element from the kth eigenvector u_{jk} to the singular value $\lambda_k^{1/2}$.

The major contributions to the variance of a regression coefficient occur when the coefficient in the eigenvector is large in absolute value, and the eigenvalue is small. A large coefficient u_{jk} indicates that the jth independent variable is a major contributor to the kth principal component. The small eigenvalues identify the near-singularities that are the source of the instability in the least squares estimates. Not all regression coefficients need be affected. If the jth variable is not significantly involved in the near-singularity, its coefficient in the kth eigenvector u_{jk} will be near zero and its regression coefficient will remain stable even in the presence of the collinearity.

Variance Decomposition Proportions

It is helpful to express each of the contributions as a proportion of the total variance for that particular regression coefficient. These partitions

$$\left(\frac{u_{jk}^2/\lambda_k}{\sum_i \left(u_{ji}^2/\lambda_i\right)} \right)$$

of the variances are called the **variance decomposition proportions**.

Example 11.13

The variance decomposition proportions for the data in Example 11.11 are given in Table 11.2. The entries in any one column show the proportion of the variance for that regression coefficient that comes from the principal component indicated on the left. For example, 34% of the variance of $\widehat{\beta}_3$ comes from the fourth principal component, 65% from the third, and only slightly over 1% from the first and second principal components. ■

Interpretation

The critical information in Table 11.2 is how the variances are being affected by the last principal component, the one with the least dispersion and the greatest impact on the collinearity problem. For reference, if the columns of $\boldsymbol{X}$ were orthogonal, the variance decomposition proportions would be all 0 except for a single 1 in each row and column. That is, each principal component would contribute to the variance of only one regression coefficient. Serious collinearity problems are indicated when a

TABLE 11.2. *Variance decomposition proportions for Example 11.11 using all principal components (upper half of table) and with the fourth principal component deleted (lower half).*

Principal	*Variance Proportion*			
Component	*Intercept*	X_1	X_2	X_3
1	.0000[a]	.0000[a]	.0000[a]	.0102
2	.0000[a]	.0000[a]	.0055	.0008
3	.0001	.0001	.0003	.6492
4	.9999	.9998	.9942	.3398
	1.0000	1.0000	1.0000	1.0000
1	.070	.060	.001	.015
2	.002	.002	.942	.001
3	.928	.939	.057	.983
	1.000	1.000	1.000	1.000

[a]Variance proportions are less than 10^{-4}.

principal component with a small eigenvalue contributes heavily—more than 50%—to two or more regression coefficients.

Example 11.14

(*Continuation of Example* 11.13) The fourth principal component is responsible for over 99% of $s^2(\widehat{\beta}_0)$, $s^2(\widehat{\beta}_1)$, and $s^2(\widehat{\beta}_2)$. The fourth principal component had a condition index of $\delta_4 = 274$, well above the critical point for severe collinearity. The fourth principal component identifies a nearly singular dimension of the X-space that is causing severe variance inflation of these three regression coefficients. Notice, however, that the variance of $\widehat{\beta}_3$ is not seriously affected by this near-singularity. This implies that X_3 is not a major component of the near-singularity defined by the fourth principal component. ■

The interpretation of the variance decomposition proportions required these conditions for the result to be an indication of serious collinearity:

1. the condition index for the principal component must be "large"; and
2. the variance decomposition proportions must show that the principal component is a major contributor ($> 50\%$) to at least two regression coefficients.

Multiple Near-Singularities

More than one near-singularity may be causing variance inflation problems. In such a situation, the variance decomposition table will be dominated by the principal component with the smallest eigenvalue so that the effect of other near-singularities may not be apparent. The variance contributions of the next principal component are found by rescaling each column

so that the proportions add to 100% *without* the last principal component. This approximates what would happen to the variance proportions if the fourth principal component were "removed."

Example 11.15

Since the condition index ($\delta_k = 5.5$) for the third principal component from Example 11.14 is not in the critical range, the analysis of the variance proportions normally would not proceed any further. However, to illustrate the process, we give in the lower portion of Table 11.2 the variance decomposition proportions for the example without the fourth principal component. If the condition index for the third principal component had been sufficiently high, this result would be suggesting that this dimension also was causing variance inflation problems. ■

Linear Functions

The variance decomposition proportions provide useful information when the primary interest is in the regression coefficients per se. When the primary objective of the regression analysis is the use of the estimated regression coefficients in some linear function, such as in a prediction equation, it is more relevant to measure the contributions of the principal components to the variance of the linear function of interest. Let $c = \boldsymbol{K}'\widehat{\boldsymbol{\beta}}$ be the linear function of interest. The variance of c is

$$\sigma^2(c) = \boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}\sigma^2, \tag{11.28}$$

which can be decomposed into the contributions from each of the principal components as

$$\sigma^2(c) = \sum_k \left(\frac{(\boldsymbol{K}'\boldsymbol{u}_k)^2}{\lambda_k} \right) \sigma^2. \tag{11.29}$$

Each term reflects the contribution to the variance of the corresponding principal component.

Example 11.16

(*Continuation of Example* 11.15) Suppose the linear funtion of interest is $c = \boldsymbol{K}'\widehat{\boldsymbol{\beta}}$, where

$$\boldsymbol{K}' = (\,1 \quad 25 \quad 0 \quad 3\,).$$

Then the variance of $c = \boldsymbol{K}'\widehat{\boldsymbol{\beta}}$ is $\sigma^2(c) = 0.0597\sigma^2$. The partitions of this variance into the contributions from the four principal components and the variance proportions are given in Table 11.3. For this (deliberately chosen) linear function, the fourth principal component, which was causing the collinearity problem and the severe variance inflation of the regression coefficients, is having almost no impact. Thus, if this particular linear function were the primary objective of the analysis, the near-singularity identified by the fourth principal component could be termed a "nonessential collinearity." This can be viewed as a generalization of the concept of "nonessential

TABLE 11.3. *The variance partitions and the variance proportions for the linear function* $\boldsymbol{K}'\widehat{\boldsymbol{\beta}}$ *where* $\boldsymbol{K}' = (1\ 25\ 0\ 3)$.

Principal Component		*Variance Partition*	*Variance Proportion*
1		.0451	.7542
2		.0003	.0050
3		.0142	.2375
4		.0002	.0033
	Total	.0597	1.0000

ill-conditioning" used by Marquardt (1980) to refer to near-singularities involving the intercept. ■

11.3.4 Summary of Collinearity Diagnostics

Collinearity Diagnostic	*Formula*	*Collinear if*
Condition Index, δ_k	$\left[\frac{\lambda_{max}}{\lambda_k}\right]^{1/2}$	$\begin{cases} 30 \leq \delta_k \leq 100 \text{ (moderate)} \\ \delta_k > 100 \text{ (strong)} \end{cases}$
mci	$\sum_{j=1}^{p'}\left[\frac{\lambda_{p'}}{\lambda_j}\right]^2$	$\begin{cases} mci \leq 2 \\ mci \approx 1 \text{ (strong)} \end{cases}$
Variance Inflation Factor, VIF	$\frac{1}{1-R_j^2}$	$VIF > 10$

11.4 Regression Diagnostics on the Linthurst Data

Example 11.17

The Linthurst data were used in Chapter 5 to illustrate the choice of variables in a model-building process. In that exercise, the modeling started with five independent variables, $SALINITY$, pH, K, Na, and Zn, and ended with a model that contained two variables. The usual assumptions of ordinary least squares were made and all of the variables were assumed to be related linearly to the dependent variable *BIOMASS*. In this section, the regression diagnostics are presented for the Linthurst data for the five-variable regression model.

The residuals e_i standardized residuals r_i (called STUDENT residual in PROC REG), and Cook's D for the regression of *BIOMASS* on the five independent variables were obtained from the RESIDUAL option in

PROC REG (SAS Institute Inc., 1989b) and are given in Table 11.4. The Studentized residuals r_i^* (called RSTUDENT in PROC REG), and several influence statistics were obtained from the INFLUENCE option and are given in Tables 11.5 and 11.6 (pages 380 and 381).

Residuals

The standardized residual for observation 34 is the largest with a value of $r_{34} = 2.834$; this residual is 2.834 standard deviations away from zero. When expressed as the Studentized residual, its value is $r_{34}^* = 3.14$. Four other Studentized residuals are greater than 2.0 in absolute value. This frequency of large residuals (11%) is higher than might be expected from a sample size of 45. An approximate chi-square test, however, does not show a significant departure from an expected 5% frequency of residuals greater than 2.0 in absolute value. (This test has an additional approximation compared to the conventional goodness-of-fit test because the residuals are not independent.)

These large residuals must not be interpreted, however, as indicating that these points are in error or that they do not belong to the population sampled. Of course, the data should be carefully checked to verify that there are no errors and that the points represent legitimate observations. But as a general rule, outlier points should not be dropped from the data set unless they are found to be in error and the error cannot be corrected. An excessively high frequency of large residuals on a carefully edited data set is probably an indication of an inadequate model. The model and the system being modeled should be studied carefully. Perhaps an important independent variable has been overlooked or the relationships are not linear as has been assumed.

11.4.1 Plots of Residuals

e Versus $\widehat{Y}$

The plot of the ordinary least squares residuals against the predicted values, Figure 11.17(a), shows the presence of five predicted values that are greater than 2,000, much larger than any of the others. Four of the five residuals associated with these points are not particularly notable, but the fifth point is the largest negative residual, −748 or a standardized residual of $r_{29} = 2.0804$. A second point of interest in Figure 11.17(a) is the apparently greater spread among the positive residuals than among the negative residuals. This suggests that the distribution of the residuals might be skewed. The skewness is seen more clearly in a frequency polygon of the residuals, Figure 11.18 (page 383). There are four residuals greater than 2.0 but only one less than −2.0 and there is a high frequency of relatively small negative residuals.

Normal Probability Plot

The normal probability plot of the standardized residuals, Figure 11.19, shows a distinct curvature rather than the straight line expected of normally distributed data. The shape of this normal plot, except for the additional bend caused by the four most negative residuals, is consistent with the positively skewed distribution suggested by the frequency polygon. The

TABLE 11.4. *Residuals analysis from the regression of BIOMASS on the five independent variables SAL, pH, K, Na, and Zn (from SAS PROC REG, option R). An * on the measure of influence indicates that the value exceeds the reference value.*

Obs.	Y_i	$\widehat{Y}_i$	$s(\widehat{Y}_i)$	e_i	$s(e_i)$	r_i	*Cook's D*
1	676	724	176	−48	357	−.135	.001
2	516	740	142	−224	372	−.601	.009
3	1,052	691	127	361	378	.956	.017
4	868	815	114	53	382	.140	.000
5	1,008	1,063	321	−56	235	−.236	.017
6	436	958	126	−522	378	−1.381	.035
7	544	527	214	17	336	.050	.000
8	680	827	141	−147	373	−.394	.004
9	640	676	174	−36	358	−.101	.000
10	492	911	165	−419	362	−1.155	.046
11	984	1,166	167	−182	362	−.503	.009
12	1,400	573	147	827	370	2.232	.130*
13	1,276	816	153	460	368	1.252	.045
14	1,736	953	137	783	374	2.093	.099*
15	1,004	898	166	106	362	.293	.003
16	396	355	135	41	375	.109	.000
17	352	577	127	−225	377	−.595	.007
18	328	586	139	−258	373	−.691	.011
19	392	586	118	−194	380	−.511	.004
20	236	494	131	−258	376	−.687	.010
21	392	596	122	−204	379	−.537	.005
22	268	570	120	−302	380	−.795	.010
23	252	584	124	−332	378	−.877	.014
24	236	479	100	−243	386	−.631	.004
25	340	425	131	−85	376	−.226	.001
26	2,436	2,296	170	140	360	.388	.006
27	2,216	2,202	196	14	347	.040	.000
28	2,096	2,230	187	−134	351	−.381	.007
29	1,660	2,408	171	−748	360	−2.080	.163*
30	2,272	2,369	168	−97	361	−.270	.003
31	824	1,110	115	−286	381	−.750	.008
32	1,196	982	118	214	381	.562	.005
33	1,960	1,155	120	805	380	2.120	.075
34	2,080	1,008	124	1072	378	2.834	.145*
35	1,764	1,254	136	510	374	1.363	.041
36	412	959	111	−547	383	−1.431	.029
37	416	626	133	−210	376	−.558	.006
38	504	624	107	−120	384	−.313	.001
39	492	588	99	−96	386	−.250	.001
40	636	837	95	−201	387	−.521	.003
41	1,756	1,526	129	230	377	.610	.007
42	1,232	1,298	97	−66	386	−.171	.000
43	1,400	1,401	106	−1	384	−.004	.000
44	1,620	1,306	113	314	382	.822	.010
45	1,560	1,265	90	295	388	.759	.005

TABLE 11.5. *Residuals and influence statistics from the regression of BIOMASS on the five independent variables SAL, pH, K, Na, and Zn (from SAS's PROC REG, option INFLUENCE). An * on the measure of influence indicates that the value exceeds the reference value.*

Obs.	e_i	r_i^*	v_{ii}	*COV-RATIO*	*DF-FITS*
1	−48	−.133	.195	1.447∗	−.065
2	−224	−.596	.127	1.266	−.228
3	361	.955	.101	1.128	.321
4	53	.138	.082	1.269	.041
5	−55	−.233	.651∗	3.318∗	−.318
6	−522	−1.398	.100	.961	−.466
7	17	.050	.289∗	1.642∗	.032
8	−147	−.390	.125	1.304	−.147
9	−36	−.100	.191	1.443∗	−.049
10	−419	−1.160	.172	1.146	−.529
11	−182	−.498	.175	1.362	−.229
12	827	2.359	.135	.595∗	.934∗
13	460	1.261	.148	1.073	.526
14	783	2.193	.119	.649	.806∗
15	106	.289	.173	1.395	.132
16	41	.107	.115	1.317	.039
17	−225	−.590	.102	1.232	−.199
18	−258	−.687	.121	1.235	−.255
19	−194	−.506	.088	1.230	−.157
20	−258	−.682	.108	1.218	−.238
21	−204	−.532	.094	1.234	−.172
22	−302	−.791	.090	1.165	−.249
23	−332	−.874	.097	1.149	−.287
24	−243	−.626	.063	1.173	−.162
25	−85	−.224	.108	1.300	−.078
26	140	.384	.181	1.395	.181
27	14	.039	.243	1.543∗	.022
28	−134	−.376	.222	1.468∗	−.201
29	−748	−2.177	.184	.708	−1.034∗
30	−97	−.267	.178	1.406∗	−.124
31	−286	−.745	.083	1.168	−.224
32	214	.557	.087	1.219	.172
33	805	2.225	.091	.617	.704
34	1,072	3.140	.098	.325∗	1.032∗
35	510	1.379	.117	.988	.502
36	−547	−1.451	.078	.917	−.421
37	−210	−.553	.111	1.253	−.196
38	−120	−.309	.072	1.241	−.086
39	−96	−.247	.062	1.235	−.064
40	−201	−.516	.057	1.188	−.127
41	230	.605	.106	1.233	.208
42	−66	−.168	.060	1.237	−.043
43	−1	−.004	.070	1.257	−.001
44	314	.819	.081	1.144	.242
45	295	.755	.051	1.127	.176

TABLE 11.6. *Influence statistics (DFBETAS) from the regression of BIOMASS on the five independent variables SAL, pH, K, Na, and Zn (from SAS's PROC REG, option INFLUENCE). An * on the measure of influence indicates that the value exceeds the reference value.*

	DFBETAS					
Obs.	X_0	*SAL*	*pH*	*K*	*Na*	*Zn*
1	.010	−.004	−.004	−.002	−.032	.001
2	.074	−.086	−.014	−.081	−.016	−.007
3	.123	−.094	−.166	−.005	.152	−.171
4	.020	−.020	−.019	−.010	.027	−.021
5	.065	−.030	−.108	.245	−.244	−.083
6	.054	−.069	.022	−.220	.007	.078
7	−.019	.022	.009	.026	−.021	.013
8	−.075	.069	.810	−.030	−.041	.091
9	.029	−.034	−.014	−.017	.004	−.014
10	−.310*	.285	.317*	−.068	−.177	.378*
11	−.174	.116	.172	.004	.022	.180
12	−.151	.442*	−.150	−.294	.092	.020
13	.307*	−.126	−.398*	−.052	−.023	−.351*
14	.133	.165	−.346*	−.041	−.090	−.331*
15	.107	−.076	−.104	−.062	.042	−.098
16	−.014	.013	.010	−.011	.005	.024
17	−.020	.027	.000	.081	−.028	−.061
18	.013	−.032	−.010	.084	.024	−.093
19	.008	−.056	.036	.041	.006	−.007
20	.043	−.118	.046	.039	.006	−.014
21	−.100	.070	.104	.106	−.084	.069
22	−.022	.012	.017	.074	.008	−.069
23	.010	−.075	.054	−.069	.163	−.044
24	.011	−.043	.030	−.014	.050	−.037
25	.041	−.057	−.012	−.007	.022	−.037
26	.074	−.074	−.006	−.047	.025	−.091
27	−.011	.012	.013	.005	−.010	.006
28	.090	−.094	−.118	.011	.037	−.042
29	−.130	.154	−.250	.235	−.010	.247
30	−.023	.026	−.024	.033	−.012	.038
31	−.141	.174	.069	.052	−.108	.097
32	−.066	.060	.059	.126	−.139	.078
33	−.044	−.179	.291	.027	.048	.249
34	.584*	−.752*	−.309*	−.183	.533*	−.406*
35	−.125	.041	.213	.307*	−.341*	.210
36	−.119	.206	.015	−.114	.039	−.002
37	.060	−.023	−.069	−.079	.076	−.119
38	−.026	.035	.020	−.023	.011	−.002
39	−.001	.009	−.001	−.015	.009	−.020
40	−.059	.065	.047	−.043	.018	.033
41	.033	−.081	.058	.017	−.044	.026
42	.010	.001	−.024	−.004	.009	−.020
43	.000	.000	−.001	−.000	.000	−.000
44	−.127	.075	.180	.080	−.105	.159
45	−.056	.013	.109	.025	−.024	.083

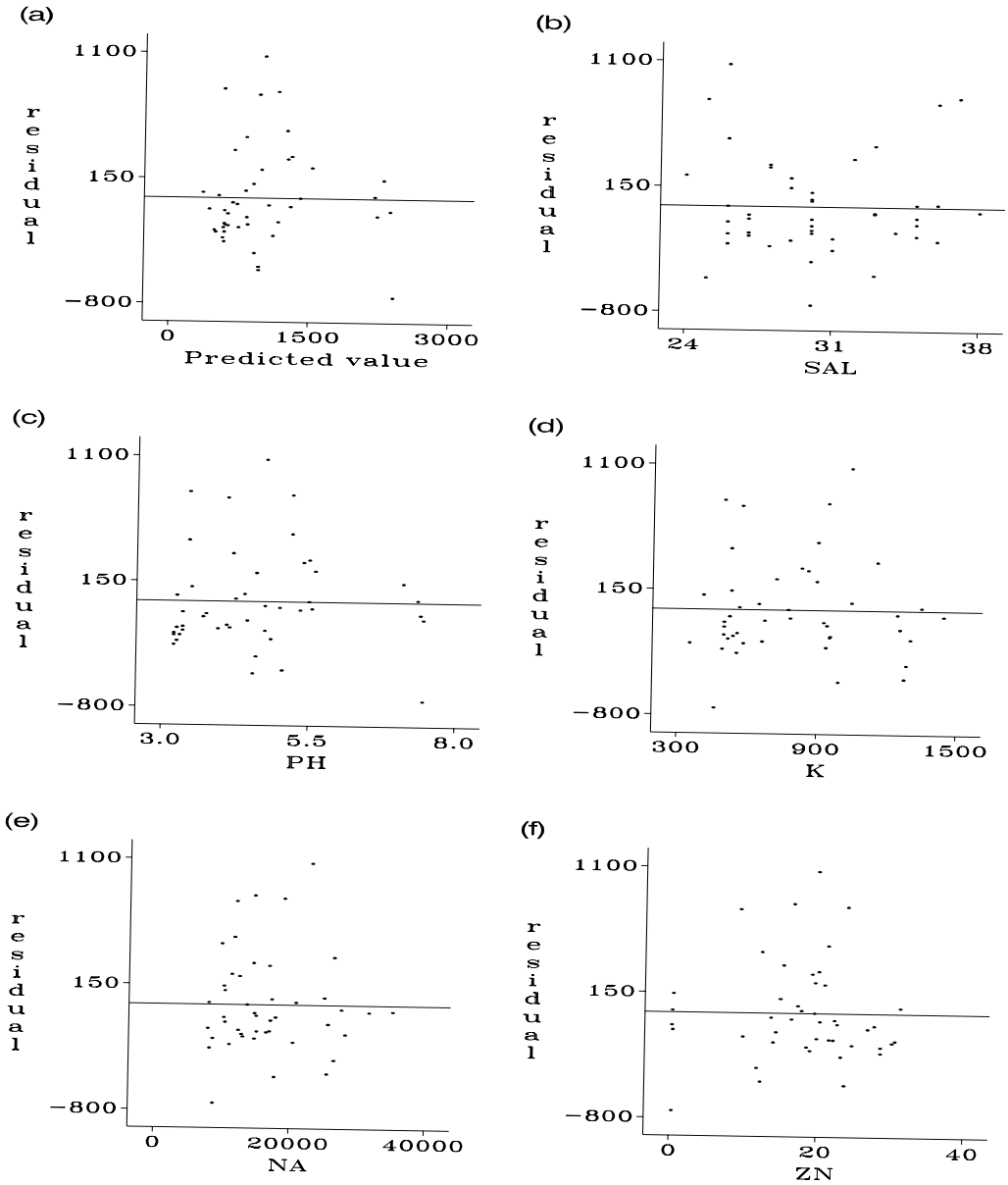

FIGURE 11.17. *Least squares residuals plotted against the predicted values (a) and each of the five independent variables [(b)–(f)] for the Linthurst September data.*

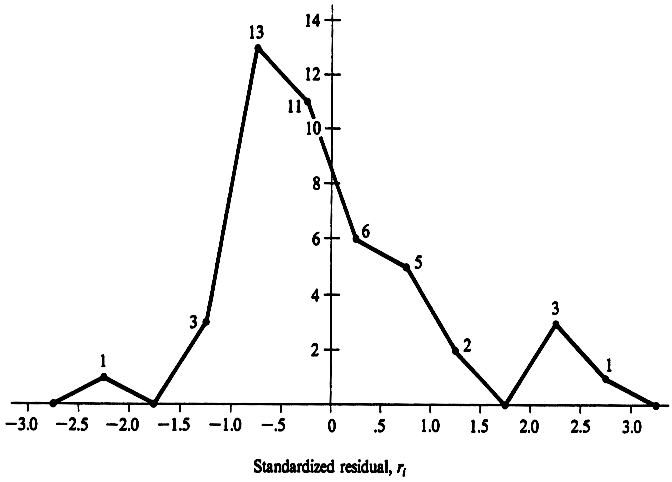

FIGURE 11.18. *Frequency polygon of the standardized residuals from the regression of BIOMASS on the five independent variables SALINITY, pH, K, Na, and Zn for the Linthurst September data.*

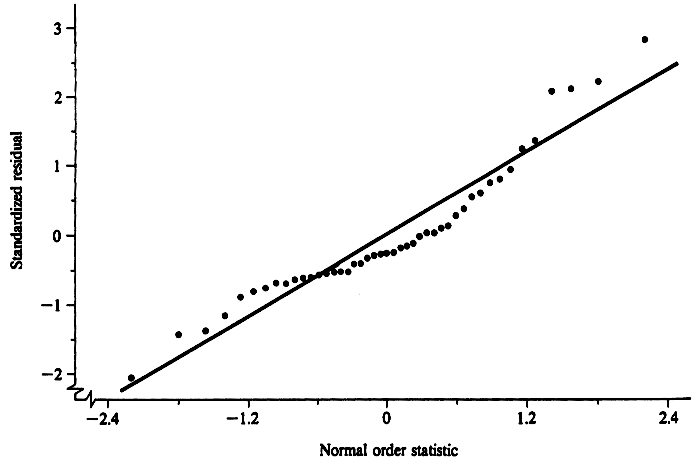

FIGURE 11.19. *Normal plot of the standardized residuals from the regression of BIOMASS on the five independent variables for the Linthurst September data.*

most negative residual, $r_{29} = -2.080$, is sufficiently larger in magnitude than the other negative residuals to raise the possibility that it might be an outlier. (A extreme standardized residual of -2.080 is not large for a normal distribution but seems large in view of the positive skewness and the fact that the next largest negative residual is -1.431.) The overall behavior of the residuals suggests that they may not be normally distributed. A transformation of the dependent variable might improve the symmetry of the distribution.

Standardized Residuals Versus $\widehat{Y}$

The values for the dependent variable *BIOMASS* cover a wide range from 236 to 2,436, Table 11.4. In such cases it is not uncommon for the variance of the dependent variable to increase with the increasing level of performance. The plot of the standardized residuals against $\widehat{Y}_i$ does not suggest any increase in dispersion for the larger $\widehat{Y}_i$. The five random samples taken at each of the nine sites, however, provide independent estimates of variation for *BIOMASS*. These "within-sampling-site" variances are not direct estimates of σ^2 because the five samples at each site are not true replicates; the values of the independent variables are not the same in all samples. They do provide, however, a measure of the differences in variance at very different levels of *BIOMASS*.

Standard Deviation Versus Mean

The plot of the standard deviation from each site versus the mean *BIOMASS* at each site, Figure 11.20, suggests that the standard deviation increases at a rate approximately proportional to the mean. As shown in Chapter 12, this suggests the logarithmic transformation of the dependent variable to stabilize the variance. The logarithmic transformation would also reduce the positive skewness noticed earlier. Continued analysis of these data would entail a transformation of *BIOMASS* to ln(*BIOMASS*), or some other similar transformation, and perhaps a change in the model as a result of the transformation. For the present purpose, however, the analysis is continued on the original scale.

Residuals Versus X_j

Inspection of the remaining plots in Figure 11.17—the residuals versus the independent variables—provides only one suggestion that the relationship of *BIOMASS* with the independent variable is other than linear. The residuals plot for *SALINITY*, Figure 11.17(b), suggests a slight curvilinear relationship between *BIOMASS* and *SALINITY*. A quadratic term for *SALINITY* in the model might be helpful. The five extreme points noticed in Figure 11.17(a) appear again as high values for pH, Figure 11.17(c), and as low values for Zn, Figure 11.17(f). These points are the five points from one sampling site, observations 26 to 30, and they are clearly having a major impact on the regression results. This site had very high *BIOMASS*, high pH, and low Zn.

Partial Regression Leverage Plots

The effects of the other independent variables may obscure relationships in plots of the residuals against any one independent variable. The partial regression leverage plots, Figure 11.21, are intended to avoid this problem. Each partial regression leverage plot shows the relationship between the dependent variable and one of the independent variables (including the

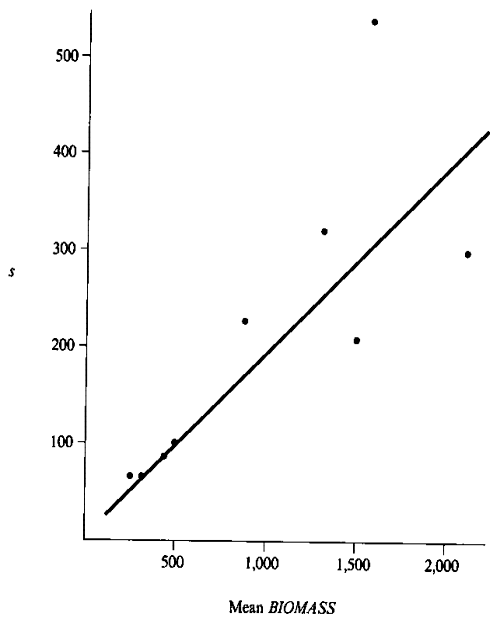

FIGURE 11.20. *The standard deviation among observations within sites plotted against the mean BIOMASS from the five observations at each site for the Linthurst September data.*

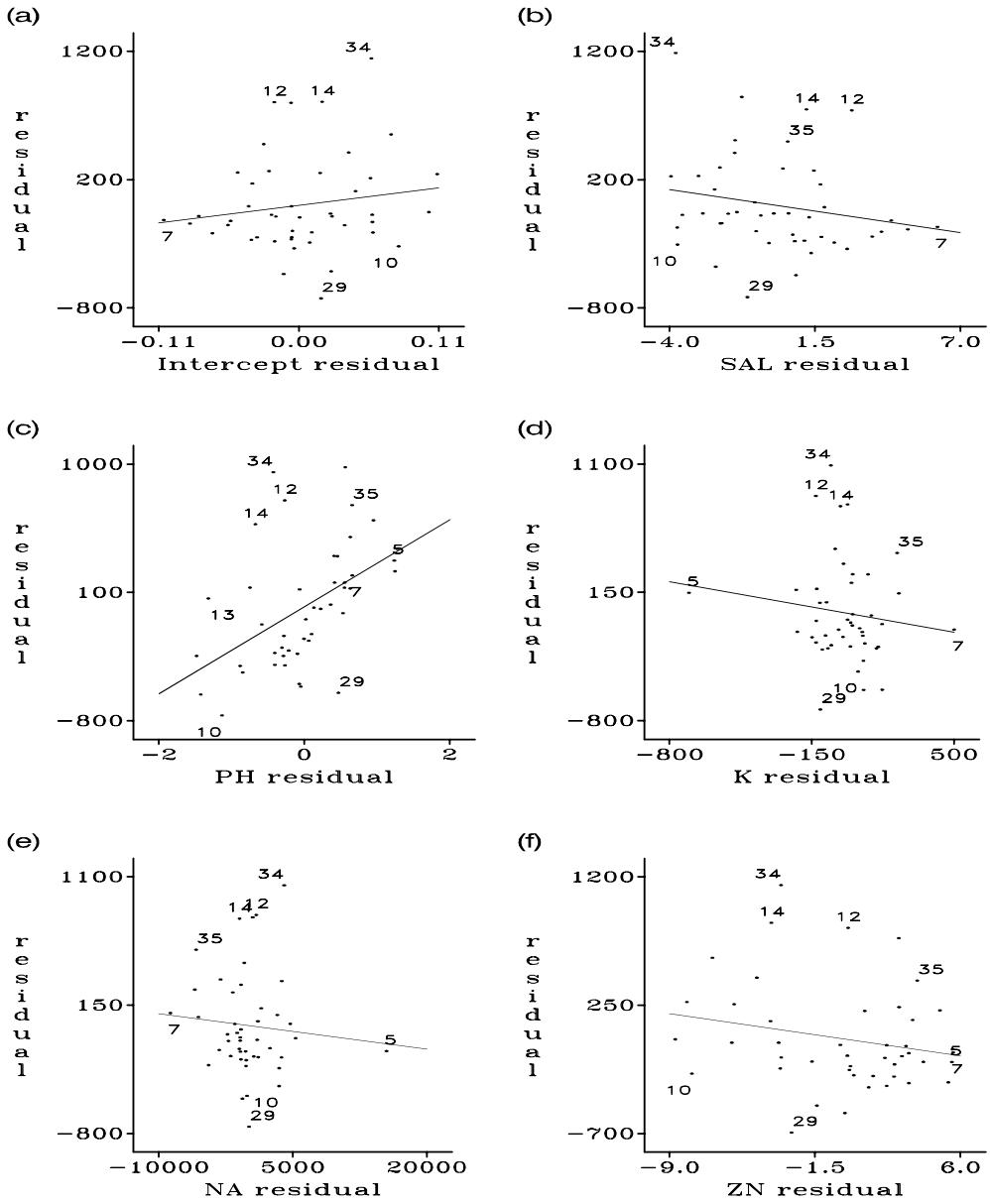

FIGURE 11.21. *The partial regression leverage plots from the regression of BIOMASS on the intercept and five independent variables for the Linthurst data. The slope of the plotted line is the partial regression coefficient for that variable. Numbers associated with specific points refer to observation number.*

intercept as an independent variable) after both have been adjusted for the effects of the other independent variables. The partial regression coefficient for the independent variable is shown by the slope of the relationship in the partial residuals plots, and any highly influential points will stand out as points around the periphery of the plot. Some of the critical observations in the plots have been labeled with their observation numbers for easier reference.

The following points are notable from the partial regression leverage plots.

1. The partial plot for *SALINITY* seems to indicate that if there is any curvilinear relationship as suggested by Figure 11.17(b) it is largely due to the influence of Observation 34 and, possibly, Observation 12.

2. Observation 34 repeatedly has a large positive residual for *BIOMASS* and may be having a marked influence on several regression coefficients. This is also the observation with the largest standardized residual, $r_{34} = 2.834$. It is important that the data for this point be verified.

3. The partial plots for K, Figure 11.21(d), and Na, Figure 11.21(e), show that points 5 and 7 are almost totally responsible for any significant relationship between *BIOMASS* and K and *BIOMASS* and Na. Without these two points in the data set, there would be no obvious relationship in either case.

4. Several other data points repeatedly occur on the periphery of the plots but not in such extreme positions. Point 29, the observation with the largest negative residual, always has a small partial residual for the independent variable. That is, Point 29 never deviates far from the zero mean for each independent variable after adjustment for the other variables. It is therefore unlikely that this observation has any great impact on any of the partial regression coefficients in this model. Nevertheless, it would be wise to recheck the data for this observation also.

5. One of the inadequacies of the influence statistics for detecting influential observations is illustrated with Points 5, 27, and 28 in the partial plot for pH, Figure 11.21(c). (Points 5, 27, and 28 are the cluster of three points where Observation 5 is labeled.) These three points have the largest partial residual for pH and would appear to have a major impact on the regression coefficient for pH. (Visualize what the slope of the regression would be if all three points were missing.) However, dropping only one of the three points may not appreciably affect the slope since the other two points are still "pulling" the line in the same direction. This illustrates that the simple influence statistics, where only one observation is dropped at a time, may not

detect influential observations when several points are having similar influence. The partial residuals plots show these jointly influential points.

Except for pH, these partial plots do not show any relationship between Y and the independent variable. This is consistent with the regression results using these five variables; only pH had a partial regression coefficient significantly different from zero (Table 5.2, page 165). In the all-possible regressions (Table 7.1, page 212), K and Na were about equally effective as the second variable in a two-variable model. The failure to see any association between Y and either of these two variables in the partial plots results from the collinearity in these data. (The collinearity is shown in Section 11.4.3.) Collinearity among the independent variables will tend to obscure regression relationships in the partial plots.

11.4.2 Influence Statistics

The influence statistics have been presented in Table 11.4 (Cook's D) and Table 11.5. The reference values for the influence statistics for this example, $p' = 6$ and $n = 45$, are as follows.

- v_{ii}, elements of $\boldsymbol{P}$ (called HAT DIAG in PROC REG): Average value is $p'/n = 6/45 = .133$. A point is potentially influential if $v_{ii} \geq 2p'/n = .267$.
- Cook's D: Cutoff value for Cook's D is $4/n = 4/45 = .09$ if the relationship to DFFITS is used.
- DFFITS: Absolute values greater than $2\sqrt{p'/n} = 2\sqrt{6/45} = .73$ indicate influence on $\widehat{Y}_i$.
- DFBETAS$_j$: Absolute values greater than $2/\sqrt{n} = .298$ indicate influence on $\widehat{\beta}_j$.
- COVRATIO: Values outside the interval $1 \pm 3p'/n = (.6, 1.4)$ indicate a major effect on the generalized variance.

The points that exceed these limits are marked with an asterisk in Tables 11.4 (Cook's D) and 11.5. Nine observations appear potentially influential, based on values of v_{ii}, or influential by Cook's D, DFFITS, or one or more of the DFBETAS$_j$; COVRATIO is ignored for the moment. These nine points are summarized in Table 11.7.

$\boldsymbol{v_{ii}}$

The influence statistics need to be studied in conjunction with the partial regression leverage plots, Figure 11.21. The plots give insight into why certain observations are influential and others are not. The ith diagonal elements of $\boldsymbol{P}$, v_{ii}, relate to the relative distance the ith observation is from the centroid of the sample X-space and, hence, that point's potential for

TABLE 11.7. *Nine observations showing potential influence (v_{ii}) or influence in the Linthurst data. The asterisk in the column indicates that the measure exceeded its cutoff point.*

				DFBETAS					
Obs.	v_{ii}	*Cook's D*	*DFFITS*	*Intercept*	*SAL*	*pH*	*K*	*Na*	*Zn*
5	*								
7	*								
10				*		*			*
12		*	*		*				
13				*		*			*
14		*	*			*			*
29		*	*						
34		*	*	*	*	*		*	*
35							*	*	

influencing the regression results. Two observations, 5 and 7, are flagged by v_{ii} meaning that these two points are the most "distant" in the sense of being on the fringe of the cloud of sample points. This is difficult to detect from simple inspection of the data, Table 5.1. Although both points have values near the extremes for one or more of the variables, neither has the most extreme value for any of the variables. They do, however, appear as extreme points in several of the residuals plots, particularly the plots for *K* and *Na*. Note, however, that neither observation is detected as being influential by any of the measures of influence. This appears to be a contradiction, but the measures of influence show the impact when only that one observation is dropped from the analysis. In the partial plots for *K* and *Na* it is clear that the two observations are operating in concert; eliminating either 5 or 7 has little effect on the regression coefficient because of the influence of the remaining observation. Similarly, and as noted earlier, the cluster of four points 5, 7, 16, and 37 (only 5 and 7 are labeled) are operating together in the partial plot for *Zn* to mask the effect of eliminating one of these points. In other cases, as with Point 5 in the partial plot for *SALINITY* or Point 7 in the partial plot for *pH*, the potentially influential point is not an extreme point in that dimension and is, in fact, not influential for that particular regression coefficient.

Cook's *D*, DFFITS, and DFBETAS

Cook's *D* and DFFITS are very similar measures and identify the same four observations as being influential: Observations 12, 14, 29, and 34. Dropping any one of these four points causes a relatively large shift in $\widehat{\boldsymbol{\beta}}$ or $\widehat{\boldsymbol{Y}}$, depending on the interpretation used. They are consistently on the periphery of the partial plots. Point 33 is also on the periphery in all plots but was not flagged by either Cook's *D* or DFFITS. However, its value for both measures is only slightly below the cutoff. Of these four points, only 34 has influence on most of the individual regression coefficients; only

DFBETAS for K is not flagged. This is consistent with the position of 34 in the partial plots.

Finally, there are three observations, 10, 13, and 35, that have been flagged as having influence on one or more regression coefficients but which were not detected by any of the general influence measures, v_{ii}, Cook's D, or DFFITS. In these cases, however, the largest $\text{DFBETAS}_{j(i)}$ was .3457, only slightly above the critical value of .298.

COVRATIO

The COVRATIO statistic identifies nine observations as being influential with respect to the variance–covariance matrix of $\widehat{\boldsymbol{\beta}}$; all but two of these nine points increase the precision of the estimates. The two points, 12 and 34, whose presence inflates the generalized variance (COVRATIO $<$ 1.0) are two points that were influential for several regression coefficients. These two points have the largest standardized residuals, so when they are eliminated the estimate of σ^2 and the generalized variance decrease. Thus, in this case, the low COVRATIO might be reflecting inadequacies in the model.

Discussion

What is gained from the partial regression leverage plots and the influence measures? They must be viewed as diagnostic techniques, as methods for studying the relationship between the regression equation and the data. These are not tests of significance, and flagging an observation as influential does not imply that the observation is somehow in error. Of course, an error in the data can make an observation very influential and, therefore, careful editing of the data should be standard practice. Detection of a highly influential point suggests that the editing of the data, and perhaps the protocol for collecting the data, be rechecked.

A point may be highly influential because, due to inadequate sampling, it is the only observation representing a particular region of the X-space. Is this the reason Points 5 and 7 are so influential? They are the two most "remote" points and are almost totally responsible for the estimates of the regression coefficients for K and Na. More data might "fill in the gaps" in the X-space between these two points and the remaining sample points and, as a result, tend to validate these regression estimates. Alternatively, more data might confirm that these two points are anomalies for the population and, hence, invalidate the present regression estimates. If one is forced to be content with this set of data, it would be prudent to be cautious regarding the importance of K and Na since they are so strongly influenced by these two data points.

The purpose of the diagnostic techniques is to identify weaknesses in the regression model or the data. Remedial measures, correction of errors in the data, elimination of true outliers, collection of better data, or improvement of the model, will allow greater confidence in the final product.

TABLE 11.8. *Collinearity diagnostics for the regression of BIOMASS on the five independent variables SAL, pH, K, Na, and Zn, Linthurst data (from SAS PROC REG, option COLLIN).*

Prin. Comp. Dimen.	*Eigen-values*	*Cond. Index*	*Variance Decompositon Proportion*					
			Inter.	*SAL*	*pH*	*K*	*Na*	*Zn*
1	5.57664	1.000	.0001	.0002	.0006	.0012	.0013	.0011
2	.21210	5.128	.0000	.0007	.0265	.0004	.0000	.1313
3	.15262	6.045	.0015	.0032	.0141	.0727	.1096	.0155
4	.03346	12.910	.0006	.0713	.1213	.2731	.2062	.0462
5	.02358	15.380	.0024	.0425	.1655	.5463	.5120	.0497
6	.00160	58.977	.9954	.8822	.6719	.1062	.1709	.7561

11.4.3 Collinearity Diagnostics

The collinearity diagnostics (Table 11.8) were obtained from the "COLLIN" option in PROC REG (SAS Institute Inc., 1989b). The collinearity measures are obtained from the eigenanalysis of the standardized $\boldsymbol{X}'\boldsymbol{X}$; the sum of squares for each column is unity and the eigenvalues must add to $p' = 6$. The condition number for $\boldsymbol{X}$ is 58.98, an indication of moderate to strong collinearities. The condition indices for the fourth and fifth dimensions are greater than 10, indicating that these two dimensions of the X-space may also be causing some collinearity problems.

Variance Decomposition Proportions

The variance decomposition proportions show that the sixth principal component dimension is accounting for more than 50% of the variance in four of the six regression coefficients. Thus, the intercept, *SALINITY*, *pH*, and *Zn* are the four independent variables primarily responsible for the near-singularity causing the collinearity problem. (The eigenvectors would be required to determine the specific linear function of the $\boldsymbol{X}$ vectors that causes the near-singularity.)

If the sixth principal component dimension is eliminated from consideration and the variance proportions of the remaining dimensions restandardized to add to one, the variance proportions associated with the fifth principal component dimension account for more than 50% of the remaining variance for four of the six regression coefficients. Similarly, eliminating the fifth principal component dimension leaves the fourth principal component dimension accounting for more than 50% of the variance of four of the six regression coefficients.

Thus, it appears that the three last principal component dimensions may be contributing to instability of the regression coefficients. The course of action to take in the face of this problem is discussed in Chapter 13. ■

11.5 Exercises

11.1. Plot the following Studentized residuals against $\widehat{Y}_i$. Does the pattern suggest any problem with the model or the data?

r_i^*	$\widehat{Y}_i$	r_i^*	$\widehat{Y}_i$	r_i^*	$\widehat{Y}_i$	r_i^*	$\widehat{Y}_i$
−.53	10	−.92	11	−1.55	15	−.82	18
.23	19	−.45	23	−1.00	26	.47	32
−.36	38	.75	41	1.27	43	1.85	48
1.16	49	.04	49	.96	51	−1.03	60
−.25	65	−.92	67	−1.84	69	.52	73
−.80	76	−.88	79	.57	85	−.25	90
1.51	93	1.62	99	.65	100		

11.2. Plot the following Studentized residuals against the corresponding $\widehat{Y}_i$. What does the pattern in the residuals suggest?

r_i^*	$\widehat{Y}_i$	r_i^*	$\widehat{Y}_i$	r_i^*	$\widehat{Y}_i$	r_i^*	$\widehat{Y}_i$
−.53	60	−.92	81	−1.55	83	−.82	78
.23	19	−.45	53	−1.00	63	.47	42
−.36	48	.75	41	1.27	23	1.85	98
1.16	29	.04	49	.96	21	−1.03	80
−.25	65	−.92	57	−1.84	72	.52	33
−.80	76	−.88	69	.57	65	−.25	30
1.51	13	1.62	19	.65	25		

11.3. For each of the following questions, choose the *one* you would use (for example, a plot or an influence statistic) to answer the question. Describe your choice and what you would expect to see if there were no problem.

(a) Do the ϵ_i have homogeneous variance?

(b) Is the regression being unduly influenced by the 11th observation?

(c) Is the regression on X_3 really linear as the model states?

(d) Is there an observation that does not seem to fit the model?

(e) Has an important independent variable been omitted from the model?

11.4. For each of the following diagnostic tools, indicate what aspects of ordinary least squares are being checked and how the results might indicate problems.

(a) Normal plot of r_i^*.

(b) Plot of $\boldsymbol{e}$ versus $\widehat{\boldsymbol{Y}}$.

(c) Cook's D.

(d) v_{ii}, the diagonal elements of $\boldsymbol{P}$.

(e) DFBETAS$_j$.

11.5. The collinearity diagnostics in PROC REG in SAS gave the eigenvalues 2.1, 1.7, .8, .3, and .1 for a set of data.

(a) Compute the condition number for the matrix and the condition index for each principal component dimension.

(b) Compute Thisted's measure of collinearity *mci*. Does the value of *mci* indicate a collinearity problem?

11.6. A regression problem gave largest and smallest eigenvalues of 3.29 and .02, and the following variance decomposition proportions corresponding to the last principal component.

Parameter:	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$
Variance Proportion:	.72	.43	.18	.85	.71	.02

(a) Do these results indicate collinearity problems?

(b) Which $\widehat{\beta}$s, if any, are "suffering" from collinearity? Explain the basis for your conclusion.

11.7. PROC REG (in SAS) was run on a set of data with $n = 40$ observations on Y and three independent variables. The collinearity diagnostics gave the following results.

Number	*Eigenvalue*	*Cond. Index*	*Variance Proportions*			
			Intercept	X_1	X_2	X_3
1	3.84682	1.000	.0007	.0010	.0043	.0075
2	.09992	6.205	.0032	.0059	.1386	.8647
3	.04679	9.067	.0285	.0942	.7645	.0912
4	.00647	24.379	.9676	.8990	.0926	.0366

(a) What is the rank of $\boldsymbol{X}$ in this model?

(b) What is the condition number for $\boldsymbol{X}$? What does that say about the potential for collinearity problems?

(c) Interpret the variance proportions for the fourth principal component. Is there variance inflation from the collinearity? Which regression coefficients are being affected most?

(d) Compute the variance proportions for the *third* principal component after the fourth has been removed. Considering the condition index and the variance proportions for the third principal component, is there variance inflation from the third component?

11.8. An experiment was designed to estimate the response surface relating Y to two quantitative independent variables. A 4×4 factorial set of treatments was used with $X_1 = 1, 2, 3,$ and 4, and $X_2 = 65, 70, 75,$ and 80.

(a) Set up $\boldsymbol{X}$ for the linear model,

$$Y_{ij} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_{ij}.$$

(You need only use the 16 distinct rows of $\boldsymbol{X}$.) Do the singular value decomposition on the scaled $\boldsymbol{X}$. Is there any indication of collinearity problems?

(b) Redefine the model so that X_1 and X_2 are both expressed as deviations from their means. Redo the singular value decomposition. Have the collinearity diagnostics changed? Explain the differences, if any.

(c) Use the centered Xs but include squares of the Xs in the model. Redo the singular value decomposition. Have the collinearity diagnostics changed? Explain the changes.

11.9. The following are the results of a principal component analysis, on $\boldsymbol{Z}$, of data collected from a fruit fly experiment attempting to relate a measure of fly activity, WFB = wing beat frequency, to the chemical activity of four enzymes, SDH, FUM, GH, and GO. Measurements were made on $n = 21$ strains of fruit fly. (Data courtesy of Dr. Laurie Alberg, North Carolina State University.)

Eigenvalues: 2.1970 1.0790 .5479 .1762

	Eigenvectors			
Variable	1*st*	2*nd*	2*rd*	4*th*
SDH	.547	−.465	−.252	−.649
FUM	.618	−.043	−.367	.694
GH	.229	.870	−.306	−.312
GO	.516	.158	.842	−.005

(a) Compute the proportion of the dispersion in the X-space accounted for by each principal component.

(b) Compute the condition number for $\boldsymbol{Z}$ and the condition index for each principal component. What do the results suggest about possible variance inflation from collinearity?

(c) Describe the first principal component in terms of the original centered and standardized variables. Describe the second principal component.

(d) The sum of the variances of the estimates of the least squares regression coefficients, $\text{tr}[\mathbf{Var}(\widehat{\boldsymbol{\beta}})] = \sum(1/\lambda_j)\sigma^2$, must be larger than σ^2/λ_4. Compute this minimum (in terms of σ^2). How does this compare to the minimum if the four variables had been orthogonal?

11.10. The following questions relate to the residuals analysis reported in Tables 11.4 and 11.5.

(a) Compute $s^2(\widehat{Y}_i) + s^2(e_i)$ for several choices of i. How do you explain the fact that you obtain very nearly the same number each time?

(b) Find the largest and smallest $s(\widehat{Y}_i)$ and the largest and smallest v_{ii}. Explain why they derive from the same observations in each case.

(c) A COVRATIO equal to 1.0 implies that the ith point has no real impact on the overall precision of the estimates. A COVRATIO less than 1.0 indicates that the presence of the ith observation has *decreased* the precision of the estimates (e.g., Observation 12). How do you explain the *presence* of an additional observation causing *less* precision?

(d) Cook's D provides a measure of the shift in $\widehat{\boldsymbol{\beta}}$. The DFBETAS measure shifts in the individual $\widehat{\beta}_j$. How do you explain the fact that Observation 29, which has the largest value of Cook's D, has no DFBETAS_j that exceed the cutoff point, whereas Observation 34, which has the next to the largest value of Cook's D, shows major shifts in all but one of the regression coefficients? Conversely, explain why Observation 10 has a small Cook's D but shows major shifts in the intercept and the regression coefficients for pH and Zn.

11.11. The accompanying table reports data on percentages of sand, silt, and clay at 20 sites. [The data are from Nielsen, Biggar, and Erh (1973), as presented by Andrews and Herzberg (1985). The depths 1, 2, and 3 correspond to depths 1, 6, and 12 in Andrews and Herzberg.] Use sand, silt, and clay percentages at the three depths as nine columns

of an $\boldsymbol{X}$ matrix.

Plot	*Depth* 1			*Depth* 2			*Depth* 3		
No.	*Sand*	*Silt*	*Clay*	*Sand*	*Silt*	*Clay*	*Sand*	*Silt*	*Clay*
1	27.3	25.3	47.4	34.9	24.2	40.7	20.7	36.7	42.6
2	40.3	20.4	39.4	42.0	19.8	38.2	45.0	25.3	29.8
3	12.7	30.3	57.0	25.7	25.4	49.0	13.1	37.6	49.3
4	7.9	27.9	64.2	8.0	26.6	64.4	22.1	30.8	47.1
5	16.1	24.2	59.7	14.3	30.4	55.3	5.6	33.4	61.0
6	10.4	27.8	61.8	18.3	27.6	54.1	8.2	34.4	57.4
7	19.0	33.5	47.5	27.5	37.6	34.9	.0	30.1	69.9
8	15.5	34.4	50.2	11.9	38.8	49.2	4.4	40.8	54.8
9	21.4	27.8	50.8	20.2	30.3	49.3	18.9	36.1	45.0
10	19.4	25.1	55.5	15.4	35.7	48.9	3.2	44.4	52.4
11	39.4	25.5	35.6	42.6	23.6	33.8	38.4	32.5	29.1
12	32.3	32.7	35.0	20.6	28.6	50.8	26.7	37.7	35.6
13	35.7	25.0	39.3	42.5	20.1	37.4	60.7	13.0	26.4
14	35.2	19.0	45.8	32.5	27.0	40.5	20.5	42.5	37.0
15	37.8	21.3	40.9	44.2	19.1	36.7	52.0	21.2	26.8
16	30.4	28.7	40.9	30.2	32.0	37.8	11.1	45.1	43.8
17	40.3	16.1	43.6	34.9	20.8	44.2	5.4	44.0	50.6
18	27.0	28.2	44.8	37.9	30.3	31.8	8.9	57.8	32.8
19	32.8	18.0	49.2	23.2	26.3	50.5	33.2	26.8	40.0
20	26.2	26.1	47.7	29.5	34.9	35.6	13.2	34.8	52.0

(a) From the nature of the variables, is there any reason to expect a collinearity problem if these nine variables were to be used as independent variables in multiple regression analysis?

(b) Center and scale the variables and do a singular value decomposition on $\boldsymbol{Z}$. Does the SVD indicate the presence of a collinearity problem? Would you have obtained the same results if the variables had not been centered and the intercept included? Explain.

12

TRANSFORMATION OF VARIABLES

Several methods for detecting problem areas were discussed in Chapter 11 and their applications to real data were demonstrated.

This chapter discusses the use of transformations of variables to simplify relationships, to stabilize variances, and to improve normality. Weighted least squares and generalized least squares are presented as methods of handling the problems of heterogeneous variances and lack of independence.

There are many situations in which transformations of the dependent or independent variables are helpful in least squares regression. Chapter 10 suggested transformation of the *dependent* variable as a possible remedy for some of the problems in least squares. In this chapter, the reasons for making transformations, including transformations on the independent variables, and the methods used to choose the appropriate transformations are discussed more fully. Generalized least squares and weighted least squares are included in this chapter because they can be viewed as ordinary least squares regression on a transformed dependent variable.

12.1 Reasons for Making Transformations

There are three basic reasons for transforming variables in regression. Transformations of the dependent variable were indicated in Chapter 10 as possi-

ble remedies for nonnormality and for heterogeneous variances of the errors. A third reason for making transformations is to simplify the relationship between the dependent variable and the independent variables.

The Simplest Model

A basic rule of science says that, all other things being equal, the simplest model that describes the observed behavior of the system should be adopted. Simple relationships are more easily understood and communicated to others. With statistical models, the model with the fewest parameters is considered the simplest, straight-line relationships are considered simpler than curvilinear relationships, and models linear in the parameters are simpler than nonlinear models.

Curvilinear Relationships

Curvilinear relationships between two variables frequently can be simplified by a transformation on either one or both of the variables. The power family of transformations and a few of the two-bend transformations are discussed for this purpose (Section 12.2).

Nonlinear Models

Many models nonlinear in the parameters can be linearized, reexpressed as a linear function of the parameters, by appropriate transformations. For example, the relationship

$$Y = \alpha X^{\beta}$$

is linearized by taking the logarithm of both sides of the equality giving

$$\ln(Y) = \ln(\alpha) + \beta[\ln(X)]$$

or

$$Y^* = \alpha^* + \beta X^*.$$

The nonlinear relationship between Y and X is represented by the linear relationship between Y^* and X^*.

Heterogeneous Variances and Nonnormality

The effects of heterogeneous variances and nonnormality on least squares regression have already been noted (Chapter 10). Transformation of the dependent variable was indicated as a possible remedy for both. Sections 12.3 and 12.4 discuss the choice of transformations for these two situations. Alternatively, weighted least squares or its more general version, generalized least squares, can be used to account for different degrees of precision in the observations. These methods are discussed in Section 12.5.

Throughout this discussion, it should be remembered that it may not be possible to find a set of transformations that will satisfy all objectives. A transformation on the dependent variable to simplify a nonlinear relationship will destroy both homogeneous variances and normality if these assumptions were met with the original dependent variable. Or, a transformation to stabilize variance may cause nonnormality. Fortunately, transformations for homogeneity of variance and normality tend to go hand-in-hand so that often both assumptions are more nearly satisfied after an appropriate transformation (Bartlett, 1947). If one must make a choice, stabilizing variance is usually given precedence over improving normality. Many recommend that simplifying the relationship should take precedence over all.

The latter would seem to depend on the intrinsic value and the general acceptance of the nonlinear relationship being considered. If a nonlinear model is meaningful and is readily interpreted, a transformation to linearize the model would not seem wise if it creates heterogeneous variance or nonnormality.

12.2 Transformations to Simplify Relationships

It is helpful to differentiate two situations where transformations to simplify relationships might be considered. In the first case, there is *no* prior idea of the form the model should take. The objective is to empirically determine mathematical forms of the dependent and independent variables that allow the observed relationship to be represented in the simplest form, preferably a straight line. The model is to be linear in the parameters; only the form in which the variables are expressed is being considered.

In the second case, prior knowledge of the system suggests a nonlinear mathematical function, nonlinear in the parameters, for relating the dependent variable to the independent variable(s). The purpose of the transformation in this case is to reexpress the nonlinear model in a form that is linear in the parameters and for which ordinary least squares can be used. Such linearization of nonlinear models is not always possible but when it is possible the transformation to be used is dictated by the functional form of the model.

"One-Bend" Transformations

The power family of transformations $X^* = X^k$ or $Y^* = Y^k$ provides a useful set of transformations for "straightening" a single bend in the relationship between two variables. These are referred to as the **"one-bend" transformations** (Tukey, 1977; Mosteller and Tukey, 1977) and can be used on either X or Y. Ordering the transformations according to the exponent k gives a sequence of power transformations, which Mosteller and Tukey (1977) call the *ladder of reexpressions*. The common powers considered are

$$k = -1,\ -\frac{1}{2},\ 0,\ \frac{1}{2},\ 1,\ 2,$$

where the power transformation $k = 0$ is to be interpreted as the logarithmic transformation. The power $k = 1$ implies no transformation.

Ladder of Transformations

The rule for straightening a "one-bend" relationship is to move up or down the ladder of transformations according to the direction in which the bulge of the curve of Y versus X points. For example, if the bulge in the curve points toward lower values of Y, as in the exponential decay and growth curves shown in Figure 12.1, moving down the ladder of transformations to $\sqrt{Y}, \ln(Y)$, and $1/Y$ will tend to straighten the relationship. [In the specific case of the exponential function, it is known that the logarithmic transformation ($k = 0$) will give a linear relationship.] For the exponential decay curve, the bulge also points toward lower values of X.

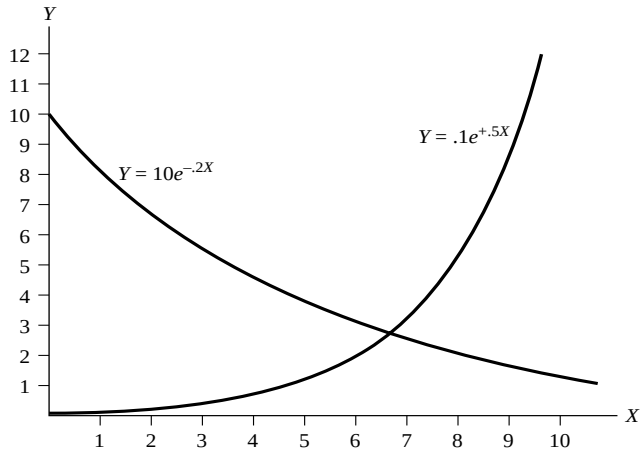

FIGURE 12.1. *Examples of the exponential growth curve and the exponential decay curve.*

Therefore, moving down the ladder for a power transformation of X will also tend to straighten the relationship. For the exponential growth curve, however, one must move up the ladder to X^2 or X^3 for a power transformation on X to straighten the relationship; the bulge points upward with respect to X. The inverse polynomial curve (Figure 12.2) points upward with respect to Y and downward with respect to X. Therefore, *higher* powers of Y or *lower* powers of X will tend to straighten the relationship.

How far one moves on the ladder of transformations depends on the sharpness of the curvature. This is easily determined when only one independent variable is involved by trying several transformations on a few observations covering the range of the data and then choosing that transformation which makes the points most nearly collinear. Several independent variables make the choice more difficult, particularly when the data are not balanced or when there are interactions among the independent variables. The partial regression leverage plots for the first-degree polynomial model will show the relationship between Y and a particular independent variable *after* adjustment for all other independent variables, and should prove helpful in determining the power transformation. Since only one transformation on Y can be used in any one analysis, attention must focus on transformations of the independent variables when several independent variables are involved.

Box–Tidwell Method

Box and Tidwell (1962) give a computational method for determining the power transformations on *independent* variables such that lower-order polynomial models of the transformed variables might be used. They assume that the usual least squares assumptions are well enough satisfied on the present scale of Y (perhaps after some transformation) so that further

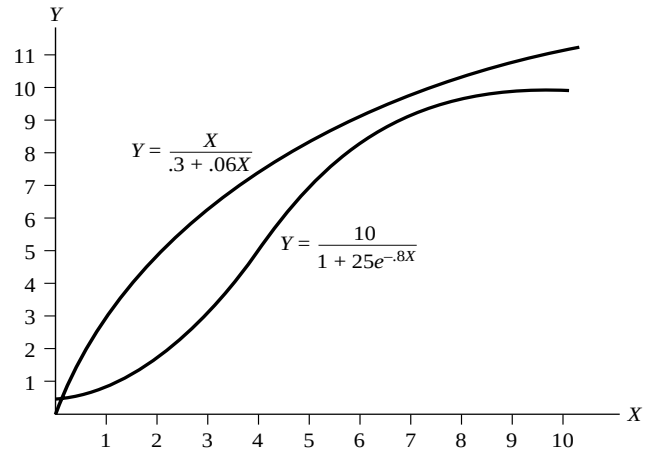

FIGURE 12.2. *Examples of the inverse polynomial model and the logistic model.*

transformations to simplify relationships must be done on the independent variables. The **Box–Tidwell method** is a general method applicable to any model and any class of transformations. However, its consideration here is restricted to the polynomial model and power transformations on individual Xs. The steps of the Box–Tidwell method are given for a full, second-degree polynomial model in two variables. The simplifications of the procedure and an illustration for the first-degree polynomial model are given.

Procedure

The proposed second degree model is

$$\begin{aligned} Y_i &= F(\boldsymbol{U}, \boldsymbol{\beta}) + \epsilon \\ &= \beta_0 + \beta_1 U_{i1} + \beta_2 U_{i2} + \beta_{11} U_{i1}^2 + \beta_{22} U_{i2}^2 + \beta_{12} U_{i1} U_{i2} + \epsilon_i, \end{aligned}$$

where $i = 1, \ldots, n$ and $j = 1, 2$. The U_{ij} are power transformations on X_{ij}:

$$U_{ij} = \begin{cases} X_{ij}^{\alpha_j} & \text{if } \alpha_j \neq 0 \\ \ln(X_{ij}) & \text{if } \alpha_j = 0. \end{cases} \tag{12.1}$$

The objective is to find the α_1 and α_2 for transforming X_{i1} and X_{i2} to U_{i1} and U_{i2}, respectively, that provide the best fit of $F(\boldsymbol{U}, \widehat{\boldsymbol{\beta}})$ to Y. The steps in the Box–Tidwell method to approximate the α_j are as follows.

1. Fit the polynomial model to Y to obtain the regression equation in the *original* variables $\widehat{\boldsymbol{Y}} = F(\boldsymbol{X}, \widehat{\boldsymbol{\beta}})$.

2. Differentiate $\widehat{\boldsymbol{Y}}$ with respect to each independent variable and evaluate the partial derivatives for each of the n observations to obtain

$W_{ij} = \partial(\widehat{\boldsymbol{Y}})/\partial \boldsymbol{X}_j$, $i = 1, \ldots, n$. For the quadratic model,

$$W_{i1} = \widehat{\beta}_1 + 2\widehat{\beta}_{11}X_{i1} + \widehat{\beta}_{12}X_{i2}$$

and

$$W_{i2} = \widehat{\beta}_2 + 2\widehat{\beta}_{22}X_{i2} + \widehat{\beta}_{12}X_{i1}.$$

(For the first degree polynomial model, the partial derivatives are simply the constants $W_{i1} = \widehat{\beta}_1$ and $W_{i2} = \widehat{\beta}_2$.)

3. Create two new independent variables Z_{i1} and Z_{i2} by multiplying each W_{ij} by the corresponding values of $X_{ij}[\ln(X_{ij})]$, $j = 1,\ 2$.

4. Refit the polynomial model augmented with the two new variables Z_1 and Z_2. Let $\widehat{\gamma}_j$ be the partial regression coefficient obtained for Z_j.

5. Compute the desired power transformations as $\widehat{\alpha}_j = \widehat{\gamma}_j + 1$, $j = 1, 2$.

This is the end of the first round of iteration to approximate the coefficients for the power transformation. The α_j are then used to transform the original Xs (according to Equation 12.1) and the process is repeated using the power-transformed variables as if they were the original variables. The α_j obtained on the second iteration are used to make a power transformation on the *previously* transformed variables. (This is equivalent to transforming the original variables using the product of α_j from the first and second steps as the power on the jth variable.) The iteration terminates when α_j converges close enough to 1.0 to cause only trivial changes in the power transformation.

Example 12.1

The Box–Tidwell method is illustrated using data from an experiment to test tolerance of certain families of pine to salt water flooding (Land, 1973). Three seedlings from each of eight families of pine were subjected to 0, 72, or 144 hours of flooding in a completely random experimental design. The data are given in Table 12.1. The response variable is the chloride content (% dry matter) of the pine needles. (The $Y = .00\%$ chloride measurement for Family 3 was changed to $Y = .01$ and $X = 0$ hours flooding was changed to $X = 1$ hour. Both changes were made to avoid problems with taking logarithms in the Box–Tidwell method and in the Box–Cox method used in Exercise 12.1.)

The regression of $Y =$ (% Chloride) on $X =$ hours of exposure, and allowing a different intercept for each family, required a quadratic polynomial to adequately represent the relationship. The Box–Tidwell method is used to search for a power transformation on X that allows the relationship to be represented by a straight line. The first step fits the model

$$Y_{ijk} = \beta_{0i} + \beta X_j + \epsilon_{ijk},$$

TABLE 12.1. *Chloride content (percent dry weight) of needles of pine seedlings exposed to 0, 72, or 144 hours of flooding with sea water. Nine seedlings of each of eight genetic families were used in a completely random experimental design. (Data from S. B. Land, Jr., 1973, Ph.D. Thesis, N.C. State University, and used with permission.)*

	Hours of Flooding with Saltwater								
Family	0			72			144		
1	.36	.47	.30	3.54	4.35	4.88	6.13	6.49	7.04
2	.32	.63	.51	4.95	4.45	1.50	6.46	4.35	2.18
3	.00	.43	.72	4.26	3.89	6.54	5.93	6.29	9.62
4	.54	.70	.49	3.69	2.81	4.08	5.68	4.68	5.79
5	.44	.42	.39	3.01	4.08	4.54	6.06	6.05	6.97
6	.55	.57	.45	2.32	3.57	3.59	4.32	6.11	6.49
7	.20	.51	.27	3.16	3.17	3.75	4.79	5.74	5.95
8	.31	.44	.84	2.80	2.96	2.04	10.58	4.44	1.70

where $i = 1, \ldots, 8$ designates the family, X_j is the number of hours of flooding, $j = 1,\ 2,\ 3$, and $k = 1,\ 2,\ 3$ designates the seedling within each $i,\ j$ combination. The estimate of the regression coefficient is $\widehat{\beta} = .01206$. This is the partial derivative of $\widehat{Y}_{ijk}$ with respect to X when the model is linear in X; therefore, $W_i = \widehat{\beta}$ in step two. Thus, the new independent variable is

$$Z_j = 0.01206 X_j [\ln(X_j)].$$

The model is augmented with Z_j to give

$$Y_{ijk} = \beta_{0i} + \beta X_j + \gamma Z_j + \epsilon_{ijk}.$$

Fitting this model gives $\widehat{\gamma} = -.66971$; thus, $\widehat{\alpha} = \widehat{\gamma} + 1 = .33029$ is the estimated power transformation on X from the first iteration. The cycle is repeated using the transformed $X(1)_j = (X_j)^{.33029}$ in place of X_j.

The second iteration gives $\widehat{\beta} = 0.41107$, $\widehat{\gamma} = .22405$, and $\widehat{\alpha} = 1.22405$. Thus, the power transformation on $X(1)_j$ is $X(2)_j = (X(1)_j)^{1.22405}$. The third iteration uses $X(2)_j$ in place of $X(1)_j$.

The third iteration gives $\widehat{\beta} = .26729$, $\widehat{\gamma} = .00332$, and $\widehat{\alpha} = .99668$. If the iterations were to continue, the new independent variable would be $X(3)_j = (X(2)_j)^{.99668}$. Since $\widehat{\alpha}$ is very close to 1.0, giving only trivial changes in $X(2)_j$, the iterations can stop. The estimated power transformation on X is the product of the three αs, $(.33029)(1.22405)(.99668) = .4023$, which is close to the square root transformation on X. In this example, a linear model using the transformed $X^* = X^{.4023}$ provides the same degree of fit as a quadratic model using the original X_j; the residual sums of squares from the two models are very nearly identical. ■

Estimating Power with Nonlinear Regression

An alternative method of determining the power transformations is to include the powers on the independent variables as parameters in the model and use nonlinear least squares to simultaneously estimate all parameters (Chapter 15). This may, in some cases, lead to overparameterization of the model and failure of the procedure to find a solution. There is no assurance that appropriate power transformations will exist to make the chosen polynomial fit the data. The usual precautions should be taken to verify that the model is adequate for the purpose.

Transformations and Model Assumptions

The objective to this point has been to find the power transformation of *either* Y or X that most nearly straightens the relationship. However, any transformation on the *dependent* variable will also affect the distributional properties of Y. Hence, the normality and common variance assumptions on ϵ must be considered at the same time as transformations to simplify relationships. The power family of transformations on the *dependent* variable is considered in Section 12.4, where the criteria are to have $\mathcal{E}(\boldsymbol{Y})$ adequately represented by a relatively simple model *and* the assumptions of normality and constant variance approximately satisfied (Box and Cox, 1964).

Two-Bend Transformations

Relationships that show more than one bend, such as the classical S-shaped growth curve (see the logistic curve in Figure 12.2), cannot be straightened with the power family of transformations. A few commonly used two-bend transformations are:

1. logit: $Y^* = \frac{1}{2}\log[p/(1-p)]$,
2. arcsin (or angular): $Y^* = \arcsin(\sqrt{p})$,
3. probit: $Y^* = \Phi^{-1}(p)$, where $\Phi^{-1}(p)$ is the standard normal deviate that gives a cumulative probability of p.

These transformations are generally applied to situations where the variable p is the proportion of "successes" and consequently bounded by 0 and 1. The effect of the transformation in all three cases is to "stretch" the upper and lower tails, the values of p near one and zero, making the relationship more nearly linear (Bartlett, 1947). The logit is sometimes preferred as a means of simplifying a model that involves products of probabilities. The probit transformation arises as the logical transformation when, for example, the chance of survival of an organism to a toxic substance is related to the dose, or ln(dose), of the toxin through a normal probability distribution of sensitivities. That is, individuals in the population vary in their sensitivities to the toxin and the threshold dose (perhaps on the logarithmic scale) that "kills" individuals has a normal distribution. In such case, the probit transformation translates the proportion affected into a linear relationship with dose, or ln(dose). The logit transformation has a similar interpretation but where the threshold distribution is the logistic distribution.

Intrinsically Linear Models

Nonlinear models that can be linearized are called **intrinsically linear**. The function $Y = \alpha X^{\beta}$ in Section 12.1 was linearized by taking the logarithm of both Y and X. If a positive multiplicative random error is incorporated to make it a statistical model, the model becomes

$$Y_i = \alpha X_i^{\beta} \epsilon_i. \tag{12.2}$$

The linearized form of this model is

$$\ln(Y_i) \quad = \quad \ln(\alpha) + \beta[\ln(X_i)] + \ln(\epsilon_i)$$

or

$$Y_i^* \quad = \quad \alpha^* + \beta X_i^* + \epsilon_i^*, \tag{12.3}$$

where $\alpha^* = \ln(\alpha)$, $X_i^* = \ln(X_i)$, and $\epsilon_i^* = \ln(\epsilon_i)$. This transformation is repeated here to emphasize the impact of the transformation of Y on the random errors. The least squares model assumes that the random errors are *additive*. Thus, in order for the random error to be additive on the log scale, they must have been *multiplicative* on the original scale. Furthermore, the ordinary least squares assumptions of normality and homogeneous variances apply to the $\epsilon_i^* = \ln(\epsilon_i)$, not to the ϵ_i. The implication is that linearization of models, and transformations in general, must also take into account the least squares assumptions. It may be better in some cases, for example, to forgo linearization of a model if the transformation destroys normality or homogeneous variances. Likewise, it may not be desirable to go to extreme lengths to achieve normality or homogeneous variances if it entails the use of an excessively complicated model.

Exponential Growth Model

Another example of an intrinsically linear model is the **exponential growth model**,

$$Y_i \quad = \quad \alpha e^{\beta X_i} \epsilon_i. \tag{12.4}$$

This growth function starts at $Y_i = \alpha$ when $X = 0$ and increases exponentially with a *relative* rate of growth equal to β ($\alpha > 0,\ \beta > 0$). The **exponential decay model** has the same form but with a negative exponential term. The decay model starts at $Y_i = \alpha$ when $X = 0$ and declines at a relative rate equal to β. The two exponential functions are illustrated in Figure 12.1. Both are linearized with the logarithmic transformation. Thus, for the growth model,

$$Y_i^* = \alpha^* + \beta X_i + \epsilon_i^*,$$

where Y_i^*, α, and ϵ_i^* are the natural logarithms of the corresponding quantities in the original model.

One version of the **inverse polynomial model** has the form

Inverse Polynomial Model

$$Y_i = \frac{X_i}{\alpha + \beta X_i + \epsilon_i}. \tag{12.5}$$

This function, illustrated in Figure 12.2, is a monotonically increasing function of X that very slowly approaches the asymptote $Y = 1/\beta$. The reciprocal transformation on Y, $Y^* = 1/Y$, gives

$$Y_i^* = \beta + \alpha\left(\frac{1}{X_i}\right) + \epsilon_i^*.$$

Thus, Y^* is a first-degree polynomial in $1/X$ with intercept β and slope α. Values of X equal to zero must be avoided for this transformation to work.

The frequently used **logistic growth model** is

Logistic Model

$$Y_i = \frac{\alpha}{1 + \gamma e^{-\beta X_i}\epsilon_i}. \tag{12.6}$$

This function gives the characteristic growth curve starting at $Y = \alpha/(1+\gamma)$ at $X = 0$ and asymptoting to $Y = \alpha$ as X gets large (Figure 12.2). The function is intrinsically linear only if the value of α is known, as is the case, for example, when the dependent variable is the proportion of individuals showing reaction to a treatment. If α is known, the model is linearized by defining

$$Y^* = \ln\left(\frac{\alpha}{Y} - 1\right)$$

and the model becomes

$$Y_i^* = \gamma^* - \beta X_i + \epsilon_i^*,$$

where $\gamma^* = \ln(\gamma)$ and $\epsilon_i^* = \ln(\epsilon_i)$.

In these examples, the placement of the error in the original model was such that the transformed model had an *additive* error. If there were reason to believe that the errors were additive in the original models, all would have become intrinsically nonlinear. The least squares assumptions on the behavior of the errors applies to the errors *after* transformation. Decisions as to how the errors should be incorporated into the models will depend on one's best judgment as to how the system operates and the analysis of the behavior of the residuals before and after transformation.

Approximating Functions with Polynomials

Any mathematical function relating Y to one or more independent variables can be approximated to any degree of precision desired with an appropriate polynomial in the independent variables. This is the fundamental reason polynomial models have proven so useful in regression, although seldom would one expect a polynomial model to be the true model for a physical, chemical, or biological process. Even intrinsically nonlinear models can be simplified, if need be, in the sense that they can be approximated

with polynomial models, which are linear in the parameters. (Some caution is needed in using a polynomial to approximate a nonlinear response that has an asymptote. The polynomial will tend to oscillate about the asymptote and eventually diverge.) The regression coefficients in the polynomial model will usually be nonlinear functions of the original parameters. This will make it more difficult to extract the physical meaning from the polynomial model than from the original nonlinear model. Nevertheless, polynomial models will continue to serve as very useful approximations, at least over limited regions of the X-space, of the more complicated, and usually unknown, true models.

12.3 Transformations to Stabilize Variances

Links Between Mean and Variance

The variance and the mean are independent in the normal probability distribution. All other common distributions have a direct link between the mean and the variance. For example, the variance is equal to the mean in the Poisson distribution, the distribution frequently associated with count data. The plot of the Poisson variance against the mean would be a straight line with a slope of one. The variance of the count of a binomially distributed random variable is $np(1-p)$ and the mean is np. The plot of the binomial variance against the mean would show zero variance at $p = 0$ and $p = 1$ and maximum variance at $p = 1/2$. The variance of a chi-square distributed random variable is equal to twice its mean. As with the Poisson, this is a linear relationship between the variance and the mean but with a steeper slope. A priori, one should expect variances to be heterogeneous when the random variable is not normally distributed.

Even in cases where there is no obvious reason to suspect nonnormality, there often is an association between the mean and the variance. Most commonly, the variance increases as the mean increases. It is prudent to suspect heterogeneous variances if the data for the dependent variable cover a wide range, such as a doubling or more in value between the smallest and largest observations.

General Transformation to Stabilize Variance

If the functional relationship between the variance and the mean is known, a transformation exists that will make the variance (approximately) constant (Bartlett, 1947). Let

$$\sigma^2 = \Omega(\mu),$$

where $\Omega(\mu)$ is the function of the mean μ that gives the variance. Let $f(\mu)$ be the transformation needed to stabilize the variance. Then $f(\mu)$ is the indefinite integral

$$f(\mu) = \int \frac{1}{[\Omega(\mu)]^{1/2}}\, d\mu.$$

(See Exercise 12.21.)

Example 12.2

For example, if σ^2 is proportional to μ, $\sigma^2 = c\mu$ as in the case of a Poisson random variable,

$$f(\mu) = \int \frac{1}{[(c\mu)]^{1/2}}\, d\mu = 2c^{-1/2}\sqrt{\mu}.$$

Thus, except for a proportionality constant and the constant of integration, the square-root transformation on the dependent variable would stabilize the variance in this case. ■

Variance Proportional to Power of Mean

In general, if the variance is (approximately) proportional to μ^{2k}, the appropriate transformation to stabilize the variance is $Y^* = Y^{1-k}$. (See Exercise 12.22.) In the Poisson example, $k = \frac{1}{2}$. When $k = 1$, the variance is proportional to the square of the mean and the logarithmic transformation is appropriate; Y^0 is interpreted as the logarithmic transformation. When the relationship between the mean and the variance is not known, empirical results can be used to approximate the relationship and suggest a transformation.

When the variance is proportional to a power of the mean, the transformation to stabilize the variance is a power transformation on the dependent variable—the same family of transformations used for "straightening" one-bend relationships. Thus, a possible course of action is to use a power transformation on the *dependent* variable to stabilize the variance and another power transformation on the *independent* variable to "straighten" the relationship.

Arcsin Transformation

The variance may not be proportional to a power of the mean. A binomially distributed random variable, for example, has maximum variance at $p = \frac{1}{2}$ with decreasing variance as p goes toward either zero or one, $\sigma^2(\widehat{p}) = p(1-p)/n$. The transformation that approximately stabilizes the variance is the arcsin transformation, $Y^* = \arcsin(\sqrt{\widehat{p}}) = \sin^{-1}\sqrt{\widehat{p}}$. See Exercise 12.22. This assumes that the number of Bernoulli trials in each $\widehat{p}_i$ is constant. Although the arcsin transformation is designed for binomial data, it seems to stabilize the variance sufficiently in many cases where the variance is not entirely binomial in origin.

The arcsin transformation is the only one of the three two-bend transformations given in Section 12.2 that also stabilizes the variance (if the data are binomially distributed). The other two, the logit and the probit, although they are generally applied to binomial data, will not stabilize the variance.

A word of caution is in order regarding transformation of proportional data. Not all such data are binomially distributed, and therefore they should not be automatically subjected to the arcsin transformation. For example, chemical proportions that vary over a relatively narrow range, such as the oil content in soybeans, may be very nearly normally distributed with constant variance.

12.4 Transformations to Improve Normality

Transformations to improve normality have generally been given lower priority than transformations to simplify relationships or stabilize variance. Even though least squares estimation per se does not require normality and moderate departures from normality are known not to be serious (Bartlett, 1947), there are sufficient reasons to be concerned about normality (see Section 10.2).

Fortunately, transformations to stabilize variance often have the effect of also improving normality. The logit, arcsin, and probit transformations that are used to stabilize variance and straighten relationships also make the distribution more normal-like by "stretching" the tails of the distribution, values near zero or one, to give a more bell-shaped distribution. Likewise, the power family of transformations, which have been discussed for straightening one-bend relationships and stabilizing variance, are also useful for increasing symmetry (decreasing skewness) of the distribution. The expectation is that the distribution will also be more nearly normal. The different criteria for deciding which transformation to make will not necessarily lead to the same choice, but it often happens that the optimum transformation for one will improve the other.

Box–Cox Method

Box and Cox (1964) present a computational method for determining a power transformation for the *dependent* variable where the objective is to obtain a simple, normal, linear model that satisfies the usual least squares assumptions. The Box–Cox criterion combines the objectives of the previous sections—simple relationship and homogeneous variance—with the objective of improving normality. The method is presented in this section because it is the only approach that directly addresses normality. The Box–Cox method results in estimates of the power transformation (λ), σ^2, and $\boldsymbol{\beta}$ that make the distribution of the transformed data as close to normal as possible [at least in large samples and as measured by the Kullback–Leibler information number (Hernandez and Johnson, 1980)]. However, normality is not guaranteed to result from the Box–Cox transformation and all the usual precautions should be taken to check the validity of the model.

The Box–Cox method uses the parametric family of transformations defined, in standardized form, as

$$Y_i^{(\lambda)} = \begin{cases} \dfrac{Y_i^{\lambda}-1}{\lambda(\dot{Y})^{(\lambda-1)}} & \text{for } \lambda \neq 0 \\ \dot{Y}\ln(Y_i) & \text{for } \lambda = 0, \end{cases} \tag{12.7}$$

where $\dot{Y}$ is the geometric mean of the original observations,

$$\dot{Y} = exp\sum[\ln(Y_i)]/n.$$

The method assumes that for some λ the $Y_i^{(\lambda)}$ satisfy all the normal-theory assumptions of least squares; that is, they are independently and normally

distributed with mean $\boldsymbol{X\beta}$ and common variance σ^2. With these assumptions, the maximum likelihood estimates of λ, $\boldsymbol{\beta}$, and σ^2 are obtained. [Hernandez and Johnson (1980) point out that this is not a valid likelihood because $Y_i^{(\lambda)}$ cannot be normal except in the special case of the original distribution being log-normal. Nevertheless, the Box–Cox method has proven to be useful.]

Estimating λ

The maximum likelihood solution is obtained by doing the least squares analysis on the transformed data for several choices of λ from, say $\lambda = -1$ to 1. Let SS[Res(λ)] be the residual sum of squares from fitting the model to $Y_i^{(\lambda)}$ for the given choice of λ and let $\sigma^2(\lambda) = \{SS[Res(\lambda)]\}/n$. The likelihood for each choice of λ is given by

$$L_{max} \quad = \quad -\frac{1}{2}\ln[\widehat{\sigma}^2(\lambda)]. \tag{12.8}$$

Maximizing the likelihood is equivalent to minimizing the residual sum of squares. The maximum likelihood solution for $\widehat{\lambda}$, then, is obtained by plotting SS[Res(λ)] against λ and reading off the value where the minimum, SS[Res(λ)]$_{min}$, is reached. It is unlikely that the exact power transformation defined by $\widehat{\lambda}$ will be used. It is more common to use one of the standard power transformations, $\lambda = \frac{1}{2}$, 0,$-\frac{1}{2}$, -1, in the vicinity of $\widehat{\lambda}$.

Confidence Intervals on λ

Approximate confidence intervals on λ can be determined by drawing a horizontal line on the graph at

$$SS[Res(\lambda)]_{min}\left(1 + \frac{t^2_{(\alpha/2,\nu)}}{\nu}\right), \tag{12.9}$$

where ν is the degrees of freedom for SS[Res(λ)]$_{min}$ and $t_{(\alpha/2,\nu)}$ is the critical value of Student's t with $\alpha/2$ probability in each tail. Confidence limits on λ are given as the values of λ where the horizontal line intersects the SS[Res(λ)] curve (Box, Hunter, and Hunter, 1978).

Considerations Before Using Box–Cox

The functional relationship between Y and the independent variables is specified in $\boldsymbol{X\beta}$ before the maximum likelihood estimate of λ is obtained. Thus, the solution obtained, $\widehat{\lambda}$, is conditional on, and can be sensitive to, the presumed form of the model (Cook and Wang, 1983). The Box–Cox method is attempting to simultaneously satisfy the three objectives, $\mathcal{E}(Y^{(\lambda)}) = \boldsymbol{X\beta}$, constant variance, and normality. The relative weights given to satisfying the three objectives will depend on which will yield the greatest impact on the likelihood function. For example, if $\boldsymbol{X\beta}$ specifies a linear relationship between $Y^{(\lambda)}$ and X when the observed relationship between Y and X is very curvilinear, it is likely that pressure to "straighten" the relationship will dominate the solution. The transformed data can be even more nonnormal and their variances more heterogeneous.

If emphasis is to be placed on improving normality or constancy of variance, the functional form of the model specified by $\boldsymbol{X\beta}$ should be flexible

enough to provide a reasonable fit to a range of transformations, including no transformation. For example, suppose the data show a curvilinear relationship that could be straightened with an appropriate power transformation. Specifying $\boldsymbol{X\beta}$ as a linear model would force the Box–Cox transformation to try to straighten the relationship. On the other hand, a quadratic model for $\boldsymbol{X\beta}$ would reduce the pressure to straighten the relationship and allow more pressure on improving normality and constancy of variance. Box and Cox (1964) show how to partition the effects of simple model, constant variance, and normality on the likelihood estimate of λ.

Example 12.3

The following example of a Box–Cox transformation is from a combined analysis of residuals from four studies on the effects of ozone and sulfur dioxide on soybean yields.[1] Each of the studies was subjected to the appropriate analysis of variance for the experimental design for that year. The observed residuals were pooled for checking model assumptions. There were a total of 174 residuals and 80 degrees of freedom for the pooled residual sum of squares.

Plots of the residuals suggested an increase in variance associated with increased yield (Figure 12.3). The normal plot of residuals was only slightly S-shaped with suggestive slightly heavy tails, but not sufficiently nonnormal to give concern. The Box–Cox standardized transformation, Equation 12.7, was applied for $\lambda = -1,\ -\frac{1}{2},\ 0,\ \frac{1}{2},\ 1$, and the analyses of variance repeated for each λ. The plot of the pooled residual sum of squares against λ, Figure 12.4 (page 412), suggested $\widehat{\lambda} = -.05$ with 95% confidence limits of approximately $-.55$ to .40. The confidence limits on λ overlap both $\lambda = 0$ and $\lambda = -.5$ but, since $\widehat{\lambda}$ was much nearer 0 than .5, the logarithmic transformation was adopted. The plot of the residuals of the log-transformed data showed no remaining trace of heterogeneous variance or nonnormality (Figure 12.5) and the normal plot of the residuals was noticeably straighter. ■

12.5 Generalized Least Squares

There will be cases where it is necessary, or at least deemed desirable, to use a dependent variable that does not satisfy the assumption of homogeneous variances. The transformation required to stabilize the variances may not be desirable because it destroys a good relationship between Y and X, or it destroys the additivity and normal distribution of the residuals.

[1] Analyses by V. M. Lesser on data courtesy of A. S. Heagle, North Carolina State University.

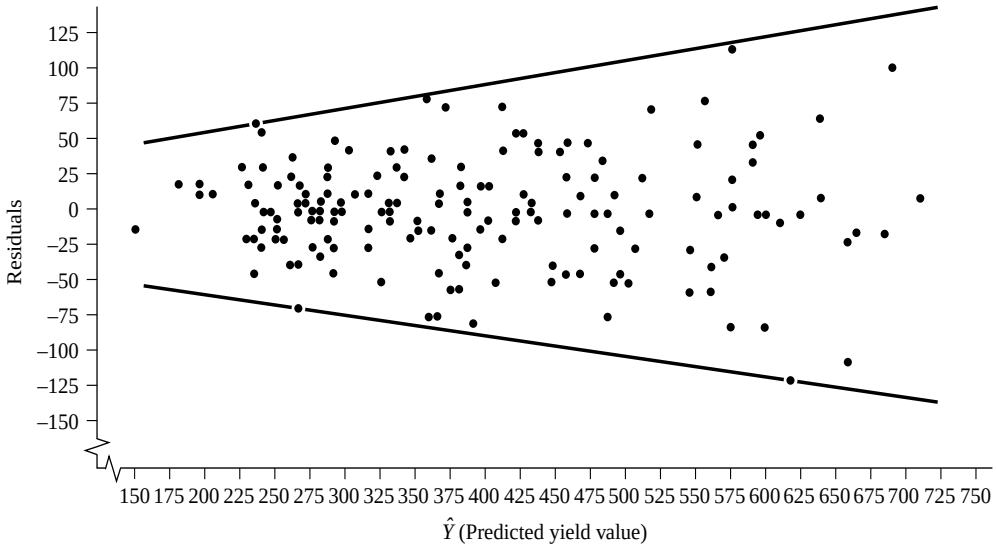

FIGURE 12.3. *Plot of e_i versus $\widehat{Y}_i$ (untransformed) from the combined analysis of four experiments on the effects of ozone and sulfur dioxide on soybean yields.*

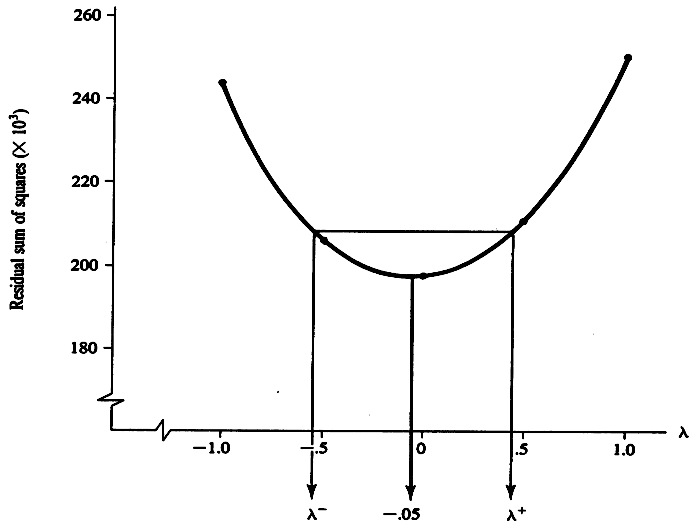

FIGURE 12.4. *Residual sum of squares plotted against λ for the Box–Cox transformation in the soybean experiments. The upper and lower limits of the approximate 95% confidence interval estimate of λ are shown by λ^- and λ^+, respectively.*

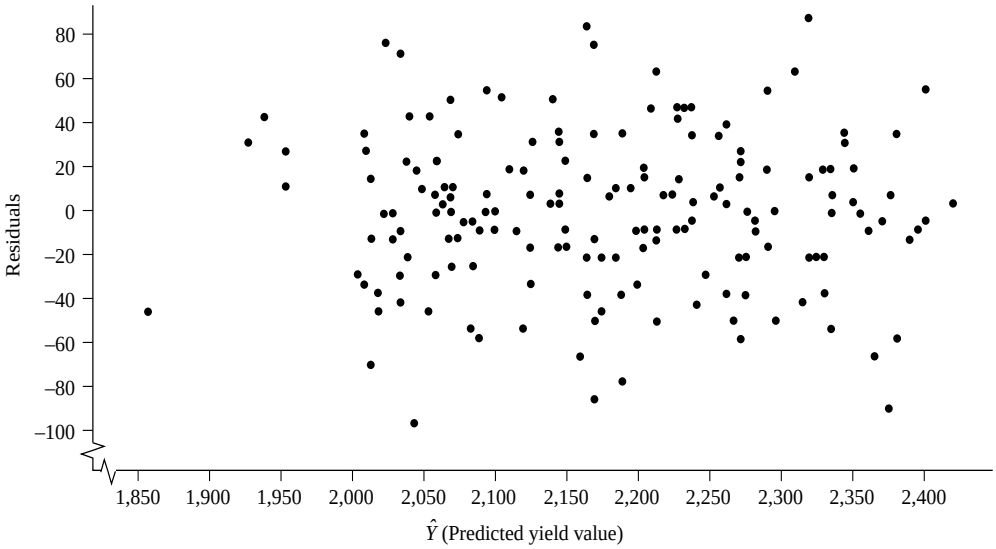

FIGURE 12.5. *Plot of e_i versus $\widehat{Y}_i$, after the logarithmic transformation, from the combined analysis of four experiments on the effects of ozone and sulfur dioxide on soybean yields.*

It may be that no transformation adequately stabilized the variances, or a transformation made to simplify a relationship left heterogeneous variances. The logit and probit transformations, for example, do not stabilize the variances. The arcsin transformation of binomial proportions will stabilize the variances only if the sample sizes n_i are equal. Otherwise, the variances will be proportional to $1/n_i$ and remain unequal after transformation. If treatment means are based on unequal numbers of observations, the variances will differ even if the original observations had homogeneous variances. Analysis on the original scale is preferred in such cases.

Weighted Versus Generalized Least Squares

Ordinary least squares estimation does not provide minimum variance estimates of the parameters when $\mathbf{Var}(\boldsymbol{\epsilon}) \neq \boldsymbol{I}\sigma^2$. This section presents the estimation procedure that does provide minimum variance linear unbiased estimates when the variance–covariance matrix of the errors is an arbitrary symmetric positive definite matrix $\mathbf{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$. This procedure is considered in two steps although the same principle is involved in both. First, the case is considered where the ϵ_i have unequal variances but are independent; $\sigma^2\boldsymbol{V}$ is a diagonal matrix of the unequal variances. Secondly, the general case is considered where, in addition to heterogeneous variances, the errors are not independent. Convention labels the first case **weighted least squares** and the second more general case **generalized least squares**.

12.5.1 Weighted Least Squares

The linear model is assumed to be

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{12.10}$$

with

$$\begin{aligned} \mathbf{Var}(\boldsymbol{\epsilon}) &= \boldsymbol{V}\sigma^2 \\ &= \text{Diag}\,(\, a_1^2 \quad a_2^2 \quad \cdots \quad a_n^2\,)\,\sigma^2. \end{aligned}$$

The variance of ϵ_i and Y_i is $a_i^2\sigma^2$, and all covariances are zero.

General Principle

The variance of a random variable is changed when the random variable is multiplied by a constant:

$$\begin{aligned} \sigma^2(cZ) &= \text{Var}(cZ) = c^2\text{Var}(Z) \\ &= c^2[\sigma^2(Z)], \end{aligned} \tag{12.11}$$

where c is a constant. If the constant is chosen to be proportional to the reciprocal of the standard deviation of Z, $c = k/\sigma(Z)$, the variance of the rescaled variable is k^2:

$$\sigma^2(cZ) = \left(\frac{k}{\sigma(Z)}\right)^2 \sigma^2(Z) = k^2. \tag{12.12}$$

Thus, if each observation in $\boldsymbol{Y}$ is divided by the proportionality factors a_i, the rescaled dependent variables will have equal variances σ^2 and ordinary least squares can be applied.

This is the principle followed in weighted least squares. The dependent variable is rescaled such that $\boldsymbol{V} = \boldsymbol{I}$ after rescaling. Then ordinary least squares is applied to the rescaled variables. (The same principle is used in generalized least squares although the weighting is more complicated.) This rescaling gives weight to each observation proportional to the reciprocal of its standard deviation. The points with the greater precision (smaller standard deviation) receive the greater weight.

Consider, for example, the model

$$Y_i = 1\beta_0 + X_{i1}\beta_1 + \cdots + X_{ip}\beta_p + \epsilon_i, \tag{12.13}$$

where the ϵ_i are uncorrelated random variables with mean zero. Suppose the variance of ϵ_i is $a_i^2\sigma^2$. Then, consider the rescaled model

$$\frac{Y_i}{a_i} = \left(\frac{1}{a_i}\right)\beta_0 + \left(\frac{1}{a_i}X_{i1}\right)\beta_1 + \cdots + \left(\frac{1}{a_i}X_{ip}\right)\beta_p + \frac{\epsilon_i}{a_i}$$

or

$$Y_i^* = X_{i0}^*\beta_0 + X_{i1}^*\beta_1 + \cdots + X_{ip}^*\beta_p + \epsilon_i^*. \tag{12.14}$$

Notice that ϵ_i^* in equation 12.14 have constant variance σ^2. In fact, the ϵ_i^*s are uncorrelated (0, σ^2) random variables. Therefore, we can obtain the best linear (in Y_i^*) unbiased estimators of $\beta_0, \ldots, \beta_p$ by using ordinary least squares regression of Y_i^* on $X_{i0}^*, \ldots, X_{ip}^*$. Since any linear function of Y_i^* is a linear function of Y_i (and vice versa), these are also the best linear (in Y_i) unbiased estimates of $\beta_0, \ldots, \beta_p$ in equation 12.13.

Matrix Formulation

The matrix formulation of weighted regression is as follows. Define the matrix $\boldsymbol{V}^{1/2}$ to be the diagonal matrix consisting of the square roots of the diagonal elements of $\boldsymbol{V}$, so that $\boldsymbol{V}^{1/2}\boldsymbol{V}^{1/2} = \boldsymbol{V}$. The weighting matrix $\boldsymbol{W}$ that rescales $\boldsymbol{Y}$ to have common variances is

$$\begin{aligned} \boldsymbol{W} &= (\boldsymbol{V}^{1/2})^{-1} \\ &= \begin{bmatrix} 1/a_1 & 0 & \cdots & 0 \\ 0 & 1/a_2 & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/a_n \end{bmatrix}, \end{aligned} \tag{12.15}$$

where the a_i are constants which reflect the proportional differences in the variances of ϵ_i. Notice that $\boldsymbol{W}\boldsymbol{W} = \boldsymbol{V}^{-1}$. Premultiplying both sides of the model by $\boldsymbol{W}$ gives

$$\boldsymbol{W}\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{\epsilon} \tag{12.16}$$

or

$$\boldsymbol{Y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*, \tag{12.17}$$

where $\boldsymbol{Y}^* = \boldsymbol{W}\boldsymbol{Y}$, $\boldsymbol{X}^* = \boldsymbol{W}\boldsymbol{X}$, and $\epsilon^* = \boldsymbol{W}\boldsymbol{\epsilon}$. The variance of $\boldsymbol{\epsilon}^*$ is, from the variances of linear functions,

$$\mathbf{Var}(\boldsymbol{\epsilon}^*) = \boldsymbol{W}[\mathbf{Var}(\boldsymbol{\epsilon})]\boldsymbol{W}' = \boldsymbol{W}\boldsymbol{V}\boldsymbol{W}\sigma^2 = \boldsymbol{I}\sigma^2, \tag{12.18}$$

since $\boldsymbol{W}\boldsymbol{V}\boldsymbol{W} = (\boldsymbol{V}^{1/2})^{-1}\boldsymbol{V}^{1/2}\boldsymbol{V}^{1/2}(\boldsymbol{V}^{1/2})^{-1} = \boldsymbol{I}$. The usual assumption of equal variances is met and ordinary least squares can be used on $\boldsymbol{Y}^*$ and $\boldsymbol{X}^*$ to estimate $\boldsymbol{\beta}$.

The weighted least squares estimate of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}_W$

$$\widehat{\boldsymbol{\beta}}_W = (\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{Y}^* \tag{12.19}$$

or, expressed in terms of the original $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_W &= (\boldsymbol{X}'\boldsymbol{W}'\boldsymbol{W}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{W}'\boldsymbol{W}\boldsymbol{Y}) \\ &= (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}). \end{aligned} \tag{12.20}$$

The variance of $\widehat{\boldsymbol{\beta}}_W$ is

$$\mathbf{Var}(\widehat{\boldsymbol{\beta}}_W) = (\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\sigma^2 = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\sigma^2. \tag{12.21}$$

Weighted least squares, which is equivalent to applying ordinary least squares to the transformed variables, finds the solution $\widehat{\boldsymbol{\beta}}_W$ that minimizes $\boldsymbol{e}^{*\prime}\boldsymbol{e}^* = \boldsymbol{e}'\boldsymbol{V}^{-1}\boldsymbol{e}$, not $\boldsymbol{e}'\boldsymbol{e}$. The analysis of variance of interest is the analysis of $\boldsymbol{Y}^*$. The fitted values $\widehat{\boldsymbol{Y}}^*$ and the residuals $\boldsymbol{e}^*$ on the transformed scale are the appropriate quantities to inspect for behavior of the model. Not all regression programs automatically provide the weighted residuals $\boldsymbol{e}^*$; BMDP does (Dixon, 1981). Usually, the regression results will be presented on the original scale so that some of the following results are given for both scales. The transformation between scales for the fitted values and for the residuals is the same as the original transformation between $\boldsymbol{Y}$ and $\boldsymbol{Y}^*$.

$\widehat{\boldsymbol{Y}}^*$ and $\widehat{\boldsymbol{Y}}_W$

The fitted values on the transformed scale are obtained by

$$\begin{aligned}\widehat{\boldsymbol{Y}}^* &= \boldsymbol{X}^*\widehat{\boldsymbol{\beta}}_W \\ &= \boldsymbol{X}^*(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{Y}^* = \boldsymbol{P}^*\boldsymbol{Y}^*, \end{aligned} \tag{12.22}$$

where $\boldsymbol{P}^*$ is the projection matrix for projecting $\boldsymbol{Y}^*$ onto the space defined by $\boldsymbol{X}^*$. The $\widehat{\boldsymbol{Y}}^*$ are transformed back to the original scale by

$$\widehat{\boldsymbol{Y}}_W = \boldsymbol{W}^{-1}\widehat{\boldsymbol{Y}}^* = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_W. \tag{12.23}$$

Their respective variances are

$$\mathbf{Var}(\widehat{\boldsymbol{Y}}^*) = \boldsymbol{X}^*(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{*\prime}\sigma^2 = \boldsymbol{P}^*\sigma^2 \tag{12.24}$$

and

$$\mathbf{Var}(\widehat{\boldsymbol{Y}}_W) = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2. \tag{12.25}$$

$\boldsymbol{e}^*$ and $\boldsymbol{e}$

The observed residuals are $\boldsymbol{e}^* = \boldsymbol{Y}^* - \widehat{\boldsymbol{Y}}^*$ on the transformed scale and $\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}_W$ on the original scale. Note that $\boldsymbol{e} = \boldsymbol{W}^{-1}\boldsymbol{e}^*$. Their variances are

$$\begin{aligned}\mathbf{Var}(\boldsymbol{e}^*) &= [\boldsymbol{I} - \boldsymbol{X}^*(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{*\prime}]\sigma^2 \\ &= (\boldsymbol{I} - \boldsymbol{P}^*)\sigma^2 \end{aligned} \tag{12.26}$$

and

$$\mathbf{Var}(\boldsymbol{e}) = [\boldsymbol{V} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}']\sigma^2. \tag{12.27}$$

Note that the usual properties of ordinary least squares apply to the transformed variables $\boldsymbol{Y}^*$, $\boldsymbol{e}^*$, and $\boldsymbol{X}^*$.

Example 12.4

For illustration, suppose the dependent variable is a vector of treatment means with unequal numbers r_i of observations per mean. If the original

observations have equal variances, the means will have variances σ^2/r_i. Thus,

$$\boldsymbol{V}\sigma^2 = \begin{bmatrix} 1/r_1 & 0 & \cdots & 0 \\ 0 & 1/r_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/r_n \end{bmatrix}. \tag{12.28}$$

The weighting matrix that gives $\mathbf{Var}(\boldsymbol{\epsilon}^*) = \boldsymbol{I}\sigma^2$ is

$$\boldsymbol{W} = \begin{bmatrix} \sqrt{r_1} & 0 & \cdots & 0 \\ 0 & \sqrt{r_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{r_n} \end{bmatrix}. \tag{12.29}$$

See also Exercise 12.24. ■

Estimating the Weights

In Example 12.4, it is clear that the variances of the dependent variable will not be equal and what the weighting matrix should be. In other cases, the variances may not be known a priori and their relative sizes will have to be determined from the data. If true replicates were available in the data set, the different variances could be estimated from the variance among the replicates for each group. In the absence of true replication, one might estimate the variances by using "near" replicates, groups of observations having nearly the same level of the independent variable(s). The variances of the "near" replicates might be plotted against the means of the "near" replicates, from which the relationship between the variance and the mean might be deduced and used to approximate the variance for each Y_i.

Computer Programs

A weighted least squares procedure is available in most least squares computer programs. Care must be used to specify the appropriate weights for the specific program. The weights in PROC GLM and PROC REG (SAS Institute, Inc., 1989b), for example, must be specified as a column vector of the squares of the diagonal elements in $\boldsymbol{W}$.

12.5.2 Generalized Least Squares

Generalized least squares extends the usual linear model to allow for an arbitrary positive definite variance–covariance matrix of $\boldsymbol{\epsilon}$, $\mathbf{Var}(\boldsymbol{\epsilon}) = \boldsymbol{V}\sigma^2$. The diagonal elements need not be equal and the off-diagonal elements need not be zero. The positive definite condition ensures that it is a proper variance matrix; that is, any linear function of the observations will have a positive variance. As with weighted least squares, a linear transformation is made on Y such that the transformed model will satisfy the least squares assumption of $\mathbf{Var}(\boldsymbol{\epsilon}^*) = \boldsymbol{I}\sigma^2$.

Finding the Weighting Matrix

For any positive definite matrix $\boldsymbol{V}$ it is possible to find a nonsingular matrix $\boldsymbol{T}$ such that

$$\boldsymbol{T}\boldsymbol{T}' = \boldsymbol{V}. \tag{12.30}$$

For example, if we express $\boldsymbol{V}$ as $\boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}'$ where $\boldsymbol{Z}$ is the matrix of eigenvectors of $\boldsymbol{V}$ and $\boldsymbol{L}$ is a diagonal matrix of eigenvalues (see equation 2.18), then $\boldsymbol{T} = \boldsymbol{Z}\boldsymbol{L}^{1/2}\boldsymbol{Z}'$ satisfies equation 12.30. Note that $\boldsymbol{T}$ in equation 12.30 is not unique. If $\boldsymbol{T}$ satisfies equation 12.30, then $\boldsymbol{T}\boldsymbol{Q}$, where $\boldsymbol{Q}$ is an orthogonal matrix, also satisfies equation 12.30. Since $\boldsymbol{T}$ is nonsingular, it has an inverse $\boldsymbol{T}^{-1}$. Premultiplying the model by $\boldsymbol{T}^{-1}$ gives

$$\boldsymbol{Y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*, \tag{12.31}$$

where $\boldsymbol{Y}^* = \boldsymbol{T}^{-1}\boldsymbol{Y}$, $\boldsymbol{X}^* = \boldsymbol{T}^{-1}\boldsymbol{X}$, and $\boldsymbol{\epsilon}^* = \boldsymbol{T}^{-1}\boldsymbol{\epsilon}$. With this transformation,

$$\mathbf{Var}(\boldsymbol{\epsilon}^*) = \boldsymbol{T}^{-1}\boldsymbol{V}(\boldsymbol{T}^{-1})'\sigma^2 = \boldsymbol{I}\sigma^2. \tag{12.32}$$

Note that ordinary least squares estimation is again appropriate for $\boldsymbol{Y}^*$ and $\boldsymbol{X}^*$ in the model in equation 12.31, and is given by

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_G &= (\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{Y}^* \\
&= (\boldsymbol{X}'(\boldsymbol{T}^{-1})'\boldsymbol{T}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{T}^{-1})'\boldsymbol{T}^{-1}\boldsymbol{Y} \\
&= [\boldsymbol{X}'(\boldsymbol{T}\boldsymbol{T}')^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}'(\boldsymbol{T}\boldsymbol{T}')^{-1}\boldsymbol{Y} \\
&= [\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}.
\end{aligned} \tag{12.33}$$

Note that $\widehat{\boldsymbol{\beta}}_G$ is invariant to the choice of $\boldsymbol{T}$ that satisfies equation 12.30. That is, even though the transformed vector $\boldsymbol{Y}^*$ may be different for different choices of $\boldsymbol{T}$ satisfying equation 12.30, we get the same estimate of $\widehat{\boldsymbol{\beta}}_G$ for $\boldsymbol{\beta}$. Recall that $\widehat{\boldsymbol{\beta}}_G$ minimized $\boldsymbol{e}^{*\prime}\boldsymbol{e}^* = \boldsymbol{e}'\boldsymbol{V}^{-1}\boldsymbol{e}$ (see Exercise 12.25). $\widehat{\boldsymbol{\beta}}_G$ is called the **generalized least squares** estimate of $\boldsymbol{\beta}$. The variance of $\widehat{\boldsymbol{\beta}}_G$ is given by

$$\begin{aligned}
\mathbf{Var}(\widehat{\boldsymbol{\beta}}_G) &= (\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\sigma^2 \\
&= (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\sigma^2.
\end{aligned}$$

Note that weighted least squares is a special case of generalized least squares. If $\boldsymbol{V}$ is a diagonal matrix, the appropriate $\boldsymbol{T}^{-1}$ is $\boldsymbol{W}$ as defined in equation 12.15.

Many of the least squares regression computer programs are not designed to handle generalized least squares. It is always possible, however, to make the indicated transformations, equation 12.31, and use ordinary least squares, or to resort to a matrix algebra computer program to do generalized least squares.

Example 12.5

When variables, like weight and blood pressure, are measured on a single individual over time, we expect the observations to be correlated over time. Consider the model

$$Y_i = \beta_0 + \epsilon_i, \quad i = 1, \ldots, n \tag{12.34}$$

for my weight over n consecutive days. Clearly, we do not expect $\mathbf{Var}(\boldsymbol{\epsilon})$ to be $\boldsymbol{I}\sigma^2$. Rather, we anticipate the weight measurements to be correlated over time and, furthermore, we expect measurements on two consecutive days to be more highly correlated than measurements that are further apart in time. One of the models that is used to model such behavior is a first-order autoregressive model:

$$\epsilon_i = \rho\epsilon_{i-1} + \eta_i, \quad i = 1, \ldots, n, \tag{12.35}$$

where η_i are uncorrelated $(0,\ \sigma^2)$ random variables. Assuming that ϵ_1 has mean zero, variance $\sigma^2/(1-\rho^2)$, is independent of η_i for $i \geq 2$, and that $|\rho| < 1$, it can be shown that

$$\begin{aligned} \mathbf{Var}(\boldsymbol{\epsilon}) &= \frac{1}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}\sigma^2 \\ &= \boldsymbol{V}\sigma^2. \end{aligned} \tag{12.36}$$

See Fuller (1996). Also, it can be shown that

$$\boldsymbol{T}^{-1} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix} \tag{12.37}$$

is such that $\boldsymbol{T}^{-1}\boldsymbol{V}\boldsymbol{T}^{-1'} = \boldsymbol{I}$ and $\boldsymbol{T}\boldsymbol{T}' = \boldsymbol{V}$.

Therefore, model 12.34 is transformed by premultiplying by $\boldsymbol{T}^{-1}$ to give

$$\begin{aligned} \begin{pmatrix} Y_1^* \\ Y_2^* \\ \vdots \\ Y_n^* \end{pmatrix} &= \begin{pmatrix} \sqrt{1-\rho^2}Y_1 \\ Y_2 - \rho Y_1 \\ \vdots \\ Y_n - \rho Y_{n-1} \end{pmatrix} \\ &= \begin{pmatrix} \sqrt{1-\rho^2} \\ 1-\rho \\ \vdots \\ 1-\rho \end{pmatrix}\beta_0 + \begin{pmatrix} \epsilon_1^* \\ \eta_2 \\ \vdots \\ \eta_n \end{pmatrix} \\ &= \boldsymbol{X}^*\beta_0 + \boldsymbol{\epsilon}^*, \end{aligned} \tag{12.38}$$

and $\mathbf{Var}(\boldsymbol{\epsilon}^*) = \boldsymbol{I}\sigma^2$. The generalized least squares estimate of β_0 is obtained by regressing $\boldsymbol{Y}^*$ on $\boldsymbol{X}^*$ and is given by

$$\begin{aligned}\widehat{\beta}_{0,G} &= (\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{Y}^* \\ &= \frac{(1-\rho^2)Y_1 + (1-\rho)(Y_2 - \rho Y_1) + \cdots + (1-\rho)(Y_n - \rho Y_{n-1})}{(1-\rho^2) + (1-\rho)^2 + \cdots + (1-\rho)^2} \\ &= \frac{Y_1 + (1-\rho)[Y_2 + \cdots + Y_{n-1}] + Y_n}{1 + (1-\rho)(n-2) + 1}. \end{aligned} \tag{12.39}$$

The variance of $\widehat{\beta}_{0,G}$ is given by

$$\begin{aligned}Var(\widehat{\beta}_{0,G}) &= (\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\sigma^2 \\ &= \frac{\sigma^2}{(1-\rho^2) + (n-1)(1-\rho)^2}. \end{aligned} \tag{12.40}$$

For the model in equation 12.34, the ordinary least squares estimator of β_0 is

$$\begin{aligned}\widehat{\beta}_0 &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\ &= \overline{Y}. \end{aligned} \tag{12.41}$$

Note that, since $\mathbf{Var}(\boldsymbol{\epsilon}) \neq \boldsymbol{I}\sigma^2$, we have

$$\begin{aligned}\mathrm{Var}(\widehat{\beta}_0) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{Var}(\boldsymbol{\epsilon})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2 \\ &\neq (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2 = \sigma^2/n. \end{aligned} \tag{12.42}$$

Since $\boldsymbol{X}$ in equation 12.34 is a column of ones, $\mathrm{Var}(\widehat{\beta}_0)$ reduces to

$$\begin{aligned}\mathrm{Var}(\widehat{\beta}_0) &= \frac{\sigma^2}{n(1-\rho)^2}\left[1 - \frac{2\rho(1-\rho^n)}{n(1-\rho^2)}\right] \\ &\neq \frac{\sigma^2}{n}. \end{aligned} \tag{12.43}$$

It can be shown that $\mathrm{Var}(\widehat{\beta}_0) \geq \mathrm{Var}(\widehat{\beta}_{0,G})$ for all values of n and ρ. Table 12.2 gives a comparison of relative efficiency of $\widehat{\beta}_0$ for various values of n and ρ. [The relative efficiency of two estimates $\widehat{\theta}_1$ to $\widehat{\theta}_2$ is measured as the ratio of variances R.E. $= s^2(\widehat{\theta}_2)/s^2(\widehat{\theta}_1)$.]

From Table 12.2, we observe that the relative efficiency of the ordinary least squares estimator is small for large values of ρ. Also, as the sample size increases, generally the relative efficiency increases. In this example, it can be shown that, for any fixed ρ, the relative efficiency converges to one as the sample size n tends to infinity. For some regression models, the relative efficiency of the ordinary least squares estimates may be quite small

TABLE 12.2. *Relative efficiency of ordinary least squares estimator with respect to the generalized least squares estimator of β_0 in an AR(1) model.*

n	.1	.3	.5	.7	.9	.95
25	.999	.993	.978	.947	.897	.909
50	1.000	.996	.988	.968	.906	.887
75	1.000	.997	.992	.977	.923	.890
100	1.000	.998	.994	.982	.936	.899

compared to the generalized least squares estimates. For example, see page 715 of Fuller (1996). ■

In this example, we have assumed that the correlation ρ between two consecutive observations is *known*. However, in practice ρ is unknown. An estimate of ρ is given by the sample correlation of consecutive observations:

$$\widehat{\rho} = \frac{\sum_{i=2}^{n}(Y_t - \overline{Y})(Y_{t-1} - \overline{Y})}{\sqrt{\sum_{t=2}^{n}(Y_{t-1} - \overline{Y})^2}\sqrt{\sum_{t=2}^{n}(Y_t - \overline{Y})^2}}. \tag{12.44}$$

When ρ is unknown, it is common to replace ρ in the transformations given in equation 12.38 with $\widehat{\rho}$. The *estimated* generalized least squares estimate, $\widetilde{\beta}_{0,EG}$, obtained by replacing ρ with $\widehat{\rho}$ in equation 12.39,

$$\tilde{\beta}_{0,EG} = \frac{Y_1 + (1-\widehat{\rho})[Y_2 + \cdots + Y_{n-1}] + Y_n}{1 + (1-\widehat{\rho})(n-2) + 1}, \tag{12.45}$$

is not necessarily a better estimator than the ordinary least squares estimator $\widehat{\beta}_0$.

Warnings

We need to emphasize that one must be somewhat cautious in the use of generalized least squares. The point made relative to equation 12.45 that the estimated generalized least squares estimate is not necessarily a better estimator than the ordinary least squares estimator applies in general. As with weighted least squares, the sum of squares $\boldsymbol{e}^{*\prime}\boldsymbol{e}^{*}$ is minimized and $\widehat{\boldsymbol{\beta}}_G$ is the best linear unbiased estimator of $\boldsymbol{\beta}$ *if* $\boldsymbol{V}$ is known. In most cases, however, $\boldsymbol{V}$ is unknown and must be estimated from the data. When an estimate of $\boldsymbol{V}$ is used, the solution obtained, called the **estimated generalized least squares** estimate, is no longer the minimum variance solution. In the worst cases where there is limited information with which to estimate $\boldsymbol{V}$, the estimated generalized least squares estimators can have larger variances than the ordinary least squares estimators. (This comment also applies to weighted least squares, but there the estimation problem is much less difficult.) Furthermore, it is possible for the generalized least squares regression line, if plotted on the original scale, to "miss" the data. That is, all of the observed data points can fall on one side of the regression

line. The necessary condition for this to occur is sufficiently large positive off-diagonal elements in $\boldsymbol{V}$. This does not depend on whether $\boldsymbol{V}$ is known or estimated. Estimation of $\boldsymbol{V}$, however, will likely cause the problem to occur more frequently. Such a result is not a satisfactory solution to a regression problem even though it may be the best linear unbiased estimate (as it is when $\boldsymbol{V}$ is known). Plotting the data and the regression line on the original scale will make the user aware of any such results.

Example 12.6

The example used to illustrate weighted and generalized least squares comes from an effort to develop a prediction equation for tree diameter at 54 inches above the ground (DBH) based on data from diameters at various stump heights. The objective was to predict amount of timber illegally removed from a tract of land and DBH was one of the measurements needed. Diameter at 54 inches (DBH) and stump diameters (SD) at stump heights (SHt) of 2, 4, 6, 8, 10, and 12 inches above ground were measured on 100 standing trees in an adjacent, similar stand. The trees were grouped into 2-inch DBH classes. There were $n = 4$, 16, 42, 26, 9, and 3 trees in DBH classes 6, 8, 10, 12, 14, and 16 inches, respectively.

It was argued that the ratio of DBH to the stump diameter at a particular height should be a monotonically decreasing function approaching one as the stump height approached 54 inches. This relationship has the form of an exponential decay function but with much sharper curvature than the exponential function allows. These considerations led to a model in which the dependent variable was defined as

$$Y_{ijk} = [\ln(SD_{ijk}) - \ln(DBH_{ik})]$$

and the independent variable as

$$X_j = [54^c - (SHt_j)^c],$$

where i is the DBH class ($i = 1, \ldots, 6$); j is the stump height class $j = 1, \ldots, 6$); k is the tree within each DBH class ($k = 1, \ldots, n_i$); and $\ln(SD_{ijk})$ and $\ln(DBH_{ik})$ are the logarithms of stump diameters and DBH. The averages of Y_{ijk} over k for each DBH–stump height category are given in Table 12.3. The exponent c, applied to the stump heights, was used to straighten the relationship (on the logarithmic scale) and was chosen by finding the value $c = 0.1$ that minimized the residual sum of squares for the linear relationship. Thus, the model is

$$\overline{Y}_{ij.} = \beta X_j + \bar{\epsilon}_{ij.},$$

a no-intercept model, where the $\overline{Y}_{ij.}$ are the DBH–stump height cell means of Y_{ijk} given in Table 12.3. Thus, $\boldsymbol{Y}$ is a 36×1 vector of the six values of $\overline{Y}_{1j.}$ in the first row of Table 12.3 followed by the six values of $\overline{Y}_{2j.}$ in

TABLE 12.3. *Averages by DBH class of logarithms of the ratios of stump diameter to diameter at 54 inches of 100 pine trees* $\overline{Y}_{ij\cdot}$*, where* $Y_{ijk} = ln(SD_{ijk}) - ln(DBH_{ik})$*. The values for the independent variable* $X_j = 54^c - SHt^c$ *for* $c = .1$ *are shown in the last row. (Data from B. J. Rawlings, unpublished.)*

DBH	*No.*	*Stump Height (Inches Above Ground)*					
(in.)	*Trees*	2	4	6	8	10	12
6	4	.3435	.3435	.2715	.1438	.0719	.0719
8	16	.3143	.2687	.2548	.2294	.1674	.1534
10	42	.2998	.2514	.2083	.1733	.1463	.1209
12	26	.3097	.2705	.2409	.1998	.1790	.1466
14	9	.2121	.1859	.1597	.1449	.1039	.1039
16	3	.2549	.2549	.1880	.1529	.1529	.1529
X_j		.4184	.3415	.2940	.2590	.2313	.2081

the second row, and so on. The $\boldsymbol{X}$ vector consists of six repeats of the six values of X_j corresponding to the six stump heights.

It is not appropriate to assume $\mathbf{Var}(\boldsymbol{\epsilon}) = \boldsymbol{I}\sigma^2$ in this example for two reasons: the dependent variable consists of averages of differing numbers of trees within each *DBH* class ranging from $n = 3$ to $n = 42$; and all Y_{ijk} from the same tree (same i and j) are correlated due to the fact that DBH_{ij} is involved in the definition of Y_{ijk} in each case. Also, the stump diameters at different heights on the same tree are expected to be correlated. Observations in different *DBH* classes are independent since different trees are involved. It is assumed that the variance–covariance matrix of the observations within each *DBH* class is the same over *DBH* classes. Thus, the 36×36 variance–covariance matrix $\mathbf{Var}(\boldsymbol{\epsilon})$ will have the form

$$\mathbf{Var}(\boldsymbol{\epsilon}) = \begin{bmatrix} \boldsymbol{B}/4 & 0 & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{B}/16 & 0 & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{B}/42 & 0 & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{B}/26 & 0 & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{B}/9 & 0 \\ 0 & 0 & 0 & 0 & 0 & \boldsymbol{B}/3 \end{bmatrix}, \quad (12.46)$$

where $\boldsymbol{B}$ is the 6×6 variance–covariance matrix for Y_{ijk} from the same tree. That is, the diagonal elements of $\boldsymbol{B}$ are variances of Y_{ijk} for a given stump height and the off-diagonal elements are covariances between Y_{ijk} at two different stump heights for the same tree.

The estimate of $\boldsymbol{B}$ was obtained by defining 6 variables from the Y_{ijk}, one for each stump height (level of j). Thus, the matrix $\boldsymbol{Y}$ of data is 100×6 (there were 100 trees), with each column containing the measurements from one of the 6 stump heights. The variance–covariance matrix $\boldsymbol{B}$ was

estimated as

$$\begin{aligned}\widehat{\boldsymbol{B}} &= [\boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{J}/n)\boldsymbol{Y}]/99 \\ &= \begin{bmatrix} 86.2 & 57.2 & 63.0 & 53.9 & 48.9 & 52.5 \\ 57.2 & 71.4 & 59.5 & 45.2 & 35.0 & 39.3 \\ 63.0 & 59.5 & 100.2 & 73.8 & 51.8 & 50.6 \\ 53.9 & 45.2 & 73.8 & 97.3 & 62.9 & 53.7 \\ 48.9 & 35.0 & 51.8 & 62.9 & 76.5 & 59.3 \\ 52.5 & 39.3 & 50.6 & 53.7 & 59.3 & 78.6 \end{bmatrix} 10^{-4}, \quad (12.47)\end{aligned}$$

where $\boldsymbol{J}$ is a 100×100 matrix of ones and $n = 100$. The correlations in $\widehat{\boldsymbol{B}}$ range from .47 to .77. It is likely that the form of $\widehat{\boldsymbol{B}}$ could be simplified by assuming, for example, a common variance or equality of subsets of the correlations. This would improve the estimates of the weights if the simplications were justified. For this example, the general covariance matrix was used.

Generalized least squares was used to estimate β and its standard error. $\widehat{\boldsymbol{B}}$ was multiplied by (99×10^2), rounded to two digits, and then substituted for $\boldsymbol{B}$ in Equation 12.46 to give the weighting matrix for generalized least squares. The computations were done with IML (SAS Institute, Inc., 1989d) which is an interactive matrix program. The regression equation obtained was

$$\widehat{Y}_{ij} = .7277 X_j$$

with $s(\widehat{\beta}_{EG}) = .0270$, where $\widehat{\beta}_{EG}$ is the estimated generalized least squares estimate of β. The regression coefficient is significantly different from zero.

For comparison, the unweighted regression and the weighted regression using only the numbers of trees in the *DBH* classes as weights were also run. The resulting regression equations differed little from the generalized regression results but the computed variances of the estimates were very different. The computed results from the two regressions were as follows.

Unweighted:

$$\widehat{Y}_{ij} = .6977 X_j \qquad \text{with } s(\widehat{\beta}) = .0237.$$

Weighted by n_i:

$$\widehat{Y}_{ij} = .7147 X_j \qquad \text{with } s(\widehat{\beta}_W) = .0148.$$

Comparison of the standard errors appears to indicate a *loss* in precision from using generalized least squares. However, the variances computed by the standard regression formulae assumed that $\mathbf{Var}(\boldsymbol{\epsilon}) = \boldsymbol{I}\sigma^2$ in the unweighted case and $\mathbf{Var}(\boldsymbol{\epsilon}) = \text{Diag}(\{1/n_i\})\sigma^2$ in the weighted regression, neither of which is correct in this example.

The correct variance when ordinary least squares is used but where $\mathbf{Var}(\boldsymbol{\epsilon}) \neq \boldsymbol{I}\sigma^2$ is given by

$$\sigma^2(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'[\mathbf{Var}(\boldsymbol{\epsilon})]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{12.48}$$

When weighted least squares is used but with an incorrect weight matrix $\boldsymbol{W}$, the correct variance is given by

$$\sigma^2(\widehat{\boldsymbol{\beta}}_W) = (\boldsymbol{X}'\boldsymbol{W}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}'[\mathbf{Var}(\boldsymbol{\epsilon})]\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}'\boldsymbol{W}\boldsymbol{X})^{-1}. \tag{12.49}$$

When $\widehat{\boldsymbol{B}}$ (equation 12.47) is substituted in equation 12.46 to give an estimate of $\mathbf{Var}(\boldsymbol{\epsilon})$, equations 12.48 and 12.49 give estimates of the variances of the regression coefficients for the unweighted and weighted (by n_i) analyses. The resulting standard errors of $\widehat{\beta}$ are as follows.

Unweighted:

$$s(\widehat{\beta}) = .04850.$$

Weighted by n_i:

$$s(\widehat{\beta}_W) = .03215.$$

The efficiency of estimated generalized least squares relative to unweighted least squares and to weighting by n_i is 3.22 and 1.42, respectively, in this example. These relative efficiencies are biased in favor of generalized least squares regression since an estimated variance–covariance matrix has been used in place of the true variance–covariance matrix. Nevertheless, in this example they show major increases in precision that result from accounting for unequal variances and correlation structure in the data. Comparison of the standard errors computed from the unweighted analysis and the weighted analysis with the results of equations 12.47 and 12.48 illustrates the underestimation of variances that commonly occurs when positively correlated errors in the data are ignored. ■

Example 12.7

The following example illustrates another important problem related to correlated errors. If autocorrelation exists but is ignored, the computed standard errors will be incorrect.

Consider the model in Example 12.5 given by

$$Y_i = \beta_0 + \epsilon_i, \qquad i = 1, \ldots, n,$$

and

$$\epsilon_i = \rho\epsilon_{i-1} + \eta_i,$$

where $-1 < \rho < 1$ and $\eta_i \sim \text{NID}(0, \sigma^2)$. In this case, one might use the ordinary least squares estimator $\widehat{\beta}_0 = \overline{Y}$ for β_0 and mistakenly use s^2/n as

an estimator of the variance of $\overline{Y}$, where $s^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2/(n-1)$ is the residual mean square error. When $\rho \neq 0$, we have seen that the ordinary least squares estimate is inefficient, but the efficiency is close to 1 in large samples. A more serious problem is the estimate of variance of $\overline{Y}$. When n is large, we have seen that

$$\mathrm{Var}(\overline{Y}) \approx \frac{\sigma^2}{n(1-\rho)^2}.$$

Also, it can be shown that s^2 is not an unbiased estimate of σ^2, but is a very good estimate of $\mathrm{Var}(Y_i) = \sigma^2/(1-\rho^2)$. Therefore, s^2/n under (over) estimates $\mathrm{Var}(\overline{Y})$ by a factor of $(1-\rho)/(1+\rho)$, approximately, when $\rho > 0$ (< 0). For example, for $\rho = .8$ the ordinary least squares standard error $\sqrt{s^2/n}$ is expected to be only $\sqrt{(1-.8)/(1+.8)} = \frac{1}{3}$ of the true standard error.

12.6 Summary

The first sections of this chapter discussed transformations of the independent and dependent variables to make the model simpler in some sense, or to make the assumptions of homogeneous variance and normality more nearly satisfied. Transformations on the independent variable affect only the form of the model. Transformations to stabilize variances or to more nearly satisfy normality must be made on the dependent variable. The power family of transformations plays an important role in all three cases.

The ladder of transformations and the rules for determining the transformation are easily applied as long as the model is reasonably simple. In more complex cases, the Box–Tidwell method provides power transformations on the independent variables that give the best fit to a particular model; the result is dependent on the model chosen. The Box–Cox transformation provides a power transformation on the dependent variable with the more general criterion of satisfying all aspects of the distributional assumption on $\boldsymbol{Y}$; $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{X\beta},\ \boldsymbol{I}\sigma^2)$. The result and the relative emphasis the method gives to simplifying the model, stabilizing variance, and improving normality is dependent on the choice of $\boldsymbol{X\beta}$. In no case are we assured that the appropriate power transformation exists to satisfy all criteria. All precautions should be taken to verify the adequacy of the model and the least squares results.

The last section covered weighted least squares and generalized least squares methods. These methods address the specific situation where the scale of the dependent variable has already been decided but where the basic assumption of $\mathbf{Var}(\boldsymbol{\epsilon}) = \boldsymbol{I}\sigma^2$ is not satisfied. In such cases, the minimum variance estimators are obtained only if the true $\mathbf{Var}(\boldsymbol{\epsilon})$ is taken into account by using weighted least squares or generalized least squares, as the situation requires.

12.7 Exercises

12.1. This exercise uses Land's data on tolerance of certain families of pine to salt water flooding given in Table 12.1. For this exercise, replace Hours = 0 with 1 and $Y = .00$ in Family 3 with .01 to avoid problems with taking logarithms.

(a) Plot Y = chloride content against X = Hours. Summarize what the plot suggests about homogeneous variances, about normality, and about the type of response curve needed if no transformations are made.

(b) Use the plot of the data and the ladder of transformations to suggest a transformation on Y that might straighten the relationship. Suggest a transformation on X that might straighten the relationship. In view of your answer to Part (a), would you prefer the transformation on Y or on X?

(c) Assume a common quadratic relationship of $Y^{(\lambda)}$ with X for all families, but allow each family to have its own intercept. Use the Box–Cox transformation for $\lambda = 0,\ .2,\ .3,\ .4,\ .5,\ .7$, and 1.0 and plot the residual sum of squares in each case against λ. At what value of λ does the minimum residual sum of squares occur? Graphically determine 95% confidence limits on λ. What power transformation on Y do you choose?

(d) Repeat Part (c) using a *linear* relationship between $Y^{(\lambda)}$ and X. Show how this changes the Box–Cox results and explain (in words) why the results differ.

(e) Use the Box–Cox transformation adopted in Part (c) as the dependent variable. If $Y^{(\lambda)}$ is regressed on X using the quadratic model in Part (c), the quadratic term is highly significant. Use the Box–Tidwell method to find a power transformation on X that will straighten the relationship. Plot the residuals from the regression of $Y^{(\lambda)}$ on X^{α}, the Box–Tidwell transformation on X, against $\widehat{Y}$ and in a normal plot. Do you detect any problems?

12.2. The Land data given in Table 12.1 are percentage data. Are they binomially distributed data? Would you a priori expect the arcsin transformation to work?

12.3. A replicated corn yield trial (25 entries in 3 blocks) grown at five locations gave data in which the response variable (yield) varied from 55 bu/acre in a particularly dry location to 190 bu/acre in the most favorable environment. The mean yields and the experimental error variances (each with 48 degrees of freedom) for the five locations were as follows.

Mean Yield	*Error Variance*
55	68
105	139
131	129
148	325
190	375

Consider these options for handling the heterogeneous variances in a combined analysis of variance: (1) an appropriate transformation on Y and (2) weighted least squares.

(a) What transformation would you suggest from inspection of the relationship between the mean and the variance?

(b) Explain what your weighting matrix would be if you used weighted least squares. This will be a very large matrix. Explain how you could do the weighting without forming this matrix.

(c) A third option would be to ignore the heterogeneous variances and proceed with the combined analysis. Discuss the merits of the three alternatives and how you would decide which to use.

12.4. The monomolecular growth model has the form

$$Y = \alpha(1 - \beta e^{-kt}).$$

Is this model nonlinear in the parameters? Can it be linearized with an appropriate transformation on Y? Can it be linearized if α is known?

12.5. A dose response model based on the Weibull function can be written as

$$Y = \alpha\{\exp[-(X/\gamma)^{\delta}]\}.$$

Does taking the logarithm of Y linearize this model?

12.6. A nonlinear model for a chemical reaction rate can be formulated as

$$Y = \alpha X_1/(1 + \beta X_1 + \gamma X_2).$$

Does the reciprocal transformation on Y give a model that is linear in the parameters? Does a redefinition of the parameters make the model linear in the parameters?

12.7. The water runoff data in Exercise 5.1 were analyzed using $\ln(Q)$ where Q was the peak rate of flow. Use the Box–Cox method with a linear model containing the logarithms of all nine independent variables to determine the transformation on Q. Is $\lambda = 0$ within the 95% confidence interval estimate of λ ?

12.8. The following growth data (Y = dry weight in grams) were taken on four independent experimental units at each of six different ages (X = age in weeks).

	X (*Weeks of Age*)					
Item	1	2	3	5	7	9
1	8	35	57	68	76	85
2	10	38	63	76	95	98
3	12	42	68	86	103	105
4	15	48	74	90	105	110

(a) Plot Y versus X. Use the ladder of transformations to determine a power transformation on Y that will straighten the relationship. Determine a power transformation on X that will straighten the relationship.

(b) Use the Box–Tidwell method to determine a power transformation on X for the linear model. Does this differ from what you decided using the ladder of transformations? Is there any problem with the behavior of the residuals?

(c) Observe the nature of the dispersion of Y for each level of X. Does there appear to be any problem with respect to the least squares assumption of constant variance? Will either of your transformations in (a) improve the situation? (Do trial transformations on Y for the first, fourth, and sixth levels of X, ages 1, 5, and 9, and observe the change in the dispersion.)

12.9. Use the data in Exercise 12.8 and the Box–Cox method to arrive at a transformation on Y. Recall that the Box–Cox method assumes a particular model $\mathcal{E}(\boldsymbol{Y}) = \boldsymbol{X\beta}$. For this exercise, use $\mathcal{E}(Y_i) = \beta_0 + \beta_1 X_i$. Plot SS[Res($\lambda$)] versus λ, find the minimum, and determine approximate 95% confidence limits on λ. What choice of λ does the Box–Cox method suggest for this model? Fit the resulting regression equation, plot the transformed data and the regression equation, and observe the nature of the residuals. Does the transformation appear to be satisfactory with respect to the straight-line relationship? With respect to the assumption of constant variance? (*Note*: The purposes of Exercises 12.9 to 12.12 are, in addition to demonstrating the use of the Box–Cox transformation, to show the dependence of the method on the assumed model and to illustrate that obtaining the power transformation via the Box–Cox method does not guarantee either that the model fits or that the usual least squares assumptions are automatically satisfied.)

12.10. Repeat Exercise 12.9 using the quadratic polynomial model in X. Using this model, to which transformation does the Box–Cox method lead and does it appear satisfactory?

12.11. Repeat Exercise 12.9 using $X^* = \ln(X)$ in the linear model. What transformation do you obtain this time and is it satisfactory?

12.12. Repeat Exercise 12.11 but allow a quadratic model in $X^* = \ln(X)$. What transformation do you obtain and does it appear to be more satisfactory?

12.13. The corn borer survival data, number of larvae surviving 3, 6, 9, 12, and 21 days after inoculation, in Exercise 9.4 were analyzed without transformation. "Number of larvae" might be expected not to have homogeneous variance. Plot the residuals from the analysis of variance against $\widehat{Y}$. Do they provide any indication of a problem? Use the Box–Cox method to estimate a transformation for "number of larvae" where $\boldsymbol{X\beta}$ is defined for the analysis of variance model. Is a transformation suggested? If so, do the appropriate transformation and summarize the results.

12.14. Show that $\boldsymbol{P}^*$ in $\widehat{\boldsymbol{Y}}^* = \boldsymbol{P}^*\boldsymbol{Y}^*$, equation 12.22, is idempotent.

12.15. Use equation 12.23 to obtain the coefficient matrix on $\boldsymbol{Y}$, the original variable, that gives $\widehat{\boldsymbol{Y}}_W$. Show that this matrix is idempotent.

12.16. Use equation 12.23 to express $\widehat{\boldsymbol{Y}}'_W\widehat{\boldsymbol{Y}}_W$ as a quadratic function of $\boldsymbol{Y}$. Likewise, obtain $\boldsymbol{e}'\boldsymbol{e}$, where $\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}_W$, as a quadratic function of $\boldsymbol{Y}$. Show that:

(a) neither coefficient matrix is idempotent;

(b) the two coefficient matrices are not orthogonal.

What are the implications of these results?

12.17. Use the variance of linear functions to derive $\mathbf{Var}(\widehat{\boldsymbol{\beta}}_W)$, equation 12.21.

12.18. Use the variance of linear functions to derive $\mathbf{Var}(\widehat{\boldsymbol{Y}}^*)$, equation 12.24.

12.19. Derive $\mathbf{Var}(\widehat{\boldsymbol{\beta}})$ when ordinary least squares is used to estimate $\boldsymbol{\beta}$ but where $\mathbf{Var}(\boldsymbol{\epsilon}) \neq \boldsymbol{I}\sigma^2$, equation 12.48.

12.20. The data used in the generalized least squares analysis in the text to develop a model to relate *DBH* (diameter at breast height, 54 inches) to diameters at various stump heights, Example 12.6, are given in Table 12.3. The numbers in the table are $\overline{Y}_{ij.}$, where Y_{ijk} and X_j are defined in the text. The estimated variance–covariance matrix is shown in equation 12.47.

(a) Use a matrix computer program to do the generalized least squares analysis on these data as outlined in Example 12.6. Notice that the model contains a zero intercept. Give the regression equation, the standard error of the regression coefficient, and the analysis of variance summary. (Your answers may differ slightly from those in Example 12.6 unless the variance–covariance matrix is rounded as described.)

(b) It would appear reasonable to simplify the variance–covariance matrix, equation 12.46, by assuming homogeneous variances and common covariances. Average the appropriate elements of $\widehat{\boldsymbol{B}}$ to obtain a common variance and a common covariance. Redo the generalized regression with $\boldsymbol{B}$ redefined in this way. Compare the results with the results in (a) and the unweighted regression results given in Example 12.6.

12.21 Consider a random variable Y with mean μ and variance σ^2. Suppose $\sigma^2 = \Omega(\mu)$. Consider a transformation $f(Y)$ of Y to stabilize the variance. Using the first-order Taylor series approximation,

$$f(Y) \doteq f(\mu) + f'(\mu)(Y - \mu),$$

where $f'(\mu)$ is the first derivative of $f(\cdot)$ at μ. This suggests

$$\text{Var}[f(Y)] \doteq [f'(\mu)]^2 \Omega(\mu).$$

Show that if $f(\mu) = \int (1/[\Omega(\mu)]^{1/2})\, d\mu$, then the variance of $f(Y)$ is constant, approximately.

12.22. Consider $\Omega(\mu)$ and $f(\mu)$ defined in Exercise 12.21.

(a) Suppose $\Omega(\mu) = \mu^{2k}$; then show that $f(\mu)$ is proportional to μ^{1-k}, for $k \neq 1$.

(b) Suppose $\Omega(\mu) = \mu^2$; then show that $f(\mu)$ is proportional to $\ln(\mu)$.

(c) When do you use the inverse transformation? [That is, for what function $\Omega(\mu)$ is $f(\mu) = \mu^{-1}$?]

(d) If $\Omega(\mu) = \mu(1-\mu)$, show that $f(\mu)$ is proportional to $\sin^{-1}(\sqrt{\mu})$.

(e) If Y has a chi-square distribution with degrees of freedom ν, what transformation of Y would approximately stabilize the variance?

(f) Suppose Y corresponds to a sample variance s^2, based on n independent $N(\mu_0, \sigma_0^2)$ variables. What transformation would you recommend to stabilize the variance of $Y = s^2$?

12.23. Consider the Box–Cox transformation given in equation 12.7. Assume that the linear model includes an intercept term. For $\lambda \neq 0$, show that $\widehat{\sigma}^2(\lambda)$ obtained using $Y_i^{(\lambda)}$ is proportional to $\tilde{\sigma}^2(\lambda)$ obtained from Y_i^{λ}, where the proportionality constant is $1/[\lambda^2 \dot{Y}^{2(\lambda-1)}]$.

12.24. Consider the model

$$\text{Model (a): } Y_{ij} = \beta_0 + X_{i1}\beta_1 + \cdots + X_{ip}\beta_p + \epsilon_{ij}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, r_i$, where we have r_i replicates at each vector $(1, X_{i1}, \ldots, X_{ip})$ of independent variables. Assume that ϵ_{ij} are uncorrelated random variables with mean zero and variance σ^2. Consider also the model

$$\text{Model (b): } \overline{Y}_{i\cdot} = \beta_0 + X_{i1}\beta_1 + \cdots + X_{ip}\beta_p + \overline{\epsilon}_{i\cdot}, \qquad i = 1, \ldots, n,$$

where $\overline{Y}_{i\cdot}$ is the mean of the r_i replicates.

(a) Show that the weighted least squares estimator of $(\beta_0 \quad \beta_1 \quad \cdots \quad \beta_p)'$ in Model (b) is also the ordinary least squares estimator of $(\beta_0 \quad \beta_1 \quad \cdots \quad \beta_p)'$ in Model (a).

(b) Show also that they coincide with the ordinary least squares estimates in the rescaled model:

$$\sqrt{r_i}\overline{Y}_{i\cdot} = (\sqrt{r_i})\beta_0 + (\sqrt{r_i}X_{i1})\beta_1 + \cdots + (\sqrt{r_i}X_{ip})\beta_p + \sqrt{r_i}\overline{\epsilon}_{i\cdot},$$

for $i = 1, \ldots, n$.

12.25. Show that $\boldsymbol{\epsilon}^{*\prime}\boldsymbol{\epsilon}^* = \boldsymbol{\epsilon}'\boldsymbol{V}^{-1}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}^* = \boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}$ as given in equation 12.31.

12.26. Consider the model

$$Y_i = \beta_0 + \epsilon_i\ , \qquad i = 1, \ldots, n,$$

where $\epsilon_i = \epsilon_1^* + \epsilon_2^* + \cdots + \epsilon_i^*$ and ϵ_i^*s are uncorrelated $(0, \sigma^2)$ random variables.

(a) Find the ordinary least squares estimator $\widehat{\beta}_0$. Compute $\text{Var}(\widehat{\beta}_0)$.

(b) Find the appropriate transformation Y^*. (Hint: Consider $Y_i - Y_{i-1}$.)

(c) Find the generalized least squares estimator $\widehat{\beta}_{0,G}$ of $\widehat{\beta}_0$.

(d) Compute $\text{Var}(\widehat{\beta}_{0,G})$ and compare it with $\text{Var}(\widehat{\beta}_0)$.

13
COLLINEARITY

Chapters 10 through 12 have outlined the problem areas, discussed methods of detecting the problems, and discussed the use of transformations to alleviate the problems.

This chapter addresses the collinearity problem, with the emphasis on understanding the relationships among the independent variables rather than on the routine application of biased regression methods. Principal component analysis and Gabriel's biplots are used to explore the correlational structure. One of the biased regression methods, the principal component regression, is presented and its limitations are discussed.

Origins of Collinearity

The collinearity problem in regression arises when at least one linear function of the independent variables is very nearly equal to zero. (Technically, a set of vectors is collinear when a linear function is *exactly* equal to zero. In general discussions of the collinearity problem, the term "collinear" is often used to apply to linear functions that are only *approximately* zero. This convention is followed in this text.) This near-singularity may arise in several ways.

1. An inbuilt mathematical constraint on variables that forces them to add to a constant will generate a collinearity. For example, frequencies of alleles at a locus will add to one if the frequencies of all alleles are recorded, or nearly to one if a rare allele is not scored. Generating

new variables as transformations of other variables can produce a collinearity among the set of variables involved. Ratios of variables or powers of variables frequently will be nearly collinear with the original variables.

2. Component variables of a system may show near-linear dependencies because of the biological, physical, or chemical constraints of the system. Various measures of size of an organism will show dependencies as will amounts of chemicals in the same biological pathway, or measures of rainfall, temperature, and elevation in an environmental system. Such correlational structures are properties of the system and can be expected to be present in all observational data obtained from the system.

3. Inadequate sampling may generate data in which the near-linear dependencies are an artifact of the data collection process. Unusual circumstances also can cause unlikely correlations to exist in the data, correlations that may not be present in later samplings or samplings from other similar populations.

4. A bad experimental design may cause some model effects to be nearly completely confounded with others. This is the result of choosing levels of the experimental factors in such a way that there are near linear dependencies among the columns of X representing the different factors. Usually, experimental designs are constructed so as to ensure that the different treatment factors are orthogonal, or very nearly orthogonal, to each other.

One may not always be able to clearly identify the origin of the collinearity problem but it is important to understand its nature as much as possible. Knowing the nature of the collinearity problem will often suggest to the astute researcher its origin and, in turn, appropriate ways of handling the problem and of interpreting the regression results.

The first section of this chapter discusses methods of analyzing the correlational structure of the X-space with a view toward understanding the nature of the collinearity. The second section introduces biased regression as one of the classical methods of handling the collinearity problem. For all discussions in this chapter, the **matrix of centered and scaled independent variables** $\boldsymbol{Z}$ is used so that $\boldsymbol{Z}'\boldsymbol{Z}$ is the correlation matrix. The artificial data set used in Section 11.3 to illustrate the measures of collinearity is again used here. Chapter 14 is a case study using the methods discussed in this chapter.

TABLE 13.1. *Correlation matrix of the independent variables for the artificial data set demonstrating collinearity.*

	X_1	X_2	X_3
X_1	1.000	.996	.290
X_2	.996	1.000	.342
X_3	.290	.342	1.000

13.1 Understanding the Structure of the X-Space

Correlation Matrix $Z'Z$

The matrix of sums of squares and products of the centered and scaled independent variables $\boldsymbol{Z}'\boldsymbol{Z}$, scaled so that the sum of squares of each variable is unity, is a useful starting point for understanding the structure of the X-space. (This is the *correlation matrix* if the independent variables are random variables and, for convenience, is referred to as the correlation matrix even when the Xs are fixed constants.) The off-diagonal elements of this matrix are the cosines of the angles between the corresponding centered and scaled vectors in X-space. Values near 1.0 or -1.0 indicate nearly collinear vectors; values near 0 indicate nearly orthogonal vectors.

Example 13.1

The correlation matrix for the artificial data from Example 11.11 shows a very high correlation between X_1 and X_2 of $r_{12} = .996$ (Table 13.1). This indicates a near-linear dependency, which is known to exist from the manner in which the data were constructed. The relatively low correlations of X_1 and X_2 with X_3 suggest that X_3 is not involved in the collinearity problem. ■

Correlations will reveal linear dependencies involving two variables, but they frequently will not reveal linear dependencies involving several variables. Individual pairwise correlations can be relatively small when several variables are involved in a linear dependency. Thus, the absence of high correlations cannot be interpreted as an indication of no collinearity problems.

Detecting Near-Singularities

Near-linear dependencies involving any number of variables are revealed with a **singular value decomposition** of the matrix of independent variables, or with an **eigenanalysis** of the sums of squares and products matrix. (See Sections 2.7 and 2.8 for discussions of eigenanalysis and singular value decomposition.) For the purpose of detecting near-singularities, the independent variables should always be scaled so that the vectors are of equal length. In addition, the independent variables are often centered to remove collinearities with the intercept. (Refer to Section 11.3 for discussion on this point.) The discussion here is presented in terms of the centered and scaled independent variables $\boldsymbol{Z}$. The eigenvectors of $\boldsymbol{Z}'\boldsymbol{Z}$ that corre-

TABLE 13.2. *Eigenvalues and eigenvectors of the correlation matrix of independent variables for the artificial data set.*

Eigenvalue	*Eigenvectors*
λ_1=2.166698	$\boldsymbol{v}_1'$= (.65594 .66455 .35793)
λ_2= .830118	$\boldsymbol{v}_2'$=(−.28245 −.22365 .93285)
λ_3= .002898	$\boldsymbol{v}_3'$= (.69998 −.71299 .04100)

spond to the smaller eigenvalues identify the linear functions of the Zs that show least dispersion. It is these specific linear functions that are causing the collinearity problem if one exists.

Example 13.2

The results of the eigenanalysis of the correlation matrix for Example 13.1 are shown in Table 13.2. The eigenvalues reflect a moderate collinearity problem, with the condition number being $(2.16698/.00290)^{1/2} = 27.3$. (This differs from the results in Section 11.3 since collinearities involving the intercept have been eliminated by centering the variables.) The eigenvector corresponding to the smallest eigenvalue defines the third principal component, the dimension causing the collinearity problem, as

$$\boldsymbol{W}_3 = .69998\boldsymbol{Z}_1 - .71299\boldsymbol{Z}_2 + .04100\boldsymbol{Z}_3.$$

The variables primarily responsible for the near-singularity are $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ as shown by their relatively large coefficients in the third eigenvector. The coefficient for Z_3 is relatively close to zero. The coefficients on $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are very similar in magnitude but opposite in sign, suggesting that the near-singularity is due to $(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)$ being nearly zero. This is known to be true from the way the data were constructed; $\boldsymbol{X}_2$ was defined as $(\boldsymbol{X}_1 - 25)$ with 2 of the 20 numbers changed by 1 digit to avoid a complete singularity. After centering and scaling, $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are very nearly identical so that their difference is almost 0.

Inspection of the first eigenvector shows that the major dispersion in the Z-space is in the dimension defined as a weighted average of the three variables

$$\boldsymbol{W}_1 = .65594\boldsymbol{Z}_1 + .66455\boldsymbol{Z}_2 + .35793\boldsymbol{Z}_3$$

with $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ receiving nearly twice as much weight as $\boldsymbol{Z}_3$. $\boldsymbol{W}_1$ is the first principal component. The second dimension, the second principal component, is dominated by $\boldsymbol{Z}_3$,

$$\boldsymbol{W}_2 = -.28245\boldsymbol{Z}_1 - 0.22365\boldsymbol{Z}_2 + 0.93285\boldsymbol{Z}_3.$$

■

Gabriel's Biplot

The correlational structure of the independent variables is displayed with Gabriel's biplot (Gabriel 1971, 1972, 1978). This is an informative plot

that shows (1) the relationships among the independent variables, (2) the relative similarities of the individual data points, and (3) the relative values of the observations for each independent variable. The name "biplot" comes from this simultaneous presentation of both row (observation) and column (variable) information in one plot.

The biplot uses the singular value decomposition of $\boldsymbol{Z}$, $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{L}^{1/2}\boldsymbol{V}'$. The matrices $\boldsymbol{L}^{1/2}$ and $\boldsymbol{V}$ can be obtained from the results of the eigenanalysis of $\boldsymbol{Z}'\boldsymbol{Z}$ shown in Table 13.2. $\boldsymbol{V}$ is the matrix of eigenvectors, each column being an eigenvector, and $\boldsymbol{L}$ is the diagonal matrix of the positive square roots of the eigenvalues. More computations are required to obtain $\boldsymbol{U}$. If the dispersion in Z-space can be adequately represented by two dimensions, one biplot using the first and second principal component information will convey most of the information in $\boldsymbol{Z}$. If needed, additional biplots of first and third, and second and third principal components can be used. Each biplot is the projection of the dispersion in Z-space onto the plane defined by the two principal components being used in the biplot.

Example 13.3

(Continuation of Example 13.2*)* The first two principal component dimensions account for

$$\frac{\lambda_1 + \lambda_2}{\sum \lambda_i} = \frac{2.16698 + .83012}{3} = .999 \tag{13.1}$$

or 99.9% of the total dispersion in the three dimensions. Therefore, a single biplot of the first and second principal components suffices; only .1% of the information in $\boldsymbol{Z}$ is ignored by not using the third principal component.

The biplot using the first two principal component dimensions is shown in Figure 13.1. The vectors in the figure are the vectors of the independent variables *as seen in this two-dimensional projection*. The coordinates for the endpoints of the vectors, which are called *column markers*, are obtained from $\boldsymbol{L}^{1/2}\boldsymbol{V}'$,

$$\boldsymbol{L}^{1/2}\boldsymbol{V}' = \begin{bmatrix} .9656 & .9783 & .5269 \\ -.2573 & -.2038 & .8499 \\ .0377 & -.0384 & .0022 \end{bmatrix}. \tag{13.2}$$

The first and second elements in column 1 are the coordinates for the $\boldsymbol{Z}_1$ vector in the biplot using the first and second principal components, the first and second elements in column 2 are the coordinates of the $\boldsymbol{Z}_2$ vector, and so on. The third number in each column of $\boldsymbol{L}^{1/2}\boldsymbol{V}'$ gives the coordinate in the third dimension for each vector, which is being ignored in this biplot. Notice, however, that none of the variable vectors are very far from zero in the third dimension. This reflects the small amount of dispersion in that dimension. ■

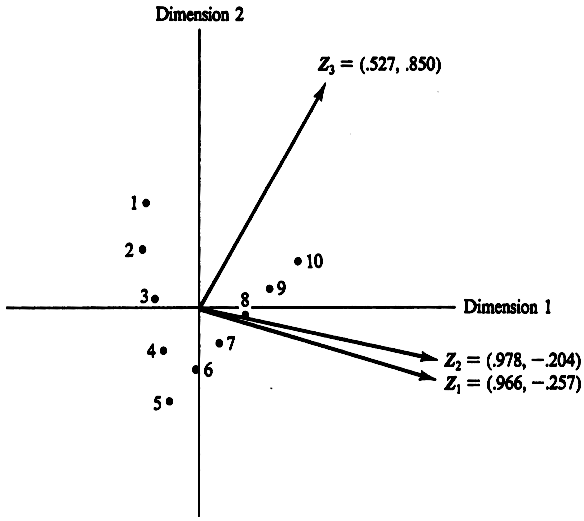

FIGURE 13.1. *Gabriel's biplot of the first two principal component dimensions for Example 13.3.*

Variable Information

Since the $\boldsymbol{Z}_j$ vectors were scaled to have unit length in the original n-dimensional space, the deviation of each vector length from unity in the biplot provides a direct measure of how far the original vector is from the plane being plotted. Thus, plotted vectors that are close to having unit length are well represented by the biplot and relationships among such vectors are accurately displayed. Conversely, plotted vectors that are appreciably shorter than unity are not well represented in that particular biplot; other biplots should be used to study relationships involving these vectors. In this example, all three plotted vectors are very close to having unit length.

Observation Information

The dots in the biplot represent the observations. The coordinates for the observations, called *row markers*, are the elements of $\boldsymbol{U}$ from the singular value decomposition. Recall that the principal components can be written as $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{L}^{1/2}$. Thus, each column of $\boldsymbol{U}$ is one of the principal components rescaled to remove λ_j.

Example 13.4

(Continuation of Example 13.3*)* The first ten rows of $\boldsymbol{U}$ are

$$\boldsymbol{U} = \begin{bmatrix} -.2350 & .4100 & -.4761 \\ -.2365 & .2332 & .4226 \\ -.2011 & .0363 & .2355 \\ -.1656 & -.1605 & .0485 \\ -.1302 & -.3574 & -.1385 \\ -.0222 & -.2491 & -.0985 \\ .0857 & -.1407 & -.0584 \\ .1937 & -.0323 & -.0184 \\ .3016 & .0761 & .0217 \\ .4096 & .1844 & .0617 \\ \vdots & \vdots & \vdots \end{bmatrix}. \tag{13.3}$$

The second 10 rows in $\boldsymbol{U}$ duplicate the first 10 in this example. The first and second columns are the first and second principal components, respectively, except for multiplication by λ_1 and λ_2. These two columns are the coordinates for the observations in the biplot (Figure 13.1). The first observation, for example, has coordinates (-.2350, .4100). The horizontal and vertical scales for plotting the row markers need not be the same as the scales for the column markers. Often the scales for the row markers will be shown across the top and across the right side of the plot as illustrated later in Figure 13.2. ■

The following are the key elements for the interpretation of Gabriel's biplot.

1. The length of the variable vector in a biplot, relative to its length in the original n-space, indicates how well the two-dimensional biplot represents that vector. Vectors that do not lie close to the plane defined by the two principal components being used in the biplot will project onto the biplot as much shorter vectors than they are in n-space. For such variables, that particular biplot will be a poor representation of the relationship among the variables and interpretations involving them should be avoided.

2. The angle between two variable vectors reflects their pairwise correlation as seen in this two-dimensional projection. The correlation is the cosine of the angle. Hence, a 90° angle indicates 0 correlation; a 0° or 180° degree angle indicates correlations of 1.0 and −1.0, respectively. [The angles between the vectors translate into correlations only because the variables have been centered before the eigenanalysis was done. The biplot is also used for some purposes on uncentered and/or unscaled data. See Bradu and Gabriel (1974, 1978), Gabriel (1971, 1972, 1978), and Corsten and Gabriel (1976) for examples.]

3. The spatial proximity of individual observations reflects their similarities with respect to this set of independent variables and as seen in

the two dimensions being plotted. Points close together have similar values and vice versa.

4. The relative values of the observations for a particular variable are seen by projecting the observation points onto the variable vector, extended as need be in either the positive or negative direction. The vector points in the direction of the largest values for the variable.

Example 13.5

The biplot of Example 13.4 shows that $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are very highly positively correlated; the angle between the two vectors is close to 0. $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are nearly orthogonal to $\boldsymbol{Z}_3$ since both angles are close to 90°. One would have to conclude from this biplot that $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are providing essentially the same information.

Although the three variables technically define a three-dimensional space, two of the vectors are so nearly collinear that the third dimension is almost nonexistent. No regression will be able to separate the effects of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ on $\boldsymbol{Y}$ from this set of data; the data are inadequate for this purpose. Furthermore, if the collinearity between $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ is a reflection of the innate properties of the system, additional data collected in the same way will show the same collinearity, and clear separation of their effects on any dependent variable will not be possible. When that is the case, it is probably best to define a new variable that reflects the ($\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$)-axis and avoid the use of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ per se. On the other hand, if the collinearity between $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ is a result of inadequate sampling or a bad experimental design, additional data will remove the collinearity and then separation of the effects of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ on the dependent variable might be possible.

The proximity of the observations (points) to each other reflects their similarities for the variables used in the biplot. For example, Points 1 and 2 are very much alike but quite different from Points 10 or 5. Real data will frequently show clusters of points that reflect meaningful groupings of the observations.

The perpendicular projections of the observations (points) onto one of the vectors, extended in either direction as needed, gives the *relative* values of the observations for that variable. If the projection of the observations onto the Z_1 or Z_2 axes is visualized, the points as numbered monotonically increase in value. Projection of the observations onto the Z_3 vector shows that their values for Z_3 decrease to the fifth point and then increase to the tenth point. (Recall that Points 11 to 20 are a repeat of 1 to 10.) This pattern is a direct reflection of the original values for the three variables. ■

Example 13.6

A second example of a biplot is taken from Shy-Modjeska, Riviere, and Rawlings (1984). This biplot, shown in Figure 13.2, displays the relationships among nephrotoxicity, physiological, and pharmacokinetic variables. The study used 24 adult female beagles which were subtotally nephrectomized (3/4 or 7/8 of the kidneys were surgically removed) and assigned

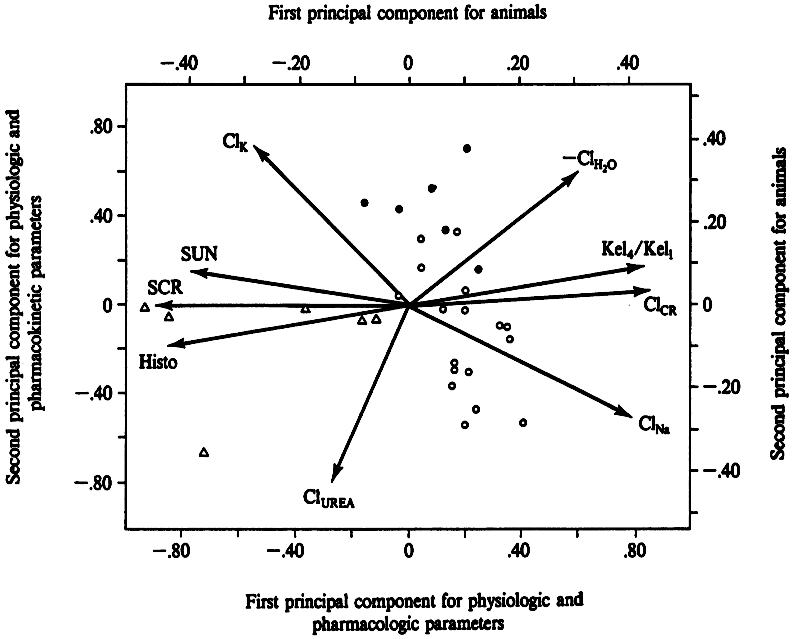

FIGURE 13.2. *Biplot of transformed physiologic data, Kel ratio, and histopathologic index for 23 subtotally nephrectomized and 6 control animals. The first two dimensions accounted for 76% of the dispersion in the full matrix. Triangles designate dogs that developed toxicity, open circles designate dogs that were nephrectomized but did not develop toxicity, and the closed circles designate control animals. (Used with permission.)*

to one of four different treatments. A control group of 6 dogs was used. Nine variables measuring renal function are used in this biplot. Complete data were obtained on 29 of the 30 animals. Six of the 24 nephrectomized animals developed toxicity. The biplot presents the information from the first two principal component dimensions of the 29×9 data matrix. These two dimensions account for 76% of the total dispersion in Z-space.

The biplot represents most of the vectors reasonably well. The shortest vectors are Cl_{UREA} and Cl_{H_2O}. All other vectors are at least 80% of their original length. The complex of five variables labeled *SUN*, *SCR*, *Histo*, Kel_4/Kel_1, and Cl_{CR} comprises a highly correlated system in these data. The first three are highly positively correlated as are the last two (the vectors point in the same direction), whereas there are high negative correlations between the two groups (the vectors point in different directions). The variable Cl_{NA}, on the other hand, is reasonably highly negatively correlated with Cl_K. Cl_{UREA} and Cl_{H_2O} also appear to be highly negatively

correlated but, recall, these are the two shortest vectors and may not be well represented in this biplot.

The horizontal axis across the bottom and the vertical axis on the left of Figure 13.2 are the scales for the column markers (variables) and row markers (animals), respectively, for the first principal component. The horizontal axis across the top and the vertical axis on the right are the scales for the column and row markers, respectively, for the second principal component. The vectors for the complex of five variables first mentioned are closely aligned with the axis of the first principal component; the first principal component is defined primarily by these five variables. Variation along the second principal component axis is primarily due to the variables Cl_{UREA}, Cl_K, Cl_{H_2O}, and Cl_{NA}, although these four variables are not as closely aligned with the axis.

The observations, the animals, tend to cluster according to the treatment received. Visualizing the projections of these points onto the vectors displays how the animals differ for these nine variables. The major differences among the animals will be along the first principal component axis and are due to the difference between toxic and nontoxic animals. The toxic animals tend to have high values for *SUN*, *SCR*, and *Histo* and low values for Kel_4/Kel_1 and Cl_{CR}. This suggests that these are the key variables to study as indicators of toxicity. (Which of the five variables caused the toxicity or are a direct result of the toxicity cannot be determined from these data. The biplot is simply showing the association of variables.) One toxic animal, the triangle in the lower left quadrant, is very different from all other animals. It has high values for the toxicity variables and a very high level of Cl_{UREA}. This would suggest a review of the data for this particular animal to ensure correctness of the values. If all appears to be in order, the other characteristics of the animal need to be studied to try to determine why it is responding so differently. The control animals separate from the nontoxic animals in the dimension of the second principal component. They have higher values for Cl_K and lower values for Cl_{UREA} and Cl_{NA} than the nontoxic animals.

This biplot accounts for 76% of the dispersion. Although this is the major part of the variation, a sizable proportion is being ignored. In this case, one would also study the information provided by the third dimension by biplotting the first and third and, perhaps, the second and third dimensions. These plots would reveal whether the negative correlation between Cl_{UREA} and Cl_{H_2O} is as strong as the first biplot suggests. ■

Overview of Gabriel's Biplot

Gabriel's biplot is a graphical technique for revealing relationships in a matrix of data. It is an exploratory tool and is not intended to provide estimates of parameters or tests of significance. Its graphical presentation of (1) the correlational structure among the variables, (2) the similarity of the observations, and (3) the relative values of the data points for the

variables measured can be most helpful in understanding a complex set of data.

For the artificial data set used in Examples 13.1 through 13.5, it is clear from the correlation matrix and the biplot that there is sufficient collinearity to cause a severe problem for ordinary least squares. With the high degree of collinearity between Z_1 and Z_2, it is unreasonable to expect any regression method to identify properly the contributions to $\boldsymbol{Y}$ of these two independent variables. Similarly, the biplot from Shy-Modjeska et al. (1984) showed a highly correlated complex of five variables that appeared to separate toxic from nontoxic animals. However, any regression analysis that attempts to assign relative importance to the five variables can be expected to be very misleading. "Seeing" the nature of the correlational structure in these data enhances the understanding of the problem and should introduce caution into the use of regression results. If it is important that effects of the individual variables be identified, data must be obtained in which the strong dependencies among the independent variables have been sufficiently weakened so that the collinearity problem no longer exists. In cases where the structure in the data is intrinsic to the system as it may be in the toxicity study of Example 13.6, it will be necessary to obtain data using experimental protocols that will disrupt the natural associations among the variables before reliable estimates of the effects can be obtained.

13.2 Biased Regression

Biased Estimators

The least squares estimators of the regression coefficients are the best linear unbiased estimators. That is, of all possible estimators that are both linear functions of the data and unbiased for the parameters being estimated, the least squares estimators have the smallest variance. In the presence of collinearity, however, this minimum variance may be unacceptably large. Relaxing the least squares condition that estimators be *unbiased* opens for consideration a much larger set of possible estimators from which one with better properties in the presence of collinearity might be found. **Biased regression** refers to this class of regression methods in which *unbiasedness* is no longer required. Such methods have been suggested as a possible solution to the collinearity problem. (See Chapters 10 and 11.) The motivation for biased regression methods rests in the potential for obtaining estimators that are closer, on average, to the parameter being estimated than are the least squares estimators.

13.2.1 Explanation

Mean Squared Error

A measure of average "closeness" of an estimator to the parameter being estimated is the **mean squared error (MSE)** of the estimator. If $\tilde{\theta}$ is an

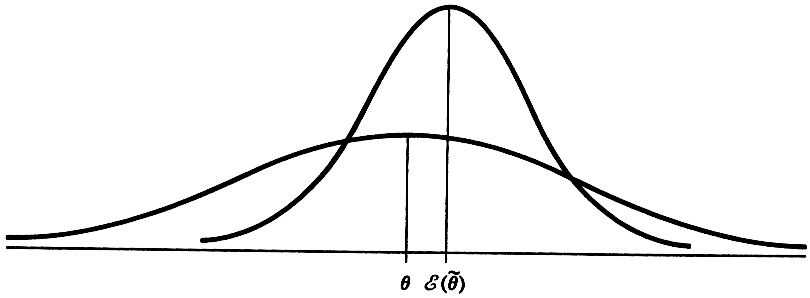

FIGURE 13.3. *Illustration of a biased estimator having smaller mean squared error than an unbiased estimator.*

estimator of θ, the mean squared error of $\tilde{\theta}$ is defined as

$$\text{MSE}(\tilde{\theta}) \quad = \quad \mathcal{E}(\tilde{\theta} - \theta)^2. \tag{13.4}$$

Recall that variance of an estimator $\tilde{\theta}$ is defined as

$$\text{Var}(\tilde{\theta}) \quad = \quad \mathcal{E}[\tilde{\theta} - \mathcal{E}(\tilde{\theta})]^2. \tag{13.5}$$

Note that MSE is the average squared deviation of the estimator from the *parameter* being estimated, whereas variance is the average squared deviation of the estimator from its *expectation*. If the estimator is unbiased, $\mathcal{E}(\tilde{\theta}) = \theta$ and $\text{MSE}(\tilde{\theta}) = \sigma^2(\tilde{\theta})$. Otherwise, MSE is equal to the variance of the estimator plus the square of its bias, where $\text{Bias}(\tilde{\theta}) = \mathcal{E}(\tilde{\theta}) - \theta$. See Exercise 13.1. It is possible for the variance of a biased estimator to be sufficiently smaller than the variance of an unbiased estimator to more than compensate for the bias introduced. In such a case, the biased estimator is closer on average to the parameter being estimated than is the unbiased estimator. Such is the hope with the biased regression techniques.

Potential for Biased Estimators

The possible advantage of biased estimators is illustrated in Figure 13.3. The normal curve centered at $\mathcal{E}(\tilde{\theta})$ represents the probability distribution of an unbiased estimator $\widehat{\theta}$ of θ; the bias is the difference between $\mathcal{E}(\tilde{\theta})$ and θ. The smaller spread in this distribution reflects its smaller variance. By allowing some bias, it may be possible to find an estimator for which the sum of its variance and squared bias, MSE, is smaller than the variance of the unbiased estimator.

Example 13.7

To illustrate the concept of biased regression methods, consider the linear model

$$Y_i \quad = \quad \beta_0 + Z_{i1}\beta_1 + Z_{i2}\beta_2 + \epsilon_i, \qquad i = 1, \ldots, n, \tag{13.6}$$

where Z_{i1} and Z_{i2} are centered and scaled and $\epsilon_i \sim \text{NID}(0, \sigma^2)$. That is, $\sum_{i=1}^n Z_{i1} = \sum_{i=1}^n Z_{i2} = 0$ and $\sum_{i=1}^n Z_{i1}^2 = \sum_{i=1}^n Z_{i2}^2 = 1$. Let ρ denote

$\sum_{i=1}^{n} Z_{i1}Z_{i2}$. Since

$$\begin{bmatrix} n & 0 & 0 \\ 0 & \sum Z_{i1}^2 & \sum Z_{i1}Z_{i2} \\ 0 & \sum Z_{i1}Z_{i2} & \sum Z_{i2}^2 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} \frac{1}{n} & 0 & 0 \\ 0 & (1-\rho^2)^{-1} & -\rho(1-\rho^2)^{-1} \\ 0 & -\rho(1-\rho^2)^{-1} & (1-\rho^2)^{-1} \end{bmatrix},$$

the variance of the ordinary least squares estimators $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$ of β_0, β_1, and β_2 are σ^2/n, $\sigma^2/(1-\rho^2)$, and $\sigma^2/(1-\rho^2)$, respectively. When ρ is close to one, then the variables Z_{i1} and Z_{i2} are highly correlated and we have a collinearity problem. Notice that, when ρ is close to one, the variance of both $\widehat{\beta}_1$ and $\widehat{\beta}_2$ is $\sigma^2/(1-\rho^2)$ which is very large. Even though $\widehat{\beta}_1$ is the best linear unbiased estimator of β_1, we may be able to find a biased estimator that has smaller mean square error. Consider, for example, an estimator of β_1 given by

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} Z_{i1}Y_i}{\sum_{i=1}^{n} Z_{i1}^2}. \tag{13.7}$$

Note that $\tilde{\beta}_1$ is the ordinary least squares estimator of β_1 if we assume that $\beta_2 = 0$ in model 13.6. The estimator $\tilde{\beta}_1$ is not unbiased for β_1 since

$$\mathcal{E}(\tilde{\beta}_1) = \beta_1 + \rho\beta_2. \tag{13.8}$$

The bias of $\tilde{\beta}_1$ is $\mathcal{E}(\tilde{\beta}_1) - \beta_1 = \rho\beta_2$ and the variance of $\tilde{\beta}_1$ is σ^2.

Therefore, the mean squared error of $\tilde{\beta}_1$ is

$$\begin{aligned} \text{MSE}(\tilde{\beta}_1) &= Var(\tilde{\beta}_1) + [\text{Bias}(\tilde{\beta}_1)]^2 \\ &= \sigma^2 + \rho^2\beta_2^2. \end{aligned}$$

For small values of β_2, $\text{MSE}(\tilde{\beta}_1)$ may be smaller than $\text{MSE}(\widehat{\beta}_1)$. ■

Several biased regression methods have been proposed as solutions to the collinearity problem; Stein shrinkage (Stein, 1960), ridge regression (Hoerl and Kennard, 1970a, 1970b), and principal component regression and variations of principal component regression (Lott, 1973; Hawkins, 1973; Hocking, Speed, and Lynn, 1976; Marquardt, 1970; Webster, Gunst, and Mason, 1974). Although ridge regression has received the greatest acceptance, all have been used with apparent success in various problems. Nevertheless, biased regression methods have not been universally accepted and should be used with caution. The MSE justification for biased regression methods makes it clear that such methods can provide better *estimates* of the parameters in the sense of mean squared. It does not necessarily follow that

a biased regression solution is acceptable or even "better" than the least squares solution for purposes other than estimation of parameters.

Although collinearity does not affect the precision of the estimated *responses* (and predictions) at the observed points in the X-space, it does cause variance inflation of estimated responses at other points. Park (1981) shows that the restrictions on the parameter estimates implicit in principal component regression are also optimal in the MSE sense for estimation of responses over certain regions of the X-space. This suggests that biased regression methods may be beneficial in certain cases for estimation of responses also. However, caution must be exercised when using collinear data for estimation and prediction of responses for points other than the observed sample points.

The biased regression methods do not seem to have much to offer when the objective is to assign some measure of "relative importance" to the independent variables involved in a collinearity. In essence, the biased estimators of the regression coefficients for the variables involved in the collinearity are weighted averages of the least squares regression coefficients for those variables. Consequently, each is reflecting the joint effects of all variables in the complex. (This is illustrated later with the data from Examples 13.1 through 13.5). The best recourse to the collinearity problem when the objective is to assign relative importance is to recognize that the data are inadequate for the purpose and obtain better data, perhaps from controlled experiments.

Ridge regression and principal component regression are two commonly used biased regression methods. The biased regression methods attack the collinearity problem by computationally suppressing the effects of the collinearity. Ridge regression does this by reducing the apparent magnitude of the correlations. Principal component regression attacks the problem by regressing Y on the important principal components and then parceling out the effect of the principal component variables to the original variables. We briefly discuss the principal component regression here and refer the readers to Hoerl and Kennard (1970a, 1970b), Hoerl, Kennard, and Baldwin (1975), Marquardt and Snee (1975), and Smith and Campbell (1980) for ridge regression.

13.2.2 Principal Component Regression

Principal component regression approaches the collinearity problem from the point of view of eliminating from consideration those dimensions of the X-space that are causing the collinearity problem. This is similar, in concept, to dropping an independent variable from the model when there is insufficient dispersion in that variable to contribute meaningful information on Y. However, in principal component regression the dimension dropped from consideration is defined by a linear combination of the variables rather than by a single independent variable.

Eigenvectors of $\boldsymbol{Z}$

Principal component regression builds on the principal component analysis of the matrix of centered and scaled independent variables $\boldsymbol{Z}$. The SVD of $\boldsymbol{Z}$ has been used in the analysis of the correlational structure of the X-space and in Gabriel's biplot. This section continues with the notation and results defined in that section. The SVD of $\boldsymbol{Z}$ is used to give

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{L}^{1/2}\boldsymbol{V}', \tag{13.9}$$

where $\boldsymbol{U}$ $(n \times p)$ and $\boldsymbol{V}$ $(p \times p)$ are matrices containing the left and right eigenvectors, respectively, and $\boldsymbol{L}^{1/2}$ is the diagonal matrix of singular values. The singular values and their eigenvectors are ordered so that $\lambda_1^{1/2} > \lambda_2^{1/2} > \cdots > \lambda_p^{1/2}$. The eigenvectors are pairwise orthogonal and scaled to have unit length so that

$$\boldsymbol{U}'\boldsymbol{U} \quad = \quad \boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}. \tag{13.10}$$

Principal Components of $\boldsymbol{Z}$

The principal components of $\boldsymbol{Z}$ are defined as the linear functions of the Z_j specified by the coefficients in the column vectors of $\boldsymbol{V}$. The first eigenvector in $\boldsymbol{V}$ (first column) defines the first principal component, the second eigenvector in $\boldsymbol{V}$ defines the second principal component, and so forth. Each principal component is a linear function of all independent variables. The principal components $\boldsymbol{W}$ are also given by the columns of $\boldsymbol{U}$ multiplied by the corresponding $\lambda_j^{1/2}$. Thus,

$$\boldsymbol{W} \quad = \quad \boldsymbol{Z}\boldsymbol{V}$$

or

$$\boldsymbol{W} \quad = \quad \boldsymbol{U}\boldsymbol{L}^{1/2} \tag{13.11}$$

is the matrix of principal component variables. Each column in $\boldsymbol{W}$ gives the values for the n observations for one of the principal components.

$\boldsymbol{W}'\boldsymbol{W}$

The sum of squares and products matrix of the principal component variables $\boldsymbol{W}$ is the diagonal matrix of the eigenvalues,

$$\boldsymbol{W}'\boldsymbol{W} \quad = \quad (\boldsymbol{U}\boldsymbol{L}^{1/2})'(\boldsymbol{U}\boldsymbol{L}^{1/2}) = \boldsymbol{L}^{1/2}\boldsymbol{U}'\boldsymbol{U}\boldsymbol{L}^{1/2} = \boldsymbol{L}, \tag{13.12}$$

where $\boldsymbol{L} = \text{Diag}(\lambda_1 \quad \lambda_2 \quad \cdots \quad \lambda_p)$. Thus, the principal components are orthogonal to each other, since all sums of products are zero, and the sum of squares of each principal component is equal to the corresponding eigenvalue λ_j. The first principal component has the largest sum of squares, λ_1. The principal components corresponding to the smaller eigenvalues are the dimensions of the Z-space having the least dispersion. These dimensions of the Z-space with the limited dispersion are responsible for the collinearity problem if one exists.

Example 13.8

(Continuation of Example 13.7*)* Recall that λ_1 and λ_2 are eigenvalues of

$\boldsymbol{Z}'\boldsymbol{Z}$ and $\boldsymbol{V}$ consists of the corresponding eigenvectors. Assume that $\rho > 0$. In our example,

$$\boldsymbol{Z}'\boldsymbol{Z} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \tag{13.13}$$

$$\lambda_1 = 1 + \rho > 1 - \rho = \lambda_2, \qquad \text{and} \tag{13.14}$$

$$\boldsymbol{V} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \tag{13.15}$$

The principal components $\boldsymbol{W}$ are given by $\boldsymbol{W} = \boldsymbol{Z}\boldsymbol{V}$. That is,

$$W_{i1} = (Z_{i1} + Z_{i2})/\sqrt{2} \quad \text{and} \quad W_{i2} = (Z_{i1} - Z_{i2})/\sqrt{2}. \tag{13.16}$$

Note that

$$\boldsymbol{W}'\boldsymbol{W} = \begin{bmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{bmatrix}. \tag{13.17}$$

■

Linear Model

The linear model

$$\boldsymbol{Y} = \mathbf{1}\beta_0 + \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{13.18}$$

can be written in terms of the principal components $\boldsymbol{W}$ as

$$\boldsymbol{Y} = \mathbf{1}\beta_0 + \boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}. \tag{13.19}$$

This uses the fact that $\boldsymbol{V}\boldsymbol{V}' = \boldsymbol{I}$ to transform $\boldsymbol{Z}\boldsymbol{\beta}$ into $\boldsymbol{W}\boldsymbol{\gamma}$:

$$\boldsymbol{Z}\boldsymbol{\beta} = \boldsymbol{Z}\boldsymbol{V}\boldsymbol{V}'\boldsymbol{\beta} = \boldsymbol{W}\boldsymbol{\gamma}. \tag{13.20}$$

Notice that $\boldsymbol{\gamma} = \boldsymbol{V}'\boldsymbol{\beta}$ is the vector of regression coefficients for the principal components; $\boldsymbol{\beta}$ is the vector of regression coefficients for the Zs. The translation of $\boldsymbol{\gamma}$ back to $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} = \boldsymbol{V}\boldsymbol{\gamma}. \tag{13.21}$$

Example 13.9

(Continuation of Example 13.8*)* Consider a reparameterization of model (13.6) given by

$$Y_i = \beta_0 + W_{i1}\gamma_1 + W_{i2}\gamma_2 + \epsilon_i. \tag{13.22}$$

Using equation 13.16,

$$\begin{aligned} Y_i &= \beta_0 + \frac{1}{\sqrt{2}}(Z_{i1} + Z_{i2})\gamma_1 + \frac{1}{\sqrt{2}}(Z_{i1} - Z_{i2})\gamma_2 + \epsilon_i \\ &= \beta_0 + Z_{i1}\frac{1}{\sqrt{2}}(\gamma_1 + \gamma_2) + Z_{i2}\frac{1}{\sqrt{2}}(\gamma_1 - \gamma_2) + \epsilon_i \end{aligned} \tag{13.23}$$

and, comparing this with equation 13.6 we see that

$$\beta_1 = \frac{1}{\sqrt{2}}\gamma_1 + \frac{1}{\sqrt{2}}\gamma_2 \quad \text{and} \tag{13.24}$$

$$\beta_2 = \frac{1}{\sqrt{2}}\gamma_1 - \frac{1}{\sqrt{2}}\gamma_2. \tag{13.25}$$

Also, note that

$$\gamma_1 = \frac{1}{\sqrt{2}}\beta_1 + \frac{1}{\sqrt{2}}\beta_2 \quad \text{and} \tag{13.26}$$

$$\gamma_2 = \frac{1}{\sqrt{2}}\beta_1 - \frac{1}{\sqrt{2}}\beta_2. \tag{13.27}$$

■

Solution

Ordinary least squares using the principal components as the independent variables gives

$$\widehat{\boldsymbol{\gamma}} = (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{Y} = \boldsymbol{L}^{-1}\boldsymbol{W}'\boldsymbol{Y} \tag{13.28}$$

$$= \begin{bmatrix} (\sum_i W_{i1}Y_i) \,/\, \lambda_1 \\ (\sum_i W_{i2}Y_i) \,/\, \lambda_2 \\ \vdots \\ (\sum_i W_{ip}Y_i) \,/\, \lambda_p \end{bmatrix} = \begin{pmatrix} \widehat{\gamma}_1 \\ \widehat{\gamma}_2 \\ \vdots \\ \widehat{\gamma}_p \end{pmatrix}. \tag{13.29}$$

The regression coefficients for the principal components can be computed individually since the principal components are orthogonal; $\boldsymbol{W}'\boldsymbol{W}$ is a diagonal matrix. Likewise, the variance–covariance matrix of $\widehat{\boldsymbol{\gamma}}$ is the diagonal matrix

$$\mathbf{Var}(\widehat{\boldsymbol{\gamma}}) = \boldsymbol{L}^{-1}\sigma^2. \tag{13.30}$$

That is, the variance of $\widehat{\gamma}_j$ is $\sigma^2(\widehat{\gamma}_j) = \sigma^2/\lambda_j$, and all covariances are zero. Because of the orthogonality of the principal components, the partial and sequential sums of squares for each principal component are equal and each regression sum of squares can be computed individually as

$$SS(\gamma_j) = \widehat{\gamma}_j^2\lambda_j. \tag{13.31}$$

Relationship to Ordinary Least Squares

If all principal components are used, the results are equivalent to ordinary least squares regression. The estimate of $\boldsymbol{\beta}$ is obtained from $\widehat{\boldsymbol{\gamma}}$ as

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{V}\widehat{\boldsymbol{\gamma}} \tag{13.32}$$

and the regression equation can be written as either

$$\widehat{\boldsymbol{Y}} = \mathbf{1}\overline{Y} + \boldsymbol{W}\widehat{\boldsymbol{\gamma}}$$

or

$$\widehat{\boldsymbol{Y}} = \mathbf{1}\overline{Y} + \boldsymbol{Z}\widehat{\boldsymbol{\beta}}. \tag{13.33}$$

Example 13.10

(Continuation of Example 13.9*)* From equation 13.22, the ordinary least squares estimators of γ_1 and γ_2 are

$$\begin{pmatrix} \widehat{\gamma}_1 \\ \widehat{\gamma}_2 \end{pmatrix} = \begin{pmatrix} \sum_i W_{i1}Y_i/(1+\rho) \\ \sum_i W_{i2}Y_i/(1-\rho) \end{pmatrix}. \tag{13.34}$$

The ordinary least squares estimators of β_1 and β_2 are given by

$$\widehat{\beta}_1 = \frac{1}{\sqrt{2}}\widehat{\gamma}_1 + \frac{1}{\sqrt{2}}\widehat{\gamma}_2 \qquad \text{and} \tag{13.35}$$

$$\widehat{\beta}_2 = \frac{1}{\sqrt{2}}\widehat{\gamma}_1 - \frac{1}{\sqrt{2}}\widehat{\gamma}_2. \tag{13.36}$$

Note that $Var(\widehat{\gamma}_1) = \sigma^2/(1+\rho)$ and $Var(\widehat{\gamma}_2) = \sigma^2/(1-\rho)$. When ρ is close to 1 (or -1), the variance of $\widehat{\gamma}_2$ (or $\widehat{\gamma}_1$) will be very large. ■

Eliminating Principal Components

The idea behind principal component regression, however, is to eliminate those dimensions that are causing the collinearity problem, those dimensions for which the λ_j are very small. Assume it has been decided to eliminate s principal components, usually those having the s smallest eigenvalues, and retain g principal components for the analysis ($g+s=p$). The subscript (g) is used on $\boldsymbol{V}$, $\boldsymbol{L}$, $\boldsymbol{W}$, and $\widehat{\gamma}$ to designate the partitions of the corresponding matrices that relate to the g principal component dimensions retained in the analysis. Thus, $\boldsymbol{V}_{(g)}$ is the $p \times g$ matrix of retained eigenvectors, $\boldsymbol{W}_{(g)}$ is the $n \times g$ matrix of the corresponding principal components, and $\widehat{\gamma}_{(g)}$ is the vector of their estimated regression coefficients. The subscript (g) is used on other results to designate the number of principal components retained in the analysis.

Recall that the principal component regression coefficients, their variances, and the sums of squares attributable to each can be computed independently since the principal components are orthogonal. Therefore, $\widehat{\gamma}_{(g)}$ is obtained from $\widehat{\gamma}$ by simply extracting the g elements corresponding to the retained principal components. The **principal component regression**

estimate of $\boldsymbol{\beta}$, the regression coefficients for the Zs, is given by

$$\underset{(p\times 1)}{\boldsymbol{\beta}^{+}_{(g)}} = \underset{(p\times g)}{\boldsymbol{V}_{(g)}}\ \underset{(g\times 1)}{\widehat{\boldsymbol{\gamma}}_{(g)}} \tag{13.37}$$

The notation $\boldsymbol{\beta}^{+}$ is used in place of $\widehat{\boldsymbol{\beta}}$ to distinguish the principal component estimates of $\boldsymbol{\beta}$ from the least squares estimates. Notice that there are p elements in $\boldsymbol{\beta}^{+}_{(g)}$, even though there are only g elements in $\widehat{\gamma}$.

The variance of $\boldsymbol{\beta}^{+}_{(g)}$ is

$$\mathbf{Var}[\boldsymbol{\beta}^{+}_{(g)}] = \boldsymbol{V}_{(g)}\boldsymbol{L}^{-1}_{(g)}\boldsymbol{V}'_{(g)}\sigma^2. \tag{13.38}$$

These variances involve the reciprocals of only the larger eigenvalues. The smaller ones causing the variance inflation in the ordinary least squares solution have been eliminated.

Example 13.11

(Continuation of Example 13.10*)* Suppose we decide to eliminate the second principal component $\boldsymbol{W}_2$ and retain only the first principal component $\boldsymbol{W}_1$. Since $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are orthogonal to each other, the estimator of γ_1 obtained by eliminating $\boldsymbol{W}_2$ (or setting $\gamma_2 = 0$) is the same as $\widehat{\gamma}_1$ given in equation 13.34. Then, the principal component estimators of β_1 and β_2 are

$$\beta^{+}_1 = \frac{1}{\sqrt{2}}\widehat{\gamma}_1 \qquad \text{and} \tag{13.39}$$

$$\beta^{+}_2 = \frac{1}{\sqrt{2}}\widehat{\gamma}_1. \tag{13.40}$$

Note that

$$\mathrm{Var}(\beta^{+}_1) = \mathrm{Var}(\beta^{+}_2) = \mathrm{Cov}(\beta^{+}_1, \beta^{+}_2) = \frac{\sigma^2}{2(1+\rho)}. \tag{13.41}$$

These variances are always smaller than $\mathrm{Var}(\widehat{\beta}_1) = \mathrm{Var}(\widehat{\beta}_2) = \sigma^2/(1-\rho^2)$ and are much smaller when ρ is close to 1. ■

SS(Regr)

The sum of squares due to regression is the sum of the contributions from the g principal components retained and has g degrees of freedom:

$$\mathrm{SS(Regr)} = \sum_{j\in\{g\}} \mathrm{SS}(\gamma_j), \tag{13.42}$$

where summation is over the subset of g principal components retained in the model.

Regression Equations

The regression equation can be written either as

$$\begin{aligned}\widehat{\boldsymbol{Y}}_{(g)} &= \mathbf{1}\overline{Y} + \boldsymbol{Z}\boldsymbol{\beta}^{+}_{(g)} \qquad \text{or} \\ \widehat{\boldsymbol{Y}}_{(g)} &= \mathbf{1}\overline{Y} + \boldsymbol{W}_{(g)}\widehat{\boldsymbol{\gamma}}_{(g)}, \end{aligned} \tag{13.43}$$

where $\boldsymbol{W}_{(g)}$ is the matrix of retained principal components; $\widehat{\beta}_0 = \overline{Y}$ and is orthogonal to each $\boldsymbol{\beta}^{+}_{(g)}$.

The variance of $\widehat{\boldsymbol{Y}}_{(g)}$ can be written in several forms. Perhaps the simplest is

$$\mathbf{Var}\left[\widehat{\boldsymbol{Y}}_{(g)}\right] = \left[\frac{\boldsymbol{J}}{n} + \boldsymbol{W}_{(g)}\boldsymbol{L}^{-1}_{(g)}\boldsymbol{W}'_{(g)}\right]. \tag{13.44}$$

Relationship of $\boldsymbol{\beta}^{+}_{(g)}$ to $\widehat{\boldsymbol{\beta}}$

The principal component regression coefficients can be expressed as linear functions of the least squares estimates:

$$\begin{aligned}\boldsymbol{\beta}^{+}_{(g)} &= \boldsymbol{V}_{(g)}\boldsymbol{V}'_{(g)}\widehat{\boldsymbol{\beta}} \\ &= [\boldsymbol{I} - \boldsymbol{V}_{(s)}\boldsymbol{V}'_{(s)}]\widehat{\boldsymbol{\beta}}, \end{aligned} \tag{13.45}$$

where $\boldsymbol{V}_{(s)}$ is the matrix of s eigenvectors that were dropped from the analysis. Since $\widehat{\boldsymbol{\beta}}$ is unbiased, the expectation and bias of the principal component regression coefficients follow from equation 13.45;

$$\mathcal{E}[\boldsymbol{\beta}^{+}_{(g)}] = \boldsymbol{\beta} - \boldsymbol{V}_{(s)}\boldsymbol{V}'_{(s)}\boldsymbol{\beta}$$

or the bias is

$$\text{Bias} = \mathcal{E}[\boldsymbol{\beta}^{+}_{(g)}] - \boldsymbol{\beta} = -\boldsymbol{V}_{(s)}\boldsymbol{V}'_{(s)}\boldsymbol{\beta}. \tag{13.46}$$

Linear Restrictions on $\boldsymbol{\beta}^{+}_{(g)}$

The fact that $\boldsymbol{\beta}^{+}_{(g)}$ has p elements, a regression coefficient for each independent variable, even though only g regression coefficients $\boldsymbol{\gamma}_{(g)}$ were estimated implies that there are linear restrictions on $\boldsymbol{\beta}^{+}_{(g)}$. There is one linear restriction for each eliminated principal component. The linear restrictions on $\boldsymbol{\beta}^{+}_{(g)}$ are defined by $\boldsymbol{V}_{(s)}$ as

$$\boldsymbol{V}'_{(s)}\boldsymbol{\beta} = 0. \tag{13.47}$$

Example 13.12

(Continuation of Example 13.11*)* Using β^{+}_1 and β^{+}_2, the regression equation can be written as

$$\begin{aligned}\widehat{Y}_{i(1)} &= \overline{Y} + Z_{i1}\widehat{\beta}^{+}_1 + Z_{i2}\widehat{\beta}^{+}_2 \\ &= \overline{Y} + W_{i1}\widehat{\gamma}_1. \end{aligned}$$

The variance of the vector $\widehat{\boldsymbol{Y}}_{(1)}$ is given by

$$\mathbf{Var}(\widehat{\boldsymbol{Y}}_{(1)}) = \sigma^2 \left[\frac{\boldsymbol{J}}{n} + \frac{1}{(1+\rho)} \boldsymbol{W}_1 \boldsymbol{W}_1' \right].$$

Even though the estimates of $\boldsymbol{\beta}_1^+$ and $\boldsymbol{\beta}_2^+$ have smaller variance than $\widehat{\beta}_1$ and $\widehat{\beta}_2$, respectively, they are biased. Note that

$$\begin{aligned} \mathcal{E}[\boldsymbol{\beta}_1^+] &= \frac{1}{\sqrt{2}} \mathcal{E}[\widehat{\gamma}_1] \\ &= \frac{1}{\sqrt{2}} \gamma_1 \end{aligned}$$

and hence the bias of $\boldsymbol{\beta}_1^+$ is

$$\begin{aligned} \mathcal{E}[\boldsymbol{\beta}_1^+] &= \frac{1}{\sqrt{2}} \gamma_1 - (\frac{1}{\sqrt{2}} \gamma_1 + \frac{1}{\sqrt{2}} \gamma_2) \\ &= -\frac{1}{\sqrt{2}} \gamma_2 \\ &= -\frac{1}{2} (\beta_1 - \beta_2). \end{aligned}$$

Similarly, the bias of $\boldsymbol{\beta}_2^+$ is

$$\mathcal{E}[\boldsymbol{\beta}_2^+] = \frac{1}{2} (\beta_1 - \beta_2).$$

Finally, note that the principal component regression coefficients satisfy a linear restriction: $\boldsymbol{\beta}_1^+ - \boldsymbol{\beta}_2^+ = 0$. ■

Eliminating Principal Components

It is best to be conservative in eliminating principal components since each one eliminated introduces another constraint on the estimates and another increment of bias. The bias term, equation 13.46, can also be expressed as $-V_{(s)}\gamma_{(s)}$, where $\boldsymbol{\gamma}_{(s)}$ is the set of principal component regression coefficients dropped. Hence, one does not want to eliminate a principal component for which γ_j is very different from zero. A good working rule seems to be to eliminate only those principal components that

1. have small enough eigenvalues to cause serious variance inflation (see Section 11.3) *and*
2. for which the estimated regression coefficient $\widehat{\gamma}_j$ is not significantly different from zero.

One may wish to use a somewhat lower level of significance (say $\alpha = .10$ or .20) for testing the principal component regression coefficients in order to allow for the low power that is likely to be present for the dimensions that have limited dispersion.

The key steps in principal component regression are the following.

1. Obtain the singular value decomposition on the matrix of centered and scaled independent variables $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{L}^{1/2}\boldsymbol{V}'$.

2. The principal components are given by $\boldsymbol{W} = \boldsymbol{Z}\boldsymbol{V}$ or $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{L}^{1/2}$.

3. Regress $\boldsymbol{Y}$ on $\boldsymbol{W}$ to obtain the estimates of the regression coefficients for the p principal components $\widehat{\boldsymbol{\gamma}}$, their estimated variances $\boldsymbol{s}^2(\widehat{\boldsymbol{\gamma}})$, and the sums of squares due to regression $\text{SS}(\widehat{\gamma}_j)$. The residual mean square from the full model is used as the estimate of σ^2 in $\boldsymbol{s}^2(\widehat{\boldsymbol{\gamma}})$.

4. Test $H_0 : \gamma_j = 0$ for each j using Student's t or F. Eliminate from the regression all principal components that

   a. are causing a collinearity problem (condition index > 10, for example) *and*

   b. do not make a significant contribution to the regression.

5. $\widehat{\boldsymbol{\gamma}}_{(g)}$ is the vector of estimated regression coefficients retained. SS(Regr) $= \sum \text{SS}(\gamma_j)$, where summation is over the g components retained. SS(Regr) has g degrees of freedom.

6. Convert the regression coefficients for the principal components to the regression coefficients for the original independent variables (centered and scaled) by
$$\boldsymbol{\beta}^{+}_{(g)} = \boldsymbol{V}_{(g)}\widehat{\boldsymbol{\gamma}}_{(g)},$$
which has estimated variance
$$\boldsymbol{s}^2[\boldsymbol{\beta}^{+}_{(g)}] = \boldsymbol{V}_{(g)}\boldsymbol{L}^{-1}_{(g)}\boldsymbol{V}'_{(g)}\boldsymbol{s}^2.$$

7. The regression equation is either
$$\widehat{\boldsymbol{Y}}_{(g)} = \boldsymbol{1}\overline{Y} + \boldsymbol{Z}\boldsymbol{\beta}^{+}_{(g)} \quad \text{or} \quad \widehat{\boldsymbol{Y}}_{(g)} = \boldsymbol{1}\overline{Y} + \boldsymbol{W}\widehat{\boldsymbol{\gamma}}_{(g)}.$$

Example 13.13

The principal component regression analysis for the example begins with the principal component analysis using the data from Example 13.2. The singular values λ_j showed that the dimension defined by the third principal component accounted for less than .1% of the total dispersion of the centered and standardized variables $\boldsymbol{Z}$. The second dimension accounted for 28% of the total dispersion.

The estimates of the regression coefficients for the principal components and the sum of squares attributable to each are shown in Table 12.5. The total sum of squares accounted for by the three principal components equals the total sum of squares due to regression of the original variables, SS(Regr)= 20.784. The regression coefficients for the first two principal components are highly significant; the regression coefficient for the third component is not significant even at $\alpha = .20$. Consequently, no important

TABLE 13.3. *Estimated regression coefficients for the principal components, their standard errors, and the sum of squares attributable to each.*

Principal Component j	*Regression Coefficients* $\widehat{\gamma}_j$	*Standard Errors*	*Sum of Squares*[a]
1	2.3473	.598	11.940**
2	3.0491	.967	7.718**
3	19.7132	16.361	1.126

[a] ** indicates sum of squares is significant at the .01 level of probability. Each sum of squares has 1 degree of freedom and was tested against the residual mean square from the full model, $s^2 = .776$ with 16 degrees of freedom.

information on Y would be lost if the third principal component were to be dropped from the regression. The very large standard error on $\widehat{\gamma}_3$ reflects the extremely small amount of variation in the dimension defined by the third principal component.

The principal component analysis and Gabriel's biplot showed that the first principal component is defined primarily by Z_1 and Z_2 with a much smaller contribution from Z_3. This particular linear function of Z_1, Z_2, and Z_3 contains information on Y as shown by its significance. Likewise, the second principal component dominated by Z_3 is important for Y. However, the third principal component, essentially the difference between Z_1 and Z_2, does not make a significant contribution to the regression. This does *not* imply that the difference between Z_1 and Z_2 is unimportant in the process being studied. In fact, the equation used to generate Y in this artificial example gives greater weight to the difference than it gives to the sum of Z_1 and Z_2. In this particular set of data, however, Z_1 and Z_2 are so nearly collinear that their difference is always very close to being a constant and, therefore, the impact of the difference is estimated only with very low precision.

The principal component regression estimate of $\boldsymbol{\beta}$ (Table 13.4) using all principal components ($g = 3$) reproduces the ordinary least squares result. The estimate of $\boldsymbol{\beta}$ using only the first two principal components $\boldsymbol{\beta}^{+}_{(2)}$ shows a marked change toward zero in the first two regression coefficients, and a marked decrease in their standard errors. The change is small in the third regression coefficient and its standard error. The large changes associated with Z_1 and Z_2 and the small change associated with Z_3 directly reflect the relative involvement of the independent variables in the near-singularity shown by the third principal component. The coefficient of determination for the principal component regression using the first two principal components is $R^2_{(2)} = .592$, only slightly less than $R^2 = .626$ for ordinary least squares. The regression equation estimated from principal

TABLE 13.4. *Principal component regression estimates of the regression coefficients for the original variables using all principal components ($g = 3$) and omitting the third principal component ($g = 2$).*

Scaled Variable Z_j	*Regression Coefficients Using g Principal Components*[a] $g = 3$	$g = 2$	*Mean Squared Error*
1	14.480 (11.464)	.678 (0.478)	14.83
2	−13.182 (11.676)	0.878 (0.452)	15.34
3	4.493 (1.144)	3.685 (0.927)	1.16

[a]Standard errors given in parentheses. The mean squared errors are for the g = 2 principal component solution.

component regression with $g = 2$ is

$$\widehat{Y}_{(g)i} = 21.18 + .678Z_{i1} + .878Z_{i2} + 3.685Z_{i3}.$$

Since the parameters $\boldsymbol{\beta}$ are known in this artificial example, the mean squared errors for the principal component regression are computed and given in the last column of Table 13.4. The mean squared errors for the variables involved in the near-singularity are an order of magnitude smaller than for ordinary least squares. Comparison with the variances of the estimated regression coefficients shows that most of MSE for $\beta^+_{(2)1}$ and $\beta^+_{(2)2}$ is due to bias.

The relationship between the principal component regression estimates and the least squares estimates for this example are shown by evaluating equation 13.45. This gives

$$\begin{pmatrix} \beta^+_{(2)1} \\ \beta^+_{(2)2} \\ \beta^+_{(2)3} \end{pmatrix} = \begin{bmatrix} .510 & .499 & -.029 \\ .499 & .492 & .029 \\ -.029 & .029 & .998 \end{bmatrix} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{pmatrix}.$$

The principal component estimates of β_1 and β_2 are very nearly simple averages of the corresponding least squares estimates. The principal component estimate of β_3 is nearly identical to the least squares estimate. This illustrates a general result of principal component regression: the estimated coefficients for any variables that are nearly orthogonal to the axes causing the collinearity problems are nearly identical to the least squares estimates. However, for variables involved in the collinearity problem, their estimates given by principal component regression are weighted averages of the least squares regression coefficients of *all* variables involved in the collinearity. Principal component regression provides no information on the relative contribution (to the response variable) of variables involved in the collinearity.

For illustration, it is helpful to follow up on the obvious suggestion from the principal component analysis and the biplot that Z_1 and Z_2, for all practical purposes, present the same information. If the two variables are redundant, a logical course of action is to use only one of the two or their average. The regression analysis was repeated using $\overline{Z} = (Z_1 + Z_2)/2$ as one variable, rescaled to have unit length, and Z_3 as the second variable. Of course, the collinearity problem disappeared. This regression analysis gave $R^2 = .593$, essentially the same as the principal component regression result. The least squares regression coefficient for $\overline{Z}$ was 1.55 (with a standard error of .93). This is almost exactly the sum of the two regression coefficients for Z_1 and Z_2 estimated from the principal component regression using $g = 2$. Thus, the principal component regression analysis replaces the correlated complex of variables causing the near-singularity with a surrogate variable, the principal component, and then "parcels out" the estimated effect of the surrogate variable among the variables that made up the complex. ■

13.3 General Comments on Collinearity

The course of action in the presence of collinearity depends on the nature and origin of the collinearity and on the purpose of the regression analysis. If the regression analysis is intended solely for prediction of the dependent variable, the presence of near singularities in the data does not create serious problems as long as certain very important conditions are met:

Less Serious for Prediction

1. The collinearity shown in the data is a reflection of the correlational structure of the X-space. It must not be an artifact of the sampling process or due to outliers in the data. [Mason and Gunst (1985) discuss the effects and detection of collinearities induced by outliers.]

2. The system continues to operate in the same manner as when the data were generated so that the correlational structure of the X-space remains consistent. This implies that the regression equation is not to be used to predict the response to some modification of the system even if the prediction point is in the sample X-space (Condition 3).

3. Prediction is restricted to points within the sample X-space. Extrapolation beyond the data is dangerous in any case but can quickly lead to serious errors of prediction when the regression equation has been estimated from highly collinear data.

These conditions are very limiting and simply reflect the extreme sensitivity of ordinary least squares when collinearity is present. Nevertheless, the

impact of collinearity for prediction is much less than it is for estimation (Thisted, 1980). Any variable selection process for model building will tend to select one independent variable from each correlated set to act as a surrogate variable for the complex. The remaining variables in that complex will be dropped. It does not matter for prediction purposes whether the retained variable is a causal variable in the process; it is only important that the system continue to "act" as it did when the data were collected so that the surrogate variable continues to adequately represent the complex of variables.

Serious Problems

On the other hand, collinearity creates serious problems if the purpose of the regression is to understand the process, to identify important variables in the process, or to obtain meaningful estimates of the regression coefficients. The ordinary least squares estimates can be far from the true values. In the numerical example, the true values of the regression coefficients were 5.138, −2.440, and 2.683 compared to the estimated values of 14.5, −13.2, and 4.49. Although there is always uncertainty with observational data regarding the *true* importance of a variable in the process being studied, the presence of collinearity almost ensures that the identification of important variables will be wrong. If all potentially important variables are retained in the model, all variables in any correlated complex will appear to be unimportant because any one of them, important or not to the process, can usurp the function of the others in the regression equation. Furthermore, any variable selection process to choose the best subset of variables will almost certainly "discard" important variables and the variable retained to represent each correlated complex may very well be unimportant to the process. For these purposes, it is extremely important that the presence of collinearity be recognized and its nature understood.

Some degree of collinearity is expected with observational data. "Seeing" the correlational structure should alert the researcher to the cases where the collinearity is the result of inadequate or erroneous data. The solution to the problem is obvious for these cases; near-singularities that result from inadequate sampling or errors in the data will disappear with more and better data. It may be necessary to change sampling strategy to obtain data points in regions of the X-space not previously represented. Correlations inherent to the system will persist. Analysis of the correlational structure should provide insight to the researcher on how the system operates and may suggest alternative parameterizations and models. In the final analysis, it will probably be necessary to resort to controlled experimentation to separate the effects of highly collinear variables. Collinearity should seldom be a problem in controlled experiments. The choice of treatment levels for the experiment should be such that the factors are orthogonal, or nearly so.

13.4 Summary

The purposes of this chapter were to emphasize the importance of understanding the nature of any near-singularities in the data that might cause problems with the ordinary least squares regression, to introduce principal component analysis and Gabriel's biplots as tools for aiding this understanding, and to acquaint the reader with one (of the several) biased regression methods. All of the biased regression methods are developed on the premise that estimators with smaller mean squared errors can be found if unbiasedness of the estimators is not required. As with many regression techniques, the reader is cautioned against indiscriminate use of biased regression methods. Every effort should be made to understand the nature and origin of the problem and to correct it with better data if possible.

13.5 Exercises

13.1. Use the definition of mean squared error in equation 13.4 to show that MSE is the variance of the estimator plus the square of the bias.

13.2. Use the variance of linear functions and $\widehat{\boldsymbol{\gamma}} = \boldsymbol{L}^{-1}\boldsymbol{W}'\boldsymbol{Y}$ to show that $\mathbf{Var}(\widehat{\boldsymbol{\gamma}}) = \boldsymbol{L}^{-1}\sigma^2$, equation 13.30.

13.3. Use equation 13.37 and the variance of linear functions to derive $\mathbf{Var}(\boldsymbol{\beta}^{+}_{(g)})$, equation 13.38.

13.4. Show that the sum of the variances of $\boldsymbol{\beta}^{+}_{(g)j}$ is equal to the sum of the variances of $\widehat{\boldsymbol{\gamma}}_j$. That is, show that $\text{tr}\{\mathbf{Var}[\boldsymbol{\beta}^{+}_{(g)}]\} = \text{tr}\{\mathbf{Var}[\widehat{\boldsymbol{\gamma}}_{(g)}]\}$.

13.5. Show that the length of the $\boldsymbol{\beta}^{+}_{(g)}$ vector is the same as the length of $\widehat{\boldsymbol{\gamma}}_{(g)}$.

13.6. Use the logarithms of the nine independent variables in the peak flow runoff data from Exercise 5.1.

(a) Center and scale the independent variables to obtain $\boldsymbol{Z}$ and $\boldsymbol{Z}'\boldsymbol{Z}$, the correlation matrix.

(b) Do the singular value decomposition on $\boldsymbol{Z}$ and construct the biplot for the first and second principal component dimensions. What proportion of the dispersion in the X-space is accounted for by these first two dimensions?

(c) Use the correlation matrix and the biplot to describe the correlational structure of the independent variables.

13.7. Do principal component regression on the peak flow runoff data (Exercise 5.1) to estimate the regression equation using the logarithms of all independent variables and $\ln(Q)$ as the dependent variable.

(a) Which principal components are causing a collinearity problem?

(b) Test the significance of the individual principal component regression coefficients. Which principal components will you retain for your regression?

(c) Convert the results to $\boldsymbol{\beta}^{+}_{(g)}$, compute estimates of their variances, and give the final regression equation (in terms of the Zs).

(d) Compute R^2.

13.8. Use the data from Andrews and Herzberg (1985) on percentages of sand, silt, and clay in soil at 20 sites given in Exercise 11.11.

(a) Do the singular value decomposition on $\boldsymbol{Z}$, the centered and scaled variables, and construct Gabriel's biplot of the data.

(b) How many principal components must be used in order to account for 80% of the dispersion?

(c) Interpret the results of the biplot (of the first and second principal components) in terms of (i) which variable vectors are not well represented by the biplot, (ii) the correlational structure of the variables, (iii) how the 20 sites tend to cluster, and (iv) which site has very low sand content at depths 1 and 2 but moderately high sand content at depth 3.

13.9. This exercise is a continuation of the Laurie-Alberg experiment on relating the activity of fruit flies to four enzymes (Exercise 11.9). The results of the SVD on $\boldsymbol{Z}$ are given in Exercise 11.9. Some of the results from principal component regression are given in the accompanying tables.

Estimates of the regression coefficients (for Zs) retaining the indicated principal components:

	Principal Components Retained			
Variable	*All*	1, 2, 3	1, 2	1
Intercept	13.118	13.118	13.118	13.118
SDH	−1.594	2.700	2.472	4.817
FUM	10.153	5.560	5.229	5.444
GH	4.610	6.676	6.400	4.543
GO	4.547	4.580	5.340	4.543

Variances of estimated regression coefficients retaining the indicated principal components:

	Principal Components Retained			
Variable	*All*	1, 2, 3	1, 2	1
Intercept	.05	.05	.05	.05
SDH	56.8	9.0	6.74	2.72
FUM	63.1	8.4	3.51	3.47
GH	29.0	17.9	14.50	.48
GO	28.8	28.8	2.89	2.42
$tr\{\mathbf{Var}[\boldsymbol{\beta}^{+}_{(g)}]\}$	177.6	64.1	27.64	9.10

(a) From the SVD in Exercise 11.9, are any principal components cause for concern in variance inflation? Which Zs are heavily involved in the fourth principal component?

(b) From inspection of the behavior of the variances as the principal components are dropped, which variables are heavily involved in the fourth principal component? Which are involved in the third principal component?

(c) Which principal component regression solution would you use? The variances continue to decrease as more principal components are dropped from the solution. Why would you not use the solution with *only* the first principal component?

(d) Do a t-test of the regression coefficients for your solution. (There were $n = 21$ observations in the data set.) State your conclusions.

13.10. Consider the model in equation 13.22 given by

$$Y_i = \beta_0 + W_{i1}\gamma_1 + W_{i2}\gamma_2 + \epsilon_i,$$

where $\epsilon_i \sim \text{NID}(0, \sigma^2)$, $\sum_i W_{i1} = \sum_i W_{i2} = \sum_i W_{i1}W_{i2} = 0$, $\sum_i W_{i1}^2 = (1+\rho)$, and $\sum_i W_{i2}^2 = (1-\rho)$. Consider the estimators

$$\begin{aligned} \tilde{\gamma}_{1(k_1)} &= \sum W_{i1}Y_i \,/\, \left(\sum W_{i1}^2 + k_1\right), \text{ and} \\ \tilde{\gamma}_{2(k_2)} &= \sum W_{i2}Y_i \,/\, \left(\sum W_{i2}^2 + k_2\right). \end{aligned}$$

(a) For $k_1 > 0$ and $k_2 > 0$, show that $\tilde{\gamma}_1(k_1)$ and $\tilde{\gamma}_2(k_2)$ are biased estimates of γ_1 and γ_2.

(b) Find the mean squared errors of $\tilde{\gamma}_1(k_1)$ and $\tilde{\gamma}_2(k_2)$.

[These are called the **generalized ridge regression estimators**. When $k_1 = k_2$, they are called the ridge regression estimators. Hoerl, Kennard, and Baldwin (1975) suggest the use of

$$k_1 = k_2 = 2s^2/(\hat{\gamma}_1^2 + \hat{\gamma}_2^2),$$

where $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ are the ordinary least squares estimators of γ_1 and γ_2, and s^2 is the residual mean square error from the least squares regression.]

14

CASE STUDY: COLLINEARITY PROBLEMS

Chapter 13 discussed methods of handling the collinearity problem. This chapter uses the Linthurst data to illustrate the behavior of ordinary least squares when collinearity is a problem. The correlational structure is then analyzed using principal component analysis and Gabriel's biplots. Finally, principal component regression is used and its limitations for the objective of this study are discussed.

This chapter gives the analysis of a set of observational data where collinearity is a problem. The purpose of this case study is (1) to demonstrate the inadequacies of ordinary least squares in the presence of collinearity, (2) to show the value of analyzing the correlational structure of the data, and (3) to demonstrate the use, and limitations, of principal component regression.

14.1 The Problem

This analysis is a continuation of the first case study (Chapter 5) which used five variables from the September sampling of the Linthurst data on Spartina BIOMASS production in the Cape Fear Estuary of North Car-

olina.[1] The objective of the study was to identify physical and chemical properties of the substrate that are influential in determining the widely varying aerial biomass production of Spartina in the Cape Fear Estuary. The sampling plan included three marshes in the estuary and three sites in each marsh representing three ecosystems: an area where the Spartina had previously died but had recently regenerated, an area consisting of short Spartina, and an area consisting of tall Spartina. In each of the nine sites, five random sampling points were chosen from which aerial biomass and the following physicochemical properties of the substrate were measured on a monthly schedule.

1. free sulfide (H_2S), moles
2. salinity (SAL), $^\circ/_{\circ\circ}$
3. redox potentials at pH 7 ($Eh7$), mv
4. soil pH in water (pH), 1:1 soil/water
5. buffer acidity at pH 6.6 (BUF), meg/100 cm^3
6. phosphorus concentration (P), ppm
7. potassium concentration (K), ppm
8. calcium concentration (Ca), ppm
9. magnesium concentration (Mg), ppm
10. sodium concentration (Na), ppm
11. manganese concentration (Mn), ppm
12. zinc concentration (Zn), ppm
13. copper concentration (Cu), ppm
14. ammonium concentration (NH_4), ppm.

Table 5.1 (page 163) gives the "Loc" and "Type" codes, the data for aerial biomass, and the five substrate variables used in that case study. Table 14.1 contains the data for the nine additional substrate variables for the September sampling date. The "Loc" and "Type" codes in Table 5.1 identify, respectively, the three islands in the Cape Fear Estuary and the nature of the Spartina vegetation at each sampling site; DVEG labels the recently regenerated areas, and TALL and SHRT identify the commonly labeled tall and short Spartina areas, respectively.

---

[1]The authors appreciate Dr. Rick A. Linthurst's permission to use the data and his contributions to this discussion.

TABLE 14.1. *Nine additional physicochemical properties of the substrate in the Cape Fear Estuary of North Carolina. Refer to Table 5.1 for aerial biomass (BIO) and the five physicochemical variables previously used. (Data used with permission of Dr. R. A. Linthurst.)*

OBS	H_2S	$Eh7$	BUF	P	Ca	Mg	Mn	Cu	NH_4
1	−610	−290	2.34	20.238	2,150.00	5,169.05	14.2857	5.02381	59.524
2	−570	−268	2.66	15.591	1,844.76	4,358.03	7.7285	4.19019	51.378
3	−610	−282	4.18	18.716	1,750.36	4,041.27	17.8066	4.79221	68.788
4	−560	−232	3.60	22.821	1,674.36	3,966.08	49.1538	4.09487	82.256
5	−610	−318	1.90	37.843	3,360.02	4,609.39	30.5229	4.60131	70.904
6	−620	−308	3.22	27.381	1,811.11	4,389.84	9.7619	4.50794	54.206
7	−590	−264	4.50	21.284	1,906.63	4,579.33	25.7371	4.91093	84.982
8	−610	−340	3.50	16.511	1,860.29	3,983.09	10.0267	5.11364	53.275
9	−580	−252	2.62	18.199	1,799.02	4,142.40	9.0074	4.64461	47.733
10	−610	−288	3.04	19.321	1,796.66	4,263.93	12.7140	4.58761	60.674
11	−540	−294	4.66	16.622	1,019.56	1,965.95	31.4815	1.74582	65.875
12	−560	−278	5.24	22.629	1,373.89	2,366.73	64.4393	3.21729	104.550
13	−570	−248	6.32	13.015	1,057.40	2,093.10	48.2886	2.97695	75.612
14	−580	−314	4.88	13.678	1,111.29	2,108.47	22.5500	2.71841	59.888
15	−640	−328	4.70	14.663	843.50	1,711.42	33.4330	1.85407	77.572
16	−610	−328	6.26	60.862	1,694.01	3,018.60	52.7993	3.72767	102.196
17	−600	−374	6.36	77.311	1,667.42	2,444.52	60.4025	2.99087	96.418
18	−630	−356	5.34	73.513	1,455.84	2,372.91	66.3797	2.41503	88.484
19	−640	−354	4.44	56.762	2,002.44	2,241.30	56.8681	2.45754	91.758
20	−600	−348	5.90	39.531	1,427.89	2,778.22	64.5076	2.82948	101.712
21	−640	−390	7.06	39.723	1,339.26	2,807.64	56.2912	3.43709	179.809
22	−650	−358	7.90	55.566	1,468.69	2,643.62	58.5863	3.47090	168.098
23	−630	−332	7.72	35.279	1,377.06	2,674.65	56.7497	3.60202	210.316
24	−640	−314	8.14	97.695	1,747.56	3,060.10	57.8526	3.92552	211.050
25	−630	−332	7.44	99.169	1,526.85	2,696.80	45.0128	4.23913	185.454
26	−620	−338	−0.42	3.718	6,857.39	1,778.77	16.4856	3.41143	16.497
27	−620	−268	−1.04	2.703	7,178.00	1,837.54	11.4075	3.43998	13.655
28	−570	−300	−1.12	2.633	6,934.67	1,586.49	7.9561	3.29673	17.627
29	−620	−328	−0.86	3.148	6,911.54	1,483.41	10.4945	3.11813	15.291
30	−570	−374	−0.90	2.626	6,839.54	1,631.32	9.4637	2.79145	14.750
31	−620	−336	3.72	16.715	1,564.84	3,828.75	10.3375	5.76402	95.721
32	−630	−342	4.90	16.377	1,644.37	3,486.84	21.6672	5.36276	86.955
33	−630	−328	2.78	21.593	1,811.00	3,517.16	13.0967	5.48042	83.935
34	−630	−332	3.90	18.030	1,706.36	4,096.67	15.6061	5.27273	104.439
35	−610	−322	3.60	34.693	1,642.51	3,593.05	6.9786	5.71123	79.773
36	−640	−290	3.58	28.956	2,171.35	3,553.17	57.5856	3.68392	118.178
37	−610	−352	5.58	25.741	1,767.63	3,359.17	72.5160	3.91827	123.538
38	−600	−280	6.58	25.366	1,654.63	3,545.32	64.4146	4.06829	135.268
39	−620	−290	6.80	17.917	1,620.83	3,467.92	53.9583	3.89583	115.417
40	−590	−328	5.30	20.259	1,446.30	3,170.65	22.6657	4.70368	108.406
41	−560	−332	1.22	134.426	2,576.08	2,467.52	51.9258	4.11065	57.315
42	−550	−276	1.82	35.909	2,659.36	2,772.99	75.1471	4.09826	77.193
43	−550	−282	1.60	38.719	2,093.57	2,665.02	71.0254	4.31487	68.294
44	−540	−370	1.26	33.562	2,834.25	2,991.99	70.1465	6.09432	71.337
45	−570	−290	1.56	36.346	3,459.26	3,059.73	89.2593	4.87407	79.383

Analysis of the full data set showed a serious collinearity problem in the data for every sampling date. The five variables used in Chapter 5—SAL, pH, K, Na, and Zn, were chosen from the larger data set to preserve some of the collinearity problem and yet reduce the dimension of the problem to a more convenient size for presentation. The multiple regression analysis of that subset of data with the five variables in the model showed significance only for pH. Backward elimination of one variable at a time led to a final model containing pH and K. In Chapter 7, all-possible regressions showed pH and Na to be the best two-variable model. Section 11.4 gave the residuals analysis, influence statistics, and the collinearity diagnostics for the model with these five variables.

In this chapter, *BIOMASS* is used as the dependent variable but all 14 physicochemical variables are investigated as independent variables. The primary objective of this research was to study the observed relationships of *BIOMASS* with the substrate variables with the purpose of identifying substrate variables that with further study might prove to be causal. As in Chapter 5, this analysis concentrates on the *total* variation over the 9 sites. The analysis of the "among-site" variation is left as exercises at the end of this chapter. The "within-site" variation can be studied in a similar manner.

Ordinary least squares is perhaps the most commonly used statistical tool for assessing importance of variables, and was the first method applied by the researcher. The results obtained, and reported here for the September data, were typical of ordinary least squares results in the presence of collinearity; the inadequacies of the method were evident. Principal component analysis and Gabriel's biplot are used here to develop an understanding of the correlational structure of the independent variables. To complete the case study, principal component regression is applied to the data to illustrate its use, and to show that biased regression methods suffer some of the same inadequacies as least squares when the purpose of the analysis is to identify "important" variables.

Although more and better data is the method of first choice for solving the collinearity problem, there will be situations where (1) it is not economically feasible with observational studies to obtain the kind of data needed to disrupt the near-singularities or (2) the near-singularities are a product of the system and will persist regardless of the amount of data collected. One purpose of this case study is to raise flags of caution on the use of least squares and biased regression methods in such cases. Biased regression methods can have advantages over least squares for estimation of the individual parameters, in terms of mean squared error, but suffer from the same problems as least squares when the purpose is identification of "important" variables.

TABLE 14.2. *Ordinary least squares regression of aerial biomass on 14 soil variables and stepwise regression results using the maximum R-squared option in PROC REG (SAS Institute Inc., 1989b). All independent variables are centered and standardized to have unit length vectors.*

Multiple Regression			*Maximum R-Squared*			
Soil Variable				*Variable Added (+)*		
X_j	$\widehat{\beta}_j$	$s(\widehat{\beta}_j)$	*Step*	*Removed (−)*	C_p	R^2
H_2S	89	610	1	$+pH$	21.4	.599
SAL	−591	646	2	$+Mg$	14.1	.659
$Eh7$	626	493	3	$+Ca$	5.7	.726
pH	2006	2764	4	$+Cu$	4.0	.750
BUF	−115	2059	5	$+P$	3.8	.764
P	−311	483	6	$+K, -P, +Zn$	3.8	.777
K	−2066	952	7	$+NH_4$	4.1	.788
Ca	−1322	1432	8	$+Eh7, -Zn, +P$	4.7	.797
Mg	−1746	1710	9	$+Zn, -P, +SAL$	5.6	.804
Na	203	1129	10	$+P$	7.2	.806
Mn	−272	873	11	$+Mn$	9.1	.807
Zn	−1032	1196	12	$+Na$	11.0	.807
Cu	2374	771	13	$+H_2S$	13.0	.807
NH_4	−848	1015	14	$+BUF$	15.0	.807

14.2 Multiple Regression: Ordinary Least Squares

The purpose of presenting this analysis is to illustrate the behavior of ordinary least squares in the presence of collinearity and to demonstrate the misleading nature of the results both for estimation of regression coefficients and for identification of important variables in the system.

Full Model Results

Ordinary least squares regression of *BIOMASS* on all 14 variables gave $R^2 = .807$. The regression coefficients and their standard errors are given in the first three columns of Table 14.2. Only 2 variables, K and Cu, have regression coefficients differing from zero by more than twice their standard error. Taken at face value, these results would seem to suggest that K and Cu are the only important variables in "determining" *BIOMASS*. However, the magnitude of the regression coefficients and their standard errors in any nonorthogonal data set depends on which other variables are included in the model. [Recall that pH was the only significant variable in the regression on the 5 variables (Chapter 5: salinity, pH, K, Na, and Zn.)] The conclusion that K and Cu are the only important variables is not warranted.

Stepwise Regression

To demonstrate the dependence of the least squares results on the method used, three stepwise variable selection options in PROC REG (SAS Institute, Inc., 1989b), maximum R-square (MAXR), backwards elimination

TABLE 14.3. *Regression of aerial biomass on 14 soil variables using backward elimination and stepwise regression options in PROC REG (SAS). All independent variables are centered and standardized to have unit length vectors.*

Backward Elimination				*Stepwise*			
Step	*Variable Removed(−)*	C_p	*Prob* $> F$	*Step*	*Variable Added (+) Removed(−)*	C_p	*Prob* $> F$
1	$-BUF$	13.0031	.9559	1	$+PH$	64.3294	.0001
2	$-H_2S$	11.0208	.8935	2	$+MG$	7.4217	.0094
3	$-NA$	9.0745	.8123	3	$+CA$	9.9068	.0031
4	$-MN$	7.1585	.7634	4	$+CU$	3.8339	.0572
5	$-P$	5.5923	.4891	5	$+P$	2.2881	.1384
6	$-SAL$	4.7860	.2505	6	$-P$		
7	$-EH7$	4.0826	.2335				
8	$-NH_4$	3.8077	.1731				
9	$-K$	4.3928	.1012				
10	$-ZN$	3.9776	.2061				

(BACKWARD), and STEPWISE, were used to simplify the model and select "important" variables. The results of the MAXR option are shown in the last columns of Table 14.2. The MAXR option follows a sequence of adding (and deleting) variables until all variables are included in the model. In this selection option, the fourth step where Cu was added was the first step for which the C_p statistic was less than p' ($C_p = 4.0$ with $p' = 5$). The major increases in R^2 had been realized at this point ($R^2 = .75$). Based on these results, one would choose the four-variate model consisting of pH, Mg, Ca, and Cu. Note that the two variables that were the *only* significant variables in the full model, Cu and K, entered in the fourth and sixth steps in the maximum R-square stepwise regression option.

The selection paths for backward elimination (with $SLS = .10$) and the stepwise options (with $SLE = .15$ and $SLS = .10$) are shown in Table 14.3. These two selection procedures, with the specified values of SLS and SLE, terminated at the same four-variate model consisting of pH, Mg, Ca, and Cu. Notice that the stepwise option would have retained P in the model if the default option of $SLS = .15$ had been used.

With the other 10 variables dropped, the magnitudes of the regression coefficients for the 4 retained variables and their standard errors changed considerably (Table 14.4). The coefficient for pH more than doubled, the coefficients for Mg and Ca nearly doubled, and the coefficient for Cu was halved. The standard errors for pH and Mg were reduced by $\frac{2}{3}$, and the standard errors for Cu and Ca by $\frac{1}{4}$ and $\frac{1}{3}$, respectively. Of these 4 variables, pH and Mg appear to be the more important, as judged by their early entry into the model and the ratio of their coefficients to their standard errors.

TABLE 14.4. *Estimated regression coefficients and their standard errors for the 4 independent variables chosen by the stepwise regression procedures compared to the estimates from the 14-variable model.*

	Estimates from 14-Variable Model		*Estimates from 4-Variable Model*	
Variable	$\widehat{\beta}_j$	$s(\widehat{\beta}_j)$	$\widehat{\beta}_j$	$s(\widehat{\beta}_j)$
pH	2,006	2,764	4,793	894
Mg	−1,746	1,710	−2,592	564
Ca	−1,322	1,432	−2,350	908
Cu	2,374	771	1,121	573

Correlation Matrix

Inspection of the correlation matrix, Table 14.5, reveals five variables with reasonably high correlations with *BIOMASS*; pH, *BUF*, Ca, Zn, and NH_4. Each of these five variables would appear important if used as the only independent variable, but none of these five were identified in the full model and only pH and Ca were revealed as important in stepwise regression. The other two variables declared important by the stepwise procedure, Mg and Cu, had correlations with *BIOMASS* of only −.38 and .09, respectively. The second most highly correlated variable with *BIOMASS*, *BUF*, was the last of the 14 variables to enter the model in the "MAXR" variable selection option.

Inconsistencies

The two stepwise regression methods suggest that future studies concentrate on pH, Mg, Ca, and Cu. On the other hand, ordinary least squares regression using all variables identified only K and Cu as the important variables, and simple regressions on one variable at a time identify pH, *BUF*, Ca, Zn, and NH_4. None of the results were satisfying to the biologist; the inconsistencies of the results were confusing and variables expected to be biologically important were not showing significant effects.

Ordinary least squares regression tends either to indicate that none of the variables in a correlated complex are important when all variables are in the model, or to arbitrarily choose one of the variables to represent the complex when an automated variable selection technique is used. A truly important variable may appear unimportant because its contribution is being usurped by variables with which it is correlated. Conversely, unimportant variables may appear important because of their associations with the real causal factors. It is particularly dangerous in the presence of collinearity to use the regression results to impart a "relative importance," whether in a causal sense or not, to the independent variables.

These seemingly inconsistent results are typical of ordinary least squares regression when there are high correlations or, more generally, near-linear dependencies among the independent variables. Inspection of the correlation matrix shows several pairs of independent variables with reasonably high correlations and three with $|r| \geq .90$. The largest absolute correlation,

TABLE 14.5. *Product moment correlations among all variables in the Linthurst September data.*

	BIO	H_2S	SAL	$Eh7$	pH	BUF	P	K
BIO	1.00							
H_2S	.33	1.00						
SAL	−.10	.10	1.00					
$Eh7$	.05	.40	.31	1.00				
pH	.77	.27	−.05	.09	1.00			
BUF	−.73	−.37	−.01	−.15	−.95	1.00		
P	−.35	−.12	−.19	−.31	−.40	.38	1.00	
K	−.20	.07	−.02	.42	.02	−.07	−.23	1.00
Ca	.64	.09	.09	−.04	.88	−.79	−.31	−.26
Mg	−.38	−.11	−.01	.30	−.18	.13	−.06	.86
Na	−.27	−.00	.16	.34	−.04	−.06	−.16	.79
Mn	−.35	.14	−.25	−.11	−.48	.42	.50	−.35
Zn	−.62	−.27	−.42	−.23	−.72	.71	.56	.07
Cu	.09	.01	−.27	.09	.18	−.14	−.05	.69
NH_4	−.63	−.43	−.16	−.24	−.75	.85	.49	−.12

	Ca	Mg	Na	Mn	Zn	Cu	NH_4
Ca	1.00						
Mg	−.42	1.00					
Na	−.25	.90	1.00				
Mn	−.31	−.22	−.31	1.00			
Zn	−.70	.35	.12	.60	1.00		
Cu	−.11	.71	.56	−.23	.21	1.00	
NH_4	−.58	.11	−.11	.53	.72	.93	1.00

r = -0.95, is between pH and buffer acidity, the first and last variables to enter the model in the "maximum R-squared" stepwise analysis. Any inference that pH is an important variable and buffer acidity is not is clearly an unacceptable conclusion. Other less obvious near-linear dependencies among the independent variables may also be influencing the inclusion or exclusion of variables from the model. The correlational structure of the independent variables makes any simple interpretation of the regression analyses unacceptable.

14.3 Analysis of the Correlational Structure

Purpose

The purpose of the analysis of the correlational structure is to gain insight into the relationships among the variables being studied and the causes of the collinearity problem. The analysis may suggest ways of removing some of the collinearity problem by obtaining more data or redefining variables. The improved understanding will identify the systems of variables that are closely related to the variation in the dependent variable and, hence, which sets of variables merit further study.

Principal Component Analysis

Inspection of the correlations among the independent variables in Table 14.5 reveals several reasonably high correlations. However, the correlations reveal only pairwise associations and provide an adequate picture of the correlational structure only in the simplest cases. A more complete understanding is obtained by using principal component analysis, or singular value decomposition, of the $n \times p$ matrix of the independent variables. For this purpose, the independent variables are centered and scaled so that the sum of squares of each independent variable is one; the vectors have unit length in n-space. (Refer to Sections 2.7, 2.8, and 13.1 for review of eigenanalysis, singular value decomposition, and construction of the principal component variables.)

The eigenvalues (λ_j) and eigenvectors (v_j) for these data are given in Table 14.6. The first principal component accounts for 35% of the dispersion in Z-space, $\lambda_1 / \sum \lambda_j = .35$, and is defined primarily by the complex of variables pH, BUF, Ca, Zn, and NH_4; these are the variables with the largest coefficients in the first eigenvector v_1. The second principal component, defined primarily by K, Mg, Na, and Cu, accounts for 26% of the dispersion. The four dimensions with eigenvalues greater than 1.0 account for 83% of the dispersion. (If all independent variables had been orthogonal, all eigenvalues would have been 1.0 and each would have accounted for 7% of the dispersion.)

With the singular value decomposition, the measures of collinearity can be used to assess the extent of the collinearity problem. The full impact will not be seen from the singular value decomposition of the centered and scaled matrix since collinearities involving the intercept have been

TABLE 14.6. *Eigenvalues and eigenvectors of the $\boldsymbol{Z}'\boldsymbol{Z}$ matrix for the 14 independent variables in the Linthurst September data. All variables were centered and standardized so that $\boldsymbol{Z}'\boldsymbol{Z}$ is the correlation matrix.*

Eigen-	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
values	4.925	3.696	1.607	1.335	.692	.500	.385
Eigen-							
vectors[a]	v_1	v_2	v_3	v_4	v_5	v_6	v_7
H_2S	.164	−.009	.232	−.690	.014	.422	−.293
SAL	.108	−.017	.606	.271	.509	−.008	−.389
$Eh7$	.124	−.225	.458	−.301	−.166	−.598	.308
pH	.408	.028	−.283	−.082	.092	−.190	−.056
BUF	−.412	−.000	.205	.166	−.162	.024	−.110
P	−.273	.111	−.160	−.200	.747	.018	.357
K	.034	−.488	−.023	−.043	−.062	.016	.073
Ca	.358	.181	−.207	.054	.206	−.427	−.117
Mg	−.078	−.499	−.050	.037	.103	−.034	.036
Na	.018	−.470	.051	.055	.240	.059	.160
Mn	−.277	.182	.020	−.483	.039	−.300	−.152
Zn	−.404	−.089	−.176	−.150	−.008	−.036	.062
Cu	.011	−.392	−.377	−.102	.064	−.075	−.549
NH_4	−.399	.026	−.011	.104	−.005	−.378	−.388

Eigen-	λ_8	λ_9	λ_{10}	λ_{11}	λ_{12}	λ_{13}	λ_{14}
values	.381	.166	.143	.0867	.0451	.0298	.0095
Eigen-							
vectors	v_8	v_9	v_{10}	v_{11}	v_{12}	v_{13}	v_{14}
H_2S	.087	.169	.296	.221	−.015	−.007	.080
SAL	−.081	−.174	−.227	.090	−.155	.095	−.089
$Eh7$	.299	−.225	.084	−.023	.055	.033	.023
pH	.033	.024	.147	.042	−.332	−.025	−.750
BUF	.159	.097	.103	.340	.455	−.354	−.478
P	.381	.077	−.018	−.035	.064	−.066	−.015
K	.112	.560	−.554	.219	−.029	.249	−.073
Ca	−.179	.189	.076	.508	.348	−.082	.306
Mg	−.173	−.012	.111	.119	−.400	−.689	.193
Na	−.459	.088	.439	−.219	.363	.275	−.144
Mn	−.524	.086	−.363	−.270	.076	−.172	−.141
Zn	−.211	−.438	.016	.572	−.217	.396	−.042
Cu	.305	−.376	−.129	−.194	.304	.000	.043
NH_4	.165	.420	.394	.132	.303	.232	.118

[a]The sum of squares of the elements in each eigenvector is 1. Thus, if a particular variable's contribution were spread equally over all components, the coefficients would be approximately $\pm 1/\sqrt{14} = \pm .27$.

eliminated. Nevertheless, the smaller eigenvalues (Table 14.6) show that there is very little dispersion in several dimensions. The last 4 principal component dimensions together account for only 1% of the dispersion in the Z-space; the last 6 principal components account for 3.4% of the total dispersion in the Z-space. Thus, there is very little dispersion in at least 6 dimensions of a nominal 14-dimensional space. The dimension with the least dispersion, $\lambda = .0095$, is due primarily to a linear restriction on pH, BUF, and Ca. The correlation between pH and BUF, $-.95$, is the correlation of highest magnitude among the independent variables (Table 14.5).

Based on a result of Hoerl and Kennard (1970a), the lower bound on the sum of the variances of estimated coefficients is $\sigma^2/\lambda_{14} = 105\sigma^2$. This is compared to $14\sigma^2$ if all independent variables had been pairwise orthogonal. The condition number for the matrix of centered variables is 22.8, above the value of 10 suggested as the point above which collinearity can be expected to cause problems. Thisted's (1980) measure of collinearity is

$$mci = \sum_{j=1}^{14} \lambda_j^{-2} \lambda_{14}^2 = 1.17$$

indicating severe collinearity. (Values of mci near 1.0 indicate high collinearity; values greater than 2.0 indicate little or no collinearity.) The variance inflation factors (VIF), the diagonal elements of $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$, also show the effects of collinearity. The largest VIF is 62 for pH, followed by 34.5 for BUF, 23.8 for Mg, 16.6 for Ca, and 11.6 for Zn. The smallest are 1.9 for P and 2.0 for $EH7$; these two variables are not seriously involved in the near-singularities. (If all independent variables were orthogonal, all VIFs would be 1.0.)

In summary, the dispersion of the sample points in at least four principal component dimensions is trivial, accounting for only 1.2% of the total dispersion. This limited dispersion in these principal component dimensions inflates the variances of regression coefficients for *all* independent variables involved in the near-singularities. The observed instability of the least squares regression estimates was to be expected.

Gabriel's Biplot

The major patterns of variation in the Z-space can be displayed by plotting the information contained in the major principal components. Gabriel's (1971) biplot using the first two principal components shows the structure of the Z matrix as "seen" in these two dimensions, Figure 14.1. This biplot of the first and second principal components accounts for 61% of the dispersion in the original 14-dimensional Z-space.

Vectors in the biplot are projections of the original variable vectors (in the 14-dimensional subspace they define) onto the plane defined by the first two principal components. The original vectors were scaled to have unit length. Therefore, the length of each *projected* vector is its correlation with the original vector and reflects the closeness of the original vector to the plane. Thus, the longest vectors, Ca, pH, Mg, Na, K, Zn, BUF,

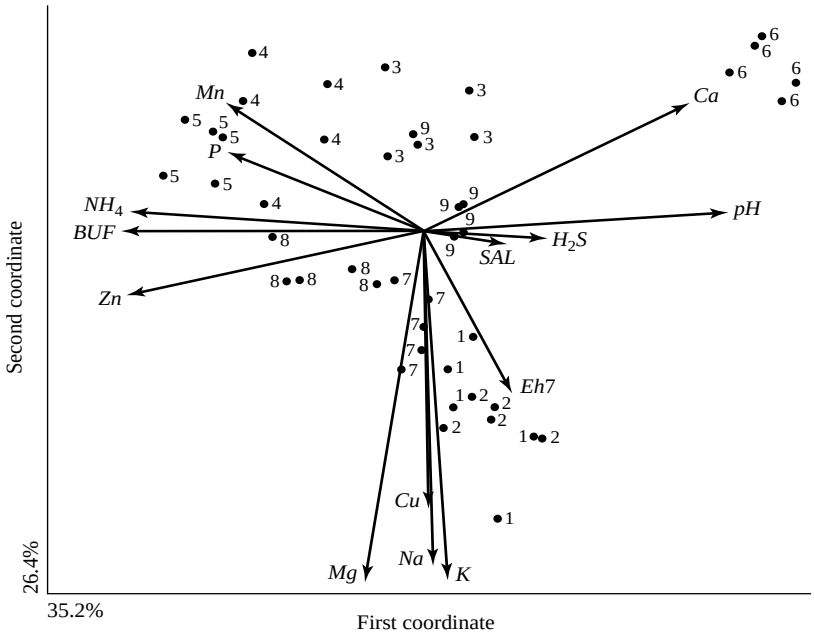

FIGURE 14.1. *Gabriel's biplot of the first and second principal components of the 14 marsh substrate variables. The variables have been centered and scaled so that all vectors have unit length in the original 14-dimensional Z-space. The first and second components account for 35.2% and 26.4% of the dispersion in the Z-space. Column markers are shown with the vectors, row markers with* •.

and NH_4, indicate the variables are close to the plane being plotted and, consequently, their relationships are well represented by the biplot. The shorter vectors in the biplot, H_2S, SAL, and $EH7$, identify variables that are more nearly orthogonal to this plane and, therefore, not well represented by this biplot. The other vectors, Mn, P, and Cu, are intermediate and relationships in this biplot involving these variables should be interpreted with caution.

The near-zero angle between the Ca and pH vectors, Figure 14.1, shows that the two variables are highly positively correlated ($r = .88$, Table 14.5) as are the three variables NH_4, BUF, and Zn ($r \geq .71$) and the three variables Mg, Na, and K ($r \geq .79$). Ca is highly negatively correlated with BUF and Zn ($r \leq -.70$), as is pH with BUF, Zn, and NH_4 ($r \leq -.72$); the angles are nearly 180°. On the other hand, pH, NH_4, BUF, and Zn are nearly orthogonal to K and Na. The angles between these vectors are close to 90° and the highest correlation is $r = .12$. The Cu and Na vectors illustrate the caution needed in interpreting associations between vectors that are not close to unity in length. Even though the angle between the two vectors is close to zero in this biplot, the correlation beteen Cu and Na is only .56 (Table 14.5). This apparent inconsistency is because the Cu vector is not well represented by this biplot as indicated by the projected Cu vector being appreciably shorter than unity.

More important than the pairwise associations are the two systems of variables revealed by this biplot. The five variables Ca, pH, NH_4, BUF, and Zn strongly associated with the first principal component axis behave as one system; the three variables Mg, Na, and K, which are strongly associated with the second principal component axis, behave as another. The two sets of variables are nearly orthogonal to each other.

The points in the biplot reflect the relative spatial similarities of the observations (or rows) of the $\boldsymbol{Z}$ matrix. The number label indicates the sampling site. This biplot indicates that the five samples labeled 6 are very similar to each other and very different from all other samples. The other points also show a distinct tendency to group according to sampling site. The perpendicular projection of the points onto each variable vector shows the relative values of the observations for that variable. Thus, the observations labeled 6 differ from the other points primarily because of their much higher values of Ca and pH and lower values of NH_4, BUF, and Zn. On the other hand, observations labeled 1 and 2 tend to be high and observations labeled 3, 4, 5, and 6 tend to be low in Mg, Na, and K.

Since the first two dimensions account for only 61% of the dispersion in Z-space, it is of interest to study the behavior in the third dimension. The first three dimensions account for 73% of the dispersion. Gabriel's biplot of the first and third dimensions (Figure 14.2) shows how the vectors in Figure 14.1 deviate above and below the plane representing the first two dimensions. The vectors primarily responsible for defining the second dimension now appear very short because the perspective in Figure 14.2

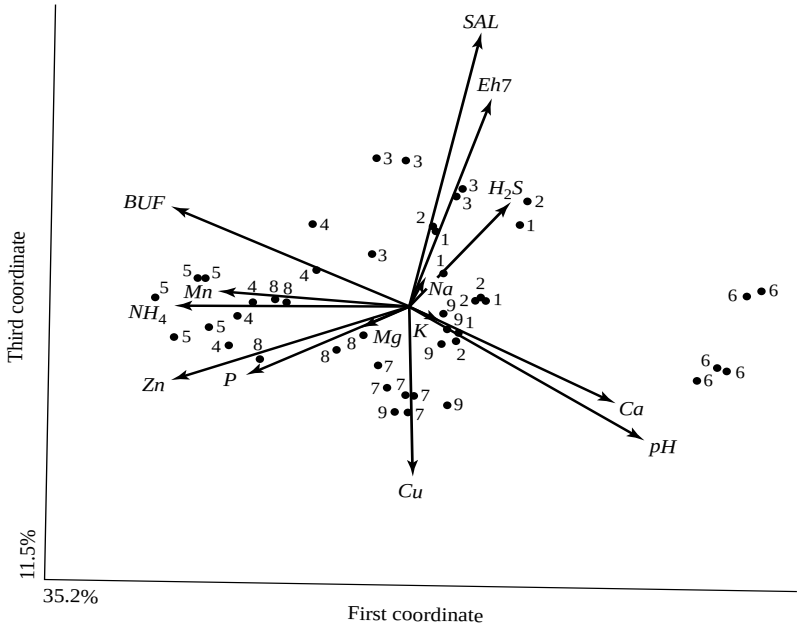

FIGURE 14.2. *Gabriel's biplot of the first and third principal components of the 14 marsh substrate variables. The variables are centered and scaled so that all vectors have unit length in 14-dimensional Z-space. The third principal component accounted for 11.5% of the dispersion in the Z-space. Column markers are shown with the vectors, row markers with the* •.

is down the second axis; only the deviations from the plane of the Na, K, and Mg vectors are observed. The third dimension is defined primarily by SAL with some impact from $EH7$ and Cu.

The fourth principal component is dominated by H_2S and Mn and accounts for 10% of the dispersion. The fifth is dominated by P and SAL and accounts for 5% of the dispersion, and so on. Gabriel's biplot also could be used to view these dimensions. The principal component analysis and Gabriel's biplots show that the major variation in the Z-space is accounted for by relatively few complexes of substrate variables, variables that have a strong tendency to vary together. Interpretation of the associations with *BIOMASS* should focus on these complexes rather than on the individual variables.

Regression of BIOMASS on Principal Components

The relationship of *BIOMASS* to the principal component variables can be determined by regressing *BIOMASS* on the principal components. (This is the first step of principal component regression but is presented here to see how *BIOMASS* fits into the principal component structure. Conversion of the regression coefficients for the principal components to the regression

TABLE 14.7. *Regression of aerial biomass on individual principal components W_j. (Linthurst September data.)*

W_j	*SS(Regr)*	$F\ (df = 1, 30)^a$
1	10,117,269	82.2∗
2	1,018,472	8.3∗
3	1,254,969	10.2∗
4	496,967	4.0
5	215,196	1.8
6	10,505	.1
7	267,907	2.2
8	675,595	5.5∗
9	803,786	6.5∗
10	110,826	.9
11	430,865	3.5
12	40,518	.3
13	2,160	.0
14	34,892	.3

[a]A * indicates significance at $\alpha = .05$ using as error the residual mean square from the full model.

coefficients for the original variables are completed in Section 14.4.) The sums of squares due to regression and the tests of significance of the principal components are given in Table 14.7. The first principal component W_1 dominates the regression, accounting for 65% of the regression sum of squares. (Note that the first principal component is defined so as to account for the greatest dispersion in the Z-space, but it does not follow that W_1 will necessarily be the best predictor of *BIOMASS*.) W_2, W_3, W_8, and W_9 also account for significant ($\alpha = .05$) amounts of variation in *BIOMASS*.

Correlations of Independent Variables with Principal Components

For ease of relating the principal components to the original variables, the correlations of each of the 14 original independent variables with these five principal components are given in Table 14.8. The importance of the first principal component in the regression strongly suggests that the pH–BUF–Ca–Zn–NH_4 complex of 5 variables be given primary consideration in future studies of the causes of variation in *BIOMASS* production of Spartina. Perhaps P and Mn should be included in this set for consideration because of their reasonably high correlations with W_1. Variables of secondary importance are K, Mg, Na, and Cu, which are highly correlated with W_2, and SAL and $EH7$, which are reasonably highly correlated with W_3. W_2 and W_3 account for 7% and 8%, respectively, of the regression sum of squares for *BIOMASS*. H_2S is the only variable not highly correlated with at least one of the three most predictive principal components.

Dispersion in Z-Space

The principal component analysis of the centered and standardized independent variables has demonstrated that most of the dispersion in Z-space

TABLE 14.8. *Correlations between original independent variables X_k and the significant principal components W_j: $\widehat{\rho}(X_k, W_j) = v_{jk}\lambda_j$.*

Variable	*Principal Component*				
X_k	W_1	W_2	W_3	W_8	W_9
H_2S	.364	−.017	.294	.054	.069
SAL	.240	−.033	.768	−.050	−.071
$Eh7$	.275	−.433	.581	.185	−.092
pH	.905	.054	−.359	.020	.010
BUF	−.914	.000	.260	.098	.040
P	.606	.213	−.203	.235	.031
K	.075	−.938	−.029	.069	.228
Ca	.794	.348	−.262	−.110	.077
Mg	−.173	−.959	−.063	−.107	−.005
Na	.040	−.904	.065	−.283	.036
Mn	−.615	.350	.025	−.323	.035
Zn	−.897	−.171	−.223	−.130	−.178
Cu	.024	−.754	−.478	.188	−.153
NH_4	−.885	.050	−.014	.102	.171

can be described by a few complexes of correlated variables. One of these systems, W_1, accounts for a major part of the variation in *BIOMASS*. This complex includes the five variables most highly correlated individually with *BIOMASS*. Three other complexes account for significant but much smaller amounts of variation in *BIOMASS*. The analysis does *not* identify which variable in the complex is responsible for the association. The principal component analysis shows that the data do not contain information that will allow separation of the effects of the individual variables in each complex.

Eliminating Variables to Control Collinearity

A pseudosolution to the collinearity problem would be to eliminate from the regression model enough independent variables to remove the collinearity. This would be equivalent to retaining one independent variable to represent each major dimension of the original X-space. Variable selection techniques in ordinary least squares regression are, in effect, doing this in a somewhat arbitrary manner. Eliminating variables is not a viable solution when the primary interest is in identifying the important variables. The correlated complexes of variables still exist in nature; it is only that they are no longer "seen" by the regression analysis. It is likely that some of the truly important variables will be lost with such a procedure.

14.4 Principal Component Regression

Principal component regression has been suggested as a means of obtaining estimates with smaller mean squared errors in the presence of collinearity. Results from principal component regression are presented for this example to illustrate the impact the method has on stability of the estimates and the inadequacy of the method for assigning relative importance to the independent variables. The reader is referred to Section 13.2 for a review of this method.

Deleting Principal Components

The principal component analysis for these data revealed that the six dimensions of the Z-space having the least dispersion accounted for only 3.4% of the total dispersion in Z-space. Regression of *BIOMASS* on the principal components and tests of significance of the *principal component* regression coefficients revealed that, of these six, only W_9 had significant predictive value for *BIOMASS* (Table 14.7). Using the rule that principal components which have small eigenvalues and contain no predictive information for Y should be eliminated, the five principal components corresponding to the five smaller eigenvalues W_{10} to W_{14} were deleted for the principal component regression; the first nine principal components, $g = 9$, were retained.

Deleting these five principal components results in a loss of 2.2% of the dispersion in Z-space, a loss in predictive value of Y from $R^2 = .807$ to $R^2 = .7754$, a decrease in $\text{tr}[\mathbf{Var}(\boldsymbol{\beta}^+)]/\sigma^2$ from 196 to 17, and a decrease in $(\boldsymbol{\beta}^{+\prime}\boldsymbol{\beta}^+)^{1/2}$ from 4,636 to 3,333 (Table 14.9). The stability of the regression estimates increased greatly with an acceptable loss in apparent predictability of *BIOMASS*.

It is of interest to follow the sequential change in these quantities as individual principal components are deleted from the regression (Table 14.9). There is virtually no loss in predictability when W_{12}, W_{13}, and W_{14} are deleted (see R^2, Table 14.9). The variances of the estimates decrease dramatically, particularly with elimination of the 14th principal component (see $\text{tr}[\mathbf{Var}(\boldsymbol{\beta}^+)]/\sigma^2$, Table 14.9). Since W_9 and W_8 are significant, none of the results where W_9 to W_1 have been eliminated would be used. They are presented here only to show the entire pattern.

Regression with Nine Principal Components

The first 9 principal components, $g = 9$, were used in principal component regression. The regression coefficients for the 9 principal components were converted to estimates of the regression coefficients for the 14 original variables, $\boldsymbol{\beta}^+_{(g)} = \boldsymbol{V}_{(g)}\widehat{\boldsymbol{\gamma}}_{(g)}$. The results are given in the last two columns of Table 14.10. Eight of the 14 regression coefficients for the independent variables are significant, pH, BUF, K, Mg, Na, Mn, Cu, and NH_4. (Results from ordinary least squares, $g = 14$, and from the first 11 principal components, $g = 11$, are included for comparison.) The variables pH, BUF, and NH_4 are significant primarily because of their contribution to W_1; K, Mg, and Na are significant primarily through W_2. The significance of Cu

TABLE 14.9. *Cumulative effect of deleting principal components in principal component regression starting with the principal component with the least dispersion, W_{14}. (Linthurst September data.)*

Component Deleted	*Information Loss in $\boldsymbol{X'X}$ (%)*	$100R^2$	$\frac{tr[\mathbf{Var}(\boldsymbol{\beta}^+)]}{\sigma^2}$	$(\boldsymbol{\beta}^{+\prime}\boldsymbol{\beta}^+)^{1/2}$
None(OLS)	.0	80.7	196	4,636
14	.1	80.6	91	4,226
13	.3	80.6	57	4,218
12	.6	80.3	35	4,111
11	1.2	78.1	23	3,451
10	2.2	77.5	17	3,333
9	3.4	73.3	10	2,507
8	6.2	69.8	8	2,151
7	8.9	68.4	5	1,952
6	12.5	68.3	3	1,948
5	17.4	67.2	1.8	1,866
4	26.8	64.6	1.1	1,763
3	38.4	58.1	.5	1,526
2	64.8	52.8	.2	1,434

and Mn appears to come through their contributions to several principal components. On the other hand, even though Ca and Zn are major components of W_1 and SAL is a major component of W_3, their contributions to *BIOMASS* through several W_j apparently tend to cancel and make then nonsignificant.

Comparison with Ordinary Least Squares

The increased stability of the principal component regression estimates compared to ordinary least squares is evident in Table 14.10. The cost of the increased stability is a loss in R^2 from .807 to .775, and an introduction of an unknown amount of bias. It is hoped that the decrease in variance is sufficient to more than compensate for the bias so that the principal component estimates will have smaller mean squared error. The large decreases in variance for several of the coefficients makes this a reasonable expectation.

The principal component regression has little impact on the regression coefficients for the variables that are not involved in the near-singularities. The regression coefficients and standard errors for $EH7$ and P change relatively little. These two variables have small coefficients for all five principal components eliminated from the principal component regression. All other variables are involved in one or more of the near-singularities.

Judging Importance of Variables

The purpose of this study was to identify "important" variables for further study of the causal mechanisms of *BIOMASS* production. It is dan-

TABLE 14.10. *Principal component regression estimates of regression coefficients and standard errors using $g = 14$ (OLS), 11 and 9 principal components. (Linthurst September data.)*

	$g = 14$ *(OLS)*[a]		$g = 11$		$g = 9$	
Variable	$\widehat{\boldsymbol{\beta}}$	$\boldsymbol{s}(\widehat{\boldsymbol{\beta}})$	$\boldsymbol{\beta}^+$	$\boldsymbol{s}(\boldsymbol{\beta}^+)$	$\boldsymbol{\beta}^+$	$\boldsymbol{s}(\boldsymbol{\beta}^+)$
H_2S	88	610	257	538	489	379
SAL	−591	645	−639	458	−238	393
$Eh7$	626	493	609	473	482	465
pH	2005	2763	896∗	210	858∗	152
BUF	−117	2058	−1364∗	459	−685∗	183
P	−312	483	−383	449	−445	446
K	−2069∗	952	−2247∗	761	−1260∗	495
Ca	−1325	1431	−1046	690	30	317
Mg	−1744	1709	−817∗	228	−652∗	145
Na	203	1128	−488	577	−1365∗	317
Mn	−274	872	−570	604	−848∗	385
Zn	−1031	1195	−1005	791	251	410
Cu	2374∗	771	2168∗	563	1852∗	500
NH_4	−847	1015	−400	621	−1043∗	479
R^2	.807		.803		.775	
VIF_{max}	62.1		5.1		2.0	

[a]A * indicates the estimate exceeds twice its standard error.

gerous to attempt to assign "relative importance" to the variables based on the relative magnitudes of their partial regression coefficients. This is the case whether the estimates are from ordinary least squares or principal component regression. The least squares estimates are too unstable in this example to give meaningful results. Principal component regression estimates are a pooling of the least squares estimates for all variables involved in the strong collinearities (see equation 13.45). The greater stability of the biased regression estimates can be viewed as coming from this "averaging" of information from the correlated variables. However, this does not prove helpful in judging the relative importance of variables in the same correlated complex.

Complexes of Variables

Principal component analysis has shown that the independent variables in this set of data behave as correlated complexes of variables with meaningful variation in only 9 dimensions of the 14-dimensional space. The W_1 complex of variables, for example, behaves more or less as a unit in this data set, and it would be inappropriate to designate any one of the five variables as "the variable of importance." It is the complex that must, for the moment at least, be considered of primary importance insofar as *BIOMASS* is concerned. Further research under controlled conditions where the effect of the individual variables in the complex can be disassociated is needed before specific causal relationships can be defined.

14.5 Summary

The classical results of ordinary least squares regression in the presence of collinearity are demonstrated with the Linthurst data; either all variables of a correlated complex appear insignificant, if a full multiple regression model is fit, or only one variable in each correlated complex is retained if some stepwise regression procedure is used. In either case, any inference as to which are the "important" variables can be very misleading. The apparent insignificance of the variables arises from the fact that the near-singularities in $\boldsymbol{X}$, reflected in the near-zero eigenvalues, cause the ordinary least squares estimates of the regression coefficients to be very unstable. Geometrically, there is only trivial dispersion of the data in one or more dimensions of the $\boldsymbol{Z}$-space and, consequently, the impact of these dimensions on the dependent variable is determined only with very low precision. Conversely, the dimensions of the Z-space showing major dispersion are defined by sets of correlated variables. Ordinary least squares somewhat arbitrarily picks one of the variables to represent the complex. If the objective is simply to predict *BIOMASS*, such a procedure is satisfactory as long as care is taken in making predictions. However, when the objective is to identify "important" variables, such a procedure will be misleading.

Principal component analysis and Gabriel's biplot clarify the complex relationships among the independent variables. Correlated complexes of variables can be identified and their associations with the dependent variable assessed. The primary variables in the complexes that have predictive value can then be studied under controlled conditions to determine their effects on the dependent variable. Principal component regression, although it may be useful in some cases for estimating regression coefficients, does not prove helpful in assigning relative importance to the independent variables involved in the near-singularities.

14.6 Exercises

The singular value decomposition of the Linthurst data in this case study was run on the 45×14 matrix of individual observations on the 14 independent variables. That analysis operated on the total variation within and among sampling sites. The following exercises study the correlational structure among the independent variables and their relationship to *BIOMASS* production using only the variation *among* sampling sites. The data to be used are the sampling site means for all variables computed from the data in Table 14.1. The "Loc–Type" codes identify the nine sampling sites.

14.1. Compute the 9×15 matrix of sampling site means for *BIOMASS* and the 14 independent variables. Center and standardize the matrix of means and compute the correlation matrix of all 15 variables. Which independent variables appear to be most highly correlated with *BIOMASS*? Identify insofar as possible the subsets of independent variables that are highly correlated with each other. Are there any independent variables that are nearly independent of the others?

14.2. Extract from the 9×15 matrix of centered and standardized variables the 14 independent variables to obtain $\boldsymbol{Z}$. Do the principal component analysis on this matrix. Explain why only eight eigenvalues are nonzero. Describe the composition (in terms of the original variables) of the three principal components that account for the most dispersion. What proportion of the dispersion do they account for? Compare these principal components to those given for the case study using all observations.

14.3. Drop *BUF* and NH_4 from the data set and repeat Exercise 14.2. Describe how the principal components change with these two variables omitted. Notice that the two variables dropped were primary variables in the first principal component computed with all variables, Exercise 14.2.

14.4. Use the principal components defined in Exercise 14.2 to construct Gabriel's biplot. Use enough dimensions to account for 75% of the

dispersion in Z-space. Interpret the biplots with respect to the correlation structure of the variables, the similarity of the sampling sites, and the major differences in the sampling sites.

14.5. Use the first eight principal components defined in Exercise 14.2 as independent variables and the sampling site means for *BIOMASS* as the dependent variable. Regress *BIOMASS* on the principal components (plus an intercept) and compute the sum of squares attributable to each principal component. These sums of squares, multiplied by five to put them on a "per observation" basis, are an orthogonal partitioning of the "among site" sum of squares. Compute the analysis of variance for the original data to obtain the "among site" and "within site" sums of squares. Verify that the "among site" sums of squares computed by the two methods agree. Test the significance of each principal component using the "within site" mean square as the estimate of σ^2. Which principal component dominates the regression and which variables does this result suggest might be most important? Which principal component is nearly orthogonal to *BIOMASS* and what does this imply, if anything, about some of the variables?

15

MODELS NONLINEAR IN THE PARAMETERS

Chapter 14 completed the series of chapters devoted to problem areas in least squares regression. This chapter returns to regression methods for fitting a variety of models. Chapter 8 introduced the use of polynomial and trigonometric response models for characterizing responses that cannot be adequately represented by straight-line relationships. This chapter extends those ideas to the large class of usually more realistic models that are nonlinear in the parameters. First, several examples of nonlinear models are given. Then regression methods for fitting these models are presented.

The models considered to this point have been linear functions of the parameters. This means that each (additive) term in the model contains only one parameter and only as a multiplicative constant on the independent variable (or function of the independent variable). This restriction excludes many useful mathematical forms, including nearly all models developed from principles of behavior of the system being studied. These linear models should be viewed as first-order approximations to the true relationships.

In this chapter, the class of models is extended to the potentially more realistic models that are **nonlinear** in the parameters. Emphasis is placed on the functional form of the model, the part of the model that gives the relationship between the expectation of the dependent variable and the independent variables. Whenever model development goes beyond the simple

summarization of the relationships exhibited in a set of data, it is likely that models nonlinear in the parameters will come under consideration. The use of prior information on the behavior of a system in building a model will often lead to nonlinear models. This prior information may be nothing more than recognizing the general shape the response curve (surface) should take. For example, it may be that the response variable should not take negative values, or the response should approach an asymptote for high or low values of an independent variable. Imposing these constraints on a system will usually lead to nonlinear models.

At the other extreme, prior information on the behavior of a system may include minute details on the physical and chemical interactions in each of several different components of the system and on how these components interact to produce the final product. Such models can become extremely complex and most likely cannot be written as a single functional relationship between $\mathcal{E}(Y)$ and the independent variables. The detailed growth models that predict crop yields based on daily, or even hourly, data on the environmental and cultural conditions during the growing season are examples of such models. (The development of such models is not pursued in this text. They are mentioned here as an indication of the natural progression of the use of prior information in model building.)

Although this chapter does not dwell on the behavior of the residuals, it is important that the assumptions of least squares be continually checked. Growth data, for example, often will not satisfy the homogeneous variance assumption, and will contain correlated errors if the data are collected as repeated measurements over time on the same experimental units.

15.1 Examples of Nonlinear Models

Form of the Model

The more general class of models that are nonlinear in the parameters allows the mean of the dependent variable to be expressed in terms of any function $f(\boldsymbol{x}_i'; \boldsymbol{\theta})$ of the independent variables and the parameters. The model becomes

$$Y_i = f(\boldsymbol{x}_i'; \boldsymbol{\theta}) + \epsilon_i, \tag{15.1}$$

where $f(\boldsymbol{x}_i'; \boldsymbol{\theta})$ is the nonlinear function relating $\mathcal{E}(Y)$ to the independent variable(s), $\boldsymbol{x}_i'$ is the row vector of observations on k independent variables for the ith observational unit, and $\boldsymbol{\theta}$ is the vector of p parameters. (It is common in nonlinear least squares to use $\boldsymbol{\theta}$ as the vector of parameters rather than $\boldsymbol{\beta}$.) The usual assumptions are made on the random errors. That is, ϵ_is are assumed to be independent $N(0, \sigma^2)$ random variables.

A sample of nonlinear models is presented to illustrate the types of functions that have proven useful and to show how information on the system can be used to develop more realistic models. Nonlinear models are usually chosen because they are more realistic in some sense or because the

functional form of the model allows the response to be better characterized, perhaps with fewer parameters. The procedures for estimating the parameters, using the least squares criterion, are discussed in Section 15.2.

Exponential Decay Model

In many cases the rate of change in the mean level of a response variable at any given point in time (or value of the independent variable) is expected to be proportional to its value or some function of its value. Such information can be used to develop a response model. Models developed in this manner often involve exponentials in some form. For example, assume that the concentration of a drug in the bloodstream is being measured at fixed time points after the drug was injected. The response variable is the concentration of the drug; the independent variable is time (t) after injection. If the rate at which the drug leaves the bloodstream is assumed to be proportional to the mean concentration of the drug in the bloodstream at that point in time, the derivative of $\mathcal{E}(Y)$, drug concentration, with respect to time t is

$$\frac{\partial \mathcal{E}(Y)}{\partial t} = -\beta \mathcal{E}(Y). \tag{15.2}$$

Integrating this differential equation, and imposing the condition that the concentration of the drug at the beginning ($t = 0$) was α gives

$$\mathcal{E}(Y) = \alpha\, e^{-\beta t}. \tag{15.3}$$

This is the **exponential decay curve**. If additive errors are assumed, the nonlinear model for a process that operates in this manner would be

$$Y_i = \alpha e^{-\beta t_i} + \epsilon_i. \tag{15.4}$$

This is a two-parameter model with $\boldsymbol{\theta}' = (\alpha\ \beta)$. If multiplicative errors are assumed,

$$Y_i = \alpha(e^{-\beta t_i})\epsilon_i. \tag{15.5}$$

The latter is intrinsically linear and is linearized by taking logarithms as discussed in Section 12.2. The model with additive errors, however, cannot be linearized with any transformation and, hence, is intrinsically nonlinear. The remaining discussions in this chapter assume the errors are additive.

Exponential Growth Model

The rate of growth of bacterial colonies might be expected to be proportional to the size of the colony if all cells are actively dividing. The partial derivative in this case would be

$$\frac{\partial \mathcal{E}(Y)}{\partial t} = \beta \mathcal{E}(Y). \tag{15.6}$$

This is the positive version of equation 15.2, reflecting the expected *growth* of this system. This differential equation yields the **exponential growth model**

$$Y_i = \alpha e^{\beta t_i} + \epsilon_i, \tag{15.7}$$

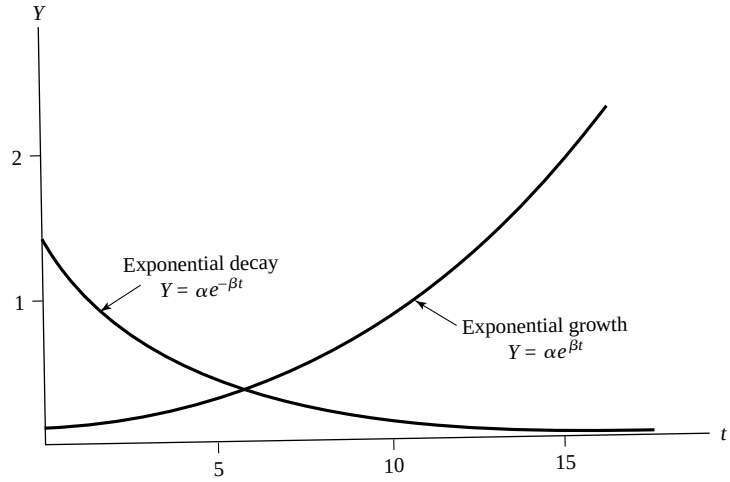

FIGURE 15.1. *Typical forms for the exponential decay model and the exponential growth model. The parameter β is positive in both cases.*

where α is the size of the colony at $t = 0$. In both models β is positive; the sign in front of β indicates whether it is an exponential *decay* process or an exponential *growth* process. Their general shapes are shown in Figure 15.1.

Two-Term Exponential Model

A **two-term exponential model** results when, for example, a drug in the bloodstream is being monitored and the amount in the bloodstream depends on two processes, the movement into the bloodstream from muscle tissue or the digestive system and removal from the bloodstream by, say the kidneys. Let the amount of the drug in the source tissue be $\mathcal{E}(Y_m)$ and that in the blood be $\mathcal{E}(Y_b)$. Suppose the drug moves into the bloodstream from the muscle at a rate proportional to its amount in the muscle $\theta_1\mathcal{E}(Y_m)$ and is removed from the bloodstream by the kidneys at a rate proportional to its amount in the bloodstream $-\theta_2\mathcal{E}(Y_b)$. Assume $\theta_1 > \theta_2 > 0$. The net rate of change of the drug in the bloodstream is

$$\frac{\partial\mathcal{E}(Y_b)}{\partial t} = \theta_1\mathcal{E}(Y_m) - \theta_2\mathcal{E}(Y_b). \tag{15.8}$$

Assume the initial amount in the muscle (at $t = 0$) is $\mathcal{E}(Y_{m_0}) = 1$. This process models the amount in the blood stream as

$$Y_{bi} = \frac{\theta_1}{\theta_1 - \theta_2}\left(e^{-\theta_2 t_i} - e^{-\theta_1 t_i}\right) + \epsilon_i. \tag{15.9}$$

This response curve shows an increasing amount of the drug in the blood in the early stages, which reaches a maximum and then declines asymptotically toward zero as the remnants of the drug are removed. This model

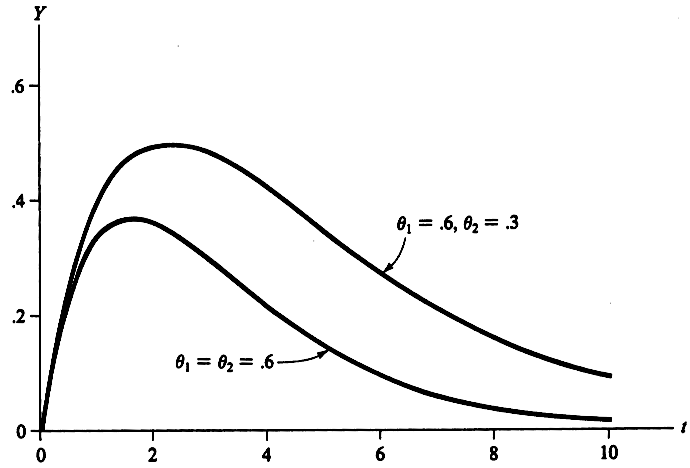

FIGURE 15.2. *The two-term exponential model with* $\theta_1 = .6$ *and* $\theta_2 = .3$*, equation 15.9; and its simpler form with* $\theta_1 = \theta_2 = .6$*, equation 15.10.*

would also apply to a process where one chemical is being formed by the decay of another, at reaction rate θ_1 and is itself decaying at reaction rate θ_2. If $\theta_1 = \theta_2$, the solution to the differential equations gives the model

$$Y_i = \theta_1 t_i e^{-\theta_1 t_i} + \epsilon_i. \tag{15.10}$$

The forms of these models are shown in Figure 15.2.

Mitscherlich Growth Model

When the increase in yield (of a crop) per unit of added nutrient X is proportional to the difference between the maximum attainable yield α and the actual yield, the partial derivative of Y with respect to X is

$$\frac{\partial \mathcal{E}(Y)}{\partial X} = \beta[\alpha - \mathcal{E}(Y)]. \tag{15.11}$$

This partial derivative generates the model known as the **Mitscherlich equation** (Mombiela and Nelson, 1981):

$$Y_i = \alpha[1 - e^{-\beta(X_i + \delta)}] + \epsilon_i, \tag{15.12}$$

where δ is the equivalent nutrient value of the soil. This model gives an estimated mean yield of

$$\widehat{Y} = \widehat{\alpha}(1 - e^{-\widehat{\beta\delta}}) \tag{15.13}$$

Monomolecular Growth Model

with no added fertilizer and an asymptotic mean yield of $\widehat{Y} = \widehat{\alpha}$ when the amount of added fertilizer is very high. If $\gamma = e^{-\beta\delta}$ is substituted in equation 15.12, this model takes the more familiar form known as **monomolecular growth model**. The form of the Mitscherlich equation is shown in Figure 15.3.

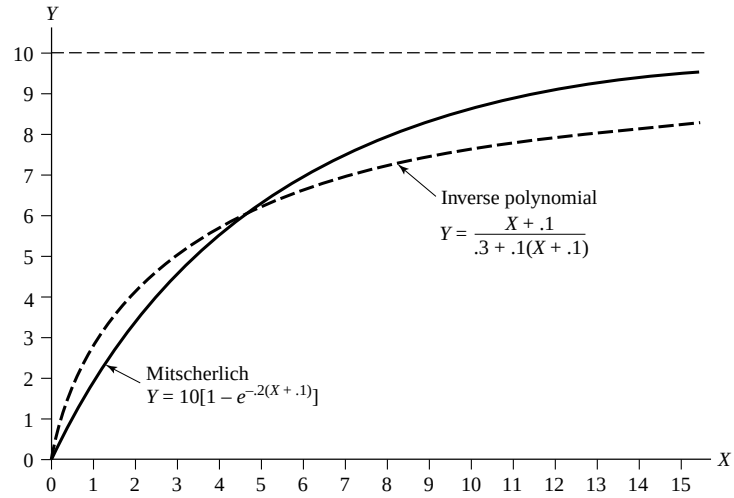

FIGURE 15.3. *The form of the Mitscherlich and inverse polynomial models. The parameter α in the Mitscherlich equation is the upper asymptote and β controls the rate at which the asymptote is approached. The inverse polynomial model approaches its asymptote of $1/\beta_1$ very slowly and at a decreasing rate determined by $\beta_0/[\beta_0 + \beta_1(X + \delta)]^2$.*

If the rate of increase in yield is postulated to be proportional to the square of $[\alpha - \mathcal{E}(Y)]$, one obtains the **inverse polynomial model** (Nelder, 1966),

Inverse Polynomial Model

$$Y_i = \frac{X_i + \delta}{\beta_0 + \beta_1(X_i + \delta)}. \tag{15.14}$$

The inverse polynomial model is also shown in Figure 15.3.

The **logistic** or **autocatalytic growth function** results when the rate of growth is proportional to the product of the size at the time and the amount of growth remaining:

Logistic Growth Model

$$\frac{\partial \mathcal{E}(Y)}{\partial t} = \frac{\beta \mathcal{E}(Y)[\alpha - \mathcal{E}(Y)]}{\alpha}. \tag{15.15}$$

This differential equation gives the model

$$Y_i = \frac{\alpha}{1 + \gamma e^{-\beta t_i}} + \epsilon_i, \tag{15.16}$$

which has the familiar S-shape associated with growth curves. The curve starts at $\alpha/(1 + \gamma)$ when $t = 0$ and increases to an upper limit of α when t is large.

The **Gompertz growth model** results from a rate of growth given by

Gompertz Growth Model

$$\frac{\partial \mathcal{E}(Y)}{\partial t} = \beta \mathcal{E}(Y) \left\{ \ln \left[\frac{\alpha}{\mathcal{E}(Y)} \right] \right\} \tag{15.17}$$

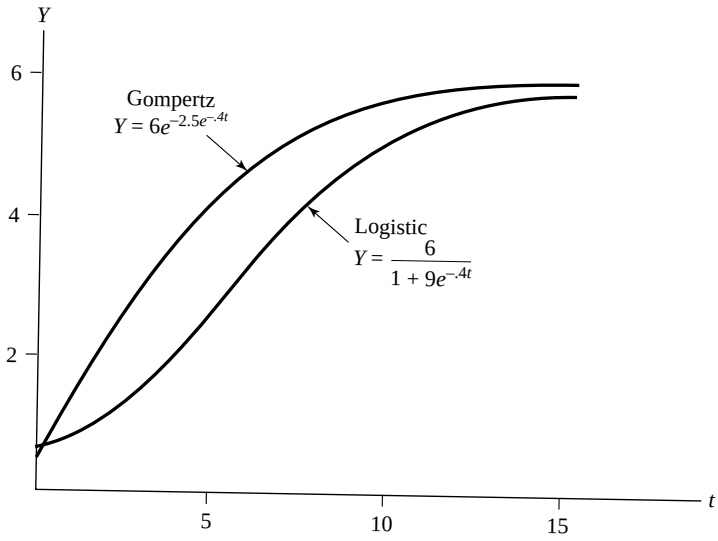

FIGURE 15.4. *The general form of the logistic function and the Gompertz growth model.*

and has the double exponential form

$$Y_i = \alpha e^{-\gamma e^{\beta t_i}} + \epsilon_i. \tag{15.18}$$

Examples of the logistic and Gompertz curves are given in Figure 15.4.

Von Bertalanffy's Model

Von Bertalanffy's model is a more general four-parameter model that yields three of the previous models by appropriate choice of values for the parameter m:

$$Y_i = (\alpha^{1-m} - \theta e^{-\beta t_i})^{1/(1-m)} + \epsilon_i. \tag{15.19}$$

When $m = 0$, this becomes the monomolecular model with $\theta = \alpha e^{-\beta\delta}$, equation 15.12. When $m = 2$, it simplifies to the logistic model with $\theta = -\gamma/\alpha$, equation 15.16, and if m is allowed to go to unity, the limiting form of Von Bertalanffy's model is the Gompertz model, equation 15.18.

Toxicity Studies

Another class of nonlinear models arises when individuals in a population are being scored for their reaction to some substance, and the individuals differ in their sensitivities to the substance. Such models have been developed most extensively in toxicity studies where it is of interest to determine the dose of a substance that causes a certain proportion of injuries or deaths. It is assumed that there is an underlying probability distribution, called the threshold distribution, of sensitivities of individuals to the toxin. The response curve for the proportion of individuals affected at various doses then follows the cumulative probability distribution of the underlying threshold distribution.

Probit and Logit Models

If the threshold distribution is the normal probability distribution, the proportion of individuals affected at dose X follows the cumulative normal distribution. This model leads to the **probit analysis** common in toxicology. Frequently, the response data are better characterized by the normal distribution after dose has been transformed to 1n(dose). Thus, the threshold distribution on the original dose metric is the log-normal distribution. The **logit transformation** results when the underlying threshold distribution is the logistic probability distribution. The probit and logit transformations linearize the corresponding response curves. Alternatively, nonlinear least squares can be used to estimate the parameters of the logistic function. (Weighting should be used to take into account the heterogeneous variances of percentage data.) The cumulative normal distribution has no closed form so that nonlinear least squares cannot be applied directly to estimate the normal parameters.

Weibull Model

The **Weibull probability distribution** is common as the underlying distribution for time-to-failure studies of, for example, electrical systems. The distribution is generated by postulating that a number of individual components must fail, or a number of independent "hits" are needed, before the system fails. Recently, the cumulative form of the Weibull probability distribution has been found to be useful for modeling plant disease progression (Pennypacker, Knoble, Antle, and Madden, 1980) and crop responses to air pollution (Rawlings and Cure, 1985; Heck et al., 1984). The cumulative Weibull probability function is

$$F(X;\mu,\gamma,\delta) = 1 - e^{-[(X_i-\mu)/\delta]^{\gamma}}, \tag{15.20}$$

where μ is the lower limit on X. The two parameters δ and γ control the shape of the curve. This function is an increasing function approaching the upper limit of $F = 1$ when X is large.

As a response model, the asymptote can be made arbitrary by introducing another parameter α as a multiplicative constant, and the function can be turned into a monotonically decreasing function by subtracting from α. Thus, the form of the Weibull function used to model crop response to increasing levels of pollution is

$$Y_i = \alpha e^{-(X_i/\delta)^{\gamma}} + \epsilon_i. \tag{15.21}$$

This form assumes that the minimum level of X is zero. The vector of parameters is $\boldsymbol{\theta}' = (\,\alpha \quad \delta \quad \gamma\,)$. Other experimental design effects such as block effects, cultivar effects, and covariates can be introduced into the Weibull model by expanding the α parameter to include a series of additive terms (Rawlings and Cure, 1985).

Choosing a Nonlinear Model

These examples of nonlinear models illustrate the variety of functional forms available when one is not restricted to linear additive models. There are many other mathematical functions that might serve as useful models. Ideally, the functional form of a model has some theoretical basis as illustrated with the partial derivatives. On the other hand, a nonlinear model

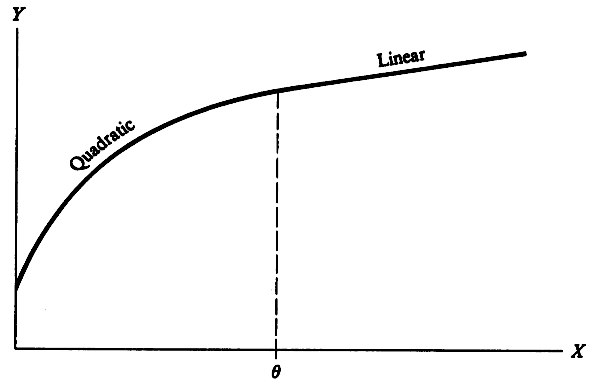

FIGURE 15.5. *An illustration of a quadratic–linear segmented polynomial response curve.*

might be adopted for no other reason than that it is a simple, convenient representation of the responses being observed. The Weibull model was adopted for characterizing crop losses from ozone pollution because it had a biologically realistic form and its flexibility allowed the use of a common model for all studies at different sites and on different crop species.

Segmented Polynomial Models

In some cases, it is simpler to model a complicated response by using different polynomial equations in different regions of the X-space. Usually constraints are imposed on the polynomials to ensure that they meet in the appropriate way at the "join" points. Such models are called **segmented polynomial models**. When the join points are known, the segmented polynomial models are linear in the parameters and can be fitted using ordinary least squares. However, when the join points must be estimated, the models become nonlinear.

Quadratic-Linear Segmented Polynomial

This class of models is illustrated with the quadratic-linear segmented polynomial model. Assume the first part of the response curve is adequately represented by a quadratic or second-degree polynomial, but at some point the response continues in a linear manner. The value of X at which the two polynomials meet, the "join" point, is labeled θ (Figure 15.5). Thus, the quadratic-linear model is

$$Y_i = \begin{cases} \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i & \text{if } X_i \le \theta \\ \gamma_0 + \gamma_1 X_i + \epsilon_i & \text{if } X_i > \theta. \end{cases} \tag{15.22}$$

This equation contains six parameters, β_0, β_1, β_2, γ_0, γ_1, and θ. Estimating all six parameters, however, puts no constraints on how, or even if, the two segments meet at the join point. It is common to impose two constraints. The two polynomials should meet when $X = \theta$ and the transition from one polynomial to the other should be smooth. The first requirement

implies that

$$\beta_0 + \beta_1\theta + \beta_2\theta^2 = \gamma_0 + \gamma_1\theta. \tag{15.23}$$

The second constraint requires the first derivatives of the two functions to be equal at $X = \theta$; that is, the slopes of both segments must be the same at the join point. Thus,

$$\left.\frac{\partial Y(X \le \theta)}{\partial X}\right|_{X=\theta} = \left.\frac{\partial Y(X > \theta)}{\partial X}\right|_{X=\theta}$$

or

$$\beta_1 + 2\beta_2\theta = \gamma_1. \tag{15.24}$$

The second constraint requires that γ_1 be a function of θ, β_1, and β_2. Substituting this result into the first constraint and solving for γ_0 gives

$$\gamma_0 = \beta_0 - \beta_2\theta^2. \tag{15.25}$$

Imposing these two constraints on the original model gives

$$Y_i = \begin{cases} \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i & \text{if } X \le \theta \\ (\beta_0 - \beta_2\theta^2) + (\beta_1 + 2\beta_2\theta)X_i + \epsilon_i & \text{if } X > \theta. \end{cases} \tag{15.26}$$

There are four parameters to be estimated.

This model can be written in one statement if a dummy variable is defined to identify when X is less than θ or greater than θ. Let $T = 0$ if $X \le \theta$ and $T = 1$ if $X > \theta$. Then,

$$\begin{aligned} Y_i &= (1-T)(\beta_0 + \beta_1 X_i + \beta_2 X_i^2) + T[(\beta_0 - \beta_2\theta^2) + (\beta_1 + 2\beta_2\theta)X_i] \\ &= \beta_0 + \beta_1 X_i + \beta_2[X_i^2 - T(X_i - \theta)^2]. \end{aligned} \tag{15.27}$$

This model is nonlinear in the parameters because the products $\beta_2\theta$ and $\beta_2\theta^2$ are present. Also, note that the dummy variable T is a function of θ. If θ is known, the model becomes linear in the parameters. The reader is referred to Anderson and Nelson (1975) and Gallant and Fuller (1973) for more discussion on segmented polynomial models.

15.2 Fitting Models Nonlinear in the Parameters

Least Squares Principle

The least squares principle is used to estimate the parameters in nonlinear models just as in the linear models case. The least squares estimate of $\boldsymbol{\theta}$, labeled $\widehat{\boldsymbol{\theta}}$, is the choice of parameters that minimizes the sum of squared residuals

$$\text{SS}[\text{Res}(\widehat{\boldsymbol{\theta}})] = \sum_{i=1}^{n}[Y_i - f(\boldsymbol{x}_i'; \widehat{\boldsymbol{\theta}})]^2$$

or, in matrix notation,

$$\text{SS}[\text{Res}(\widehat{\boldsymbol{\theta}})] = [\boldsymbol{Y} - \boldsymbol{f}(\widehat{\boldsymbol{\theta}})]'[\boldsymbol{Y} - \boldsymbol{f}(\widehat{\boldsymbol{\theta}})], \tag{15.28}$$

where $\boldsymbol{f}(\widehat{\boldsymbol{\theta}})$ is the $n \times 1$ vector of $f(\boldsymbol{x}_i'; \widehat{\boldsymbol{\theta}})$ evaluated at the n values of $\boldsymbol{x}_i'$. Under the assumption that the random errors in equation 15.1 are independent $N(0, \sigma^2)$ variables, the least squares estimate of $\boldsymbol{\theta}$ is also the maximum likelihood estimate of $\boldsymbol{\theta}$. The partial derivatives of $\text{SS}[\text{Res}(\widehat{\boldsymbol{\theta}})]$, with respect to each $\widehat{\theta}_j$ in turn, are set equal to zero to obtain the p normal equations. The solution to the normal equations gives the least squares estimate of $\boldsymbol{\theta}$.

Form of the Normal Equations

Each normal equation has the general form

$$\frac{\partial\{\text{SS}[\text{Res}(\widehat{\boldsymbol{\theta}})]\}}{\partial\widehat{\theta}_j} = -\sum_{i=1}^{n}[Y_i - f(\boldsymbol{x}_i'; \widehat{\boldsymbol{\theta}})]\left[\frac{\partial f(\boldsymbol{x}_i'; \widehat{\boldsymbol{\theta}})}{\partial\widehat{\theta}_j}\right] = 0, \tag{15.29}$$

where the second set of brackets contains the partial derivative of the functional form of the model. Unlike linear models, the partial derivatives of a nonlinear model are functions of the parameters. The resulting equations are nonlinear equations and, in general, cannot be solved to obtain explicit solutions for $\widehat{\boldsymbol{\theta}}$.

The normal equations for a nonlinear model are illustrated using the exponential growth model $Y_i = \alpha[\exp(\beta t_i)] + \epsilon_i$, equation 15.7. The partial derivatives of the model with respect to the two parameters are

$$\frac{\partial f}{\partial \alpha} = \frac{\partial(\alpha e^{\beta t_i})}{\partial \alpha} = e^{\beta t_i}$$

and

$$\frac{\partial f}{\partial \beta} = \frac{\partial(\alpha e^{\beta t_i})}{\partial \beta} = \alpha t_i e^{\beta t_i}. \tag{15.30}$$

The two normal equations for this model are

$$\sum_{i=1}^{n}(Y_i - \widehat{\alpha}e^{\widehat{\beta}t_i})(e^{\widehat{\beta}t_i}) = 0$$

and

$$\sum_{i=1}^{n}(Y_i - \widehat{\alpha}e^{\widehat{\beta}t_i})(\widehat{\alpha}t_i e^{\widehat{\beta}t_i}) = 0. \tag{15.31}$$

Solving the Normal Equations

A difficulty with nonlinear least squares arises in trying to solve the normal equations for $\widehat{\boldsymbol{\theta}}$. There is no explicit solution even in this simple example. Since explicit solutions cannot be obtained, iterative numerical methods are used. These methods require initial guesses, or starting values,

for the parameters; the starting values are labeled $\boldsymbol{\theta}^0$. The initial guesses are substituted for $\boldsymbol{\theta}$ to compute the residual sum of squares *and* to compute adjustments to $\boldsymbol{\theta}^0$ that will reduce SS(Res) and (it is hoped) move $\boldsymbol{\theta}^0$ closer to the least squares solution. The new estimates of the parameters are then used to repeat the process until a sufficiently small adjustment is being made at each step. When this happens, the process is said to have converged to a solution.

Grid Search Method

Several methods for finding a solution to the normal equations are used in various nonlinear least squares computer programs. The simplest conceptual method of finding the solution is a **grid search** over the region of possible values of the parameters for that combination of values that gives the smallest residual sum of squares. This method can be used to provide reasonable starting values for other methods or, if repeated on successively finer grids, to provide the final solution. Such a procedure is not efficient.

Gauss–Newton Method

Four other methods of solving the normal equations are commonly used. The **Gauss–Newton method** uses a Taylor's expansion of $f(\boldsymbol{x}_i';\boldsymbol{\theta})$ about the starting values $\boldsymbol{\theta}^0$ to obtain a linear approximation of the model in the region near the starting values. That is, $f(\boldsymbol{x}_i';\boldsymbol{\theta})$ is replaced with

$$f(\boldsymbol{x}_i';\boldsymbol{\theta}) \doteq f(\boldsymbol{x}_i';\boldsymbol{\theta}^0) + \sum_{j=1}^{p} \left(\frac{\partial f(\boldsymbol{x}_i';\boldsymbol{\theta}^0)}{\partial \theta_j} \right) (\theta_j - \theta_j^0)$$

or

$$\boldsymbol{f}(\boldsymbol{\theta}) \doteq \boldsymbol{f}(\boldsymbol{\theta}^0) + \boldsymbol{F}(\boldsymbol{\theta}^0)(\boldsymbol{\theta} - \boldsymbol{\theta}^0), \tag{15.32}$$

where $\boldsymbol{F}(\boldsymbol{\theta}^0)$ is the $n \times p$ matrix of partial derivatives, evaluated at $\boldsymbol{\theta}^0$ and the n data points $\boldsymbol{x}_i'$. $\boldsymbol{F}(\boldsymbol{\theta}^0)$ has the form

$$\boldsymbol{F}(\boldsymbol{\theta}^0) = \begin{bmatrix} \frac{\partial [f(\boldsymbol{x}_1';\boldsymbol{\theta}^0)]}{\partial \theta_1} & \frac{\partial [f(\boldsymbol{x}_1';\boldsymbol{\theta}^0)]}{\partial \theta_2} & \cdots & \frac{\partial [f(\boldsymbol{x}_1';\boldsymbol{\theta}^0)]}{\partial \theta_p} \\ \frac{\partial [f(\boldsymbol{x}_2';\boldsymbol{\theta}^0)]}{\partial \theta_1} & \frac{\partial [f(\boldsymbol{x}_2';\boldsymbol{\theta}^0)]}{\partial \theta_2} & \cdots & \frac{\partial [f(\boldsymbol{x}_2';\boldsymbol{\theta}^0)]}{\partial \theta_p} \\ \vdots & \vdots & & \vdots \\ \frac{\partial [f(\boldsymbol{x}_n';\boldsymbol{\theta}^0)]}{\partial \theta_1} & \frac{\partial [f(\boldsymbol{x}_n';\boldsymbol{\theta}^0)]}{\partial \theta_2} & \cdots & \frac{\partial [f(\boldsymbol{x}_n';\boldsymbol{\theta}^0)]}{\partial \theta_p} \end{bmatrix}. \tag{15.33}$$

Linear least squares is used on the linearized model to estimate the shift in the parameters, or the amount to adjust the starting values. That is, the shift in the parameters $(\boldsymbol{\theta} - \boldsymbol{\theta}^0)$ is obtained by regressing $\boldsymbol{Y} - \boldsymbol{f}(\boldsymbol{\theta}^0)$ on $\boldsymbol{F}(\boldsymbol{\theta}^0)$. New values of the parameters are obtained by adding the estimated shift to the initial values. The model is then linearized about the new values of the parameters and linear least squares is again applied to find the second set of adjustments, and so forth, until the desired degree of convergence is attained. The adjustments obtained from the Gauss–Newton method can be too large and bypass the solution, in which case the residual sum of squares may increase at that step rather than decrease. When this happens,

a modified Gauss–Newton method can be used that successively halves the adjustment until the residual sum of squares is smaller than in the previous step (Hartley, 1961).

Method of Steepest Descent

A second method, the **method of steepest descent**, finds the path for amending the initial estimates of the parameters that gives the most rapid decrease in the residual sum of squares (as approximated by the linearization). After each change in the parameter values, the residual sum of squares surface is again approximated in the vicinity of the new solution and a new path is determined. Although the method of steepest descent may move rapidly in the initial stages, it can be slow to converge (Draper and Smith, 1981).

Marquardt's Compromise

The third method, called **Marquardt's compromise** (Marquardt, 1963) is designed to capitalize on the best features of the previous two methods. The adjustment computed by Marquardt's method tends toward the Gauss–Newton adjustment if the residual sum of squares is reduced at each step, and toward the steepest descent adjustment if the residual sum of squares increases in any step. This method appears to work well in most cases.

Derivative-Free Method

These three methods require the partial derivatives of the model with respect to each of the parameters. Alternatively, a derivative-free method (Ralston and Jennrich, 1978) can be used in which numerical estimates of the derivatives are computed from observed shifts in $\widehat{\boldsymbol{Y}}$ as the values of the θ_j are changed. The derivative-free method appears to work well as long as the data are "rich enough" for the model being fit. There have been cases with relatively limited data where the derivative-free method did not appear to work as well as the derivative methods. Convergence was either not obtained, was not as fast, or the "solution" did not appear to be as good.

Summary of Methods

The details of the numerical methods for finding the least squares solution are not discussed in this text. Gallant (1987) presents a thorough discussion of the theory and methods of nonlinear least squares including the methods of estimation. It is sufficient for now to understand that (1) the least squares principle is being used to find the estimates of the parameters, (2) the nonlinear least squares methods are iterative and use various numerical methods to arrive at the solution, and (3) apparent convergence of the estimates to a solution does not necessarily imply that the solution is, in fact, the optimum. The methods differ in their rates of convergence to a solution and, in some cases, whether a solution is obtained. No one method can be proclaimed as universally best and it may be desirable in some difficult cases to try more than one method.

Starting Values

It is important that the starting values in nonlinear regression be reasonably good. Otherwise, convergence may be slow or not attained. In addition, there may be local minima on the residual sum of squares surface, and poor starting values for the parameters increase the chances that the iterative

process will converge to a local minimum rather than the global minimum. To protect against convergence to a local minimum, different sets of starting values can be used to see if convergence is to the same solution in all cases. Plotting the resulting response function with the data superimposed is particularly important in nonlinear regression to ensure that the solution is reasonable.

Nonconvergence

Convergence to a solution may not be obtained in some cases. One reason for nonconvergence is that the functional form of the model is inconsistent with the observed response. For example, an exponential decay model cannot be made to adequately characterize a logistic growth model. "Convergence" in such cases, if attained, would be meaningless. Errors in specification of the derivatives is another common reason for lack of convergence.

Even with an appropriate form for the model and correct derivatives, convergence may not be attained. The reason for lack of convergence can be stated in several ways. (1) The model may be overdefined, meaning the model has more parameters or is more complex than need be for the process. The two-term exponential in the previous section is an overdefined model if the two rate constants are the same. When the two rate constants are too close to the same value, the estimation process will begin to behave as an overdefined model. (2) There may not be sufficient data to fully characterize the response curve. This implies that the model is correct; it is the data that are lacking. Of course, a model may appear to be overdefined because there are not sufficient data to show the complete response curve. (3) The model may be poorly parameterized with two (or more) parameters playing very similar roles in the nonlinear function. Thus, very nearly the same fitted response curve can be obtained by very different combinations of values of the parameters. These situations are reflected in the estimates of the parameters being very highly correlated, perhaps .98 or higher. This is analogous to the collinearity problem in linear models and has similar effects.

15.3 Inference in Nonlinear Models

Distribution of $\widehat{\boldsymbol{\theta}}$ and SS(Res)

Confidence intervals and hypothesis testing for parameters in nonlinear models are based on the approximate distribution of the nonlinear least squares estimator. The familiar properties of linear least squares apply only approximately or asymptotically for nonlinear least squares. The matrix $\boldsymbol{F}(\boldsymbol{\theta}) = \boldsymbol{F}$, equation 15.33, plays the role in nonlinear least squares that $\boldsymbol{X}$ plays in linear least squares. Gallant (1987) shows that if $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{I}\sigma^2)$, $\widehat{\boldsymbol{\theta}}$ is *approximately* normally distributed with mean $\boldsymbol{\theta}$ and $\mathbf{Var}(\widehat{\boldsymbol{\theta}}) = (\boldsymbol{F}'\boldsymbol{F})^{-1}\sigma^2$:

$$\widehat{\boldsymbol{\theta}} \dot{\sim} N[\boldsymbol{\theta}, (\boldsymbol{F}'\boldsymbol{F})^{-1}\sigma^2], \tag{15.34}$$

where the symbol "$\dot{\sim}$" is read "approximately distributed." The residual sum of squares SS[Res($\widehat{\boldsymbol{\theta}}$)], when divided by σ^2, has *approximately* a chi-squared distribution with $(n-p)$ degrees of freedom. Alternatively, asymptotic arguments can be used to show asymptotic normality of $\widehat{\boldsymbol{\theta}}$ as n gets large, without the normality assumption on $\boldsymbol{\epsilon}$ (Gallant, 1987).

In practice, $\boldsymbol{F}(\boldsymbol{\theta})$ is computed as $\boldsymbol{F}(\widehat{\boldsymbol{\theta}})$, which is labeled $\widehat{\boldsymbol{F}}$ for brevity, and σ^2 is estimated with $s^2 = \text{SS[Res}(\widehat{\boldsymbol{\theta}})]/(n-p)$, so that the estimated asymptotic variance–covariance matrix for $\widehat{\boldsymbol{\theta}}$ is

$$\boldsymbol{s}^2(\widehat{\boldsymbol{\theta}}) = (\widehat{\boldsymbol{F}}'\widehat{\boldsymbol{F}})^{-1}s^2. \tag{15.35}$$

Standard errors given in computer programs are based on this approximation. [Some computer programs for nonlinear least squares give only the standard errors of $\widehat{\theta}_j$ and the estimated correlation matrix for $\widehat{\boldsymbol{\theta}}$, labeled $\widehat{\boldsymbol{\rho}}$. The variance–covariance matrix can be recovered as

$$\boldsymbol{s}^2(\widehat{\boldsymbol{\theta}}) = \boldsymbol{S}\widehat{\boldsymbol{\rho}}\boldsymbol{S}, \tag{15.36}$$

where $\boldsymbol{S}$ is the $p \times p$ diagonal matrix of standard errors of $\widehat{\boldsymbol{\theta}}$.]

Confidence Intervals and Tests of Significance

The approximate normality of $\widehat{\boldsymbol{\theta}}$ and chi-squared distribution of $(n-p)s^2/\sigma^2$ [and their independence (Gallant, 1987)] permit the usual computations of confidence limits and tests of significance of the $\widehat{\theta}_j$ and functions of $\widehat{\theta}_j$. Let $C = \boldsymbol{K}'\boldsymbol{\theta}$ to be any *linear* function of interest. The point estimate of C is $\widehat{C} = \boldsymbol{K}'\widehat{\boldsymbol{\theta}}$ with (approximate) standard error $s(\widehat{C}) = \{\boldsymbol{K}'[\boldsymbol{s}^2(\widehat{\boldsymbol{\theta}})]\boldsymbol{K}\}^{1/2}$. The 95% confidence interval estimate of C is

$$\widehat{C} \pm t_{[\alpha/2,(n-p)]}s(\widehat{C}). \tag{15.37}$$

A test statistic for the null hypothesis that $C = C_0$ is

$$t = \frac{\widehat{C} - C_0}{s(\widehat{C})}, \tag{15.38}$$

which is distributed approximately as Student's t with $(n-p)$ degrees of freedom.

Nonlinear Functions of $\widehat{\boldsymbol{\theta}}$

Usually the function of interest in nonlinear regression is a *nonlinear* function of $\boldsymbol{\theta}$, which is estimated with the same nonlinear function of $\widehat{\boldsymbol{\theta}}$. For example, the fitted values of the response variable $\widehat{\boldsymbol{Y}} = \boldsymbol{f}(\widehat{\boldsymbol{\theta}})$ are nonlinear functions of $\widehat{\boldsymbol{\theta}}$. Let $h(\boldsymbol{\theta})$ be any nonlinear function of interest. Gallant (1987) shows that $h(\widehat{\boldsymbol{\theta}})$ is approximately normally distributed with mean $h(\boldsymbol{\theta})$ and variance $\boldsymbol{H}(\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{H}'\sigma^2$; that is,

$$h(\widehat{\boldsymbol{\theta}}) \dot{\sim} N[h(\boldsymbol{\theta}), \boldsymbol{H}(\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{H}'\sigma^2], \tag{15.39}$$

where

$$\boldsymbol{H} = \left(\frac{\partial[h(\boldsymbol{\theta})]}{\partial\theta_1} \quad \frac{\partial[h(\boldsymbol{\theta})]}{\partial\theta_2} \quad \cdots \quad \frac{\partial[h(\boldsymbol{\theta})]}{\partial\theta_p} \right) \tag{15.40}$$

is the row vector of partial derivatives of the function $h(\boldsymbol{\theta})$ with respect to each of the parameters. This result uses the first-order terms of a Taylor's series expansion to approximate $h(\boldsymbol{\theta})$ with a linear function. Thus, $h(\widehat{\boldsymbol{\theta}})$ is (approximately) an unbiased estimate of $h(\boldsymbol{\theta})$. Letting $\widehat{\boldsymbol{H}} = \boldsymbol{H}(\widehat{\boldsymbol{\theta}})$ and $\widehat{\boldsymbol{F}} = \boldsymbol{F}(\widehat{\boldsymbol{\theta}})$, we can estimate the variance of $h(\widehat{\boldsymbol{\theta}})$ by

$$s^2[h(\widehat{\boldsymbol{\theta}})] = \left[\widehat{\boldsymbol{H}}(\widehat{\boldsymbol{F}}'\widehat{\boldsymbol{F}})^{-1}\widehat{\boldsymbol{H}}'\right] s^2. \tag{15.41}$$

The approximate $100(1-\alpha)\%$ confidence interval estimate of $h(\boldsymbol{\theta})$ is

$$h(\widehat{\boldsymbol{\theta}}) \pm t_{[\alpha/2,(n-p)]}\left[\widehat{\boldsymbol{H}}(\widehat{\boldsymbol{F}}'\widehat{\boldsymbol{F}})^{-1}\widehat{\boldsymbol{H}}' s^2\right]^{1/2} \tag{15.42}$$

and an approximate test statistic for the null hypothesis that $h(\boldsymbol{\theta}) = h_0$ is

$$t = \frac{h(\widehat{\boldsymbol{\theta}}) - h_0}{s[h(\widehat{\boldsymbol{\theta}})]}, \tag{15.43}$$

which is distributed approximately as Student's t with $(n-p)$ degrees of freedom.

Several Functions

If there are q functions of interest, $h(\boldsymbol{\theta})$ becomes a vector of order q and $\boldsymbol{H}$ becomes a $q \times p$ matrix of partial derivatives with each row being the derivatives for one of the functions. Assume that the rank of $\boldsymbol{H}$ is q. The composite hypothesis

$$H_0: \ \boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{0}$$

is tested against the two-tailed alternative hypothesis with an approximate test referred to as the **Wald statistic** (Gallant, 1987);

$$W = \frac{\boldsymbol{h}(\widehat{\boldsymbol{\theta}})'\left[\widehat{\boldsymbol{H}}(\widehat{\boldsymbol{F}}'\widehat{\boldsymbol{F}})^{-1}\widehat{\boldsymbol{H}}'\right]^{-1}\boldsymbol{h}(\widehat{\boldsymbol{\theta}})}{qs^2}. \tag{15.44}$$

Notice the similarity in the form of W to the F-statistic in general linear hypotheses. W is approximately distributed as F with q and $(n-p)$ degrees of freedom.

Note that if the functions of interest are the n values of $\widehat{Y}_i$, then $\boldsymbol{h}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{f}(\widehat{\boldsymbol{\theta}})$ and $\boldsymbol{H}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{F}(\widehat{\boldsymbol{\theta}})$, so that

$$s^2(\widehat{\boldsymbol{Y}}) = \left[\widehat{\boldsymbol{F}}(\widehat{\boldsymbol{F}}'\widehat{\boldsymbol{F}})^{-1}\widehat{\boldsymbol{F}}'\right] s^2. \tag{15.45}$$

Wald and Likelihood Ratio

The matrix $\left[\widehat{\boldsymbol{F}}(\widehat{\boldsymbol{F}}'\widehat{\boldsymbol{F}})^{-1}\widehat{\boldsymbol{F}}'\right]$ is analogous to $\boldsymbol{P}$, the projection matrix, in linear least squares.

In general, the confidence limits in equation 15.42, the t-test in equation 15.43, and W in equation 15.44 are referred to as the Wald methodology. The Wald approximation appears to work well in most cases in that

the stated probability levels are sufficiently close to the true levels (Gallant, 1987). However, Gallant has shown cases where the Wald approach can be seriously wrong. Tests and confidence intervals based on the more difficult likelihood ratio test, however, gave results consistent with the stated probabilities in all cases investigated. For this reason, Gallant recommends that the Wald results be compared to the likelihood ratio results for some cases in each problem to verify that the simpler Wald approach is adequate.

The approximate joint $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}$, based on likelihood ratio theory, is defined as that set of $\boldsymbol{\theta}$ for which

$$\text{SS}[\text{Res}(\boldsymbol{\theta})] - \text{SS}[\text{Res}(\widehat{\boldsymbol{\theta}})] \leq p\widehat{\sigma}^2 F_{(\alpha;p,\nu)}, \tag{15.46}$$

where $\widehat{\sigma}$ is an estimate of σ^2 based on ν degrees of freedom. The reader is referred to Gallant (1987) for discussion of the likelihood ratio procedure.

The validity of the Wald approach depends on how well $f(\boldsymbol{x}_i'; \boldsymbol{\theta})$ is represented by the linear approximation in $\boldsymbol{\theta}$. This depends on the parameterization of the model and is referred to as **parameter effects curvature**. Clarke (1987) defined components of overall parameter effects curvature that could be identified with each parameter. These component measures of curvature are then used to define severe curvature, cases in which the Wald methodology may not be adequate for the particular parameters, and to provide higher-order correction terms for the confidence interval estimates. The reader is referred to Clarke (1987) for details.

Example 15.1

The example to illustrate nonlinear regression comes from calcium ion experiments for biochemical analysis of intracellular storage and transport of Ca^{++} across the plasma membrane. The study was run by Howard Grimes, Botany Department, North Carolina State University, and is used with his permission. The data consist of amount of radioactive calcium in cells (nmole/mg) that had been in "hot" calcium suspension for given periods of time (minutes). Data were obtained on 27 independent cell suspensions with times ranging from .45 to 15.00 minutes (Table 15.1). The kinetics involved led the researchers to postulate that the response would follow the nonlinear model

Proposed Model

$$Y_i = \alpha_1[1 - \exp(-\lambda_1 t_i)] + \alpha_2[1 - \exp(-\lambda_2 t_i)] + \epsilon_i, \tag{15.47}$$

where Y is nmoles/mg of Ca^{++}. (This model is referred to as the Michaelis–Menten model.) The partial derivatives for this model are

$$\begin{aligned}
\frac{\partial f}{\partial \alpha_1} &= 1 - e^{-\lambda_1 t}, \\
\frac{\partial f}{\partial \lambda_1} &= t\alpha_1 e^{-\lambda_1 t}, \\
\frac{\partial f}{\partial \alpha_2} &= 1 - e^{-\lambda_2 t}, \text{ and}
\end{aligned}$$

TABLE 15.1. *Calcium uptake of cells suspended in a solution of radioactive calcium. (Data from H. Grimes, North Carolina State University, and used with permission.)*

Suspen. Number	*Time (min)*	*Calcium (nmoles/mt)*	*Suspen. Number*	*Time (min)*	*Calcium (nmoles/mt)*
1	.45	.34170	15	6.10	2.67061
2	.45	−.00438	16	8.05	3.05959
3	.45	.82531	17	8.05	3.94321
4	1.30	1.77967	18	8.05	3.43726
5	1.30	0.95384	19	11.15	4.80735
6	1.30	0.64080	20	11.15	3.35583
7	2.40	1.75136	21	11.15	2.78309
8	2.40	1.27497	22	13.15	5.13825
9	2.40	1.17332	23	13.15	4.70274
10	4.00	3.12273	24	13.15	4.25702
11	4.00	2.60958	25	15.00	3.60407
12	4.00	2.57429	26	15.00	4.15029
13	6.10	3.17881	27	15.00	3.42484
14	6.10	3.00782	–	–	–

$$\frac{\partial f}{\partial \lambda_2} = t\alpha_2 e^{-\lambda_2 t}.$$

In this case, a derivative-free method was used to fit the data in Table 15.1 to the two-term exponential model. The starting values used for the four parameters were

$$\boldsymbol{\theta}^0 = \begin{pmatrix} \alpha_1^0 \\ \lambda_1^0 \\ \alpha_2^0 \\ \lambda_2^0 \end{pmatrix} = \begin{pmatrix} .05 \\ .09 \\ .20 \\ .20 \end{pmatrix}.$$

These were not well-chosen starting values because α_1 and α_2 are the upper asymptotes of the two exponential functions and their sum should be near the upper limits of the data, approximately 4.5. Likewise, the rate constants were chosen quite arbitrarily. This was simply an expedient; if there appeared to be convergence problems or a logical inconsistency in the final model, more effort would be devoted to choice of starting values. We have used the NLIN procedure in SAS (SAS Institute Inc., 1989b) to obtain these results.

Convergence Not Attained

A solution *appeared* to have been obtained. The residual sum of squares decreased from SS[Res($\boldsymbol{\theta}^0$)] = 223 with the starting values $\boldsymbol{\theta}^0$ to SS[Res($\widehat{\boldsymbol{\theta}}$)] = 7.4645 with the final solution $\widehat{\boldsymbol{\theta}}$. A plot of $\widehat{\boldsymbol{Y}}$ superimposed on the data appeared reasonable. However, the results raised several flags. First, a pro-

TABLE 15.2. *Nonlinear regression results from the Grimes data using the two-term exponential model.*

Analysis of Variance: *Source*	*d.f.*	*Sum of Squares*	*Mean Square*
Model	3[a]	240.78865	80.26288
Residual	24	7.46451	0.31102
Uncorr. total	27	248.25315	
mboxCorr.total	26	53.23359	

		Asymptotic	*Asymptotic 95%*	
Parameter	*Estimate*	*Std. Error*	*Lower*	*Upper*
$\widehat{\alpha}_1$	.000100	.0000000	.0000000	.0000000
$\widehat{\lambda}_1$	4,629.250	12,091.767	−20,326.728	29,585.229
$\widehat{\alpha}_2$	4.310418	.9179295	2.4159195	6.2049156
$\widehat{\lambda}_2$	.208303	.0667369	.0705656	.3460400

Asymptotic Correlation Matrix of the Parameters

	$\widehat{\alpha}_1$	$\widehat{\lambda}_1$	$\widehat{\alpha}_2$	$\widehat{\lambda}_2$
$\widehat{\alpha}_1$	.0000	.0000	.0000	.0000
$\widehat{\lambda}_1$	.0000	1.0000	−.5751	−1.0000
$\widehat{\alpha}_2$	.0000	−.5751	1.0000	.5751
$\widehat{\lambda}_2$	.0000	−1.0000	.5751	1.0000

[a]Although the model contained four parameters, the convergence of $\widehat{\alpha}_1$ to the lower bound of .0001 has effectively removed it as a parameter to be estimated.

gram message "**CONVERGENCE ASSUMED**" indicated that the convergence criterion had not been attained. The iterations had terminated because no further progress in reducing the residual sum of squares had been realized during a sequence of halving the size of the parameter changes. Furthermore, the estimates of the parameters and their correlation matrix revealed an overdefined model (Table 15.2). $\widehat{\alpha}_1$ converged to the lower bound imposed to keep the estimate positive, .0001, and its standard error was reported as zero. $\widehat{\lambda}_1$ converged to a very high value with an extremely large standard error and confidence interval. The correlation matrix for the parameter estimates showed other peculiarities. The zeros for the first row and column of the correlation matrix are reflections of the zero *approximated* variance for $\widehat{\alpha}_1$. The correlation matrix showed $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ to be perfectly negatively correlated, and the correlations of $\widehat{\alpha}_2$ with $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ were identical in magnitude.

These results are a reflection of the model being overly complex for the response shown in the data. The first exponential component of the model,

when evaluated using the parameter estimates, goes to $\widehat{\alpha}_1$ for extremely small values of t. For all practical purposes, the first term is contributing only a constant to the overall response curve. This suggests that a single-term exponential model would adequately characterize the behavior of these data.

To verify that these results were not a consequence of the particular starting values, another analysis was run with $\boldsymbol{\theta}^{0'} = (1.0\ 3.9\ .50\ .046)$. Again, the "**CONVERGENCE ASSUMED**" message was obtained and the residual sum of squares was slightly larger, SS(Res) = 7.4652. The solution, however, was very different (results not given). Now, the estimate of λ_2 and its standard error were exceptionally large, but the correlation matrix appeared quite reasonable. Evaluation of the first exponential term produced very nearly the same numerical results as the second did in the first analysis and the second exponential term converged to $\widehat{\alpha}_2$ for very small values of t. The model was simplified to contain only one exponential process,

Simplified Model

$$Y_i = \alpha\{1 - \exp[-(t/\delta)^{\gamma}]\}.$$

This is the Weibull growth model, with an upper asymptote of α, and reduces to the exponential growth model if $\gamma = 1.0$. The presence of γ in the model permits greater flexibility than the simple exponential and can be used to test the hypothesis that the exponential growth model is adequate, $H_0 : \gamma = 1.0$. (Notice that δ in this model is equivalent to $1/\lambda$ in the previous exponential models.)

The convergence criterion was met for this model with SS[Res($\widehat{\boldsymbol{\theta}}$)] = 7.4630, even slightly smaller than that obtained with the two-term exponential model. The key results are shown in Table 15.3. There are no indications of any problems with the model. The standard errors and confidence limits on the parameter estimates are reasonable and the correlation matrix shows no extremely high correlations. The Wald t-test of the null hypothesis $H_0 : \gamma = 1.0$ can be inferred from the confidence limits on γ; $\widehat{\gamma}$ is very close to 1.0 and the 95% confidence interval (.55, 1.48) overlaps 1.0. These results indicate that a simple exponential growth model would suffice.

The logical next step in fitting this model would be to set $\gamma = 1.0$ and fit the simple one-term exponential model

$$Y_i = \alpha[1 - \exp(-t_i/\delta)] + \epsilon_i.$$

Rather than proceed with that analysis, we use the present analysis to show the recovery of $\boldsymbol{s}^2(\widehat{\boldsymbol{\theta}})$ from the correlation matrix, and the computation of approximate variances and standard errors for nonlinear functions of the parameters.

The estimated variance–covariance matrix of the parameter estimates, $\boldsymbol{s}^2(\widehat{\boldsymbol{\theta}})$

TABLE 15.3. *Nonlinear regression results from the Weibull growth model applied to the Grimes data.*

Analysis of Variance:			
Source	*d.f.*	*Sum of Squares*	*Mean Square*
Model	3	240.79017	80.2634
Residual	24	7.46297	.3110
Uncorr. total	27	248.25315	
Corr. total	26	53.23359	

		Asymptotic	*Asymptotic 95%*	
Parameter	*Estimate*	*Std. Error*	*Lower*	*Upper*
$\widehat{\alpha}$	4.283429	.4743339	3.3044593	5.2623977
$\widehat{\delta}$	4.732545	1.2700253	2.1113631	7.3537277
$\widehat{\gamma}$	1.015634	.2272542	.5466084	1.4846603

Asymptotic Correlation Matrix of the Parameters

	$\widehat{\alpha}$	$\widehat{\delta}$	$\widehat{\gamma}$
$\widehat{\alpha}$	1.0000	.9329	−.7774
$\widehat{\delta}$	.9329	1.0000	−.7166
$\widehat{\gamma}$	−.7774	−.7166	1.0000

equation 15.35, is recovered from the correlation matrix $\widehat{\boldsymbol{\rho}}$ by

$$\boldsymbol{s}^2(\widehat{\boldsymbol{\theta}}) = \boldsymbol{S}\widehat{\boldsymbol{\rho}}\boldsymbol{S},$$

where $\boldsymbol{S}$ is the diagonal matrix of standard errors of the estimates from Table 15.3,

$$\boldsymbol{S} = \begin{bmatrix} .47433392 & 0 & 0 \\ 0 & 1.27002532 & 0 \\ 0 & 0 & .22725425 \end{bmatrix}.$$

The resulting asymptotic variance–covariance matrix, $\boldsymbol{s}^2(\widehat{\boldsymbol{\theta}}) = (\widehat{\boldsymbol{F}}'\widehat{\boldsymbol{F}})^{-1}s^2$, equation 15.35, is

$$\boldsymbol{s}^2(\widehat{\boldsymbol{\theta}}) = \begin{bmatrix} .2250 & .5620 & -.0838 \\ .5620 & 1.6130 & -.2068 \\ -.0838 & -.2068 & .0516 \end{bmatrix}.$$

Proportional Response Estimates

To illustrate the computation of approximate variances and confidence limits for nonlinear functions of the parameters, assume that the functions of interest are the estimated responses as a proportion of the upper asymptote, α for $t = 1$, 5, and 15 minutes. That is, the function of interest is

$$h(t, \boldsymbol{\theta}) = 1 - \exp[-(t/\delta)^{\gamma}]$$

evaluated at $t = 1$, 5, and 15. Writing $h(t, \theta)$ for the three values of t as a column vector and substituting $\widehat{\boldsymbol{\theta}} = (\,4.2834 \quad 4.7325 \quad 1.0156\,)'$ from Table 15.3 for $\boldsymbol{\theta}$ gives

$$\boldsymbol{h}(\widehat{\boldsymbol{\theta}}) = \begin{pmatrix} h(1,\ \widehat{\boldsymbol{\theta}}) \\ h(5,\ \widehat{\boldsymbol{\theta}}) \\ h(15,\ \widehat{\boldsymbol{\theta}}) \end{pmatrix} = \begin{pmatrix} .1864 \\ .6527 \\ .9603 \end{pmatrix}$$

as the point estimates of the proportional responses.

The partial derivatives of $\boldsymbol{h}(\boldsymbol{\theta})$ are needed to obtain the variance–covariance matrix for $\boldsymbol{h}(\widehat{\boldsymbol{\theta}})$ equations 15.40 and 15.41. The partial derivatives are

$$\begin{aligned} \frac{\partial h}{\partial \alpha} &= 0 \\ \frac{\partial h}{\partial \delta} &= -\left(\frac{\gamma}{\delta}\right)\left(\frac{t}{\delta}\right)^{\gamma}\left\{\exp\left[-\left(\frac{t}{\delta}\right)^{\gamma}\right]\right\} \\ \frac{\partial h}{\partial \gamma} &= \left(\frac{t}{\delta}\right)^{\gamma}\left[\ln\left(\frac{t}{\delta}\right)\right]\left\{\exp\left[-\left(\frac{t}{\delta}\right)^{\gamma}\right]\right\}. \end{aligned}$$

Writing the partial derivatives as a row vector, equation 15.40, substituting $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$, and evaluating the vector for each value of t gives the matrix $\widehat{\boldsymbol{H}}$:

$$\widehat{\boldsymbol{H}} = \begin{bmatrix} 0 & -.03601 & -.26084 \\ 0 & -.07882 & .02019 \\ 0 & -.02747 & .14768 \end{bmatrix}.$$

The variance–covariance matrix for the predictions, equation 15.41, is

$$\begin{aligned} \boldsymbol{s}^2(\boldsymbol{h}) &= \widehat{\boldsymbol{H}}[\boldsymbol{s}^2(\widehat{\boldsymbol{\theta}})]\widehat{\boldsymbol{H}}' \\ &= \begin{bmatrix} .001720 & .000204 & -.000776 \\ .000204 & .010701 & .006169 \\ -.000776 & .006169 & .004002 \end{bmatrix}. \end{aligned}$$

The square roots of the diagonal elements are the standard errors of the estimated levels of Ca^{++} relative to the upper limit at $t = 1$, 5, and 15 minutes. The 95% confidence interval estimates of these increases are given by the Wald approximation as $h(\widehat{\boldsymbol{\theta}}) \pm s[h(\widehat{\boldsymbol{\theta}})]t_{(.05/2,24)}$ since the residual mean square had 24 degrees of freedom. The Wald confidence limits are summarized as follows:

$\boldsymbol{t}$	$\boldsymbol{h}(\widehat{\boldsymbol{\theta}})$	*Lower Limit*	*Upper Limit*
1	.186	.101	.272
5	.653	.439	.866
15	.960	.829	1.091

Note that the upper limit on the interval for $t = 15$ exceeds 1.0, the logical upper bound on a proportion. This reflects inadequacies in the Wald

approximation as the limits are approached. Simultaneous confidence intervals based on Bonferroni and Scheffé methods can be computed using formulas given in Section 4.6.2. ■

15.4 Violation of Assumptions

In Section 15.3, we have assumed that the random errors ϵ_i in equation 15.1 are independent $N(0, \sigma^2)$ variables. For inferences on the parameters of the nonlinear model, equation 15.1, the assumption of normality is not essential provided the sample size is large and some other mild assumptions are met. However, the violation of assumptions regarding homogeneous and uncorrelated errors has an impact on the inferences in nonlinear models. When errors are heterogeneous and/or correlated, the least squares estimators are inefficient and the estimated variances of $\widehat{\boldsymbol{\theta}}$ given in equation 15.35 are not appropriate.

15.4.1 Heteroscedastic Errors

Consider the model given in equation 15.1 where the ϵ_is are independent $(0, \sigma_i^2)$ variables (but not necessarily normally distributed). Under some weak regularity conditions on σ_i^2 and $f(\boldsymbol{x}_i'; \boldsymbol{\theta})$, Gallant (1987) shows that the least squares estimator $\widehat{\boldsymbol{\theta}}$ is approximately normally distributed with mean $\boldsymbol{\theta}$ and variance $\mathbf{Var}(\widehat{\boldsymbol{\theta}}) = (\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{F}'\boldsymbol{V}\boldsymbol{F}(\boldsymbol{F}'\boldsymbol{F})^{-1}$, where $\boldsymbol{V} = \text{diag}\,(\,\sigma_1^2 \quad \sigma_2^2 \quad \cdots \quad \sigma_n^2\,)$. When $\boldsymbol{V} \neq \sigma^2\boldsymbol{I}$, the variance estimator given in equation 15.35 is not appropriate.

In practice, several types of models are assumed for the behavior of σ_i^2. One such model is given by $\sigma_i^2 = \sigma^2/w_i$ where the w_is are assumed to be known constants. For example, if Y_i is the mean of n_i measurements Y_{ij}, $j = 1, \ldots, n_i$, where we assume that $\text{var}(Y_{ij}) = \sigma^2$, then $Y_i = n_i^{-1}\sum_{j=1}^{n_i} Y_{ij}$ has variance σ^2/w_i, where $w_i = n_i$. As in the case of linear models, the transformed model

$$w_i^{1/2}Y_i = w_i^{1/2}f(\boldsymbol{x}_i'; \boldsymbol{\theta}) + w_i^{1/2}\epsilon_i \tag{15.48}$$

has homogeneous errors. The least squares estimator of $\boldsymbol{\theta}$ in this model, equation 15.48, minimizes

$$S_w(\boldsymbol{\theta}) = \sum_{i=1}^{n} w_i[Y_i - f(\boldsymbol{x}_i'; \boldsymbol{\theta})]^2. \tag{15.49}$$

The estimator $\widehat{\boldsymbol{\theta}}_w$ that minimizes $S_w(\boldsymbol{\theta})$ is known as the **weighted least squares estimator** of $\boldsymbol{\theta}$. If the ϵ_is are independent $N(0, \sigma^2/w_i)$, then $\widehat{\boldsymbol{\theta}}_w$ is also the maximum likelihood estimator of $\boldsymbol{\theta}$. Under mild regularity

Weighted Least Squares

conditions, $\widehat{\boldsymbol{\theta}}_w$ is approximately normally distributed with mean $\boldsymbol{\theta}$ and $\mathbf{Var}(\widehat{\boldsymbol{\theta}}_w) = (\boldsymbol{F}'\boldsymbol{W}\boldsymbol{F})^{-1}\sigma^2$, where $\boldsymbol{W} = \text{diag}\,(\,w_1 \quad w_2 \quad \cdots \quad w_n\,)$ (Gallant, 1987).

If σ^2 is unknown, but an estimate s_i^2 is available for $i = 1, \ldots, n$, then we may use an **estimated generalized least squares estimator** obtained as the value of $\boldsymbol{\theta}$ that minimizes $\sum_{i=1}^{n} s_i^{-2}[Y_i - f(\boldsymbol{x}_i'; \boldsymbol{\theta})]^2$. For example, if Y_i is the mean of n_i measurements Y_{ij}, then an estimate of σ_i^2 is given by $s_i^2 = (n_i - 1)^{-1} \sum_{j=1}^{n_i}(Y_{ij} - Y_i)^2$. If the n_is are small, then s_i^2 may not estimate σ_i^2 very well. In such cases, another estimator of σ_i^2 that is commonly used is given by

$$s_i^2 = n_i^{-1} \sum_{j=1}^{n_i} [Y_{ij} - f(\boldsymbol{x}_i'; \widehat{\boldsymbol{\theta}})]^2, \tag{15.50}$$

where $\widehat{\boldsymbol{\theta}}$ is the least squares estimate of $\boldsymbol{\theta}$.

Variance Related to the Mean

Another class of heteroscedastic variance models has $\sigma_i^2 = h(f(\boldsymbol{x}_i'; \boldsymbol{\theta}))$, where $h(\cdot)$ is a known function. That is, the variance of the response variable is a function of its mean. Consider for example, a binary response variable Y_i that takes the value 1 or 0 depending on whether the ith patient receiving a dose of x_i units is disease-free or not. Let p_i denote $P(Y_i = 1)$ and assume p_i is a function $f(\boldsymbol{x}_i'; \boldsymbol{\theta})$ of x_i that is nonlinear in the parameters. Note that, in this case,

$$Y_i = f(\boldsymbol{x}_i'; \boldsymbol{\theta}) + \epsilon_i,$$

where $\mathcal{E}(Y_i) = f(\boldsymbol{x}_i'; \boldsymbol{\theta})$ and $\text{Var}(\epsilon_i) = \text{Var}(Y_i) = f(\boldsymbol{x}_i'; \boldsymbol{\theta})[1 - f(\boldsymbol{x}_i'; \boldsymbol{\theta})]$. Here, $\text{Var}(\epsilon_i)$ is a known function of the mean function $f(\boldsymbol{x}_i'; \boldsymbol{\theta})$. Similarly, if Y_i is a count variable that has a Poisson distribution with mean $f(\boldsymbol{x}_i'; \boldsymbol{\theta})$, then the variance of Y_i is also $f(\boldsymbol{x}_i'; \boldsymbol{\theta})$.

For models with $\sigma_i^2 = h(f(\boldsymbol{x}_i'; \boldsymbol{\theta}))$, a weighted least squares estimator of $\widehat{\boldsymbol{\theta}}_h$ is obtained as the value of $\boldsymbol{\theta}$ that minimizes

$$\sum_{i=1}^{n} \left[h(f(\boldsymbol{x}_i'; \boldsymbol{\theta}))\right]^{-1} \left[Y_i - f(\boldsymbol{x}_i'; \boldsymbol{\theta})\right]^2. \tag{15.51}$$

It can be shown that $\widehat{\boldsymbol{\theta}}_h$ is not necessarily the maximum likelihood estimator of $\boldsymbol{\theta}$ even if ϵ_i are assumed to be normally distributed. Under some weak regularity conditions, Gallant (1987) shows the $\widehat{\boldsymbol{\theta}}_h$ is approximately normally distributed. van Houwelingen (1988) shows that $\widehat{\boldsymbol{\theta}}_h$ and the maximum likelihood estimator may be inconsistent when the variance function $h(\cdot)$ is misspecified.

Iterative Reweighted Least Squares

Iterative methods are used to compute $\widehat{\boldsymbol{\theta}}_h$. One approach is to obtain $\widehat{\boldsymbol{\theta}}_h^{(j+1)}$ as the value of $\boldsymbol{\theta}$ that minimizes

$$\sum_{i=1}^{n} \left[h(f(\boldsymbol{x}_i'; \widehat{\boldsymbol{\theta}}^{(j)}))\right]^{-1} \left[Y_i - f(\boldsymbol{x}_i'; \boldsymbol{\theta})\right]^2, \tag{15.52}$$

where $\widehat{\boldsymbol{\theta}}_h^{(1)} = \widehat{\boldsymbol{\theta}}$, the least squares estimate of $\boldsymbol{\theta}$. This procedure is repeated until $\widehat{\boldsymbol{\theta}}_h^{(j)}$ converges, and is called **iteratively reweighted least squares**.

Carroll and Ruppert (1988) present the properties of iterative reweighted least squares estimators. They also discuss the Box–Cox transformations and power transformations on both sides of the model.

15.4.2 Correlated Errors

In growth curve models, where data are observed on a single animal or an individual, the errors may exhibit significant serial correlation over time. Also, some economic data may be serially correlated over time. In such cases, $\mathbf{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$ is not $\sigma^2\boldsymbol{I}$. If $\boldsymbol{V}$ were known, the **generalized least squares estimator** $\widehat{\boldsymbol{\theta}}_V$ is the value of θ that minimizes

Generalized Least Squares

$$S_V(\boldsymbol{\theta}) = \left[\boldsymbol{Y} - f(\boldsymbol{x}_i'; \boldsymbol{\theta})\right]' \boldsymbol{V}^{-1} \left[\boldsymbol{Y} - f(\boldsymbol{x}_i'; \boldsymbol{\theta})\right]. \tag{15.53}$$

If $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{V})$, then $\widehat{\boldsymbol{\theta}}_V$ corresponds to the maximum likelihood estimator of $\boldsymbol{\theta}$. Under some regularity conditions, Gallant (1987) shows that $\widehat{\boldsymbol{\theta}}_V$ is approximately normally distributed with mean $\boldsymbol{\theta}$ and $\mathbf{Var}(\widehat{\boldsymbol{\theta}}_V) = (\boldsymbol{F}'\boldsymbol{V}^{-1}\boldsymbol{F})^{-1}$. Iterative methods are used to obtain $\widehat{\boldsymbol{\theta}}_V$.

AR(1) Errors

In some cases, $\boldsymbol{V}$ may be a known function of some unknown parameters $\boldsymbol{\delta}$. For example, consider the first-order autoregressive model given in equation 12.35. For this case, $\boldsymbol{V}$ is given in equation 12.36 and $\boldsymbol{\delta} = \rho$. If ρ is unknown, an estimate $\widehat{\rho}$ of ρ may be obtained as

$$\widehat{\rho} = \frac{\sum_{t=2}^{n} \widehat{\epsilon}_{t-1}\widehat{\epsilon}_t}{\sum_{t=2}^{n} \widehat{\epsilon}_{t-1}^2}, \tag{15.54}$$

where $\widehat{\epsilon}_t = Y_t - f(\boldsymbol{x}_i'; \widehat{\boldsymbol{\theta}})$. An estimated generalized least squares estimate of $\boldsymbol{\theta}$ is obtained by minimizing $S_{\widehat{V}}(\boldsymbol{\theta})$, where $S_V(\boldsymbol{\theta})$ is given in equation 15.53 and $\widehat{\boldsymbol{V}}$ is $\boldsymbol{V}$ given in equation 12.36 with ρ replaced by $\widehat{\rho}$. Care must be used when $\boldsymbol{\delta}$ includes some or all of $\boldsymbol{\theta}$. See Carroll and Ruppert (1988) and Gallant (1987) for details. Warnings described in Section 12.5 for linear models are also appropriate for nonlinear models.

Estimated Generalized Least Squares

15.5 Logistic Regression

We now consider a particular nonlinear regression model where the variance of the response variable is a function of its mean. Consider a binary response variable Y_i that takes the values 0 and 1. For example, $Y_i = 1$ or 0 depending on whether the ith patient has a certain disease. In this case,

$$\mathcal{E}(Y_i) = P[Y_i = 1] = p_i$$

and

$$\text{Var}(Y_i) = p_i(1 - p_i).$$

We wish to relate p_i to certain explanatory variables. For example, if we are interested in studying heart disease, p_i may be related to the ith individual's age, cholesterol level, sex, race, and so on. The relationship between p_i and the explanatory variables may not be linear. Several models are proposed in the literature for p_i as a function of the explanatory variables. One such function is given by the **logistic regression model**,

$$\begin{aligned} p_i &= f(\boldsymbol{x}_i'; \boldsymbol{\theta}) && (15.55) \\ &= \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\theta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\theta})}. && (15.56) \end{aligned}$$

Note that $0 \leq p_i \leq 1$ for all values of $\boldsymbol{\theta}$ and $\boldsymbol{x}_i$. Also, it can be shown that p_i is a monotone function of each explanatory variable, when all other explanatory variables are fixed. This is called the logistic regression model since the **logit**, the log odds ratio,

The Logit

$$\log\left(\frac{p_i}{1 - p_i}\right) = \boldsymbol{x}_i'\boldsymbol{\theta} \tag{15.57}$$

is linear in the parameters $\boldsymbol{\theta}$. [The ratio $p_i/(1-p_i) = P(Y_i = 1)/P(Y_i = 0)$ is known as the **odds ratio**.]

As seen in Section 15.4, the logistic regression may be viewed as a nonlinear model with heteroscedastic errors. In particular,

$$Y_i = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\theta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\theta})} + \epsilon_i, \tag{15.58}$$

where $\mathcal{E}(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i) = p_i(1 - p_i)$. An iteratively reweighted least squares estimator of $\boldsymbol{\theta}$ is obtained by minimizing

$$\sum_{i=1}^{n} \frac{1}{\widehat{p}_i(1 - \widehat{p}_i)} \left[Y_i - \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\theta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\theta})}\right]^2,$$

where $\widehat{p}_i$ is the value of p_i evaluated at the current estimate of $\boldsymbol{\theta}$. We initiate the process with $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, the least squares estimate of $\boldsymbol{\theta}$. These estimates can be obtained using the CATMOD procedure in SAS (SAS Institute Inc., 1989a).

Agresti (1990) presents the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{ML}$ of $\boldsymbol{\theta}$ and shows that $\widehat{\boldsymbol{\theta}}_{ML}$ is the value of $\boldsymbol{\theta}$ that maximizes

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} Y_i \boldsymbol{x}_i'\boldsymbol{\theta} - \sum_{i=1}^{n} \log[1 + \exp(\boldsymbol{x}_i'\boldsymbol{\theta})].$$

He presents iterative procedures to obtain $\widehat{\boldsymbol{\theta}}_{ML}$ and shows that the estimator is approximately normally distributed with mean $\boldsymbol{\theta}$ and variance $\left[\boldsymbol{X}' \text{ diag}\left(\widehat{p}_i(1-\widehat{p}_i)\right)\boldsymbol{X}\right]^{-1}$. The asymptotic distribution of $\widehat{\boldsymbol{\theta}}_{ML}$ and likelihood ratio tests can be used to test relevant hypotheses regarding the parameter $\boldsymbol{\theta}$.

15.6 Exercises

15.1. The data in the accompanying table were taken to develop standardized soil moisture curves for each of six soil types. Percent soil moisture is determined at each of six pressures. The objective is to develop a response curve for prediction of soil moisture from pressure readings. (Data courtesy of Joanne Rebbeck, North Carolina State University.)

Pressure	*Soil Type*					
(Bars)	I	II	III	IV	V	VI
0.10	15.31	17.32	14.13	16.75	14.07	14.15
0.33	11.59	14.88	10.58	14.20	11.39	10.57
0.50	9.74	13.17	8.71	12.07	9.40	9.27
1.00	9.5	12.44	7.62	11.38	8.62	8.73
5.00	6.09	10.08	5.30	9.62	5.17	5.32
15.00	4.49	8.75	4.09	8.59	3.92	4.08

(a) Plot percent moisture against pressure for each soil type. Search for a transformation on X or Y or both that linearizes the relationship for all soils. Fit your transformed data and test homogeneity of the responses over the six soil types.

(b) Use the nonlinear model $Y_{ij} = \alpha_j + \beta_j X_i^{\gamma} + \epsilon_{ij}$ to summarize the relationship between Y and X on the original scale, where Y = moisture, X = pressure, and j indexes soil types. (*Caution*: Your nonlinear program may not be able to iterate γ across $\gamma = 0$ and you may have to try both $\gamma > 0$ and $\gamma < 0$.) The full model allows for a value of α_j and β_j for each soil. Fit the reduced model for $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6$ and test this composite null hypothesis. Plot the residuals for your adopted model and summarize the results. What does the estimate of γ suggest about the adequacy in Part (a) of only a logarithmic transformation on pressure?

15.2. What model is obtained if $\theta_2 = 0$ in the two-term exponential model, equation 15.9?

15.3. Use the data in Exercise 8.8 to fit the nonlinear Mitscherlich model, equation 15.12 with $\delta = 0$, to describe the change in algae density

with time. Allow each treatment to have its own response. Then fit reduced models to test (1) the composite hypothesis that all β_j are equal, and (2) the composite hypothesis that all α_j are equal. Summarize the results and state your conclusions.

15.4 This exercise uses the data from Exercise 8.9. Use the nonlinear model $Y = \alpha X^{\gamma} + \epsilon$ to represent the relationship between Y (on the original scale, Y = dry weight) and volume. Divide volume by 1,000 to make the numbers more manageable. Fit the model, plot the residuals, and summarize the results. Define a reduced model that will test $H_0 : \gamma = 1$. Is this reduced model *nonlinear*? Complete the test and state your conclusion.

15.5. Use the data in Exercise 8.7 to fit the two-term exponential model (equation 15.9) to the data from each environment separately. Use the derivative-free method and $\theta_1 = .2$ and $\theta_2 = .02$ as starting values. Do you get convergence with all six data sets? Plot the response curves and the data. Do the solutions appear reasonable?

15.6. The following data are from a study of the colony-forming activity of six bacterial strains (only strain 3 reported here) under exposure to three *pH* levels (4.5, 6.5, 8.5) and three concentrations of chlorine dioxide CLO_2 in phosphate buffer (20, 50, 80 ppm). (Chlorine dioxide is important in sanitation for controlling bacterial growth.) After suspension of bacteria in the solutions, colony counts were taken on samples from the solutions at recorded time intervals. Use Y = ln(count) in all analyses. (The data from Vipa Hemstapat, North Carolina State University, used with permission.)

(a) Characterize the response of the bacterial strain to CLO_2 for each of the nine pH $\times$ CLO_2 combinations by fitting the Weibull model using Y = ln(count) as the dependent variable and time as the independent variable. You should get convergence in all cases with reasonable starting values; try $\alpha = 20$, $\delta = 20$, and $\gamma = 2$. Summarize your results with a 3×3 table of the estimates of the parameters.

(b) Verify algebraically that the time to 50% decline in the colony is estimated by

$$t_{50} = \widehat{\delta}(.693)^{(1/\widehat{\gamma})}.$$

Use your fitted Weibull response curves to estimate t_{50} in each case. Do an analysis of variance of the 3×3 table of "times to 50% count." (You do not have an estimate of error with which to test the main effects of concentration and pH, but the analysis will show the major patterns.) Summarize the results.

Colony forming activity of six bacterial strains.

	$pH = 4.5$		$pH = 6.5$		$pH = 8.5$	
CLO_2	*Time*	*Colony*	*Time*	*Colony*	*Time*	*Colony*
(ppm)	*(min)*	*Count*	*(min)*	*Count*	*(min)*	*Count*
80	0	2,700,000	0	3,100,000	0	2,400,000
80	5	2,300,000	6	1,700,000	5	2,100,000
80	10	610,000	11	180,000	10	730,000
80	15	140,000	15	13,000	15	130,000
80	20	142	20	1	20	186
50	0	7,500,000	0	2,900,000	0	720,000
50	10	2,800,000	10	2,600,000	10	220,000
50	20	670,000	20	1,300,000	20	8,000
50	30	89,000	30	400,000	30	260
50	40	20	40	94	40	1
50	50	2	50	1		
20	0	16,000,000	0	2,400,000	0	2,100,000
20	10	13,000,000	10	2,800,000	10	2,500,000
20	20	11,000,000	20	2,400,000	20	2,300,000
20	30	6,300,000	30	2,500,000	30	2,000,000
20	40	5,900,000	40	1,800,000	50	440,000
20	50	3,400,000	50	970,000	60	260,000
20	60	1,500,000	60	250,000	70	120,000
20	70	340,000	70	240,000	80	46
20	80	1	80	840	90	24
			90	12		

(c) The nonlinear function of interest in Part (b) is t_{50}. Use the Wald procedure to find the approximate standard error and 95% confidence interval estimate of t_{50} for the middle cell of your 3×3 table. You will have to obtain the partial derivatives of t_{50} with respect to the three parameters and recover the variance–covariance matrix for $\widehat{\boldsymbol{\theta}}$, and then use these results in equation 15.41.

15.7. Fit a polynomial model to the Grimes data, Table 15.1, where $Y =$ Calcium (nmoles/mg) and $X =$ time. (The description of the study is given in Example 15.1.) Is there a reason to force β_0 to be zero in this case? Plot your polynomial response curve and the Weibull response curve given in the text, and superimpose the observed data. Compare the two curves. Does one appear to provide a better fit than the other? If so, in what ways?

15.8. In his famous experiments on gravity and motion in 1608, Galileo rolled a ball down a ramp that was sitting at the edge of a table, recording the release height above the table top H, and the horizontal distance D, from the end of the table at which the ball hit the floor. Our modern knowledge of physics implies the model

$$D^2 - \gamma H\, D - \delta H = 0,$$

where γ and δ are constants that are functions of the table height, ramp angle, and acceleration of gravity. Galileo carefully controlled H while simply observing D so H should be thought of as the independent variable and D as the dependent variable. Solving for D and adding an error term, we find

$$D = \gamma H/2 + \sqrt{\gamma^2 H^2/4 + \delta H} + \epsilon.$$

The data are from Drake (1978) and are in *punti* (points):

D	573	534	495	451	395	337	253
H	1000	800	600	450	300	200	100

(a) Regress D^2 on HD and H (with no intercept). Even though D is the independent variable this should give rough initial estimates of γ and δ.

(b) Now we want to fit the model using the estimates from (a) as initial values. Compute the partial derivatives of $\mathcal{E}(D) = \gamma H/2 + \sqrt{(\gamma^2 H^2/4 + \delta H)}$ with respect to the parameters γ and δ. Fit the correct version of the model in which D is treated as the dependent variable as shown previously.

16

CASE STUDY: RESPONSE CURVE MODELING

Chapters 8 and 15 discussed the use of polynomial and nonlinear response models, respectively. This chapter uses polynomial models and the nonlinear Weibull model to characterize the seed yield response of soybeans to levels of ozone pollution in one experiment. Then data from four experiments on yield response to ozone are combined, the residuals are inspected, and the response variable is transformed as indicated by the analysis. The response models are fit to the transformed data.

The data used in this case study came from research on the effects of air pollutants on crop yields conducted by Dr. A. S. Heagle, Professor of Plant Pathology, North Carolina State University and USDA. The pollutant of primary interest is ozone. Ozone has been shown to cause crop yield losses and the purpose of this research, as part of a nationwide program, was to quantify the effects of air pollutants on the agricultural industry. Of critical importance in the assessment are the possible interactive effects of ozone with other pollutants and environmental factors. The data from the 1981–1984 studies on soybeans, cultivar Davis, are used in this case study. The studies included effects of sulfur dioxide in 1981, different methods of

dispensing ozone in 1982, and different levels of moisture stress in 1983 and 1984.[1]

Description of Experiments

The pollution studies are conducted in the field using open-top chambers to partially contain the pollutants so that higher than ambient levels of the pollutant gases can be maintained. The air flow through the open-top chamber is sufficient to avoid temperature buildup; plant growth within the chambers is normal. There are measurable chamber effects but they are relatively small. The pollutant levels are controlled by dispensing the gas for 7 hours daily, 10:00 A.M. to 5:00 P.M., into the air stream being forced through the chamber. The level of pollutant in the chamber is continuously monitored and dispensing is adjusted to meet the target value. Since the target value of pollutant is never precisely met, treatments with the same target level have slightly different levels of the gas in different replicates.

The basic details of the four experiments are as follows.

1981. The purpose of the 1981 study was to investigate the bivariate response surface of two pollutant gases—ozone and sulfur dioxide. The experimental design was a randomized complete block design with two blocks and 24 treatments per block. The 24 treatments were all combinations of six levels of ozone and four levels of sulfur dioxide. The six levels of ozone were charcoal-filtered air (CF) which gives about .025 ppm ozone; nonfiltered air (NF), which gives the ambient level of ozone; and constant additions to ambient levels of ozone of .020, .030, .050, and .070 ppm. The constant addition treatments are labeled CA20, CA30, CA50, and CA70, respectively. The four levels of SO_2 were ambient air (NF) and constant additions of .030, .090, and .350 ppm, which are labeled S1, S2, and S3, respectively.

1982. The 1982 study had the purposes of developing more information on the ozone dose–response curves and of investigating possible effects of different methods of dispensing pollutant into the chambers. Prior to 1982, the target with ozone dispensing was to add a constant amount to the ambient levels at any given time. It was believed by some that a proportional increase in the gas at any given time would give more realistic distributions of the pollutant and that differences in distributions of the pollutant might affect plant response. Therefore, the treatments in this study included, in addition to CF and NF, both constant additions of .020, .040, and .060 ppm and proportional increases of 30, 60, and 90% of ambient. The proportional treatments are labeled P13, P16, and P19, respectively. There were a total of eight treatments in a randomized complete block design with two blocks.

[1]Some of the analyses in this case study were done by V. M. Lesser, N. C. State University.

1983. The purpose of the 1983 study was to investigate the effects of moisture stress to the plants on their response to ozone. In addition, physiological data were taken on half the plants in each plot so that yield was reported for only one-half plot per chamber. There were two levels of moisture stress and four levels of ozone, CF, NF, CA30, and CA60, giving eight treatments. The experimental design was a randomized complete block design with three blocks.

1984. This was a continuation of the 1983 moisture stress study with, again, only half the plot being used for yield measurement. There were two levels of moisture and six levels of ozone, CF, NF, CA15, CA30, CA45, and CA60, giving 12 treatments in a randomized complete block design with two replications.

Two distinct analyses are presented in this case study. First, the 1981 data alone are analyzed. The bivariate response surface is fit using a polynomial response model and a nonlinear response model. Then, all four years of data are combined in an analysis of the residuals. The residuals analysis suggests a transformation of the data, and a nonlinear response model involving ozone, sulfur dioxide, and moisture level is fit to the transformed data.

16.1 The Ozone–Sulfur Dioxide Response Surface (1981)

The objective is to develop a bivariate response surface model to characterize the 1981 yield response of soybeans, cultivar Davis, to pollutant mixtures of ozone and sulfur dioxide. The yield data and the observed seasonal averages of ozone and sulfur dioxide for each experimental unit are given in Table 16.1. The north and south halves of the experimental plots are recorded separately as Y_1 and Y_2, respectively. This was done to investigate the possibility of an effect of position within the chamber on the response to the pollutant. Preliminary analyses indicated that although there was a north–south position effect within the chambers, there was no position by treatment interaction effect. Therefore, all analyses reported in this section use the average of Y_1 and Y_2 for each experimental unit.

Analysis of Variance

The analysis of variance for the 1981 soybean data is given in Table 16.2. The model for this analysis is

$$Y_{ijk} = \mu + \rho_i + \tau_j + \gamma_k + (\tau\gamma)_{jk} + \epsilon_{ijk}, \tag{16.1}$$

where ρ_i, τ_j, and γ_k are the block, ozone treatment, and sulfur dioxide treatment effects, respectively. All effects are assumed to be fixed; ϵ_{ijk} are assumed to be normally and independently distributed with zero mean and common variance σ^2.

TABLE 16.1. *Yields of soybean (grams per meter row) following exposure to ozone (O_3) and sulfur dioxide (SO_2) for seven hours daily during the growing season. Ozone and sulfur dioxide levels (ppm) are seasonal averages during the exposure period. (Data courtesy A. S. Heagle, Plant Pathologist, N. C. State University and USDA; data used with permission.)*

Treatment		*Block* 1				*Block* 2			
O_3	SO_2	O_3	SO_2	Y_1 [a]	Y_2	O_3	SO_2	Y_1	Y_2
CF	NF	.025	.000	516.5	519.5	.025	.000	603.0	635.0
CF	$S1$	.023	.022	552.0	596.0	.022	.015	796.0	454.5
CF	$S2$	.028	.075	569.0	500.5	.018	.100	597.5	697.0
CF	$S3$	.029	.389	419.0	358.5	.025	.380	458.0	365.5
NF	NF	.059	.000	503.5	449.5	.051	.000	652.0	496.0
NF	$S1$	.058	.016	411.0	484.0	.052	.028	590.5	292.5
NF	$S2$	.058	.070	502.5	477.0	.055	.092	440.0	427.5
NF	$S3$	.058	.350	353.0	338.5	.051	.341	487.0	284.0
$CA20$	NF	.068	.000	449.5	480.5	.067	.000	533.5	321.5
$CA20$	$S1$	.073	.016	472.5	478.0	.066	.023	486.0	317.0
$CA20$	$S2$	.072	.085	382.5	411.5	.069	.104	420.5	456.0
$CA20$	$S3$	.068	.395	291.0	266.5	.068	.377	271.0	280.5
$CA30$	NF	.084	.000	399.0	414.5	.089	.000	390.5	324.5
$CA30$	$S1$	.086	.034	321.5	336.5	.087	.040	373.0	320.5
$CA30$	$S2$	.082	.067	373.0	384.5	.085	.091	321.0	246.0
$CA30$	$S3$	.090	.350	269.0	303.0	.083	.379	246.5	274.0
$CA50$	NF	.105	.000	438.0	345.0	.110	.000	307.0	281.5
$CA50$	$S1$	.111	.018	346.5	347.5	.107	.047	387.5	329.5
$CA50$	$S2$	.108	.084	297.0	316.5	.100	.098	270.0	246.0
$CA50$	$S3$	.106	.369	242.5	244.0	.100	.362	197.5	196.0
$CA70$	NF	.123	.000	342.5	331.5	.121	.000	275.0	278.5
$CA70$	$S1$	.131	.021	269.0	298.5	.125	.028	266.0	243.5
$CA70$	$S2$	.126	.056	297.5	308.5	.127	.099	303.0	215.5
$CA70$	$S3$	.123	.345	211.0	227.0	.122	.355	283.5	208.0

[a] Y_1 and Y_2 are the yields from the north and south halves of the plot, respectively.

TABLE 16.2. *Analysis of variance of 1981 soybean yield following exposure to ozone and sulfur dioxide pollutants.*

Source	*d.f.*	*Sum of Squares*	*Mean Square*	*F*	*Prob > F*
Total	47	606,481			
Block	1	467	467		
Ozone	5	408,117	81,623	46.01	.0001
Sulfur Dioxide	3	126,697	42,232	23.81	.0001
Ozone × Sulfur	15	30,400	2,027	1.14	.3766
Error	23	40,799	1,774		

TABLE 16.3. *Soybean mean yields (grams per meter row) for the 1981 ozone by sulfur dioxide study.*

SO_2	*Ozone treatment*						
Trt.	*CF*	*NF*	*CA*20	*CA*30	*CA*50	*CA*70	*Mean*
NF	568.5[a]	525.3	446.3	382.1	342.9	306.9	428.6
*S*1	599.6	444.5	438.4	337.9	352.8	269.3	407.1
*S*2	591.0	461.8	417.6	331.1	282.4	281.1	394.2
*S*3	400.3	365.6	277.3	273.1	220.0	232.4	294.8
Mean	539.8	449.3	394.9	331.1	299.5	272.4	381.2

[a] $s(\overline{Y}_{\cdot jk}) = 29.8$ is the standard error for the cell means. $s(\overline{Y}_{\cdot j\cdot}) = 14.9$ and $s(\overline{Y}_{\cdot\cdot k}) = 12.2$ are the standard errors for the ozone and sulfur dioxide marginal means, respectively.

The analysis of variance shows that there are highly significant ozone and sulfur dioxide effects on soybean seed yield but gives no indication that the two pollutants interact (Table 16.2). The treatment means, Table 16.3, show a 30% change in yield over the sulfur dioxide treatments and a 45% change over the ozone treatments. The standard errors of the treatment means are given in the footnote of Table 16.3. It is this joint response to the two pollutants that is to be characterized with an appropriate response model. For this purpose, the quantitative levels of the pollutants for each plot, rather than the treatment codes, are used.

Pollutant Levels

Using the quantitative levels of the pollutant in each plot introduces a problem that is somewhat unique to these studies. The specified treatments are target levels of the pollutant to be added to ambient air levels. Due to some imprecision in both the monitoring and the dispensing systems, the target levels are not precisely attained. These small discrepancies cause a slight imbalance in the study when the treatments are viewed in terms of the quantitative levels attained. (The effects of imbalance are discussed in

Chapter 17. In general, imbalance in an experiment causes the analysis of variance to be inappropriate in that the sum of squares due to one factor will contain effects of other factors.)

In this particular case, the discrepancies in the pollutant levels are relatively minor, Table 16.1, and the analysis of variance can be viewed as a close approximation to the effects of the pollutants. Nevertheless, the ozone treatment sum of squares may contain some bias due to differences in sulfur dioxide levels and vice versa, the ozone by sulfur dioxide interaction sum of squares may contain main effects of the two pollutants, and experimental error may be biased upward by the effects of the pollutants. Thus, the analysis of variance is used only as a guide to what to expect in the response surface modeling. The lack of fit of the polynomial model cannot be judged solely on how much of the treatment sums of squares is not explained, and experimental error from the analysis of variance is not used as the unbiased estimate of σ^2.

16.1.1 Polynomial Response Model

Second-Degree Polynomial

The analysis of variance showed significant main effects for both ozone and sulfur dioxide but no indication of an interaction between the two gases (Table 16.2). Therefore, the first polynomial model tried was a second-degree polynomial in both pollutants but with no product, or interaction, term:

$$\begin{aligned} Y_{ijk} &= \beta_0 + \rho D_i + \beta_1 X_{ijk1} + \beta_{11} X_{ijk1}^2 + \beta_2 X_{ijk2} \\ &\quad + \beta_{22} X_{ijk2}^2 + \epsilon_{ijk}, \end{aligned} \tag{16.2}$$

where D_i is a dummy variable coded $+1$ and -1 to identify the two blocks, ρ is the regression coefficient to account for the block effect, and X_{ijk1} and X_{ijk2} are the observed seasonal averages of ozone and sulfur dioxide, respectively, for the ijkth experimental unit. $\boldsymbol{X}$ for this model is of order 48×6 and consists of the column of ones for the intercept, a column for the dummy variable D_i, and the four columns of X_1, X_1^2, X_2, and X_2^2. The analysis for this model is summarized in Table 16.4. [The analysis was obtained using PROC GLM (SAS Institute, Inc., 1989b).]

Nonorthogonality

The nonorthogonality of the data, due to the variable treatment levels, is evident. The "SO_2 linear" sum of squares (135,161 in Table 16.4) exceeds the total treatment sum of squares for sulfur dioxide (126,697 in Table 16.2), and the residual mean square from the regression analysis is appreciably smaller than experimental error in the analysis of variance, 1,494 versus 1,774. Neither result is possible in the balanced case. Also, differences between sequential (Type I) and partial (Type III) sums of squares for "Block" and "O_3 quadratic" show that the replication effects are not orthogonal to the realized levels of ozone and sulfur dioxide, and that ozone levels are not orthogonal to sulfur dioxide levels.

TABLE 16.4. *Analysis of variance for the second degree polynomial model in both gases with no interaction.*

Source	*d.f.*	*Sum of Squares*	*Mean Square*	*F*	*Prob > F*
Total	47	606,481			
Regression	5	543,713	108,743	72.76	.0001
Residual	42	62,768	1,494		

SS(Regr) partition:

Source	*d.f.*	*Sequential SS*	*F*	*Prob > F*	*Partial SS*	*F*	*Prob > F*
Block	1	467	.31	.5791	1,792	1.20	.2798
O_3 linear	1	397,665	266.09	.0001	54,922	36.75	.0001
SO_2 linear	1	135,161	90.44	.0001	2,613	1.75	.1933
O_3 quadratic	1	10,281	6.88	.0121	10,295	6.89	.0120
SO_2 quadratic	1	138	.09	.7630	138	.09	.7630

Summary

The key results from the analysis of the first polynomial model (Table 16.4) can be summarized as follows.

1. The quadratic term for sulfur dioxide makes no significant contribution and can be dropped from the model.

2. The quadratic term for ozone is significant in both the sequential and partial sums of squares and, consequently, will remain significant even after the "SO_2 quadratic" term is dropped.

3. The sequential sum of squares for "SO_2 linear" is highly significant, and even exceeds the total sulfur dioxide treatment sum of squares. Although it is very likely "SO_2 linear" will remain significant after "SO_2 quadratic" has been dropped, one cannot be certain that it will from this analysis (since the sequential sum of squares for "SO_2 linear" has not been adjusted for "O_3 quadratic"). The nonsignificant partial sum of squares for "SO_2 linear" should be ignored; remember that it has been adjusted for the higher-degree "SO_2 quadratic" term.

4. Block effects are nonsignificant but, since they were part of the basic experimental design, they will be retained in the model. Dropping the block effects, in this case, causes only trivial changes in the final model.

Modifying the Model

Comparison of the sums of squares for the polynomial model with the corresponding treatment sums of squares for ozone and sulfur (remember that in these data they are not precisely comparable) suggests that there

TABLE 16.5. *Analysis of variance for the polynomial model allowing a quadratic response for ozone, linear response for sulfur dioxide, and a linear-by-linear interaction.*

Source	*d.f.*	*Sum of Squares*	*Mean Square*	*F*	*Prob > F*
Total	47	606,481			
Regression	5	550,004	110,001	81.80	.0001
Residual	42	56,477	1,345		

SS(Regr) partitions:

Source	*d.f.*	*Sequential SS*	*F*	*Prob > F*	*Partial SS*	*F*	*Prob > F*
Block	1	467	.35	.5587	1,709	1.27	.2659
O_3 linear	1	397,665	295.73	.0001	60,087	44.69	.0001
SO_2 linear	1	135,161	100.52	.0001	46,385	34.50	.0001
O_3 quadratic	1	10,281	7.65	.0084	10,756	8.00	.0071
Linear × Linear	1	6,429	4.79	.0344	6,429	4.78	.0344

is nothing to be gained by expanding the polynomial model to include cubic terms in either variable. On the other hand, there may be some improvement in the model from a second-degree product term, the "O_3 linear × SO_2 linear" interaction term. Even though the interaction sum of squares in the analysis of variance was not significant, it is possible for a single degree-of-freedom contrast to be significant. Hence, the second polynomial model to be fitted dropped the quadratic term for sulfur dioxide and added the linear-by-linear product term:

$$\begin{aligned} Y_{ijk} &= \beta_0 + \rho D_i + \beta_1 X_{ijk1} + \beta_2 X_{ijk2} + \beta_{11} X_{ijk1}^2 \\ &\quad + \beta_{12} X_{ijk1} X_{ijk2} + \epsilon_{ijk}. \end{aligned} \tag{16.3}$$

The analysis of this model is summarized in Table 16.5.

All terms in this model are significant and will be retained. There remains the possibility that a higher-order product term would contribute significantly to the model. The most logical possibility is the "O_3 quadratic × SO_2 linear" interaction term, $X_1^2 X_2$, since there is significant quadratic response to ozone and the analysis of variance interaction sum of squares is the largest partition not explained by the present model. It is left as an exercise for the student to show whether this term is needed. Although a plot of the data superimposed on the response surface showed considerable dispersion about the surface, there was no apparent pattern suggesting inadequacies in this model. Likewise, the plot of the residuals versus $\widehat{\boldsymbol{Y}}$ and the normal plot appeared reasonable.

Final Response Surface

This polynomial model, equation 16.3, is adopted as a reasonable characterization of the ozone–sulfur dioxide response surface in these data. The

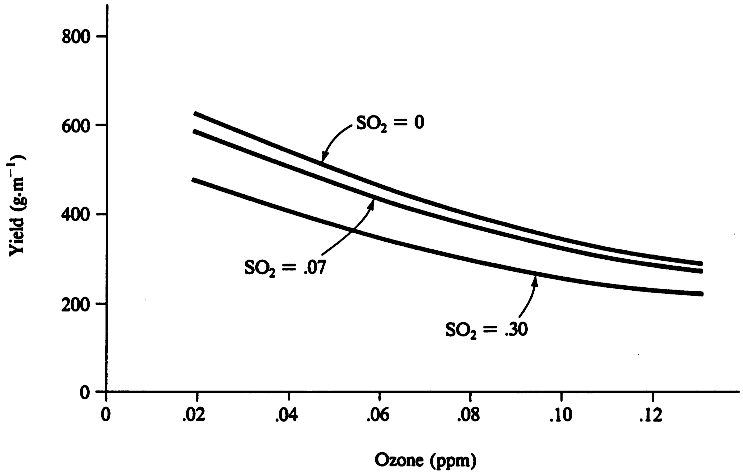

FIGURE 16.1. *The bivariate polynomial response surface for yield of soybeans exposed to chronic doses of ozone and sulfur dioxide. The surface is represented by three traces from the surface for different levels of* SO_2.

final response surface equation, averaged over the block effects, is

$$\begin{aligned} \widehat{Y} &= 724 - 5,152X_1 + 13,944X_1^2 - 543X_2 + 2,463X_1X_2 \quad (16.4) \\ &\quad (28) \quad (771) \qquad (4930) \qquad (92) \qquad (1126). \end{aligned}$$

The standard errors of the regression coefficients are shown in parentheses. The response surface is shown in Figure 16.1 as a series of three response curves for ozone at three levels of SO_2.

Understanding the Response

The response surface has a negative slope with respect to both ozone and sulfur dioxide at near-zero pollution. Thus, there is evidence that increasing levels of either pollutant causes yield of Davis soybean to decline in this environment. The positive sign of the quadratic regression coefficient $\widehat{\beta}_{11}$ indicates that the rate of decline in yield is decreasing with increasing ozone and the polynomial response curve will eventually reach a minimum with yield appearing to increase for levels of ozone beyond that point. The minimum point on the ozone response curve for a given level of sulfur dioxide is obtained by setting the partial derivative of Y with respect to X_1 equal to zero and solving for X_1. The partial derivative is

$$\frac{\partial Y}{\partial X_1} = -5,152 + 2(13,944)X_1 + 2,463X_2. \quad (16.5)$$

Setting this equation equal to zero and solving for $X_{1\,min}$ gives

$$X_{1\,min} = \frac{(5,152 - 2,463X_2)}{2(13,944)}.$$

$X_{1\,min}$ ranges from .1847 for $X_2 = 0$ to .1582 for $X_2 = .3$. These levels of ozone are beyond the limits of the experiment since the average ozone level for CA70 was .125 and, consequently, any inference that sufficiently high levels of ozone would cause yield to increase would be an inappropriate extrapolation.

The interaction term has the effect of decreasing the rate of decline in yield as the level of the other pollutant increases. The impact of SO_2 at the highest level of O_3 is approximately half, in absolute terms, what it is at the low level of O_3. This diminished effect of one pollutant at higher levels of the other is reasonable since there is less yield to be lost at the higher levels.

Extrapolations

Within the limits of the levels of pollutant in this experiment, the polynomial model provides a reasonable characterization of the response surface. Any extrapolation beyond the limits of the experiment encounters biologically inconsistent predictions: minimum yield in the vicinity of .16 ppm ozone with predictions of increasing yields at higher levels, and predictions of negative yields when SO_2 is sufficiently high, approximately 1.3 ppm.

16.1.2 Nonlinear Weibull Response Model

A nonlinear response model based on the functional form of the Weibull probability distribution has been used as a dose–response model in the ozone pollution research simply because it has a biologically realistic form with sufficient flexibility to cover the range of responses encountered for the various crop species and environmental conditions. A single flexible form facilitates comparing responses and summarizing the results with a minimum number of response equations.

Form of the Model

The Weibull model in its simplest form was given in equation 15.21. For this experiment, the α term in that model must be extended to account for additional effects—block effects and the effect of sulfur dioxide. Thus, the Weibull model takes the form

$$Y_{ijk} = (\alpha_1 + \alpha_2 D_i + \beta X_{ijk2}) e^{-(X_{ijk1}/\delta)^\gamma} + \epsilon_{ijk}, \qquad (16.6)$$

where the exponential term controls the relative response to ozone, decreasing from 1 at $X_1 = 0$ to a limit of zero when X_1 is large. If $\gamma = 1$, this becomes the exponential decay curve. The three terms in parentheses in front of the exponential term control the yield level under the hypothetical situation of $X_1 = 0$, which is expressed here as an overall constant α_1, a block effect α_2, and a linear adjustment for the level of sulfur dioxide βX_2. The dummy variable D is defined as 1 if the observation is from block 1 and -1 if the observation is from block 2. Thus, setting $D = 0$ gives an average result for the two blocks so that α_1 is the expected yield for this environment with $X_1 = X_2 = 0$. (On the basis of the polynomial results, SO_2 is handled with a linear response in this model.)

Fitting the Model

The derivative-free method of PROC NLIN in SAS (SAS Institute Inc., 1989b) was used to fit this model. The program statements that generated the analysis are as follows:

```
PROC NLIN METHOD=DUD;

PARMS A1=700 A2=0 B= −0.5 DELTA=0.14 GAMMA=1;

MODEL PODWT=(A1 + A2*D + B*X2)
  *EXP(−(X1/DELTA)**GAMMA);

OUTPUT OUT=OUT.R5 P=PWHAT R=PWRESID;
```

(A1, A2, B, DELTA, and GAMMA are used in place of α_1, α_2, β, δ, and γ, respectively, because the programming language will not accomodate Greek letters.) The starting values for the parameters are given in the PARMS statement. These values were chosen on the basis of a preliminary plot of the data. The highest yields for the low ozone treatment were in the vicinity of $\alpha_1 = 700$; thus, $\alpha_1^0 = 700$. The "block" effects were small, suggesting $\alpha_2^0 = 0$. The starting value for β, $\beta^0 = -.5$, resulted from a visual assessment of the change in yield per unit change in SO_2 but contained an error in placement of the decimal. The value should have been $\beta^0 = -500$. The parameter δ is interpreted as the dose at which yield has been reduced to the fraction e^{-1} of what it is at zero ozone. The starting value was read from a plot of the data as $\delta^0 = .14$. Finally, $\gamma^0 = 1$ was chosen because the plot appeared to be similar in shape to an exponential decay curve.

Solution

In spite of a very poor starting value for β, convergence was quickly attained. The summary of this analysis is given in Table 16.6. The residual sum of squares is SS(Res) = 59,049 with 43 degrees of freedom, compared to SS(Res) = 56,478 with 42 degrees of freedom for the final polynomial model. The corresponding mean squares are 1,373 and 1,345. Thus, the nonlinear model with five parameters fits the data nearly as well as the polynomial model with six parameters. (*Note:* The difference in the residual sums of squares for the two models cannot be tested as previously done since neither model is "nested" in the other.) The resulting nonlinear response equation is

$$\widehat{Y} = (759.4 + 3.7D - 631X_2)e^{-(X_1/.134)^{.88}}. \tag{16.7}$$

Checking the Equation

The plot of this response equation (not given here) is almost indistinguishable, within the limits of the design space, from the plot for the polynomial response model given in Figure 16.1. Estimated responses for the two equations are compared in Table 16.7. The nonlinear equation has slightly less curvature except at the low levels of ozone when sulfur dioxide is near zero. The plot of the residuals against $\widehat{Y}$, Figure 16.2, and the normal plot of the residuals, Figure 16.3, give no reason for concern about the adequacy of the model. (These plots are very similar to the corresponding

TABLE 16.6. *Nonlinear regression results from fitting the Weibull model to the 1981 yield data of soybeans following exposure to ozone and sulfur dioxide.*

Source	*d.f.*	*Sum of Squares*	*Mean Square*
Model	5	7,521,067	1,504,213
Residual	43	59,049	13,73
Uncorrected total	48	7,580,116	
(Corrected total)	47	606,481	

			Asymptotic 95% Confidence Interval	
Parameter	*Estimate*	*Asymptotic Std. Error*	*Lower*	*Upper*
α_1	759.4479	88.2776	581.4198	937.4761
α_2	3.6723	9.4117	−15.3082	22.6529
β	−631.2867	93.9163	−820.6862	−441.8871
δ	0.1336	.0145	.1044	.1629
γ	0.8788	.2248	.4255	1.3320

TABLE 16.7. *Estimated responses for the nonlinear model and the polynomial model for the 1981 soybean yield response to ozone and sulfur dioxide.*

Ozone	$SO_2 = 0$ *ppm*		$SO_2 = .30$ *ppm*	
(ppm)	*Nonlinear*	*Polynomial*	*Nonlinear*	*Polynomial*
.02	629.0	626.2	472.2	478.1
.04	537.1	539.8	403.1	406.5
.06	463.1	464.7	347.6	346.2
.08	401.6	400.7	301.5	296.9
.10	349.9	347.8	262.6	258.8
.12	305.8	306.1	229.5	231.9

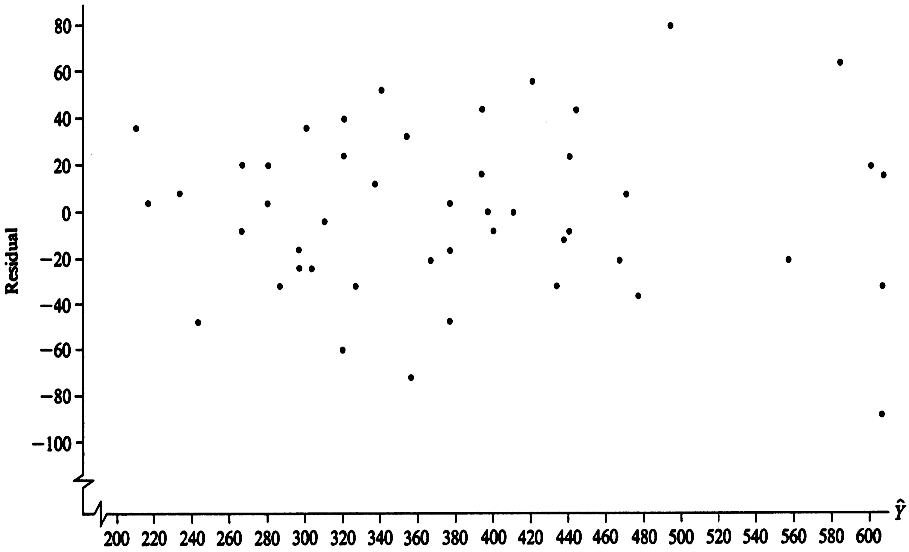

FIGURE 16.2. *The residuals from the nonlinear model for the 1981 soybean response to ozone and sulfur dioxide plotted against the estimated yield.*

plots for the polynomial model. For that reason, the plots are given only for the nonlinear model.)

Setting $\gamma = 1$

The standard error on $\widehat{\gamma}$ and the confidence interval estimate of γ (Table 16.6) suggest that the exponential decay model for ozone effects $\gamma = 1$ would be adequate. The next step in the model-building process would be to fit the model with $\gamma = 1$. The nonlinear model would be reduced to four parameters that, it appears, would provide nearly the same fit as the polynomial model with six parameters. This step of the model building is left as an exercise for the student, and the current five-parameter nonlinear response equation is used for interpretation.

$\widehat{\rho}$ and $s^2(\widehat{\theta})$

The (asymptotic) correlation matrix for the estimates of the parameters $\widehat{\boldsymbol{\theta}}' = \begin{pmatrix} \widehat{\alpha}_1 & \widehat{\alpha}_2 & \widehat{\beta} & \widehat{\delta} & \widehat{\gamma} \end{pmatrix}$ is

$$\widehat{\boldsymbol{\rho}} = \begin{bmatrix} 1 & .121322 & -.790652 & -.914682 & -.960015 \\ .121322 & 1 & -.072519 & -.128904 & -.108028 \\ -.790652 & -.072519 & 1 & .683393 & .716938 \\ -.914682 & -.128904 & .683393 & 1 & .799745 \\ -.960015 & -.108028 & .716938 & .799745 & 1 \end{bmatrix}$$

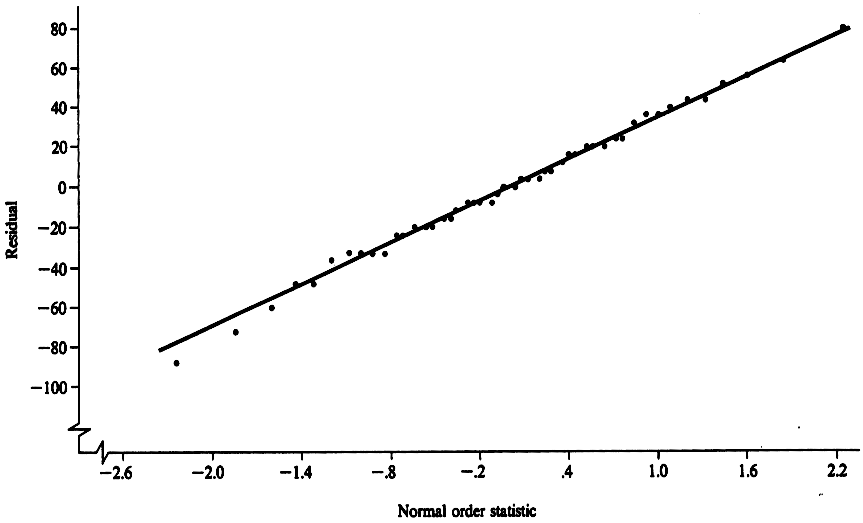

FIGURE 16.3. *The normal plot of the residuals from the nonlinear model for the 1981 soybean response to ozone and sulfur dioxide.*

The variance–covariance matrix for the estimates of the parameters, reconstructed from the correlation matrix, is

$$s^2(\widehat{\boldsymbol{\theta}}) = \begin{bmatrix} 7,792.94 & 100.800 & -6,555.065 & -1.170042 & -19.04732 \\ 100.800 & 88.5810 & -64.1007 & -.017580 & -.228514 \\ -6,555.06 & -64.1007 & 8,820.27 & .930020 & 15.13310 \\ -1.17004 & -.01758 & .93002 & .000210 & .002605 \\ -19.0473 & -.22851 & 15.1331 & .002646 & .050514 \end{bmatrix}$$

The variance–covariance matrix is needed to compute approximate standard errors of any quantities computed from the regression results.

Estimated Yields and Yield Losses

The quantities of particular interest are the estimated yields at specific levels of ozone and sulfur dioxide and the relative yield losses for given *changes* in the level of ozone or sulfur dioxide pollution. The use of the regression equation and the determination of variances of the estimated quantities are illustrated for

1. the estimated yield level for $X_1 = .05$ ppm and $X_2 = .10$ ppm, and
2. the relative yield losses expected from a change in the ozone level from $X_{1r} = .025$ ppm to $X_{1o} = .06$ ppm and from $X_{1r} = .025$ ppm to $X_{1o} = .08$ ppm. (X_{1r} and X_{1o} designate the reference level and the postulated new level of ozone, respectively.)

The estimated yield level for $X_1 = .05$ and $X_2 = .10$ is obtained by substitution of these values in the regression equation, along with $D = 0$

to give the average for the two blocks. This gives $\widehat{Y} = 456.83\ \mathrm{g\,m^{-1}}$. The variance is approximated by applying equation 15.41. This requires the partial derivatives of the nonlinear function with respect to each parameter, which for $\widehat{Y}$ (with $D = 0$) are

$$\begin{aligned}
\frac{\partial \widehat{Y}}{\partial \alpha_1} &= E, \\
\frac{\partial \widehat{Y}}{\partial \alpha_2} &= 0, \\
\frac{\partial \widehat{Y}}{\partial \beta} &= X_2 E, \\
\frac{\partial \widehat{Y}}{\partial \delta} &= (\alpha_1 + \beta X_2) E \left(\frac{\gamma}{\delta}\right) \left(\frac{X_1}{\delta}\right)^{\gamma}, \text{ and} \\
\frac{\partial \widehat{Y}}{\partial \gamma} &= -(\alpha_1 + \beta X_2) E \left(\frac{X_1}{\delta}\right)^{\gamma} \left[\ln\left(\frac{X_1}{\delta}\right)\right],
\end{aligned} \tag{16.8}$$

where

$$E = \exp\left[-\left(\frac{X_1}{\delta}\right)^{\gamma}\right].$$

Evaluating the partial derivatives by substituting the estimates of the parameters $X_1 = .05$ and $X_2 = .10$, and arranging them in a column vector, gives

$$\widehat{\boldsymbol{H}} = (\ .65606 \quad 0 \quad .065606 \quad 1,266.172 \quad -189.3025\ )'.$$

Thus, the variance of $\widehat{Y}$ is approximated by

$$\begin{aligned}
s^2(\widehat{Y}) &= \widehat{\boldsymbol{H}}'[\boldsymbol{s}^2(\widehat{\boldsymbol{\theta}})]\widehat{\boldsymbol{H}} \\
&= 78.6769
\end{aligned}$$

and so the estimated standard error is $s(\widehat{Y}) = 8.87$.

The estimated relative yield loss (RYL) resulting from a change in ozone pollution from X_{1r} to X_{1o} is

$$\begin{aligned}
RYL(X_{1r}, X_{1o}) &= \frac{\widehat{Y}(X_{1r}) - \widehat{Y}(X_{1o})}{\widehat{Y}(X_{1r})} \\
&= 1 - \frac{\exp[-(X_{1o}/\widehat{\delta})^{\widehat{\gamma}}]}{\exp[-(X_{1r}/\widehat{\delta})^{\widehat{\gamma}}]} \\
&= 1 - \exp(-DIF),
\end{aligned} \tag{16.9}$$

where

$$DIF = \left(\frac{X_{1o}}{\widehat{\delta}}\right)^{\widehat{\gamma}} - \left(\frac{X_{1r}}{\widehat{\delta}}\right)^{\widehat{\gamma}}.$$

For $(X_{1r}, X_{1o}) = (.025, .06)$, $RYL = .233$. That is, there is estimated to be a 23% loss in yield associated with an increase in ozone level from .025 ppm to .06 ppm. For $(X_{1r}, X_{1o}) = (0.025, .08)$, $RYL = .335$ or a 34% loss.

Variances of Relative Yield Losses

The partial derivatives of RYL are needed to obtain approximate variances of the estimated relative yield losses. The partial derivatives with respect to α_1, α_2, and β are zero since the function does not involve these parameters. The partial derivatives with respect to δ and γ are

$$\begin{aligned}\frac{\partial(RYL)}{\partial\delta} &= \left(\frac{\gamma}{\delta}\right)(DIF)\exp(-DIF), \text{ and} \\ \frac{\partial(RYL)}{\partial\gamma} &= \exp(-DIF)\left\{\left(\frac{X_{1o}}{\delta}\right)^{\gamma}\left[\ln\left(\frac{X_{1o}}{\delta}\right)\right] - \left(\frac{X_{1r}}{\delta}\right)^{\gamma}\left[\ln\left(\frac{X_{1r}}{\delta}\right)\right]\right\},\end{aligned} \tag{16.10}$$

where DIF is as defined following equation 16.9. Evaluating the derivatives at $\widehat{\theta}$ with $X_{1r} = .025$ and $X_{1o} = .06$ gives

$$\widehat{\boldsymbol{H}} = (\,0 \quad 0 \quad 0 \quad -1.338825 \quad -0.0091626\,)'$$

and

$$\begin{aligned}s^2(RYL) &= \widehat{\boldsymbol{H}}'[\boldsymbol{s}^2(\widehat{\theta})]\widehat{\boldsymbol{H}} \\ &= .0004445,\end{aligned}$$

or an estimated standard error of

$$s(RYL) = .0211.$$

For estimated relative yield loss for the $(X_{1r}, X_{1o}) = (0.025, .08)$ interval,

$$\widehat{\boldsymbol{H}} = (\,0 \quad 0 \quad 0 \quad -1.783608 \quad .0381499\,)'$$

and $s(RYL) = .033$. These estimated relative yield losses are summarized in the following table.

X_{1r}	X_{1o}	RYL	$s(RYL)$	*95% Confidence Interval*
.025	.06	.233	.0211	(0.191, .276)
.025	.08	.335	.0197	(0.295, .375)

16.2 Analysis of the Combined Soybean Data

The purpose of this analysis is to use the combined information from the four years of experiments, 1981 to 1984, to produce a response equation characterizing the response of Davis soybeans to ozone pollution, sulfur dioxide pollution, and moisture stress. First, the combined data are used

TABLE 16.8. *Soybean yield data, cultivar Davis, from the 1982, 1983, and 1984 studies on the effects of ozone, dispensing method, and moisture stress. (Data courtesy of Dr. A. S. Heagle, Plant Pathologist, N.C. State University and USDA; used with permission).*

1982:

	Block 1			*Block* 2		
Treatment	*Ozone*	Y_1	Y_2	*Ozone*	Y_1	Y_2
$CA20$	.0674	487.80	476.40	.0637	511.15	423.00
$CA40$	.0866	499.95	377.20	.0863	479.50	382.45
$CA60$	.1135	398.95	283.00	.1051	344.25	266.40
CF	.0149	653.30	583.40	.0222	652.70	600.70
NF	.0406	671.75	525.30	.0483	724.70	627.45
$P13$	.0635	599.65	412.15	.0672	620.85	513.55
$P16$	.0798	395.40	378.40	.0817	518.20	438.35
$P19$	.0933	354.55	288.85	.0902	419.25	325.50

1983:

Moisture	*Ozone*	*Block* 1		*Block* 2		*Block* 3	
Stress	*Trt.*	*Ozone*	Y	*Ozone*	Y	*Ozone*	Y
W	$CA30$	.0755	477.9	.0773	512.6	.0756	487.2
W	$CA60$	.0975	395.7	.1010	415.6	.1025	498.0
W	CF	.0299	535.9	.0277	642.0	.0255	639.5
W	NF	.0526	565.4	.0517	493.4	.0488	706.4
D	$CA30$	.0779	344.0	.0758	225.6	.0753	238.3
D	$CA60$	.0980	248.4	.1004	237.1	.0947	299.0
D	CF	—	—	.0314	448.8	.0293	282.5
D	NF	.0523	271.9	.0533	211.2	.0520	255.3

1984:

Moisture	*Ozone*	*Block* 1		*Block* 2	
Stress	*Trt.*	*Ozone*	Y	*Ozone*	Y
W	CF	.024	344	.024	416
W	NF	.043	438	.045	428
W	$CA15$	.065	268	.069	283
W	$CA30$	.082	293	.082	344
W	$CA45$	.087	297	.095	231
W	$CA60$	.104	249	.112	214
D	CF	—	—	.027	297
D	NF	.043	279	.047	330
D	$CA15$	.066	254	.064	363
D	$CA30$	.077	202	.081	213
D	$CA45$	.095	215	.093	229
D	$CA60$	.107	138	.105	216

TABLE 16.9. *Pooled residual sums of squares for several choices of λ for the Box–Cox transformation on the 1981 to 1984 soybean experiments.*

λ	*Pooled SS*
-1	$243,433$
$-.5$	$207,201$
$0 = \ln(Y)$	$198,633$
$.5$	$212,027$
1	$249,122$

to check the validity of the assumptions of normality and constant variance. The 1981 data were given in Table 16.1. The 1982, 1983, and 1984 data are given in Table 16.8.

Checking Normality and Constancy of Variance

The individual yearly experiments do not provide sufficient information to critically check normality and constancy of variance. Therefore, data from all experiments were combined to check these assumptions. In 1983 and 1984, half of each chambered plot was used destructively for physiological measurements and, consequently, yield was measured on only the remaining half. In order to keep plot sizes comparable over years, all analyses used the "half plot" yield as the basic unit. Thus, the north (N) and south (S) halves of each plot in 1981 and 1982 were used as different data sets. (The correlations between the subsets of data in the two experiments were ignored for this analysis of residuals.) The appropriate analysis of variance was run on each data set and the residuals from all analyses were combined to study their behavior. The combined data set has a total of 174 observations and the pooled residuals have 80 degrees of freedom. A missing observation in each of 1983 and 1984 made the data unbalanced from the analysis of variance point of view. (The analysis of unbalanced data is discussed in Chapter 17.) For present purposes, the effects and dummy variables are defined so as to give a full rank model and regression analyses are used.

The plot of the residuals from the analyses of variance versus $\widehat{Y}$, Figure 16.4, showed a tendency for increased dispersion at the higher values of $\widehat{Y}$. The normal plot of the residuals, Figure 16.5, showed a very slight S-shaped curvature. On the basis of these graphical results, the Box–Cox method was used to find a transformation on Y that would improve normality and constancy of variance.

Logarithmic Transformation Used

The criterion used for choice of power transformation was minimum pooled residual sum of squares from the analyses of variance for the four years of data. The pooled residual sums of squares for several choices of λ in the Box–Cox transformation are given in Table 16.9. Quadratic interpolation using the three middle points indicated that the minimum was near $\lambda = -.05$ with SS[Res(λ)]$= 198,471$. The plot of these residual sums

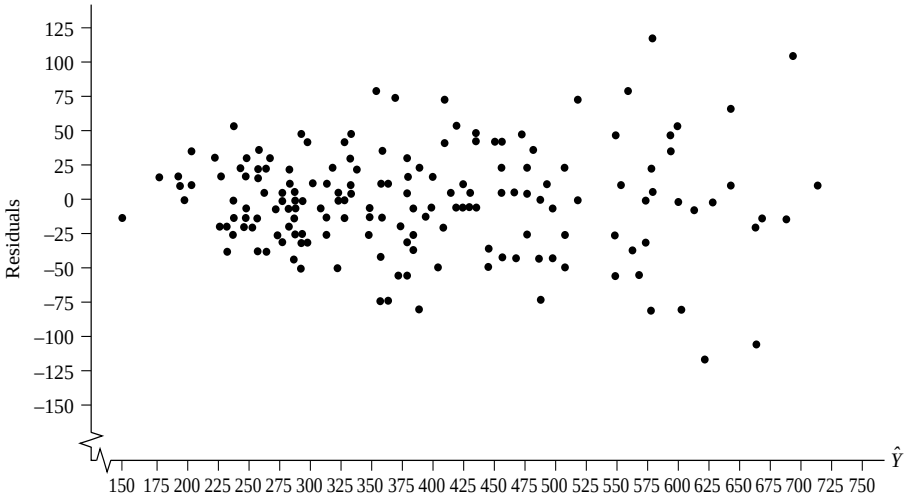

FIGURE 16.4. *Pooled residuals from the separate analyses of variance of yield for the 1981 to 1984 soybean studies plotted against* $\widehat{Y}$.

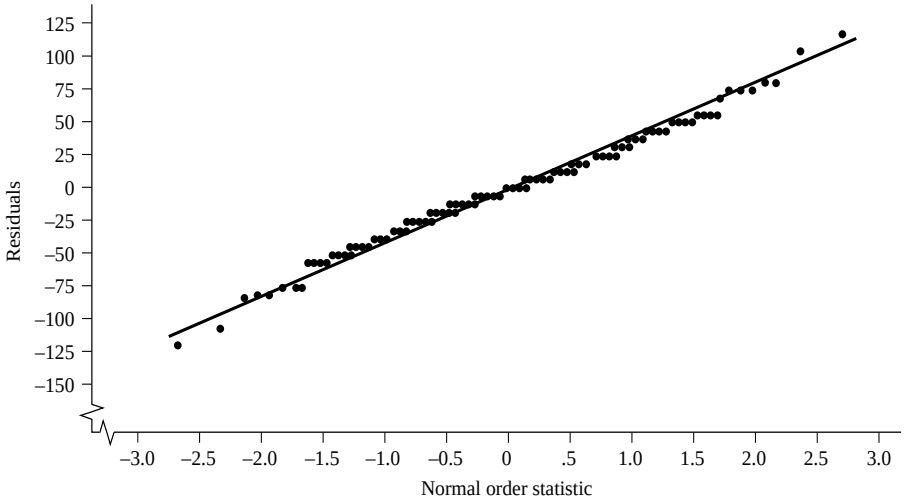

FIGURE 16.5. *Normal plot of the pooled residuals from the analyses of variance of yield for the 1981 to 1984 soybean studies.*

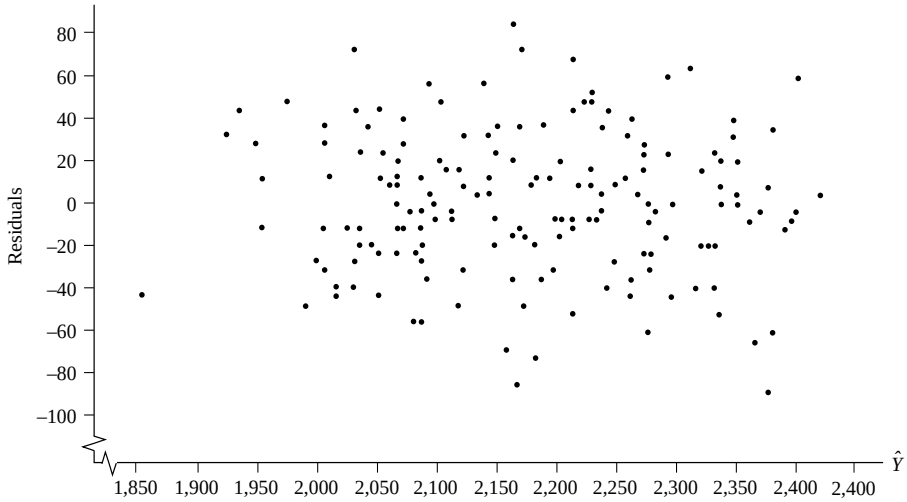

FIGURE 16.6. *Pooled residuals from the analyses of variance of ln(Y) for the 1981 to 1984 soybean studies plotted against* $\widehat{Y}$.

of squares and the confidence interval estimate of λ, presented in Chapter 12, Figure 12.4, suggested a logarithmic transformation. The analyses of variance were repeated using $\ln(Y)$ as the dependent variable. The pooled residuals obtained from the analyses on $\ln(Y)$ showed better behavior both with respect to constancy of variance, Figure 16.6, and normality, Figure 16.7. Consequently, the response model for the combined data is developed using $\ln(Y)$ as the dependent variable.

Full Model

A complete model for the combined 1981 to 1984 soybean experiments needs to account for differences among years, differences among blocks in years, the joint ozone and sulfur dioxide response in 1981, the joint ozone and method of dispensing effects in 1982, the joint ozone and moisture stress effects in 1983 and 1984, and possible ozone by year, ozone by dispensing method, ozone by moisture, and ozone by sulfur dioxide interaction effects. However, previous analyses had shown the main and interaction effects due to ozone dispensing methods not to be significant and, consequently, these effects are not included. The year, block, and moisture stress effects are incorporated in the model with the use of dummy variables. A plot of the data suggested that a linear regression term would adequately account for the average sulfur dioxide effects. The logarithm of the exponential component in the original Weibull model gives $-(X_1/\delta)^\gamma$, suggesting that the ozone response on the logarithmic scale can be characterized by a nonlinear term $\beta(X_1)^\gamma$, where $\beta = -(1/\delta)^\gamma$. Thus, a power parameter γ on the level of ozone is included in the full model. The interaction effects are incorporated as product terms in the usual way.

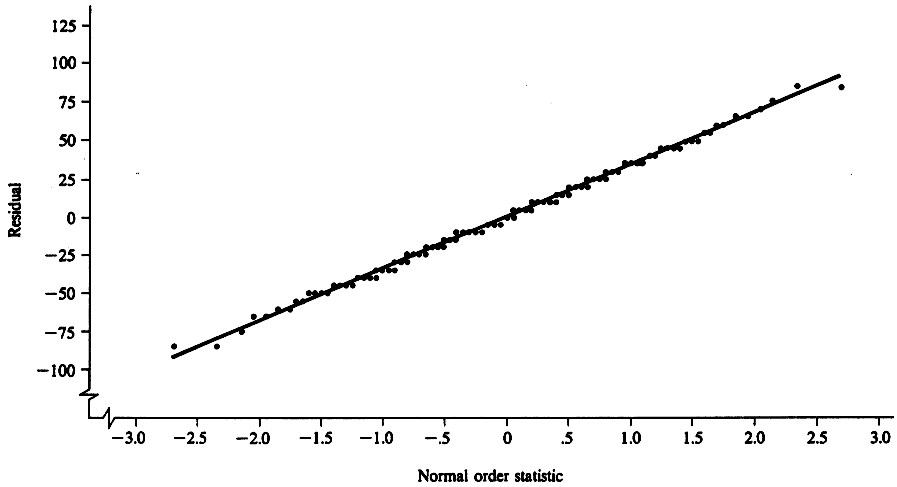

FIGURE 16.7. *Normal plot of the pooled residuals from the analyses of variance of ln(Y) for the 1981 to 1984 soybean studies.*

Let T_1, T_2, T_3, and T_4 be dummy variables identifying the four years, respectively, by taking the value of 1 if the observation is from the year indicated by the subscript and 0 otherwise. Let R_{11}, R_{21}, R_{31}, R_{32}, and R_{41} be dummy variables to account for block differences within each year. Each R_{ij} takes the value 1 if the observation is from the jth block in the ith year and 0 otherwise. Notice that there is one less R_{ij} dummy variable for each year than the number of blocks in that year. The moisture-stressed plots are identified by $M = 1$ and the well-watered plots with $M = 0$. Let MI be a dummy variable to allow for a moisture stress by year interaction between 1983 and 1984, taking the value of 1 if the plot is a moisture-stressed plot in 1983, -1 if it is a moisture-stressed plot in 1984, and 0 otherwise. Thus, the full model, without subscripts to identify the experimental unit, is

$$\begin{aligned} \ln(Y) \quad = \quad & \beta_1 T_1 + \beta_2 T_2 + \beta_3 T_3 + \beta_4 T_4 \\ & + \beta_5 R_{11} + \beta_6 R_{21} + \beta_7 R_{31} + \beta_8 R_{32} + \beta_9 R_{41} \\ & + \beta_{10} M + \beta_{11} MI + \beta_{12} X_2 + \beta_{13} X_1^\gamma \\ & + \beta_{14} X_2 X_1^\gamma + \beta_{15} M X_1^\gamma + \beta_{16} T_1 X_1^\gamma \\ & + \beta_{17} T_2 X_1^\gamma + \beta_{18} T_3 X_1^\gamma + \epsilon, \end{aligned} \tag{16.11}$$

where X_2 is the level of sulfur dioxide and X_1 is the level of ozone. The product term MX_1^γ allows the moisture-stressed plots to have a different response to ozone, and the last three terms allow for year by ozone interactions. This is a nonlinear model only because of the power parameter on X_1.

Fitting the Model

This model was fitted using the derivative-free option in PROC NLIN (SAS Institute Inc., 1989b). The starting values for the parameters were $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 6.5$, $\beta_{12} = -1$, $\beta_{13} = -5$, $\gamma = 1$, and all others zero. Although convergence was obtained, the derivative-free method appeared to be inefficient. With 19 parameters in the model, 20 iterations are required with the derivative-free method before the numerical estimates of all derivatives can be computed. In this particular case, 7 additional iterations were made and then iterations were restarted with a smaller grid around the current estimates. This required an additional 20 iterations to recompute the numerical derivatives and a final 5 iterations to reach convergence. Thus, there were a total of 52 iterations to find the solution. Except for the terms involving X_1^{γ}, this model is linear in the parameters. In models that are "nearly" linear in the parameters, convergence is usually fairly rapid when the derivatives are specified. It is left as an exercise for the reader to fit this model using derivatives.

Summary of the Analysis

The summary of this analysis is shown in Table 16.10. The asymptotic confidence intervals can be used as guides to the significance of the various parameters. This is equivalent to testing the corresponding hypotheses using the Wald statistics. The year parameters β_1 to β_4 are different from zero, as expected, and are retained in the model. The block differences within years are not significantly different from zero as shown by the confidence intervals for β_5 to β_9 overlapping zero. However, the block effects are part of the original experimental designs and are kept in the model. The average moisture stress effect β_{10}, the moisture stress by year interaction effect β_{11}, and the regression coefficients for sulfur dioxide β_{12}, and ozone β_{13}, are significantly different from zero. The analysis gives no indication of an ozone by sulfur dioxide interaction β_{14}, a moisture stress by ozone interaction β_{15}, nor any year by ozone interactions β_{16}, β_{17}, and β_{18}.

Rather than dropping all nonsignificant interaction terms at one time, the analysis proceeds more cautiously by dropping first the year by ozone interaction effects and then dropping other interaction effects if they remain unimportant. This protects against dropping effects that may become significant after other effects in the model have been dropped, and it provides the opportunity to test the significance of the effects with the likelihood ratio test using the difference in residual sums of squares from the two models.

Dropping Year by Ozone Interactions

The residual sum of squares from the model in which all year by ozone interaction effects β_{16}, β_{17}, and β_{18} are set equal to zero is SS(Res) = 3.8088 with 158 degrees of freedom. Comparing this to the residual sum of squares from the full model, Table 16.8, and computing the F-statistic gives

$$F = \frac{(\mathrm{SS}(\mathrm{Res}_{\mathrm{reduced}}) - \mathrm{SS}(\mathrm{Res}_{\mathrm{full}}))/q}{\mathrm{SS}(\mathrm{Res}_{\mathrm{full}})/(n-p)}$$

TABLE 16.10. *Summary of the nonlinear least squares analysis of ln(seed yield) for the 1981–1984 soybean data using the full model.*

Source	*d.f.*	*Sum of Squares*	*Mean Square*
Model	19	6,089.9388	320.5231
Residual	155	3.6824	.0238
Uncorrected total	174	6,093.6212	
(Corrected total)	173	20.0321	

	Parameter	*Estimate*	*Asymptotic Std. Error*	*Asymptotic 95% Confidence Interval*
β_1	Years	6.4828	.0948	$(6.2956, 6.6700)^a$
β_2		6.6811	.1092	$(6.4654, 6.8968)^a$
β_3		6.5359	.1129	$(6.3129, 6.7589)^a$
β_4		6.2472	.1278	$(5.9948, 6.4997)^a$
β_5	Blocks/Years	.0604	.0315	$(-.0018, .1227)$
β_6		$-.0614$	.0546	$(-.1693, .0465)$
β_7		$-.0346$	.0805	$(-.1936, 0.1245)$
β_8		$-.0552$	.0771	$(-0.2075, .0971)$
β_9		$-.1005$	.0647	$(-.2284, 0.0273)$
β_{10} :	M	$-.4712$	.1206	$(-.7094, -.2330)^a$
β_{11} :	M × Yr	$-.2059$	.0460	$(-.2967, -.1151)^a$
β_{12} :	SO_2	-10194	.2563	$(-1.5257, -.5130)^a$
β_{13} :	O_3	-9.3216	4.4734	$(-18.1584, -.4848)^a$
β_{14} :	$SO_2 \times O_3$	.3332	4.0286	$(-7.6249, 8.2913)$
β_{15} :	M × O_3	.7626	2.1813	$(-3.5464, 5.0716)$
β_{16}	Yr × O_3	.4954	1.9172	$(-3.2918, 4.2827)$
β_{17}		$-.9537$	2.2033	$(-5.3061, 3.3987)$
$\beta 18$		3.9134	2.7087	$(-1.4375, 9.2642)$
γ :	O_3 power	1.1287	.2330	$(.6684, 1.5889)^a$

[a]95% confidence interval does not overlap zero.

$$= \frac{(3.8088 - 3.6824)/3}{3.6824/155} = 1.77,$$

where $q = 3$ is the number of constraints placed on the parameters. This is an approximate F-test with q and $n-p$ degrees of freedom and is nonsignificant. Gallant (1987) shows that this is equivalent to the likelihood ratio test. This confirms the decision based on the Wald statistic that β_{16}, β_{17}, and β_{18} are not different from zero. The reduced model continues to show that β_{14} and β_{15}, the sulfur dioxide by ozone interaction and the moisture stress by ozone interaction, are not different from zero.

Moisture Stress by Ozone Interaction

The model without β_{16}, β_{17}, and β_{18} is adopted as the full model for testing the significance of β_{15}. The reduced model, with β_{15} set equal to zero, gives $\text{SS}(\text{Res}_{reduced}) = 3.8447$ with 159 degrees of freedom. The likelihood ratio test of $H_0 : \beta_{15} = 0$ gives

$$F = \frac{3.8447 - 3.8088}{3.8088/158} = 1.49$$

which, with 1 and 158 degrees of freedom, is not significant and β_{15} is dropped from the model.

Sulfur Dioxide by Ozone Interaction

The Wald confidence interval for this model with β_{15} dropped continues to indicate that β_{14} is not significantly different from zero. The model was further reduced by setting $\beta_{14} = 0$. This gives $\text{SS}(\text{Res}_{reduced}) = 3.8449$ with 160 degrees of freedom (Table 16.11). Comparing this to the residual sum of squares for the previous model gives

$$F = \frac{3.8449 - 3.8447}{3.8447/159} = .01,$$

which is nonsignificant. Thus, the sulfur dioxide by ozone interaction effect is also not important and can be dropped from the model. The only interaction effect remaining is the moisture stress by year interaction β_{11}, which is significant in this reduced model. Likewise, the moisture stress effect, the sulfur dioxide effect, and the ozone effect remain significant as judged by their 95% approximate confidence interval estimates.

Setting $\gamma = 1$

The final stage in simplifying this model relates to the power parameter on X_1. The logical null hypothesis for γ is $H_0 : \gamma = 1.0$ which, if true, removes the nonlinearity of the model. The point estimate of γ (in the last reduced model) is $\widehat{\gamma} = 1.078$ and the 95% confidence interval estimate is (0.625, 1.530). There appears to be no reason to reject the null hypothesis that $\gamma = 1.0$. Since the model with $\gamma = 1$ is linear in the parameters, PROC GLM with the no-intercept option is used to fit this final reduced model. The results for this model are summarized in Table 16.12. The likelihood ratio test of the null hypothesis that $\gamma = 1.0$ gives

$$F = \frac{3.8479 - 3.8449}{3.8449/160} = .12,$$

TABLE 16.11. *Summary of the nonlinear least squares analysis of ln(seed yield) for the 1981 to 1984 soybean data using the reduced model.*

Source	*d.f.*	*Sum of Squares*	*Mean Square*
Model	14	6,089.7763	434.9840
Residual	160	3.8449	.0240
Uncorrected total	174	6,093.6212	
(Corrected total)	173	20.0321	

	Parameter	*Estimate*	*Asymptotic Std. Error*	*Asymptotic 95% Confidence Interval*
β_1	Years	6.4910	.0870	$(6.3191, 6.6629)^a$
β_2		6.6180	.0986	$(6.4233, 6.8127)^a$
β_3		6.6931	.1065	$(6.4828, 6.9034)^a$
β_4		6.2316	.1022	$(6.0299, 6.4333)^a$
β_5	Blocks/Years	.0603	.0317	$(-.0023, .1228)$
β_6		$-.0616$	.0549	$(-.1701, .0469)$
β_7		$-.0143$	.0805	$(-.1732, .1446)$
β_8		$-.0496$	.0775	$(-.2027, .1035)$
β_9		$-.0987$	.0648	$(-.2267, .0293)$
β_{10} :	M	$-.4283$	.0459	$(-.5190, -.3376)^a$
β_{11} :	M × Yr	$-.2020$	.0458	$(-0.2925, -0.1114)^a$
β_{12} :	SO_2	$-.9996$	.1083	$(-1.2135, -.7856)^a$
β_{13} :	O_3	-7.9230	3.2129	$(-14.2682, -1.5778)^a$
γ :	O_3 power	1.0778	.2292	$(0.6253, 1.55304)^a$

[a]95% confidence interval does not overlap zero.

TABLE 16.12. *Summary of the analysis of ln(seed yield) for the 1981 to 1984 soybean data using the final linear model.*

Source	*d.f.*	*Sum of Squares*	*Mean Square*
Model	13	6,089.7733	468.4441
Residual	161	3.8479	.0239
Uncorrected total	174	6,093.6212	
(Corrected total)	173	20.0321	

	Parameter	*Estimate*	*Asymptotic Std. Error*	*Asymptotic 95% Confidence Interval*
β_1	Years	6.5193	.0394	$(6.4421, 6.5966)^a$
β_2		6.6486	.0473	$(6.5559, 6.7414)^a$
β_3		6.7227	.0678	$(6.5899, 6.8555)^a$
β_4		6.2611	.0611	$(6.1413, 6.3809)^a$
β_5	Blocks/Years	.0605	.0316	$(-.0014, .1224)$
β_6		$-.0627$	.0547	$(-.1699, .0446)$
β_7		$-.0133$	.0802	$(-.1704, .1439)$
β_8		$-.0494$	.0773	$(-.2009, .1021)$
β_9		$-.0981$	.0646	$(-.2247, .0285)$
β_{10} :	M	$-.4275$	.0457	$(-.5171, -.3378)^a$
β_{11} :	M × Yr	$-.2019$	.0457	$(-.2915, -0.1123)^a$
β_{12} :	SO_2	$-.9977$	.1079	$(-1.2091, -.7862)^a$
β_{13} :	O_3	-6.9170	.3869	$(-7.6753, -6.1587)^a$

[a]95% confidence interval does not overlap zero.

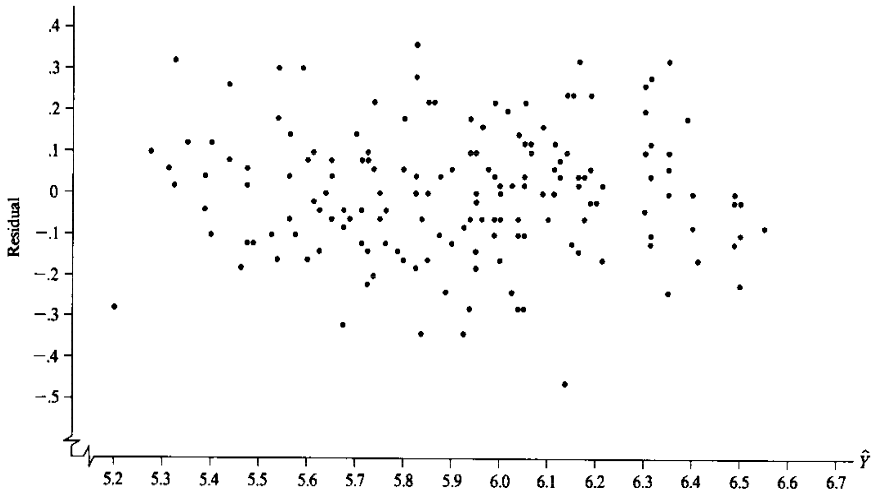

FIGURE 16.8. *Pooled residuals from the final response model for the 1981 to 1984 soybean data plotted against* $\widehat{Y}$.

which is clearly nonsignificant. All the remaining terms in this model, except the block effects, are significant. The plot of the residuals versus $\widehat{Y}$ (Figure 16.8) and the normal plot of the residuals (Figure 16.9) give no reason for concern about inadequacies in the model.

Relative Yield Losses

Thus, the final model to represent the 1981 to 1984 soybean response to sulfur dioxide and ozone shows a decline in $\ln(Y)$ of 6.9 units per ppm increase in ozone and a decline of 1.0 unit per ppm increase in sulfur dioxide. Translating this regression equation back to the original scale, by taking the antilog, and computing the relative yield loss for changes in ozone gives

$$\begin{aligned} RYL &= 1 - \exp[\widehat{\beta}_{13}(X_{1o} - X_{1r})] \\ &= 1 - \exp[-6.917(X_{1o} - X_{1r})]. \end{aligned}$$

The partial derivatives of RYL with respect to the parameters in the model are all zero except for the partial derivative with respect to β_{13},

$$\frac{\partial(RYL)}{\partial\beta_{13}} = \{-\exp[\beta_{13}(X_{1o} - X_{1r})]\}(X_{1o} - X_{1r}).$$

Thus, $s^2(RYL)$ involves only the one variance $s^2(\widehat{\beta}_{13})$, multiplied by the square of the partial derivative evaluated at $\widehat{\beta}_{13}$. The estimated relative yield losses (RYL), their approximate standard errors, and the 95% approximate confidence interval estimates for several choices of X_{1o} are given in Table 16.13.

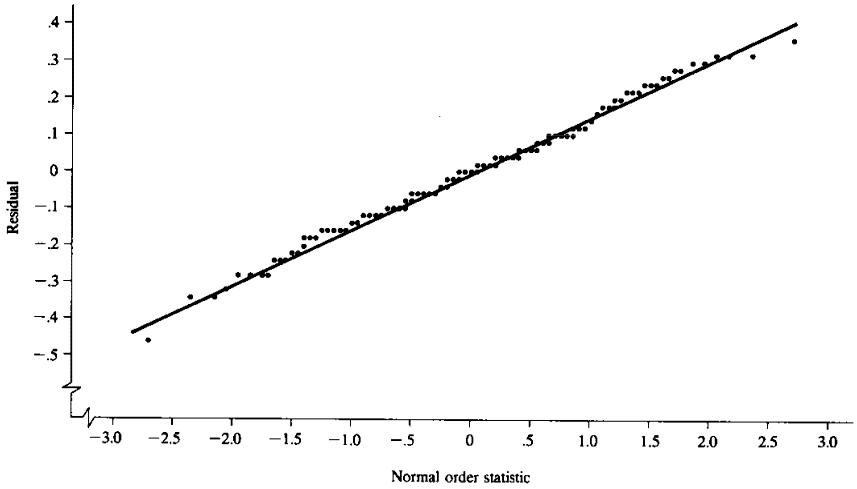

FIGURE 16.9. *Normal plot of the residuals from the final response model for the 1981 to 1984 soybean data.*

TABLE 16.13. *Estimates of relative yield losses, their approximate standard errors, and approximate 95% confidence interval estimates.*

Estimation Interval			*Approx.*	*Approximate 95%*
X_{1r}	X_{1o}	*RYL*	*s*(*RYL*)	*Confidence Interval*
.025	.03	.034	.0019	(.030, .038)
	.04	.099	.0052	(.088, .109)
	.05	.158	.0081	(.143, .175)
	.06	.215	.0106	(.194, .236)
	.07	.267	.0128	(.242, .292)
	.08	.316	.0145	(.288, .345)

An alternative approach to obtain confidence interval estimates of *RYL* in this example is to first compute the confidence interval estimates of

$$\ln(1 - RYL) = \beta_{13}(X_{1o} - X_{1r})$$

as

$$[\widehat{\beta}_{13} \pm t_{(\alpha/2,\nu)} s(\widehat{\beta}_{13})](X_{1o} - X_{1r})$$

and then transform the limits. The antilogs of these limits subtracted from unity give the limits on *RYL*. In this example, the limits obtained in this way agreed to the third decimal with those in Table 16.13 in all cases except for a difference of one in the third decimal when $X_{1o} = .08$.

The estimates of relative yield losses are very similar to those obtained from the 1981 data alone, .215 versus .233 for $X_{1o} = .06$ and .316 versus .335 for $X_{1o} = .08$. The standard errors are appreciably smaller as expected from the use of additional information, .011 versus .018 and .015 versus .033.

Consequences of Setting $\gamma = 1$

Most of the point estimates of the parameters changed only slightly when γ was set equal to 1.0 in the last step of developing this model. The estimate of β_{13} changed most noticeably from -7.92 to -6.924, but this was to be expected since β_{13} is now the coefficient on X_1, not $X^\gamma - 1$. The standard error on β_{13}, however, decreased to only one-tenth its previous value when γ was set equal to 1.0. This greatly increased precision in the estimate of β_{13} is the result of eliminating a collinearity problem; the correlation between $\widehat{\beta}_{13}$ and $\widehat{\gamma}$ was .990. This high negative correlation means that changes in one parameter could be offset by compensating changes in the other parameter; the joint confidence region for the two parameters would be a very elongated ellipse.

16.3 Exercises

16.1. The polynomial response model adopted for the 1981 soybean data did not use the O_3 quadratic $\times$ SO_2 linear interaction term but the text suggested that it would be the next most logical term to test. Add the term $X_1^2 X_2$ to the model shown in equation 16.4 and fit the 1981 soybean data (Table 16.1). Compare these results to those obtained from the model shown in equation 16.4 and test the significance of the new term. State your conclusions.

16.2. Determine whether cubic terms in either ozone or sulfur dioxide would have significantly improved the polynomial response model, equation 16.4, for the 1981 soybean data.

16.3. The sums of squares due to the polynomial terms in the analysis of the 1981 data were not partitions of the analysis of variance due to the

fact that a given pollutant treatment was not constant over the levels of the other pollutant and the two replications. Rerun the polynomial analysis using the *mean* ozone level for each ozone treatment and the *mean* sulfur dioxide level for each sulfur dioxide treatment; that is, use $\overline{X}_{i\cdot\cdot 1}$ and $\overline{X}_{\cdot j\cdot 2}$. How does this change your results? What polynomial model do you adopt? Are the sums of squares due to the polynomial terms in ozone level partitions of the ozone treatment sum of squares? Are the sequential sums of squares due to the polynomial terms in sulfur dioxide level partitions of the sulfur dioxide treatment sum of squares?

16.4. Refit the Weibull model, equation 16.6, to the 1981 soybean data using one of the methods that require derivatives. Compare your results to those reported in the text for the derivative-free method (Table 16.6).

16.5. Use the likelihood ratio test with the 1981 data to test the null hypothesis that the parameter γ in the Weibull model is equal to 1. (Refit the nonlinear model you obtain from the Weibull model by setting $\gamma = 1$. Test the increase in residual sums of squares of this "reduced" model over the "full" model against the residual mean square from the "full" model using an F-test.) Is the result of this test consistent with the conclusion you reach if you use the Wald test?

16.6. The nonlinear model used in relating $\ln(Y)$ to the treatment variables in the combined 1981–1984 data, equation 16.11, was fit using the derivative-free method. Convergence was slow because of the large number of parameters in the model. Refit the model using one of the methods requiring derivatives. Use the same starting values used in the text. Was convergence obtained or assumed? How many iterations were required? Does the solution agree with that from the derivative-free method, Table 16.10? Does it appear reasonable from these results to set $\gamma = 1$? On what do you base your answer?

16.7. The nonlinear model used in relating $\ln(Y)$ to the treatment variables in the combined 1981–1984 data, equation 16.11, can also be fit using linear least squares. If γ is fixed at some value, the model is linear in the parameters. Fitting this linear model gives a residual sum of squares that is conditional on the chosen value of γ. Repeating the analysis for a series of values of γ from which the one with the minimum residual sum of squares is chosen will eventually lead to the least squares solution if small enough steps in γ are used. Obtain the least squares solution by this grid search method and compare your results with those obtained from nonlinear least squares. Use $\gamma = 1.0$, 1.1, 1.12, 1.13, 1.14, 1.20 as trial values.

17

ANALYSIS OF UNBALANCED DATA

Chapter 9 introduced the use of class variables, with which the classical analyses of variance for balanced data became special cases of least squares regression.

This chapter discusses the analysis of unbalanced data using least squares regression with class variables. Emphasis is on understanding estimability and the estimable functions of the parameters that are tested by the various sums of squares. Treatment means adjusted for the effects of imbalance are defined.

Definition of "Balance"

The classical analyses of variance for the standard experimental designs are appropriate only for data from balanced experiments. The common definition of balance is that an experiment is **balanced** if all cells of the data table have equal numbers of observations. Critical to this definition is the understanding, which is often not stated, that the "cells" of the data table must include a cell for every possible combination of the levels of all treatment factors and, if blocking is used, for each combination of treatments and blocks. These conditions imply that every possible multiway table involving different treatment factors (and blocks) will have the same number of observations in all cells of the table.

The balance in the data allows contrasts, and sums of squares associated with the contrasts, to be computed directly from corresponding marginal data tables. (Marginal data tables are constructed by summing across factors not involved in the contrast of immediate interest.) Without balance,

contrasts on the marginal sums (or means) will include unwanted effects of other treatment factors. This leads to a "working" definition of balance:

> Data are balanced if the contrasts of interest, and sums of squares for the contrasts, can be computed directly from the marginal sums (or means) for the factors involved in the contrast.

[There are other definitions of balance; see, for example, Basson (1965). The definition given here is more restrictive than necessary. Unequal but proportional numbers, for example, may be sufficient for some cases.]

Methods for Unbalanced Data

In this chapter, methods of analyzing unbalanced data are discussed. The first two methods attempt to avoid the effects of imbalance by applying least squares analysis to cell means. (The analysis of cell means is not to be confused with the use of the means model.) The third method applies least squares principles to obtain estimates of estimable functions of the parameters and sums of squares for relevant testable hypotheses. The emphasis in this text is on the application of least squares to the classical effects models. The reader is referred to Hocking (1985) for a thorough discussion of the alternative of using means models.

Many procedures for the analysis of unbalanced data concentrate more on partitioning sums of squares than on the hypotheses being tested. Consequently, the hypotheses often are not the most meaningful and may not even be clearly specified. [See Hocking and Speed (1975), Speed and Hocking (1976), and Speed, Hocking, and Hackney (1978), for extensive discussions on analysis of unbalanced data.] The emphasis in this text is on estimable functions and testable hypotheses in order to enhance the reader's understanding of the analyses. The general linear models procedure, PROC GLM (SAS Institute Inc., 1989b), is used extensively. This procedure computes four types of sums of squares, which include most of the options usually considered, and provides the estimable functions of the parameters being tested by these sums of squares. This book concentrates on the SAS Type I and Type III testable hypotheses and sums of squares. [The reader is referred to Freund, Littell, and Spector (1986) and Searle and Henderson (1979) for more discussion on PROC GLM.]

17.1 Sources Of Imbalance

Imbalance in data can arise for different reasons and at different "levels" in the experiment. The imbalance may be deliberate in the design of the experiment or it may be the result of failure to give adequate consideration to the design. Certain treatment combinations, such as simultaneous high temperature and high pressure, may not be possible for the particular system being studied, or limited resources may restrict the number of treatment combinations that can be handled.

Most often, however, unequal numbers arise due to accidents during the experiment; contamination of material or mortality of animals or plants causes the loss of experimental units, sample material is lost or handled incorrectly before it can be analyzed and data recorded, or data are recorded incorrectly and subsequently have to be discarded. The loss of data may occur at the sampling unit level (if sampling units are used), at the experimental unit level, or at the treatment level. The loss of an entire treatment will cause confounding of effects if the treatment is one of a factorial set of treatments.

Although imbalance is occasionally deemed necessary because of the nature of the system being studied and often occurs accidentally, the availability of computing power and general analysis programs such as PROC GLM should never be the justification for conducting an unbalanced experiment. As shown, the analysis and interpretation of results are much more difficult for unbalanced data and, frequently, the imbalance will result in the loss of important information.

17.2 Effects Of Imbalance

Two-Way Model

The confounding effects of imbalance are illustrated with a 2×3 factorial set of treatments in a completely random experimental design. The effects model for this case is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \tag{17.1}$$

where α_i and β_j are the effects of the ith and jth levels of treatment factors A and B, respectively; γ_{ij} is the interaction effect between the ith level of A and the jth level of B, and ϵ_{ijk} is the random error associated with the observation from the kth experimental unit receiving the ijth treatment combination.

Balanced Data: Expectations of Cell Means

When the data are balanced, the sums of squares for the standard analysis of variance are computed directly from contrasts on the treatment means. Functions of the squared differences among the A treatment means generate the sum of squares for the A treatment factor unconfounded by the effects of factor B, and vice versa. The simplicity of the analysis of variance is a direct result of the balance in the data. The reason is evident from the expectations of the cell and marginal means (Table 17.1). Expectations of the cell means are obtained by averaging the fixed effects in the model, equation 17.1, over subscript k, the observations within each cell. In this case, the fixed effects do not involve the subscript k so that the expectation for the ijth cell mean is

$$\mathcal{E}(\overline{Y}_{ij.}) = \mu + \alpha_i + \beta_j + \gamma_{ij}.$$

The expectations of the marginal means are obtained by averaging the cell expectations over each row or column, as the case may be, giving equal

TABLE 17.1. *The expectations of the cell means and the marginal means for a* 2×3 *factorial in a completely random experimental design. The marginal means are computed assuming equal numbers of observations in each cell.*

	B			
A	1	2	3	$\mathcal{E}(\overline{Y}_{i\cdot\cdot})^a$
1	$\mu+\alpha_1$ $+\beta_1+\gamma_{11}$	$\mu+\alpha_1$ $+\beta_2+\gamma_{12}$	$\mu+\alpha_1$ $+\beta_3+\gamma_{13}$	$\mu+\alpha_1$ $+\overline{\beta}_{\cdot}+\overline{\gamma}_{1\cdot}$
2	$\mu+\alpha_2$ $+\beta_1+\gamma_{21}$	$\mu+\alpha_2$ $+\beta_2+\gamma_{22}$	$\mu+\alpha_2$ $+\beta_3+\gamma_{23}$	$\mu+\alpha_2$ $+\overline{\beta}_{\cdot}+\overline{\gamma}_{2\cdot}$
$\mathcal{E}(\overline{Y}_{\cdot j\cdot})$	$\mu+\overline{\alpha}_{\cdot}$ $+\beta_1+\overline{\gamma}_{\cdot 1}$	$\mu+\overline{\alpha}_{\cdot}$ $+\beta_2+\overline{\gamma}_{\cdot 2}$	$\mu+\overline{\alpha}_{\cdot}$ $+\beta_3+\overline{\gamma}_{\cdot 3}$	$\mu+\overline{\alpha}_{\cdot}$ $+\overline{\beta}_{\cdot}+\overline{\gamma}_{\cdot\cdot}$

[a]The bar over the symbol indicates the average over the subscript that has been replaced with a dot.

weight to each cell. The equal weight for each cell simulates the averaging one would do if all cells had the same number of observations.

Balanced Data: Expectations of Contrasts

The expectations of all marginal means for the B factor contain exactly the same function of the α_i effects (Table 17.1). Thus, all α_i effects will cancel in the expectation of any *contrast* on the marginal means for the B factor. For example, the contrast between levels 1 and 2 for the B factor has expectation

$$\mathcal{E}(\overline{Y}_{\cdot 1\cdot} - \overline{Y}_{\cdot 2\cdot}) = \beta_1 - \beta_2 + (\overline{\gamma}_{\cdot 1} - \overline{\gamma}_{\cdot 2}), \tag{17.2}$$

which involves no α_i. The result is that any contrast of interest on the β_j effects is estimated with the same contrast on the marginal means for the B factor and is not *confounded* with the effects of the A factor. Similarly, any contrast of interest on the α_i effects is estimated with the same contrast on the marginal means for the A factor without being confounded with β_j effects. It follows that the sums of squares for contrasts among the A factor means will not involve the β_j effects and sums of squares for contrasts among the B factor means will not involve the α_i effects when the data are balanced.

Balanced Data: Interaction Effects

The interaction effects γ_{ij} do not cancel in contrasts on the marginal means in balanced data, but they are present in very specific ways. The expectation of any contrast on marginal means in balanced data involves the same contrast on the simple marginal averages of the γ_{ij} effects. There is no function of the data that will estimate a contrast on main effects without involving interaction effects, if the model contains interaction effects, unless constraints are imposed on the parameters. In this discussion, all results are presented in terms of the full model without constraints. Thus, contrasts involving *only* main effects, $\alpha_1 - \alpha_2$, for example, are nonestimable.

Unbalanced Data: Expectations

The effect of imbalance is illustrated by considering the same set of factorial treatments but with unequal cell numbers. Let

$$\begin{array}{lll} n_{11}=1, & n_{12}=2, & n_{13}=1, \\ n_{21}=3, & n_{22}=1, & n_{23}=1. \end{array} \tag{17.3}$$

The expectations of the cell means remain as shown in Table 17.1. However, the expectations of the marginal means now are weighted averages of the expectations of the cell means, where the weighting is by n_{ij}. Thus,

$$\begin{aligned} \mathcal{E}(\overline{Y}_{1\cdot\cdot}) &= \frac{[\mathcal{E}(\overline{Y}_{11\cdot}) + 2\mathcal{E}(\overline{Y}_{12\cdot}) + \mathcal{E}(\overline{Y}_{13\cdot})]}{4} \\ &= \mu + \alpha_1 + \frac{\beta_1 + 2\beta_2 + \beta_3}{4} + \frac{\gamma_{11} + 2\gamma_{12} + \gamma_{13}}{4} \end{aligned} \tag{17.4}$$

and

$$\begin{aligned} \mathcal{E}(\overline{Y}_{2\cdot\cdot}) &= \frac{[3\mathcal{E}(\overline{Y}_{21\cdot}) + \mathcal{E}(\overline{Y}_{22\cdot}) + \mathcal{E}(\overline{Y}_{23\cdot})]}{5} \\ &= \mu + \alpha_2 + \frac{3\beta_1 + \beta_2 + \beta_3}{5} + \frac{3\gamma_{21} + \gamma_{22} + \gamma_{23}}{5}. \end{aligned} \tag{17.5}$$

The marginal means for the A factor now involve different functions of the β_j so that they will not cancel in a contrast on the A treatment means:

$$\begin{aligned} \mathcal{E}(\overline{Y}_{1\cdot\cdot} - \overline{Y}_{2\cdot\cdot}) &= \alpha_1 - \alpha_2 + \frac{(-7\beta_1 + 6\beta_2 + \beta_3)}{20} \\ &\quad + \left[\frac{(\gamma_{11} + 2\gamma_{12} + \gamma_{13})}{4} - \frac{(3\gamma_{21} + \gamma_{22} + \gamma_{23})}{5}\right] \end{aligned} \tag{17.6}$$

Similarly, contrasts on the B treatment means will be confounded with α_i effects. Furthermore, the expectations contain different functions of the interaction effects from the balanced case. Simple contrasts on the treatment means, and sums of squares for these contrasts, no longer provide direct estimates of the appropriate functions of the parameters. Other approaches become necessary.

Empty Cells

This illustration assumed that the unequal numbers did not create any empty cells, cells with $n_{ij} = 0$. As long as there are no empty cells, all functions of the parameters that were estimable with balanced data remain estimable in the unbalanced data. However, when there are empty cells, some additional functions may become nonestimable and it may be impossible to obtain estimates of some functions of interest.

17.3 Analysis of Cell Means

Averaging

The method of **unweighted analysis of cell means** is an attempt to

avoid the effects of imbalance by replacing the unequal numbers of observations with their cell means. The method is dependent on there being no empty cells. If the imbalance arises from unequal numbers of sampling units within experimental units, the available sampling observations from each experimental unit are averaged to obtain a mean response for each experimental unit. The analysis is then conducted on these experimental unit means, as if there had been no sampling. If the imbalance arises from experimental units being lost, data from the available experimental units for each treatment are averaged and then used for the analysis of treatment effects.

Model for Observations

The analysis of cell means is described in terms of a completely random experimental design with a 2×3 factorial set of treatments. Let n_{ij} be the number of experimental units receiving the ijth treatment combination. The effects model for the individual observations is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \tag{17.7}$$

where α_i $(i = 1, \ldots, a)$ is the effect of the ith level of factor A, β_j $(j = 1, \ldots, b)$ is the effect of the jth level of factor B, and γ_{ij} is the interaction effect between the ith level of factor A and the jth level of factor B. The subscript k designates the observation receiving the ijth treatment combination $(k = 1, \ldots, n_{ij})$. The usual least squares assumptions apply to ϵ_{ijk}. The data are unbalanced if the n_{ij} are not equal.

Model for Cell Means

The cell means are obtained by averaging over the n_{ij} observations receiving the ijth treatment,

$$\overline{Y}_{ij.} = \frac{1}{n_{ij}} \left(\sum_{k=1}^{n_{ij}} Y_{ijk} \right). \tag{17.8}$$

The model in terms of these cell means is

$$\overline{Y}_{ij.} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \overline{\epsilon}_{ij.}. \tag{17.9}$$

If the variance–covariance matrix of the ϵ_{ijk} in the original data is $\mathbf{Var}(\epsilon) = \boldsymbol{I}\sigma^2$, the variance–covariance matrix for the $\overline{\epsilon}_{ij.}$ in the cell means model will be

$$\mathbf{Var}(\overline{\boldsymbol{\epsilon}}) = \begin{bmatrix} 1/n_{11} & 0 & \cdots & 0 \\ 0 & 1/n_{12} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/n_{ab} \end{bmatrix}. \tag{17.10}$$

The unweighted analysis of cell means ignores these unequal variances and proceeds as if $\mathbf{Var}(\overline{\epsilon}_{ij.}) = \boldsymbol{I}\sigma^2$.

Expectations of Means

The expectations of the cell means, given by the first four terms in the model, equation 17.9, and the expectations of the marginal means, obtained

TABLE 17.2. *Degrees of freedom and mean square expectations for the unweighted analysis of cell means for an $A \times B$ factorial with n_{ij} observations per treatment in a completely random design; all $n_{ij} > 0$.*

Source	*d.f.*	$\mathcal{E}$(*Mean Square*)[a]
Total	$ab - 1$	
A	$a - 1$	$\sigma^2 + \overline{n}_h\theta^2_\gamma + b\overline{n}_h\theta^2_\alpha$
B	$b - 1$	$\sigma^2 + \overline{n}_h\theta^2_\gamma + a\overline{n}_h\theta^2_\beta$
$A \times B$	$(a-1)(b-1)$	$\sigma^2 + \overline{n}_h\theta^2_\gamma$
Exp. error	$n_{\cdot\cdot} - ab$	σ^2

[a]The θ^2 terms are quadratic forms of the fixed effects indicated by the subscript.

by unweighted averaging of the cell means, have the same composition of all fixed effects as with balanced data (Table 17.1).

Analysis of Variance

The analysis of variance of the $a \times b$ table of cell means, with each sum of squares multiplied by $\overline{n}_h$, the harmonic mean of the numbers of observations per cell, gives the SS(A), SS(B), and SS(AB). The harmonic mean is

$$\overline{n}_h = \frac{ab}{\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{1}{n_{ij}}} \tag{17.11}$$

which simplifies to n when all $n_{ij} = n$. The mean squares estimate the same functions of the fixed effects as the corresponding analysis with balanced data except the coefficient n is replaced with $\overline{n}_h$ (Table 17.2).

Variances

The estimate of σ^2 is obtained from a separate computation of the variances among experimental units within treatments and pooled over the ab treatments. Thus,

$$\text{MS(Error)} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\left[\sum_{k=1}^{n_{ij}}(Y_{ijk} - \overline{Y}_{ij.})^2\right]}{\nu}, \tag{17.12}$$

where

$$\nu = \sum_{i=1}^{a}\sum_{j=1}^{b}(n_{ij} - 1) = n_{\cdot\cdot} - ab \tag{17.13}$$

is the degrees of freedom.

The variance of the ijth treatment mean is σ^2/n_{ij} as shown in equation 17.10. The variance of a marginal treatment mean, computed as the unweighted average of cell means, is σ^2/k, where the divisor k is the product of the number of cell means in the average and the harmonic mean of the n_{ij} *for those cells*. The variance of the difference between two unweighted marginal treatment means is the sum of the variances of the

two means. Consider for example a 2×3 factorial experiment with n_{ij} given in equation 17.3. Consider the unweighted averages of cell means $(\overline{Y}_{11.} + \overline{Y}_{12.} + \overline{Y}_{13.})/3$ and $(\overline{Y}_{21.} + \overline{Y}_{22.} + \overline{Y}_{23.})/3$ for the two levels of factor A. These two unweighted averages of cell means have variances $\sigma^2\left(\frac{1}{1} + \frac{1}{2} + \frac{1}{1}\right)/9$ and $\sigma^2\left(\frac{1}{3} + \frac{1}{1} + \frac{1}{1}\right)/9$, respectively. Also, the variance of the difference between these two means is given by the sum of the variances of the two means.

The analysis of cell means will avoid the confounding of effects associated with imbalance only in those cases where the averaging is over observations that have the same expectation. Or, equivalently, the averaging must be over observations that differ only in random elements. Averaging over unequal numbers of sampling units always provides unbiased estimates of treatment comparisons. Averaging over experimental units to obtain cell means, however, requires care to avoid confounding fixed effects in the final analysis. If the experimental design is a completely random design or if the experimental design is a randomized complete block design with *random* block effects, the analysis of cell means will yield unbiased comparisons of treatment effects. However, some of the efficiency of blocking will be lost because variances of treatment comparisons will involve the component of variance due to random block effects. If the block effects are *fixed* effects, treatment comparisons based on unweighted means will be confounded with block effects.

Inefficiency

Although the unweighted analysis of cell means is simple, it is not an efficient analysis since unequal variances (of the cell means) are being ignored. Furthermore, the sums of squares that are generated are not distributed as chi-squared random variables and, hence, the conventional tests of significance are only approximate. With the computing facilities generally available, the simplicity of the unweighted analysis of cell means does not justify its use (Speed, Hocking, and Hackney, 1978).

Weighted Analysis

The **weighted analysis of cell means** uses weighted least squares to take into account the unequal variances of the cell means. The relative sizes of the variances of the cell means are determined by $1/n_{ij}$, equation 17.10, so that the appropriate weighting matrix is a diagonal matrix of the n_{ij}. Note that if we consider the transformed model

$$n_{ij}^{1/2}\overline{Y}_{ij.} = n_{ij}^{1/2}\mu + n_{ij}^{1/2}\alpha_i + n_{ij}^{1/2}\beta_j + n_{ij}^{1/2}\gamma_{ij} + n_{ij}^{1/2}\overline{\epsilon}_{ij.},$$

then the errors $n_{ij}^{1/2}\overline{\epsilon}_{ij.}$ have equal variances. The least squares estimates from the transformed model of estimable functions of the parameters are best linear unbiased estimates. The sums of squares obtained correspond to those obtained from the general linear models analysis of the original observations discussed in the next section.

17.4 Linear Models for Unbalanced Data

Least squares regression with linear models containing class variables reproduces the analyses of variance for the standard experimental designs when the data are balanced (Chapter 9). The general linear models approach, however, does not require balanced data. As long as the parametric functions of interest remain estimable, the general linear models approach will provide estimates of the functions and sums of squares for tests of significance of any testable hypotheses. This section discusses the use of least squares regression with class variables for the analysis of unbalanced data.

General Procedure

The general procedure is as discussed in Chapter 9. To review briefly, a linear model is constructed using dummy variables in $\boldsymbol{X}$ to bring in the effects of class variables, such as treatments. Each set of dummy variables introduces at least one linear dependency among the columns of $\boldsymbol{X}$ so that the model is not of full rank and the unique inverse does not exist. The general linear models approach uses a generalized inverse of $\boldsymbol{X}'\boldsymbol{X}$ to obtain one of the nonunique solutions to the normal equations,

$$\boldsymbol{\beta}^0 = (\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{Y}, \tag{17.14}$$

where $(\boldsymbol{X}'\boldsymbol{X})^-$ is a generalized inverse of $\boldsymbol{X}'\boldsymbol{X}$. Even though $\boldsymbol{\beta}^0$ is not unique, it can be used to obtain a unique estimate of any estimable function of the parameters and a unique sum of squares for any testable hypothesis. That is, if $\boldsymbol{K}'\boldsymbol{\beta}$ is an estimable function of $\boldsymbol{\beta}$, it is uniquely estimated with $\boldsymbol{K}'\boldsymbol{\beta}^0$, where $\boldsymbol{\beta}^0$ is one of the nonunique solutions. Furthermore, if $\boldsymbol{K}'\boldsymbol{\beta}$ is estimable and $\boldsymbol{K}'$ is of full row rank, then $\boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{0}$ is a testable hypothesis for which the unique sum of squares is

$$Q = (\boldsymbol{K}'\boldsymbol{\beta}^0)'[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{K}]^{-1}(\boldsymbol{K}'\boldsymbol{\beta}^0) \tag{17.15}$$

with $r(\boldsymbol{K}')$ degrees of freedom.

Estimable Functions

The specific linear functions of parameters that are estimable play a dominant role in the the analysis of models of less than full rank. This was indicated in the discussion of the analysis of balanced data (Chapter 9), but the specific form of the estimable functions was not critical to that discussion and was not pursued at that time. In the analysis of unbalanced data, however, the form of the estimable functions *defines* different types of sums of squares that might be computed and serves as a convenient vehicle for describing these differences. First, and for background, the general form of the estimable functions and the specific forms that generate the sums of squares in the analysis of variance of *balanced data* are presented. Then, the estimable functions that generate the sums of squares for two classes of hypotheses with unbalanced data are discussed. The two classes of hypotheses with which we are concerned are labeled Types I and III in the general linear models program PROC GLM (SAS Institute Inc., 1989b). Type I hypotheses and their sums of squares are generated by sequentially

testing model effects as they are added to the model. These correspond to what we have labeled as the **sequential** hypotheses and sums of squares. The Type III hypotheses and their sums of squares generated by PROC GLM are one of many possible types of hypotheses one could generate where effects of interest have been *adjusted* (according to specific rules) for other effects in the model. These correspond to what we have labeled as the **partial** hypotheses and sums of squares. Other types of hypotheses are discussed by Speed, Hocking, and Hackney (1978, Table 7) for the two-way classified model.

17.4.1 Estimable Functions with Balanced Data

General Form

A general form $\boldsymbol{L}'\boldsymbol{\beta}$ that encompasses all linear estimable functions can be obtained from the $\boldsymbol{X}$ matrix. The coefficients in each row of $\boldsymbol{X}$ define an estimable function of $\boldsymbol{\beta}$. This follows from the fact that each observation in $\boldsymbol{Y}$ is an unbiased estimate of the particular function of $\boldsymbol{\beta}$ defined by the corresponding row of $\boldsymbol{X}$. That is, $\mathcal{E}(Y_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$, where $\boldsymbol{x}_i'$ is the ith row of $\boldsymbol{X}$. It also follows that any linear function of the rows of $\boldsymbol{X}$ also defines an estimable function of $\boldsymbol{\beta}$.

This principle is used to generate, by row operations on $\boldsymbol{X}$, a general form that encompasses all estimable functions for a given model and set of data. Only the unique rows of $\boldsymbol{X}$ need to be considered. That is, no new estimable function is generated by an additional observation that has the same expectation (identical values of $\boldsymbol{X}$) as a previously considered observation. (A corollary of this statement is that imbalance in data does not change the set of estimable functions as long as none of the unique rows of $\boldsymbol{X}$ has been lost. This requires that there be at least one observation in every cell.)

Illustration with CRD

Derivation of the general form of the estimable functions is illustrated for the completely random experimental design with $t = 4$ treatments. The general linear model is

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \qquad (i = 1, \ldots, 4;\ j = 1, \ldots, n_i)$$

from which the unique rows of $\boldsymbol{X}$ are

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The linear functions of the parameters defined by $\boldsymbol{A}\boldsymbol{\beta}$ are estimable. To obtain the general form of estimable functions as given by PROC GLM,

row operations on $\boldsymbol{A}$ are used to reduce it to a simpler form given by

$$\boldsymbol{A}^* = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}.$$

The row operations on $\boldsymbol{A}$ are linear operators so that all linear functions defined by $\boldsymbol{A}^*$ are also estimable. The first row of $\boldsymbol{A}^*$ says that $(\mu + \tau_4)$ is estimable, the second row says that $(\tau_1 - \tau_4)$ is estimable, and so forth.

Furthermore, any arbitrary linear function of these estimable functions will be estimable. Let the arbitrary linear function be defined by the coefficients

$$\boldsymbol{C}' = (\, C_1 \quad C_2 \quad C_3 \quad C_4 \,).$$

Thus, the general form that encompasses all estimable functions for this example is

$$\boldsymbol{C}'\boldsymbol{A}^*\boldsymbol{\beta} = C_1\mu + C_2\tau_1 + C_3\tau_2 + C_4\tau_3 + (C_1 - C_2 - C_3 - C_4)\tau_4$$

or, letting $\boldsymbol{L}' = \boldsymbol{C}'\boldsymbol{A}^*$,

$$\begin{aligned} \boldsymbol{L}' &= [\, L_1 \quad L_2 \quad L_3 \quad L_4 \quad L_5 \,] \\ &= [\, C_1 \quad C_2 \quad C_3 \quad C_4 \quad (C_1 - C_2 - C_3 - C_4) \,]. \end{aligned} \tag{17.16}$$

Notice the fifth element L_5 of $\boldsymbol{L}$, the coefficient of τ_4, is a linear function of other L_j. This reflects the over-parameterization of the model.

Any choice of values for the L_j yields an estimable function of the parameters as long as L_5 satisfies the relationship in equation 17.16. For example, setting $L_1 = 1$, $L_2 = 1$, and all others equal to zero gives $(\mu + \tau_1)$, which is the expectation of the mean of the first treatment. Setting $L_1 = 1$ and $L_2 = L_3 = L_4 = \frac{1}{4}$ (and $L_5 = \frac{1}{4}$) shows that $(\mu + \overline{\tau}_\cdot)$ is estimable.

Estimable Contrasts for CRD

To obtain an estimable contrast on the treatment effects, L_1 must be set to zero to avoid having μ involved. There are three remaining "free" coefficients in $\boldsymbol{L}$ involving only the τ_i so that there are a maximum of three linearly independent estimable functions of the τ_i. (This is why three degrees of freedom are assigned to the treatment sum of squares.) Setting $L_2 = 1$ and $L_3 = L_4 = 0$ (hence $L_5 = -1$) gives $(\tau_1 - \tau_4)$ as one estimable contrast. Similarly, setting $L_3 = 1$ and $L_2 = L_4 = 0$ (hence $L_5 = -1$) gives $(\tau_2 - \tau_4)$ and setting $L_4 = 1$ and $L_2 = L_3 = 0$ (and hence $L_5 = -1$) gives $(\tau_3 - \tau_4)$. If these three choices of the L_j are combined in one matrix,

$$\boldsymbol{K}' = \begin{bmatrix} 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}, \tag{17.17}$$

then $\boldsymbol{K}'\boldsymbol{\beta}$ is a set of linearly independent *estimable* functions (contrasts) involving the τ_i. The composite hypothesis that all τ_i are equal, or that

TABLE 17.3. *The general form for estimable functions in a* 2×3 *factorial (with no empty cells) and choices of* L_k *that give the conventional analysis of variance results with balanced data.*

Parameters	*Coefficients for General Form*[a]	*Specific Estimable Functions*				
		for αs	*for* βs		*for* γs	
			(1)	(2)	(1)	(2)
μ	L_1	0	0	0	0	0
α_1	L_2	1	0	0	0	0
α_2	$L_3 = L_1 - L_2$	-1	0	0	0	0
β_1	L_4	0	1	0	0	0
β_2	L_5	0	0	1	0	0
β_3	$L_6 = L_1 - L_4 - L_5$	0	-1	-1	0	0
γ_{11}	L_7	$\frac{1}{3}$	$\frac{1}{2}$	0	1	0
γ_{12}	L_8	$\frac{1}{3}$	0	$\frac{1}{2}$	0	1
γ_{13}	$L_9 = L_2 - L_7 - L_8$	$\frac{1}{3}$	$-\frac{1}{2}$	$-\frac{1}{2}$	-1	-1
γ_{21}	$L_{10} = L_4 - L_7$	$-\frac{1}{3}$	$\frac{1}{2}$	0	-1	0
γ_{22}	$L_{11} = L_5 - L_8$	$-\frac{1}{3}$	0	$\frac{1}{2}$	0	-1
γ_{23}	$L_{12} = L_1 - L_2 - L_4 - L_5 + L_7 + L_8$	$-\frac{1}{3}$	$-\frac{1}{2}$	$-\frac{1}{2}$	1	1

[a]The subscripts on the L coefficients correspond to the sequence of the parameters. The coefficients, L_3, L_6, L_9, L_{10}, L_{11}, and L_{12} are constrained by the design and the model as shown.

there are no differences among the treatments, can be written as H_0 : $\boldsymbol{K}'\boldsymbol{\beta} = \mathbf{0}$. This is a testable hypothesis since each row vector in $\boldsymbol{K}'$ defines an estimable function of $\boldsymbol{\beta}$.

General Form for 2×3 Factorial

Return now to the 2×3 factorial in a completely random experimental design, which was used to illustrate the effects of imbalance (Section 17.2). The general form for all estimable functions for the 2×3 factorial with interaction is given in the second column of Table 17.3. The last five columns give the specific estimable functions that generate the sums of squares for the conventional analysis of variance with *balanced data*.

SS($\boldsymbol{A}$)

The estimable function (contrast) of the αs that generates SS($\boldsymbol{A}$) (column 3 of Table 17.3) is obtained by setting L_1 equal to zero to remove μ from the contrast, the remaining free coefficient on the α_i, L_2, equal to unity, and L_4 and L_5 equal to zero to remove the β_j effects from the contrast. This leaves L_7 and L_8 to be determined. When the data are balanced, comparisons on the marginal means for the A factor involve the same comparisons on the row averages of the γ_{ij} effects. That result is obtained by

setting $L_7 = L_8 = \frac{1}{3}L_2$ (see Table 17.1). The divisor of 3 comes from the number of levels of the B factor being averaged across.

SS(B)

Two linearly independent contrasts, two degrees of freedom, are required to generate SS(B), the variation due to the β_j. This is evident in the general form by the two "free" coefficients L_4 and L_5 associated with the β_j. There are several ways contrasts can be defined whenever more than one degree of freedom is involved. It is only necessary that the contrasts be linearly independent. The contrasts on β_j require that $L_1 = L_2 = 0$ to avoid confounding the contrast with μ and α_i. The first contrast in Table 17.3 sets $L_4 = 1$ and $L_5 = 0$; the second contrast is the converse where $L_4 = 0$ and $L_5 = 1$. The L_7 and L_8 coefficients are chosen in each case so as to give the same contrast on the column averages of the interaction effects. The sum of squares due to the composite hypothesis that both contrasts are zero is SS(B) in the analysis of variance of balanced data.

SS(AB)

Finally, contrasts for the γ_{ij} effects require that all L_k except L_7 and L_8 be zero to avoid confounding the interaction contrasts with μ, α_i, and β_j. This leaves two free coefficients L_7 and L_8, and hence two linearly independent contrasts to be defined. The first contrast in Table 17.3 uses $L_7 = 1$ and $L_8 = 0$; the second uses the converse $L_7 = 0$ and $L_8 = 1$. The sum of squares due to the composite hypothesis that both contrasts are zero is SS(AB) in the analysis of variance of balanced data.

Properties for Balanced Data

This illustrates the general nature of the estimable functions or the testable hypotheses that generate the sums of squares in the conventional analyses of variance for balanced data (Table 17.3). These linear functions define the hypotheses being tested *with balanced data* and they provide a guide for the kinds of hypotheses that might be considered in the analysis of unbalanced data. They possess the following properties that can be used to define various types of hypotheses, and their sums of squares, for unbalanced data.

Property 1: No estimable function for generating a main effect sum of squares, such as the contrast on α_i or the contrasts on the β_j, involves main effects of the other factor. Each does, however, contain a contrast on higher order interaction effects involving the same factor. This illustrates the more general result:

> Estimable functions for the sum of squares for any one class of effects, main effects or interaction effects, will not involve any other class of effects except those that are higher-order interaction effects or higher-level nested effects of the same factor.

For example, the estimable functions for the $A \times B$ interaction sum of squares in a *three*-factor factorial will have zero coefficients on all main effects and the $A \times C$ and $B \times C$ interaction effects. They will have nonzero coefficients on

the $A \times B \times C$ interaction effects since this is a higher-level interaction effect involving $A \times B$. The $A \times B \times C$ interaction effect is said to "contain" (in notation) the $A \times B$ interaction effect. Thus, estimable functions for any class of effects will have zero coefficients on all other classes of effects that do not *contain* the effects being contrasted.

Property 2: An estimable function for the sum of squares for one class of main effects includes the same contrast on averages of the corresponding interaction effects. In effect, the coefficient on each main effect is divided and *equitably distributed* over the interaction effects associated with the same cells as the main effect. For example, the "-1" coefficient on α_2 in the first contrast (Table 17.3) is distributed equally over the three interaction effects γ_{21}, γ_{22}, and γ_{23}, with a coefficient of $-\frac{1}{3}$ on each. In multifactor experiments, this property of "equitable distribution" of coefficients extends to all higher-order interaction effects that *contain* the class of effects on which the estimable function is being constructed. This is referred to as the **equitable distribution property** of the coefficients and is always obtained in balanced data.

Property 3: The estimable function for the sum of squares for the α_i effects is *orthogonal* to both estimable functions constructed for the sum of squares for the γ_{ij} effects. Similarly, the two estimable functions constructed for the sum of squares for the β_j effects are pairwise orthogonal to the two estimable functions constructed for the γ_{ij} effects. [The sum of products of the coefficients in any one of columns 3, 4, or 5 with the coefficients in either one of columns 6 or 7 is zero (Table 17.3).] This is referred to as the **orthogonality property** and is always obtained in balanced data. More generally, the orthogonality property states that:

> The estimable functions, or the testable hypotheses, constructed for the sum of squares for any class of effects are pairwise orthogonal to the estimable functions constructed for the sum of squares for any class of effects that contain them.

17.4.2 Estimable Functions with Unbalanced Data

Effect of Imbalance on Estimable Functions

Imbalance in the data does not change the general form of estimable functions as long as all cells of the table have at least one observation. When there are empty cells, the general form of estimable functions will change,

and some additional linear functions will become nonestimable if the missing data have caused the loss of one or more of the unique rows of $\boldsymbol{X}$. Even if the general form of estimable functions has not changed, imbalance does change the functions being estimated by the standard analysis of variance sums of squares, and there are different methods of adjusting for the confounding of effects that results. These different methods of adjusting are equivalent to imposing different conditions on the choice of coefficients in the general form of estimable functions.

PROC GLM (SAS Institute Inc., 1989b) is programed to compute four types of sums of squares for unbalanced data, all of which might be considered logical extensions in one way or another of the analysis of variance for balanced data to the unbalanced case. In all cases, the sums of squares are conveniently described in terms of the testable hypotheses they represent. Type I and Type II sums of squares can be described solely in terms of the other effects in the model for which the sum of squares has been adjusted. (If a sum of squares has been adjusted for a particular class of effects, the testable hypotheses for that sum of squares have zero coefficients on that class of effects.) For both the Type I and Type II sums of squares, no control is exercised over the coefficients on classes of effects for which the sum of squares has *not* been adjusted. The Type III and Type IV sums of squares differ from Type I and Type II in that regard; constraints are imposed on the coefficients of the classes of effects for which the sum of squares has *not* been adjusted. Constraints are imposed so that the underlying hypotheses possess the orthogonality property Type III, the equitable distribution property Type IV, or both.

Other analysis programs compute various ones of these four types or variations of these. The reader is referred to Speed, Hocking, and Hackney (1978) for a summary of the hypotheses being tested by the sums of squares from various programs. Speed, Hocking, and Hackney specify their hypotheses in terms of the full-rank means model, but there is an equivalence to the classical effects model (Speed and Hocking, 1976). In this text we are concerned only with the Type I (sequential) and Type III (partial) hypotheses and sums of squares.

Sequential sums of squares: Type I

The Type I sums of squares are the classical sequential sums of squares obtained from adding the terms to the model in some logical sequence. The sum of squares for each class of effects is adjusted for only those effects that precede it in the model. Thus, the sums of squares and their expectations are dependent on the order in which the model is specified. Using the 2×3 factorial for illustration, adding the terms to the model in the order A, B, AB would generate Type I sums of squares described with the R-notation as

$$\mathrm{SS}(A) \quad = \quad R(\boldsymbol{\alpha}|\mu)$$

$$\begin{aligned} \text{SS(B)} &= R(\boldsymbol{\beta}|\boldsymbol{\alpha}\,\mu) \\ \text{SS}(AB) &= R(\boldsymbol{\gamma}|\boldsymbol{\alpha}\,\boldsymbol{\beta}\,\mu). \end{aligned}$$

SS(A)

The sum of squares for the α effects SS(A) has been adjusted only for μ. It is computed as the (corrected) sum of squares among the A treatment totals giving no consideration to the β_j and γ_{ij} effects. The estimable function that generates this Type I sum of squares is obtained from the general form, Table 17.3, by setting $L_2 = 1$, to give a contrast on α_1 and α_2, and $L_1 = 0$, to remove the effect of μ. All other coefficients in the general estimable form take the values that result from computing the minimum variance unbiased estimate of this contrast on the α_i adjusted for μ. These coefficients will be functions of the n_{ij}, the numbers of observations in the cells. SS(A) will almost certainly be confounded with β_j effects in unbalanced data. It is often referred to as the sum of squares for A *ignoring* B.

SS(B)

The Type I sum of squares for the β effects SS(B) is adjusted for both μ and the α_i effects, since these effects precede B in the model statement. It is computed as the sum of squares for differences among the levels of the B factor but further adjusted to remove any α_i effects. The presence of the γ_{ij} effects is ignored. The two estimable functions that generate this sum of squares have $L_1 = L_2 = 0$, to remove μ and the α_i, and L_4 and L_5 chosen to specify two contrasts on the β_j as in Table 17.3. The free coefficients on the γ_{ij}, L_7 and L_8, however, take whatever values the minimum variance unbiased estimators of the two β contrasts happen to have and, again, are functions of the numbers of observations. Thus, the Type I SS(B) is not confounded with the α_i effects but the function of the γ_{ij} effects contained in the contrasts is not as shown in the balanced data example of Table 17.3.

SS(AB)

The Type I sum of squares for interaction SS(AB) is adjusted for all other effects in the model since it occurs last in the model statement. The estimable functions that generate this sum of squares are the same as those shown in Table 17.3 for balanced data.

Uses of Type I SS

Because of the sequential manner in which the Type I sums of squares are adjusted, they are not appropriate for many hypotheses used in analysis of variance problems. They are appropriate sums of squares for testing hypotheses when there is some logic in the particular sequence of adjustments such as, for example, the contributions of successively higher degree terms in a polynomial model or the sequential terms in a purely nested model. Sums of Type I sums of squares are useful for testing composite hypotheses of several class effects if appropriately ordered in the model. In general, however, the Type I sums of squares should be used with caution.

Partial Sums of Squares: Type III

The Type III sums of squares is a partial sums of squares in that each is adjusted for all other classes of effects in the model according to two general rules. First, the estimable functions that generate the sum of squares for

one class of effects will not involve any other classes of effects except those that "contain" the class of effects in question. This is the first general property noted in Section 17.4.1 on the nature of estimable functions in balanced data. Thus, Type III sums of squares are defined so as to test hypotheses that contain the same classes of effects as the corresponding hypotheses in balanced data. For example, the estimable functions that generate SS(AB) in a *three*-factor factorial will have zero coefficients on all main effects and the $A \times C$ and $B \times C$ interaction effects. They will contain nonzero coefficients on the $A \times B \times C$ interaction effects, since the $A \times B \times C$ interaction "contains" the $A \times B$ interaction.

Secondly, the Type III sums of squares require the coefficients on the higher-order interaction or nested effects that contain the effects in question to satisfy the **orthogonality property**. The coefficients on these effects are no longer functions of the n_{ij} and, consequently, are the same for all designs with the same general form of estimable functions. If there are no empty cells, no $n_{ij} = 0$, the Type III sums of squares also satisfy the **equitable distribution** property and the hypotheses being tested are the same as when the data are balanced.

Which Type is Appropriate

When data are balanced, the four types of sums of squares computed by PROC GLM are the same and identical to the conventional analysis of variance for the particular design. When the data are unbalanced, the four types of sums of squares and the hypotheses being tested may differ. Decisions as to which are the appropriate sums of squares to use should be based on which sums of squares test the most meaningful hypotheses. A Type I sum of squares, being a sequential sum of squares adjusted only for effects that precede it in the model, is usually not appropriate for the classical analysis of variance hypotheses. They are appropriate in special cases as already noted.

The Type III sums of squares are adjusted so that the classes of effects involved (those that have nonzero coefficients) in each sum of squares are the same as in the sums of squares for balanced data. [This is also true for the Type II and Type IV sums of squares computed by PROC GLM but their adjustment for the higher-order interaction effects that contain the effects in question is either not done (Type II) or done so as to satisfy the equitable distribution property (Type IV). We consider the Type II and IV sums of squares to have limited usefulness and do not discuss them in this text.] The Type III sums of squares adjust the nonzero coefficients on the higher-order effects to satisfy the orthogonality property that is present when data are balanced. The hypotheses being tested by the Type III sums of squares are no longer dependent on the particular n_{ij} as they are for the Type II sums of squares and would appear to be the more appropriate for testing the usual hypotheses associated with analysis of variance problems. [The reader is referred to Freund, Littell, and Spector (1986), and *SAS/STAT User's Guide* (SAS Institute Inc., 1989a) for more discussion of the four types of sums of squares.)

Unbalanced Data: An Example

Example 17.1

The differences in the Type I and Type III sums of squares and their estimable functions are illustrated using a specific unbalanced case of the 2×3 factorial. The example is taken from Searle and Henderson (1979), and is used with their permission. The data and numbers of observations per cell are as follows.

Data:

		Factor B		
		1	2	3
Factor A	1	2, 4, 6	4, 6	5
	2	12, 8	11, 7	—

$\boldsymbol{n_{ij}}$**:**

		Factor B			
		1	2	3	$n_{i.}$
Factor A	1	3	2	1	6
	2	2	2	0	4
	$n_{.j}$	5	4	1	$n_{..} = 10$

The data contain one missing cell: $n_{23} = 0$. The numbers of observations for the other cells vary from $n_{13} = 1$ to $n_{11} = 3$. The model is the same as used earlier, equation 17.7,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

where α_i and β_j are the main effects and the γ_{ij} the interaction effects. The difference is that the $(i, j) = (2, 3)$ combination does not occur since that cell is empty. The sequential (Type I) and partial (Type III) sums of squares (computed using PROC GLM with the model specified in the order A, B, AB) are given in Table 17.4.

General Form

The general form of the estimable functions for this set of data differs from that for the balanced 2×3 factorial, Table 17.3, only because of the empty cell. The general form for the estimable functions is obtained by row operations on the unique rows of $\boldsymbol{X}$. The absence of an observation in cell (2, 3) caused the loss of the row of $\boldsymbol{X}$ containing γ_{23} and, consequently, must affect the estimable functions. The general coefficients for the α_i and β_j effects remain as shown in Table 17.3; the general coefficients on the interaction effects change to the following:

$$\begin{aligned}
\gamma_{11}: &\quad L_7 \\
\gamma_{12}: &\quad -L_1 + L_2 + L_4 + L_5 - L_7 \\
\gamma_{13}: &\quad L_1 - L_4 - L_5 \\
\gamma_{21}: &\quad L_4 - L_7 \\
\gamma_{22}: &\quad L_1 - L_2 - L_4 + L_7.
\end{aligned} \tag{17.18}$$

TABLE 17.4. *Analysis of data for an unbalanced* 2×3 *factorial with one empty cell. (From Searle and Henderson, BU-641-M, May* 1979*. Used with permission.)*

Source	*d.f.*	*Sum of Squares*	*Mean Square*
Model	4	62.5	15.625
Error	5	26.0	5.200
Total	9	88.5	

		Sum of Squares	
Source	*d.f.*	*Type I*	*Type III*
A	1	60.00	54.55
B	2	0.32	0.21
$A \times B$	1	2.18	2.18

The absence of γ_{23} in this list should be interpreted as the coefficient on γ_{23} always being zero. Note that the linear function for γ_{13} is the same as that for β_3 (Table 17.3) which implies that these two parameters always have the same coefficient in any estimable function of the parameters. Thus, no estimable function can separate β_3 and γ_{13}.

Estimable Functions for SS

The differences between the Type I and Type III sums of squares are illustrated by the estimable function(s) being considered in each case (Table 17.5). The estimable functions for the A sum of squares show (1) that the Type I sum of squares involves β_j effects whereas the Type III sum of squares involves only contrasts on the α_i and γ_{ij}, and (2) the coefficients on the γ_{ij} for the Type I sum of squares are functions of the n_{ij} whereas those for Type III are not.

SS(A)

The Type I SS(A) is inappropriate for testing hypotheses about α_i; it is confounded with the β_j effects. The Type III sum of squares for A is based on an estimable function similar in form to that in the balanced case. It differs from the balanced case in that there is no information on γ_{23}.

The estimable functions for SS(B) sums of squares are shown in the middle portion of Table 17.5. There are two degrees of freedom for SS(B), there are two "free" coefficients in the general form, so that two linear contrasts are required. The Type I sum of squares for B does not involve α_i effects, whereas the Type I sum of squares for A does involve β_j effects. This results from B occurring after A in the model and reflects the sequential nature of the Type I sums of squares. The Type I and Type III sums of squares still differ in their coefficients on the γ_{ij} effects with those for Type I being functions of the n_{ij}.

SS(AB)

Only one estimable function exists for SS(AB) and it is the same for both Type I and Type III. The contrast is shown in the lower portion of Table 17.5. This contrast involves only the effects in the 2×2 part of the table that does not involve the missing cell. The orthogonality criterion of the Type III sums of squares can be verified by computing the sum of

TABLE 17.5. *Estimable functions for the Type I and Type III sums of squares from the* 2×3 *factorial with cell* $(2, 3)$ *missing.*

Type	*Parameter*											
SS	μ	α_1	α_2	β_1	β_2	β_3	γ_{11}	γ_{12}	γ_{13}	γ_{21}	γ_{22}	γ_{23}
SS(*A*)												
I	0	1	−1	0	$-\frac{1}{6}$	$\frac{1}{6}$	$\frac{3}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$-\frac{1}{2}$	$-\frac{1}{2}$	0
III	0	1	−1	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$-\frac{1}{2}$	0
SS(*B*)												
I	0	0	0	1	0	−1	$\frac{9}{11}$	$\frac{2}{11}$	−1	$\frac{2}{11}$	$-\frac{2}{11}$	0
	0	0	0	0	1	−1	$\frac{3}{11}$	$\frac{8}{11}$	−1	$-\frac{3}{11}$	$\frac{3}{11}$	0
III	0	0	0	1	0	−1	$\frac{3}{4}$	$\frac{1}{4}$	−1	$\frac{1}{4}$	$-\frac{1}{4}$	0
	0	0	0	0	1	−1	$\frac{1}{4}$	$\frac{3}{4}$	−1	$-\frac{1}{4}$	$\frac{1}{4}$	0
SS(*AB*)												
I & *III*	0	0	0	0	0	0	1	−1	0	−1	1	0

products of the coefficients for the A and B Type III contrasts with the $A \times B$ Type III contrast (Table 17.5).

This discussion and Example 17.1 have centered on the factorial model. Models with nested effects or both nested and cross-classified effects follow much the same rules. The general form of the estimable functions for any specific case can be determined from the unique rows of the $\boldsymbol{X}$ matrix before reparameterization [see *SAS User's Guide*, SAS Institute Inc., 1989a] and can be requested as the E option in the model statement in PROC GLM (SAS Institute Inc., 1989b).

17.4.3 Least Squares Means

Definition

The marginal means in an unbalanced set of data do not in general provide meaningful comparisons. The least squares solution to the normal equations, however, can be used to obtain estimates of the *same* linear functions of effects as provided by the corresponding means in balanced data *if* these functions are estimable. These estimates can be thought of as adjusted means, adjusted to remove the unwanted confounding effects. They are called the **least squares means** and are designated with "LS" in front of the usual mean notation.

The particular linear functions of $\boldsymbol{\beta}$ that must be estimated to obtain the least squares means are defined by the expectations of the corresponding means for balanced data. These expectations are called **population marginal means** (Searle, Speed, and Milliken, 1980). The population

marginal means are obtained by averaging the fixed effects in the model in the manner specified by the particular mean being considered. Thus, the expectation of the mean is completely defined by the subscript–dot notation used to define the mean. The rules for writing the expectation for a particular mean when the data are balanced are given in the box.

Rules for Obtaining Population Marginal Means

1. Specify the desired mean using the dot notation.
2. Include in its expectation a term for each class of fixed effects in the model. Drop all random effects. (See Chapter 18.)
3. On each fixed effects term, replace each subscript in the model with the specific number or dot consistent with the notation for the particular mean of interest.
4. Any fixed effect that contains a dot in its subscript is an average of effects as indicated by the dot notation. Place a "bar" over the effect to denote a mean.
5. Any covariable (continuous variable) in the expectation is replaced with its mean value.

Illustration with 2 × 3 Factorial

To illustrate, consider the expectation of the marginal means for an $A \times B$ balanced factorial with interaction effects in the model. The model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}.$$

Assume there are two levels of factor A and three levels of factor B. To obtain $\mathcal{E}(\overline{Y}_{1..})$, drop the random effects term ϵ_{ijk}, replace the subscript i on α_i and γ_{ij} with 1 and the subscript j on β_j and γ_{ij} with ".", and place a bar over all terms with a dot. Thus,

$$\mathcal{E}(\overline{Y}_{1..}) = \mu + \alpha_1 + \overline{\beta}_{.} + \overline{\gamma}_{1.} = \mathcal{E}_B(\overline{Y}_{1..}). \tag{17.19}$$

Similarly,

$$\begin{aligned}
\mathcal{E}(\overline{Y}_{2..}) &= \mu + \alpha_2 + \overline{\beta}_{.} + \overline{\gamma}_{2.} = \mathcal{E}_B(\overline{Y}_{2..}),\\
\mathcal{E}(\overline{Y}_{.1.}) &= \mu + \overline{\alpha}_{.} + \beta_1 + \overline{\gamma}_{.1} = \mathcal{E}_B(\overline{Y}_{.1.}),\\
\mathcal{E}(\overline{Y}_{.2.}) &= \mu + \overline{\alpha}_{.} + \beta_2 + \overline{\gamma}_{.2} = \mathcal{E}_B(\overline{Y}_{.2.}), \text{ and}\\
\mathcal{E}(\overline{Y}_{.3.}) &= \mu + \overline{\alpha}_{.} + \beta_3 + \overline{\gamma}_{.3} = \mathcal{E}_B(\overline{Y}_{.3.}).
\end{aligned} \tag{17.20}$$

These parametric functions obtained from balanced data, called the marginal population means, are usually the functions of interest even in the unbalanced case. However, their estimators in the unbalanced case in general will not be the simple marginal means of the data, and when at least

one cell is empty, not all of the populational marginal means are estimable. We use the notation $\mathcal{E}_B(\overline{Y}_{i..})$ and $\mathcal{E}_B(\overline{Y}_{.j.})$ to denote the marginal population means for the balanced case.

Marginal Means

The least squares marginal treatment means for this model, $LS\overline{Y}_{i..}$ and $LS\overline{Y}_{.j.}$, are defined as the best linear unbiased estimates of the corresponding linear functions of the parameters in $\mathcal{E}_B(\overline{Y}_{i..})$ and $\mathcal{E}_B(\overline{Y}_{.j.})$, equation 17.20. All are estimable if there are no empty cells. When cell (2, 3) is empty, as in Example 17.1, there is no information on γ_{23} and, therefore, any expectation involving γ_{23} must be a nonestimable function. Thus, it is not possible in Example 17.1 to compute $LS\overline{Y}_{2..}$ and $LS\overline{Y}_{.3.}$ since the functions they are supposed to be estimating involve γ_{23}. Although the concept of estimability applies to linear functions of the parameters, for convenience the terms "estimable" and "nonestimable" are attached to the least squares means according to whether the corresponding population marginal means are estimable or nonestimable. [SAS Institute Inc. (1989b) defines the expectation to be estimated by the least squares means as the average of the expectations over *only* the cells that contain data.]

Estimability of Population Marginal Means

If a population marginal mean is estimable, its expectation can be obtained from the general form of estimable functions for that specific case with proper choice of coefficients.

Example 17.2

This is illustrated with the 2×3 factorial with cell (2, 3) empty (Example 17.1). The particular linear function of the parameters contained in the expectation of $\overline{Y}_{1..}$, equation 17.19, is obtained from the general form by setting $L_1 = L_2 = 1$ and $L_4 = L_5 = L_7 = \frac{1}{3}$. [Combine equation 17.18 with Table 17.3 to obtain the general linear form for the case with cell (2, 3) empty.] Therefore, $LS\overline{Y}_{1..}$ is an estimable least squares mean in this example. $\mathcal{E}_B(\overline{Y}_{.1.})$ is obtained by setting $L_1 = L_4 = 1$, $L_2 = L_7 = \frac{1}{2}$, and $L_5 = 0$ and, therefore, $LS\overline{Y}_{.1.}$ is estimable. On the other hand, $\mathcal{E}_B(\overline{Y}_{2..})$ cannot be obtained by any choice of coefficients, and therefore $LS\overline{Y}_{2..}$ is nonestimable. (PROC GLM informs the user when least squares means are nonestimable.) The population means for the individual cells of the table have expectations

$$\mathcal{E}(LS\overline{Y}_{ij.}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

which can be obtained from the general form for all (i, j) except $(i = 2, j = 3)$. Therefore, all $LS\overline{Y}_{ij.}$ except $LS\overline{Y}_{23.}$ are estimable. ■

Estimability Dependent on the Model

The estimability of the population marginal means for a particular set of data is dependent on the model being used. This is illustrated in the 2×3 example by noting that all marginal means become estimable if the model does not contain interaction effects γ_{ij} even though cell (2,3) is empty. The general form of estimable functions is as before but with the γ_{ij} coefficients

TABLE 17.6. *The GLM solution to the* 2×3 *factorial with cell* (2, 3) *empty and the expectations of the corresponding estimators.*

GLM Results				
Parameter		*Estimate*		*Expectation of the Estimator*
Intercept		9.0	B^a	$\mu + \alpha_2 + \beta_3 - \gamma_{12} + \gamma_{13} + \gamma_{22}$
A	1	-4.0	B	$\alpha_1 - \alpha_2 + \gamma_{12} - \gamma_{22}$
	2	0	B	0
B	1	1.0	B	$\beta_1 - \beta_3 + \gamma_{12} - \gamma_{13} + \gamma_{21} - \gamma_{22}$
	2	0	B	$\beta_2 - \beta_3 + \gamma_{12} - \gamma_{13}$
	3	0	B	0
AB	11	-2.0	B	$\gamma_{11} - \gamma_{12} - \gamma_{21} + \gamma_{22}$
	12	0	B	0
	13	0	B	0
	21	0	B	0
	22	0	B	0

[a]The "B" is part of the SAS output to remind the user that the estimators are biased for the corresponding parameter.

dropped. $\mathcal{E}_B(\overline{Y}_{.3.}) = \mu + \overline{\alpha}_{.} + \beta_3$ is estimable and is obtained from the general linear form by setting $L_1 = 1$, $L_2 = \frac{1}{2}$, and $L_4 = L_5 = 0$. Note also that the population cell mean $\mathcal{E}_B(\overline{Y}_{23.}) = \mu + \alpha_2 + \beta_3$ for the missing cell (2,3) is estimable and is obtained by setting $L_1 = 1$ and $L_2 = L_4 = L_5 = 0$.

Computation

The least squares means are computed as linear functions of one of the nonunique solutions $\boldsymbol{\beta}^0$ to the normal equations. The least squares estimate $\boldsymbol{\beta}^0$ is biased, $\mathcal{E}(\boldsymbol{\beta}^0) \neq \boldsymbol{\beta}$, since $\boldsymbol{X}$ is not of full rank. However, the best linear unbiased estimate of any *estimable function* of $\boldsymbol{\beta}$ is given by the same linear function of the least squares solution.

Example 17.3

The vector of parameters in Example 17.1 is

$$\boldsymbol{\beta}' = (\,\mu \quad \alpha_1 \quad \alpha_2 \quad \beta_1 \quad \beta_2 \quad \beta_3 \quad \gamma_{11} \quad \gamma_{12} \quad \gamma_{13} \quad \gamma_{21} \quad \gamma_{22} \quad \text{—}\,)\,.$$

Notice that γ_{23} is missing since cell (2,3) is empty. A dash has been inserted in its place so that it is not forgotten. The estimates $\boldsymbol{\beta}^0$ computed by PROC GLM and their expectations are given in Table 17.6. Note that the first expectation in Table 17.6 is obtained by setting $L_1 = 1$, $L_2 = L_4 = L_5 = L_7 = 0$ in the general form for estimable functions (Table 17.3 and equation 17.18; the second by setting $L_2 = 1$, $L_1 = L_4 = L_5 = L_7 = 0$; the fourth by setting $L_4 = 1$, $L_1 = L_2 = L_5 = L_7 = 0$; the fifth by setting $L_5 = 1$, $L_1 = L_2 = L_4 = L_7 = 0$; and the seventh by setting $L_7 = 1$, $L_1 = L_2 = L_4 = L_5 = 0$. The $\mathcal{E}(LS\overline{Y}_{1..})$ in equation 17.20 is written in vector notation, $\mathcal{E}(LS\overline{Y}_{1..}) = \boldsymbol{K}_1'\boldsymbol{\beta}$, where $\boldsymbol{K}_1'$ is

$$\boldsymbol{K}_1' = (\,1 \quad 1 \quad 0 \quad \tfrac{1}{3} \quad \tfrac{1}{3} \quad \tfrac{1}{3} \quad \tfrac{1}{3} \quad \tfrac{1}{3} \quad \tfrac{1}{3} \quad 0 \quad 0 \quad 0\,)\,.$$

Thus, the least squares mean for the first level of factor A is computed from Table 17.6 as

$$\begin{aligned} LS\overline{Y}_{1..} &= \boldsymbol{K}_1'\boldsymbol{\beta}^0 \\ &= 9 + (-4) + (\frac{1}{3}) + (-\frac{2}{3}) = 4.667. \end{aligned}$$

$LS\overline{Y}_{2..}$ would be computed as $\boldsymbol{K}_2'\boldsymbol{\beta}^0$ where

$$\boldsymbol{K}_2' = (\,1 \quad 0 \quad 1 \quad \tfrac{1}{3} \quad \tfrac{1}{3} \quad \tfrac{1}{3} \quad 0 \quad 0 \quad 0 \quad \tfrac{1}{3} \quad \tfrac{1}{3} \quad \tfrac{1}{3}\,)$$

except for the fact that the last element in $\boldsymbol{K}_2'$ is the coefficient on the missing γ_{23}. Therefore, $LS\overline{Y}_{2..}$ cannot be computed, or $LS\overline{Y}_{2..}$ is nonestimable. Any least squares mean that has a nonzero coefficient on γ_{23} is nonestimable in this example.

Variances of LS Means

The variances for the least squares means that are estimable are obtained by applying the rule for variances of linear functions. For example, $\mathrm{Var}(LS\overline{Y}_{1..}) = \boldsymbol{K}_1'(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{K}_1\sigma^2$. The estimate of the variance is obtained by substituting s^2 for σ^2. The standard deviations of the least squares means are available on request as one of the options in PROC GLM. They are invariant to which generalized variance of $\boldsymbol{X}'\boldsymbol{X}$ is used.

Example 17.4

(Continuation of Example 17.1*)* Table 17.7 gives the estimates of the estimable least squares means, their standard errors, and the linear functions of the parameters being estimated. The least squares means for the individual cells of the table, the $A \times B$ means, are the same as the unadjusted means and their variances are σ^2/n_{ij}. This will always be the case for the smallest subdivision of a factorial table. ■

Summary

Section 17.4 has been a general introduction to the analysis of unbalanced data and has concentrated on the results obtained from PROC GLM. Freund, Littell, and Spector (1986) and Searle and Henderson (1979) are recommended reading for other applications and more detailed discussions of the use of PROC GLM. There are other computer programs that treat the analysis of unbalanced data, for f , BMDP (Dixon, 1981) and SPSS (Norusis, 1985). It is not always clear what sums of squares are being computed by the various programs and, therefore, what hypotheses are being tested. It is important that the user understand the program and its output in order to avoid misinterpretation of the results. Other references are Myers (1990) and Searle (1986).

17.5 Exercises

17.1 Unequal numbers of observations may be designed into an experiment. Discuss a situation in which it might be desirable to have

TABLE 17.7. *Least Squares means, their standard errors, and expectations for the* 2×3 *factorial example with cell* (2,3) *empty.*

		LSMEAN	*Std. Error*	*Expectations*
LSMEANS for A factor:				
A	1	4.67	1.029	$\mu + \alpha_1 + \overline{\beta}_{\cdot} + \overline{\gamma}_{1\cdot}$
	2	Nonest.	—	
LSMEANS for B factor:				
B	1	7.00	1.041	$\mu + \overline{\alpha}_{\cdot} + \beta_1 + \overline{\gamma}_{\cdot 1}$
	2	7.00	1.140	$\mu + \overline{\alpha}_{\cdot} + \beta_2 + \overline{\gamma}_{\cdot 2}$
	3	Nonest.	—	
LSMEANS for $A \times B$ factor:				
A	B			
1	1	4.00	1.317	$\mu + \alpha_1 + \beta_1 + \gamma_{11}$
1	2	5.00	1.612	$\mu + \alpha_1 + \beta_2 + \gamma_{12}$
1	3	5.00	2.280	$\mu + \alpha_1 + \beta_3 + \gamma_{13}$
2	1	10.00	1.612	$\mu + \alpha_2 + \beta_1 + \gamma_{21}$
2	2	9.00	1.612	$\mu + \alpha_2 + \beta_2 + \gamma_{22}$

unequal numbers of observations. Give the outline of the analysis—model, sources of variation, and degrees of freedom—and discuss whether Type I or Type III hypotheses would be more meaningful.

17.2 Table 17.1 gives the expectations of the cell means for a 2×3 factorial in a completely random experimental design. Construct a similar $A \times B$ table but for a randomized complete block design with balanced data. Assume the block effects are fixed effects. Include $A \times B$ interactions but do not include interactions with blocks. Demonstrate that the expectation of any contrast on treatment means, cell means, or marginal means does not involve block effects.

17.3 Reconstruct the table developed in Exercise 17.2 assuming there are three blocks but that treatment (2,3) is missing in Block 3. Identify the contrasts on cell treatment means and on marginal treatment means that are free of block effects. Would the analysis of cell means be appropriate for these data? Show why or why not.

17.4 Exercise 9.13 used data on survival time of patients with different types of cancer (Cameron and Pauling, 1978). The data are cross-classified with unequal numbers if both sex of patient and cancer type are considered. Use the logarithm of the ratio of days survival of the treated patient to the mean days survival of his or her control group as the dependent variable. (In the following analyses, include interaction effects between sex of patient and type of cancer in your models, but ignore differences in age.)

(a) Do an unweighted analysis of cell means to investigate the effects of sex, cancer type, and their interaction. Compute the within-cell variance and the harmonic mean of the numbers of observations, and summarize the results in an analysis of variance table. Note that Type I and Type III sums of squares are equal and that the ordinary means are equal to the least squares means. Can you explain why?

(b) Do a weighted analysis of cell means, weighted by n_{ij}, using PROC GLM or a similar program. Do any of the sums of squares agree with those obtained from the unweighted analysis of cell means? Do the ordinary means or the least squares means agree with those from the unweighted analysis?

(c) Use the general linear models approach (PROC GLM or similar program) to analyze the data. Compare this analysis with the weighted analysis of cell means. Compare the least squares means with those from the unweighted and weighted analysis of cell means.

17.5. Repeat Exercise 17.4 with the "Type × Sex" interactions omitted from all models. Compare the sums of squares, the ordinary means, and the least squares means with those obtained with interaction effects in the models.

17.6. In the weighted analysis of cell means, weighting was determined by the n_{ij}. This resulted from the assumption of constant variance for the ϵ_{ijk}; that is, $\mathbf{Var}(\boldsymbol{\epsilon}) = \boldsymbol{I}\sigma^2$. (See equation 17.9.) Suppose the variances for the observations as well as the numbers of observations differed from cell to cell. Let the variance of cell (i, j) be σ_{ij}^2. What would be an appropriate weighting for the weighted analysis of cell means? How would you determine numerical values for the weights?

17.7. Construct the general form of estimable functions $\boldsymbol{L}'\boldsymbol{\beta}$ for the nested model

$$Y_{ijk} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk},$$

where $i = 1, 2$ and $j = 1, 2$. Assume all effects are fixed effects and

$$\boldsymbol{\beta}' = (\,\mu \quad \alpha_1 \quad \alpha_2 \quad \beta_{11} \quad \beta_{12} \quad \beta_{21} \quad \beta_{22}\,).$$

(You need to define $\boldsymbol{X}$ for this model, eliminate any nonunique rows, and then use row operations to reduce $\boldsymbol{X}$ to the "near identity" form.) You should obtain

$$\boldsymbol{L}' = (L_1, \quad L_2, \quad (L_1 - L_2), \quad L_4, \quad (L_2 - L_4), \quad L_6, \quad (L_1 - L_2 - L_6)).$$

17.8. Use the general form of estimable functions in Exercise 17.7 to determine if each of the following is an estimable function. Give the choice

of coefficients that generates the linear function if it is an estimable function.

(a) $\mu + \alpha_1 + \beta_{11}$
(b) $\beta_{11} - \beta_{12}$
(c) $\beta_{11} + \beta_{12}$
(d) $\alpha_1 - \alpha_2$
(e) $\beta_{11} + \beta_{21} - \beta_{12} - \beta_{22}$
(f) $\mu + \alpha_1 + \frac{1}{2}(\beta_{11} + \beta_{12})$
(g) $\alpha_1 - \alpha_2 + \frac{1}{2}(\beta_{11} + \beta_{12} - \beta_{21} - \beta_{22})$

17.9. Use the model and the general form of estimable functions in Exercise 17.7 to answer each of the following. In each case, explain how you arrived at your answer. (*Note:* In the nested model, the nested effects β_{ij} "contain" the α_i effects.)

(a) How many degrees of freedom are there for SS(A)?
(b) How many degrees of freedom are there for SS($B(A)$)?
(c) Which coefficients will be zero for the Type I sum of squares SS(A)? For the Type III SS(A)?
(d) Which coefficients will be zero for the Type I sum of squares SS($B(A)$)? For the Type III SS($B(A)$)?

17.10. Construct the expectations of the least squares means for A and $B(A)$ for the nested model in Exercise 17.8. Are they all estimable if there are no empty cells?

17.11. Construct an artificial set of data for the nested model in Exercise 17.7 with $n_{11} = 2$, $n_{12} = 3$, $n_{21} = 1$, and $n_{22} = 2$ and use PROC GLM or a similar program to obtain the general linear form and the specific estimable functions for the sums of squares. Request the LSMEANS for A and $B(A)$. Compare the results with your answers to Exercises 17.8 through 17.10. (It does not matter what you use for the values of the dependent variable since the estimable functions depend only on $\boldsymbol{X}$.) Use PROC GLM to determine how SAS defines the least squares mean for the first level of factor A when cell (1,2) is empty.

17.12. The 1983 soybean data from Heagle (Table 16.8, page 531), contain one missing observation. Do the analysis of variance, using the general linear models approach (PROC GLM). Include block, ozone, moisture, and ozone × moisture interaction effects in the model. (Use the ozone treatment codes and ignore the slight differences in realized ozone levels.) Use the Type III sums of squares to interpret the results. Are all relevant least squares means estimable?

17.13. Do Exercise 17.12 using the 1984 soybean data from Heagle (Table 16.8).

17.14. Use the corn borer data in Exercise 9.4. Make the data unbalanced by assuming the first two observations in Days = 3 (the 17 and 22) and in Days = 6 (the 37 and 26) are missing. Analyze the data using (a) unweighted analysis of cell means, (b) weighted analysis of cell means, and (c) a general linear models procedure such as PROC GLM. Obtain the simple treatment means and the least squares treatment means. Do they differ? Why or why not?

17.15. The Weber data, Exercise 9.7, is a $2 \times 2 \times 5$ factorial in a randomized complete block design with $r = 2$ blocks. Make the data unbalanced by assuming that the two highest concentrations (80 and 100) of herbicide B could not be used at the high temperature (55° C). (Call all treatment factors class variables.) Include block effects, treatment main effects, and treatment interaction effects in the model. Use PROC GLM to analyze the data and obtain the simple and least squares treatment means.

(a) Which sums of squares will you use for testing hypotheses about the treatment effects? Explain why you choose the particular set you do.

(b) Which least squares means are nonestimable? Explain why these particular means are nonestimable. Do the results of the analysis let you simplify the model so that all relevant means are estimable?

(c) Summarize the results with tables of relevant least squares means and their standard errors.

18
MIXED EFFECTS MODELS

The models considered in all of the previous chapters contain only one random element, the random error. Many situations call for models in which there is more than one random term. This chapter introduces mixed models that contain both fixed effects and several random effects. Analysis of variance models for randomized block designs and split-plot experiments and models for repeated measurement data are special cases of mixed effects models. Hypothesis testing based on generalized least squares (GLS), maximum likelihood (ML), and restricted maximum likelihood (REML) are discussed.

Fixed Models

The classical least squares model contains only one random element, the random error; all other effects are assumed to be fixed constants. For this class of models, the assumption of independence of the ϵ_i implies independence of the Y_i. That is, if $\mathbf{Var}(\boldsymbol{\epsilon}) = \boldsymbol{I}\sigma^2$, then $\mathbf{Var}(\boldsymbol{Y}) = \boldsymbol{I}\sigma^2$ also. Such models are called **fixed effects models** or more simply, **fixed models**.

Random Models

Many situations call for models in which there is more than one random term. The classical **variance components problems**, in which the purpose is to estimate components of variance rather than specific treatment effects, is one example. In these cases, the "treatment effects" are assumed to be a random sample from a population of such effects and the goal of the study is to estimate the variance among these effects in the population. The individual effects that happen to be observed in the study are not

of any particular interest except for the information they provide on the variance component. Models in which all effects are assumed to be random effects are called **random models**. Observational studies often involve a hierarchy of nested effects that represent "levels" of random sampling of some population, such as random homes in random counties in random states.

The sampling of environments in which controlled experiments are conducted, locations and years, often are regarded as a random sampling of environmental conditions. The purpose is to infer behavior of the fixed treatments over some population of environments, rather than just to the particular set of environments encountered in the experiments. In such cases, the treatment effects may be fixed and the environments assumed to be random. Models that contain both fixed and random effects are called **mixed models**. The appropriate model for the commonly used split-plot experimental design specifies two random terms in the model, the whole-plot error and the subplot error, and hence is a mixed model if treatment effects are assumed fixed.

Mixed Models

The net effect of more than one random term in the model is that $\mathbf{Var}(\boldsymbol{Y}) \neq \boldsymbol{I}\sigma^2$ even if $\mathbf{Var}(\boldsymbol{\epsilon}) = \boldsymbol{I}\sigma^2$. The random elements shared by observations introduce nonzero covariances among all observations having common "levels" of the random effects.

As discussed in Chapter 12, if $\mathbf{Var}(\boldsymbol{Y}) \neq \boldsymbol{I}\sigma^2$, the ordinary least squares estimator of the fixed effects may be inefficient and the standard errors computed using $\mathbf{Var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$ are inappropriate. In mixed effects models, $\mathbf{Var}(\boldsymbol{Y})$ is modeled as a function of some unknown variance–covariance parameters. Estimation and hypothesis testing regarding the variance–covariance parameters are also of interest in practice. Estimates of the variance–covariance parameters are used to obtain the estimated generalized least squares estimates of the fixed effects. First, we present examples and traditional analysis of variance methods for balanced mixed effects models. Then, we present the analyses based on maximum likelihood and restricted maximum likelihood estimation methods for general mixed linear models.

18.1 Random Effects Models

As an example, suppose you want to investigate the magnitude of genetic variability for a particular characteristic present in a collection of soybean cultivars in the genetic seed bank. (One enitity in the collection of genetic material is called a cultivar.) The total collection contains thousands of cultivars of which a researcher will test a random sample in a completely randomized design with n replicate plots of each of a cultivars. The particular characteristic of interest, say seed yield, is measured for each plot.

Let Y_{ij} denote the yield for the ith variety from the jth plot. Here, the researcher is more interested in studying the variability among the cultivars in the entire collection, or population, than in the effects of the few cultivars selected by chance to be included in the study. The cultivar effects are considered to be random and the quantity of interest is the estimate of variance among cultivars.

One-Way ANOVA Model

An appropriate model in this case is the one-way analysis of variance (ANOVA) model:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \begin{array}{l} i = 1, \ldots, a: \text{ cultivars}, \\ j = 1, \ldots, n: \text{ plots}, \end{array} \tag{18.1}$$

where the α_i are assumed to be independent $N(0, \sigma_\alpha^2)$, the ϵ_{ij} are assumed to be independent $N(0, \sigma^2)$, and $\{\alpha_i\}$ and $\{\epsilon_{ij}\}$ are independent.

Scheffé (1959) gives a motivation for the model in equation 18.1. Consider the model

$$Y_{lj} = M_l + \epsilon_{lj}, \tag{18.2}$$

where M_l is the mean yield of the lth cultivar in the *population*. The variability around its mean M_l for the lth cultivar in the population is measured by the variance σ^2. Assume that the population is large. Let μ and σ_α^2 denote the mean and the variance of M_l in the population. Then, from equation 18.2,

$$Y_{lj} = \mu + A_l + \epsilon_{lj},$$

where $A_l = M_l - \mu$ has mean zero and variance σ_α^2. Since a random sample of cultivars is selected from the population, the α_i in equation 18.1 may be viewed as a random sample from the population of A_l. That is, the α_i can be assumed to be independent random variables with mean zero and variance σ_α^2.

Variance Components

The model in equation 18.1 contains two random components $\{\alpha_i\}$ and $\{\epsilon_{ij}\}$. Note that

$$\text{Var}(Y_{ij}) = \sigma_\alpha^2 + \sigma^2 \tag{18.3}$$

and, hence, σ_α^2 and σ^2 are called the **components of variance** or **variance components**. Also, note that

$$\text{Cov}(Y_{ij}, Y_{st}) = \begin{cases} \sigma_\alpha^2, & i = s, j \neq t \\ 0, & i \neq s. \end{cases} \tag{18.4}$$

Therefore, for the model in equation 18.1, $\mathbf{Var}(\boldsymbol{Y}) \neq \boldsymbol{I}\sigma^2$. In this model it is of primary interest to estimate σ_α^2 and σ^2 and, secondarily, to test the hypothesis that $\sigma_\alpha^2 = 0$ (i.e., no variability among the cultivars).

Analysis of Variance Approach

The conventional least squares approach, sometimes called the **analysis of variance approach**, to estimate the variance components in random

TABLE 18.1. *One-way analysis of variance and mean square expectations for a random effects model.*

Source	*d.f.*	*Sum of Squares*	$\mathcal{E}$(*Mean Square*)	*F for testing* $H_0 : \sigma_\alpha^2 = 0$
Cultivars	$a-1$	$n\sum_{i=1}^{a}(\overline{Y}_{i.} - \overline{Y}_{..})^2$	$\sigma^2 + n\sigma_\alpha^2$	$\frac{\text{MS(Cultivars)}}{\text{MS(Res)}}$
Error	$a(n-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i.})^2$	σ^2	

effects models is to calculate sums of squares as though all effects, other than the unique error assigned to each observation, were fixed effects. These sums of squares and their expectations under the *random* model are then used to estimate the variance components.

Consider the analysis of variance in Table 18.1 for the model in equation 18.1 where α_i are considered "fixed." Note that

$$\begin{aligned}\mathcal{E}[\text{MS(Cultivars)}] &= \mathcal{E}\left[n\sum_{i=1}^{a}\left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2/(a-1)\right] \\ &= n\mathcal{E}\left[\sum_{i=1}^{a}\left(Z_i - \overline{Z}\right)^2/(a-1)\right] = n\sigma_Z^2, \quad (18.5)\end{aligned}$$

where $Z_i = \alpha_i + \overline{\epsilon}_{i.}$ are independent random variables with mean zero and variance $\sigma_Z^2 = \sigma_\alpha^2 + \sigma^2/n$. Therefore, the expectation of the cultivar effects mean square is $\sigma^2 + n\sigma_\alpha^2$. Similarly, the expectation of the residual mean square is

$$\begin{aligned}\mathcal{E}[\text{MS(Res)}] &= \mathcal{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i.})^2/[a(n-1)]\right] \\ &= \frac{1}{a}\sum_{i=1}^{a}\mathcal{E}\left[\sum_{j=1}^{n}\left(\epsilon_{ij} - \overline{\epsilon}_{i.}\right)^2/(n-1)\right] = \sigma^2. \quad (18.6)\end{aligned}$$

The analysis of variance estimators of σ_α^2 and σ^2 are given by equating the mean squares to their expectations and solving the set of equations. Thus,

$$\begin{aligned}\widehat{\sigma}_\alpha^2 &= \left[\text{MS(Cultivars)} - \text{MS(Res)}\right]/n \\ \widehat{\sigma}^2 &= \text{MS(Res)}. \quad (18.7)\end{aligned}$$

From equations 18.5 and 18.6, it is clear that the estimators $\widehat{\sigma}_\alpha^2$ and $\widehat{\sigma}^2$ in equation 18.7 are unbiased for σ_α^2 and σ^2, respectively. In some samples, it is possible that $\widehat{\sigma}_\alpha^2$ will be negative. This analysis of variance method is an example of the "method of moments" estimation.

Since $\text{Cov}(\bar{\epsilon}_{i\cdot}, \epsilon_{ij} - \bar{\epsilon}_{i\cdot}) = 0$, it follows that under the normality assumption, $\{\alpha_i + \bar{\epsilon}_{i\cdot}\}$ are independent of $\{\epsilon_{ij} - \bar{\epsilon}_{i\cdot}\}$. Hence, we have

$$\begin{aligned} \frac{(a-1)\text{MS(Cultivars)}}{\sigma^2 + n\sigma_\alpha^2} &\sim \chi^2_{a-1} \\ \frac{a(n-1)\text{MS(Res)}}{\sigma^2} &\sim \chi^2_{a(n-1)} \end{aligned} \tag{18.8}$$

and MS(Cultivars) and MS(Res) are independent of each other. The variances of the estimators of σ_α^2 and σ^2 (equation 18.7) are computed as the variance of linear functions of mean squares. Since $\text{Var}(\chi^2_\nu) = 2\nu$, we have

Variances of Estimators

$$\text{Var}(\widehat{\sigma}^2) = \frac{2\sigma^4}{a(n-1)} \tag{18.9}$$

and

$$\text{Var}(\widehat{\sigma}_\alpha^2) = \frac{2\left(\sigma^2 + n\sigma_\alpha^2\right)^2}{(a-1)n^2} + \frac{2\sigma^4}{a(n-1)n^2}. \tag{18.10}$$

Testing $\sigma_\alpha^2 = 0$

Since the scaled mean squares have chi-square distributions, equation 18.8, and are mutually independent, the variance ratio

$$F = \frac{\text{MS(Cultivars)}}{\text{MS(Res)}} \tag{18.11}$$

has an F distribution with numerator degrees of freedom $(a-1)$ and denominator degrees of freedom $a(n-1)$ if $H_0 : \sigma_\alpha^2 = 0$ is true. Thus, an α-level test criterion for testing that there is no cultivar effect is a test that $\sigma_\alpha^2 = 0$. The test rejects the null hypothesis if $F > F_{(\alpha;a-1,a(n-1))}$. Note that this is the same test criterion that would have been used had the cultivar (or treatment) effects been considered fixed. Other approaches for estimation and hypothesis testing are discussed in Section 18.4.

Randomized Block Design

Consider a randomized complete block design for the previous investigation of the genetic variance among cultivars in the soybean seed bank. Suppose a random sample of locations (used as blocks) is selected and within each location a plots are used, one for each of the a selected cultivars. Let Y_{ij} denote the yield from the ith cultivar in the jth location. A model that may be appropriate in this case is

$$\begin{aligned} Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad & i = 1, \ldots, a : \text{cultivars}, \\ & j = 1, \ldots, n : \text{locations}, \end{aligned} \tag{18.12}$$

where the cultivar effects $\{\alpha_i\}$ are independent $N(0, \sigma_\alpha^2)$, the location effects $\{\beta_j\}$ are independent $N(0, \sigma_\beta^2)$; $\{\epsilon_{ij}\}$ are independent $N(0, \sigma^2)$, and $\{\alpha_i\}$, $\{\beta_j\}$, and $\{\epsilon_{ij}\}$ are mutually independent. Note that

$$\text{Var}(Y_{ij}) = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma^2.$$

TABLE 18.2. *Two-way analysis of variance for a random effects model.*

Source	*d.f.*	*Sum of Squares*	$\mathcal{E}$(*Mean Square*)
Cultivars	$a-1$	$n\sum_{i=1}^{a}(\overline{Y}_{i.}-\overline{Y}_{..})^2$	$\sigma^2+n\sigma_\alpha^2$
Locations	$n-1$	$a\sum_{j=1}^{n}(\overline{Y}_{.j}-\overline{Y}_{..})^2$	$\sigma^2+a\sigma_\beta^2$
Error	$(a-1)(n-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{i.}-\overline{Y}_{.j}+\overline{Y}_{..})^2$	σ^2

Also, note that

$$\mathrm{Cov}(Y_{ij},Y_{st})=\begin{cases}\sigma_\alpha^2+\sigma_\beta^2+\sigma^2 & , \quad i=s, j=t\\ \sigma_\alpha^2 & , \quad i=s, j\neq t\\ \sigma_\beta^2 & , \quad i\neq s, j=t\\ 0 & , \quad \text{otherwise}\end{cases}$$

and, hence, $\mathbf{Var}(\boldsymbol{Y})\neq \boldsymbol{I}\sigma^2$.

The analysis of variance table, including expected mean squares, is given in Table 18.2. Equating the mean squares to their expectations gives the analysis of variance estimators of σ_α^2, σ_β^2, and σ^2:

$$\begin{aligned}\widehat{\sigma}_\alpha^2 &= \frac{\mathrm{MS(Cultivars)}-\mathrm{MS(Res)}}{n}\\ \widehat{\sigma}_\beta^2 &= \frac{\mathrm{MS(Locations)}-\mathrm{MS(Res)}}{a}, \text{ and}\\ \widehat{\sigma}^2 &= \mathrm{MS(Res)},\end{aligned} \tag{18.13}$$

where

$$\begin{aligned}\mathrm{MS(Res)} &= \sum_{i=1}^{a}\sum_{j=1}^{n}\left(Y_{ij}-\overline{Y}_{i.}-\overline{Y}_{.j}+\overline{Y}_{..}\right)^2/[(a-1)(n-1)],\\ \mathrm{MS(Cultivars)} &= n\sum_{i=1}^{a}\left(\overline{Y}_{i.}-\overline{Y}_{..}\right)^2/(a-1), \text{ and}\\ \mathrm{MS(Locations)} &= a\sum_{j=1}^{n}\left(\overline{Y}_{.j}-\overline{Y}_{..}\right)^2/(n-1).\end{aligned}$$

As in the case of the completely randomized design, it can be shown that these three mean squares are mutually independent and, when properly normalized, are distributed as chi-square random variables. Variances of estimators in equation 18.13 can be obtained as in equations 18.9 and 18.10. Hypotheses $H_0:\sigma_\alpha^2=0$ (no variance among cultivars) and $H_0:\sigma_\beta^2=0$

(no variance among locations) can be tested using the F statistics that are used in the case of fixed effects models:

$$H_0 : \sigma_\alpha^2 = 0 \qquad F = \frac{\text{MS(Cultivars)}}{\text{MS(Res)}}$$
$$H_0 : \sigma_\beta^2 = 0 \qquad F = \frac{\text{MS(Locations)}}{\text{MS(Res)}}.$$

In general, however, the appropriate F-tests will not be the same for the fixed and random models. For example, in the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

with all effects but μ random, a similar analysis shows that the appropriate denominator mean square for the F-tests for the null hypotheses $\sigma_\alpha^2 = 0$ and $\sigma_\beta^2 = 0$ is MS(Cultivar × Location). In the fixed model, MS(Residual) is the appropriate denominator mean square in both tests.

18.2 Fixed and Random Effects

In many situations, some of the effects are fixed and some others are random effects. For example, consider a randomized block experiment where the treatments (or varieties) are fixed but block effects are random. An appropriate model for this experiment is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad \begin{array}{lll} i = 1, \ldots, a & : & \text{treatments} \\ j = 1, \ldots, n & : & \text{blocks,} \end{array} \tag{18.14}$$

where the α_i are fixed effects, $\beta_j \sim \text{NID}(0, \sigma_\beta^2)$, $\epsilon_{ij} \sim \text{NID}(0, \sigma^2)$, and $\{\beta_j\}$ and $\{\epsilon_{ij}\}$ are independent. Such models, where some effects are fixed and others are random are called **mixed effects models**. The degrees of freedom and sums of squares presented in Table 18.2 are also appropriate for the mixed effects model, equation 18.14, but the expectation of the mean square for treatments (cultivars) will be $\sigma^2 + n\sum(\alpha_i - \overline{\alpha}_\cdot)^2/(a-1)$.

Mixed Effects Models

Another example of a mixed effects model is that for the split-plot experiment where a whole-plot treatments are each applied to n whole-plot units. Within each whole-plot, b split-plot treatments are applied in a completely random fashion to b subunits. An appropriate model for the response Y_{ijk}, from the subunit receiving the kth split-plot treatment in the jth whole-plot receiving the ith whole-plot treatment is

Split-Plot Experiment

$$Y_{ijk} = \mu + \alpha_i + \delta_{ij} + \beta_k + \gamma_{ik} + \epsilon_{ijk}, \quad \begin{array}{l} i = 1, \ldots, a; \\ j = 1, \ldots, n; \\ k = 1, \ldots, b, \end{array} \tag{18.15}$$

TABLE 18.3. *Degrees of freedom and expected mean squares for the split-plot analysis of variance.*

Source[a]	*d.f.*	$\mathcal{E}$(*Mean squares*)
Treatment A	$a-1$	$\sigma^2 + b\sigma_\delta^2 + \frac{nb}{(a-1)}\sum_{i=1}^{a}(\alpha_i - \overline{\alpha}_\cdot)^2$
Error (a)	$a(n-1)$	$\sigma^2 + b\sigma_\delta^2$
Treatment B	$b-1$	$\sigma^2 + \frac{na}{(b-1)}\sum_{k=1}^{b}(\beta_k + \overline{\gamma}_{\cdot k} - \overline{\beta}_\cdot - \overline{\gamma}_{\cdot\cdot})^2$
Interaction	$(a-1)(b-1)$	$\sigma^2 + \frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{k=1}^{b}(\gamma_{ik} - \overline{\gamma}_{i\cdot} - \overline{\gamma}_{\cdot k} + \overline{\gamma}_{\cdot\cdot})^2$
Error (b)	$a(n-1)(b-1)$	σ^2

[a]Treatment A and Treatment B are the whole-plot and subplot treatments, respectively.

where

α_i is the effect of the ith whole-plot treatment,
δ_{ij} is the whole-plot error,
β_k is the effect of the kth subplot treatment,
γ_{ik} is the interaction effect due to the ith and kth levels of the treatments, and
ϵ_{ijk} is the subplot error.

The treatment effects $\{\alpha_i\}$, $\{\beta_k\}$, and $\{\gamma_{ik}\}$ are assumed fixed; the errors $\{\delta_{ij}\}$ and $\{\epsilon_{ijk}\}$ are considered random. It is assumed that $\delta_{ij} \sim \text{NID}(0, \sigma_\delta^2)$, $\epsilon_{ijk} \sim \text{NID}(0, \sigma^2)$, and $\{\delta_{ij}\}$ and $\{\epsilon_{ijk}\}$ are independent. The analysis of variance approach for estimation and hypothesis testing are summarized in Table 18.3.

The different sums of squares in Table 18.3 are mutually independent and, when properly normalized, are distributed as chi-square random variables. From Table 18.3 we note that the appropriate denominator sum of squares for testing null hypothesis of no whole-plot treatment effect is the Error(a) sum of squares, whereas for testing the null hypothesis of no subplot treatment effects the Error(b) sum of squares is appropriate.

Limitations of Analysis of Variance Approach

With balanced data, the method of moments estimation (equation 18.7) generates the conventional analysis of variance for the design and, with the appropriate adjustment of the mean square expectations for the random effects, gives the same results as would be obtained with a full generalized least squares analysis. The generalized least squares analysis is not obtained by this method, however, when the data are not balanced. Nevertheless, the analysis of variance approach has traditionally been used for unbalanced data with the variance component estimates obtained by equating observed mean squares to their expectations.

All computations in the classical approach to the analysis of mixed models begin as though the model were fixed—having only one random element. As already noted, the estimates of $\boldsymbol{\beta}$ and all linear functions of $\boldsymbol{\beta}$ obtained by this approach will not be the best linear unbiased estimates since the true variance–covariance structure is not being taken into account. The estimates are unbiased but there will be some loss in precision. In addition and perhaps more critically, if no adjustments are made for random effects, the tests of significance and the computed measures of precision $\boldsymbol{s}(\boldsymbol{\beta}^0)$, $\boldsymbol{s}(\widehat{\boldsymbol{Y}})$, and standard errors of the least squares means will be incorrect. That is, it is incorrect to compute measures of precision as if $\boldsymbol{I}\sigma^2$ were the true variance–covariance matrix of $\boldsymbol{Y}$ rather than the more general $\mathbf{Var}(\boldsymbol{Y})$. If ordinary least squares is to be used for the analysis of models with more than one random component, adjustments to the tests of significance and the estimates of the standard errors must be considered.

Expectations of Mean Squares

Adjustments to tests of significance are made by "constructing" an error mean square that has the proper expectation with respect to the random elements. This requires the expectations of the mean squares under the random model. For balanced data the mean square expectations are easily obtained and are reported in many places [e.g., Searle (1971, 1986), and Steel, Torrie, and Dickey (1997)]. For unbalanced data, computer programs provide the expectations. The "RANDOM" statement in PROC GLM prompts the program to provide the mean square expectations under a mixed model in which the random effects are specified in the "RANDOM" statement. (The "MODEL" statement in GLM specifies all classes of effects, fixed and random, except for the unique random element associated with each observation.) The expectations are given for any of four types of sums of squares available in PROC GLM and all contrasts used in the analysis. The expectations are expressed in terms of linear functions of the variance components for the random effects plus general symbols representing the fixed effects involved in the quadratic functions. The specific quadratic functions of the fixed effects can also be obtained, if needed.

Example 18.1

To illustrate the use of the results provided by the "RANDOM" statement, the mean square expectations are given here for the Type III sums of squares for the whole plot treatment factor and the whole plot error for the unbalanced data analyzed in Chapter 19. The experiment is a split-plot experiment with the whole-plot treatments [a factorial set of treatments involving two factors, tillage (TILL) and herbicide (HERB)] arranged in a randomized complete block experimental design. The estimate of the whole-plot error, Error (a), was computed from the pooled "Block × TILL × HERB," "Block × TILL," and "Block × HERB" interaction sums of squares. (You are referred to Chapter 19 for the details of the experiment.) The expectations for the Type III mean squares for treatment factor TILL

and Error (a) are as follows.

Source	*Type III Expected Mean Square*
TILL	$\sigma^2 + 1.0909\,\sigma^2_\delta + Q(TILL, TILL \times HERB)$
Error (a)	$\sigma^2 + 1.9048\,\sigma^2_\delta$

The expectation of the residual mean square is σ^2. The $Q(.)$ function indicates that the mean square expectation is a quadratic function of the "$TILL$" and "$TILL \times HERB$" treatment effects. For simplicity, let $E_a = \text{MS(Error (a))}$ and $E_b = \text{MS(Res)}$ denote the Error (a) and Error (b) mean squares.

If these data had been balanced, the coefficient on σ^2_δ would have been 2 in each case, the number of levels of the subplot treatment factor, and E_a would have been the appropriate error for testing the null hypothesis that $Q(TILL, TILL \times HERB) = 0$. With the imbalance, the coefficients on σ^2_δ differ and, consequently, E_a is not the appropriate error for the test. An approximate test is obtained by constructing a mean square that has the correct expectation. The test for $H_0 : Q(TILL, TILL \times HERB) = 0$ requires a denominator mean square whose expectation is $\sigma^2 + 1.0909\sigma^2_\delta$. Such a mean square is constructed as a linear function of E_a and E_b as follows.

$$E' = \left(\frac{1.0909}{1.9048}\right) E_a + \left(1 - \frac{1.0909}{1.9048}\right) E_b. \tag{18.16}$$

The approximate test of $H_0 : Q(TILL, TILL \times HERB)$ is then

$$F' = \text{MS}(TILL)/E'.$$

■

Distribution of F'

The constructed variance ratio F' in Example 18.1 is only approximately distributed as an F-statistic for the following reasons. First, a linear function of mean squares does not behave quite like a chi-square random variable as is required for the F-test. The degrees of freedom f' for E' are determined so as to minimize this problem (Satterthwaite, 1946). Second, the Type III sums of squares, in general, are not orthogonal partitions of the model sum of squares and, hence, the numerator and denominator mean squares in F' are not independent. This lack of independence is ignored in the test of significance.

The Satterthwaite (1946) approximation for the degrees of freedom for a linear function of mean squares $\sum a_i \text{MS}_i$ is

$$f' = \frac{\left(\sum_i a_i\,\text{MS}_i\right)^2}{\sum_i \left(\frac{a_i^2\,\text{MS}_i^2}{f_i}\right)}, \tag{18.17}$$

where f_i is the degrees of freedom of MS_i. This approximation for the degrees of freedom is obtained by equating the mean and the variance of $\sum a_i \mathrm{MS}_i$ to that of a constant multiple of a chi-square random variable. See Exercise 18.9. For E' in equation 18.16, $a_1 = (1.0909)/(1.9048)$, $a_2 = 1-(1.0909)/(1.9048)$, and f_1 and f_2 are the degrees of freedom for Error (a) and Error (b), respectively.

Differences in Defining Interaction Effects

A word of warning is needed on the use of the mean square expectations. There are differences of opinion on how interaction effects between a fixed and a random factor are to be handled in deriving mean square expectations. Some argue that if one of the factors involved in the interaction is a random factor, the interaction effects should be treated as completely random variables with no constraints imposed on their behavior. In such cases, the interaction component of variance is present in the expectations of the interaction mean square and both main effects mean squares. SAS uses this procedure in deriving expectations in the "RANDOM" statement in PROC GLM (SAS Institute Inc., 1989b).

The classical approach to handling interaction effects is to impose the constraint that the interaction effects sum to zero over the levels of the *fixed* factor; that is, the effects sum to zero in the fixed direction of the two-way table of effects. This causes the interaction component of variance to "drop out" of the mean square expectation for the *random* main effect. These expectations are the logical extension of those derived under a two-dimensional finite sampling model in which the samples of effects for factor A and factor B are assumed to have resulted from taking random samples from the two finite populations of effects. Let N_a and N_b be the two population sizes and n_a and n_b be the respective sample sizes, $n_a \leq N_a$ and $n_b \leq N_b$. The mean square expectations for the *mixed* model are then obtained from this finite model by letting the population size go to infinity for the random factor and letting the number of levels sampled equal the number of population levels for the fixed factor. The covariances among the effects due to the finiteness of the population cause the interaction effects to drop out of the mean square expectation for the random factor. See Exercises 18.5 and 18.6.

These differences in philosophy do not enter into the present split-plot example since all treatment factors are assumed to be fixed. The differences will affect the choice of error in many cases and the reader needs to be aware of the problem. The reader is referred to Speed and Hocking (1976) for more discussion on this point.

Adjusting Measures of Precision

Two methods of adjusting the measures of precision obtained from the standard least squares analysis might be used. If the generalized inverse of $\boldsymbol{X}'\boldsymbol{X}$ is available from the computer program, the correct variance–covariance matrix for any linear function of $\boldsymbol{\beta}^0$ can be computed using a matrix program such as IML (SAS Institute Inc., 1989d). Let the rows of $\boldsymbol{L}'\boldsymbol{\beta}^0$ be k linear functions of $\boldsymbol{\beta}^0$ of interest and let $\boldsymbol{s}(\boldsymbol{Y})$ be an estimate of $\mathbf{Var}(\boldsymbol{Y})$, the variance–covariance matrix of $\boldsymbol{Y}$. The true variance of $\boldsymbol{L}'\boldsymbol{\beta}^0$,

when $\boldsymbol{\beta}^0$ is the least squares estimator $(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{Y}$, computed under the *incorrect* assumption that $\mathbf{Var}(\boldsymbol{Y}) = \boldsymbol{I}\sigma^2$, is

$$\mathbf{Var}(\boldsymbol{L}'\boldsymbol{\beta}^0) = \boldsymbol{L}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'[\mathbf{Var}(\boldsymbol{Y})]\boldsymbol{X}[(\boldsymbol{X}'\boldsymbol{X})^-]'\boldsymbol{L} \tag{18.18}$$

and is estimated by substituting $\boldsymbol{s}^2(\boldsymbol{Y})$ for $\mathbf{Var}(\boldsymbol{Y})$. This gives $\boldsymbol{s}^2(\boldsymbol{\beta}^0)$ if $\boldsymbol{L}'$ is the identity matrix, $\boldsymbol{s}^2(\widehat{\boldsymbol{Y}})$ if $\boldsymbol{L}'$ is $\boldsymbol{X}$, and $\boldsymbol{s}^2(\mathbf{LSMEANS})$ if $\boldsymbol{L}'$ consists of row vectors of the estimable functions for the least squares means.

As an alternative to computing the exact variances, expectations of the mean squares can be used to make approximate adjustments to standard errors of the least squares means. The expectation for the random elements of a particular mean square provides an *average* variance for the class of means involved in that mean square. As with the tests of significance, a mean square can be constructed that has this expectation. Multiplication of the standard errors reported for any particular class of means by

$$\text{Ratio} = \left[\frac{\text{Constructed MS}}{\text{MS(Residual)}}\right]^{1/2}$$

provides reasonable approximations of the standard errors. (A comparison of the two methods is given for the case study in Chapter 19.)

18.3 Random Coefficient Regression Models

In biological, medical, agricultural, and clinical studies several measurements are often taken on the same experimental unit over time or under different experimental conditions with the objective of fitting a response curve to the data. **Random coefficient regression models** have been used to analyze such data. Consider, for example, a study where n individuals are selected from a population. For each individual different doses of pain relief medication are given on different days. The response time, time until the individual felt pain relief, is recorded. Let X_{ij} and Y_{ij} denote the dosage and the response times of the ith indivdual on the jth day. An appropriate model for such data is

$$\begin{aligned} Y_{ij} = \alpha_i + \beta_i X_{ij} + \epsilon_{ij}, \;\; & i = 1, \ldots, n: \;\; \text{individuals}, \\ & j = 1, \ldots, r: \;\; \text{days}, \end{aligned} \tag{18.19}$$

where α_i and β_i are the intercept and the slope of the ith individual. That is, we think that the relationship between the response time and the dosage is of the same form for all individuals, but the parameters (coefficients) of the relationship may differ among individuals. Since individuals are assumed to be a random sample from a population, it is common to assume that

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \;\sim\; \text{NID}\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{bmatrix}\right),$$

$$\epsilon_{ij} \sim \text{NID}(0, \sigma^2),$$

and $\{(\alpha_i, \beta_i)'\}$ and $\{\epsilon_{ij}\}$ are independent. (Note that an analysis of covariance model with random treatment effects is a special case of the model in equation 18.19 with $\sigma_\beta^2 = 0$.)

The model in equation 18.19 can also be written as

$$Y_{ij} = [\alpha + \beta X_{ij}] + [(\alpha_i - \alpha) + (\beta_i - \beta)X_{ij} + \epsilon_{ij}], \tag{18.20}$$

where α and β correspond to the fixed population average response and $(\alpha_i - \alpha)$, $(\beta_i - \beta)$, and ϵ_{ij} are the random deviations of individual responses from the average population response. Here

$$\text{Var}(Y_{ij}) = \sigma_\alpha^2 + 2\sigma_{\alpha\beta}X_{ij} + \sigma_\beta^2 X_{ij}^2 + \sigma^2$$

and for $j \neq l$

$$\text{Cov}(Y_{ij}, Y_{il}) = \sigma_\alpha^2 + \sigma_\beta^2 X_{ij} X_{il}.$$

That is, $\mathbf{Var}(\boldsymbol{Y}) \neq \boldsymbol{I}\sigma^2$ in this case either.

A simple extension of the analysis of variance estimation is to estimate the individual coefficients α_i and β_i using least squares and then use the individual estimates to estimate α, β, σ_α^2, $\sigma_{\alpha\beta}$, σ_β^2, and σ^2.

Gumpertz and Pantula (1989) consider a general random coefficient model for the case where t observations are measured on each of the n experimental units, given by

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, n, \tag{18.21}$$

where $\boldsymbol{Y}_i = (\, Y_{i1} \quad \cdots \quad Y_{it} \,)'$ is a $t \times 1$ vector of responses for the ith individual, $\boldsymbol{X}_i$ is a $t \times k$ matrix of observations on k explanatory variables, $\boldsymbol{\beta}_i$ is $k \times 1$ vector of coefficients unique to the ith experimental unit, and $\boldsymbol{\epsilon}_i$ is a $t \times 1$ vector of errors. It is assumed that $\boldsymbol{\beta}_i \sim \text{NID}(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, $\boldsymbol{\epsilon}_i \sim \text{NID}(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$, and $\{\boldsymbol{\beta}_i\}$ and $\{\boldsymbol{\epsilon}_i\}$ are independent. It is of interest to estimate, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ and test hypotheses regarding these parameters. Assuming that $\boldsymbol{X}_i'\boldsymbol{X}_i$ is nonsingular, we can obtain least squares estimates $\widehat{\boldsymbol{\beta}}_i$ of $\boldsymbol{\beta}_i$ for each individual. That is,

$$\widehat{\boldsymbol{\beta}}_i = (\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{Y}_i. \tag{18.22}$$

It is easy to see that

$$\mathbf{Var}(\boldsymbol{Y}_i) = \boldsymbol{X}_i\boldsymbol{\Sigma}\boldsymbol{X}_i' + \boldsymbol{I}\sigma^2 \tag{18.23}$$

and

$$\mathbf{Cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_l) = \boldsymbol{0}, \qquad \text{for } i \neq l.$$

Note that

$$\widehat{\boldsymbol{\beta}}_i \sim \text{NID}(\boldsymbol{\beta}, (\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\sigma^2 + \boldsymbol{\Sigma}). \tag{18.24}$$

Gumpertz and Pantula (1989) consider the simple estimator

$$\widehat{\boldsymbol{\beta}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{\beta}}_i$$

for $\boldsymbol{\beta}$ and suggest test criteria for testing hypotheses regarding $\boldsymbol{\beta}$. They also show that

$$\widehat{\boldsymbol{\Sigma}} = (n-1)^{-1}\sum_{i=1}^{n}(\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}})' - \widehat{\sigma}^2\, n^{-1}\sum_{i=1}^{n}(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}$$

and

$$\widehat{\sigma}^2 = [n(t-k)]^{-1}\sum_{i=1}^{n}[\boldsymbol{Y}_i'\boldsymbol{Y}_i - \widehat{\boldsymbol{\beta}}_i'\boldsymbol{X}_i'\boldsymbol{Y}_i]$$

are unbiased for $\boldsymbol{\Sigma}$ and σ^2.

As in the case of random and mixed effects models, the simple approaches are reasonable for balanced data. When the data are not balanced or have missing values, such approaches may be infeasible and/or may lead to inefficient estimates. In the next section, the maximum likelihood and restricted maximum likelihood methods that are more appropriate for mixed effects models are discussed.

18.4 General Mixed Linear Models

The models considered in Sections 18.1 through 18.3 are special cases of the **general mixed linear model** given by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\nu} + \boldsymbol{\epsilon}, \tag{18.25}$$

where $\boldsymbol{X}$ is an $N \times p$ matrix of known constants, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed parameters ("effects"), $\boldsymbol{Z}$ is a $N \times q$ matrix of known constants, $\boldsymbol{\nu}$ is a $q \times 1$ vector of unknown random effects, and $\boldsymbol{\epsilon}$ is the $N \times 1$ vector of random errors. Assume that

$$\begin{bmatrix}\boldsymbol{\nu}\\ \boldsymbol{\epsilon}\end{bmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{bmatrix}\boldsymbol{G} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{R}\end{bmatrix}\right), \tag{18.26}$$

where $\boldsymbol{G}$ and $\boldsymbol{R}$ are matrices of known form, but depend on some unknown parameters $\boldsymbol{\theta}$. Note that $\mathbf{Var}(\boldsymbol{Y}) = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R}$.

Before discussing estimation and hypothesis testing methods, we show that the models in the previous sections are special cases of this model. The least squares fixed effects model does not have the random component $\boldsymbol{\nu}$ ($\boldsymbol{G} = \mathbf{0}$) and $\boldsymbol{R} = \boldsymbol{I}\sigma^2$. The random effects model, equation 18.12, is a

special case of the general model, equation 18.25, with $\boldsymbol{X}$ being a column of 1s, $\boldsymbol{\beta}$ being μ, $\boldsymbol{\nu} = (\,\alpha_1 \quad \cdots \quad \alpha_a \quad \beta_1 \quad \cdots \quad \beta_n\,)'$,

$$\boldsymbol{Z} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{I} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} & \boldsymbol{I} \\ \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} & \boldsymbol{I} \end{bmatrix}, \quad \boldsymbol{G} = \begin{bmatrix} \boldsymbol{I}_a\sigma_\alpha^2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_n\sigma_\beta^2 \end{bmatrix},$$

and $\boldsymbol{R} = \boldsymbol{I}_{an}\sigma^2$. (Note that model 18.1 is a special case of model 18.12 where the terms involving β_j are not included.) On the other hand, model 18.14 is a special case of model 18.25 with

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \mathbf{1} & \cdots & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{1} & \mathbf{0} & \cdots & \mathbf{1} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_a \end{pmatrix},$$

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{I} \\ \boldsymbol{I} \\ \vdots \\ \boldsymbol{I} \end{pmatrix}, \qquad \boldsymbol{\nu} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix},$$

$$\boldsymbol{G} = \boldsymbol{I}_n\sigma_\beta^2 \qquad \text{and} \qquad \boldsymbol{R} = \boldsymbol{I}_{an}\sigma^2.$$

Similarly, the split-plot model in equation 18.15 is a special case of model 18.25.

Now, consider the random coefficient model

$$\begin{aligned} \boldsymbol{Y}_i &= \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i \\ &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{X}_i(\boldsymbol{\beta}_i - \boldsymbol{\beta}) + \boldsymbol{\epsilon}_i. \end{aligned} \tag{18.27}$$

Note that model 18.27 is a special case of model 18.25 where

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_n \end{pmatrix}, \qquad \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{X}_n \end{bmatrix},$$

$$\boldsymbol{\nu} = \begin{pmatrix} \beta_1 - \boldsymbol{\beta} \\ \beta_2 - \boldsymbol{\beta} \\ \vdots \\ \beta_n - \boldsymbol{\beta} \end{pmatrix}, \quad \boldsymbol{G} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{bmatrix}$$

and $\boldsymbol{R} = \boldsymbol{I}\sigma^2$. In some cases, where measurements are observed over time on the same experimental unit, it may not be reasonable to assume that $\epsilon_{i1}, \ldots, \epsilon_{it}$ are uncorrelated. Time series correlation functions considered

in Chapter 12 may be more appropriate. For example, Pantula and Pollock (1985) consider the first order autoregressive model for the errors with

Autocorrelated Errors

$$\boldsymbol{R} = \frac{\sigma^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{t-1} \\ \rho & 1 & \rho & \cdots & \rho^{t-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{t-1} & \rho^{t-2} & \rho^{t-3} & \cdots & 1 \end{bmatrix}. \tag{18.28}$$

Estimation of general mixed linear models involves not only the mean parameters $\boldsymbol{\beta}$, but also the variance parameters $\boldsymbol{\theta}$. Also, one may be interested in testing hypotheses not only about $\boldsymbol{\beta}$, but also about $\boldsymbol{\theta}$. For example, in equation 18.12, one may wish to test $H_0; \sigma^2_\beta = 0$, or in the model containing $\boldsymbol{R}$ in equation 18.28 it may be of interest to test $H_0 : \rho = 0$.

Estimation

In model 18.26, $\mathbf{Var}(\boldsymbol{Y}) = \boldsymbol{ZGZ}' + \boldsymbol{R} = \boldsymbol{V}_Y$. Because $\boldsymbol{V}_Y$ is not $\boldsymbol{I}\sigma^2$, ordinary least squares does not necessarily yield the best estimates of $\boldsymbol{\beta}$. The generalized least squares approach that minimizes

$$(\boldsymbol{Y} - \boldsymbol{X\beta})'\boldsymbol{V}_Y^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta}) \tag{18.29}$$

is more appropriate. However, $\boldsymbol{V}_Y$ is not known because it depends on unknown parameters $\boldsymbol{\theta}$. One approach is to find a reasonable estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, then use $\widehat{\boldsymbol{V}}_Y$, obtained by replacing $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}$ in $\boldsymbol{V}_Y$, to minimize 18.29. This is called the **estimated generalized least squares** estimate of $\boldsymbol{\beta}$. Two methods of estimation that are commonly used are maximum likelihood and restricted maximum likelihood estimation.

As discussed in Chapter 12, maximum likelihood estimators are obtained by maximizing the likelihood function with respect to the parameters. For model 18.26, with the assumption of normal errors, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{ML}$ of $\boldsymbol{\theta}$ is obtained by minimizing

$$-2\log\lambda(\boldsymbol{\theta}) = \log|\boldsymbol{V}_Y| + N\log(\widehat{\boldsymbol{\epsilon}}'\boldsymbol{V}_Y^{-1}\widehat{\boldsymbol{\epsilon}}), \tag{18.30}$$

where $\lambda(\boldsymbol{\theta})$ is the likelihood function, and

$$\widehat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}_Y^{-1}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{V}_Y^{-1}\boldsymbol{Y}. \tag{18.31}$$

The maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_{ML}$ of $\boldsymbol{\beta}$ is the same as the estimated generalized least squares estimator of $\boldsymbol{\beta}$ where $\widehat{\boldsymbol{V}}_Y$ is computed at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{ML}$. In most situations, no closed forms exist for $\widehat{\boldsymbol{\theta}}_{ML}$ and $\widehat{\boldsymbol{\beta}}_{ML}$. Iterative methods are used to compute these estimates. For example, PROC MIXED in SAS (SAS Institute Inc., 1997; Littell et al., 1996) uses the Newton–Raphson method to obtain these estimates.

Restricted Maximum Likelihood Estimation

Maximum likelihood estimators of $\boldsymbol{\theta}$, although efficient, generally are biased. A less biased estimator of $\boldsymbol{\theta}$ is obtained by minimizing the function

$$-2\log\lambda_R(\boldsymbol{\theta}) = \log|\boldsymbol{V}_Y| + \log|\boldsymbol{X}'\boldsymbol{V}_Y^{-1}\boldsymbol{X}| + (N-r)\log(\widehat{\boldsymbol{\epsilon}}'\boldsymbol{V}_Y^{-1}\widehat{\boldsymbol{\epsilon}}). \tag{18.32}$$

This estimate is called the **restricted maximum likelihood** estimate (REML) $\widehat{\boldsymbol{\theta}}_{REML}$ of the vector $\boldsymbol{\theta}$. Here $r = rank(\boldsymbol{X})$ and $\widehat{\boldsymbol{\epsilon}}$ is defined in equation 18.31. Note that equation 18.32 differs from equation 18.30 in two respects: N is replaced with $(N-r)$ in the last term; and there is an additional term $\log|\boldsymbol{X}'\boldsymbol{V}_Y^{-1}\boldsymbol{X}|$. Here $\lambda_R(\boldsymbol{\theta})$ is the likelihood function of $(N-r)$ residual variables that have a distribution free of $\boldsymbol{\beta}$. As with the maximum likelihood estimation, estimates of $\boldsymbol{\beta}$ are obtained by minimizing 18.29 with $\boldsymbol{V}_Y$ replaced by $\widehat{\boldsymbol{V}}_Y = \boldsymbol{V}_Y(\widehat{\boldsymbol{\theta}}_{REML})$. Numerical optimization methods are required to obtain $\widehat{\boldsymbol{\theta}}_{REML}$.

Hypothesis Testing

Hypotheses of the form $H_0 : \boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m}$ may be tested using the test statistic

$$\boldsymbol{T} = (\boldsymbol{K}'\widehat{\boldsymbol{\beta}}_{ML} - \boldsymbol{m})' \left[\boldsymbol{K}'\widehat{\mathbf{Var}}(\widehat{\boldsymbol{\beta}}_{ML})\boldsymbol{K}\right]^{-1} (\boldsymbol{K}'\widehat{\boldsymbol{\beta}}_{ML} - \boldsymbol{m}). \qquad (18.33)$$

Under H_0, $\boldsymbol{T}$ is approximately distributed as a chi-square with degrees of freedom rank$(\boldsymbol{K}')$. Similarly, one can compute $\boldsymbol{T}$ using $\widehat{\boldsymbol{\beta}}_{REML}$ in place of $\widehat{\boldsymbol{\beta}}_{ML}$. Iterative algorithms provide an estimate of the variance of the estimators. PROC MIXED prints an estimate of $\mathbf{Var}(\widehat{\boldsymbol{\beta}}_{ML})$.

If the hypothesis $H_0 : \boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m}$ can be used to obtain a reduced model with fewer parameters, then likelihood ratio tests may be used to test $H_0 : \boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m}$. Let $\widehat{\boldsymbol{\theta}}_{ML}^{FULL}$ and $\widehat{\boldsymbol{\theta}}_{ML}^{RED}$ be estimates of the parameters under the full and reduced model, respectively. Under some regularity conditions on the model, it can be shown that

$$[-2\log\lambda(\widehat{\boldsymbol{\theta}}_{ML}^{RED})] - [-2\log\lambda(\widehat{\boldsymbol{\theta}}_{ML}^{FULL})] \sim \chi^2_{r(\boldsymbol{K})}, \qquad (18.34)$$

approximately. Similarly, $\widehat{\boldsymbol{\theta}}_{ML}$ and $\widehat{\boldsymbol{\theta}}_{REML}$ can be used to test hypotheses regarding $\boldsymbol{\theta}$. PROC MIXED reports the values of $\lambda(\widehat{\boldsymbol{\theta}}_{ML})$ and $\lambda(\widehat{\boldsymbol{\theta}}_{REML})$, that can be used to test the relevant hypothesis. It is not appropriate to use equation 18.34 with $\widehat{\boldsymbol{\theta}}_{REML}$ if the hypothesis of interest involves $\boldsymbol{\beta}$. PROC MIXED procedure also provides the AIC and SBC criteria discussed in Chapter 11. These criteria can be used to compare different models. An example of analysis using PROC MIXED is presented in the next chapter.

18.5 Exercises

18.1. In the split-plot example in Section 18.2 it was stated the $\sigma^2(Y_{ijk}) = \sigma^2 + \sigma^2_\delta$. Derive this result using the definition of variance

$$\sigma^2(Y_{ijk}) = \mathcal{E}\{[Y_{ijk} - \mathcal{E}(Y_{ijk})]^2\}$$

and the split-plot model given in the text. Derive the covariance of Y_{ijk} and $Y_{ijk'}$ using the definition

$$\mathrm{Cov}(Y_{ijk}, Y_{ijk'}) = \mathcal{E}\{[Y_{ijk} - \mathcal{E}(Y_{ijk})][Y_{ijk'} - \mathcal{E}(Y_{ijk'})]\}.$$

18.2. You have a completely random experimental design with t treatments and r experimental units per treatment. The response of each experimental unit was determined by measuring the response variable on each of s random samples. This gives the model

$$Y_{ijk} = \mu + \tau_i + \gamma_{ij} + \epsilon_{ijk},$$

where τ_i are fixed treatment effects and γ_{ij} and ϵ_{ijk} are random experimental unit and sampling unit effects with zero means and variances σ_γ^2 and σ^2, respectively.

(a) What is $\sigma^2(Y_{ijk})$? What is $\text{Cov}(Y_{ijk}, Y_{ijk'})$? Show the form of the variance–covariance matrix $\mathbf{Var}(\boldsymbol{Y})$.

(b) What is the form of $\mathbf{Var}(\boldsymbol{Y})$ *if* the mean of all samples within each experimental unit is used as the response variable?

(c) If the $\boldsymbol{Y}_{ijk}$ are used in the analysis using PROC GLM in SAS, how are the standard errors of the treatment means given by GLM computed? Are they correct? If not, how can they be corrected?

(d) If the experimental unit means are used in the analysis, how are the standard errors of the treatment means computed in PROC GLM? Are they correct? What if the numbers of samples per experimental unit are not constant?

(e) Explain the differences in assumptions between doing the analysis with a general linear models program such as PROC GLM and with a program such as PROC MIXED.

18.3. Consider a two-level nested model given by

$$Y_{ijk} = \mu + \alpha_i + \gamma_{ij} + \epsilon_{ijk}, \quad \begin{array}{l} i = 1, \ldots, a; j = 1, \ldots, n; \\ k = 1, \ldots, r, \end{array}$$

where $\alpha_i \sim \text{NID}(0, \sigma_\alpha^2)$, $\gamma_{ij} \sim \text{NID}(0, \sigma_\gamma^2)$, $\epsilon_{ijk} \sim \text{NID}(0, \sigma^2)$, and $\{\alpha_i\}$, $\{\gamma_{ij}\}$, and $\{\epsilon_{ijk}\}$ are independent.

(a) Give the ANOVA table and compute the expected mean squares.

(b) Use the expected mean squares to derive unbiased estimators of the variance components.

(c) Derive the standard errors of the unbiased estimators in (b).

18.4. Consider a two-way cross-classified model given by

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad \begin{array}{l} i = 1, \ldots, a; \\ j = 1, \ldots, n, \end{array}$$

where $\alpha_i \sim \text{NID}(0, \sigma_\alpha^2)$, $\beta_j \sim \text{NID}(0, \sigma_\beta^2)$, $\epsilon_{ij} \sim \text{NID}(0, \sigma^2)$, and $\{\alpha_i\}$, $\{\beta_j\}$, and $\{\epsilon_{ij}\}$ are mutually independent.

(a) Give the ANOVA table and compute the expected mean squares.

(b) Show that the mean squares in the ANOVA table are independent of each other and, when properly normalized, are distributed as chi-square random variables.

(c) Derive the standard errors of unbiased estimators of σ_α^2, σ_β^2, and σ^2 given in equation 18.13.

18.5. Consider a two-way cross-classified model given by

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \quad \begin{array}{l} i = 1, \ldots, a; \\ j = 1, \ldots, n; \\ k = 1, \ldots, r, \end{array}$$

where α_i are fixed, $\beta_j \sim \text{NID}(0, \sigma_\beta^2)$, $\gamma_{ij} \sim \text{NID}(0, \sigma_\gamma^2)$, $\epsilon_{ij} \sim \text{NID}(0, \sigma^2)$, and $\{\beta_j\}$, $\{\gamma_k\}$, and $\{\epsilon_{ijk}\}$ are mutually independent.

(a) Give the ANOVA table and compute the expected mean squares.

(b) Use the expected mean squares to derive the unbiased estimators of the variance components.

(c) Derive the standard errors of the unbiased estimators in (b).

(d) Give the test statistics for testing $H_0 : \alpha_1 = \cdots = \alpha_a$ and $H_0 : \sigma_\beta^2 = 0$.

18.6. Consider a two-way cross-classified model given by

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijk},$$

where $\delta_{ij} = \gamma_{ij} - \overline{\gamma}_{.j}$ and $\{\alpha_i\}$, $\{\beta_j\}$, $\{\gamma_k\}$, and $\{\epsilon_{ijk}\}$ are as defined in Exercise 18.5. That is, here we are assuming that the interaction effects sum to zero ($\sum_{i=1}^{a} \delta_{ij} = 0$) over the index for the fixed effects. Do Parts (a) through (d) in Exercise 18.5.

18.7. Consider a split-plot ANOVA model given by

$$Y_{ijk} = \mu + \alpha_i + \rho_j + \delta_{ij} + \beta_k + \gamma_{ik} + \epsilon_{ijk}, \quad \begin{array}{l} i = 1, \ldots, a; \\ j = 1, \ldots, n; \\ k = 1, \ldots, b, \end{array}$$

where $\rho_j \sim \text{NID}(0, \sigma_\rho^2)$, $\delta_{ij} \sim \text{NID}(0, \sigma_\delta^2)$, $\epsilon_{ijk} \sim \text{NID}(0, \sigma^2)$, and $\{\alpha_i\}$, $\{\beta_k\}$, and $\{\gamma_{ik}\}$ are fixed and $\{\rho_j\}$, $\{\delta_{ij}\}$, and $\{\epsilon_{ijk}\}$ are mutually independent.

(a) Give the ANOVA table and the expected mean squares.

(b) Use the expected mean squares to derive unbiased estimators of the variance components.

(c) Derive the standard errors of the unbiased estimators in (b).

18.8. Consider a one-way analysis of covariance model given by

$$Y_{ij} = \mu + \beta X_{ij} + \alpha_i + \epsilon_{ij}, \quad \begin{matrix} i = 1, \ldots, a; \\ j = 1, \ldots, n, \end{matrix}$$

where $\alpha_i \sim \text{NID}(0, \sigma_\alpha^2)$, $\epsilon_{ij} \sim \text{NID}(0, \sigma^2)$, and $\{\alpha_i\}$, and $\{\epsilon_{ij}\}$ are independent.

(a) Show that this model is a special case of the general mixed linear model in equation 18.25.

(b) Give the "analysis of variance" type estimators of μ, β, σ_α^2, and σ^2.

(c) Give test statistics for testing $H_0 : \beta = 0$ and $H_0 : \sigma_\alpha^2 = 0$.

18.9. Consider the linear combination $Z^2 = \sum_i c_i Z_i^2$, where Z_i^2 are independent chi-square random variables with degrees of freedom f_i. Satterthwaite (1946) approximates the distribution of Z^2 by that of $c\chi_f^2$.

(a) Show that $\mathcal{E}(Z^2) = \sum c_i f_i$ and $\text{Var}(Z^2) = 2\sum c_i^2 f_i$.

(b) Show that $\mathcal{E}[c\chi_f^2] = cf$ and $\text{Var}(c\chi_f^2) = 2c^2 f$.

(c) Equate the mean and variance of Z^2 with that of $c\chi_f^2$ to obtain

$$c = \frac{\sum c_i^2 f_i}{\sum c_i f_i} \quad \text{and} \quad f = \frac{(\sum c_i f_i)^2}{\sum c_i^2 f_i}.$$

These results can be related to Satterthwaite's approximation in equation 18.17 by appropriate definitions of c_i and substitution of observed mean squares for unknown variances.

19

CASE STUDY: ANALYSIS OF UNBALANCED DATA

Chapters 17 and 18 discussed the analysis of unbalanced data and introduced mixed models—models with more than one random effect.

This case study illustrates the analysis of unbalanced data where the model contains more than one random effect. First, the classical analysis of variance approach with a less-than-full-rank effects model is presented. This is followed with an analysis using a program designed to handle mixed models.

The data for this example are from a study of several management systems for corn production (courtesy of Dr. Gar House, North Carolina State University). The set of treatments was intended to be the $2 \times 2 \times 2$ factorial from the 3 factors method of tillage ($TILL$), herbicide application ($HERB$), and additional removal of weeds by hand ($CULT$). The levels of the treatment factors were conventional tillage (CT) and no tillage (NT) for the factor $TILL$, a recommended level of herbicide (H) and no herbicide (NOH) for the factor $HERB$, and hand weeding (C) and no hand weeding (NOC) for the factor $CULT$. The experimental design was a split-plot design with whole plots in a randomized complete block design with 4 blocks. The whole-plot treatments were the 4 $TILL$–$HERB$ treatment combinations; the subplot treatments were the 2 levels of $CULT$. There are a total of $2^3 \times 4 = 32$ experimental units.

Cause of Imbalance

The data are unbalanced because the hand weeding (C) was not done on the no-tillage plots (NT) and, hence, the C level became an NOC treatment

TABLE 19.1. *Yield in bushels per acre for the unbalanced* $2\times2\times2$ *factorial study of cultural practices on corn yield. (Data courtesy of Dr. Gar House, N.C. State University; used with permission.)*

Treatment			*BLOCK*[a]			
TILL	*HERB*	*CULT*	1	2	3	4
CT	*H*	*C*	75.38	92.11	79.59	94.22
CT	*H*	*NOC*	–	39.80	51.54	51.05
CT	*NOH*	*C*	16.59	61.88	68.06	94.50
CT	*NOH*	*NOC*	5.34	25.88	8.57	39.24
NT	*H*	*NOC*	51.47	71.16	45.84	77.06
NT	*H*	*NOC*	55.13	55.13	63.84	74.40
NT	*NOH*	*NOC*	.00	7.31	.00	58.22
NT	*NOH*	*NOC*	.00	.00	.00	31.78

[a]The zeros represent zero yield and not missing values.

for those plots. In addition, the *NOC–H–CT* observation in block 1 is missing. (This observation was dropped for this case study to introduce more imbalance.) The number of observations per treatment are as follows.

	TILL	*CT*		*NT*	
	HERB	*H*	*NOH*	*H*	*NOH*
CULT	*C*	4	4	0	0
	NOC	3	4	8	8

The missing observation in the lower left-hand cell is from Block 1. Otherwise all treatments were equally represented in each block. The data, yield of corn in bushels per acre, are given in Table 19.1.

The linear effects model for the full $2\times2\times2$ factorial in a split-plot arrangement is

$$\begin{aligned} Y_{ijkl} &= \mu + B_i + T_j + H_k + TH_{jk} + \delta_{ijk} + C_l + TC_{jl} + HC_{kl} \\ &\quad + THC_{jkl} + \epsilon_{ijkl}, \end{aligned} \tag{19.1}$$

where B_i, T_j, H_k, and C_l are block, tillage, herbicide, and cultivation effects, respectively, and products designate the respective interaction effects; $i = 1,2,3,4$; $j = k = l = 1,2$. In this study, however, the absence of the C level of the cultivation treatment factor when the tillage treatment is *NT* makes it impossible to estimate any $TILL\times CULT$ or $TILL\times HERB\times CULT$ interactions. Therefore, the TC_{jl} and THC_{jkl} terms are dropped from the model, which is equivalent to imposing the constraints that these effects are zero. These constraints are reflected in the analysis. In this case, the *full* $2\times2\times2$ factorial model gives somewhat larger partial (Type III) SS(*HERB*) than the simpler model, and most of the least

squares means are nonestimable, because the required two- and three-factor interaction effects are nonestimable.

Fixed Effects

The T_j, H_k, and C_l effects and any interaction involving only these effects are regarded as fixed effects in all analyses. On the other hand, there is room for discussion as to whether the block effects B_j should be treated as random with variance σ_B^2 or as fixed effects. Clearly, from an inferential point of view, it is desirable to be able to infer that the observed treatment effects apply to a population of block effects that presumably have been sampled by this study. The disadvantage of treating block effects as random is that the variances of treatment means then will take into account the added uncertainty due to sampling blocks and will include a fraction of the component of variance due to blocks σ_B^2. This is appropriate if we regard a treatment mean as an estimate of the performance of the treatment over repeated sampling of blocks. Almost always, however, our interest is in estimating *differences* among the treatment means, not in the absolute level of performance of any one treatment. The differencing of the means removes from the variance of the mean difference the covariance between two treatment means that arises from the block component of variance σ_B^2. Thus, the standard errors of mean differences cannot be safely approximated from the standard errors of treatment means as we are used to doing in the conventional analyses. On the other hand, treating block effects as fixed in the analysis gives estimated variances of treatment means such that the sum of two variances closely approximates the variance of the difference between the two treatment means. This is simply an expedient to obtain quick estimates of variances of treatment differences. We illustrate this in the mixed model analysis, Section 19.4.

Random Errors

The random error associated with subplots is designated by ϵ_{ijkl} and the whole-plot error is designated by δ_{ijk}. Both are assumed to be normally distributed with variances σ^2 and σ_δ^2, respectively. The presence of several zero yields in the *NT–NOH* treatment (five out of the eight are zero) raises the possibility that assumptions of normality and common variance over all treatments may not be satisfied. The large readings for the fourth block, however, show that the variation for these two treatments is comparable to that for the others. It is likely that, with the wide range in yields observed in this study, the variance will be associated with the mean yield level. For the purpose of demonstrating the analysis of unbalanced data, common variance and normality are assumed. It is left as an exercise for the student to investigate the need for a transformation to stabilize the variance.

Logical Comparisons

Due to the empty cells, the treatments are more appropriately described as the 2×2 factorial for *HERB* and *CULT* conducted at *TILL* = *CT*, and the 2×2 factorial for *TILL* and *HERB* conducted at *CULT* = *NOC*, with two treatments being common to the two sets. From this perspective, it is clear that the *HERB* effect, *CULT* effect, and *HERB* × *CULT* interaction effect can be estimated from the two-way table for *TILL* = *CT*, and the *TILL* effect, *HERB* effect, and *TILL* × *HERB* interaction effect can be

estimated at the *NOC* level of the factor *CULT*. Notice that the *HERB* effect is estimated in both tables. These are logical contrasts one might generate if the analysis were approached from the cell means model point of view (Hocking, 1985). This case study emphasizes the analysis using the effects model.

Outline of the Analysis

In the first analysis, the general linear model analysis for fixed models, assuming for the moment that the δ_{ijk} are fixed effects, is used to partition the sums of squares and obtain the least squares means. [This ignores the covariance structure that exists among the Y_{ijkl} due to observations having common δ_{ijk} (and common B_i if block effects are also random).] Then, the expectations of the mean squares are determined with δ_{ijk} and ϵ_{ijkl} assumed to be random variables. The mean square expectations are used to determine appropriate (approximate) tests of significance and to obtain better approximations of the standard errors of the least squares means. PROC GLM (SAS Institute Inc., 1989b) is used for the analysis with the RANDOM option providing the expectations of the Type III (partial) sums of squares. An interactive matrix language program, [IML (SAS Institute Inc., 1989d)] is used to determine the correct variances of the least squares means.

In the second analysis, estimation of the fixed effects and the variance components for the random effects are considered jointly in an iterative manner. First the fixed effects are estimated with an assumption of a simple variance–covariance structure and then the variance components are estimated from information contained in the residuals. The estimated variance components are used to construct the estimated variance–covariance matrix. In the second iteration, the fixed effects are reestimated using the updated variance–covariance matrix and the variance components are reestimated from the residuals. This iteration process continues until some measure of convergence is met.

19.1 The Analysis Of Variance

The class and model statements for PROC GLM are

```
PROC GLM;CLASS BLOCK TILL HERB CULT;
MODEL Y =BLOCK TILL HERB TILL*HERB
         BLOCK*TILL*HERB CULT HERB*CULT / E E1 E3;
```

The sum of squares for the whole plot error is computed as the three-factor interaction *BLOCK* × *TILL* × *HERB*. The sum of squares for the subplot error appears as the residual sum of squares (labeled ERROR in PROC GLM). The options E, E1, and E3 request the general form of the estimable functions and the specific form of the estimable functions for each of the sequential and partial sums of squares, respectively. (These

TABLE 19.2. *Analysis of variance from the cultural practices study on yield, from PROC GLM, SAS.*

SOURCE	*d.f.*	*Sum of Squares*	*Mean Square*
Model	17	27,760.81	1,632.99
Error	13	1,554.23	119.56

		Sum of Squares	
SOURCE	*d.f.*	*Sequential*	*Partial*
BLOCK	3	5,213.91	4,830.67
TILL	1	1,890.70	109.35
HERB	1	1,1621.47	7,431.66
TILL × *HERB*	1	961.45	692.73
BLOCK × *TILL* × *HERB*	9	2,249.50	1,677.09
CULT	1	5,823.38	5,718.07
HERB × *CULT*	1	0.39	0.39

options generate several pages of results and should not be requested unless needed for understanding the analysis.) The default option in PROC GLM produces the sequential (Type I) and partial (Type III) sums of squares.

Analysis of Variance

The results of this analysis are summarized in Table 19.2. The sum of squares denoted "MODEL" in PROC REG and PROC GLM (SAS Institute Inc., 1989b) is SS(Regr) in the notation of this text. The sum of squares labeled "ERROR" is the residual sum of squares which in the split-plot analysis is an estimate of the subplot error. The bottom portion of Table 19.2 gives the sequential and partial sums of squares for each class of effects in the model. The discussion in Chapter 17 noted that partial (Type III) sums of squares tested the most reasonable hypotheses in most cases of unbalanced data.

Estimating Whole-Plot Error

The degrees of freedom for *BLOCK* × *TILL* × *HERB* and Error sources of variation need explanation. Usually, an interaction sum of squares has degrees of freedom equal to the corresponding product of the degrees of freedom of the component main effects which, in this case, would be three for the BLOCK × *TILL* × *HERB* interaction. However, the two-factor interactions BLOCK × *TILL* and BLOCK × *HERB* are not specified in the model and both are contained in the three-factor interaction. Consequently, the degrees of freedom and sums of squares for these two-factor interactions are absorbed by the three-factor interaction. The interactions of the whole-plot treatments with blocks in the split-plot model are estimates of whole-plot error and this specification of the model is a convenient technique of pooling these sums of squares.

Degrees of Freedom for Subplot Error

The residual sum of squares in the conventional split-plot design would have degrees of freedom determined by the pooling of the sums of squares for the interactions between block effects and subplot treatment and in-

teraction effects. This would give 12 degrees of freedom if the data were balanced. In this case, the residual sum of squares is the pooling of *CULT* × BLOCK, with 3 degrees of freedom, *HERB* × *CULT* × BLOCK, with 3 degrees of freedom, and differences between duplicate plots of the *NT*-*NC* treatment in each level of *HERB* in each of the four blocks, 8 degrees of freedom, minus 1 degree of freedom for the missing plot.

Comparison of Sums of Squares

It is evident from the sums of squares that the data are not balanced since the sequential and partial sums of squares differ. The largest adjustments in the sums of squares are for SS(*TILL*) and for SS(*HERB*). The difference between the simple averages of all plots receiving the *CT* treatment and all plots receiving the *NT* treatment is reflecting primarily the confounded cultivation effect, *C* versus *NOC*. Recall that none of the *NT* treated plots received the *C* cultivation treatment.

General Form of Estimable Functions

The estimable functions explicitly define the differences in the types of sums of squares. The general form for estimable functions for this model and this set of data is given in Table 19.3. The specific forms for the estimable functions for the sequential (Type I) and partial (Type III) sums of squares are given for each source of variation in Tables 19.4 through 19.9.

The number of *free* coefficients in the general form of estimable functions, Table 19.3, for any particular class of effects shows the number of linearly independent contrasts for that class and the number of degrees of freedom for its sum of squares. The free coefficients for any class of effects are those coefficients in that class that are not involved in any other classes of effects except those that "contain" the effects in question. Thus, there are three "free" coefficients for the *BLOCK* effects, L_2, L_3, and L_4; the other coefficient in that set, L_1, is involved in the Intercept and, therefore, is not a free coefficient. L_2, L_3, and L_4 are involved in the *BLOCK* × *TILL* × *HERB* interaction but this is a class of effects that contains the *BLOCK* effects. There are nine free coefficients for the *BLOCK* × *TILL* × *HERB* effects, L_{14} to L_{24} excluding L_{17} and L_{21}, and, hence, nine linearly independent contrasts and nine degrees of freedom for its sums of squares. The remaining coefficients in the *BLOCK* × *TILL* × *HERB* effects must be set equal to zero to remove all other effects. There are no other classes of effects that contain this class of effects.

Specific Estimable Functions for SS

The specific estimable functions in Tables 19.4 through 19.9 are determined from this general form. For example, the sequential (Type I) estimable function for *BLOCK* sum of squares, Table 19.4, is obtained by

1. setting $L_1 = 0$ (to remove the intercept);
2. leaving L_2, L_3, and L_4 general as the free coefficients; and
3. setting all other coefficients to multiples of L_2, L_3, and L_4; $L_6 = L_8 = -.0714\,L_2$, $L_{10} = -.1071\,L_2$, $L_{14} = .1429\,L_2$, and so forth. These nonzero coefficients are functions of the numbers of observations and result from the computations adjusting the *BLOCK* sum of squares for μ. It is important

TABLE 19.3. *The general form of estimable functions for the unbalanced split-plot study.*

Effect		*Coefficients*
Intercept		L_1
BLOCK	1	L_2
	2	L_3
	3	L_4
	4	$L_5 = L_1 - L_2 - L_3 - L_4$
TILL	*CT*	L_6
	NT	$L_7 = L_1 - L_6$
HERB	*H*	L_8
	NOH	$L_9 = L_1 - L_8$
TILL × *HERB*	*CT H*	L_{10}
	CT NOH	$L_{11} = L_6 - L_{10}$
	NT H	$L_{12} = L_8 - L_{10}$
	NT NOH	$L_{13} = L_1 - L_6 - L_8 + L_{10}$
BLOCK × *TILL*	1 *CT H*	L_{14}
× *HERB*	1 *CT NOH*	L_{15}
	1 *NT H*	L_{16}
	1 *NT NOH*	$L_{17} = L_2 - L_{14} - L_{15} - L_{16}$
	2 *CT H*	L_{18}
	2 *CT NOH*	L_{19}
	2 *NT H*	L_{20}
	2 *NT NOH*	$L_{21} = L_3 - L_{18} - L_{19} - L_{20}$
	3 *CT H*	L_{22}
	3 *CT NOH*	L_{23}
	3 *NT H*	L_{24}
	3 *NT NOH*	$L_{25} = L_4 - L_{22} - L_{23} - L_{24}$
	4 *CT H*	$L_{26} = L_{10} - L_{14} - L_{18} - L_{22}$
	4 *CT NOH*	$L_{27} = L_6 - L_{10} - L_{15} - L_{19} - L_{23}$
	4 *NT H*	$L_{28} = L_8 - L_{10} - L_{16} - L_{20} - L_{24}$
	4 *NT HOH*	$L_{29} = L_1 - L_2 - L_3 - L_4 - L_6 - L_8$ $+ L_{10} + L_{14} + L_{15} + L_{16} + L_{18}$ $+ L_{19} + L_{20} + L_{22} + L_{23} + L_{24}$
CULT	*C*	L_{30}
	NOC	$L_{31} = L_1 - L_{30}$
HERB × *CULT*	*H C*	L_{32}
	H NOC	$L_{33} = L_8 - L_{32}$
	NOH C	$L_{34} = L_{30} - L_{32}$
	NOH NOC	$L_{35} = L_1 - L_8 - L_{30} + L_{32}$

TABLE 19.4. *The estimable functions for BLOCK sums of squares.*

Effect		*Coefficients*	
		Sequential	*Partial*
Intercept		0	0
BLOCK	1	L_2	L_2
	2	L_3	L_3
	3	L_4	L_4
	4	$-L_2 - L_3 - L_4$	$-L_2 - L_3 - L_4$
TILL	*CT*	$-.0714\,L_2$	0
	NT	$.0714\,L_2$	0
HERB	*H*	$-.0714\,L_2$	0
	NOH	$.0714\,L_2$	0
TILL	*CT H*	$-.1071\,L_2$	0
× *HERB*	*CT NOH*	$.0357\,L_2$	0
	NT H	$.0357\,L_2$	0
	NT NOH	$.0357\,L_2$	0
BLOCK	1 *CT H*	$.1429\,L_2$	$.25\,L_2$
× *HERB*	1 *CT NOH*	$.2857\,L_2$	$.25\,L_2$
× *TILL*	1 *NT H*	$.2857\,L_2$	$.25\,L_2$
	1 *NT NOH*	$.2857\,L_2$	$.25\,L2$
	2 *CT H*	$.25\,L_3$	$.25\,L_3$
	2 *CT NOH*	$.25\,L_3$	$.25\,L_3$
	2 *NT H*	$.25\,L_3$	$.25\,L_3$
	2 *NT NOH*	$.25\,L_3$	$.25\,L_3$
	3 *CT H*	$.25\,L_4$	$.25\,L_4$
	3 *CT NOH*	$.25\,L_4$	$.25\,L_4$
	3 *NT H*	$.25\,L_4$	$.25\,L_4$
	3 *NT NOH*	$.25\,L_4$	$.25\,L_4$
	4 *CT H*	$-.25(L_2 + L_3 + L_4)$	$-.25(L_2 + L_3 + L_4)$
	4 *CT NOH*	$-.25(L_2 + L_3 + L_4)$	$-.25(L_2 + L_3 + L_4)$
	4 *NT H*	$-.25(L_2 + L_3 + L_4)$	$-.25(L_2 + L_3 + L_4)$
	4 *NT HOH*	$-.25(L_2 + L_3 + L_4)$	$-.25(L_2 + L_3 + L_4)$
CULT	*C*	$.0375\,L_2$	0
	NOC	$-.0375\,L_2$	0
HERB	*H C*	$.0179\,L_2$	0
× *CULT*	*H NOC*	$-.0893\,L_2$	0
	NOH C	$.0179\,L_2$	0
	NOH NOC	$.0536\,L_2$	0

TABLE 19.5. *The estimable functions for TILL sums of squares.*

Effect		*Coefficients*	
		Sequential	*Partial*
Intercept		0	0
BLOCK	1	0	0
	2	0	0
	3	0	0
	4	0	0
TILL	*CT*	L_6	L_6
	NT	$-L_6$	$-L_6$
HERB	*H*	$-.037\,L_6$	0
	NOH	$.037\,L_6$	0
TILL × *HERB*	*CT H*	$.463\,L_6$	$.5\,L_6$
	CT NOH	$.537\,L_6$	$.5\,L_6$
	NT H	$-.5\,L_6$	$-.5\,L_6$
	NT NOH	$-.5\,L_6$	$-.5\,L_6$
BLOCK × *TILL* × *HERB*	1*CT H*	$.0741\,L_6$	$.125\,L_6$
	1*CT NOH*	$.1481\,L_6$	$.125\,L_6$
	1*NT H*	$-.1111\,L_6$	$-.125\,L_6$
	1*NT NOH*	$-.1111\,L_6$	$-.125\,L_6$
	2*CT H*	$.1296\,L_6$	$.125\,L_6$
	2*CT NOH*	$.1296\,L_6$	$.125\,L_6$
	2*NT H*	$-.1296\,L_6$	$-.125\,L_6$
	2*NT NOH*	$-.1296\,L_6$	$-.125\,L_6$
	3*CT H*	$.1296\,L_6$	$.125\,L_6$
	3*CT NOH*	$.1296\,L_6$	$.125\,L_6$
	3*NT H*	$-.1296\,L_6$	$-.125\,L_6$
	3*NT NOH*	$-.1296\,L_6$	$-.125\,L_6$
	4*CT H*	$.1296\,L_6$	$.125\,L_6$
	4*CT NOH*	$.1296\,L_6$	$.125\,L_6$
	4*NT H*	$-.1296\,L_6$	$-.125\,L_6$
	4*NT HOH*	$-.1296\,L_6$	$-.125\,L_6$
CULT	*C*	$.537\,L_6$	0
	NOC	$-.537\,L_6$	0
HERB × *CULT*	*H C*	$.2685\,L_6$	0
	H NOC	$-.3056\,L_6$	0
	NOH C	$.2685\,L_6$	0
	NOH NOC	$-.2315\,L_6$	0

TABLE 19.6. *The estimable functions for HERB sums of squares.*

Effect		*Coefficients*	
		Sequential	*Partial*
Intercept		0	0
BLOCK	1	0	0
	2	0	0
	3	0	0
	4	0	0
TILL	*CT*	0	0
	NT	0	0
HERB	*H*	L_8	L_8
	NOH	$-L_8$	$-L_8$
TILL × *HERB*	*CT H*	$.4808\,L_8$	$.5\,L_8$
	CT NOH	$-.4808\,L_8$	$-.5\,L_8$
	NT H	$.5192\,L_8$	$.5\,L_8$
	NT NOH	$-.5192\,L_8$	$-.5\,L_8$
BLOCK × *TILL*	1*CT H*	$.0769\,L_8$	$.125\,L_8$
× *HERB*	1*CT NOH*	$-.1058\,L_8$	$-.125\,L_8$
	1*NT H*	$.1442\,L_8$	$.125\,L_8$
	1*NT NOH*	$-.1154\,L_8$	$-.125\,L_8$
	2*CT H*	$.1346\,L_8$	$.125\,L_8$
	2*CT NOH*	$-.125\,L_8$	$-.125\,L_8$
	2*NT H*	$.125\,L_8$	$.125\,L_8$
	2*NT NOH*	$-.1346\,L_8$	$-.125\,L_8$
	3*CT H*	$.1346\,L_8$	$.125\,L_8$
	3*CT NOH*	$-.125\,L_8$	$-.125\,L_8$
	3*NT H*	$.125\,L_8$	$.125\,L_8$
	3*NT NOH*	$-.1346\,L_8$	$-.125\,L_8$
	4*CT H*	$.1346\,L_8$	$.125\,L_8$
	4*CT NOH*	$-.125\,L_8$	$-.125\,L_8$
	4*NT H*	$.125\,L_8$	$.125\,L_8$
	4*NT HOH*	$-.1346\,L_8$	$-.125\,L_8$
CULT	*C*	$.0385\,L_8$	0
	NOC	$-.0385\,L_8$	0
HERB × *CULT*	*H C*	$.2788\,L_8$	$.5\,L_8$
	H NOC	$.7212\,L_8$	$.5\,L_8$
	NOH C	$-.2404\,L_8$	$-.5\,L_8$
	NOH NOC	$-.7596\,L_8$	$-.5\,L_8$

TABLE 19.7. *The estimable functions for TILL*HERB sums of squares.*

Effect		*Coefficients*	
		Sequential	*Partial*
Intercept		0	0
BLOCK	1	0	0
	2	0	0
	3	0	0
	4	0	0
TILL	*CT*	0	0
	NT	0	0
HERB	*H*	0	0
	NOH	0	0
TILL × *HERB*	*CT H*	L_{10}	L_{10}
	CT NOH	$-L_{10}$	$-L_{10}$
	NT H	$-L_{10}$	$-L_{10}$
	NT NOH	L_{10}	L_{10}
BLOCK × *TILL* × *HERB*	1*CT H*	$.16\,L_{10}$	$.25\,L_{10}$
	1*CT NOH*	$-.22\,L_{10}$	$-.25\,L_{10}$
	1*NT H*	$-.22\,L_{10}$	$-.25\,L_{10}$
	1*NT NOH*	$.28\,L_{10}$	$.25\,L_{10}$
	2*CT H*	$.28\,L_{10}$	$.25\,L_{10}$
	2*CT NOH*	$-.26\,L_{10}$	$-.25\,L_{10}$
	2*NT H*	$-.26\,L_{10}$	$-.25\,L_{10}$
	2*NT NOH*	$.24\,L_{10}$	$.25\,L_{10}$
	3*CT H*	$.28\,L_{10}$	$.25\,L_{10}$
	3*CT NOH*	$-.26\,L_{10}$	$-.25\,L_{10}$
	3*NT H*	$-.26\,L_{10}$	$-.25\,L_{10}$
	3*NT NOH*	$.24\,L_{10}$	$.25\,L_{10}$
	4*CT H*	$.28\,L_{10}$	$.25\,L_{10}$
	4*CT NOH*	$-.26\,L_{10}$	$-.25\,L_{10}$
	4*NT H*	$-.26\,L_{10}$	$-.25\,L_{10}$
	4*NT HOH*	$.24\,L_{10}$	$.25\,L_{10}$
CULT	*C*	$.08\,L_{10}$	0
	NOC	$-.08\,L_{10}$	0
HERB × *CULT*	*H C*	$.58\,L_{10}$	0
	H NOC	$-.58\,L_{10}$	0
	NOH C	$-.5\,L_{10}$	0
	NOH NOC	$.5\,L_{10}$	0

TABLE 19.8. *The estimable functions for BLOCK*TILL*HERB sums of squares.*

Effect		*Coefficients*	
		Sequential	*Partial*
Intercept		0	0
BLOCK	1	0	0
	2	0	0
	3	0	0
	4	0	0
TILL	*CT*	0	0
	NT	0	0
HERB	*H*	0	0
	NOH	0	0
TILL	*CT H*	0	0
× *HERB*	*CT NOH*	0	0
	NT H	0	0
	NT NOH	0	0
BLOCK	1*CT H*	L_{14}	L_{14}
× *TILL*	1*CT NOH*	L_{15}	L_{15}
× *HERB*	1*NT H*	L_{16}	L_{16}
	1*NT NOH*	$-L_{14} - L_{15} - L_{16}$	$-L_{14} - L_{15} - L_{16}$
	2*CT H*	L_{18}	L_{18}
	2*CT NOH*	L_{19}	L_{19}
	2*NT H*	L_{20}	L_{20}
	2*NT NOH*	$-L_{18} - L_{19} - L_{20}$	$-L_{18} - L_{19} - L_{20}$
	3*CT H*	L_{22}	L_{22}
	3*CT NOH*	L_{23}	L_{23}
	3*NT H*	L_{24}	L_{24}
	3*NT NOH*	$-L_{22} - L_{23} - L_{24}$	$-L_{22} - L_{23} - L_{24}$
	4*CT H*	$-L_{14} - L_{18} - L_{22}$	$-L_{14} - L_{18} - L_{22}$
	4*CT NOH*	$-L_{15} - L_{19} - L_{23}$	$-L_{15} - L_{19} - L_{23}$
	4*NT H*	$-L_{16} - L_{20} - L_{24}$	$-L_{16} - L_{20} - L_{24}$
	4*NT NOH*	$L_{14} + L_{15} + L_{16} + L_{18} + L_{19} + L_{20} + L_{22} + L_{23} + L_{24}$	$L_{14} + L_{15} + L_{16} + L_{18} + L_{19} + L_{20} + L_{22} + L_{23} + L_{24}$
CULT	*C*	$.5\,L_{14}$	0
	NOC	$-.5\,L_{14}$	0
HERB	*H C*	$.5\,L_{14}$	0
× *CULT*	*H NOC*	$-.5\,L_{14}$	0
	NOH C	0	0
	NOH NOC	0	0

TABLE 19.9. *The estimable functions for CULT sums of squares.*

Effect		*Coefficients*	
		Sequential	*Partial*
Intercept		0	0
BLOCK	1	0	0
	2	0	0
	3	0	0
	4	0	0
TILL	*CT*	0	0
	NT	0	0
HERB	*H*	0	0
	NOH	0	0
TILL × *HERB*	*CT H*	0	0
	CT NOH	0	0
	NT H	0	0
	NT NOH	0	0
BLOCK × *TILL*	1*CT H*	0	0
× *HERB*	1*CT NOH*	0	0
	1*NT H*	0	0
	1*NT NOH*	0	0
	2*CT H*	0	0
	2*CT NOH*	0	0
	2*NT H*	0	0
	2*NT NOH*	0	0
	3*CT H*	0	0
	3*CT NOH*	0	0
	3*NT H*	0	0
	3*NT NOH*	0	0
	4*CT H*	0	0
	4*CT NOH*	0	0
	4*NT H*	0	0
	4*NT NOH*	0	0
CULT	*C*	L_{30}	L_{30}
	NOC	$-L_{30}$	$-L_{30}$
HERB × *CULT*	*H C*	$.4286\,L_{30}$	$.5\,L_{30}$
	H NOC	$-.4286\,L_{30}$	$-.5\,L_{30}$
	NOH C	$.5714\,L_{30}$	$.5\,L_{30}$
	NOH NOC	$-.5714\,L_{30}$	$-.5\,L_{30}$

to note which coefficients are nonzero and that they are functions of the numbers of observations.

The partial (Type III) estimable functions for the *BLOCK* sum of squares are obtained by

1. setting $L_1 = L_6 = L_8 = L_{10} = L_{30} = L_{32} = 0$ (to remove all other effects that do not contain *BLOCK* effects);
2. leaving L_2, L_3, and L_4 general; and
3. setting all other coefficients, L_{14} to L_{24}, to multiples of L_2, L_3, and L_4. The multiples for the partial (Type III) estimable functions are chosen to satisfy the orthogonality property.

Nonzero coefficients for *TILL*, *HERB*, *TILL* × *HERB*, *CULT*, and *HERB* × *CULT* effects in the estimable function for the sequential (Type I) sum of squares for *BLOCK* (Table 19.4) result from the fact that they are *sequential*. The sequential sums of squares for a particular effect are adjusted only for effects that precede it in the model. Consequently, the *BLOCK* sum of squares, being first in the model statement, is adjusted only for μ. Clearly, the sequential *BLOCK* sum of squares is confounded with all other effects in the model. On the other hand, the partial (Type III) *BLOCK* sum of squares has been adjusted for all effects that do not contain the *BLOCK* effects by setting L_6, L_8, L_{10}, L_{30}, and L_{32} equal to zero. The multiples of L_2, L_3, and L_4 are chosen to satisfy the orthogonality property for the higher-order interaction effects that contain *BLOCK* effects.

Each of Tables 19.4 through 19.9 contains the estimable functions for the sequential and partial sums of squares for one source of variation. The sequence of the tables corresponds to the order in which the class variables were entered into the model statement. Thus, comparison of the sequential estimable functions from table to table shows the sequential nature of these sums of squares. The sequential estimable function for *BLOCK* sum of squares, Table 19.4, contains nonzero coefficients for all effects other than the intercept; it is confounded with all other effects. The sequential estimable function for *TILL*, Table 19.5, has zero coefficients for *BLOCK* effects but nonzero coefficients for all succeeding classes of effects; this sum of squares is adjusted for *BLOCK* effects but is confounded with all classes of effects that follow *TILL* in the model statement. Inspection of the remaining tables shows that this pattern continues for successive terms in the model. The estimable function for the last term in the model, *HERB* × *CULT*, is the same for all types of sums of squares and is not given in a separate table. Being the last term in the model, the sequential *HERB* × *CULT* estimable function has zero coefficients for all other effects.

Summary

In summary, the sequential (Type I) estimable function for each class of effects is adjusted only for other classes of effects that precede it in the

model statement and, consequently, is confounded with all classes of effects that follow it in the model; the coefficients on the effects for which it is not adjusted are dependent on the cell numbers and the coefficients do not have the orthogonality property. This confounding of different classes of effects and the dependence of the coefficients on the numbers of observations makes the sequential sums of squares inappropriate for testing hypotheses in this example. In contrast, the partial (Type III) hypotheses have nonzero coefficients only on higher-order interaction effects that "contain" the effects being tested and possess the orthogonality property. These hypotheses are the same as those being tested by the analysis of variance sums of squares in balanced data. Thus, the partial sums of squares are appropriate in this example for testing hypotheses that various classes of effects are zero.

19.2 Mean Square Expectations and Choice of Errors

Before turning to interpretation of the analysis of variance, the analysis based on a fixed effects model must be reconciled with the fact that the correct model contains two random effects, the whole-plot effect δ_{ijk} and the subplot effect ϵ_{ijkl}. With balanced data, the whole-plot error is estimated with the interaction mean square between blocks and the whole-plot treatments, in this case, the $BLOCK \times TILL \times HERB$ mean square. With unbalanced data, the expectations of the mean squares must be used to determine proper error terms. The RANDOM statement in PROC GLM was used to obtain these expectations. The expectations of the partial (Type III) mean squares in the analysis are given in Table 19.10. Also, these expectations may be obtained using the formulae for expectations of quadratic forms. The residual mean square always has expectation σ^2 where σ^2 is the true variance of the unique random element in the model, ϵ_{ijkl} in this case. Thus, $s^2 = 119.56$ with 13 degrees of freedom is the estimate of the subplot error variance (Table 19.2). Equating the expectations of ERROR A and ERROR B to their partial (Type III) sums of squares (Table 19.2) gives two equations with which the components of variance can be estimated. These equations give $\widehat{\sigma}^2 = 119.56$ and $\widehat{\sigma}^2_\delta = 35.06$.

Variance Component Estimates

The only random component in the expectations of $CULT$ and $HERB \times CULT$ mean squares is σ^2. This confirms that the subplot error, ERROR B, is the appropriate error term for testing hypotheses about $CULT$ and $HERB \times CULT$ effects, the subplot treatment comparisons, as is the case with balanced data.

Expectations Involving Only σ^2

The variance ratio for $HERB \times CULT$ interaction is less than unity indicating that the herbicide effects and the cultivation effects are additive. The variance ratio for $CULT$ effects is $F = 5,718.07/119.56 = 47.8$, which

Tests Using Error B

TABLE 19.10. *Expectations of partial (Type III) mean squares for the split-plot experiment using the RANDOM option in PROC GLM.*

Mean Square	*Expectation of Mean Square*[a]
BLOCK	$\sigma^2 + 1.8667\,\sigma^2_\delta + Q(BLOCK)$
TILL	$\sigma^2 + 1.0909\,\sigma^2_\delta + Q(T, T\times H)$
HERB	$\sigma^2 + 1.3333\,\sigma^2_\delta + Q(H, T\times H, H\times C)$
TILL × *HERB*	$\sigma^2 + 1.0909\,\sigma^2_\delta + Q(T\times H)$
ERROR A[b]	$\sigma^2 + 1.9048\,\sigma^2_\delta$
CULT	$\sigma^2 + Q(C, H\times C)$
HERB × *CULT*	$\sigma^2 + Q(H\times C)$
ERROR B	σ^2

[a] $Q(\cdot)$ is a quadratic function of the effects in parentheses. T = Till, H = HERB, and C = CULT.

[b] ERROR A = MS(*BLOCK* × *TILL* × *HERB*).

is highly significant; that is, the average difference in yield between the *CULT* treatments is too large to be explained by random variation. The absence of an *HERB* × *CULT* interaction indicates that this effect of hand weeding is consistent over both herbicide levels. Recall that the information on the *HERB* × *CULT* interaction effects and the *CULT* effects comes only from data on conventional tillage, *TILL* = *CT*. These conclusions can be extended to the *TILL* = *NT* treatment only if there is no interaction of these effects with *TILL*. This was implicitly assumed when the *TILL* × *CULT* and *TILL* × *CULT* × *HERB* interaction effects were dropped from the model, but these assumptions cannot be tested with these data.

Expectations Involving Both σ^2 and σ^2_δ

The random components in the expectations of the remaining mean squares are not the same as in balanced data. If the data were balanced, the expectation of the mean square for the whole-plot error (ERROR A) would contain $\sigma^2 + k\sigma^2_\delta$, where k is the number of subplots per whole-plot. The expectations of all whole-plot treatment mean squares also would contain $\sigma^2 + k\sigma^2_\delta$, plus a quadratic function of fixed effects, so that ERROR A would be the appropriate error mean square for all tests of whole-plot treatment effects. With this unbalanced example, the coefficients on σ^2_δ for *TILL*, *HERB*, and *TILL* × *HERB* differ from that for ERROR A (Table 19.10). Thus, ERROR A is not the appropriate error for tests of significance. (If the coefficients were very similar, one might be content to use ERROR A in approximate tests of hypotheses about whole-plot treatment effects. In this case, the coefficients are quite different, 1.0909 versus 1.9048, so that tests using ERROR A could be seriously biased unless σ^2_δ were close to zero.)

Constructed Error Mean Squares

When the coefficients are more than trivially different, it is better to construct for each F-test an error mean square that has the same expectation for the random elements as the numerator mean square. The constructed error mean square for testing *TILL* and *TILL* × *HERB* is that linear function of ERROR A (E_a) and ERROR B (E_b) that has expectation $\sigma^2 + 1.0909\sigma_\delta^2$. Thus,

$$E' = \frac{1.0909}{1.9048} E_a + \left(1 - \frac{1.0909}{1.9048}\right) E_b = 157.81.$$

The degrees of freedom for this estimate of error are approximated with Satterthwaite's approximation as

$$\begin{aligned} f' &= \frac{(\sum a_i \mathrm{MS}_i)^2}{\sum (a_i^2 \mathrm{MS}_i^2 / f_i)} \\ &= \frac{(157.8080)^2}{\left(\frac{1.0909}{1.9048}\right)^2 \frac{(186.34)^2}{9} + \left(1 - \frac{1.0909}{1.9048}\right)^2 \frac{(119.56)^2}{13}} \\ &= 16.98 \quad \text{or 17 degrees of freedom,} \end{aligned}$$

where a_i is the coefficient of MS_i and f_i is the degrees of freedom for MS_i.

With this constructed error term, the variance ratio for *TILL* × *HERB* is $F' = 4.39$ which just misses being significant at $\alpha = .05$, $F_{(.05;1,17)} = 4.45$. If one adheres strictly to the chosen α, the interaction effect between *TILL* and *HERB* would be declared unimportant. However, one would probably report the herbicide effects at each tillage level and then point out that the differences were not quite significant (at $\alpha = .05$). The variance ratio for the test of *TILL* effects averaged over the levels of *HERB* is $F' = .69$ which is not significant. This does not imply that the tillage effects are negligible *within* each herbicide treatment.

The constructed error term for testing *HERB* effects is $E' = 166.31$ with approximate degrees of freedom $f' = 14$. The variance ratio for this test is $F' = 44.68$, far exceeding the critical level for $\alpha = .01$. Unlike the *TILL* and *CULT* main effects, information on the *HERB* effect comes from both two-way tables. This *average* herbicide effect, averaged over *TILL* and *CULT* treatments, is significantly different from zero but the (nearly) significant *TILL* × *HERB* interaction suggests that the herbicide effect may not be the same for the two tillage levels.

Summary of Analysis of Variance

To summarize the results of the analysis of variance, the near significance of the interaction between *TILL* and *HERB* suggests that the yield response to herbicide depends on whether conventional tillage or no tillage is used. The *average* herbicide effect is significant but is somewhat difficult to interpret since it is an average from the 2 two-way factorials, one of which shows an interaction. The average cultivation effect is different from zero and its effects are relatively constant over levels of *HERB* as observed under the *TILL* = CT treatment. These results suggest that the effects of

the treatments can be summarized in the two-way table of *TILL* × *HERB* means and the marginal means for *CULT*.

19.3 Least Squares Means and Standard Errors

The least squares means are estimated as the linear functions of $\boldsymbol{\beta}^0$ that have the same expectations as the corresponding means in balanced data, the population marginal means. In the tillage–herbicide–cultivation study, there are no empty cells for any of the effects defined in the model, so that all least squares means are estimable. (It was recognized at the beginning of the case study that there was no information in the data on two of the interactions, and their effects were dropped from the model. If these effects had been retained in the model, many of the least squares means would not have been estimable.)

Expectations of Least Squares Means

The expectations of the least squares marginal means for the herbicide treatments *HERB* and the cultivation treatments *CULT* are given in Table 19.11. (For comparison, the expectation of the *unadjusted* mean for the *C* level of *CULT* is also given. The differences in coefficients between the last column and the third column show the nature of the confounding in this unadjusted mean. The coefficient of 1.0 on the *CT* effect of the *TILL* factor shows that the unadjusted *C* mean is completely confounded with the *CT* effect.) The estimable functions for the two-way table of *TILL* × *HERB* means are given in Table 19.12. The coefficients in each column of Tables 19.11 and 19.12 define the linear functions of $\boldsymbol{\beta}^0$ that must be computed to obtain the least squares mean.

Estimates of Means and Interpretations

The least squares marginal means for all three treatment factors and the two-way *TILL* × *HERB* treatment means are given in the first column of data in Table 19.13. The *unadjusted* treatment means are given in the last column of the table for comparison only. All interpretations should be based on the least squares means. The tests of significance have indicated that the *CT* and the *NT* means for tillage are not different. (The unadjusted tillage means, on the other hand, were very different — 53.58 versus 36.96. The adjustment is primarily on the *NT* treatment mean and is reflecting its total confounding with the *NOC* treatment. The *NT* treatment did not involve any plots on which there was additional hand weeding.)

The difference between the herbicide treatment means is significant; the presence of herbicide more than doubled yield in this experiment. Similarly, additional hand weeding *C* doubled yield. It must not be overlooked, however, that there was no measure of the interaction between *CULT* and *TILL* since hand weeding *C* was done only on the no-tillage *NT* plots. Thus, it would be an extrapolation to imply that hand weeding would have this same effect on the conventional-tillage plots.

TABLE 19.11. *The estimable functions for the least squares means for levels of herbicide (HERB) and cultivation (CULT). The unadjusted C mean is given for comparison.*

Effect		*CULT*		*HERB*		*Unadj.*
		C	*NOC*	*H*	*NOH*	*C Mean*
Intercept		1	1	1	1	1
BLOCK	1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
	2	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
	3	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
	4	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
TILL	*CT*	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1
	NT	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0
HERB	*H*	$\frac{1}{2}$	$\frac{1}{2}$	1	0	$\frac{1}{2}$
	NOH	$\frac{1}{2}$	$\frac{1}{2}$	0	1	$\frac{1}{2}$
TILL × *HERB*	*CT H*	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	0	$\frac{1}{2}$
	CT NOH	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{2}$	$\frac{1}{2}$
	NT H	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	0	0
	NT NOH	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{2}$	0
BLOCK × *TILL*	1*CT H*	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	0	$\frac{1}{8}$
× *HERB*	1*CT NOH*	$\frac{1}{16}$	$\frac{1}{16}$	0	$\frac{1}{8}$	$\frac{1}{8}$
	1*NT H*	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	0	0
	1*NT NOH*	$\frac{1}{16}$	$\frac{1}{16}$	0	$\frac{1}{8}$	0
	2*CT H*	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	0	$\frac{1}{8}$
	2*CT NOH*	$\frac{1}{16}$	$\frac{1}{16}$	0	$\frac{1}{8}$	$\frac{1}{8}$
	2*NT H*	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	0	0
	2*NT NOH*	$\frac{1}{16}$	$\frac{1}{16}$	0	$\frac{1}{8}$	0
	3*CT H*	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	0	$\frac{1}{8}$
	3*CT NOH*	$\frac{1}{16}$	$\frac{1}{16}$	0	$\frac{1}{8}$	$\frac{1}{8}$
	3*NT H*	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	0	0
	3*NT NOH*	$\frac{1}{16}$	$\frac{1}{16}$	0	$\frac{1}{8}$	0
	4*CT H*	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	0	$\frac{1}{8}$
	4*CT NOH*	$\frac{1}{16}$	$\frac{1}{16}$	0	$\frac{1}{8}$	$\frac{1}{8}$
	4*NT H*	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	0	0
	4*NT HOH*	$\frac{1}{16}$	$\frac{1}{16}$	0	$\frac{1}{8}$	0
CULT	*C*	1	0	$\frac{1}{2}$	$\frac{1}{2}$	1
	NOC	0	1	$\frac{1}{2}$	$\frac{1}{2}$	0
HERB × *CULT*	*H C*	$\frac{1}{2}$	0	$\frac{1}{2}$	0	$\frac{1}{2}$
	H NOC	0	$\frac{1}{2}$	$\frac{1}{2}$	0	0
	NOH C	$\frac{1}{2}$	0	0	$\frac{1}{2}$	$\frac{1}{2}$
	NOH NOC	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0

TABLE 19.12. *The estimable functions for the two-way table of least squares means for levels of tillage (TILL) and herbicide (HERB).*

Effect		*CT H*	*CT NOH*	*NT H*	*NT NOH*
Intercept		1	1	1	1
BLOCK	1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
	2	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
	3	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
	4	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
TILL	*CT*	1	1	0	0
	NT	0	0	1	1
HERB	*H*	1	0	1	0
	NOH	0	1	0	1
TILL × *HERB*	*CT H*	1	0	0	0
	CT NOH	0	1	0	0
	NT H	0	0	1	0
	NT NOH	0	0	0	1
BLOCK × *TILL*	1*CT H*	$\frac{1}{4}$	0	0	0
× *HERB*	1*CT NOH*	0	$\frac{1}{4}$	0	0
	1*NT H*	0	0	$\frac{1}{4}$	0
	1*NT NOH*	0	0	0	$\frac{1}{4}$
	2*CT H*	$\frac{1}{4}$	0	0	0
	2*CT NOH*	0	$\frac{1}{4}$	0	0
	2*NT H*	0	0	$\frac{1}{4}$	0
	2*NT NOH*	0	0	0	$\frac{1}{4}$
	3*CT H*	$\frac{1}{4}$	0	0	0
	3*CT NOH*	0	$\frac{1}{4}$	0	0
	3*NT H*	0	0	$\frac{1}{4}$	0
	3*NT NOH*	0	0	0	$\frac{1}{4}$
	4*CT H*	$\frac{1}{4}$	0	0	0
	4*CT NOH*	0	$\frac{1}{4}$	0	0
	4*NT H*	0	0	$\frac{1}{4}$	0
	4*NT NOH*	0	0	0	$\frac{1}{4}$
CULT	*C*	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
	NOC	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
HERB × *CULT*	*H C*	$\frac{1}{2}$	0	$\frac{1}{2}$	0
	H NOC	$\frac{1}{2}$	0	$\frac{1}{2}$	0
	NOH C	0	$\frac{1}{2}$	0	$\frac{1}{2}$
	NOH NOC	0	$\frac{1}{2}$	0	$\frac{1}{2}$

TABLE 19.13. *Least squares means, standard errors of least squares means as given by GLM, GLM standard errors adjusted for the mean square expectations, "exact" standard errors of least squares means, standard errors of mean differences, and unadjusted treatment means.*

	Least				*S.E.*	
	Squares	*Standard Errors*			*Mean*	*Unadj.*
Treatment	*Means*	*GLM*	*GLM ADJ*	*EXACT*	*Diff.*[a]	*Means*
TILL:						
CT	52.37	2.95	3.39	3.62		53.58
NT	57.38	4.02	4.62	4.54	6.01	36.96
HERB:						
H	73.54	3.35	3.95	3.95		65.18
NOH	36.21	3.35	3.95	3.95	5.58	26.09
CULT:						
C	75.29	4.67	4.67	4.89		72.79
NOC	34.46	2.62	2.62	3.01	5.91	35.34
*TILL*HERB:*						
CT H	64.74	4.46	5.12	5.35		69.10
CT NOH	40.01	3.87	4.47	4.87	7.23	40.01
NT H	82.34	5.91	6.79	6.62		61.75
NT NOH	32.41	5.47	6.29	6.22	9.07	12.16

[a]Standard errors of differences between adjacent pairs of treatment means using the EXACT computations.

The two-way *TILL* × *HERB* means are given because the interaction was close to significance at $\alpha = .05$. The pattern of the means in this two-way table suggests that no tillage *NT* is better than conventional tillage *CT* when herbicide is being used but is slightly worse if no herbicide is used. The herbicide effect is positive under both types of tillage but the difference is much larger in the *NT* treatment. It appears from this study that it is better to use herbicide and, if herbicide is to be used, to also use the no-tillage method.

Standard Errors

Columns 3 through 5 in Table 19.13 give standard errors of the least squares means computed according to different rules. The first column of standard errors, labeled "GLM," are as given by PROC GLM. The GLM standard errors are computed as if $\mathbf{Var}(\boldsymbol{Y}) = \boldsymbol{I}\sigma^2$ and the residual mean square, ERROR B = 119.56, is used as the estimate of σ^2.

Adjusted Standard Errors

The second column of standard errors, labeled "GLM ADJ," has been computed from the "GLM" standard errors by multiplying each by the square root of the ratio of the constructed error mean square to ERROR B. This approach still assumes $\mathbf{Var}(\boldsymbol{Y}) = \boldsymbol{I}\sigma^2$ but replaces σ^2 with an average variance of the means in that class of means; the average is taken from the expectations of the partial (Type III) mean squares given by the RANDOM option. The estimates of the error components of variance are computed from the PROC GLM partial (Type III) sums of squares. For example, the GLM standard errors for the *TILL* means have been multiplied by $\sqrt{157.8/119.6} = 1.149$ to obtain GLM ADJ. The 157.8 is the error mean square constructed as the appropriate denominator for the F-test of tillage effects.

"Exact" Standard Errors

The third column of standard errors, labeled "EXACT," uses the estimated variance–covariance matrix for $\boldsymbol{Y}$, which takes into account the covariances of observations due to the presence of more than one random element, and the PROC GLM algebra to compute correct estimated standard errors of the least squares means (see equation 18.18, page 584). The estimates of the variance components used to obtain $\boldsymbol{s}^2(\boldsymbol{Y})$ were computed from GLM partial (Type III) sums of squares.

The standard errors reported by PROC GLM will not in general be correct when the model involves more than one random element (Table 19.13). (This is true whether or not the data are balanced.) In this case study, the GLM standard errors for the whole-plot treatment means (*TILL* and *HERB*) varied from 81 to 89% of the "EXACT" standard errors. The standard errors for the subplot treatment means (*CULT*), which contain only the one variance component, varied from 87 to 96 % of the "EXACT" standard errors. The GLM ADJ standard errors provide better agreement with the "EXACT" for the whole-plot treatment means. This adjustment has no effect on the standard errors for the subplot treatment means.

The need for correcting the GLM standard errors will depend on the relative magnitudes of the components of variance in the model. Multiplying by the square root of the ratio of the appropriate error mean squares

GLM ADJ is a simple adjustment and is recommended in all cases where computation of the "EXACT" standard errors does not seem practical. Adjustments to the standard errors are necessary even when the data are balanced. In the balanced case, the "GLM ADJ" procedure gives the "EXACT" result.

Standard Errors of Mean Differences

The standard errors of the mean differences, column 6 of Table 19.13, are given to emphasize that, with unbalanced data, variances of differences cannot in general be computed simply as the sum of the variances; the least squares means are *not* independent. The standard errors of the mean differences given in Table 19.13 are computed using the exact method that takes the covariances into account. The mean difference between the *CULT* treatments, 40.83, has a standard error of 5.91 if computed with the exact method but 5.74 if computed by summing the GLM variances as if the means were independent. Of the marginal treatment means, only the *H* and *NOH* treatment means for the *HERB* treatment factor are independent. The variance of the difference between the *H* and *NOH* means is equal to the sum of the two variances. Within the two-way table of *TILL* × *HERB* means, all means are independent except the *CT–H* mean and the *NT–H* mean.

All least squares means were estimable in this case because it was recognized in advance that the data contained no information on interactions between *CULT* and *TILL* and these interaction effects were left out of the model. Had this not been done, any least squares means involving the nonestimable higher-order interactions in their expectations would not have been estimable. Nonestimability of least squares means is a common problem in the analysis of unbalanced data when the model includes higher-order interactions. In such cases, it is sometimes necessary to simplify the model by dropping interaction effects to make the means estimable. If the interactions are significant, this creates problems with interpretation.

19.4 Mixed Model Analysis

The analysis in the previous sections, Sections 19.1 through 19.3, ignored the fact that the δ_{ijk} (and possibly the B_i) were random effects and used least squares estimation to produce an analysis of variance and adjusted treatment means. Only then was the randomness of the δ_{ijk} taken into account to construct tests of significance and appropriate measures of precision. Relatively recent developments in computing power and software have made it practical to attack the analysis of mixed models as described in Chapter 18. This section presents the results of the analysis of these data using the SAS program PROC MIXED (SAS Institute Inc., 1997).

The mixed model for these data is as presented in equation 19.1 where all effects are fixed effects except the δ_{ijk} and ϵ_{ijkl}. The latter are assumed

to be normally distributed random effects with zero mean and variances σ_δ^2 and σ^2, respectively. For illustration, we also present results for the analysis where the B_i (*BLOCK* effects) also are considered to be normally distributed random effects with variance σ_B^2. Initial mixed model analyses showed the *HERB* × *CULT* effects to be trivial (which is in the analysis of variance results). Consequently, these interaction effects have been dropped from the model for the mixed model analysis results presented.

The PROC MIXED program statements (with *BLOCK* effects fixed) that generated the results presented are as follows.

```
PROC MIXED DATA= filename;
  CLASS BLOCK TILL HERB CULT;
  MODEL YIELD = BLOCK TILL HERB TILL*HERB CULT /
   ddfm=SATTERTHWAITE;
  RANDOM BLOCK*TILL*HERB;
  LSMEANS TILL HERB CULT TILL*HERB;
  ESTIMATE 'TILL CT-NT' TILL 1 -1 / CL;
  ESTIMATE 'HERB H-NOH' HERB 1 -1 / CL;
  ESTIMATE 'CULT C-NOC' CULT 1 -1 / CL;
RUN;
```

The MODEL statement contains only the fixed effects; the random effects are listed in the RANDOM statement. Note that the residuals ϵ_{ijkl} are always assumed to be random effects. The three-way interaction in the RANDOM statement identifies the δ_{ijk} effects. Only the MODEL and RANDOM statements need to be changed in order to treat *BLOCK* effects as random:

```
MODEL YIELD = TILL HERB TILL*HERB CULT;
RANDOM BLOCK BLOCK*TILL*HERB;
```

The REML (Restricted Estimated Maximum Likelihood) method of estimation was used. Convergence to a solution is usually quick. In this case, the convergence criterion was met in two iterations when *BLOCK* effects were fixed and in three iterations when *BLOCK* effects were random. The estimates of the variance components and the F-tests of the fixed effects are shown in Table 19.14. The "ddfm=SATTERTHWAITE" option in the model statement specifies that the Satterthwaite approximation is to be used for the denominator degrees of freedom for any F-tests. Both models (BLOCK fixed and random) give very similar results with respect to estimates of the variance components σ_δ^2 and σ^2. Recall that the estimates of the variance components obtained from the partial (Type III) sums of squares in the analysis of variance were $\widehat{\sigma}_\delta^2 = 35.06$ and $\widehat{\sigma}^2 = 119.56$. These estimates came from the model with the *HERB* × *CULT* interaction effects included. The analysis of variance estimates with these interaction effects dropped from the model are $\widehat{\sigma}_\delta^2 = 39.69$ and $\widehat{\sigma}^2 = 111.04$, much closer to

TABLE 19.14. *Estimates of variance components and F-tests of fixed effects from mixed model analysis using REML estimation.*

Variance	*BLOCK Effects Fixed*		*BLOCK Effects Random*	
Component	*Estimate*		*Estimate*	
$\sigma_B^2 (= BLOCK)$	—		201.16	
σ_δ^2	41.14		42.01	
σ^2	109.67		109.24	
Fixed Effect	*Type III F*	*Pr > F*	*Type III F*	*Pr > F*
BLOCK	9.09	.0042	—	—
TILL	.92	.3532	.85	.3709
HERB	55.05	.0001	54.99	.0001
TILL × *HERB*	6.54	.0304	6.39	.0319
CULT	56.58	.0001	56.30	.0001

the mixed model analysis results. Also, the tests of significance of the fixed effects are only trivially different between the two mixed models. (In these tests of significance, the Type III or partial sum of squares for the particular fixed effect is used for the numerator and an appropriate error mean square is computed for the denominator.) In the mixed model analyses, the *TILL* × *HERB* interaction is significant. It was approaching significance in the analysis of variance approach.

As with the analysis of variance approach, the treatment means can be adequately summarized with the marginal means for the factor *CULT* and the two-way table of *TILL* × *HERB* means, in all cases adjusted for the imbalance in the data. These least squares means and their standard errors are given in Table 19.15.

The least squares means are trivially different between the two models, *BLOCK* effects fixed or random. The striking difference in the two models is in the much larger standard errors of the treatment means when *BLOCK* effects are random. This is the direct contribution of σ_B^2 to the variance of the treatment means, and is appropriate if the means are to be viewed as estimates of the treatment means averaged over repeated samplings of blocks. However, these standard errors are much too large if one were to (mistakenly) compute the variance of a difference between two of the treatment means by adding the squares of these standard errors. To illustrate this, the estimates of the differences between the two levels of each of the three factors and the appropriate standard errors for the mean differences are given in the bottom portion of Table 19.15. As expected, the mean contrasts are trivially different between the two models but now the standard errors of the mean differences are also almost identical. Furthermore, they are similar to the results one would obtain if they were to be approximated using the standard errors of the means and assuming

TABLE 19.15. *Least squares means and standard errors estimated from the mixed model analyses with BLOCK effects fixed and random.*

		Block Effects Fixed		*Block Effects Random*	
Factor	*Level*	*Mean*	*Std. Err.*	*Mean*	*Std. Err.*
TILL	*CT*	52.09	3.56	52.19	7.94
	NT	57.66	4.42	57.56	8.36
HERB	*H*	73.32	3.73	73.37	8.02
	NOH	36.43	3.73	36.38	8.02
CULT	*C*	75.57	4.71	75.47	8.51
	NOC	34.18	2.90	34.28	7.66
TILL	*CT H*	64.18	5.18	64.38	8.79
$\times$*HERB*	*CT NOH*	40.01	4.90	40.01	8.63
	NT H	82.45	5.62	82.35	9.05
	NT NOH	32.86	5.50	32.76	9.05
CONTRAST:					
TILL	*CT−NT*	−5.56	5.81	−5.36	5.82
HERB	*H−NOH*	36.88	4.97	36.98	4.99
CULT	*C−NOC*	41.39	5.50	41.20	5.49

independence between the two means. For example, for the *TILL* contrast one obtains $\sqrt{3.56^2 + 4.42^2} = 5.67$ using the standard errors for the two *TILL* treatments versus the correct standard error of 5.81.

The differences between the PROC MIXED and PROC GLM results are small in this example, as they usually will be when the imbalance in the data is limited. The advantage of the PROC MIXED procedure is that the variance–covariance information is being utilized. This will produce more precise estimates and more powerful tests of significance if the information on the components of variance is reliable.

19.5 Exercises

19.1. Investigate whether a transformation of the data in this case study might be desirable. Use the Box–Cox transformation on yield for several values of λ. You will need to add a small constant to avoid problems with the zero yields. Run PROC GLM (or a similar program) for each transformed yield variable and plot the residual sums of squares against λ. Construct the confidence interval on λ. What transformation is suggested?

19.2. The partial (Type III) sums of squares from the analysis of variance, Table 19.2, and the mean square expectations, Table 19.10,

have been used to estimate the two components of variance σ^2 and σ_δ^2. Compute standard errors for each. (Assume each mean square is distributed as a chi-squared random variable scaled by $\mathcal{E}(MS)/d.f.$ so that its variance is $2[\mathcal{E}(MS)]^2/d.f$ and that the two mean squares are independent. The estimate of the variance of a chi-squared random variable is obtained by substituting the observed mean square for its expectation.) Compare these estimated standard errors with those given for the PROC MIXED solution, Section 19.4.

19.3. Verify that the constructed error mean square for testing *HERB* effects in the analysis of variance approach is $E' = 166.31$ and that its approximate degrees of freedom are $f' = 14$.

19.4. Determine the estimable functions for the population marginal means for a 2×3 factorial set of treatments in a randomized complete block design with $r = 4$ blocks. Include $A \times B$ interactions in your model. Give the estimable functions for the six treatment means and for the marginal treatment means for each treatment factor. How do the estimable functions change if there are no interactions in the model? Suppose cell (1, 2) is empty. Which means become nonestimable if there are interactions in the model? If there are no interactions in the model?

19.5 In Exercise 17.2 fixed block effects were assumed. Show how the expectations change if block effects are assumed to be random variables with zero mean and variance σ_b^2. Show how this changes your conclusions when the numbers are unequal as in Exercise 17.3.

Appendix A
APPENDIX TABLES

TABLE A.1. *Upper-tail probabilities for the t distribution.*

	Probability t > table entry								
d.f.	.25	.2	.15	.1	.05	.025	.01	.005	.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.460
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

TABLE A.2. *Percentage points for the F-distribution—Upper 10% points.*

	ν_1 = *Numerator Degrees of Freedom*									
ν_2[a]	1	2	3	4	5	6	7	8	9	10
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60

[a]Denominator degrees of freedom.

TABLE A.2. *(Continued).*

	ν_1 = *Numerator Degrees of Freedom*								
ν_2[a]	12	15	20	24	30	40	60	120	∞
1	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.11
6	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

[a]Denominator degrees of freedom.

TABLE A.3. *Percentage points for the F-distribution—Upper 5% points.*

	ν_1 = *Numerator Degrees of Freedom*									
ν_2[a]	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

[a]Denominator degrees of freedom.

TABLE A.3. *(Continued).*

	ν_1 = *Numerator Degrees of Freedom*								
ν_2[a]	12	15	20	24	30	40	60	120	∞
1	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

[a]Denominator degrees of freedom.

TABLE A.4. *Percentage points for the F-distribution—Upper 1% points.*

	ν_1 = *Numerator Degrees of Freedom*									
ν_2[a]	1	2	3	4	5	6	7	8	9	10
1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
∞	6.64	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

[a]Denominator degrees of freedom.

TABLE A.4. *(Continued).*

	ν_1 = *Numerator Degrees of Freedom*								
ν_2[a]	12	15	20	24	30	40	60	120	∞
1	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

[a]Denominator degrees of freedom.

TABLE A.5. *Bonferroni critical values* ($t_{(\alpha/2p;\nu)}$, $\alpha = .05$).

ν	*Number of tests* (p)								
d.f.	2	3	4	5	6	7	8	9	10
1	25.452	38.188	50.923	63.657	76.390	89.123	101.856	114.589	127.321
2	6.205	7.649	8.860	9.925	10.886	11.769	12.590	13.360	14.089
3	4.177	4.857	5.392	5.841	6.232	6.580	6.895	7.185	7.453
4	3.495	3.961	4.315	4.604	4.851	5.068	5.261	5.437	5.598
5	3.163	3.534	3.810	4.032	4.219	4.382	4.526	4.655	4.773
6	2.969	3.287	3.521	3.707	3.863	3.997	4.115	4.221	4.317
7	2.841	3.128	3.335	3.499	3.636	3.753	3.855	3.947	4.029
8	2.752	3.016	3.206	3.355	3.479	3.584	3.677	3.759	3.833
9	2.685	2.933	3.111	3.250	3.364	3.462	3.547	3.622	3.690
10	2.634	2.870	3.038	3.169	3.277	3.368	3.448	3.518	3.581
11	2.593	2.820	2.981	3.106	3.208	3.295	3.370	3.437	3.497
12	2.560	2.779	2.934	3.055	3.153	3.236	3.308	3.371	3.428
13	2.533	2.746	2.896	3.012	3.107	3.187	3.256	3.318	3.372
14	2.510	2.718	2.864	2.977	3.069	3.146	3.214	3.273	3.326
15	2.490	2.694	2.837	2.947	3.036	3.112	3.177	3.235	3.286
16	2.473	2.673	2.813	2.921	3.008	3.082	3.146	3.202	3.252
17	2.458	2.655	2.793	2.898	2.984	3.056	3.119	3.173	3.222
18	2.445	2.639	2.775	2.878	2.963	3.034	3.095	3.149	3.197
19	2.433	2.625	2.759	2.861	2.944	3.014	3.074	3.127	3.174
20	2.423	2.613	2.744	2.845	2.927	2.996	3.055	3.107	3.153
21	2.414	2.601	2.732	2.831	2.912	2.980	3.038	3.090	3.135
22	2.405	2.591	2.720	2.819	2.899	2.965	3.023	3.074	3.119
23	2.398	2.582	2.710	2.807	2.886	2.952	3.009	3.059	3.104
24	2.391	2.574	2.700	2.797	2.875	2.941	2.997	3.046	3.091
25	2.385	2.566	2.692	2.787	2.865	2.930	2.986	3.035	3.078
26	2.379	2.559	2.684	2.779	2.856	2.920	2.975	3.024	3.067
27	2.373	2.552	2.676	2.771	2.847	2.911	2.966	3.014	3.057
28	2.368	2.546	2.669	2.763	2.839	2.902	2.957	3.004	3.047
29	2.364	2.541	2.663	2.756	2.832	2.894	2.949	2.996	3.038
30	2.360	2.536	2.657	2.750	2.825	2.887	2.941	2.988	3.030
40	2.329	2.499	2.616	2.704	2.776	2.836	2.887	2.931	2.971
60	2.299	2.463	2.575	2.660	2.729	2.785	2.834	2.877	2.915
120	2.270	2.428	2.536	2.617	2.683	2.737	2.783	2.824	2.860
∞	2.241	2.394	2.498	2.576	2.638	2.690	2.734	2.773	2.807

TABLE A.6. *Bonferroni critical values* ($t_{(\alpha/2p;\nu)}$, $\alpha = .01$).

ν	*Number of tests* (p)								
d.f.	2	3	4	5	6	7	8	9	10
1	127.321	190.984	254.647	318.309	381.971	445.633	509.295	572.957	636.619
2	14.089	17.277	19.962	22.327	24.464	26.429	28.258	29.975	31.599
3	7.453	8.575	9.465	10.215	10.869	11.453	11.984	12.471	12.924
4	5.598	6.254	6.758	7.173	7.529	7.841	8.122	8.376	8.610
5	4.773	5.247	5.604	5.893	6.138	6.352	6.541	6.713	6.869
6	4.317	4.698	4.981	5.208	5.398	5.563	5.709	5.840	5.959
7	4.029	4.355	4.595	4.785	4.944	5.082	5.202	5.310	5.408
8	3.833	4.122	4.334	4.501	4.640	4.759	4.864	4.957	5.041
9	3.690	3.954	4.146	4.297	4.422	4.529	4.622	4.706	4.781
10	3.581	3.827	4.005	4.144	4.259	4.357	4.442	4.518	4.587
11	3.497	3.728	3.895	4.025	4.132	4.223	4.303	4.373	4.437
12	3.428	3.649	3.807	3.930	4.031	4.117	4.192	4.258	4.318
13	3.372	3.584	3.735	3.852	3.948	4.030	4.101	4.164	4.221
14	3.326	3.530	3.675	3.787	3.880	3.958	4.026	4.086	4.140
15	3.286	3.484	3.624	3.733	3.822	3.897	3.963	4.021	4.073
16	3.252	3.444	3.581	3.686	3.773	3.846	3.909	3.965	4.015
17	3.222	3.410	3.543	3.646	3.730	3.801	3.862	3.917	3.965
18	3.197	3.380	3.510	3.610	3.692	3.762	3.822	3.874	3.922
19	3.174	3.354	3.481	3.579	3.660	3.727	3.786	3.837	3.883
20	3.153	3.331	3.455	3.552	3.630	3.697	3.754	3.804	3.850
21	3.135	3.310	3.432	3.527	3.604	3.669	3.726	3.775	3.819
22	3.119	3.291	3.412	3.505	3.581	3.645	3.700	3.749	3.792
23	3.104	3.274	3.393	3.485	3.560	3.623	3.677	3.725	3.768
24	3.091	3.258	3.376	3.467	3.540	3.603	3.656	3.703	3.745
25	3.078	3.244	3.361	3.450	3.523	3.584	3.637	3.684	3.725
26	3.067	3.231	3.346	3.435	3.507	3.567	3.620	3.666	3.707
27	3.057	3.219	3.333	3.421	3.492	3.552	3.604	3.649	3.690
28	3.047	3.208	3.321	3.408	3.479	3.538	3.589	3.634	3.674
29	3.038	3.198	3.310	3.396	3.466	3.525	3.575	3.620	3.659
30	3.030	3.189	3.300	3.385	3.454	3.513	3.563	3.607	3.646
40	2.971	3.122	3.227	3.307	3.372	3.426	3.473	3.514	3.551
60	2.915	3.057	3.156	3.232	3.293	3.344	3.388	3.426	3.460
120	2.860	2.995	3.088	3.160	3.217	3.265	3.306	3.342	3.373
∞	2.807	2.935	3.023	3.090	3.144	3.189	3.227	3.261	3.291

TABLE A.7. *Significance points of the d_L and d_U for the Durbin–Watson test for correlation.*

5%[a]

	$p=1$		$p=2$		$p=3$		$p=4$		$p=5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	.95	1.54	.82	1.75	.69	1.97	.56	2.21
16	1.10	1.37	.98	1.54	.86	1.73	.74	1.93	.62	2.15
17	1.13	1.38	1.02	1.54	.90	1.71	.78	1.90	.67	2.10
18	1.16	1.39	1.05	1.53	.93	1.69	.82	1.87	.71	2.06
19	1.18	1.40	1.08	1.53	.97	1.68	.86	1.85	.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	.90	1.83	.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	.93	1.81	.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	.96	1.80	.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	.99	1.79	.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

[a]Reproduced in part from Tables 4 and 6 of Durbin and Watson (1951) with permission of the Biometrika Trustees.

TABLE A.7. *(Continued).*

1%

	$p=1$		$p=2$		$p=3$		$p=4$		$p=5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	.81	1.07	.70	1.25	.59	1.46	.49	1.70	.30	1.96
16	.84	1.09	.74	1.25	.63	1.44	.53	1.66	.44	1.90
17	.87	1.10	.77	1.25	.67	1.43	.57	1.63	.48	1.85
18	.90	1.12	.80	1.26	.71	1.42	.61	1.60	.52	1.80
19	.93	1.13	.83	1.26	.74	1.41	.65	1.58	.56	1.77
20	.95	1.15	.86	1.27	.77	1.41	.68	1.57	.60	1.74
21	.97	1.16	.89	1.27	.80	1.41	.72	1.55	.63	1.71
22	1.00	1.17	.91	1.28	.83	1.40	.75	1.54	.66	1.69
23	1.02	1.19	.94	1.29	.86	1.40	.77	1.53	.70	1.67
24	1.04	1.20	.96	1.30	.88	1.41	.80	1.53	.72	1.66
25	1.05	1.21	.98	1.30	.90	1.41	.83	1.52	.75	1.65
26	1.07	1.22	1.00	1.31	.93	1.41	.85	1.52	.78	1.64
27	1.09	1.23	1.02	1.32	.95	1.41	.88	1.51	.81	1.63
28	1.10	1.24	1.04	1.32	.97	1.41	.90	1.51	.83	1.62
29	1.12	1.25	1.05	1.33	.99	1.42	.92	1.51	.85	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	.94	1.51	.88	1.61
31	1.15	1.27	1.08	1.34	1.02	1.42	.96	1.51	.90	1.60
32	1.16	1.28	1.10	1.35	1.04	1.43	.98	1.51	.92	1.60
33	1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	.94	1.59
34	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	.95	1.59
35	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	.97	1.59
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	.99	1.59
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

TABLE A.8. *Empirical percentage points of the approximate W' test.*

	P^a										
n	.01	.05	.10	.15	.20	.50	.80	.85	.90	.95	.99
35	.919	.943	.952	.956	.964	.976	.982	.985	.987	.989	.992
50	.935	.953	.963	.968	.971	.981	.987	.988	.990	.991	.994
51	.935	.954	.964	.968	.971	.981	.988	.989	.990	.992	.994
53	.938	.957	.964	.969	.972	.982	.988	.989	.990	.992	.994
55	.940	.958	.965	.971	.973	.983	.988	.990	.991	.992	.994
57	.944	.961	.966	.971	.974	.983	.989	.990	.991	.992	.994
59	.945	.962	.967	.972	.975	.983	.989	.990	.991	.992	.994
61	.947	.963	.968	.973	.975	.984	.990	.990	.991	.992	.994
63	.947	.964	.970	.973	.976	.984	.990	.991	.992	.993	.994
65	.948	.965	.971	.974	.976	.985	.990	.991	.992	.993	.995
67	.950	.966	.971	.974	.977	.985	.990	.991	.992	.993	.995
69	.951	.966	.972	.976	.978	.986	.990	.991	.992	.993	.995
71	.953	.967	.972	.976	.978	.986	.990	.991	.992	.994	.995
73	.956	.968	.973	.976	.979	.986	.991	.992	.993	.994	.995
75	.956	.969	.973	.976	.979	.986	.991	.992	.993	.994	.995
77	.957	.969	.974	.977	.980	.987	.991	.992	.993	.994	.996
79	.957	.970	.975	.978	.980	.987	.991	.992	.993	.994	.996
81	.958	.970	.975	.979	.981	.987	.992	.992	.993	.994	.996
83	.960	.971	.976	.979	.981	.988	.992	.992	.993	.994	.996
85	.961	.972	.977	.980	.981	.988	.992	.992	.993	.994	.996
87	.961	.972	.977	.980	.982	.988	.992	.993	.994	.994	.996
89	.961	.972	.977	.981	.982	.988	.992	.993	.994	.995	.996
91	.962	.973	.978	.981	.983	.989	.992	.993	.994	.995	.996
93	.963	.973	.979	.981	.983	.989	.992	.993	.994	.995	.996
95	.965	.974	.979	.981	.983	.989	.993	.993	.994	.995	.996
97	.965	.975	.979	.982	.984	.989	.993	.993	.994	.995	.996
99	.967	.976	.980	.982	.984	.989	.993	.994	.994	.995	.996

[a]Reproduced with permission from Shapiro and Francia (1972).

TABLE A.9. *Runs test—critical number of runs in a sample of n for a 5% significance level. The null hypothesis is rejected if the observed number of runs is less than or equal to the tabled value.*

	n_a = *number in smaller category*								
n[a]	2	3	4	5	6	7	8	9	10
8	∗[b]	2	2						
9	∗	2	2						
10	2	2	2	3					
11	2	2	3	3					
12	2	2	3	3	3				
13	2	2	3	3	3				
14	2	3	3	4	4	4			
15	2	3	3	4	4	4			
16	2	3	3	4	4	5	5		
17	2	3	4	4	5	5	5		
18	2	3	4	4	5	5	5	5	
19	2	3	4	4	5	5	6	6	
20	2	3	4	5	5	6	6	6	6

[a]5% significance cannot be achieved if $n < 8$.
[b]Not even as few as 2 runs is significant.

TABLE A.10. *Runs test—critical number of runs in a sample of n for a 1% significance level. The null hypothesis is rejected if the observed number of runs is less than or equal to the tabled value.*

	n_a = *number in smaller category*								
n[a]	2	3	4	5	6	7	8	9	10
10	∗[b]	∗	2	2					
11	∗	∗	2	2					
12	∗	2	2	2	2				
13	∗	2	2	2	3				
14	∗	2	2	3	3	3			
15	∗	2	2	3	3	3			
16	∗	2	3	3	3	3	3		
17	∗	2	3	3	3	4	4		
18	∗	2	3	3	4	4	4	4	
19	∗	2	3	3	4	4	4	5	
20	∗	2	3	4	4	4	5	5	5

[a]1% significance cannot be achieved if $n < 10$.
[b]Not even as few as 2 runs is significant.

REFERENCES

[1] A. Agresti. *Categorical Data Analysis*. Wiley, New York, 1990.

[2] H. Akaike. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21:243–247, 1969.

[3] D. F. Alderdice. Some effects of simultaneous variation in salinity, temperature and dissolved oxygen on the resistance of young coho salmon to a toxic substance. *Journal of the Fisheries Research Board of Canada*, 20:525–475, 1963.

[4] D. M. Allen. Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13:469–475, 1971a.

[5] D. M. Allen. The prediction sum of squares as a criterion for selection of predictor variables. Technical Report 23, Department of Statistics, University of Kentucky, 1971b.

[6] R. L. Anderson and L. A. Nelson. A family of models involving intersecting straight lines and concomitant experimental designs useful in evaluating response to fertilizer nutrients. *Biometrics*, 31:303–318, 1975.

[7] T. W. Anderson. *The Statistical Analysis of Time Series*. Wiley, New York, 1971.

[8] D. F. Andrews and A. M. Herzberg. *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. Springer-Verlag, New York, 1985.

[9] F. J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27:17–21, 1973.

[10] A. C. Atkinson. Diagnostic regression analysis and shifted power transformations. *Technometrics*, 25:23–33, 1983.

[11] M. S. Bartlett. The use of transformations. *Biometrics*, 3:39–53, 1947.

[12] M. S. Bartlett. Fitting a straight line when both variables are subject to error. *Biometrics*, 5:207–212, 1949.

[13] R. P. Basson. *On unbiased estimation in variance component models*. PhD thesis, Iowa State University of Science and Technology, 1965.

[14] D. A. Belsley. Demeaning conditioning diagnostics through centering (with discussion). *The American Statistician*, 38:73–77, 1984.

[15] D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York, 1980.

[16] R. B. Bendel and A. A. Afifi. Comparison of stopping rules in forward "stepwise" regression. *Journal of the American Statistical Association*, 72:46–53, 1977.

[17] K. N. Berk. Tolerance and condition in regression computations. *Journal of the American Statistical Association*, 72:863–866, 1977.

[18] K. N. Berk. Comparing subset regression procedures. *Technometrics*, 20:1–6, 1978.

[19] G. Blom. *Statistical Estimates and Transformed Beta Variates*. Wiley, New York, 1958.

[20] P. Bloomfield. *Fourier Analysis of Time Series: An Introduction*. Wiley, New York, 1976.

[21] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26:211–243, 1964.

[22] G. E. P. Box and N. R. Draper. *Empirical Model-Building and Response Surfaces*. Wiley, New York, 1987.

[23] G. E. P. Box and P. W. Tidwell. Transformation of the independent variables. *Technometrics*, 4:531–550, 1962.

[24] G. E. P. Box, W. G. Hunter, and J. S. Hunter. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. Wiley, New York, 1978.

[25] D. Bradu and K. R. Gabriel. Simultaneous statistical inference on interaction in two-way analysis of variance. *Journal of the American Statistical Association*, 69:428–436, 1974.

[26] D. Bradu and K. R. Gabriel. The biplot as a diagnostic tool for models of two-way tables. *Technometrics*, 20:47–68, 1978.

[27] R. L. Brown, J. Durbin, and J. M. Evans. Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society, Series B*, 37:149–192, 1975.

[28] O. Bunke and B. Droge. Estimators of the mean squared error of prediction in linear regression. *Technometrics*, 26:145–155, 1984.

[29] D. A. Buonagurio, S. Nakada, J. D. Parvin, M. Krystal, P. Palese, and W. M. Fitch. Evolution of human influenza A viruses over 50 years: Rapid, uniform rate of change in NS gene. *Science*, 232:980–982, 1986.

[30] E. Cameron and L. Pauling. Supplemental ascorbate in the supportive treatment of cancer: Reevaluation of prolongation of survival times in terminal human cancer. *Proceedings of the National Academy of Sciences U.S.A.*, 75:4538–4542, 1978.

[31] R. J. Carroll and D. Ruppert. *Transformations and Weighting in Regression*. Chapman & Hall, London, 1988.

[32] R. J. Carroll, D. Ruppert, and L. A. Stefanski. *Measurement Error in Nonlinear Models*. Chapman & Hall, London, 1995.

[33] R. L. Carter and W. A. Fuller. Instrumental variable estimation of the simple errors-in-variables model. *Journal of the American Statistical Association*, 75:687–692, 1980.

[34] G. P. Y. Clarke. Marginal curvatures in the analysis of nonlinear regression models. *Journal of the American Statistical Association*, 82:844–850, 1987.

[35] W. G. Cochran. *Planning and Analysis of Observational Studies*. Wiley, New York, 1983.

[36] J. Cook and L. A. Stefanski. A simulation extrapolation method for parametric measurement error models. *Journal of the American Statistical Association*, 89:1314–1328, 1995.

[37] R. D. Cook. Detection of influential observations in linear regression. *Technometrics*, 19:15–18, 1977.

[38] R. D. Cook. Influential observations in linear regression. *Journal of the American Statistical Association*, 74:169–174, 1979.

[39] R. D. Cook. Comment [to Belsley, D. A. (1984)]. *The American Statistician*, 38:78–79, 1984.

[40] R. D. Cook and P. Prescott. On the accuracy of Bonferroni significance levels for detecting outliers in linear models. *Technometrics*, 23:59–63, 1981.

[41] R. D. Cook and P. C. Wang. Transformations and influential cases in regression. *Technometrics*, 25:337–343, 1983.

[42] R. D. Cook and S. Weisberg. *Residuals and Influence in Regression*. Chapman & Hall, London, 1982.

[43] L. C. A. Corsten and K. R. Gabriel. Graphical exploration in comparing variance matrices. *Biometrics*, 32:851–863, 1976.

[44] H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, New Jersey, 1946.

[45] C. Daniel and F. S. Wood. *Fitting Equations to Data: Computer Analysis of Multifactor Data*. Wiley, New York, 2nd edition, 1980.

[46] W. J. Dixon, editor. *BMDP Statistical Software 1981*. University of California Press, Berkeley, California, 1981.

[47] S. Drake. *Galileo at Work*. University of Chicago Press, Chicago, 1978.

[48] N. Draper and H. Smith. *Applied Regression Analysis*. Wiley, New York, 2nd edition, 1981.

[49] J. Durbin and G. S. Watson. Testing for serial correlation in least squares regression. II. *Biometrika*, 38:159–178, 1951.

[50] J. Durbin and G. S. Watson. Testing for serial correlation in least squares regression. III. *Biometrika*, 58:1–19, 1971.

[51] M. Feldstein. Errors in variables: A consistent estimator with smaller MSE in finite samples. *Journal of the American Statistical Association*, 69:990–996, 1974.

[52] R. J. Freund, R. C. Littell, and P. C. Spector. *SAS System for Linear Models*. SAS Institute, Inc., Cary, North Carolina, 2nd edition, 1986.

[53] W. A. Fuller. *Measurement Error Models*. Wiley, New York, 1987.

[54] W. A. Fuller. *Introduction to Statistical Time Series*. Wiley, New York, 1996.

[55] G. M. Furnival. All possible regressions with less computation. *Technometrics*, 13:403–408, 1971.

[56] G. M. Furnival and R. B. Wilson. Regression by leaps and bounds. *Technometrics*, 16:499–511, 1974.

[57] K. R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58:453–467, 1971.

[58] K. R. Gabriel. Analysis of meteorological data by means of canonical decomposition and biplots. *Journal of Applied Meteorology*, 11:1071–1077, 1972.

[59] K. R. Gabriel. Least squares approximation of matrices by additive and multiplicative models. *Journal of the Royal Statistical Society, Series B*, 40:186–196, 1978.

[60] A. R. Gallant. *Nonlinear Statistical Models.* Wiley, New York, 1987.

[61] A. R. Gallant and W. A. Fuller. Fitting segmented polynomial models whose join points have to be estimated. *Journal of the American Statistical Association*, 68:144–147, 1973.

[62] J. S. Galpin and D. M. Hawkins. The use of recursive residuals in checking model fit in linear regression. *The American Statistician*, 38:94–105, 1984.

[63] F. A. Graybill. *An Introduction to Linear Statistical Models.* McGraw-Hill, New York, 1961.

[64] M. L. Gumpertz and S. G. Pantula. A simple approach to inferences in random coefficient models. *The American Statistician*, 43:203–210, 1989.

[65] R. F. Gunst. Comment: Toward a balanced assessment of collinearity diagnostics. *The American Statistician*, 38:79–82, 1984.

[66] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics, The Approach Based on Influence Functions.* Wiley, New York, 1986.

[67] H. O. Hartley. The modified Gauss–Newton method for the fitting of nonlinear regression functions by least-squares. *Technometrics*, 3:269–280, 1961.

[68] C. M. Hawkins. On the investigation of alternative regressions by principal component analysis. *Applied Statistics*, 22:275–286, 1973.

[69] W. W. Heck, W. W. Cure, J. O. Rawlings, L. J. Zaragosa, A. S. Heagle, H. E. Heggestad, R. J. Kohut, L. W. Kress, and P. J. Temple. Assessing impacts of ozone on agricultural crops: II. *Journal of the Air Pollution Control Association*, 34:810–817, 1984.

[70] A. Hedayat and D. S. Robson. Independent stepwise residuals for testing homoscedasticity. *Journal of the American Statistical Association*, 65:1573–1581, 1970.

[71] F. Hernandez and R. A. Johnson. The large-sample behavior of transformations to normality. *Journal of the American Statistical Association*, 75:855–861, 1980.

[72] R. R. Hocking. The analysis and selection of variables in linear regression. *Biometrics*, 32:1–49, 1976.

[73] R. R. Hocking. *The Analysis of Linear Models*. Brooks/Cole, Monterey, California, 1985.

[74] R. R. Hocking and F. M. Speed. A full-rank analysis of some linear model problems. *Journal of the American Statistical Association*, 70:706–712, 1975.

[75] R. R. Hocking, F. M. Speed, and M. J. Lynn. A class of biased estimators in linear regression. *Technometrics*, 18:425–437, 1976.

[76] A. E. Hoerl and R. W. Kennard. Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12:69–82, 1970a.

[77] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970b.

[78] A. E. Hoerl, R. W. Kennard, and K. F. Baldwin. Ridge regression: Some simulations. *Communications in Statistics*, 4:105–124, 1975.

[79] A. S. Householder and G. Young. Matrix approximation and latent roots. *American Mathematical Monthly*, 45:165–171, 1938.

[80] C. J. Huang and B. W. Bolch. On testing of regression disturbances for normality. *Journal of the American Statistical Association*, 69:330–335, 1974.

[81] P. J. Huber. *Robust Statistics*. Wiley, New York, 1981.

[82] G. G. Judge, W. E. Griffiths, R. C. Hill, and T. Lee. *The Theory and Practice of Econometrics*. Wiley, New York, 1980.

[83] W. J. Kennedy and T. A. Bancroft. Model-building for prediction in regression based on repeated significance tests. *Annals of Mathematical Statistics*, 42:1273–1284, 1971.

[84] S. B. Land. *Sea water flood tolerance of some Southern pines*. PhD thesis, Department of Forestry and Department of Genetics, North Carolina State University, 1973.

[85] R. A. Linthurst. *Aeration, nitrogen, pH and salinity as factors affecting Spartina Alterniflora growth and dieback*. PhD thesis, North Carolina State University, 1979.

[86] R. C. Littell, G. A. Milliken, W. W. Stroup, and R. D. Wolfinger. *SAS System for Mixed Models*. SAS Institute Inc., Cary, North Carolina, 1996.

[87] W. F. Lott. The optimal set of principal component restrictions on a least squares regression. *Communications in Statistics*, 2:449–464, 1973.

[88] A. Madansky. The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association*, 54:173–205, 1959.

[89] C. L. Mallows. Data analysis in a regression context. In W. L. Thompson and F. B. Cady, editors, *University of Kentucky Conference on Regression with a Large Number of Predictor Variables*, Department of Statistics, University of Kentucky, 1973a.

[90] C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973b.

[91] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11:431–441, 1963.

[92] D. W. Marquardt. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12:591–612, 1970.

[93] D. W. Marquardt. Comment: You should standardize the predictor variables in your regression models. *Journal of the American Statistical Association*, 75:87–91, 1980.

[94] D. W. Marquardt and R. D. Snee. Ridge regression in practice. *The American Statistician*, 29:3–19, 1975.

[95] R. L. Mason and R. F. Gunst. Outlier-induced collinearities. *Technometrics*, 27:401–407, 1985.

[96] R. G. Miller, Jr. *Simultaneous Statistical Inference*. Springer-Verlag, New York, 2nd edition, 1981.

[97] R. A. Mombiela and L. A. Nelson. Relationships among some biological and empirical fertilizer response models and use of the power family of transformations to identify an appropriate model. *Agronomy Journal*, 73:353–356, 1981.

[98] R. Mosteller and J. W. Tukey. *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley, Reading, Massachusetts, 1977.

[99] R. H. Myers. *Classical and Modern Regression with Applications*. PWS-KENT, Boston, 2nd edition, 1990.

[100] J. A. Nelder. Inverse polynomials, a useful group of multifactor response functions. *Biometrics*, 22:128–140, 1966.

[101] W. R. Nelson and D. W. Ahrenholz. Population and fishery characteristics of Gulf Menhaden, Brevoortia patronus. *Fishery Bulletin*, 84:311–325, 1986.

[102] D. R. Nielsen, J. W. Biggar, and E. T. Erh. Spatial variability of field-measured soil-water properties. *Hilgardia*, 42:215–259, 1973.

[103] M. J. Norusis. *SPSS-X Advanced Statistics Guide*. McGraw-Hill, Chicago, 1985.

[104] S. G. Pantula and K. H. Pollock. Nested analyses of variance with autocorrelated errors. *Biometrics*, 41:909–920, 1985.

[105] S. H. Park. Collinearity and optimal restrictions on regression parameters for estimating responses. *Technometrics*, 23:289–295, 1981.

[106] E. S. Pearson and H. O. Hartley. *Biometrika Tables for Statisticians*, volume 1. Cambridge University Press, London, 3rd edition, 1966.

[107] S. P. Pennypacker, H. D. Knoble, C. E. Antle, and L. V. Madden. A flexible model for studying plant disease progression. *Phytopathology*, 70:232–235, 1980.

[108] Pharos Books. *1993 Almanac and Book of Facts*. Scripps Howard Company, New York, 1993.

[109] D. A. Pierce and R. J. Gray. Testing normality of errors in regression models. *Biometrika*, 69:233–236, 1982.

[110] D. A. Pierce and K. J. Kopecky. Testing goodness of fit for the distribution of errors in regression models. *Biometrika*, 66:1–5, 1979.

[111] C. P. Quesenberry. Some transformation methods in goodness-of-fit. In R. B. D'Agostino and M. A. Stephens, editors, *Goodness of Fit Techniques*. Chapter 6. Marcel Dekker, New York, 1986.

[112] C. P. Quesenberry and C. Quesenberry, Jr. On the distribution of residuals from fitted parametric models. *Journal of Statistical Computation and Simulation*, 15:129–140, 1982.

[113] M. L. Ralston and R. I. Jennrich. Dud, a derivative-free algorithm for nonlinear least squares. *Technometrics*, 20:7–14, 1978.

[114] C. R. Rao. *Linear Statistical Inference and Its Applications*. Wiley, New York, 2nd edition, 1973.

[115] J. O. Rawlings and W. W. Cure. The Weibull function as a dose-response model for air pollution effects on crop yields. *Crop Science*, 25:807–814, 1985.

[116] D. S. Riggs, J. A. Guarnieri, and S. Addelman. Fitting straight lines when both variables are subject to error. *Life Sciences*, 22:1305–1360, 1978.

[117] F. J. Rohlf and R. R. Sokal. *Statistical Tables*. W. H. Freeman, San Francisco, 2nd edition, 1981.

[118] M. Saeed and C. A. Francis. Association of weather variables and genotype × environment interactions in grain sorghum. *Crop Science*, 24:13–16, 1984.

[119] SAS Institute Inc. *SAS/STAT User's Guide, Version 6, Volume I*. SAS Institute Inc., Cary, North Carolina, 4th edition, 1989a.

[120] SAS Institute Inc. *SAS/STAT User's Guide, Version 6, Volume II*. SAS Institute Inc., Cary, North Carolina, 4th edition, 1989b.

[121] SAS Institute Inc. *SAS Language and Procedures: Usage, Version 6*. SAS Institute Inc., Cary, North Carolina, 1st edition, 1989c.

[122] SAS Institute Inc. *SAS/IML Software: Usage and Reference, Version 6*. SAS Institute Inc., Cary, North Carolina, 1st edition, 1989d.

[123] SAS Institute Inc. *SAS Procedures Guide, Version 6*. SAS Institute Inc., Cary, North Carolina, 3rd edition, 1990.

[124] SAS Institute Inc. *SAS/STAT Software: Changes and Enhancements Through Release 6.12*. SAS Institute Inc., Cary, North Carolina, 1997.

[125] F. E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2:110–114, 1946.

[126] H. Scheffé. A method for judging all contrasts in the analysis of variance. *Biometrika*, 40:87–104, 1953.

[127] H. Scheffé. *The Analysis of Variance*. Wiley, New York, 1959.

[128] H. Schneeweiss. Consistent estimation of a regression with errors in the variables. *Metrika*, 23:101–116, 1976.

[129] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[130] S. R. Searle. *Linear Models*. Wiley, New York, 1971.

[131] S. R. Searle. *Matrix Algebra Useful for Statistics*. Wiley, New York, 1982.

[132] S. R. Searle. *Linear Models for Unbalanced Data*. Wiley, New York, 1986.

[133] S. R. Searle and W. H. Hausman. *Matrix Algebra for Business and Economics*. Wiley, New York, 1970.

[134] S. R. Searle and H. V. Henderson. Annotated computer output for analyses of unbalanced data: SAS GLM. Technical Report BU-641-M, Biometrics Unit, Cornell University, 1979.

[135] S. R. Searle, F. M. Speed, and G. A. Milliken. Population marginal means in the linear model: An alternative to least squares means. *The American Statistician*, 34:216–221, 1980.

[136] S. S. Shapiro and R. S. Francia. An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67:215–216, 1972.

[137] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611, 1965.

[138] J. S. Shy-Modjeska, J. S. Riviere, and J. O. Rawlings. Application of biplot methods to the multivariate analysis of toxicological and pharmacokinetic data. *Toxicology and Applied Pharmacology*, 72:91–101, 1984.

[139] G. Smith and F. Campbell. A critique of some ridge regression methods. *Journal of the American Statistical Association*, 75:74–81, 1980.

[140] G. W. Snedecor and W. G. Cochran. *Statistical Methods*. Iowa State University Press, Ames, Iowa, 8th edition, 1989.

[141] R. D. Snee. Validation of regression models: Methods and examples. *Technometrics*, 19:415–428, 1977.

[142] R. D. Snee and D. W. Marquardt. Comment: Collinearity diagnostics depend on the domain of prediction, the model, and the data. *The American Statistician*, 38:83–87, 1984.

[143] F. M. Speed and R. R. Hocking. The use of the R(·)-notation with unbalanced data. *The American Statistician*, 30:30–33, 1976.

[144] F. M. Speed, R. R. Hocking, and O. P. Hackney. Methods of analysis of linear models with unbalanced data. *Journal of the American Statistical Association*, 73:105–112, 1978.

[145] R. G. D. Steel, J. H. Torrie, and D. A. Dickey. *Principles and Procedures of Statistics: A Biometrical Approach*. McGraw-Hill, New York, 3rd edition, 1997.

[146] C. M. Stein. Multiple regression. In *Contributions to Probability and Statistics, Essays in Honor of Harold Hotelling*. Stanford University Press, Stanford, California, 1960.

[147] G. W. Stewart. *Introduction to Matrix Computations*. Academic Press, New York, 1973.

[148] F. S. Swed and C. Eisenhart. Tables for testing randomness of grouping in a sequence of alternatives. *Annals of Mathematical Statistics*, 14:66–87, 1943.

[149] H. Theil. *Principles of Econometrics*. Wiley, New York, 1971.

[150] R. A. Thisted. Comment: A critique of some ridge regression methods. *Journal of the American Statistical Association*, 75:81–86, 1980.

[151] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts, 1977.

[152] J. C. van Houwelingen. Use and abuse of variance models in regression. *Biometrics*, 44:1073–1081, 1988.

[153] A. Wald. The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics*, 11:284–300, 1940.

[154] J. T. Webster, R. F. Gunst, and R. L. Mason. Latent root regression analysis. *Technometrics*, 16:513–522, 1974.

[155] S Weisberg. An empirical comparison of the percentage points of W and W′. *Biometrika*, 61:644–646, 1974.

[156] S. Weisberg. Comment on White and MacDonald (1980). *Journal of the American Statistical Association*, 75:28–31, 1980.

[157] S. Weisberg. A statistic for allocating C_p to individual cases. *Technometrics*, 23:27–31, 1981.

[158] S. Weisberg. *Applied Linear Regression*. Wiley, New York, 2nd edition, 1985.

[159] H. White and G. M. MacDonald. Some large-sample tests for nonnormality in the linear regression model (with comment by S. Weisberg). *Journal of the American Statistical Association*, 75:16–31, 1980.

[160] F. S. Wood. Comment: Effect of centering on collinearity and interpretation of the constant. *The American Statistician*, 38:88–90, 1984.

[161] H. Working and H. Hotelling. Application of the theory of error to the interpretation of trends. *Journal of the American Statistical Association, Supplement (Proceedings)*, 24:73–85, 1929.

AUTHOR INDEX

SUBJECT INDEX

Springer Texts in Statistics *(continued from page ii)*

Nguyen and Rogers: Fundamentals of Mathematical Statistics: Volume I: Probability for Statistics
Nguyen and Rogers: Fundamentals of Mathematical Statistics: Volume II: Statistical Inference
Noether: Introduction to Statistics: The Nonparametric Way
Nolan and Speed: Stat Labs: Mathematical Statistics Through Applications
Peters: Counting for Something: Statistical Principles and Personalities
Pfeiffer: Probability for Applications
Pitman: Probability
Rawlings, Pantula and Dickey: Applied Regression Analysis
Robert: The Bayesian Choice: A Decision-Theoretic Motivation
Robert: The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, Second Edition
Robert and Casella: Monte Carlo Statistical Methods
Santner and Duffy: The Statistical Analysis of Discrete Data
Saville and Wood: Statistical Methods: The Geometric Approach
Sen and Srivastava: Regression Analysis: Theory, Methods, and Applications
Shao: Mathematical Statistics
Shorack: Probability for Statisticians
Shumway and Stoffer: Time Series Analysis and Its Applications
Terrell: Mathematical Statistics: A Unified Introduction
Whittle: Probability via Expectation, Fourth Edition
Zacks: Introduction to Reliability Analysis: Probability Models and Statistical Methods