

APPLIED REGRESSION ANALYSIS AND OTHER MULTIVARIABLE METHODS

FIFTH EDITION



KLEINBAUM / KUPPER / NIZAM / ROSENBERG

APPLIED REGRESSION ANALYSIS AND OTHER MULTIVARIABLE METHODS

FIFTH EDITION

DAVID G. KLEINBAUM
EMORY UNIVERSITY

LAWRENCE L. KUPPER
UNIVERSITY OF NORTH CAROLINA, CHAPEL HILL

AZHAR NIZAM
EMORY UNIVERSITY

ELI S. ROSENBERG
EMORY UNIVERSITY



Australia • Brazil • Mexico • Singapore • United Kingdom • United States

This is an electronic version of the print textbook. Due to electronic rights restrictions, some third party content may be suppressed. Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. The publisher reserves the right to remove content from this title at any time if subsequent rights restrictions require it. For valuable information on pricing, previous editions, changes to current editions, and alternate formats, please visit www.cengage.com/highered to search by ISBN#, author, title, or keyword for materials in your areas of interest.

**Applied Regression Analysis and Other
Multivariable Methods, Fifth Edition**

Kleinbaum, Kupper, Nizam, Rosenberg

Senior Product Manager: Molly Taylor

Senior Content Developer: Laura Wheel

Media Developer: Andrew Coppola

Market Development Manager: Ryan Ahern

Senior Content Project Manager:
Jessica Rasile

Art Director: Linda May

Manufacturing Planner: Sandee Milewski

Rights Acquisition Specialist: Shalice
Shah-CaldwellProduction Service/Compositor:
Cenveo® Publisher Services

Cover Designer: Pier1 Design

Cover Images: Flying Colours Ltd/Photodisc/
Getty Images; © iStockPhoto.com/exdez;
© iStockPhoto.com/peepo; © iStockPhoto
.com/iPandastudio

© 2014, 2008, 1998 Cengage Learning

ALL RIGHTS RESERVED. No part of this work covered by the copyright herein may be reproduced, transmitted, stored, or used in any form or by any means graphic, electronic, or mechanical, including but not limited to photocopying, recording, scanning, digitizing, taping, web distribution, information networks, or information storage and retrieval systems, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the publisher.

For product information and technology assistance, contact us at
Cengage Learning Customer & Sales Support, 1-800-354-9706

For permission to use material from this text or product,

submit all requests online at www.cengage.com/permissions.Further permissions questions can be emailed to
permissionrequest@cengage.com.

Library of Congress Control Number: 2013946149

ISBN-13: 978-1-285-05108-6

ISBN-10: 1-285-05108-4

Cengage Learning20 Channel Center Street
Boston, MA 02210
USACengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil and Japan. Locate your local office at international.cengage.com/region.Cengage Learning products are represented in Canada by
Nelson Education, Ltd.For your course and learning solutions, visit www.cengage.com.Purchase any of our products at your local college store or at our preferred online store www.cengagebrain.com.**Instructors:** Please visit login.cengage.com and log in to access instructor-specific resources.

Printed in the United States of America

1 2 3 4 5 6 7 17 16 15 14 13

DEDICATIONS

David Kleinbaum:

To my wonderful wife, Edna, for her love and fun approach to life, and to Dr. Viola Vaccarino, my department chair in Epidemiology at RSPH/Emory, for her leadership and continued support.

Larry Kupper:

To Sandy (the light of my life), to Mark and Chieko (my wonderful son and daughter-in-law), and to Dr. William Mendenhall (a fabulous teacher and mentor).

Azhar Nizam:

To the family that inspires me: my loving and wonderful wife, Janet, and children Zainab and Sohail.

Eli S. Rosenberg:

To my partner and best friend, Abby; my mother, Frances; and my father, Gabriel, the first and most influential statistician in my life.

All authors:

To all teachers who deserve recognition for the long hours of course preparation, development of innovative course materials and textbooks, mentoring of students, and devotion to successful and creative teaching.

CONTENTS

1 Concepts and Examples of Research

1

1.1 Concepts	1
1.2 Examples	2
1.3 Concluding Remarks	5
References	6

2 Classification of Variables and the Choice of Analysis

7

2.1 Classification of Variables	7
2.2 Overlapping of Classification Schemes	11
2.3 Choice of Analysis	12
References	15

3 Basic Statistics: A Review

16

3.1 Preview	16
3.2 Descriptive Statistics	17
3.3 Random Variables and Distributions	19
3.4 Sampling Distributions of t , χ^2 , and F	24
3.5 Statistical Inference: Estimation	27
3.6 Statistical Inference: Hypothesis Testing	30
3.7 Error Rates, Power, and Sample Size	35
Problems	37
References	40

4 Introduction to Regression Analysis

41

4.1 Preview	41
4.2 Association versus Causality	42
4.3 Statistical versus Deterministic Models	45
4.4 Concluding Remarks	45
References	46

5 Straight-line Regression Analysis

47

5.1 Preview	47
5.2 Regression with a Single Independent Variable	47
5.3 Mathematical Properties of a Straight line	50
5.4 Statistical Assumptions for a Straight-line Model	51
5.5 Determining the Best-fitting Straight Line	55
5.6 Measure of the Quality of the Straight-line Fit and Estimate of σ^2	60
5.7 Inferences about the Slope and Intercept	61
5.8 Interpretations of Tests for Slope and Intercept	64
5.9 The Mean Value of Y at a Specified Value of X	66
5.10 Prediction of a New Value of Y at X_0	68
5.11 Assessing the Appropriateness of the Straight-line Model	70
5.12 Example: BRFSS Analysis	71
Problems	74
References	107

6 The Correlation Coefficient and Straight-line Regression Analysis

108

6.1 Definition of r	108
6.2 r as a Measure of Association	109
6.3 The Bivariate Normal Distribution	112
6.4 r^2 and the Strength of the Straight-line Relationship	113
6.5 What r^2 Does Not Measure	115
6.6 Tests of Hypotheses and Confidence Intervals for the Correlation Coefficient	117
6.7 Testing for the Equality of Two Correlations	120
6.8 Example: BRFSS Analysis	122
6.9 How Large Should r^2 Be in Practice?	123
Problems	125
References	127

7 The Analysis-of-Variance Table 129

7.1 Preview	129
7.2 The ANOVA Table for Straight-line Regression Problems	129 133

8 Multiple Regression Analysis: General Considerations 136

8.1 Preview	136
8.2 Multiple Regression Models	137
8.3 Graphical Look at the Problem	138
8.4 Assumptions of Multiple Regression	141
8.5 Determining the Best Estimate of the Multiple Regression Equation	143
8.6 The ANOVA Table for Multiple Regression	145
8.7 Example: BRFSS Analysis	146
8.8 Numerical Examples Problems References	148 151 164

9 Statistical Inference in Multiple Regression 165

9.1 Preview	165
9.2 Test for Significant Overall Regression	166
9.3 Partial <i>F</i> Test	167
9.4 Multiple Partial <i>F</i> Test	172
9.5 Strategies for Using Partial <i>F</i> Tests	175
9.6 Additional Inference Methods for Multiple Regression	180
9.7 Example: BRFSS Analysis Problems References	186 188 198

10 Correlations: Multiple, Partial, and Multiple Partial 199

10.1 Preview	199
10.2 Correlation Matrix	200
10.3 Multiple Correlation Coefficient	201
10.4 Relationship of $R_{Y X_1, X_2, \dots, X_k}$ to the Multivariate Normal Distribution	203

10.5 Partial Correlation Coefficient	204
10.6 Alternative Representation of the Regression Model	212
10.7 Multiple Partial Correlation	212
10.8 Concluding Remarks	214
Problems	215
References	225

11 Confounding and Interaction in Regression 226

11.1 Preview	226
11.2 Overview	226
11.3 Interaction in Regression	228
11.4 Confounding in Regression	236
11.5 Summary and Conclusions	242
Problems	242
References	256

12 Dummy Variables in Regression 257

12.1 Preview	257
12.2 Definitions	257
12.3 Rule for Defining Dummy Variables	258
12.4 Comparing Two Straight-line Regression Equations: An Example	259
12.5 Questions for Comparing Two Straight Lines	261
12.6 Methods of Comparing Two Straight Lines	262
12.7 Method I: Using Separate Regression Fits to Compare Two Straight Lines	263
12.8 Method II: Using a Single Regression Equation to Compare Two Straight Lines	268
12.9 Comparison of Methods I and II	271
12.10 Testing Strategies and Interpretation: Comparing Two Straight Lines	272
12.11 Other Dummy Variable Models	273
12.12 Comparing Four Regression Equations	275
12.13 Comparing Several Regression Equations Involving Two Nominal Variables	277
Problems	283
References	307

13 Analysis of Covariance and Other Methods for Adjusting Continuous Data 308

13.1 Preview	308
13.2 Adjustment Problem	309

13.3	Analysis of Covariance	310
13.4	Assumption of Parallelism: A Potential Drawback	313
13.5	Analysis of Covariance: Several Groups and Several Covariates	314
13.6	Analysis of Covariance: Several Nominal Independent Variables	316
13.7	Comments and Cautions	318
13.8	Summary	321
	Problems	321
	References	338

14 Regression Diagnostics

339

14.1	Preview	339
14.2	Simple Approaches to Diagnosing Problems in Data	340
14.3	Residual Analysis: Detecting Outliers and Violations of Model Assumptions	347
14.4	Strategies for Addressing Violations of Regression Assumptions	355
14.5	Collinearity	358
14.6	Diagnostics Example	372
	Problems	382
	References	399

15 Polynomial Regression

401

15.1	Preview	401
15.2	Polynomial Models	402
15.3	Least-squares Procedure for Fitting a Parabola	402
15.4	ANOVA Table for Second-order Polynomial Regression	404
15.5	Inferences Associated with Second-order Polynomial Regression	405
15.6	Example Requiring a Second-order Model	406
15.7	Fitting and Testing Higher-order Models	410
15.8	Lack-of-fit Tests	410
15.9	Orthogonal Polynomials	412
15.10	Strategies for Choosing a Polynomial Model	422
	Problems	423

16 Selecting the Best Regression Equation

438

16.1	Preview	438
16.2	Steps in Selecting the Best Regression Equation: Prediction Goal	439
16.3	Step 1: Specifying the Maximum Model: Prediction Goal	439
16.4	Step 2: Specifying a Criterion for Selecting a Model: Prediction Goal	442
16.5	Step 3: Specifying a Strategy for Selecting Variables: Prediction Goal	444
16.6	Step 4: Conducting the Analysis: Prediction Goal	454

16.7 Step 5: Evaluating Reliability with Split Samples: Prediction Goal	454
16.8 Example Analysis of Actual Data	457
16.9 Selecting the Most Valid Model	463
Problems	466
References	480

17 One-way Analysis of Variance

481

17.1 Preview	481
17.2 One-way ANOVA: The Problem, Assumptions, and Data Configuration	484
17.3 Methodology for One-way Fixed-effects ANOVA	488
17.4 Regression Model for Fixed-effects One-way ANOVA	494
17.5 Fixed-effects Model for One-way ANOVA	497
17.6 Random-effects Model for One-way ANOVA	500
17.7 Multiple-comparison Procedures for Fixed-effects One-way ANOVA	503
17.8 Choosing a Multiple-comparison Technique	515
17.9 Orthogonal Contrasts and Partitioning an ANOVA Sum of Squares	516
Problems	522
References	543

18 Randomized Blocks: Special Case of Two-way ANOVA

545

18.1 Preview	545
18.2 Equivalent Analysis of a Matched-pairs Experiment	549
18.3 Principle of Blocking	553
18.4 Analysis of a Randomized-blocks Study	555
18.5 ANOVA Table for a Randomized-blocks Study	557
18.6 Regression Models for a Randomized-blocks Study	561
18.7 Fixed-effects ANOVA Model for a Randomized-blocks Study	565
Problems	566
References	578

19 Two-way ANOVA with Equal Cell Numbers

579

19.1 Preview	579
19.2 Using a Table of Cell Means	581
19.3 General Methodology	586
19.4 <i>F</i> Tests for Two-way ANOVA	592
19.5 Regression Model for Fixed-effects Two-way ANOVA	594

19.6	Interactions in Two-way ANOVA	599
19.7	Random- and Mixed-effects Two-way ANOVA Models	607
	Problems	610
	References	629

20 Two-way ANOVA with Unequal Cell Numbers 630

20.1	Preview	630
20.2	Presentation of Data for Two-way ANOVA: Unequal Cell Numbers	630
20.3	Problem with Unequal Cell Numbers: Nonorthogonality	632
20.4	Regression Approach for Unequal Cell Sample Sizes	637
20.5	Higher-way ANOVA	641
	Problems	642
	References	659

21 The Method of Maximum Likelihood 661

21.1	Preview	661
21.2	The Principle of Maximum Likelihood	661
21.3	Statistical Inference Using Maximum Likelihood	665
21.4	Summary	677
	Problems	678
	References	680

22 Logistic Regression Analysis 681

22.1	Preview	681
22.2	The Logistic Model	681
22.3	Estimating the Odds Ratio Using Logistic Regression	683
22.4	A Numerical Example of Logistic Regression	689
22.5	Theoretical Considerations	698
22.6	An Example of Conditional ML Estimation Involving Pair-matched Data with Unmatched Covariates	704
22.7	Summary	707
	Problems	708
	References	712

23 Polytomous and Ordinal Logistic Regression 714

23.1	Preview	714
23.2	Why Not Use Binary Regression?	715

23.3 An Example of Polytomous Logistic Regression: One Predictor, Three Outcome Categories	715
23.4 An Example: Extending the Polytomous Logistic Model to Several Predictors	721
23.5 Ordinal Logistic Regression: Overview	726
23.6 A “Simple” Example: Three Ordinal Categories and One Dichotomous Exposure Variable	727
23.7 Ordinal Logistic Regression Example Using Real Data with Four Ordinal Categories and Three Predictor Variables	731
23.8 Summary	737
Problems	738
References	742

24 Poisson Regression Analysis

743

24.1 Preview	743
24.2 The Poisson Distribution	743
24.3 An Example of Poisson Regression	745
24.4 Poisson Regression	748
24.5 Measures of Goodness of Fit	753
24.6 Continuation of Skin Cancer Data Example	756
24.7 A Second Illustration of Poisson Regression Analysis	762
24.8 Summary	765
Problems	766
References	780

25 Analysis of Correlated Data Part 1: The General Linear Mixed Model

781

25.1 Preview	781
25.2 Examples	784
25.3 General Linear Mixed Model Approach	792
25.4 Example: Study of Effects of an Air Pollution Episode on FEV1 Levels	806
25.5 Summary—Analysis of Correlated Data: Part 1	818
Problems	819
References	824

26 Analysis of Correlated Data Part 2: Random Effects and Other Issues

825

26.1 Preview	825
26.2 Random Effects Revisited	825

26.3	Results for Models with Random Effects Applied to Air Pollution Study Data	829
26.4	Second Example—Analysis of Posture Measurement Data	839
26.5	Recommendations about Choice of Correlation Structure	859
26.6	Analysis of Data for Discrete Outcomes	861
	Problems	862
	References	882

27 Sample Size Planning for Linear and Logistic Regression and Analysis of Variance

883

27.1	Preview	883
27.2	Review: Sample Size Calculations for Comparisons of Means and Proportions	884
27.3	Sample Size Planning for Linear Regression	886
27.4	Sample Size Planning for Logistic Regression	889
27.5	Power and Sample Size Determination for Linear Models: A General Approach	893
27.6	Sample Size Determination for Matched Case-control Studies with a Dichotomous Outcome	908
27.7	Practical Considerations and Cautions	910
	Problems	911
	References	913

Appendix A—Tables

915

A.1	Standard Normal Cumulative Probabilities	916
A.2	Percentiles of the <i>t</i> Distribution	919
A.3	Percentiles of the Chi-square Distribution	920
A.4	Percentiles of the <i>F</i> Distribution	921
A.5	Values of $\frac{1}{2} \ln \frac{1+r}{1-r}$	928
A.6	Upper α Point of Studentized Range	930
A.7	Orthogonal Polynomial Coefficients	932
A.8A	Bonferroni Corrected Jackknife Residual Critical Values	933
A.8B	Bonferroni Corrected Studentized Residual Critical Values	933
A.9	Critical Values for Leverages	934
A.10	Critical Values for the Maximum of n Values of Cook's $(n - k - 1) d_i$	936

Appendix B—Matrices and Their Relationship to Regression Analysis

937

Appendix C—SAS Computer Appendix	949
---	------------

Appendix D—Answers to Selected Problems	991
--	------------

Index	1037
--------------	-------------

PREFACE

This is the fourth revision of our second-level statistics text, originally published in 1978 and revised in 1987, 1998, and 2008. As with previous versions, this text is intended primarily for advanced undergraduates, graduate students, and working professionals in the health, social, biological, and behavioral sciences who engage in applied research in their fields. The text may also provide professional statisticians with some new insights into the application of advanced statistical techniques to realistic research problems.

We have attempted in this revision to retain the basic structure and flavor of the earlier editions, while at the same time making changes to keep pace with current analytic practices and computer usage in applied research. Notable changes in this fifth edition, discussed in more detail later, include

- i. Clarification of content and/or terminology as suggested by reviewers and readers, including revision of variable and subscript notation used for predictor variables and regression coefficients to provide consistency over different chapters.
- ii. Expanded and updated coverage of some content areas (e.g., confounding and interaction in regression in Chapter 11, selecting the best regression equation in Chapter 16, sample size determination in Chapter 27).
- iii. A new linear regression example that is carried through and expanded upon in Chapters 5, 6, 8, 9, 11, 12, 13, and 16.
- iv. Some new exercises at the end of selected chapters, including exercises related to the new example described in item (iii) above.
- v. Updated SAS computer output using SAS 9.3 that reflects improvements in output styling.
- vi. Two computer appendices on programming procedures for multiple linear regression models, logistic regression models, Poisson regression models, and mixed linear models:
 - a. In-text: SAS
 - b. Online: SPSS, STATA, and R

In this fifth edition, as in our previous versions, we emphasize the intuitive logic and assumptions that underlie the techniques covered, the purposes for which these techniques are designed, the advantages and disadvantages of these techniques, and valid interpretations

based on these techniques. Although we describe the statistical calculations required for the techniques we cover, we rely on computer output to provide the results of such calculations so the reader can concentrate on how to apply a given technique rather than how to carry out the calculations. The mathematical formulas that we do present require no more than simple algebraic manipulations. Proofs are of secondary importance and are generally omitted. Calculus is not explicitly used anywhere in the main text. We introduce matrix notation to a limited extent in Chapters 25 and 26 because we believe that the use of matrices provides a more convenient way to understand some of the complicated mathematical aspects of the analysis of correlated data. We also have continued to include an appendix on matrices for the interested reader.

This edition, as with the previous editions, is *not* intended to be a general reference work dealing with all the statistical techniques available for analyzing data involving several variables. Instead, we focus on the techniques we consider most essential for use in applied research. We want the reader to understand the concepts and assumptions involved in these techniques and how these techniques can be applied in practice, including how computer packages can help make it easier to perform the analysis of one's data.

The most notable features of this fifth edition, including the material that has not been modified from the previous edition, are the following:

1. Regression analysis (Chapters 1–16) and analysis of variance (Chapters 17–20) are discussed in considerable detail and with pedagogical care that reflects the authors' extensive experience and insight as teachers of such material.
2. A new linear regression example based on a complex survey design is carried through and expanded upon in several chapters, including new exercises involving the dataset for this example. To obtain the most valid estimates of regression coefficients, weighting and stratification schemes involved in the survey design should be taken into account. Although it is beyond the scope of this text to describe regression methods for analyzing complex survey designs, we discuss and illustrate the extent to which results from using such "weighted" methods may differ from results from using the "unweighted" methods emphasized in this text.
3. The relationship between regression analysis and analysis of variance is highlighted.
4. The connection between multiple regression analysis and multiple and partial correlation analysis is discussed in detail.
5. Several advanced topics are presented in a unique, nonmathematical manner, including chapters on maximum likelihood (ML) methods (21), binary logistic regression (22), polytomous and ordinal logistic regression (23), and Poisson regression (24) and two chapters (25–26) on the analysis of correlated data (described further below). The material on ML methods in Chapters 21–26 provides a strong foundation for understanding why ML estimation is the most widely used method for fitting mathematical models involving several variables.
6. An up-to-date discussion of the issues and procedures involved in fine-tuning a regression analysis is presented on confounding and interaction in regression (Chapter 11), selecting the best regression model (Chapter 16), and regression diagnostics (Chapter 14).

7. Chapter 23 on polytomous and ordinal logistic regression methods extends the standard (binary) logistic model to outcome variables that have more than two categories. Polytomous logistic regression is used when the outcome categories do not have any natural order, whereas ordinal logistic regression is appropriate when the outcome categories have a natural order.
8. Chapters 25 and 26 on the analysis of correlated data describe the ML/REML linear mixed model approach incorporated into SAS's MIXED procedure. Since ML estimation is assumed, these chapters are logically ordered after the current Chapter 21 on ML estimation. In Chapter 25, we describe the general form of the linear mixed model, introduce the terms *correlated structure* and *robust/empirical standard errors*, and illustrate how to model correlated data when only fixed effects are considered. In Chapter 26, which serves as Part 2 of this topic, we focus on linear mixed models that contain random effects. Chapter 26 also provides a link to ANOVA Chapters 17–20, alternatively formulating the linear mixed model approach in terms of an ANOVA that partitions sources of variation from various predictors into the sums of squares and corresponding mean square terms of a summary ANOVA table.
9. Chapter 27 on sample size determination for linear and logistic regression models describes two approaches for sample size calculation, the first being an approximate approach that yields fairly accurate sample sizes and requires only manual computation. The second approach is based on more traditional theory for sample size determination and is best implemented using computer software. This chapter has been updated to reflect updated SAS 9.3 and PASS 11 output.
10. Representative computer results from SAS 9.3 are used to illustrate concepts in the body of the text, as well as to provide a basis for exercises for the reader. In this edition, we revised the computer output to reflect the most recent version of SAS, and, in many instances, we annotated comments on the output so that it is easier to read.
11. Numerous examples and exercises illustrate applications to real studies in a wide variety of disciplines. New exercises have been added to several chapters.
12. Solutions to selected exercises are provided in Appendix D. An Instructor's Solutions Manual containing solutions to all exercises is also available with the fifth edition. In addition, a Student Solutions Manual containing complete solutions to selected problems is available for students.
13. Computer Appendix C is a new addition to the text that describes how to use Version 9.3 of SAS to carry out linear regression, logistic regression, Poisson regression, and correlated data analysis of linear models.
14. Links to freely downloadable datasets; a computer appendix on the use of STATA, SPSS, and R packages to carry out linear regression modeling; updates on errata; and other information are available at CengageBrain.com.
15. The computer appendix mentioned in item (14) will be a freely downloadable electronic document providing computer guidelines for multiple linear regression models. (Other textbooks by Kleinbaum and Klein have computer appendices for SAS, STATA, and SPSS use with logistic models and Cox proportional

hazards and extended Cox models for survival data.) The computer appendix will provide a quick and easy reference guide to help the reader avoid having to spend a lot of time finding information from sometimes confusing help guides in packages like SAS.

Suggestions for Instructors or Individual Learners

For formal classroom instruction and/or individual/distance learning, the chapters fall naturally into four clusters:

Course 1: Chapters 4–16, on linear regression analysis

Course 2: Chapters 17–20, on the analysis of variance

Course 3: Chapters 21–24, on maximum likelihood methods and important applications involving logistic and Poisson regression modeling

Course 4: Chapters 25–26, on the analysis of correlated data involving linear mixed models

Portions of Chapter 27 on sample size determination could be added, as appropriate, to Courses 1–3 above. Courses 1 and 2 have often been combined into one course on regression and ANOVA methods. For a first course in regression analysis, some of Chapters 11 through 16 may be considered too specialized. For example, Chapter 15 on selecting the best regression model and Chapter 16 on regression diagnostics might be used in a continuation course on regression modeling, which might also include some of the advanced topics covered in Chapters 21–27.

Acknowledgments

We wish to acknowledge several people who contributed to the development of this text, including early editions as well as this fifth edition. Drs. Kleinbaum and Kupper continue to be indebted to John Cassel and Bernard Greenberg, two mentors who have provided us with inspiration and the professional and administrative guidance that enabled us at the beginning of our careers to gain the broad experience necessary to write this text.

Dr. Kleinbaum also wishes to thank John Boring, former Chair of the Department of Epidemiology at Emory University, for his strong support and encouragement during the writing of the third and fourth editions and for his deep commitment to teaching excellence. Dr. Kleinbaum also wishes to thank Dr. Mitch Klein of Emory's Department of Epidemiology for his colleagueship, including thoughtful suggestions on and review of previous editions. Dr. Kleinbaum also thanks Dr. Viola Vaccarino, Chair of the Department of Epidemiology at Emory University, for continued support and encouragement of his academic life at the Rollins School of Public Health at Emory University.

Dr. Kupper will forever be indebted to Dr. William Mendenhall, founder and longtime Chair of the University of Florida Department of Statistics. Dr. Mendenhall gave Dr. Kupper his start in the field of statistics, and he served as a perfect example of an inspiring teacher and a caring mentor.

Mr. Nizam wishes to thank Dr. Lance Waller, Chair of the Department of Biostatistics and Bioinformatics at Emory University, for his strong support and Dr. John Spurrier of the Department of Statistics at the University of South Carolina for being a wonderful teacher, advisor, and mentor.

We thank Julia Labadie for her assistance in preparing SAS computer output for this edition. We also thank Dr. Keith Muller for his contributions to earlier editions as one of our coauthors.

We thank our spouses—Edna Kleinbaum, Sandy Martin, Janet Nizam, and Abby Horowitz—for their encouragement and support during the writing of various revisions.

We thank our reviewers of the fifth edition for their helpful suggestions:

Joseph Glaz, University of Connecticut

Lynn Kuo, University of Connecticut

Robert Paige, Missouri University of Science and Technology

Debaraj Sen, Concordia University

Po Yang, DePaul University

We thank the Cengage Learning Statistics and Mathematics team, especially Molly Taylor, Senior Product Manager, and Laura Wheel, Senior Content Developer, for guiding us through the publication process for the fifth edition, as well as Jessica Rasile, Content Project Manager, and Tania Andrabi, Production Manager.

David G. Kleinbaum

Lawrence L. Kupper

Azhar Nizam

Eli S. Rosenberg

1

Concepts and Examples of Research

1.1 Concepts

The purpose of most empirical research is to assess relationships among a set of variables, which are factors that are distinctly measured on observational units (or subjects). *Multivariable*¹ techniques are concerned with the statistical analysis of such relationships, particularly when at least three variables are involved. Regression analysis, our primary focus, is one type of multivariable technique. Other techniques will also be described in this text. Choosing an appropriate technique depends on the purpose of the research and on the types of variables under investigation (a subject discussed in Chapter 2).

Research may be classified broadly into three types: *experimental*, *quasi-experimental*, or *observational*. Multivariable techniques are applicable to all such types, yet the confidence one may reasonably have in the results of a study can vary with the research type. In most types, one variable is usually taken to be a *response* or *dependent variable*—that is, a variable to be predicted from other variables. The other variables are called *predictor* or *independent variables*.

If observational units (subjects) are randomly assigned to levels of important predictors, the study is usually classified as an *experiment*. Experiments are the most controlled type of study; they maximize the investigator's ability to isolate the observed effect of the predictors from the distorting effects of other (independent) variables that might also be related to the response.

¹ The term *multivariable* is preferable to *multivariate*. Statisticians generally use the term *multivariate analysis* to describe a method in which several dependent variables can be considered simultaneously. Researchers in the biomedical and health sciences who are not statisticians, however, use this term to describe any statistical technique involving several variables, even if only one dependent variable is considered at a time. In this text, we prefer to avoid the confusion by using the term *multivariable analysis* to denote the latter, more general description.

If subjects are assigned to treatment conditions without randomization, the study is called *quasi-experimental* (Campbell and Stanley 1963). Such studies are often more feasible and less expensive than experimental studies, but they provide less control over the study situation.

Finally, if all observations are obtained without either randomization or artificial manipulation (i.e., allocation) of the predictor variables, the study is said to be *observational*. Experiments offer the greatest potential for drawing definitive conclusions, and observational studies the least; however, experiments are the most difficult studies to implement, and observational studies the easiest. A researcher must consider this trade-off between interpretive potential and complexity of design when choosing among types of studies (Kleinbaum, Kupper, and Morgenstern 1982, Chapter 3).

To assess a relationship between two variables, one must measure both of them in some manner. Measurement inherently and unavoidably involves error. The need for statistical design and analysis emanates from the presence of such error. Traditionally, statistical inference has been divided into two kinds: estimation and hypothesis testing. *Estimation* refers to describing (i.e., quantifying) characteristics and strengths of relationships. *Testing* refers to specifying hypotheses about relationships, making statements of probability about the reasonableness of such hypotheses, and then providing practical conclusions based on such statements.

This text focuses on regression and correlation methods involving one response variable and one or more predictor variables. In these methods, a mathematical model is specified that describes how the variables of interest are related to one another. The model must somehow be developed from study data, after which inference-making procedures (e.g., testing hypotheses and constructing confidence intervals) are conducted about important parameters of interest. Although other multivariable regression methods will be discussed, linear regression techniques are emphasized for three reasons: they have wide applicability; they can be the most straightforward to implement; and other, more complex statistical procedures can be better appreciated once linear regression methods are understood.

1.2 Examples

The examples that follow concern *real* problems from a variety of disciplines and involve variables to which the methods described in this book can be applied. We shall return to these examples later when illustrating various methods of multivariable analysis.

- **Example 1.1** Study of the associations among the physician–patient relationship, perception of pregnancy, and outcome of pregnancy, illustrating the use of regression analysis and logistic regression analysis.

Thompson (1972) and Hulka and others (1971) looked at both the process and the outcomes of medical care in a cohort of 107 pregnant married women in North Carolina. The data were obtained through patient interviews, questionnaires completed by physicians, and a review of medical records. Several variables were recorded for each patient.

One research goal of primary interest was to determine what association, if any, existed between satisfaction with medical care and a number of variables meant to describe patient

perception of pregnancy and the physician–patient relationship. Three perception-of-pregnancy variables measured the patient’s worry during pregnancy, her desire for the baby, and her concern about childbirth. Two other variables measured the physician–patient relationship in terms of informational communication concerning prescriptions and affective communication concerning perceptions. Other variables considered were age, social class, education, and parity.

Regression analysis was used to describe the relationship between scores measuring patient satisfaction with medical care and the preceding variables. From this analysis, variables found not to be related to medical care satisfaction could be eliminated, while those found to be associated with satisfaction could be ranked in order of importance. Also, the effects of confounding variables such as age and social class could be considered, to three ends: any associations found could not be attributed solely to such variables; measures of the strength of the relationship between satisfaction and other variables could be obtained; and a functional equation predicting level of patient satisfaction in terms of the other variables found to be important in describing satisfaction could be developed.

Another question of interest in this study was whether patient perception of pregnancy and/or the physician–patient relationship was associated with complications of pregnancy. A variable describing complications was defined so that the value 1 could be assigned if the patient experienced one or more complications of pregnancy and 0 if she experienced no complications. *Logistic regression analysis* was used to evaluate the relationship between the occurrence of complications of pregnancy and other variables. This method, like regression analysis, allows the researcher to determine and rank important variables that can distinguish between patients who have complications and patients who do not. ■

■ **Example 1.2** Study of race and social influence in cooperative problem-solving dyads, illustrating the use of analysis of variance and analysis of covariance.

James (1973) conducted an experiment on 140 seventh- and eighth-grade males to investigate the effects of two factors—race of the experimenter (E) and race of the comparison norm (N)—on social influence behaviors in three types of dyads: white–white; black–black; and white–black. Subjects played a game of strategy called Kill the Bull, in which 14 separate decisions must be made for proceeding toward a defined goal on a game board. In the game, each pair of players (dyad) must reach a consensus on a direction at each decision step, after which they signal the E, who then rolls a die to determine how far they can advance along their chosen path of six squares. Photographs of the current champion players (N) (either two black youths [black norm] or two white youths [white norm]) were placed above the game board.

Four measures of social influence activity were used as the outcome variables of interest. One of these, called performance output, was a measure of the number of times a given subject attempted to influence his dyad to move in a particular direction.

The major research question focused on the outcomes for biracial dyads. Previous research of this type had used only white investigators and implicit white comparison norms, and the results indicated that the white partner tended to dominate the decision making. James’s study sought to determine whether such an “interaction disability,” previously attributed to blacks, would be maintained, removed, or reversed when the comparison norm, the experimenter, or both were black. One approach to analyzing this problem was to

perform a *two-way analysis of variance* on social-influence-activity difference scores between black and white partners, to assess whether such differences were affected by either the race of E or the race of N. No such significant effects were found, however, implying that neither E nor N influenced biracial dyad interaction. Nevertheless, through use of *analysis of covariance*, it was shown that, controlling for factors such as age, height, grade, and verbal and mathematical test scores, there was no statistical evidence of white dominance in any of the experimental conditions.

Furthermore, when combined output scores for both subjects in same-race dyads (white-white or black-black) were analyzed using a *three-way analysis of variance* (the three factors being race of dyad, race of E, and race of N), subjects in all-black dyads were found to be more verbally active (i.e., exhibited a greater tendency to influence decisions) under a black E than under a white E; the same result was found for white dyads under a white E. This property is generally referred to in statistical jargon as a “race of dyad” by “race of E” interaction. The property continued to hold up after *analysis of covariance* was used to control for the effects of age, height, and verbal and mathematical test scores. ■

■ Example 1.3 Study of the relationship of cultural change to health, illustrating the use of analysis of variance.

Patrick and others (1974) studied the effects of cultural change on health in the U.S. Trust Territory island of Ponape. Medical and sociological data were obtained on a sample of about 2,000 people by means of physical exams and a sociological questionnaire. This Micronesian island has experienced rapid Westernization and modernization since American occupation in 1945. The question of primary interest was whether rapid social and cultural change caused increases in blood pressure and in the incidence of coronary heart disease. A specific hypothesis guiding the research was that persons with high levels of cultural ambiguity and incongruity and low levels of supportive affiliations with others have high levels of blood pressure and are at high risk for coronary heart disease.

A preliminary step in the evaluation of this hypothesis involved measuring three variables: attitude toward modern life; preparation for modern life; and involvement in modern life. Each of these variables was created by isolating specific questions from a sociological questionnaire. Then a *factor analysis*² determined how best to combine the scores on specific questions into a single overall score that defined the variable under consideration. Two cultural incongruity variables were then defined. One involved the discrepancy between attitude toward modern life and involvement in modern life; the other was defined as the discrepancy between preparation for modern life and involvement in modern life.

These variables were then analyzed to determine their relationship, if any, to blood pressure and coronary heart disease. Individuals with large positive or negative scores on either of the two incongruity variables were hypothesized to have high blood pressure and to be at high risk for coronary heart disease.

One approach to analysis involved categorizing both discrepancy scores into high and low groups. Then a *two-way analysis of variance* could be performed using blood pressure

² Factor analysis was described in Chapter 24 of the second edition of this text, but this topic is not included as a topic in this (fifth) edition.

as the outcome variable. We will see later that this problem can also be described as a regression problem. ■

■ **Example 1.4** Study of the association between alcohol consumption frequency and body-mass index (BMI) in the Behavioral Risk Factor Surveillance System (BRFSS).

The BRFSS is a large and ongoing surveillance project managed by the U.S. Centers for Disease Control and Prevention (CDC) and conducted by state health departments as telephone-based interviews, based on random-digit dialing. Its purpose is to “generate information about health risk behaviors, clinical preventive practices, and health care access and use primarily related to chronic diseases and injury”(CDC 2012).

The unpublished example considered here examines the relationship between frequency of alcohol use in the previous 30 days and the response variable of BMI, a common measure of body fat defined as $(\text{weight in kg})/(\text{height in m})^2$. Dozens of studies have demonstrated cardiovascular benefits of red wine consumption. Yet the relationship between alcohol consumption and BMI, an important risk factor for numerous chronic diseases, is less clear. An analysis of data from the National Health Interview Survey found a moderate reduction in BMI associated with increasing drinking frequency, yet an increase in BMI with greater drinking volume (Breslow and Smothers 2005). These relationships were different for males and females (an example of *interaction*; see Chapter 11), who are known to metabolize alcohol differently.

This analysis of drinking frequency and BMI considers females who live in the state of Georgia and who consume nonheavy amounts of alcohol (for the 2010 BRFSS data collection year). *Straight-line regression analysis* is used to quantify the same negative association between drinking frequency and BMI found by others. *Multiple regression analysis* and *analysis of covariance* are used to additionally consider the effects of age and other health behaviors (e.g., sleep quality, exercise, and tobacco use) that are known to be associated with BMI.

This example is unique in that it provides key illustrations of the objectives of regression techniques for the analysis of public health surveillance data on a health outcome with numerous determinants. These objectives can differ from those used for the analysis of data emanating from more controlled health studies (such as randomized controlled clinical trials). In particular, the large sample size associated with the BRFSS provides opportunities for the detection of statistically significant (and sometimes both unexpected and meaningful) associations between certain determinants and BMI that might otherwise be challenging to detect. Such hypothesis-generating regression findings can suggest avenues for further research. It is important to mention that such surveillance studies limit causal interpretations of the findings. These and related issues are discussed further in several chapters that follow.

1.3 Concluding Remarks

The four examples described in Section 1.2 indicate the variety of research questions to which multivariable statistical methods are applicable. In Chapter 2, we will provide a broad overview of such techniques; in the remaining chapters, we will discuss each technique in detail.

References

- Breslow, R. A., and Smothers, B. A. 2005. "Drinking Patterns and Body Mass Index in Never Smokers: National Health Interview Survey, 1997–2001." *American Journal of Epidemiology* 161(4): 368–76.
- Campbell, D. T., and Stanley, J. C. 1963. *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand McNally.
- CDC Office of Surveillance, Epidemiology, and Laboratory Services. 2012. "Behavioral Risk Factor Surveillance System: BRFSS Frequently Asked Questions (FAQs)." <http://www.cdc.gov/brfss/faqs.htm>.
- Hulka, B. S.; Kupper, L. L.; Cassel, J. C.; and Thompson, S. J. 1971. "A Method for Measuring Physicians' Awareness of Patients' Concerns." *HSMHA Health Reports* 86: 741–51.
- James, S. A. 1973. "The Effects of the Race of Experimenter and Race of Comparison Norm on Social Influence in Same Race and Biracial Problem-Solving Dyads." Ph.D. dissertation, Department of Clinical Psychology, Washington University, St. Louis, Mo.
- Kleinbaum, D. G.; Kupper, L. L.; and Morgenstern, H. 1982. *Epidemiologic Research*. Belmont, Calif.: Lifetime Learning Publications.
- Patrick, R.; Cassel, J. C.; Tyroler, H. A.; Stanley, L.; and Wild, J. 1974. "The Ponape Study of Health Effects of Cultural Change." Paper presented at the annual meeting of the Society for Epidemiologic Research, Berkeley, Calif.
- Thompson, S. J. 1972. "The Doctor–Patient Relationship and Outcomes of Pregnancy." Ph.D. dissertation, Department of Epidemiology, University of North Carolina, Chapel Hill.

2

Classification of Variables and the Choice of Analysis

2.1 Classification of Variables

Variables can be classified in a number of ways. Such classifications are useful for determining which method of data analysis to use. In this section, we describe three methods of classification: by gappiness, by level of measurement, and by descriptive orientation.

2.1.1 Gappiness

In the classification scheme we call *gappiness*, we determine whether gaps exist between successively observed values of a variable (Figure 2.1). If gaps exist between observations, the variable is said to be *discrete*; if no gaps exist, the variable is said to be *continuous*. To speak more precisely, a variable is discrete if, between any two potentially observable values, a value exists that is not possibly observable. A variable is continuous if, between any two potentially observable values, another potentially observable value exists.

Examples of continuous variables are age, blood pressure, cholesterol level, height, and weight. Discrete variables are often counts, such as of the numbers of deaths or car accidents. Additionally, nonnumeric information is often numerically coded in data sources using discrete variables. Examples of this are sex (e.g., 0 if male and 1 if female), group identification (e.g., 1 if group A and 2 if group B), and state of disease (e.g., 1 if a coronary heart disease case and 0 if not a coronary heart disease case).

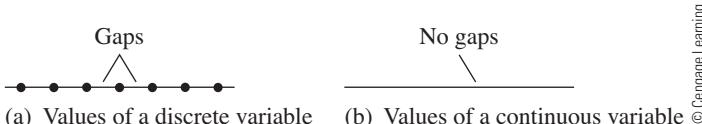
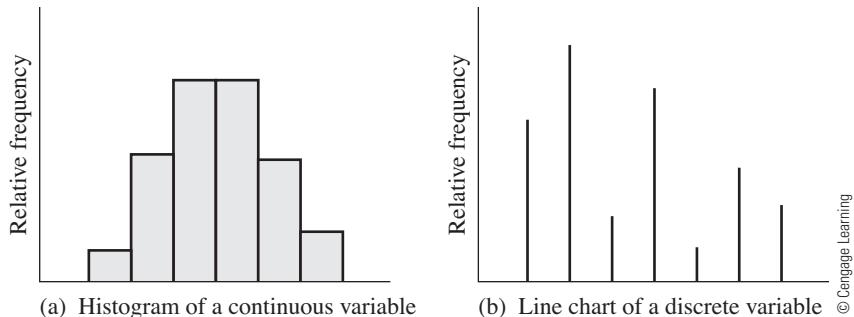


FIGURE 2.1 Discrete versus continuous variables

**FIGURE 2.2** Sample frequency distributions of a continuous and a discrete variable

In analyses of actual data, the sampling frequency distributions for continuous variables are represented differently from those for discrete variables. Data on a continuous variable are usually *grouped* into class intervals, and a relative frequency distribution is determined by counting the proportion of observations in each interval. Such a distribution is usually represented by a histogram, as shown in Figure 2.2(a). Data on a discrete variable, on the other hand, are usually not grouped but are represented instead by a line chart, as shown in Figure 2.2(b).

Discrete variables can sometimes be treated for analysis purposes as continuous variables. This is possible when the values of such a variable, even though discrete, are not far apart and cover a wide range of numbers. In such a case, the possible values, although technically gappy, show such small gaps between values that a visual representation would approximate an interval (Figure 2.3).

Furthermore, a line chart, like the one in Figure 2.2(b), representing the frequency distribution of data on such a variable would probably show few frequencies greater than 1 and thus would be uninformative. As an example, the variable “social class” is usually measured as discrete; one measure of social class¹ takes on integer values between 11 and 77. When data on this variable are grouped into classes (e.g., 11–15, 16–20, etc.), the resulting frequency histogram gives a clearer picture of the characteristics of the variable than a line chart does. Thus, in this case, treating social class as a continuous variable is sometimes more useful than treating it as discrete.

Just as it is often useful to treat a discrete variable as continuous, some fundamentally continuous variables may be grouped into categories and treated as discrete variables in a given analysis. For example, the variable “age” can be made discrete by grouping its values into two categories, “young” and “old.” Similarly, “blood pressure” becomes a discrete variable if it is categorized into “low,” “medium,” and “high” groups or into deciles.

**FIGURE 2.3** Discrete variable that may be treated as continuous (© Cengage Learning)

¹ Hollingshead’s “Two-Factor Index of Social Position,” a description of which can be found in Green (1970).

The decision to categorize a continuous variable into discrete levels is nuanced, requiring consideration of both pros and cons. On the one hand, a discrete version of a variable might make the data easier to collect and summarize. This often, in turn, aids in the presentation of results to colleagues. Yet these advantages must be balanced against the loss of information that comes with converting a continuous variable into a discrete one. The choice of variable type often impacts the type of analysis that can ultimately be conducted, and the desire to use a certain analysis technique may drive decisions about the treatment of variables.

A further consideration concerns when to categorize continuous data. One may categorize a continuous variable either at the time of data collection or at the time of data analysis. The former choice often allows cheaper, quicker, and/or less precise methodology for data collection to be employed. Yet this may also introduce human error (e.g., when a clinician is given the extra step of classifying a continuous reading into one of several groups). Categorization at the time of analysis reduces the likelihood of human error and also allows for multiple classification schemes to be later considered, since the original continuous data have not been forfeited.

A related issue is that both continuous and discrete variables can be error-prone. Continuous variables can be measured with error, and discrete variables can be misclassified. When such error-prone variables are used in regression analyses, incorrect statistical conclusions can be made (i.e., statistical validity can be compromised). In this textbook, it will be assumed that variables to be considered are not subject to either measurement error or misclassification error. A discussion of rigorous statistical methods for dealing with error-prone variables in regression analyses is beyond the scope of this textbook, but Gustafson (2004) provides numerous relevant references to such methods.

2.1.2 Level of Measurement

A second classification scheme deals with the preciseness of measurement of the variable. There are three such levels: nominal, ordinal, and interval.

The numerically weakest level of measurement is the *nominal* level. At this level, the values assumed by a variable simply indicate different categories. The variable “sex,” for example, is nominal: by assigning the numbers 1 and 0 to denote male and female, respectively, we distinguish the two sex categories. A variable that describes treatment group is also nominal, provided that the treatments involved cannot be ranked according to some criterion (e.g., dosage level).

A somewhat higher level of measurement allows not only *grouping* into separate categories but also *ordering* of categories. This level is called *ordinal*. The treatment group may be considered ordinal if, for example, different treatments differ by dosage. In this case, we could tell not only which treatment group an individual falls into but also who received a heavier dose of the treatment. Social class is another ordinal variable, since an ordering can be made among its different categories. For example, all members of the upper middle class are higher in some sense than all members of the lower middle class.

A limitation—perhaps debatable—in the preciseness of a measurement such as social class is the amount of information supplied by the magnitude of the differences between different categories. Thus, although upper middle class is higher than lower middle class, it is debatable *how much* higher.

A variable that can give not only an ordering but also a meaningful measure of the distance between categories is called an *interval* variable. To be interval, a variable must be expressed in terms of some standard or well-accepted physical unit of measurement. Height, weight, blood pressure, and number of deaths all satisfy this requirement, whereas subjective measures such as perception of pregnancy, personality type, prestige, and social stress do not.

An interval variable that has a scale with a true zero is occasionally designated as a *ratio* or *ratio-scale variable*. An example of a ratio-scale variable is the height of a person. Temperature is commonly measured in degrees Celsius, an interval scale. Measurement of temperature in degrees Kelvin is based on a scale that begins at absolute zero and thus is a ratio variable. An example of a ratio variable common in health studies is the concentration of a substance (e.g., cholesterol) in the blood.

Ratio-scale variables often involve measurement errors that follow a nonnormal distribution and are proportional to the size of the measurement. We will see in Chapter 5 that such proportional errors violate an important assumption of linear regression—namely, equality of error variance for all observations. Hence, the presence of a ratio variable is a signal to be on guard for a possible violation of this assumption. In Chapter 14 (on regression diagnostics), we will describe methods for detecting and dealing with this problem.

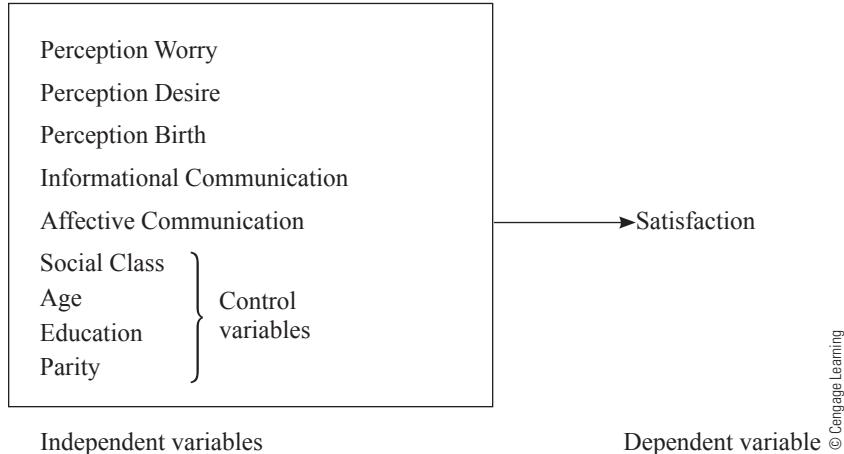
As with variables in other classification schemes, the same variable may be considered at one level of measurement in one analysis and at a different level in another analysis. Thus, “age” may be considered as interval in a regression analysis or, by being grouped into categories, as nominal in an analysis of variance.

The various levels of mathematical precision are cumulative. An ordinal scale possesses all the properties of a nominal scale plus ordinality. An interval scale is also nominal and ordinal. The cumulativeness of these levels allows the researcher to drop back one or more levels of measurement in analyzing the data. Thus, an interval variable may be treated as nominal or ordinal for a particular analysis, and an ordinal variable may be analyzed as nominal.

2.1.3 Descriptive Orientation

A third scheme for classifying variables is based on whether a variable is intended to *describe* or *be described* by other variables. Such a classification depends on the study objectives rather than on the inherent mathematical structure of the variable itself. If the variable under investigation is to be described in terms of other variables, we call it a *response* or *dependent variable*, typically denoted by the letter *Y*. If we are using the variable in conjunction with other variables to describe a given response variable, we call it a *predictor*, a *regressor*, or an *independent variable*,² typically denoted by the letter *X*. Some independent variables may

² The term *independent variable* is a historical term meant to evoke the notion that these measured factors may freely vary from subject to subject, whereas changes in the *dependent variable* are thought to depend on and be determined by the values of a subject's independent variables. This usage of the term *independent* differs from the statistical concept of independence. Two variables are statistically independent when the statistical behavior of one variable is completely unaffected by the statistical behavior of the other variable. When two variables are independent, they are uncorrelated, although zero correlation does not imply independence. In most regression analysis situations, there are nonzero correlations among the independent (or predictor) variables. Though not ideal terminology, the phrase *independent variable* is still commonly used in practice to denote a predictor variable in regression analysis, and we use this standard terminology in this textbook.



© Cengage Learning

FIGURE 2.4 Descriptive orientation for Thompson's (1972) study of satisfaction with medical care

affect relationships among other independent variables and/or the dependent variables but be of no intrinsic interest in a particular study. Such variables may be referred to as *control* or *nuisance variables* or, in some contexts, as *covariates* or *confounders*.

For example, in Thompson's (1972) study of the relationship between patient perception of pregnancy and patient satisfaction with medical care, the perception variables are independent variables (or regressors), and the satisfaction variable is the dependent (or response) variable (Figure 2.4).

Usually, the distinction between independent and dependent variables is clear, as it is in the examples we have given. Nevertheless, a variable considered as dependent for purposes of evaluating one study objective may be considered as independent for purposes of evaluating a different objective. For example, in Thompson's study, in addition to determining the relationship of perceptions as independent variables to patient satisfaction, the researcher sought to determine the relationships of social class, age, and education to perceptions treated as dependent variables.

2.2 Overlapping of Classification Schemes

The three classification schemes described in Section 2.1 overlap in the sense that any variable can be labeled according to each scheme. "Social class," for example, may be considered as ordinal, discrete, and independent in a given study; "blood pressure" may be considered interval, continuous, and dependent in the same or another study.

The overlap between the level-of-measurement classification and the gappiness classification is shown in Figure 2.5. The diagram does not include classification into dependent or independent variables because that dimension is entirely a function of the study objectives

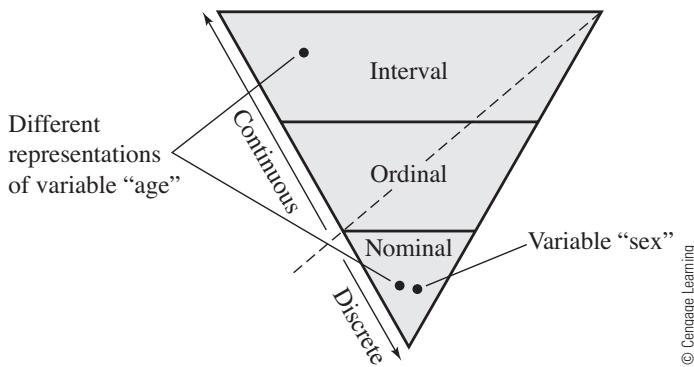


FIGURE 2.5 Overlap of variable classifications

and not of the variable itself. In reading the diagram, one should consider any variable as being representable by some point within the triangle. If the point falls below the dashed line within the triangle, it is classified as discrete; if it falls above that line, it is continuous. Also, a point that falls into the area marked “interval” is classified as an interval variable, and similarly for the other two levels of measurement.

As Figure 2.5 indicates, any nominal variable must be discrete, but a discrete variable may be nominal, ordinal, or interval. Also, a continuous variable must be either ordinal or interval, although ordinal or interval variables may exist that are not continuous. For example, “sex” is nominal and discrete; “age” may be considered interval and continuous or, if grouped into categories, nominal and discrete; and “social class,” depending on how it is measured and on the viewpoint of the researcher, may be considered ordinal and continuous, ordinal and discrete, or nominal and discrete.

2.3 Choice of Analysis

Any researcher faced with the need to analyze data requires a rationale for choosing a particular method of analysis. Four considerations should enter into such a choice: the purpose of the investigation; the mathematical characteristics of the variables involved; the statistical assumptions made about these variables; and how the data are collected (e.g., the sampling procedure). The first two considerations are generally sufficient to determine an appropriate analysis. However, the researcher must consider the latter two items before finalizing initial recommendations.

Here we focus on the use of variable classification, as it relates to the first two considerations noted at the beginning of this section, in choosing an appropriate method of analysis. Table 2.1 provides a rough guide to help the researcher in this choice when several variables are involved. The guide distinguishes among various multivariable methods.

TABLE 2.1 Rough guide to multivariable methods

Classification of Variables			
Method	Dependent	Independent	General Purpose
Multiple linear regression analysis	Continuous	Classically all continuous, but in practice any type(s) can be used	To describe the extent, direction, and strength of the relationship between several independent variables and a continuous dependent variable
Analysis of variance	Continuous	All nominal	To describe the relationship between a continuous dependent variable and one or more nominal independent variables
Analysis of covariance	Continuous	Mixture of nominal variables and continuous variables (the latter used as control variables)*	To describe the relationship between a continuous dependent variable and one or more nominal independent variables, controlling for the effect of one or more continuous independent variables
Logistic regression analysis	Dichotomous	A mixture of various types can be used	To determine how one or more independent variables are related to the probability of the occurrence of one of two possible outcomes
Poisson regression analysis	Discrete	A mixture of various types can be used	To determine how one or more independent variables are related to the rate of occurrence of some outcome

*Generally, a control variable is a variable that must be considered before any relationships of interest can be quantified; this is because a control variable may be related to the variables of primary interest and must be taken into account in studying the relationships among the primary variables. For example, in describing the relationship between blood pressure and physical activity, we would probably consider "age" and "sex" as control variables because they are related to blood pressure and physical activity and, unless taken into account, could confound any conclusions regarding the primary relationship of interest.

© Cengage Learning

It considers the types of variable sets usually associated with each method and gives a general description of the purposes of each method. In addition to using the table, however, one must carefully check the statistical assumptions being made. These assumptions will be described fully later in the text. Table 2.2 shows how these guidelines can be applied to the examples given in Chapter 1.

Several methods for dealing with multivariable problems are *not* included in Table 2.1 or in this text—among them, nonparametric methods of analysis of variance, multivariate multiple regression, and multivariate analysis of variance (which are extensions of the corresponding methods given here that allow for *several* dependent variables), as well as methods of cluster analysis. In this book, we will cover only the multivariable techniques used most often by health and social researchers.

TABLE 2.2 Application of Table 2.1 to examples in Chapter 1

Study	Multivariable Method	Dependent Variable	Independent Variables	Purpose
Example 1.1	Multiple linear regression analysis	Patient satisfaction with medical care, a continuous variable	Mother's desire for baby, worry during pregnancy, concern about childbirth, age, social class, education, and parity; patient–doctor informational and affective communication	To describe the relationship between mother's satisfaction and her desire for baby, worry during pregnancy, etc.
Example 1.1	Logistic regression analysis	Complications of pregnancy (0 = no, 1 = yes), a nominal variable	Same as above	To determine whether and to what extent the independent variables are related to the probability of having pregnancy complications
Example 1.2	Analysis of covariance	Social influence activity score, a continuous variable	Race of subject (e.g., 1 = white, 2 = black), age, height, etc.	To determine whether one racial group dominates the other in biracial dyads, after controlling for age, height, etc.
Example 1.2	Two-way analysis of variance	Social influence activity difference score between black and white partners in biracial dyads, a continuous variable	Race of experimenter (e.g., 1 = white, 2 = black), race of comparison norm (e.g., 1 = white, 2 = black)	To determine whether the experimenter's race and the comparison norm's race have any effect on the difference score
Example 1.3	Two-way analysis of variance	Systolic blood pressure (SBP), a continuous variable	Discrepancy between attitude toward and involvement in modern life, categorized as "high" or "low"; discrepancy between preparation for and involvement in modern life, categorized as "high" or "low"	To describe the relationship between nominal discrepancy scores and SBP
Example 1.4	Multiple linear regression analysis	BMI, a continuous variable	Drinking frequency, age, sleep quality	To describe the relationship between BMI and drinking frequency, age, and sleep quality
Example 1.4	Analysis of covariance	BMI, a continuous variable	Exercise (0 = no, 1 = yes), tobacco use (0 = no, 1 = yes), drinking frequency, sleep quality	To understand whether BMI levels are different for those who do and do not exercise and those who use and do not use tobacco, controlling for drinking frequency and sleep quality

References

- Green, L. W. 1970. "Manual for Scoring Socioeconomic Status for Research on Health Behaviors." *Public Health Reports* 85: 815–27.
- Gustafson, P. 2004. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Boca Raton, Fla.: Chapman & Hall, CRC Press.
- Thompson, S. J. 1972. "The Doctor–Patient Relationship and Outcomes of Pregnancy." Ph.D. dissertation, Department of Epidemiology, University of North Carolina, Chapel Hill.

3

Basic Statistics: A Review

3.1 Preview

This chapter reviews the fundamental statistical concepts and methods that are needed to understand the more sophisticated multivariable techniques discussed in this text. Through this review, we shall introduce the statistical notation (using conventional symbols whenever possible) employed throughout the text.

The broad area associated with the word *statistics* involves the methods and procedures for collecting, classifying, summarizing, and analyzing data. We shall focus on the latter two activities here. The primary goal of most statistical analysis is to make *statistical inferences*—that is, to draw valid conclusions about a *population* of items or measurements based on information contained in a *sample* from that population.

A *population* is any set of items or measurements of interest, and a *sample* is any subset of items selected from that population. Any characteristic of that population is called a *parameter*, and any characteristic of the sample is termed a *statistic*. A statistic may be considered an estimate of some population parameter, and its accuracy of estimation may be good or bad.

Once sample data have been collected, it is useful, prior to analysis, to examine the data using tables, graphs, and *descriptive statistics*, such as the sample mean and the sample variance. Such descriptive efforts are important for representing the essential features of the data in easily interpretable terms.

Following such examination, statistical inferences are made through two related activities: *estimation* and *hypothesis testing*. The techniques involved here are based on certain assumptions about the probability pattern (or *distribution*) of the (*random*) variables being studied.

Each of the preceding key terms—*descriptive statistics*, *random variables*, *probability distribution*, *estimation*, and *hypothesis testing*—will be reviewed in the sections that follow.

3.2 Descriptive Statistics

A *descriptive statistic* may be defined as any single numerical measure computed from a set of data that is designed to describe a particular aspect or characteristic of the data set. The most common types of descriptive statistics are measures of *central tendency* and of *variability* (or *dispersion*).

The central tendency in a sample of data is the “average value” of the variable being observed. Of the several measures of central tendency, the most commonly used is the sample mean, which we denote by \bar{X} whenever our underlying variable is called X . The formula for the sample mean is given by

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

where n denotes the sample size; X_1, X_2, \dots, X_n denote the n measurements (or observed values) of X ; and \sum denotes summation. The sample mean \bar{X} —in contrast to other measures of central tendency, such as the median or mode—uses in its computation all the observations in the sample. This property means that \bar{X} is necessarily affected by the presence of extreme X -values, so in some cases it may be preferable to use the median instead of the mean.

Measures of central tendency (such as \bar{X}) do not, however, completely summarize all features of the data. Obviously, two sets of data with the same mean can differ widely in appearance (e.g., an \bar{X} of 4 results both from the values 4, 4, and 4 and from the values 0, 4, and 8). Thus, we customarily consider, in addition to \bar{X} , measures of variability, which tell us the extent to which the values of the measurements in the sample differ from one another.

The two measures of variability most often considered are the *sample variance* and the *sample standard deviation*. These are given by the following formulas when considering observations X_1, X_2, \dots, X_n on a single variable X :

$$\text{Sample variance} = S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (3.1)$$

$$\text{Sample standard deviation} = S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \quad (3.2)$$

The formula for S^2 describes variability in terms of an average of squared deviations from the sample mean—although $(n-1)$ is used as the divisor instead of n , due to considerations that make S^2 a good estimator of the variability in the entire population.

A drawback to the use of S^2 is that it is expressed in squared units of the underlying variable X . To obtain a measure of dispersion that is expressed in the same units as X , we simply take the square root of S^2 and call it the sample standard deviation S . Using S in combination with \bar{X} thus gives a fairly succinct picture of both the amount of spread and the center of the data, respectively.

When more than one variable is being considered in the same analysis (as will be the case throughout this text), we will use different letters and/or different subscripts to differentiate among the variables, and we will modify the notations for mean and variance accordingly. For example, if we are using X to stand for age and Y to stand for systolic blood pressure, we will denote the sample mean and the sample standard deviation for each variable as (\bar{X}, S_x) and (\bar{Y}, S_y) , respectively.

All statistical analyses begin with an examination of descriptive statistics computed from the data set at hand. However, often the most direct and revealing way to examine the data is to make a series of plots. We describe three types of simple but useful plots: histograms (especially stem-and-leaf versions), schematic plots, and normal probability plots.

Suppose that we have collected data on the amount of error that occurs in measurements taken with a particular type of instrument. We think the error may be related to the age of the instrument; therefore, readings are taken with 17 instruments of varying ages; the age of each instrument and the error in its measurement are recorded.

In our descriptive analysis of these data, first we examine a frequency histogram of the measurement errors, shown in Figure 3.1(a). We observe that the errors appear to be quite symmetrically distributed around 0 (i.e., the mean and the median error are roughly 0) and that the picture approximates a bell-shaped curve. (See Section 3.3.2 for more information on data that follow this pattern.) No *outliers* (data points that are extreme in value and that may represent data errors) or other anomalies appear to be present.

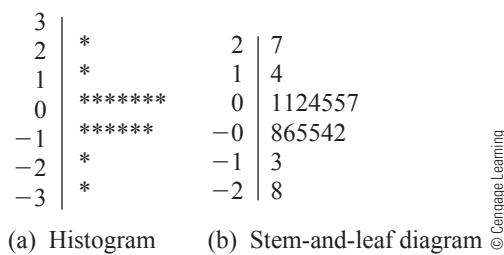


FIGURE 3.1 Frequency histogram and stem-and-leaf diagram of instrument error data ($n = 17$)

The frequency histogram conveys even more information if it is converted into a stem-and-leaf diagram, as in Figure 3.1(b), which shows the actual data values while maintaining the shape of the histogram. In the stem-and-leaf diagram, the top-most value has a *stem* of 2 and a *leaf* of 7, indicating that the original data value is 2.7. Beneath that is a value of 1.4 (stem 1, leaf 4); after that are two values that both share a stem of 0 and have leaves equal to 1 (i.e., both values are 0.1), followed by a value of 0.2, and so on. The last value shown in the plot is −2.8 (stem −2, leaf 8).

The second kind of useful plot is a schematic plot. Figure 3.2 presents a schematic plot of the measurement error data. A schematic plot is based entirely on the order of the values in the sample. *Quartiles* are the most important order-based statistics for the schematic plot. The *first quartile*, or 25th percentile, is the value at or below which 25% of the data values lie; the *second quartile*, or 50th percentile (or median), is the value at or below which 50% of the data values lie; the *third quartile*, or 75th percentile, is the value at or below which 75% of the data values

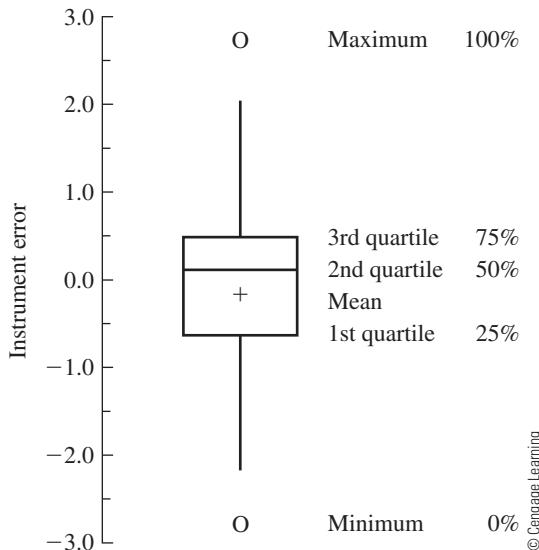


FIGURE 3.2 Schematic plot of instrument error data ($n = 17$)

lie. The *interquartile range* (IQR), calculated as the value of the third quartile minus the value of the first quartile, is a measure of the spread of a distribution, like the variance. One important difference between the IQR and the variance, however, is illustrated by the fact that, whereas doubling the largest value in the sample would, in general, increase variance dramatically, it would not change the IQR. For the error measurements, the first, second, and third quartiles are approximately -0.6 , 0.1 , and 0.5 , respectively, with an interquartile range of 1.1 .

A schematic plot is sometimes called a *box-and-whisker plot*, or simply a *boxplot*, due to its appearance. The box is outlined by three horizontal lines, which mark the values of the first quartile, the second quartile (the median), and the third quartile (see Figure 3.2). The scale is determined by the units and range of the data. The mean is indicated by a + on the backbone of the plot. If the data are symmetric, the mean and median will be close in value (i.e., the + will be marked on or close to the middle horizontal line), and the distances between the first and second quartiles and between the second and third quartiles will be similar in size. The whiskers (vertical lines) extend from the box as far as the data extend up or down, to a limit of 1.5 IQRs (in the vertical direction). An O at the end of a whisker indicates a moderate outlier. Referring to Figure 3.2, we see one positive moderate outlier and one negative moderate outlier.

3.3 Random Variables and Distributions

The term *random variable* is used to denote a variable whose observed values may be considered outcomes of a stochastic or random experiment (e.g., the drawing of a random sample). The values of such a variable in a particular sample, then, cannot be anticipated with certainty before the sample is gathered. Thus, if we select a random sample of persons

from some community and determine the systolic blood pressure (W), cholesterol level (X), race (Y), and sex (Z) of each person, then W , X , Y , and Z are four random variables whose particular realizations (or observed values) for a given person in the sample cannot be known for sure beforehand. In this text, we shall denote random variables by capital italic letters.

The probability pattern that gives the relative frequencies associated with all the possible values of a random variable in a population is generally called the *probability distribution* of the random variable. We represent such a distribution by a table, graph, or mathematical expression that provides the probabilities corresponding to the different values or ranges of values taken by a random variable.

Discrete random variables (such as the number of deaths in a sample of patients or the number of arrivals at a clinic), whose possible values are countable, have (gappy) distributions that are graphed as a series of vertical lines; the heights of these lines represent the probabilities associated with the various possible discrete outcomes (Figure 3.3(a)). *Continuous* random variables (such as blood pressure and weight), whose possible values are uncountable, have (nongappy) distributions that are graphed as smooth curves; an *area* under such a curve represents the probability associated with a *range of values* of the continuous variable (Figure 3.3(b)). We note in passing that the probability of a continuous random variable taking one particular value is 0 because there can be no area above a single point. For discrete distributions, the sum of the probabilities for all possible values of X is equal to 1. For continuous distributions, the total area under the curve representing the distribution is equal to 1.

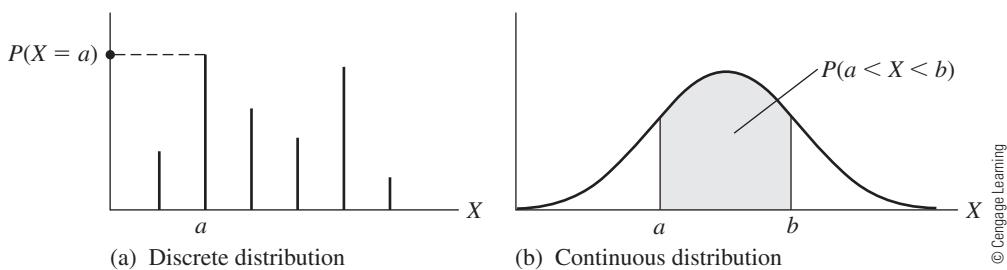


FIGURE 3.3 Discrete and continuous distributions: $P(X = a)$ is read: “The probability that X takes the value a ”

© Cengage Learning

In the next two subsections, we will discuss two particular distributions of enormous practical importance: the binomial (a discrete distribution) and the normal (a continuous distribution).

3.3.1 The Binomial Distribution

A *binomial* random variable describes the number of occurrences of a particular event in a series of n trials, under the following four conditions:

1. The n trials are conducted identically.
2. There are two possible outcomes of each trial: “success” (i.e., the event of interest occurs) or “failure” (i.e., the event of interest does not occur), with probabilities π and $1 - \pi$, respectively.

3. The outcome of any one trial is independent of (i.e., is not affected by) the outcome of any other trial.
4. The probability of success, π , remains the same for all trials.

For example, the distribution of the number of lung cancer deaths in a random sample of $n = 400$ persons would be considered binomial only if the four conditions were all satisfied, as would the distribution of the number of persons in a random sample of $n = 70$ who favor a certain form of legislation.

The two elements of the binomial distribution that one must specify to determine the precise shape of the probability distribution and to compute binomial probabilities are the sample size n and the parameter π . The usual notation for this distribution is, therefore, $B(n, \pi)$. If X has a binomial distribution, it is customary to write

$$X \sim B(n, \pi)$$

where \sim stands for “is distributed as.” The probability formula for this discrete random variable X is given by the expression

$$P(X = j) = {}_n C_j \pi^j (1 - \pi)^{n-j} \quad j = 0, 1, \dots, n$$

where ${}_n C_j = n!/[j!(n-j)!]$ denotes the number of combinations of n distinct objects selected j at a time.

3.3.2 The Normal Distribution

The *normal distribution*, denoted as $N(\mu, \sigma)$, where μ and σ are the two parameters, is described by the well-known bell-shaped curve (Figure 3.4). The parameters μ (the mean) and σ (the standard deviation) characterize the center and the spread, respectively, of the distribution. We generally attach a subscript to the parameters μ and σ to distinguish among variables; that is, we often write

$$X \sim N(\mu_X, \sigma_X)$$

to denote a normally distributed X .

An important property of any normal curve is its *symmetry*, which distinguishes it from some other continuous distributions that we will discuss later. This symmetry property is quite helpful when using tables to determine probabilities or percentiles of the normal distribution.

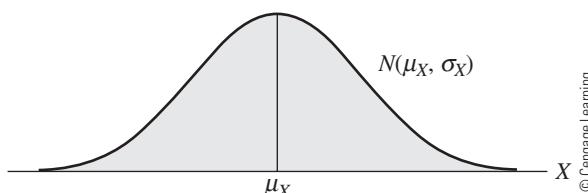


FIGURE 3.4 A normal distribution

Probability statements about a normally distributed random variable X that are of the form $P(a \leq X \leq b)$ require for computation the use of a single table (Table A.1 in Appendix A). This table gives the probabilities (or areas) associated with the *standard normal distribution*, which is a normal distribution with $\mu = 0$ and $\sigma = 1$. It is customary to denote a standard normal random variable by the letter Z , so we write

$$Z \sim N(0, 1)$$

To compute the probability $P(a \leq X \leq b)$ for an X that is $N(\mu_X, \sigma_X)$, we must transform (i.e., *standardize*) X to Z by applying the conversion formula

$$Z = \frac{X - \mu_X}{\sigma_X} \quad (3.3)$$

to each of the elements in the probability statement about X , as follows:

$$P(a \leq X \leq b) = P\left(\frac{a - \mu_X}{\sigma_X} \leq Z \leq \frac{b - \mu_X}{\sigma_X}\right)$$

We then look up the equivalent probability statement about Z in the $N(0, 1)$ tables.

For random samples, this rule also applies to the sample mean \bar{X} whenever the underlying variable X is normally distributed or whenever the sample size is moderately large (by the Central Limit Theorem). But because the standard deviation of \bar{X} is σ_X/\sqrt{n} , the conversion formula has the form

$$Z = \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}$$

A *percentile* is a value of a random variable X below which the area under the probability distribution has a certain specified value. We denote the $(100p)$ th percentile of X by X_p and picture it as in Figure 3.5, where p is the amount of area under the curve to the left of X_p . In determining X_p for a given p , we must again use the conversion formula (3.3). Since the procedure requires that we first determine Z_p and then convert back to X_p , however, we generally rewrite the conversion formula as

$$X_p = \mu_X + \sigma_X Z_p \quad (3.4)$$

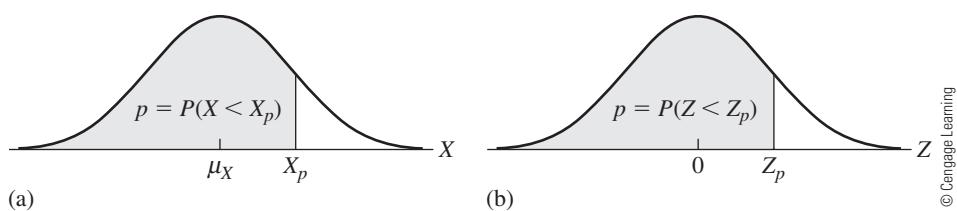


FIGURE 3.5 The $(100p)$ th percentiles of X and Z

For example, if $\mu_X = 140$ and $\sigma_X = 40$, and we want to find $X_{0.95}$, the $N(0, 1)$ table first gives us $Z_{0.95} = 1.645$, which we convert back to $X_{0.95}$ as follows:

$$X_{0.95} = 140 + (40)Z_{0.95} = 140 + 40(1.645) = 205.8$$

Formulas (3.3) and/or (3.4) can also be used to approximate probabilities and percentiles for the binomial distribution $B(n, \pi)$ whenever n is moderately large (e.g., $n > 20$). Two conditions are usually required for this approximation to be accurate: $n\pi > 5$ and $n(1 - \pi) > 5$. Under such conditions, the mean and the standard deviation of the approximating normal distribution are

$$\mu = n\pi \quad \text{and} \quad \sigma = \sqrt{n\pi(1 - \pi)}$$

A *normal probability plot* assesses how well the sample data adhere to a normal distribution, in order to help infer whether the data are sampled from a normally distributed population. The ordered data values are plotted against corresponding percentiles from an estimated normal distribution. Plots that are linear in appearance are consistent with the assumption of normality, since the cumulative relative frequencies for a normal distribution plot as a straight line. For example, in Figure 3.6, plot (a) supports the assumption that the data constitute a random sample from a normal distribution; the other plots suggest deviations from this assumption.

The *skewness* and *kurtosis* statistics can also be helpful in assessing normality. *Skewness* indicates the degree of asymmetry of a distribution. Just as variance is the average squared deviation of observations about the mean, skewness is the average of cubed deviations about the mean. To simplify comparisons between samples and to help account for estimation in small samples, skewness is usually computed as

$$\text{sk}(X) = \left(\frac{n}{n-2} \right) \left(\frac{1}{n-1} \right) \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{S_X} \right)^3$$

For large n , $\text{sk}(X)$ should be approximately equal to 0 for a random sample size of n from any symmetric probability distribution (such as a normal distribution). Positive values of $\text{sk}(X)$ indicate that relatively more values are above the mean than below it; the sample values are thus said to be “positively skewed.” A negative value for $\text{sk}(X)$ indicates that relatively more values are below the mean than above it.

Kurtosis indicates the heaviness of the tails relative to the middle of a distribution. Because kurtosis is the average of the fourth power of the deviations about the mean, it is always nonnegative. Standardized kurtosis may be computed as

$$\text{Kur}(X) = \left[\frac{n(n+1)}{(n-2)(n-3)} \right] \left(\frac{1}{n-1} \right) \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{S_X} \right)^4$$

The term in brackets, which approaches 1 as n increases, helps to account for estimation based on a small sample. Since standardized kurtosis for a standard normal distribution is 3, this value is often subtracted from $\text{Kur}(X)$. The resulting statistic can be as small as -3 for flat distributions with short tails; it is approximately zero for moderate to large random samples from a normal distribution, and it is positive for heavy-tailed distributions. Thus, the positive

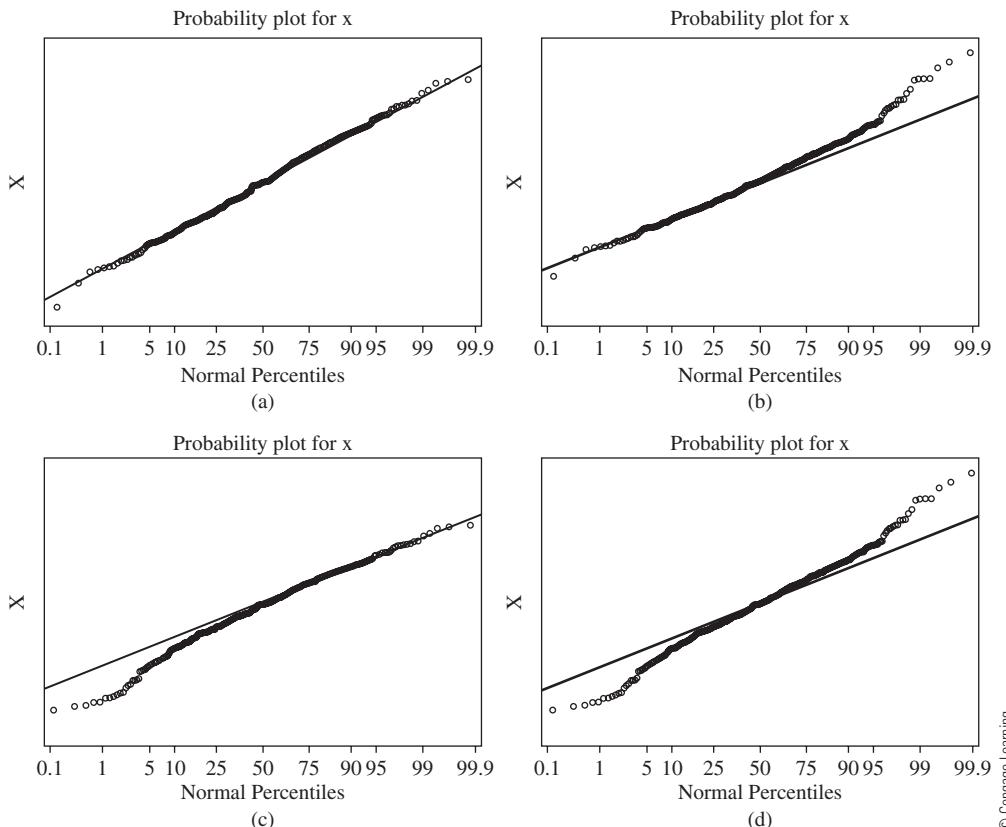


FIGURE 3.6 Normal probability plots

© Cengage Learning

kurtosis value in our example (the reader is encouraged to do the required calculations) suggests a distribution with tails heavier than for a normal distribution. Skewness and kurtosis statistics are highly variable in small samples and hence are often difficult to interpret.

3.4 Sampling Distributions of t , χ^2 , and F

The Student's t , chi-square (χ^2), and Fisher's F distributions are particularly important in statistical inference making.

The (*Student's*) t distribution (Figure 3.7(a)), which like the standard normal distribution is symmetric about 0, was originally developed to describe the behavior of the random variable

$$T = \frac{\bar{X} - \mu_X}{\frac{S_X}{\sqrt{n}}} \quad (3.5)$$

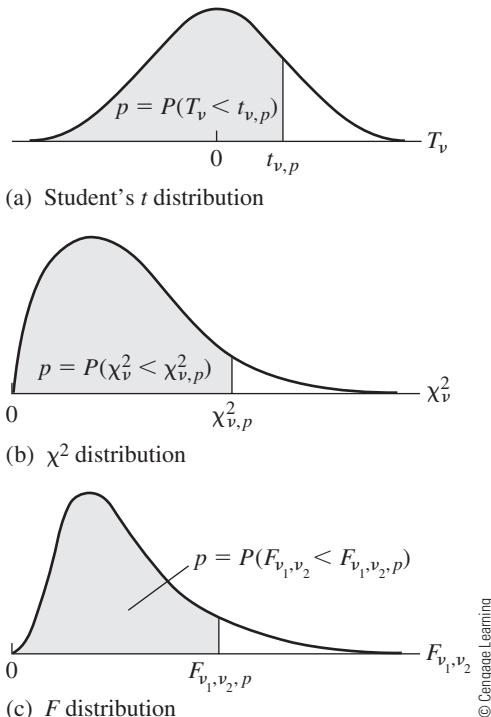


FIGURE 3.7 The t , χ^2 , and F distributions

© Cengage Learning

which represents an alternative to

$$Z = \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}$$

whenever the population variance σ_X^2 is unknown and is estimated by S_X^2 . The denominator of (3.5), S_X/\sqrt{n} , is the *estimated standard error of \bar{X}* . When the underlying distribution of X is normal and when \bar{X} and S_X^2 are calculated using a random sample from that normal distribution, then (3.5) has the *t distribution with $n - 1$ degrees of freedom*, where $n - 1$ is the quantity that must be specified in order to look up tabulated percentiles of this distribution. We denote all this by writing

$$T = \frac{\bar{X} - \mu_X}{\frac{S_X}{\sqrt{n}}} \sim t_{n-1}$$

It has generally been shown by statisticians that the t distribution is sometimes appropriate for describing the behavior of a random variable of the general form

$$T = \frac{\hat{\theta} - \mu_{\hat{\theta}}}{S_{\hat{\theta}}} \quad (3.6)$$

where $\hat{\theta}$ is any random variable that is normally distributed with mean $\mu_{\hat{\theta}}$ and standard deviation $\sigma_{\hat{\theta}}$, where $S_{\hat{\theta}}$ is the estimated standard error of $\hat{\theta}$, and where $\hat{\theta}$ and $S_{\hat{\theta}}$ are statistically independent. For example, when random samples are taken from two normally distributed populations with the same standard deviation (e.g., from $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$), and we consider $\hat{\theta} = \bar{X}_1 - \bar{X}_2$ in (3.6), we can write

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}$$

where

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \quad (3.7)$$

estimates the common variance σ^2 in the two populations. The quantity S_p^2 is called a *pooled sample variance*, since it is calculated by pooling the data from both samples in order to estimate the common variance σ^2 .

The *chi-square* (or χ^2) *distribution* (Figure 3.7(b)) is a nonsymmetric distribution and describes, for example, the behavior of the nonnegative random variable

$$\frac{(n - 1)S^2}{\sigma^2} \quad (3.8)$$

where S^2 is the sample variance based on a random sample of size n from a normal distribution. The variable given by (3.8) has the chi-square distribution with $n - 1$ degrees of freedom:

$$\frac{(n - 1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Because of the nonsymmetry of the chi-square distribution, both upper and lower percentage points of the distribution need to be tabulated, and such tabulations are solely a function of the degrees of freedom associated with the particular χ^2 distribution of interest. The chi-square distribution has widespread application in analyses of categorical data.

The *F distribution* (Figure 3.7(c)), which like the chi-square distribution is skewed to the right, is often appropriate for modeling the probability distribution of the ratio of independent estimators of two population variances. For example, given random samples of sizes n_1 and n_2 from $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$, respectively, so that estimates S_1^2 and S_2^2 of σ_1^2 and σ_2^2 can be calculated, it can be shown that

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \quad (3.9)$$

has the F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, which are called the *numerator* and *denominator* degrees of freedom, respectively. We write this as

$$\frac{S_1^2 \sigma_2^2}{S_2^2 \sigma_1^2} \sim F_{n_1 - 1, n_2 - 1}$$

The F distribution can also be related to the t distribution when the numerator degrees of freedom equal 1; that is, the square of a variable distributed as Student's t with v degrees of freedom has the F distribution with 1 and v degrees of freedom. In other words,

$$T^2 \sim F_{1, v} \text{ if and only if } T \sim t_v$$

Percentiles of the t , χ^2 , and F distributions may be obtained from Tables A.2, A.3, and A.4 in Appendix A. The shapes of the curves that describe these probability distributions, together with the notation we will use to denote their percentile points, are given in Figure 3.7.

3.5 Statistical Inference: Estimation

Two general categories of statistical inference—estimation and hypothesis testing—can be distinguished by their differing purposes: estimation is concerned with quantifying the specific value of an unknown population parameter; hypothesis testing is concerned with making a decision about a hypothesized value of an unknown population parameter.

In estimation, which we focus on in this section, we want to estimate an unknown parameter θ by using a random variable $\hat{\theta}$ (“theta hat,” called a *point estimator* of θ). This point estimator takes the form of a formula or rule. For example,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{or} \quad S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

tells us how to calculate a specific point estimate, given a particular set of data.

To estimate a parameter of interest (e.g., a population mean μ , a binomial proportion π , a difference between two population means $\mu_1 - \mu_2$, or a ratio of two population standard deviations σ_1/σ_2), the usual procedure is to select a random sample from the population or populations of interest, calculate the point estimate of the parameter, and then associate with this estimate a measure of its variability, which usually takes the form of a confidence interval for the parameter of interest.

As its name implies, a *confidence interval* (often abbreviated CI) consists of two random boundary points between which we have a certain specified *level of confidence* that the population parameter lies. More specifically, a 95% confidence interval for a parameter θ consists of lower and upper limits determined so that, in many repeated sets of samples of the same size, about 95% of all such intervals would be expected to contain the parameter θ . Care must be taken when interpreting such a confidence interval not to consider θ a random variable that either falls or does not fall in the calculated interval; rather, θ is a fixed (unknown) constant, and the random quantities are the lower and upper limits of the confidence interval, which vary from sample to sample.

We illustrate the procedure for computing a confidence interval with two examples using random samples from normally distributed populations—one involving estimation of a single population mean μ and one involving estimation of the difference between two population means $\mu_1 - \mu_2$. In each case, the appropriate confidence interval has the following general form:

$$\left(\begin{array}{l} \text{Point estimate of} \\ \text{the parameter} \end{array} \right) \pm \left[\left(\begin{array}{l} \text{Percentile of} \\ \text{the } t \text{ distribution} \end{array} \right) \left(\begin{array}{l} \text{Estimated standard} \\ \text{error of the estimate} \end{array} \right) \right] \quad (3.10)$$

This general form also applies to confidence intervals for other parameters considered in the remainder of the text (e.g., those considered in multiple regression analysis).

■ Example 3.1 Suppose that we have determined the Quantitative Graduate Record Examination (QGRE) scores for a random sample of nine student applicants to a certain graduate department in a university and that we have found $\bar{X} = 520$ and $S = 50$. If we want to estimate with 95% confidence the population mean QGRE score (μ) for all such applicants to the department, and we are willing to assume that the population of such scores from which our random sample was selected is approximately normally distributed, the confidence interval for μ is given by the general formula

$$\bar{X} \pm t_{n-1, 1-\alpha/2} \left(\frac{S}{\sqrt{n}} \right) \quad (3.11)$$

which gives the $100(1 - \alpha)\%$ (small-sample) confidence interval for μ when σ is unknown. In our problem, $\alpha = 1 - .95 = .05$ and $n = 9$; therefore, by substituting the given information into (3.11), we obtain

$$520 \pm t_{8, 0.975} \left(\frac{50}{\sqrt{9}} \right)$$

Since $t_{8, 0.975} = 2.3060$, this formula becomes

$$520 \pm 2.3060 \left(\frac{50}{\sqrt{9}} \right)$$

or

$$520 \pm 38.43$$

Our 95% confidence interval for μ is thus given by

$$(481.57, 558.43)$$

If we wanted to use this confidence interval to help determine whether 600 is a likely value for μ (i.e., if we were interested in making a decision about a specific value for μ), we would conclude that 600 is not a likely value, since it is not contained in the 95% confidence interval for μ just developed. This helps clarify the connection between estimation and hypothesis testing. ■

■ **Example 3.2** Suppose that we want to compare the change in health status of two groups of mental patients who are undergoing different forms of treatment for the same disorder. Suppose that we have a measure of change in health status based on a questionnaire given to each patient at two different times and that we are willing to assume this measure of change in health status is approximately normally distributed and has the same variance in the populations of patients from which we selected our independent random samples. The data obtained are summarized as follows:

$$\text{Group 1: } n_1 = 15, \bar{X}_1 = 15.1, S_1 = 2.5$$

$$\text{Group 2: } n_2 = 15, \bar{X}_2 = 12.3, S_2 = 3.0$$

where the underlying variable X denotes the change in health status between time 1 and time 2.

A 99% confidence interval for the true mean difference ($\mu_1 - \mu_2$) in health status change between these two groups is given by the following formula, which assumes equal population variances (i.e., $\sigma_1^2 = \sigma_2^2$):

$$(\bar{X}_1 - \bar{X}_2) \pm t_{n_1+n_2-2, 1-\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (3.12)$$

where S_p is the pooled standard deviation derived from S_p^2 , the pooled sample variance given by (3.7). Here we have

$$S_p^2 = \frac{(15-1)(2.5)^2 + (15-1)(3.0)^2}{15+15-2} = 7.625$$

so

$$S_p = \sqrt{7.625} = 2.76$$

Since $\alpha = .01$, our percentile in (3.12) is given by $t_{28, 0.995} = 2.7633$. So the 99% confidence interval for $\mu_1 - \mu_2$ is given by

$$(15.1 - 12.3) \pm 2.7633(2.76) \sqrt{\frac{1}{15} + \frac{1}{15}}$$

which reduces to

$$2.80 \pm 2.78$$

yielding the following 99% confidence interval for $\mu_1 - \mu_2$:

$$(0.02, 5.58)$$

Since the value 0 is not contained in this interval, we conclude that there is statistical evidence of a difference in health status change between the two groups. ■

3.6 Statistical Inference: Hypothesis Testing

Although closely related to confidence interval estimation, hypothesis testing has a slightly different orientation. When developing a confidence interval, we use our sample data to estimate what we think is a *likely* set of values for the parameter of interest. When performing a statistical test of a null hypothesis concerning a certain parameter, we use our sample data to *test* whether our estimated value for the parameter is *different enough* from the hypothesized value to support the conclusion that the null hypothesis is *unlikely* to be true.

The general procedure used in testing a statistical null hypothesis remains basically the same, regardless of the parameter being considered. This procedure (which we will illustrate by example) consists of the following seven steps:

1. Check the assumptions regarding the properties of the underlying variable(s) being measured that are needed to justify use of the testing procedure under consideration.
2. State the null hypothesis H_0 and the alternative hypothesis H_A .
3. Specify the significance level α .
4. Specify the test statistic to be used and its distribution under H_0 .
5. Form the decision rule for rejecting or not rejecting H_0 (i.e., specify the rejection and nonrejection regions for the test, based on both H_A and α).
6. Compute the value of the test statistic from the observed data.
7. Draw conclusions regarding rejection or nonrejection of H_0 .

■ **Example 3.3** Let us again consider the random sample of nine student applicants with mean QGRE score $\bar{X} = 520$ and standard deviation $S = 50$. The department chairperson suspects that, because of the declining reputation of the department, this year's applicants are not quite as good quantitatively as those from the previous five years for whom the average QGRE score was 600. If we assume that the population of QGRE scores from which our random sample has been selected is normally distributed, we can test the null hypothesis that the population mean score associated with this year's applicants is 600 versus the alternative hypothesis that it is less than 600. The *null hypothesis*, in mathematical terms, is $H_0: \mu = 600$, which asserts that the population mean μ for this year's applicants does not differ from what it has generally been in the past. The *alternative hypothesis* is stated as $H_A: \mu < 600$, which asserts that the QGRE scores, on average, have gotten worse.

We have thus far considered the first two steps of our testing procedure:

1. Assumptions: The variable QGRE score has a normal distribution, from which a random sample has been selected.
2. Hypotheses: $H_0: \mu = 600$; $H_A: \mu < 600$.

Our next step is to decide what error or probability we are willing to tolerate for incorrectly rejecting H_0 (i.e., making a Type I error, as discussed later in this chapter). We call this probability of making a Type I error the *significance level* α .¹

We usually assign a value such as .1, .05, .025, or .01 to α . Suppose, for now, that we choose $\alpha = .025$. Then Step 3 is

3. Use $\alpha = .025$.

Step 4 requires us to specify the test statistic that will be used to test H_0 . In this case, with $H_0: \mu = 600$, we have

$$4. T = \frac{\bar{X} - 600}{S/\sqrt{9}} \sim t_8 \text{ under } H_0: \mu = 600.$$

Step 5 requires us to specify the decision rule that we will use to reject or not reject H_0 . In determining this rule, we divide the possible values of T into two sets: the *rejection region* (or *critical region*), which consists of values of T for which we reject H_0 ; and the *nonrejection region*, which consists of those T -values for which we do not reject H_0 . If our computed value of T falls in the rejection region, we conclude that the observed results deviate far enough from H_0 to cast considerable doubt on the validity of the null hypothesis.

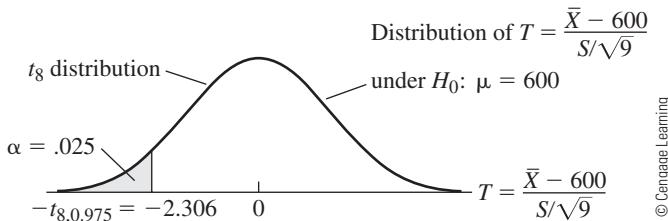
In our example, we determine the critical region by choosing from t tables a point called the *critical point*, which defines the boundary between the nonrejection and rejection regions. The alternative hypothesis (H_A) informs the determination of the rejection region. Because our H_A states that the true mean is *less than* 600, an observed sample mean sufficiently less than 600 would be needed to support this alternative hypothesis. Accordingly, the test statistic T above would need to be negative, and thus all values of T in the rejection region would be negative. The value we choose is

$$-t_{8, 0.975} = -2.306$$

in which case the probability that the test statistic takes a value of less than -2.306 under H_0 is exactly $\alpha = .025$, the significance level (Figure 3.8). We thus have the following decision rule:

$$5. \text{ Reject } H_0 \text{ if } T = \frac{\bar{X} - 600}{S/\sqrt{9}} < -2.306; \text{ do not reject } H_0 \text{ otherwise.}$$

¹ Two types of errors can be made when performing a statistical test. A Type II error occurs if we fail to reject H_0 when H_0 is actually false. We denote the probability of a Type II error as β and call $(1 - \beta)$ the *power* of the test. For a fixed sample size, α and β for a given test are inversely related; that is, lowering one has the effect of increasing the other. In general, the power of any statistical test can be raised by increasing the sample size. These issues are described further in Section 3.7.

**FIGURE 3.8** The critical point for Example 3.3

If H_A stated that $\mu \neq 600$ (a two-sided hypothesis), then either an extremely negative or an extremely positive value of T would support this alternative hypothesis. The rejection region would, therefore, include negative and positive values of T . Since the region would be two-tailed, α would be split between the two tails ($\alpha/2 = 0.0125$ here), yielding a decision rule of rejecting H_0 if $T < -t_{8, 0.9875}$ or if $T > t_{8, 0.9875}$.

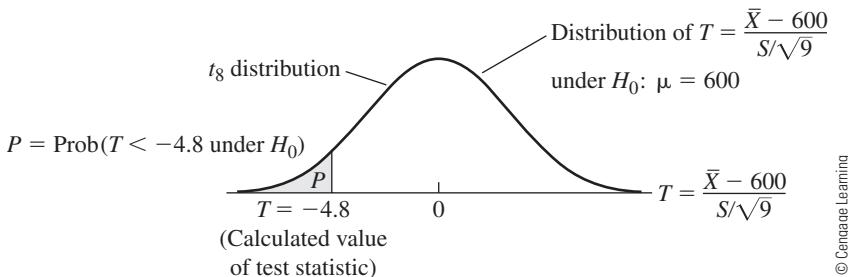
Now we simply apply the decision rule to our data by computing the observed value of T . In our example, since $\bar{X} = 520$ and $S = 50$, our computed T is

$$6. \quad T = \frac{\bar{X} - 600}{S/\sqrt{9}} = \frac{520 - 600}{50/3} = -4.8.$$

The last step is to make the decision about H_0 based on the rule given in Step 5:

7. Since $T = -4.8$, which lies below -2.306 , we reject H_0 at significance level $.025$ and conclude that there is evidence that students currently applying to the department have QGRE scores that are, on average, *lower* than 600.

In addition to performing the procedure just described, we often want to compute a *P-value*, which quantifies *exactly how unusual the observed results would be if H_0 were true*. An equivalent way of describing the *P-value* is as follows: *The P-value gives the probability of obtaining a value of the test statistic that is at least as unfavorable to H_0 as the observed value, assuming that H_0 is true* (Figure 3.9).

**FIGURE 3.9** The *P*-value

To get an idea of the approximate size of the P -value in this example, our approach is to determine from the table of the distribution of T under H_0 the two percentiles that bracket the observed value of T . In this case, the two percentiles are

$$-t_{8,0.995} = -3.355 \quad \text{and} \quad -t_{8,0.9995} = -5.041$$

Since the observed value of T lies between these two values, we conclude that the area P we seek lies between the two areas corresponding to these two percentiles:

$$.0005 < P < .005$$

In interpreting this inequality, we observe that the P -value is *quite small*, indicating that we have observed a highly unusual result if H_0 is true. In fact, this P -value is so small as to lead us to reject H_0 . Furthermore, the size of this P -value means that we would reject H_0 even for an α as small as .005.

For the general computation of a P -value, the appropriate P -value for a two-tailed test is twice that for the corresponding one-tailed test. If an investigator wants to draw conclusions about a test on the basis of the P -value (e.g., in lieu of specifying α a priori), the following guidelines are recommended:

1. If P is small (less than .01), reject H_0 .
2. If P is large (greater than .1), do not reject H_0 .
3. If $.01 < P < .1$, the significance is borderline, since we reject H_0 for $\alpha = .1$ but not for $\alpha = .01$.

Notice that, if we actually do specify α a priori, we reject H_0 when $P < \alpha$. ■

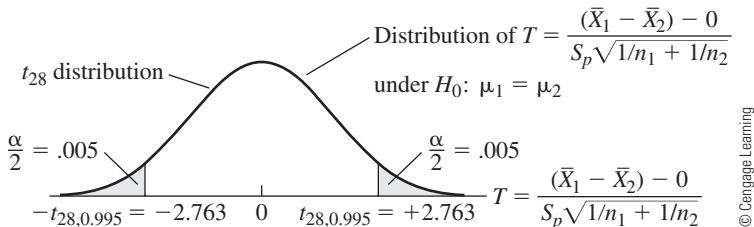
■ **Example 3.4** We now look at one more worked example about hypothesis testing—this time involving a comparison of two means, μ_1 and μ_2 . Consider the following data on health status change, which were discussed earlier:

$$\begin{aligned} \text{Group 1: } n_1 &= 15, \bar{X}_1 = 15.1, S_1 = 2.5 & (S_p = 2.76) \\ \text{Group 2: } n_2 &= 15, \bar{X}_2 = 12.3, S_2 = 3.0 \end{aligned}$$

Suppose that we want to test at significance level .01 whether the true average change in health status differs between the two groups. The steps required to perform this test are as follows:

1. Assumptions: We have independent random samples from two normally distributed populations. The population variances are assumed to be equal.
2. Hypotheses: $H_0: \mu_1 = \mu_2$; $H_A: \mu_1 \neq \mu_2$.
3. Use $\alpha = .01$.

$$4. T = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{28} \text{ under } H_0.$$



© Cengage Learning

FIGURE 3.10 Critical region for the health status change example

5. Reject H_0 if $|T| \geq t_{28,0.995} = 2.763$; do not reject H_0 otherwise (Figure 3.10).

$$6. T = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{15.1 - 12.3}{2.76 \sqrt{\frac{1}{15} + \frac{1}{15}}} = 2.78.$$

7. Since $T = 2.78$ exceeds $t_{28,0.995} = 2.763$, we reject H_0 at $\alpha = .01$ and conclude that there is evidence the true average change in health status differs between the two groups.

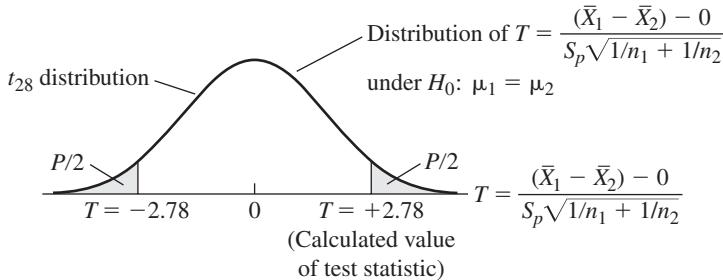
The P -value for this test is given by the shaded area in Figure 3.11. For the t distribution with 28 degrees of freedom, we find that $t_{28,0.995} = 2.763$ and $t_{28,0.9995} = 3.674$. Thus, $P/2$ is given by the inequality

$$1 - .9995 < \frac{P}{2} < 1 - .995$$

so

$$.001 < P < .01$$

■



© Cengage Learning

FIGURE 3.11 P -value for the health status change example

3.7 Error Rates, Power, and Sample Size

Table 3.1 summarizes the decisions that result from hypothesis testing. If the true state of nature is that the null hypothesis is true and if the decision is made that the null hypothesis is true, then a correct decision has been made. Similarly, if the true state of nature is that the alternative hypothesis is true and if the decision is made that the alternative is true, then a correct decision has been made. On the other hand, if the true state of nature is that the null hypothesis is true but the decision is made to choose the alternative, then a false positive error (commonly referred to as a *Type I error*) has been made. And if the true state of nature supports the alternative hypothesis but the decision is made that the null hypothesis is true, then a false negative error (commonly referred to as a *Type II error*) has been made.

TABLE 3.1 Outcomes of hypothesis testing

Hypothesis Chosen	True State of Nature	
	H_0	H_A
H_0	Correct decision	False negative decision (Type II error)
H_A	False positive decision (Type I error)	Correct decision

© Cengage Learning

Table 3.2 summarizes the probabilities associated with the outcomes of hypothesis testing just described. If the true state of nature corresponds to the null hypothesis but the alternative hypothesis is chosen, then a Type I error has been made, with probability denoted by the symbol α . Hence, the probability of making a correct choice of H_0 given that H_0 is true must be $1 - \alpha$. In turn, if the actual state of nature is that the alternative hypothesis is true but the null hypothesis is chosen, then a Type II error has occurred, with probability denoted by β . In turn, $1 - \beta$ is the probability of choosing the alternative hypothesis given that it is true, and this probability is often called the *power of the test*.

TABLE 3.2 Probabilities of outcomes of hypothesis testing

Hypothesis Chosen	True State of Nature	
	H_0	H_A
H_0	$1 - \alpha$	β
H_A	α	$1 - \beta$

© Cengage Learning

When we design a research study, we would like to use statistical tests for which both α and β are small (i.e., for which there is a small chance of making either a Type I or a Type II error). For a given α , we can sometimes determine the sample size required in the study to ensure that β is no larger than some desired value for a particular alternative hypothesis of interest. Such a design consideration generally involves the use of a *sample size formula* pertinent to the research question(s). This formula usually requires the researcher to make educated guesses about the values of some of the unknown parameters to be estimated in the study (see Cohen 1977; Muller and Peterson 1984; Kupper and Hafner 1989).

For example, the classical sample size formula used for a one-sided test of $H_0: \mu_1 = \mu_2$ versus $H_A: \mu_2 > \mu_1$, when a random sample of size n is selected from each of two normal populations with common variance σ^2 , is as follows:

$$n \geq \frac{2(Z_{1-\alpha} + Z_{1-\beta})^2\sigma^2}{\Delta^2}$$

For chosen values of α , β , and σ^2 , this formula provides the minimum sample size n required to detect a specified difference $\Delta = \mu_2 - \mu_1$ between μ_1 and μ_2 (i.e., to reject $H_0: \mu_2 - \mu_1 = 0$ in favor of $H_A: \mu_2 - \mu_1 = \Delta > 0$ with power $1 - \beta$). Thus, in addition to picking α and β , the researcher must specify the size of the population variance σ^2 and specify the difference Δ to be detected. An educated guess about the value of the unknown parameter σ^2 can sometimes be made by using information obtained from related research studies. To specify Δ intelligently, the researcher has to decide on the smallest population mean difference ($\mu_2 - \mu_1$) that is practically (as opposed to statistically) meaningful for the study.

For a fixed sample size, α and β are inversely related in the following sense, illustrated in Figure 3.12. If one tries to guard against making a Type I error by choosing a small rejection region, the nonrejection region (and hence β) will be large. Conversely, protecting against a Type II error necessitates using a large rejection region, leading to a large value for α . Increasing the sample size generally decreases the standard deviation of the test statistic (standard error) and accordingly decreases β ; of course, α remains unaffected. A detailed discussion about power and sample size determination for statistical methods taught in this text is provided in Chapter 27.

It is common practice to conduct several statistical tests using the same data set. If such a data-set-specific series of tests is performed and each test is based on a size α rejection region,

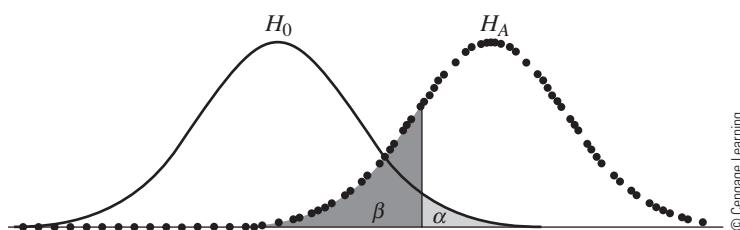


FIGURE 3.12 Distributions of a test statistic under the null (H_0) and alternative (H_A) hypotheses, displaying the relationship between α and β

the probability of making at least one Type I error will be much larger than α . This multiple-testing problem is pervasive and bothersome. One simple—but not optimal—method for addressing this problem is to employ the so-called *Bonferroni correction*. For example, if k tests are to be conducted and if the overall Type I error rate (i.e., the probability of making at least one Type I error in k tests) is to be no more than α , then a rule of thumb is to conduct each individual test at a Type I error rate of α/k .

This simple adjustment ensures that the overall Type I error rate will (at least approximately) be no larger than α . In many situations, however, this correction leads to such a small rejection region for each individual test that the power of each test may be too low to detect important deviations from the null hypotheses being tested. Resolving this antagonism between Type I and Type II error rates requires a conscientious study design and carefully considered error rates for planned analyses.

Problems

1.
 - a. Give two examples of discrete random variables.
 - b. Give two examples of continuous random variables.
2. Name the four levels of measurement, and give an example of a variable at each level.
3. Assume that Z is a normal random variable with mean 0 and variance 1.
 - a. $P(Z \geq -1) = ?$
 - b. $P(Z \leq ?) = .20$
4.
 - a. $P(\chi^2_7 \geq ?) = .01$
 - b. $P(\chi^2_{12} \leq 14) = ?$
5.
 - a. $P(T_{13} \leq ?) = .10$
 - b. $P(|T_{28}| \geq 2.05) = ?$
6.
 - a. $P(F_{6, 24} \geq ?) = .05$
 - b. $P(F_{5, 40} \geq 2.9) = ?$
7. What are the (a) mean, (b) median, and (c) mode of the standard normal distribution?
8. An $F_{1, v}$ random variable can be thought of as the square of what kind of random variable?
9. Find the (a) mean, (b) median, and (c) variance for the following set of scores:
$$\{0, 2, 5, 6, 3, 3, 3, 1, 4, 3\}$$
 - d. Find the set of Z scores for the data.
10. Which of the following statements about descriptive statistics is correct?
 - a. All of the data are used to compute the median.
 - b. The mean should be preferred to the median as a measure of central tendency if the data are noticeably skewed.
 - c. The variance has the same units of measurement as the original observations.
 - d. The variance can never be 0.
 - e. The variance is like an average of squared deviations from the mean.

11. Suppose that the weight W of male patients registered at a diet clinic has the normal distribution with mean 190 and variance 100.
- For a random sample of patients of size $n = 25$, the expression $P(\bar{W} < 180)$, in which \bar{W} denotes the sample mean weight, is equivalent to saying $P(Z > ?)$.
[Note: Z is a standard normal random variable.]
 - Find an interval (a, b) such that $P(a < \bar{W} < b) = .80$ for the same random sample in part (a).
12. The limits of a 95% confidence interval for the mean μ of a normal population with unknown variance are found by adding to and subtracting from the sample mean a certain multiple of the estimated standard error of the sample mean. If the sample size on which this confidence interval is based is 28, the *multiple* referred to in the previous sentence is the number _____.
13. A random sample of 32 persons attending a certain diet clinic was found to have lost (over a three-week period) an average of 30 pounds, with a sample standard deviation of 11. For these data, a 99% confidence interval for the true mean weight loss by all patients attending the clinic would have the limits $(?, ?)$.
14. From two normal populations assumed to have the same variance, independent random samples of sizes 15 and 19 were drawn. The first sample (with $n_1 = 15$) yielded mean and standard deviation 111.6 and 9.5, respectively, while the second sample ($n_2 = 19$) gave mean and standard deviation 100.9 and 11.5, respectively. The estimated standard error of the difference in sample means is _____.
15. For the data of Problem 14, suppose that a test of $H_0: \mu_1 = \mu_2$ versus $H_A: \mu_1 > \mu_2$ yielded a computed value of the appropriate test statistic equal to 2.55.
 - What conclusions should be drawn for $\alpha = .05$?
 - What conclusions should be drawn for $\alpha = .01$?
16. Test the null hypothesis that the true population average body weight is the same for two independent diagnosis groups from one hospital versus the alternative hypothesis that these two population averages are different, using the following data:

Diagnosis group 1 data: {132, 145, 124, 122, 165, 144, 151}

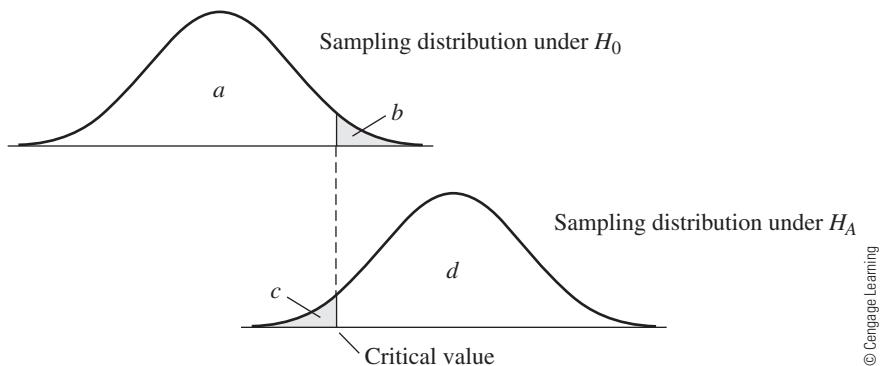
Diagnosis group 2 data: {141, 139, 172, 131, 150, 125}

You may assume that the populations from which the data come are each normally distributed, with equal population variances. What conclusion should be drawn, with $\alpha = .05$?

17. Independent random samples are drawn from two normal populations, which are assumed to have the same variance. One sample (of size 5) yields mean 86.4 and standard deviation 8.0, and the other sample (of size 7) has mean 78.6 and standard deviation 10. The limits of a 99% confidence interval for the difference in population means are found by adding to and subtracting from the difference in sample means a certain multiple of the estimated standard error of this difference. This *multiple* is the number _____.

- 18.** If a 99% confidence interval for $\mu_1 - \mu_2$ is 4.8 to 9.2, which of the following conclusions can be drawn *based on this interval?*
- Do not reject $H_0: \mu_1 = \mu_2$ at $\alpha = .05$ if the alternative is $H_A: \mu_1 \neq \mu_2$.
 - Reject $H_0: \mu_1 = \mu_2$ at $\alpha = .01$ if the alternative is $H_A: \mu_1 \neq \mu_2$.
 - Reject $H_0: \mu_1 = \mu_2$ at $\alpha = .01$ if the alternative is $H_A: \mu_1 < \mu_2$.
 - Do not reject $H_0: \mu_1 = \mu_2$ at $\alpha = .01$ if the alternative is $H_A: \mu_1 \neq \mu_2$.
 - Do not reject $H_0: \mu_1 = \mu_2 + 3$ at $\alpha = .01$ if the alternative is $H_A: \mu_1 \neq \mu_2 + 3$.
- 19.** Assume that we gather data, compute a T , and reject the null hypothesis. If, in fact, the null hypothesis is true, we have made (a) _____. If the null hypothesis is false, we have made (b) _____. Assume instead that our data lead us to not reject the null hypothesis. If, in fact, the null hypothesis is true, we have made (c) _____. If the null hypothesis is false, we have made (d) _____.
- 20.** Suppose that the critical region for a certain test of hypothesis is of the form $|T| \geq 2.5$ and that the computed value of T from the data is -2.75 . Which, if any, of the following statements is correct?
- H_0 should be rejected.
 - The significance level α is the probability that, under H_0 , T is either greater than 2.75 or less than -2.75 .
 - The nonrejection region is given by $-3.5 < T < 3.5$.
 - The nonrejection region consists of values of T above 3.5 or below -3.5 .
 - The P -value of this test is given by the area to the right of $T = 3.5$ for the distribution of T under H_0 .
- 21.** Suppose that $\bar{X}_1 = 125.2$ and $\bar{X}_2 = 125.4$ are the mean systolic blood pressures for two random samples of workers from different plants in the same industry. Suppose, further, that a test of $H_0: \mu_1 = \mu_2$ using these samples is rejected for $\alpha = .001$. Which of the following conclusions is most reasonable?
- There is a meaningful difference (clinically speaking) in population means but not a statistically significant difference.
 - The difference in population means is both statistically and meaningfully significant.
 - There is a statistically significant difference but not a meaningfully significant difference in population means.
 - There is neither a statistically significant nor a meaningfully significant difference in population means.
 - The sample sizes used must have been quite small.
- 22.** The choice of an alternative hypothesis (H_A) should depend primarily on (choose all that apply)
- the data obtained from the study.
 - what the investigator is interested in determining.
 - the critical region.
 - the significance level.
 - the power of the test.

23. For each of the areas in the accompanying figure, labeled a , b , c , and d , select an answer from the following: α , $1 - \alpha$, β , $1 - \beta$.



© Cengage Learning

24. Suppose that $H_0: \mu_1 = \mu_2$ is the null hypothesis and that $.10 < P < .25$. What is the most appropriate conclusion?
25. Suppose that $H_0: \mu_1 = \mu_2$ is the null hypothesis and that $.005 < P < .01$. Which of the following conclusions is most appropriate?
- Do not reject H_0 because P is small.
 - Reject H_0 because P is small.
 - Do not reject H_0 because P is large.
 - Reject H_0 because P is large.
 - Do not reject H_0 at $\alpha = .01$.

References

- Cohen, J. 1977. *Statistical Power Analysis for the Behavioral Sciences*, 2d ed. New York: Academic Press.
- Kupper, L. L., and Hafner, K. B. 1989. "How Appropriate Are Popular Sample Size Formulas?" *The American Statistician* 43(2): 101–5.
- Muller, K. E., and Peterson, B. L. 1984. "Practical Methods for Computing Power in Testing the Multivariate General Linear Hypothesis." *Computational Statistics and Data Analysis* 2: 143–58.

4

Introduction to Regression Analysis

4.1 Preview

Regression analysis is a statistical tool for evaluating the relationship of one or more independent variables X_1, X_2, \dots, X_k to a single, continuous dependent variable Y . It is most often used when the independent variables cannot be controlled, as when they are collected in a sample survey or other observational study. Nevertheless, it is equally applicable to more-controlled experimental situations.

In practice, a regression analysis is appropriate for several possibly overlapping situations, including the following:

Application 1 You want to *characterize the relationship* between the dependent and independent variables by determining the extent, direction, and strength of the association. For example ($k = 2$), in Thompson's (1972) study described in Chapter 1, one of the primary questions was to describe the extent, direction, and strength of the association between "patient satisfaction with medical care" (Y) and the variables "affective communication between patient and physician" (X_1) and "informational communication between patient and physician" (X_2).

Application 2 You seek a *quantitative formula* or equation to describe (e.g., predict) the dependent variable Y as a function of the independent variables X_1, X_2, \dots, X_k . For example ($k = 1$), a quantitative formula may be desired for a study of the effect of dosage of a blood-pressure-reducing treatment (X_1) on blood pressure change (Y).

Application 3 You want to describe quantitatively or qualitatively the relationship between X_1, X_2, \dots, X_k and Y but *control for the effects of still other variables* $X_{k+1}, X_{k+2}, \dots, X_{k+p}$, which you believe have an important relationship with the dependent variable. For example ($k = 2, p = 2$), a study of the epidemiology of chronic diseases might describe the relationship of blood pressure (Y) to smoking habits (X_1) and social class

(X_2), controlling for age (X_3) and weight (X_4). In Chapter 11, we will introduce an alternate notation that indicates such control variables with the letter C .

Application 4 You want to *determine which of several independent variables are important and which are not* for describing or predicting a dependent variable. You may want to control for other variables. You may also want to *rank* independent variables in their order of importance. In Thompson's (1972) study, for example ($k = 4$, $p = 2$), the researcher sought to determine for the dependent variable "satisfaction with medical care" (Y) which of the following independent variables were important descriptors: worry (X_1), desire (X_2), informational communication (X_3), and affective communication (X_4). It was also considered necessary to control for age (X_5) and education (X_6).

Application 5 You want to *determine the best mathematical model* for describing the relationship between a dependent variable and one or more independent variables. Any of the previous examples can be used to illustrate this goal.

Application 6 You want to *compare several derived regression relationships*. An example would be a study to determine whether smoking (X_1) is related to blood pressure (Y) in the same way for males as for females, controlling for age (X_2).

Application 7 You want to *assess the interactive effects of two or more independent variables* with regard to a dependent variable. For example, you may want to determine whether the relationship of alcohol consumption (X_1) to blood pressure level (Y) is different depending on smoking habits (X_2). In particular, the relationship between alcohol and blood pressure might be quite strong for heavy smokers but very weak for non-smokers. If so, we would say that there is *interaction* between alcohol and smoking. Then any conclusions about the relationship between alcohol and blood pressure must take into account whether—and possibly how much—a person smokes. More generally, if X_1 and X_2 interact in their joint effect on Y , then the relationship of either X variable to Y depends on the value of the other X variable.

Application 8 You want to *obtain a valid and precise estimate of one or more regression coefficients* from a larger set of regression coefficients in a given model. For example, you may want to obtain an accurate estimate of the coefficient of a variable measuring alcohol consumption (X_1) in a regression model that relates hypertension status (Y), a dichotomous response variable, to X_1 and several other control variables (e.g., age and smoking status). Such an estimate may be used to quantify the effect of alcohol consumption on hypertension status after adjustment for the effects of certain control variables also in the model.

4.2 Association versus Causality

A researcher must be cautious about interpreting the results obtained from a regression analysis or, more generally, from any form of analysis seeking to quantify an association (e.g., via a correlation coefficient) among two or more variables. Although the statistical computations used to produce an estimated measure of association may be correct, the estimate itself may be biased. Such bias may result from the method used to select subjects for the study, from errors in the information used in the statistical analyses, or even from other variables that

can account for the observed association but that have not been measured or appropriately considered in the analysis. (See Kleinbaum, Kupper, and Morgenstern 1982 or Kleinbaum 2002 for a discussion of validity in epidemiologic research.)

For example, if diastolic blood pressure and physical activity level were measured on a sample of individuals at a particular time, a regression analysis might suggest that, on the average, blood pressure decreases with increased physical activity; further, such an analysis may provide evidence (e.g., based on a confidence interval) that this association is of moderate strength and is statistically significant. If the study involved only healthy adults, however, or if physical activity level was measured inappropriately, or if such other factors as age, race, and sex were not correctly taken into account, the above conclusions might be rendered invalid or at least questionable.

Continuing with the preceding example, if the investigators were satisfied that the findings were basically valid (i.e., the observed association was not spurious), could they then conclude that a low level of physical activity is a cause of high blood pressure? The answer is an unequivocal *no!*

The finding of a “statistically significant” association in a particular study (no matter how well done) does not establish a causal relationship. To evaluate claims of causality, the investigator must consider criteria that are external to the specific characteristics and results of any single study.

It is beyond the scope of this text to discuss causal inference making. Nevertheless, we will briefly review some key ideas on this subject. Most strict definitions of causality (e.g., Blalock 1971; Susser 1973) require that a *change* in one variable (X) can *produce* a change in another variable (Y).¹ This suggests that, to demonstrate a cause–effect relationship between X and Y , *experimental proof* is required that a change in Y results from a change in X . Although it is needed, such experimental evidence is often impractical, infeasible, or even unethical to obtain, especially when considering risk factors (e.g., cigarette smoking or exposure to chemicals) that are potentially harmful to human subjects. Consequently, alternative criteria based on information *not* involving direct experimental evidence are typically employed when attempting to make causal inferences regarding variable relationships in human populations.

The recent decades have seen a large growth in methods and models for understanding causal relationships in the context of individual studies. Two graphical frameworks have been commonly employed to understand the relationships among concepts (and variables) under study. The first, known as the *sufficient component cause model (SCC)*, uses circular “pies” to represent unique causes of a disease outcome and slices of the pies to represent the components of those causes (Rothman, Greenland, and Lash 2008). When all component causes of a disease are present, that outcome will occur. SCC models are helpful for conceptualizing hypothesized causes of, and their relative and joint impacts on, the occurrence of a disease outcome (e.g., dependent variable) but are limited for describing situations with multiple outcomes and those involving confounding, an important determinant of spurious noncausal associations discussed in Chapter 11. A second graphical model, known as the *directed acyclic graph (DAG)*, has come to supersede the *path diagram* (Greenland, Pearl,

¹ An imperfect approximation to this ideal for real-world phenomena might be that, on the average, a change in Y is produced by a change in X .

and Robins 1999). In such a graph, hypothesized causal relationships between variables are mapped out using arrows that indicate which variable has a causal effect on the other. Multiple causes of an outcome, and even multiple outcomes, can be visualized, and confounding can be understood, even in the absence of data. The relationships suggested by a DAG can be used to inform more quantitative analyses. A DAG may help guide the appropriate selection of independent variables for regression analyses, which may then provide estimates of the hypothesized relationships. Other techniques, such as *path analysis* and the more general *structural equation modeling*, use serial correlation and regression-type analyses to more comprehensively understand the relationships implied in a DAG and the appropriateness of the model (Blalock 1971; Bollen 1989).

A widely used approach for making causal conjectures, particularly in the health and medical sciences, employs a judgmental (and more qualitative than quantitative) evaluation of the combined results from several studies, using a set of operational criteria generally agreed on as necessary (but not sufficient) for supporting a given causal theory. Efforts to define such a set of criteria were made in the late 1950s and early 1960s by investigators reviewing research on the health hazards of smoking. A list of general criteria for assessing the extent to which available evidence supports a causal relationship was formalized by Bradford Hill (1965), and this list has subsequently been adopted by many epidemiologic researchers. The list contains nine criteria:

1. *Strength of association.* The stronger an observed association appears over a series of different studies, the less likely it is that this association is spurious because of bias.
2. *Consistency of the findings.* Most or all studies concerned with a given causal hypothesis produce similar results. Of course, studies dealing with a given question may all have serious bias problems that can diminish the importance of observed associations.
3. *Specificity of the association.* The study factor (i.e., the suspected cause) is associated with only one effect (e.g., a specific disease). Many study factors have multiple effects, however, and most diseases have multiple causes.
4. *Lack of temporal ambiguity.* The hypothesized cause precedes the occurrence of the effect. (The ability to establish this time pattern depends on the study design used.)
5. *Dose-response effect.* The value of the dependent variable (e.g., the rate of disease development) changes in a meaningful pattern (e.g., increases) with the dose (or level) of the suspected causal agent under study. This is also known as a *biological gradient*.
6. *Biological and theoretical plausibility of the hypothesis.* The hypothesized causal relationship is consistent with current biological and theoretical knowledge. The current state of knowledge may nonetheless be insufficient to explain certain findings.
7. *Coherence of the evidence.* The findings do not seriously conflict with accepted facts about the outcome variable being studied (e.g., knowledge about the natural history of some disease).

8. *Experimental evidence.* The findings are supported by experimental or quasi-experimental evidence. Studies in which subjects have been observed with and without the condition or treatment, to either remove or put in place the outcome, provide the strongest caliber of evidence for causation. As mentioned previously, ideal randomized experimental studies are often not practical. Yet in the observational setting, quasi-experimental information may be a convincing substitute. An example of this would be an observed reduction in lung cancer deaths following the enactment of regulations that limit smoking.
9. *Analogy.* The observed association may be similar in nature to another setting where a causal link has been established.

Clearly, applying the above criteria to a given causal hypothesis is hardly a straightforward matter. Even if these criteria are all satisfied, a causal relationship cannot be claimed with complete certainty. Nevertheless, in the absence of solid experimental evidence, the use of such criteria may be a logical and practical way to address the issue of causality, especially with regard to studies on human populations.

4.3 Statistical versus Deterministic Models

Although *causality* cannot be established by statistical analyses, associations among variables can be well quantified in a *statistical* sense. With proper statistical design and analysis, an investigator can model the extent to which changes in independent variables are related to changes in dependent variables. However, *statistical models* developed by using regression or other multivariable methods must be distinguished from *deterministic models*.

The law of falling bodies in physics, for example, is a deterministic model that assumes an ideal setting: the dependent variable varies in a completely prescribed way according to a perfect (error-free) mathematical function of the independent variables.

Statistical models, on the other hand, allow for the possibility of error in describing a relationship. For example, in a study relating blood pressure to age, persons of the same age are unlikely to have exactly the same observed blood pressure. Nevertheless, with proper statistical methods, we might be able to conclude that, on the average, blood pressure increases with age. Further, appropriate statistical modeling can permit us to predict the expected blood pressure for a given age and to associate a measure of variability with that prediction. Through the use of probability and statistical theory, such statements take into account the uncertainty of the real world by means of measurement error and individual variability. Of course, because such statements are necessarily nondeterministic, they require careful interpretation. Unfortunately, such interpretation is often quite difficult to make.

4.4 Concluding Remarks

In this short chapter, we have informally introduced the general regression problem and indicated a variety of situations to which regression modeling can be applied. We have also cautioned the reader about the types of conclusions that can be drawn from such modeling efforts.

We now turn to the actual quantitative details involved in fitting a regression model to a set of data and then in estimating and testing hypotheses about important parameters in the model. In the next chapter, we will discuss the simplest form of regression model—a straight line. In subsequent chapters, we will consider more complex forms.

References

- Blalock, H. M., Jr., ed. 1971. *Causal Models in the Social Sciences*. Chicago: Aldine Publishing.
- Bollen, K. A. 1989. *Structural Equations with Latent Variables*. New York: John Wiley & Sons.
- Greenland, S.; Pearl, J.; and Robins, J. M. 1999. "Causal Diagrams for Epidemiologic Research." *Epidemiology* 10: 37–48.
- Hill, A. B. 1965. "The Environment and Disease: Association or Causation?" *Proceedings of the Royal Society of Medicine* 58: 295–300.
- Kleinbaum, D. G. 2002. *ActivEpi*. New York and Berlin: Springer Publishers.
- Kleinbaum, D. G.; Kupper, L. L.; and Morgenstern, H. 1982. *Epidemiologic Research: Principles and Quantitative Methods*. Belmont, Calif.: Lifetime Learning Publications.
- Rothman, K. J.; Greenland, S.; and Lash, T. L. 2008. *Modern Epidemiology*. Philadelphia: Lippincott, Williams, & Wilkins.
- Susser, M. 1973. *Causal Thinking in the Health Sciences*. New York: Oxford University Press.
- Thompson, S. J. 1972. "The Doctor–Patient Relationship and Outcomes of Pregnancy." Ph.D. dissertation, Department of Epidemiology, University of North Carolina, Chapel Hill.

5

Straight-line Regression Analysis

5.1 Preview

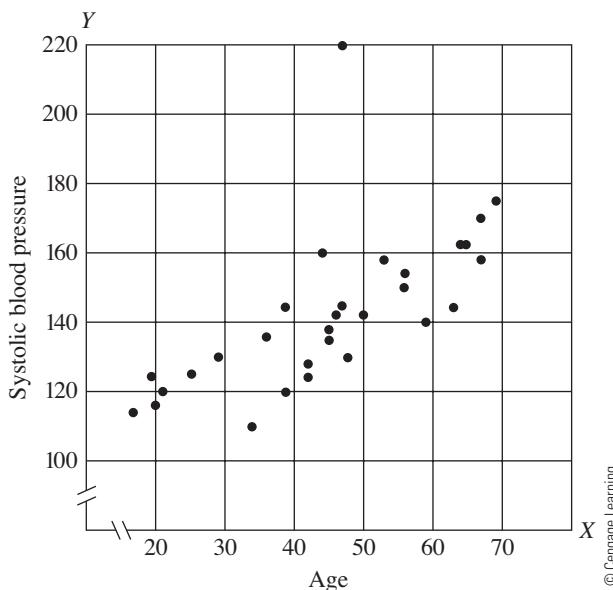
The simplest (but by no means trivial) form of the general regression problem deals with one dependent variable Y and one independent variable X . We have previously described the general problem in terms of k independent variables X_1, X_2, \dots, X_k . Let us now restrict our attention to the special case $k = 1$ but denote X_1 as X to keep our notation as simple as possible. To clarify the basic concepts and assumptions of regression analysis, we find it useful to begin with a single independent variable. Furthermore, researchers often begin by looking at one independent variable at a time even when several independent variables are eventually jointly considered.

5.2 Regression with a Single Independent Variable

We begin this section by describing the statistical problem of finding the *curve* (straight line, parabola, etc.) that *best fits* the data, closely approximating the true (but unknown) relationship between X and Y .

5.2.1 The Problem

Given a sample of n individuals (or other study units, such as animals, plants, geographical locations, time points, or pieces of physical material), we observe for each a value of X and a value of Y . We thus have n pairs of observations that can be denoted by $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, where the subscripts now refer to different individuals rather than different variables. Because these pairs may be considered as points in two-dimensional space, we can plot them on a graph. Such a graph is called a *scatter diagram*. For example, measurements of age and systolic blood pressure for 30 individuals might yield the scatter diagram given in Figure 5.1.



© Cengage Learning

FIGURE 5.1 Scatter diagram of age and systolic blood pressure

5.2.2 Basic Questions to Be Answered

Two basic questions must be dealt with in any regression analysis:

1. What is the most appropriate mathematical model to use—a straight line, a parabola, a log function, or what?
2. Given a specific model, what do we mean by and how do we determine the best-fitting model for the data? In other words, if our model is a straight line, how do we find the best-fitting line?

5.2.3 General Strategy

Several general strategies can be used to study the relationship between two variables by means of regression analysis. The most common of these is called the *forward method*. This strategy begins with a simply structured model—usually a straight line—and adds more complexity to the model in successive steps, if necessary. Another strategy, called the *backward method*, begins with a complicated model—such as a high-degree polynomial—and successively simplifies it, if possible, by eliminating unnecessary terms. A third approach uses a *model suggested from experience or theory*, which is revised either toward or away from complexity, as dictated by the data.

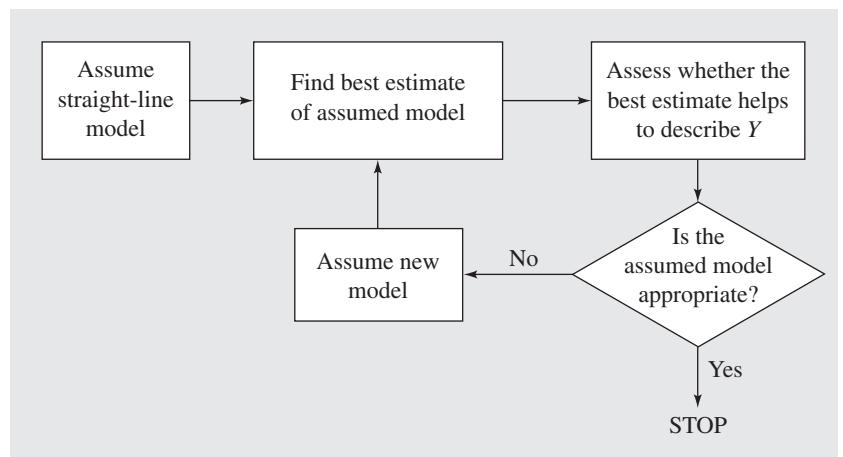
The strategy chosen depends on the type of problem and on the data; there are no hard-and-fast rules. The quality of the results often depends more on the skill with which a strategy is applied than on the particular strategy chosen. It is often tempting to try many strategies and then to use the results that provide the most “reasonable” interpretation of the relationship between the response and predictor variables. This exploratory approach demands particular care to ensure the reliability of any conclusions.

In Chapter 16, we will discuss in detail the issue of choosing a model-building strategy. For reasons discussed there, we often prefer the backward strategy. The forward method, however, corresponds more naturally to the usual development of theory from simple to complex. In some simple situations, forward and backward strategies lead to the same final model. In general, however, this is not the case!

Since it is the simplest method to understand and can, therefore, be used as a basis for understanding other methods, we begin by offering a step-by-step description of the forward strategy as applied to a regression model containing a single predictor.

1. Assume that a straight line is the appropriate model. Later the validity of this assumption can be investigated.
2. Find the best-fitting straight line, which is the line among all possible straight lines that best agrees (as will be defined later) with the data.
3. Determine whether the straight line found in Step 2 significantly helps to describe the dependent variable Y . Here it is necessary to check that certain basic statistical assumptions (e.g., normality) are met. These assumptions will be discussed in detail subsequently.
4. Examine whether the assumption of a straight-line model is correct. One approach for doing this is called *testing for lack of fit*, although other approaches can be used instead.
5. If the assumption of a straight line is found to be invalid in Step 4, fit a new model (e.g., a parabola) to the data, determine how well it describes Y (i.e., repeat Step 3), and then decide whether the new model is appropriate (i.e., repeat Step 4).
6. Continue to try new models until an appropriate one is found.

A flow diagram for this strategy is given in Figure 5.2.



© Cengage Learning

FIGURE 5.2 Flow diagram of the forward method

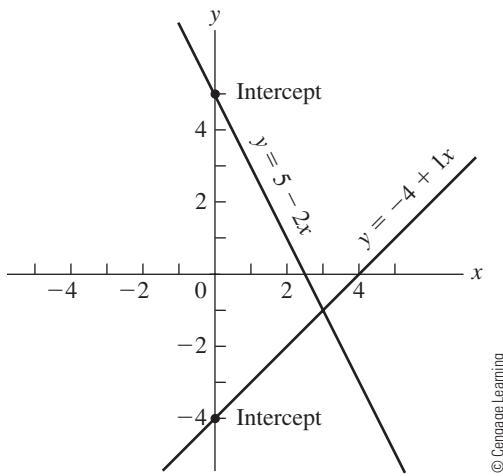
Since the usual (forward) approach to regression analysis with a single independent variable begins with the assumption of a straight-line model, we will consider this model first. Before describing the *statistical* methodology for this special case, let us review some basic straight-line mathematics. You may want to skip the next section if you are already familiar with its contents.

5.3 Mathematical Properties of a Straight Line

Mathematically, a straight line can be described by an equation of the form

$$y = \beta_0 + \beta_1 x \quad (5.1)$$

We have used lowercase letters y and x , instead of capital letters, in this equation to emphasize that we are treating these variables in a purely mathematical, rather than statistical, context. The symbols β_0 and β_1 have constant values for a given line and are, therefore, not considered variables; β_0 is called the *y-intercept* of the line, and β_1 is called the *slope*. Thus, $y = 5 - 2x$ describes a straight line with intercept 5 and slope -2 , whereas $y = -4 + 1x$ describes a different line with intercept -4 and slope 1. These two lines are shown in Figure 5.3.



© Cengage Learning

FIGURE 5.3 Straight-line plots

The intercept β_0 is the value of y when $x = 0$. For the line $y = 5 - 2x$, $y = 5$ when $x = 0$. For the line $y = -4 + 1x$, $y = -4$ when $x = 0$. The slope β_1 is the amount of change in y for each 1-unit increase in x . For any given straight line, this rate of change is always constant. Thus, for the line $y = 5 - 2x$, when x changes 1 unit from 3 to 4, y changes -2 units (the value of the slope) from $5 - 2(3) = -1$ to $5 - 2(4) = -3$; and when x changes from 1 to 2, also 1 unit, y changes from $5 - 2(1) = 3$ to $5 - 2(2) = 1$, also -2 units.

The properties of any straight line can be viewed graphically as in Figure 5.3. To graph a given line, plot any two points on the line and then connect them with a ruler. One of the two points often used is the *y-intercept*. This point is given by $(x = 0, y = 5)$ for the line

$y = 5 - 2x$ and by $(x = 0, y = -4)$ for $y = -4 + 1x$. The other point for each line may be determined by arbitrarily selecting an x and finding the corresponding y . An x of 3 was used in our two examples. Thus, for $y = 5 - 2x$, an x of 3 yields a y of $5 - 2(3) = -1$; and for $y = -4 + 1x$, an x of 3 yields a y of $-4 + 1(3) = -1$. The line $y = 5 - 2x$ can then be drawn by connecting the points $(x = 0, y = 5)$ and $(x = 3, y = -1)$, and the line $y = -4 + 1x$ can be drawn from the points $(x = 0, y = -4)$ and $(x = 3, y = -1)$.

As Figure 5.3 illustrates, in the equation $y = 5 - 2x$, y decreases as x increases. Such a line is said to have *negative* slope. Indeed, this definition agrees with the sign of the slope -2 in the equation. Conversely, the line $y = -4 + 1x$ is said to have *positive* slope, since y increases as x increases.

5.4 Statistical Assumptions for a Straight-line Model

Suppose that we have tentatively assumed a straight-line model as the first step in the forward method for determining the best model to describe the relationship between X and Y . We now want to determine the best-fitting line. Certainly, we will have no trouble deciding what is meant by “best fitting” if the data allow us to draw a single straight line through every point in the scatter diagram. Unfortunately, this will never happen with real-life data. For example, persons of the same age are unlikely to have the same blood pressure, height, or weight.

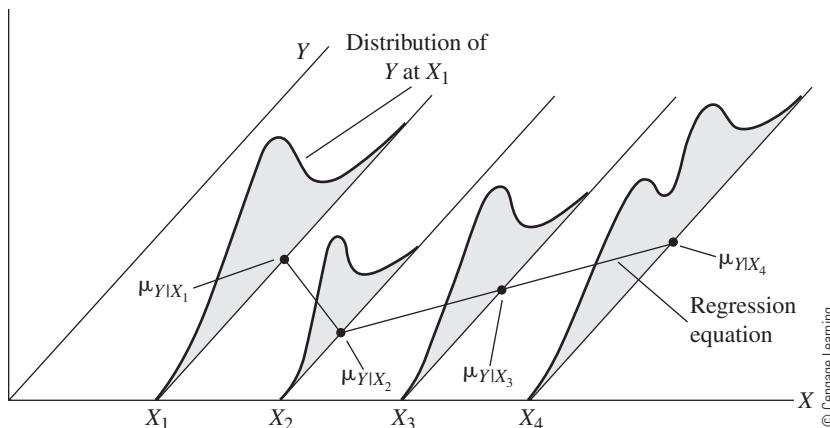
Thus, the straight line we seek can only approximate the true state of affairs and cannot be expected to predict precisely each individual's Y from that individual's X . In fact, this need to approximate would exist even if we measured X and Y on the whole population of interest instead of on just a sample from that population. In addition, the fact that the line is to be determined from the sample data and not from the population requires us to consider the problem of how to estimate unknown population parameters.

What are these parameters? The ones of primary concern at this point are the intercept β_0 and the slope β_1 of the straight line of the general mathematical form of (5.1) that best fits the X - Y data for the entire population. To make inferences from the sample about this population line, we need to make five statistical assumptions covering existence, independence, linearity, homoscedasticity, and normality.

5.4.1 Statement of Assumptions

Assumption 1: Existence For any fixed value of the variable X , Y is a random variable with a certain probability distribution having finite mean and variance. The (population) mean of this distribution will be denoted as $\mu_{Y|X}$ and the (population) variance as $\sigma_{Y|X}^2$. The notation “ $Y|X$ ” indicates that the mean and the variance of the random variable Y depend on the value of X .

This assumption applies to any regression model, whether a straight line or not. Figure 5.4 illustrates the assumption. The different distributions are drawn vertically to correspond to different values of X . The dots denoting the mean values $\mu_{Y|X}$ at different X 's have been connected to form the *regression equation*, which models how the population mean of Y varies with X and which is to be estimated from the data.

**FIGURE 5.4** General regression equation

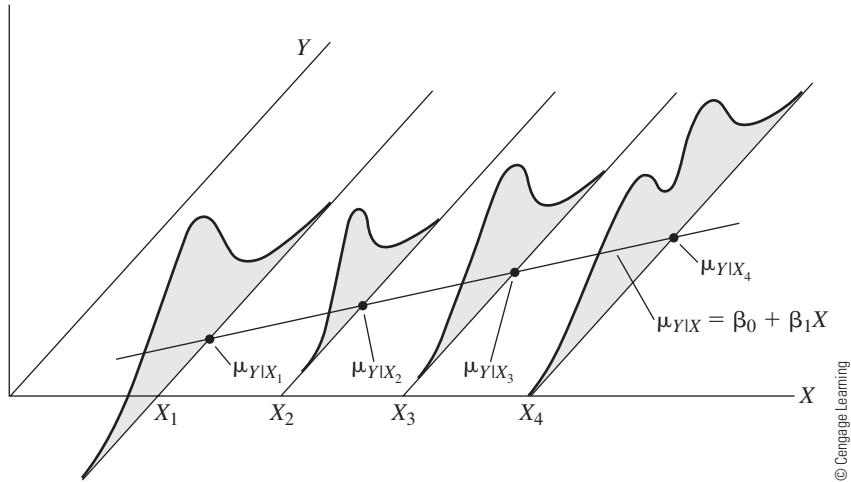
Assumption 2: Independence *The Y-values are statistically independent of one another.* This assumption is appropriate in many, but not all, situations. In particular, Assumption 2 is usually violated when different observations are made on the same individual at different times. For example, if weight is measured on an individual at different times (i.e., longitudinally over time), we can expect that the weight at one time is related to the weight at a later time. As another example, if blood pressure is measured on a given individual longitudinally over time, we can expect the blood pressure value at one time to be in the same range as the blood pressure value at the previous or following time. When Assumption 2 is not satisfied, ignoring the dependency among the Y-values can often lead to invalid statistical conclusions.

When the Y-values are not independent—that is, when they are “correlated”—special methods can be used to find the best-fitting model and to make valid statistical inferences. The method chosen depends on the characteristics of the response variable, the type of dependence, and the complexity of the problem. In Chapter 26, we describe a “mixed model” analysis of variance approach for designs involving repeated measurements on study subjects. In some cases, multivariate linear models are appropriate. See Morrison (1976) or Timm (1975) for a general introduction to multivariate linear models. More recently, Zeger and Liang (1986) introduced the “generalized estimating equations” (GEE) approach for analyzing correlated response data, and an excellent book on this very general and useful methodology is available (Diggle et al. 2002).

Assumption 3: Linearity *The mean value of Y, $\mu_{Y|X}$, is a straight-line function of X.* In other words, if the dots denoting the different mean values $\mu_{Y|X}$ are connected, a straight line is obtained. This assumption is illustrated in Figure 5.5.

Using mathematical symbols, we can describe Assumption 3 by the equation

$$\mu_{Y|X} = \beta_0 + \beta_1 X \quad (5.2)$$



© Cengage Learning

FIGURE 5.5 Straight-line assumption

where β_0 and β_1 are the intercept and the slope of this (population) straight line, respectively. Equivalently, we can express (5.2) in the form

$$Y = \beta_0 + \beta_1X + E \quad (5.3)$$

where E denotes a random variable that has mean 0 at fixed X (i.e., $\mu_{E|X} = 0$ for any X). More specifically, since X is fixed and not random, (5.3) represents the dependent variable Y as the sum of a constant term ($\beta_0 + \beta_1X$) and a random variable (E).¹ Thus, the probability distributions of Y and E differ only in the value of this constant term; that is, since E has mean 0, Y must have mean $\beta_0 + \beta_1X$.

Equations (5.2) and (5.3) describe a *statistical* model. These equations should be distinguished from the *mathematical* model for a straight line described by (5.1), which does not consider Y as a random variable.

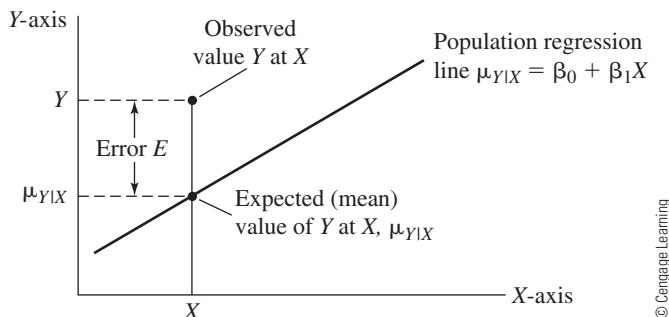
The variable E describes how distant an individual's response can be from the population regression line (Figure 5.6). In other words, what we observe at a given X (namely, Y) is in *error* from that expected on the average (namely, $\mu_{Y|X}$) by an amount E , which is random and varies from individual to individual. For this reason, E is commonly referred to as the *error component* in the model (5.3). Mathematically, E is given by the formula

$$E = Y - (\beta_0 + \beta_1X)$$

or by

$$E = Y - \mu_{Y|X}$$

¹ For our purposes, the practical implication of the statement "X is fixed and not random" for making statistical inferences from sample to population is that the *only* random component on the right-hand side of (5.3) is E when X is fixed.



© Cengage Learning

FIGURE 5.6 Error component E

This concept of an error component is particularly important for defining a good-fitting line, since, as we will see in the next section, a line that fits data well ought to have small deviations (or errors) between what is observed and what is predicted by the fitted model.

Assumption 4: Homoscedasticity *The variance of Y is the same for any X .* (*Hom-* means “same,” and *-scedastic* means “scattered.”) An example of the violation of this assumption (called *heteroscedasticity*) is shown in Figure 5.5, where the distribution of Y at X_1 has considerably more spread than the distribution of Y at X_2 . This means that $\sigma_{Y|X_1}^2$, the variance of Y at X_1 , is greater than $\sigma_{Y|X_2}^2$, the variance of Y at X_2 .

In mathematical terms, the homoscedastic assumption can be written as

$$\sigma_{Y|X}^2 \equiv \sigma^2$$

for all X . This formula is a shorthand way of saying that, since $\sigma_{Y|X_i}^2 = \sigma_{Y|X_j}^2$ for any two different values of X , we might as well simplify our notation by giving the common variance a single name—say, σ^2 —that does not involve X at all.

A number of techniques of varying statistical sophistication can be used to determine whether the homoscedastic assumption is satisfied. Some of these procedures will be discussed in Chapter 14.

Assumption 5: Normal Distribution *For any fixed value of X , Y has a normal distribution.* This assumption makes it possible to evaluate the statistical significance (e.g., by means of confidence intervals and tests of hypotheses) of the relationship between X and Y , as reflected by the fitted line.

Figure 5.5 provides an example in which this assumption is violated. In addition to the variances not being all equal in this figure, the distributions of Y at X_3 and at X_4 are not normal. The distribution at X_3 is skewed, whereas the normal distribution is symmetric. The distribution at X_4 is bimodal (two humps), whereas the normal distribution is unimodal (one hump). Methods for determining whether the normality assumption is tenable are described in Chapter 14.

If the normality assumption is not *badly* violated, the conclusions reached by a regression analysis in which normality is assumed will generally be reliable and accurate. This stability property with respect to deviations from normality is a type of *robustness*. Consequently, we

recommend giving considerable leeway before deciding that the normality assumption is so badly violated as to require alternative inference-making procedures.

If the normality assumption is deemed unsatisfactory, the Y -values may be transformed by using a log, square root, or other function to see whether the new set of observations is approximately normal. Care must be taken when using such transformations to ensure that other assumptions, such as variance homogeneity, are not violated for the transformed variable. Fortunately, in practice such transformations usually help satisfy both the normality and the variance homogeneity assumptions.

5.4.2 Summary and Comments

The assumptions of homoscedasticity and normality apply to the distribution of Y when X is fixed (i.e., Y given X) and not to the distributions of Y associated with different X -values. Many people find it more convenient to describe these two assumptions in terms of the error E . It is sufficient to say that the random variable E has a normal distribution with mean 0 and variance σ^2 for all observations. Of course, the linearity, existence, and independence assumptions must also be specified.

It is helpful to maintain distinctions among such concepts as *random variables*, *parameters*, and *point estimates*. The variable Y is a random variable, and an observation of it yields a particular value or “realization”; the variable X is assumed to be measured without error. The constants β_0 and β_1 are parameters with unknown but specific values for a particular population. The variable E is a random, unobservable variable. Using some estimation procedure (e.g., least squares), one constructs point estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1 , respectively. Once $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained, a point estimate of E at the value X is calculated as

$$\hat{E} = Y - \hat{Y} = Y - (\hat{\beta}_0 + \hat{\beta}_1 X)$$

The estimated error \hat{E} is typically called a *residual*. If there are n (X, Y) pairs $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, then there are n residuals $\hat{E}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$, $i = 1, 2, \dots, n$.

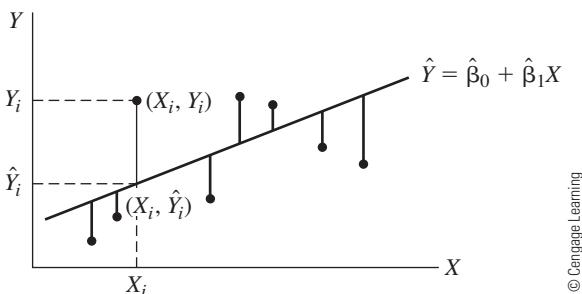
Some statisticians refer to a normally distributed random variable as having a Gaussian distribution. This terminology avoids confusing *normal* with its other meaning of “customary” or “usual”; it emphasizes the fact that the term *Gaussian* refers to a *particular* bell-shaped function; and it appropriately honors the mathematician Carl Gauss (1777–1855).

5.5 Determining the Best-fitting Straight Line

By far the simplest and quickest method for determining a straight line is to choose the line that can best be drawn by eye. Although this method often paints a reasonably good picture, it is extremely subjective and imprecise and is worthless for statistical inference. We now consider two analytical approaches for finding the best-fitting straight line.

5.5.1 The Least-squares Method

The *least-squares method* determines the best-fitting straight line as the line that minimizes the sum of squares of the lengths of the vertical-line segments (Figure 5.7) drawn from the observed data points on the scatter diagram to the fitted line. The idea here is that the smaller



© Cengage Learning

FIGURE 5.7 Deviations of observed points from the fitted regression line

the deviations of observed values from this line (and consequently the smaller the sum of squares of these deviations), the closer or “snugger” the best-fitting line will be to the data.

In mathematical notation, the least-squares method is described as follows. Let \hat{Y}_i denote the estimated response at X_i based on the fitted regression line; in other words, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the intercept and the slope of the fitted line, respectively. The vertical distance between the observed point (X_i, Y_i) and the corresponding point (X_i, \hat{Y}_i) on the fitted line is given by the absolute value $|Y_i - \hat{Y}_i|$ or $|Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)|$. The sum of the squares of all such distances is given by

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \text{SSE}$$

The least-squares solution is defined to be the choice of $\hat{\beta}_0$ and $\hat{\beta}_1$ for which the sum of squares just described is a minimum. In standard jargon, $\hat{\beta}_0$ and $\hat{\beta}_1$ are termed the *least-squares estimates* of the parameters β_0 and β_1 , respectively, in the statistical model given by (5.3).

The *minimum sum of squares* corresponding to the least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ is usually called the *sum of squares about the regression line*, the *residual sum of squares*, or the *sum of squares due to error* (SSE). The measure SSE is of great importance in assessing the quality of the straight-line fit, and its interpretation will be discussed in Section 5.6.

Mathematically, the essential property of the measure SSE can be stated in the following way. If β_0^* and β_1^* denote any other possible estimators of β_0 and β_1 , we must have

$$\text{SSE} = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \leq \sum_{i=1}^n (Y_i - \beta_0^* - \beta_1^* X_i)^2$$

5.5.2 The Minimum-Variance Method

The *minimum-variance method* is more classically statistical than the method of least squares, which can be viewed as a purely mathematical algorithm. In this second approach, determining the best fit becomes a statistical estimation problem. The goal is to find point estimators of β_0 and β_1 with good statistical properties. In this regard, under the previous assumptions, the best line is determined by the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that are *unbiased* for their unknown population counterparts β_0 and β_1 , respectively, and have minimum variance among all unbiased (linear) estimators of β_0 and β_1 .

5.5.3 Solution to the Best-fit Problem

Fortunately, both the least-squares method and the minimum-variance method yield exactly the same solution, which we will state without proof.²

Let \bar{Y} denote the sample mean of the observations on Y , and let \bar{X} denote the sample mean of the values of X . Then the best-fitting straight line is determined by the formulas

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (5.4)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (5.5)$$

In calculating $\hat{\beta}_0$ and $\hat{\beta}_1$, we recommend using a computer program for regression modeling from a convenient computer package. Many computer packages with regression programs are now available, and the packages SAS, SPSS, STATA, and R are introduced in appendices associated with this text. In this text, we will use SAS exclusively to present computer output, although we recognize that other packages may be preferred by particular users.

The least-squares line may generally be represented by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \quad (5.6)$$

or equivalently by

$$\hat{Y} = \bar{Y} + \hat{\beta}_1(X - \bar{X}) \quad (5.7)$$

Either (5.6) or (5.7) may be used to determine predicted Y 's that correspond to X 's actually observed or to other X -values in the region of experimentation. Simple algebra can be used to demonstrate the equivalence of (5.6) and (5.7). The right-hand side of (5.7), $\bar{Y} + \hat{\beta}_1(X - \bar{X})$, can be written as $\bar{Y} + \hat{\beta}_1X - \hat{\beta}_1\bar{X}$, which, in turn, equals $\bar{Y} - \hat{\beta}_1\bar{X} + \hat{\beta}_1X$, which is equivalent to (5.6), since $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$ from (5.5).

Table 5.1 lists observations on systolic blood pressure and age for a sample of 30 individuals. The scatter diagram for this sample was presented in Figure 5.1. For this data set, the output from the SAS package's PROC REG routine is also shown.

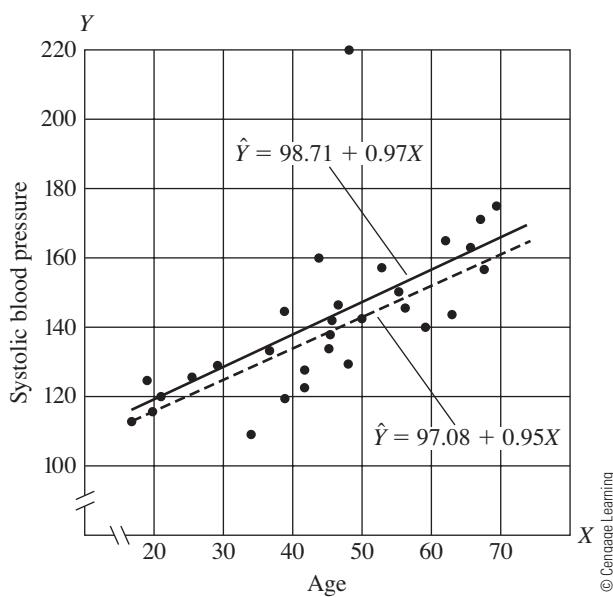
The estimated line (5.6) is computed to be $\hat{Y} = 98.71 + 0.97X$ and is graphed as the solid line in Figure 5.8. This line reflects the clear trend that systolic blood pressure increases

² Another general method of parameter estimation is called *maximum likelihood*. Under the assumptions of a Gaussian distribution for Y , given X fixed and mutual independence of the Y 's, the maximum-likelihood estimates of β_0 and β_1 are exactly the same as the least-squares and minimum-variance estimates. A general discussion of maximum-likelihood methods is given in Chapter 21.

TABLE 5.1 Observations on systolic blood pressure (SBP) and age for a sample of 30 individuals

Individual (<i>i</i>)	SBP (<i>Y</i>)	Age (<i>X</i>)	Individual (<i>i</i>)	SBP (<i>Y</i>)	Age (<i>X</i>)
1	144	39	16	130	48
2	220	47	17	135	45
3	138	45	18	114	17
4	145	47	19	116	20
5	162	65	20	124	19
6	142	46	21	136	36
7	170	67	22	142	50
8	124	42	23	120	39
9	158	67	24	120	21
10	154	56	25	160	44
11	162	64	26	158	53
12	150	56	27	144	63
13	140	59	28	130	29
14	110	34	29	125	25
15	128	42	30	175	69

© Cengage Learning



© Cengage Learning

FIGURE 5.8 Best-fitting straight line to age–systolic blood pressure data of Table 5.1

Edited SAS Output (PROC REG) for Data of Table 5.1

Regression of SBP (Y) on Age (X)

DESCRIPTIVE STATISTICS					
Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	30.00000	1.00000	30.00000	0	0
X	1354.00000	$\bar{X} \rightarrow 45.13333$	67894	$S_x^2 \rightarrow 233.91264$	15.29420
Y	4276.00000	$\bar{Y} \rightarrow 142.53333$	624260	$S_y^2 \rightarrow 509.91264$	22.58125

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6394.02269	6394.02269	21.33	<.0001
Error	28	8393.44398	$S_{Y X}^2 \rightarrow 299.76586$		
Corrected Total	29	14787			

Root MSE	$S_{Y X} \rightarrow 17.31375$	R-Square	0.4324
Dependent Mean	142.53333	Adj R-Sq	0.4121
Coeff Var	12.14716		

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	$\hat{\beta}_0 \rightarrow 98.71472$	10.00047	9.87	<.0001
X	1	$\hat{\beta}_1 \rightarrow 0.97087$	0.21022	4.62	<.0001

© Cengage Learning

as age increases. Notice that one point, (47, 220), seems quite out of place with the other data; such an observation is often called an *outlier*. Because an outlier can affect the least-squares estimates, the determination whether an outlier should be removed from the data is important. Usually, this decision can be made only after thorough evaluation of the experimental conditions, the data collection process, and the data themselves. (See Chapter 14 for further discussion of the treatment of outliers.) If the decision is difficult, one can always determine the effect of removing the outlier by refitting the model to the remaining data. In this case, the resulting least-squares line is

$$\hat{Y} = 97.08 + 0.95X$$

and is shown on the graph in Figure 5.8 as the dashed line. As might be expected, this line is slightly below the one obtained by using all the data.

5.6 Measure of the Quality of the Straight-line Fit and Estimate of σ^2

Once the least-squares line is determined, we want to evaluate whether the fitted line actually aids in predicting Y —and if so, to what extent. A measure that helps to answer these questions is provided by

$$\text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. Clearly, if SSE = 0, the straight line fits perfectly; that is, $Y_i = \hat{Y}_i$ for each i , and every observed point lies on the fitted line. As the fit gets worse, SSE gets larger, since the deviations of points from the regression line become larger.

Two possible scenarios contribute to the inflation of SSE. First, there may be a lot of random variation in the response Y ; that is, σ^2 may be large. Second, the assumption of a straight-line model may not be appropriate. It is important, therefore, to determine the separate effects of each of these scenarios, since they address decidedly different issues with regard to the fit of the model. For the time being, we will assume that the second factor is not at issue. Thus, assuming that the straight-line model is appropriate, we can obtain an estimate of σ^2 by using SSE. Such an estimate is needed for making statistical inferences about the true (i.e., population) straight-line relationship between X and Y . This estimate of σ^2 is given by the formula

$$S_{Y|X}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \frac{1}{n-2} \text{SSE} \quad (5.8)$$

■ **Example 5.1** From the computer output provided on p. 59, we find

$$S_{Y|X}^2 = 299.76586$$

Readers may wonder why $S_{Y|X}^2$ estimates σ^2 —especially since, at first glance, (5.8) looks different from the formula usually used for the sample variance of Y —namely, $S_Y^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 / (n - 1)$. The latter formula is appropriate when the Y 's are mutually independent, with the same mean μ and variance σ^2 . Since μ is unknown in this case (its estimate being, of course, \bar{Y}), we must divide by $n - 1$ instead of n to make the sample variance an unbiased estimator of σ^2 . To put it another way, we subtract 1 from n because to estimate σ^2 we had to estimate *one* other parameter first, μ .

If a straight-line model is appropriate, the population mean response $\mu_{Y|X}$ changes with X . For example, using the least-squares line (5.6) as an approximation to the population line for the age–systolic blood pressure data, the estimated mean of the Y 's at $X = 40$ is approximately 138, whereas the estimated mean of the Y 's at $X = 70$ is close to 167. Therefore, instead of subtracting \bar{Y} from each Y_i when estimating σ^2 , we should subtract \hat{Y}_i from Y_i because $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ is the estimate of $\mu_{Y|X}$. Furthermore, we subtract 2 from n in the

denominator of our estimate, since the determination of \hat{Y}_i requires the estimation of two parameters, β_0 and β_1 .

When we discuss testing for lack of fit of the assumed model, we will show how it is possible in some circumstances to obtain a reliable estimate of σ^2 that does not assume the correctness of the straight-line model. ■

5.7 Inferences about the Slope and Intercept

To assess whether the fitted line helps to describe the relationship between X and Y , as well as to take into account the uncertainties in sample data, it is standard practice to compute confidence intervals and/or test statistical hypotheses about the unknown parameters in the assumed straight-line model. Such confidence intervals and tests require, as described in Section 5.4, the assumptions that the random variable Y has a normal distribution at each fixed value of X and that Y_1, Y_2, \dots, Y_n are mutually independent given X_1, X_2, \dots, X_n . Working from these assumptions, one can deduce that the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are each normally distributed, with respective means β_0 and β_1 when (5.2) holds and with easily derivable variances.³ These estimators, together with estimators of their variances, can then be used to form confidence intervals and test statistics based on the *t distribution*.

More specifically, to test the hypothesis $H_0: \beta_1 = \beta_1^{(0)}$, where $\beta_1^{(0)}$ is some hypothesized value for β_1 , the test statistic used is

$$T = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{S_{\hat{\beta}_1}}$$

The denominator $S_{\hat{\beta}_1}$ in this test statistic is an estimate of the unknown standard error of the estimator $\hat{\beta}_1$ —namely,

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{S_X \sqrt{n-1}}$$

³ An important property that allows the normality assumption on Y to carry over to $\hat{\beta}_0$ and $\hat{\beta}_1$ is that these estimators are *linear functions* of the mutually independent Y 's. Such a function is defined by a formula of the structure

$$L = \sum_{i=1}^n c_i Y_i$$

or, equivalently, $L = c_1 Y_1 + c_2 Y_2 + \dots + c_n Y_n$, where the c_i 's are constants not involving the Y 's. A simple example of a linear function is \bar{Y} , which can be written as

$$\sum_{i=1}^n \frac{1}{n} Y_i$$

Here the c_i 's equal $1/n$ for each i . The normality of $\hat{\beta}_0$ and $\hat{\beta}_1$ derives from a statistical theorem stating that linear functions of mutually independent normally distributed observations are themselves normally distributed. This property of linear functions of mutually independent normal variables is utilized in later chapters (e.g., Chapter 22).

This unknown parameter is estimated from a sample using

$$S_{\hat{\beta}_1} = \frac{S_{Y|X}}{S_X \sqrt{n - 1}}$$

so that

$$T = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{\left(\frac{S_{Y|X}}{S_X \sqrt{n - 1}} \right)} \quad (5.9)$$

The test statistic for testing $H_0: \beta_0 = \beta_0^{(0)}$ is structured identically to the test statistic (5.9) for β_1 . In particular, the test statistic denominator estimates its standard error

$$\sigma_{\hat{\beta}_0} = \sigma \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n - 1)S_X^2}}$$

with

$$S_{\hat{\beta}_0} = S_{Y|X} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n - 1)S_X^2}}$$

producing the test statistic

$$T = \frac{\hat{\beta}_0 - \beta_0^{(0)}}{S_{Y|X} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n - 1)S_X^2}}} \quad (5.10)$$

Both test statistics are thus structured as the ratio of a normally distributed random variable minus its hypothesized mean divided by an independent estimator of its standard error. Such statistics have a t distribution for the kinds of situations encountered in this text; in this case, both test statistics have a t distribution with $n - 2$ degrees of freedom when H_0 is true.

The reason why both test statistics (5.9) and (5.10) have $n - 2$ degrees of freedom is that both involve $S_{Y|X}^2$, which itself has $n - 2$ degrees of freedom and is the only random component in the denominator of both statistics.

In testing either of the preceding hypotheses at a significance level α , we should reject H_0 whenever any of the following occur:

$$\begin{cases} T \geq t_{n-2, 1-\alpha} & \text{for an upper one-tailed test} \\ T \leq -t_{n-2, 1-\alpha} & \text{for a lower one-tailed test} \\ |T| \geq t_{n-2, 1-\alpha/2} & \text{for a two-tailed test} \end{cases} \quad \begin{array}{l} (\text{i.e., } H_A: \beta_1 > \beta_1^{(0)} \text{ or } H_A: \beta_0 > \beta_0^{(0)}) \\ (\text{i.e., } H_A: \beta_1 < \beta_1^{(0)} \text{ or } H_A: \beta_0 < \beta_0^{(0)}) \\ (\text{i.e., } H_A: \beta_1 \neq \beta_1^{(0)} \text{ or } H_A: \beta_0 \neq \beta_0^{(0)}) \end{array}$$

TABLE 5.2 Confidence intervals, tests of hypotheses, and prediction intervals for straight-line regression analysis

Quantity Estimated	100(1 - α)% Interval Estimate	H_0	Test Statistic (T)	Distribution under H_0
β_1	$\hat{\beta}_1 \pm t_{n-2, 1-\alpha/2} S_{\hat{\beta}_1}$	$\beta_1 = \beta_1^{(0)}$	$T = \frac{(\hat{\beta}_1 - \beta_1^{(0)})}{S_{\hat{\beta}_1}}$	t_{n-2}
β_0	$\hat{\beta}_0 \pm t_{n-2, 1-\alpha/2} S_{\hat{\beta}_0}$	$\beta_0 = \beta_0^{(0)}$	$T = \frac{(\hat{\beta}_0 - \beta_0^{(0)})}{S_{\hat{\beta}_0}}$	t_{n-2}
$\mu_{Y X_0}$	$\bar{Y} + \hat{\beta}_1(X_0 - \bar{X}) \pm t_{n-2, 1-\alpha/2} S_{\hat{Y} X_0}$	$\mu_{Y X_0} = \mu_{Y X_0}^{(0)}$	$T = \frac{\bar{Y} + \hat{\beta}_1(X_0 - \bar{X}) - \mu_{Y X_0}^{(0)}}{S_{\hat{Y} X_0}}$	t_{n-2}
Y_{X_0}	$\bar{Y} + \hat{\beta}_1(X_0 - \bar{X}) \pm t_{n-2, 1-\alpha/2} S_{Y X} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)S_X^2}}$			Prediction interval

Note: $\mu_{Y|X} = \beta_0 + \beta_1 X$ is the assumed true regression model.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\mu}_{Y|X_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0 = \bar{Y} + \hat{\beta}_1(X_0 - \bar{X})$$

Y_{X_0} = the predicted value of Y at $X = X_0 = \hat{\mu}_{Y|X_0}$

$t_{n-2, 1-\alpha/2}$ is the $100(1 - \alpha/2)\%$ point of the t distribution with $n - 2$ degrees of freedom.

$$S_{Y|X}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad S_{\hat{\beta}_1} = \frac{S_{Y|X}}{S_X \sqrt{n-1}}$$

$$S_{\hat{\beta}_0}^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad S_{\hat{\beta}_0} = S_{Y|X} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}}$$

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad S_{\hat{Y}|X_0} = S_{Y|X} \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)S_X^2}}$$

© Cengage Learning

where $t_{n-2, 1-\alpha}$ denotes the $100(1 - \alpha)\%$ point of the t distribution with $n - 2$ degrees of freedom. As an alternative to using a specified significance level, we may compute P -values based on the calculated value of the test statistic T .

Table 5.2 summarizes the formulas needed for performing statistical tests and computing confidence intervals for β_0 and β_1 . Also given in this table are formulas for inference-making procedures concerned with prediction using the fitted line; these formulas are described in Sections 5.9 and 5.10. Table 5.3 gives examples illustrating the use of each formula in Table 5.2, using the age-systolic blood pressure data previously considered.

TABLE 5.3 Sample calculations of confidence intervals, tests of hypotheses, and prediction intervals for the age–systolic blood pressure data of Table 5.1

Quantity Estimated	100 (1 – α)% Interval Estimate	H_0	Test Statistic (T)
β_1	For $\alpha = .05$: $0.97 \pm 2.0484(0.21)$ or (0.54, 1.40)	$\beta_1 = 0$	$T = \frac{(0.97 - 0)}{0.21} = 4.62$ Reject H_0 at $\alpha = .05$ (two-tailed test), since $t_{28, 0.975} = 2.0484$ ($P < .001$).
β_0	For $\alpha = .05$: $98.71 \pm (2.0484)(10.00)$ or (78.23, 119.20)	$\beta_0 = 75$	$T = \frac{(98.71 - 75)}{10.00} = 2.37$ Reject H_0 at $\alpha = .05$ (two-tailed test), since $t_{28, 0.975} = 2.0484$ ($.02 < P < .05$).
$\mu_{Y X_0} = 65$	For $\alpha = .10$: $142.53 + (0.97)(65 - 45.13) \pm (1.7011)(5.24)$ or (152.89, 170.72)	$\mu_{Y X_0=65} = 147$	$T = \frac{142.53 + (0.97)(65 - 45.13) - 147}{5.24}$ = 2.82 Reject H_0 at $\alpha = .10$ (two-tailed test), since $t_{28, 0.95} = 1.7011$ ($.001 < P < .01$).
$Y_{X_0} = 65$	For $\alpha = .10$: $142.53 + (0.97)(65 - 45.13)$ $\pm (1.7011)(17.31)\sqrt{1 + \frac{1}{30} + \frac{(65 - 45.13)^2}{(30 - 1)(15.29)^2}}$ or (131.04, 192.5)		

Note: $n = 30$, $\hat{\beta}_0 = 98.71$, $\hat{\beta}_1 = 0.97$, $\bar{Y} = 142.53$, $\bar{X} = 45.13$, $S_{Y|X} = 17.31$, $S_X = 15.29$,

$$S_{\hat{\beta}_1} = \frac{17.31}{15.29\sqrt{30 - 1}} = 0.21, S_{\hat{\beta}_0} = 17.31\sqrt{\frac{1}{30} + \frac{(45.13)^2}{(30 - 1)(15.29)^2}} = 10.00, S_{Y_{X_0}} = 17.31\sqrt{\frac{1}{30} + \frac{(65 - 45.13)^2}{(30 - 1)(15.29)^2}} = 5.24$$

© Cengage Learning

5.8 Interpretations of Tests for Slope and Intercept

Researchers often make errors when interpreting the results of tests regarding the slope and the intercept. In this section, we discuss conclusions that can be drawn based on non-rejection or rejection of the most common null hypotheses involving the slope and the

intercept.⁴ In our discussion, we assume that the usual assumptions about normality, independence, and variance homogeneity are not violated. If these assumptions do not hold, any conclusions based on testing procedures developed under the assumptions are suspect.

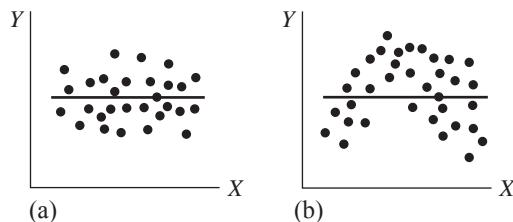
5.8.1 Test for Zero Slope

The most important test of hypothesis dealing with the parameters of the straight-line model relates to whether the slope of the regression line differs significantly from zero or, equivalently, whether X helps to predict Y using a straight-line model. The appropriate null hypothesis for this test is $H_0: \beta_1 = 0$. Care must be taken in interpreting the result of the test of this hypothesis.

If we ignore for now the ever-present possibilities of making a Type I error (i.e., rejecting a true H_0) or a Type II error (i.e., not rejecting a false H_0), we can make the following interpretations:

1. If $H_0: \beta_1 = 0$ is not rejected, one of the following is true:
 - a. For a true underlying straight-line model, X provides little or no help in predicting Y ; that is, \bar{Y} is essentially as good as $\bar{Y} + \hat{\beta}_1(X - \bar{X})$ for predicting Y (Figure 5.9(a)).
 - b. The true underlying relationship between X and Y is *not* linear; that is, the true model may involve quadratic, cubic, or other more complex functions of X (Figure 5.9(b)).

Examples when $H_0: \beta_1 = 0$ is not rejected



Examples when $H_0: \beta_1 = 0$ is rejected

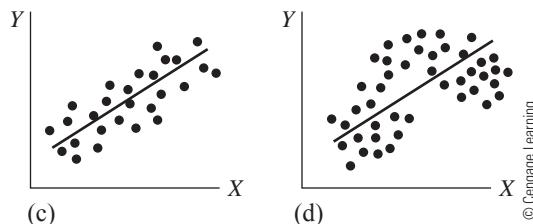


FIGURE 5.9 Interpreting the test for zero slope

⁴ Statistically speaking, “nonrejection of H_0 ” really means “determination that there is insufficient evidence to reject H_0 .”

Combining (a) and (b), we can say that *not rejecting $H_0: \beta_1 = 0$ implies that a straight-line model in X is not the best model to use and does not provide much help for predicting Y .*

2. If $H_0: \beta_1 = 0$ is rejected, one of the following is true:
 - a. X provides significant information for predicting Y ; that is, the model $\bar{Y} + \hat{\beta}_1(X - \bar{X})$ is far better than the naive model \bar{Y} for predicting Y (Figure 5.9(c)).
 - b. A better model might have, for example, a curvilinear term (e.g., Figure 5.9(d)), although there is statistical evidence of a linear component.

Combining (a) and (b), we can say that *rejecting $H_0: \beta_1 = 0$ implies that a straight-line model in X is better than a model that does not include X at all, although it may well represent only a linear approximation to a truly nonlinear relationship.*

An important point implied by these interpretations is that *whether or not the hypothesis $H_0: \beta_1 = 0$ is rejected, a straight-line model may not be appropriate; instead, some other curve may describe the relationship between X and Y better.*

5.8.2 Test for Zero Intercept

Another hypothesis sometimes tested involves whether the population straight line goes through the origin—that is, whether its Y -intercept β_0 is zero. The null hypothesis here is $H_0: \beta_0 = 0$. If this null hypothesis is not rejected, it may be appropriate to remove the constant from the model, provided that previous experience or a relevant theory suggests that the line may go through the origin and provided that observations are taken around the origin to improve the estimate of β_0 . Forcing the fitted line through the origin merely because $H_0: \beta_0 = 0$ cannot be rejected may give a spurious appearance to the regression line. In any case, this hypothesis is rarely of interest in most studies because data are not usually gathered near the origin. For example, when dealing with age (X) and blood pressure (Y), we are not interested in knowing what happens at $X = 0$, and we rarely choose values of X near 0.

5.9 The Mean Value of Y at a Specified Value of X

In addition to making inferences about the slope and the intercept, we may also want to perform tests and/or compute confidence intervals concerning the regression line itself. More specifically, for a given $X = X_0$, we may want to find a confidence interval for $\mu_{Y|X_0}$, the mean value of Y at X_0 .⁵ We may also be interested in testing the hypothesis $H_0: \mu_{Y|X_0} = \mu_{Y|X_0}^{(0)}$ where $\mu_{Y|X_0}^{(0)}$ is some hypothesized value of interest.

⁵ The point $(X_0, \mu_{Y|X_0})$, of course, lies on the population regression line.

The test statistic to use for the hypothesis $H_0: \mu_{Y|X_0} = \mu_{Y|X_0}^{(0)}$ is given by the formula

$$T = \frac{\hat{Y}_{X_0} - \mu_{Y|X_0}^{(0)}}{S_{\hat{Y}_{X_0}}} \quad (5.11)$$

where $\hat{Y}_{X_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0 = \bar{Y} + \hat{\beta}_1(X_0 - \bar{X})$ is the predicted value of Y at X_0 and

$$S_{\hat{Y}_{X_0}} = S_{Y|X} \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n - 1)S_X^2}} \quad (5.12)$$

This test statistic, like those for slope and intercept, has the t distribution with $n - 2$ degrees of freedom when H_0 is true. The denominator $S_{\hat{Y}_{X_0}}$ is an estimate of the standard error of \hat{Y}_{X_0} , which is given by

$$\sigma_{\hat{Y}_{X_0}} = \sigma \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n - 1)S_X^2}}$$

The corresponding confidence interval for $\mu_{Y|X_0}$ at a given $X = X_0$ is given by the formula

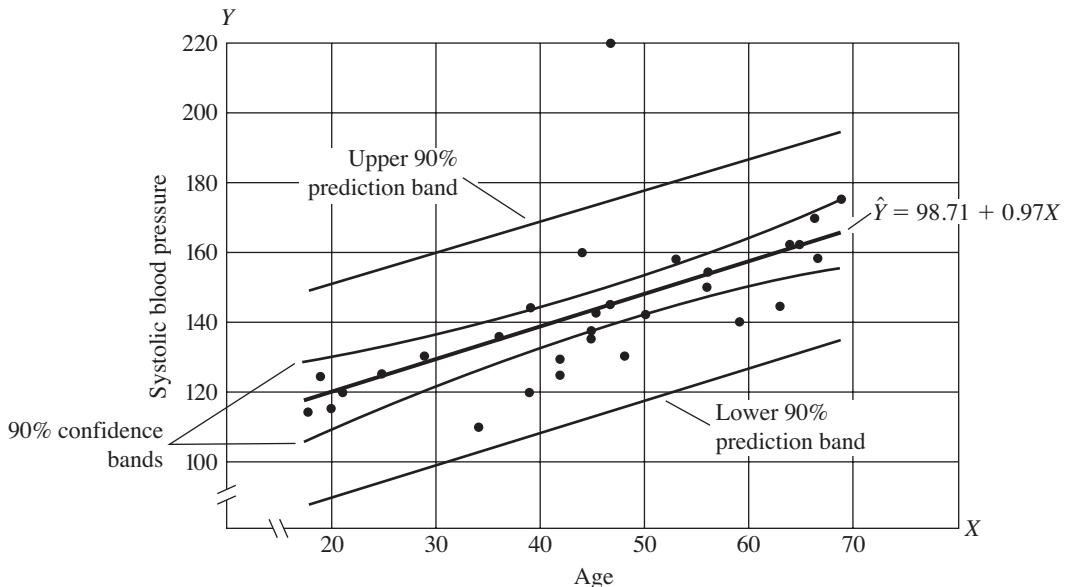
$$\hat{Y}_{X_0} \pm t_{n-2, 1-\alpha/2} S_{\hat{Y}_{X_0}} \quad (5.13)$$

In addition to drawing inferences about specific points on the regression line, researchers often find it useful to construct a confidence interval for the regression line over the entire range of X -values. The most convenient way to do this is to plot the upper and lower confidence limits obtained for several specified values of X and then to sketch the two curves that connect these points. Such curves are called *confidence bands* for the regression line. The 90% confidence bands for the data of Table 5.1 are indicated in Figure 5.10.

Sketching confidence bands by hand calculation can be a painful job. Instead, we generally recommend using a computer program for regression analysis to compute confidence intervals for a range of X_0 values and then to plot these intervals on the same graph that contains the fitted regression line. A convenient way to choose X_0 values is to use $X_0 = \bar{X}$, $X_0 = \bar{X} \pm k$, $X_0 = \bar{X} \pm 2k$, $X_0 = \bar{X} \pm 3k$, and so on, where k is chosen so that the range of X -values in the data is uniformly covered.

■ Example 5.2 For 90% confidence bands for our age–systolic blood pressure data, confidence interval formula (5.13) simplifies to

$$142.53 + (0.97)(X_0 - 45.13) \pm (1.7011)(17.31) \sqrt{0.033 + \frac{(X_0 - 45.13)^2}{6,783.48}} \quad (5.14)$$



© Cengage Learning

FIGURE 5.10 90% confidence and prediction bands for age–systolic blood pressure data of Table 5.1

At $X_0 = \bar{X} = 45.13$, the formula simplifies to $142.53 \pm (1.7011)(17.31)\sqrt{0.033}$, which yields a lower limit of 137.18 and an upper limit of 147.90. Notice that the minimum-width confidence interval is always obtained at $X_0 = \bar{X}$, since the second term under the square root sign in (5.14) is zero.

At $X_0 = \bar{X} \pm k$, the confidence interval formula becomes

$$142.53 \pm (0.97)k \pm (1.7011)(17.31)\sqrt{0.033 + \frac{k^2}{6,783.48}}$$

Thus, when $k = 10$, the confidence limits are 145.78 and 158.70 for $X_0 = \bar{X} + 10$ and 126.37 and 139.28 for $X_0 = \bar{X} - 10$. Figure 5.10 shows the 90% confidence bands for these data, together with the fitted model.

In SAS, PROC REG will compute 95% confidence intervals for $\mu_{Y|X}$, using the observed values of X as different X_0 values in the formula (5.13). In the SAS output on the next page, the lower and upper 95% confidence limits for the systolic blood pressure–age example appear in the columns labeled “95% CL Mean.” ■

5.10 Prediction of a New Value of Y at X_0

We have just dealt with estimating the mean $\mu_{Y|X_0}$ at $X = X_0$. In practice, we may want instead to estimate the response Y of a single individual, based on the fitted regression line; that is, we may want to predict an individual's Y given his or her $X = X_0$. The obvious point estimate to use in this case is $\hat{Y}_{X_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0$. In SAS, PROC REG will compute 95% confidence bands for $\mu_{Y|X_0}$ using all the observed values of X .

Edited SAS Output Showing 95% Confidence and Prediction Intervals

Regression of SBP (Y) on Age (X)							
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict	
				\hat{Y}_{X_0}	$S_{\hat{Y}_{X_0}}$	OUTPUT STATISTICS	95% confidence and prediction intervals
1	144.0000	136.5787	3.4139	129.5857	143.5717	100.4302	172.7271
2	220.0000	144.3456	3.1853	137.8208	150.8704	108.2848	180.4064
3	138.0000	142.4039	3.1612	135.9285	148.8792	106.3520	178.4558
4	145.0000	144.3456	3.1853	137.8208	150.8704	108.2848	180.4064
5	162.0000	161.8213	5.2377	151.0923	172.5502	124.7684	198.8742
6	142.0000	143.3748	3.1663	136.8889	149.8606	107.3210	179.4285
7	170.0000	163.7630	5.5787	152.3356	175.1905	126.5018	201.0242
8	124.0000	139.4913	3.2289	132.8771	146.1055	103.4142	175.5684
9	158.0000	163.7630	5.5787	152.3356	175.1905	126.5018	201.0242
10	154.0000	153.0835	3.9001	145.0946	161.0724	116.7292	189.4377
11	162.0000	160.8504	5.0717	150.4616	171.2393	123.8945	197.8063
12	150.0000	153.0835	3.9001	145.0946	161.0724	116.7292	189.4377
13	140.0000	155.9961	4.2999	147.1881	164.8041	119.4531	192.5391
14	110.0000	131.7243	3.9332	123.6676	139.7810	95.3551	168.0935
15	128.0000	139.4913	3.2289	132.8771	146.1055	103.4142	175.5684
16	130.0000	145.3165	3.2180	138.7248	151.9082	109.2435	181.3895
17	135.0000	142.4039	3.1612	135.9285	148.8792	106.3520	178.4558
18	114.0000	115.2195	6.7058	101.4832	128.9558	77.1867	153.2523
19	116.0000	118.1321	6.1568	105.5204	130.7439	80.4909	155.7734
20	124.0000	117.1613	6.3382	104.1781	130.1444	79.3939	154.9286
21	136.0000	133.6661	3.6984	126.0901	141.2420	97.4003	169.9318
22	142.0000	147.2582	3.3225	140.4525	154.0640	111.1455	183.3709
23	120.0000	136.5787	3.4139	129.5857	143.5717	100.4302	172.7271
24	120.0000	119.1030	5.9774	106.8588	131.3472	81.5833	156.6227
25	160.0000	141.4330	3.1700	134.9395	147.9265	105.3779	177.4882
26	158.0000	150.1708	3.5675	142.8632	157.4785	113.9602	186.3815
27	144.0000	159.8796	4.9090	149.8238	169.9353	123.0159	196.7432
28	130.0000	126.8700	4.6362	117.3731	136.3668	90.1549	163.5851
29	125.0000	122.9865	5.2825	112.1657	133.8072	85.9069	160.0661
30	175.0000	165.7048	5.9299	153.5579	177.8517	128.2167	203.1929

© Cengage Learning

Of course, some bounds (i.e., limits) must be placed on this estimate to take its variability into account. Here, however, we cannot say that we are constructing a confidence interval for Y , since Y is a random variable and not a parameter. The term used to describe

the “hybrid limits” we require is the *prediction interval* (PI), which is given by

$$\bar{Y} + \hat{\beta}_1(X_0 - \bar{X}) \pm t_{n-2, 1-\alpha/2} S_{Y|X} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)S_X^2}} \quad (5.15)$$

We first note that an estimator of an individual’s response should naturally have more variability than an estimator of a group’s mean response. This is reflected by the extra term 1 under the square root sign in (5.15), which is not found in the square root part of the confidence interval formula for $\mu_{Y|X}$ (see (5.12) and (5.13)). To be more specific, in predicting an actual observed Y for a given individual, there are two sources of error operating: individual error as measured by σ^2 and the error in estimating $\mu_{Y|X_0}$ using \hat{Y}_{X_0} . More precisely, this can be expressed by the equation

$$\underbrace{Y - \hat{Y}_{X_0}}_{\text{Error in predicting an individual's } Y \text{ at } X_0} = \underbrace{(Y - \mu_{Y|X_0})}_{\text{Deviation of an individual's } Y \text{ from true mean at } X_0} + \underbrace{(\mu_{Y|X_0} - \hat{Y}_{X_0})}_{\text{Deviation of } \hat{Y}_{X_0} \text{ from true mean at } X_0}$$

This representation allows us to write the variance of an individual’s predicted response at X_0 as

$$\text{Var } Y + \text{Var } \hat{Y}_{X_0} = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)S_X^2} \right]$$

This variance expression is estimated by replacing σ^2 by its estimate $S_{Y|X}^2$. This accounts for the expression on the right-hand side of the prediction interval in (5.15).

Prediction bands, used to describe individual predictions over the entire range of X -values, may be determined in a manner analogous to that by which confidence bands are computed. Figure 5.10 gives 90% prediction bands for the age–systolic blood pressure data. As expected, the 90% prediction bands in this figure are wider than the corresponding 90% confidence bands.

Once again, SAS can be used to compute 95% prediction intervals. In the previous computer output, the lower and upper limits for these intervals are given in the two columns labeled “95% CL Predict.” SAS uses the sample values of the independent variable X_0 .

5.11 Assessing the Appropriateness of the Straight-line Model

In Section 5.2, we noted that the usual strategy for regression with a single independent variable is to assume that the straight-line model is appropriate. This assumption is then rejected if the data indicate that a more complex model is warranted.

Many methods may be used to assess whether the straight-line assumption is reasonable. These will be discussed separately later. The basic techniques include tests for lack of fit and are understood most easily in terms of polynomial regression models (Chapter 15). Many regression diagnostics (Chapter 14) also help in evaluating the straight-line assumption, either explicitly or implicitly. With the linear model, the assumptions of linearity, homoscedasticity, and normality are so intertwined that they often are met or violated as a set.

5.12 Example: BRFSS Analysis

In the examination of 30-day alcohol consumption frequency and BMI among female low-quantity alcohol drinkers in the BRFSS (Example 1.6), the investigators began their regression analysis by examining the straight-line relationship between these two factors.⁶ This analysis is restricted to females only ($n = 1,099$). Edited SAS output, using the 1,056 records with no missing data, and a scatterplot of a subset of the data are provided in Figure 5.11. This output is from PROC MEANS and PROC GLM, where the latter procedure is an alternate regression procedure to PROC REG, and is discussed in later chapters and in Appendix C. The presentation of the output is similar to that from PROC REG shown in Section 5.5.

The initial questions are:

- a. What is the form and interpretation of the calculated straight-line fit for alcohol consumption frequency (“drink days”) as a predictor of BMI? (*Section 5.5*)
- b. Is there a statistically significant linear relationship between alcohol frequency and BMI? (*Sections 5.7 and 5.8*)

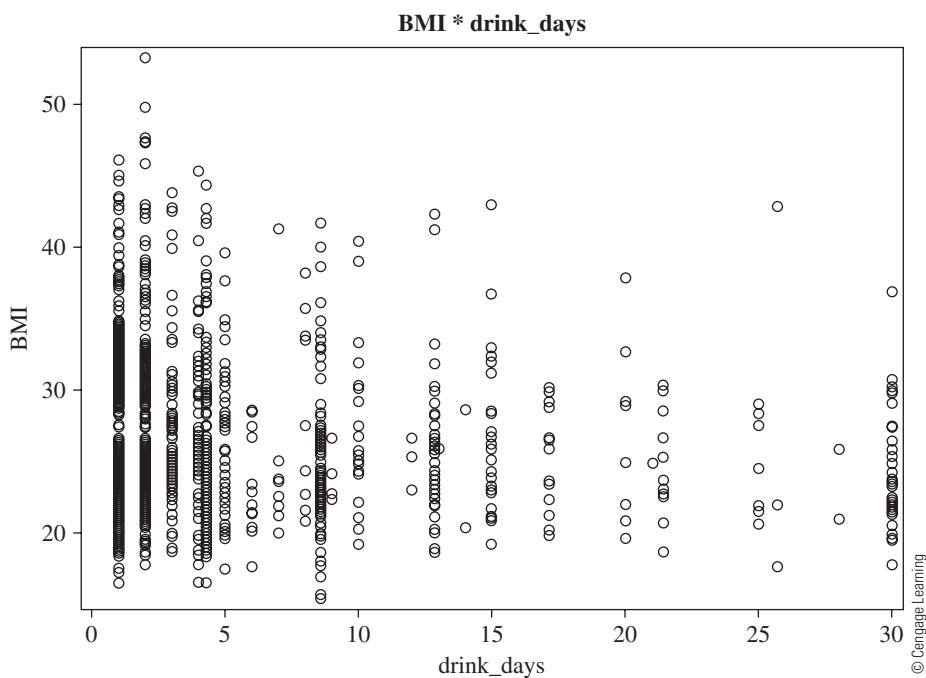
⁶ For all examples in this text, we implicitly assume that all individuals from the population of interest have an equal probability of being included in the study. This is formally known as the simple random sampling (SRS) design, and it is a commonly used design for health research. In order to be theoretically valid, most analyses in this text assume SRS.

In contrast, the BRFSS is implemented as a complex probability survey. This means that a number of techniques are employed that depart from SRS in order to obtain a more representative and/or efficient sample. Specifically, even though random-digit dialing is used, this does not guarantee equal representation of individual people. This is because multiple people might live in a single household that is dialed and because a household with two phone lines is twice as likely to be selected as a household with one line. To adjust for this potential noncomparability among household observations and to better describe the population, weights are used to “scale up” or “scale down” an observation based on what proportion of individuals it represents. Additional corrections to the weights are made for those who do not answer and for households without telephones or with cell phones. Further, the BRFSS design involves *stratified sampling*, meaning that certain, often less populated, regions are oversampled in order to ensure better representation of groups within that region.

To obtain the most valid model estimates, the weighting and stratification schemes should be taken into account. Weighted-analysis analogues exist for many of the regression techniques presented in this book, but their descriptions are beyond this text’s scope. Nevertheless, it is common for standard, unweighted analyses to be conducted as interim or precursor analyses prior to more sophisticated, computationally intensive, and statistically preferred weighted analyses. In practice, global statistical conclusions often do not change substantially between the two analysis types. Compared to those of their unweighted counterparts, results of weighted analyses often result in larger estimated standard errors. Seen another way, unweighted analyses are at risk of having artificially inflated power and so can lead to erroneously rejecting true null hypotheses. Problem 21 discusses the results of the weighted straight-line regression analysis for the example considered in this section, allowing for comparisons to be made.

Edited SAS Output (PROC REG) for BRFSS Example

ANALYSIS VARIABLE : DRINK_DAYS				
N	Mean	Std Dev	Variance	
1056	6.254	7.053	49.74	
DEPENDENT VARIABLE: BMI				
Source	DF	Sum of Squares		
Model	1	1175.78658		
Error	1054	36377.54464		
Corrected Total	1055	37553.33122		
Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	27.77496308	0.24167816	114.93	<.0001
drink_days	-0.14968629	0.02564568	-5.84	<.0001

**FIGURE 5.11** Scatterplot of BMI and days of alcohol drinking in the previous month

- c. What is the estimated mean BMI and corresponding 95% confidence interval for a respondent who drinks the median number of 4 days per month? For one who drinks 15 days per month (90th percentile of frequency)? (*Section 5.9*)

The answers to the above questions are given as follows:

- a. Substituting the parameter estimates from the computer output into the regression equation (5.6), we find

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 (\text{drink days}) = 27.775 - 0.150 (\text{drink days})$$

The Y -intercept estimate $\hat{\beta}_0 = 27.775$ gives the predicted BMI for a female who has had alcohol on 0 days in the previous month. Since the analysis is limited to those who have consumed at least some alcohol in the previous month (i.e., > 0 alcohol days), interpreting this value literally is an extrapolation that may lead to erroneous statements about nondrinkers, particularly if nondrinkers differ by other attributes that may affect BMI. The estimated slope $\hat{\beta}_1 = -0.150$ means that the predicted BMI decreases by 0.150 for each increase of 1 alcohol-consuming day per month. This finding may seem counter-intuitive, but other recently published results have supported this negative association between alcohol intake frequency and BMI (Breslow and Smothers 2005). Subsequent analyses discussed in later chapters of this text will incorporate other factors and will attempt to validly quantify the true BMI-alcohol consumption frequency relationship.

We would like to add a note of caution about the interpretation of the estimated slope and the nature of the causal relationship between drinking and BMI. We would not advise a woman seeking to decrease her BMI by 3 units to drink an additional 20 days per month $[-0.150(20) = -3]$! The cross-sectional design of the study simply allows for making statistical inferences about the BMI levels of women who happen to also be more frequent or less frequent drinkers of alcohol. Frequent drinkers may be very different from infrequent drinkers with respect to nonanalyzed factors that may also be strong determinants of BMI (such as family history, diet, exercise, and smoking habits). This phenomenon is called confounding and is expanded upon in Chapter 11. Without analyses that control for relevant confounders, or better yet a more sophisticated study design (such as a randomized controlled clinical trial), it is difficult to say what would happen to a woman's BMI if she became a more or less frequent drinker.

- b. To test whether there is statistical evidence for a linear relationship between alcohol consumption frequency and BMI, we perform a hypothesis test regarding the slope parameter, testing $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$. To determine the value of the test statistic given by equation (5.9), we need to first determine the numerical value of $S_{\hat{\beta}_1}$, using equation (5.8) and the formula presented above equation (5.9):

$$S_{\hat{\beta}_1} = \frac{S_{Y|X}}{S_X \sqrt{n-1}} = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{36,377.545}{1,054}} = 0.0256$$

Using equation (5.9), we find

$$T = \frac{\hat{\beta}_1 - 0}{S_{\hat{\beta}_1}} = \frac{-0.150}{0.0256} = -5.8$$

To perform a two-tailed test at the $\alpha = .05$ level, we examine whether $|T| \geq t_{1054, 0.975} \approx 1.96$. Indeed, 5.8 exceeds 1.96, so we reject the null hypothesis that there is no linear relationship between alcohol consumption frequency and BMI. The probability of observing a t_{1054} value of at least 5.8 in absolute value, given that H_0 is true, is $< .0001$. Note that $S_{\hat{\beta}_1}$, the value of T , and the P -value are readily obtained from the computer output.

- c. The estimated mean value of BMI at any level of drinking frequency X_0 is simply the value of \hat{Y} found by evaluating the estimated regression equation at X_0 . To compute the 95% confidence interval for the mean BMI, we use equations (5.12) and (5.13).

Thus, we estimate the mean BMI for a female who drinks 4 days per month to be

$$\hat{Y} = 27.775 - 0.150(4) = 27.18. \text{ We next calculate } S_{\hat{Y}_{X_0}}:$$

$$\begin{aligned} S_{\hat{Y}_{X_0}} &= S_{Y|X} \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)S_X^2}} = \sqrt{\frac{36,377.545}{1,054}} \times \sqrt{\frac{1}{1,056} + \frac{(4 - 6.254)^2}{(1,055)(49.74)}} \\ &= 0.1898 \end{aligned}$$

Hence, the 95% confidence interval for the mean BMI for a female who drinks 4 days per month is $\hat{Y}_{X_0} \pm t_{n-2, 1-\frac{\alpha}{2}}(S_{\hat{Y}_{X_0}}) = 27.18 \pm 1.96(0.1898) = (26.80, 27.55)$

To compute the 95% confidence interval for a woman who drinks 15 days/month, we find $S_{\hat{Y}_{X_0}} = 0.2881$. The corresponding 95% confidence interval is then $25.53 \pm 1.96(0.2881) = (24.96, 26.10)$. Note that the width (1.14) of this confidence interval is larger than the width (0.75) of the previously computed confidence interval, the reason being that the frequency of 15 days/month is farther from the sample mean of 6.254 days/month than is 4 days/month.

If we had instead desired 95% prediction intervals for BMI, we would have adjusted our calculation of $S_{\hat{Y}_{X_0}}$ by incorporating the addition of the number 1 in the second square-root term, as shown in Section 5.10. For those women who have 4 drinking days/month, this leads to an estimated standard error of

$$\sqrt{\frac{36,377.545}{1,054}} \times \sqrt{1 + \frac{1}{1,056} + \frac{(4 - 6.254)^2}{(1,055)(49.74)}} = 5.8779$$

The corresponding prediction interval is then (15.64, 38.71). Note that this prediction interval is *much* wider than the confidence interval. This reflects the substantial (and not surprising) between-individual variability in BMI in the population.

Methods for the computer generation of these interval estimates are additionally discussed in Appendix C.

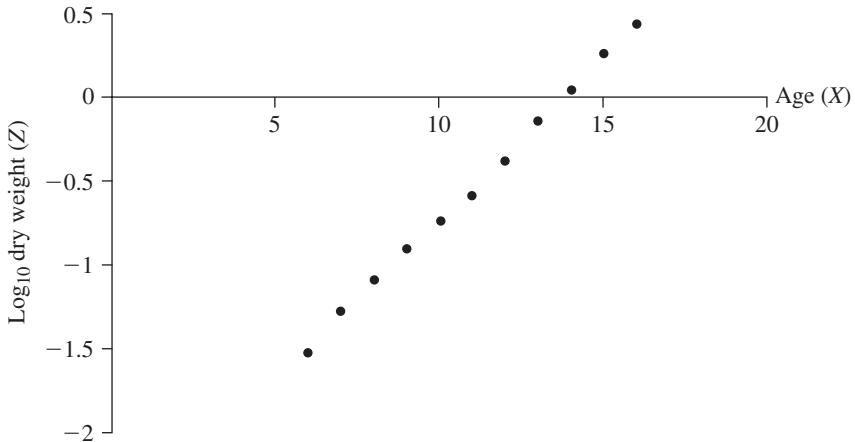
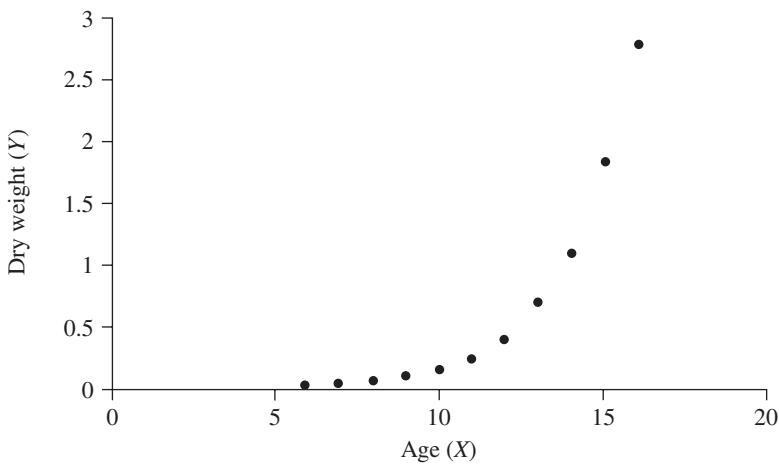
Problems

SAS computer output is provided for most of the problems in this chapter. The same is true for many of the subsequent chapters. In some problems, using the output can significantly reduce the computational and programming effort required to answer the questions. Since the actual data are also often provided, however, instructors and students may choose to do their own computations and/or programming to perform the necessary analyses.

1. The accompanying table gives the dry weights (Y) of 11 chick embryos ranging in age from 6 to 16 days (X). Also given in the table are the values of the common logarithms of the weights (Z).

Age (X) (days)	6	7	8	9	10	11
Dry Weight (Y)	0.029	0.052	0.079	0.125	0.181	0.261
\log_{10} Dry Weight (Z)	-1.538	-1.284	-1.102	-0.903	-0.742	-0.583
Age (X) (days)	12	13	14	15	16	
Dry Weight (Y)	0.425	0.738	1.130	1.882	2.812	
\log_{10} Dry Weight (Z)	-0.372	-0.132	0.053	0.275	0.449	

- a. Observe the following two scatter diagrams. Describe the relationships between age (X) and dry weight (Y) and between age and \log_{10} dry weight (Z).



© Cengage Learning

- b. State the simple linear regression models for these two regressions: Y regressed on X and Z regressed on X .
 c. Determine the least-squares estimates of each of the regression lines in part (b).

- d. Sketch each estimated line on the appropriate scatter diagram. Which of the two regression lines has the better fit? Based on your answers to parts (a)–(c), is it more appropriate to run a linear regression of Y on X or of Z on X ? Explain.
- e. For the regression that you chose as being more appropriate in part (d), find 95% confidence intervals for the true slope and intercept. Interpret each interval with regard to the null hypothesis that the true parameter is 0.
- f. For the regression that you chose as being more appropriate in part (d), find and sketch 95% confidence and prediction bands. Using your sketch, find and interpret an approximate 95% confidence interval for the mean response of an 8-day-old chick.

Edited SAS Output (PROC REG) for Problem 1

Regression of Dry Weight (Y) on Age (X)

[Portion of output omitted]

Root MSE	0.48185	R-Square	0.7442
----------	---------	----------	--------

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-1.88453	0.52584	-3.58	0.0059
X	1	0.23507	0.04594	5.12	0.0006

OUTPUT STATISTICS

Obs	Dependent Variable	Predicted Value	Std Error	95% CL Mean		95% CL Predict	
			Mean Predict				
1	0.0290	-0.4741	0.2718	-1.0889	0.1408	-1.7256	0.7774
2	0.0520	-0.2390	0.2343	-0.7690	0.2909	-1.4510	0.9730
3	0.0790	-0.003945	0.2003	-0.4570	0.4491	-1.1844	1.1765
4	0.1250	0.2311	0.1719	-0.1577	0.6200	-0.9262	1.3884
5	0.1810	0.4662	0.1524	0.1215	0.8109	-0.6770	1.6094
6	0.2610	0.7013	0.1453	0.3726	1.0299	-0.4372	1.8398
7	0.4250	0.9363	0.1524	0.5917	1.2810	-0.2069	2.0796
8	0.7380	1.1714	0.1719	0.7826	1.5603	0.0141	2.3287
9	1.1300	1.4065	0.2003	0.9535	1.8595	0.2261	2.5869
10	1.8820	1.6416	0.2343	1.1116	2.1715	0.4296	2.8536
11	2.8120	1.8766	0.2718	1.2618	2.4915	0.6252	3.1281

(continued)

Regression of Log10 Dry Weight (Z) on Age (X)

Root MSE		0.02807		R-Square		0.9983	
PARAMETER ESTIMATES							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t		
Intercept	1	-2.68925	0.03064	-87.78	<.0001		
X	1	0.19589	0.00268	73.18	<.0001		

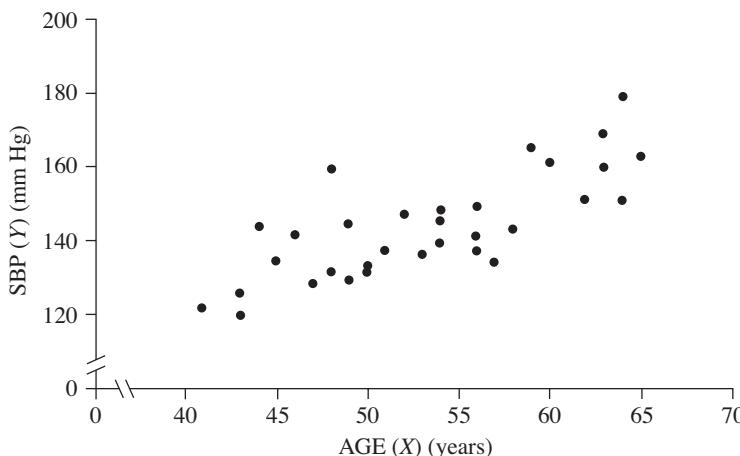
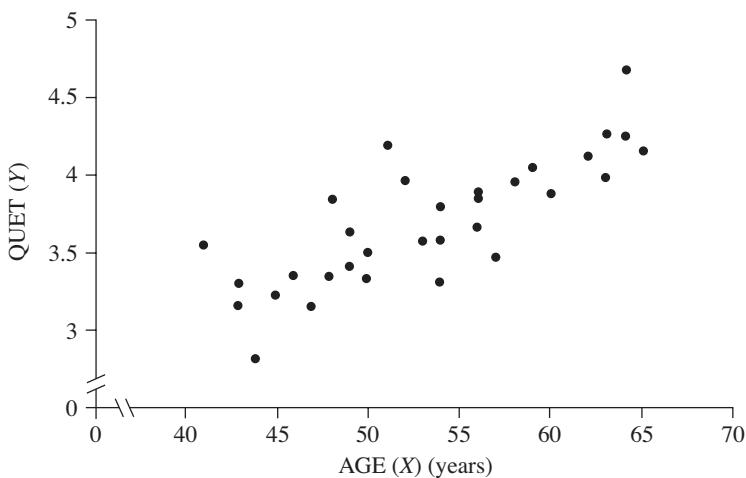
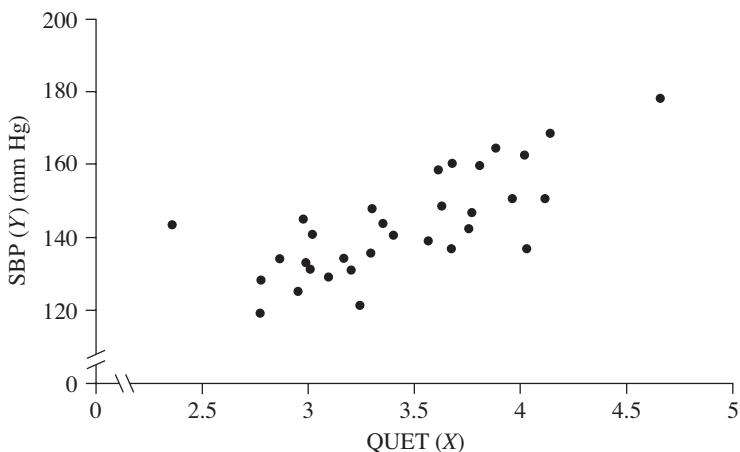
OUTPUT STATISTICS							
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict	
1	-1.5380	-1.5139	0.0158	-1.5497	-1.4781	-1.5868	-1.4410
2	-1.2840	-1.3180	0.0136	-1.3489	-1.2871	-1.3886	-1.2474
3	-1.1020	-1.1221	0.0117	-1.1485	-1.0957	-1.1909	-1.0534
4	-0.9030	-0.9262	0.0100	-0.9489	-0.9036	-0.9937	-0.8588
5	-0.7420	-0.7303	0.008878	-0.7504	-0.7103	-0.7970	-0.6637
6	-0.5830	-0.5345	0.008465	-0.5536	-0.5153	-0.6008	-0.4681
7	-0.3720	-0.3386	0.008878	-0.3586	-0.3185	-0.4052	-0.2720
8	-0.1320	-0.1427	0.0100	-0.1653	-0.1200	-0.2101	-0.0752
9	0.0530	0.0532	0.0117	0.0268	0.0796	-0.0156	0.1220
10	0.2750	0.2491	0.0136	0.2182	0.2800	0.1785	0.3197
11	0.4490	0.4450	0.0158	0.4092	0.4808	0.3721	0.5179

2. The following table gives the systolic blood pressure (SBP), body size (QUET),⁷ age (AGE), and smoking history (SMK = 0 if a nonsmoker, SMK = 1 if a current or previous smoker) for a hypothetical sample of 32 white males over 40 years old from the town of Angina.

Person	SBP	QUET	AGE	SMK	Person	SBP	QUET	AGE	SMK	Person	SBP	QUET	AGE	SMK
1	135	2.876	45	0	12	138	4.032	51	1	23	137	3.296	53	0
2	122	3.251	41	0	13	152	4.116	64	0	24	132	3.210	50	0
3	130	3.100	49	0	14	138	3.673	56	0	25	149	3.301	54	1
4	148	3.768	52	0	15	140	3.562	54	1	26	132	3.017	48	1
5	146	2.979	54	1	16	134	2.998	50	1	27	120	2.789	43	0
6	129	2.790	47	1	17	145	3.360	49	1	28	126	2.956	43	1
7	162	3.668	60	1	18	142	3.024	46	1	29	161	3.800	63	0
8	160	3.612	48	1	19	135	3.171	57	0	30	170	4.132	63	1
9	144	2.368	44	1	20	142	3.401	56	0	31	152	3.962	62	0
10	180	4.637	64	1	21	150	3.628	56	1	32	164	4.010	65	0
11	166	3.877	59	1	22	144	3.751	58	0					

⁷ QUET stands for "quetelet index," a measure of size defined by QUET = 100 (weight/height²).

- a. On each of the accompanying scatter diagrams, sketch by eye a line that fits the data reasonably well. Comment on the relationships described.



- b.**
- (1) Determine the least-squares estimates of the slope (β_1) and intercept (β_0) for the straight-line regression of SBP (Y) on QUET (X).
 - (2) Sketch the estimated regression line on the scatter diagram involving SBP and QUET. Compare this new line with the line you drew in part (a).
 - (3) Test the null hypothesis of zero slope; be sure to interpret the result.
 - (4) Based on your test in part (b)(3), would you conclude that blood pressure increases as body size increases?
 - (5) Find and sketch 95% prediction bands on the appropriate scatter diagram.
 - (6) Using your answer to part (b)(5), find an approximate 95% prediction interval for an individual with QUET = 3.4 (the sample mean value of QUET). Interpret your answer.
 - (7) Are any assumptions for straight-line regression clearly not satisfied in this example?
- c.** Repeat questions (1) through (4) in part (b) for the regression of QUET on AGE.
- d.** Repeat questions (1) through (4) in part (b) for the regression of SBP on AGE.
- e.**
- (1) Determine the least-squares estimates of the slope and intercept for the straight-line regression of SBP (Y) on SMK (X).
 - (2) Compare the value of $\hat{\beta}_0$ with the mean SBP for nonsmokers. Compare the values of $\hat{\beta}_0 + \hat{\beta}_1$ with the mean SBP for smokers. Explain the results of these comparisons.
 - (3) Test the null hypothesis that the true slope is 0; be sure to interpret the result.
 - (4) Is the test in part (e)(3) equivalent to the usual two-sample t test for the equality of two population means, assuming equal but unknown variances? Demonstrate your answer numerically.

Edited SAS Output (PROC REG) for Problem 2

Regression of SBP (Y) on QUET (Z)

[Portion of output omitted]

Root MSE	9.81160	R-Square	0.5506
----------	---------	----------	--------

PARAMETER ESTIMATES

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	70.57640	12.32187	5.73	<.0001
QUET	1	21.49167	3.54515	6.06	<.0001

OUTPUT STATISTICS

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
1	135.0000	132.3864	2.6499	111.6306	153.1423	2.6136
2	122.0000	140.4458	1.8608	120.0507	160.8409	-18.4458
3	130.0000	137.2006	2.1144	116.7026	157.6985	-7.2006

(continued)

OUTPUT STATISTICS						
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
4	148.0000	151.5570	2.0860	131.0712	172.0428	-3.5570
5	146.0000	134.6001	2.3858	113.9782	155.2219	11.3999
6	129.0000	130.5382	2.8873	109.6506	151.4257	-1.5382
7	162.0000	149.4078	1.9119	128.9930	169.8227	12.5922
8	160.0000	148.2043	1.8372	127.8181	168.5905	11.7957
9	144.0000	121.4687	4.1810	99.6873	143.2501	22.5313
10	180.0000	170.2333	4.5807	148.1191	192.3475	9.7667
11	166.0000	153.8996	2.3230	133.3077	174.4915	12.1004
12	138.0000	157.2308	2.7197	136.4373	178.0243	-19.2308
13	152.0000	159.0361	2.9552	138.1090	179.9632	-7.0361
14	138.0000	149.5153	1.9194	129.0975	169.9331	-11.5153
15	140.0000	147.1297	1.7866	126.7623	167.4972	-7.1297
16	134.0000	135.0084	2.3401	114.4085	155.6084	-1.0084
17	145.0000	142.7884	1.7581	122.4313	163.1455	2.2116
18	142.0000	135.5672	2.2792	114.9957	156.1387	6.4328
19	135.0000	138.7265	1.9812	118.2841	159.1689	-3.7265
20	142.0000	143.6696	1.7403	123.3189	164.0203	-1.6696
21	150.0000	148.5482	1.8567	128.1546	168.9418	1.4518
22	144.0000	151.1917	2.0531	130.7197	171.6636	-7.1917
23	137.0000	141.4129	1.8091	121.0372	161.7887	-4.4129
24	132.0000	139.5647	1.9182	119.1473	159.9820	-7.5647
25	149.0000	141.5204	1.8042	121.1465	161.8943	7.4796
26	132.0000	135.4168	2.2954	114.8378	155.9958	-3.4168
27	120.0000	130.5167	2.8901	109.6275	151.4058	-10.5167
28	126.0000	134.1058	2.4425	113.4563	154.7553	-8.1058
29	161.0000	152.2447	2.1511	131.7309	172.7586	8.7553
30	170.0000	159.3800	3.0013	138.4255	180.3344	10.6200
31	152.0000	155.7264	2.5335	135.0312	176.4216	-3.7264
32	164.0000	156.7580	2.6601	135.9967	177.5193	7.2420

Regression of QUET (Y) on AGE (X)

[Portion of output omitted]

Root MSE	0.30131	R-Square	0.6444
----------	---------	----------	--------

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.38645	0.41769	0.93	0.3622
AGE	1	0.05736	0.00778	7.37	<.0001

(continued)

Regression of SBP (Y) on AGE (X)

[Portion of output omitted]

Root MSE	9.24543	R-Square	0.6009
----------	---------	----------	--------

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	59.09163	12.81626	4.61	<.0001
AGE	1	1.60450	0.23872	6.72	<.0001

Regression of SBP (Y) on SMK (X)

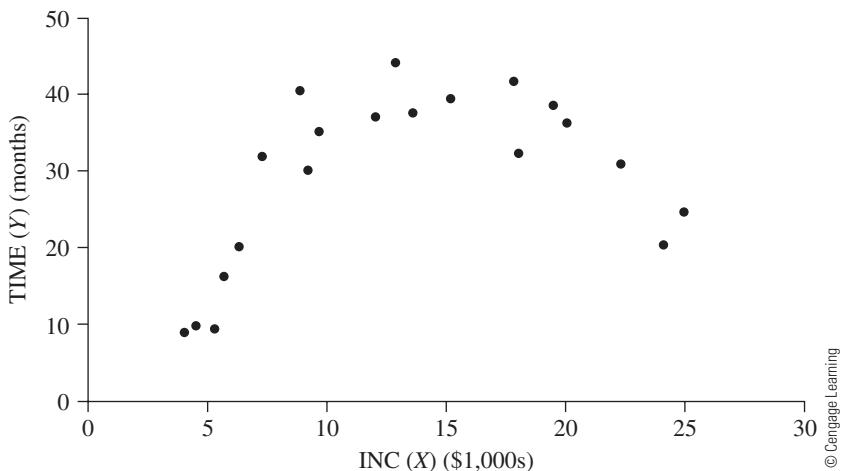
[Portion of output omitted]

Root MSE	14.18082	R-Square	0.0612
----------	----------	----------	--------

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	140.80000	3.66147	38.45	<.0001
SMK	1	7.02353	5.02350	1.40	0.1723

3. For married couples with one or more offspring, a demographic study was conducted to determine the effect of the husband's annual income (at marriage) on the time (in months) between marriage and the birth of the first child. The following table gives the husband's annual income (INC) and the time between marriage and the birth of the first child (TIME) for a hypothetical sample of 20 couples.

INC	TIME	INC	TIME
5,775	16.20	4,608	9.70
9,800	35.00	24,210	20.00
13,795	37.20	19,625	38.20
4,120	9.00	18,000	41.25
25,015	24.40	13,000	44.00
12,200	36.75	5,400	9.20
7,400	31.75	6,440	20.00
9,340	30.00	9,000	40.20
20,170	36.00	18,180	32.00
22,400	30.80	15,385	39.20



© Cengage Learning

- On the scatter diagram above, sketch by eye a line that fits the data reasonably well. Comment on the relationship between TIME (Y) and INC (X).
- Determine the least-squares estimates of the slope (β_1) and intercept (β_0) for the straight-line regression of TIME (Y) on INC (X).
- Draw the estimated regression line on the accompanying scatter diagram. Comment on how well this line fits the data.
- Are any of the assumptions for straight-line regression clearly not satisfied in this example?
- Test the null hypothesis that the true slope is 0. Interpret the results of this test.
- Can you suggest a model that would describe the TIME–INC relationship better than a straight line does?

Edited SAS Output (PROC REG) for Problem 3

Regression of TIME (Y) on INC (X)

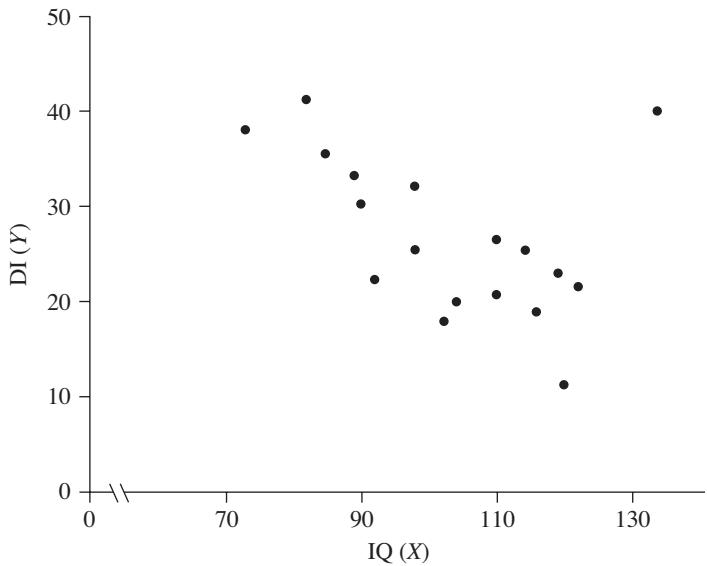
[Portion of output omitted]

Root MSE		10.49580	R-Square		0.1853
PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	19.62575	5.21291	3.76	0.0014
X	1	0.00071376	0.00035281	2.02	0.0582

- A sociologist assigned to a correctional institution was interested in studying the relationship between intelligence and delinquency. A delinquency index (ranging from

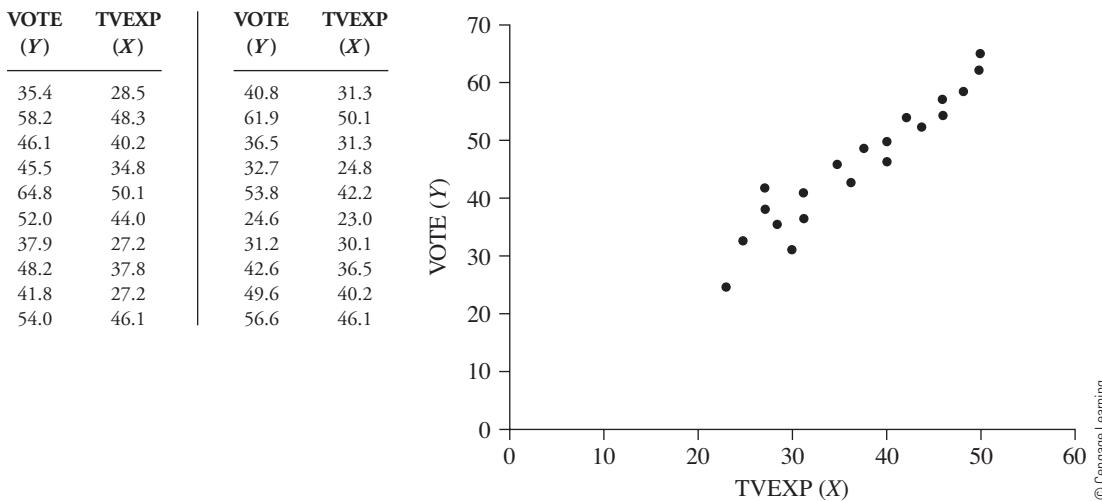
0 to 50) was formulated to account for both the severity and the frequency of crimes committed, while intelligence was measured by IQ. The following table displays the delinquency index (DI) and IQ of a sample of 18 convicted minors.

DI (Y)	IQ (X)	DI (Y)	IQ (X)
26.20	110	22.10	92
33.00	89	18.60	116
17.50	102	35.50	85
25.25	98	38.00	73
20.30	110	30.00	90
31.90	98	19.70	104
21.10	122	41.10	82
22.70	119	39.60	134
10.70	120	25.15	114



- Given that $\hat{\beta}_1 = -0.249$ and $\hat{\beta}_0 = 52.273$, draw the estimated regression line on the accompanying scatter diagram.
- How do you account for the fact that $\hat{Y} = 52.273$ when IQ = 0, even though the delinquency index goes no higher than 50?
- Find a 95% confidence interval for the true slope β_1 using the fact that $S_{Y|X} = 7.704$ and $S_X = 16.192$.
- Interpret this confidence interval with regard to testing the null hypothesis of zero slope at the $\alpha = .05$ level.
- Notice that the convicted minor with IQ = 134 and DI = 39.6 appears to be quite out of place in the data. Decide whether this outlier has any effect on your estimate of the IQ–DI relationship by looking at the graph for the fitted line obtained when the outlier is omitted. (Note that $\hat{\beta}_0 = 70.846$ and $\hat{\beta}_1 = -0.444$ when the outlier is removed.)
- Test the null hypothesis of zero slope when the outlier is removed, given that $S_{Y|X} = 4.933$, $S_X = 14.693$, and $n = 17$. (Use $\alpha = .05$.)
- For these data, would you conclude that the delinquency index decreases as IQ increases?
- Following a recent congressional election, a political scientist attempted to investigate the relationship between campaign expenditures on television advertisements and subsequent voter turnout. The following table presents the percentage of total campaign expenditures delegated to television advertisements (TVEXP) and the percentage of registered voter turnout (VOTE) for a hypothetical sample of 20 congressional districts.

- Determine the least-squares line of VOTE on TVEXP, and draw the estimated line on the accompanying scatter diagram.
- Are any of the assumptions for straight-line regression clearly *not* satisfied in this example?
- Test the null hypothesis that the true slope is 0; be sure to interpret your result.
- Test the hypothesis $\mu_{Y|X_0} = 45$ for $X_0 = \bar{X} = 36.99$. Interpret your result.
- Calculate the corresponding 95% confidence interval for $\mu_{Y|36.99}$, and interpret your result.



© Cengage Learning

Edited SAS Output (PROC REG) for Problem 5

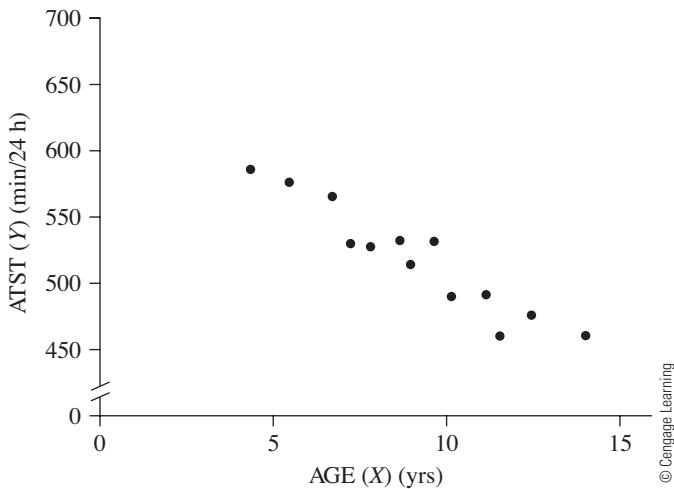
Regression of VOTE (Y) on TVEXP (X)

[Portion of output omitted]

Root MSE		3.33177	R-Square	0.9101	
PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	2.17407	3.30974	0.66	0.5196
X	1	1.17696	0.08718	13.50	<.0001

6. A group of 13 children and adolescents (considered healthy) participated in a psychological study designed to analyze the relationship between age and average total sleep time (ATST). To obtain a measure for ATST (in minutes), recordings were taken on each subject on three consecutive nights and then averaged. The results obtained are displayed in the following table.

ATST (min/24 h)	AGE	ATST (min/24 h)	AGE
586.00	4.40	515.20	8.90
461.75	14.00	493.00	11.10
491.10	10.10	528.30	7.75
565.00	6.70	575.90	5.50
462.00	11.50	532.50	8.60
532.10	9.60	530.50	7.20
477.60	12.40		



© Cengage Learning

- Determine the least-squares estimates of the slope and intercept for the straight-line regression of ATST (Y) on AGE (X). Draw the estimated line on the accompanying scatter diagram, and comment on the fit.
- Are any of the assumptions for straight-line regression clearly *not* satisfied in this example?
- Test the null hypothesis that the true slope is 0; be sure to interpret your result.
- Obtain a 95% confidence interval for β_1 . Interpret your result.
- Would you reject the null hypothesis $H_0: \beta_1 = 0$ based on the confidence interval you calculated in part (d)? Explain.
- Determine and sketch 95% confidence bands on the accompanying scatter diagram. Use your diagram to estimate the mean ATST when AGE = 10. Interpret your result.

Edited SAS Output (PROC REG) for Problem 6

Regression of ATST (Y) on AGE (X)

[Portion of output omitted]

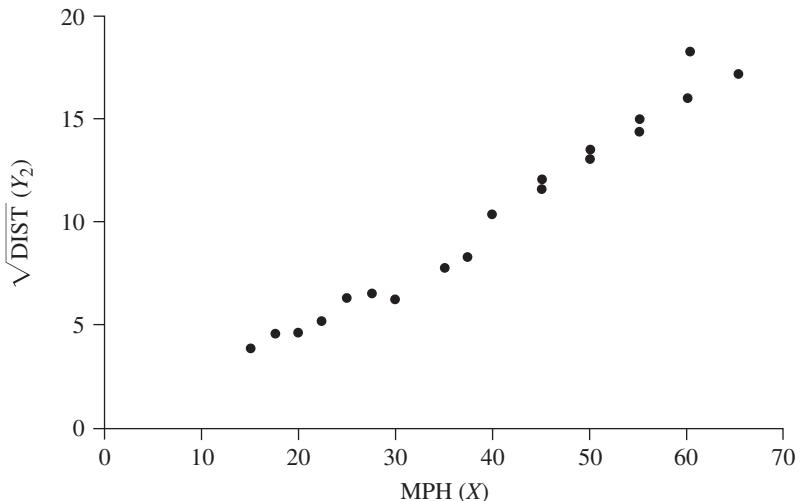
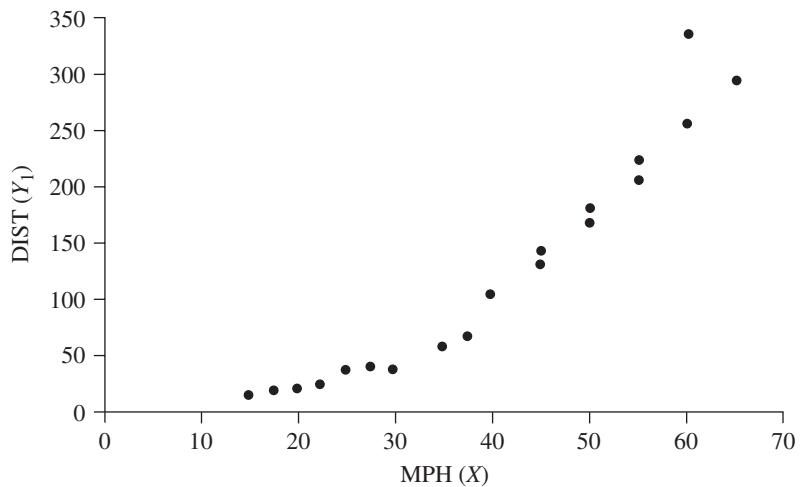
Root MSE		13.15238		R-Square	0.9054
PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	646.48334	12.91773	50.05	<.0001
X	1	-14.04105	1.36812	-10.26	<.0001

OUTPUT STATISTICS						
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		Residual
1	586.0000	584.7027	7.3425	568.5420	600.8635	1.2973
2	461.7500	449.9087	7.6829	432.9988	466.8185	11.8413
3	491.1000	504.6688	3.9166	496.0483	513.2892	-13.5688
4	565.0000	552.4083	4.8694	541.6908	563.1258	12.5917
5	462.0000	485.0113	4.9468	474.1234	495.8992	-23.0113
6	532.1000	511.6893	3.7225	503.4961	519.8824	20.4107
7	477.6000	472.3743	5.8494	459.4998	485.2488	5.2257
8	515.2000	521.5180	3.6542	513.4752	529.5608	-6.3180
9	493.0000	490.6277	4.5950	480.5143	500.7411	2.3723
10	528.3000	537.6652	4.0629	528.7228	546.6076	-9.3652
11	575.9000	569.2576	6.0826	555.8700	582.6452	6.6424
12	532.5000	525.7303	3.7012	517.5841	533.8765	6.7697
13	530.5000	545.3878	4.4459	535.6024	555.1731	-14.8878

7. Several research workers associated with the Office of Highway Safety were evaluating the relationship between driving speed (MPH) and the distance a vehicle travels once brakes are applied (DIST). The results of 19 experimental tests are displayed in the following table.
- Determine the least-squares estimates of the slope and intercept for each of the following straight-line regressions: Y_1 (DIST) on X and $Y_2(\sqrt{DIST})$ on X. Draw the estimated lines on the appropriate scatter diagrams.
 - Which of the two variable pairs mentioned in part (a) seems to be better suited for straight-line regression?
 - For the variable pair Y_2 and X, test the hypothesis that the true slope is equal to 1 ($\alpha = .01$). Be sure to interpret your result.
 - Construct a 99% confidence interval for the true slope in part (c). Interpret your result.

- e. For the same variable pair considered in parts (c) and (d), calculate and sketch 95% confidence bands on the appropriate scatter diagram. Using the confidence bands, estimate the mean value of Y_2 when $X = 45$. Interpret your result.

MPH (X)	DIST (Y_1)	$\sqrt{\text{DIST}} (Y_2)$	MPH (X)	DIST (Y_1)	$\sqrt{\text{DIST}} (Y_2)$
25.0	37.4	6.12	50.0	170.0	13.04
35.0	57.7	7.60	20.0	20.0	4.47
60.0	337.6	18.37	15.0	13.5	3.67
45.0	142.5	11.94	27.5	40.8	6.39
50.0	182.4	13.51	55.0	207.8	14.42
37.5	67.5	8.22	40.0	105.0	10.25
30.0	37.5	6.12	45.0	132.6	11.52
55.0	225.0	15.00	17.5	19.1	4.37
60.0	258.1	16.07	22.5	25.0	5.00
65.0	297.4	17.25			



Edited SAS Output (PROC REG) for Problem 7

Regression of DIST (Y1) on MPH (X)

[Portion of output omitted]

Root MSE		31.64950	R-Square	0.9106	
PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-122.34459	20.15624	-6.07	<.0001
X	1	6.22708	0.47319	13.16	<.0001

Regression of Sqrt[DIST] (Y2) on MPH (X)

[Portion of output omitted]

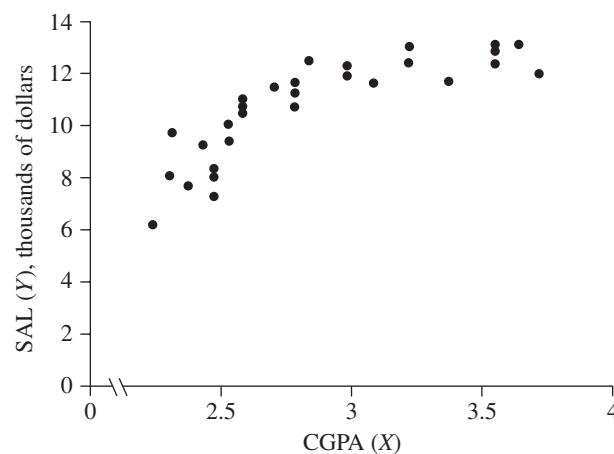
Root MSE		0.81097	R-Square	0.9728	
PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-1.69712	0.51647	-3.29	0.0044
X	1	0.29878	0.01212	24.64	<.0001

OUTPUT STATISTICS					
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean	
1	6.1200	5.7723	0.2580	5.2280	6.3165
2	7.6000	8.7600	0.1947	8.3492	9.1708
3	18.3700	16.2294	0.3082	15.5792	16.8796
4	11.9400	11.7478	0.1967	11.3328	12.1627
5	13.5100	13.2416	0.2238	12.7694	13.7139
6	8.2200	9.5070	0.1880	9.1103	9.9036
7	6.1200	7.2661	0.2203	6.8013	7.7310
8	15.0000	14.7355	0.2624	14.1819	15.2892
9	16.0700	16.2294	0.3082	15.5792	16.8796
10	17.2500	17.7233	0.3584	16.9671	18.4794
11	13.0400	13.2416	0.2238	12.7694	13.7139
12	4.4700	4.2784	0.3031	3.6389	4.9179
13	3.6700	2.7845	0.3529	2.0399	3.5292
14	6.3900	6.5192	0.2380	6.0171	7.0213
15	14.4200	14.7355	0.2624	14.1819	15.2892
16	10.2500	10.2539	0.1861	9.8613	10.6465
17	11.5200	11.7478	0.1967	11.3328	12.1627
18	4.3700	3.5314	0.3276	2.8403	4.2226
19	5.0000	5.0253	0.2798	4.4350	5.6157

8. The following table presents the starting annual salaries (SAL) of a group of 30 college graduates who have recently entered the job market, along with their cumulative grade-point averages (CGPA).
- Determine the least-squares estimates of the slope and intercept for the straight-line regression of SAL (Y) on CGPA (X). Draw the estimated line on the accompanying scatter diagram, and comment on the fit.
 - Are any of the assumptions for straight-line regression clearly *not* satisfied in this example?

SAL (Y)	CGPA (X)
10455	2.58
9680	2.31
7300	2.47
9388	2.52
12496	3.22
11812	3.37
9224	2.43
11725	3.08
11320	2.78
12000	2.98
12500	3.55
13310	3.64
12105	3.72
6200	2.24
11522	2.70

SAL (Y)	CGPA (X)
8000	2.30
12548	2.83
7700	2.37
10028	2.52
13176	3.22
13255	3.55
13004	3.55
8000	2.47
8224	2.47
10750	2.78
11669	2.78
12322	2.98
11002	2.58
10666	2.58
10839	2.58



- Obtain a 95% confidence interval for β_1 .
- Would you reject the null hypothesis $H_0: \beta_1 = 4,000$ at the $\alpha = .05$ level?
- Find and graph 95% confidence and prediction bands.
- Would you reject the hypothesis $H_0: \mu_{Y|X} = 11,500$ at $X_0 = 2.75$?

Edited SAS Output (PROC REG) for Problem 8

Regression of SAL (Y) on CGPA (X)

[Portion of output omitted]

Root MSE	1124.71499	R-Square	0.6845
----------	------------	----------	--------

PARAMETER ESTIMATES

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	435.92357	1337.85967	0.33	0.7470
X	1	3630.56128	465.76874	7.79	<.0001

(continued)

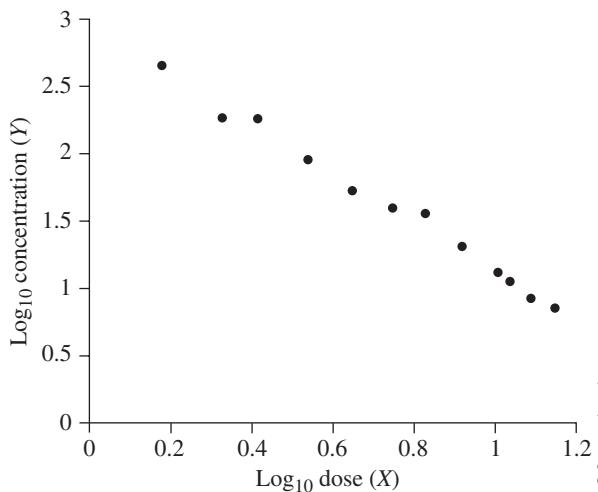
OUTPUT STATISTICS							
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict	
1	10455	9803	237.9998	9315	10290	7448	12158
2	9680	8823	320.5028	8166	9479	6427	11218
3	7300	9403	267.5786	8855	9952	7035	11772
4	9388	9585	253.2786	9066	10104	7223	11947
5	12496	12126	271.6022	11570	12683	9756	14496
6	11812	12671	321.6964	12012	13330	10275	15067
7	9224	9258	279.8892	8685	9832	6884	11632
8	11725	11618	234.1710	11138	12098	9265	13971
9	11320	10529	207.1336	10105	10953	8186	12872
10	12000	11255	215.6850	10813	11697	8909	13601
11	12500	13324	389.9229	12526	14123	10886	15763
12	13310	13651	426.1305	12778	14524	11187	16115
13	12105	13942	459.1316	13001	14882	11453	16430
14	6200	8568	346.1668	7859	9277	6158	10979
15	11522	10238	215.2151	9798	10679	7893	12584
16	8000	8786	324.0927	8122	9450	6389	11184
17	12548	10710	205.3806	10290	11131	8368	13052
18	7700	9040	299.5814	8427	9654	6656	11425
19	10028	9585	253.2786	9066	10104	7223	11947
20	13176	12126	271.6022	11570	12683	9756	14496
21	13255	13324	389.9229	12526	14123	10886	15763
22	13004	13324	389.9229	12526	14123	10886	15763
23	8000	9403	267.5786	8855	9952	7035	11772
24	8224	9403	267.5786	8855	9952	7035	11772
25	10750	10529	207.1336	10105	10953	8186	12872
26	11669	10529	207.1336	10105	10953	8186	12872
27	12322	11255	215.6850	10813	11697	8909	13601
28	11002	9803	237.9998	9315	10290	7448	12158
29	10666	9803	237.9998	9315	10290	7448	12158
30	10839	9803	237.9998	9315	10290	7448	12158

9. In an experiment designed to describe the dose-response curve for vitamin K, individual rats were depleted of their vitamin K reserves and then fed dried liver for 4 days at different dosage levels.⁸ The response of each rat was measured as the concentration of a clotting agent needed to clot a sample of its blood in 3 minutes. The results of the experiment on 12 rats are given in the following table; values are expressed in common logarithms for both dose and response.

⁸ Adapted from a study by Schonheyder (1936).

- a. Determine the least-squares estimates of the slope (β_1) and the intercept (β_0) for the straight-line regression of Log_{10} concentration (Y) on Log_{10} dose (X).

Rat	Log_{10} Concentration (Y)	Log_{10} Dose (X)
1	2.65	0.18
2	2.25	0.33
3	2.26	0.42
4	1.95	0.54
5	1.72	0.65
6	1.60	0.75
7	1.55	0.83
8	1.32	0.92
9	1.13	1.01
10	1.07	1.04
11	0.95	1.09
12	0.88	1.15



© Cengage Learning

- b. Draw the estimated regression line on the accompanying scatter diagram. How well does this line fit the data?
 c. Determine and sketch 95% confidence bands based on the estimated regression line.
 d. Convert the fitted straight line into an equation in the original units $Y' = 10^Y$ and $X' = 10^X$.
 e. For the converted equation obtained in part (d), determine 99% confidence intervals for the true mean responses at the maximum and minimum doses used in the experiment.
 f. If the values for X and Y on each rat are converted to their original units X' and Y' , the following fitted straight line is obtained: $Y' = 237.16095 - 21.32117X'$. How would you evaluate whether using the variables X' and Y' is better or worse than using X and Y for the regression analysis?

Edited SAS Output (PROC REG) for Problem 9

Regression of Log10[Conc] (Y) on Log10[Dose] (X)

[Portion of output omitted]

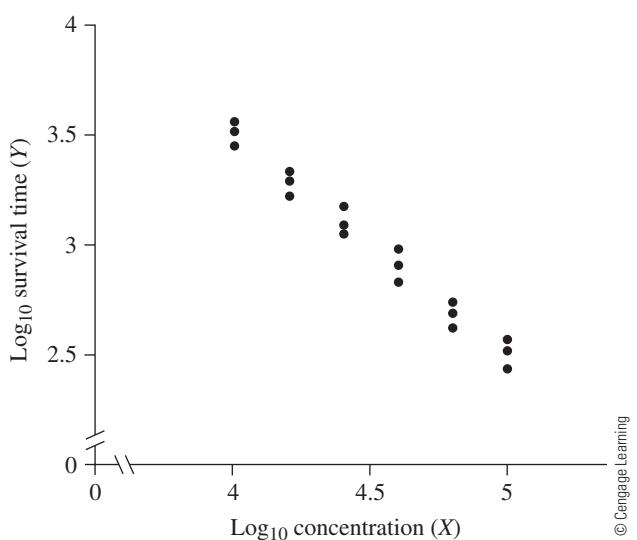
Root MSE	0.05589	R-Square	0.9914		
PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	2.93620	0.04230	69.41	<.0001
X	1	-1.78501	0.05267	-33.89	<.0001

(continued)

OUTPUT STATISTICS						
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		Residual
1	2.6500	2.6149	0.0337	2.5397	2.6901	0.0351
2	2.2500	2.3472	0.0271	2.2869	2.4074	-0.0972
3	2.2600	2.1865	0.0234	2.1343	2.2387	0.0735
4	1.9500	1.9723	0.0193	1.9292	2.0154	-0.0223
5	1.7200	1.7759	0.0169	1.7384	1.8135	-0.0559
6	1.6000	1.5974	0.0161	1.5615	1.6334	0.002554
7	1.5500	1.4546	0.0168	1.4173	1.4920	0.0954
8	1.3200	1.2940	0.0186	1.2524	1.3355	0.0260
9	1.1300	1.1333	0.0214	1.0856	1.1811	-0.003343
10	1.0700	1.0798	0.0225	1.0297	1.1299	-0.009792
11	0.9500	0.9905	0.0244	0.9362	1.0449	-0.0405
12	0.8800	0.8834	0.0268	0.8236	0.9433	-0.003441

10. The susceptibility of catfish to a certain chemical pollutant was determined by immersing individual fish in 2 liters of an emulsion containing the pollutant and measuring the survival time in minutes.⁹ The data in the following table give the common log of survival time (Y) and the common log of concentration (X) of the pollutant in parts per million for 18 fish.
- Determine and draw the estimated straight line of Y regressed on X on the accompanying scatter diagram. Comment on the fit.
 - Test for the significance of the straight-line regression. Interpret your result.
 - Determine 95% confidence intervals for the true mean survival time $\mu_{Y|X}$ (where $Y = 10^Y$) at values of $X = 5$, 4.5, and 4. (Note $\bar{X} = 4.5$.) Interpret these intervals.

Fish	Log ₁₀ Survival Time (Y)	Log ₁₀ Concentration (X)
1	2.516	5.0
2	2.572	5.0
3	2.438	5.0
4	2.621	4.8
5	2.742	4.8
6	2.689	4.8
7	2.830	4.6
8	2.910	4.6
9	2.983	4.6
10	3.175	4.4
11	3.056	4.4
12	3.095	4.4
13	3.332	4.2
14	3.221	4.2
15	3.293	4.2
16	3.447	4.0
17	3.523	4.0
18	3.551	4.0



© Cengage Learning

⁹ Adapted from a study by Nagasawa, Asano, and Kondo (1964).

Edited SAS Output (PROC REG) for Problem 10

Regression of Log₁₀[Surv Time] (Y) on Log₁₀[Conc] (X)

[Portion of output omitted]

Root MSE		0.05597	R-Square		0.9766
PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	7.49110	0.17431	42.97	<.0001
X	1	-0.99810	0.03863	-25.84	<.0001

OUTPUT STATISTICS						
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean	Residual	
1	2.5160	2.5006	0.0234	2.4510	2.5502	0.0154
		...				
9	2.9830	2.8999	0.0137	2.8707	2.9290	0.0831
10	3.1750	3.0995	0.0137	3.0703	3.1286	0.0755
		...				
18	3.5510	3.4987	0.0234	3.4491	3.5483	0.0523

11. An experiment was conducted to determine the extent to which the growth rate of a certain fungus could be affected by filling test tubes containing the same medium at the same temperature with different inert gases.¹⁰ Three such experiments were performed for each of six gases, and the average growth rate over these three tests was used as the response. The following table gives the molecular weight (X) of each gas used and the average growth rate (Y) in milliliters per hour for the three tests.

Gas	Average Growth Rate (Y)	Molecular Weight (X)
A	3.85	4.0
B	3.48	20.2
C	3.27	28.2
D	3.08	39.9
E	2.56	83.8
F	2.21	131.3

- a. Find the least-squares estimates of slope and intercept for the straight-line regression of Y on X, and draw the estimated straight line on a scatter diagram for this data set.
- b. Test for significant slope of the fitted straight line.
- c. What information has not been used that might improve the sensitivity of the analysis?

¹⁰ Adapted from a study by Schreiner, Gregoire, and Lawrie (1962).

- d. What is the 90% confidence interval for the true average growth rate if the gas used has a molecular weight of 100?
- e. Why would it be inappropriate to use the fitted line to estimate the growth rate of a gas whose molecular weight is 200?
- f. Based on the choice of X -values used in this study, how would you criticize the accuracy of prediction obtained in this experiment by using the fitted straight line?

Edited SAS Output (PROC REG) for Problem 11

Regression of Avg. Growth Rate (Y) on Molecular Weight (X)

DESCRIPTIVE STATISTICS					
Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	6.00000	1.00000	6.00000	0	0
X	307.40000	51.23333	27073	2264.85867	47.59053
Y	18.45000	3.07500	58.54990	0.36323	0.60269

[Portion of output omitted]

Root MSE	0.15123	R-Square	0.9496
----------	---------	----------	--------

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	3.70727	0.09546	38.83	<.0001
X	1	-0.01234	0.00142	-8.68	0.0010

12. Consider the data in the following table.¹¹

Age			Age		
Years	Months	Vocabulary Size	Years	Months	Vocabulary Size
0	8	0	3	0	896
0	10	1	3	6	1,222
1	0	3	4	0	1,540
1	3	19	4	6	1,870
1	6	22	5	0	2,072
1	9	118	5	6	2,289
2	0	272	6	0	2,562
2	6	446			

Note: Data are from M. E. Smith (1926).

¹¹ Taken from Bourne, Ekstrand, and Dominowski (1971, Table 14.3).

- First convert ages to decimal years (e.g., 1 year 6 months gives 1.5 years). Draw a scatter diagram for the variable pair vocabulary size (Y) and age (X).
- Calculate the least-squares estimates of the parameters of the regression line. Sketch this regression line on the scatter diagram.
- Add a new observation to the vocabulary data, with values 0.00 years and 0 words. Plot this new point. Logically, this value should be on the line of vocabulary growth. Is it near the fitted line from part (b)?
- Recompute the least-squares estimates of the regression line to include the (0, 0) observation ($n = 16$). Sketch the new line distinctly on the scatter diagram.
- Does either fitted line describe the data pattern adequately? If not, sketch your idea of the true relationship. If the data included observations through age 30 years, what would the extrapolated curve look like?
- The data appear to be from one child. If this is true, what assumption of the least-squares approach is most likely violated, and why?

Edited SAS Output (PROC REG) for Problem 12

Regression of Vocab (Y) on Age (X)

[Portion of output omitted]

Root MSE		148.07090	R-Square	0.9776	
PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-621.12595	74.08216	-8.38	<.0001
X	1	526.71836	22.13536	23.80	<.0001

Regression of Vocab (Y) on Age (X), Y on X, Including 16th OBS.

[Portion of output omitted]

Root MSE		205.90999	R-Square	0.9558	
PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-496.77535	92.13222	-5.39	<.0001
X	1	494.89316	28.43144	17.41	<.0001

- The following table gives rat body weights (in grams) and latency to seizure (in minutes), following injection of 40 mg/kg of body weight of metrazol.

Latency	Weight	Latency	Weight
2.30	348	2.00	409
1.95	372	1.70	413
2.90	378	2.00	415
2.30	390	2.95	423
1.10	392	1.25	428
2.50	395	2.05	464
1.30	400	3.70	468

- Draw a scatter diagram on graph paper, with latency plotted as a function of weight.
- Determine the least-squares estimates of slope and intercept for the straight-line regression of latency on weight.
- Test whether the slope is equal to 0. Use $\alpha = .10$.
- Test whether the intercept is equal to 0. Use $\alpha = .10$.
- Sketch the estimated regression line.
- Distinctly sketch your choice for a regression line if it differs from that of part (e). Explain why you agree or disagree with part (e), noting whether any assumptions appear to be violated.

Edited SAS Output (PROC REG) for Problem 13

Regression of Latency on Weight

[Portion of output omitted]

Root MSE		0.72779	R-Square	0.0519	
PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.11598	2.50749	0.05	0.9639
Weight	1	0.00498	0.00615	0.81	0.4333

- Stevens (1966), citing Dimmick and Hubbard (1939), reported data from 20 studies of the color perception of unitary hues. The wavelength (in millimeters) of light called green by subjects in each experiment is given in the following table, along with the year the study was conducted.

Study	Wavelength	Date	Study	Wavelength	Date	Study	Wavelength	Date
1	532	1874	8	506	1909	15	500	1928
2	535	1884	9	509	1911	16	506	1931
3	495	1888	10	520	1912	17	528	1934
4	527	1890	11	514	1920	18	530	1935
5	505	1898	12	504	1922	19	512	1935
6	505	1898	13	515	1926	20	515	1939
7	503	1907	14	498	1927			

- These data stimulate the question “Is there any linear trend over time in the wavelength of light called green?” Evaluate this question by finding the least-squares estimates of the straight-line regression function for predicting wavelength from year. [Hint: Subtract 500 from the wavelengths and 1850 from the year.]
- Find a 95% confidence interval for the true slope.
- Draw a scatter diagram on graph paper.

Edited SAS Output (PROC REG) for Problem 14

Regression of Wavelength on Date

[Portion of output omitted]

Root MSE		12.17807	R-Square	0.0310	
PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	19.88171	9.52906	2.09	0.0514
Date	1	-0.10933	0.14403	-0.76	0.4576

15. The data listed in the following table are from a study by Benignus and others (1981). Blood and brain levels of toluene (a commonly used solvent) were measured in rats following a 3-hour inhalation exposure to 50, 100, 500, or 1,000 parts per million (ppm) toluene (PPM_TOLU). Blood toluene (BLOODTOL) is expressed in parts per million, weight in grams, and age in days.

Rat	BLOODTOL	BRAINTOL	PPM_TOLU	WEIGHT	AGE	LN_BLDTL	LN_BRNTL	LN_PPMTL
1	0.553	0.481	50	393	95	-0.593	-0.732	3.912
2	0.494	0.584	50	378	95	-0.706	-0.538	3.912
3	0.609	0.585	50	450	95	-0.495	-0.536	3.912
4	0.763	0.628	50	439	95	-0.270	-0.465	3.912
5	0.420	0.533	50	397	95	-0.868	-0.629	3.912
6	0.397	0.490	50	301	84	-0.923	-0.713	3.912
7	0.503	0.719	50	406	84	-0.687	-0.330	3.912
8	0.534	0.585	50	302	84	-0.628	-0.536	3.912
9	0.531	0.675	50	382	84	-0.633	-0.393	3.912
10	0.384	0.442	50	355	84	-0.957	-0.816	3.912
11	0.215	0.492	50	405	85	-1.536	-0.709	3.912
12	0.552	0.859	50	405	85	-0.595	-0.152	3.912
13	0.420	0.650	50	387	85	-0.868	-0.431	3.912
14	0.324	0.528	50	358	85	-1.127	-0.639	3.912
15	0.387	0.546	50	311	85	-0.949	-0.605	3.912
16	1.036	1.262	100	355	86	0.035	0.233	4.605
17	1.065	1.584	100	440	86	0.063	0.460	4.605
18	1.084	1.773	100	421	86	0.081	0.573	4.605
19	0.944	1.307	100	370	86	-0.058	0.268	4.605

Rat	BLOODTOL	BRAINTOL	PPM_TOLU	WEIGHT	AGE	LN_BLDTL	LN_BRNTL	LN_PPMTL
20	0.994	1.338	100	375	86	-0.006	0.291	4.605
21	1.146	1.180	100	368	83	0.136	0.166	4.605
22	1.167	1.108	100	321	83	0.154	0.103	4.605
23	0.833	0.939	100	359	83	-0.183	-0.063	4.605
24	0.630	0.909	100	367	83	-0.462	-0.095	4.605
25	0.955	1.078	100	363	83	-0.046	0.075	4.605
26	0.687	1.152	100	388	86	-0.376	0.141	4.605
27	0.723	1.796	100	404	86	-0.324	0.586	4.605
28	0.705	1.262	100	454	86	-0.349	0.233	4.605
29	0.696	1.865	100	389	86	-0.363	0.623	4.605
30	0.868	1.892	100	352	86	-0.142	0.638	4.605
31	8.223	19.843	500	367	83	2.107	2.988	6.215
32	10.604	24.450	500	406	83	2.361	3.197	6.215
33	12.085	29.297	500	371	83	2.492	3.377	6.215
34	7.936	18.098	500	408	83	2.071	2.896	6.215
35	11.164	25.196	500	305	83	2.413	3.227	6.215
36	10.289	18.266	500	391	84	2.331	2.905	6.215
37	11.140	19.486	500	396	84	2.411	2.970	6.215
38	9.647	18.479	500	347	84	2.267	2.917	6.215
39	13.343	21.920	500	372	84	2.591	3.087	6.215
40	11.292	20.861	500	331	84	2.424	3.038	6.215
41	7.524	22.130	500	365	85	2.018	3.097	6.215
42	10.783	18.301	500	348	85	2.378	2.907	6.215
43	8.595	17.038	500	416	85	2.151	2.835	6.215
44	9.616	22.423	500	344	85	2.263	3.110	6.215
45	11.956	15.452	500	398	85	2.481	2.738	6.215
46	30.274	44.900	1000	417	93	3.410	3.804	6.908
47	32.923	35.500	1000	351	93	3.494	3.570	6.908
48	28.619	30.800	1000	378	93	3.354	3.428	6.908
49	28.761	38.500	1000	338	93	3.359	3.651	6.908
50	25.402	31.500	1000	433	93	3.235	3.450	6.908
51	35.464	42.330	1000	342	85	3.569	3.745	6.908
52	32.706	34.030	1000	319	85	3.488	3.527	6.908
53	29.347	30.760	1000	440	85	3.379	3.426	6.908
54	26.481	32.360	1000	363	85	3.276	3.477	6.908
55	33.401	41.830	1000	336	85	3.509	3.734	6.908
56	39.541	54.930	1000	378	86	3.677	4.006	6.908
57	28.155	39.780	1000	420	86	3.338	3.683	6.908
58	25.629	49.290	1000	346	86	3.244	3.898	6.908
59	33.188	47.490	1000	413	86	3.502	3.861	6.908
60	33.505	42.660	1000	432	86	3.512	3.753	6.908

- a. Provide a scatter diagram, with BLOODTOL as the response and PPM_TOLU as the predictor.
- b. Compute least-squares estimates of the straight-line regression coefficients. Plot the line on the scatter diagram.
- c. Repeat part (a) for the natural logarithms LN_BLDTL and LN_PPMTL, the respective natural logarithms of BLOODTOL and PPM_TOLU.
- d. Repeat part (b) for the natural logarithms.
- e. Which transformation leads to the best representation of the data? Note in your comments the validity of the regression assumptions.

Edited SAS Output (PROC REG) for Problem 15

Regression BLOODTOL on PPM_TOLU

[Portion of output omitted]

Root MSE	2.85301	R-Square	0.9497
----------	---------	----------	--------

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-2.54610	0.54254	-4.69	<.0001
PPM_TOLU	1	0.03196	0.00096570	33.09	<.0001

Regression of LN_BLDTL on LN_PPMTL

[Portion of output omitted]

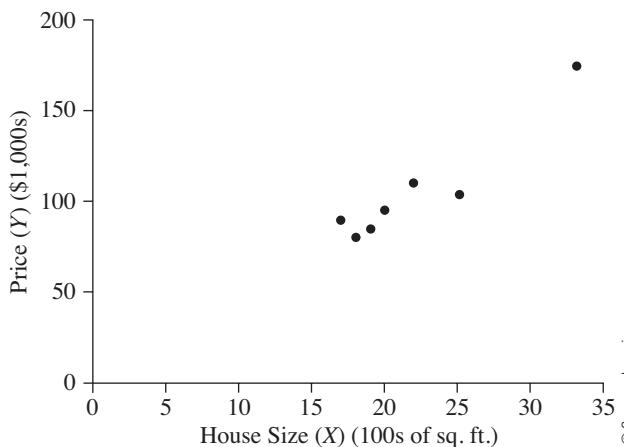
Root MSE	0.24144	R-Square	0.9813
----------	---------	----------	--------

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-6.53158	0.14365	-45.47	<.0001
LN_PPMTL	1	1.43045	0.02592	55.19	<.0001

16. Real estate prices depend, in part, on property size. The house size X (in hundreds of square feet) and house price Y (in thousands of dollars) of a random sample of houses in a certain county were recorded as in the following table.

House	1	2	3
X	18	20	25
Y	80	95	104

House	4	5	6	7
X	22	33	19	17
Y	110	175	85	89



- On the accompanying scatter diagram, sketch by eye a line that fits the data reasonably well. Comment on the relationship between house size and house price.
- Determine the least-squares estimates of the slope (β_1) and the intercept (β_0) for the straight-line regression of Y on X .
- Draw the estimated regression line on the scatter diagram. Comment on how well the line fits the data.
- Test the null hypothesis that the true slope is 0. Interpret the results of this test.

Edited SAS Output (PROC REG) for Problem 16

Regression of Price (Y) on House Size (X)

[Portion of output omitted]

Root MSE	10.70973	R-Square	0.9091
----------	----------	----------	--------

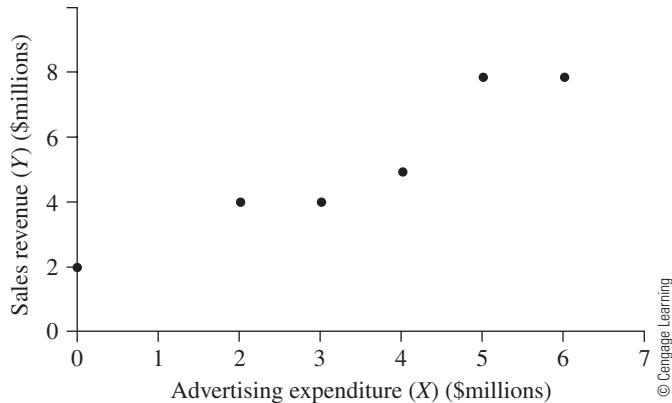
PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-17.36491	17.83513	-0.97	0.3750
X	1	5.58152	0.78953	7.07	0.0009

OUTPUT STATISTICS						
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		Residual
1	80.0000	83.1025	5.1341	69.9048	96.3002	-3.1025
2	95.0000	94.2655	4.3450	83.0964	105.4347	0.7345
3	104.0000	122.1731	4.6900	110.1172	134.2291	-18.1731
4	110.0000	105.4286	4.0479	95.0231	115.8340	4.5714
5	175.0000	166.8253	9.5819	142.1944	191.4563	8.1747
6	85.0000	88.6840	4.6900	76.6281	100.7399	-3.6840
7	89.0000	77.5210	5.6542	62.9865	92.0554	11.4790

17. Sales revenue (Y) and advertising expenditure (X) data for a large retailer for the period 1988–1993 are given in the following table.

Year	1988	1989	1990
Sales revenue (\$millions)	4	8	2
Advertising expenditure (\$millions)	2	5	0

Year	1991	1992	1993
Sales revenue (\$millions)	8	5	4
Advertising expenditure (\$millions)	6	4	3



- Does the plot of Y versus X suggest that a linear relationship exists between X and Y ?
- Calculate the least-squares estimates of the parameters of the regression line, and draw the estimated line on the accompanying scatter diagram. Does the line appear to fit the data well?
- Find a 95% confidence interval for the slope parameter. Based on your interval, is sales revenue linearly related to advertising expenditure? Explain.
- Would it be appropriate to use the estimated regression line in part (b) to estimate the sales for a new year in which an advertising expenditure of \$10 million is planned? Why or why not?

Edited SAS Output (PROC REG) for Problem 17

Regression of Sales (Y) on Adv (X)

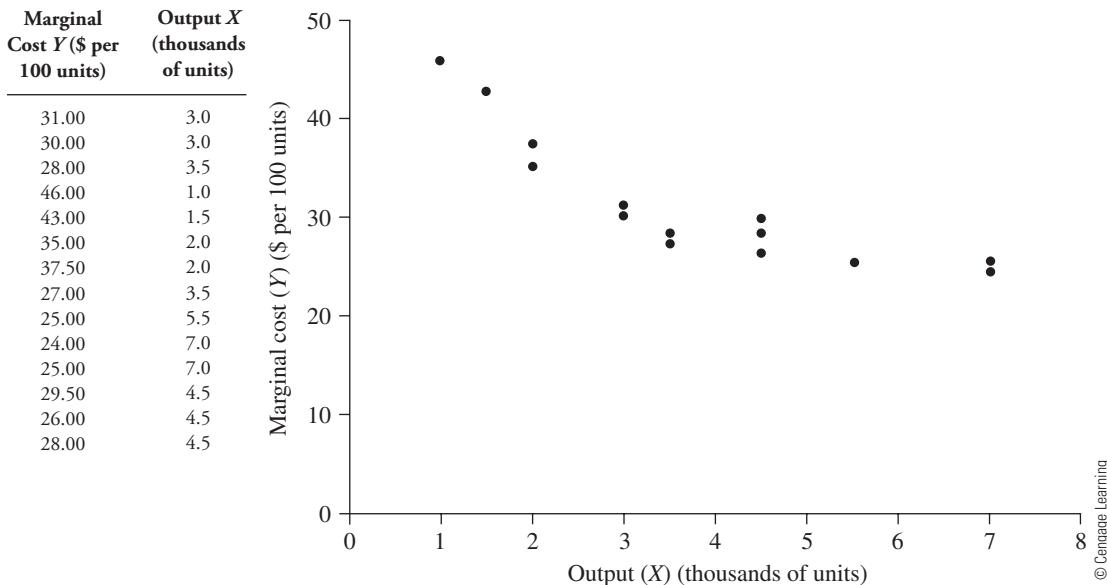
[Portion of output omitted]

Root MSE	0.83023	R-Square	0.9044
----------	---------	----------	--------

PARAMETER ESTIMATES

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	1.64286	0.66567	2.47	0.0691
X	1	1.05714	0.17187	6.15	0.0035

- The production manager of a plant that manufactures syringes records the marginal cost at various levels of output for 14 randomly selected months. The data are shown here:



© Cengage Learning

- On the accompanying scatter diagram of marginal cost (Y) versus output (X), sketch by eye a line that fits the data reasonably well.
- Find the estimated least-squares equation for the regression of marginal cost on output.
- Sketch the estimated line on the scatter diagram. Does it seem to fit the data well?
- Test the null hypothesis that the true slope is zero, at the $\alpha = .05$ significance level. Interpret your result.
- Can you suggest a model that would describe the marginal cost–output relationship for this manufacturer better than a straight line does?

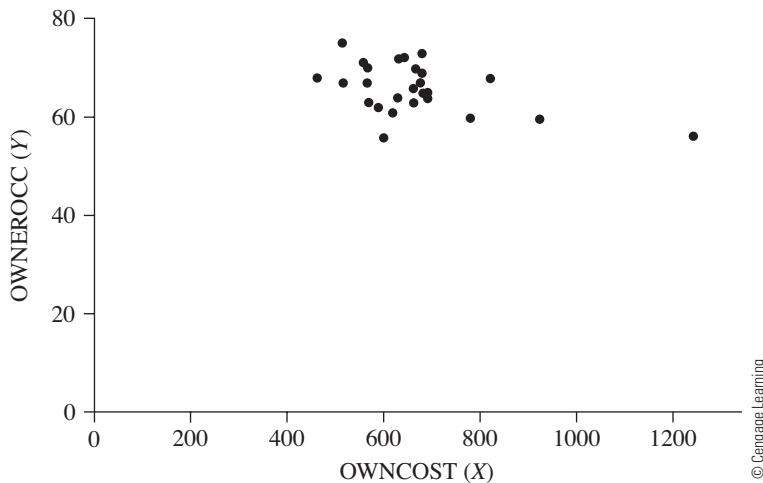
Edited SAS Output (PROC REG) for Problem 18

Regression of Marginal Cost (Y) on Output (X)

[Portion of output omitted]

Root MSE		3.63338	R-Square		0.7405
PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	42.84255	2.23377	19.18	<.0001
Output	1	-3.13896	0.53644	-5.85	<.0001

19. The data shown in the following table were obtained from the 1990 Census.¹² Included is information on 26 randomly selected Metropolitan Statistical Areas (MSAs). Of interest are factors that potentially are associated with the rate of owner occupancy of housing units. Two variables are included in the data set: OWNEROCC = percentage of housing units that are owner-occupied (as opposed to renter-occupied); and OWN-COST = median selected monthly ownership costs (in \$).
- a. Based on the accompanying scatter diagram of OWNEROCC versus OWN-COST, does there appear to be a linear relationship between these two variables?



© Cengage Learning

MSA	OWNEROCC (Y)	OWNCOST(X)
Abilene, TX	62	583
Burlington, NC	72	627
Daytona Beach, FL	72	636
Grand Rapids, MI	73	677
Laredo, TX	61	614
Louisville, KY-IN	67	561
Oklahoma City, OK	64	627
Pine Bluff, AR	67	513
San Francisco-Oakland-San Jose, CA	57	1,234
Wichita Falls, TX	63	565
Albany, GA	56	597
Canton, OH	71	555
Des Moines, IA	67	673
Jacksonville, FL	65	687
Johnstown, PA	75	510
Medford, OR	66	660
Omaha, NE-IA	64	687
Provo-Orem, UT	63	659
Williamsburg, PA	70	564

¹² The data were randomly sampled on June 20, 1996, from the 1990 U.S. Census.

MSA	OWNEROCC (Y)	OWNCOST(X)
Appleton–Oshkosh–Neenah, WI	70	663
Melbourne–Titusville–Palm Bay, FL	69	675
Redding, CA	65	682
Worcester, MA	60	918
Milwaukee–Racine, WI	60	777
Rochester, NY	68	818
St. Joseph, MO	68	457

- b. State the model for the straight-line regression of OWNEROCC (Y) on OWCOST (X). Determine the least-squares estimates for this regression line. Interpret the estimated values of the slope and the intercept in the context of the problem.
- c. Sketch the estimated line on the scatter diagram and assess the fit.
- d. Test for the significance of the slope parameter of the model in part (b). Interpret your result.
- e. Determine a 95% confidence interval for the true slope in part (d). Interpret your result with regard to the test mentioned in part (d).

Edited SAS Output (PROC REG) for Problem 19

Regression of OWNEROCC (Y) on OWCOST (X)

[Portion of output omitted]

Root MSE	4.41749	R-Square	0.2207		
PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	76.00764	3.94979	19.24	<.0001
Owncost	1	-0.01517	0.00582	-2.61	0.0155

- 20. Researchers have studied the ecology of ponds in rural Bangladesh. Of particular interest in such studies are the zooplankton, phytoplankton, and copepod counts (per liter of water) in these ponds. Copepods are a particular type of zooplankton that are thought to be a natural reservoir of cholera bacteria, but copepod counts can be difficult to obtain in laboratory analyses; total zooplankton counts are easier to obtain. It is also believed that copepod counts will be related to phytoplankton counts since copepods feed on phytoplankton.

The data here show zooplankton, phytoplankton, and copepod counts for 100 water samples from ponds in rural Bangladesh.

Obs	Copepods	Zooplankton	Phytoplankton	Obs	Copepods	Zooplankton	Phytoplankton
1	585	1,560	475.54	51	150	228	493.92
2	111	1,191	96.96	52	60	108	63.54
3	48	537	521.31	53	279	804	24.59
4	33	597	26.30	54	84	252	34.16
5	372	1,710	49.02	55	36	279	49.53
6	3	993	86.77	56	240	1,317	7,069.31
7	7	86	225.45	57	26	100	839.62
8	117	468	209.75	58	42	57	701.06
9	123	714	101.46	59	258	285	48.51
10	153	836	89.52	60	3	597	1,031.35
11	18	684	1,133.49	61	186	3,780	1,195.33
12	15	36	69.18	62	573	4,932	42.70
13	21	489	84.04	63	521	715	252.26
14	129	888	684.61	64	39	234	28.01
15	60	369	827.06	65	3	84	80.79
16	144	258	360.92	66	12	660	302.68
17	30	45	26.99	67	81	225	17.25
18	1,080	1,308	42.02	68	36	126	69.69
19	18	321	33.14	69	570	900	1,316.26
20	108	210	99.41	70	78	4,119	50.22
21	146	1,052	65.01	71	420	486	117.52
22	33	264	125.37	72	297	744	44.06
23	258	1,527	213.63	73	15	207	13.66
24	39	2,151	4,554.82	74	27	93	28.69
25	264	696	548.64	75	171	258	235.72
26	120	430	971.28	76	561	1,473	141.77
27	269	827	63.04	77	405	465	393.38
28	226	285	419.37	78	54	66	72.77
29	90	174	328.98	79	87	204	5.29
30	237	351	1,361.87	80	81	468	449.74
31	708	936	80.62	81	162	189	287.64
32	315	351	62.17	82	402	466	139.86
33	12	336	49.19	83	429	468	43.73
34	74	117	28.02	84	42	340	9.19
35	66	156	81.99	85	252	324	20.84
36	141	156	403.45	86	408	671	227.18
37	192	3,483	110.00	87	171	408	104.71
38	12	284	117.52	88	30	432	332.39
39	120	1,182	63.19	89	162	261	156.28
40	129	2,436	7,710.88	90	150	1,431	287.64
41	18	504	1,075.25	91	66	69	36.33
42	222	264	481.86	92	54	287	233.00
43	18	24	37.58	93	276	360	131.52
44	378	732	35.19	94	48	345	89.50
45	66	924	312.75	95	150	336	99.92
46	3	46	174.00	96	96	873	521.65
47	879	1,221	44.07	97	90	450	1,122.39
48	60	324	53.63	98	237	276	279.79
49	228	1,641	2,646.53	99	143	657	21.45
50	729	1,809	87.11	100	243	426	39.29

- a. Find the estimated least-squares equation for the regression of copepod count on zooplankton count.
 - b. Test the null hypothesis that the true slope is zero, at the $\alpha = .05$ significance level. Interpret your result.
 - c. Find a 95% confidence interval for the true slope in part (a). Interpret your result.
 - d. Find the estimated least-squares equation for the regression of copepod count on phytoplankton count.
 - e. Test the null hypothesis that the true slope in part (d) is zero, at the $\alpha = .05$ significance level. Interpret your result.
21. The BRFSS linear regression model of Section 5.12 was fit accounting for the sampling design and weights (using PROC SURVEYREG), with relevant output shown here.
- Answer the same set of questions previously addressed in Section 5.12 using the output for this new weighted analysis. Note that the degrees of freedom are different for this analysis ($df = 420,993$) and that relevant standard errors are provided, since they are more challenging to compute manually.
- a. What is the form of the calculated straight-line fit using alcohol frequency (“drink days”) as a predictor of BMI?
 - b. Determine if there is a statistically significant linear relationship between alcohol frequency and BMI.
 - c. What is the estimated mean BMI and corresponding 95% confidence interval for a woman who drinks the median number of 4 days per month? For one who drinks 15 days per month (90th percentile of consumption frequency)?
 - d. Comment on how these results do, or do not, differ from those for the unweighted analysis.

Edited SAS Output (PROC SURVEYREG) for Problem 21

The SURVEYREG Procedure

Domain Regression Analysis for Variable bmi

ESTIMATED REGRESSION COEFFICIENTS		
Parameter	Estimate	Standard Error
Intercept	27.5302751	0.34731875
drink_days	-0.1332682	0.03673852
ESTIMATE		
Label	Standard Error	
4 drink days / month	0.2658	
15 drink days / month	0.3950	

Standard errors for
confidence intervals
for the mean BMI

Note: The degrees of freedom for the t tests is 420,993.

References

- Benignus, V. A.; Muller, K. E.; Barton, C. N.; and Bittekofer, J. A. 1981. "Toluene Levels in Blood and Brain of Rats during and after Respiratory Exposure." *Toxicology and Applied Pharmacology* 61: 326–34.
- Bourne, L. E.; Ekstrand, B. E.; and Dominowski, R. L. 1971. *The Psychology of Thinking*. Englewood Cliffs, N.J.: Prentice-Hall.
- Breslow, R. A., and Smothers, B. A. 2005. "Drinking Patterns and Body Mass Index in Never Smokers: National Health Interview Survey, 1997–2001." *American Journal of Epidemiology* 161(4): 368–76.
- Diggle, P. J.; Heagerty, P.; Liang, K. Y.; and Zeger, S. L. 2002. *Analysis of Longitudinal Data*. New York: Oxford University Press.
- Dimmick, F. L., and Hubbard, M. R. 1939. "The Spectral Location of Psychologically Unique Yellow, Green, and Blue." *American Journal of Psychology* 52: 242.
- Morrison, D. F. 1976. *Multivariate Statistical Methods*. New York: McGraw-Hill.
- Nagasawa, S.; Osana, S.; and Kondo, K. 1964. "An Analytical Method for Evaluating the Susceptibility of Fish Species to an Agricultural Chemical." *Japanese Journal of Applied Enterological Zoology* 8: 118–22.
- Schønheyder, F. 1936. "The Quantitative Determination of Vitamin K." *Biochemistry Journal* 30: 890–96.
- Schreiner, H. R.; Gregoine, R. C.; and Lawrie, J. A. 1962. "New Biological Effects of the Gases on the Helium Group." *Science* 136: 653–54.
- Siegel, S. 1956. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.
- Smith, M. E. 1926. "An Investigation of the Development of the Sentence and the Extent of Vocabulary in Young Children." *Studies in Child Welfare* 3: 5.
- Stevens, S. S. 1966. *Handbook of Experimental Psychology*. New York: John Wiley & Sons.
- Timm, N. H. 1975. *Multivariate Analysis with Applications in Education and Psychology*. Monterey, Calif.: Brooks/Cole.
- Zeger, S. L., and Liang, K. Y. 1986. "Longitudinal Data Analysis for Discrete and Continuous Outcomes" *Biometrics* 42: 121–30.

6

The Correlation Coefficient and Straight-line Regression Analysis

6.1 Definition of r

The correlation coefficient is an often-used statistic that provides a measure of how two random variables are linearly associated in a sample and has properties closely related to those of straight-line regression. We define the *sample correlation coefficient* r for two variables X and Y by the formula

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\left(\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2 \right)^{1/2}} = \frac{\text{SSXY}}{\sqrt{\text{SSX} \cdot \text{SSY}}} \quad (6.1)$$

where $\text{SSXY} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$, $\text{SSX} = \sum_{i=1}^n (X_i - \bar{X})^2$, and $\text{SSY} = \sum_{i=1}^n (Y_i - \bar{Y})^2$.

An equivalent formula for r that illustrates its mathematical relationship to the least-squares estimate of the slope of a fitted regression line is¹

$$r = \frac{S_X \hat{\beta}_1}{S_Y} \quad (6.2)$$

¹ $S_X^2 = \frac{1}{n-1} \text{SSX}$ and $S_Y^2 = \frac{1}{n-1} \text{SSY}$ are the estimated sample variances of the X and Y variables, respectively.

■ **Example 6.1** For the age–systolic blood pressure data in Table 5.1, r is 0.66. This value can be obtained from the SAS output on page 59 by taking the square root of the R -square value of 0.4324.

Alternatively, using (6.2), we have

$$r = \frac{15.29}{22.58}(0.97) = 0.66$$

Three important mathematical properties are associated with r :

1. The possible values of r range from -1 to 1 .
2. r is a dimensionless quantity; that is, r is independent of the units of measurement of X and Y .
3. r is positive, negative, or zero as $\hat{\beta}_1$ is positive, negative, or zero; and vice versa. This property follows directly, of course, from (6.2). ■

6.2 r as a Measure of Association

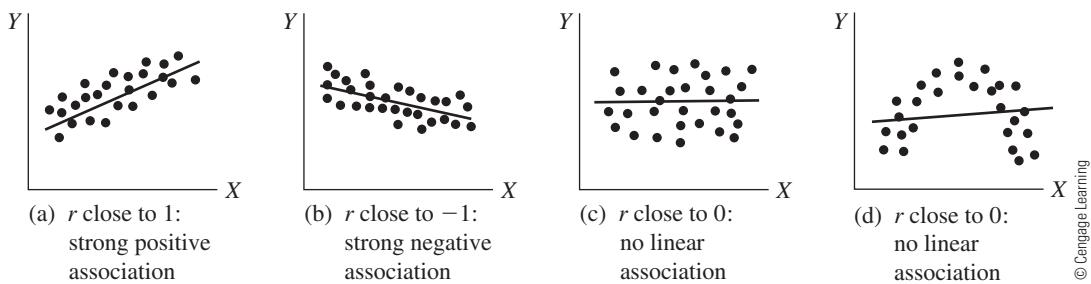
In the statistical assumptions for straight-line regression analysis discussed earlier, we did not specifically consider the variable X to be a random variable. Nevertheless, it often makes sense to view the regression problem as one in which both X and Y are random variables. The measure r can then be interpreted as an *index of linear association* between X and Y , in the following sense:

1. The more positive r is, the more positive the association is. This means that, when r is close to 1 , an individual with a high value for one variable will likely have a high value for the other, and an individual with a low value for one variable will likely have a low value for the other (Figure 6.1(a)).
2. The more negative r is, the more negative the association is; that is, an individual with a high value for one variable will likely have a low value for the other when r is close to -1 , and conversely (Figure 6.1(b)).
3. If r is close to 0 , there is little, if any, *linear* association between X and Y (Figure 6.1(c) or 6.1(d)).²

By *association*, we mean the lack of statistical independence between X and Y . More loosely, the lack of an association means that the value of one variable cannot be reasonably anticipated from knowing the value of the other variable.

Since r is an index obtained from a *sample* of n observations, it can be considered as an estimate of an unknown population parameter. This unknown parameter, called the *population correlation coefficient*, is generally denoted by the symbol ρ_{XY} or more simply ρ (if it is clearly understood which two variables are being considered). We shall agree to use ρ unless confusion is possible. The parameter ρ_{XY} is defined as $\rho_{XY} = \sigma_{XY}/\sigma_X\sigma_Y$, where σ_X and σ_Y

² Later we will see that a value of r close to 0 does not rule out a possible *nonlinear* association.

**FIGURE 6.1** Correlation coefficient as a measure of association

denote the population standard deviations of the random variables X and Y and where σ_{XY} is called the population covariance between X and Y . The covariance σ_{XY} is a population parameter describing the average amount by which two variables covary. In actuality, it is the population mean of the random variable $SSXY/(n - 1)$.

Figure 6.2 provides informative examples of scatter diagrams. Data were generated (via computer simulation) to have means and variances similar to those of the age–systolic blood pressure data of Chapter 5. The six samples observed here were produced by selecting 30 paired observations at random from each of six populations for which the population correlation coefficient ρ varied in value.

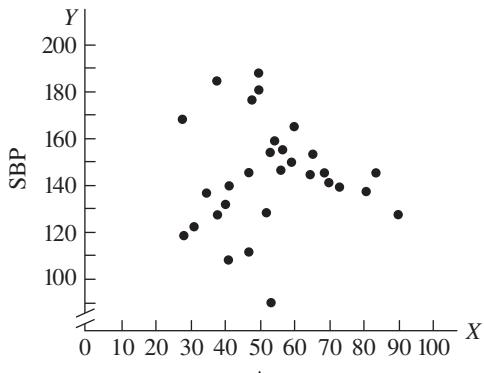
In Figure 6.2, the sample correlations range in value from 0.037 to 0.894. It should be clear that an eyeball analysis of the relative strengths of association is difficult, even though $n = 30$. For example, the difference between $r = 0.037$ in Figure 6.2(a) and $r = 0.220$ in Figure 6.2(b) is apparently due to the influence of just a few points. The study of so-called influential data points will be described in Chapter 14 on regression diagnostics.

In evaluating a scatter diagram, we find it helpful to include reference lines at the X and Y means, as in Figure 6.2(f). Roughly speaking, the proportions of observations in each quadrant reflect the strength of association. Notice that most of the observations in this figure are located in quadrants B and C, which are often referred to as the *positive quadrants*. Quadrants A and D are called the *negative quadrants*. When more observations are in positive quadrants than in negative quadrants, the sample correlation coefficient r is usually positive. On the other hand, if more observations are in negative quadrants, r is usually negative.

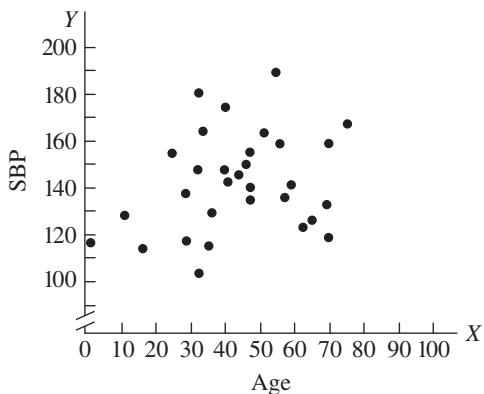
To understand why this is so, we need to examine the numerator part of equation (6.1)—namely,

$$\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

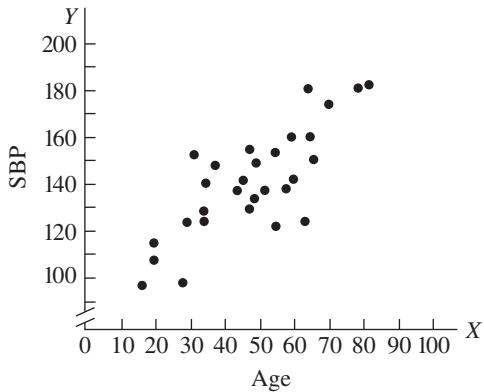
(Notice that the denominator in (6.1) is simply a positive scaling factor ensuring that r both is dimensionless and satisfies the inequality $-1 \leq r \leq 1$.) The numerator describes how X and Y covary in terms of the n cross-products $(X_i - \bar{X})(Y_i - \bar{Y})$, where $i = 1, 2, \dots, n$. For a given i , such a cross-product term is either positive or negative (or zero), depending on how X_i compares with \bar{X} and how Y_i compares with \bar{Y} . In particular, if the i th observation (X_i, Y_i) is in quadrant B, then $X_i > \bar{X}$ and $Y_i > \bar{Y}$; hence, the product of $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ must be positive. Similarly, if (X_i, Y_i) is in quadrant C, $X_i < \bar{X}$ and $Y_i < \bar{Y}$, so $(X_i - \bar{X})(Y_i - \bar{Y})$



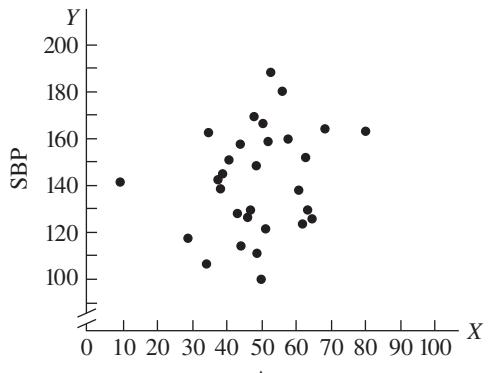
(a) $r = 0.037$: $\bar{X} = 52.77$, $\bar{Y} = 145.12$
 $S_X = 15.91$, $S_Y = 22.68$



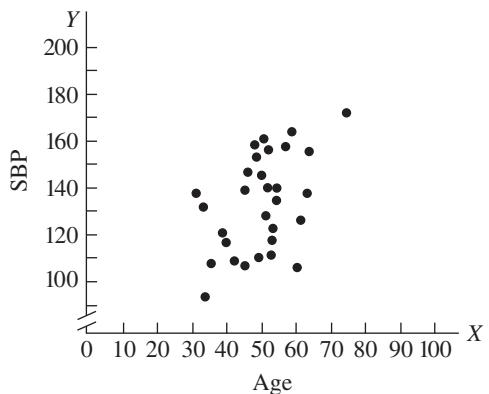
(c) $r = 0.342$: $\bar{X} = 48.28$, $\bar{Y} = 143.66$
 $S_X = 13.38$, $S_Y = 22.09$



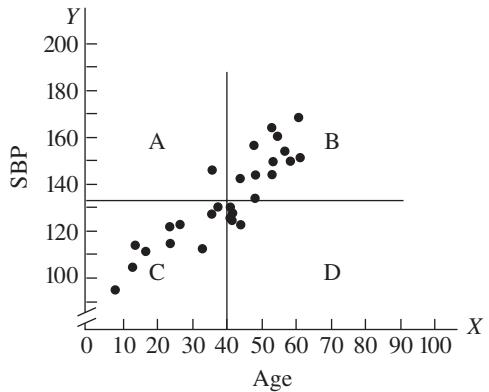
(e) $r = 0.814$: $\bar{X} = 48.95$, $\bar{Y} = 143.48$
 $S_X = 18.37$, $S_Y = 23.43$



(b) $r = 0.220$: $\bar{X} = 43.67$, $\bar{Y} = 143.01$
 $S_X = 18.19$, $S_Y = 20.11$



(d) $r = 0.485$: $\bar{X} = 50.12$, $\bar{Y} = 134.05$
 $S_X = 9.72$, $S_Y = 20.65$



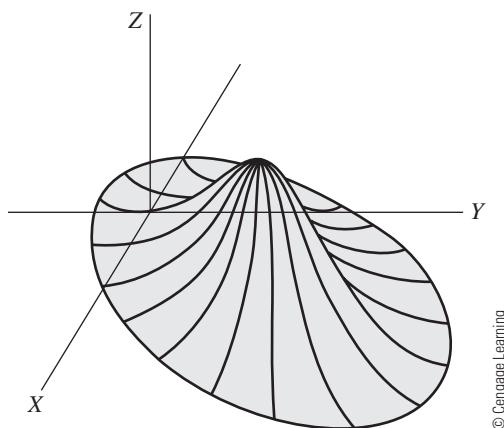
(f) $r = 0.894$: $\bar{X} = 40.33$, $\bar{Y} = 135.7$
 $S_X = 15.07$, $S_Y = 18.67$

FIGURE 6.2 Examples of a range of observed correlations between age and systolic blood pressure (SBP) for simulated data

is again positive. Thus, observations in the positive quadrants B and C contribute positive values to the numerator of (6.1). Conversely, observations in the negative quadrants A and D contribute negative values to this numerator. So, roughly speaking, the sign of the correlation coefficient reflects the distribution of observations in these positive and negative quadrants.

6.3 The Bivariate Normal Distribution³

Another way of looking at straight-line regression is to consider X and Y as random variables having the *bivariate normal distribution*, which is a generalization of the *univariate normal distribution*. Just as the univariate normal distribution is described by a density function that appears as a bell-shaped curve when plotted in two dimensions, the bivariate normal distribution is described by a *joint density function* whose plot looks like a bell-shaped surface in three dimensions (Figure 6.3).



© Cengage Learning

FIGURE 6.3 The bivariate normal distribution

One property of the bivariate normal distribution that has implications for straight-line regression analysis is the following: if the bell-shaped surface is cut by a plane *parallel* to the YZ -plane and passing through a specific X -value, the curve, or *trace*, that results is a normal distribution. In other words, the distribution of Y for fixed X is univariate-normal. We call such a distribution the *conditional distribution* of Y at X , and we denote the corresponding random variable as Y_X . Let us denote the mean of this distribution as $\mu_{Y|X}$ and the variance as $\sigma_{Y|X}^2$. Then it follows from statistical theory that the mean and the variance, respectively, of Y_X can be written in terms of μ_X , μ_Y , σ_X^2 , σ_Y^2 , and ρ_{XY} as follows:

$$\mu_{Y|X} = \mu_Y + \rho_{XY} \frac{\sigma_Y}{\sigma_X} (X - \mu_X) \quad (6.3)$$

³ This section is not essential for understanding the correlation coefficient as it relates to regression analysis.

and

$$\sigma_{Y|X}^2 = \sigma_Y^2(1 - \rho_{XY}^2) \quad (6.4)$$

Now suppose that we let $\beta_1 = \rho_{XY}(\sigma_Y/\sigma_X)$ and $\beta_0 = \mu_Y - \beta_1\mu_X$. Then (6.3) has been transformed into the familiar expression for a straight-line model; that is, $\mu_{Y|X} = \beta_0 + \beta_1 X$. Furthermore, if we substitute the estimators \bar{X} , \bar{Y} , S_X , S_Y , and r for their respective parameters μ_X , μ_Y , σ_X , σ_Y , and ρ_{XY} in (6.3), we obtain the formula

$$\hat{\mu}_{Y|X} = \bar{Y} + r \frac{S_Y}{S_X}(X - \bar{X})$$

The right-hand side of this equation is exactly equivalent to the expression for the least-squares straight line given in (5.7), since

$$\hat{\beta}_1 = r \frac{S_Y}{S_X}$$

Thus, the least-squares formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ can be developed by assuming that X and Y are random variables having the bivariate normal distribution and by substituting the usual estimates of μ_X , μ_Y , σ_X , σ_Y , and ρ_{XY} into the expression for $\mu_{Y|X}$, the conditional mean of Y given X .

Our estimate $S_{Y|X}^2$ of $\sigma_{Y|X}^2$ can also be obtained by substituting the estimates S_Y^2 and r for σ_Y^2 and ρ_{XY} in (6.4). Thus, we obtain

$$S_{Y|X}^2 = S_Y^2(1 - r^2)$$

Finally, (6.4) can be algebraically manipulated into the form

$$\rho_{XY}^2 = \frac{\sigma_Y^2 - \sigma_{Y|X}^2}{\sigma_Y^2} \quad (6.5)$$

This equation describes the square of the population correlation coefficient as the proportionate reduction in the variance of Y due to conditioning on X . The importance of (6.5) in describing the strength of the straight-line relationship will be discussed in the next section.

6.4 r^2 and the Strength of the Straight-line Relationship

To quantify what we mean by the *strength* of the linear relationship between X and Y , we should first consider what our predictor of Y would be if we did not use X at all. The best predictor in this case would simply be \bar{Y} , the sample mean of the Y 's. The sum of the squares

of deviations associated with the naive predictor \bar{Y} would then be given by the formula

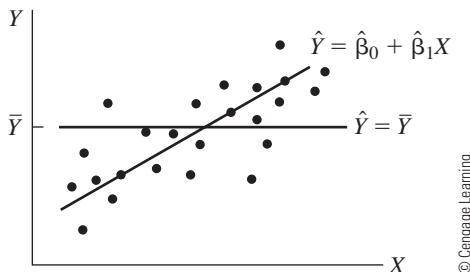
$$\text{SSY} = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

Now, if the variable X is of any value in predicting the variable Y , the residual sum of squares given by

$$\text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

should be considerably less than SSY. If so, the least-squares model $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ fits the data better than does the horizontal line $\hat{Y} = \bar{Y}$ (Figure 6.4). A quantitative measure of the improvement in the fit⁴ obtained by using X is given by the *square of the sample correlation coefficient r*, which can be equivalently expressed by using (6.1) and some algebra or by substituting estimates for parameters in (6.5):

$$r^2 = \frac{\text{SSY} - \text{SSE}}{\text{SSY}} \quad (6.6)$$



© Cengage Learning

FIGURE 6.4 Predictions of Y using and not using X

This quantity naturally varies between 0 and 1, since r itself varies between -1 and 1 .

What interpretation can be given to the quantity r^2 ? To answer this question, we first note that the difference, or *reduction*, in SSY due to using X to predict Y may be measured by $(\text{SSY} - \text{SSE})$, which is always nonnegative. Furthermore, the *proportionate reduction in SSY* due to using X to predict Y is this difference divided by SSY. Thus, r^2 measures the strength of the linear relationship between X and Y in the sense that it gives the proportionate reduction in the sum of the squares of vertical deviations obtained by using the least-squares line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ instead of the naive model $\hat{Y} = \bar{Y}$ (the predictor of Y if X is ignored).

⁴ When a straight-line model is being considered, r^2 provides both a quantitative measure of the improvement in fit obtained by using X in the model and a measure of the strength of the linear relationship between X and Y . When we discuss more general multiple linear regression models (starting in Chapter 8), we will see that a generalized version of r^2 , the squared multiple correlation coefficient R^2 , provides an overall measure of the ability of the fitted model to predict the response Y .

The larger the value of r^2 , the greater the reduction in SSE relative to $\sum_{i=1}^n (Y_i - \bar{Y})^2$, and the stronger the linear relationship between X and Y .

As stated earlier, the largest value that r^2 can attain is 1, which occurs when $\hat{\beta}_1$ is nonzero and when $SSE = 0$ (i.e., when a perfect positive or negative straight-line relationship exists between X and Y). By “perfect,” we mean that *all* the data points lie on the fitted straight line. In other words, when $Y_i = \hat{Y}_i$ for all i , we must have

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 0$$

so

$$r^2 = \frac{SSY - SSE}{SSY} = \frac{SSY}{SSY} = 1$$

Figure 6.5 illustrates examples of perfect positive and perfect negative linear association.

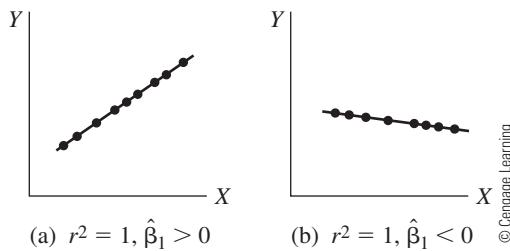


FIGURE 6.5 Examples of perfect linear association

The smallest value that r^2 may take, of course, is 0. This value means that using X offers no improvement in predictive power; that is, $SSE = SSY$. Furthermore, appealing to (6.2), we see that a correlation coefficient of 0 implies an estimated slope of 0 and consequently the absence of any linear relationship (although a nonlinear relationship is still possible).

Finally, one should *not* be led to a false sense of security by considering the magnitude of r , rather than of r^2 , when assessing the strength of the linear association between X and Y . For example, when r is 0.5, r^2 is only 0.25, and it takes $r > 0.7$ to make $r^2 > 0.49$. Also, when r is 0.3, r^2 is 0.09, which indicates that only 9% of the variation in Y is explained with the help of X . For the age–systolic blood pressure data, r^2 is 0.43, compared with an r of 0.66. The r^2 value also appears on the SAS output on page 59.

6.5 What r^2 Does Not Measure

Two common misconceptions about r^2 occasionally lead a researcher to make spurious interpretations of the relationship between X and Y . The correct notions are as follows:

1. r^2 is not a measure of the magnitude of the slope of the regression line. Even when the value of r^2 is high (i.e., close to 1), the magnitude of the slope $\hat{\beta}_1$ is not necessarily

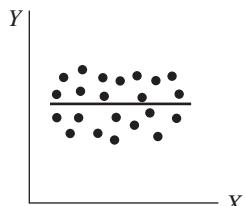
large. This phenomenon is illustrated in Figure 6.5. Notice that r^2 equals 1 in both parts, despite the fact that the slopes are different. This is because, in both graphs, there is no deviation of any points from the fitted straight lines, so that SSE = 0 in both cases. In general, the more the data points cluster closely to the fitted straight line, the larger the value of r^2 will be. Another way to understand this, using (6.2), is

$$\hat{\beta}_1^2 = \frac{S_Y^2}{S_X^2} \quad \text{when } r^2 = 1$$

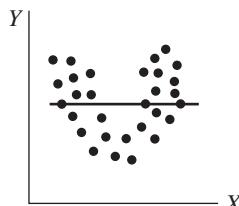
Thus, if two different sets of data have the same amount of X variation, but the first set has less Y variation than the second set, the magnitude of the slope for the first set is smaller than that for the second.

2. r^2 is not a measure of the appropriateness of the straight-line model. Note that $r^2 = 0$ in parts (a) and (b) of Figure 6.6 even though no evidence of association between X and Y exists in (a) and strong evidence of a nonlinear association exists in (b). Conversely, r^2 is high in parts (c) and (d), even though a straight-line model is quite appropriate in (c) but not entirely appropriate in (d).

Examples when $r^2 = 0$

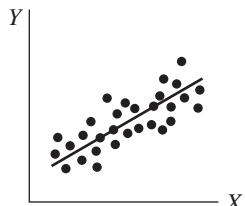


(a) No association between X and Y

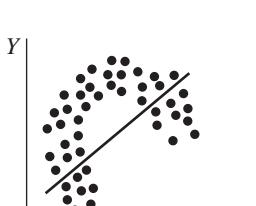


(b) No linear association between X and Y

Examples when r^2 is high



(c) Straight-line association between X and Y



(d) Curvilinear association between X and Y

© Cengage Learning

FIGURE 6.6 Examples showing that r^2 is not a measure of the appropriateness of the straight-line model

6.6 Tests of Hypotheses and Confidence Intervals for the Correlation Coefficient

Researchers interested in the association between two interval variables X and Y often want to test the null hypothesis $H_0: \rho = 0$.

6.6.1 Test of $H_0: \rho = 0$

A test of $H_0: \rho = 0$ turns out to be mathematically equivalent to the test of the null hypothesis $H_0: \beta_1 = 0$ described in Section 5.8. This equivalence is suggested by the formulas $\beta_1 = \rho\sigma_Y/\sigma_X$ and $\hat{\beta}_1 = rS_Y/S_X$, which tell us, for example, that β_1 is positive, negative, or zero as ρ is positive, negative, or zero and that an analogous relationship exists between $\hat{\beta}_1$ and r . The test statistic for the hypothesis $H_0: \rho = 0$ can be written entirely in terms of r and n , so we can perform the test without having to fit the straight line. This test statistic is given by the formula

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (6.7)$$

which has the t distribution with $n - 2$ degrees of freedom when the null hypothesis $H_0: \rho = 0$ (or equivalently, $H_0: \beta_1 = 0$) is true. Formula (6.7) yields exactly the same numerical answer as does (5.9), given by

$$T = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{S_{\hat{\beta}_1}} \quad \text{when } \beta_1^{(0)} = 0$$

Example 6.2 For the age–systolic blood pressure data of Table 5.1, for which $r = 0.66$, the statistic in (6.7) is calculated as follows:

$$T = \frac{0.66\sqrt{30-2}}{\sqrt{1-(0.66)^2}} = 4.62$$

which is the same value as obtained for the test for slope in Table 5.3. ■

6.6.2 Test of $H_0: \rho = \rho_0$, When $\rho_0 \neq 0$

Sometimes we wish to assess whether the population correlation coefficient is equal to a specific nonzero value. A test concerning the null hypothesis $H_0: \rho = \rho_0 (\rho_0 \neq 0)$ cannot be directly related to a test concerning β_1 since the ratio σ_Y/σ_X is an unknown parameter. Nevertheless, a test of $H_0: \rho = \rho_0 (\rho_0 \neq 0)$ is meaningful when previous experience or theory suggests a particular value to use for ρ_0 .

The test statistic in this case can be obtained by considering the distribution of the sample correlation coefficient r . This distribution happens to be symmetric, like the normal

distribution, *only* when ρ is 0. When ρ is nonzero, the distribution of r is skewed. This lack of normality prevents us from using a test statistic of the usual form, which has a normally distributed estimator in the numerator and an independent estimator of its standard deviation in the denominator. But through an appropriate transformation, r can be changed into a statistic that is often approximately normal. This transformation is called *Fisher's Z transformation*.⁵ The formula for this transformation is

$$\frac{1}{2} \ln \frac{1+r}{1-r} \quad (6.8)$$

This quantity has approximately the normal distribution, with mean $\frac{1}{2} \ln [(1+\rho)/(1-\rho)]$ and variance $1/(n-3)$ when n is not too small (e.g., $n \geq 20$). In testing the null hypothesis $H_0: \rho = \rho_0$ ($\rho_0 \neq 0$), we can then use the test statistic

$$Z = \frac{\frac{1}{2} \ln [(1+r)/(1-r)] - \frac{1}{2} \ln [(1+\rho_0)/(1-\rho_0)]}{1/\sqrt{n-3}} \quad (6.9)$$

This test statistic has approximately the standard normal distribution (i.e., $Z \sim N(0, 1)$) under H_0 . To test $H_0: \rho = \rho_0$ ($\rho_0 \neq 0$), therefore, we use one of the following critical regions for significance level α :

$$Z \geq z_{1-\alpha} \quad (\text{upper one-tailed alternative } H_A: \rho > \rho_0)$$

$$Z \leq -z_{1-\alpha} \quad (\text{lower one-tailed alternative } H_A: \rho < \rho_0)$$

$$|Z| \geq z_{1-\alpha/2} \quad (\text{two-tailed alternative } H_A: \rho \neq \rho_0)$$

where $z_{1-\alpha}$ denotes the $100(1-\alpha)\%$ point of the standard normal distribution. Computation of Z can be aided by using Appendix Table A.5, which gives values of $\frac{1}{2} \ln [(1+r)/(1-r)]$ for given values of r .

Example 6.3 Suppose that from previous experience we can hypothesize that the true correlation between age and systolic blood pressure is $\rho_0 = 0.85$. To test the null hypothesis $H_0: \rho = 0.85$ against the two-sided alternative $H_A: \rho \neq 0.85$, we perform the following calculations using $r = 0.66$, $\rho_0 = 0.85$, and $n = 30$:

$$\frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0} = \frac{1}{2} \ln \frac{1+0.85}{1-0.85} = 1.2561 \quad (\text{from Table A.5})$$

$$\frac{1}{2} \ln \frac{1+r}{1-r} = \frac{1}{2} \ln \frac{1+0.66}{1-0.66} = 0.7928 \quad (\text{from Table A.5})$$

$$Z = \frac{0.7928 - 1.2561}{1/\sqrt{30-3}} = -2.41$$

⁵ Named after R. A. Fisher, who introduced it in 1925.

For $\alpha = .05$, the critical region is

$$|Z| \geq z_{.975} = 1.96$$

Since $|Z| = 2.41$ exceeds 1.96, the hypothesis $H_0: \rho_0 = 0.85$ is rejected at the .05 significance level. Further calculations show that the P -value for this test is $P = .0151$, which tells us that the result is not significant at $\alpha = .01$. ■

6.6.3 Confidence Interval for ρ

A $100(1 - \alpha)\%$ confidence interval for ρ can be obtained by using Fisher's Z transformation (6.8) as follows. First, compute a $100(1 - \alpha)\%$ confidence interval for the parameter $\frac{1}{2} \ln [(1 + \rho)/(1 - \rho)]$ using the formula

$$\frac{1}{2} \ln \frac{1 + r}{1 - r} \pm \frac{z_{1-\alpha/2}}{\sqrt{n-3}} \quad (6.10)$$

where $z_{1-\alpha/2}$ is as defined previously.

Denote the lower limit of the confidence interval (6.10) by L_Z and the upper limit by U_Z ; then use Appendix Table A.5 (in reverse) to determine the lower and upper confidence limits L_ρ and U_ρ for the confidence interval for ρ . In other words, determine L_ρ and U_ρ from the following formulas⁶:

$$L_Z = \frac{1}{2} \ln \frac{1 + L_\rho}{1 - L_\rho} \quad \text{and} \quad U_Z = \frac{1}{2} \ln \frac{1 + U_\rho}{1 - U_\rho}$$

■ **Example 6.4** Suppose that we seek a 95% confidence interval for ρ based on the age-systolic blood pressure data for which $r = 0.66$ and $n = 30$. A 95% confidence interval for $\frac{1}{2} \ln [(1 + \rho)/(1 - \rho)]$ is given by

$$\frac{1}{2} \ln \frac{1 + 0.66}{1 - 0.66} \pm \frac{1.96}{\sqrt{30-3}}$$

which is equal to

$$0.793 \pm 0.377$$

providing a lower limit of $L_Z = 0.416$ and an upper limit of $U_Z = 1.170$.

To transform these L_Z and U_Z values into lower and upper confidence limits for ρ , we determine the values of L_ρ and U_ρ that satisfy

$$0.416 = \frac{1}{2} \ln \frac{1 + L_\rho}{1 - L_\rho} \quad \text{and} \quad 1.170 = \frac{1}{2} \ln \frac{1 + U_\rho}{1 - U_\rho}$$

⁶ L_ρ and U_ρ can also be calculated directly using the conversion formulas:

$$L_\rho = \frac{e^{2L_Z} - 1}{e^{2L_Z} + 1} \quad \text{and} \quad U_\rho = \frac{e^{2U_Z} - 1}{e^{2U_Z} + 1}$$

Using Table A.5, we see that a value of 0.416 corresponds to an r of about 0.394, so $L_p = 0.394$. Similarly, a value of 1.170 corresponds to an r of about 0.824, so $U_p = 0.824$. The 95% confidence interval for ρ thus has a lower limit of 0.394 and an upper limit of 0.824.

Notice that the interval (0.394, 0.824) does not contain the value 0.85, which agrees with the conclusion of the previous section that $H_0: \rho = 0.85$ is to be rejected at the 5% level (two-tailed test). ■

6.7 Testing for the Equality of Two Correlations

Suppose that independent random samples of sizes n_1 and n_2 are selected from two populations. Further, suppose that we want to test $H_0: \rho_1 = \rho_2$ versus, say, $H_A: \rho_1 \neq \rho_2$. An appropriate test statistic can be developed based on the results given in Section 6.6. In this section, we will also consider the situation in which the sample correlations to be compared are calculated by using the same data set; in this case, the sample correlations are themselves “correlated.”

6.7.1 Test of $H_0: \rho_1 = \rho_2$ Using Independent Random Samples

Let us assume that independent random samples of sizes n_1 and n_2 have been selected from two populations. For each population, the straight-line regression analysis assumptions given in Chapter 5, including that of normality, are assumed to hold.

An approximate test of $H_0: \rho_1 = \rho_2$ can be based on the use of Fisher’s Z transformation. Let r_1 be the sample correlation calculated by using the n_1 observations from the first population, and let r_2 be defined similarly. Using (6.8), let

$$Z_1 = \frac{1}{2} \ln \frac{1 + r_1}{1 - r_1} \quad (6.11)$$

and

$$Z_2 = \frac{1}{2} \ln \frac{1 + r_2}{1 - r_2} \quad (6.12)$$

Appendix Table A.5 can be used to determine Z_1 and Z_2 .

To test $H_0: \rho_1 = \rho_2$, we can compute the test statistic

$$Z = \frac{Z_1 - Z_2}{\sqrt{1/(n_1 - 3) + 1/(n_2 - 3)}} \quad (6.13)$$

For large n_1 and n_2 , this test statistic has (approximately) the standard normal distribution when H_0 is true. Hence, the following critical regions for significance level α should

be used:

$$Z \geq z_{1-\alpha} \quad (\text{upper one-tailed alternative } H_A: \rho_1 > \rho_2)$$

$$Z \leq -z_{1-\alpha} \quad (\text{lower one-tailed alternative } H_A: \rho_1 < \rho_2)$$

$$|Z| \geq z_{1-\alpha/2} \quad (\text{two-tailed alternative } H_A: \rho_1 \neq \rho_2)$$

To illustrate this procedure, let us test whether the data sets plotted in Figures 6.2(b) and 6.2(c) reflect populations with different correlations. In other words, we want to test $H_0: \rho_1 = \rho_2$ versus the two-sided alternative $H_A: \rho_1 \neq \rho_2$.

For the data in Figure 6.2(b), $r_1 = 0.220$; for the Figure 6.2(c) data, $r_2 = 0.342$. Using Fisher's Z transformation and Table A.5, we can calculate Z_1 and Z_2 as

$$Z_1 = \frac{1}{2} \ln \frac{1 + r_1}{1 - r_1} = \frac{1}{2} \ln \frac{1 + 0.220}{1 - 0.220} = 0.2237$$

and

$$Z_2 = \frac{1}{2} \ln \frac{1 + r_2}{1 - r_2} = \frac{1}{2} \ln \frac{1 + 0.342}{1 - 0.342} = 0.3564$$

Then the test statistic (6.13) takes the value

$$Z = \frac{0.2237 - 0.3564}{\sqrt{1/(30 - 3) + 1/(30 - 3)}} = \frac{-0.1327}{0.2722} = -0.488$$

For $\alpha = .01$, the critical region is

$$|Z| \geq z_{0.005} = 2.576$$

Since $|Z| = 0.488$ is less than 2.576, we cannot reject $H_0: \rho_1 = \rho_2$ at $\alpha = .01$.

6.7.2 Single Sample Test of $H_0: \rho_{12} = \rho_{13}$

Consider testing the null hypothesis that the correlation ρ_{12} of variable 1 with variable 2 is the same as the correlation ρ_{13} of variable 1 with variable 3. Let us assume that a single random sample of n subjects is selected and that the three sample correlations— r_{12} , r_{13} , and r_{23} —are calculated. Clearly, these sample correlations are not independent, since they are computed using the same data set. Under the usual straight-line regression analysis assumptions, it can be shown (we omit the details) that an appropriate large-sample test statistic for testing $H_0: \rho_{12} = \rho_{13}$ is

$$Z = \frac{(r_{12} - r_{13})\sqrt{n}}{\sqrt{(1 - r_{12}^2)^2 + (1 - r_{13}^2)^2 - 2r_{23}^2 - (2r_{23} - r_{12}r_{13})(1 - r_{12}^2 - r_{13}^2 - r_{23}^2)}} \quad (6.14)$$

For large n , this test statistic has approximately the standard normal distribution under $H_0: \rho_{12} = \rho_{13}$ (Olkin and Siotani 1964; Olkin 1967).

■ **Example 6.5** Assume that the weight, height, and age have been measured for each member of a sample of 12 nutritionally deficient children. Such a small sample brings into question the normal approximation involved in the use of (6.14). The data to be analyzed appear in Table 8.1 in Chapter 8. For these data, the three sample correlations are

$$r_{12} = r_{(\text{weight}, \text{height})} = 0.814$$

$$r_{13} = r_{(\text{weight}, \text{age})} = 0.770$$

$$r_{23} = r_{(\text{height}, \text{age})} = 0.614$$

We want to test whether height and age are equally correlated with weight (i.e., $H_0: \rho_{12} = \rho_{13}$) versus the two-tailed alternative that they are not (i.e., $H_A: \rho_{12} \neq \rho_{13}$). Using (6.14), the test statistic takes the value

$$\begin{aligned} Z &= \frac{(0.814 - 0.770)\sqrt{12}}{\sqrt{[1 - (0.814)^2]^2 + [1 - (0.770)^2]^2 - 2(0.614)^3}} \\ &= \frac{0.1524}{\sqrt{0.1968}} = 0.3435 \end{aligned}$$

It is clear that, for these data, we cannot reject the null hypothesis of equal correlation of weight with height and age. ■

6.8 Example: BRFSS Analysis

Returning to the BRFSS analysis of alcohol consumption frequency and BMI (Example 1.5), we illustrate the calculations of r and r^2 for the subgroup analysis of the data on 1,056 females examined in Chapter 5. As described in this chapter, there are multiple ways to compute r and r^2 , some involving the fitting of a linear regression equation and some not.

For the 1,056 complete observations, the mean drinking days \bar{X} is 6.2541 and the mean BMI \bar{Y} is 26.8388. Using these two values and the observed data X_i and Y_i , we find $\text{SS}_X = 52,476.45$, $\text{SS}_Y = 37,553.33$, and $\text{SS}_{XY} = -7,855.01$. We can then calculate r using equation (6.1):

$$r = \frac{\text{SS}_{XY}}{\sqrt{\text{SS}_X \cdot \text{SS}_Y}} = \frac{-7,855.01}{\sqrt{(52,476.45)(37,553.33)}} = -0.1769$$

The estimated correlation coefficient of -0.1769 confirms our earlier finding of a negative estimated association between BMI and drinking frequency, although the magnitude of this estimated correlation coefficient indicates that this estimated inverse relationship between BMI and alcohol consumption frequency is not a strong one. The $\hat{\beta}_1$ value of -0.150 obtained by fitting the straight-line regression model can also be computed using equation (6.2) and the value for r computed earlier:

$$\hat{\beta}_1 = \frac{r(S_Y)}{S_X} = \frac{r\sqrt{\frac{SSY}{n-1}}}{\sqrt{\frac{SSX}{n-1}}} = \frac{-0.1769\sqrt{\frac{37,553.33}{1,055}}}{\sqrt{\frac{52,476.45}{1,055}}} = -0.150$$

Knowing the value of r , we readily compute r^2 as $(0.1769)^2 = 0.0313$. Using the regression computer output in the example in Section 5.12, we could have found r^2 another way by using SSY (labeled as the sum of squares for the “corrected total”), SSE, and equation (6.6):

$$r^2 = \frac{SSY - SSE}{SSY} = \frac{37,553.33 - 36,377.54}{37,553.33} = 0.0313$$

We may interpret this value of r^2 to mean that using alcohol consumption frequency as the sole predictor of BMI in a straight-line regression model yields about a 3% reduction in SSY. In other words, 3% of the variability in BMI is explained by the linear relationship with drinking frequency.

Having computed the correlation coefficient among female drinkers, a natural follow-up question is whether the correlation is the same among men. To address this question, we consider the null hypothesis $H_0: \rho_f = \rho_m$. Among the 849 men in the sample, the estimated value of ρ_m , denoted as r_m , is -0.1559 . We can test H_0 using the Fisher Z transformation method given by equations (6.11)–(6.13):

$$\begin{aligned} Z &= \frac{Z_f - Z_m}{\sqrt{\frac{1}{n_f - 3} + \frac{1}{n_m - 3}}} = \frac{\frac{1}{2}\ln\left(\frac{1 + r_f}{1 - r_f}\right) - \frac{1}{2}\ln\left(\frac{1 + r_m}{1 - r_m}\right)}{\sqrt{\frac{1}{n_f - 3} + \frac{1}{n_m - 3}}} \\ &= \frac{\frac{1}{2}\ln\left(\frac{1 - 0.1769}{1 + 0.1769}\right) - \frac{1}{2}\ln\left(\frac{1 - 0.1559}{1 + 0.1559}\right)}{\sqrt{\frac{1}{1,056 - 3} + \frac{1}{849 - 3}}} = -0.468 \end{aligned}$$

With $\alpha = 0.05$, the two-tailed critical region is $|Z| \geq 1.96$. Thus, the test statistic is not significant, and so we do not reject the null hypothesis of no difference in the correlations of BMI with drinking frequency for men and women.

6.9 How Large Should r^2 Be in Practice?

One might naturally ask how large a value of r^2 is considered “good” for a regression model. There are no strict statistical guidelines for deciding whether or not a particular r^2 value is large enough, since such a decision will often depend on the research question under study. If, for example, the purpose of a research study is to develop a regression model used for the prediction of BMI, with the resulting model to be deployed by clinicians performing

patient assessments, then a model with $r^2 = 0.03$ would almost universally not be considered useful.⁷ For this situation, a model that explains 50% ($r^2 = 0.50$) or even 75% ($r^2 = 0.75$) of the variability in the outcome would be preferable.

Nevertheless, a complex indicator of health status such as BMI has numerous contributing factors, so that any particular factor would not be expected to add substantially to the prediction of BMI. In order to achieve a model with a large r^2 for predicting BMI, one would likely need to include a large number of predictors. Moreover, if a diverse population (such as the set of all females in Georgia) is under study in an observational research setting, a wider range of factors would likely be required to meaningfully explain variations in BMI levels for these subjects as opposed to subjects enrolled in a randomized controlled clinical trial involving strict eligibility criteria.

The occurrence of a low r^2 value is common in studies involving population-based surveillance and in other observational studies of BMI (Li et al. 2010), yet a low r^2 does not necessarily invalidate certain regression analysis conclusions. For example, we may be primarily interested in assessing the presence, magnitude, and direction of the association between two factors, rather than the prediction of the outcome (as assessed using r^2), as in the current example involving BMI and drinking frequency.

As described later, the large sample sizes found in surveillance studies often result in the detection of a significant predictor-outcome relationship despite a low r^2 value. Such large studies provide power to identify possibly important and rarely occurring predictors of an outcome that are hard to detect in small research studies. Newly found associations in such hypothesis-generating studies may then be targeted for investigation in more controlled research settings. As an important caveat, there is also the potential to find statistically significant but etiologically unimportant associations. To ensure that a statistically significant finding is truly etiologically relevant, one can assess the association by examining the corresponding estimated effect size (e.g., an estimated regression coefficient). In the BRFSS example, an average change of -0.15 unit of BMI per extra day of drinking is arguably meaningful. Also, from a policy perspective, a factor that may only explain 3% of the total variation in the outcome might translate into a large number of people affected at the population level. Intervening on a factor with a low r^2 may impact a large number of individuals and may be desirable if cost-effective.

In general, we may observe factors with a meaningful and statistically significant effect but low predictive value if *at least one* of the following study conditions is satisfied: the study sample is heterogeneous (such as encountered in population-based surveillance projects like BRFSS); the outcome is complex (i.e., has multiple, possibly interacting causes); important values of some independent variables under study are uncommon. An example of the third situation is characterized by a binary predictor that represents a rare gene X that regulates BMI. We may observe a meaningful and statistically significant relationship (i.e., a significant estimate of β) between BMI and the presence or absence of this factor. However, the estimated strength of the relationship (i.e., r^2) between X and BMI will be low simply because the gene is rarely present in the subjects under study.

We conclude this discussion of the importance of r^2 by considering the statistical circumstances that can lead to the finding of a statistically significant but weak linear association

⁷ In multiple regression (introduced in Chapter 8), we might also consider whether an additional predictor producing an increase of 0.03 in r^2 should be considered a useful addition to the regression model.

between a response Y and a predictor X . Consider the following structure of the test statistic T used to test the null hypothesis $H_0: \beta_1 = 0$, an expression that can be obtained by rewriting the two components of equation (5.9) as

$$T = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\frac{S_{Y|X}}{\sqrt{S_X^2(n-1)}}} = \frac{\hat{\beta}_1(S_X\sqrt{n-1})}{S_{Y|X}}$$

Also, consider equation (6.7), which is used to test $H_0: \rho = 0$ (or, equivalently, $H_0: \beta_1 = 0$) and which is reproduced here:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

For fixed values of all other variables in these two formulas, the presence of the sample size n in the numerator of both formulas shows that these test statistics increase in magnitude as n increases, leading to the important observation that statistically significant findings are more likely for large sample sizes. As an example, when fitting a straight-line regression model, small observed values of r^2 will often be statistically significant in studies with moderate sample sizes. For example, a study with $n = 100$ will find an r^2 of 0.04 to be statistically significant at $\alpha = .05$, and one with $n = 1,000$ will result in statistical significance for an r^2 of 0.005.⁸

Given the fact that a large sample size can lead to the finding that a small r^2 value is statistically significant, it is important to carefully consider etiologic interpretations for any statistically significant finding, as done for the BRFSS example that we are considering.

For further reading about the interpretation of low r^2 values, see Newman and Newman (2000).

Problems

1. Using the data set of Problem 1 in Chapter 5, perform the following operations.
 - a. Determine the sample correlation coefficients of (1) age with dry weight and (2) age with \log_{10} dry weight. Interpret your results.
 - b. Using Fisher's Z transformation, obtain a 95% confidence interval for ρ based on each of the correlations obtained in part (a).
 - c. For each straight-line regression, determine r^2 directly by squaring the r obtained in part (a); also determine r^2 from the computer output or from the formula $r^2 = (\text{SSY} - \text{SSE})/\text{SSY}$. Interpret your results.
 - d. Based on the preceding results, which of the two regression lines provides the better fit? Explain. Does this agree with your earlier conclusion in Problem 1(d) of Chapter 5?

⁸Similarly, the larger n is, the greater the probability will be that a small value of $\hat{\beta}$ will be found statistically significant. This relationship is explored in more depth in Chapter 27.

2. Examine the five pairs of data points given in the following table.

i	1	2	3	4	5
X_i	-2	-1	0	1	2
Y_i	4	1	0	1	4

- a. What is the mathematical relationship between X and Y ?
 - b. Show by computation that, for the straight-line regression of Y on X , $\hat{\beta}_1 = 0$.
 - c. Show by computation that $r = 0$.
 - d. Why is there apparently no relationship between X and Y , as indicated by the estimates of $\hat{\beta}_1$ and ρ ?
3. Consider the data in the following table.
- | i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-------|---|---|---|---|---|---|---|---|---|----|
| X_i | 1 | 1 | 1 | 2 | 2 | 2 | 3 | 3 | 3 | 20 |
| Y_i | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 20 |
- a. Find the sample correlation coefficient r . Interpret your result.
 - b. Show that the test statistic $T = \hat{\beta}_1/(S_{Y|X}/S_X\sqrt{n-1})$ for testing $H_0: \beta_1 = 0$ (based on a straight-line regression relationship between Y and X) is exactly equivalent to the test statistic $T' = r\sqrt{n-2}/\sqrt{1-r^2}$ for testing $H_0: \rho = 0$. [Hint: Use $\hat{\beta}_1 = rS_Y/S_X$ and $S_{Y|X}^2 = [(n-1)/(n-2)](S_Y^2 - \hat{\beta}_1^2 S_X^2)$.]
 - c. Using T' , test $H_0: \rho = 0$ versus $H_A: \rho \neq 0$.
 - d. Despite the conclusion you obtained in part (c), why should you be reluctant to conclude that the two variables are linearly related? (Hint: "A graph is worth a thousand words.")
- 4–6. Answer the following questions concerning the straight-line regressions of Y on X referred to in parts (c), (d), and (e) of Problem 2 in Chapter 5.
- a. Determine r and r^2 , and interpret your results.
 - b. Find a 99% confidence interval for ρ , and interpret your result with regard to the test of $H_0: \rho = 0$ versus $H_A: \rho \neq 0$ at $\alpha = .01$.
- 7–12. Answer the following questions for each of the data sets of Problems 3–8 in Chapter 5.
- a. Determine r and r^2 for each variable pair, and interpret your results.
 - b. Test $H_0: \rho = 0$ versus $H_A: \rho \neq 0$, and interpret your findings.
 - c. Find a 95% confidence interval for ρ . Interpret your result with regard to the test in part (b).
13. Suppose that, in a study on geographic variation in a certain species of beetle,⁹ the mean tibia length (U) and the mean tarsus length (V) were obtained for samples of size 50 from each of 10 different regions spanning five southern states. Suppose further that the results were as given in the following table.

Region	1	2	3	4	5	6	7	8	9	10
U	7.500	7.164	7.512	8.544	7.380	7.860	7.836	8.100	7.584	7.344
V	1.680	1.596	1.680	1.908	1.632	1.752	1.776	1.860	1.692	1.680

⁹ Adapted from a study by Sokal and Thomas (1965).

- a. Determine the sample correlation coefficient between mean tarsus length and mean tibia length.
 - b. Find a 99% confidence interval for ρ . Interpret your results with respect to the hypotheses $H_0: \rho = 0$ versus $H_A: \rho \neq 0$.
14. In a sample of 23 young adult men, the correlation between total hemoglobin (THb) measured from venipuncture and measured from a finger needle puncture was 0.82. For a sample of 32 women of similar age, the correlation was 0.74. The two samples from each person were collected within 1 hour of each other. Assume that the straight-line regression assumptions hold.
- a. Test the hypothesis that the two population correlations are equal. Use a two-tailed test. What do you conclude?
 - b. If the experimenter had planned to do so before collecting the data, a valid one-tailed test could have been conducted. With this assumption, repeat part (a), but use a one-tailed test to assess whether the correlation for women is lower than that for men. What do you conclude?
 - c. Assume that the researcher had planned to conduct a one-tailed test of the hypothesis that the correlation for women is higher than that for men. What test should be conducted? What do you conclude?
15. A university admissions officer regularly administers a test to all entering freshmen. A new version of the test is marketed by the testing company. To evaluate the new form, the admissions officer has 121 freshmen take both the old and the new versions. After the end of the school year, the admissions officer correlates the two scores with each other and with the students' freshman grade-point averages (GPAs). With 1 indicating the old test version, 2 the new test version, and G the GPA,

$$r_{12} = 0.6969 \quad r_{1G} = 0.5514 \quad r_{2G} = 0.4188$$

Test the hypothesis that the two forms of the test are equally correlated with GPA. Use a two-tailed test with $\alpha = .05$. Assume that the straight-line regression assumptions hold.

- 16.–26. Answer the following questions for each of the data sets of Problems 12(a), 12(c), 13, 14, 15(a), 15(c), 16, 17, 18, 19, and 20 in Chapter 5.
- a. Determine r and r^2 for each variable pair, and interpret your results.
 - b. Test $H_0: \rho = 0$ versus $H_A: \rho \neq 0$ and interpret your findings.
 - c. Find a 95% confidence interval for ρ . Interpret your result with respect to the test in part (b).

References

- Li, S.; Zhao, J. H.; Luan, J.; Ekelund, U.; Luben, R. N.; Khaw, K.; Wareham, N. J.; and Loos, R. J. F. 2010. "Physical Activity Attenuates the Genetic Predisposition to Obesity in 20,000 Men and Women from EPIC-Norfolk Prospective Population Study." *PLoS Medicine* 7(8): e1000332. doi:10.1371/journal.pmed.1000332
- Newman, I., and Newman, C. 2000. "A Discussion of Low r-Squares: Concerns and Uses." *Educational Research Quarterly* 24(2): 3–9.

- Olkin, I. 1967. "Correlation Revisited." In Julian C. Stanley, ed., *Improving Experimental Design and Statistical Analysis*. Chicago: Rand McNally.
- Olkin, I., and Siotani, M. 1964. *Asymptotic Distribution Functions of a Correlation Matrix*. (Report No. 6). Stanford. Calif.: Stanford University Laboratory for Quantitative Research in Education.
- Sokal, R. R., and Thomas, P. A. 1965. "Geographic Variation of *Pemphigus populitransversus* in Eastern North America: Stem Mothers and New Data on Alates." *University of Kansas Scientific Bulletin* 46: 201–52.

7

The Analysis-of-Variance Table

7.1 Preview

An overall summary of the results of any regression analysis, whether straight-line or not, can be provided by a table called an *analysis-of-variance (ANOVA) table*. This name derives primarily from the fact that the basic information in an ANOVA table consists of several estimates of variance. These estimates, in turn, can be used to answer the principal inferential questions of regression analysis. In the straight-line case, there are three such questions: (1) Is the true slope β_1 zero? (2) What is the strength of the straight-line relationship? (3) Is the straight-line model appropriate?

Historically, the name “analysis-of-variance table” was coined to describe the overall summary table for the statistical procedure known as *analysis of variance*. As we observed in Chapter 2 and will see later when discussing the ANOVA method, regression analysis and analysis of variance are closely related. More precisely, analysis-of-variance problems can be expressed in a regression framework. Thus, such a table can be used to summarize the results obtained from either method.

7.2 The ANOVA Table for Straight-line Regression

Various textbooks, researchers, and computer program printouts have slightly different ways of presenting the ANOVA table associated with straight-line regression analysis. This section describes the most common form.

The simplest version of the ANOVA table for straight-line regression is given in the accompanying SAS computer output, as applied to the age–systolic blood pressure data of Table 5.1. Within each row, the mean-square term is obtained by dividing the sum of

squares by its degrees of freedom. The F statistic (i.e., F value in the output) is obtained by dividing the regression (i.e., model) mean square by the residual (i.e., error) mean square.

Edited SAS Output (PROC REG) for an ANOVA Table Based on Table 5.1 Data

Regression of Sbp (Y) on Age (X)

(SSY – SSE) ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model (Regression)	1	6394.02269	6394.02269	21.33	<.0001
Error	28	8393.44398	299.76586		
Corrected Total	29	14787			

© Cengage Learning

In Chapter 6, when describing the correlation coefficient, we observed in (6.6) that

$$r^2 = \frac{SSY - SSE}{SSY}$$

where $SSY = \sum_{i=1}^n (Y_i - \bar{Y})^2$ is the sum of the squares of deviations of the observed Y 's from the mean \bar{Y} and $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ is the sum of the squares of deviations of the observed Y 's from the fitted regression line. Since SSY represents the total variation of Y before accounting for the linear effect of the variable X , we usually call SSY the *total unexplained variation* or the *total sum of squares about (or corrected for) the mean*. Because SSE measures the amount of variation in the observed Y 's that remains after accounting for the linear effect of the variable X , we refer to the quantity $(SSY - SSE)$ as the variation that is due to regression. This is also commonly called the *sum of squares due to (or explained by) regression* or the *regression sum of squares*. It turns out that $(SSY - SSE)$ is mathematically equivalent to the expression

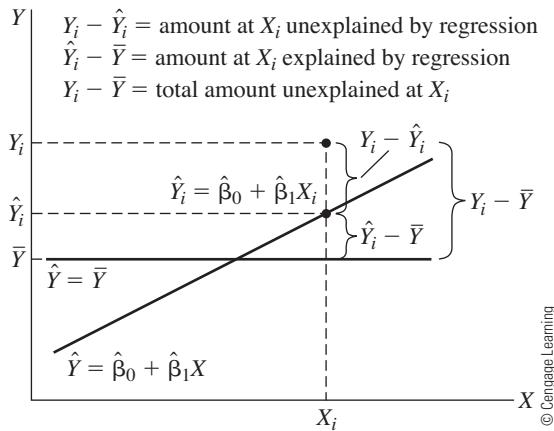
$$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

which represents the sum of squares of deviations of the predicted values from the mean \bar{Y} . We thus have the following mathematical result:

$$\begin{aligned} \text{Total unexplained variation} &= \text{Variation due to regression} \\ &\quad + \text{Residual variation after regression} \end{aligned}$$

or

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ \text{Total sum of} &\quad \text{Regression sum} \quad \text{Residual sum of} \\ \text{squares} &\quad \text{of squares} \quad \text{squares} \end{aligned} \tag{7.1}$$



© Cengage Learning

FIGURE 7.1 Variation explained and unexplained by straight-line regression

Equation (7.1), which is often called the *fundamental equation of regression analysis*, holds for any general regression situation. Figure 7.1 illustrates this equation.

The mean-square error value of 299.77 is simply the estimate $S^2_{Y|X}$ presented earlier. If the true regression model is a straight line, then, as mentioned in Section 5.6, $S^2_{Y|X}$ is an estimate of σ^2 . On the other hand, the mean-square regression value of 6,394.02 provides an estimate of σ^2 only if the variable X does not help to predict the dependent variable Y —that is, only if the hypothesis $H_0: \beta_1 = 0$ is true. If, in fact, $\beta_1 \neq 0$, the mean-square regression term will be inflated in proportion to the magnitude of β_1 and will correspondingly overestimate σ^2 .

It can be shown that the mean-square residual and mean-square regression terms are *statistically independent* of one another. Thus, if $H_0: \beta_1 = 0$ is true, the ratio of these terms represents the ratio of two independent estimates of the same variance σ^2 . Under the normality and independence assumptions about the Y 's, such a ratio has the F distribution, and this F statistic (with the value 21.33 in the accompanying SAS computer output) can be used to test the hypothesis H_0 : “No significant straight-line relationship of Y on X ” (i.e., $H_0: \beta_1 = 0$ or $H_0: \rho = 0$).

Fortunately, this way of testing H_0 is *equivalent* to using the two-sided t test previously discussed. This is so because, for v degrees of freedom,

$$F_{1, v} = T_v^2 \quad (7.2)$$

so

$$F_{1, v, 1-\alpha} = t_{v, 1-\alpha/2}^2 \quad (7.3)$$

The expression in (7.3) states that the $100(1 - \alpha)\%$ point of the F distribution with 1 and v degrees of freedom is exactly the same as the square of the $100(1 - \alpha/2)\%$ point of the t distribution with v degrees of freedom.

To illustrate the equivalence of the F and t tests, we can see from our age–systolic blood pressure example that $F = 21.33$ and $T^2 = 4.62^2 = 21.33$, where 4.62 is the figure obtained for T at the end of Section 6.6.1. Also, it can be seen that $F_{1, 28, 0.95} = 4.20$ and that $t_{28, 0.975}^2 = (2.05)^2 = 4.20$.

As these equalities establish, the critical region

$$|T| > t_{28, 0.975} = 2.05$$

for testing $H_0: \beta_1 = 0$ against the two-sided alternative $H_A: \beta_1 \neq 0$ is exactly the same as the critical region

$$F > F_{1, 28, 0.95} = 4.20$$

Hence, if $|T|$ exceeds 2.05, then F will exceed 4.20. Similarly, if F exceeds 4.20, then $|T|$ will exceed 2.05. Thus, the null hypothesis $H_0: \beta_1 = 0$ (or, equivalently, H_0 : “No significant straight-line relationship of Y on X ”) is rejected at the $\alpha = .05$ level of significance.

An alternative but less common representation of the ANOVA table is given in Table 7.1. This table differs from the SAS output table only in that it splits up the total sum of squares corrected for the mean, SSY, into its two components: the *total uncorrected sum of squares*, $\sum_{i=1}^n Y_i^2$; and the *correction factor*, $(\sum_{i=1}^n Y_i)^2/n$. The relationship between these components is given by the equation

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n}$$

In the total (uncorrected) sum of squares $\sum_{i=1}^n Y_i^2$, the n observations on Y are considered before any estimation of the population mean of Y . The “Regression \bar{Y} ” listed in Table 7.1 refers to the variability explained by using a model involving only β_0 (which is estimated by \bar{Y}). This is necessarily the same amount of variability as is explained by using only \bar{Y} to predict Y , without attempting to account for the linear contribution of X to the prediction of Y . The “Regression $X|\bar{Y}$ ” describes the contribution of the variable X to predicting Y over and above that contributed by \bar{Y} alone. Usually, “Regression $X|\bar{Y}$ ” is written simply as

TABLE 7.1 Alternative ANOVA table for age–systolic blood pressure data of Table 5.1

Source	Degrees of Freedom (d.f.)	Sum of Squares (SS)	Mean Square (MS)	Variance Ratio (F)
Regression	1	$\frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} = 609,472.53$		
	1	6,394.02	6,394.02	21.33 ($P < .001$)
Residual	28	8,393.44	299.77	
Total	30	$\sum_{i=1}^n Y_i^2 = 624,260.00$		

“Regression X ,” the “given \bar{Y} ” part being suppressed for notational simplicity. We will see more of this notation when we discuss multiple regression in subsequent chapters.

Problems

1. Use the data set of Problem 1 in Chapter 5 to answer the following questions.
 - a. Determine the ANOVA tables for the following regressions: (1) dry weight (Y) on age (X) and (2) \log_{10} dry weight (Z) on age (X). The following results will be helpful in reducing computation time:

$$S_{Y|X}^2 = 0.23218 \quad SSY = 8.16811 \quad S_{Z|X}^2 = 0.0007838 \quad SSZ = 4.2276839$$

- b. Use the tables in part (a) to perform the F test for the significance of each straight-line regression. Interpret your results.
- 2.-4. Answer the same questions as in parts (a) and (b) of Problem 1 for each regression of Y on X , using the data in parts (b), (c), and (d) of Problem 2 in Chapter 5. The following results will be helpful in reducing computation time:

$$\text{SBP } (Y) \text{ regressed on QUET } (X): \quad S_{Y|X}^2 = 96.26743 \quad SSY = 6,425.96875$$

$$\text{QUET } (Y) \text{ regressed on AGE } (X): \quad S_{Y|X}^2 = 0.09079 \quad SSY = 7.65968$$

$$\text{SBP } (Y) \text{ regressed on AGE } (X): \quad S_{Y|X}^2 = 85.47795 \quad SSY = 6,425.96875$$

5. Use the data of Problem 3 in Chapter 5 to answer the following questions.
 - a.-b. Answer the same questions as in parts (a) and (b) of Problem 1 for the regression of TIME (Y) on INC (X). The following results will be helpful:

$$S_{Y|X}^2 = 110.16190 \quad SSY = 2,433.78137$$

- c. Compare the value of the test statistic F obtained in part (b) with the value of T^2 , the square of the test statistic for testing $H_0: \beta_1 = 0$ that was required in part (g) of Problem 3 in Chapter 5.
 - d. The P -values for the F test in part (b) and for the t test in part (g) of Problem 3 in Chapter 5 are the same. Intuitively, why does this make sense? (Hint: Compare the hypotheses for each of the tests.)

- 6.-10. Answer the same questions as in parts (a) and (b) of Problem 1 for each of the regressions of Y on X in Problems 5 through 8 and Problem 10 of Chapter 5. The following results will be helpful:

	$S_{Y X}^2$	SSY
Chapter 5, Problem 5:	11.101	2,223.018
Chapter 5, Problem 6:	172.985	20,123.382
Chapter 5, Problem 7 (Y1):	1,001.691	190,502.800
Chapter 5, Problem 7 (Y2):	0.658	410.531
Chapter 5, Problem 8:	1,264,983.805	112,278,032.670
Chapter 5, Problem 10:	0.003	2.142

11. A biologist wanted to study the effects of the temperature of a certain medium on the growth of human amniotic cells in a tissue culture. Using the same parent batch, she conducted an experiment in which five cell lines were cultured at each of four temperatures. The procedure involved initially inoculating a fixed number (0.25 million) of cells into a fresh culture flask and then, after 7 days, removing a small sample from the growing surface to use in estimating the total number of cells in the flask. The results are given in the following table, together with a computer printout for straight-line regression.

Number of Cells ($\times 10^{-6}$) after 7 days (Y)	Temperature (X)	Number of Cells ($\times 10^{-6}$) after 7 days (Y)	Temperature (X)
1.13	40	2.30	80
1.20	40	2.15	80
1.00	40	2.25	80
0.91	40	2.40	80
1.05	40	2.49	80
1.75	60	3.18	100
1.45	60	3.10	100
1.55	60	3.28	100
1.64	60	3.35	100
1.60	60	3.12	100

- Complete the ANOVA table shown in the accompanying computer output.
- Perform the F test for the significance of the straight-line regression of Y on X , and interpret your results.

Edited SAS Output (PROC REG) for Problem 11

Regression of # of Cells (Y) on Temperature (X)

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1				<.0001
Error	18	0.36618			
Corrected Total	19	13.19690			

Root MSE	0.14263	R-Square	0.9723
Dependent Mean	2.04500	Adj R-Sq	0.9707
Coeff Var	6.97454		

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-0.46240	0.10481	-4.41	0.0003
X	1	0.03582	0.00143	25.11	<.0001

12–21. Answer the same questions as in parts (a) and (b) of Problem 1 for each of the data sets in Problems 12(a), 12(c), 13, 14, 15(a), 15(c), 16, 17, 18, and 19 in Chapter 5. The following results will be useful:

	$S^2_{Y X}$	SSY
Chapter 5, Problem 12(a):	21,924.992	12,699,346.400
Chapter 5, Problem 12(c):	42,398.926	13,439,939.000
Chapter 5, Problem 13:	0.530	6.704
Chapter 5, Problem 14:	384.701	3,770.900
Chapter 5, Problem 15(a):	8.140	9,386.654
Chapter 5, Problem 15(c):	0.058	180.911
Chapter 5, Problem 16:	114.698	6,305.714
Chapter 5, Problem 17:	0.689	28.833
Chapter 5, Problem 18:	13.201	610.429
Chapter 5, Problem 19:	19.514	600.962

8

Multiple Regression Analysis: General Considerations

8.1 Preview

Multiple regression analysis can be looked upon as an extension of straight-line regression analysis (which involves only one independent variable) to the situation in which more than one independent variable must be considered. Several general applications of multiple regression analysis¹ were described in Chapter 4, and specific examples were given in Chapter 1. In this chapter, we will describe the multiple regression method in detail, stating the required assumptions, describing the procedures for estimating important parameters, explaining how to make and interpret inferences about these parameters, and providing examples that illustrate how to use the techniques of multiple regression analysis. Dealing with several independent variables simultaneously in a regression analysis is considerably more difficult than dealing with a single independent variable for the following reasons:

1. It is more difficult to choose the best model, since several reasonable candidates may exist.
2. It is more difficult to visualize what the fitted model looks like (especially if there are more than two independent variables), since it is not possible to plot either the data or the fitted model directly in more than three dimensions.
3. It is sometimes more difficult to interpret what the best-fitting model means in real-life terms.
4. Computations are virtually impossible without access to a high-speed computer and a reliable packaged computer program.

¹ We shall generally refer to multiple regression analysis simply as “regression analysis” throughout the remainder of the text.

8.2 Multiple Regression Models

One example of a multiple regression model is given by any second- or higher-order polynomial. Adding higher-order terms (e.g., an X^2 or X^3 term) to a model can be considered as equivalent to adding new independent variables. Thus, if we rename X as X_1 and X^2 as X_2 , the second-order model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + E$$

can be rewritten as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$$

Of course, in polynomial regression we have only one basic independent variable, the others being simple mathematical functions of this basic variable. In more general multiple regression problems, however, the number of basic independent variables may be greater than one. The general form of a regression model for k independent variables is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + E$$

where $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ are the *regression coefficients* that need to be estimated. The *independent* variables X_1, X_2, \dots, X_k may all be separate basic variables, or some may be functions of a few basic variables.

■ Example 8.1 Suppose that we want to investigate how weight (WGT) varies with height (HGT) and age (AGE) for children with a particular kind of nutritional deficiency.² The dependent variable here is $Y = \text{WGT}$, and our two basic independent variables are $X_1 = \text{HGT}$ and $X_2 = \text{AGE}$.

Suppose that, as outlined in Example 6.5, a random sample consists of 12 children who attend a certain clinic. The WGT, HGT, and AGE data obtained for each child are given in Table 8.1.

TABLE 8.1 WGT, HGT, and AGE of a random sample of 12 nutritionally deficient children

Child	1	2	3	4	5	6	7	8	9	10	11	12
WGT (Y)	64	71	53	67	55	58	77	57	56	51	76	68
HGT (X_1)	57	59	49	62	51	50	55	48	42	42	61	57
AGE (X_2)	8	10	6	11	8	7	10	9	10	6	12	9

© Cengage Learning

² Perhaps the main question associated with this type of study is whether the relationship for nutritionally deficient children is the same as that for "normal" children. To answer this question would require additional data on normal children and some kind of comparison of the models obtained for each group. Although we will learn how to deal with this kind of question in Chapter 12, we focus here on the methods needed to describe the relationship of weight to height and age for this single group of nutritionally deficient children.

In describing the relationship of WGT to HGT and AGE, we may want to consider the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$$

if we are interested only in first-order terms. If we want to consider, in addition, the higher-order term X_1^2 , our model is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$$

where $X_3 = X_1^2$. To consider all possible first- and second-order terms, we look at the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + E$$

where $X_3 = X_1^2$, $X_4 = X_2^2$, and $X_5 = X_1 X_2$ or, equivalently,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + E$$

If we want to find the best predictive model, we might consider all of the preceding models (as well as some others) and then choose the best model according to some reasonable criterion.

We discuss the question of model selection in Chapter 16; the interpretation of product terms such as $X_1 X_2$ as interaction effects is explained in Chapter 11. For now, our focus is on the methods used and the interpretations that can be made when the choice of independent variables to use in the model is not at issue. ■

8.3 Graphical Look at the Problem

When we are dealing with only one basic independent variable, our problem can easily be described graphically as that of finding the curve that best fits the scatter of points (X_1, Y_1) , (X_2, Y_2) , ..., (X_n, Y_n) obtained on n individuals. Thus, we have a *two-dimensional* representation involving a plot of the form shown in Figure 8.1. Furthermore, the *regression equation* for this problem is defined as the path described by the mean values of the distribution of Y when X is allowed to vary.

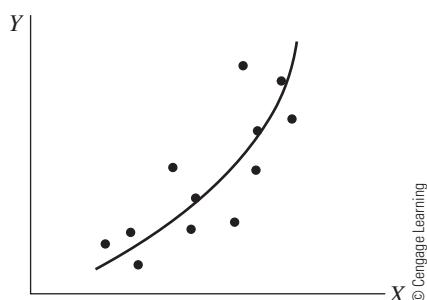
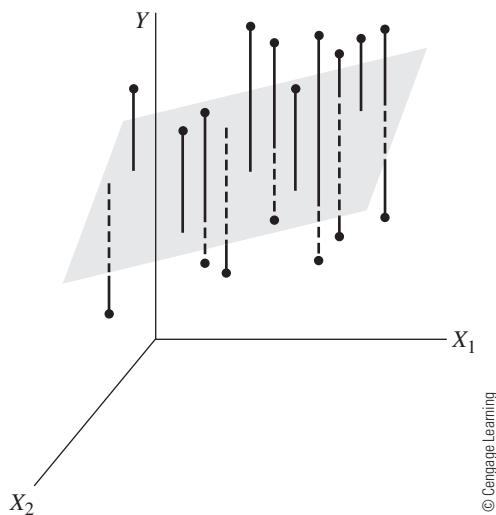


FIGURE 8.1 Scatterplot for a single independent variable

When the number k of (basic) independent variables is two or more, the (graphical) dimension of the problem increases. The regression equation ceases to be a curve in two-dimensional space and becomes instead a *hypersurface in $(k + 1)$ -dimensional space*. Obviously, we will not be able to represent in a single plot either the scatter of data points or the regression equation if more than two basic independent variables are involved. In the special case $k = 2$, as in the example just given, where $X_1 = \text{HGT}$, $X_2 = \text{AGE}$, and $Y = \text{WGT}$, the problem is to find the *surface* in three-dimensional space that best fits the scatter of points $(X_{11}, X_{21}, Y_1), (X_{21}, X_{22}, Y_2), \dots, (X_{n1}, X_{n2}, Y_n)$, where (X_{i1}, X_{i2}, Y_i) denotes the X_1 -, X_2 -, and Y -values for the i th individual in the sample. The *regression equation* in this case, therefore, is the surface described by the mean values of Y at various combinations of values of X_1 and X_2 ; that is, corresponding to each distinct pair of values of X_1 and X_2 is a distribution of Y values with mean $\mu_{Y|X_1, X_2}$ and variance $\sigma_{Y|X_1, X_2}^2$.

Just as the simplest curve in two-dimensional space is a straight line, the simplest surface in three-dimensional space is a *plane*, which has the statistical model form $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$. Thus, finding the best-fitting plane is frequently the first step in determining the best-fitting surface in three-dimensional space when two independent variables are relevant, just as fitting the best straight line is the first step when one independent variable is involved. A graphical representation of a planar fit to data in the three-dimensional situation is given in Figure 8.2.



© Cengage Learning

FIGURE 8.2 Best-fitting plane for three-dimensional data

For the three-dimensional case, the least-squares solution that gives the best-fitting plane is determined by minimizing the sum of squares of the distances between the observed values Y_i and the corresponding predicted values $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2}$, based on the fitted plane. In other words, the quantity

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2})^2$$

is minimized to find the least-squares estimates $\hat{\beta}_0$ of β_0 , $\hat{\beta}_1$ of β_1 , and $\hat{\beta}_2$ of β_2 .

How much can one learn by considering the independent variables in the multivariable problem separately? Probably the best answer is that we can learn something about what is going on, but there are too many separate (univariable) pieces of information to permit us to complete the (multivariable) puzzle. For example, consider the data previously given for $Y = \text{WGT}$, $X_1 = \text{HGT}$, and $X_2 = \text{AGE}$. If we plot separate scatter diagrams of WGT on HGT, WGT on AGE, and AGE on HGT, we get the results shown in Figure 8.3.

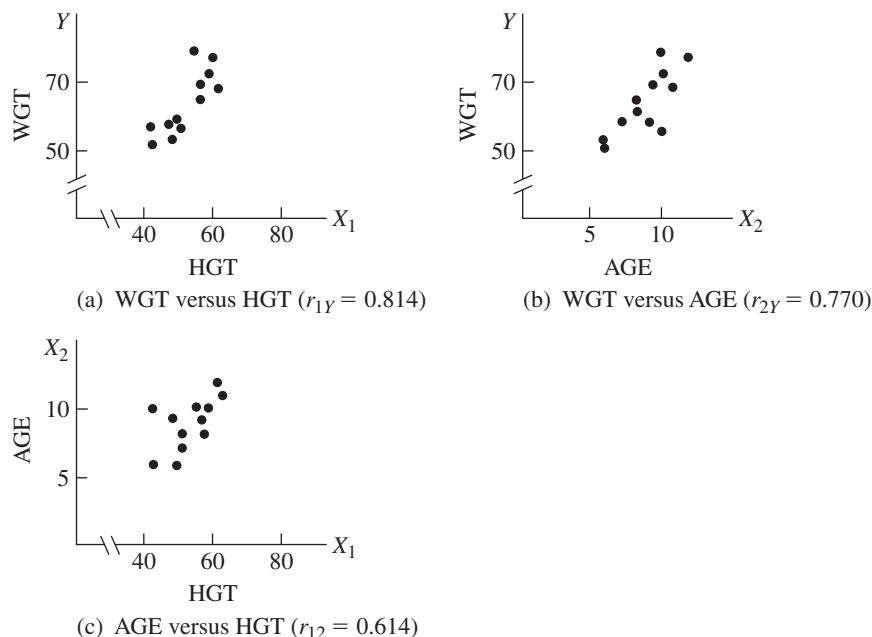


FIGURE 8.3 Separate scatter diagrams of WGT versus HGT, WGT versus AGE, and AGE versus HGT

© Cengage Learning

HGT is highly positively correlated with WGT ($r_{1Y} = 0.814$), as is AGE ($r_{2Y} = 0.770$). Thus, if we used each of these independent variables separately, we would likely find two separate, significant straight-line regressions. Does this mean that the best-fitting plane with both variables in the model together will also have significant predictive ability? The answer is probably yes. But what will the plane look like? This is difficult to say. We can get some idea of the difficulty if we consider the plot of HGT versus AGE in part (c), which reflects a positive correlation ($r_{12} = 0.614$). If, instead, these two variables were negatively correlated, we would expect a different orientation of the plane, although we could not clearly quantify either orientation. Thus, treating each independent variable separately does not help very much because the relationships between the independent variables themselves are not taken directly into account. The techniques of multiple regression, however, account for all these intercorrelations with regard to both estimation and inference making.

8.4 Assumptions of Multiple Regression

In the previous section, we described the multiple regression problem in some generality and also hinted at some of the assumptions involved. We now state these assumptions somewhat more formally.

8.4.1 Statement of Assumptions

Assumption 1: Existence For each specific combination of values of the (basic) independent variables X_1, X_2, \dots, X_k (e.g., $X_1 = 57, X_2 = 8$ for the first child in Example 8.1), Y is a (univariate) random variable with a certain probability distribution having finite mean and variance.

Assumption 2: Independence The Y observations are statistically independent of one another. As with straight-line regression, this assumption is usually violated when several Y observations are made on the same subject. Methods for dealing with regression modeling of correlated data include *generalized estimating equations* (GEE) techniques (Zeger and Liang 1986; Diggle et al. 2002) and *mixed model* techniques (described in Chapters 25 and 26) using SAS's MIXED procedure (SAS release 9.3, SAS Institute 2011).

Assumption 3: Linearity The mean value of Y for each specific combination of X_1, X_2, \dots, X_k is a linear function³ of $\beta_0, \beta_1, \dots, \beta_k$. That is,

$$\mu_{Y|X_1, X_2, \dots, X_k} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (8.1)$$

or

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + E \quad (8.2)$$

where E is the error component reflecting the difference between an individual's observed response Y and the true average response $\mu_{Y|X_1, X_2, \dots, X_k}$. Some comments are in order regarding Assumption 3:

1. The surface described by (8.1) is called the *regression equation* (or *response surface* or *regression surface*).
2. If some of the independent variables are higher-order functions of a few basic independent variables (e.g., $X_3 = X_1^2, X_5 = X_1 X_2$), the expression

³The techniques of multiple regression that we will be describing are applicable as long as the model under consideration is *inherently linear* in the regression coefficients (regardless of how the independent variables are defined). For example, a model of the form $\mu_{Y|X} = \beta_0 e^{(\beta_1 X + \beta_2 X^2)}$ is inherently linear because it can be transformed into the equivalent form $\mu_{Y|X}^* = \beta_0^* + \beta_1 X + \beta_2 X^2$, where $\mu_{Y|X}^* = \ln \mu_{Y|X}$ and $\beta_0^* = \ln \beta_0$. However, the model $\mu_{Y|X_1, X_2} = e^{\beta_1 X_1} + e^{\beta_2 X_2}$ cannot be transformed directly into a form that is linear in β_1 and β_2 ; thus, estimating β_1 and β_2 requires the use of *nonlinear regression* procedures (see, e.g., Gallant 1975). A discussion of these procedures is beyond the scope of this text.

$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$ is nonlinear in the basic variables (we use the word *surface* rather than *plane*, because curvature is now present).

3. Consonant with its meaning in straight-line regression, E is the amount by which any individual's observed response deviates from the response surface. Thus, E is the *error component* in the model.

Assumption 4: Homoscedasticity *The variance of Y is the same for any fixed combination of X_1, X_2, \dots, X_k . That is,*

$$\sigma_{Y|X_1, X_2, \dots, X_k}^2 = \text{Var}(Y|X_1, X_2, \dots, X_k) \equiv \sigma^2 \quad (8.3)$$

As before, this is called the assumption of homoscedasticity. An alternative (but equivalent) definition of homoscedasticity, based on (8.2), is that

$$\sigma_{E|X_1, X_2, \dots, X_k}^2 \equiv \sigma^2$$

This assumption may seem very restrictive. But variance heteroscedasticity needs to be considered only when the data show very obvious and significant departures from homogeneity. In general, mild departures do not have significant adverse effects on the results.

Assumption 5: Normality *For any fixed combination of X_1, X_2, \dots, X_k , the variable Y is normally distributed.* In other words,

$$Y \sim N(\mu_{Y|X_1, X_2, \dots, X_k}, \sigma^2)$$

or, equivalently,

$$E \sim N(0, \sigma^2) \quad (8.4)$$

This assumption is not necessary for the least-squares fitting of the regression model, but it is required in general for inference-making. The usual parametric tests of hypotheses and confidence intervals used in a regression analysis are robust in the sense that only extreme departures of the distribution of Y from normality yield spurious results. (This statement is based on both theoretical and experimental evidence.) If the normality assumption does not hold, one typically seeks a transformation of Y —say, $\log Y$ or \sqrt{Y} —to produce a transformed set of Y observations that are approximately normal (see Section 14.4.1). If the Y variable is either categorical or ordinal, however, alternative regression methods such as logistic regression (for binary Y 's) or Poisson regression (for discrete Y 's) are typically required (see Chapters 22 and 24).

8.4.2 Summary and Comments

Our assumptions for simple linear (i.e., straight-line) regression analysis can be generalized to multiple linear regression analysis. Here homoscedasticity and normality apply to $Y|X_1, X_2, \dots, X_k$ rather than to Y (i.e., to the conditional distribution of Y given X_1, X_2, \dots, X_k rather than to the so-called unconditional or marginal distribution of Y).

The assumptions for multiple linear regression analysis dictate that the random error component E have a normal distribution with mean 0 and variance σ^2 . Of course, the linearity, existence, and independence assumptions must also hold.

Again, Y is an observable random variable, while X_1, X_2, \dots, X_k are assumed to be measured without error. The constants $\beta_0, \beta_1, \dots, \beta_k$ are unknown population parameters, and E is an unobservable random variable. If one estimates $\beta_0, \beta_1, \dots, \beta_k$ with $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$, then an acceptable estimate of E_i for the i th subject is

$$\hat{E}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \dots + \hat{\beta}_k X_{ik})$$

The estimated error \hat{E} is usually called a *residual*.

The assumption of a normal (i.e., Gaussian) distribution is needed to justify the use of procedures of statistical inference involving the t and F distributions.

8.5 Determining the Best Estimate of the Multiple Regression Equation

As with straight-line regression, there are two basic approaches to estimating a multiple regression equation: the least-squares approach and the minimum-variance approach. In the straight-line case, both approaches yield the same solution. (We are assuming, as previously noted, that we already know the best form of regression model to use; that is, we have already settled on a fixed set of k independent variables X_1, X_2, \dots, X_k . The problem of determining the best model form via algorithms for choosing the most important independent variables will be discussed in detail in Chapter 16.) The multiple regression model may also be fitted by using other statistical methodology, such as maximum likelihood (see Chapter 21). Given fixed X 's and assuming Y_i is Gaussian with the Y_i 's mutually independent, then the least-squares estimates of the regression coefficients are identical to the maximum-likelihood estimates.

8.5.1 Least-Squares Approach

In general, the least-squares method chooses as the best-fitting model the one that minimizes the sum of squares of the distances between the observed responses and those predicted by the fitted model. Again, the better the fit, the smaller the deviations of observed from predicted values. Thus, if we let

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k$$

denote the fitted regression model, the sum of squares of deviations of observed Y -values from corresponding values predicted by using the fitted regression model is given by

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \dots - \hat{\beta}_k X_{ik})^2 \quad (8.5)$$

The least-squares solution then consists of the values $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ (called the “least-squares estimates”) for which the sum in (8.5) is a minimum. This minimum sum of squares is generally called the *residual sum of squares* (or, equivalently, the *error sum of squares* or the *sum of squares about regression*); as in the case of straight-line regression, it is referred to as the SSE.

8.5.2 Minimum-Variance Approach

As in the straight-line case, the minimum-variance approach to estimating the multiple regression equation identifies as the best-fitting surface the one utilizing the minimum-variance (linear) unbiased estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ of $\beta_0, \beta_1, \dots, \beta_k$, respectively.

8.5.3 Comments on the Least-Squares Solutions

In this text, we do not present matrix formulas for calculating the least-squares estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$, since computer programs are readily available to perform the necessary calculations. Even so, we provide in Appendix B a discussion of matrices and their use in regression analysis; by using matrix mathematics, one can represent the general regression model and the associated least-squares methodology in compact form. Also, an understanding of the matrix formulation for regression analysis carries over to more complex modeling problems, such as those involving multivariate data (i.e., data relating to two or more dependent variables).

The least-squares solutions have several important properties:

1. Each of the estimators $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ is a linear function of the Y -values. This linearity property makes determining the statistical properties of these estimators fairly straightforward. In particular, since the Y -values are assumed to be normally distributed and to be statistically independent of one another, each of the estimators $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ will be normally distributed, with easily computable standard errors.
2. The least-squares regression equation $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k$ is the unique linear combination of the independent variables X_1, X_2, \dots, X_k that has maximum possible correlation with the dependent variable. In other words, of all possible linear combinations of the form $b_0 + b_1 X_1 + b_2 X_2 + \dots + b_k X_k$, the linear combination \hat{Y} is such that the correlation

$$r_{Y, \hat{Y}} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2 \sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})^2}} \quad (8.6)$$

is a maximum, where \hat{Y}_i is the predicted value of Y for the i th individual and $\bar{\hat{Y}}$ is the mean of the \hat{Y}_i 's. Incidentally, it is always true that $\bar{\hat{Y}} = \bar{Y}$; that is, the mean of the predicted values is equal to the mean of the observed values. The quantity $r_{Y, \hat{Y}}$ is called the *multiple correlation coefficient*.

3. Just as straight-line regression is related to the bivariate normal distribution, multiple regression can be related to the multivariate normal distribution. We will return to this point in Section 10.4 of Chapter 10.

■ **Example 8.2** For the data given in Table 8.1 on the variables $Y = \text{WGT}$, $X_1 = \text{HGT}$, and $X_2 = \text{AGE}$, the least-squares algorithm applied to the model

$$\text{WGT} = \beta_0 + \beta_1 \text{HGT} + \beta_2 \text{AGE} + \beta_3 (\text{AGE})^2 + E$$

produces the estimated equation

$$\text{WGT} = 3.438 + 0.724 \text{HGT} + 2.777 \text{AGE} - 0.042(\text{AGE})^2$$

$$\text{so } \hat{\beta}_0 = 3.438, \quad \hat{\beta}_1 = 0.724, \quad \hat{\beta}_2 = 2.777, \text{ and } \hat{\beta}_3 = -0.042. \blacksquare$$

8.6 The ANOVA Table for Multiple Regression

As with straight-line regression, an ANOVA table can be used to provide an overall summary of a multiple regression analysis. The particular form of an ANOVA table may vary, depending on how the contributions of the independent variables are to be considered (e.g., individually or collectively in some fashion). A simple form reflects the contribution that all independent variables considered collectively make to prediction. For example, consider Table 8.2, an ANOVA table based on the use of HGT, AGE, and $(\text{AGE})^2$ as independent variables for the data of Table 8.1.

As before, the term $\text{SSY} = \sum_{i=1}^n (Y_i - \bar{Y})^2 = 888.25$ is called the *total sum of squares* (corrected for the mean), and this figure represents the total variability in the Y observations before accounting for the joint effect of using the independent variables HGT, AGE, and $(\text{AGE})^2$. The term $\text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 195.19$ is the *residual sum of squares* (or the *sum of squares due to error*), which represents the amount of Y variation left unexplained after the independent variables have been used in the regression equation to predict Y . Finally, $\text{SSY} - \text{SSE} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = 693.06$ is called the *regression sum of squares* and measures

TABLE 8.2 ANOVA table for WGT regressed on HGT, AGE, and $(\text{AGE})^2$

Source	d.f.	SS	MS	F	R ²
Regression	$k = 3$	$\text{SSY} - \text{SSE} = 693.06$	231.02	9.47**	0.7802
Residual	$n - k - 1 = 8$	SSE = 195.19	24.40		
Total	$n - 1 = 11$	SSY = 888.25			

Note: The ** next to the computed F denotes significance at the .01 level; i.e., $P < .01$.

the reduction in variation (or the variation explained) due to the independent variables in the regression equation. We thus have the familiar partition

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Total sum of squares *Regression sum of squares* *Residual sum of squares*

In Table 8.2, as in ANOVA tables for straight-line regression, the SS column identifies the various sums of squares. The d.f. column gives the corresponding degrees of freedom: the regression degrees of freedom is k (the number of independent variables in the model); the residual degrees of freedom is $n - k - 1$; and the total degrees of freedom is $n - 1$. The MS column contains the mean-square terms, obtained by dividing the sum-of-squares terms by their corresponding degrees-of-freedom values. The F ratio is obtained by dividing the mean-square regression by the mean-square residual; the interpretation of this F ratio will be discussed in Chapter 9 on hypothesis testing.

The R^2 in Table 8.2 (with the value 0.7802) provides a quantitative measure of how well the fitted model containing the variables HGT, AGE, and $(AGE)^2$ predicts the dependent variable WGT. The computational formula for R^2 is

$$R^2 = \frac{\text{SSY} - \text{SSE}}{\text{SSY}} \quad (8.7)$$

This formula is essentially the same one provided for r^2 in straight-line regression (Section 6.4); however, R^2 is a more general measure that is used for multiple linear regression models. Like r^2 , the quantity R^2 lies between 0 and 1. If the value is 1, we say that the fit of the model is perfect. R^2 always increases as more variables are added to the model, but a very small increase in R^2 may be neither practically nor statistically important. Additional properties of R^2 are discussed in Chapter 10.

8.7 Example: BRFSS Analysis

Returning to the BRFSS Example 1.4, a multiple regression model was next fit that included two additional continuous variables thought to predict BMI. The first is age, in whole years, and the other is sleep quality, defined as the number of days of insufficiently restful sleep in the past month (range 0–30). For this analysis 1,049 observations with no missing values are considered. Based on the information given below, we wish to

- Find the estimated multivariable regression equation. (Section 8.5)
- Use this equation to predict \hat{Y} for females who drink 10 days per month, are 30 years old, and have 5 days of poor sleep in the past month. (Section 8.5)
- Complete the ANOVA table. (Section 8.6)
- Calculate R^2 for this model. (Section 8.6)

Parameter	Estimate
Intercept	26.66350677
drink_days	-0.14945246
Age	0.01404807
poor_sleep_days	0.04593183

- a. From the description given, the form of the estimated regression equation is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 (\text{drink_days}) + \hat{\beta}_2 (\text{age}) + \hat{\beta}_3 (\text{poor_sleep_days})$$

Using the parameter estimates from the computer output, we write the fitted model as

$$\hat{Y} = 26.664 - 0.150 (\text{drink_days}) + 0.014 (\text{age}) + 0.046(\text{poor_sleep_days})$$

- b. To obtain the estimated average BMI of females who drink 10 days per month, are 30 years old, and have 5 days of poor sleep in the past month, we find $\hat{Y} = 26.664 - 0.150(4) + 0.014(30) + 0.046(5) = 25.814$.
- c. We have inserted the numerical values obtained for SSE and SSY in the ANOVA table below. We can use these values to compute the other quantities in the table.

Source	DF	Sum of Squares	Mean Square	F Value
Model (Regression)	[a] k	[d] $\text{SSY} - \text{SSE}$	[g] $\frac{\text{SS Regression}}{\text{DF Regression}}$	[i] $\frac{\text{MS Regression}}{\text{MS Residual [MSE]}}$
Error (Residual)	[b] $n - k - 1$	[e] $\text{SSE} = 36065$	[h] $\frac{\text{SSE}}{\text{DF Residual}}$	
Corrected Total	[c] $n - 1$	[f] $\text{SSY} = 37355$		

- The model contains $k = 3$ predictors, and thus the regression degrees of freedom [a] = 3. With $n = 1,049$, the total degrees of freedom [c] is $n - 1 = 1,048$.
- Since the residual error degrees of freedom [b] is obtained by computing $n - k - 1$, we subtract [a] from [c] to obtain $1,048 - 3 = 1,045$.
- The regression sum of squares [d] is found by subtracting, within the Sum of Squares column, the residual sum of squares (SSE) from the total sum of squares (SST)—namely, [f] – [e]—to obtain $37,355 - 36,065 = 1,290$.
- Each mean-square value is found by dividing the relevant sum of squares by its degrees of freedom. The mean-square regression [g] is thus $[d]/[a] = 1,290/3 = 430$, and the mean-square residual [h] is $\text{SSE}/[b] = 36,065/1,045 = 34.5$.

5. The F statistic [i] is found by taking the ratio of the two means squares, $[g]/[h] = 12.5$. Many computer packages also place the two-sided P -value for the F statistic in another column to the right of this value, providing an overall test for the significance of the regression model. We will discuss this test further in the next chapter.

Our ANOVA table is now complete:

Source	DF	Sum of Squares	Mean Square	F Value
Model (Regression)	3	1290	430	12.5
Error (Residual)	1045	36065	34.5	
Corrected Total	1048	37355		

- d. The R^2 for this model is found by dividing the regression sum of squares by SSY, giving $1,290/37,355 = 0.0345$. Comparing this value to the value of 0.0313 found in Chapter 6 for the straight-line model with drinking days as the only independent variable, we can see that the inclusion of age and sleep quality in the model negligibly improves the prediction of BMI.

8.8 Numerical Examples

We conclude this chapter with some examples of the type of computer output to be expected from a typical regression program. This output generally consists of the values of the estimated regression coefficients, their estimated standard errors, the associated partial F (or T^2) statistics,⁴ and an ANOVA table. For the data of Table 8.1, the six models that follow are by no means the only ones possible; for instance, no product (i.e., interaction) terms were included.

Model 1 $\text{WGT} = \beta_0 + \beta_1 \text{HGT} + \epsilon$

Edited SAS Output (PROC REG) for Regression of WGT on HGT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	588.9225232	588.9225232	19.67	0.0013
Error	10	299.3274768	29.9327477		
Corrected Total	11	888.2500000			

↓
F statistic for overall test

SSY SSE SSY-SSE

P-value for overall test

(continued)

⁴ Partial F statistics will be discussed in Chapter 9 on hypothesis testing.

R-Square	Coeff Var	Root MSE	WGT Mean
0.663014	8.718857	5.471083	62.75000

 R^2 *[Portion of output omitted]* $\hat{\beta}_0$

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	6.189848707	12.84874620	0.48	0.6404
HGT	1.072230356	0.24173098	4.44	0.0013

 $\hat{\beta}_1$

$$\text{Model 2 } \text{WGT} = \beta_0 + \beta_2 \text{AGE} + \epsilon$$

Edited SAS Output (PROC REG) for Regression of WGT on AGE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	526.3928571	526.3928571	14.55	0.0034
Error	10	361.8571429	36.1857143		
Corrected Total	11	888.2500000			

R-Square	Coeff Var	Root MSE	WGT Mean
0.592618	9.586385	6.015456	62.75000

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	30.57142857	8.61370526	3.55	0.0053
AGE	3.64285714	0.95511512	3.81	0.0034

$$\text{Model 3 } \text{WGT} = \beta_0 + \beta_3 (\text{AGE})^2 + \epsilon$$

Edited SAS Output (PROC REG) for Regression of WGT on (AGE)²

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	521.9320473	521.9320473	14.25	0.0036
Error	10	366.3179527	36.6317953		
Corrected Total	11	888.2500000			

R-Square	Coeff Var	Root MSE	WGT Mean
0.587596	9.645292	6.052421	62.75000

(continued)

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	45.99764279	4.76964028	9.64	<.0001
AGESQ	0.20597161	0.05456692	3.77	0.0036

$$\text{Model 4 } \text{WGT} = \beta_0 + \beta_1 \text{HGT} + \beta_2 \text{AGE} + \epsilon$$

Edited SAS Output (PROC REG) for Regression of WGT on HGT and AGE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	692.8226065	346.4113033	15.95	0.0011
Error	9	195.4273935	21.7141548		
Corrected Total	11	888.2500000			

R-Square	Coeff Var	Root MSE	WGT Mean
0.779986	7.426048	4.659845	62.75000

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	6.553048251	10.94482708	0.60	0.5641
HGT	0.722037958	0.26080506	2.77	0.0218
AGE	2.050126352	0.93722561	2.19	0.0565

Test statistics and P-values for partial tests on model parameters (see Section 9.3)

$$\text{Model 5 } \text{WGT} = \beta_0 + \beta_1 \text{HGT} + \beta_3 (\text{AGE})^2 + \epsilon$$

Edited SAS Output (PROC REG) for Regression of WGT on HGT and (AGE)²

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	689.6499511	344.8249755	15.63	0.0012
Error	9	198.6000489	22.0666721		
Corrected Total	11	888.2500000			

R-Square	Coeff Var	Root MSE	WGT Mean
0.776414	7.486084	4.697518	62.75000

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	15.11753900	11.79690059	1.28	0.2321
HGT	0.72597651	0.26333057	2.76	0.0222
AGESQ	0.11480164	0.05373319	2.14	0.0614

$$\text{Model 6} \quad \text{WGT} = \beta_0 + \beta_1 \text{HGT} + \beta_2 \text{AGE} + \beta_3 (\text{AGE})^2 + \epsilon$$

Edited SAS Output (PROC REG) for Regression of WGT on HGT, AGE, and (AGE)²

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	693.0604634	231.0201545	9.47	0.0052
Error	8	195.1895366	24.3986921		
Corrected Total	11	888.2500000			
R-Square		Coeff Var	Root MSE	WGT Mean	
0.780254		7.871718	4.939503	62.75000	
Parameter	Estimate	Standard Error	t Value	Pr > t	
Intercept	3.438426001	33.61081984	0.10	0.9210	
HGT	0.723690241	0.27696316	2.61	0.0310	
AGE	2.776874563	7.42727877	0.37	0.7182	
AGESQ	-0.041706699	0.42240715	-0.10	0.9238	

Although we will discuss model selection more fully in Chapter 16, it may already be clear from these results that model 4, involving HGT and AGE, is the best of the lot if we use R^2 and model simplicity as our criteria for selecting a model. The R^2 -value of 0.7800 achieved by using this model is, for all practical purposes, the same as the maximum R^2 -value of 0.7803 obtained by using all three variables.

Problems

1. The multiple regression relationships of SBP (Y) to AGE (X_1), SMK (X_2), and QUET (X_3) are studied using the data in Problem 2 of Chapter 5. Three regression models are considered, yielding least-squares estimates and ANOVA tables as shown in the accompanying SAS output.
 - a. Use the model that includes all three independent variables (X_1 , X_2 , and X_3) to answer the following questions. (1) What is the predicted SBP for a 50-year-old smoker with a quetelet index of 3.5? (2) What is the predicted SBP for a 50-year-old nonsmoker with a quetelet index of 3.5? (3) For 50-year-old smokers, estimate the change in SBP corresponding to an increase in quetelet index from 3.0 to 3.5.
 - b. Using the computer output, determine and compare the R^2 -values for the three models. If you use R^2 and model simplicity as the criteria for selecting a model, which of the three models appears to be the best?

Edited SAS Output (PROC REG) for Problem 1

Regression of SBP on AGE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3861.630375	3861.630375	45.18	<.0001
Error	30	2564.338375	85.477946		
Corrected Total	31	6425.968750			

[Portion of output omitted]

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	59.09162500	12.81626145	4.61	<.0001
AGE	1.60450000	0.23871593	6.72	<.0001

Regression of SBP on AGE and SMK

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4689.684229	2344.842114	39.16	<.0001
Error	29	1736.284521	59.871880		
Corrected Total	31	6425.968750			

[Portion of output omitted]

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	48.04960299	11.12955962	4.32	0.0002
AGE	1.70915965	0.20175872	8.47	<.0001
SMK	10.29439180	2.76810685	3.72	0.0009

Regression of SBP on AGE, SMK, and QUET

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4889.825697	1629.941899	29.71	<.0001
Error	28	1536.143053	54.862252		
Corrected Total	31	6425.968750			

[Portion of output omitted]

(continued)

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	45.10319242	10.76487511	4.19	0.0003
AGE	1.21271462	0.32381922	3.75	0.0008
SMK	9.94556782	2.65605655	3.74	0.0008
QUET	8.59244866	4.49868122	1.91	0.0664

2. A psychiatrist wants to know whether the level of pathology (Y) in psychotic patients 6 months after treatment can be predicted with reasonable accuracy from knowledge of pretreatment symptom ratings of thinking disturbance (X_1) and hostile suspiciousness (X_2). The following table lists data collected on 53 patients.

Patient	Y	X_1	X_2	Patient	Y	X_1	X_2	Patient	Y	X_1	X_2
1	44	2.80	6.1	19	21	2.81	6.0	37	50	2.90	6.7
2	25	3.10	5.1	20	22	2.80	6.4	38	9	2.74	5.5
3	10	2.59	6.0	21	60	3.62	6.8	39	13	2.70	6.9
4	28	3.36	6.9	22	10	2.74	8.4	40	22	3.08	6.3
5	25	2.80	7.0	23	60	3.27	6.7	41	23	2.18	6.1
6	72	3.35	5.6	24	12	3.78	8.3	42	31	2.88	5.8
7	45	2.99	6.3	25	28	2.90	5.6	43	20	3.04	6.8
8	25	2.99	7.2	26	39	3.70	7.3	44	65	3.32	7.3
9	12	2.92	6.9	27	14	3.40	7.0	45	9	2.80	5.9
10	24	3.23	6.5	28	8	2.63	6.9	46	12	3.29	6.8
11	46	3.37	6.8	29	11	2.65	5.8	47	21	3.56	8.8
12	8	2.72	6.6	30	7	3.26	7.2	48	13	2.74	7.1
13	15	3.47	8.4	31	23	3.15	6.5	49	10	3.06	6.9
14	28	2.70	5.9	32	16	2.60	6.3	50	4	2.54	6.7
15	26	3.24	6.0	33	26	2.74	6.8	51	18	2.78	7.2
16	27	2.65	6.0	34	8	2.72	5.9	52	10	2.81	5.2
17	4	3.41	7.6	35	11	3.11	6.8	53	7	3.26	6.6
18	14	2.58	6.2	36	12	2.79	6.7				

- a. The least-squares equation involving both independent variables is given by $\hat{Y} = -0.0635 + 23.451X_1 - 7.073X_2$. Using this equation, determine the predicted level of pathology for a patient with pretreatment scores of 2.80 on thinking disturbance and 7.0 on hostile suspiciousness. How does this predicted value compare with the value actually obtained for patient 5?
- b. Sums of squares are shown next for three regression models. Determine the R^2 -values for each of these models. If you use R^2 and model simplicity as selection criteria, which model appears to be the best?

Y regressed on X_1 : $SSY = 13,791.1698$, $SSE = 12,255.3128$

Y regressed on X_2 : $SSE = 13,633.3225$

Y regressed on X_1 and X_2 : $SSE = 11,037.2985$

3. The following table presents the weight (X_1), age (X_2), and plasma lipid levels of total cholesterol (Y) for a hypothetical sample of 25 patients suffering from hyperlipoproteinemia, before drug therapy.

- a. Three estimated regression models, along with their sums-of-squares results, are as follows:

$$\hat{Y} = 77.98310.417X_1 + 5.217X_2 \quad \text{SSY} = 145,377.0400, \quad \text{SSE} = 42,806.2254$$

$$\hat{Y} = 199.2975 + 1.622X_1 \quad \text{SSE} = 135,145.3138$$

$$\hat{Y} = 102.5751 + 5.321X_2 \quad \text{SSE} = 43,444.3743$$

For each of these models, determine the predicted cholesterol level (Y) for patient 4, and compare these predicted cholesterol levels with the observed values. Comment on your findings.

Patient	Total Cholesterol (Y) (mg/100 ml)	Weight (X_1) (kg)	Age (X_2) (yr)	Patient	Total Cholesterol (Y) (mg/100 ml)	Weight (X_1) (kg)	Age (X_2) (yr)
1	354	84	46	14	254	57	23
2	190	73	20	15	395	59	60
3	405	65	52	16	434	69	48
4	263	70	30	17	220	60	34
5	451	76	57	18	374	79	51
6	302	69	25	19	308	75	50
7	288	63	28	20	220	82	34
8	385	72	36	21	311	59	46
9	402	79	57	22	181	67	23
10	365	75	44	23	274	85	37
11	209	27	24	24	303	55	40
12	290	89	31	25	244	63	30
13	346	65	52				

- b. Determine R^2 -values for each of the three models considered in part (a). If you use R^2 and model simplicity as selection criteria, which model appears to be the best predictive model?
4. A sociologist investigating the recent increase in the incidence of homicide throughout the United States studied the extent to which the homicide rate per 100,000 population (Y) is associated with the city's population size (X_1), the percentage of families with yearly income less than \$5,000 (X_2), and the rate of unemployment (X_3). Data are provided in the following table for a hypothetical sample of 20 cities.

City	Y	X_1 (thousands)	X_2	X_3	City	Y	X_1 (thousands)	X_2	X_3
1	11.2	587	16.5	6.2	11	14.5	7,895	18.1	6.0
2	13.4	643	20.5	6.4	12	26.9	762	23.1	7.4
3	40.7	635	26.3	9.3	13	15.7	2,793	19.1	5.8
4	5.3	692	16.5	5.3	14	36.2	741	24.7	8.6
5	24.8	1,248	19.2	7.3	15	18.1	625	18.6	6.5
6	12.7	643	16.5	5.9	16	28.9	854	24.9	8.3
7	20.9	1,964	20.2	6.4	17	14.9	716	17.9	6.7
8	35.7	1,531	21.3	7.6	18	25.8	921	22.4	8.6
9	8.7	713	17.2	4.9	19	21.7	595	20.2	8.4
10	9.6	749	14.3	6.4	20	25.7	3,353	16.9	6.7

- a. Using the regression results presented next, determine the R^2 -values for each two-variable model, and comment on which appears to be the best.

Y regressed on X_1 and X_2 : $\text{SSY} = 1,855.2020$, $\text{SSE} = 537.4036$

Y regressed on X_1 and X_3 : $\text{SSE} = 431.9700$

Y regressed on X_2 and X_3 : $\text{SSE} = 367.3426$

- b. When Y is regressed on X_1 , X_2 , and X_3 , we get $\text{SSE} = 337.0571$. Determine and comment on the increase in R^2 in going from a model with just X_2 and X_3 to a model that includes all three independent variables.
- c. Consider the model with independent variables X_2 , X_3 , and X_4 , where $X_4 = X_2X_3$.⁵ The ANOVA table for this model, from SAS's GLM procedure, is given next. Does the addition of X_4 lead to a large improvement in fit over the model with X_2 and X_3 as the only independent variables? Explain.

Edited SAS Output (PROC GLM) for Problem 4

Regression of Y on X_2 , X_3 , X_4

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1490.571279	496.857093	21.80	<.0001
Error	16	364.630721	22.789420		
Corrected Total	19	1855.202000			

5. A panel of educators in a large urban community wanted to evaluate the effects of educational resources on student performance. They examined the relationship between 12th-grade mean math SAT scores (Y) and the following independent variables for a random sample of 25 high schools: X_1 = per pupil expenditure (in dollars); X_2 = percentage of teachers with a master's degree or higher; and X_3 = pupil-teacher ratio. The sums of squares shown next can be used to summarize the key results from the regression of Y on X_1 , X_2 , and X_3 :

$\text{SSY} = 28,222.23$ $\text{SSE} = 2,248.23$

- a. Determine the ANOVA table for the regression of Y on X_1 , X_2 , and X_3 .
- b. Determine the R^2 -value for the model in part (a). Based on this value, comment on whether the three educational resource variables appear to be associated with student performance.
6. A team of environmental epidemiologists used data from 23 counties to investigate the relationship between respiratory cancer mortality rates (Y) for a given year and

⁵ The coefficient of the product term X_4 measures an *interaction effect* associated with the variables X_2 and X_3 , which concerns whether the relationship between Y and one of these two variables depends on the levels of the other variable. A more detailed discussion of the concept of interaction is given in Chapter 11.

the following three independent variables: X_1 = air pollution index for the county; X_2 = mean age (over 21) for the county; and X_3 = percentage of workforce in the county employed in a certain industry. The sums of squares shown next can be used to summarize the key results from the regression of Y on X_1 , X_2 , and X_3 :

$$\text{SSY} = 2,387.653 \quad \text{SSE} = 551.723$$

- a. Determine the ANOVA table for the regression of Y on X_1 , X_2 , and X_3 .
 - b. Determine the R^2 -value for the model in part (a). Based on this value, comment on whether the pollution level, mean age, and percentage of people working in the particular industry appear to be related (as a group) to respiratory cancer mortality rates.
7. In an experiment to describe the toxic action of a certain chemical on silkworm larvae, the relationship of \log_{10} dose and \log_{10} larva weight to \log_{10} survival time was sought.⁶ The data, obtained by feeding each larva a precisely measured dose of the chemical in an aqueous solution and then recording the survival time (i.e., time until death), are given in the following table. Relevant computer results are also provided.

Larva	1	2	3	4	5	6	7	8
\log_{10} survival time (Y)	2.836	2.966	2.687	2.679	2.827	2.442	2.421	2.602
\log_{10} dose (X_1)	0.150	0.214	0.487	0.509	0.570	0.593	0.640	0.781
\log_{10} weight (X_2)	0.425	0.439	0.301	0.325	0.371	0.093	0.140	0.406

Larva	9	10	11	12	13	14	15
\log_{10} survival time (Y)	2.556	2.441	2.420	2.439	2.385	2.452	2.351
\log_{10} dose (X_1)	0.739	0.832	0.865	0.904	0.942	1.090	1.194
\log_{10} weight (X_2)	0.364	0.156	0.247	0.278	0.141	0.289	0.193

- a. Compute R^2 for each of the three models. Which independent variable appears to be the single best predictor of survival time?
- b. Which model involving one or both of the independent variables do you prefer? Why?

Edited SAS Output (PROC GLM) for Problem 7

Regression of Y on X1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.36327405	0.36327405	31.91	<.0001
Error	13	0.14799288	0.01138407		
Corrected Total	14	0.51126693			

(continued)

⁶ Adapted from a study by Bliss (1936).

[Portion of output omitted]

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	2.952199058	0.07355501	40.14	<.0001
X1	-0.549855934	0.09733756	-5.65	<.0001

Regression of Y on X2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.33667410	0.33667410	25.07	0.0002
Error	13	0.17459283	0.01343022		
Corrected Total	14	0.51126693			

[Portion of output omitted]

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	2.184705838	0.08199583	26.64	<.0001
X2	1.375578800	0.27474016	5.01	0.0002

Regression of Y on X1 and X2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.46420205	0.23210102	59.18	<.0001
Error	12	0.04706489	0.00392207		
Corrected Total	14	0.51126693			

[Portion of output omitted]

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	2.588995674	0.08360793	30.97	<.0001
X1	-0.378480493	0.06637411	-5.70	<.0001
X2	0.874974778	0.17248352	5.07	0.0003

8. An experiment to evaluate the effects of certain variables on soil erosion was performed on 10-foot-square plots of sloped farmland subjected to 2 inches of artificial rain applied over a 20-minute period.⁷ The data and related ANOVA table are as follows:

⁷ Adapted from a study by Packer (1951).

Plot	1	2	3	4	5	6	7	8	9	10	11
SL (Y)	27.1	35.6	31.4	37.8	40.2	39.8	55.5	43.6	52.1	43.8	35.7
SG (X_1)	0.43	0.47	0.44	0.48	0.48	0.49	0.53	0.50	0.55	0.51	0.48
LOBS (X_2)	1.95	5.13	3.98	6.25	7.12	6.50	10.67	7.08	9.88	8.72	4.96
PGC (X_3)	0.34	0.32	0.29	0.30	0.25	0.26	0.10	0.16	0.19	0.18	0.28

Note: SL denotes soil lost (in pounds/acre), SG denotes slope gradient of the plot, LOBS denotes length (in inches) of the largest opening of bare soil on any boundary, and PGC denotes percentage of ground cover.

Source	d.f.	SS
Regression	3	680.4912
Residual	7	16.0942
Total	10	696.5855

- Compute R^2 , and comment on the fit of the model.
- The fitted model involving all three independent variables is given by $\hat{Y} = -1.879 + 77.326X_1 + 1.559X_2 - 23.904X_3$. Compute and compare observed and predicted values of Y for plots 1, 5, and 7.
- In a study by Yoshida (1961), the oxygen consumption of wireworm larva groups was measured at five temperatures. The rate of oxygen consumption per larva group (in milliliters per hour)—the dependent variable—was transformed to 0.5 less than the common logarithm. Another independent variable (other than temperature) of importance was larva group weight, which was also transformed to common logarithms. The data are given in the following table.

Oxygen Consumption (Y) (log ml/hr - 0.5)	Larva Group Weight (X_1) (log cg)	Temperature (X_2) (°C)	Oxygen Consumption (Y) (log ml/hr - 0.5)	Larva Group Weight (X_1) (log cg)	Temperature (X_2) (°C)
0.054	0.130	15.5	0.482	0.053	30.0
0.154	0.215	15.5	0.477	0.114	30.0
0.073	0.250	15.5	0.551	0.137	30.0
0.182	0.267	15.5	0.516	0.190	30.0
0.241	0.389	15.5	0.561	0.210	30.0
0.316	0.490	15.5	0.588	0.230	30.0
0.290	0.491	15.5	0.561	0.240	30.0
0.061	0.004	20.0	0.580	0.260	30.0
0.143	0.164	20.0	0.674	0.389	30.0
0.188	0.225	20.0	0.718	0.470	30.0
0.176	0.314	20.0	0.754	0.521	30.0
0.248	0.447	20.0	0.800	0.544	30.0
0.357	0.477	20.0	0.654	-0.004	35.0
0.403	0.505	20.0	0.744	0.033	35.0
0.342	0.537	20.0	0.711	0.049	35.0
0.335	-0.046	25.0	0.855	0.140	35.0
0.408	0.176	25.0	0.932	0.204	35.0
0.366	0.199	25.0	0.927	0.210	35.0
0.482	0.292	25.0	0.914	0.215	35.0
0.545	0.380	25.0	0.914	0.265	35.0
0.596	0.483	25.0	0.973	0.346	35.0
0.590	0.491	25.0	1.000	0.462	35.0
0.631	0.491	25.0	0.998	0.468	35.0
0.610	0.519	25.0			

- a. The fitted multiple regression model containing both X_1 and X_2 is given by

$$\hat{Y} = -0.6838 + 0.5921X_1 + 0.0394X_2$$

On the basis of this fitted model, how much of a change in oxygen consumption would be predicted for a larva group with fixed weight X_1 if the temperature were increased from $X_2 = 20$ to $X_2 = 25$?

- b. For a temperature of 20°C, compute and compare the predicted values of \hat{Y} for weights of 0.250 and 0.500.
 c. What is R^2 for each of the three models?

Edited SAS Output (PROC GLM) for Problem 9

Regression of Y on X1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.06607190	0.06607190	0.89	0.3505
Error	45	3.33995010	0.07422111		
Corrected Total	46	3.40602200			

Regression of Y on X2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.77418353	2.77418353	197.58	<.0001
Error	45	0.63183847	0.01404085		
Corrected Total	46	3.40602200			

Regression of Y on X1 and X2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3.21119764	1.60559882	362.62	<.0001
Error	44	0.19482436	0.00442783		
Corrected Total	46	3.40602200			

10. Residential real estate prices depend, in part, on property size and number of bedrooms. The house size X_1 (in hundreds of square feet), number of bedrooms X_2 , and house price Y (in thousands of dollars) of a random sample of houses in a certain county were observed. The data are listed in the following table.

House	1	2	3	4	5	6	7
House size (X_1)	18	20	25	22	33	19	17
Number of bedrooms (X_2)	3	3	4	4	5	4	3
House price (Y)	80	95	104	110	175	85	89

Use the accompanying output to answer the following questions.

- Determine the least-squares estimates for the model in which house price is regressed on both house size and number of bedrooms. Find the predicted average price of a 2,500-square-foot home that has four bedrooms.
- Compare the R^2 -value for the regression in part (a) with the r^2 value for the data in Problem 16 of Chapter 5. Does adding X_2 to a model that already contains X_1 appear to be useful in predicting the house price?

Edited SAS Output (PROC GLM) for Problem 10

Regression of Y on X1 and X2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5733.321290	2866.660645	20.03	0.0082
Error	4	572.392996	143.098249		
Corrected Total	6	6305.714286			
Parameter		Estimate	Standard Error	t Value	Pr > t
Intercept		-16.09338521	24.64693805	-0.65	0.5494
X1		5.72178988	1.82778969	3.13	0.0352
X2		-1.17315175	13.38993701	-0.09	0.9344

- Data on sales revenues Y , television advertising expenditures X_1 , and print media advertising expenditures X_2 for a large retailer for the period 1988–1993 are given in the following table:

Year	1988	1989	1990	1991	1992	1993
Sales (\$millions)	4.0	8.0	2.0	8.0	5.0	4.0
TV advertising (\$millions)	1.5	4.5	0.0	5.0	3.0	1.5
Print advertising (\$millions)	0.5	0.5	0.0	1.0	1.0	1.5

- State the model for the regression of sales revenue on television advertising expenditures and print advertising expenditures.

Use the accompanying computer output to answer the following questions.

- State the estimate of the model in part (a). What is the estimated change in sales revenue for every \$1,000 increase in television advertising expenditures? What is the estimated change in sales revenue for every \$1,000 increase in print advertising expenditures?
- Find the R^2 -value for the regression of Y on X_1 and X_2 . Interpret your result.
- Predict the sales for a year in which \$5 million is spent on TV advertising and \$1 million is spent on print advertising.

Edited SAS Output (PROC GLM) for Problem 11

Regression of Y on X1 and X2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	28.11853189	14.05926594	59.01	0.0039
Error	3	0.71480144	0.23826715		
Corrected Total	5	28.83333333			

[Portion of output omitted]

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	2.104693141	0.42196591	4.99	0.0155
X1	1.241877256	0.11913357	10.42	0.0019
X2	-0.194945848	0.43944079	-0.44	0.6874

12. Radial keratotomy is a type of refractive surgery in which radial incisions are made in the cornea of myopic (nearsighted) patients in an effort to reduce their myopia. Theoretically, the incisions allow the curvature of the cornea to become less steep, thereby reducing the refractive error of the patient. This and other vision correction surgery grew in popularity in the 1980s and 1990s, both among the public and among ophthalmologists.

Edited SAS Output (PROC GLM) for Problem 12

Regression of Y on X1 and X2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	17.62277191	8.81138596	7.18	0.0018
Error	51	62.63016929	1.22804254		
Corrected Total	53	80.25294120			

R-Square	Coeff Var	Root MSE	Y Mean
0.219590	28.90811	1.108171	3.833426

[Portion of output omitted]

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	12.36001523	5.08621177	2.43	0.0187
X1	-0.29160125	0.09165295	-3.18	0.0025
X2	-0.22039615	0.11482391	-1.92	0.0605

The Prospective Evaluation of Radial Keratotomy (PERK) study began in 1983 to investigate the effects of radial keratotomy. Lynn et al. (1987) examined the factors associated with the 5-year postsurgical change in refractive error (Y , measured in diopters, D). Two independent variables under consideration were baseline refractive error (X_1 , in diopters) and baseline curvature of the cornea (X_2 , in diopters). (Note: Myopic patients have negative refractive errors. Patients who are farsighted have positive refractive errors. Patients who are neither near- nor farsighted have zero refractive error.)

The accompanying computer output is based on data adapted from the PERK study. Use it to answer the following questions.

- a. State the estimated least-squares equation for the regression of change in refractive error (Y) on baseline refractive error (X_1) and baseline curvature (X_2).
 - b. Using your answer to part (a), give a point estimate for the change in refractive error for a patient who, at baseline, has a refractive error of -8.00D and a corneal curvature of 44D . (Note: Myopic patients have negative refractive errors.)
 - c. Find the R^2 -value for the regression in part (a), and comment on the fit of the model.
13. In 1990, *Business Week* magazine compiled financial data on the 1,000 companies that had the biggest impact on the U.S. economy.⁸ Data sampled from the top 500 companies in *Business Week's* report are presented in the following table. In addition to the company name, four variables are shown:
- 1990 rank: Based on company's market value (share price on March 16, 1990, multiplied by available common shares outstanding)
- 1989 rank: Rank in 1989 compilation
- P-E ratio: Price-to-earnings ratio, based on 1989 earnings and March 16, 1990, share price
- Yield: Annual dividend rate as a percentage of March 16, 1990, share price

Company	1990 Rank	1989 Rank	P-E Ratio	Yield	Company	1990 Rank	1989 Rank	P-E Ratio	Yield
AT&T	4	4	17	2.87	ITT	81	57	8	2.98
Merck	7	7	19	2.54	Humana	162	209	15	2.62
Boeing	27	41	24	1.72	Salomon	236	172	7	2.91
American Home Products	32	37	14	4.26	Walgreen	242	262	17	1.87
Walt Disney	33	42	23	0.41	Lincoln National	273	274	9	4.73
Pfizer	46	46	14	4.10	Citizens Utilities	348	302	21	0.00
MCI Communications	52	72	19	0.00	MNC Financial	345	398	6	5.46
Dunn & Bradstreet	55	48	15	4.27	Bausch & Lomb	354	391	15	1.99
United Telecommunications	63	93	22	2.61	Medtronic	356	471	16	1.10
Warner Lambert	77	91	17	2.94	Circuit City	497	514	14	0.33

⁸ "The Business Week 1000" 1990.

Answer the following questions about these data.

- What is the estimated least-squares equation for the regression of yield (Y) on 1989 rank (X_2) and P-E ratio (X_3)?
- Using your answer to part (a), give a point estimate for the yield for a company that had a 1989 ranking of 200 and a P-E ratio of 10.
- Find the R^2 -value for the regression in part (a), and comment on the fit of the model.

Edited SAS Output (PROC GLM) for Problem 13

Regression of Y on X2 and X3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	26.58272135	13.29136068	10.51	0.0011
Error	17	21.49757365	1.26456316		
Corrected Total	19	48.08029500			

R-Square	Coeff Var	Root MSE	Y Mean
0.552882	45.24353	1.124528	2.485500

[Portion of output omitted]

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	6.873525074	0.99001534	6.94	<.0001
X2	-0.004138085	0.00164749	-2.51	0.0224
X3	-0.234451675	0.05292277	-4.43	0.0004

14. This problem refers to the 1990 Census data presented in Problem 19 of Chapter 5. In addition to median selected monthly ownership costs (OWNCOST), another independent variable studied was the proportion of the total metropolitan statistical area (MSA) population living in urban areas (URBAN).

Use the accompanying computer output to answer the following questions about the regression of OWNEROCC on OWNCOST and URBAN.

- What is the estimated least-squares equation for the regression of OWNEROCC on OWNCOST and URBAN?
- Using your answer to part (a), give a point estimate for the rate of owner occupancy for an MSA for which the urban proportion is 0.73 and the median selected monthly ownership cost is \$513.
- Find the R^2 -value for the regression in part (a), and comment on the fit of the model.

Edited SAS Output (PROC REG) for Problem 14

Regression of OWNROCC on OWNCOST and URBAN

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	255.77851	127.88925	8.52	0.0017
Error	23	345.18303	15.00796		
Corrected Total	25	600.96154			

Root MSE	3.87401	R-Square	0.4256
Dependent Mean	65.96154	Adj R-Sq	0.3757
Coeff Var	5.87314		

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	86.75877	5.10721	16.99	<.0001
OWNCOST	1	-0.01113	0.00529	-2.10	0.0467
URBAN	1	-16.87557	5.89098	-2.86	0.0088

15. This problem refers to the pond ecology data of Chapter 5, Problem 20.
- What is the estimated least-squares equation for the multiple regression of copepod count on zooplankton and phytoplankton counts.
 - Find the R^2 -value for the model in part (a), and comment on the fit of the model.

References

- Bliss, C. I. 1936. "The Size Factor in Action of Arsenic upon Silkworms' Larvae." *Journal of Experimental Biology* 13: 95–110.
- "The Business Week 1000, America's Most Valuable Companies." 1990. *Business Week*, special issue, April 13.
- Diggle, P. J.; Heagerty, P.; Liang, K. Y.; and Zeger, S. L. 2002. *Analysis of Longitudinal Data*. New York: Oxford University Press.
- Gallant, A. R. 1975. "Non-linear Regression." *American Statistician* 29: 73–81.
- Lynn, M. J.; Waring, G. O.; and Sperduto, R. D. 1987. "Factors Affecting Outcome and Predictability of Radial Keratotomy in the PERK Study." *Archives of Ophthalmology* 105: 42–51.
- Packer, P. E. 1951. "An Approach to Watershed Protection Criteria." *Journal of Forestry* 49: 638–44.
- Yoshida, M. 1961. "Ecological and Physiological Researches on the Wireworm, *Melanotus caudex* Lewis. Iwata." *Shizuoka Pref.*, Japan.
- Zeger, S. L., and Liang, K. Y. 1986. "Longitudinal Data Analysis for Discrete and Continuous Outcomes." *Biometrics* 42: 121–30.

9

Statistical Inference in Multiple Regression

9.1 Preview

Once we have fit a multiple regression model and obtained estimates for the various parameters of interest, we want to answer questions about the contributions of various independent variables to the prediction of Y . Such questions raise the need for three basic types of tests:

1. *Overall test.* Taken collectively, does the *entire set* of independent variables (or, equivalently, the fitted model itself) contribute significantly to the prediction of Y ?
2. *Test for addition of a single variable.* Does the addition of *one* particular independent variable of interest add significantly to the prediction of Y achieved by other independent variables already present in the model?
3. *Test for addition of a group of variables.* Does the addition of some *group* of independent variables of interest add significantly to the prediction of Y obtained through other independent variables already present in the model?

These questions are typically answered by performing statistical tests of hypotheses. The null hypotheses for the tests can be stated in terms of the unknown parameters (the regression coefficients) in the model. The form of these hypotheses differs depending on the question being asked. (In Chapter 10, we will look at alternative but equivalent ways to state such null hypotheses in terms of population correlation coefficients.)

In the sections that follow, we will describe the statistical test appropriate for each of the preceding questions. Each of these tests can be expressed as an F test; that is, the test statistic will have an F distribution when the stated null hypothesis is true. In some cases, the test may be equivalently expressed as a t test. (For a review of material concerning the F and t distributions, refer to Chapter 3.)

All F tests used in regression analyses involve a ratio of two independent estimates of variance—say, $F = \hat{\sigma}_0^2/\hat{\sigma}^2$. Under the assumptions for the standard multiple linear regression analysis given earlier, the term $\hat{\sigma}_0^2$ estimates σ^2 if H_0 is true; the term $\hat{\sigma}^2$ estimates σ^2 whether

H_0 is true or not. The specific forms that these variance estimates take will be described in subsequent sections. In general, each is a mean-square term that can be found in an appropriate ANOVA table. If H_0 is not true, then $\hat{\sigma}_0^2$ estimates some quantity larger than σ^2 . Thus, we would expect a value of F that is close to 1 ($= \sigma^2/\sigma^2$) if H_0 is true but larger than 1 if H_0 is not true. The larger the value of F , then, the likelier H_0 is to be untrue.

Another general characteristic of the tests to be discussed in this chapter is that *each test can be interpreted as a comparison of two models*. One of these models will be referred to as the *full* or *complete* model; the other will be called the *reduced* model (i.e., the model to which the complete model reduces under the null hypothesis).

As a simple example, consider the following two models:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$$

and

$$Y = \beta_0 + \beta_1 X_1 + E$$

Under $H_0: \beta_2 = 0$, the larger (full) model reduces to the smaller (reduced) model. A test of $H_0: \beta_2 = 0$ is thus essentially equivalent to determining which of these two models is more appropriate.

As this example suggests, the set of independent variables in the reduced model (namely, X_1) is a subset of the independent variables in the full model (namely, X_1 and X_2). This is a characteristic common to all the basic types of tests to be described in this chapter. (More generally, this subset characteristic need not always be present. Suppose, for example, that we have $H_0: \beta_1 = \beta_2 [= \beta, \text{say}]$. Then, the reduced model may be written as $Y = \beta_0 + \beta X + E$, with $\beta = \beta_1 = \beta_2$ and $X = X_1 + X_2$.)

9.2 Test for Significant Overall Regression

We now reconsider our first question, regarding an overall test for a model containing k independent variables—say,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + E$$

The null hypothesis for this test may be generally stated as H_0 : “All k independent variables considered together do not explain a significant amount of the variation in Y .” Equivalently, we may state the null hypothesis as H_0 : “There is no significant overall regression using all k independent variables in the model,” or as $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$. Under this last version of H_0 , the full model is reduced to a model that contains only the intercept term β_0 .

To perform the test, we use the mean-square quantities provided in our ANOVA table (see Table 8.2 of Chapter 8). We calculate the F statistic

$$F = \frac{\text{MS Regression}}{\text{MS Residual}} = \frac{(\text{SSY} - \text{SSE})/k}{\text{SSE}/(n - k - 1)} \quad (9.1)$$

where $\text{SSY} = \sum_{i=1}^n (Y_i - \bar{Y})^2$ and $\text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ are the total and error sums of squares, respectively. The computed value of F can then be compared with the critical point $F_{k, n-k-1, 1-\alpha}$, where α is the preselected significance level. We would reject H_0 if the

computed F exceeded the critical point. Alternatively, we could compute the P -value for this test as the area under the curve of the $F_{k,n-k-1}$ distribution to the right of the computed F statistic. It can be shown that an equivalent expression for (9.1) in terms of R^2 is

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)} \quad (9.2)$$

For the example summarized in Table 8.2, which concerns the regression of WGT on HGT, AGE, and $(AGE)^2$ for a sample of $n = 12$ children, we have $k = 3$, MS Regression = 231.02, MS Residual (MS Error) = 24.40, and $R^2 = 0.7802$, so that

$$F = \frac{231.02}{24.40} = \frac{0.7802/3}{(1 - 0.7802)/(12 - 3 - 1)} = 9.47$$

The critical point for $\alpha = .01$ is $F_{3,8,0.99} = 7.59$. Thus, we would reject H_0 at $\alpha = .01$ because the P -value is less than .01. (We usually denote $P < .01$ by putting a double ** next to the computed F , as in Table 8.2. When $.01 < P < .05$, we usually use only one*.)

In interpreting the results of this test, we can conclude that, based on the observed data, the set of variables HGT, AGE, and $(AGE)^2$ significantly helps to predict WGT. This conclusion does not mean that *all three* variables are needed for significant prediction of Y ; perhaps only one or two of them are sufficient. In other words, a more parsimonious model than the one involving all three variables may be adequate. To determine this requires further tests, to be described in the next section.

The mean-square residual term in the overall F test, which is the denominator of the F in (9.1), is given by the formula

$$\frac{1}{n - k - 1} SSE = \frac{1}{n - k - 1} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

This quantity provides an estimate of σ^2 under the assumed model. The mean-square regression term $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2/k$, which is the numerator of the F in (9.1), provides an independent estimate of σ^2 only if the null hypothesis of no significant overall regression is true. However, if H_0 is not true (i.e., one or more of the $\beta_1, \beta_2, \dots, \beta_k$ are not equal to zero), then the numerator overestimates σ^2 ; this is why an F -value that is “too large” favors rejection of H_0 . Thus, the F statistic (9.1) is the ratio of two independent estimates of the same variance only if the null hypothesis $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ is true.

9.3 Partial F Test

Some important additional information regarding the fitted regression model can be obtained by presenting the ANOVA table as shown in Table 9.1. In this representation, we have partitioned the regression sum of squares into three components:

1. $SS(X_1)$: the sum of squares explained by using only $X_1 = \text{HGT}$ to predict Y
2. $SS(X_2 | X_1)$: the extra sum of squares explained by using $X_2 = \text{AGE}$ in addition to X_1 to predict Y

TABLE 9.1 ANOVA table for WGT regressed on HGT, AGE, and (AGE)², containing components of the regression sum of squares

Source	d.f.	SS	MS	F	R ²
Regression	X ₁	1	588.92	588.92	19.67**
	X ₂ X ₁	1	103.90	103.90	4.78 (.05 < P < .10)
	X ₃ X ₁ , X ₂	1	0.24	0.24	0.01
Residual	8	195.19	24.40		
Total	11	888.25			

© Cengage Learning

3. SS(X₃|X₁, X₂): the extra sum of squares explained by using X₃ = (AGE)² in addition to X₁ and X₂ to predict Y

We can use the extra information in Table 9.1 to answer the following questions:

1. Does X₁ = HGT alone significantly aid in predicting Y?
2. Does the addition of X₂ = AGE significantly contribute to the prediction of Y after we account (or control) for the contribution of X₁?
3. Does the addition of X₃ = (AGE)² significantly contribute to the prediction of Y after we account for the contribution of X₁ and X₂?

To answer question 1, we simply fit a straight-line regression model, using X₁ = HGT as the single independent variable. The value 588.92, therefore, is the regression sum of squares for this straight-line regression model. The SSE for this model can be obtained from Table 9.1 by adding 195.19, 0.24, and 103.90 together, which yields the sum of squares 299.33, having 10 degrees of freedom (i.e., 10 = 8 + 1 + 1). The F statistic for testing whether there is significant straight-line regression when we use only X₁ = HGT is then given by $F = (588.92/1)/(299.33/10) = 19.67$, which has a P-value of less than .01 (i.e., X₁ contributes significantly to the linear prediction of Y).

To answer questions 2 and 3, we must use what is called a *partial F test*. This test assesses whether the addition of any specific independent variable, given others already in the model, significantly contributes to the prediction of Y. The test, therefore, allows for the deletion of variables that do not help in predicting Y and thus enables one to reduce the set of possible independent variables to an economical set of “important” predictors.

9.3.1 The Null Hypothesis

Suppose that we want to test whether adding a variable X^* significantly improves the prediction of Y given that variables X₁, X₂, ..., X_p are already in the model. The null hypothesis may then be stated as H_0 : “X* does not significantly add to the prediction of Y given that X₁, X₂, ..., X_p are already in the model” or, equivalently, as $H_0: \beta^* = 0$ in the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \beta^* X^* + E$.

As can be inferred from the second statement, the test procedure essentially compares two models: the *full* model contains X_1, X_2, \dots, X_p and X^* as independent variables; the *reduced* model contains X_1, X_2, \dots, X_p but not X^* (since $\beta^* = 0$ under the null hypothesis). The goal is to determine which model is more appropriate based on how much additional information X^* provides about Y over that already provided by X_1, X_2, \dots, X_p . In the next chapter, we shall see that an equivalent statement of H_0 can be given in terms of a partial correlation coefficient.

9.3.2 The Procedure

To perform a partial F test involving a variable X^* , given that variables X_1, X_2, \dots, X_p are already in the model, we must first compute the extra sum of squares from adding X^* given X_1, X_2, \dots, X_p , which we place in our ANOVA table under the source heading “Regression $X^* | X_1, X_2, \dots, X_p$.” This sum of squares is computed by the formula

$$\left[\begin{array}{l} \text{Extra sum of} \\ \text{squares from} \\ \text{adding } X^* \text{ given} \\ X_1, X_2, \dots, X_p \end{array} \right] = \left[\begin{array}{l} \text{Regression sum} \\ \text{of squares when} \\ X_1, X_2, \dots, X_p \\ \text{and } X^* \text{ are all} \\ \text{in the model} \end{array} \right] - \left[\begin{array}{l} \text{Regression sum} \\ \text{of squares when} \\ X_1, X_2, \dots, X_p \\ (\text{and not } X^*) \text{ are} \\ \text{in the model} \end{array} \right] \quad (9.3)$$

or, more compactly,

$$\begin{aligned} SS(X^* | X_1, X_2, \dots, X_p) &= \text{Regression SS}(X_1, X_2, \dots, X_p, X^*) \\ &\quad - \text{Regression SS}(X_1, X_2, \dots, X_p) \end{aligned}$$

[Recall from Section 8.6 that for any linear regression model, $\sum_{i=1}^n (Y_i - \bar{Y})^2$ can be split into two components—the regression sum of squares and the residual sum of squares. Therefore,

$$SS(X^* | X_1, X_2, \dots, X_p) = \text{Residual SS}(X_1, X_2, \dots, X_p) - \text{Residual SS}(X_1, X_2, \dots, X_p, X^*)$$

is an equivalent expression.]

Thus, for our example,

$$\begin{aligned} SS(X_2 | X_1) &= \text{Regression SS}(X_1, X_2) - \text{Regression SS}(X_1) \\ &= 692.82 - 588.92 \\ &= 103.90 \end{aligned}$$

and

$$\begin{aligned} SS(X_3 | X_1, X_2) &= \text{Regression SS}(X_1, X_2, X_3) - \text{Regression SS}(X_1, X_2) \\ &= 693.06 - 692.82 \\ &= 0.24 \end{aligned}$$

To test the null hypothesis H_0 : “The addition of X^* to a model already containing X_1, X_2, \dots, X_p does not significantly improve the prediction of Y ,” we compute

$$F(X^*|X_1, X_2, \dots, X_p) = \frac{\text{Extra sum of squares from adding } X^* \text{ given } X_1, X_2, \dots, X_p}{\begin{matrix} \text{MS Residual for the model} \\ \text{containing all the variables } X_1, X_2, \dots, X_p, X^* \end{matrix}}$$

or, more compactly,

$$F(X^*|X_1, X_2, \dots, X_p) = \frac{\text{SS}(X^*|X_1, X_2, \dots, X_p)}{\text{MS Residual } (X_1, X_2, \dots, X_p, X^*)} \quad (9.4)$$

This F statistic has an F distribution with 1 and $n - p - 2$ degrees of freedom under H_0 , so we should reject H_0 if the computed F exceeds $F_{1,n-p-2,1-\alpha}$. For our example, the partial F statistics are (from Table 9.1).

$$F(X_2|X_1) = \frac{\text{SS}(X_2|X_1)}{\text{MS Residual } (X_1, X_2)} = \frac{103.90}{(195.19 + 0.24)/9} = 4.78$$

and

$$F(X_3|X_1, X_2) = \frac{\text{SS}(X_3|X_1, X_2)}{\text{MS Residual } (X_1, X_2, X_3)} = \frac{0.24}{24.40} = 0.01$$

The quantity MS Residual (X_1, X_2) can be obtained directly from the ANOVA table for only X_1 and X_2 or indirectly from the partitioned ANOVA table for X_1, X_2 , and X_3 by using the formula

$$\text{MS Residual } (X_1, X_2) = \frac{\text{Residual SS}(X_1, X_2, X_3) + \text{SS}(X_3|X_1, X_2)}{8 + 1}$$

The statistic $F(X_2|X_1) = 4.78$ has a P -value satisfying $.05 < P < .10$, since $F_{1,9,0.90} = 3.36$ and $F_{1,9,0.95} = 5.12$. Thus, we should reject H_0 at $\alpha = .10$ and conclude that the addition of X_2 after accounting for X_1 significantly adds to the prediction of Y at the $\alpha = .10$ level. At $\alpha = .05$, however, we would not reject H_0 .

The statistic $F(X_3|X_1, X_2)$ equals 0.01, so obviously H_0 should not be rejected, regardless of the significance level; we, therefore, conclude that, once $X_1 = \text{HGT}$ and $X_2 = \text{AGE}$ are in the model, the addition of $X_3 = (\text{AGE})^2$ is superfluous.

9.3.3 The t Test Alternative

An equivalent way to perform the partial F test for the variable added last is to use a t test. (You may recall that an F statistic with 1 and w degrees of freedom is the square of a t statistic with w degrees of freedom.) The t -test alternative focuses on a test of the null hypothesis

$H_0: \beta^* = 0$, where β^* is the coefficient of X^* in the regression equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \beta^* X^* + E$. The equivalent statistic for testing this null hypothesis is

$$T = \frac{\hat{\beta}^*}{S_{\hat{\beta}}^*} \quad (9.5)$$

where $\hat{\beta}^*$ is the corresponding estimated coefficient and $S_{\hat{\beta}}^*$ is the estimate of the standard error of $\hat{\beta}^*$, both of which are produced by standard regression programs.

In performing this test, we reject $H_0: \beta^* = 0$ if

$$\begin{cases} |T| > t_{n-p-2, 1-\alpha/2} & \text{(two-sided test; } H_A: \beta^* \neq 0) \\ T > t_{n-p-2, 1-\alpha} & \text{(upper one-sided test; } H_A: \beta^* > 0) \\ T < -t_{n-p-2, 1-\alpha} & \text{(lower one-sided test; } H_A: \beta^* < 0) \end{cases}$$

It can be shown that a two-sided t test is equivalent to the partial F test described earlier. For example, in testing $H_0: \beta_3 = 0$ in the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$ fit to the data in Table 8.1, we compute

$$T = \frac{\hat{\beta}_3}{S_{\hat{\beta}_3}} = \frac{-0.0417}{0.4224} = -0.10$$

Squaring, we get

$$T^2 = 0.01 = \text{partial } F(X_3 | X_1, X_2)$$

from Table 9.1.

The t -test approach presented here may be additionally adapted to test $H_0: \beta^* = \beta^{*(0)}$, where $\beta^{*(0)}$ is a specific nonzero value of interest. This may be done by adapting equation (9.5) as follows:

$$T = \frac{\hat{\beta}^* - \beta^{*(0)}}{S_{\hat{\beta}}^*}$$

9.3.4 Comments

An important general application of the partial F test concerns the control of extraneous variables (e.g., confounders, which will be discussed in Chapter 11). Consider, for example, a situation with one main study variable of interest, X^* , and q control variables C_1, C_2, \dots, C_q . The effect of X^* on the outcome variable Y , controlling for C_1, C_2, \dots, C_q , may be assessed by considering the model

$$Y = \beta_0 + \beta_1 C_1 + \beta_2 C_2 + \cdots + \beta_q C_q + \beta_1^* X^* + E$$

The appropriate null hypothesis is $H_0: \beta_1^* = 0$. The partial F statistic in this situation is given by $F(X^* | C_1, C_2, \dots, C_q)$ using (9.4) with $C_i = X_i$, for $i = 1, 2, \dots, q$.

When several study variables (i.e., several X^* 's) are involved, the task includes determining which of the X^* 's are important and perhaps even rank-ordering them by their relative importance. Such a task amounts to finding a best model, a topic we will address in Chapter 16 (where the term *best* will be carefully defined). For now, we note that one strategy (detailed in Chapter 16) is to work backward by *deleting* study (X^*) variables, one at a time, until a best model is obtained. This requires performing several partial F tests (as described in Chapter 16). If the starting model of interest is

$$Y = \beta_0 + \beta_1 C_1 + \cdots + \beta_q C_q + \beta_1^* X_1^* + \cdots + \beta_s^* X_s^* + E,$$

then the first backward step involves considering s partial F tests, $F(X_i^* | C_1, C_2, \dots, C_q)$, all X^* variables except X_i^*), where $i = 1, 2, \dots, s$. The corresponding (separate) null hypotheses are $H_0: \beta_i^* = 0$, where $i = 1, 2, \dots, s$. The usual backward procedure identifies the variable X_I^* associated with the smallest partial F value. This variable becomes the first to be deleted from the model, provided that its partial F is not significant. Then the elimination process starts all over again for the reduced model with X_I^* removed. Of course, if the smallest partial F value is significant, no X^* variables are deleted.

Each partial F test made at the first backward step weighs the contribution of a specific X^* variable given that it is the last X^* variable to enter the model. It is, therefore, inappropriate to delete more than one X^* variable at this first step. For example, it is inappropriate to delete simultaneously *all* X^* variables from the model if all partial F 's are nonsignificant at this first step. This is because, given that one particular X^* variable (say, X_I^*) is deleted, the remaining X^* variables may become important (based on consideration of their partial F 's under the reduced model).

For example, suppose that we fit the model

$$Y = \beta_0 + \beta_1 C_1 + \beta_1^* X_1^* + \beta_2^* X_2^* + E$$

and obtain the following partial F results:

$$F(X_1^* | C_1, X_2^*) = 0.01 (P = .90)$$

$$F(X_2^* | C_1, X_1^*) = 0.85 (P = .25)$$

Then X_1^* is “less significant” than X_2^* , controlling for C_1 and the other X^* variable. Under the strategy of backward elimination, X_1^* should be deleted before the elimination of X_2^* is considered; however, to delete both X_1^* and X_2^* at this point would be incorrect. In fact, when considering the reduced model $Y = \beta_0 + \beta_1 C_1 + \beta_2^* X_2^* + E$, the partial F statistic $F(X_2^* | C_1)$ may be highly significant. In other words, if X_1^* is not significant given X_2^* and C_1 , and if X_2^* is not significant given X_1^* and C_1 , it does not necessarily follow that X_2^* is unimportant in a reduced model containing X_2^* and C_1 but not X_1^* .

9.4 Multiple Partial F Test

This testing procedure addresses the more general problem of assessing the additional contribution of two or more independent variables over and above the contribution made by other variables already in the model. For the example involving $Y = \text{WGT}$, $X_1 = \text{HGT}$,

$X_2 = \text{AGE}$, and $X_3 = (\text{AGE})^2$, we may be interested in testing whether the AGE variables, taken collectively, significantly improve the prediction of WGT given that HGT is already in the model. In contrast to the partial F test discussed in Section 9.3, the multiple partial F test addresses the simultaneous addition of two or more variables to a model. Nevertheless, the test procedure is a straightforward extension of the partial F test.

9.4.1 The Null Hypothesis

We want to test whether the addition of the s variables $X_1^*, X_2^*, \dots, X_s^*$ significantly improves the prediction of Y given that the q variables X_1, X_2, \dots, X_q are already in the model. The (full) model of interest is thus

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q + \beta_1^* X_1^* + \cdots + \beta_s^* X_s^* + E$$

Then the null hypothesis of interest may be stated as H_0 : “ $X_1^*, X_2^*, \dots, X_s^*$ do not significantly add to the prediction of Y given that X_1, X_2, \dots, X_q are already in the model” or, equivalently as $H_0: \beta_1^* = \beta_2^* = \cdots = \beta_s^* = 0$ in the (full) model.¹

From the second version of H_0 , it follows that the *reduced* model is of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q + E$$

(i.e., terms involving X_1^*, \dots, X_s^* are dropped from the full model).

For the preceding example, the (full) model is

$$\text{WGT} = \beta_0 + \beta_1 \text{HGT} + \beta_1^* \text{AGE} + \beta_2^* (\text{AGE})^2 + E$$

The null hypothesis here is $H_0: \beta_1^* = \beta_2^* = 0$.

9.4.2 The Procedure

As in the case of the partial F test, we must compute the extra sum of squares due to the addition of terms involving X_1^*, \dots, X_s^* to the model. In particular, we have

$$\begin{aligned} & \text{SS}(X_1^*, X_2^*, \dots, X_s^* | X_1, X_2, \dots, X_q) \\ &= \text{Regression SS}(X_1, X_2, \dots, X_q, X_1^*, X_2^*, \dots, X_s^*) - \text{Regression SS}(X_1, X_2, \dots, X_q) \\ &= \text{Residual SS}(X_1, X_2, \dots, X_q) - \text{Residual SS}(X_1, X_2, \dots, X_q, X_1^*, X_2^*, \dots, X_s^*) \end{aligned}$$

Using this extra sum of squares, we obtain the following F statistic:

$$F(X_1^*, X_2^*, \dots, X_s^* | X_1, X_2, \dots, X_q) = \frac{\text{SS}(X_1^*, X_2^*, \dots, X_s^* | X_1, X_2, \dots, X_q) / s}{\text{MS Residual}(X_1, X_2, \dots, X_q, X_1^*, X_2^*, \dots, X_s^*)} \quad (9.6)$$

In (9.6), we must divide the extra sum of squares by s , the number of regression coefficients specified to be zero under the null hypothesis of interest. Under $H_0: \beta_1^* = \beta_2^* = \cdots =$

¹ In Chapter 10, an equivalent expression for this null hypothesis will be given in terms of a multiple partial correlation coefficient.

$\beta_s^* = 0$, this number s is also the numerator degrees of freedom for the F statistic. The denominator of the F statistic is the MS Residual for the full model; its degrees of freedom is $n - (q + s + 1)$, which is $(n - 1)$ minus the number of predictor variables in this model (namely, $q + s$ [or p , or $k - 1$]).

An alternative way to write this F statistic is

$$\begin{aligned} F(X_1^*, X_2^*, \dots, X_s^* | X_1, X_2, \dots, X_q) &= \frac{[\text{Regression SS(full)} - \text{Regression SS(reduced)}]/s}{\text{MS Residual}} \\ &= \frac{[\text{Residual SS(reduced)} - \text{Residual SS(full)}]/s}{\text{MS Residual}} \end{aligned}$$

Using the information in Table 9.1 involving WGT, HGT, AGE, and $(\text{AGE})^2$, we can test $H_0: \beta_1^* = \beta_2^* = 0$. in the model $\text{WGT} = \beta_0 + \beta_1 \text{HGT} + \beta_1^* \text{AGE} + \beta_2^* (\text{AGE})^2 + E$, as follows:

$$\begin{aligned} F[\text{AGE}, (\text{AGE})^2 | \text{HGT}] \\ &= \frac{\{\text{Regression SS}[\text{HGT, AGE, } (\text{AGE})^2] - \text{Regression SS}(\text{HGT})\}/2}{\text{MS Residual}[\text{HGT, AGE, } (\text{AGE})^2]} \\ &= \frac{[(588.92 + 103.90 + 0.24) - 588.92]/2}{24.40} \\ &= 2.13 \end{aligned}$$

For $\alpha = .05$, the critical point is

$$F_{s, n-q-s-1, 0.95} = F_{2, 12-1-2-1, 0.95} = 4.46$$

so H_0 would not be rejected at $\alpha = .05$.

In the preceding calculation, we used the relationship

$$\begin{aligned} \text{Regression SS}[\text{HGT, AGE, } (\text{AGE})^2] \\ &= \text{Regression SS}(\text{HGT}) + \text{Regression SS}(\text{AGE} | \text{HGT}) \\ &\quad + \text{Regression SS}[(\text{AGE})^2 | \text{HGT, AGE}] \\ &= 588.92 + 103.90 + 0.24 \end{aligned}$$

Alternatively, we could form two ANOVA tables (Table 9.2)—one for the full model and one for the reduced model—and then extract the appropriate regression and/or residual sum-of-squares values from these tables. More examples of partial F calculations are given at the end of this chapter.

9.4.3 Comments

The multiple partial F test is useful for assessing the simultaneous importance of particular subsets of a set of predictor variables. In particular, it is often used to test whether a “chunk” (i.e., a group) of variables having some trait in common is important when considered together. An example of a chunk is a collection of variables that are all of a certain order [e.g., $(\text{AGE})^2$, $\text{HGT} \times \text{AGE}$, and $(\text{HGT})^2$ are all of order 2].

TABLE 9.2 ANOVA tables for WGT regressed on HGT, AGE, and (AGE)²

Full Model				Reduced Model			
Source	d.f.	SS	MS	Source	d.f.	SS	MS
Regression (X_1, X_2, X_3)	3	693.06	231.02	Regression (X_1)	1	588.92	588.92
Residual	8	195.19	24.40	Residual	10	299.33	29.93
Total	11	888.25		Total	11	888.25	

© Cengage Learning

Another example is a collection of two-way product terms (e.g., X_1X_2, X_1X_3, X_2X_3); this latter group is sometimes referred to as a set of interaction variables (see Chapter 11). It is often of interest to assess the importance of interaction effects collectively before trying to consider individual interaction terms in a model. In fact, the initial use of such a chunk test can reduce the total number of tests to be performed, since variables may be dropped from the model as a group. This, in turn, helps provide better control of overall Type I error rates, which may be inflated due to multiple testing (Kupper, Stewart, and Williams 1976; Abt 1981).

9.5 Strategies for Using Partial F Tests

In applying the ideas presented in this chapter, readers will typically use a computer program to carry out the numerical calculations required. Therefore, we will briefly describe the computer output for typical regression programs. To help readers understand and use such output, we discuss two strategies for using partial F tests: *variables-added-in-order tests* and *variables-added-last tests*.

The accompanying computer output is from a typical regression computer program² for the model

$$\text{WGT} = \beta_0 + \beta_1 \text{HGT} + \beta_2 \text{AGE} + \beta_3 (\text{AGE})^2 + E$$

The results here were computed with *centered* predictors³ (see Section 14.5.2), so $(\text{HGT} - 52.75)$, $(\text{AGE} - 8.833)$, and $(\text{AGE} - 8.833)^2$ were used, with mean $\text{HGT} = 52.75$ and mean $\text{AGE} = 8.833$. The computer output consists of five sections, labeled A through E. Section A provides the overall ANOVA table for the regression model. Computer output typically presents numbers with far more significant digits than can be justified. Section B provides the multiple R^2 -value, the coefficient of variation ($100S/\bar{Y}$), the WGT residual standard deviation S (i.e., “Root MSE”), and the mean (\bar{Y}) of the dependent variable WGT.

² This particular output was produced by the SAS program GLM.

³ *Centering* is the process of transforming a variable—say, AGE—by subtracting from AGE its sample mean, e.g., 8.833 in the preceding data. The newly defined centered variable $(\text{AGE} - 8.833)$ will then have zero as its overall mean. Centering is frequently done to gain computational accuracy when estimating higher-order (e.g., polynomial) terms in a regression model. See Section 14.5.2 for further discussion.

Edited SAS Output for Data from Table 8.1 (HGT and AGE Centered)

Dependent Variable: Weight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	693.0604634	231.0201545	9.47	0.0052
Error	8	195.1895366	24.3986921		
Corrected Total	11	888.2500000			
R-Square		Coeff Var	Root MSE	WGT Mean	
0.780254		7.871718	4.939503	62.75000	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
HGT	1	588.9225232	588.9225232	24.14	0.0012
AGE	1	103.9000834	103.9000834	4.26	0.0730
AGESQ	1	0.2378569	0.2378569	0.01	0.9238
Source	DF	Type III SS	Mean Square	F Value	Pr > F
HGT	1	166.5819549	166.5819549	6.83	0.0310
AGE	1	101.8175237	101.8175237	4.17	0.0753
AGESQ	1	0.2378569	0.2378569	0.01	0.9238
Parameter	Estimate	Standard Error	t Value	Pr > t	
Intercept	62.88718379	1.99573190	31.51	<.0001	
HGT	0.72369024	0.27696316	2.61	0.0310	
AGE	2.04008401	0.99866553	2.04	0.0753	
AGESQ	-0.04170670	0.42240715	-0.10	0.9238	

© Cengage Learning

Section C provides certain (Type I) tests for assessing the importance of each predictor in the model; section D provides a different set of (Type III) tests regarding these predictors; and section E provides yet a third set of (t) tests.

9.5.1 Basic Principles

Two methods (or strategies) are widely used for evaluating whether a variable should be included in a model: partial (Type I) F tests for variables added in order and partial (Type III) F tests for variables added last.⁴ For the first (variables-added-in-order) method, the following procedure is employed: (1) an order for adding variables one at a time is specified; (2) the significance of the (straight-line) model involving only the variable ordered first is assessed;

⁴ Searle (1971), among others, refers to these methods as "ignoring" and "eliminating" tests, respectively.

Instead of considering variables added in order, it may be of interest to consider *variables deleted in order*. The latter strategy would apply, for example, in polynomial regression, where a backward selection algorithm is used to determine the proper degree of the polynomial. As can be seen from the accompanying computer output, the same set of sums of squares is produced whether the variables are considered to be added in one order or deleted in the reverse order.

(3) the significance of adding the second variable to the model involving only the first variable is assessed; (4) the significance of adding the third variable to the model containing the first and second variables is assessed; and so on.

For the second (variables-added-last) method, the following procedure is used: (1) an initial model containing two or more variables is specified; (2) the significance of each variable in the initial model is assessed separately, as if it were the last variable to enter the model (i.e., if k variables are in the initial model, then k variables-added-last tests are conducted). In either method, each test is conducted using a partial *F* test for the addition of a single variable.

Variables-added-in-order tests can be illustrated with the weight example. One possible ordering is HGT first, followed by AGE and then $(AGE)^2$. For this ordering, the smallest model considered is

$$WGT = \beta_0 + \beta_1 HGT + E$$

The overall regression *F* test of $H_0: \beta_1 = 0$ is used to assess the contribution of HGT. Next, the model

$$WGT = \beta_0 + \beta_1 HGT + \beta_2 AGE + E$$

is fit. The significance of adding AGE to a model already containing HGT is then assessed by using the partial $F(AGE|HGT)$. Finally, the full model is fit by using HGT, AGE, and $(AGE)^2$. The importance of the last variable is tested with the partial $F[(AGE)^2|HGT, AGE]$. The tests used are those discussed in this chapter and summarized in Table 9.1. These are related to the tests provided in section C of the earlier computer output (using Type I sums of squares). However, each test in Table 9.1 involves a different residual sum of squares, while those in the computer output use a common residual sum of squares. More will be said about this issue shortly.

To describe variables-added-last tests, consider again the full model

$$WGT = \beta_0 + \beta_1 HGT + \beta_2 AGE + \beta_3 (AGE)^2 + E$$

The contribution of HGT, when added last, is assessed by comparing the full model to the model with HGT deleted—namely,

$$WGT = \beta_0 + \beta_2 AGE + \beta_3 (AGE)^2 + E$$

The partial *F* statistic, based on (9.4), has the form $F[HGT|AGE, (AGE)^2]$. The sum of squares for HGT added last is then the appropriate difference in the error sum of squares (or the regression sum of squares) for the two preceding models. Similarly, the reduced model with AGE deleted is

$$WGT = \beta_0 + \beta_1 HGT + \beta_3 (AGE)^2 + E$$

for which the corresponding partial *F* statistic is $F[AGE|HGT, (AGE)^2]$, and the reduced model with $(AGE)^2$ omitted is

$$WGT = \beta_0 + \beta_1 HGT + \beta_2 AGE + E$$

for which the partial F statistic is $F[(\text{AGE})^2 | \text{HGT}, \text{AGE}]$. The three F statistics just described are provided in section D of the earlier computer output (using Type III sums of squares).⁵

An important characteristic of variables-added-in-order sums of squares is that they decompose the regression sum of squares into a set of mutually exclusive and exhaustive pieces. For example, the sums of squares provided in section C of the computer output (588.922, 103.900, and 0.238) add to 693.060, which is the regression sum of squares given in section A. The variables-added-last sums of squares do not generally have this property (e.g., the sums of squares given in section D of the computer output do not add to 693.060).

Each of these two testing strategies has its own advantages, and the situation being considered determines which is preferable. For example, if all variables are considered to be of equal importance, the variables-added-last tests are usually preferred. Such tests treat all variables equally in the sense that each variable is assessed as if it were the last variable to enter the model.

In contrast, if the order in which the predictors enter the model is an important consideration, the variables-added-in-order testing approach may be better. An example where the entry order is important is one where main effects (e.g., X_1 , X_2 , and X_3) are forced into the model, followed by their cross-products (X_1X_2 , X_1X_3 , and X_2X_3) or so-called interaction terms (see Chapter 11). Such tests evaluate the contribution of one or more variables and adjust *only* for the variables just preceding them into the model.

9.5.2 Commentary

As discussed in the preceding subsection, section C of the earlier computer output provides variables-added-in-order tests, which are also given for the same data in Table 9.1. Section D of the computer output provides variables-added-last tests. Finally, section E provides t tests (which are equivalent to the variables-added-last F tests in section D), as well as regression coefficient estimates and their standard errors, for the *centered* predictor variables.

Table 9.3 gives an ANOVA table for the variables-added-last tests for the weight example. (We recommend that readers consider how this table was extracted from the earlier computer output.) The variables-added-last tests usually give a different ANOVA table from one based on the variables-added-in-order tests. A different residual sum of squares is used for each variables-added-in-order test in Table 9.1, while the same residual sum of squares [based on the three-variable model involving the *centered* versions of HGT, AGE, and $(\text{AGE})^2$] is used for all the variables-added-last tests in Table 9.3.

An argument can be made that it is preferable to use the residual sum of squares for the three-variable model (i.e., the largest model containing all candidate predictors) for all tests. This is because the error variance σ^2 will *not* be correctly estimated by a model that ignores important predictors, but it will be correctly estimated (under the usual regression assumptions) by a model that contains all candidate predictors (even if some are not important). In other words, overfitting a model in estimating σ^2 is safer than underfitting it. Of course,

⁵ It is important to remember that the preceding computer output was based on the use of the *centered* predictor variables ($\text{HGT} - 52.75$), ($\text{AGE} - 8.833$), and ($\text{AGE} - 8.833$)² and not on the use of the original (uncentered) predictor variables HGT, AGE, and $(\text{AGE})^2$. As a result, certain numerical results (specifically, some of those in sections D and E) for the centered predictors differ from those for the uncentered predictors (see, e.g., the computer output for Model 6 on page 151 in Chapter 8).

TABLE 9.3 ANOVA table for WGT regressed on HGT, AGE, and (AGE)², using variables-added-last tests

Source	d.f.	SS (Type III)	MS	F	R ²
X ₁ X ₂ , X ₃	1	166.58	166.58	6.83*	0.7802
X ₂ X ₁ , X ₃	1	101.81	101.82	4.17	
X ₃ X ₁ , X ₂	1	0.24	0.24	0.01	
Residual	8	195.19	24.40		
Total	11	888.25			

*Exceeds .05 critical value of 5.32 for F with 1 and 8 degrees of freedom.

© Cengage Learning

extreme overfitting results in lost precision, but it still provides a valid estimate of residual variation. We generally prefer using the residual sum of squares based on fitting the “largest” model, although some statisticians disagree.

9.5.3 Models Underlying the Source Tables

Tables 9.4 and 9.5 present the models being compared based on the earlier computer output and the associated ANOVA tables. Table 9.4 summarizes the models and residual sums of squares needed to conduct variables-added-last tests for the full model containing the *centered* versions of HGT, AGE, and (AGE)². Table 9.5 lists the models that must be fitted to provide variables-added-in-order tests for the order of entry HGT, then AGE, and then (AGE)², again using *centered* versions of these predictors.

Table 9.6 details computations of regression sums of squares for both types of tests. For example, the first line, where 24,226.2637 = 24,421.4532 – 195.1895, is the difference in the error sums of squares for models 1 and 5 given in Table 9.4. These results can then be used to produce any of the F tests given in Tables 9.1 and 9.2 and in the computer output.

TABLE 9.4 Variables-added-last regression models and residual sums of squares for data from Table 8.1

Model No.	Model	SSE
1	WGT = β_1 HGT + β_2 AGE + β_3 (AGE) ² + E	24,421.4532
2	WGT = β_0 + β_2 AGE + β_3 (AGE) ² + E	361.7715
3	WGT = β_0 + β_1 HGT + β_3 (AGE) ² + E	297.0071
4	WGT = β_0 + β_1 HGT + β_2 AGE + E	195.4274
5	WGT = β_0 + β_1 HGT + β_2 AGE + β_3 (AGE) ² + E	195.1895

© Cengage Learning

TABLE 9.5 Variables-added-in-order regression models and residual sums of squares for data from Table 8.1

Model No.	Model	SSE
6	WGT = + E	48,139.0000
7	WGT = β_0 + E	888.2500
8	WGT = $\beta_0 + \beta_1 HGT$ + E	299.3275
4	WGT = $\beta_0 + \beta_1 HGT + \beta_2 AGE$ + E	195.4274
5	WGT = $\beta_0 + \beta_1 HGT + \beta_2 AGE + \beta_3 (AGE)^2$ + E	195.1895

© Cengage Learning

TABLE 9.6 Computations for regression sums of squares for data from Table 8.1

Parameter	Variable	Regression SS	
		Added Last	In Order
β_0	Intercept	SSE(1) – SSE(5) = 24,226.2637	SSE(6) – SSE(7) = 47,250.7500
β_1	HGT	SSE(2) – SSE(5) = 166.5820	SSE(7) – SSE(8) = 588.9225
β_2	AGE	SSE(3) – SSE(5) = 101.8176	SSE(8) – SSE(4) = 103.9001
β_3	(AGE) ²	SSE(4) – SSE(5) = 0.2379	SSE(4) – SSE(5) = 0.2379

© Cengage Learning

9.6 Additional Inference Methods for Multiple Regression

Until this point, our inferential methods for multiple linear regression have been focused on hypothesis tests concerning whether an independent variable (or a set of independent variables) significantly improves prediction of the response variable. In this section, we consider hypothesis testing and confidence interval methods concerning individual regression coefficients, linear functions of regression coefficients, and mean and predicted responses.

9.6.1 Tests Involving the Intercept

Inferences about the intercept β_0 are occasionally of interest in multiple regression analysis. A test of $H_0: \beta_0 = 0$ is usually carried out with an intercept-added-last test, although an intercept-added-in-order test is also feasible (where the intercept is the first term added to the model). Many computer programs provide only a t test involving the intercept. The t -test

statistic for the intercept in the earlier computer output corresponds exactly to a partial F test for adding the intercept last. The two models being compared are

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + E$$

and

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + E$$

The null hypothesis of interest is $H_0: \beta_0 = 0$ versus $H_A: \beta_0 \neq 0$. The test is computed as

$$F = \frac{(\text{SSE without } \beta_0 - \text{SSE with } \beta_0)/1}{(\text{SSE with } \beta_0)/(n - k - 1)}$$

This F statistic has 1 and $n - k - 1$ degrees of freedom and is equal to the square of the t statistic used for testing $H_0: \beta_0 = 0$. For the weight example involving the *centered* predictors, an intercept-added-last test is reported in the output on page 176 as a t test. The corresponding partial F equals $(31.51)^2 = 992.88$ and has 1 and 8 degrees of freedom.

An intercept-added-in-order test can also be conducted. In this case, the two models being compared are

$$Y = E$$

and

$$Y = \beta_0 + E$$

Again, the null hypothesis is $H_0: \beta_0 = 0$ versus the alternative $H_A: \beta_0 \neq 0$. The special nature of this test leads to the simple expression

$$F = \frac{n\bar{Y}^2}{\text{SSY}/(n - 1)}$$

which represents an F statistic with 1 and $n - 1$ degrees of freedom. This statistic involves SSY, the residual sum of squares from a model with just an intercept (such as model 7 in Table 9.5). Alternatively, the residual sum of squares from the “largest” model may be used. When we use the latter approach, the F statistic for the weight data becomes (see Table 9.5)

$$\begin{aligned} F &= \frac{[\text{SSE}(6) - \text{SSE}(7)]/1}{\text{SSE}(5)/8} \\ &= \frac{(48,139.00 - 888.25)/1}{195.19/8} \\ &= 1936.61 \end{aligned}$$

where $\text{SSE}(5)$ denotes the residual (i.e., error) sum of squares for model 5 (which is the largest of the models in Table 9.5); this F statistic has 1 and 8 degrees of freedom. In general, using the residual from the largest model (with k predictors) gives $n - k - 1$ error degrees of freedom, so the F statistic is compared to a critical value with 1 and $n - k - 1$ degrees of freedom.

9.6.2 Confidence Intervals about Regression Coefficients

Recall, from Section 9.3, that a variables-added-last partial F test can be used to assess whether a given predictor significantly improves the prediction of Y , given that other predictors are already in the model. Further, this partial F test is equivalent to a t test of the null hypothesis that the β coefficient for that added-last predictor is equal to 0. In addition to this partial F test (or the equivalent t test), it is strongly recommended that a confidence interval for the parameter β be constructed. The important advantage of a confidence interval over a hypothesis test is that the confidence interval provides a quantitative measure of the precision with which the parameter β is being estimated (i.e., the wider the confidence interval, the less the precision).

The $100(1 - \alpha)\%$ confidence interval for the coefficient β^* for a given predictor X^* is computed using the formula

$$\hat{\beta}^* \pm t_{n-k-1, 1-\frac{\alpha}{2}}(S_{\beta^*}) \quad (9.7)$$

Recall the nutritional deficiency example, in which Model 4 considered WGT as a function of HGT and AGE according to the model $WGT = \beta_0 + \beta_1 HGT + \beta_2 AGE + E$; numerical results based on Model 4 are given in Section 8.7. For this model, we calculate the 95% confidence interval for β_2 (a parameter representing the change in WGT [in pounds] for a 1-inch change in HGT) as follows:

$$95\% \text{ CI} = 2.0501 \pm t_{9, 0.975}(0.9372) = (-0.0698, 4.1700)$$

Note that the value 0 is barely contained in this 95% confidence interval; this finding supports the P -value of .0595 for the partial F test for the improvement in the prediction of WGT when adding AGE to a model already containing the predictor HGT. Sometimes the confidence interval for a 1-unit increase in a factor may not be of practical relevance (e.g., HGT if expressed in mm). If a confidence interval for any specified change in a predictor is desired (such as a 5-unit change or 100-unit change), one simply multiplies the lower and upper confidence interval limits for a 1-unit change by the specified value of the change of interest. As we will see later in this section, confidence interval construction is more complicated when the parameter of interest is a linear combination of regression coefficients.

9.6.3 The Mean Value of Y at Specified Values of X_1, X_2, \dots, X_k

As in the special case of straight-line regression (see Section 5.9), we may wish to compute the estimated value of the response Y for some specific set of values $X_{1,0}, X_{2,0}, \dots, X_{k,0}$ for the predictor variables; this estimated response is represented notationally as $\hat{Y}_{X_{1,0}, X_{2,0}, \dots, X_{k,0}}$. We can then use this estimated response, along with appropriate standard error estimates, to make statistical inferences about the true mean response and predicted response for any specified set of values of the predictor variables. In contrast to simple linear regression (see Section 5.9), the plotting of confidence bands and prediction bands in multiple linear regression is quite complicated, since the number of variables involved makes the setting multidimensional; such multivariable graphical methods are beyond the scope of this textbook.

Like their straight-line regression counterparts, the hypothesis test and confidence interval procedures for making statistical inferences about the mean value of Y , or the predicted

value of Y , involve the t distribution. In the multivariable situation, the computation of the required estimated standard errors for these procedures is not straightforward, and computer packages are typically used to compute these estimated standard errors.

Hypothesis Tests

To test the null hypothesis

$$H_0: \mu_{Y|X_1, 0, X_2, 0, \dots, X_k, 0} = \mu_{Y|X_1, 0, X_2, 0, \dots, X_k, 0}^{(0)}$$

that the mean value of Y equals a hypothesized value at a given set of values for the independent variables, we use the following t -test statistic, which has $(n - k - 1)$ degrees of freedom:

$$T = \frac{\hat{Y}_{X_1, 0, X_2, 0, \dots, X_k, 0} - \mu_{Y|X_1, 0, X_2, 0, \dots, X_k, 0}^{(0)}}{S_{\hat{Y}_{X_1, 0, X_2, 0, \dots, X_k, 0}}} \quad (9.8)$$

Referring again to Model 4 for the nutritional deficiency example in Section 8.2, suppose that we want to test the null hypothesis that the true average weight is 60 pounds for nutritionally deficient 9-year-old members of the sampled population who are 55 inches tall. Using the numerical results, we compute the predicted response as

$$\begin{aligned}\hat{Y}_{X_1, 0, X_2, 0} &= \widehat{\text{WGT}}_{\text{HGT}_0, \text{AGE}_0} = \hat{\beta}_0 + \hat{\beta}_1(\text{HGT}_0) + \hat{\beta}_2(\text{AGE}_0) \\ &= \widehat{\text{WGT}}_{(\text{HGT}=55, \text{AGE}=9)} = 6.5530 + 0.7220(55) + 2.0501(9) = 64.7139\end{aligned}$$

Using computer-generated standard error estimates (see the SAS Computer Appendix C for details), we find that $S_{\hat{Y}_{X_1, 0, X_2, 0, \dots, X_k, 0}} = S_{\widehat{\text{WGT}}_{(\text{HGT}=55, \text{AGE}=9)}} = 1.4373$

The t -test statistic then has the value

$$T = \frac{\hat{Y}_{X_1, 0, X_2, 0, \dots, X_k, 0} - \mu_{Y|X_1, 0, X_2, 0, \dots, X_k, 0}^{(0)}}{S_{\hat{Y}_{X_1, 0, X_2, 0, \dots, X_k, 0}}} = \frac{64.7139 - 60}{1.4373} = 3.2797$$

For the t_9 distribution, this test statistic has a P -value $< .005$, and thus we would reject the null hypothesis of a mean weight of 60 using a two-sided test at the $\alpha = .005$ level.

Confidence Intervals

The $100(1 - \alpha)\%$ confidence interval for $\mu_{Y|X_1, 0, X_2, 0, \dots, X_k, 0}$ is constructed using the formula

$$\hat{Y}_{X_1, 0, X_2, 0, \dots, X_k, 0} \pm t_{n-k-1, 1-\frac{\alpha}{2}}(S_{\hat{Y}_{X_1, 0, X_2, 0, \dots, X_k, 0}}) \quad (9.9)$$

Continuing the above example, we compute the 95% confidence interval for the true mean weight of nutritionally deficient 9-year-olds who are 55 inches tall as follows:

$$\hat{Y}_{X_1, 0, X_2, 0, \dots, X_k, 0} \pm t_{n-k-1, 1-\frac{\alpha}{2}}(S_{\hat{Y}_{X_1, 0, X_2, 0, \dots, X_k, 0}}) = 64.7139 \pm 2.2622(1.4373) = (61.46, 67.97)$$

9.6.4 Prediction of a New Value of Y at $X_{1,0}, X_{2,0}, \dots, X_{k,0}$

Rather than calculate a confidence interval for the true mean of Y at a specified set of predictor values, we may also want to compute an interval (called a *prediction interval*) for a new (as yet unobserved) value of Y .

As explained in Section 5.10, the estimated mean response $\hat{Y}_{X_{0,1}, X_{0,2}, \dots, X_{0,k}}$ computed from the regression equation at the desired values of $X_{1,0}, X_{2,0}, \dots, X_{k,0}$ is also the estimated predicted value of Y at those values. And as discussed, the variability of predicted individual values is necessarily greater than that of estimated mean values.

This larger variance may be partitioned into two sources of variability: the distance from the individual response to the true population regression line (deviation of Y from the true mean of Y at the specified X -values) and the inaccuracy of the estimated regression model (deviation of $\hat{Y}_{X_{0,1}, X_{0,2}, \dots, X_{0,k}}$ from the true mean of Y at the specified X -values). As a simple generalization of results given in Section 5.10, the $100(1 - \alpha)\%$ prediction interval is

$$\hat{Y}_{X_{1,0}, X_{2,0}, \dots, X_{k,0}} \pm t_{n-k-1, 1-\frac{\alpha}{2}} \sqrt{(S^2_{Y|X_{1,0}, X_{2,0}, \dots, X_{k,0}}) + (S^2_{\hat{Y}_{X_{1,0}, X_{2,0}, \dots, X_{k,0}}})} \quad (9.10)$$

where $S^2_{Y|X_{1,0}, X_{2,0}, \dots, X_{k,0}}$ is equivalent to MS Residual, as described earlier.

Returning to the nutritional deficiency example involving Model 4 in Chapter 8, the computed 95% prediction interval for the weight of a randomly selected 9-year-old who is 55 inches tall is

$$64.7139 \pm 2.2622 \sqrt{(21.7142) + (1.4373)^2} = (53.68, 75.75)$$

9.6.5 Inference Methods for Linear Functions of Regression Coefficients

It is often of interest to consider statistical inferences about a *linear sum*

$$L = c_1\beta_1 + c_2\beta_2 + \dots + c_k\beta_k$$

of regression coefficients, where c_1, c_2, \dots, c_k are specified constants.

Consider a study where varying doses of two drugs A and B in milligrams (mg) were prescribed to lower systolic blood pressure (SBP). Consider the following situations:

- Suppose that it is conjectured that the amount of drug A given is twice as effective as the amount of drug B with regard to lowering SBP. This conjecture can be addressed by considering the null hypothesis $H_0: \beta_A = 2\beta_B$.
- We may be interested in estimating *how much* SBP decreases, on average, as a result of 1 mg (or larger) doses of both drugs. In the 1 mg case, we would want to estimate the quantity $(\beta_A + \beta_B)$ and construct a confidence interval for this unknown parameter.
- We may wish to assess whether or not the joint effect of 1 mg doses of both drugs equals a hypothesized number q , where q is a specified mean decrease in SBP. This assessment could be made via a hypothesis test of $H_0: \beta_A + \beta_B = q$.

All of the above situations can be addressed using the same general approach. The quantity of interest is expressed as a linear function L of the β coefficients, with the estimate of L denoted as \hat{L} . In the case of hypothesis tests (situations 1 and 3 above), we typically rearrange L as an expression equal to 0. For example, for situation 1, $L = \beta_A - 2\beta_B = 0$ and $\hat{L} = \hat{\beta}_A - 2\hat{\beta}_B$. Thus the c_i values corresponding to β_A and β_B would be 1 and -2 , respectively, and are equal to 0 for any remaining *regression* coefficients considered in the model. More generally, when $(c_1 + c_2 + \dots + c_k) = 0$, the linear function is called a *linear contrast*.

Hypothesis tests about, and confidence intervals for, L utilize the estimated value \hat{L} and its estimated standard error $S_{\hat{L}}$. Hypothesis tests use the test statistic $\frac{\hat{L}}{S_{\hat{L}}}$, which follows the t_{n-k-1} distribution under the null hypothesis that $L = 0$. Confidence intervals are constructed using the formula $\hat{L} \pm t_{n-k-1, 1-\frac{\alpha}{2}}(S_{\hat{L}})$.

A key difference from earlier is that the standard error of \hat{L} , $S_{\hat{L}}$ —or, equivalently, its variance, $S_{\hat{L}}^2$ —is not readily obtained from standard output and requires extra calculations and/or programming. The formula for $S_{\hat{L}}^2 = \widehat{\text{Var}}(\hat{L})$ is

$$S_{\hat{L}}^2 = \sum_{i=1}^k c_i^2 S_{\hat{\beta}_i}^2 + 2 \sum_{i=1}^{k-1} \sum_{j=i+1}^k c_i c_j \widehat{\text{cov}}(\hat{\beta}_i, \hat{\beta}_j) \quad (9.11)$$

Note that this complicated expression involves the k estimated variances of the estimated regression coefficients and the estimated covariances⁶ for each of the $k(k-1)/2$ pairs of estimated regression coefficients. The individual variance estimates may be found by squaring the estimated standard errors of the estimated regression coefficients that are given in standard computer output, but the estimated covariances need to be obtained by further requesting the estimated variance–covariance matrix.⁷

We now provide an application of linear contrasts, again using Model 4 for the nutritional deficiency example. Suppose we want to assess the increase in weight associated with both a 1-inch increase in HGT and a 1-year increase in AGE. Thus, the parameter of interest is the linear function $L = (\beta_1 + \beta_2)$. From the Model 4 output, $\hat{L} = 0.7220 + 2.0501 = 2.7721$. Squaring the estimates of the standard errors of the estimated regression coefficients provided in the output, we obtain $S_{\hat{\beta}_1}^2 = 0.0680$ and $S_{\hat{\beta}_2}^2 = 0.8784$. Using the variance–covariance matrix output below, we locate the cell representing HGT and AGE (i.e., either the cell entry for row 2 and column 3 or the cell entry for row 3 and column 2) to find $\widehat{\text{cov}}(\hat{\beta}_1, \hat{\beta}_2) = -0.1500$. Note that the values of $S_{\hat{\beta}_1}^2$ and $S_{\hat{\beta}_2}^2$ also appear on the diagonal of this matrix.

Using this information, we find $S_{\hat{L}}^2 = 0.0680 + 0.8784 + 2(-0.1500) = 0.6464$, so that the 95% confidence interval for L is $2.7721 \pm 2.2622\sqrt{0.6464} = (0.9533, 4.5909)$.

⁶ The covariance between two variables is defined as the expected value of the product of the two variables minus the product of their individual expected values (see Chapter 25 for more details).

⁷ The estimated variance–covariance matrix for a regression model containing $(k+1)$ predictors is a $(k+1) \times (k+1)$ symmetric matrix; the diagonal elements of this square matrix are the values of the estimated variances of the estimated regression coefficients, and the off-diagonal elements are the values of the estimated covariances (see Chapter 25 and Appendix B for further discussions).

COVARIANCE OF ESTIMATES			
Variable	Intercept	HGT	AGE
Intercept	119.78923971	-2.262641341	0.1556155374
HGT	-2.262641341	0.0680192799	-0.150042529
AGE	0.1556155374	-0.150042529	0.8783918353

Fortunately for the analyst, most regression packages, including SAS, are capable of conducting hypothesis testing and interval estimation for linear functions, some of which are described in Appendix C.

Linear contrasts appear in many applications in other regression contexts, particularly in epidemiological examples and those involving nominal variables. We revisit linear contrasts in the discussions of ANACOVA in Chapter 13, ANOVA in Chapter 17, and logistic regression in Chapter 22.

9.7 Example: BRFSS Analysis

In the multiple regression model fit to the BRFSS data at the end of Chapter 8, drinking frequency, age, and sleep quality were together considered in the prediction of BMI. Several questions come to mind that can be answered using the hypothesis testing techniques covered in this chapter:

- a. Do these three factors together significantly predict BMI? (*Section 9.2*)
- b. For a model containing only the single predictor drinking frequency, does adding age and sleep quality significantly improve the prediction of BMI? (*Sections 9.3–9.5*)
- c. If the answer is yes to the question in part (b), does one of these two predictors appear to be more important than the other? (*Sections 9.3–9.5*)

Additional model output is provided to answer these questions:

Source	DF	Sum of Squares	Mean Square	F Value
Model	3	1289.79375	429.93125	12.46
Error	1045	36065.26233	34.51221	
Corrected Total	1048	37355.05608		

Source	DF	Type I SS	Type III SS
drink_days	1	1073.403915	1068.207272
AGE	1	11.595517	42.497569
poor_sleep_days	1	204.794319	204.794319

- a. To assess whether the model containing all three predictors significantly predicts BMI, the overall F test should be used. Representing the fitted model as

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1(\text{drink_days}) + \hat{\beta}_2(\text{age}) + \hat{\beta}_3(\text{poor_sleep_days})$, the null hypothesis is $H_0: \beta_1 = \beta_2 = \beta_3 = 0$, and the appropriate test statistic is given by equation (9.1). For these data, the test statistic value is

$$F = \frac{\text{MS Regression}}{\text{MS Residual}} = \frac{429.93}{34.51} = 12.46$$

If the null hypothesis were true, this F statistic would follow an $F_{3, 1045}$ distribution. From Table A.4 of Appendix A, we see that the $\alpha = .001$ cutoff is approximately 5.46, meaning the observed value of 12.46 highly supports rejecting H_0 and concluding that one or more of these factors is significantly associated with BMI.

- b. We next perform a test of significance for the addition of both age and sleep quality given a model that already contains the predictor drinking frequency. More formally, this is a test of $H_0: \beta_2 = \beta_3 = 0$ in the full model containing all three predictors. We can perform this hypothesis test using a multiple partial F test, comparing the regression sums of squares in models with and without the new factors, using the formulas presented in Section 9.4.2 and the variables-added-in-order sums of squares (Type I). The numerical value of this multiple partial F test is

$$\begin{aligned} & F(\text{age, sleep quality} | \text{drinking freq}) \\ &= \frac{[\text{Regression SS(full)} - \text{Regression SS(drinking freq only)}]/2}{\text{MS Residual(full)}} \\ &= \frac{(1,289.79 - 1,073.40)/2}{34.51} = 3.14 \end{aligned}$$

Note that we can compute an equivalent test statistic by using an alternative formulation that directly evaluates the two new independent variables under consideration. The numerator in this test statistic utilizes the fact that the Type I sums of squares represent a perfect partitioning of the full model regression sums of squares into the unique sequentially ordered regression-sum-of-squares contribution of each predictor:

$$\begin{aligned} & \frac{[\text{Regression SS(sleep quality} | \text{age, drinking freq}) + \text{Regression SS(age} | \text{drinking freq})]/2}{\text{MS Residual(full)}} \\ &= \frac{(204.79 + 11.60)/2}{34.51} = 3.14 \end{aligned}$$

Since there are $k = 2$ parameters being set equal to zero under the null hypothesis, the test statistic follows an $F_{2, 1045}$ distribution if the null hypothesis is true. The P -value is between .025 and .05, per Table A.4 of Appendix A (with a more precise P -value computed to be .044). Thus, we would reject H_0 at an α level of .05 but

recognize that this test statistic is only marginally significant and far less so than that of the overall F test (which considered drinking frequency as a third predictor). Taken together, these results suggest that drinking frequency was responsible for the highly significant overall F -test finding and so is perhaps the most significant predictor of BMI.

- c. Having found age and sleep quality to be collectively marginally significant for the prediction of BMI, a natural next question concerns whether or not both predictors are actually needed in the model. It is best to make this assessment using the second method of Section 9.5.1, the variables-added-last tests, to evaluate the significance of each factor given that the other two predictors are already in the model. These two partial F tests use the Type III sums of squares and are evaluated using the $F_{1, 1045}$ distribution:

$$F(\text{age}|\text{drinking freq, sleep quality}) = \frac{\text{SS}(\text{age}|\text{drinking freq, sleep quality})}{\text{MS Residual(full)}} \\ = \frac{42.50}{34.51} = 1.23$$

$$F(\text{sleep quality}|\text{drinking freq, age}) = \frac{\text{SS}(\text{sleep quality}|\text{drinking freq, age})}{\text{MS Residual(full)}} \\ = \frac{204.79}{34.51} = 5.93$$

These two test statistics have P -values of .27 and .015, respectively, and thus we conclude that sleep quality is the only factor of the two that is significant after adjustment for the other two predictors. Based on this information, many researchers would not include the factor age in the final model that is reported. Various strategies used for variable selection in multiple regression models are discussed further in Chapter 16.

Problems

1. Use the data of Problem 1 in Chapter 8 to answer the following questions.
 - a. Conduct the overall F tests for significant regression for the three models in Problem 1 of Chapter 8. Be sure to state the null and alternative hypotheses for each test and interpret each result.
 - b. Based on your results in part (a), which of the three models is best? Compare your answer here with your answer to Problem 1(b) in Chapter 8.
2. Use the information given in Problem 2 of Chapter 8, as well as the computer output given here, to answer the following questions about the data from that problem.
 - a. Conduct overall regression F tests for these three models: Y regressed on X_1 and X_2 ; Y regressed on X_1 alone; and Y regressed on X_2 alone. Based on your

answers, how would you rate the importance of the two variables in predicting Y ? Comment on how your answer compares with the answer to Problem 2(b) in Chapter 8.

- b. Provide variables-added-in-order tests for both variables, with thinking disturbances (X_1) added first. Use $\alpha = .05$.
- c. Provide variables-added-in-order tests for both variables, with hostile suspiciousness (X_2) added first. Use $\alpha = .05$.
- d. Provide a table of variables-added-last tests.
- e. What, if any, differences are present in the approaches in parts (a) through (d)?
- f. Which predictors appear to be necessary? Why?

Edited SAS Output (PROC GLM) for Problem 2

Y Regressed on X1 and X2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2753.87136	1376.93568	6.24	0.0038
Error	50	11037.29845	220.74597		
Corrected Total	52	13791.16981			
<hr/>					
R-Square	Coeff Var	Root MSE	Y Mean		
0.199684	65.45708	14.85752	22.69811		
Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	1535.856965	1535.856965	6.96	0.0111
X2	1	1218.014394	1218.014394	5.52	0.0228
Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	2596.024025	2596.024025	11.76	0.0012
X2	1	1218.014394	1218.014394	5.52	0.0228
Parameter		Estimate	Standard Error	t Value	Pr > t
Intercept		-0.63535020	20.96833367	-0.03	0.9759
X1		23.45143521	6.83850964	3.43	0.0012
X2		-7.07260920	3.01092434	-2.35	0.0228

3. A psychologist examined the regression relationship between anxiety level (Y)—measured on a scale ranging from 1 to 50, as the average of an index determined at three points in a 2-week period—and the following three independent variables: X_1 = systolic blood pressure; X_2 = IQ; and X_3 = job satisfaction (measured on a scale ranging from 1 to 25). The following ANOVA table summarizes results obtained from a variables-added-in-order regression analysis on data involving 22 outpatients who were undergoing therapy at a certain clinic.

Source	d.f.	SS
Regression	1	981.326
	1	190.232
	1	129.431
Residual	18	442.292

- a. Test for the significance of each independent variable as it enters the model. State the null hypothesis for each test in terms of regression coefficient parameters.
- b. Test for the significance of adding both X_2 and X_3 to a model already containing X_1 . State the null hypothesis in terms of regression coefficient parameters.
- c. In terms of regression sums of squares, identify the test that corresponds to comparing the two models

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$$

and

$$Y = \beta_0 + \beta_3 X_3 + E$$

Why can't this test be done by using the ANOVA table? Describe the appropriate test procedure.

- d. Based on the tests made, what would you recommend as the most appropriate statistical model? Use $\alpha = .05$.
- 4. An educator examined the relationship between the number of hours devoted to reading each week (Y) and the independent variables social class (X_1), number of years of school completed (X_2), and reading speed (X_3), in pages read per hour. The following ANOVA table was obtained from a stepwise regression analysis on data for a sample of 19 women over 60.

Source	d.f.	SS
Regression	1	1,058.628
	1	183.743
	1	37.982
Residual	15	363.300

- a. Test for the significance of each variable as it enters the model.
- b. Test $H_0: \beta_1 = \beta_2 = 0$ in the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$.
- c. Why can't we test $H_0: \beta_1 = \beta_3 = 0$ by using the ANOVA table given? What formula would you use for this test?
- d. Based on your results in parts (a) and (b), what is the most appropriate model to use?
- 5. An experiment was conducted regarding a quantitative analysis of factors found in high-density lipoprotein (HDL) in a sample of human blood serum. Three variables thought to be predictive of, or associated with, HDL measurement (Y) were the total cholesterol (X_1) and total triglyceride (X_2) concentrations in the sample, plus the presence or absence of a certain sticky component of the serum called sinking pre-beta or SPB (X_3), coded as 0 if absent and 1 if present. The data obtained are shown in the following table.

- a. Test whether X_1 , X_2 , or X_3 alone significantly helps to predict Y .
 b. Test whether X_1 , X_2 , and X_3 taken together significantly help to predict Y .

Y	X_1	X_2	X_3	Y	X_1	X_2	X_3	Y	X_1	X_2	X_3
47	287	111	0	63	339	168	1	36	318	180	0
38	236	135	0	40	161	68	1	42	270	134	0
47	255	98	0	59	324	92	1	41	262	154	0
39	135	63	0	56	171	56	1	42	264	86	0
44	121	46	0	76	265	240	1	39	325	148	0
64	171	103	0	67	280	306	1	27	388	191	0
58	260	227	0	57	248	93	1	31	260	123	0
49	237	157	0	57	192	115	1	39	284	135	0
55	261	266	0	42	349	408	1	56	326	236	1
52	397	167	0	54	263	103	1	40	248	92	1
49	295	164	0	60	223	102	1	58	285	153	1
47	261	119	1	33	316	274	0	43	361	126	1
40	258	145	1	55	288	130	0	40	248	226	1
42	280	247	1	36	256	149	0	46	280	176	1

Edited SAS Output (PROC GLM) for Problem 5

Y Regressed on X_2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	21.339709	21.339709	0.19	0.6687
Error	40	4592.279339	114.806983		
Corrected Total	41	4613.619048			

R-Square	Coeff Var	Root MSE	Y Mean
0.004625	22.43378	10.71480	47.76190

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X2	1	21.33970871	21.33970871	0.19	0.6687

Y Regressed on X_3

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X3	1	735.2054113	735.2054113	7.58	0.0088

Y Regressed on X_1 , X_2 , X_3 , X_1X_2 , and X_2X_3

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	46.2355575	46.2355575	0.45	0.5078
X2	1	89.1465681	89.1465681	0.86	0.3591
X3	1	684.3651973	684.3651973	6.62	0.0143
X1X3	1	48.6904324	48.6904324	0.47	0.4968
X2X3	1	26.1296111	26.1296111	0.25	0.6181

- c. Test whether the true coefficients of the product terms X_1X_3 and X_2X_3 are simultaneously zero in the model containing X_1 , X_2 , and X_3 plus these product terms. Specify the two models being compared, and state the null hypothesis in terms of regression coefficients. If this test is not rejected, what can you conclude about the relationship of Y to X_1 and X_2 when X_3 equals 1 compared to the same relationship when X_3 equals 0?
- d. Test (at $\alpha = .05$) whether X_3 is associated with Y , after taking into account the combined contribution of X_1 and X_2 . What does your result, together with your answer to part (c), tell you about the relationship of Y with X_1 and X_2 when SPB is present compared to the same relationship when it is absent?
- e. What overall conclusion can you draw about the association of Y with the three independent variables for this data set? Specify the two models being compared, and state the appropriate null hypothesis in terms of regression coefficients.
6. Use the results from Problem 3 in Chapter 8, as well as the computer output given here, to answer the following questions about the data from that problem.
- a. Conduct overall regression F tests for these three models: Y regressed on weight and age; Y regressed on weight alone; and Y regressed on age alone. Based on your answers, how would you rate the importance of the two variables in predicting Y ? Comment on how your answer here compares with your answer to Problem 3(b) in Chapter 8.

Edited SAS Output (PROC GLM) for Problem 6

Total Cholesterol (Y) Regressed on Weight (X1)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	10231.7262	10231.7262	1.74	0.2000
Error	23	135145.3138	5875.8832		
Corrected Total	24	145377.0400			

Total Cholesterol (Y) Regressed on Age (X2)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	101932.6657	101932.6657	53.96	<.0001
Error	23	43444.3743	1888.8858		
Corrected Total	24	145377.0400			

Total Cholesterol (Y) Regressed on Weight (X1) and Age (X2)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	102570.8147	51285.4073	26.36	<.0001
Error	22	42806.2253	1945.7375		
Corrected Total	24	145377.0400			

- b. Provide variables-added-in-order tests for both variables, with weight (X_1) added first. Use $\alpha = .05$.
- c. Provide variables-added-in-order tests for both variables, with age (X_2) added first. Use $\alpha = .05$.
- d. Provide a table of variables-added-last tests for both weight and age.
- e. What, if any, differences are discernible in the results of the approaches in parts (a) through (d)?
7. Use the results from Problem 4 in Chapter 8, as well as the computer output given here, to answer the following questions about the data from that problem.
- a. Conduct the overall regression F test for the model where Y is regressed on X_1 , X_2 , and X_3 . Use $\alpha = .05$. Interpret your result.
- b. Provide variables-added-in-order tests for the order X_2 , X_1 , and X_3 .

Edited SAS Output (PROC GLM) for Problem 7

Y Regressed on X2 and X1

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X2	1	1308.339479	1308.339479	41.39	<.0001
X1	1	9.458932	9.458932	0.30	0.5915

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X2	1	1309.445931	1309.445931	41.42	<.0001
X1	1	9.458932	9.458932	0.30	0.5915

Y Regressed on X3 and X1

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X3	1	1387.599718	1387.599718	54.61	<.0001
X1	1	35.632307	35.632307	1.40	0.2526

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X3	1	1414.879544	1414.879544	55.68	<.0001
X1	1	35.632307	35.632307	1.40	0.2526

Y Regressed on X3 and X2

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X3	1	1387.599718	1387.599718	64.22	<.0001
X2	1	100.259687	100.259687	4.64	0.0459

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X3	1	179.5199260	179.5199260	8.31	0.0103
X2	1	100.2596869	100.2596869	4.64	0.0459

(continued)

Y Regressed on X1, X2, and X3

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	8.352481	8.352481	0.40	0.5378
X2	1	1309.445931	1309.445931	62.16	<.0001
X3	1	200.346530	200.346530	9.51	0.0071

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	30.2855361	30.2855361	1.44	0.2480
X2	1	94.9129162	94.9129162	4.51	0.0497
X3	1	200.3465296	200.3465296	9.51	0.0071

- c. Provide variables-added-in-order tests for the order X_3 , X_1 , and X_2 .
- d. List all orders that can be tested, using the computer results below and in Problem 4 of Chapter 8. List all orders that *cannot* be so computed.
- e. Provide variables-added-last tests for X_1 , X_2 , and X_3 .
- f. Provide the variables-added-last test for $X_4 = X_2 X_3$ given that X_2 and X_3 are already in the model. Does X_4 significantly improve the prediction of Y given that X_2 and X_3 are already in the model?
8. The following ANOVA table is based on the data discussed in Problem 5 of Chapter 8. Use $\alpha = .05$.

Source	d.f.	SS
Regression	(X_1)	18,953.04
	($X_3 X_1$)	7,010.03
	($X_2 X_1, X_3$)	10.93
Residual	21	2,248.23
Total	24	28,222.23

- a. Provide a test to compare the following two models:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$$

and

$$Y = \beta_0 + \beta_1 X_1 + E$$

- b. Provide a test to compare the following two models:

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + E$$

and

$$Y = \beta_0 + E$$

- c. State which two models are being compared in the computation

$$F = \frac{(18,953.04 + 7,010.03 + 10.93)/3}{2,248.23/21}$$

9. Residential real estate prices are thought to depend, in part, on property size and number of bedrooms. The house size X_1 (in hundreds of square feet), number of bedrooms X_2 , and house price Y (in thousands of dollars) of a random sample of houses in a certain county were observed. The resulting data and some associated computer output were presented in Problem 10 of Chapter 8. Additional portions of the output are shown here. Use all of the output to answer the following questions.
- Perform the overall F test for the regression of Y on both independent variables. Interpret your result.
 - Perform variables-added-in-order tests for both independent variables, with X_1 added first.
 - Perform variables-added-in-order tests for both independent variables, with X_2 added first.
 - Provide a table of variables-added-last tests.
 - Which predictors appear to be necessary? Why?

Edited SAS Output (PROC GLM) for Problem 9

Y Regressed on X1 and X2

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	5732.222826	5732.222826	40.06	0.0032
X2	1	1.098464	1.098464	0.01	0.9344

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	1402.315337	1402.315337	9.80	0.0352
X2	1	1.098464	1.098464	0.01	0.9344

10. Data on sales revenue Y , television advertising expenditures X_1 , and print media advertising expenditures X_2 for a large retailer for the period 1988–1993 were given in Problem 11 of Chapter 8. Use the computer output for that problem, along with the additional portions of the output shown here, to answer the following questions.
- Perform the overall F test for the regression of Y on both independent variables. Interpret your result.
 - Perform variables-added-in-order tests for both independent variables, with X_1 added first.
 - Perform variables-added-in-order tests for both independent variables, with X_2 added first.
 - Provide a table of variables-added-last tests.
 - Which predictors appear to be necessary? Why?

Edited SAS Output (PROC GLM) for Problem 10

Y Regressed on X1 and X2

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	28.07164068	28.07164068	117.82	0.0017
X2	1	0.04689121	0.04689121	0.20	0.6874
Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	25.89125916	25.89125916	108.66	0.0019
X2	1	0.04689121	0.04689121	0.20	0.6874

11. Use the computer output for the radial keratotomy data of Problem 12 in Chapter 8, along with the additional output here, to answer the following questions.
- Perform the overall F test for the regression of Y on both independent variables. Interpret your result.
 - Perform variables-added-in-order tests for both independent variables, with X_1 added first.
 - Perform variables-added-in-order tests for both independent variables, with X_2 added first.
 - Provide a table of variables-added-last tests.
 - Which predictors appear to be necessary? Why?

Edited SAS Output (PROC GLM) for Problem 11

Y Regressed on X1 and X2

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	13.09841552	13.09841552	10.67	0.0020
X2	1	4.52435640	4.52435640	3.68	0.0605
Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	12.43080596	12.43080596	10.12	0.0025
X2	1	4.52435640	4.52435640	3.68	0.0605

12. Use the computer output for the *Business Week* magazine data of Problem 13 in Chapter 8, as well as the additional output here, to answer the following questions.
- Perform the overall F test for the regression of Y on both independent variables, X_2 and X_3 . Interpret your result.
 - Perform variables-added-in-order tests for both independent variables, with X_2 added first.

- c. Perform variables-added-in-order tests for both independent variables, with X_3 added first.
- d. Provide a table of variables-added-last tests.
- e. Which predictors appear to be necessary? Why?

Edited SAS Output (PROC GLM) for Problem 12

Y Regressed on X2 and X3

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X2	1	1.76498949	1.76498949	1.40	0.2537
X3	1	24.81773186	24.81773186	19.63	0.0004

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X2	1	7.97795163	7.97795163	6.31	0.0224
X3	1	24.81773186	24.81773186	19.63	0.0004

13. This problem refers to the 1990 Census data presented in Problem 19 of Chapter 5 and in Problem 14 of Chapter 8.

Use the computer output from Problem 14 of Chapter 8, along with the additional output shown here, to answer the following questions about the regression of OWNEROCC on OWNCOST and URBAN.

- a. Perform the overall F test for the regression of OWNEROCC on OWNCOST and URBAN. Interpret your result.
- b. Perform variables-added-in-order tests for both independent variables, with OWNCOST added first.
- c. Perform variables-added-in-order tests for both independent variables, with URBAN added first.
- d. Provide a table of variables-added-last tests.
- e. Which predictors appear to be necessary? Why?

Edited SAS Output (PROC GLM) for Problem 13

Ownerocc Regressed on OWNCOST and URBAN

Source	DF	Type I SS	Mean Square	F Value	Pr > F
OWNCOST	1	132.6203242	132.6203242	8.84	0.0068
URBAN	1	123.1581842	123.1581842	8.21	0.0088

Source	DF	Type III SS	Mean Square	F Value	Pr > F
OWNCOST	1	66.3318457	66.3318457	4.42	0.0467
URBAN	1	123.1581842	123.1581842	8.21	0.0088

14. This problem refers to the pond ecology data of Chapter 5, Problem 20.
- Perform the overall F test for the multiple regression of copepod count on zooplankton and phytoplankton counts.
 - Perform variables-added-in-order tests for both independent variables, with zooplankton count added first.
 - Perform variables-added-last tests for the regression in part (a).
 - Which variables appear to be necessary? Why?
15. Consider once again the nutritional deficiency example and Model 4 from Chapter 8. In Section 9.6.5, the parameter $L = (\beta_1 + \beta_2)$ associated with one-unit increases in height and age was considered, and a 95% confidence interval for this linear function was computed. For this problem, we now consider the effect of a 3-inch height increase and a 2-year age increase.
- What is the new value of \hat{L} for a 3-inch increase in height and a 2-year increase in age?
 - What are the corresponding values of $S_{\hat{L}}^2$ and $S_{\hat{L}}$?
 - What is the 95% confidence interval for L ?
16. The hypothetical study of two drugs A and B designed to lower systolic blood pressure has now been conducted, using a design where individuals are randomly assigned to take specific doses of both drugs. The researchers would like to know if 5 mg increases in each drug can decrease systolic blood pressure by *more than* 10 mmHg. Given $\hat{\beta}_A = -2.12$, $\hat{\beta}_B = -1.48$, $S_{\hat{\beta}_A}^2 = 0.11$, $S_{\hat{\beta}_B}^2 = 0.03$, $\widehat{\text{cov}}(\hat{\beta}_A, \hat{\beta}_B) = 0.02$, and d.f. = 197, perform and interpret an $\alpha = .01$ one-sided hypothesis test that would be important to these researchers.
17. The numerical output in Section 9.7 was useful for answering questions (a)–(c) in the BRFSS example, but not every significance test about the independent variables can be performed using this single model's results. Which tests of hypotheses about the three predictors cannot be conducted using the given output?

References

- Abt, K. 1981. "Problems of Repeated Significance Testing." *Controlled Clinical Trials* 1: 377–81.
- Kupper, L. L.; Stewart, J. R.; and Williams, K. A. 1976. "A Note on Controlling Significance Levels in Stepwise Regression." *American Journal of Epidemiology* 103(1): 13–15.
- Searle, S. R. 1971. *Linear Models*. New York: John Wiley & Sons.

10

Correlations: Multiple, Partial, and Multiple Partial

10.1 Preview

We saw in Chapter 5 that the following essential features of straight-line regression (excluding the quantitative prediction formula provided by the fitted regression equation) can also be described in terms of the correlation coefficient r . These features are summarized as follows:

1. r (or r_{XY}) is an estimate of the population parameter ρ (or ρ_{XY}), which describes the correlation between X and Y , both considered as random variables.
2. r can be used as a general index of linear association between two random variables, in the following sense:
 - a. The more highly positive r is, the more positive the linear association is; that is, an individual with a high value of one variable will likely have a high value of the other, and an individual with a low value of one variable will probably have a low value of the other.
 - b. The more highly negative r is, the more negative the linear association is; that is, an individual with a high value of one variable will likely have a low value of the other, and conversely.
 - c. If r is close to 0, there is little evidence of linear association, which indicates that there is a nonlinear association or no association at all.
3. $r = \hat{\beta}_1(S_X/S_Y)$, where $\hat{\beta}_1$ is the estimated slope of the regression line.
4. The squared correlation coefficient r^2 measures the strength of the linear relationship between the dependent variable Y and the independent variable X . The closer r^2 is to 1, the stronger the linear relationship; the closer r^2 is to 0, the weaker the linear relationship.

5. $r^2 = (\text{SSY} - \text{SSE})/\text{SSY}$ is the proportionate reduction in the total sum of squares achieved by using a straight-line model in X to predict Y .
6. Assuming that X and Y have a bivariate normal distribution with parameters μ_X , μ_Y , σ_X^2 , σ_Y^2 , and ρ_{XY} , the conditional distribution of Y given X is $N(\mu_{Y|X}, \sigma_{Y|X}^2)$, where

$$\mu_{Y|X} = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(X - \mu_X) \quad \text{and} \quad \sigma_{Y|X}^2 = \sigma_Y^2(1 - \rho^2)$$

Here r^2 estimates ρ^2 , which can be expressed as

$$\rho^2 = \frac{\sigma_Y^2 - \sigma_{Y|X}^2}{\sigma_Y^2}$$

This connection between regression and correlation can be extended to the multiple regression case, as we will discuss in this chapter. When several independent variables are involved, however, the essential features of regression are described not by a single correlation coefficient, as in the straight-line case, but by several correlations. These include a set of zero-order correlations such as r , plus a whole group of higher-order indices called *multiple correlations*, *partial correlations*, and *multiple partial correlations*.¹ These higher-order correlations allow us to answer many of the same questions that can be answered by fitting a multiple regression model. In addition, the correlation analog has been found to be particularly useful in uncovering spurious relationships among variables, identifying intervening variables, and making certain types of causal inferences.²

10.2 Correlation Matrix

When dealing with more than one independent variable, we can represent the collection of all zero-order correlation coefficients (i.e., the r 's between all possible pairs of variables) most compactly in *correlation matrix form*. For example, given $k = 3$ independent variables X_1 , X_2 , and X_3 and one dependent variable Y , there are $C_2^4 = 6$ zero-order correlations, and the correlation matrix has the general form

$$\begin{array}{ccccc} & Y & X_1 & X_2 & X_3 \\ Y & \left[\begin{array}{cccc} 1 & r_{YX_1} & r_{YX_2} & r_{YX_3} \\ & 1 & r_{X_1X_2} & r_{X_1X_3} \\ & & 1 & r_{X_2X_3} \\ & & & 1 \end{array} \right] \end{array}$$

¹ The *order* of a correlation coefficient, as the term is used here, is the number of variables being controlled or adjusted for (see Section 10.5).

² See Blalock (1971) and Bollen (1989) for discussions of techniques for causal inference using regression modeling.

Here r_{YX_j} ($j = 1, 2, 3$) is the correlation between Y and X_j , and $r_{X_i X_j}$ ($i, j = 1, 2, 3$) is the correlation between X_i and X_j .

For the data in Table 8.1, this matrix takes the form

	WGT	HGT	AGE	$(AGE)^2$
WGT	1	0.814	0.770	0.767
HGT		1	0.614	0.615
AGE			1	0.994
$(AGE)^2$				1

Taken separately, each of these correlations describes the strength of the linear relationship between the two variables involved. In particular, the correlations $r_{YX_1} = 0.814$, $r_{YX_2} = 0.770$, and $r_{YX_3} = 0.767$ measure the strength of the linear association with the dependent variable WGT for each of the independent variables taken separately. As we can see, HGT ($r_{Y1} = 0.814$) is the independent variable with the strongest linear relationship to WGT, followed by AGE and then $(AGE)^2$.

Nevertheless, these zero-order correlations do not describe (1) the overall relationship of the dependent variable WGT to the independent variables HGT, AGE, and $(AGE)^2$ considered together; (2) the relationship between WGT and AGE after controlling for³ HGT; or (3) the relationship between WGT and the combined effects of AGE and $(AGE)^2$ after controlling for HGT. The measure that describes relationship (1) is called the *multiple correlation coefficient* of WGT on HGT, AGE, and $(AGE)^2$. The measure that describes relationship (2) is called the *partial correlation coefficient* between WGT and AGE controlling for HGT. Finally, the measure that describes relationship (3) is called the *multiple partial correlation coefficient* between WGT and the combined effects of AGE and $(AGE)^2$ controlling for HGT.

Even though AGE and $(AGE)^2$ are very highly correlated in our example, it is possible—if the general relationship of AGE to WGT is nonlinear—that $(AGE)^2$ will be significantly correlated with WGT even after AGE has been controlled for. In fact, this is what happens in general when the addition of a second-order term in polynomial regression significantly improves the prediction of the dependent variable.

10.3 Multiple Correlation Coefficient

The *multiple correlation coefficient*, denoted as $R_{Y|X_1, X_2, \dots, X_k}$, is a measure of the overall *linear association* of one (dependent) variable Y with several other (independent) variables X_1, X_2, \dots, X_k . By “linear association,” we mean that $R_{Y|X_1, X_2, \dots, X_k}$ measures the strength of the association between Y and the best-fitting linear combination of the X ’s, which is the least-squares solution $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k$. In fact, no other linear combination of the X ’s will have as great a correlation with Y . Also, $R_{Y|X_1, X_2, \dots, X_k}$ is always nonnegative.

³ In this case, the phrase “controlling for” pertains to determining the extent to which the variables WGT and AGE are related after removing the effect of HGT on WGT and AGE.

The multiple correlation coefficient is thus a direct generalization of the simple correlation coefficient r to the case of several independent variables. We have dealt with this measure up to now under the name R^2 , which is the square of the multiple correlation coefficient.

Two computational formulas provide useful interpretations of the multiple correlation coefficient $R_{Y|X_1, X_2, \dots, X_k}$ and its square:

$$R_{Y|X_1, X_2, \dots, X_k} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2 \sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})^2}} \quad (10.1)$$

and

$$R_{Y|X_1, X_2, \dots, X_k}^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\text{SSY} - \text{SSE}}{\text{SSY}}$$

where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \dots + \hat{\beta}_k X_{ik}$ (the predicted value for the i th individual) and $\bar{\hat{Y}} = \sum_{i=1}^n \hat{Y}_i / n$. The second formula in (10.1), which we have seen several times before (as R^2), is most useful for assessing the fit of the regression model. The first formula in (10.1) indicates that $R_{Y|X_1, X_2, \dots, X_k} = r_{Y, \hat{Y}}$, the simple linear correlation between the observed values Y and the predicted values \hat{Y} , regardless of how many predictors are contained in the regression equation used to predict Y .

As a numerical example, let us again consider the data of Table 8.1, where $X_1 = \text{HGT}$, $X_2 = \text{AGE}$, and $Y = \text{WGT}$. Using only X_1 and X_2 in the model, the fitted regression equation is $\hat{Y} = 6.553 + 0.722X_1 + 2.050X_2$, and the observed and predicted values are as given in Table 10.1.

One can check that $\bar{Y} = \bar{\hat{Y}} = 62.75$ (that is, the mean of the observed Y values equals the mean of the predicted values of Y). As we mentioned in Chapter 8, this is no coincidence; it is a mathematical fact that $\bar{Y} = \bar{\hat{Y}}$.

The following SAS output shows that the computed value of R^2 is 0.7800 for the model containing X_1 and X_2 . This tells us that 78% of the variation in Y is explained by the regression model. The corresponding multiple correlation coefficient $R_{Y|X_1, X_2} = R$ is 0.8832, since R is defined as the *positive* square root of R^2 .

TABLE 10.1 Observed and predicted values for the regression of WGT on HGT and AGE

Child	1	2	3	4	5	6	7	8	9	10	11	12
Observed	64	71	53	67	55	58	77	57	56	51	76	68
Predicted	64.11	69.65	54.23	73.87	59.78	57.01	66.77	59.66	57.38	49.18	75.20	66.16

Edited SAS Output (PROC REG) for Regression of WGT on HGT and AGE

Dependent Variable: WGT

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	692.82261	346.41130	15.953	0.0011
Error	9	195.42739	21.71415		
Corrected Total	11	888.25000			

Root MSE	4.65984	R-Square	0.7800	
Dependent Mean	62.75000	Adj R-Sq	0.7311	← Corrects for # of predictors in model
Coeff Var	7.42605			

© Cengage Learning

10.4 Relationship of $R_{Y|X_1, X_2, \dots, X_k}$ to the Multivariate Normal Distribution⁴

An informative way of looking at the sample multiple correlation coefficient $R_{Y|X_1, X_2, \dots, X_k}$ is to consider it as an estimator of a population parameter characterizing the joint distribution of all the variables Y, X_1, X_2, \dots, X_k taken together. When we had two variables X and Y and assumed that their joint distribution was bivariate normal $N_2(\mu_Y, \mu_X, \sigma_Y^2, \sigma_X^2, \rho_{XY})$, we saw that r_{XY} estimated ρ_{XY} , which satisfied the formula $\rho_{XY}^2 = (\sigma_Y^2 - \sigma_{Y|X}^2)/\sigma_Y^2$, where $\sigma_{Y|X}^2$ was the variance of the conditional distribution of Y given X . Now, when we have k independent variables and one dependent variable, we get an analogous result if we assume that their joint distribution is *multivariate normal*. Let us now consider what happens with just two independent variables. In this case, the *trivariate normal distribution* of Y, X_1 , and X_2 can be described as

$$N_3(\mu_Y, \mu_{X_1}, \mu_{X_2}, \sigma_Y^2, \sigma_{X_1}^2, \sigma_{X_2}^2, \rho_{Y1}, \rho_{Y2}, \rho_{12})$$

where μ_Y, μ_{X_1} , and μ_{X_2} are the three (unconditional) means; $\sigma_Y^2, \sigma_{X_1}^2$, and $\sigma_{X_2}^2$ are the three (unconditional) variances; and ρ_{Y1}, ρ_{Y2} , and ρ_{12} are the three correlation coefficients. The *conditional distribution of Y given X_1 and X_2* is then a univariate normal distribution with a (conditional) mean denoted by $\mu_{Y|X_1, X_2}$ and a (conditional) variance denoted by $\sigma_{Y|X_1, X_2}^2$; we usually write this compactly as

$$Y|X_1, X_2 \sim N(\mu_{Y|X_1, X_2}, \sigma_{Y|X_1, X_2}^2)$$

⁴ This section is not essential for the application-oriented reader.

In fact, it turns out that

$$\mu_{Y|X_1, X_2} = \mu_Y + \rho_{Y1|2} \frac{\sigma_{Y|X_2}}{\sigma_{X_1|X_2}} (X_1 - \mu_{X_1}) + \rho_{Y2|1} \frac{\sigma_{Y|X_1}}{\sigma_{X_2|X_1}} (X_2 - \mu_{X_2})$$

and

$$\sigma_{Y|X_1, X_2}^2 = (1 - \rho_{Y, \mu_{Y|X_1, X_2}}^2) \sigma_Y^2$$

where $\rho_{Y, \mu_{Y|X_1, X_2}}$ is the population correlation coefficient between the random variables Y and $\mu_{Y|X_1, X_2} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ (where we are considering X_1 and X_2 as random variables) and where $\rho_{Y1|2}$ and $\rho_{Y2|1}$ are partial correlations (to be discussed in Section 10.5). Also, $\sigma_{Y|X_2}^2$, $\sigma_{X_1|X_2}^2$, and $\sigma_{X_2|X_1}^2$ are the conditional variances, respectively, of Y given X_2 , Y given X_1 , X_1 given X_2 , and X_2 given X_1 .

The parameter $\rho_{Y, \mu_{Y|X_1, X_2}}$ is the *population analog of the sample multiple correlation coefficient* $R_{Y|X_1, X_2}$, and we write $\rho_{Y, \mu_{Y|X_1, X_2}}$ simply as $\rho_{Y|X_1, X_2}$. Furthermore, from the formula for $\sigma_{Y|X_1, X_2}^2$, we can confirm (with a little algebra) that

$$\rho_{Y|X_1, X_2}^2 = \frac{\sigma_Y^2 - \sigma_{Y|X_1, X_2}^2}{\sigma_Y^2}$$

which is the *proportionate reduction in the unconditional variance of Y due to conditioning on X_1 and X_2* .

Generalizing these findings to the case of k independent variables, we may summarize the characteristics of the multiple correlation coefficient $R_{Y|X_1, X_2, \dots, X_k}$ as follows:

1. $R_{Y|X_1, X_2, \dots, X_k}^2$ measures the proportionate reduction in the total sum of squares $\sum_{i=1}^n (Y_i - \bar{Y})^2$ to $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ due to the multiple linear regression of Y on X_1, X_2, \dots, X_k .
2. $R_{Y|X_1, X_2, \dots, X_k}$ is the correlation $r_{Y, \hat{Y}}$ of the observed values (Y) with the predicted values (\hat{Y}), and this correlation is always nonnegative.
3. $R_{Y|X_1, X_2, \dots, X_k}$ is an estimate of $\rho_{Y|X_1, X_2, \dots, X_k}$, the correlation of Y with the true regression equation $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$, where the X 's are considered to be random.
4. $R_{Y|X_1, X_2, \dots, X_k}^2$ is an estimate of the proportionate reduction in the unconditional variance of Y due to conditioning on X_1, X_2, \dots, X_k ; that is, it estimates

$$\rho_{Y|X_1, X_2, \dots, X_k}^2 = \frac{\sigma_Y^2 - \sigma_{Y|X_1, X_2, \dots, X_k}^2}{\sigma_Y^2}$$

10.5 Partial Correlation Coefficient

The *partial correlation coefficient* is a measure of the strength of the linear relationship between two variables after we control for the effects of other variables. If the two variables of interest are Y and X and the control variables are C_1, C_2, \dots, C_q , then we denote the corresponding

partial correlation coefficient by $r_{YX|C_1, C_2, \dots, C_q}$. The order of the partial correlation depends on the number of variables that are being controlled for. Thus, *first-order* partials have the form $r_{YX|C_1}$, *second-order* partials have the form $r_{YX|C_1, C_2}$, and in general, q th-order partials have the form $r_{YX|C_1, C_2, \dots, C_q}$.

For the three independent variables HGT, AGE, and $(AGE)^2$ in our example, the highest-order partial possible is second order. The values of most of the partial correlations that can be computed from this data set are given in Table 10.2.

The easiest way to obtain a partial correlation is to use a standard computer program. Formulas that help highlight the structure of the partial correlation coefficient will be given in Sections 10.5.2 and 10.5.3. First, however, let us see how we can use the information in Table 10.2 to describe our data.

Looking back at our (zero-order) correlation matrix, we see that the variable most highly correlated with WGT is HGT ($r_{Y1} = 0.814$). Thus, of the three independent variables we are considering, HGT is the most important according to the strength of its linear relationship with WGT.

After HGT, what is the next most important variable for the linear prediction of WGT? Since the first-order partial $r_{WGT, AGE|HGT} = 0.589$ is larger than $r_{WGT, (AGE)^2|HGT} = 0.580$, it makes sense to conclude that AGE is next in importance, after we have accounted for HGT. (If we wanted to test the significance of this partial correlation coefficient, we would use a partial F test, as described in Chapter 9. We shall return to this point shortly.)

The only variable left to consider is $(AGE)^2$. But once we have accounted for HGT and AGE, does $(AGE)^2$ add anything to our prediction of WGT? To answer this, we look at the second-order partial correlation $r_{WGT, (AGE)^2|HGT, AGE} = -0.035$. Notice that the magnitude

TABLE 10.2 Partial correlations for the WGT, HGT, and AGE data of Table 8.1

Order	Controlling Variables	Form of Correlation	Computed Value
1	HGT	$r_{WGT, AGE HGT}$	0.589
1	HGT	$r_{WGT, (AGE)^2 HGT}$	0.580
1	HGT	$r_{AGE, (AGE)^2 HGT}$	0.991
1	AGE	$r_{WGT, HGT AGE}$	0.678
1	AGE	$r_{WGT, (AGE)^2 AGE}$	0.015
1	AGE	$r_{HGT, (AGE)^2 AGE}$	0.060
1	$(AGE)^2$	$r_{WGT, HGT (AGE)^2}$	0.677
1	$(AGE)^2$	$r_{WGT, AGE (AGE)^2}$	0.111
1	$(AGE)^2$	$r_{HGT, AGE (AGE)^2}$	0.022
2	HGT, AGE	$r_{WGT, (AGE)^2 HGT, AGE}$	-0.035
2	HGT, $(AGE)^2$	$r_{WGT, AGE HGT, (AGE)^2}$	0.131
2	AGE, $(AGE)^2$	$r_{WGT, HGT AGE, (AGE)^2}$	0.679

of this partial correlation is very small. Thus, we would be inclined to conclude that $(AGE)^2$ provides essentially no additional information about WGT once HGT and AGE have been used together as predictors. (This can be verified with a formal partial F test.)

The procedure for selecting variables just described—starting with the most important variable and continuing step by step to possibly add variables in descending order of importance while controlling for variables already selected—is called a *forward selection* procedure. Alternatively, we could have handled the variable selection problem by working backward—starting with all the variables and deleting (step by step) variables that do not contribute much to the description of the dependent variable. We shall discuss procedures for selecting variables further in Chapter 16.

Output from SAS's REG procedure can be used to determine the partial correlations mentioned in the preceding discussion. The following output is for the model with WGT as the dependent variable and HGT, AGE, and $(AGE)^2$ as the independent variables. The output provides both the variables-added-in-order (i.e., sequential) squared partial correlations (labeled “Squared Partial Corr Type I”) and the variables-added-last squared partial correlations (labeled “Squared Partial Corr Type II”). The output reports that $r_{WGT, AGE|HGT}^2 = 0.3471$. To determine $r_{WGT, AGE|HGT}$, the square root of 0.3471 is first taken. The sign of the partial correlation is then determined from the sign of the slope estimate for AGE in the regression of WGT on AGE and HGT (see output for model 4 on page 150 in Chapter 8). Since the sign of the slope estimate for AGE is positive in this regression, the sign of $r_{WGT, AGE|HGT}$ is also positive. Hence, $r_{WGT, AGE|HGT} = +0.589$.

The other partial correlations can be determined similarly. For example, $r_{WGT, (AGE)^2|HGT, AGE}^2 = 0.0012$. Since the sign of the slope estimate for $(AGE)^2$, given that HGT and AGE are in the model, is negative, we have

$$r_{WGT, (AGE)^2|HGT, AGE} = -\sqrt{0.0012} = -0.035$$

In addition to the use of PROC REG, Appendix C also illustrates how to use PROC CORR to compute any particular type of partial correlation coefficient.

Edited SAS Output (PROC REG) for Regression of WGT on HGT, AGE, and $(AGE)^2$

Dependent Variable: WGT

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	693.06046	231.02015	9.47	0.0052
Error	8	195.18954	24.39869		
Corrected Total	11	888.25000			

Root MSE	4.93950	R-Square	0.7803
Dependent Mean	62.75000	Adj R-Sq	0.6978
Coeff Var	7.87172		

(continued)

PARAMETER ESTIMATES							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1	3.43843	33.61082	0.10	0.9210		
HGT	1	0.72369	0.27696	2.61	0.0310	$r_{WGT, HGT}^2 \rightarrow 0.66301$	$0.46046 \leftarrow r_{WGT, HGT AGE, (AGE)^2}^2$
AGE	1	2.77687	7.42728	0.37	0.7182	$r_{WGT, AGE HGT}^2 \rightarrow 0.34711$	$0.01717 \leftarrow r_{WGT, AGE HGT, (AGE)^2}^2$
AGESQ	1	-0.04171	0.42241	-0.10	0.9238	0.00122	0.00122

$r_{WGT, (AGE)^2|HGT, AGE}^2$

© Cengage Learning

10.5.1 Tests of Significance for Partial Correlations

Regardless of the procedure we use to select variables, we must decide at each step whether a particular partial correlation coefficient is significantly different from zero. We already described how to test for such significance in a slightly different context, in connection with the various uses of the ANOVA table in regression analysis. When we wanted to test whether adding a variable to the regression model was worthwhile, given that certain other variables were already in the model, we used a partial F test. It can be shown that this partial F test is exactly equivalent to a test of significance for the corresponding partial correlation coefficient. Thus, to test whether $r_{YX|C_1, C_2, \dots, C_q}$ is significantly different from zero, we compute the corresponding partial $F(X|C_1, C_2, \dots, C_q)$ and reject the null hypothesis if this F statistic exceeds an appropriate critical value of the $F_{1, n-q-2}$ distribution. For example, in testing whether $r_{WGT, (AGE)^2|HGT, AGE}$ is significant, we find that the partial $F[(AGE)^2|HGT, AGE] = 0.010$ does not exceed $F_{1, 12-2-2, 0.90} = F_{1, 8, 0.90} = 3.46$. Therefore, we conclude that this partial correlation is not significantly different from zero, so $(AGE)^2$ does not contribute to the prediction of WGT once we have accounted for HGT and AGE.

The null hypothesis for this test can be stated more formally by considering the population analog of the sample partial correlation coefficient $r_{YX|C_1, C_2, \dots, C_q}$. This corresponding population parameter, usually denoted by $\rho_{YX|C_1, C_2, \dots, C_q}$, is called the *population partial correlation coefficient*. The null hypothesis can then be stated as $H_0: \rho_{YX|C_1, C_2, \dots, C_q} = 0$, and the associated alternative hypothesis as $H_A: \rho_{YX|C_1, C_2, \dots, C_q} \neq 0$.

10.5.2 Relating the Test for Partial Correlation to the Partial F Test

The structure of the population partial correlation helps us relate this form of higher-order correlation to regression. For simplicity, let us consider this relationship for the special case of two independent variables. The formula for the square of $\rho_{YX_1|X_2}$ can be written as

$$\rho_{YX_1|X_2}^2 = \frac{\sigma_{Y|X_2}^2 - \sigma_{Y|X_1, X_2}^2}{\sigma_{Y|X_2}^2}$$

Thus, the square of the sample partial correlation $r_{YX_1|X_2}$ is an estimate of the proportionate reduction in the conditional variance of Y given X_2 due to conditioning on both X_1 and X_2 .⁵

It then follows that an analogous formula for the squared sample partial correlation coefficient is

$$\begin{aligned} r_{YX_1|X_2}^2 &= \frac{\left[\begin{array}{l} \text{Residual SS (using only } X_2 \text{ in the model)} \\ - \text{Residual SS (using } X_1 \text{ and } X_2 \text{ in the model)} \end{array} \right]}{\text{Residual SS (using only } X_2 \text{ in the model)}} \\ &= \frac{\left[\begin{array}{l} \text{Extra sum of squares due to adding } X_1 \text{ to the model} \\ \text{given that } X_2 \text{ is already in the model} \end{array} \right]}{\text{Residual SS (using only } X_2 \text{ in the model)}} \end{aligned} \quad (10.2)$$

It should be clear from the structure of (10.2) and from the discussion of partial F statistics in Chapters 8 and 9 why the test of $H_0: \rho_{YX_1|X_2} = 0$ is performed using $F(X_1|X_2)$ as the test statistic.

10.5.3 Another Way of Describing Partial Correlations

Another way to compute a first-order partial correlation is to use the formula

$$r_{YX|Z} = \frac{r_{YX} - r_{YZ}r_{XZ}}{\sqrt{(1 - r_{YZ}^2)(1 - r_{XZ}^2)}} \quad (10.3)$$

For example, to compute $r_{WGT, AGE|HGT}$, we calculate

$$\begin{aligned} \frac{r_{WGT, AGE} - r_{WGT, HGT}r_{AGE, HGT}}{\sqrt{(1 - r_{WGT, HGT}^2)(1 - r_{AGE, HGT}^2)}} &= \frac{0.770 - (0.814)(0.614)}{\sqrt{[1 - (0.814)^2][1 - (0.614)^2]}} \\ &= \frac{0.770 - 0.500}{\sqrt{0.337(0.623)}} = 0.589 \end{aligned}$$

Notice that the first correlation in the numerator is the simple zero-order correlation between Y and X . The *control variable* Z appears in the second expression in the numerator (where it is correlated separately with each of the variables Y and X) and in both terms in the denominator. By using (10.3), we can interpret the partial correlation coefficient as an adjustment of the simple correlation coefficient to take into account the effect of the control variable. For example, if r_{YZ} and r_{XZ} have opposite signs, controlling for Z always

⁵ The partial correlation $\rho_{YX_1|X_2}$ can also be described as a zero-order correlation for a conditional bivariate distribution. If the joint distribution of Y , X_1 , and X_2 is trivariate normal, the conditional joint distribution of Y and X_1 given X_2 is bivariate normal. The zero-order correlation between Y and X_1 for this conditional distribution is what we call $\rho_{YX_1|X_2}$; this is exactly the partial correlation between X_1 and Y , controlling for X_2 .

increases r_{YX} . However, if r_{YZ} and r_{XZ} have the same sign, then $r_{YX|Z}$ could be either larger or smaller than r_{YX} .

To compute higher-order partial correlations, we simply reapply this formula using the appropriate next-lower-order partials. For example, the second-order partial correlation is an adjustment of the first-order partial, which, in turn, is an adjustment of the simple zero-order correlation. In particular, we have the following general formula for a second-order partial correlation:

$$r_{YX|Z,W} = \frac{r_{YX|Z} - r_{YW|Z}r_{XW|Z}}{\sqrt{(1 - r_{YW|Z}^2)(1 - r_{XW|Z}^2)}} = \frac{r_{YX|W} - r_{YZ|W}r_{XZ|W}}{\sqrt{(1 - r_{YZ|W}^2)(1 - r_{XZ|W}^2)}} \quad (10.4)$$

To compute $r_{WGT, (AGE)^2|HGT, AGE}$, for example, we have

$$\begin{aligned} \frac{r_{WGT, (AGE)^2|HGT} - r_{WGT, AGE|HGT}r_{(AGE)^2, AGE|HGT}}{\sqrt{(1 - r_{WGT, AGE|HGT}^2)(1 - r_{(AGE)^2, AGE|HGT}^2)}} &= \frac{0.580 - (0.589)(0.991)}{\sqrt{[1 - (0.589)^2][1 - (0.991)^2]}} \\ &= -0.035 \end{aligned}$$

10.5.4 Partial Correlation as a Correlation of Residuals of Regression

There is still another important interpretation concerning partial correlations. For the variables Y , X , and C , suppose that we fit the two straight-line regression equations $Y = \beta_0 + \beta_1 C + E$ and $X = \hat{\beta}_0 + \hat{\beta}_1 C + E$. Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 C$ be the fitted line of Y on C , and let $\hat{X} = \hat{\beta}_0^* + \hat{\beta}_1^* C$ be the fitted line of X on C . Then, the n pairs of deviations, or residuals, $(Y_i - \hat{Y}_i)$ and $(X_i - \hat{X}_i)$, $i = 1, 2, \dots, n$, represent what remains unexplained after the variable C has explained all the variation it can in the variables Y and X separately.

If we now correlate these n pairs of residuals (i.e., find $r_{Y-\hat{Y}, X-\hat{X}}$), we obtain a measure that is independent of the effects of C . It can be shown that *the partial correlation between Y and X , controlling for C , can be defined as the correlation of the residuals of the straight-line regressions of Y on C and of X on C ; that is, $r_{YX|C} = r_{Y-\hat{Y}, X-\hat{X}}$.*

10.5.5 Semipartial Correlations

An alternative form of partial correlation is sometimes considered. The partial correlation $r_{YX|C}$ was just characterized as a correlation between Y adjusted for C and X adjusted for C . Some statisticians refer to this as a *full partial*, since both variables being correlated have been adjusted for C .

The *semipartial correlation* (or “part” correlation) may be characterized as the correlation between two variables when only one of the two has been adjusted for a third variable. For example, one may consider the semipartial correlation between Y and X with only X adjusted for C or with only Y adjusted for C . The first will be denoted by $r_{Y(X|C)}$, and the second by $r_{X(Y|C)}$. Thus, we have $r_{Y(X|C)} = r_{Y, X-\hat{X}}$ and $r_{X(Y|C)} = r_{Y-\hat{Y}, X}$, where \hat{X} and \hat{Y} are obtained from straight-line regressions on C .

Another way of describing these semipartialis is in terms of zero-order correlations, as follows:

$$r_{Y(X|C)} = \frac{r_{YX} - r_{YC}r_{XC}}{\sqrt{1 - r_{XC}^2}} \quad (10.5)$$

and

$$r_{X(Y|C)} = \frac{r_{YX} - r_{YC}r_{XC}}{\sqrt{1 - r_{YC}^2}} \quad (10.6)$$

It is instructive to compare these formulas with formula (10.3) for the full partial. The numerator is the same in all three expressions: the partial covariance between Y and X adjusted for C (with all three variables standardized to have variance 1). Clearly, then, these correlations all have the same sign; and if any one equals 0, they all do. For significance testing, it is appropriate to use the extra-sum-of-squares test described earlier.

These three correlations have different interpretations. They each describe the relationship between Y and X , but with adjustment for different quantities. The semipartial $r_{Y(X|C)}$ is the correlation between Y and X with X adjusted for C ; the semipartial $r_{X(Y|C)}$ is the correlation between Y and X with Y adjusted for C . Finally, the full partial $r_{YX|C}$ is the correlation between Y and X with both Y and X adjusted for C .

Choosing the proper correlation coefficient depends on the relationship among the three variables X , Y , and C (the nuisance variable). Table 10.3 shows the four possible types of relationships. Case 1 involves assessing the relationship between X and Y without a nuisance variable present; here the simple correlation r_{XY} should be used. Case 2 illustrates the situation in which the nuisance variable C is related to both X and Y , so that the use of $r_{YX|C}$ is appropriate. In cases 3 and 4, the nuisance variable affects just one of the two variables X and Y . In these cases, semipartial correlations permit just one of two primary variables to be adjusted for the effects of a nuisance variable.

10.5.6 Summary of the Features of the Partial Correlation Coefficient

1. The partial correlation $r_{YX|C_1, C_2, \dots, C_q}$ measures the strength of the linear relationship between two variables X and Y while controlling for variables C_1, C_2, \dots, C_q .
2. The square of the partial correlation $r_{YX|C_1, C_2, \dots, C_q}^2$ measures the proportion of the residual sum of squares that is accounted for by the addition of X to a regression model already involving C_1, C_2, \dots, C_q ; that is,

$$r_{YX|C_1, C_2, \dots, C_q}^2 = \frac{\left[\begin{array}{l} \text{Extra sum of squares due to adding} \\ X \text{ to a model already containing } C_1, C_2, \dots, C_q \end{array} \right]}{\text{Residual SS (using only } C_1, C_2, \dots, C_q \text{ in the model})}$$

TABLE 10.3 Possible relationships among variables X and Y and nuisance variable C

Case	Nuisance Relationship	Diagram	Preferred Correlation
1	Neither X nor Y affected by C	$X \longleftrightarrow Y$	r_{XY}
2	Both X and Y affected by C	$\begin{array}{ccc} X & \xleftarrow{\quad} & Y \\ & \swarrow & \searrow \\ & C & \end{array}$	$r_{YX C}$
3	Only X affected by C	$\begin{array}{ccc} X & \xleftarrow{\quad} & Y \\ & \swarrow & \downarrow \\ & C & \end{array}$	$r_{Y(X C)}$
4	Only Y affected by C	$\begin{array}{ccc} X & \xleftarrow{\quad} & Y \\ & \uparrow & \searrow \\ & C & \end{array}$	$r_{X(Y C)}$

© Cengage Learning

3. The partial correlation coefficient $r_{YX|C_1, C_2, \dots, C_q}$ is an estimate of the population parameter $\rho_{YX|C_1, C_2, \dots, C_q}$, which is the correlation between Y and X in the conditional joint distribution of Y and X given C_1, C_2, \dots, C_q . The square of this population partial correlation coefficient is given by the equivalent formula

$$\rho_{YX|C_1, C_2, \dots, C_q}^2 = \frac{\sigma_{Y|C_1, C_2, \dots, C_q}^2 - \sigma_{Y|X, C_1, C_2, \dots, C_q}^2}{\sigma_{Y|C_1, C_2, \dots, C_q}^2}$$

where $\sigma_{Y|C_1, C_2, \dots, C_q}^2$ is the variance of the conditional distribution of Y given C_1, C_2, \dots, C_q (and where $\sigma_{Y|X, C_1, C_2, \dots, C_q}^2$ is similarly defined).

4. The partial F statistic $F(X|C_1, C_2, \dots, C_q)$ is used to test $H_0: \rho_{YX|C_1, C_2, \dots, C_q} = 0$.
5. The (first-order) partial correlation coefficient $r_{YX|C}$ is an adjustment of the (zero-order) correlation r_{XY} that takes into account the effect of the control variable C . This can be seen from the formula

$$r_{YX|C} = \frac{r_{XY} - r_{YC}r_{XC}}{\sqrt{(1 - r_{YC}^2)(1 - r_{XC}^2)}}$$

Higher-order partial correlations are computed by reapplying this formula, using the next-lower-order partials.

6. The partial correlation $r_{YX|C}$ can be defined as the correlation of the residuals of the straight-line regressions of Y on C and of X on C ; that is, $r_{YX|C} = r_{Y-\hat{Y}, X-\hat{X}}$.

10.6 Alternative Representation of the Regression Model

With the correlation analog to multiple regression, we can express the regression model $\mu_{Y|X_1, X_2, \dots, X_k} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$ in terms of partial correlation coefficients and conditional variances. When $k = 3$, this representation takes the form

$$\begin{aligned}\mu_{Y|X_1, X_2, X_3} &= \mu_Y + \rho_{YX_1|X_2, X_3} \left(\frac{\sigma_{Y|X_2, X_3}}{\sigma_{X_1|X_2, X_3}} \right) (X_1 - \mu_{X_1}) \\ &\quad + \rho_{YX_2|X_1, X_3} \left(\frac{\sigma_{Y|X_1, X_3}}{\sigma_{X_2|X_1, X_3}} \right) (X_2 - \mu_{X_2}) \\ &\quad + \rho_{YX_3|X_1, X_2} \left(\frac{\sigma_{Y|X_1, X_2}}{\sigma_{X_3|X_1, X_2}} \right) (X_3 - \mu_{X_3})\end{aligned}\tag{10.7}$$

where

$$\beta_1 = \rho_{YX_1|X_2, X_3} \left(\frac{\sigma_{Y|X_2, X_3}}{\sigma_{X_1|X_2, X_3}} \right) \quad \beta_2 = \rho_{YX_2|X_1, X_3} \left(\frac{\sigma_{Y|X_1, X_3}}{\sigma_{X_2|X_1, X_3}} \right) \quad \beta_3 = \rho_{YX_3|X_1, X_2} \left(\frac{\sigma_{Y|X_1, X_2}}{\sigma_{X_3|X_1, X_2}} \right)$$

Notice the similarity between this representation and the one for the straight-line case, where β_1 is equal to $\rho(\sigma_Y/\sigma_X)$. Also, here

$$\beta_0 = \mu_Y - \beta_1 \mu_{X_1} - \beta_2 \mu_{X_2} - \beta_3 \mu_{X_3}$$

An equivalent method to that of least squares for estimating the coefficients $\beta_0, \beta_1, \beta_2$, and β_3 is to substitute appropriate estimates of the population parameters in the preceding formulas:

$$\begin{aligned}\hat{\mu}_Y &= \bar{Y} & \hat{\mu}_{X_1} &= \bar{X}_1 & \hat{\mu}_{X_2} &= \bar{X}_2 & \hat{\mu}_{X_3} &= \bar{X}_3 \\ \hat{\beta}_1 &= r_{YX_1|X_2, X_3} \left(\frac{S_{Y|X_2, X_3}}{S_{X_1|X_2, X_3}} \right) & \hat{\beta}_2 &= r_{YX_2|X_1, X_3} \left(\frac{S_{Y|X_1, X_3}}{S_{X_2|X_1, X_3}} \right) & \hat{\beta}_3 &= r_{YX_3|X_1, X_2} \left(\frac{S_{Y|X_1, X_2}}{S_{X_3|X_1, X_2}} \right)\end{aligned}$$

10.7 Multiple Partial Correlation

10.7.1 The Coefficient and Its Associated F Test

The *multiple partial correlation coefficient* is used to describe the overall relationship between a dependent variable and two or more independent variables while controlling for still other variables. For example, suppose that we consider the variable $X_1^2 = (\text{HGT})^2$ and the product term $X_1 X_2 = \text{HGT} \times \text{AGE}$, in addition to the independent variables $X_1 = \text{HGT}$, $X_2 = \text{AGE}$, and $X_2^2 = (\text{AGE})^2$. Our complete regression model is then of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + E$$

We call such a model a *complete second-order model*, since it includes all possible variables up through second-order terms. For such a complete model, we frequently want to know whether any of the second-order terms are important—in other words, whether a first-order model involving only X_1 and X_2 (i.e., $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$) is adequate. There are two equivalent ways to represent this question as a hypothesis-testing problem: one is to test $H_0: \beta_{11} = \beta_{22} = \beta_{12} = 0$ (i.e., all second-order coefficients are zero); the other is to test the hypothesis $H_0: \rho_{Y(X_1^2, X_2^2, X_1X_2)|X_1, X_2} = 0$, where $\rho_{Y(X_1^2, X_2^2, X_1X_2)|X_1, X_2}$ is the population multiple partial correlation of Y with the second-order variables, controlling for the effects of the first-order variables. (In general, we write the multiple partial as $\rho_{Y(X_1, X_2, \dots, X_k)|Z_1, Z_2, \dots, Z_p}$.) This parameter is estimated by the sample multiple partial correlation $r_{Y(X_1^2, X_2^2, X_1X_2)|X_1, X_2}$, which describes the overall multiple contribution of adding the second-order terms to the model after the effects of the first-order terms are partialled out or controlled for (hence the term *multiple partial*). Two equivalent formulas for $r_{Y(X_1^2, X_2^2, X_1X_2)|X_1, X_2}^2$ are

$$\begin{aligned} r_{Y(X_1^2, X_2^2, X_1X_2)|X_1, X_2}^2 &= \frac{\left[\begin{array}{l} \text{Residual SS (only } X_1 \text{ and } X_2 \text{ in the model)} \\ - \text{Residual SS (all first-and second-order terms in the model)} \end{array} \right]}{\text{Residual SS (only } X_1 \text{ and } X_2 \text{ in the model)}} \\ &= \frac{\left[\begin{array}{l} \text{Extra sum of squares due to the addition of the second-order terms } X_1^2, X_2^2, \text{ and} \\ X_1X_2 \text{ to a model containing only the first-order terms } X_1 \text{ and } X_2 \end{array} \right]}{\text{Residual SS (only } X_1 \text{ and } X_2 \text{ in the model)}} \end{aligned} \quad (10.8)$$

and

$$r_{Y(X_1^2, X_2^2, X_1X_2)|X_1, X_2}^2 = \frac{R_{Y|X_1, X_2, X_1^2, X_2^2, X_1X_2}^2 - R_{Y|X_1, X_2}^2}{1 - R_{Y|X_1, X_2}^2}$$

In most applications, estimating a multiple partial correlation is rarely of interest, so the preceding formulas are infrequently used. Nevertheless, there often is interest in testing hypotheses about a collection (or “chunk”) of higher-order terms, and such tests involve the multiple partial correlation. For instance, to test $H_0: \rho_{Y(X_1^2, X_2^2, X_1X_2)|X_1, X_2} = 0$ (or, equivalently, $H_0: \beta_{11} = \beta_{22} = \beta_{12} = 0$), we calculate the multiple partial F statistic given in general form by expression (9.6) in Chapter 9. For this example, the formula becomes

$$\begin{aligned} F(X_1^2, X_2^2, X_1X_2|X_1, X_2) &= \frac{[\text{Regression SS}(X_1, X_2, X_1^2, X_2^2, X_1X_2) - \text{Regression SS}(X_1, X_2)]/3}{\text{MS Residual}(X_1, X_2, X_1^2, X_2^2, X_1X_2)} \\ &= \frac{[\text{Residual SS}(X_1, X_2) - \text{Residual SS}(X_1, X_2, X_1^2, X_2^2, X_1X_2)]/3}{\text{MS Residual}(X_1, X_2, X_1^2, X_2^2, X_1X_2)} \end{aligned}$$

We would reject H_0 at the α level of significance if the calculated value of $F(X_1^2, X_2^2, X_1X_2 | X_1, X_2)$ exceeded the critical value $F_{3, n-6, 1-\alpha}$.

In general, the null hypothesis $H_0: \rho_{Y(X_1, X_2, \dots, X_s) | C_1, C_2, \dots, C_q} = 0$ is equivalent to the hypothesis $H_0: \beta_1^* = \beta_2^* = \dots = \beta_s^* = 0$ in the model $Y = \beta_0 + \beta_1 C_1 + \beta_2 C_2 + \dots + \beta_q C_q + \beta_1^* X_1 + \beta_2^* X_2 + \dots + \beta_s^* X_s + E$. The appropriate test statistic is the multiple partial F given by

$$F(X_1, X_2, \dots, X_s | C_1, C_2, \dots, C_q)$$

which has the F distribution with s and $n - q - s - 1$ degrees of freedom under H_0 .

The general F statistic given by (9.6) may be expressed in terms of squared multiple correlations (R^2 terms) involving the two models being compared. The general form for this alternative expression is

$$F = \frac{[R^2(\text{larger model}) - R^2(\text{smaller model})] / [\text{Regression d.f. (larger model)} - \text{Regression d.f. (smaller model)}]}{[1 - R^2(\text{larger model})] / \text{Residual d.f. (larger model)}} \quad (10.9)$$

10.8 Concluding Remarks

As we have seen throughout this chapter, a regression F test is associated with many equivalent null hypotheses. As an example, the following null hypotheses all make the same statement but in different forms:

1. H_0 : “Adding variables to the smaller model to form the larger model does not significantly improve the prediction of Y .”
2. H_0 : “The population regression coefficients for the variables in the larger model but not in the smaller model are all equal to 0.”
3. H_0 : “The population multiple partial correlation between Y and variables added to produce the larger model, controlling for the variables in the smaller model, is 0.”
4. H_0 : “The value of the population squared multiple correlation coefficient for the larger model is not greater than the value of that parameter for the smaller model.”

The first two null hypotheses involve prediction, while the latter two involve association. The investigator, for interpretive purposes, can choose either group depending on whether his or her focus is on the predictive ability of the regression model or on a particular association of interest.

Problems

1. The correlation matrix obtained for the variables SBP (Y), AGE (X_1), SMK (X_2), and QUET (X_3), using the data from Problem 2 in Chapter 5, is given by

	SBP	AGE	SMK	QUET
SBP	1	0.7752	0.2473	0.7420
AGE	0.7752	1	-0.1395	0.8028
SMK	0.2473	-0.1395	1	-0.0714
QUET	0.7420	0.8028	-0.0714	1

- a. Based on this matrix, which of the independent variables AGE, SMK, and QUET explains the largest proportion of the total variation in the dependent variable SBP?
- b. Using the available computer outputs, determine the partial correlations $r_{\text{SBP}, \text{SMK}|\text{AGE}}$ and $r_{\text{SBP}, \text{QUET}|\text{AGE}}$.
- c. Test for the significance of $r_{\text{SBP}, \text{SMK}|\text{AGE}}$ using the ANOVA results given in Problem 1 of Chapter 8. Express the appropriate null hypothesis in terms of a population partial correlation coefficient.

Edited SAS Output (PROC REG) for Problem 1

Regression of SBP on AGE and SMK

PARAMETER ESTIMATES							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1	48.04960	11.12956	4.32	0.0002		
AGE	1	1.70916	0.20176	8.47	<.0001	0.60094	0.71220
SMK	1	10.29439	2.76811	3.72	0.0009	0.32291	0.32291

Regression of SBP on AGE and QUET

PARAMETER ESTIMATES							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1	55.32344	12.53475	4.41	0.0001		
AGE	1	1.04516	0.38606	2.71	0.0113	0.60094	0.20175
QUET	1	9.75073	5.40246	1.80	0.0815	0.10099	0.10099

(continued)

Regression of SBP on AGE, SMK, and QUET

PARAMETER ESTIMATES							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1	45.10319	10.76488	4.19	0.0003	.	.
AGE	1	1.21271	0.32382	3.75	0.0008	0.60094	0.33373
SMK	1	9.94557	2.65606	3.74	0.0008	0.32291	0.33367
QUET	1	8.59245	4.49868	1.91	0.0664	0.11527	0.11527

- d. Determine the second-order partial $r_{\text{SBP}, \text{QUET}|\text{AGE}, \text{SMK}}$ and test for the significance of this partial correlation (again, using the computer output here and in Chapter 8, Problem 1).
- e. Based on the results you obtained in parts (a) through (d), how would you rank the independent variables in terms of their importance in predicting Y ? Which of these variables are relatively unimportant?
- f. Compute the squared multiple partial correlation $r_{\text{SBP}(\text{QUET}, \text{SMK})|\text{AGE}}^2$ using the output here and in Problem 1 of Chapter 8. Test for the significance of this correlation. Does this test result alter your decision in part (e) about which variables to include in the regression model?
- 2. An (equivalent) alternative to performing a partial F test for the significance of adding a new variable to a model while controlling for variables already in the model is to perform a t test using the appropriate partial correlation coefficient. If the dependent variable is Y , the independent variable of interest is X , and the controlling variables are Z_1, Z_2, \dots, Z_p , then the t test for $H_0: \rho_{YX|Z_1, Z_2, \dots, Z_p} = 0$ versus $H_A: \rho_{YX|Z_1, Z_2, \dots, Z_p} \neq 0$ is given by the test statistic

$$T = r_{YX|Z_1, Z_2, \dots, Z_p} \frac{\sqrt{n - p - 2}}{\sqrt{1 - r_{YX|Z_1, Z_2, \dots, Z_p}^2}}$$

which has a t distribution under H_0 with $n - p - 2$ degrees of freedom. The critical region for this test is therefore given by

$$|T| \geq t_{n-p-2, 1-\alpha/2}$$

Two variables X and Y are said to have a *spurious correlation* if their correlation solely reflects each variable's relationship to a third (antecedent) variable Z (and possibly to other variables). For example, the correlation between the total annual income (from all sources) of members of the U.S. Congress (Y) and the number of persons owning color television sets (X) is quite high. Simultaneously, however, a general upward trend has occurred in buying power (Z_1) and in wages of all types (Z_2), which would naturally be reflected in increased purchases of color TVs, as well as in increased income of members of Congress. Thus, the high correlation between X

and Y probably only reflects the influence of inflation on each of these two variables. Therefore, this correlation is spurious because it misleadingly suggests a relationship between color TV sales and the income of members of Congress.

- How would you attempt to detect statistically whether a correlation between X and Y like the one described is spurious?
- In a hypothetical study investigating socioecological determinants of respiratory morbidity for a sample of 25 communities, the following correlation matrix was obtained for four variables.

	Unemployment Level (X_1)	Average Temperature (X_2)	Air Pollution Level (X_3)	Respiratory Morbidity Rate (Y)
Unemployment level (X_1)	1	0.51	0.41	0.35
Average temperature (X_2)	—	1	0.29	0.65
Air pollution level (X_3)	—	—	1	0.50
Respiratory morbidity rate (Y)	—	—	—	1

- Determine the partial correlations $r_{YX_1|X_2}$, $r_{YX_1|X_3}$, and $r_{YX_1|X_2, X_3}$.
- Use the results in part (1) to determine whether the correlation of 0.35 between unemployment level (X_1) and respiratory morbidity rate (Y) is spurious. (Use the testing formula given in Problem 2 to make the appropriate tests.)
- Describe a relevant example of spurious correlation in your field of interest. (Use only interval variables, and define them carefully.)
- a. Using the information provided in Problem 2 of Chapter 9, determine the proportion of residual variation that is explained by the addition of X_2 to a model already containing X_1 ; that is, compute
$$Q = \frac{\text{Regression SS}(X_1, X_2) - \text{Regression SS}(X_1)}{\text{Residual SS}(X_1)}$$
b. How is the formula given in part (a) related to the partial correlation $r_{YX_2|X_1}$?
c. Test the hypothesis $H_0: \rho_{YX_2|X_1} = 0$ using both an F test and a two-sided t test. Check to confirm that these tests are equivalent.
- Refer to Problem 7 of Chapter 9 to answer the following questions about the relationship of homicide rate (Y) to city population size (X_1), percentage of families with yearly incomes less than \$5,000 (X_2), and unemployment rate (X_3).
 - Determine the squared partial correlations $r^2_{YX_1|X_3}$ and $r^2_{YX_2|X_3}$ using the computer output here. Check the computation of $r^2_{YX_2|X_3}$ by means of an alternative formula, using the information that $r_{YX_2} = 0.8398$, $r_{YX_3} = 0.8648$, and $r_{X_2X_3} = 0.8154$.
 - Based on the results you obtained in part (a), which variable (if any) should next be considered for entry into the model given that X_3 is already in the model?
 - Test $H_0: \rho_{YX_2|X_3} = 0$ using the t test described in Problem 2.
 - Determine the squared partial correlation $r^2_{YX_1|X_2, X_3}$ from the output here and/or in Problem 7 of Chapter 9, and then test $H_0: \rho_{YX_1|X_2, X_3} = 0$.

- e. Determine the squared multiple partial correlation $r_{Y(X_1, X_2)|X_3}^2$, and test $H_0: \rho_{Y(X_1, X_2)|X_3} = 0$.
- f. Based on the results you obtained in parts (a) through (e), which variables would you include in your final regression model? Use $\alpha = .05$.

Edited SAS Output (PROC REG) for Problem 4

Regression of Y on X3, X2, and X1

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1518.14494	506.04831	24.02	<.0001
Error	16	337.05706	21.06607		
Corrected Total	19	1855.20200			

Root MSE	4.58978	R-Square	0.8183
Dependent Mean	20.57000	Adj R-Sq	0.7843
Coeff Var	22.31297		

PARAMETER ESTIMATES							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1	-36.76493	7.01093	-5.24	<.0001	.	.
X3	1	4.71982	1.53048	3.08	0.0071	0.74795	0.37280
X2	1	1.19217	0.56165	2.12	0.0497	0.21441	0.21972
X1	1	0.00076294	0.00063630	1.20	0.2480	0.08244	0.08244

5. Using the ANOVA table given in Problem 8 of Chapter 9, which deals with the regression relationship of 12th-grade mean verbal SAT scores (Y) to per pupil expenditures (X_1), percentage of teachers with advanced degrees (X_2), and pupil–teacher ratio (X_3), test the following null hypotheses:
- $H_0: \rho_{YX_3|X_1} = 0$
 - $H_0: \rho_{YX_2|X_1, X_3} = 0$
 - $H_0: \rho_{Y(X_2, X_3)|X_1} = 0$
 - Based on these results, and assuming that X_1 is an important predictor of Y , what additional variables would you include in your regression model?
6. Using the following ANOVA table based on data in Problem 6 of Chapter 8 about the regression relationship of respiratory cancer mortality rates (Y) to air pollution index (X_1), mean age (X_2), and percentage of workforce employed in a certain industry (X_3), test the following hypotheses:
- $H_0: \rho_{YX_2|X_1} = 0$
 - $H_0: \rho_{YX_3|X_1, X_2} = 0$

Source	d.f.	SS
X_1	1	1,523.658
$X_2 X_1$	1	181.743
$X_3 X_1, X_2$	1	130.529
Residual	19	551.723
Total	22	2,387.653

- c. H_0 : "The addition of X_2 and X_3 to a model already containing X_1 does not significantly improve the prediction of Y ."
- d. Based on these results, which variables are important predictors of Y ? Use $\alpha = .05$.
- e. State the results in part (a) in terms of equivalent tests about population semipartial correlations.
7. Refer to the following ANOVA tables and to SAS output given here (from data in Problem 8 of Chapter 8) to answer the following questions dealing with factors related to soil erosion.
- a. Using the accompanying SAS output, compute $r_{YX_2|X_1}$ and $r_{YX_3|X_1}$.
- b. Based on your results in part (a), which variable (if any) should next be entered into a regression model that already contains X_1 ?
- c. Test $H_0: \rho_{YX_2|X_1} = 0$ using the t test described in Problem 2.
- d. Determine the squared multiple partial correlation $r^2_{Y(X_2, X_3)|X_1}$, and test $H_0: \rho_{Y(X_2, X_3)|X_1} = 0$.

ANOVA Tables for Problem 7

Source	d.f.	SS	Source	d.f.	SS
X_2	1	667.7279	X_1	1	640.4249
$X_3 X_2$	1	5.8228	$X_2 X_1$	1	32.7819
$X_1 X_2, X_3$	1	6.9406	$X_3 X_1, X_2$	1	7.2844
Residual	7	16.0942	Residual	7	16.0942

Edited SAS Output (PROC REG) for Problem 7

Regression of Y on X1 and X2

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	673.20680	336.60340	115.18	<.0001
Error	8	23.37865	2.92233		
Corrected Total	10	696.58545			

Root MSE	1.70948	R-Square	0.9664
Dependent Mean	40.23636	Adj R-Sq	0.9580
Coeff Var	4.24860		

(continued)

PARAMETER ESTIMATES							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1	-8.08481	20.06315	-0.40	0.6975		
X1	1	68.25068	49.84521	1.37	0.2081	0.91938	0.18986
X2	1	2.29387	0.68488	3.35	0.0101	0.58372	0.58372

Regression of Y on X1 and X3

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	670.13094	335.06547	101.33	<.0001
Error	8	26.45452	3.30681		
Corrected Total	10	696.58545			

Root MSE	1.81846	R-Square	0.9620
Dependent Mean	40.23636	Adj R-Sq	0.9525
Coeff Var	4.51946		

PARAMETER ESTIMATES							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1	-26.70451	16.62113	-1.61	0.1468		
X1	1	157.26713	28.44374	5.53	0.0006	0.91938	0.79259
X3	1	-39.92595	13.32103	-3.00	0.0171	0.52895	0.52895

Regression of Y on X1, X2, and X3

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	680.49122	226.83041	98.66	<.0001
Error	7	16.09423	2.29918		
Corrected Total	10	696.58545			

Root MSE	1.51630	R-Square	0.9769
Dependent Mean	40.23636	Adj R-Sq	0.9670
Coeff Var	3.76849		

PARAMETER ESTIMATES							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1	-1.87932	18.13420	-0.10	0.9204		
X1	1	77.32578	44.50547	1.74	0.1259	0.91938	0.30131
X2	1	1.55910	0.73447	2.12	0.0714	0.58372	0.39163
X3	1	-23.90378	13.42936	-1.78	0.1183	0.31158	0.31158

8. Use the computer results from Problem 9 of Chapter 8 to answer the following questions.
 - a. Test $H_0: \rho_{YX_1} = 0$ and $H_0: \rho_{YX_2} = 0$.
 - b. Test $H_0: \rho_{YX_1|X_2} = 0$ and $H_0: \rho_{YX_2|X_1} = 0$.
 - c. Based on your results in parts (a) and (b), which variables (if any) should be included in the regression model, and what is their order of importance?
9. Use the correlation matrix from Problem 1 to answer the following questions.
 - a. Compute the semipartial $r_{\text{SBP}(\text{SMK}|\text{AGE})}$.
 - b. Compute the semipartial $r_{\text{SMK}(\text{SBP}|\text{AGE})}$.
 - c. Compare these correlations to the (full) partial $r_{\text{SBP}, \text{SMK}|\text{AGE}}$ computed in Problem 1.
10. For the data discussed in Problem 6, provide numerical values for the following quantities:
 - a. $r_{YX_1}^2$
 - b. $R_{Y|X_1, X_2}^2$
 - c. $R_{Y|X_1, X_2, X_3}^2$
 - d. $r_{YX_3|X_1, X_2}^2$
 - e. $r_{YX_2|X_1}^2$
11. Use the results in Problem 7 to answer the following questions.
 - a. Provide a numerical value for $r_{YX_1|X_2, X_3}^2$.
 - b. Which two models should you compare in testing whether the correlation in part (a) is zero in the population?
 - c. Provide a numerical value for $r_{YX_1}^2$.
 - d. What does the difference between parts (a) and (c) say about the relationships among the three predictor variables?
 - e. Which two models should you compare in testing whether the correlations in parts (a) and (c) differ in the population?
12. The accompanying SAS computer output relates to the house price data of Problem 10 in Chapter 8. Use this output and, if necessary, the output associated with Problem 9 in Chapter 9 to answer the following questions.
 - a. Determine $r_{Y|X_1, X_2}^2$, the squared multiple correlation between house price (Y) and the independent variables house size (X_1) and number of bedrooms (X_2).
 - b. Determine $r_{YX_2|X_1}$, the partial correlation of Y with X_2 given that X_1 is in the model.
 - c. Determine $r_{YX_1|X_2}$, the partial correlation of Y with X_1 given that X_2 is in the model.
 - d. Using the t test technique of Problem 2, test $H_0: \rho_{YX_2|X_1} = 0$. Compare your test statistic value with the partial t test statistic for X_2 shown on the SAS output here.
 - e. Using the t test technique of Problem 2, test $H_0: \rho_{YX_1|X_2} = 0$. Compare your test statistic value with the partial t test statistic for X_1 shown on the SAS output here.
 - f. Based on your answers to parts (a) through (e), which variables (if any) should be included in the regression model, and what is their order of importance?

Edited SAS Output (PROC REG) for Problem 12

Regression of Y on X1 and X2

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5733.32129	2866.66064	20.03	0.0082
Error	4	572.39300	143.09825		
Corrected Total	6	6305.71429			

Root MSE	11.96237	R-Square	0.9092
Dependent Mean	105.42857	Adj R-Sq	0.8638
Coeff Var	11.34642		

PARAMETER ESTIMATES							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1	-16.09339	24.64694	-0.65	0.5494	.	.
X1	1	5.72179	1.82779	3.13	0.0352	0.90905	0.71014
X2	1	-1.17315	13.38994	-0.09	0.9344	0.00192	0.00192

13. The accompanying SAS computer output relates to the sales revenue data from Problem 11 in Chapter 8. Use this output and, if necessary, the output for Problem 10 in Chapter 9 to answer the following questions.
- Determine $r^2_{Y|X_1, X_2}$, the squared multiple correlation between sales revenue (Y) and the independent variables TV advertising expenditures (X_1) and print advertising expenditures (X_2).
 - For the sales revenue data, answer the questions posed in Problem 12, parts (b) through (f).

Edited SAS Output (PROC REG) for Problem 13

Regression of Y on X1 and X2

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	28.11853	14.05927	59.01	0.0039
Error	3	0.71480	0.23827		
Corrected Total	5	28.83333			

(continued)

Root MSE	0.48813	R-Square	0.9752
Dependent Mean	5.16667	Adj R-Sq	0.9587
Coeff Var	9.44760		

PARAMETER ESTIMATES							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t 	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1	2.10469	0.42197	4.99	0.0155	.	.
X1	1	1.24188	0.11913	10.42	0.0019	0.97358	0.97313
X2	1	-0.19495	0.43944	-0.44	0.6874	0.06156	0.06156

14. The accompanying SAS computer output relates to the radial keratotomy data from Problem 12 in Chapter 8. Use this output and, if necessary, the output from Problem 11 in Chapter 9 to answer the following questions.
- Determine $r^2_{Y|X_1, X_2}$, the squared multiple correlation between change in refractive error (Y) and the independent variables baseline refractive error (X_1) and baseline curvature (X_2).
 - For the radial keratotomy data, answer the questions posed in Problem 12, parts (b) through (f).

Edited SAS Output (PROC REG) for Problem 14

Regression of Y on X1 and X2

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	17.62277	8.81139	7.18	0.0018
Error	51	62.63017	1.22804		
Corrected Total	53	80.25294			

Root MSE	1.10817	R-Square	0.2196
Dependent Mean	3.83343	Adj R-Sq	0.1890
Coeff Var	28.90811		

PARAMETER ESTIMATES							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t 	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1	12.36002	5.08621	2.43	0.0187	.	.
X1	1	-0.29160	0.09165	-3.18	0.0025	0.16321	0.16561
X2	1	-0.22040	0.11482	-1.92	0.0605	0.06737	0.06737

15. The accompanying SAS computer output relates to the *Business Week* data from Problem 13 in Chapter 8. Use this output and, if necessary, the output from Problem 12 in Chapter 9 to answer the following questions.
- Determine $r^2_{Y|X_2, X_3}$, the squared multiple correlation between the yield (Y) and the independent variables 1989 rank (X_2) and P-E ratio (X_3).
 - For the *Business Week* data, answer the questions posed in Problem 12, parts (b) through (f).

Edited SAS Output (PROC REG) for Problem 15

Regression of Y on X2 and X3

PARAMETER ESTIMATES			
Variable	DF	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1		
X2	1	0.03671	0.27066
X3	1	0.53584	0.53584

16. The accompanying SAS computer output relates to the 1990 Census data from Problem 14 in Chapter 8. Use this output and, if necessary, the output from Problem 13 in Chapter 9 to answer the following questions.
- Determine $r^2_{Y|X_1, X_2}$, the squared multiple correlation between the rate of owner occupancy ($Y = \text{OWNEROCC}$) and the independent variables for monthly ownership costs ($X_1 = \text{OWNCOST}$) and proportion of population living in urban areas ($X_2 = \text{URBAN}$).
 - For the Census data, answer the questions posed in Problem 12, parts (b) through (f).

Edited SAS Output (PROC REG) for Problem 16

Regression of OWNEROCC on OWN COST and URBAN

PARAMETER ESTIMATES			
Variable	DF	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1		
OWNCOST	1	0.22068	0.16119
URBAN	1	0.26297	0.26297

17. This problem refers to the pond ecology data of Chapter 5, Problem 20.
- Determine the squared multiple correlation for the regression of copepod count on zooplankton and phytoplankton counts.
 - Determine the partial correlation between each independent variable in (a) and copepod count, controlling for the other independent variable. Using the technique of Problem 2, perform tests of significance of the partial correlations.
 - Based on your answers in (b), which independent variables should be included in the regression model, and in what order of importance?

References

- Blalock, H. M., Jr., ed. 1971. *Causal Models in the Social Sciences*. Chicago: Aldine Publishing.
Bollen, K. A. 1989. *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

11

Confounding and Interaction in Regression

11.1 Preview

A regression analysis may have two different goals: to predict the dependent variable by using a set of independent variables and to quantify the relationship of one or more independent variables to a dependent variable. The first of these goals focuses on finding a model that fits the observed data and predicts future data as well as possible, whereas the second pertains to producing valid and precise statistical inferences about one or more regression coefficients in the model. The second goal is of particular interest when the research question concerns clarifying a causal process, such as trying to identify one or more determinants of a disease or other health-related outcome.

Confounding and *interaction* are the two methodological concepts most relevant to attaining the second goal. In this chapter, we use regression terminology to describe these concepts. More general discussions of this subject can be found elsewhere (e.g., Kleinbaum, Kupper, and Morgenstern 1982; Rothman, Greenland, and Lash 2008) within the context of epidemiological research, which typically addresses etiologic questions involving the second goal. We begin here with a general overview of these concepts, after which we discuss the regression formulation of each concept separately. In Chapter 13, we describe a popular regression procedure, analysis of covariance (ANACOVA), which may be used to adjust or correct for problems of confounding. Subsequently, in Chapter 16, we briefly describe a strategy for obtaining a “best” regression model that incorporates the assessment of both confounding and interaction.

11.2 Overview

The assessment of both confounding and interaction involves the quantification of an association between two or more variables so that additional variables that may affect this association are appropriately taken into account. The measure of association chosen usually

depends on the characteristics of the variables of interest. For example, if both variables are continuous, as in the classic linear regression context, the measure of association is typically a regression coefficient. The additional variables to be considered are synonymously referred to as *extraneous variables*, *control variables*, or *covariates*. The essential questions about these variables are whether and how they should be incorporated into a model that can be used to validly estimate the association of interest.

Suppose that we are conducting an observational study of adults to assess whether physical activity level (PAL) is associated with systolic blood pressure (SBP), accounting (i.e., controlling) for AGE. The extraneous variable here is AGE. We need to determine whether we can ignore AGE in our analysis and still correctly assess the PAL–SBP association. In particular, we need to address the following two questions: (1) Is the estimate of the association between PAL and SBP meaningfully different depending on whether or not we ignore AGE? (2) Is the estimate of the association between PAL and SBP meaningfully different for different values of AGE? The first question is concerned with confounding; the second question focuses on interaction.

In general, *confounding exists if meaningfully different interpretations of the relationship of interest result when an extraneous variable is ignored or included in the data analysis*. In practice, the assessment of confounding requires a comparison between a *crude* estimate of an association (which ignores the extraneous variable[s]) and an *adjusted* estimate of the association (which accounts in some way for the extraneous variable[s]). If the crude and adjusted estimates are meaningfully different, confounding is present, and one or more extraneous variables must be included in our data analysis. This definition does not require a statistical test but rather a comparison of estimates obtained from the data (see Kleinbaum, Kupper, and Morgenstern 1982, chap. 13, or Kleinbaum 2002, Lesson 10, for further discussion of this point).

For example, a crude estimate of the relationship between PAL and SBP (ignoring AGE) is given by the estimated regression coefficient—say, $\hat{\beta}_1$ —of the variable PAL in the straight-line model that predicts SBP using just PAL. In contrast, an adjusted estimate is given by the estimated regression coefficient $\hat{\beta}_1^*$ of the same variable PAL in the multiple regression model that uses both PAL and AGE to predict SBP. In particular, if PAL is defined dichotomously (e.g., PAL = 1 or 0 for high or low physical activity, respectively), then the crude estimate is simply the crude difference between the mean SBPs in each physical activity group, and the adjusted estimate represents an adjusted difference in these two mean SBPs that controls for AGE. In general, confounding is present if any meaningful difference exists between the crude and adjusted estimates.

In this particular example, it is known that PAL decreases with AGE and that SBP increases with AGE, so that AGE is associated with both PAL and SBP. Thus, any estimated association between PAL and SBP would be affected by the age distribution of the subjects in the study, and as a result, not adjusting appropriately for AGE could lead to an erroneous conclusion about the true PAL–SBP relationship. More generally, a variable (like AGE) may be a confounder if it is associated in the observed data both with the predictor of interest (like PAL) and with the outcome of interest (like SBP). Other definitions for confounding exist, and Kleinbaum, Kupper, and Morgenstern (1982) and Rothman, Greenland, and Lash (2008) discuss this in greater detail.

Interaction is the condition in which the relationship of interest is different at different levels (i.e., values) of the extraneous variable(s). In contrast to confounding, the assessment

of interaction does not consider either a crude estimate or an (overall) adjusted estimate; instead, it focuses on describing the relationship of interest at different values of the extraneous variables. For example, in assessing interaction due to AGE in describing the PAL–SBP relationship, we must determine whether some description (i.e., estimate) of the relationship varies with different values of AGE (e.g., whether the relationship is strong at older ages and weak at younger ages). If the PAL–SBP relationship does vary with AGE, then we say that an AGE \times (read “by”) PAL interaction exists. To assess interaction, we may employ a statistical test in addition to subjective evaluation of the meaningfulness (e.g., clinical importance) of an estimated interaction effect. Again, for further discussion, see Kleinbaum, Kupper, and Morgenstern (1982) or Kleinbaum (2002).

When both confounding and interaction are considered for the same data set, using an overall (adjusted) estimate as a summary index of the relationship of interest tends to mask any (strong) interaction effects that may be present. For example, if the PAL–SBP association differs meaningfully at different values of AGE, using a single overall estimate, such as the regression coefficient of PAL in a multiple regression model containing both AGE and PAL, would hide this interaction finding. This illustrates the following important principle: *interaction should be assessed before confounding is assessed; the use of a summary (adjusted) estimate that controls for confounding is recommended only when there is no meaningful interaction* (Kleinbaum, Kupper, and Morgenstern 1982, chap. 13).

A variable may manifest both confounding and interaction, neither, or only one of the two. But if strong interaction is found, an adjustment for confounding is inappropriate.

We are now ready to use regression terminology to address how these concepts can be employed, assuming a linear model and a continuous dependent variable. In epidemiologic and clinical research studies, the dependent variable is often a dichotomous health outcome or state (e.g., develops diabetes or not). A regression analog for a dichotomous outcome variable might, for example, involve a logistic rather than a linear model. Logistic regression analysis is discussed in detail in Chapter 22. A more detailed discussion in which confounding and interaction are considered can be found in Kleinbaum, Kupper, and Morgenstern (1982, chaps. 20–24) or Kleinbaum and Klein (2002, chaps. 6–7).

11.3 Interaction in Regression

In this section, we describe how two independent variables can interact to affect a dependent variable and how such an interaction can be represented by an appropriate regression model.

11.3.1 First Example

To illustrate the concept of interaction, let us consider the following simple example. Suppose that we wish to determine how two independent variables—temperature (T) and catalyst concentration (C)—jointly affect the growth rate (Y) of organisms in a certain biological system. Further, suppose that two particular temperature levels (T_0 and T_1) and two particular levels of catalyst concentration (C_0 and C_1) are to be examined and that an experiment is performed in which an observation on Y is obtained for each of the four combinations of temperature–catalyst concentration level: (T_0, C_0) , (T_0, C_1) , (T_1, C_0) , and (T_1, C_1) .

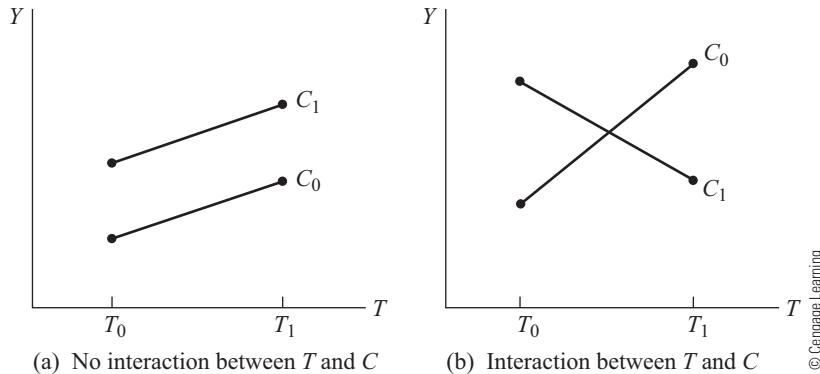


FIGURE 11.1 Graphs of noninteracting and interacting independent variables

C_1). In statistical parlance, this experiment is called a *complete factorial experiment* because observations on Y are obtained for all combinations of settings for the independent variables (or factors). The advantage of a complete factorial experiment is that any existing interaction effects can be detected and estimated efficiently.

Now, let us consider two graphs based on two hypothetical data sets for the experimental scheme just described. Because both lines are parallel, Figure 11.1(a) suggests that the rate of change¹ in the growth rate as a function of temperature remains the same, regardless of the level of catalyst concentration; in other words, the relationship between Y and T does not in any way depend on C .

We are not saying that Y and C are unrelated but that the relationship between Y and T does not vary as a function of C . When this is the case, we say that T and C do not interact or, equivalently, that there is no $T \times C$ interaction effect. Practically speaking, this means that we can investigate the effects of T and C on Y independently of one another and that we can legitimately talk about the separate effects (sometimes called the *main effects*) of T and C on Y .

One way to quantify the relationship depicted in Figure 11.1(a) is with a regression model of the form

$$\mu_{Y|T,C} = \beta_0 + \beta_1 T + \beta_2 C \quad (11.1)$$

Here the change in the mean of Y for a one-unit change in T is equal to β_1 , regardless of the level of C . In fact, changing the level of C in (11.1) has only the effect of shifting the straight line relating $\mu_{Y|T,C}$ and T either up or down, without affecting the value of the slope β_1 , as seen in Figure 11.1(a). In particular, $\mu_{Y|T,C_0} = (\beta_0 + \beta_2 C_0) + \beta_1 T$ and $\mu_{Y|T,C_1} = (\beta_0 + \beta_2 C_1) + \beta_1 T$.

In general, then, we might say that no interaction is synonymous with parallelism, in the sense that the response curves of Y versus T for fixed values of C are parallel; in other words,

¹ For readers familiar with calculus, the phrase “rate of change” is related to the notion of a derivative of a function. In particular, Figure 11.1(a) portrays a situation in which the partial derivative with respect to T of the response function relating the mean of Y to T and C is independent of C .

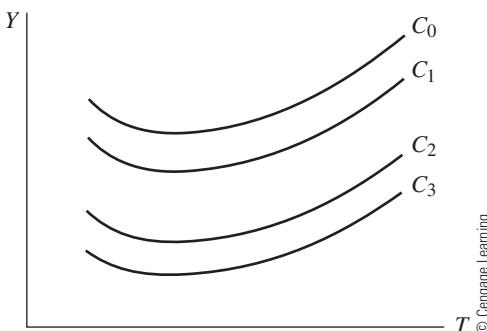


FIGURE 11.2 Response curves illustrating no interaction between T and C

these response curves (which may be linear or nonlinear) all have the same general shape, differing from one another only by additive constants independent of T (see, e.g., Figure 11.2).

In contrast, Figure 11.1(b) depicts a situation in which the relationship between Y and T depends on C ; in particular, Y appears to increase with increasing T when $C = C_0$ but to decrease with increasing T when $C = C_1$. In other words, the behavior of Y as a function of temperature cannot be considered in isolation from catalyst concentration. When this is the case, we say that T and C interact or, equivalently, that there is a $T \times C$ interaction effect. Practically speaking, this means that it does not make much sense to talk about the separate (or main) effects of T and C on Y , since T and C do not operate independently of one another in their effects on Y .

One way to represent such interaction effects mathematically is to use a regression model of the form

$$\mu_{Y|T,C} = \beta_0 + \beta_1 T + \beta_2 C + \beta_{12} T C \quad (11.2)$$

Here the change in the mean value of Y for a one-unit change in T is equal to $\beta_1 + \beta_{12}C$, which clearly depends on the level of C . In other words, introducing a product term such as $\beta_{12}TC$ in a regression model of the type given in (11.2) is one way to account for the fact that two such factors as T and C do not operate independently of one another. For our particular example, when $C = C_0$, model (11.2) can be written as

$$\mu_{Y|T,C_0} = (\beta_0 + \beta_2 C_0) + (\beta_1 + \beta_{12} C_0)T$$

and when $C = C_1$, model (11.2) becomes

$$\mu_{Y|T,C_1} = (\beta_0 + \beta_2 C_1) + (\beta_1 + \beta_{12} C_1)T$$

A negative interaction effect is to be expected here, since Figure 11.1(b) suggests that the slope of the linear relationship between Y and T decreases (goes from positive to negative in sign) as C changes from C_0 to C_1 . This implies that the slope $(\beta_1 + \beta_{12} C_0)$ at C_0 is positive and the slope $(\beta_1 + \beta_{12} C_1)$ at C_1 is negative. For this to be true, the interaction effect β_{12} must be negative. Of course, it is possible for β_{12} to be positive, in which case the interaction effect manifests itself as a larger positive value for the slope when $C = C_1$ than when $C = C_0$.

11.3.2 Interaction Modeling in General

As the preceding illustration suggests, interaction among independent variables can generally be described in terms of a regression model that involves product terms. Although no precise rules exist for specifying which terms to include, we offer three complementary approaches below.

In the first approach, one includes only interactions that are reasonable a priori, based on evaluation of the literature and one's expertise. These need not be well-established interactions but rather ones that are worth considering in the model. This approach helps to keep regression models parsimonious and to ensure that the model's explanatory variables are readily interpretable.

Oftentimes, a priori knowledge of possible interactions is unavailable or incomplete, and a second approach is to specify a model with a full set of product terms. For example, if interactions among three variables X_1 , X_2 , and X_3 are of interest, such a model to consider is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 + \beta_7 X_1 X_2 X_3 + E \quad (11.3)$$

Here the two-factor products of the form $X_i X_j$ are often referred to as *first-order interactions*; the three-factor products, like $X_1 X_2 X_3$, are called *second-order interactions*; and so on for higher-order products. The higher the order of interaction, the more difficult it becomes to interpret its meaning. As a general rule, if a higher-order interaction is specified in a model, all lower-order terms contained in it must also be included in the model. This is known as ensuring that a model is *hierarchically well-formulated* (HWF). Failure to maintain a regression model that is HWF may lead to misleading findings. The presence of the second-order interaction $X_1 X_2 X_3$ requires that all first-order interactions and the main effects X_1 , X_2 , and X_3 also appear in the model. If it is determined that the second-order interaction is important and should be retained in the model, then all associated lower-order terms must also be retained.

Model (11.3) is not the most general model possible for considering the three variables X_1 , X_2 , and X_3 . Additional terms such as X_i^2 , X_j^2 , $X_i X_j^2$, $X_i X_j^3$, $X_i^2 X_j^2$, and so on can be included. Nevertheless, there is a limit on the total number of such terms: a model with an intercept (β_0) term cannot contain more than $n - 1$ independent variables when n is the total number of observations in the data. Moreover, it may not even be possible to fit a model with fewer than $n - 1$ variables reliably if some of the variables (e.g., higher-order products) are highly correlated with other variables in the model, as would be the case when the model contains several interaction terms. This problem, called *collinearity*, is discussed in Chapter 14.

The third approach deals with a situation where the association between the dependent variable and a particular factor (or factors) is of primary interest. In this case, model (11.3) may be considered too general, and one may choose to include only interactions with the primary factor(s). For example, if the purpose of one's study is to describe the relationship between X_1 and Y , controlling for the possible confounding and/or interaction effects of X_2 and X_3 , the following simpler model may be of more interest than (11.3):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + E \quad (11.4)$$

The terms X_1X_2 and X_1X_3 describe the interactions of X_2 and X_3 , respectively, with X_1 . In contrast, the term X_2X_3 , which is not contained in model (11.4), has no relevance to interaction involving X_1 .

In using statistical testing to evaluate interaction for a given regression model, we have a number of available options. (A more detailed discussion of how to select variables is given in Chapter 16.) One approach is to test globally for the presence of any kind of interaction and then, if significant interaction is found, to identify particular interaction terms of importance by using other tests. For example, in considering model (11.3), we could first test $H_0: \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$ using the multiple partial F statistic

$$F(X_1X_2, X_1X_3, X_2X_3, X_1X_2X_3 | X_1, X_2, X_3)$$

which has an $F_{4, n - 8}$ distribution when H_0 is true. If this F statistic is found to be significant, individually important interaction terms may then be identified by using selected partial F tests.

A second way to assess interaction is to test for interaction in a hierarchical sequence, beginning with the highest-order terms and then proceeding sequentially through lower-order terms if higher-order terms are not significant. Using model (11.3), for example, we might first test $H_0: \beta_7 = 0$, which considers the second-order interaction, and then test $H_0: \beta_4 = \beta_5 = \beta_6 = 0$ in a reduced model (excluding the three-way product term $X_1X_2X_3$) if the result of the first test is nonsignificant.

11.3.3 Second Example

We present here an example of how to assess interaction by using computer output (from SAS). In Chapters 5 through 7, the age–systolic blood pressure example was used to illustrate the main principles and methods of straight-line regression analysis. These hypothetical data were shown, upon analysis, to support the commonly found observation that blood pressure increases with age.

Another question that can be answered by such data is whether an interaction exists between age and sex: Does the slope of the straight line relating systolic blood pressure to age significantly differ for males and for females? We will continue with our age–systolic blood pressure example to investigate this possible interaction.

In Chapter 5, we observed that the data point ($AGE = 47$, $SBP = 220$) is an outlier quite distinct from the rest of the data. We will discard this data point in all further analyses; henceforth, we will assume that all 29 remaining observations on age and systolic blood pressure considered previously were made on females and that a second sample of observations on age and systolic blood pressure was collected on 40 males. The data set for all 69 individuals is given in Table 11.1. Note that each individual's sex is recorded in the binary 0/1 variable SEX . These so-called dummy variables are formally introduced in Chapter 12.

The accompanying SAS computer output is given for the following regression model:

$$Y = \beta_0 + \beta_1X + \beta_2Z + \beta_3XZ + E$$

where Z represents SEX ($Z = 0$ if male, $Z = 1$ if female) and XZ is the interaction between AGE and SEX . To evaluate the importance of the interaction term XZ , recall that the model, in order to be HWF, must also contain the lower-order main-effect terms X and Z .

TABLE 11.1 Data on systolic blood pressure (SBP) and age (AGE) for 40 males and 29 females (SEX), together with associated results for comparing two straight-line regression equations

Obs.	SEX (Z)	SBP (Y)	AGE (X)												
1	0	158	41	19	0	130	22	37	0	148	35	55	1	130	48
2	0	185	60	20	0	138	21	38	0	140	33	56	1	135	45
3	0	152	41	21	0	150	38	39	0	132	26	57	1	114	17
4	0	159	47	22	0	156	52	40	0	169	61	58	1	116	20
5	0	176	66	23	0	134	41	41	1	144	39	59	1	124	19
6	0	156	47	24	0	134	18	42	1	138	45	60	1	136	36
7	0	184	68	25	0	174	51	43	1	145	47	61	1	142	50
8	0	138	43	26	0	174	55	44	1	162	65	62	1	120	39
9	0	172	68	27	0	158	65	45	1	142	46	63	1	120	21
10	0	168	57	28	0	144	33	46	1	170	67	64	1	160	44
11	0	176	65	29	0	139	23	47	1	124	42	65	1	158	53
12	0	164	57	30	0	180	70	48	1	158	67	66	1	144	63
13	0	154	61	31	0	165	56	49	1	154	56	67	1	130	29
14	0	124	36	32	0	172	62	50	1	162	64	68	1	125	25
15	0	142	44	33	0	160	51	51	1	150	56	69	1	175	69
16	0	144	50	34	0	157	48	52	1	140	59				
17	0	149	47	35	0	170	59	53	1	110	34				
18	0	128	19	36	0	153	40	54	1	128	42				

© Cengage Learning

From the output, we see that when $Z = 0$ (i.e., when SEX = male), the estimated model reduces to

$$\begin{aligned} Y &= 110.04 + 0.96X + (-12.96)(0) + (-0.01)X(0) \\ &= 110.04 + 0.96X \end{aligned}$$

This is the estimated regression line for males. When $Z = 1$, the estimated model gives the estimated regression line for females:

$$\begin{aligned} Y &= 110.04 + 0.96X + (-12.96)(1) + (-0.01)X(1) \\ &= 97.08 + 0.95X \end{aligned}$$

Plotting these two lines (Figure 11.3), we see that they appear to be almost parallel, indicating that there is probably no statistically significant interaction. We can confirm this lack of significance by inspecting the output: the partial F test for the significance of β_3 , given that X and Z are in the model, has a P -value of .9342. Therefore, the slopes of the straight lines

Edited SAS Output

SBP Regressed on AGE (X), SEX (Z), and AGE \times SEX Interaction (XZ)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	18010.32871	6003.44290	75.02	<.0001
Error	65	5201.43940	80.02214		
Corrected Total	68	23211.76812			
R-Square	Coeff Var	Root MSE	Y Mean		
0.775914	6.014814	8.945510	148.7246		
Source	DF	Type I SS	Mean Square	F Value	Pr > F
X	1	14951.25461	14951.25461	186.84	<.0001
Z	1	3058.52475	3058.52475	38.22	<.0001
XZ	1	0.54936	0.54936	0.01	0.9342
Source	DF	Type III SS	Mean Square	F Value	Pr > F
X	1	7971.007140	7971.007140	99.61	<.0001
Z	1	273.443297	273.443297	3.42	0.0691
XZ	1	0.549356	0.549356	0.01	0.9342
Parameter	Estimate	Standard Error	t Value	Pr > t	
Intercept	110.0385285	4.73610350	23.23	<.0001	
X	0.9613526	0.09632327	9.98	<.0001	
Z	-12.9614443	7.01172459	-1.85	0.0691	
XZ	-0.0120301	0.14519328	-0.08	0.9342	

© Cengage Learning

relating systolic blood pressure and age do not statistically significantly differ for males and for females: there is no interaction between age and sex in this situation. These tests of *parallelism* are revisited in Chapters 12 and 13.

11.3.4 Third Example

We now consider a study to assess physical activity level (PAL) as a predictor of systolic blood pressure (SBP), controlling for AGE and SEX. A model that allows for possible interactions of both AGE with PAL and SEX with PAL is given by

$$\begin{aligned} \text{SBP} = & \beta_0 + \beta_1(\text{PAL}) + \beta_2(\text{AGE}) + \beta_3(\text{SEX}) + \beta_4(\text{PAL} \times \text{AGE}) \\ & + \beta_5(\text{PAL} \times \text{SEX}) + E \end{aligned}$$

Notice the absence of any term involving AGE \times SEX; such a term does not indicate interaction associated with the study variable of interest (PAL).

To assess interaction for this model, we might first perform a multiple partial *F* test of $H_0: \beta_4 = \beta_5 = 0$; if this test was found to be significant, we could conduct partial *F* tests to

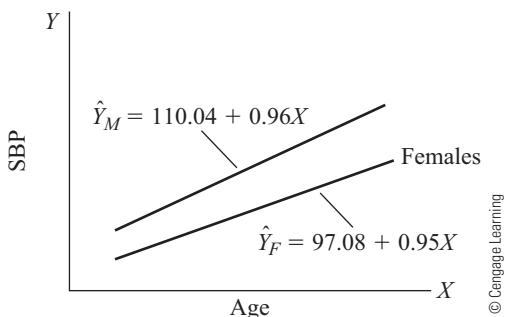


FIGURE 11.3 Comparison by sex of straight-line regressions of systolic blood pressure on age

determine whether one or more of these product terms should be kept in the model. If the first test was found to be nonsignificant, we could then simplify the full model by removing these two product terms entirely, leaving the reduced model $SBP = \beta_0 + \beta_1(PAL) + \beta_2(AGE) + \beta_3(SEX) + E$. At this point, the interaction phase of model building would be complete, and the next step would involve the assessment of confounding, as discussed in the next section.

11.3.5 Interaction versus Effect Modification

The term *effect modification* (also called *effect measure modification*) is often used interchangeably with the term *interaction*, particularly for health-related research situations considered in the field of epidemiology.² From an epidemiological perspective, effect modification is used to describe an association between a predictor (i.e., exposure) of interest and a health outcome that is “modified” (i.e., is different), depending on the value of one or more “control” variables. Such control variables are called *effect modifiers* of the relationship between exposure and outcome. In contrast, the term *interaction* is used to describe the statistical property of a mathematical model containing a predictor variable defined as the product of two or more component predictor variables (e.g., $X_k = X_i \times X_j$). Such a product term represents a combination (i.e., *interaction*) of its component terms to form a predictor that explains additional statistical variation in the outcome variable over and above the independent contributions of each component predictor.

To illustrate effect modification, consider the regression model provided in the previous section to describe the association of the predictor, physical activity level (PAL), with the health outcome, systolic blood pressure (SBP), controlling for AGE and SEX:

$$SBP = \beta_0 + \beta_1(PAL) + \beta_2(AGE) + \beta_3(SEX) + \beta_4(PAL \times AGE) + \beta_5(PAL \times SEX) + E$$

This model contains two interaction terms, $PAL \times SEX$ and $PAL \times AGE$. As previously described, the assessment of interaction is typically supported using the results of statistical

² Although the terms *effect modification* and *interaction* are often used interchangeably, there is some controversy in the epidemiologic literature about the precise definitions of effect modification and interaction (see Kleinbaum, Kupper, and Morgenstern, 1982, chap. 19, and Kleinbaum, 2002, Lesson 10). One distinction frequently made is that effect modification describes a non-quantitative clinical or biological attribute of a population, whereas interaction is typically quantitative and data-specific and, in particular, depends on the scale on which the “interacting” variables are measured.

testing, where, in this case, the null hypothesis is $H_0: \beta_4 = \beta_5 = 0$. Assuming that such a statistical test is significant and that both interaction terms need to remain in the model, the fitted model would be written as

$$\widehat{\text{SBP}} = \hat{\beta}_0 + \hat{\beta}_1(\text{PAL}) + \hat{\beta}_2(\text{AGE}) + \hat{\beta}_3(\text{SEX}) + \hat{\beta}_4(\text{PAL} \times \text{AGE}) + \hat{\beta}_5(\text{PAL} \times \text{SEX})$$

We would then conclude that there is *effect modification* of the association of PAL with SBP and that AGE and SEX are *effect modifiers* of this association. In other words, the association of PAL with SBP differs (i.e., is *modified*), depending on a person's AGE and/or SEX.

For example, suppose we compare the association between PAL and SBP for a 40-year-old female with the corresponding association for a 30-year-old male. Suppose, also, that AGE is a continuous variable and SEX is coded as 1 if female and 0 if male. Then the predicted SBP for a 40-year-old female (i.e., AGE = 40, SEX = 1) is

$$\begin{aligned}\widehat{\text{SBP}}_{40F} &= \hat{\beta}_0 + \hat{\beta}_1(\text{PAL}) + 40\hat{\beta}_2 + \hat{\beta}_3 + 40\hat{\beta}_4(\text{PAL}) + \hat{\beta}_5(\text{PAL}) \\ &= (\hat{\beta}_0 + 40\hat{\beta}_2 + \hat{\beta}_3) + (\hat{\beta}_1 + 40\hat{\beta}_4 + \hat{\beta}_5)(\text{PAL}) \\ &= \hat{\beta}_{0(40F)} + \hat{\beta}_{1(40F)}\text{PAL}\end{aligned}$$

where $\hat{\beta}_{0(40F)} = (\hat{\beta}_0 + 40\hat{\beta}_2 + \hat{\beta}_3)$ and $\hat{\beta}_{1(40F)} = (\hat{\beta}_1 + 40\hat{\beta}_4 + \hat{\beta}_5)$. Note that once AGE and SEX are specified, the model simplifies to a straight line model for predicting SBP from PAL. The estimated slope for this model, $\hat{\beta}_{1(40F)}$, represents the measure of association relating PAL to SBP.

Similarly, the predicted SBP for a 30-year-old male (AGE = 30, SEX = 0) is

$$\begin{aligned}\widehat{\text{SBP}}_{30M} &= \hat{\beta}_0 + \hat{\beta}_1(\text{PAL}) + 30\hat{\beta}_2 + 0 + 30\hat{\beta}_4(\text{PAL}) + 0 \\ &= (\hat{\beta}_0 + 30\hat{\beta}_2) + (\hat{\beta}_1 + 30\hat{\beta}_4)(\text{PAL}) \\ &= \hat{\beta}_{0(30M)} + \hat{\beta}_{1(30M)}\text{PAL}\end{aligned}$$

where $\hat{\beta}_{0(30M)} = (\hat{\beta}_0 + 30\hat{\beta}_2)$ and $\hat{\beta}_{1(30M)} = (\hat{\beta}_1 + 30\hat{\beta}_4)$. This model is also a straight line model for predicting SBP from PAL, but here the slope of the model that represents the measure of association is $\hat{\beta}_{1(30M)}$. It follows that the two estimated slopes will differ—i.e., $\hat{\beta}_{(40F)} \neq \hat{\beta}_{(30M)}$ —whenever $\hat{\beta}_4 \neq 0$ and/or $\hat{\beta}_5 \neq 0$. Thus, the association between PAL and SBP varies with the values of AGE and/or SEX; i.e., the association is *modified* by AGE and SEX.

11.4 Confounding in Regression

We emphasized earlier (in Sections 11.2 and 11.3.2) that the assessment of confounding is questionable in the presence of interaction. Thus, our discussion of confounding here assumes throughout that no interaction is present.³

³ It is possible, however, to assess confounding for variables that are not components of interaction terms. For example, if we consider the model $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_1X_3 + E$, where X_1 is the study variable of interest, we might want to consider whether X_2 is a confounder, since it is not a component of X_1X_3 , the only interaction term in the model. For more realistic examples, see Kleinbaum, Kupper, and Morgenstern (1982, chap. 23).

11.4.1 Controlling for One Extraneous Variable

Suppose that we are interested in describing the relationship between an independent variable T and a continuous dependent variable Y , taking into account the possible confounding effect of a third variable C . As described in Section 11.2, the assessment of confounding requires us to compare a crude estimate of the T - Y relationship—which ignores the effect of the control variable (C)—with an estimate of the relationship that accounts (or controls) for this variable. This comparison can be expressed in terms of the following two regression models:

$$Y = \beta_0 + \beta_1 T + \beta_2 C + E \quad (11.5)$$

and

$$Y = \beta_0 + \beta_1 T + E \quad (11.6)$$

The assumption of no $T \times C$ interaction precludes the need to consider a product term of the form TC in these models.

From model (11.5), we can express the relationship between T and Y , adjusted for the variable C , in terms of the (partial) regression coefficient (β_1) of the T variable. The estimate of β_1 (which we will denote by $\hat{\beta}_{1|C}$), obtained from least-squares fitting of model (11.5), is an adjusted-effect measure in the sense that it gives the estimated change in Y per unit change in T after accounting for C (i.e., with C in the model).

A crude estimate of the T - Y relationship is the estimated coefficient of T (namely, $\hat{\beta}_1$) based on model (11.6)—a model that does not involve the variable C .

Thus, we have the following general rule for assessing the presence of confounding when only one independent variable is to be controlled: confounding is present if the estimate of the coefficient (β_1) of the study variable T meaningfully changes when the variable C is removed from model (11.5)—that is, if

$$\hat{\beta}_{1|C} \neq \hat{\beta}_1 \quad (11.7)$$

where $\hat{\beta}_{1|C}$ denotes the (adjusted for C) estimate of β_1 using model (11.5) and $\hat{\beta}_1$ denotes the (crude) estimate of β_1 using model (11.6).

The \neq sign in expression (11.7) indicates that a subjective decision is required as to whether the two estimates are meaningfully different; that is, we must determine subjectively whether the two estimates each describe a different interpretation of the T - Y association in question. A statistical test is neither required nor appropriate (Kleinbaum, Kupper, and Morgenstern 1982, chap. 13, or Kleinbaum 2002, Lesson 10).

As an example, suppose that Y denotes SBP, T denotes PAL, and C denotes AGE. For some set of data, suppose we found that

$$\hat{\beta}_{1|AGE} = 4.1 \text{ and } \hat{\beta}_1 = 15.9$$

Then, we could conclude that a 1-unit change in PAL yields about a 16-unit change in SBP when AGE is ignored and that a 1-unit change in PAL yields only a 4.1-unit change in SBP when AGE is controlled; that is, the association between PAL and SBP is much weaker after controlling for

AGE. (As a special case, if PAL is a 0–1 variable, then $\hat{\beta}_1$ gives the crude difference in mean SBPs between the two PAL groups, and $\hat{\beta}_{1|AGE}$ gives an adjusted [for AGE] difference in mean SBPs.) Thus, we would treat AGE as a confounder and control for it in the analysis.

As another example, suppose that

$$\hat{\beta}_{1|AGE} = 6.2 \text{ and } \hat{\beta}_1 = 6.1$$

Here we would be inclined to say that AGE is not a confounder because there is no meaningful difference between the estimates 6.2 and 6.1. Unfortunately, an investigator may have to deal with much more problematic comparisons, such as $\hat{\beta}_{1|AGE} = 4.1$ versus $\hat{\beta}_1 = 5.5$. In comparing such estimates numerically, one must consider the clinical importance of the numerical difference between estimates, based on (a priori) knowledge of the variable(s) involved. For instance, since the coefficients 4.1 and 5.5 estimate, respectively, adjusted and crude differences in mean SBPs between high and low PAL groups, it is important to decide whether a mean difference of 5.5 is clinically more important than a mean difference of 4.1. One approach to this problem is to control for any variable (as a confounder) that changes the crude effect estimate by some prespecified amount determined by clinical judgment (e.g., 10%).

One approach sometimes used (incorrectly) to assess confounding is, for example, to conduct a statistical test of $H_0: \beta_2 = 0$ in model (11.5). Such a test does not address confounding but rather *precision*; that is, such a test evaluates whether significant additional variation in Y is explained by adding C to a model already containing T . An almost equivalent approach is to determine whether a confidence interval for β_1 , the coefficient of T , is considerably narrower when C is in the model than when it is not. Precision is often an important issue to assess when considering extraneous factors, but it differs fundamentally from confounding. For etiologic questions, confounding, which concerns validity (i.e., do you have the right answer?), usually takes precedence over precision. Another reason for not focusing on β_2 is that, if $\hat{\beta}_2 \neq 0$, it does not follow that $\hat{\beta}_{1|C} \neq \hat{\beta}_1$. In other words, $\hat{\beta}_2 \neq 0$ is not a sufficient condition for confounding.⁴

What type of variables (i.e., covariates) should be considered for control as potential confounders? Although our answer is somewhat debatable, we consider that a list of eligible variables should be constructed based on prior knowledge and/or research about the relationship of the dependent variable to each covariate under consideration. In particular, we recommend that only variables known to be reasonably predictive of (i.e., associated with) the dependent variable be considered as potential confounders and/or effect modifiers. In epidemiological terms, such variables are generally referred to as *risk factors* (Kleinbaum 2002). The idea here is to restrict attention to controlling only the (previously studied) extraneous variables that the investigator anticipates may account for the hypothesized relationship between T and Y currently being studied. To develop such a list, the investigators have to make a subjective decision.⁵

⁴ Suppose that $n = 6$ and that we have the following data for (T, C, Y) : (1, 0, 4), (1, 1, 5), (1, 2, 6), (0, 0, 1), (0, 1, 2), and (0, 2, 3). Then unweighted least-squares fitting gives $\hat{Y} = 1 + 3T + C$ when T and C are predictors, whereas $\hat{Y} = 2 + 3T$ when C is ignored. Thus, $\hat{\beta}_2 = 1 (\neq 0)$, yet there is no confounding, since $\hat{\beta}_1 = 3 = \hat{\beta}_{1|C}$.

⁵ As a caveat to these recommendations, we note that certain variables usually referred to as *intervening variables* should not be considered as potential confounders (Kleinbaum, Kupper, and Morgenstern 1982). A variable C is said to intervene between T and Y if T causes C and then C causes Y . Controlling intervening variables may spuriously reduce or eliminate any manifestation in the data of a true association between T and Y .

11.4.2 Controlling for Several Extraneous Variables

Suppose that we want to describe the association between T and Y , taking into account several covariates C_1, C_2, \dots, C_p . Similarly to our approach for handling one covariate, we can assess confounding by comparing a crude estimate of the T - Y relationship to some adjusted estimate. As before, the crude estimate can be defined in terms of a regression model like (11.6), which describes the relationship between T and Y while ignoring all covariates. To obtain the adjusted estimate, however, we must now consider an extended model defined as

$$Y = \beta_0 + \beta_1 T + \beta_2 C_1 + \beta_3 C_2 + \cdots + \beta_{p+1} C_p + E \quad (11.8)$$

(Like model [11.5], model [11.8] assumes that there is no interaction involving T , since no product terms of the form TC_i are included.)

Using this model, we can define confounding that involves several variables as follows: confounding is present if the estimate of the regression coefficient (β_1) of T in a regression model like (11.6), which ignores the variables C_1, C_2, \dots, C_p , is meaningfully different from the corresponding estimate of β_1 based on a model like (11.8), which controls for C_1, C_2, \dots, C_p —that is, if

$$\hat{\beta}_{1|C_1, C_2, \dots, C_p} \neq \hat{\beta}_1 \quad (11.9)$$

where $\hat{\beta}_{1|C_1, C_2, \dots, C_p}$ denotes the (adjusted) estimate of β_1 using (11.8) and $\hat{\beta}_1$ is the (crude) estimate of β_1 using (11.6).

One problem with applying this definition is that it addresses the question of whether confounding is present without directly identifying specific variables to be controlled.⁶ In other words, when confounding is deemed to be present based on (11.9), it may still be the case that only a subset of C_1, C_2, \dots, C_p is required for adequate control. How does one identify such a subset? More specifically, why bother to identify such a subset rather than simply controlling for all variables C_1, C_2, \dots, C_p ?

One answer to the latter question is that, when addressing the control of covariates, we should consider the possible gains in *precision*⁷ in addition to the control of confounding. In particular, we may prefer a subset of C_j variables to the entire set because the subset may provide equivalent control of confounding (i.e., may give essentially the same adjusted estimate) while providing greater precision in estimating the adjusted association of interest. However, there is no guarantee that precision will be increased by using a subset; in fact, precision may frequently be reduced. In any case, confounding should take precedence over precision in

⁶ Another problem concerns the assessment of confounding when there are two or more study variables—say, T_1 and T_2 —of interest. For this general situation, confounding may be defined to be present if (11.9) is satisfied for the coefficient of *any* study variable of interest given a model containing all such study variables and all control variables. Unfortunately, this definition has the practical drawback of requiring the researcher to make a subjective decision for each study variable of interest.

⁷ The term *precision* refers to the size of an estimator's variance or, equivalently, the narrowness of a confidence interval for the parameter being estimated. The smaller the variance of the estimator, the higher the precision of the estimator. Equivalently, since the width of a confidence interval depends on the variance estimate, the narrower the width of the confidence interval, the higher the precision of the estimator.

the sense that no subset should be considered unless it gives almost the same adjusted-effect estimate as is obtained when the researcher controls for all C_j 's.

To illustrate, suppose that $p = 5$; that is, we consider controlling for C_1, C_2, \dots, C_5 using model (11.8). Suppose, too, that the estimate of β_1 takes on the following values, depending on which sets of C_1, C_2, \dots, C_5 are controlled:

$$\hat{\beta}_{1|C_1, C_2, \dots, C_5} = 4.0, \hat{\beta}_{1|C_1, C_2} = 4.3, \hat{\beta}_1 = 16.0$$

Then, because 16.0 is much different from 4.0, one can argue that confounding is present. Yet since 4.0 is not meaningfully different from 4.3, it can also be argued that C_3, C_4 , and C_5 do not need to be controlled, since essentially the same (adjusted) estimate is obtained when we control only for C_1 and C_2 as when we adjust for all C_j 's.

Thus, for this example, we have identified two sets of C_j variables that we can use for control of confounding. Which set do we choose? The answer depends on an evaluation of precision. One approach is to compare interval estimates for some parameter of interest—one interval being derived from a model that controls for C_1 and C_2 only and the other interval coming from a model that controls for C_1 through C_5 . The logical parameter for this example is the population regression coefficient β_1 of the variable T when controlling for a particular set of C_j 's. In other words, we may compare an interval estimate for β_1 when only C_1 and C_2 are controlled to a corresponding interval estimate for β_1 when C_1 through C_5 are controlled. The narrower interval of the two is the interval reflecting the greater precision. For example, if the two 95% interval estimates are (2.6, 7.4) for $\beta_{1|C_1, C_2}$ and (1.7, 7.6) for $\beta_{1|C_1, C_2, \dots, C_5}$, then the former interval is narrower; in this case, some precision is gained by dropping C_3, C_4 , and C_5 from the model.

An alternative, but not exactly equivalent, approach to evaluating precision is to perform a statistical test for the significance of the addition of C_3, C_4 , and C_5 to a model containing T, C_1 , and C_2 . The null hypothesis for this test may be stated as $H_0: \beta_4 = \beta_5 = \beta_6 = 0$ in model (11.8), with $p = 5$. If this test is not significant, we could argue that retaining C_3, C_4 , and C_5 does not provide additional precision (i.e., explanation of variance). This would indicate that only C_1 and C_2 should be controlled for greater precision.

Because this testing approach does not always lead to the same conclusion as the internal estimation approach, the investigator may need to choose between them. In most situations, however, both approaches lead to similar results.

How do we identify which set of C_j variables to control? We have seen, by example, that we must first identify a baseline-adjusted estimate (i.e., a “gold standard”) against which to make comparisons. The ideal gold standard is the regression coefficient estimate that controls for all C_j 's that are being considered. Then, any subset of C_j 's that gives essentially the same adjusted estimate (i.e., an estimate that is not meaningfully different from the gold standard when only the C_j 's in that subset are controlled) is a candidate set for control. Several such candidates may be possible (Kleinbaum, Kupper, and Morgenstern 1982, chap. 14, or Kleinbaum and Klein 2002, chap. 7).

Which set should we finally use? The answer, again, is based on precision: we should use the set that gives the greatest precision (e.g., the tightest confidence interval for the adjusted effect under study). (For “political” reasons—that is, to convince people that all variables have been controlled—it might be better to control for C_1, C_2, \dots, C_p unless some subset of C_j 's leads to a large increase in precision.)

TABLE 11.2 An example of candidate sets for control

Candidate Set	$\hat{\beta}_{1 \text{Candidate Set}}$	95% Confidence Interval for $\beta_{1 \text{Candidate Set}}$
C_1, C_2, C_3, C_4, C_5 (baseline)	4.8	(2.3, 7.2)
C_1, C_2	5.1	(2.6, 7.6)
C_1, C_4	4.6	(2.1, 7.0)
C_1, C_2, C_4	4.7	(2.6, 6.8)

© Cengage Learning

To illustrate, suppose that the candidate sets in Table 11.2 can be identified when $p = 5$ in model (11.8). All three proper subsets of C_1, C_2, C_3, C_4 , and C_5 may be considered candidates for control, since they all give adjusted estimates that are roughly equal to the gold standard $\hat{\beta}_{1|C_1, C_2, \dots, C_5} = 4.8$. Of these candidates, the subset involving C_1, C_2 , and C_4 gives the best precision (narrowest confidence interval); therefore, this subset can be used both to control confounding and to enhance precision.

11.4.3 An Example Revisited

In Section 11.3.3, we considered a hypothetical study to assess the relationship between physical activity level (PAL) and systolic blood pressure (SBP) while controlling for both AGE and SEX. We considered a model that allows for possible interactions of AGE and SEX with PAL, and we described methods for testing for such interactions. Assuming that no significant interaction effects are found, the resulting reduced model is

$$\text{SBP} = \beta_0 + \beta_1(\text{PAL}) + \beta_2(\text{AGE}) + \beta_3(\text{SEX}) + E$$

Given this no-interaction model, we next assess confounding: Does the coefficient of PAL change when AGE and/or SEX is dropped from the model? To answer this, we can examine the estimate of the coefficient of PAL in four models—namely, one including both AGE and SEX, one involving either AGE or SEX but not both, and one involving neither. The gold standard model for comparison contains both control variables and PAL. Then, for example, if the estimate of β_1 changes considerably when at least one control variable is dropped from this gold standard model, we need to control for both AGE and SEX. However, if we obtain essentially the same estimate of β_1 (as obtained using the gold standard model) when only AGE is in the model, we do not need to retain SEX in the model to control for confounding. Even so, including the SEX variable in addition to AGE may increase or decrease precision. Thus, the decision as to whether to control for just AGE or for both AGE and SEX depends, for example, on a comparison of confidence intervals for β_1 . If the confidence interval is considerably narrower when only AGE is controlled, we should not retain SEX in the model.

Finally, once we make a decision about which variables to control (i.e., which model is the best for providing a valid and precise estimate of the coefficient of PAL), we make statistical inferences about the true PAL–SBP relationship. Given the no-interaction model, this involves testing $H_0: \beta_1 = 0$ in the best model and then obtaining an interval estimate of β_1 .

11.5 Summary and Conclusions

Confounding and interaction are two methodological concepts that pertain to assessing a relationship between independent and dependent variables.

Interaction, which takes precedence over confounding, exists when the relationship of interest differs at different levels of extraneous (control) variables. In multiple linear regression, interaction is evaluated by using statistical tests about product terms involving basic independent variables in the model.

Confounding, which is not evaluated with statistical testing, is present when the effect of interest differs depending on whether an extraneous variable is ignored or retained in the analysis. In regression terms, confounding is assessed by comparing crude versus adjusted regression coefficients from different models.

When several potential confounders are being considered, it may be worthwhile to identify nonconfounders that can be dropped from the model to gain precision; this may not be possible (i.e., precision may be lost by dropping variables) in some situations.

When there is strong interaction involving a certain extraneous variable, the assessment of confounding for that extraneous variable becomes irrelevant. Moreover, in such a situation, the assessment of confounding involving other extraneous variables, though possible, is quite complex and extremely subjective. Consequently, the assessment of confounding is not usually recommended when important interaction effects have been identified.

Problems

1. Consider the numerical examples given in Section 8.8 of Chapter 8, involving assessment of the relationship of the independent variables HGT, AGE, and $(AGE)^2$ to the dependent variable WGT. Suppose that HGT is the independent variable of primary concern, so interest lies in evaluating the relationship of HGT to WGT, controlling for the possible confounding effects of AGE and $(AGE)^2$.
 - a. Assuming that no interaction of any kind exists, state an appropriate regression model to use as the baseline (i.e., standard) for decisions about confounding.
 - b. Using an appropriate regression coefficient given in part (a) as your measure of association, determine whether confounding exists due to AGE and/or $(AGE)^2$.
 - c. Can $(AGE)^2$ be dropped from your initial model in part (a) because it is not needed to control adequately for confounding? Explain your answer (using a regression coefficient as your measure of association).
 - d. Should $(AGE)^2$ be retained in the final model for the sake of precision? Explain.
 - e. In light of both confounding and precision, what should be your final model? Why?
 - f. How would you modify your initial model in part (a) to allow for assessing interactions?
 - g. Regarding your answer to part (f), how would you test for interaction?
2. Consider the following computer results, which describe regression analyses involving two independent variables X_1 and X_2 and a dependent variable Y . Assume that your

Edited SAS Output (PROC REG) for Problem 2

Regression of Y on X1 and X2

CORRELATION			
Variable	X1	X2	Y
X1	1.0000	0.5000	0.6527
X2	0.5000	1.0000	0.9494
Y	0.6527	0.9494	1.0000

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	268.00000	134.00000	108.88	<.0001
Error	13	16.00000	1.23077		
Corrected Total	15	284.00000			

[Portion of output omitted]

PARAMETER ESTIMATES							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1	5.00000	0.42366	11.80	<.0001		
X1	1	2.00000	0.64051	3.12	0.0081	0.42606	0.42857
X2	1	7.00000	0.64051	10.93	<.0001	0.90184	0.90184

Regression of Y on X1

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	121.00000	121.00000	10.39	0.0061
Error	14	163.00000	11.64286		
Corrected Total	15	284.00000			

[Portion of output omitted]

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	6.75000	1.20638	5.60	<.0001
X1	1	5.50000	1.70608	3.22	0.0061

goal is to assess the relationship of X_1 with Y , controlling for the possible confounding effects of X_2 .

- a. Using an appropriate regression coefficient as your measure of association, determine whether confounding exists. Explain.
 - b. Suppose that confounding was defined to require a comparison of crude versus adjusted (partial) correlation coefficients. What conclusion would you draw? Explain.
 - c. What is the moral of this example?
3. a–c. Consider the accompanying computer results, which describe regression analyses involving two independent variables X_1 and X_2 and a dependent variable Y (using a different data set from the one used in Problem 2). Answer the same questions as in Problem 2 for this new printout.
- d. What does this example illustrate about using a test of the hypothesis $H_0: \beta_2 = 0$ to assess confounding?

Edited SAS Output (PROC REG) for Problem 3

Regression of Y on X1 and X2

CORRELATION			
Variable	X1	X2	Y
X1	1.0000	0.0000	0.2649
X2	0.0000	1.0000	0.9272
Y	0.2649	0.9272	1.0000

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	106.00000	53.00000	33.13	0.0013
Error	5	8.00000	1.60000		
Corrected Total	7	114.00000			

[Portion of output omitted]

PARAMETER ESTIMATES							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1	5.00000	0.77460	6.45	0.0013		
X1	1	2.00000	0.89443	2.24	0.0756	0.07018	0.50000
X2	1	7.00000	0.89443	7.83	0.0005	0.92453	0.92453

(continued)

Regression of Y on X1

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8.00000	8.00000	0.45	0.5260
Error	6	106.00000	17.66667		
Corrected Total	7	114.00000			

[Portion of output omitted]

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	8.50000	2.10159	4.04	0.0068
X1	1	2.00000	2.97209	0.67	0.5260

4. A regression analysis of data on $n = 53$ males considered the following variables:

$Y = \text{SBPSL}$ (estimated slope based on the straight-line regression of an individual's blood pressure over time)

$X_1 = \text{SBP1}$ (initial blood pressure)

$X_2 = \text{RW}$ (relative weight)

$X_3 = X_1 X_2 = \text{SR}$ (product of SBP1 and RW)

The accompanying computer printout was obtained by using a standard stepwise regression program (SPSS). Using this output, complete the following exercises.

- a. Fill in the following ANOVA table for the fit of the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$.

Source	d.f.	SS	MS
Regression	$\begin{cases} X_1 \\ X_2 X_1 \\ X_3 X_1, X_2 \end{cases}$		
Residual			
Total	52		

- b. Test $H_0: \rho_{YX_2|X_1} = 0$.
- c. Test H_0 : "The addition of X_3 to the model, given that X_1 and X_2 are already in the model, is not significant."
- d. Test $H_0: \rho_{Y(X_2, X_3)|X_1} = 0$.
- e. Based on the tests in parts (b) through (d), what is the most appropriate regression model? Use $\alpha = .05$.
- f. Based on the information provided, can you assess whether X_1 is a confounder of the X_2-Y relationship? Explain.

Edited SPSS Output for Problem 4

DEPENDENT VARIABLE.. SBPSL		VARIABLE(S) ENTERED ON STEP NUMBER 1.. SBP 1		ANALYSIS OF VARIANCE		SUM OF SQUARES		MEAN SQUARE		F	
MULTIPLE R		0.45834		REGRESSION		1.		14.79083		13.56308	
R SQUARE		0.21007		RESIDUAL		51.		55.61661		1.09052	
<hr/>											
VARIABLES IN THE EQUATION		-----		VARIABLES NOT IN THE EQUATION		BETA		IN PARTIAL		TOLERANCE	
VARIABLE	B	BETA	STD ERROR B	F	VARIABLE	BETA	IN PARTIAL	MEAN SQUARE	TOLERANCE	F	
SBP 1	-0.04660	-0.45834	0.01265	13.563	RW	0.23166	0.26007	0.99553	0.99933	3.627	
(CONSTANT)	5.10797				SR	0.23074	0.25953	0.99933		3.611	
<hr/>											
VARIABLE(S) ENTERED ON STEP NUMBER 2 .. RW		ANALYSIS OF VARIANCE		SUM OF SQUARES		MEAN SQUARE		F			
MULTIPLE R	0.51332	REGRESSION	DF	18.55240	9.27620	8.94435					
R SQUARE	0.26350	RESIDUAL	2.								
STANDARD ERROR	1.01838		50.								
<hr/>											
VARIABLES IN THE EQUATION		-----		VARIABLES NOT IN THE EQUATION		BETA		IN PARTIAL		TOLERANCE	
VARIABLE	B	BETA	STD ERROR B	F	VARIABLE	BETA	IN PARTIAL	MEAN SQUARE	TOLERANCE	F	
SBP 1	-0.04817	-0.47382	0.01237	15.174	SR	0.04646	0.00450	0.00690	0.001		
RW	0.02252	0.23166	0.01182	3.627							
(CONSTANT)	5.38484										
<hr/>											
VARIABLE(S) ENTERED ON STEP NUMBER 3 .. SR		ANALYSIS OF VARIANCE		SUM OF SQUARES		MEAN SQUARE		F			
MULTIPLE R	0.51334	REGRESSION	DF	18.55345	6.18448	5.84409					
R SQUARE	0.26352	RESIDUAL	3.								
STANDARD ERROR	1.02871		49.								
<hr/>											
VARIABLES IN THE EQUATION		-----		VARIABLES NOT IN THE EQUATION		BETA		IN PARTIAL		TOLERANCE	
VARIABLE	B	BETA	STD ERROR B	F	VARIABLE	BETA	IN PARTIAL	MEAN SQUARE	TOLERANCE	F	
SBP 1	-0.04798	-0.47193	0.01391	11.899							
RW	0.01801	0.18527	0.14372	0.016							
SR	0.00004	0.04646	0.00122	0.001							
(CONSTANT)	5.36183										

From Nie et al., *Statistical Package for the Social Sciences*. Copyright© 1975 by McGraw-Hill, Inc. Used with permission of McGraw-Hill Book Company and Dr. Norman Nie, President, SPSS, Inc.

5. An experiment involved a quantitative analysis of factors found in high-density lipoprotein (HDL) in a sample of human blood serum. Three variables thought to be predictive of or associated with HDL measurement (Y) were the total cholesterol (X_1)

Dataset for Problem 5

Y	X_1	X_2	X_3	Y	X_1	X_2	X_3
47	287	111	0	57	192	115	1
38	236	135	0	42	349	408	1
47	255	98	0	54	263	103	1
39	135	63	0	60	223	102	1
44	121	46	0	33	316	274	0
64	171	103	0	55	288	130	0
58	260	227	0	36	256	149	0
49	237	157	0	36	318	180	0
55	261	266	0	42	270	134	0
52	397	167	0	41	262	154	0
49	295	164	0	42	264	86	0
47	261	119	1	39	325	148	0
40	258	145	1	27	388	191	0
42	280	247	1	31	260	123	0
63	339	168	1	39	284	135	0
40	161	68	1	56	326	236	1
59	324	92	1	40	248	92	1
56	171	56	1	58	285	153	1
76	265	240	1	43	361	126	1
67	280	306	1	40	248	226	1
57	248	93	1	46	280	176	1

Edited SAS Output (PROC GLM) for Problem 5

Regression of Y on X_1

Source	DF	Sum of Squares	F Value	Pr > F
Model	1	46.235557	0.40	0.5282
Error	40	4567.383490		
Corrected Total	41	4613.619048		

R-Square	Coeff Var	y Mean
0.010022	22.37289	47.76190

Source	DF	Type I SS	F Value	Pr > F
X1	1	46.23555746	0.40	0.5282

Source	DF	Type III SS	F Value	Pr > F
X1	1	46.23555746	0.40	0.5282

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	52.47018272	7.58057266	6.92	<.0001
X1	-0.01758070	0.02762815	-0.64	0.5282

(continued)

Regression of Y on X2

Source	DF	Sum of Squares	F Value	Pr > F
Model	1	21.339709	0.19	0.6687
Error	40	4592.279339		
Corrected Total	41	4613.619048		

R-Square	Coeff Var	y Mean
0.004625	22.43378	47.76190

Source	DF	Type I SS	F Value	Pr > F
X2	1	21.33970871	0.19	0.6687

Source	DF	Type III SS	F Value	Pr > F
X2	1	21.33970871	0.19	0.6687

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	46.24519280	3.88711502	11.90	<.0001
X2	0.00978223	0.02268966	0.43	0.6687

Regression of Y on X3

Source	DF	Sum of Squares	F Value	Pr > F
Model	1	735.205411	7.58	0.0088
Error	40	3878.413636		
Corrected Total	41	4613.619048		

R-Square	Coeff Var	y Mean
0.159355	20.61652	47.76190

Source	DF	Type I SS	F Value	Pr > F
X3	1	735.2054113	7.58	0.0088

Source	DF	Type III SS	F Value	Pr > F
X3	1	735.2054113	7.58	0.0088

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	43.77272727	2.09935424	20.85	<.0001
X3	8.37727273	3.04225332	2.75	0.0088

Regression of Y on X1 and X2

Source	DF	Sum of Squares	F Value	Pr > F
Model	2	135.382126	0.59	0.5595
Error	39	4478.236922		
Corrected Total	41	4613.619048		

(continued)

R-Square	Coeff Var	y Mean
0.029344	22.43570	47.76190

Source	DF	Type I SS	F Value	Pr > F
X1	1	46.23555746	0.40	0.5294
X2	1	89.14656809	0.78	0.3837

Source	DF	Type III SS	F Value	Pr > F
X1	1	114.0424168	0.99	0.3251
X2	1	89.1465681	0.78	0.3837

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	52.76400680	7.60916422	6.93	<.0001
X1	-0.03216055	0.03227093	-1.00	0.3251
X2	0.02328833	0.02643061	0.88	0.3837

Regression of Y on X1 and X3

Source	DF	Sum of Squares	F Value	Pr > F
Model	2	783.169069	3.99	0.0266
Error	39	3830.449979		
Corrected Total	41	4613.619048		

R-Square	Coeff Var	y Mean
0.169752	20.74966	47.76190

Source	DF	Type I SS	F Value	Pr > F
X1	1	46.2355575	0.47	0.4967
X3	1	736.9335114	7.50	0.0092

Source	DF	Type III SS	F Value	Pr > F
X1	1	47.9636576	0.49	0.4888
X3	1	736.9335114	7.50	0.0092

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	48.56350948	7.17377750	6.77	<.0001
X1	-0.01790642	0.02562390	-0.70	0.4888
X3	8.38720265	3.06193225	2.74	0.0092

Regression of Y on X2 and X3

Source	DF	Sum of Squares	F Value	Pr > F
Model	2	737.806891	3.71	0.0334
Error	39	3875.812157		
Corrected Total	41	4613.619048		

R-Square	Coeff Var	y Mean
0.159919	20.87216	47.76190

(continued)

Source	DF	Type I SS	F Value	Pr > F
X2	1	21.3397087	0.21	0.6457
X3	1	716.4671821	7.21	0.0106

Source	DF	Type III SS	F Value	Pr > F
X2	1	2.6014795	0.03	0.8723
X3	1	716.4671821	7.21	0.0106

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	43.26641927	3.78286563	11.44	<.0001
X2	0.00343683	0.02124209	0.16	0.8723
X3	8.32148668	3.09921587	2.69	0.0106

Regression of Y on X1, X2, X3, X1X3, and X2X3

Source	DF	Sum of Squares	F Value	Pr > F
Model	5	894.567366	1.73	0.1523
Error	36	3719.051681		
Corrected Total	41	4613.619048		

R-Square	Coeff Var	y Mean
0.193897	21.28057	47.76190

Source	DF	Type I SS	F Value	Pr > F
X1	1	46.2355575	0.45	0.5078
X2	1	89.1465681	0.86	0.3591
X3	1	684.3651973	6.62	0.0143
X1X3	1	48.6904324	0.47	0.4968
X2X3	1	26.1296111	0.25	0.6181

Source	DF	Type III SS	F Value	Pr > F
X1	1	154.2311377	1.49	0.2297
X2	1	47.4875648	0.46	0.5021
X3	1	1.9962507	0.02	0.8902
X1X3	1	74.2675600	0.72	0.4021
X2X3	1	26.1296111	0.25	0.6181

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	52.33465853	9.25548949	5.65	<.0001
X1	-0.04955819	0.04055966	-1.22	0.2297
X2	0.03188430	0.04702749	0.68	0.5021
X3	-2.07933755	14.95830269	-0.14	0.8902
X1X3	0.05438698	0.06414462	0.85	0.4021
X2X3	-0.02821521	0.05610244	-0.50	0.6181

and total triglyceride (X_2) concentrations in the sample, plus the presence or absence of a certain sticky component called sinking pre-beta, or SPB (X_3), which was coded as 0 if absent and 1 if present. The data obtained are shown in the table on page 247 and the accompanying computer results.

- a. Test whether X_1 , X_2 , or X_3 alone significantly helps in predicting Y .
 - b. Test whether X_1 , X_2 , and X_3 taken together significantly help to predict Y .
 - c. Test whether the true coefficients of the product terms X_1X_3 and X_2X_3 are simultaneously zero in the model containing X_1 , X_2 , and X_3 plus these product terms. State the null hypothesis in terms of a multiple partial correlation coefficient. If this test is not rejected, what can you conclude about the relationship of Y to X_1 and X_2 when X_3 equals 1, as compared with when X_3 equals 0?
 - d. Using $\alpha = .05$, test whether X_3 is associated with Y after the combined contribution of X_1 and X_2 is taken into account. State the appropriate null hypothesis in terms of a partial correlation coefficient. What does your result, together with your answer to part (c), tell you about the relationship of Y with X_1 and X_2 when SPB is present as compared with when it is absent?
 - e. How would you determine whether X_1 , X_2 , or both X_1 and X_2 need to be retained in the model to control for confounding and possibly to enhance precision? Assume that no interaction occurs and that the study variable of interest is X_3 .
 - f. Based on the information provided, can confounding of X_1 and/or X_2 be assessed in evaluating the relationship of X_3 to Y ? Explain.
6. Use the computer output from Problem 10 in Chapter 8 and from Problem 16 in Chapter 5 to answer the following questions. (Assume that no interaction occurs between house size [X_1] and number of rooms [X_2].)
 - a. Does the number of rooms (X_2) confound the relationship between house price (Y) and house size (X_1)?
 - b. Should the number of rooms be included in a model that already contains house size, based on considerations of precision?
 - c. In light of your answers to parts (a) and (b), should the final model include both predictors? Explain.
 7. Use the output from Problem 11 in Chapter 8 and the output given here to answer the following questions. (Assume that no interaction occurs between TV advertising expenditure [X_1] and print advertising expenditure [X_2].)
 - a. Does the print advertising expenditure confound the relationship between sales (Y) and TV advertising expenditure (X_1)?
 - b. Should print advertising expenditure be included in a model that already contains TV advertising expenditure, based on considerations of precision?
 - c. In light of your answers to parts (a) and (b), should the final model include both types of advertising expenditures as predictors? Explain.

Edited SAS Output (PROC GLM) for Problem 7

Regression of Y on X1

[Portion of output omitted]

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	2.002227171	0.31569714	6.34	0.0032
X1	1.224944321	0.10088866	12.14	0.0003

8. Use the computer output from Problem 12 in Chapter 8 and the output given here to answer the following questions.
- State the model that relates change in refraction (Y) to baseline refraction (X_1), baseline curvature (X_2), and the interaction of X_1 and X_2 . Is the partial F test for the interaction significant?
 - Is it appropriate to assess confounding given your answer to part (a)? Explain.
 - If your answer to part (b) is yes, does X_2 confound the relationship between Y and X_1 ?
 - If your answer to part (b) is yes, does X_1 confound the relationship between Y and X_2 ?
 - In light of your answers to parts (a) through (d), and on considerations of precision, which predictor(s) should be included in the model?

Edited SAS Output (PROC GLM) for Problem 8

Regression of Y on X1, X2, and X1X2

[Portion of output omitted]

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	25.85215663	16.33395716	1.58	0.1199
X1	2.69854313	3.41991591	0.79	0.4339
X2	-0.52636869	0.37006537	-1.42	0.1613
X1X2	-0.06778472	0.07750534	-0.87	0.3861

Regression of Y on X1

[Portion of output omitted]

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	2.619102900	0.41260537	6.35	0.0001
X1	-0.298729328	0.09463300	-3.16	0.0027

(continued)

Regression of Y on X2

[Portion of output omitted]

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	14.14611679	5.52569396	2.56	0.0135
X2	-0.23446227	0.12544713	-1.87	0.0674

Regression of Y on X1 and X2

Source	DF	Sum of Squares	F Value	Pr > F
Model	2	17.53454762	7.02	0.0021
Error	20	62.41955709		
Corrected Total	52	79.95410472		

R-Square	Coeff Var	y Mean
0.219308	29.22453	3.82320755

[Portion of output omitted]

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	12.29334748	5.13074580	2.40	0.0204
X1	-0.29135396	0.09241113	-3.15	0.0027
X2	-0.21905406	0.11581741	-1.89	0.0644

9. Refer to the data in Problem 19 of Chapter 5. The relationship between the rate of owner occupancy of housing units (OWNEROCC) and the median monthly ownership costs (OWNCOST) was studied in that problem in connection with a random sample of data from 26 Metropolitan Statistical Areas (MSAs). The results of the regression of OWNEROCC on OWCOST and a third variable, the median household income (INCOME), are presented next.

Use the output given here and in Problem 19 of Chapter 5 to answer the following questions.

- Test whether OWCOST and INCOME, taken together, significantly help to predict the rate of owner occupancy.
- Test whether OWCOST is associated with the rate of owner occupancy after taking into account the contribution of INCOME.
- Test whether INCOME is associated with the rate of owner occupancy after taking into account the contribution of OWCOST.
- Should INCOME be included in the model to control for confounding? Explain your answer, including discussion of any assumptions you made in reaching your answer.

Edited SAS Output (PROC GLM) for Problem 9

Regression of OWNROCC on OWNCOST and INCOME

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	214.4833906	107.2416953	6.38	0.0062
Error	23	386.4781479	16.8033977		
Corrected Total	25	600.9615385			

R-Square	Coeff Var	Root MSE	Y Mean
0.356900	6.214523	4.099195	69.96154

Source	DF	Type I SS	Mean Square	F Value	Pr > F
OWNCOST	1	132.6203242	132.603242	7.89	0.0100
INCOME	1	81.8630664	81.8630664	4.87	0.0375

Source	DF	Type III SS	Mean Square	F Value	Pr > F
OWNCOST	1	192.6978217	192.6978217	11.47	0.0025
INCOME	1	81.8630664	81.8630664	4.87	0.0375

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	70.03201034	4.55666725	15.37	0.0001
OWNCOST	-0.03334031	0.00984532	-3.39	0.0025
INCOME	0.00064489	0.00029217	2.21	0.0375

10. Refer to the *Business Week* magazine data in Problem 13 of Chapter 8. The relationships among the yield (Y), 1989 rank (X_2), and P-E ratio (X_3) were studied in that problem using a random sample of data from the magazine's compilation of information on the top 1,000 companies. The results of the regression of yield on 1989 rank, P-E ratio, and 1990 rank (X_1) are presented next. Use this output and the output from Problem 13 in Chapter 8 to answer the following questions.
- Test whether 1989 rank, P-E ratio, and 1990 rank, taken together, significantly help to predict the yield.
 - Test whether 1990 rank is associated with the yield after taking into account the contribution of 1989 rank and P-E ratio.
 - Should 1990 rank be included in the model to control for confounding? Explain your answer, including discussion of any assumptions you made in reaching your answer.

Edited SAS Output (PROC GLM) for Problem 10

Regression of Y on X1, X2, and X3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	29.97725832	9.99241944	8.83	0.0011
Error	16	18.10303668	1.13143979		
Corrected Total	19	48.08029500			

R-Square	Coeff Var	Root MSE	Y Mean
0.623483	42.79588	1.063692	2.485500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	1.82894871	1.82894871	1.62	0.2218
X2	1	0.00159198	0.00159198	0.00	0.9705
X3	1	28.14671763	28.14671763	24.88	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	3.39453697	3.39453697	3.00	0.1025
X2	1	1.35441258	1.35441258	1.20	0.2901
X3	1	28.14671763	28.14671763	24.88	0.0001

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	7.421187723	0.98839344	7.51	<.0001
X1	-0.013351684	0.00770835	-1.73	0.1025
X2	0.007620880	0.00696539	1.09	0.2901
X3	-0.261846279	0.05249866	-4.99	0.0001

11. Let us return once more to the body-mass index (BMI) example for the BRFSS data. We now extend the models involving the explanatory factors of drinking frequency, age, and sleep quality in Chapters 8–9 to consider a model that includes all first-order product terms involving these three factors. Selected output for both models is provided below.
- Perform and interpret a hypothesis test that evaluates the presence of interaction in this model.
 - If the main interest is to understand the relationship between drinking frequency and BMI, based on results obtained in Section 5.12, is controlling for age and sleep quality necessary?
 - A researcher believes that sleep quality should be controlled for given the results of the hypothesis test shown below. Do you agree with this assessment? What would you say to this researcher?

Edited SAS Output (PROC GLM) for Problem 11

Regression of BMI on DRINK_DAYS, AGE, POOR_SLEEP_DAYS, with 2-way interactions

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1409.32009	234.88668	6.81	<.0001
Error	1042	35945.73599	34.49687		
Corrected Total	1048	37355.05608			

[Portion of output omitted]

Source	DF	Type I SS	Mean Square	F Value	Pr > F
drink_days	1	1073.403915	1073.403915	31.12	<.0001
AGE	1	11.595517	11.595517	0.34	0.5622
poor_sleep_days	1	204.794319	204.794319	5.94	0.0150
drink_days*AGE	1	4.917524	4.917524	0.14	0.7058
drink_days*poor_sleep	1	8.581905	8.581905	0.25	0.6180
AGE*poor_sleep_days	1	106.026909	106.026909	3.07	0.0799

Regression of BMI on DRINK_DAYS, AGE, POOR_SLEEP_DAYS

[Portion of output omitted]

Source	DF	Type I SS	Mean Square	F Value	Pr > F
drink_days	1	1073.403915	1073.403915	31.10	<.0001
AGE	1	11.595517	11.595517	0.34	0.5623
poor_sleep_days	1	204.794319	204.794319	5.93	0.0150

[Portion of output omitted]

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	26.66350677	0.70642133	37.74	<.0001
drink_days	-0.14945246	0.02686344	-5.56	<.0001
AGE	0.01404807	0.01265963	1.11	0.2674
poor_sleep_days	0.04593183	0.01885564	2.44	0.0150

References

- Kleinbaum, D. G. 2002. *ActivEpi*, New York and Berlin: Springer.
- Kleinbaum, D. G., and Klein, M. 2002. *Logistic Regression: A Self-Learning Text*, Second Edition. New York and Berlin: Springer.
- Kleinbaum, D. G.; Kupper, L. L.; and Morgenstern, H. 1982. *Epidemiologic Research*. Belmont, Calif.: Lifetime Learning Publications.
- Nie, N., et al. 1975. *Statistical Package for the Social Sciences*. New York: McGraw-Hill.
- Rothman, K. J.; Greenland, S.; and Lash, T. L. 2008. *Modern Epidemiology*. Philadelphia: Lippincott, Williams, & Wilkins.

12

Dummy Variables in Regression

12.1 Preview

To this point, we have mainly considered only continuous variables as predictors, but the methods of regression analysis can be generalized to include categorical predictors as well. The generalization is based entirely on the use of dummy variables, the central idea of this chapter.

By using dummy variables, we can broaden the application of regression analysis. In particular, dummy variables allow us to employ regression analysis to produce the same information obtained by such seemingly distinct analytical procedures as analysis of covariance (Chapter 13) and analysis of variance (Chapters 17 through 20).

In this chapter, we focus on one important application of dummy variables: the comparison of several regression equations via the use of a single multiple regression model. We also describe an alternative method that can be used if only two equations are being compared.

12.2 Definitions

A *dummy*, or *indicator*, *variable* is any variable in a regression equation that takes on a finite number of values so that different categories of a nominal variable can be identified. The term *dummy* reflects the fact that the values taken on by such variables (usually values like 0, 1, and -1) do not indicate meaningful measurements but rather the categories of interest.

Examples of dummy variables include the following:

$$X_1 = \begin{cases} 1 & \text{if treatment A is used} \\ 0 & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if subject is female} \\ -1 & \text{if subject is male} \end{cases}$$

$$Z_1 = \begin{cases} 1 & \text{if residence is in western United States} \\ 0 & \text{if residence is in central United States} \\ -1 & \text{if residence is in eastern United States} \end{cases}$$

$$Z_2 = \begin{cases} 0 & \text{if residence is in western United States} \\ 1 & \text{if residence is in central United States} \\ -1 & \text{if residence is in eastern United States} \end{cases}$$

The variable X_1 indicates a nominal variable describing the category “treatment group” (either treatment A or not treatment A); the variable X_2 indexes the levels of the nominal variable “sex”; and variables Z_1 and Z_2 work in tandem to describe the nominal variable “geographical residence.” In the last case, the three categories of geographical residence are described by the following combination of the two variables Z_1 and Z_2 :

Residence in western United States: $Z_1 = 1, Z_2 = 0$

Residence in central United States: $Z_1 = 0, Z_2 = 1$

Residence in eastern United States: $Z_1 = -1, Z_2 = -1$

12.3 Rule for Defining Dummy Variables

The following simple rule should always be applied to avoid collinearity (see Chapter 14) in defining a dummy variable for regression analysis: *If the nominal independent variable of interest has k categories, then exactly $k - 1$ dummy variables must be defined to index these categories, provided that the regression model contains a constant term (i.e., an intercept β_0). If the regression model does not contain an intercept, then k dummy variables are needed to index the k categories of interest.* For example, given $k = 3$ categories, the number of dummy variables should be $k - 1 = 2$ for a model containing an intercept. If an intercept is not included in an overall regression model designed to compare several regression equations, however, the dummy variables can be defined so that each of the regression equations derived from the overall model has its own intercept. Thus, using an intercept in the overall model generally depends on how the investigator prefers to code the dummy variables.

Applying this rule raises several significant points:

1. If an intercept is used in the regression equation, proper definition of the $k - 1$ dummy variables automatically indexes all k categories.
2. If k dummy variables are used to describe a nominal variable with k categories in a model containing an intercept, all the coefficients in the model cannot be uniquely estimated (i.e., collinearity is present).
3. The $k - 1$ dummy variables for indexing the k categories of a given nominal variable can be properly defined in many different ways. For example, two equivalent

ways to describe the nominal variable “geographical residence” (represented earlier by Z_1 and Z_2) are

$$Z_1^* = \begin{cases} 1 & \text{if residence is in western United States} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_2^* = \begin{cases} 1 & \text{if residence is in central United States} \\ 0 & \text{otherwise} \end{cases}$$

and

$$Z'_1 = \begin{cases} 1 & \text{if residence is in western United States} \\ 0 & \text{otherwise} \end{cases}$$

$$Z'_2 = \begin{cases} 1 & \text{if residence is in eastern United States} \\ 0 & \text{otherwise} \end{cases}$$

Among the coding schemes available for regression, we recommend the method often referred to as *reference cell coding*, which uses $k - 1$ dummy variables as suggested earlier. Each variable takes on only values of 1 and 0, and each variable indicates group membership (1 for a specific group, 0 otherwise). Some computer programs use 1 and -1 for coding. Since the choice of coding scheme affects analysis and interpretation, it is important to specify, or otherwise be aware of, which coding method is being used.

We now illustrate how to use dummy variables to compare two or more regression models. We begin by considering two straight-line models, and then we extend the discussion to comparisons of more than two multiple regression models.

12.4 Comparing Two Straight-line Regression Equations: An Example

In Chapter 11, the age–systolic blood pressure example was used to assess whether there is a significant relationship between blood pressure and age. Here we use the data to investigate the common observation that males tend to have higher blood pressure than females of similar age. To do this, we must compare the straight-line regression of systolic blood pressure versus age for females against the corresponding regression for males. The entire data set for this example was presented in Table 11.1. Table 12.1 also provides the information needed to compare the two fitted straight lines. For each data set, this information consists of the sample size (n), the intercept ($\hat{\beta}_0$), the slope ($\hat{\beta}_1$), the sample mean \bar{X} of the X 's, the sample mean \bar{Y} of the Y 's, the sample variance S_X^2 of the X 's, and the residual mean-square error ($S_{Y|X}^2$). To distinguish between male and female data, we have used the subscripts M and F , respectively. Thus, n_M , $\hat{\beta}_{0M}$, $\hat{\beta}_{1M}$, and $S_{Y|X_M}^2$ denote the sample size, intercept, slope, and mean-square error for the

TABLE 12.1 Comparing results for males and females of straight-line regression of systolic blood pressure on age

Group	n	$\hat{\beta}_0$	$\hat{\beta}_1$	\bar{X}	\bar{Y}	s_x^2	$s_{Y X}^2$
Males	40	110.04	0.96	46.93	155.15	221.15	71.90
Females	29	97.08	0.95	45.07	139.86	242.14	91.46

© Cengage Learning

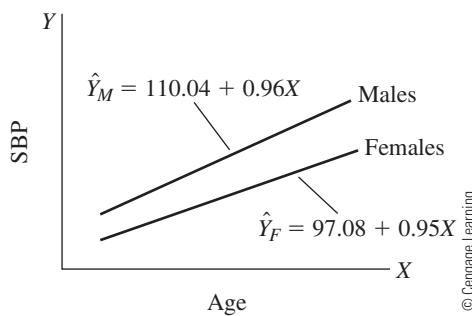
male data, whereas n_F , $\hat{\beta}_{0F}$, $\hat{\beta}_{1F}$, and $S_{Y|X_F}^2$ denote the corresponding information for the female data.

The least-squares lines are then given as follows:¹

$$\text{Males: } \hat{Y}_M = 110.04 + 0.96X$$

$$\text{Females: } \hat{Y}_F = 97.08 + 0.95X$$

These lines are sketched in Figure 12.1. It can be seen from the figure that the male line lies completely above the female line. This fact alone supports the contention that males have higher blood pressure than females over the age range being considered. Nevertheless, it is necessary to explore statistically whether the observed differences between the regression lines could have occurred by chance. In other words, to be statistically precise when comparing two regression lines, we must consider the sampling variability of the data by using statistical test(s) and/or confidence interval(s). The sections that follow describe a number of statistical procedures for dealing with this comparison problem.



© Cengage Learning

FIGURE 12.1 Comparison by sex of straight-line regressions of systolic blood pressure on age

¹ It can be shown that the straight-line model for males, like that for females, is appropriate based on a lack-of-fit test (see Chapter 15).

12.5 Questions for Comparing Two Straight Lines

There are three basic questions to consider when comparing two straight-line regression equations:

1. Are the two slopes the same or different (regardless of whether the intercepts are different)?²
2. Are the two intercepts the same or different (regardless of whether the slopes are different)?
3. Are the two lines coincident (that is, the same), or do they differ in slope and/or intercept?

Situations pertaining to these three questions are illustrated in Figure 12.2.

For our particular age–systolic blood pressure (SBP) example, concluding that the lines are parallel (Figure 12.2(a)) is equivalent to finding that one sex has a consistently higher SBP than the other at all ages but that the rate of change with respect to age is the same for both sexes. If we conclude that the two lines have a common intercept but different slopes

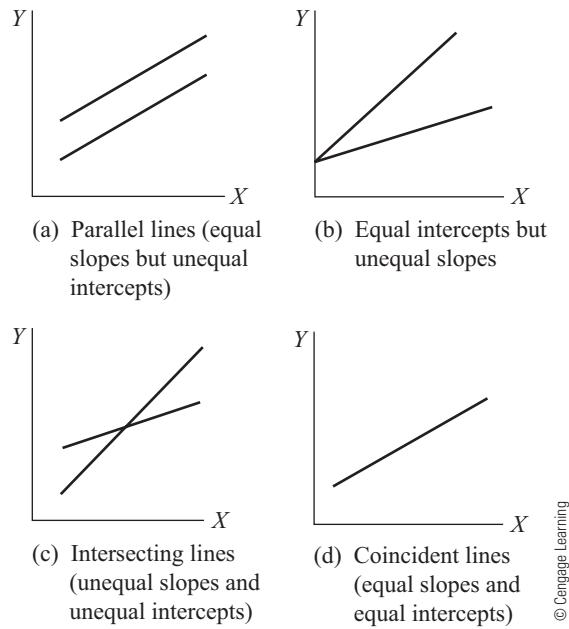


FIGURE 12.2 Possible conclusions from comparing two straight-line regressions

© Cengage Learning

²If the two slopes are not different, we say that the two lines are *parallel*.

(Figure 12.2(b)), we find that the two sexes begin at an early age with the same average SBP but that average SBP changes with respect to age at different rates for each sex. If the two lines have different slopes and different intercepts³ (Figure 12.2(c)), it means that the relationship between age and mean SBP differs for the two sexes with regard to both the origins and the rates of change. Furthermore, if the lines intersect in the range of X -values of interest, this indicates that at early ages one sex has a higher average systolic blood pressure than the other, but at later ages the other sex does.

12.6 Methods of Comparing Two Straight Lines

There are two general approaches to answering the earlier three questions related to comparing two straight lines.

Method I

Treat the male and female data separately by fitting the two separate regression equations

$$Y_M = \beta_{0M} + \beta_{1M}X + E_M \quad (12.1)$$

and

$$Y_F = \beta_{0F} + \beta_{1F}X + E_F \quad (12.2)$$

and then conduct appropriate two-sample t tests.

Method II

Define the dummy variable Z to be 0 if the subject is male and 1 if female. Thus, for the n_M observations on males, $Z = 0$; and for the n_F observations on females, $Z = 1$. Our data will then be of the form

Males: $(X_{1M}, Y_{1M}, 0), (X_{2M}, Y_{2M}, 0), \dots, (X_{n_M M}, Y_{n_M M}, 0)$

Females: $(X_{1F}, Y_{1F}, 1), (X_{2F}, Y_{2F}, 1), \dots, (X_{n_F F}, Y_{n_F F}, 1)$

Then, for the combined data above, the single multiple regression model

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + E \quad (12.3)$$

³ It is possible for two lines to have unequal slopes and unequal intercepts and yet not intersect within the range of X -values of interest. This is illustrated by our example in Figure 12.1.

yields the following two models for the two values of Z :

$$\begin{cases} Z = 0: Y_M = \beta_0 + \beta_1 X + E \\ Z = 1: Y_F = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X + E \end{cases}$$

This allows us to write the regression coefficients for the separate models for method I in terms of the coefficients of model (12.3) as follows:

$$\beta_{0M} = \beta_0, \beta_{0F} = \beta_0 + \beta_2, \beta_{1M} = \beta_1, \beta_{1F} = \beta_1 + \beta_3$$

Thus, *model (12.3) incorporates the two separate regression equations within a single model* and allows for different slopes (β_1 for males and $\beta_1 + \beta_3$ for females) and different intercepts (β_0 for males and $\beta_0 + \beta_2$ for females). We now consider the details involved in making statistical inferences for these two methods.

12.7 Method I: Using Separate Regression Fits to Compare Two Straight Lines

12.7.1 Testing for Parallelism

From (12.1) and (12.2), we can conclude that the appropriate null hypothesis for comparing the slopes (i.e., for conducting a test of parallelism) is given by

$$H_0: \beta_{1M} = \beta_{1F}$$

When the null hypothesis $H_0: \beta_{1M} = \beta_{1F}$ is true, the two regression lines simplify to $Y_M = \beta_{0M} + \beta_1 X + E_M$ for males and $Y_F = \beta_{0F} + \beta_1 X + E_F$ for females, where $\beta_1 (= \beta_{1M} = \beta_{1F})$ is the common slope. An estimate of this common slope β_1 is given by the following formula, which is a weighted average of the two separate slope estimates:

$$\hat{\beta}_1 = \frac{(n_M - 1)S_{X_M}^2 \hat{\beta}_{1M} + (n_F - 1)S_{X_F}^2 \hat{\beta}_{1F}}{(n_M - 1)S_{X_M}^2 + (n_F - 1)S_{X_F}^2}$$

Note that $\hat{\beta}_1$ equals the slope computed by fitting a straight line to the pooled data.

Any of the following three alternative hypotheses can be used:

$$H_A: \begin{cases} \beta_{1M} > \beta_{1F} & \text{(one-sided)} \\ \beta_{1M} < \beta_{1F} & \text{(one-sided)} \\ \beta_{1M} \neq \beta_{1F} & \text{(two-sided)} \end{cases}$$

The test statistic for evaluating parallelism is then given by

$$T = \frac{\hat{\beta}_{1M} - \hat{\beta}_{1F}}{S_{(\hat{\beta}_{1M} - \hat{\beta}_{1F})}} \quad (12.4)$$

where

$\hat{\beta}_{1M}$ = Least-squares estimate of the slope β_{1M} , using the n_M observations (on males)

$\hat{\beta}_{1F}$ = Least-squares estimate of the slope β_{1F} , using the n_F observations (on females)

$S_{(\hat{\beta}_{1M} - \hat{\beta}_{1F})}$ = Estimate of the standard error of the estimated difference between slopes
 $(\hat{\beta}_{1M} - \hat{\beta}_{1F})$

This standard error involves pooling and summing the estimated variances of the slopes of the fitted regression lines.⁴ It is equal to the square root of the following variance:

$$S_{(\hat{\beta}_{1M} - \hat{\beta}_{1F})}^2 = S_{P, Y|X}^2 \left[\frac{1}{(n_M - 1)S_{X_M}^2} + \frac{1}{(n_F - 1)S_{X_F}^2} \right] \quad (12.5)$$

where

$$S_{P, Y|X}^2 = \frac{(n_M - 2)S_{Y|X_M}^2 + (n_F - 2)S_{Y|X_F}^2}{n_M + n_F - 4} \quad (12.6)$$

is a pooled estimate of σ^2 based on combining residual mean-square errors for males and females and where

$S_{Y|X_M}^2$ = Residual mean-square error for the male data

$S_{Y|X_F}^2$ = Residual mean-square error for the female data

$S_{X_M}^2$ = Variance of the X 's for the male data

$S_{X_F}^2$ = Variance of the X 's for the female data

The test statistic given by (12.4) will, under the usual regression assumptions, be distributed as a Student's t with $n_M + n_F - 4$ degrees of freedom when H_0 is true. We then have the following critical regions for different hypotheses and significance level α :

$$\begin{cases} T \geq t_{n_M + n_F - 4, 1 - \alpha} & \text{for } H_A: \beta_{1M} > \beta_{1F} \\ T \leq -t_{n_M + n_F - 4, 1 - \alpha} & \text{for } H_A: \beta_{1M} < \beta_{1F} \\ |T| > t_{n_M + n_F - 4, 1 - \alpha/2} & \text{for } H_A: \beta_{1M} \neq \beta_{1F} \end{cases}$$

The associated $100(1 - \alpha)\%$ confidence interval for $(\beta_{1M} - \beta_{1F})$ is of the form

$$(\hat{\beta}_{1M} - \hat{\beta}_{1F}) \pm t_{n_M + n_F - 4, 1 - \alpha/2} S_{(\hat{\beta}_{1M} - \hat{\beta}_{1F})}$$

⁴ This pooled estimate of σ^2 is valid only if $H_0: \sigma_M^2 = \sigma_F^2$ (i.e., variance homogeneity) holds. The F statistic

$$F_{df_{MSE_M}, df_{MSE_F}} = MSE_M / MSE_F = S_{Y|X_M}^2 / S_{Y|X_F}^2$$

can be used to test this H_0 . If H_0 is rejected, one should test for equality of slopes (or intercepts) using a large-sample Z test where the estimated variance of each slope (or intercept) estimator is determined separately using the appropriate MSE value for that particular fitted straight line.

■ **Example 12.1** Using the data given in Table 12.1, we can compute the estimates $S_{P,Y|X}^2$, $S_{(\hat{\beta}_{1M} - \hat{\beta}_{1F})}^2$, $S_{\hat{\beta}_{1M}}^2$, and $S_{\hat{\beta}_{1F}}^2$ as follows:

$$\begin{aligned} S_{P,Y|X}^2 &= \frac{(n_M - 2)S_{Y|X_M}^2 + (n_F - 2)S_{Y|X_F}^2}{n_M + n_F - 4} = \frac{38(71.90) + 27(91.46)}{40 + 29 - 4} \\ &= \frac{5,201.62}{65} = 80.02 \\ S_{(\hat{\beta}_{1M} - \hat{\beta}_{1F})}^2 &= S_{P,Y|X}^2 \left[\frac{1}{(n_M - 1)S_{X_M}^2} + \frac{1}{(n_F - 1)S_{X_F}^2} \right] \\ &= 80.02 \left[\frac{1}{39(221.15)} + \frac{1}{28(242.14)} \right] \\ &= 0.021 \\ S_{\hat{\beta}_{1M}}^2 &= \frac{S_{Y|X_M}^2}{(n_M - 1)S_{X_M}^2} = \frac{71.90}{39(221.15)} = 0.0083 \\ S_{\hat{\beta}_{1F}}^2 &= \frac{S_{Y|X_F}^2}{(n_F - 1)S_{X_F}^2} = \frac{91.46}{28(242.14)} = 0.0135 \end{aligned}$$

The test statistic (12.4) is then computed as

$$T = \frac{\hat{\beta}_{1M} - \hat{\beta}_{1F}}{S_{(\hat{\beta}_{1M} - \hat{\beta}_{1F})}} = \frac{0.96 - 0.95}{\sqrt{0.021}} = \frac{0.01}{0.145} = 0.069$$

For this test statistic, the critical value for a two-sided test (i.e., $H_A: \beta_{1M} \neq \beta_{1F}$) with $\alpha = .05$ is given by

$$t_{65, 0.975} = 1.9964$$

Since $|T| = 0.069$ does not exceed 1.9964 (i.e., $P > .9$), we do not reject H_0 . Thus, we conclude that there is insufficient evidence to permit us to reject the hypothesis of parallelism (namely, that the lines for males and females have the same slope). ■

12.7.2 Comparing Two Intercepts

We now describe how to use separate regression fits to determine whether both straight lines have the same intercept, regardless of the two slopes. The null hypothesis in this case is given by

$$H_0: \beta_{0M} = \beta_{0F}$$

If $H_0: \beta_{0M} = \beta_{0F}$ is true, the two regression lines simplify to $Y_M = \beta_0 + \beta_{1M}X + E_M$ and $Y_F = \beta_0 + \beta_{1F}X + E_F$, where $\beta_0 (= \beta_{0M} = \beta_{0F})$ is the common intercept. An estimate of this common intercept β_0 is given by the equation

$$\hat{\beta}_0 = \frac{n_M \hat{\beta}_{0M} + n_F \hat{\beta}_{0F}}{n_M + n_F}$$

which is a weighted average of the two separate intercept estimates.

The test statistic in this case is given by

$$T = \frac{\hat{\beta}_{0M} - \hat{\beta}_{0F}}{S_{(\hat{\beta}_{0M} - \hat{\beta}_{0F})}} \quad (12.7)$$

where $\hat{\beta}_{0M}$ and $\hat{\beta}_{0F}$ are the intercept estimates for males and females, respectively, and where $S_{(\hat{\beta}_{0M} - \hat{\beta}_{0F})}^2$ estimates the variance of the estimated difference between the intercepts by means of the formula

$$S_{(\hat{\beta}_{0M} - \hat{\beta}_{0F})}^2 = S_{P,Y|X}^2 \left[\frac{1}{n_M} + \frac{1}{n_F} + \frac{\bar{X}_M^2}{(n_M - 1)S_{X_M}^2} + \frac{\bar{X}_F^2}{(n_F - 1)S_{X_F}^2} \right] \quad (12.8)$$

The statistic T given in (12.7) will have the t distribution with $n_M + n_F - 4$ degrees of freedom when $H_0: \beta_{0M} = \beta_{0F}$ is true and when $\sigma_M^2 = \sigma_F^2$. We, therefore, have the following critical regions for different hypotheses and significance level α :

$$\begin{cases} T \geq t_{n_M + n_F - 4, 1 - \alpha} & \text{for } H_A: \beta_{0M} > \beta_{0F} \\ T \leq -t_{n_M + n_F - 4, 1 - \alpha} & \text{for } H_A: \beta_{0M} < \beta_{0F} \\ |T| > t_{n_M + n_F - 4, 1 - \alpha/2} & \text{for } H_A: \beta_{0M} \neq \beta_{0F} \end{cases}$$

The associated $100(1 - \alpha)\%$ confidence interval for $\beta_{0M} - \beta_{0F}$ is

$$(\hat{\beta}_{0M} - \hat{\beta}_{0F}) \pm t_{n_M + n_F - 4, 1 - \alpha/2} S_{(\hat{\beta}_{0M} - \hat{\beta}_{0F})}$$

■ Example 12.2 For the data in Table 12.1 and the value of $S_{P,Y|X}^2$ obtained in Example 12.1, the estimates $S_{(\hat{\beta}_{0M} - \hat{\beta}_{0F})}^2$, $S_{\hat{\beta}_{0M}}^2$, and $S_{\hat{\beta}_{0F}}^2$ are computed as follows:

$$\begin{aligned} S_{(\hat{\beta}_{0M} - \hat{\beta}_{0F})}^2 &= S_{P,Y|X}^2 \left[\frac{1}{n_M} + \frac{1}{n_F} + \frac{\bar{X}_M^2}{(n_M - 1)S_{X_M}^2} + \frac{\bar{X}_F^2}{(n_F - 1)S_{X_F}^2} \right] \\ &= 80.02 \left[\frac{1}{40} + \frac{1}{29} + \frac{(46.93)^2}{39(221.15)} + \frac{(45.07)^2}{28(242.14)} \right] \\ &= 80.02(0.0250 + 0.0345 + 0.2554 + 0.2996) \\ &= 49.17 \end{aligned}$$

$$\begin{aligned} S_{\hat{\beta}_{0M}}^2 &= S_{Y|X_M}^2 \left[\frac{1}{n_M} + \frac{\bar{X}_M^2}{(n_M - 1)S_{X_M}^2} \right] = 71.90 \left[\frac{1}{40} + \frac{(46.93)^2}{39(221.15)} \right] \\ &= 71.90(0.0250 + 0.2554) = 20.16 \end{aligned}$$

$$\begin{aligned} S_{\hat{\beta}_{0F}}^2 &= S_{Y|X_F}^2 \left[\frac{1}{n_F} + \frac{\bar{X}_F^2}{(n_F - 1)S_{X_F}^2} \right] = 91.46 \left[\frac{1}{29} + \frac{(45.07)^2}{28(242.14)} \right] \\ &= 91.46(0.0345 + 0.2996) = 30.56 \end{aligned}$$

From these results, we compute the T statistic of (12.7) as

$$T = \frac{\hat{\beta}_{0M} - \hat{\beta}_{0F}}{S_{(\hat{\beta}_{0M} - \hat{\beta}_{0F})}} = \frac{110.04 - 97.08}{\sqrt{49.17}} = \frac{12.96}{7.01} = 1.85$$

For a two-sided test ($H_A: \beta_{0M} \neq \beta_{0F}$) with $\alpha = .05$, we find that $|T| = 1.85$ does not exceed $t_{65, 0.975} = 1.9964$ (i.e., $.05 < P < .10$). Thus, the null hypothesis of common intercepts is not rejected at $\alpha = .05$, but it is rejected at $\alpha = .10$. ■

12.7.3 Testing for Coincidence from Separate Straight-line Fits

Two straight lines are coincident if their slopes and their intercepts are equal. In considering the male–female regression equations given by (12.1) and (12.2), we may conclude that the null hypothesis of coincidence is, therefore, equivalent to testing $H_0: \beta_{0M} = \beta_{0F}$ and $\beta_{1M} = \beta_{1F}$ simultaneously. If so, the two regression models both reduce to the general form

$$Y = \beta_0 + \beta_1 X + E$$

where $\beta_0 (= \beta_{0M} = \beta_{0F})$ and $\beta_1 (= \beta_{1M} = \beta_{1F})$ are the common intercept and slope, respectively. The estimates of the common slope β_1 and common intercept β_0 are obtained by pooling all observations on males and females together and determining the usual least-squares slope and intercept estimates using the pooled data set.

A preferred way to test the null hypothesis of coincident lines is to employ a multiple regression model involving dummy variables (i.e., method II). Another, generally less efficient (e.g., not as powerful) procedure is often convenient when separate models are fit. In practice, this procedure frequently yields the same conclusion as is obtained from using dummy variables.

Using separate regression fits, we perform both the test of $H_0: \beta_{0M} = \beta_{0F}$ of equal intercepts and the test of $H_0: \beta_{1M} = \beta_{1F}$ of equal slopes. If either one or both of these null hypotheses are rejected, we can conclude that statistical evidence indicates that the two lines do not coincide. If neither is rejected, we must conclude that no evidence of noncoincidence exists in the data.

A valid criticism of this testing procedure, which calls into question its power, is that it involves two separate tests rather than a single test. This fact raises two difficulties:

1. The procedure does not precisely test for coincidence.
2. If α is the significance level of each separate test, the overall significance level for the two tests combined is greater than α ; that is, there is more chance of rejecting a true H_0 (i.e., of making a Type I error).

One reasonable (but fairly conservative) way to get around the second difficulty is to use $\alpha/2$ for each separate test to guarantee an overall significance level of no more than α (the Bonferroni correction). Nevertheless, using $\alpha/2$ for each test is conservative (i.e., makes it harder to reject either H_0), thus making it difficult to detect a real difference between the two lines.

With regard to the first difficulty, even if both tests are not rejected, it is still possible (although unlikely) that the two lines do not coincide. This is so because each separate test (e.g., the test of $H_0: \beta_{1M} = \beta_{1F}$) allows the remaining parameters (β_{0M} and β_{0F}) to be unequal. In other words, the test for equal slopes does not assume equal intercepts; nor does the test for equal intercepts assume equal slopes. The multiple regression procedure, which involves using a single model containing a dummy variable for group status, avoids this drawback and permits testing for common slope and common intercept *simultaneously*.

We saw in Examples 12.1 and 12.2 that the null hypothesis of equal slopes (regardless of the intercepts) was not rejected ($P > .90$) and that the null hypothesis of equal intercepts (regardless of the slopes) was associated with a P -value of between .05 and .10. Putting these two facts together, we would be inclined to support the conclusion that there is no *strong* evidence for noncoincidence. As we shall see shortly, the more appropriate test procedure involving a single model yields a different conclusion.

12.8 Method II: Using a Single Regression Equation to Compare Two Straight Lines

Another approach for comparing regression equations uses a single multiple regression model that contains one or more dummy variables to distinguish the groups being compared. The model for comparing two straight lines is given by (12.3), which we restate here:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + E$$

where $Y = \text{SBP}$, $X = \text{AGE}$, and Z is a dummy variable indicating sex (1 if female, 0 if male). For the data in Table 12.1 ($n_M = 40$, $n_F = 29$), the fitted model is

$$\hat{Y} = 110.04 + 0.96X - 12.96Z - 0.012XZ$$

which yields the following separate straight-line equations:

$$Z = 0: \quad \hat{Y}_M = 110.04 + 0.96X$$

$$Z = 1: \quad \hat{Y}_F = 97.08 + 0.95X$$

These two straight-line equations are identical to those obtained in Section 12.4 by fitting separate regressions.

Table 12.2 provides ANOVA results needed to answer statistical inference questions about this model. This table provides variables-added-in-order tests for the fitted regression equation and allows us to perform appropriate tests for parallelism, for equal intercepts, and for coincidence.

12.8.1 Test of Parallelism: Single-model Approach

Referring again to the dummy variable model of (12.3), we know the null hypothesis that the two regression lines are parallel is equivalent to $H_0: \beta_3 = 0$. If $\beta_3 = 0$, then the slope for

TABLE 12.2 Three models for the age–systolic blood pressure example

Source	d.f.	SS	MS	F
Regression (X)	1	14,951.25	14,951.25	121.27
Residual	67	8,260.51	123.29	
Regression (X, Z)	2	18,009.78	9,004.89	114.25
Residual	66	5,201.99	78.82	
Regression (X, Z, XZ)	3	18,010.33	6,003.44	75.02
Residual	65	5,201.44	80.02	

© Cengage Learning

females, $\beta_{1F} = \beta_1 + \beta_3$, simplifies to β_1 , which is the slope for males (i.e., the two lines are parallel). The test statistic for testing $H_0: \beta_3 = 0$ is the partial F statistic (or equivalent t test) for the significance of the addition of the variable XZ to a model already containing X and Z .⁵

In our example, this test statistic is computed as follows:

$$\begin{aligned} F(XZ|X, Z) &= \frac{\text{Regression SS}(X, Z, XZ) - \text{Regression SS}(X, Z)}{\text{MS Residual } (X, Z, XZ)} \\ &= \frac{18,010.33 - 18,009.78}{80.02} \\ &= 0.007 \quad (P = .9342) \end{aligned}$$

This F statistic, with 1 and 65 degrees of freedom, is extremely small (P is very large), so we do not reject H_0 and, therefore, have no statistical basis for believing that the two lines are not parallel. This was the same decision we reached on the basis of separate regression fits. In fact, the F computed here is (theoretically) the square of the corresponding T computed when using separate straight-line fits in Section 12.7.1, although the numerical answers may not exactly agree due to round-off errors.

12.8.2 Test of Equal Intercepts: Single-model Approach

The hypothesis that the two intercepts are equal, allowing for unequal slopes, is equivalent to $H_0: \beta_2 = 0$ for the overall model (12.3). The test compares the overall model

$$Y = \beta_0 + \beta_1X + \beta_2Z + \beta_3XZ + E$$

to the reduced model

$$Y = \beta_0 + \beta_1X + \beta_3XZ + E$$

⁵If H_0 is not rejected by the test of $H_0: \beta_3 = 0$, model (12.3) can be revised to eliminate the β_3 term. This revised (or reduced) model becomes $Y = \beta_0 + \beta_1X + \beta_2Z + E$, which has the form of an analysis-of-covariance model (see Chapter 13).

This is a variables-added-last test considering Z , the sex group dummy variable.⁶ Another approach involves a variables-added-in-order test comparing

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + E$$

to the reduced model

$$Y = \beta_0 + \beta_1 X + E$$

The latter test presumes equal slopes, so it is essentially a test for coincidence, assuming parallelism. Not surprisingly, neither test is uniformly preferred (see Section 12.10). As discussed in Chapters 9 and 15, we recommend using the residual from the full model, (12.3), for either test.

For the example under consideration, we opt for the second approach, since the slope test was not significant. The variables-added-in-order test for $H_0: \beta_2 = 0$ is then computed as follows:

$$\begin{aligned} F(Z|X) &= \frac{\text{Regression SS}(Z, X) - \text{Regression SS}(X)}{\text{MS Residual } (X, Z, XZ)} \\ &= \frac{18,009.78 - 14,951.25}{80.02} \\ &= 38.22 \end{aligned}$$

The preceding test statistic is modified from the usual partial F statistic in that we are now using for the denominator the mean-square residual for the full model, which contains X , Z , and XZ ; the usual partial F statistic would use the mean-square residual for the model containing only X and Z .

Table 12.3 indicates that, with 1 and 65 degrees of freedom, $P < .0001$. Hence, the intercepts are judged to be different for the male and female straight-line models.

TABLE 12.3 ANOVA table with variables-added-in-order tests for method II for the age–systolic blood pressure example

Source	d.f.	SS	MS	F	P
X (AGE)	1	14,951.25	14,951.25	186.84	<.0001
Z (SEX) X	1	3,058.52	3,058.52	38.22	<.0001
XZ X, Z	1	0.55	0.55	0.01	.9342
Residual	65	5,201.44	80.02		
Total (corrected)	68	23,211.76			

© Cengage Learning

⁶ The partial F statistic for the variables-added-last test for equal intercepts, $F(Z|X, XZ)$, is the square of the statistic given by (12.7) when fitting separate straight-line models.

12.8.3 Test of Coincidence: Single-model Approach

The hypothesis that the two regression lines coincide is $H_0: \beta_2 = \beta_3 = 0$. When both β_2 and β_3 are 0, the model for females, $Y_F = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X + E$, reduces to $Y_M = \beta_0 + \beta_1X + E$, the model for males (i.e., the two lines coincide). The test of $H_0: \beta_2 = \beta_3 = 0$ is thus a multiple partial F test, since it involves a subset of regression coefficients.⁷ The two models being compared are, therefore,

$$Y = \beta_0 + \beta_1X + \beta_2Z + \beta_3XZ + E$$

and

$$Y = \beta_0 + \beta_1X + E$$

For our example, the information in either Table 12.2 or Table 12.3 leads to the following computation:

$$\begin{aligned} F(XZ, Z|X) &= \frac{[\text{Regression SS}(X, Z, XZ) - \text{Regression SS}(X)]/2}{\text{MS Residual}(X, Z, XZ)} \\ &= \frac{(18,010.33 - 14,951.25)/2}{80.02} \\ &= 19.1 \end{aligned}$$

Comparing this F with $F_{2, 65, 0.999} = 7.72$, we reject H_0 with $P < .001$ and conclude that very strong evidence exists that the two lines are *not* coincident. This conclusion contradicts our earlier conclusion (Section 12.7.3) based on the results from separate tests for equal slopes and equal intercepts.

12.9 Comparison of Methods I and II

Does the method that uses dummy variables differ from the method that fits two separate regression equations? And if so, is one of the methods preferable to the other?

In deciding whether one method is preferable, we first observe that the two methods yield exactly the same estimated regression coefficients for the two straight-line models. That is, if we fit the model $Y = \beta_0 + \beta_1X + \beta_2Z + \beta_3XZ + E$ by the least-squares method to obtain estimated coefficients $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$, the straight-line equations obtained by setting Z equal to 0 and to 1 in this estimated model will be the same as those obtained by fitting the two straight lines separately. In particular, if $\hat{\beta}_{0M}$, $\hat{\beta}_{0F}$, $\hat{\beta}_{1M}$, and $\hat{\beta}_{1F}$ denote the estimated regression coefficients based on separate regression fits, then $\hat{\beta}_{0M} = \hat{\beta}_0$, $\hat{\beta}_{0F} = \hat{\beta}_0 + \hat{\beta}_2$, $\hat{\beta}_{1M} = \hat{\beta}_1$, and $\hat{\beta}_{1F} = \hat{\beta}_1 + \hat{\beta}_3$. As for statistical tests involving regression coefficients estimated by the two methods, the following two points are valid:

1. *The tests for parallel lines are exactly equivalent;* that is, the T statistic with $n_1 + n_2 - 4$ degrees of freedom computed for testing $H_0: \beta_3 = 0$ in the dummy variable model (method II) is exactly the same as the T statistic given by (12.4) for testing $H_0: \beta_{1M} = \beta_{1F}$ based on fitting two separate models (method I).

⁷ If the test for coincidence is not rejected, model (12.3) can be reduced to the form $Y = \beta_0 + \beta_1X + E$.

2. *The tests for coincident lines differ*, and the one using the dummy variable model (method II) is generally preferable. The approach using separate regressions tests $H_0: \beta_{1M} = \beta_{1F}$ and $H_0: \beta_{0M} = \beta_{0F}$ separately and then rejects the null hypothesis of coincident lines if either or both null hypotheses are rejected. This is exactly equivalent to performing two separate tests of $H_0: \beta_2 = 0$ and $H_0: \beta_3 = 0$ and using the same decision rule for the dummy variable approach; but it is not equivalent to testing the single null hypothesis $H_0: \beta_2 = \beta_3 = 0$ (i.e., testing whether β_2 and β_3 are both simultaneously 0).

As a final note, method II is often easier to implement in statistical computing packages, since only one model needs to be fit and standard output is typically sufficient to conduct the hypothesis tests described. In contrast, method I requires that two models be fit, followed by manual calculations based on the model-fitting results.

12.10 Testing Strategies and Interpretation: Comparing Two Straight Lines

Several strategies can be used to identify a best model for comparing two straight lines. Strategies for more general situations are described in Chapter 16. We prefer a backward strategy for most situations—that is, starting with the largest model of interest and then trying to reduce the model through a sequence of hypothesis tests. A flow diagram of this strategy for comparing two straight lines is given in Figure 12.3. In this case, the largest model to be considered is model (12.3), which contains X , Z , and XZ as independent variables. To reduce the model, we perform tests for coincidence, then for parallelism, and then for equal intercepts, as follows:

1. If the test for coincidence is nonsignificant, we stop further testing and believe that the best model is $Y = \beta_0 + \beta_1X + E$ (i.e., coincident lines, Figure 12.2(d)).
2. If the test for coincidence is significant and the test for parallelism is nonsignificant, the data support a finding of parallel but noncoincident lines (Figure 12.2(a)).
3. If the test for coincidence is significant and the test for parallelism is significant, we might not even be interested in a test for equal intercepts; if we are, the appropriate test procedure involves the variables-added-last statistic $F(Z|X, XZ)$, which does not assume parallel lines. If this test produces a significant result, we would argue for Figure 12.2(c); if it produces a nonsignificant result, we would tend to support Figure 12.2(b).

Applying this strategy to the age–systolic blood pressure data, we would conclude, based on the preceding tests, that the test for coincidence is significant and the test for parallelism is nonsignificant. Our overall conclusion, therefore, is that the best model has the form

$$Y = \beta_0 + \beta_1X + \beta_2Z + E$$

In words, we assume parallel (and noncoincident) lines, as shown in Figure 12.2(a).

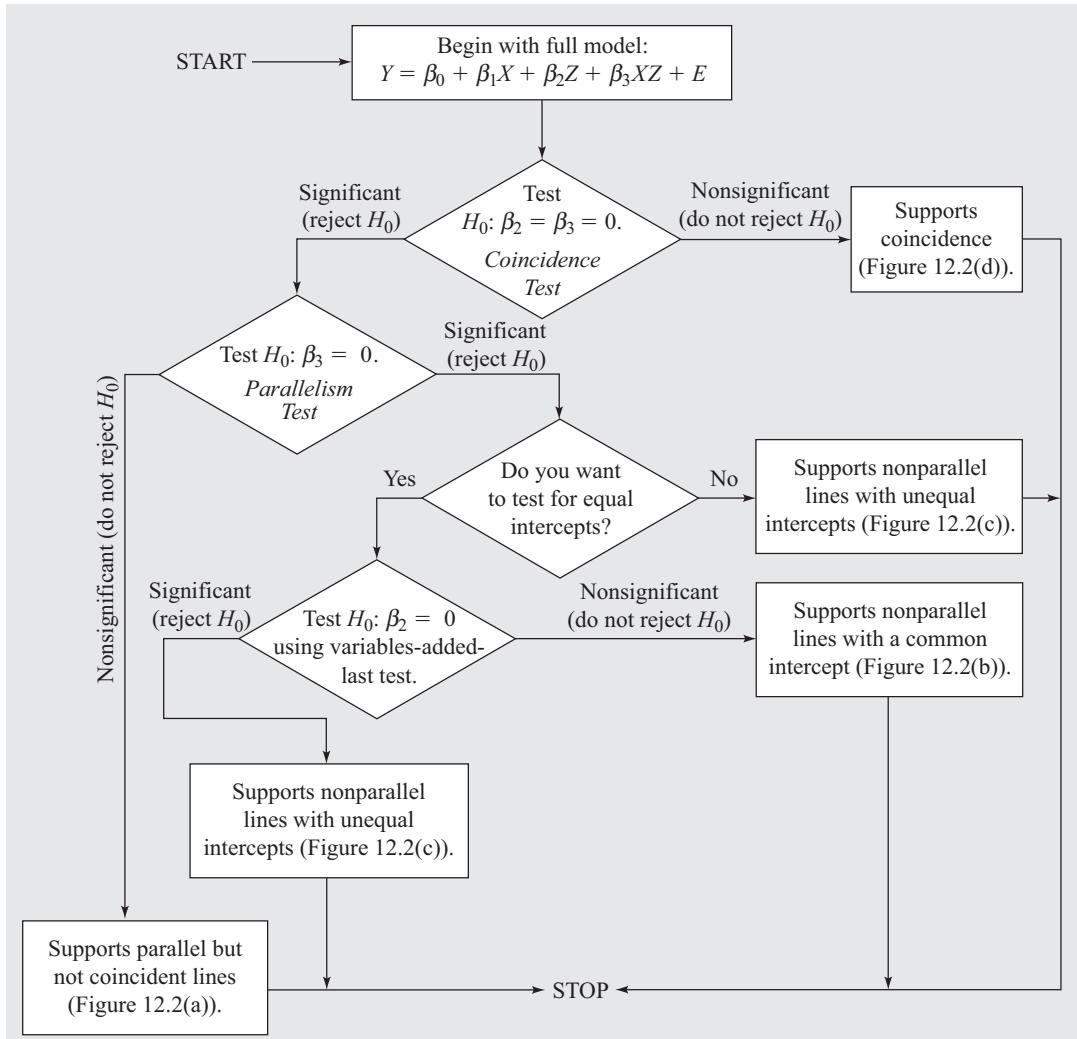


FIGURE 12.3 Comparing two straight lines: Backward testing strategy

12.11 Other Dummy Variable Models

At least two other dummy variable models could have been used instead of (12.3). One such model is given by

$$Y = \beta_0^* + \beta_1^*X + \beta_2^*Z^* + \beta_3^*XZ^* + E \quad (12.9)$$

for which

$$Z^* = \begin{cases} 1 & \text{if subject is male} \\ -1 & \text{if subject is female} \end{cases}$$

This coding scheme for Z^* is known as *effect coding*. Another possible model is

$$Y = \beta_{0M}Z'_1 + \beta_{0F}Z'_2 + \beta_{1M}XZ'_1 + \beta_{1F}XZ'_2 + E \quad (12.10)$$

for which

$$Z'_1 = \begin{cases} 1 & \text{if subject is male} \\ 0 & \text{if subject is female} \end{cases}$$

$$Z'_2 = \begin{cases} 0 & \text{if subject is male} \\ 1 & \text{if subject is female} \end{cases}$$

For model (12.9), the separate regression equations are

$$Z^* = 1: Y_M = (\beta_0^* + \beta_2^*) + (\beta_1^* + \beta_3^*)X + E$$

and

$$Z^* = -1: Y_F = (\beta_0^* - \beta_2^*) + (\beta_1^* - \beta_3^*)X + E$$

The test for parallel lines is equivalent to testing $H_0: \beta_3^* = 0$; the test for equal intercepts is equivalent to testing $H_0: \beta_2^* = 0$; and the test for coincident lines is equivalent to testing $H_0: \beta_2^* = \beta_3^* = 0$.⁸

For model (12.10), the separate regression equations are

$$Z'_1 = 1, Z'_2 = 0: Y_M = \beta_{0M} + \beta_{1M}X + E$$

and

$$Z'_1 = 0, Z'_2 = 1: Y_F = \beta_{0F} + \beta_{1F}X + E$$

The test for parallel lines here is equivalent to testing $H_0: \beta_{1M} = \beta_{1F} = \beta_1$ (not necessarily equal to 0); the test for equal intercepts is equivalent to testing $H_0: \beta_{0M} = \beta_{0F} = \beta_0$ (not necessarily equal to 0); and the test for coincident lines is equivalent to testing $H_0: \beta_{1M} = \beta_{1F} = \beta_1$ and $\beta_{0M} = \beta_{0F} = \beta_0$ simultaneously.⁹ The test of $H_0: \beta_{1M} = \beta_{1F}$ and

⁸ We can express the coefficients of model (12.9) in terms of the regression coefficients for the separate male and female models (12.1) and (12.2) as follows:

$$\beta_0^* = \frac{\beta_{0M} + \beta_{0F}}{2}, \quad \beta_1^* = \frac{\beta_{1M} + \beta_{1F}}{2}, \quad \beta_2^* = \frac{\beta_{0M} - \beta_{0F}}{2}, \quad \beta_3^* = \frac{\beta_{1M} - \beta_{1F}}{2}$$

⁹ The coefficients of model (12.10) are identical to the correspondingly labeled coefficients for the separate male and female models (12.1) and (12.2).

$\beta_{0M} = \beta_{0F}$ differs from previously discussed multiple partial F tests because the coefficients under this H_0 are not equal to 0. The testing procedure in such a case is given as follows:

1. Reduce the model according to the specifications under the null hypothesis; for the test of coincidence, for example, the full model (12.10) becomes

$$\begin{aligned} Y &= \beta_0(Z'_1 + Z'_2) + \beta_1(XZ'_1 + XZ'_2) + E \\ &= \beta_0 + \beta_1X + E \end{aligned}$$

since $Z'_1 + Z'_2 = 1$

2. Find the residual sum of squares for this reduced model.
3. Compute the following F statistic:

$$F = \frac{[\text{Residual SS(reduced model)} - \text{Residual SS(full model)}]/v^*}{\text{MS Residual(full model)}}$$

where v^* is the number of linearly independent parametric functions specified to be 0 under H_0 (in our case $v^* = 2$, since the null hypothesis of coincidence specifies that $\beta_{1M} - \beta_{1F} = 0$ and that $\beta_{0M} - \beta_{0F} = 0$).

4. Test H_0 using F tables with v^* and $n - 4$ degrees of freedom, where 4 is the number of parameters in the full model.

Model (12.10) does not include an overall intercept. Care must be taken in using this particular model with most computer programs (which generally include an overall intercept by default). Collinearity (often labeled “LESS THAN FULL RANK MODEL”) may occur.

12.12 Comparing Four Regression Equations

Suppose that we want to compare the separate multiple regressions of systolic blood pressure (SBP) on age and weight for four social class groups. For each individual in each social class group, we observe values of the variables $Y = \text{SBP}$, $X_1 = \text{AGE}$, and $X_2 = \text{WEIGHT}$. Further, let us suppose that there are n_i individuals in the i th social class (SC) group, $i = 1, 2, 3, 4$. We begin by defining three dummy variables Z_1 , Z_2 , and Z_3 :

$$\begin{aligned} Z_1 &= \begin{cases} 1 & \text{if SC2 member} \\ 0 & \text{otherwise} \end{cases} & Z_2 &= \begin{cases} 1 & \text{if SC3 member} \\ 0 & \text{otherwise} \end{cases} \\ Z_3 &= \begin{cases} 1 & \text{if SC4 member} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

The complete model to be used (if no interaction between AGE and WEIGHT is considered) is given as follows:

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3Z_1 + \beta_4Z_2 + \beta_5Z_3 + \beta_6X_1Z_1 + \beta_7X_2Z_1 + \beta_8X_1Z_2 + \beta_9X_2Z_2 + \beta_{10}X_1Z_3 + \beta_{11}X_2Z_3 + E \quad (12.11)$$

For each particular social class, model (12.11) specializes as follows:

- | | |
|-----------------------------------|---|
| SC1 ($Z_1 = Z_2 = Z_3 = 0$): | $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$ |
| SC2 ($Z_1 = 1, Z_2 = Z_3 = 0$): | $Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_6)X_1 + (\beta_2 + \beta_7)X_2 + E$ |
| SC3 ($Z_1 = Z_3 = 0, Z_2 = 1$): | $Y = (\beta_0 + \beta_4) + (\beta_1 + \beta_8)X_1 + (\beta_2 + \beta_9)X_2 + E$ |
| SC4 ($Z_1 = Z_2 = 0, Z_3 = 1$): | $Y = (\beta_0 + \beta_5) + (\beta_1 + \beta_{10})X_1 + (\beta_2 + \beta_{11})X_2 + E$ |

12.12.1 Tests of Hypotheses

The following hypotheses involving the parameters in model (12.11) are of interest.

1. *All four regression equations¹⁰ are coincident* (i.e., test $H_0: \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$). When H_0 is true, all four social class models reduce to the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$$

The test statistic is the multiple partial

$$F(Z_1, Z_2, Z_3, X_1 Z_1, X_2 Z_1, X_1 Z_2, X_2 Z_2, X_1 Z_3, X_2 Z_3 | X_1, X_2)$$

For this F statistic, the numerator degrees of freedom is 9, since 9 parameters are equal to 0 under H_0 . The denominator degrees of freedom is $n_1 + n_2 + n_3 + n_4 - 12$, which is equal to the total sample size minus the total number of parameters (12) being estimated in the full model.

2. *All four regression equations are parallel* (i.e., test $H_0: \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$). When H_0 is true, the models for the four social classes reduce to

- | | |
|------|---|
| SC1: | $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$ |
| SC2: | $Y = (\beta_0 + \beta_3) + \beta_1 X_1 + \beta_2 X_2 + E$ |
| SC3: | $Y = (\beta_0 + \beta_4) + \beta_1 X_1 + \beta_2 X_2 + E$ |
| SC4: | $Y = (\beta_0 + \beta_5) + \beta_1 X_1 + \beta_2 X_2 + E$ |

Thus, the coefficients of X_1 and X_2 are the same for each social class when H_0 is true (i.e., the four regression equations are said to be parallel). The test statistic for testing H_0 is given by the multiple partial

$$F(X_1 Z_1, X_2 Z_1, X_1 Z_2, X_2 Z_2, X_1 Z_3, X_2 Z_3 | X_1, X_2, Z_1, Z_2, Z_3)$$

which has 6 and $n_1 + n_2 + n_3 + n_4 - 12$ degrees of freedom.

12.12.2 An Alternate Dummy Variable Model

An *effect-coding* dummy variable scheme for comparing the four social class groups begins by defining

¹⁰ In the case where more than one explanatory factor is included in the model, we are no longer discussing coincident or parallel lines but rather equations that may represent planes (in the case of two explanatory factors) or high-dimensional surfaces (in the case of more than two factors).

$$Z_1^* = \begin{cases} -1 & \text{if SC1 member} \\ 1 & \text{if SC2 member} \\ 0 & \text{if SC3 member} \\ 0 & \text{if SC4 member} \end{cases} \quad Z_2^* = \begin{cases} -1 & \text{if SC1 member} \\ 0 & \text{if SC2 member} \\ 1 & \text{if SC3 member} \\ 0 & \text{if SC4 member} \end{cases}$$

$$Z_3^* = \begin{cases} -1 & \text{if SC1 member} \\ 0 & \text{if SC2 member} \\ 0 & \text{if SC3 member} \\ 1 & \text{if SC4 member} \end{cases}$$

Then we use the model

$$Y = \beta_0^* + \beta_1^* X_1 + \beta_2^* X_2 + \beta_3^* Z_1^* + \beta_4^* Z_2^* + \beta_5^* Z_3^* + \beta_6^* X_1 Z_1^* + \beta_7^* X_2 Z_1^* + \beta_8^* X_1 Z_2^* + \beta_9^* X_2 Z_2^* + \beta_{10}^* X_1 Z_3^* + \beta_{11}^* X_2 Z_3^* + E \quad (12.12)$$

For the preceding dummy variable coding, the four regression equations for the four social classes based on model (12.12) are

$$\text{SC 1}(Z_1^* = Z_2^* = Z_3^* = -1):$$

$$Y = (\beta_0^* - \beta_3^* - \beta_4^* - \beta_5^*) + (\beta_1^* - \beta_6^* - \beta_8^* - \beta_{10}^*) X_1 + (\beta_2^* - \beta_7^* - \beta_9^* - \beta_{11}^*) X_2 + E$$

$$\text{SC 2}(Z_1^* = 1, Z_2^* = Z_3^* = 0):$$

$$Y = (\beta_0^* + \beta_3^*) + (\beta_1^* + \beta_6^*) X_1 + (\beta_2^* + \beta_7^*) X_2 + E$$

$$\text{SC 3}(Z_1^* = 0, Z_2^* = 1, Z_3^* = 0):$$

$$Y = (\beta_0^* + \beta_4^*) + (\beta_1^* + \beta_8^*) X_1 + (\beta_2^* + \beta_9^*) X_2 + E$$

$$\text{SC 4}(Z_1^* = Z_2^* = 0, Z_3^* = 1):$$

$$Y = (\beta_0^* + \beta_5^*) + (\beta_1^* + \beta_{10}^*) X_1 + (\beta_2^* + \beta_{11}^*) X_2 + E$$

So the null hypotheses to be tested in connection with parallelism and coincidence (using appropriate multiple partial F tests) are

$$\text{Parallelism: } H_0: \beta_6^* = \beta_7^* = \beta_8^* = \beta_9^* = \beta_{10}^* = \beta_{11}^* = 0$$

$$\text{Coincidence: } H_0: \beta_3^* = \beta_4^* = \beta_5^* = \beta_6^* = \beta_7^* = \beta_8^* = \beta_9^* = \beta_{10}^* = \beta_{11}^* = 0$$

12.13 Comparing Several Regression Equations Involving Two Nominal Variables

Suppose that we want to compare eight regression equations of SBP (Y) on AGE (X_1) and WEIGHT (X_2), corresponding to the eight combinations of SEX (Q) and social class (SC) groups. Then the following regression model can be used:

$$\begin{aligned}
 Y = & \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 Z_1 + \beta_4 Z_2 + \beta_5 Z_3 + \beta_6 Q + \beta_7 Z_1 Q + \beta_8 Z_2 Q \\
 & + \beta_9 Z_3 Q + \beta_{10} X_1 Z_1 + \beta_{11} X_2 Z_1 + \beta_{12} X_1 Z_2 + \beta_{13} X_2 Z_2 + \beta_{14} X_1 Z_3 \\
 & + \beta_{15} X_2 Z_3 + \beta_{16} X_1 Q + \beta_{17} X_2 Q + \beta_{18} X_1 Z_1 Q + \beta_{19} X_2 Z_1 Q \\
 & + \beta_{20} X_1 Z_2 Q + \beta_{21} X_2 Z_2 Q + \beta_{22} X_1 Z_3 Q + \beta_{23} X_2 Z_3 Q + E
 \end{aligned} \tag{12.13}$$

in which the dummy variables are defined as

$$\begin{aligned}
 Z_1 &= \begin{cases} 1 & \text{if SC2 member} \\ 0 & \text{otherwise} \end{cases} & Z_2 &= \begin{cases} 1 & \text{if SC3 member} \\ 0 & \text{otherwise} \end{cases} \\
 Z_3 &= \begin{cases} 1 & \text{if SC4 member} \\ 0 & \text{otherwise} \end{cases} & Q &= \begin{cases} 1 & \text{if subject is male} \\ 0 & \text{if subject is female} \end{cases}
 \end{aligned}$$

For each SEX–SC combination, we have

$$\begin{aligned}
 \text{SC1-male: } & Y = (\beta_0 + \beta_6) + (\beta_1 + \beta_{16})X_1 + (\beta_2 + \beta_{17})X_2 + E \\
 \text{SC2-male: } & Y = (\beta_0 + \beta_3 + \beta_6 + \beta_7) + (\beta_1 + \beta_{10} + \beta_{16} + \beta_{18})X_1 \\
 & + (\beta_2 + \beta_{11} + \beta_{17} + \beta_{19})X_2 + E \\
 \text{SC3-male: } & Y = (\beta_0 + \beta_4 + \beta_6 + \beta_8) + (\beta_1 + \beta_{12} + \beta_{16} + \beta_{20})X_1 \\
 & + (\beta_2 + \beta_{13} + \beta_{17} + \beta_{21})X_2 + E \\
 \text{SC4-male: } & Y = (\beta_0 + \beta_5 + \beta_6 + \beta_9) + (\beta_1 + \beta_{14} + \beta_{16} + \beta_{22})X_1 \\
 & + (\beta_2 + \beta_{15} + \beta_{17} + \beta_{23})X_2 + E \\
 \text{SC1-female: } & Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E \\
 \text{SC2-female: } & Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_{10})X_1 + (\beta_2 + \beta_{11})X_2 + E \\
 \text{SC3-female: } & Y = (\beta_0 + \beta_4) + (\beta_1 + \beta_{12})X_1 + (\beta_2 + \beta_{13})X_2 + E \\
 \text{SC4-female: } & Y = (\beta_0 + \beta_5) + (\beta_1 + \beta_{14})X_1 + (\beta_2 + \beta_{15})X_2 + E
 \end{aligned}$$

The model, therefore, includes $\text{SEX} \times \text{SC}$ interaction but not $\text{WEIGHT} \times \text{AGE}$ interaction. It is best to make decisions about which interactions to include based on subject matter knowledge. However, if necessary, we could test whether there are no interaction effects involving SEX and SC. This is a test of

$$H_0: \beta_7 = \beta_8 = \beta_9 = \beta_{18} = \beta_{19} = \beta_{20} = \beta_{21} = \beta_{22} = \beta_{23} = 0$$

given that all the other terms in (12.13) are included in the model.

Once the decision to include the interaction effect between SEX and SC has been made, several hypotheses concerning model (12.13) are of interest.

1. All eight regression equations are coincident. This is a test of

$$H_0: \beta_3 = \beta_4 = \cdots = \beta_{23} = 0$$

given that X_1 and X_2 are in the model. If this null hypothesis is rejected, we can proceed with further testing.

2. All eight regression equations are parallel. This is a test of

$$H_0: \beta_{10} = \beta_{11} = \cdots = \beta_{23} = 0$$

given that $X_1, X_2, Z_1, Z_2, Z_3, Q, Z_1Q, Z_2Q$, and Z_3Q are in the model. If this null hypothesis is true, the eight equations are all of the form

$$Y = \beta_{0(j)} + \beta_1 X_1 + \beta_2 X_2 + E \quad \text{for } j = 1, 2, \dots, 8$$

These eight models differ only in the intercept, $\beta_{0(j)}$, which varies as SC (i.e., Z_1, Z_2, Z_3) and SEX (i.e., Q) vary.

Also, the specific effects of SEX and SC can be investigated via the following hypotheses.

1. Male and female regression equations are coincident (controlling for SC). This is a test of

$$H_0: \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{16} = \beta_{17} = \beta_{18} = \beta_{19} = \beta_{20} = \beta_{21} = \beta_{22} = \beta_{23} = 0$$

2. Male and female regression equations are parallel (controlling for SC). This is a test of

$$H_0: \beta_{16} = \beta_{17} = \beta_{18} = \beta_{19} = \beta_{20} = \beta_{21} = \beta_{22} = \beta_{23} = 0$$

When this null hypothesis is true, the eight equations above reduce to

$$\begin{cases} \text{SC1-male: } Y = (\beta_0 + \beta_6) + \beta_1 X_1 + \beta_2 X_2 + E \\ \text{SC1-female: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E \\ \text{SC2-male: } Y = (\beta_0 + \beta_3 + \beta_6 + \beta_7) + (\beta_1 + \beta_{10}) X_1 + (\beta_2 + \beta_{11}) X_2 + E \\ \text{SC2-female: } Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_{10}) X_1 + (\beta_2 + \beta_{11}) X_2 + E \\ \text{SC3-male: } Y = (\beta_0 + \beta_4 + \beta_6 + \beta_8) + (\beta_1 + \beta_{12}) X_1 + (\beta_2 + \beta_{13}) X_2 + E \\ \text{SC3-female: } Y = (\beta_0 + \beta_4) + (\beta_1 + \beta_{12}) X_1 + (\beta_2 + \beta_{13}) X_2 + E \\ \text{SC4-male: } Y = (\beta_0 + \beta_5 + \beta_6 + \beta_9) + (\beta_1 + \beta_{14}) X_1 + (\beta_2 + \beta_{15}) X_2 + E \\ \text{SC4-female: } Y = (\beta_0 + \beta_5) + (\beta_1 + \beta_{14}) X_1 + (\beta_2 + \beta_{15}) X_2 + E \end{cases}$$

Thus, within any specific social class group, the male and female regression equations are parallel (since they have the same X_1 and X_2 coefficients).

3. All four social class equations are coincident (controlling for SEX). This is a test of

$$H_0: \beta_3 = \beta_4 = \beta_5 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = \beta_{18} = \beta_{19} = \beta_{20} = \beta_{21} = \beta_{22} = \beta_{23} = 0$$

4. All four social class equations are parallel (controlling for SEX). This is a test of

$$H_0: \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = \beta_{18} = \beta_{19} = \beta_{20} = \beta_{21} = \beta_{22} = \beta_{23} = 0$$

When this hypothesis is true, the eight equations reduce to

$$\begin{cases} \text{SC1-male: } Y = (\beta_0 + \beta_6) + (\beta_1 + \beta_{16}) X_1 + (\beta_2 + \beta_{17}) X_2 + E \\ \text{SC2-male: } Y = (\beta_0 + \beta_3 + \beta_6 + \beta_7) + (\beta_1 + \beta_{16}) X_1 + (\beta_2 + \beta_{17}) X_2 + E \\ \text{SC3-male: } Y = (\beta_0 + \beta_4 + \beta_6 + \beta_8) + (\beta_1 + \beta_{16}) X_1 + (\beta_2 + \beta_{17}) X_2 + E \\ \text{SC4-male: } Y = (\beta_0 + \beta_5 + \beta_6 + \beta_9) + (\beta_1 + \beta_{16}) X_1 + (\beta_2 + \beta_{17}) X_2 + E \\ \text{SC1-female: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E \\ \text{SC2-female: } Y = (\beta_0 + \beta_3) + \beta_1 X_1 + \beta_2 X_2 + E \\ \text{SC3-female: } Y = (\beta_0 + \beta_4) + \beta_1 X_1 + \beta_2 X_2 + E \\ \text{SC4-female: } Y = (\beta_0 + \beta_5) + \beta_1 X_1 + \beta_2 X_2 + E \end{cases}$$

Thus, within any given sex group, all four regression equations have the same coefficients for X_1 and X_2 .

■ **Example 12.3** Suppose that, in a study similar to that of Chapter 5, Problem 2, we are interested in comparing the multiple regressions of systolic blood pressure (SBP) on quetelet index (QUET) for four age groups and two smoking status categories. For each individual in each age group, we observe the values of the variables $Y = \text{SBP}$, $X = \text{QUET}$, and $\text{SMK} = \text{smoking status}$. The relevant dummy variables (Z_1 , Z_2 , Z_3 , and Q) can be defined as

$$\begin{aligned} Z_1 &= \begin{cases} 1 & \text{if age } 40-49 \text{ years} \\ 0 & \text{otherwise} \end{cases} & Z_2 &= \begin{cases} 1 & \text{if age } 50-59 \text{ years} \\ 0 & \text{otherwise} \end{cases} \\ Z_3 &= \begin{cases} 1 & \text{if age } 60 \text{ or older} \\ 0 & \text{otherwise} \end{cases} & Q &= \begin{cases} 0 & \text{if never a smoker (SMK = 0)} \\ 1 & \text{if ever a smoker (SMK = 1)} \end{cases} \end{aligned}$$

Note that $Z_1 = Z_2 = Z_3 = 0$ for ages 30–39.

The complete multiple regression model to be used is as follows:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 Z_3 + \beta_5 Q \\ &\quad + \beta_6 Z_1 Q + \beta_7 Z_2 Q + \beta_8 Z_3 Q + \beta_9 XZ_1 + \beta_{10} XZ_2 + \beta_{11} XZ_3 \\ &\quad + \beta_{12} XQ + \beta_{13} XZ_1 Q + \beta_{14} XZ_2 Q + \beta_{15} XZ_3 Q + E \end{aligned} \tag{12.14}$$

The data for this hypothetical example are shown in Table 12.4.

For each age-smoking status combination, the regression model above specializes as follows:

$$\begin{aligned} 30-39 \text{ yrs, nonsmoker: } Y &= \beta_0 + \beta_1 X + E \\ 30-39 \text{ yrs, smoker: } Y &= (\beta_0 + \beta_5) + (\beta_1 + \beta_{12}) X + E \\ 40-49 \text{ yrs, nonsmoker: } Y &= (\beta_0 + \beta_2) + (\beta_1 + \beta_9) X + E \\ 40-49 \text{ yrs, smoker: } Y &= (\beta_0 + \beta_2 + \beta_5 + \beta_6) + (\beta_1 + \beta_9 + \beta_{12} + \beta_{13}) X + E \\ 50-59 \text{ yrs, nonsmoker: } Y &= (\beta_0 + \beta_3) + (\beta_1 + \beta_{10}) X + E \\ 50-59 \text{ yrs, smoker: } Y &= (\beta_0 + \beta_3 + \beta_5 + \beta_7) + (\beta_1 + \beta_{10} + \beta_{12} + \beta_{14}) X + E \\ 60+ \text{ yrs, nonsmoker: } Y &= (\beta_0 + \beta_4) + (\beta_1 + \beta_{11}) X + E \\ 60+ \text{ yrs, smoker: } Y &= (\beta_0 + \beta_4 + \beta_5 + \beta_8) + (\beta_1 + \beta_{11} + \beta_{12} + \beta_{15}) X + E \end{aligned}$$

Several hypothesis tests will be performed in order to compare the multiple regressions of SBP on QUET for different age groups and smoking status categories.

- First, we will determine whether the SBP–QUET relationship is coincident for non-smokers and smokers, within each age group, by testing the null hypothesis

$$H_0: \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = 0$$

given that X , Z_1 , Z_2 , Z_3 , XZ_1 , XZ_2 , and XZ_3 are in the model. The test statistic for this test of coincidence is

TABLE 12.4 Data for Example 12.3

Person	SBP (Y)	QUET (X)	SMK	Age Group	Person	SBP (Y)	QUET (X)	SMK	Age Group
1	120	2.789	0	1	36	160	3.612	1	2
2	152	4.116	0	4	37	142	3.401	0	3
3	135	3.171	0	3	38	140	3.562	1	3
4	180	4.637	1	4	39	144	2.368	1	1
5	137	3.296	0	2	40	142	3.024	1	1
6	162	3.668	1	4	41	142	3.401	0	3
7	130	3.100	0	2	42	161	3.800	0	4
8	162	3.668	1	4	43	130	3.100	0	2
9	122	3.251	0	1	44	144	3.751	0	3
10	132	3.210	0	2	45	146	2.979	1	3
11	138	3.673	0	3	46	166	3.877	1	4
12	180	4.637	1	4	47	129	2.790	1	1
13	135	2.876	0	1	48	138	4.032	1	2
14	137	3.296	0	2	49	148	3.768	0	2
15	120	2.789	0	1	50	132	3.017	1	2
16	149	3.301	1	3	51	150	3.628	1	3
17	145	3.360	1	2	52	129	2.790	1	1
18	161	3.800	0	4	53	166	3.877	1	4
19	122	3.251	0	1	54	170	4.132	1	4
20	145	3.360	1	2	55	134	2.998	1	2
21	144	3.751	0	3	56	138	3.673	0	3
22	140	3.562	1	3	57	134	2.998	1	2
23	132	3.017	1	2	58	164	4.010	0	4
24	164	4.010	0	4	59	126	2.956	1	1
25	152	3.962	0	4	60	170	4.132	1	4
26	135	2.876	0	1	61	149	3.301	1	3
27	126	2.956	1	1	62	160	3.612	1	2
28	150	3.628	1	3	63	142	3.024	1	1
29	148	3.768	0	2	64	132	3.210	0	2
30	135	3.171	0	3					
31	152	4.116	0	4					
32	152	3.962	0	4					
33	146	2.979	1	3					
34	138	4.032	1	2					
35	144	2.368	1	1					

Note: Age Group = 1 if age is 30–39 years
 2 if age is 40–49 years
 3 if age is 50–59 years
 4 if age is 60 + years

$$\begin{aligned}
 & F(Q, Z_1Q, Z_2Q, Z_3Q, XQ, XZ_1Q, XZ_2Q, XZ_3Q | X, Z_1, Z_2, Z_3, XZ_1, XZ_2, XZ_3) \\
 & \text{[Regression SS}(X, Z_1, Z_2, Z_3, Q, Z_1Q, Z_2Q, Z_3Q, XZ_1, XZ_2, XZ_3, XQ, XZ_1Q, XZ_2Q, XZ_3Q) \\
 & - \text{Regression SS}(X, Z_1, Z_2, Z_3, XZ_1, XZ_2, XZ_3)]/8 \\
 & = \frac{\text{MS Residual } (X, Z_1, Z_2, Z_3, Q, Z_1Q, Z_2Q, Z_3Q, XZ_1, XZ_2, XZ_3, XQ, XZ_1Q, XZ_2Q, XZ_3Q)}{8}
 \end{aligned}$$

For this example, this multiple partial F test has 8 numerator d.f. and 48 denominator d.f. and value 3.69 ($P = .002$). At the $\alpha = .05$ level of significance, we reject the null hypothesis that the regression lines for nonsmokers and smokers, controlling for age group, are coincident.

2. Next, since the regression lines for nonsmokers and smokers have been found to differ by age group, we will perform a test of parallelism. The null hypothesis of interest is

$$H_0: \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = 0$$

given that $X, Z_1, Z_2, Z_3, Q, Z_1Q, Z_2Q, Z_3Q, XZ_1, XZ_2$, and XZ_3 are in the model. The test statistic is

$$F(XQ, XZ_1Q, XZ_2Q, XZ_3Q | X, Z_1, Z_2, Z_3, Q, Z_1Q, Z_2Q, Z_3Q, XZ_1, XZ_2, XZ_3)$$

$$= \frac{[\text{Regression SS}(X, Z_1, Z_2, Z_3, Q, Z_1Q, Z_2Q, Z_3Q, XZ_1, XZ_2, XZ_3, XQ, XZ_1Q, XZ_2Q, XZ_3Q) - \text{Regression SS}(X, Z_1, Z_2, Z_3, Q, Z_1Q, Z_2Q, Z_3Q, XZ_1, XZ_2, XZ_3)] / 4}{\text{MS Residual}(X, Z_1, Z_2, Z_3, Q, Z_1Q, Z_2Q, Z_3Q, XZ_1, XZ_2, XZ_3, XQ, XZ_1Q, XZ_2Q, XZ_3Q)}$$

which, for this example, has 4 and 48 degrees of freedom and value 1.85 ($P = .14$). At the $\alpha = .05$ level of significance, we fail to reject the null hypothesis that the regression lines for nonsmokers and smokers, controlling for age group, are parallel.

Therefore, we conclude that the age group-specific straight lines for nonsmokers and smokers are parallel but not coincident.

3. We will determine whether the SBP–QUET relationship is coincident for the four age groups, controlling for smoking status, by testing the following null hypothesis:

$$H_0: \beta_2 = \beta_3 = \beta_4 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{13} = \beta_{14} = \beta_{15} = 0$$

given that X, Q , and XQ are in the model. The test statistic for the relevant multiple partial F test is

$$F(Z_1, Z_2, Z_3, Z_1Q, Z_2Q, Z_3Q, XZ_1, XZ_2, XZ_3, XZ_1Q, XZ_2Q, XZ_3Q | X, Q, XQ)$$

$$= \frac{[\text{Regression SS}(X, Z_1, Z_2, Z_3, Q, Z_1Q, Z_2Q, Z_3Q, XZ_1, XZ_2, XZ_3, XQ, XZ_1Q, XZ_2Q, XZ_3Q) - \text{Regression SS}(X, Q, XQ)] / 12}{\text{MS Residual}(X, Z_1, Z_2, Z_3, Q, Z_1Q, Z_2Q, Z_3Q, XZ_1, XZ_2, XZ_3, XQ, XZ_1Q, XZ_2Q, XZ_3Q)}$$

which, for this example, has 12 and 48 degrees of freedom and value 5.65 ($P < .0001$). At the $\alpha = .05$ level of significance, we reject the null hypothesis of coincident lines for the four different age groups, controlling for smoking status.

4. Finally, since the regression lines for the four age groups have been found to differ (controlling for smoking status), we will perform a test of parallelism. The null hypothesis is

$$H_0: \beta_9 = \beta_{10} = \beta_{11} = \beta_{13} = \beta_{14} = \beta_{15} = 0$$

given that $X, Z_1, Z_2, Z_3, Q, Z_1Q, Z_2Q, Z_3Q$, and XQ are in the model. The test statistic is

$$F(XZ_1, XZ_2, XZ_3, XZ_1Q, XZ_2Q, XZ_3Q | X, Z_1, Z_2, Z_3, Q, Z_1Q, Z_2Q, Z_3Q, XQ)$$

$$= \frac{[\text{Regression SS}(X, Z_1, Z_2, Z_3, Q, Z_1Q, Z_2Q, Z_3Q, XZ_1, XZ_2, XZ_3, XQ, XZ_1Q, XZ_2Q, XZ_3Q) - \text{Regression SS}(X, Z_1, Z_2, Z_3, Q, Z_1Q, Z_2Q, Z_3Q, XQ)]/6}{\text{MS Residual}(X, Z_1, Z_2, Z_3, Q, Z_1Q, Z_2Q, Z_3Q, XZ_1, XZ_2, XZ_3, XQ, XZ_1Q, XZ_2Q, XZ_3Q)}$$

which, for this example, has 6 and 48 degrees of freedom and value 3.22 ($P = .01$). At the $\alpha = .05$ level of significance, we reject the null hypothesis that the regression lines for the four age groups, within each of the two categories of smoking status, are parallel.

It is not of interest to test for differences in intercepts in this example; therefore, we simply conclude that the sets of straight lines, controlling for either SC or SMK, are not parallel. From the form of the complete model (12.14), each of the coefficients in the parallelism null hypotheses considered corresponds (as expected) to a product term of the general form Z_jQ or XZ_jQ . ■

Problems

- Using the data from Problem 2 in Chapter 5 and/or the SAS output given here, answer the following questions about the separate straight-line regressions of SBP on QUET for smokers (SMK = 1) and nonsmokers (SMK = 0).
 - Determine the least-squares line of SBP (Y) on QUET (X) separately for smokers and nonsmokers.
 - Test H_0 : “The slopes are the same for the populations of smokers and nonsmokers being sampled” versus H_A : “Nonsmokers have a more positive slope.”
 - Test H_0 : “The intercepts are the same for the populations of smokers and nonsmokers being sampled” versus H_A : “The intercepts are different.”
 - Test H_0 : “The straight lines coincide for the populations of smokers and nonsmokers being sampled” versus H_A : “The straight lines do not coincide.”

Edited SAS Output (PROC REG) for Problem 1

Regression of SBP on QUET: Nonsmokers

DESCRIPTIVE STATISTICS					
Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
QUET	52.17400	3.47827	183.93641	0.17581	0.41930
SBP	2112.00000	140.80000	299700	166.45714	12.90183
ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1702.83966	1702.83966	35.27	<.0001
Error	13	627.56034	48.27387		
Corrected Total	14	2330.40000			

(continued)

Root MSE	6.94794	R-Square	0.7307
Dependent Mean	140.80000	Adj R-Sq	0.7100
Coeff Var	4.93462		

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	49.31176	15.50814	3.18	0.0072
QUET	1	26.30283	4.42865	5.94	<.0001

Regression of SBP on QUET: Smokers

DESCRIPTIVE STATISTICS					
Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
QUET	57.94100	3.40829	202.63931	0.32246	0.56785
SBP	2513.00000	147.82353	375183	231.40441	15.21198

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2088.16977	2088.16977	19.40	0.0005
Error	15	1614.30082	107.62005		
Corrected Total	16	3702.47059			

Root MSE	10.37401	R-Square	0.5640
Dependent Mean	147.82353	Adj R-Sq	0.5349
Coeff Var	7.01783		

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	79.25533	15.76837	5.03	0.0002
QUET	1	20.11804	4.56719	4.40	0.0005

2. A topic of major concern to demographers and economists is the effect of a high fertility rate on per capita income. The first two accompanying tables display values of per capita income (PCI) and population percentage under age 15 (YNG) for a hypothetical sample of developing countries in Latin America and Africa, respectively. The third table summarizes the results of straight-line regressions of PCI (Y) on YNG (X) for each group of countries.
- a.–d. Repeat parts (a) through (d) of Problem 1 for the straight-line regressions of PCI (Y) on YNG (X) for Latin American and African countries.

Latin American Countries

YNG (X)	PCI (Y)	YNG (X)	PCI (Y)	YNG (X)	PCI (Y)
32.2	788	44.0	292	35.0	685
47.0	202	44.0	321	47.4	220
34.0	825	43.0	300	48.0	195
36.0	675	43.0	323	37.0	605
38.7	590	40.0	484	38.4	530
40.9	408	37.0	625	40.6	480
45.0	324	39.0	525	35.8	690
45.4	235	44.6	340	36.0	685
42.2	338	33.0	765		

African Countries

YNG (X)	PCI (Y)	YNG (X)	PCI (Y)	YNG (X)	PCI (Y)
34.0	317	41.0	188	39.0	225
36.0	270	42.0	166	39.0	232
38.2	208	45.0	132	37.0	260
43.0	150	36.0	290	37.0	250
44.0	105	42.6	160	46.0	92
44.0	128	33.0	300	45.6	110
45.0	85	33.0	320	42.0	180
48.0	75	47.0	85	38.8	235
40.0	210	47.0	75		

Summary of Separate Straight-Line Fits

Location	n	$\hat{\beta}_0$	$\hat{\beta}_1$	\bar{X}	\bar{Y}	s_x^2	$s_{Y X}^2$	r
Latin America	26	2170.67	-42.0	40.277	478.846	21.633	1391.756	-0.983
Africa	26	897.519	-17.39	40.892	186.462	20.244	188.919	-0.985

- e. Test H_0 : "The population correlation coefficients are equal for the two groups of countries under study." Use $\alpha = .05$. Does your conclusion here clash with your findings regarding the equality of slopes? (Hint: See Section 6.7.)

For each of the preceding tests, assume that the alternative hypothesis is two-sided.

- f. Comment on the validity of the homogeneous variance assumption for these data.

3. A team of anthropologists and nutrition experts investigated the influence of protein content in diet on the relationship between AGE and height (HT) for New Guinean children. The first two accompanying tables display values of HT (in centimeters) and AGE for a hypothetical sample of children with protein-rich and protein-poor diets, respectively.
- a.-d. Repeat parts (a) through (d) of Problem 1 for the straight-line regressions of HT (Y) on AGE (X) for the two diets. (Consider a two-sided alternative in each case.)
- e. Test whether the population correlation coefficient for children with a protein-rich diet differs significantly from that for children with a protein-poor diet. (Consider a two-sided alternative.)

Protein-Rich Diet														
AGE (X)	0.2	0.5	0.8	1.0	1.0	1.4	1.8	2.0	2.0	2.5	2.5	3.0	2.7	
HT (Y)	54	54.3	63	66	69	73	82	83	80.3	91	93.2	94	94	
Protein-Poor Diet														
AGE (X)	0.4	0.7	1.0	1.0	1.5	2.0	2.0	2.4	2.8	3.0	1.3	1.8	0.2	3.0
HT (Y)	52	55	61	63.4	66	68.5	67.9	72	76	74	65	69	51	77
Summary of Separate Straight-line Fits														
Diet	n	$\hat{\beta}_0$	$\hat{\beta}_1$	\bar{X}	\bar{Y}	S_x^2	$S_{Y X}^2$	r						
Protein-rich	13	50.324	16.009	1.646	76.677	0.808	5.841	0.937						
Protein-poor	14	51.225	8.686	1.650	65.557	0.873	4.598	0.969						

4. For the data involving regression of DI (Y) on IQ (X) in Problem 4 in Chapter 5, assume that the sample of 17 observations (with the outlier removed) consists of males only. Now suppose that another sample of observations on DI (Y) and IQ (X) has been obtained for 14 females. The information needed to compare the straight-line regression equations for males and females is given in the following table.

Sex Group	n	$\hat{\beta}_0$	$\hat{\beta}_1$	\bar{X}	\bar{Y}	S_x^2	$S_{Y X}^2$	r
Males	17	70.846	-0.444	101.411	25.812	215.882	24.335	-0.807
Females	14	61.871	-0.438	101.053	17.579	175.497	16.692	-0.825

- a.-d. Repeat parts (a) through (d) of Problem 1 for the straight-line regressions of DI (Y) on IQ (X) for the two sexes. (Consider two-sided alternatives.)
e. Perform a two-sided test of whether the population correlation coefficients for males and females are equal.
5. Assume that the data involving the regression of VOTE (Y) on TVEXP (X) in Problem 5 in Chapter 5 came from congressional districts in New York. Now, suppose that researchers selected a second sample of 17 congressional districts in California and recorded the same information. The following table provides the information needed to compare the straight-line regression equations for New York and California.

Location	n	$\hat{\beta}_0$	$\hat{\beta}_1$	\bar{X}	\bar{Y}	S_x^2	$S_{Y X}^2$	r
New York	20	2.174	1.177	36.99	45.71	76.870	11.101	0.954
California	17	8.030	1.036	36.371	45.706	97.335	13.492	0.945

- a.-d. Repeat parts (a) through (d) of Problem 1 for the straight-line regression of VOTE (Y) on TVEXP (X) for the two states. Consider the one-sided alternative $H_A: \beta_1(\text{CAL}) < \beta_1(\text{NY})$ for the test for slope, and consider the one-sided alternative $H_A: \beta_0(\text{CAL}) < \beta_0(\text{NY})$ for the test for intercept.

- e. Perform a two-sided test of whether the correlation coefficients for New York and California are equal.
6. The data in the following table represent four-week growth rates for depleted chicks at different dosage levels of vitamin B, by sex.¹¹

Males		Females	
Growth Rate (Y)	Log ₁₀ Dose (X)	Growth Rate (Y)	Log ₁₀ Dose (X)
17.1	0.301	18.5	0.301
14.3	0.301	22.1	0.301
21.6	0.301	15.3	0.301
24.5	0.602	23.6	0.602
20.6	0.602	26.9	0.602
23.8	0.602	20.2	0.602
27.7	0.903	24.3	0.903
31.0	0.903	27.1	0.903
29.4	0.903	30.1	0.903
30.1	1.204	28.1	0.903
28.6	1.204	30.3	1.204
34.2	1.204	33.0	1.204
37.3	1.204	35.8	1.204
33.3	1.505	32.6	1.505
31.8	1.505	36.1	1.505
40.2	1.505	30.5	1.505

Use the information provided in the next table to answer the following questions.

Sex Group	n	$\hat{\beta}_0$	$\hat{\beta}_1$	\bar{X}	\bar{Y}	S_x^2	$S_{Y X}^2$	r
Males	16	14.178	14.825	0.922	27.84	0.187	10.5553	0.807
Females	16	15.656	12.735	0.903	27.16	0.1812	8.4719	0.788

- a. Determine the dose-response straight lines separately for each sex, and plot them on the same graph.
- b. Test whether the slopes for males and females differ.
- c. Find a 99% confidence interval for the true difference between the male and female slopes.
- d. Test for coincidence of the two straight lines.
7. The results in the first of the following tables were obtained in a study of the amount of energy metabolized by two similar species of birds under constant temperature.¹² Information based on separate straight-line fits to each data set is summarized in the second table.
- a. Plot the least-squares straight lines for each species on the same graph.
- b. Test whether the two lines are parallel.
- c. Test whether the two lines have the same intercept.

¹¹ Adapted from a study by Clark, Lechyeka, and Cook (1940).

¹² Adapted from a study by David (1955).

- d. Give a 95% confidence interval for the true difference between the mean amounts of energy metabolized by each species at 15°C. [Hint: Use a confidence interval of the form

$$(\hat{Y}_{15}^A - \hat{Y}_{15}^B) \pm t_{n_A + n_B - 4, 1 - \alpha/2} \sqrt{S_{\hat{Y}_{15}^A}^2 + S_{\hat{Y}_{15}^B}^2}$$

Species A		Species B	
Calories (Y)	Temperature (X) (°C)	Calories (Y)	Temperature (X) (°C)
36.9	0	41.1	0
35.8	2	40.6	2
34.6	4	38.9	4
34.3	6	37.9	6
32.8	8	37.0	8
31.7	10	36.1	10
31.0	12	36.3	12
29.8	14	34.2	14
29.1	16	33.4	16
28.2	18	32.8	18
27.4	20	32.0	20
27.8	22	31.9	22
25.5	24	30.7	24
24.9	26	29.5	26
23.7	28	28.5	28
23.1	30	27.7	30

Species	n	$\hat{\beta}_0$	$\hat{\beta}_1$	\bar{X}	\bar{Y}	S_x^2	$S_{Y X}^2$	r
A	16	36.579	-0.4528	15.00	29.79	90.6667	0.1662	-0.9959
B	16	40.839	-0.4368	15.00	34.29	90.6667	0.1757	-0.9953

where \hat{Y}_{15}^i is the predicted value at 15°C for species i ($i = A, B$) and $S_{\hat{Y}_{15}^i}^2$ is the estimated variance of \hat{Y}_{15}^i given by the general formula (for $X = X_0$)

$$S_{\hat{Y}_{X_0}^i}^2 = S_{P, Y|X}^2 \left[\frac{1}{n_i} + \frac{(X_0 - \bar{X})^2}{(n_i - 1)S_{X_i}^2} \right].$$

8. In Problem 1, separate straight-line regressions of SBP on QUET were compared for smokers ($SMK = 1$) and nonsmokers ($SMK = 0$).
- Define a single multiple regression model that uses the data for both smokers and nonsmokers and that defines straight-line models for each group with possibly differing intercepts and slopes. Define the intercept and slope for each straight-line model in terms of the regression coefficients of the single regression model.
 - Using the accompanying computer output, determine and plot on graph paper the two fitted straight lines obtained from the fit of the regression model.

- c. Test the following null hypotheses:

H_0 : "The two lines are parallel."

H_0 : "The two lines are coincident."

For each of these tests, state the appropriate null hypothesis in terms of the regression coefficients of the regression model.

- d. Compare your answers in parts (b) and (c) with those you obtained by fitting separate regressions in Problem 1.

Edited SAS Output (PROC REG) for Problem 8

Regression of SBP on QUET, SMK, and QUET \times SMK

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4184.10759	1394.70253	17.42	<.0001
Error	28	2241.86116	80.06647		
Corrected Total	31	6425.96875			

Root MSE	8.94799	R-Square	0.6511
Dependent Mean	144.53125	Adj R-Sq	0.6137
Coeff Var	6.19104		

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	49.31176	19.97235	2.47	0.0199	668457
QUET	1	26.30283	5.70349	4.61	<.0001	3537.94574
SMK	1	29.94357	24.16355	1.24	0.2256	582.42075
QUETSMK	1	-6.18478	6.93171	-0.89	0.3799	63.74110

9. Use the computer output shown next to compare the separate regressions of SBP on AGE and QUET for smokers and nonsmokers (based on the data from Problem 2 in Chapter 5), as follows.
- State the appropriate regression model, simultaneously incorporating equations for both smokers and nonsmokers.
 - Determine the fitted regression equations for smokers and nonsmokers separately, using the fitted regression model.
 - Test for parallelism of the two models, stating the null hypothesis in terms of appropriate regression coefficients.
 - Test for coincidence of the two models, stating the null hypothesis in terms of regression coefficients.

Edited SAS Output (PROC REG) for Problem 9

Regression of SBP on AGE, QUET, SMK, AGESMK, and QUETSMK

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	4915.63040	983.12608	16.92	<.0001
Error	26	1510.33835	58.08994		
Corrected Total	31	6425.96875			

Root MSE	7.62168	R-Square	0.7650
Dependent Mean	144.53125	Adj R-Sq	0.7198
Coeff Var	5.27338		

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	48.61270	17.01537	2.86	0.0083	668457
AGE	1	1.02892	0.50177	2.05	0.0505	3861.63038
QUET	1	10.45104	9.13015	1.14	0.2628	258.96187
SMK	1	-0.53744	23.23004	-0.02	0.9817	769.23345
AGESMK	1	0.43733	0.71279	0.61	0.5449	18.92033
QUETSMK	1	-3.70682	10.76763	-0.34	0.7334	6.88437

10. The results presented in the following tables are based on data from a study by Gruber (1970) to determine how and to what extent changes in blood pressure over time depend on initial blood pressure (at the beginning of the study), the sex of the individual, and the relative weight of the individual. Data were collected on $n = 104$ persons, for which the $k = 7$ independent variables were examined by multiple regression. The following variables were used in the study:

$Y = \text{SBPSL}$ (estimated slope based on the straight-line regression of an individual's blood pressure over time)

$X_1 = \text{SBP1}$ (initial systolic blood pressure)

$X_2 = \text{SEX}$ (male = 1, female = -1)

$X_3 = \text{RW}$ (relative weight)

$X_4 = X_1X_2, \quad X_5 = X_1X_3$

$X_6 = X_2X_3, \quad X_7 = X_1X_2X_3$

Variable	$\hat{\beta}$	$S_{\hat{\beta}}$	Partial $F(\hat{\beta}^2/S_{\hat{\beta}}^2)$	
$X_1(\text{SBP1})$	-0.045	0.00762	34.987	
$X_2(\text{SEX})$	0.695	0.86644	0.643	
$X_3(\text{RW})$	0.027	0.07049	0.149	
$X_4(X_1X_2)$	-0.0029	0.00762	0.145	
$X_5(X_1X_3)$	-0.00018	0.00062	0.084	
$X_6(X_2X_3)$	-0.0092	0.07049	0.017	
$X_7(X_1X_2X_3)$	0.00022	0.00062	0.125	
(Intercept)	4.667			
Source	d.f.	SS	MS	F
Overall regression	7	37.148	5.307	$\frac{5.307}{76.246/96} = 6.68^{**}$
X_1	1	24.988	24.988	$\frac{24.988}{88.406/102} = 28.83^{**}$
$X_2 X_1$	1	7.886	7.886	$\frac{7.886}{80.520/101} = 9.89^{**}$
$X_3 X_1, X_2$	1	1.057	1.057	$\frac{1.057}{79.463/100} = 1.33$
Regression $X_4 X_1, X_2, X_3$	1	0.020	0.020	$\frac{0.020}{79.443/99} = 0.025$
$X_5 X_1, X_2, X_3, X_4$	1	0.254	0.254	$\frac{0.254}{79.189/98} = 0.314$
$X_6 X_1, X_2, X_3, X_4, X_5$	1	2.844	2.844	$\frac{2.844}{76.345/97} = 3.613$
$X_7 X_1, X_2, X_3, X_4, X_5, X_6$	1	0.099	0.099	$\frac{0.099}{76.246/96} = 0.125$
Residual	96	76.246	0.794	
Total (corrected)	103	113.394		

Note: $F_{7, 96, 0.95} = 2.11$, $F_{1, 96, 0.95} = 3.95$, $F_{1, 102, 0.95} = 3.94$.

Use only the ANOVA table to answer the following questions.

- Determine the form of the separate fitted regression models of SBPSL (Y) on SBP1 (X_1), RW (X_3), and SBP1 \times RW (X_5) for each sex in terms of the estimated regression coefficients for the fitted regression model.
 - Why can't you use this ANOVA table to test whether the two regression equations are either parallel or coincident? Describe the appropriate testing procedure in each case.
11. For the data from Problem 2 in this chapter, address the following questions, using the information provided in the accompanying SAS output.
- State the regression model that incorporates the straight-line models for each group of countries.
 - Determine and plot the separate fitted straight lines, based on the fitted regression model given in part (a).
 - Test for parallelism of the two straight lines.
 - Test for coincidence of the two straight lines.
 - Compare your results in parts (b) through (d) to those obtained in Problem 2.

Edited SAS Output (PROC REG) for Problem 11

Regression of PCI on YNG, Z, and YNG \times Z

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2218614	739538	935.72	<.0001
Error	48	37936	790.33773		
Corrected Total	51	2256550			

Root MSE	28.11295	R-Square	0.9832
Dependent Mean	332.65385	Adj R-Sq	0.9821
Coeff Var	8.45111		

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	897.51861	51.39769	17.46	<.0001	5754246
YNG	1	-17.38853	1.24965	-13.91	<.0001	1089772
Z	1	1273.15148	71.01247	17.93	<.0001	970417
YNGZ	1	-24.61627	1.73867	-14.16	<.0001	158424

12. Using the information given here, answer the same questions as in Problem 11 about the regression of height (Y) on age (X) for children in one of two diet categories. (This problem is based on the data in Problem 3.)

Edited SAS Output (PROC REG) for Problem 12

Regression of HEIGHT on AGE, Z, and AGE \times Z

Where Z = 1 if Protein Rich, 0 Otherwise

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4174.20665	1391.40222	267.98	<.0001
Error	23	119.42001	5.19217		
Corrected Total	26	4293.62667			

Root MSE	2.27863	R-Square	0.9722
Dependent Mean	70.91111	Adj R-Sq	0.9686
Coeff Var	3.21337		

(continued)

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	51.22517	1.27112	40.30	<.0001	135766
AGE	1	8.68604	0.67621	12.85	<.0001	3053.34835
Z	1	-0.90148	1.86194	-0.48	0.6329	840.45239
AGEZ	1	7.32293	0.99647	7.35	<.0001	280.40592

13. In Gruber's (1970) study of $n = 104$ individuals (discussed in Problem 10), the relationship between blood pressure change (SBPSL) and relative weight (RW), controlling for initial blood pressure (SBP1), was compared for three different geographical backgrounds and for three different psychosocial orientations, using the following 15 variables:

$$Y = \text{SBPSL}$$

$$X_1 = \text{SBP1} \text{ (initial blood pressure)}$$

$$X_2 = R \text{ (1 if rural background, 0 if town, -1 if urban)}$$

$$X_3 = T \text{ (1 if town background, 0 if rural, -1 if urban)}$$

$$X_4 = TD \text{ (1 if traditional orientation, 0 if transitional, -1 if modern)}$$

$$X_5 = TN \text{ (1 if transitional orientation, 0 if traditional, -1 if modern)}$$

$$X_6 = RW \text{ (relative weight)}$$

$$X_7 = T \times TD$$

$$X_8 = T \times TN$$

$$X_9 = R \times TD$$

$$X_{10} = R \times TN$$

$$X_{11} = R \times TD \times RW$$

$$X_{12} = R \times TN \times RW$$

$$X_{13} = T \times TD \times RW$$

$$X_{14} = T \times TN \times RW$$

A standard stepwise regression program was run using these data, yielding the following ANOVA table (variables were forced to enter in the order presented) based on the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{14} X_{14} + E$$

- Using this regression model, determine the form of the nine fitted regression equations corresponding to the nine possible combinations of background with orientation (i.e., $R = 1$ and $TD = 1$, $R = 0$ and $TD = 1$, $R = -1$ and $TD = 1$, etc.). [Note: Each of the nine equations will be of the form $\hat{Y} = \hat{\beta}_0^* + \hat{\beta}_1^*(\text{SBP1}) + \hat{\beta}_2^*(\text{RW})$.]
- Test the null hypothesis that the nine regression equations determined in part (a) are parallel. State the null hypothesis in terms of the regression coefficients of the original 14-variable regression model.
- Test the hypothesis H_0 : "The three regression equations corresponding to the three backgrounds (rural, town, and urban) are parallel (but not necessarily coincident)" against the alternative H_A : "They are not parallel."

- d. Set up the multiple partial F -test formula for testing H_0 : "The nine regression equations dealt with in part (a) are coincident" against H_A : "They are not coincident." State the null hypothesis in terms of the coefficients in the regression equation.

	Source	d.f.	SS	F
Regression	X_1	1	24.9878	28.830
	$X_2 X_1$	1	0.5218	0.600
	$X_3 X_1, X_2$	1	0.0057	0.006
	$X_4 X_1, X_2, X_3$	1	1.0520	1.199
	$X_5 X_1, X_2, X_3, X_4$	1	1.1116	1.271
	$X_6 X_1 - X_5$	1	0.8321	0.951
	$X_7 X_1 - X_6$	1	0.2919	0.331
	$X_8 X_1 - X_7$	1	1.6601	1.902
	$X_9 X_1 - X_8$	1	0.5843	0.667
	$X_{10} X_1 - X_9$	1	0.2266	0.257
	$X_{11} X_1 - X_{10}$	1	1.1916	1.355
	$X_{12} X_1 - X_{11}$	1	2.0853	2.407
	$X_{13} X_1 - X_{12}$	1	1.5915	1.854
	$X_{14} X_1 - X_{13}$	1	0.0208	0.024
Residual		89	77.2303	
Total		103	113.3934	

14. The Environmental Protection Agency conducted an experiment to assess the characteristics of sampling procedures designed to measure the suspended particulate concentration (X) in a particular city. At each of two distinct locations (designated as location 1 and location 2), two identical sampling units were set up side by side, and readings were taken on each of 10 days. The data are given here in tabular form, where X_{ij1} and X_{ij2} are, respectively, the measured concentration for samplers 1 and 2 at location i on day j ($i = 1, 2$; $j = 1, 2, \dots, 10$). Researchers hypothesized that the inherent variation in the observations depends on the level of concentration being measured. To quantify this hypothesis, they proposed to fit a model of the form

$$|d_{ij}| = \beta_0 + \beta_1 Z + \beta_2(X_{ij1} + X_{ij2}) + \beta_3(X_{ij1} + X_{ij2})Z + E$$

where Z is 1 if the observation pertains to location 1 and is 0 otherwise and where $|d_{ij}| = |X_{ij1} - X_{ij2}|$.

Day	Location 1			Location 2		
	X_{1j1}	X_{1j2}	$d_{1j} = X_{1j1} - X_{1j2}$	X_{2j1}	X_{2j2}	$d_{2j} = X_{2j1} - X_{2j2}$
1	4	3	1	6	5	1
2	8	6	2	3	1	2
3	12	16	-4	1	2	-1
4	1	1	0	10	12	-2
5	7	6	1	17	17	0
6	11	8	3	4	7	-3
7	14	10	4	8	6	2
8	10	12	-2	12	12	0
9	2	2	0	10	9	1
10	15	20	-5	20	19	1

- Using the results provided in the following table, determine and interpret the fitted straight-line relationship between $|d_{ij}|$ and $(X_{ij1} + X_{ij2})$ at each location.
- Test for parallelism of the two lines.
- Test for coincidence of the two lines.
- How would you test whether level of concentration is significantly related to inherent variation in the observations for at least one of the two locations?

Edited SAS Output (PROC REG) for Problem 14

Regression of ABS (d) on X (= X1 + X2), Z, and X × Z

Where Z = 1 if Location 1, 0 Otherwise

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	31.33952	10.44651	19.87	<.0001
Error	16	8.41048	0.52565		
Corrected Total	19	39.75000			

Root MSE	0.72502	R-Square	0.7884
Dependent Mean	1.75000	Adj R-Sq	0.7487
Coeff Var	41.42974		

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	2.00856	0.43196	4.65	0.0003	61.25000
X	1	-0.03915	0.02023	-1.94	0.0708	4.83482
Z	1	-2.42792	0.61773	-3.93	0.0012	4.61660
XZ	1	0.19506	0.03023	6.45	<.0001	21.88810

- A biologist compared the effect of temperature for each of two media on the growth of human amniotic cells in a tissue culture. The data shown in the following table were obtained.
 - Assuming that a parabolic model is appropriate for describing the relationship between Y and X for each medium, provide a single regression model that incorporates two separate parabolic models, one corresponding to each medium.
 - Use the SAS output provided here to determine and plot the separate fitted parabolas for each medium. (Note: $Z = 0$ for medium A, and $Z = 1$ for medium B.)
 - Test for “parallelism” of the two parabolas.
 - Test for coincidence of the two parabolas.
 - Is it possible to test whether a quadratic term should be included in the model for each medium using only the provided SAS output? Explain.

Medium A				Medium B			
No. of Cells $\times 10^{-6}$	Temperature (°F)	No. of Cells $\times 10^{-6}$	Temperature (°F)	No. of Cells $\times 10^{-6}$	Temperature (°F)	No. of Cells $\times 10^{-6}$	Temperature (°F)
(Y)	(X)	(Y)	(X)	(Y)	(X)	(Y)	(X)
1.13	40	2.30	80	0.98	40	2.20	80
1.20	40	2.15	80	1.05	40	2.10	80
1.00	40	2.25	80	0.92	40	2.20	80
0.91	40	2.40	80	0.90	40	2.30	80
1.05	40	2.49	80	0.89	40	2.26	80
1.75	60	3.18	100	1.60	60	3.10	100
1.45	60	3.10	100	1.45	60	3.00	100
1.55	60	3.28	100	1.40	60	3.13	100
1.64	60	3.35	100	1.50	60	3.20	100
1.60	60	3.12	100	1.56	60	3.07	100

Edited SAS Output (PROC REG) for Problem 15

Regression of Y on X, X2, Z, ZX, and ZX2

Where Z = 1 if Medium B, 0 Otherwise

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	26.06864	5.21373	583.97	<.0001
Error	34	0.30356	0.00893		
Corrected Total	39	26.37220			

Root MSE	0.09449	R-Square	0.9885
Dependent Mean	1.99275	Adj R-Sq	0.9868
Coeff Var	4.74163		

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	0.494600	0.24256234	2.04	0.0493	158.842103
X	1	0.005370	0.00745505	0.72	0.4763	25.668613
X2	1	0.000217	0.00005282	4.12	0.0002	0.290702
Z	1	-0.143700	0.34303495	-0.42	0.6779	0.109202
ZX	1	0.001235	0.01054303	0.12	0.9074	0.0000005
ZX2	1	-0.00000875	0.00007470	-0.12	0.9074	0.000122

16. Answer the following questions using the accompanying SPSS output, which is based on data on the growth rates (Y) of depleted chicks at different (log) dosage levels (X) of vitamin B for males and females.

Edited SPSS Output for Problem 16

VARIABLE(S) ENTERED ON STEP NUMBER 1.. X				ANALYSIS OF VARIANCE				MEAN SQUARE			
MULTIPLE R	0.88583	REGRESSION	F	DF	SUM OF SQUARES	1041.33728	109.33735				
R SQUARE	0.78470	RESIDUAL		1.							
ADJUSTED R SQUARE	0.77752			30.							
STANDARD ERROR	3.08611										
VARIABLE(S) ENTERED ON STEP NUMBER 2.. Z (= 1 if MALE, = 0 if FEMALE)				ANALYSIS OF VARIANCE				MEAN SQUARE			
MULTIPLE R	0.88592	REGRESSION	F	DF	SUM OF SQUARES	1041.55874	52.89862				
R SQUARE	0.78486	RESIDUAL		2.							
ADJUSTED R SQUARE	0.77002			29.							
STANDARD ERROR	3.13765										
VARIABLE(S) ENTERED ON STEP NUMBER 3.. XZ				ANALYSIS OF VARIANCE				MEAN SQUARE			
MULTIPLE R	0.88878	REGRESSION	F	DF	SUM OF SQUARES	1048.28988	349.42996				
R SQUARE	0.78993	RESIDUAL		3.							
ADJUSTED R SQUARE	0.76743			28.							
STANDARD ERROR	3.15532										
VARIABLE(S) ENTERED ON STEP NUMBER 4.. XZ				ANALYSIS OF VARIANCE				MEAN SQUARE			
MULTIPLE R	0.81424	REGRESSION	F	DF	SUM OF SQUARES	1048.28988	349.42996				
R SQUARE	0.81424	RESIDUAL		1.							
ADJUSTED R SQUARE	-0.14670			30.							
STANDARD ERROR	0.19139										
VARIABLE(S) ENTERED ON STEP NUMBER 5.. XZ				ANALYSIS OF VARIANCE				MEAN SQUARE			
MULTIPLE R	0.81424	REGRESSION	F	DF	SUM OF SQUARES	1048.28988	349.42996				
R SQUARE	0.81424	RESIDUAL		1.							
ADJUSTED R SQUARE	-0.14670			30.							
STANDARD ERROR	0.19139										
VARIABLE(S) ENTERED ON STEP NUMBER 6.. XZ				ANALYSIS OF VARIANCE				MEAN SQUARE			
MULTIPLE R	0.73852	REGRESSION	F	DF	SUM OF SQUARES	1048.28988	349.42996				
R SQUARE	0.73852	RESIDUAL		2.							
ADJUSTED R SQUARE	0.71233			29.							
STANDARD ERROR	0.676										
VARIABLE(S) ENTERED ON STEP NUMBER 7.. XZ				ANALYSIS OF VARIANCE				MEAN SQUARE			
MULTIPLE R	0.73852	REGRESSION	F	DF	SUM OF SQUARES	1048.28988	349.42996				
R SQUARE	0.73852	RESIDUAL		2.							
ADJUSTED R SQUARE	0.71233			29.							
STANDARD ERROR	0.676										
VARIABLE(S) ENTERED ON STEP NUMBER 8.. XZ				ANALYSIS OF VARIANCE				MEAN SQUARE			
MULTIPLE R	0.71233	REGRESSION	F	DF	SUM OF SQUARES	1048.28988	349.42996				
R SQUARE	0.71233	RESIDUAL		3.							
ADJUSTED R SQUARE	0.676			28.							
STANDARD ERROR	0.676										

From Nie et al., *Statistical Package for the Social Sciences*. Copyright © 1975 by McGraw-Hill, Inc. Used with permission of McGraw-Hill Book Company and Dr. Norman Nie, President, SPSS Inc.

- a. Define a single multiple regression model that incorporates different straight-line models for males and females.
 - b. Plot the fitted straight lines for each sex on graph paper.
 - c. Test for parallelism.
 - d. Test for coincidence.
17. Market research was conducted for a national retail company to compare the relationship between sales and advertising during the warm spring and summer seasons as compared with the cool fall and winter seasons. The data shown in the following table were collected over a period of several years.

Season (Warm = 0, Cool = 1)	Advertising Expenditure (\$millions)	Sales Revenue (\$millions)	Season (Warm = 0, Cool = 1)	Advertising Expenditure (\$millions)	Sales Revenue (\$millions)
0	17.0	156.1	1	10.0	131.0
0	12.5	142.6	1	13.8	136.8
0	20.5	166.8	1	15.0	141.5
0	16.0	155.4	1	19.5	151.8
0	15.0	150.5	1	17.0	148.3
0	14.5	147.5	1	12.5	133.3
0	17.5	156.9	1	14.5	138.0
0	12.5	138.8	1	12.5	135.9
0	11.5	134.3	1	12.0	132.0

- a. Identify a single regression model that uses the data for both warm and cool seasons and that defines straight-line models relating sales revenue (Y) to advertising expenditure (X) for each season.
- b. Using the computer output given next, determine and plot the fitted straight lines for each season.
- c. Test whether the straight lines for cool and warm seasons coincide.
- d. Test H_0 : “The lines are parallel” versus H_A : “The lines are not parallel.”
- e. In light of your answers to parts (c) and (d), comment on differences and similarities in the sales–advertising expenditure relationship between cooler and warmer seasons.

Edited SAS Output (PROC REG) for Problem 17

Regression of Sales Revenue (X) on Adv. Expenditure (X), Season (Z), and XZ

DESCRIPTIVE STATISTICS					
Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	18.00000	1.00000	18.00000	0	0
X	263.80000	14.65556	4002.94000	8.04732	2.83678
Z	9.00000	0.50000	9.00000	0.26471	0.51450
XZ	126.80000	7.04444	1851.44000	56.36497	7.50766
Y	2597.50000	144.30556	376639	106.17938	10.30434

(continued)

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1755.87993	585.29331	166.65	<.0001
Error	14	49.16951	3.51211		
Corrected Total	17	1805.04944			

Root MSE	1.87406	R-Square	0.9728
Dependent Mean	144.30556	Adj R-Sq	0.9669
Coeff Var	1.29868		

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	96.83045	3.56516	27.16	<.0001	374834
X	1	3.48486	0.23058	15.11	<.0001	1461.75022
Z	1	7.17269	4.88170	1.47	0.1639	260.06703
XZ	1	-1.01978	0.32746	-3.11	0.0076	34.06268

18. A testing laboratory studies and compares the relationship between tire tread wear per 1,000 miles (Y) and average driving speed (X) for two competing tire types (denoted A and B). The data shown in the following table were collected for a random sample of 20 tires, 10 of A and 10 of B.

Tire	Tread Wear per 100 Miles of Travel (% of tread thickness)	Average Speed (mph)	Tire	Tread Wear per 100 Miles of Travel (% of tread thickness)	Average Speed (mph)
A	0.5	65	B	0.7	65
A	0.4	55	B	0.7	64
A	0.6	70	B	0.8	69
A	0.6	68	B	0.7	66
A	0.6	70	B	0.9	70
A	0.5	66	B	0.9	69
A	0.4	55	B	0.5	58
A	0.3	50	B	0.6	62
A	0.5	64	B	0.6	64
A	0.6	68	B	0.8	68

- a. Identify a single regression model that uses the data for both tire types and that defines straight-line models relating tread wear (Y) to average speed (X) for each tire.
- b. Using the computer output given next, determine and plot the fitted straight lines for each tire type.
- c. Test the null hypothesis that the straight lines for the two tire types coincide.

- d. Test H_0 : “The lines are parallel” versus H_A : “The lines are not parallel.”
e. In light of your answers to parts (c) and (d), comment on differences and similarities in the tread wear–average speed relationship for the two tire types.

Edited SAS Output (PROC REG) for Problem 18

Regression of Tread Wear (Y) on Avg. Speed (X), Tire (Z), and XZ
(Z = 0 for Tire A, Z = 1 for Tire B)

DESCRIPTIVE STATISTICS					
Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	20.00000	1.00000	20.00000	0	0
X	1286.00000	64.30000	83302	32.22105	5.67636
Z	10.00000	0.50000	10.00000	0.26316	0.51299
XZ	655.00000	32.75000	43027	1135.56579	33.69816
Y	12.20000	0.61000	7.94000	0.02621	0.16190

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.47861	0.15954	131.64	<.0001
Error	16	0.01939	0.00121		
Corrected Total	19	0.49800			

Root MSE	0.03481	R-Square	0.9611
Dependent Mean	0.61000	Adj R-Sq	0.9538
Coeff Var	5.70697		

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	-0.40752	0.10313	-3.95	0.0011	7.44200
X	1	0.01438	0.00163	8.85	<.0001	0.29506
Z	1	-1.08212	0.22917	-4.72	0.0002	0.14687
XZ	1	0.01935	0.00352	5.50	<.0001	0.03668

19. A random sample of data was collected on residential sales in a large city. The following table shows the sales price Y (in \$1,000s), area X_1 (in hundreds of square feet), number of bedrooms X_2 , total number of rooms X_3 , age X_4 (in years), and location (dummy variables Z_1 and Z_2 , defined as follows: $Z_1 = Z_2 = 0$ for intown; $Z_1 = 1$, $Z_2 = 0$ for inner suburbs; $Z_1 = 0$, $Z_2 = 1$ for outer suburbs) of each house.

House	Y	X_1	X_2	X_3	X_4	Z_1	Z_2
1	84.0	13.8	3	7	10	1	0
2	93.0	19.0	2	7	22	0	1
3	83.1	10.0	2	7	15	0	1
4	85.2	15.0	3	7	12	0	1
5	85.2	12.0	3	7	8	0	1
6	85.2	15.0	3	7	12	0	1
7	85.2	12.0	3	7	8	0	1
8	63.3	9.1	3	6	2	0	1
9	84.3	12.5	3	7	11	0	1
10	84.3	12.5	3	7	11	0	1
11	77.4	12.0	3	7	5	1	0
12	92.4	17.9	3	7	18	0	0
13	92.4	17.9	3	7	18	0	0
14	61.5	9.5	2	5	8	0	0
15	88.5	16.0	3	7	11	0	0
16	88.5	16.0	3	7	11	0	0
17	40.6	8.0	2	5	5	0	0
18	81.6	11.8	3	7	8	0	1
19	86.7	16.0	3	7	9	1	0
20	89.7	16.8	2	7	12	0	0
21	86.7	16.0	3	7	9	1	0
22	89.7	16.8	2	7	12	0	0
23	75.9	9.5	3	6	6	0	1
24	78.9	10.0	3	6	11	1	0
25	87.9	16.5	3	7	15	1	0
26	91.0	15.1	3	7	8	0	1
27	92.0	17.9	3	8	13	0	1
28	87.9	16.5	3	7	15	1	0
29	90.9	15.0	3	7	8	0	1
30	91.9	17.8	3	8	13	0	1

- a. Identify a single regression model that uses the data for all three locations and that defines straight-line models relating sales price (Y) to area (X_1) for each location.

Edited SAS Output (PROC REG) for Problem 19

Regression of Sales Price (Y) on Area (X_1), Location (Z), and XZ

DESCRIPTIVE STATISTICS					
Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	30.00000	1.00000	30.00000	0	0
X1	423.90000	14.13000	6276.55000	9.89114	3.14502
Z1	7.00000	0.23333	7.00000	0.18506	0.43018
Z2	15.00000	0.50000	15.00000	0.25862	0.50855
X1Z1	100.80000	3.36000	1490.94000	39.73283	6.30340
X1Z2	204.20000	6.80667	2914.06000	52.55651	7.24959
Y	2504.90000	83.49667	212705	122.55206	11.07032

(continued)

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	3158.41441	631.68288	38.32	<.0001
Error	24	395.59526	16.48314		
Corrected Total	29	3554.00967			

Root MSE	4.05994	R-Square	0.8887
Dependent Mean	83.49667	Adj R-Sq	0.8655
Coeff Var	4.86240		

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	8.96861	6.07754	1.48	0.1530	209151
X1	1	4.80699	0.39735	12.10	<.0001	2271.71350
Z1	1	52.12239	11.22484	4.64	0.0001	0.01631
Z2	1	48.55824	7.79710	6.23	<.0001	336.00295
X1Z1	1	-3.20121	0.75896	-4.22	0.0003	89.28227
X1Z2	1	-2.80309	0.52981	-5.29	<.0001	461.39938

- b. Using the computer output given next, determine and plot the fitted straight lines for each location.
- c. Test the null hypothesis that the straight lines for the three locations coincide.
- d. Test H_0 : "The lines are parallel" versus H_A : "The lines are not parallel."
- e. In light of your answers to parts (c) and (d), comment on differences and similarities in the sales price-area relationship for the three locations.
20. In Problem 19 in Chapter 5 and Problem 14 in Chapter 8, data from the 1990 Census for 26 randomly selected Metropolitan Statistical Areas (MSAs) were discussed. Of interest were factors potentially associated with the rate of owner occupancy of housing units. The following three variables were included in the data set:

OWNEROCC: Proportion of housing units that are owner-occupied (as opposed to renter-occupied)

OWNCOST: Median selected monthly ownership costs, in \$

URBAN: Proportion of population living in urban areas

It is also of interest to see whether the owner occupancy rate–ownership cost relationship is different in metropolitan areas where 75% or more of the population lives in urban areas compared to metropolitan areas where less than 75% of the population lives in urban areas. For this purpose, the following additional variable is defined:

$$X_1 = \begin{cases} 1 & \text{if proportion of population living in urban areas} \geq 0.75 \\ 0 & \text{otherwise} \end{cases}$$

- State a single regression model that defines straight-line models relating OWN-EROCC to OWNCOST both for MSAs with high percentages ($\geq 75\%$) of urban populations and for MSAs with lower percentages of urban populations ($< 75\%$).
- Using the computer output given next, test whether the straight lines for the two MSA types described in part (a) coincide. (Note: In the accompanying SAS output, the variable OWNCOST has been centered to avoid collinearity problems.)
- Test H_0 : "The lines are parallel" versus H_A : "The lines are not parallel."
- In light of your answers to parts (b) and (c), comment on differences and similarities in the owner occupancy rate–ownership cost relationship for the two types of MSA.

Edited SAS Output (PROC REG) for Problem 20

Regression of OWNEROCC on OWNCOST, X1, and X1*OWNCOST

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	214.19496	71.39832	4.06	0.0194
Error	22	386.76658	17.58030		
Corrected Total	25	600.96154			

Regression of OWNEROCC on OWNCOST and X1

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	213.49860	106.74930	6.34	0.0064
Error	23	387.46294	16.84621		
Corrected Total	25	600.96154			

[portion of output omitted]

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	68.66546	1.47336	46.60	<.0001
OWNCOST	1	-0.01267	0.00553	-2.29	0.0314
X1	1	-3.90567	1.78250	-2.19	0.0388

- This problem involves the PERK study data discussed in Problem 12 in Chapter 8. Suppose that we want to compare the relationship between change in refraction five

years after surgery and baseline refractive error, for males and females. To this end, we define the following dummy variable:

$$Z = \begin{cases} 1 & \text{if patient is male} \\ 0 & \text{otherwise} \end{cases}$$

- a. State a single regression model that defines straight-line models relating change in refraction five years after surgery and baseline refractive error for both males and females.
- b. Using the computer output given next, test whether the straight lines for males and females coincide.
- c. Test H_0 : “The lines are parallel” versus H_A : “The lines are not parallel.”
- d. In light of your answers to parts (b) and (c), comment on differences and similarities in the change in refraction–baseline refractive error relationship for males and females.

Edited SAS Output (PROC REG) for Problem 21

Regression of Y on X1, Z, and X1Z

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	19.65170	6.55057	5.40	0.0027
Error	50	60.60124	1.21202		
Corrected Total	53	80.25294			

Root MSE	1.10092	R-Square	0.2449
Dependent Mean	3.83343	Adj R-Sq	0.1996
Coeff Var	28.71896		

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	3.17821	0.46488	6.84	<.0001
X1	1	-0.20101	0.10786	-1.86	0.0683
Z	1	-1.99513	0.88972	-2.24	0.0294
X1Z	1	-0.38383	0.20266	-1.89	0.0640

Regression of Y on X1 and Z

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	15.30414	7.65207	6.01	0.0045
Error	51	64.94880	1.27351		
Corrected Total	53	80.25294			

(continued)

Root MSE	1.12850	R-Square	0.1907
Dependent Mean	3.83343	Adj R-Sq	0.1590
Coeff Var	29.43835		

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	2.75265	0.41717	6.60	<.0001
X1	1	-0.30973	0.09360	-3.31	0.0017
Z	1	-0.41288	0.31372	-1.32	0.1940

22. Among the BRFSS participants, the linear relationship between body-mass index (BMI) BMI and drinking frequency is to be compared between those who reported any exercise in the previous month (Exercise = 1) and those who report no exercise (Exercise = 0). Output is provided for separate straight-line equations among exercisers and non-exercisers, as well as for single models that include exercise, represented as a dummy variable.
- Using method I (comparison of two lines), do you believe the lines are coincident? (*Hint:* Only one hypothesis test is necessary.)
 - Using method II (single model), do you believe the lines are coincident?
 - Do the tests of coincidence agree? Based on the magnitude of the results, is this expected?
 - Using either method, test whether the lines are parallel.
 - Using either method, test whether the lines have the same intercept.
 - Comment on the validity of the homogeneous variance assumption for these data.

Edited SAS Output (PROC REG) for Problem 22

Regression of BMI on DRINK_DAYS – Exercise = YES

DESCRIPTIVE STATISTICS					
Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	844.00000	1.00000	844.00000	0	0
drink_days	5338.28571	6.32498	75919	50.00538	7.07145
bmi	22236	26.34634	611377	30.28665	5.50333

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	921.17702	921.17702	31.52	<.0001
Error	842	24610	29.22859		
Corrected Total	843	25532			

(continued)

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	27.28133	0.24974	109.24	<.0001
drink_days	1	-0.14783	0.02633	-5.61	<.0001

Regression of BMI on DRINK_DAYS – Exercise = NO

DESCRIPTIVE STATISTICS					
Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	212.00000	1.00000	212.00000	0	0
drink_days	1266.00000	5.97170	17861	48.81881	6.98705
bmi	6105.48000	28.79943	186836	52.14243	7.22097

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	211.68169	211.68169	4.12	0.0436
Error	210	10790	51.38272		
Corrected Total	211	11002			

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	29.65549	0.64827	45.75	<.0001
drink_days	1	-0.14335	0.07063	-2.03	0.0436

Regression of BMI on DRINK_DAYS, EXERCISE

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2152.32442	1076.16221	32.01	<.0001
Error	1053	35401	33.61919		
Corrected Total	1055	37553			

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	29.67696	0.42595	69.67	<.0001
drink_days	1	-0.14695	0.02532	-5.80	<.0001
exercise	1	-2.40118	0.44553	-5.39	<.0001

(continued)

Regression of BMI on DRINK_DAYS, EXERCISE, DRINK_DAYS*EXERCISE

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2152.49001	717.49667	21.32	<.0001
Error	1052	35401	33.665099		
Corrected Total	1055	37553			

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	29.65549	0.52462	56.53	<.0001
drink_days	1	-0.14335	0.05716	-2.51	0.0123
exercise	1	-2.37416	0.58910	-4.03	<.0001
drink_days_exercise	1	-0.00447	0.06376	-0.07	0.9441

References

- Allen, D. M., and Grizzle, J. E. 1969. "Analysis of Growth and Dose Response Curves." *Biometrics* 25: 357–82.
- Armitage, P. 1971. *Statistical Methods in Medical Research*. Oxford: Blackwell Scientific Publications.
- Clark, M. E.; Lechyeka, M.; and Cook, C. A. 1940. "The Biological Assay of Riboflavin." *Journal of Nutrition* 20: 133–44.
- Davis, E. A., Jr. 1955. "Seasonal Changes in the Energy Balance of the English Sparrow." *Auk.*, 72(4): 385–411.
- Gruber, F. J. 1970. "Industrialization and Health." Ph.D. dissertation, Department of Epidemiology, University of North Carolina, Chapel Hill, N.C.
- Nie, N., et al. 1975. *Statistical Package for the Social Sciences*. New York: McGraw-Hill.

13

Analysis of Covariance and Other Methods for Adjusting Continuous Data

13.1 Preview

In Chapter 11, we discussed the issue of controlling for certain variables when assessing an association of interest. Three reasons for considering control are to assess *interaction*, to correct for *confounding*, and to increase the *precision* in estimating the association of interest. In regression, the usual approach for carrying out such control is to fit a regression model that contains as independent variables not only the study factors (exposure variables) of interest but also control variables considered to be important (and perhaps even product terms involving these variables). The focus then becomes one of determining the effects of the study factors on the response variable *adjusted* for the presence of the control variables in the model.

In this chapter, we describe how to carry out this process of adjustment by using a popular procedure for regression modeling called analysis of covariance (ANACOVA). This technique involves a multiple regression model in which the study factors of interest are all treated as nominal variables, whereas the variables being controlled—that is, the *covariates*—may be measurements on any measurement scale. As discussed in Chapter 12, nominal variables are incorporated into regression models by means of dummy variables. Thus, the general ANACOVA model usually contains a mixture of dummy variables and other types of variables, and the dependent variable is considered continuous. In Chapter 12, the main focus was on comparing multiple linear regression models involving continuous predictors across levels of nominal covariates. In ANACOVA, the same general types of regression models are considered from a different perspective: the nominal covariates are now the main predictors of interest and the (primarily) continuous covariates are included in the models for control purposes (i.e., for validity and precision considerations). In using the ANACOVA model, we also assume that there is no interaction of covariates with study variables, although (as discussed later) this assumption should be assessed in the analysis.

In addition to considering ANACOVA, we briefly review other regression-type methods for covariate control.

13.2 Adjustment Problem

Suppose, as in the example in Section 12.4, that we are considering a sample of observations made on the dependent variable systolic blood pressure and the independent variable age for $n_F = 29$ women and $n_M = 40$ men.

Two questions are often of interest in analyses of such data:

1. Is the true straight-line relationship between blood pressure and age (given that a straight-line model is adequate) the same for male and female populations?
2. Do the mean blood pressure levels for the male and female groups differ significantly from one another after taking into account (i.e., after one adjusts for or controls for) the possible confounding effect of there being differing age distributions in the two groups?

Although the statistical techniques required to answer these questions are related, the questions nevertheless differ in emphasis: the first focuses on a comparison of straight-line regression equations, whereas the second focuses on a comparison of the mean blood pressure levels in the two groups.

We have already considered question 1 (in Chapter 12) through use of the model

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + E \quad (13.1)$$

where Z is a dummy variable identifying the sex group ($Z = 1$ if female and $Z = 0$ if male). By using appropriate tests of hypotheses about the parameters in this model, we may reach one of three important conclusions regarding question 1:

1. The lines are coincident (i.e., $\beta_2 = \beta_3 = 0$).
2. The lines are parallel but not coincident (i.e., $\beta_3 = 0$, but $\beta_2 \neq 0$).
3. The lines are not parallel ($\beta_3 \neq 0$).

These conclusions greatly influence the answer to question 2. If conclusion 1 is appropriate, we say that the two sex groups do not differ in mean blood pressure level after the effect of age is controlled. If conclusion 2 holds, we say that the sex group associated with the higher straight line has a higher mean blood pressure level at all ages. If conclusion 3 is valid, we must look closer at the orientation of the two straight lines: if they do not intersect in the age range of interest, we say that the sex group associated with the higher curve has a higher mean blood pressure level at each age; if they do intersect in the age range of interest, we say that one sex group has a higher mean blood pressure level than the other group at lower ages and a lower mean blood pressure level at higher ages (i.e., there is an age–sex group interaction effect).

Thus, by considering question 1 as described, we may draw reasonable inferences about the relationship between the true average blood pressure levels in the two groups as a function of age. Nevertheless, we must take additional statistical considerations into account to estimate the true adjusted mean difference and the adjusted means for each group. In this regard, question 2 raises the problem of determining an appropriate method for adjusting the sample mean blood pressure levels to take into account the effect of age, as well as the problem of providing a statistical test to compare these adjusted means.

In the example we are considering, age is a factor known to be strongly associated with blood pressure, and the two groups, as sampled, may have widely different age distributions. Without adjusting the sample mean values to reflect any difference in the age distributions in the two groups, we could not determine (e.g., through the use of a two-sample t test based on the unadjusted sample means) whether a significant difference was primarily due to age (i.e., confounding). With adjustment, however, we could determine whether any findings were driven by, for example, the fact that the females in the sample were older than the males (or vice versa).

13.3 Analysis of Covariance

The usual statistical technique for handling the adjustment problem described in Section 13.2 is called the *analysis of covariance*. In this approach, we fit a regression model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + E \quad (13.2)$$

where X (age) is referred to as the *covariate* and Z is a dummy variable that indexes the two groups to be compared ($Z = 1$ if female, $Z = 0$ if male). This model assumes—in contrast to model (13.1), which contains a $\beta_3 XZ$ term—that the regression lines for males and females are parallel.

Under this model, the *adjusted means*¹ for males and females are defined to be the predicted values obtained by evaluating the model at $Z = 0$ and $Z = 1$ when X is set equal to the overall mean age for the two groups. A partial F test of the hypothesis $H_0: \beta_2 = 0$ is then used to determine whether these adjusted means are significantly different.

In computing these adjusted means, we need to consider the two straight lines obtained by fitting model (13.2):

$$\begin{aligned} \text{Males } (Z = 0): \quad \hat{Y}_M &= \hat{\beta}_0 + \hat{\beta}_1 X \\ \text{Females } (Z = 1): \quad \hat{Y}_F &= (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 X \end{aligned} \quad (13.3)$$

Explicit formulas for the estimated coefficients in (13.3) are

$$\begin{aligned} \hat{\beta}_1 &= \frac{(n_M - 1)S_{X_M}^2 \hat{\beta}_{1M} + (n_F - 1)S_{X_F}^2 \hat{\beta}_{1F}}{(n_M - 1)S_{X_M}^2 + (n_F - 1)S_{X_F}^2} \\ \hat{\beta}_0 &= \bar{Y}_M - \hat{\beta}_1 \bar{X}_M \\ \hat{\beta}_0 + \hat{\beta}_2 &= \bar{Y}_F - \hat{\beta}_1 \bar{X}_F \end{aligned} \quad (13.4)$$

where $\hat{\beta}_{1M}$ and $\hat{\beta}_{1F}$ are the estimated slopes based on separate straight-line fits for males and females, $S_{X_M}^2$ and $S_{X_F}^2$ are the sample variances (of X) for males and females, \bar{Y}_M and \bar{Y}_F are the

¹ Other common names for the adjusted means are *estimated marginal means*, *least-square means (LS means)*, and *adjusted mean scores*.

mean blood pressures for the male and female samples, and \bar{X}_M and \bar{X}_F are the mean ages for the male and female samples, respectively. Notice that $\hat{\beta}_1$ is a weighted average of the slopes $\hat{\beta}_{1M}$ and $\hat{\beta}_{1F}$, which are estimated separately from the male and female data sets.

Based on (13.3) and (13.4), two alternative formulas for computing adjusted means can be used:

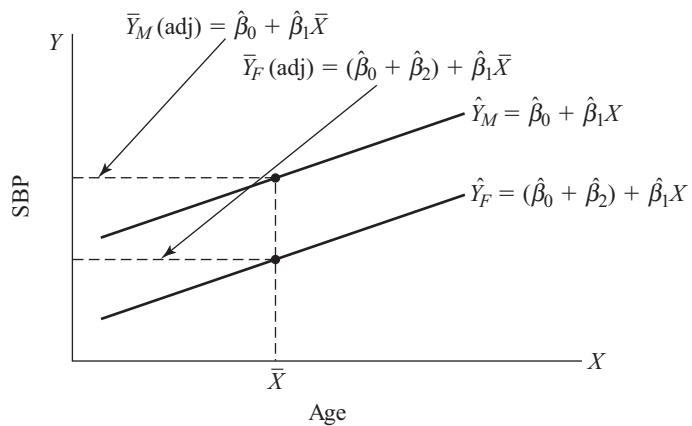
Sex	Z	Adjusted Mean	Formula 1	Formula 2	(13.5)
Male	0	$\bar{Y}_M(\text{adj})$	$\hat{\beta}_0 + \hat{\beta}_1\bar{X}$	$\bar{Y}_M - \hat{\beta}_1(\bar{X}_M - \bar{X})$	
Female	1	$\bar{Y}_F(\text{adj})$	$(\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1\bar{X}$	$\bar{Y}_F - \hat{\beta}_1(\bar{X}_F - \bar{X})$	

In this table, \bar{X} is the overall mean age for the combined data on males and females:

$$\bar{X} = \frac{n_M \bar{X}_M + n_F \bar{X}_F}{n_M + n_F}$$

Formula 1 is useful when model (13.2) has been estimated directly using a standard multiple regression program and corresponds to estimating linear functions (see Section 9.6), with $c_i = 0, 1$, or \bar{X} as appropriate. Formula 2, on the other hand, does not require the use of multiple regression procedures, although separate straight lines must be fitted to the male and female data sets.

Given that the parallel straight-line assumption of model (13.2) is appropriate (we discuss this assumption in Sections 13.4 and 13.6.2), the two formulas provide a comparison of the mean blood pressure levels for the two sex groups as if they both had the same age distribution. In this regard, the covariance approach just described attempts to artificially equate the age distributions by treating both sex groups as if they had the same mean age, the best estimate of which is \bar{X} . The adjusted means, then, represent the predicted \hat{Y} -values for each fitted line at \bar{X} , the assumed common mean age. This is depicted graphically in Figure 13.1.



©Cengage Learning

FIGURE 13.1 Adjusted systolic blood pressure (SBP) means for males and females, controlling for age by using analysis of covariance

That the partial F test of $H_0: \beta_2 = 0$ addresses the question of whether a significant difference exists between the adjusted means follows because, from formula 1 in (13.5), the difference in the two adjusted means is exactly equal to $\hat{\beta}_2$; that is,

$$\hat{\beta}_2 = \bar{Y}_F(\text{adj}) - \bar{Y}_M(\text{adj}) = [(\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 \bar{X}] - (\hat{\beta}_0 + \hat{\beta}_1 \bar{X})$$

■ **Example 13.1** The least-squares fitting of the model (13.2) for the male–female data we have been discussing yields the following estimated model:

$$\hat{Y} = 110.29 + 0.96X - 13.51Z$$

The separate fitted equations for males and females, respectively, are

$$\text{Males } (Z = 0): \quad \hat{Y}_M = 110.29 + 0.96X$$

$$\text{Females } (Z = 1): \quad \hat{Y}_F = 96.78 + 0.96X$$

The adjusted means obtained from these equations, using formula 1 of (13.5) with $\bar{X} = 46.14$, are

$$\bar{Y}_M(\text{adj}) = 110.29 + 0.96(46.14) = 154.58$$

$$\bar{Y}_F(\text{adj}) = 96.78 + 0.96(46.14) = 141.07$$

A comparison of these adjusted means with the unadjusted means yields the following statistics:

Sex	Unadjusted \bar{Y}	Adjusted
Male	155.15	154.58
Female	139.86	141.07

Notice that the adjusted mean for males is slightly lower than the unadjusted mean for males, whereas the female adjusted mean is slightly higher than its unadjusted counterpart. The direction of these changes accurately reflects the fact that, in this sample, the males are somewhat older on the average ($\bar{X}_M = 46.93$) than the females ($\bar{X}_F = 45.07$). Using the adjusted means, in effect, removes the influence of age on the comparison of mean blood pressures by considering what the mean blood pressures in the two groups would be if both groups had the same mean age ($\bar{X} = 46.14$).

But whether adjusted or not, the mean blood pressure for males in this example appears to be considerably higher than that for females. In fact, the covariance adjustment has done little to change this impression: the discrepancy between the male and female groups is 15.29 using unadjusted means and 13.51 using adjusted means. To test whether this difference in adjusted means is significant, we use the partial F test of the hypothesis $H_0: \beta_2 = 0$ based on model (13.2), which can be computed from the following analysis-of-variance presentation:

Source	SS	d.f.	MS
Reduced model ($\beta_2 = 0$) $\begin{cases} \text{Regression } (X) \\ \text{Residual} \end{cases}$	14,951.25	1	14,951.25
	8,260.51	67	123.29
Complete model (13.2) $\begin{cases} \text{Regression } (X, Z) \\ \text{Residual} \end{cases}$	18,009.78	2	9,004.89
	5,201.99	66	78.82

From this presentation, we can obtain the appropriate partial F statistic as follows:

$$\begin{aligned} F(Z|X) &= \frac{\text{Regression SS}(X, Z) - \text{Regression SS}(X)}{\text{MS Residual } (X, Z)} \\ &= \frac{18,009.78 - 14,951.25}{78.82} \\ &= 38.80 \end{aligned}$$

which has 1 and 66 degrees of freedom. The P -value for this test satisfies $P < .001$, so we reject H_0 and conclude that the two adjusted means differ significantly. ■

We may also be interested in confidence intervals based on these adjusted means. First, we might want to compute a confidence interval for the population difference between the male and female adjusted means. This would correspond to a confidence interval for β_2 and would be computed using the t -distribution methods discussed in Section 9.6.2. Alternatively, we may wish to report separate confidence intervals for the population male and female adjusted means. Since each of these adjusted mean estimates is a linear sum, methods described in Section 9.6.5 can be used to construct such confidence intervals.

13.4 Assumption of Parallelism: A Potential Drawback

A potential problem raised by using ANACOVA involves the assumption of parallelism of the regression lines. In certain applications, these regression lines may have different slopes. In such cases, the parallelism assumption is invalid, and the covariance method of adjustment just described should be avoided. *To guard against applying the covariance method of adjustment incorrectly, we recommend conducting a test for parallelism before proceeding with ANACOVA.* This amounts to testing $H_0: \beta_3 = 0$ for the complete model (13.1). We saw in Chapter 12 that the parallelism hypothesis ($H_0: \beta_3 = 0$) is not rejected for the age–systolic blood pressure data. This result supports the use of the ANACOVA model for these data (given, of course, that the assumption of variance homogeneity also holds).

If the test for parallelism supports the conclusion that the regression lines are not truly parallel, what, if anything, should be done about adjustment? Usually, no adjustment at all should be made to the sample means, since any such adjustment would be misleading; that is, a direct comparison of means is not appropriate when the true difference between the mean blood pressure levels in the two groups varies with age (i.e., when there is an age–sex interaction). In this case, since the main feature of the data is that the two regression lines describe very different relationships between age and blood pressure, an analysis that allows two separate regression lines to be fitted (without assuming parallelism) and that quantifies how the lines differ is sufficient.

13.5 Analysis of Covariance: Several Groups and Several Covariates

In the example discussed in previous sections, we compared two groups and adjusted for the single covariate, age. In general, ANACOVA may be used to provide adjusted means when there are several (say, s) groups and when it is necessary to adjust simultaneously for several (say, q) covariates. The regression model describing this general situation is written

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_q X_q + \beta_{q+1} Z_1 + \beta_{q+2} Z_2 + \dots + \beta_{q+s-1} Z_{s-1} + E, \quad (13.6)$$

where $q \geq 1$ and $s \geq 2$.

This model includes q covariates X_1, X_2, \dots, X_q and s groups, which are represented by the $s - 1$ dummy variables Z_1, Z_2, \dots, Z_{s-1} . As discussed in Chapter 12, we have some leeway in defining these dummy variables, but for our purposes we assume that the Z 's are defined as follows:

$$Z_j = \begin{cases} 1 & \text{if group } j \\ 0 & \text{otherwise} \end{cases} \quad j = 1, 2, \dots, s - 1$$

The fitted regression equations for the s different groups are then determined by specifying the appropriate combinations of values for the Z 's. These are

Group 1 ($Z_1 = 1$, other $Z_j = 0$):	$\hat{Y}_1 = (\hat{\beta}_0 + \hat{\beta}_{q+1}) + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_q X_q$
Group 2 ($Z_2 = 1$, other $Z_j = 0$):	$\hat{Y}_2 = (\hat{\beta}_0 + \hat{\beta}_{q+2}) + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_q X_q$
⋮	⋮
Group $s - 1$ ($Z_{s-1} = 1$, other $Z_j = 0$):	$\hat{Y}_{s-1} = (\hat{\beta}_0 + \hat{\beta}_{q+s-1}) + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_q X_q$
Group s (all $Z_j = 0$):	$\hat{Y}_s = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_q X_q$

(13.7)

From (13.7), we see that the corresponding coefficients of the covariates X_1, X_2, \dots, X_q in each of the s equations are identical. Thus, these regression equations represent “parallel” hypersurfaces in $(q + 1)$ dimensions, which is a natural generalization of the situation for the single covariate case. The adjusted mean for a particular group is then computed as

the predicted Y -value obtained by evaluating the fitted equation for that group at the mean values $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_q$ of the q covariates, based on the combined data for all s groups:

$$\begin{aligned}\bar{Y}_j(\text{adj}) &= (\hat{\beta}_0 + \hat{\beta}_{q+j}) + \hat{\beta}_1\bar{X}_1 + \dots + \hat{\beta}_q\bar{X}_q \quad j = 1, 2, \dots, s-1 \\ \bar{Y}_s(\text{adj}) &= \hat{\beta}_0 + \hat{\beta}_1\bar{X}_1 + \dots + \hat{\beta}_q\bar{X}_q\end{aligned}\quad (13.8)$$

To determine whether the s adjusted means differ significantly from one another, we test the null hypothesis

$$H_0: \beta_{q+1} = \beta_{q+2} = \dots = \beta_{q+s-1} = 0$$

using a multiple partial F test with $s-1$ and $n-q-s$ degrees of freedom, based on model (13.6). If H_0 is rejected, we conclude that significant differences exist among the adjusted means (although we cannot, without further inspection, determine where the major differences are).

Example 13.2 In the Ponape study (Patrick et al. 1974) of the effect of rapid cultural change on health status, one research goal was to determine whether blood pressure was associated with a measure of the strength of a Ponapean male's prestige in the modern (i.e., western) part of his culture relative to the traditional culture. A measure of prestige discrepancy (PD) was developed and then measured by questionnaire for each of 550 Ponapean males. Of particular interest was whether a higher prestige discrepancy score corresponded to higher blood pressure. The effects of the covariates age and body size were adjusted or controlled for when considering these questions.

To perform this analysis, the researchers categorized PD into three groups:

Group 1: Modern prestige much higher than traditional prestige

Group 2: Modern prestige not much different from traditional prestige

Group 3: Traditional prestige much higher than modern prestige

Then an ANACOVA was carried out with diastolic blood pressure (DBP) as the dependent variable and with AGE and quetelet index (QUET, a measure of body size) as the two covariates. An analysis of covariance model in this case is given by

$$\text{DBP} = \beta_0 + \beta_1\text{AGE} + \beta_2\text{QUET} + \beta_3Z_1 + \beta_4Z_2 + E \quad (13.9)$$

where

$$Z_1 = \begin{cases} 1 & \text{if group 1} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad Z_2 = \begin{cases} 1 & \text{if group 2} \\ 0 & \text{otherwise} \end{cases}$$

A test of the parallelism assumption implicit in model (13.9) considers the null hypothesis $H_0: \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$ in the full model

$$\begin{aligned}\text{DBP} &= \beta_0 + \beta_1\text{AGE} + \beta_2\text{QUET} + \beta_3Z_1 + \beta_4Z_2 + \beta_5Z_1\text{AGE} + \beta_6Z_2\text{AGE} \\ &\quad + \beta_7Z_1\text{QUET} + \beta_8Z_2\text{QUET} + E\end{aligned}$$

TABLE 13.1 ANACOVA example using Ponape study data

DBP	PD Group 1 ($n_1 = 87$)	PD Group 2 ($n_2 = 383$)	PD Group 3 ($n_3 = 80$)	P-value for F Test of $H_0: \beta_3 = \beta_4 = 0^*$
Unadjusted	71.68	68.16	68.55	.001 < $P < .005$
Adjusted	72.07	68.02	68.80	

* Using model (13.9) with 2 and 545 d.f.

© Cengage Learning

This multiple partial F test (with 4 and 541 degrees of freedom) was not rejected using the Ponape data, thus supporting the use of model (13.9).

The researchers then determined adjusted DBP means for the three groups by substituting the values of the overall means AGE and QUET into the fitted equations for the three groups, as follows:

$$\overline{\text{DBP}}_1(\text{adj}) = (\hat{\beta}_0 + \hat{\beta}_3) + \hat{\beta}_1 \overline{\text{AGE}} + \hat{\beta}_2 \overline{\text{QUET}}$$

$$\overline{\text{DBP}}_2(\text{adj}) = (\hat{\beta}_0 + \hat{\beta}_4) + \hat{\beta}_1 \overline{\text{AGE}} + \hat{\beta}_2 \overline{\text{QUET}}$$

$$\overline{\text{DBP}}_3(\text{adj}) = \hat{\beta}_0 + \hat{\beta}_1 \overline{\text{AGE}} + \hat{\beta}_2 \overline{\text{QUET}}$$

The test for equality of these adjusted means was based on use of a multiple partial F test of the null hypothesis $H_0: \beta_3 = \beta_4 = 0$ under model (13.9). Table 13.1 summarizes these calculations.

The results indicate the presence of highly significant differences among the adjusted blood pressure means, with group 1 having a somewhat higher adjusted mean blood pressure than the other two groups. Despite the statistical significance found, however, the adjusted mean blood pressure of 72.07 for group 1 is close enough (clinically speaking) to the adjusted means for the other groups to cast doubt on the clinical significance of these results. ■

13.6 Analysis of Covariance: Several Nominal Independent Variables

Continuing Example 1.4, involving female alcohol consumers in the BRFSS, suppose now that a different research team is primarily interested in the effects of exercise and tobacco use on body-mass index (BMI). Each of these new predictor variables is nominal with two levels: whether or not the respondent reported any exercise in the previous month (exercise) [1 = yes, 0 = no] and whether or not the respondent currently uses any tobacco products (tobacco_now) [1 = yes, 0 = no]. The researchers suspect that aspects of alcohol consumption and sleep may act as confounders of the relationships of interest. They additionally have learned from the previous researchers' findings (see the end of Chapter 9) that drinking frequency and sleep quality are significant predictors of BMI, whereas age is not, so

they would like to include these factors in regression models to increase the precision in predicting BMI.

This new analysis scenario may be conceptualized as an ANACOVA model, since nominal exposure variables are the key predictors of interest. Unlike the previous example, where several dummy variables together delineate the categories of a single underlying categorical variable, this model has two different nominal predictors considered simultaneously:

$$Y = \beta_0 + \beta_1(\text{drink_days}) + \beta_2(\text{poor_sleep_days}) + \beta_3(\text{exercise}) \\ + \beta_4(\text{tobacco_now}) + E$$

One approach to computing adjusted means using this model is to consider the four combinations of 0 and 1 for the two exposure variables, exercise and tobacco_now. Using the estimated regression coefficients and the sample means for the continuous control variables, we then compute adjusted means corresponding to these four combinations as shown here:

Variable	Mean	N
drink_days	6.177	1048
poor_sleep_days	8.783	1048

Dependent Variable: BMI

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	29.53802443	0.47751631	61.86	<.0001
drink_days	-0.14276859	0.02579711	-5.53	<.0001
poor_sleep_days	0.04027718	0.01820454	2.21	0.0271
exercise	-2.46982297	0.44928929	-5.50	<.0001
tobacco_now	-1.10760806	0.49204750	-2.25	0.0246

Adjusted mean (exercise = 1, tobacco_now = 1):

$$\hat{Y}_{1,1} = \hat{\beta}_0 + \hat{\beta}_1(\overline{\text{drink_days}}) + \hat{\beta}_2(\overline{\text{poor_sleep_days}}) + \hat{\beta}_3(1) + \hat{\beta}_4(1) \\ = (\hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_4) + \hat{\beta}_1(\overline{\text{drink_days}}) + \hat{\beta}_2(\overline{\text{poor_sleep_days}}) \\ = (29.54 - 2.47 - 1.11) - 0.14(6.18) + 0.04(8.78) = 25.45$$

Adjusted mean (exercise = 1, tobacco_now = 0):

$$\hat{Y}_{1,0} = \hat{\beta}_0 + \hat{\beta}_1(\overline{\text{drink_days}}) + \hat{\beta}_2(\overline{\text{poor_sleep_days}}) + \hat{\beta}_3(1) + \hat{\beta}_4(0) \\ = (\hat{\beta}_0 + \hat{\beta}_3) + \hat{\beta}_1(\overline{\text{drink_days}}) + \hat{\beta}_2(\overline{\text{poor_sleep_days}}) \\ = (29.54 - 2.47) - 0.14(6.18) + 0.04(8.78) = 26.56$$

Adjusted mean (exercise = 0, tobacco_now = 1):

$$\begin{aligned}\hat{Y}_{0,1} &= \hat{\beta}_0 + \hat{\beta}_1(\overline{\text{drink_days}}) + \hat{\beta}_2(\overline{\text{poor_sleep_days}}) + \hat{\beta}_3(0) + \hat{\beta}_4(1) \\ &= (\hat{\beta}_0 + \hat{\beta}_4) + \hat{\beta}_1(\overline{\text{drink_days}}) + \hat{\beta}_2(\overline{\text{poor_sleep_days}}) \\ &= (29.54 - 1.11) - 0.14(6.18) + 0.04(8.78) = 27.92\end{aligned}$$

Adjusted mean (exercise = 0, tobacco_now = 0):

$$\begin{aligned}\hat{Y}_{0,0} &= \hat{\beta}_0 + \hat{\beta}_1(\overline{\text{drink_days}}) + \hat{\beta}_2(\overline{\text{poor_sleep_days}}) + \hat{\beta}_3(0) + \hat{\beta}_4(0) \\ &= \hat{\beta}_0 + \hat{\beta}_1(\overline{\text{drink_days}}) + \hat{\beta}_2(\overline{\text{poor_sleep_days}}) \\ &= 29.54 - 0.14(6.18) + 0.04(8.78) = 29.03\end{aligned}$$

The unadjusted and adjusted means for the BRFSS data are arranged in the table below. Notice that mean BMI is lower among those who exercise and/or use tobacco; these associations are explored further in Problem 17 in this chapter. Aside from the method presented here, there are other options for computing the adjusted means when several nominal variables are considered simultaneously; some of these options are discussed in Problem 18 in this chapter.

Exercise	Tobacco_now	Unadjusted Mean	Adjusted Mean
Yes (1)	Yes (1)	25.99	25.45
Yes (1)	No (0)	26.42	26.56
No (0)	Yes (1)	27.35	27.92
No (0)	No (0)	29.29	29.03

13.7 Comments and Cautions

13.7.1 Rationale for Adjustment

ANACOVA adjusts for disparities in covariate distributions over groups by artificially assuming that all groups have the same set of mean covariate values. For example, if age and weight are the covariates and two groups are being compared, the ANACOVA adjustment procedure treats both groups as if they had the same mean age and the same mean weight.

The ANACOVA adjustment procedure is equivalent to assuming a common covariate *distribution* based on the combined sample over all groups. That is, not only are the means assumed to be equal but also the entire distribution of the covariates in the combined sample is assumed to be the same as the distribution of the covariates in each group. This process is often referred to as *standardization*, and correspondingly the common covariate distribution is referred to as the *standard distribution/population*. The adjusted mean for any group can be expressed as the average over the combined sample of the predicted means for that group; that is,

$$\hat{Y}_j(\text{adj}) = \frac{1}{n} \sum_{i=1}^n \hat{Y}_{ij} \quad j = 1, 2, \dots, s$$

where $\hat{Y}_j(\text{adj})$ is the adjusted mean for group j defined by (13.8), n is the number of subjects in the combined sample, and \hat{Y}_{ij} is the predicted response in the j th group, based on the set

of equations (13.7) and the covariate values of the i th individual in the combined sample. For example,

$$\hat{Y}_{il} = (\hat{\beta}_0 + \hat{\beta}_{q+1}) + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_q X_{iq}$$

is the predicted group 1 response for the i th individual in the combined sample, where $X_{i1}, X_{i2}, \dots, X_{iq}$ are the covariate values for that individual. Thus, the adjusted mean for a given group can be obtained by artificially assuming that all n persons in the combined sample constitute the given group; then the covariate distribution of the group is assumed to be that of the combined sample.

Thus, the method of adjustment using ANACOVA corrects the disparity in covariate distributions over groups by assuming a common distribution (not just a common set of means).

13.7.2 The Parallelism Assumption

As stated earlier, the ANACOVA method is inappropriate when the relationship between the covariates and the response is not the same in each group. Such nonparallelism or interaction might be reflected, for example, in a finding that males have higher blood pressures than females at older ages and that females have higher blood pressures than males at younger ages. Consequently, using a standard ANACOVA could lead to adjusted (for age) means for each (gender) group that are roughly equal; this would create the misleading impression that there was little difference between male and female blood pressures when, in fact, there were large differences in certain age categories. This example illustrates why we recommend that no method of adjustment be used in the presence of interaction; instead, the nature of the interaction itself should be quantified. The method for assessing interaction in this context, as previously indicated, requires first that the ANACOVA model (13.6) be expanded to include product terms between covariates and group variables and second that the coefficients of these product terms be tested for significance. The extended ANACOVA model that allows for such interaction terms has the form

$$Y = \beta_0 + \sum_{i=1}^q \beta_i X_i + \sum_{j=1}^{s-1} \beta_{q+j} Z_j + \sum_{i=1}^q \sum_{j=1}^{s-1} \gamma_{ij} X_i Z_j + E \quad (13.10)$$

In this model, a chunk test for parallelism would test H_0 : All $\gamma_{ij} = 0$ and would involve a multiple partial F statistic of the form

$$F(\text{All } X_i Z_j \text{ product terms} | \text{All } X_i \text{ and } Z_j \text{ terms})$$

which would have an F distribution under H_0 with $q(s-1)$ and $n - 1 - q - (s-1) - q(s-1)$ degrees of freedom. The use of ANACOVA-adjusted means is appropriate only when the preceding test for interaction yields the conclusion that interaction effects are not present.

13.7.3 Validity and Precision

As discussed in Chapter 11, validity and precision are two reasons to consider for controlling covariates. In using ANACOVA, validity is achieved by adjusting for confounding,

thereby obtaining an estimate of association that would have been distorted (biased) if the covariate(s) of interest had been ignored in the analysis.

Although validity should be the first consideration, it is possible to find no confounding in one's data and still control for one or the more covariates to gain precision. To assess precision, we can consider either the variances of the estimators of the association(s) of interest (the smaller these variances, the greater the precision) or confidence intervals for the association(s) of interest (the narrower the confidence intervals, the greater the precision). For example, consider an ANACOVA model involving two groups—say, males and females—with systolic blood pressure as the dependent variable and age as the only covariate of interest. If the age distribution for males were identical to that for females in the data, then age would not be a confounder. Nevertheless, because age is strongly positively associated with systolic blood pressure, precision will probably be increased by adjusting for age, even though confounding (i.e., validity) is not at issue.

13.7.4 Alternatives to ANACOVA

When adjusting for covariates, we can fit a model to contain the covariates and the study variables of primary interest (and perhaps even product terms, so interaction can be assessed) without having to make the study variables categorical. A best model would then have to be derived by means of criteria for selecting variables (see Chapter 16). If, for such a model, significant interaction is found, it is inappropriate (as noted earlier with regard to categorical study variables) to derive adjusted means. Moreover, even in the absence of interaction, it is impossible to obtain adjusted means for groups unless the study variables are defined categorically.

Nevertheless, predicted values based on the best regression model can be treated as adjusted values, since the covariates are being taken into account in the modeling process. Furthermore, adjusted means for distinct values of continuous study variables can be obtained by computing predicted values using the overall mean covariate values in the best model, as was done for ANACOVA using categorical study variables. For example, if we determine that the best model relating systolic blood pressure to age and weight is

$$\widehat{\text{SBP}} = \hat{\beta}_0 + \hat{\beta}_1 \text{AGE} + \hat{\beta}_2 \text{WEIGHT}$$

then we can compute adjusted blood pressure means for persons weighing 150 and 175 pounds (controlling for age) as

$$\widehat{\text{SBP}}_{150}(\text{adj}) = \hat{\beta}_0 + \hat{\beta}_1 \overline{\text{AGE}} + 150\hat{\beta}_2$$

and

$$\widehat{\text{SBP}}_{175}(\text{adj}) = \hat{\beta}_0 + \hat{\beta}_1 \overline{\text{AGE}} + 175\hat{\beta}_2$$

These adjusted values are similar to group adjusted means: the value of 150 can be thought of as representing a group of people whose weights are all close to 150; and a similar interpretation can be given to the 175-pound value.

In a more restrictive alternative to ANACOVA, *all* variables—even covariates—are treated as categorical. If we distinguish the set of dummy variables defining the covariates from the set of dummy variables defining the study variables, we can treat the regression model under consideration as a two-way analysis-of-variance model with unequal cell numbers (see Chapter 20). This alternative, however, might be inappropriate if the underlying means of measurement of some of the covariates are actually noncategorical (e.g., are continuous). If, for example, the inherently continuous variable age is categorized into three age groups, a completely categorical model based on that arbitrary categorization might lead to different results from those obtained by using a model that treats age continuously.

13.8 Summary

In this chapter, we have examined a long-established approach to controlling for covariates, ANACOVA. To use ANACOVA, the study variables of interest must be treated as categorical variables, whereas the covariates are not so restricted. ANACOVA also requires the assumption that there is no interaction between covariates and study variables. This assumption can be checked by testing for the significance of appropriate product terms in an extended ANACOVA model. If the test for interaction yields a finding of nonsignificance, adjusted means for different groups can be obtained by substituting the mean covariate values for the combined sample into the group-specific ANACOVA model. If the test for interaction is significant, adjusted means should not be used; instead, the nature of the interaction should be characterized.

Problems

1. Problem 8 in Chapter 12 involved comparing straight-line regression fits of SBP on QUET for smokers and nonsmokers, and it was found that these straight lines could be considered parallel. Use results based on that problem and the computer output given next (and the fact that the overall sample mean value $\bar{\text{QUET}} = 3.441$) to address the following issues.
 - a. State the appropriate ANACOVA regression model to use for comparing the mean blood pressures in the two smoking categories, controlling for QUET.
 - b. Determine the adjusted SBP means for smokers and nonsmokers. Compare these values to the unadjusted mean values

$$\bar{\text{SBP}}(\text{smokers}) = 147.823 \quad \text{and} \quad \bar{\text{SBP}}(\text{nonsmokers}) = 140.800$$
 - c. Test whether the true adjusted mean blood pressures in the two groups are equal. State the null hypothesis in terms of the regression coefficients in the ANACOVA model given in part (a).
 - d. Obtain a 95% confidence interval for the true difference in adjusted SBP means.

Edited SAS Output (PROC REG) for Problem 1

Regression of SBP on QUET and SMK

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4120.36649	2060.18325	25.91	<.0001
Error	29	2305.60226	79.50353		
Corrected Total	31	6425.96875			

Root MSE	8.91647	R-Square	0.6412
Dependent Mean	144.53125	Adj R-Sq	0.6165
Coeff Var	6.16924		

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	63.87603	11.46811	5.57	<.0001	668457
QUET	1	22.11560	3.22996	6.85	<.0001	3537.94574
SMK	1	8.57101	3.16670	2.71	0.0113	582.42075

2. **a–d.** Answer the same questions as in parts (a) through (d) in Problem 1 regarding an analysis of covariance designed to control for *both* AGE and QUET.
(Note: AGE = 53.250.) Use the results from Problem 9 in Chapter 12.
- e.** Is it necessary to control for *both* AGE and QUET as opposed to controlling just one of the two covariates?

Edited SAS Output (PROC REG) for Problem 2

Regression of SBP on QUET and SMK

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4889.82570	1629.94190	29.71	<.0001
Error	28	1536.14305	54.86225		
Corrected Total	31	6425.96875			

Root MSE	7.40691	R-Square	0.7609
Dependent Mean	144.53125	Adj R-Sq	0.7353
Coeff Var	5.12478		

(continued)

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	45.10319	10.76488	4.19	0.0003	668457
SMK	1	9.94557	2.65606	3.74	0.0008	393.09816
AGE	1	1.21271	0.32382	3.75	0.0008	4296.58607
QUET	1	8.59245	4.49868	1.91	0.0664	200.14147

3. In an experiment conducted at the National Institute of Environmental Health Sciences, the absorption (or uptake) of a chemical by a rat on one of two different diets, I or II, was known to be affected by the weight (or size) of the rat. A completely randomized design utilizing four rats on each diet was employed in the experiment, and the initial weight of each rat was recorded so that the diets could be compared after the researchers adjusted for the effect of initial weight. The data for the experiment are given in the following table.

Initial weight (X)	3	1	4	4	5	2	3	2
Diet (Z)	I	I	I	I	II	II	II	II
Response (Y)	14	13	14	15	16	15	15	14

- a. Using the initial weight as a covariate, state the ANACOVA regression model for comparing the two diets (set $Z = -1$ if diet I is used and $Z = 1$ if diet II is used).
- b. Use the computer results given next to determine the adjusted mean responses for each diet, controlling for initial weight.
- c. Use the ANACOVA regression model defined in part (a) to test whether the two diets differ significantly.
- d. Test whether the two diets differ significantly, completely ignoring the covariate. How do the two testing procedures compare?
- e. Determine a 95% confidence interval for the true difference in the adjusted mean responses.

Edited SAS Output (PROC REG) for Problem 3

Regression of Response (Y) on Initial Weight (X) and Diet (Z)

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5.00000	2.50000	12.50	0.0113
Error	5	1.00000	0.20000		
Corrected Total	7	6.00000			

(continued)

Root MSE	0.44721	R-Square	0.8333
Dependent Mean	14.50000	Adj R-Sq	0.7667
Coeff Var	3.08423		

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t 	Type I SS
Intercept	1	13.00000	0.41833	31.08	<.0001	1682.00000
X	1	0.50000	0.12910	3.87	0.0117	3.00000
Z	1	0.50000	0.15811	3.16	0.0250	2.00000

4. A political scientist developed a questionnaire to determine political tolerance scores (Y) for a random sample of faculty members at her university. She wanted to compare mean scores adjusted for age (X) in each of three categories: full professors, associate professors, and assistant professors. The data results are given in the accompanying tables (the higher the score, the more tolerant the individual).

Group 1: Full Professors ($Z_1 = 1, Z_2 = 0$)

Age (X)	65	61	47	52	49	45	41	41	40	39
Tolerance (Y)	3.03	2.70	4.31	2.70	5.09	4.02	3.71	5.52	5.29	4.62

Group 2: Associate Professors ($Z_1 = 0, Z_2 = 1$)

Age (X)	34	31	30	35	49	31	42	43	39	49
Tolerance (Y)	4.62	5.22	4.85	4.51	5.12	4.47	4.50	4.88	5.17	5.21

Group 3: Assistant Professors ($Z_1 = Z_2 = 0$)

Age (X)	26	33	48	32	25	33	42	30	31	27
Tolerance (Y)	5.20	5.86	4.61	4.55	4.47	5.71	4.77	5.82	3.67	5.29

- a. State an ANACOVA regression model that can be used to compare the three groups, controlling for age.
- b. What model should be used to check whether the ANACOVA model in part (a) is appropriate? Carry out the appropriate test. Use $\alpha = .01$.
- c. Using ANACOVA, determine adjusted mean tolerance scores for each group, and test whether these adjusted scores differ significantly from one another. Also, compare the adjusted means with the unadjusted means.

[Note: $\bar{X}(\text{overall}) = 39.667$, $\bar{Y}(\text{group 1}) = 4.10$, $\bar{Y}(\text{group 2}) = 4.86$, and $\bar{Y}(\text{group 3}) = 5.00$.]

Edited SAS Output (PROC REG) for Problem 4

Regression of Tolerance (Y) on Age (X), Faculty Categories (Z1 and Z2)

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	6.18372	0.61189	10.11	<.0001
X	1	-0.03635	0.01741	-2.09	0.0467
Z1	1	-0.33981	0.41457	-0.82	0.4199
Z2	1	0.06357	0.33234	0.19	0.8498

Regression of Y on X, Z1, Z2, XZ1, and XZ2

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	10.20934	2.04187	5.02	0.0027
Error	24	9.76275	0.40678		
Corrected Total	29	19.97210			

Root MSE	0.63779	R-Square	0.5112
Dependent Mean	4.64967	Adj R-Sq	0.4093
Coeff Var	13.71699		

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	5.42706	0.98483	5.51	<.0001	648.58200
X	1	-0.01321	0.02948	-0.45	0.6580	6.20495
Z1	1	2.78490	1.51591	1.84	0.0786	0.62494
Z2	1	-1.22343	1.50993	-0.81	0.4258	0.01847
XZ1	1	-0.07247	0.03779	-1.92	0.0671	3.14678
XZ2	1	0.03022	0.04165	0.73	0.4751	0.21420

5. A psychological experiment was performed to determine whether in problem-solving dyads containing one male and one female, “influencing” behavior depended on the sex of the experimenter. The problem for each dyad was a strategy game called “Rope a Steer,” which required 20 separate decisions about which way to proceed toward a defined goal on a game board. For each subject in the dyad, a verbal-influence activity score was derived as a function of the number of statements made by the subject to influence the dyad to move in a particular direction. The difference in verbal-influence activity scores within a dyad was denoted as the variable VIAD, which was then used as the dependent variable in an ANACOVA

designed to control for the effects of differing IQ scores of the male and female in each dyad. The relevant data are given in the following tables.

- State an ANACOVA model appropriate for these data.
- Determine adjusted mean VIAD scores for each group, and compare these to the unadjusted means for each group.
- Test whether the adjusted mean scores differ significantly.
- Find a 95% confidence interval for the true difference in adjusted mean scores.

Group 1: Male Experimenter ($Z = 0$)

VIAD	-10	-4	9	-15	-15	5	-8	-4	-1	13
IQ_M	115	112	106	123	125	105	115	122	138	110
IQ_F	100	110	108	135	115	112	121	132	135	126

Group 2: Female Experimenter ($Z = 1$)

VIAD	8	-5	2	-7	15	-10	-3	10	2	4
IQ_M	120	130	110	113	102	141	120	113	114	102
IQ_F	141	128	104	98	106	130	128	105	107	111

Edited SAS Output (PROC REG) for Problem 5

Regression of Viad on IQM, IQF, and Z

DESCRIPTIVE STATISTICS					
Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	20.00000	1.00000	20.00000	0	0
IQM	2336.00000	116.80000	275060	116.58947	10.79766
IQF	2352.00000	117.60000	279924	175.20000	13.23631
Z	10.00000	0.50000	10.00000	0.26316	0.51299
VIAD	-14.00000	-0.70000	1518.00000	79.37895	8.90949

[Portion of output omitted]

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	44.59001	17.85879	2.50	0.0238
IQM	1	-0.69301	0.19389	-3.57	0.0025
IQF	1	0.28109	0.15967	1.76	0.0974
Z	1	5.19610	3.13292	1.66	0.1167

(continued)

Regression of Viad on IQM, IQF, Z, IQM \times Z, and IQF \times Z

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	746.73852	149.34770	2.75	0.0623
Error	14	761.46148	54.39011		
Corrected Total	19	1508.20000			

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	43.99054	29.67269	1.48	0.1604	9.80000
IQM	1	-0.73495	0.31358	-2.34	0.0344	502.25850
IQF	1	0.32723	0.25688	1.27	0.2234	109.79250
Z	1	6.03875	38.43473	0.16	0.8774	131.46715
IQMZ	1	0.07613	0.41737	0.18	0.8579	0.06709
IQFZ	1	-0.08265	0.34325	-0.24	0.8132	3.15327

6. An experiment was conducted to compare the effects of four different drugs (A, B, C, and D) in delaying atrophy of denervated muscles. A certain leg muscle in each of 48 rats was deprived of its nerve supply by surgical severing of the appropriate nerves. The rats were then put randomly into four groups, and each group was assigned treatment with one of the drugs. After 12 days, four rats from each group were sacrificed, and the weight W (in grams) of the denervated muscle was obtained for each rat, as listed in the following table.

Theoretically, atrophy should be measured as the loss in weight of the muscle, but the initial weight of the muscle could not have been obtained without sacrificing the rat. Consequently, the initial total body weight X (in grams) of the rat was measured. It was assumed that this figure is closely related to the initial weight of the leg muscle.

Drugs A and C were large and small dosages, respectively, of atropine sulfate. Drug B was quinidine sulfate. Drug D acted as a control; it was simply a saline solution and could not have had any effect on atrophy. Use ANACOVA to compare the effects of the four drugs, controlling for initial total body weight (X). Use the results given in the SAS output shown next to perform your analysis. (Note: $X = 226.125$).

Drug A (Z1 = 1, Z2 = Z3 = 0)		Drug B (Z2 = 1, Z1 = Z3 = 0)		Drug C (Z3 = 1, Z1 = Z2 = 0)		Drug D (Z1 = Z2 = Z3 = 0)	
X	W	X	W	X	W	X	W
198	0.34	233	0.41	204	0.57	186	0.81
175	0.43	250	0.87	234	0.80	286	1.01
199	0.41	289	0.91	211	0.69	245	0.97
224	0.48	255	0.87	214	0.84	215	0.87

Edited SAS Output (PROC REG) for Problem 6

Regression of W on X, Z1, Z2, and Z3

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.61847	0.15462	10.71	0.0009
Error	11	0.15873	0.01443		
Corrected Total	15	0.77720			
Root MSE		0.12013	R-Square	0.7958	
Dependent Mean		0.70500	Adj R-Sq	0.7215	
Coeff Var		17.03922			

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	0.17283	0.30373	0.57	0.5808	7.95240
X	1	0.00319	0.00128	2.49	0.0299	0.32015
Z1	1	-0.39170	0.09541	-4.11	0.0017	0.20177
Z2	1	-0.22565	0.09020	-2.50	0.0294	0.06236
Z3	1	-0.13505	0.08776	-1.54	0.1521	0.03418

7. Through urine samples were analyzed for sodium content for each of two collection periods, one before and one after administration of Mercurhydrin, for each of 30 dogs. The experimenter used 7 dogs as a control group for the study; these dogs were not administered the drug, but their urine samples were collected for two similar time periods.

Experimental Group (Z = 1)			Experimental Group (Z = 1)		
Animal	First Collection Period (X) ([Na], mM/l)	Second Collection Period (Y) ([Na], mM/l)	Animal	First Collection Period (X) ([Na], mM/l)	Second Collection Period (Y) ([Na], mM/l)
1	17.5	22.1	18	6.3	12.7
2	9.4	12.0	19	9.7	17.1
3	10.0	15.2	20	7.1	9.5
4	7.4	23.1	21	7.2	11.0
5	8.8	9.8	22	5.3	8.2
6	18.9	26.9	23	14.3	15.8
7	10.8	11.1	24	7.9	9.7
8	8.8	13.6	25	14.1	14.7
9	8.8	12.8	26	12.8	17.0
10	9.2	7.5	27	12.8	20.2
11	8.1	8.1	28	10.7	13.9
12	10.3	27.5	29	5.9	11.8
13	10.1	11.2	30	3.8	9.0
14	7.3	11.0			
15	11.1	15.3	Mean	9.7	13.9
16	9.4	11.5	S.D.	3.3	5.3
17	8.2	8.4			

Control Group ($Z = 0$)		
Animal	First Collection Period (X) ([Na], mM/l)	Second Collection Period (Y) ([Na], mM/l)
1	11.1	9.4
2	5.1	5.9
3	6.5	14.8
4	17.2	15.5
5	11.8	23.4
6	6.6	7.3
7	4.1	8.2
Mean	8.9	12.1
S.D.	4.7	6.4

- a. Use ANACOVA (computer results are given next), with the first (“before”) collection period measure of sodium content as the covariate, to find the adjusted mean sodium contents for the experimental and control groups for the second (“after”) collection period.
- b. Test whether the two adjusted means differ significantly.
- c. What alternative testing approach (involving a t test) could be used for these data? Carry out this test. [Hint: Variances of before–after differences for each group are 16.785 (experimental) and 25.915 (control).]
- d. Are the two testing methods (t test versus covariance analysis) equivalent in this problem? Explain by comparing regression models appropriate for each method.
- e. What do the results of a lack-of-fit test, based on using the computer output given, indicate about the appropriateness of the covariance model used in parts (a) and (b)? The lack of fit test compares the value $6.3568 (= 21.0913/3.3179)$ to $F_{30, 4, 0.95}$; see Chapter 15 for further discussion of lack-of-fit tests.

Some Computer Results for Problem 7

Multiple $R^2 = .406$				ANOVA		
Variable	$\hat{\beta}$	S.D. of $\hat{\beta}$	$\hat{\beta}/\text{S.D.}$	Source	d.f.	MS
X	0.96155	0.20418	4.709	X	1	434.4861
Z	1.06435	1.83729	0.5793	Added by Z	1	6.3764
(intercept)	3.49992			Lack of fit	30	21.0913
				Pure error	4	3.3179

8. Consider again the data in Problem 4.
 - a. How would you compute appropriate cross-product variables to allow testing of whether an interaction exists between age and faculty rank?
 - b. State the associated regression model.
 - c. Using a computer, fit the model. Provide estimates of the regression coefficients.
 - d. Provide a multiple partial F test for interaction, controlling for age and faculty rank. Use $\alpha = .05$.
 - e. Does the F test in part (d) indicate that ANACOVA is valid?

9. This problem involves data from Problem 15 in Chapter 5. Treat LN_BRNTL as the dependent variable and dosage level of toluene (PPM_TOLU) as a categorical predictor (four levels). The experimenter wanted to explore the possibility of using WEIGHT as a control variable.
 - a. State the appropriate ANACOVA model. Use the 50-ppm exposure group as the reference group when coding dummy variables.
 - b. Use a computer program to fit the model and provide estimated regression coefficients.
 - c. Provide adjusted means. Compare these with the unadjusted means. (Do not perform any statistical tests.)
 - d. Test whether the adjusted means are all equal. Use $\alpha = .05$. State the null hypothesis in terms of regression coefficients.
10. Repeat Problem 9 using LN_BLDTL as the dependent variable.
11. a. How would you compute appropriate cross-product terms for testing the interaction of WEIGHT and dosage level of toluene for Problem 9?
 b. State the appropriate regression model.
 c. State the null hypothesis (of no interaction) in terms of regression coefficients.
12. Refer to the residential sales data in Problem 19 in Chapter 12. Use the computer output given below to answer the following questions.
 - a. State an ANACOVA regression model that can be used to compare outer suburbs with other locations (intown and inner suburbs), controlling for age. (*Hint:* Let $Z = 0$ if intown or inner suburbs, 1 if outer suburbs.)
 - b. Identify the model that should be used to check whether the ANACOVA model in part (a) is appropriate. Carry out the appropriate test.
 - c. Using ANACOVA, determine adjusted mean sales prices for the two locations, and test whether they differ significantly from one another. (*Note:* Mean house age = 10.8667; unadjusted mean sales price (in \$1,000s) for intown and inner suburbs = 82.1867; unadjusted mean sales price (in \$1,000s) for outer location = 84.8067.)

Edited SAS Output (PROC REG) for Problem 12

Regression of Sales Price (Y) on Age (X4), Location (Z), and X4 \times Z

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1646.14368	548.71456	7.48	0.0009
Error	26	1907.86599	73.37946		
Corrected Total	29	3554.00967			

(continued)

Root MSE	8.56618	R-Square	0.4632
Dependent Mean	83.49667	Adj R-Sq	0.4012
Coeff Var	10.25931		

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	55.45706	6.85947	8.08	<.0001	209151
X4	1	2.37245	0.57631	4.12	0.0003	1324.84735
Z	1	17.90588	8.90573	2.01	0.0548	114.99710
X4Z	1	-1.27910	0.76286	-1.68	0.1056	206.29922

Regression of Sales Price (Y) on Age (X4), Location (Z), and $X4 \times Z$

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1439.84445	719.92223	9.19	0.0009
Error	27	2114.16521	78.30242		
Corrected Total	29	3554.00967			

Root MSE	8.84887	R-Square	0.4051
Dependent Mean	83.49667	Adj R-Sq	0.3611
Coeff Var	10.59787		

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	63.68190	4.95305	12.86	<.0001
X4	1	1.64244	0.39005	4.21	0.0003
Z	1	3.93395	3.24618	1.21	0.2361

13. A company wants to compare three different point-of-sale promotions for its snack foods. The three promotions are

Promotion 1: Buy two items, get a third free.

Promotion 2: Mail in a rebate for \$1.00 with any \$2.00 purchase.

Promotion 3: Buy reduced-price multipacks of each snack food.

The company is interested in the average increase in sales volume due to the promotions. Fifteen grocery stores were selected in a targeted market, and each store was randomly assigned one of the promotion types. During the month-long run of the promotions, the company collected data on increase in sales volume (Y , in hundreds of units) at each store, to be gauged against average monthly sales volume (X , in hundreds of units) prior to the promotions. Let $Z_1 = 1$ if promotion type 1,

or 0 otherwise. Let $Z_2 = 1$ if promotion type 2, or 0 otherwise. The sample data are shown in the following table.

Store	Promotion	<i>Y</i>	<i>X</i>
1	1	12	39
2	1	23	42
3	2	11	23
4	3	17	39
5	3	15	37
6	3	18	31
7	1	12	36
8	2	19	38
9	3	21	33
10	1	13	44
11	1	7	26
12	2	5	20
13	2	8	32
14	3	17	36
15	2	19	29

- a. State an ANACOVA regression model for comparing the three promotion types, controlling for average pre-promotion monthly sales.
- b. Identify the model that should be used to check whether the ANACOVA model in part (a) is appropriate. Carry out the appropriate test.
- c. Using ANACOVA, determine adjusted mean increases in sales volume for the three promotions, and test whether they differ significantly from one another. (Note: Mean pre-promotional average sales volume = 33.6667; unadjusted mean increases in sales volume were 13.4 for promotion 1, 12.4 for promotion 2, and 17.6 for promotion 3.)

Edited SAS Output (PROC REG) for Problem 13

Regression of *Y* on *X*, *Z1*, *Z2*, *XZ1*, and *XZ2*

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	219.94330	43.98866	2.25	0.1369
Error	9	175.79003	19.53223		
Corrected Total	14	395.73333			

Root MSE	4.41953	R-Square	0.5558
Dependent Mean	14.46667	Adj R-Sq	0.3090
Coeff Var	30.54973		

(continued)

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	31.92157	24.43508	1.31	0.2238	3139.26667
X	1	-0.40686	0.69190	-0.59	0.5710	115.60187
Z1	1	-39.96273	27.16860	-1.47	0.1754	61.21220
Z2	1	-36.29584	26.03369	-1.39	0.1967	6.85169
XZ1	1	0.98016	0.75946	1.29	0.2290	2.41391
XZ2	1	0.99751	0.75757	1.32	0.2205	33.86363

Regression of Y on X, Z1, and Z2

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	183.66577	61.22192	3.18	0.0674
Error	11	212.06757	19.27887		
Corrected Total	14	395.73333			

Root MSE	4.39077	R-Square	0.4641
Dependent Mean	14.46667	Adj R-Sq	0.3180
Coeff Var	30.35095		

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	0.30045	7.58360	0.04	0.9691	3139.26667
X	1	0.49146	0.20810	2.36	0.0377	115.60187
Z1	1	-5.28122	2.81445	-1.88	0.0874	61.21220
Z2	1	-1.85804	3.11672	-0.60	0.5631	6.85169

14. In Problem 19 in Chapter 5 and Problem 14 in Chapter 8, data from the 1990 Census for 26 randomly selected Metropolitan Statistical Areas (MSAs) were discussed. Of interest were factors potentially associated with the rate of owner occupancy of housing units. The following three variables were included in the data set:

OWNEROCC: Proportion of housing units that are owner-occupied (as opposed to renter-occupied)

OWNCOST: Median selected monthly ownership costs, in dollars

URBAN: Proportion of population living in urban areas

It is also of interest to see whether the average owner occupancy rate differs between metropolitan areas where 75% or more of the population lives in urban areas and metropolitan areas where less than 75% of the population lives in urban areas, while

controlling for ownership costs. For this purpose, the following additional variable is defined:

$$Z = \begin{cases} 1 & \text{if proportion of population living in urban areas} \geq 0.75 \\ 0 & \text{otherwise} \end{cases}$$

- a. State an ANACOVA regression model that can be used to compare the two MSA types, controlling for average monthly ownership costs.
- b. State the model that should be used to check whether the ANACOVA model in part (a) is appropriate. Using the computer output given next, carry out the appropriate test. (*Note:* In the SAS output, the variable OWNCOST has been centered to avoid collinearity problems.)
- c. Using ANACOVA methods and the accompanying output, determine adjusted mean owner occupancy rates for the two types of MSAs, and test whether they significantly differ from one another. (*Note:* The unadjusted average owner occupancy rate for MSAs where at least 75% of the population lives in urban areas = 64.5%; the unadjusted average owner occupancy rate for MSAs where less than 75% of the population lives in urban areas = 69.25%).)

Edited SAS Output (PROC REG) for Problem 14

Regression of OWNEROCC on OWNCOST, Z, and OWNCOST*Z

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	214.19496	71.39832	4.06	0.0194
Error	22	386.76658	17.58030		
Corrected Total	25	600.96154			

Regression of OWNEROCC on OWNCOST and Z

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	213.49860	106.74930	6.34	0.0064
Error	23	387.46294	16.84621		
Corrected Total	25	600.96154			

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	68.66546	1.47336	46.60	<.0001
OWNCOST	1	-0.01267	0.00553	-2.29	0.0314
Z	1	-3.90567	1.78250	-2.19	0.0388

15. This problem refers to the radial keratotomy study data from Problem 12 in Chapter 8. Suppose that we want to compare the average change in refraction for males and females, controlling for baseline refractive error and baseline curvature. To this end, we define the following dummy variable:

$$Z = \begin{cases} 1 & \text{if patient is male} \\ 0 & \text{otherwise} \end{cases}$$

- a. State an ANACOVA regression model that can be used to compare the mean change in refraction (Y) for males and females, controlling for baseline refractive error (X_1) and baseline curvature (X_2).
- b. State the model that should be used to check whether the ANACOVA model in part (a) is appropriate. Using the computer output given next, carry out the appropriate test.
- c. Using ANACOVA methods and the SAS output, determine adjusted mean changes in refractive error for males and females, and test whether they differ significantly from one another. (Note: Unadjusted average change in refraction for males = 3.64 diopters; unadjusted average change in refraction for females = 3.965 diopters; overall mean baseline refraction = -4.03 diopters; overall mean baseline curvature = 44.02 diopters.)

Edited SAS Output (PROC REG) for Problem 15

Regression of Y on X1, X2, Z, X1 * Z, X2 * Z

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	24.71516	4.94303	4.27	0.0027
Error	48	55.53778	1.15704		
Corrected Total	53	80.25294			

Root MSE	1.07566	R-Square	0.3080
Dependent Mean	3.83343	Adj R-Sq	0.2359
Coeff Var	28.05993		

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	13.06563	6.97300	1.87	0.0671
X1	1	-0.20220	0.10539	-1.92	0.0610
X2	1	-0.22389	0.15756	-1.42	0.1618
Z	1	-0.91427	10.01057	-0.09	0.9276
X1Z	1	-0.35838	0.19864	-1.80	0.0775
X2Z	1	-0.02426	0.22572	-0.11	0.9149

(continued)

Regression of Y on X1, X2, and Z

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	20.90031	6.96677	5.87	0.0016
Error	50	59.35263	1.18705		
Corrected Total	53	80.25294			

Root MSE	1.08952	R-Square	0.2604
Dependent Mean	3.83343	Adj R-Sq	0.2161
Coeff Var	28.42156		

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	13.71986	5.06713	2.71	0.0093
X1	1	-0.30382	0.09041	-3.36	0.0015
X2	1	-0.24770	0.11408	-2.17	0.0347
Z	1	-0.50859	0.30608	-1.66	0.1028

16. For the BRFSS example discussed in Section 13.6, there are a number of inference-making procedures that may be conducted about the joint effects of the two exposure variables exercise and tobacco_now. Using the output in Section 13.6 and below for the analyses of 1,048 individuals, do the following:
- Test the null hypothesis H_0 : “The population mean BMI for an individual who exercises and doesn’t use tobacco equals that for an individual who neither exercises nor uses tobacco” against the alternative that this null hypothesis is false.
 - Assess whether there is statistical evidence that the population mean BMI for an individual who exercises and uses tobacco equals that for an individual who neither exercises nor uses tobacco.

Covariance Estimates from Problem 16

COVARIANCE OF ESTIMATES					
Variable	Intercept	drink_days	poor_sleep_days	exercise	tobacco_now
Intercept	0.2280218243	-0.004271369	-0.003156079	-0.166485871	-0.055383413
drink_days	-0.004271369	0.0006654907	0.0000211512	-0.000172153	0.0006952607
poor_sleep_days	-0.003156079	0.0000211512	0.0003314055	0.0002936177	-0.000743809
exercise	-0.166485871	-0.000172153	0.0002936177	0.201860869	0.023262661
tobacco_now	-0.055383413	0.0006952607	-0.000743809	0.023262661	0.242110744

- c. Describe how you would compute a 99% confidence interval for the population adjusted mean BMI of an individual who exercises and uses tobacco.
 - d. In a regression model not containing the two continuous covariates drink_days and poor_sleep_days, the value of $\hat{\beta}_4$ for tobacco_now changes from -1.108 to -0.855 . Do you think that drink_days and poor_sleep_days together are confounders of the relationship between tobacco_now and BMI?
17. In Section 13.6, adjusted mean BMI values were obtained for the two nominal variables exercise and tobacco_now. A drawback of the approach used is that these four adjusted means were estimated for the four combinations of these two factors considered simultaneously, leaving one without adjusted estimates for the two levels of each variable considered separately (e.g., the adjusted “marginal” mean for exercise = 1 regardless of the level of tobacco_now, etc.). Given that no product term between exercise and tobacco_now was included in the model, the computation of these four marginal adjusted means should be possible.

One approach is to use estimates for the “average” value of the nominal factor being controlled, but it is unclear which values should be used for exercise or tobacco_now, given that these dichotomous variables take on only the values 0 and 1. Two possible options are

1. *For the two levels of exercise, estimate the adjusted mean BMI assuming a value of 0.5 for tobacco_now. Similarly, for the two levels of tobacco_now, assume a value of 0.5 for exercise.* Specifically, the adjusted means for both values of exercise would be estimated by $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1(\text{drink_days}) + \hat{\beta}_2(\text{poor_sleep_days}) + \hat{\beta}_3(\text{exercise}) + \hat{\beta}_4(0.5)$. A criticism of this approach is that, since no person could have an observed value of 0.5 for either factor, using 0.5 may yield unrealistic estimates. A second issue is that, for population-based studies such as the BRFSS, values of 0.5 may represent unrealistic choices for the sample prevalences of these variables (see option 2 of this problem). This is because the sample distributions of exercise and tobacco_now are not perfectly balanced, with exactly 50% of the population having the value 1 for each variable. This approach for computing adjusted means is revisited in the discussion of ANOVA in Chapter 17.
2. *For the two levels of exercise, estimate the adjusted mean BMI assuming that the sample prevalence for current tobacco use is 16.1%. For the two levels of tobacco_now, assume that the sample prevalence for exercise is 79.9%.* This is known as the *observed margins weighting, conditional, or conditional margins approach* and is the simplest of a set of techniques known as *standardization methods* (Wilcosky and Chambliss 1985). By using sample prevalences for these dichotomous predictors, one is effectively using the analog of sample means for continuous covariates. As with option 1, option 2 assigns a non-observable value to a variable taking only the values 0 and 1.

Both of the above options are readily computed in SAS, although there are even more approaches, such as *stratified* and *marginal* methods (Wilcosky and Chambliss 1985; Flanders and Rhodes 1987). Using each of the two options described above and the output in Section 13.6, estimate the adjusted mean BMI for the levels of exercise and

tobacco_now. How do these estimates compare to each other and to the different sets of adjusted means in Section 13.6?

References

- Flanders, W. D., and Rhodes, P. H. 1987. "Large Sample Confidence Intervals for Regression Standardized Risks, Risk Ratios, and Risk Differences." *Journal of Chronic Diseases* 40(7): 697–704.
- Patrick, R.; Cassel, J. C.; Tyroler, H. A.; Stanley, L.; and Wild, J. 1974. "The Ponape Study of the Health Effects of Cultural Change." Paper presented at the annual meeting of the Society for Epidemiologic Research, Berkeley, California.
- Wilcosky, T. C., and Chambless, L. E. 1985. "A Comparison of Direct Adjustment and Regression Adjustment of Epidemiologic Measures." *Journal of Chronic Diseases* 38(10): 849–56.

14

Regression Diagnostics

14.1 Preview

This chapter provides an introduction to *regression diagnostics*—that is, statistical techniques for detecting conditions that can lead to inaccurate or invalid regression results. Such techniques include methods for detecting outliers, checking regression assumptions (see Section 8.4), and detecting the presence of collinearity.

Outliers (individual data values that are much larger or smaller than the rest of the values in the data set) may represent recording errors. Clearly, such values must be detected and corrected before the analysis proceeds. Sometimes, however, outliers are not recording errors. Even so, they may influence the fit of the regression model and, therefore, may affect our understanding of the relationship between the dependent and independent variables. As a result, outliers need to be identified, their plausibility carefully and scientifically judged, and a decision made as to whether the analysis should be revised in light of their presence.

If regression assumptions are violated, then, naturally, regression results may be invalid. Assumption checking can be a subjective, difficult, and time-consuming process, and remedies for assumption violations can themselves be difficult to implement. Nevertheless, it is essential that a thorough check of assumptions be performed so that potential limitations of the analysis are identified and attempts to address them can be made.

Collinearity is a problem that exists when there are strong linear relationships among two or more *predictor variables*. When severe collinearity exists, regression results are unstable in the sense that small, practically unimportant changes in data values can result in meaningfully large changes in the estimated model. Such instability can lead to inaccurate estimates of regression coefficients, variability, and *P*-values and is clearly undesirable.

We present a strategy for regression diagnostics, including simple methods that will help the analyst avoid and diagnose such problems. Suggestions are also made for corrective action when problems exist.

14.2 Simple Approaches to Diagnosing Problems in Data

In statistical analyses, it is important to be thoroughly familiar with the basic characteristics of the data, and regression analyses are no exception. Such familiarity helps avoid many errors. For example, it is essential to know the following:

1. The type of subject or experimental unit (e.g., small pine tree needles, elderly male humans)
2. The procedure for data collection
3. The unit of measurement for each variable (e.g., kilograms, meters, cubic centimeters)
4. A plausible range of values and a typical value for each variable

This knowledge, combined with a thorough descriptive statistical analysis of each variable, can be very useful in detecting errors in the data and pinpointing potential violations of regression assumptions. With this in mind, we recommend, as a first step in a regression diagnostics strategy, that the following simple descriptive analyses be performed¹:

1. Examine the five largest and five smallest values for every numeric variable. This is a simple but extremely powerful technique for examining data. Combined with knowledge of the basic characteristics of the data described above, examination of these extreme values may help pinpoint the most extreme recording errors and other outliers. For large data sets, it may be necessary to examine more than just these 10 values for some variables.

When an outlier is detected, the data value in question should be compared against data collection forms or other original records to determine whether a recording error has occurred. If it has, every attempt should be made to recover the correct data value before proceeding with the analysis. If the correct value cannot be recovered, the data value in question is usually set to missing in the computerized data set so that it is not inadvertently used in the fitting of the model.² Note that, since subsequent regression models may not require the use of the variable with the outlying value, the entire observation is usually not deleted—only the value of the variable in question is set to missing.

If it is determined that the data value has not been recorded in error, it is important to judge the plausibility of the value in question. For example, suppose we have collected data on body temperature in human subjects. The value 38.1 is plausible if the units are degrees Celsius. But, if the units are degrees Fahrenheit, 38.1 is implausible. More generally, one may classify any observation as being impossible, highly implausible, or plausible. Impossible values should be

¹ Indeed, every statistical analysis, not just regression analyses, should begin with steps 1 and 2.

² Modern statistical techniques for data imputation may also be employed to estimate missing data. These techniques will be useful if there are more than just a few missing values. See Rubin (1987) and Schafer (1997) for more information.

set to missing. Plausible values are not themselves changed, but it may be necessary to consider an alternative form of the model (i.e., a model with interaction or other higher-order terms) that might fit the data better. For highly implausible values, a decision should be made as to whether to remove the value in question from the analysis.³ When values are removed, this action, its impact on the final results obtained, and the scientific justification for it should be documented and presented along with any regression results.

This simple approach may not be sufficient for detecting all important outliers. We discuss additional methods focused on outlier detection in 14.3.1.

2. Examine the appropriate descriptive statistics for each variable. For categorical data, frequency tables should be produced. For continuous variables, the mean, median, standard deviation, interquartile range, and range should be examined, at a minimum. Graphical approaches such as bar charts, histograms, and box-and-whisker plots are also useful, especially when the number of variables is large. The information should be compared with what is expected from the study design and from scientific knowledge about the variables.
3. For a simple linear regression with a continuous predictor, create a scatterplot of the dependent variable versus the independent variable. Since a linear relationship is assumed, the plot should depict, at least roughly, a linear scatter of the plotted points. The Pearson correlation between the pair of variables should also be calculated to help further describe the strength and direction of the linear relationship observed. A distinct but nonlinear pattern may indicate a potential violation of the linearity assumption and a need for revision of the model. A fairly nondescript or random scatter of points may indicate that there will be no significant linear association between the independent and dependent variables in the subsequent regression analysis.

If these plots indicate significant problems, the regression model may need to be reviewed and reformulated before proceeding further. Usually, however, model reformulation is postponed until further diagnostics analyses (described below) have been performed.

Scatterplots are also useful for identifying strong outliers in the data.

4. For multiple regression models, the simple plots of the dependent variable versus individual predictor variables described in (3) above may be misleading, since they do not account for other predictor variables. In general, independent variables will not be completely uncorrelated with each other, and these associations may impact relationships with the dependent variable (i.e., confounding).

Partial regression plots do take the other predictor variables into account and should be used for multiple regression instead of, or in addition to, the simple scatter-plots

³ One needs to be very cautious about removing supposedly “highly implausible” outliers, since such a removal process increases the risk of altering data enough to produce different analysis conclusions. In the context of multiple linear regression, the removal of observations most at odds with observed data patterns may produce a significantly better-fitting (but possibly erroneous) regression model, underestimation of variability, and hence *P*-values that are too small and confidence intervals that are too narrow.

described in (3) above. In a partial regression plot for a particular predictor, we plot the dependent variable against the predictor of interest, adjusting for the remaining $(k - 1)$ predictors. In particular, assume that the overall model of primary interest is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + E$$

To produce the partial regression plot for the k th predictor, we first fit two models:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1} + E$$

and

$$X_k = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_{k-1} X_{k-1} + E$$

Then we plot the residuals from the two models against each other; that is, we plot $(Y - \hat{Y})$ versus $(X_k - \hat{X}_k)$. The residuals represent the original Y and X_k variables adjusted for the remaining $(k - 1)$ predictors. The partial correlation, $r_{YX_k|X_1, X_2, \dots, X_{k-1}}$, between Y and X_k should also be calculated to help further describe the strength and direction of the linear relationship observed.

5. Calculate pairwise correlations between independent variables. Strong linear associations between independent variables may signal collinearity problems, which we will discuss in detail later.

- Example 14.1** We will illustrate many of the techniques described in this chapter using a data set taken from Lewis and Taylor (1967). The variables WEIGHT (in kilograms), HEIGHT (in meters), and AGE (in years) were all recorded for a sample of boys and girls. We consider only the data on 127 boys. The goal of the regression analysis is to model WEIGHT, the dependent variable, as a linear function of HEIGHT and AGE.

Output for the relevant descriptive statistical analyses can be easily produced using most computer packages, and output from SAS (PROC UNIVARIATE) is shown below, in edited form. We see that the summary statistics (mean, median, standard deviation, range, and interquartile range, shown in boxes on the output) for AGE are consistent with what is expected in this study on boys; the same is true for HEIGHT. Furthermore, the five largest and five smallest data values for these variables do not seem to be outliers. For WEIGHT, the measures of central tendency look reasonable; however, the standard deviation appears somewhat large, and the largest value (88.9 kg) seems to be quite a bit larger than the next nearest value. It would be prudent to verify that no recording errors have occurred before proceeding. If not, then the plausibility of this largest value of WEIGHT should be examined. Suppose that it is decided that the value, although unusual, is scientifically plausible; it will, therefore, remain unchanged.

The plots of WEIGHT versus HEIGHT and AGE (Figures 14.1 and 14.2) support the idea that there are linear associations between WEIGHT and these predictors.⁴

⁴ The temptation to overinterpret these plots should be avoided. For example, some analysts might perceive a very subtle curvilinear pattern in Figure 14.1 and decide that the regression model should be revised to reflect that curvature.

Edited SAS Output (PROC UNIVARIATE) for Example 14.1

Variable: AGE

BASIC STATISTICAL MEASURES			
Location		Variability	
Mean	13.68110	Std Deviation	1.45449
Median	13.50000	Variance	2.11553
Mode	12.00000	Range	5.58333
		Interquartile Range	2.16667

EXTREME OBSERVATIONS			
Lowest		Highest	
Value	Obs	Value	Obs
11.5833	68	16.6667	59
11.5833	3	16.9167	58
11.6667	120	17.0833	32
11.6667	55	17.1667	19
11.6667	34	17.1667	103

Variable: HEIGHT

BASIC STATISTICAL MEASURES			
Location		Variability	
Mean	1.575760	Std Deviation	0.10980
Median	1.569720	Variance	0.01206
Mode	1.562100	Range	0.54610
		Interquartile Range	0.17780

EXTREME OBSERVATIONS			
Lowest		Highest	
Value	Obs	Value	Obs
1.28270	42	1.76530	103
1.33350	94	1.77292	114
1.35382	18	1.80340	59
1.36652	127	1.80340	76
1.39700	112	1.82880	67

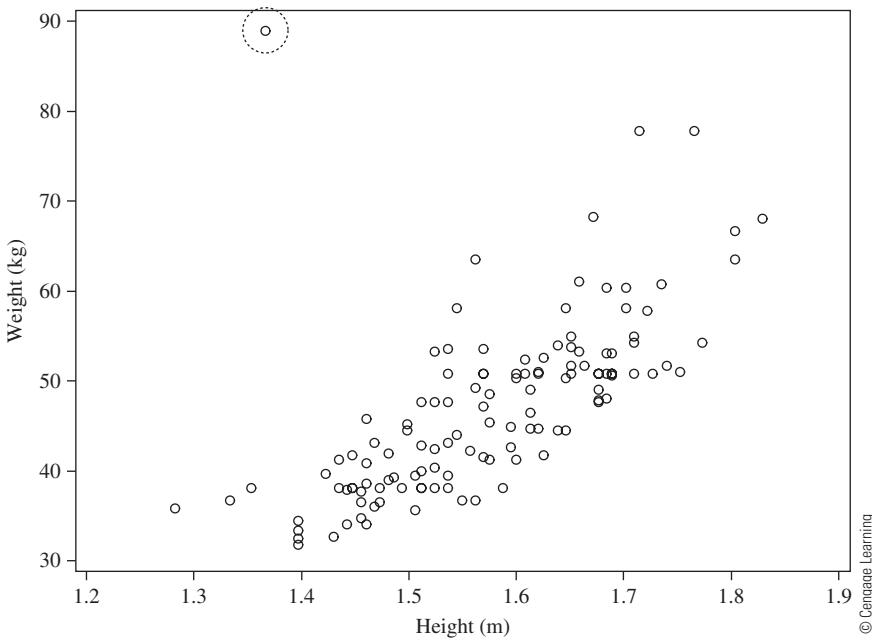
Variable: WEIGHT

BASIC STATISTICAL MEASURES			
Location		Variability	
Mean	47.25396	Std Deviation	9.76021
Median	47.62719	Variance	95.26163
Mode	50.80234	Range	57.15263
		Interquartile Range	12.47379

EXTREME OBSERVATIONS			
Lowest		Highest	
Value	Obs	Value	Obs
31.7515	112	68.0388	67
32.4318	86	68.2656	77
32.6586	85	77.7911	32
33.3390	68	77.7911	103
34.0194	123	88.9041	127

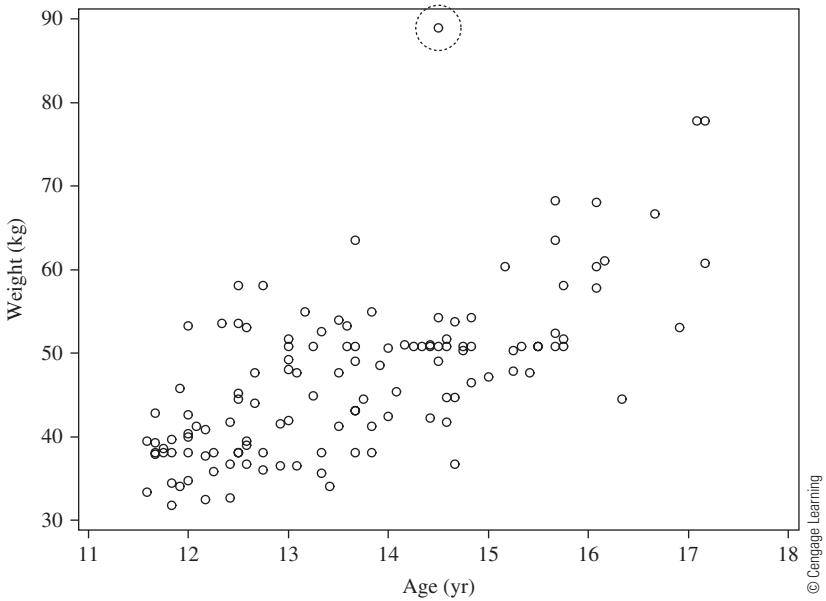
© Cengage Learning

However, while these very simple plots are useful for detecting gross departures from ideal conditions, other, more subtle departures that are apparent may simply be artifacts of sample size, scale, and style. Therefore, further diagnostics analyses should be undertaken before decisions to revise the model are made. Likewise, even if no anomalies are apparent in the plots, that does not necessarily mean that there are no problems with the data or the regression assumptions; further diagnostics analyses may reveal such problems.



© Cengage Learning

FIGURE 14.1 Children's body weight as a function of height; Lewis and Taylor data (1967) ($n = 127$)

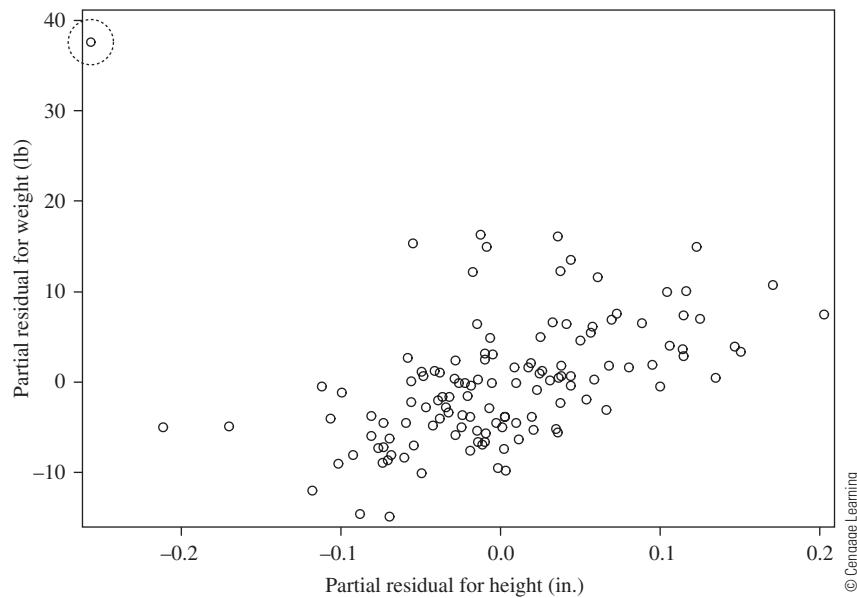


© Cengage Learning

FIGURE 14.2 Children's body weight as a function of age; Lewis and Taylor data (1967) ($n = 127$)

The plots provide a clear indication that the observation in which WEIGHT is approximately 90 kg (circled in the figures) is an outlier. This is, of course, observation 127 identified above. The partial regression plots (Figures 14.3 and 14.4), which are appropriate to examine for multiple regression models, roughly suggest linear relationships. However, the relationships are different than in Figures 14.1 and 14.2 because they have been adjusted for the effects of the second predictor. After the adjustment, the relationships look less obvious, especially for AGE. This is an indication that the apparent relationship between WEIGHT and HEIGHT in the simple scatterplot in Figure 14.1 is partly attributable to the children's ages. Similarly, the apparent relationship between WEIGHT and AGE in the simple scatterplot in Figure 14.2 is partly attributable to the children's heights. In this case, when both predictors will be included in a multiple regression model, neither will appear to have as strong a relationship with WEIGHT compared with being used separately in a simple linear regression model. (Note: The inclusion of the outlier, observation 127, in Figure 14.4 requires an extension of the vertical axis such that it is difficult to see the linear relationship between WEIGHT and AGE for the rest of the data. Plotting without the outlier reveals the linear association slightly more clearly, though it still remains weak.)

The SAS output (PROC CORR) shows that the simple Pearson correlations between WEIGHT and HEIGHT and between WEIGHT and AGE are moderately large (0.65470 and 0.66497, respectively), supporting the idea that the assumption of linearity may be reasonable. The partial correlations (output not shown) between WEIGHT and each of the predictors are $r_{\text{WEIGHT}, \text{HEIGHT} | \text{AGE}} = 0.31$ and $r_{\text{WEIGHT}, \text{AGE} | \text{HEIGHT}} = 0.35$,



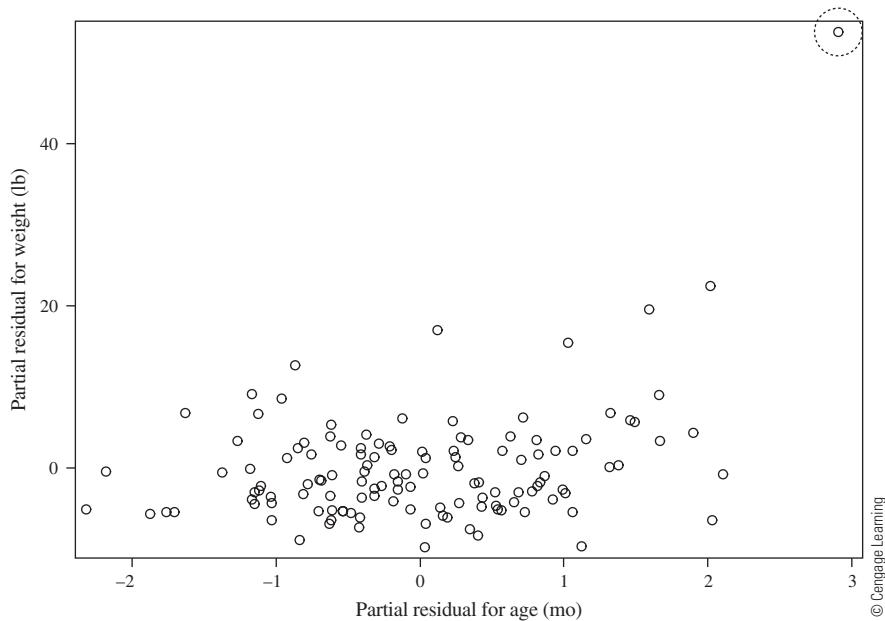
© Cengage Learning

FIGURE 14.3 Partial regression plot (adjusting for age) of children's body weight versus height; Lewis and Taylor data (1967) ($n = 127$)

Edited SAS Output (PROC CORR) for Example 14.1

3 Variables:	WEIGHT	HEIGHT	AGE
PEARSON CORRELATION COEFFICIENTS, N = 127			
WEIGHT	1.00000	0.65470	0.66497
HEIGHT	0.65470	1.00000	0.75328
AGE	0.66497	0.75328	1.00000

© Cengage Learning

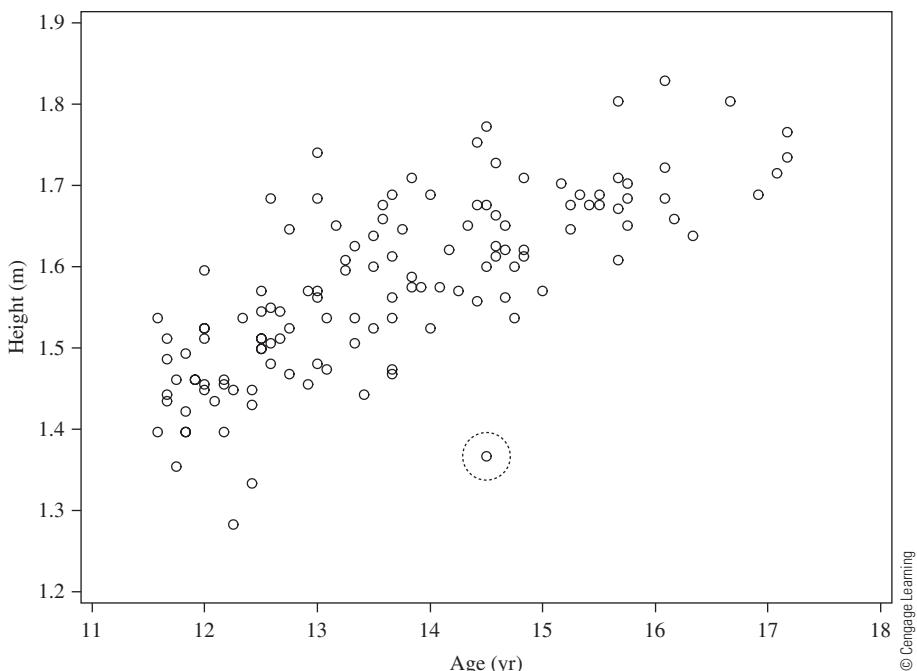


© Cengage Learning

FIGURE 14.4 Partial regression plot (adjusting for height) of children's body weight versus age; data from Lewis and Taylor (1967) ($n = 127$)

confirming the information contained in the partial regression plots: the linear associations of HEIGHT with WEIGHT and AGE with WEIGHT are not as strong once the other predictor is controlled for. (Note: Once again the outlier affects these partial correlations; without the outlier, we would find that $r_{\text{WEIGHT}, \text{HEIGHT}|\text{AGE}} = 0.55$ and $r_{\text{WEIGHT}, \text{AGE}|\text{HEIGHT}} = 0.23$.)

Both Figure 14.5 and the SAS output (PROC CORR) suggest that there is a linear association between the two independent variables ($r_{\text{HEIGHT}, \text{AGE}} = 0.75$). This is a preliminary



© Cengage Learning

FIGURE 14.5 Children's height as a function of age; Lewis and Taylor data (1967)
($n = 127$)

indication of potential collinearity problems that could cause invalid regression results. However, further examination of the problem is necessary to determine whether the collinearity problem is severe enough to warrant corrective action; relevant methods will be described in Section 14.5. ■

14.3 Residual Analysis: Detecting Outliers and Violations of Model Assumptions

Although the simple techniques described in Section 14.2 are useful for detecting the largest outliers in a data set, they often fail to detect more subtle outliers that may still be influential in the analysis. The methods of Section 14.2 may not be sufficient, by themselves, for detecting regression assumption violations. The second stage of a diagnostics analysis usually involves examination of regression residuals to detect additional outliers and assumption violations.

In Section 8.4.2, we defined the regression residual, \hat{E}_i , for the i th observation as the difference between the observed and predicted responses:

$$\hat{E}_i = Y_i - \hat{Y}_i \quad i = 1, 2, \dots, n$$

It is reasonable to expect that the residuals will be large for observations that are outliers, since the outliers lie “far” from the bulk of the data and, therefore, will also be relatively far from the best-fitting regression surface. We will describe a few methods for examining the sizes of residuals. Also, the \hat{E}_i are estimates of the unobserved model error terms, E_i , $i = 1, 2, \dots, n$. The regression assumptions, described in Chapter 8, are that the E_i are independent, have a mean of zero, have a common variance σ^2 , and follow a normal distribution. It is reasonable to expect that the residuals, \hat{E}_i , should roughly exhibit properties consistent with these assumptions. We will describe methods for assessing whether the residuals follow a normal distribution, have constant variance, and are consistent with the linearity and independence assumptions.⁵

Readers should note that, instead of examining the ordinary residuals, \hat{E}_i , which have values on the scale of the dependent variable, we often study the following functions of the residuals that scale them to be dimensionless: the *standardized* residuals, $z_i = \frac{\hat{E}_i}{S}$, where S is the square root of the mean squared error (MSE); the *studentized* residuals, $r_i = \frac{\hat{E}_i}{S\sqrt{1 - h_i}}$, where h_i is the “leverage” of the i th observation (described in Section 14.3.1); and the *jackknife* residuals, discussed in more detail next. The z_i have unit variance in the sense that

$$\frac{1}{n - k - 1} \sum_{i=1}^n z_i^2 = \frac{1}{n - k - 1} \sum_{i=1}^n \left(\frac{\hat{E}_i}{S} \right)^2 = \frac{1}{S^2} \left(\frac{1}{n - k - 1} \sum_{i=1}^n \hat{E}_i^2 \right) = 1$$

Also, if regression assumptions are satisfied, each r_i should follow a Student’s t distribution with $(n - k - 1)$ degrees of freedom.

14.3.1 Detecting Outliers

Substantial differences exist among possible types of extreme values. An *outlier* is any rare or unusual observation that appears at one of the extremes of the data range. All regression observations—and hence outliers, in particular—may be evaluated with respect to three criteria: reasonableness, given knowledge of the variable; response extremeness; and predictor extremeness. The goals in a diagnostics analysis are to identify recording errors and to identify *influential* observations that significantly affect either the choice of variables in the model or the accuracy of estimates of the regression coefficients and associated standard errors.

Methods for detecting outliers have been well documented (see, e.g., Belsley, Kuh, and Welsch 1980; Cook and Weisberg 1982; and Stevens 1984). In this section, we describe three regression diagnostic statistics for evaluating outliers: leverages, jackknife residuals, and Cook’s distance. Many other closely related statistics also exist. Most of these statistics are easily produced using SAS and other standard statistical software.

Leverages

The leverage, h_i , is a measure of the extremeness of an observation with respect to the independent variables. For a data set containing predictor variables X_1, X_2, \dots, X_k , it is

⁵ The residuals are not independent random variables—this is obvious from the fact that they sum to zero. However, in general, if the number of residuals is large relative to the number of independent variables, the dependency effect can, for all practical purposes, be ignored in any analysis of the residuals (see Anscombe and Tukey, 1963, for a discussion of the effect of this dependency on graphical procedures involving residuals).

the geometric distance of the i th predictor point $(X_{i1}, X_{i2}, \dots, X_{ik})$ from the center point $(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k)$ of the predictor space. Therefore, the larger the value, the farther the outlier is from the center of the data. Thus, whereas the scatterplot in Figure 14.5 allowed for a graphical assessment of extremeness, the leverage provides a numerical quantification of the level of extremeness. It can be shown that the average leverage value is $(k + 1)/n$ and that $0 \leq h_i \leq 1$. Hoaglin and Welsch (1978) recommend scrutinizing all observations for which $h_i > 2(k + 1)/n$.

Jackknife Residuals

Generally, large outliers will result in large \hat{E}_i values. Therefore, the sizes of the \hat{E}_i (or the standardized or studentized residuals) can be studied in order to detect outliers. Sometimes, however, outliers can mask their own effects. That is, the outlier exerts influence on the fitted regression surface, pulling the fitted regression surface away from the main body of the data and toward itself, thereby reducing the size of the associated residual. The jackknife residual avoids this problem and, therefore, unmasks outliers. For the i th observation, the jackknife residual, $r_{(-i)}$, is calculated as

$$r_{(-i)} = \frac{\hat{E}_i}{S_{(-i)} \sqrt{1 - h_i}}$$

The quantity $S_{(-i)}^2$ is the mean square error computed with the i th observation deleted. The unmasking is achieved by this deletion of the i th observation during the computation of the mean square error. The idea is that the ordinary mean square error, S^2 , will be larger than $S_{(-i)}^2$ if the outlier is masking its effect, since the outlier will pull the surface toward itself and away from the main body of the data. Therefore, for the i th observation, $r_{(-i)}$ will be larger than the studentized residual, r_i , in the case of an outlier.

From the formula for $r_{(-i)}$, we see that the jackknife residuals may be large if an observation is an outlier in the response variable Y (resulting in a large numerator) or in the predictor space of X_1, X_2, \dots, X_k (resulting in a large h_i) or if it strongly affects the fit of the model (as reflected in the difference between S^2 and $S_{(-i)}^2$). It can be shown that each jackknife residual has a t distribution with $n - k - 2$ degrees of freedom if the usual regression assumptions are met. Therefore, if the absolute value of $r_{(-i)}$ is greater than the 95th percentile of the relevant t distribution (i.e., $\alpha = .10$), then it would be reasonable to say that observation i is an outlier and should be scrutinized further.⁶

Cook's Distance

Cook's distance measures the extent to which the estimates of the regression coefficients change when an observation is deleted from the analysis. When the error terms have mean 0 and equal variance and when the model predictors are uncorrelated, Cook's distance, d_i ,

⁶ For large data sets consisting of several thousand observations, this may be an impractical criterion, as a large number of observations will need to be scrutinized. Smaller α levels are advisable in this case—for example, .05, .01, or even .005. Another approach would be to divide α by the number of observations, resulting in a Bonferroni type of correction—this is described in more detail in Section 17.7.1.

when the i th observation is deleted, is proportional to

$$\sum_{j=0}^k [\hat{\beta}_j - \hat{\beta}_{j(-i)}]^2 = [\hat{\beta}_0 - \hat{\beta}_{0(-i)}]^2 + [\hat{\beta}_1 - \hat{\beta}_{1(-i)}]^2 + \cdots + [\hat{\beta}_k - \hat{\beta}_{k(-i)}]^2$$

In general, for the i th observation, Cook's distance is given by

$$d_i = \left(\frac{1}{k+1} \right) \left(\frac{h_i}{1-h_i} \right) r_i^2$$

Clearly, d_i may be large either because the i th observation is extreme in the predictor space (i.e., resulting in a large h_i) or because it has a large studentized residual r_i (indicating an outlier with respect to either the predictors or the response). Cook and Weisberg (1982) suggested that any observation with $d_i > 1$ may deserve closer scrutiny, although more recent research indicates that this may not be the optimal critical value in all situations.⁷ In the absence of other hard and fast rules, we suggest that analysts should use the $d_i > 1$ rule but should also use Cook's distance in conjunction with other outlier statistics rather than in isolation.

As described in Section 14.2, once identified, outliers that are not easily identifiable as recording errors should be judged as to plausibility.

■ Example 14.2 For the regression of WEIGHT on HEIGHT and AGE, using the data of Example 14.1, outlier detection statistics were computed using SAS. The observations for which at least one of the diagnostic statistics exceeded the critical value (i.e., $d_i > 1$, $h_i > 2(2+1)/127 = 0.047$, and $|r_{(-i)}| > t_{127-2-2, 0.95} \approx 1.66$) are shown in Table 14.1. For each of these observations, we first determine whether any data values were recorded in error. If not, we then examine each diagnostic statistic for each observation in more detail.

For observation 5, only the leverage value is large ($h_i = 0.057$), indicating that the value of either HEIGHT or AGE (or both) should be examined. For this observation, HEIGHT = 1.68 m, which is one of the taller heights (but not one of the extremes) and so is not highly implausible, and AGE = 12.58 years, which is also not an extreme value and thus is not highly implausible for this study. Since both values are plausible, and assuming that they have been recorded correctly, no corrective action is necessary. This observation simply appears as an outlier (albeit a moderate one) because it is unusual with respect to both variables when jointly considered, representing one of the younger but taller children in the study.

Let us also consider observation 127. From Table 14.1, we see that all three outlier diagnostic statistics are large. Since the leverage is large, we should examine the HEIGHT and AGE values. Indeed, the data reveal that this child was 14.5 years old (close to the average for the study) but was only 1.37m tall (one of the shortest heights in the study).

⁷ Muller and Chen Mok (1997) performed simulations that suggested that the $d_i > 1$ rule can be inaccurate. They suggested an alternate approach, which is presented in Table A.10.

TABLE 14.1 Outliers with large Cook's distance (d_i), leverage (h_i), or jackknife residual ($r_{(-i)}$) values for regression of WEIGHT on HEIGHT and AGE using Lewis and Taylor data (1967) ($n = 127$)

Observation	d_i	h_i	$r_{(-i)}$
5	0.01235	0.05673	0.78378
19	0.00027	0.05578	-0.11583
32	0.12519	0.05586	2.57621
41	0.02372	0.01505	2.19094
42	0.00187	0.08374	0.24686
52	0.02261	0.02140	1.77651
58	0.01712	0.05475	-0.94113
67	0.01552	0.05015	0.93846
69	0.04869	0.04612	-1.75257
75	0.02172	0.01891	-1.85646
77	0.02675	0.02312	1.85939
93	0.00067	0.07225	0.16055
94	0.00012	0.05804	0.07542
103	0.09529	0.05357	2.28530
126	0.01580	0.00813	2.45374
127	2.01754	0.10997	8.96226

© Cengage Learning

The plausibility of this HEIGHT and AGE combination needs to be judged. Since Cook's distance and the jackknife residual are also quite large, it is worth examining the dependent variable WEIGHT as well (although it is possible that the independent variables alone are responsible for all the outlier statistics being large in magnitude). It turns out that this child's WEIGHT, 88.9 kg, is the largest in the data set. For this observation, the combined effect of all three variables causes it to stand out as an extreme outlier. Although we would probably decide that the observation is unusual but plausible and that nothing should be changed (indeed, that is what happens in most cases), it is possible that, on a rare occasion, the decision to remove an observation due to implausibility will be made or a decision to revise the model (e.g., by adding interaction or other higher-order terms or by applying a transformation to, say, the response variable Y [Section 14.4.1]) will be made in order to improve fit. We then examine each of the other observations in Table 14.1 in similar fashion.

Note that, for the WEIGHT/HEIGHT/AGE data, while the simple plots and univariate statistics in Section 14.2 did reveal observation 127 to be an outlier, they did not readily identify the remaining outliers shown in Table 14.1. This is precisely why the simple approaches of Section 14.2 are not, by themselves, sufficient for a reliable outlier detection analysis. ■

14.3.2 Assessing the Linearity, Homoscedasticity, and Independence Assumptions Using Graphical Analyses of Residuals

Plots of the residuals (ordinary, studentized, or jackknife) versus predicted values are very useful for checking regression assumptions. These plots are commonly referred to simply as *residual plots*. Some of the general patterns that may emerge in the plots are shown in Figure 14.6.

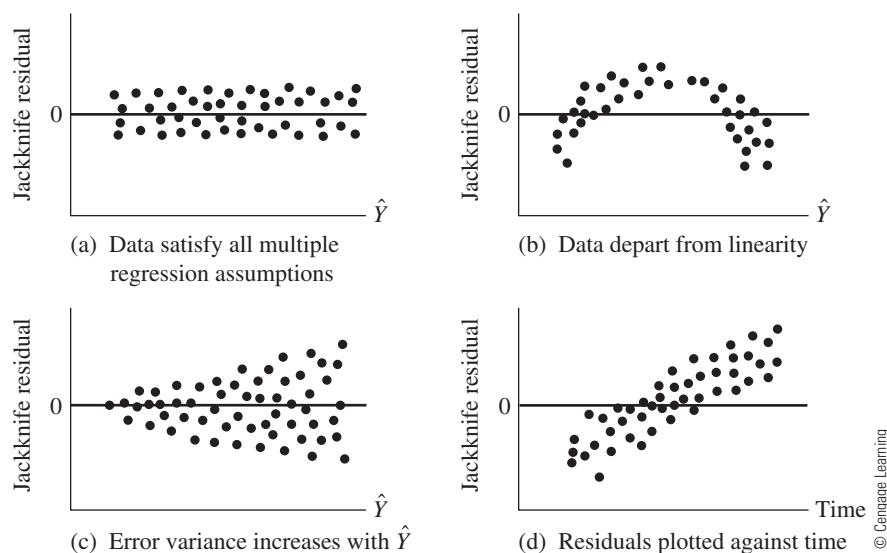


FIGURE 14.6 Typical jackknife residual plots as a function of predicted value, \hat{Y} , or time of data collection for hypothetical data

Figure 14.6(a) illustrates the type of pattern to be expected when all basic assumptions hold: since the assumption is that model errors are independent with mean 0 and homogeneous variance, a random scatter around the horizontal line at residual = 0 should be obtained with no hint of any systematic trends.

Figure 14.6(b) illustrates a systematic pattern that could occur when the data depart from linearity. In this case, the plot indicates that current variables in the model may need to be transformed or that additional curvilinear terms may need to be added to introduce curvature. Naturally, different types of model inappropriateness result in different residual patterns and different corrective actions.

The pattern in Figure 14.6(c) shows that the variance of the residuals is not constant, suggesting a violation of the homoscedasticity assumption. Transformations of the data, or advanced techniques such as weighted least-squares analysis, are often used to address heteroscedasticity (see Section 14.4).

Figure 14.6(d) is a plot of the residuals versus time. A time-related effect is clearly present. This may reflect the omission of a covariate—namely, TIME. But it may instead be suggestive of a violation of the independence assumption. There are few simple remedies for violations of the independence assumption. If corrective action does become necessary, advanced

methods (e.g., time-series analysis or correlated data analysis methods [see Chapters 25 and 26]) may be helpful in completing the analysis.

When a residual plot indicates possible assumption violations, it is useful to create further plots of the residuals versus each predictor in order to identify the specific predictors involved in the violations. Any model changes undertaken to correct violations would then focus on the dependent variable and/or on the identified predictors. Note that the residual plots also provide one more opportunity to detect outliers.

It is important to note that, sometimes, patterns apparent in these plots are not due to assumption violations but rather to small sample size or sparse data for certain combinations of predictor values. The best way to avoid such a situation is, of course, to have a large enough sample size and to design the study in such a way that a sufficient number of replicate observations is taken at each predictor value. Unfortunately, researchers usually do not have this level of control over the sampled predictor variable values in observational studies. We suggest that analysts look only for very distinct patterns or *gross* assumption violations in these plots rather than focusing on more subtle patterns and violations.

■ Example 14.3 Figure 14.7 shows the jackknife residual plot for the WEIGHT/HEIGHT/AGE example. The plot appears to reflect random scattering around the value zero, so that no assumption violations are readily apparent. One outlier is apparent (circled); further inspection will reveal that it is observation 127 (already identified in Sections 14.2 and 14.3).⁸

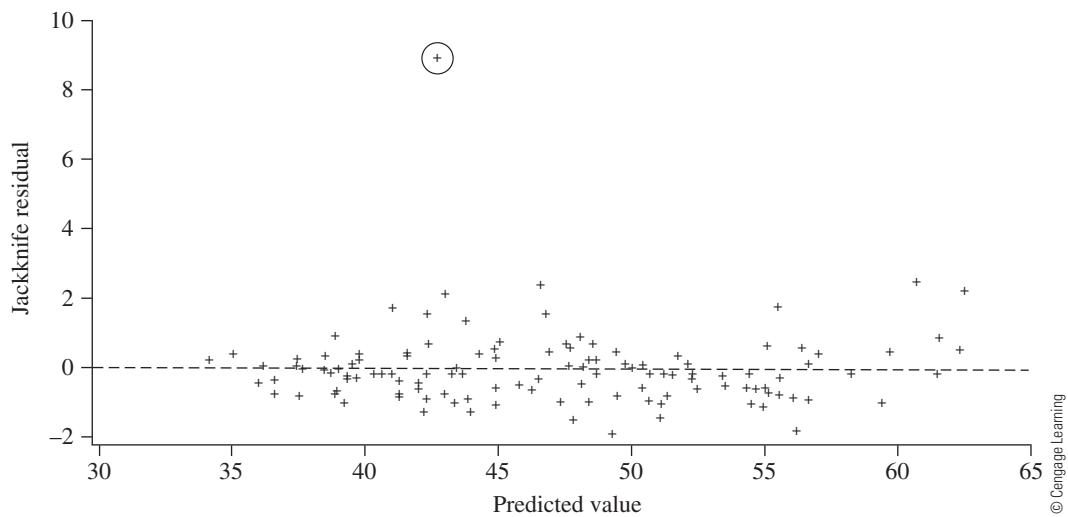


FIGURE 14.7 Jackknife residual plot for regression of children's body weight as a function of height and age; Lewis and Taylor data (1967) ($n = 127$)

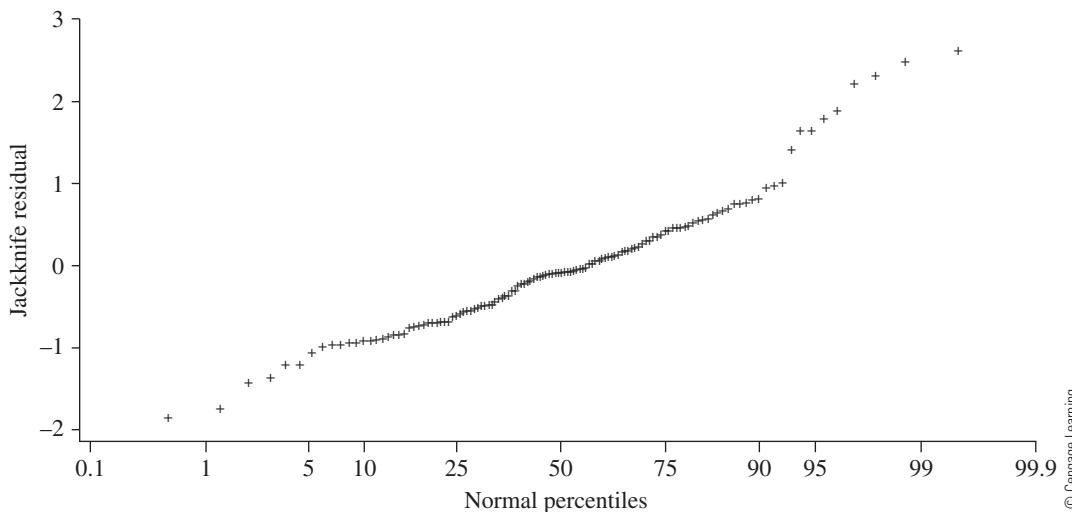
© Cengage Learning

⁸ Once again, these plots should not be overinterpreted (see footnote 4, earlier in this chapter). Some investigators might suggest that the variance of the residuals is not constant based on Figure 14.7. However, this impression is likely the result of the sparseness of the data in some areas on the plot.

14.3.3 Assessing the Normality Assumption

The normality assumption can be assessed by examining statistics such as skewness and kurtosis, as well as figures such as normal probability plots of ordinary, studentized, or jackknife residuals (see Chapter 3 for a review of these statistics and plots). Furthermore, goodness-of-fit tests such as the Kolmogorov–Smirnov test (Stephens 1974) and, in the case of small sample sizes (e.g., $n < 50$), the Shapiro–Wilks (1965) test, can be performed. These methods are covered in many introductory statistics courses and textbooks and can be easily implemented using SAS and other standard statistical software.

■ **Example 14.4** For the regression of WEIGHT on HEIGHT and AGE using the data of Example 14.1, the estimated residuals were computed and, using the UNIVARIATE procedure in SAS, descriptive statistics produced (shown below). The output shows that the skewness and kurtosis statistics are quite different from the ideal value of 0, and the stem-and-leaf and box-and-whisker plots also reflect strongly skewed data. The Kolmogorov–Smirnov test for normality has a small P -value (.0264), indicating that the null hypothesis of normality is to be rejected at a significance level of $\alpha = .05$. However, upon further inspection, this lack of normality is due to the one extreme outlier in the data set (observation 127). Without this outlier, the skewness statistic value ($= 0.87$) is reasonably close to 0, and the kurtosis statistic value ($= 1.37$) is consistent with values for t distributions (which have slightly heavier tails than normal distributions). The normal probability plot in Figure 14.8 also reveals a fairly linear pattern consistent with normality, the slight curvature at the extremes being, in part, an indication that the jackknife residuals follow a t distribution rather than a normal distribution. No gross violation of the normality assumption seems to be present for these data after eliminating the extreme outlier.



© Cengage Learning

FIGURE 14.8 SAS normal probability plot of jackknife residuals for regression of children's body weight as a function of height and age using 126 observations (excluding the outlier); Lewis and Taylor data (1967) ($n = 127$)

Edited SAS Output (PROC UNIVARIATE) for Example 14.4

The UNIVARIATE Procedure

Variable: RESIDUAL (Residual)

Moments			
N	127	Sum Weights	127
Mean	0	Sum Observations	0
Std Deviation	6.92312568	Variance	47.9296692
Skewness	2.70909774	Kurtosis	14.986965
Uncorrected SS	6039.13832	Corrected SS	6039.13832
Coeff Variation	.	Std Error Mean	0.61432806

TESTS FOR NORMALITY				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.709149	Pr < W	0.0011
Kolmogorov-Smirnov	D	0.278162	Pr > D	0.0264

© Cengage Learning

14.4 Strategies for Addressing Violations of Regression Assumptions

When one or more basic assumptions of regression are clearly not satisfied and/or when numerical problems are identified, the analyst may want to turn to alternate strategies of analysis. In the subsections that follow, we briefly describe transformations of variables that may remedy certain assumption violations, we list some alternatives to classical linear regression methods, and we briefly mention generalizations of multiple linear regression that may be appropriate in some applications.

14.4.1 Transformations

There are three primary reasons for using data transformations: (1) *to stabilize* the variance of the dependent variable if the homoscedasticity assumption is violated; (2) *to normalize* (i.e., to transform to the normal distribution) the dependent variable if the normality assumption is noticeably violated; (3) *to linearize* the regression model if the original data suggest a model that is nonlinear in the regression coefficients. Fortunately, the same transformation often helps to accomplish the first two goals, and sometimes even the third, rather than achieving one goal at the expense of one or both of the other two.

To evaluate whether a particular transformation of Y is appropriate, the linear regression equation is first fit to the transformed version of the dependent variable (such as $\sqrt{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$). The regression diagnostic procedures discussed earlier are then re-run on the new model. It is important to conduct all procedures again because it is possible to violate formerly upheld assumptions while trying to correct previously violated ones. If the transformed model meets all regression assumptions, it may be deemed suitable.

A more complete discussion of the properties of various transformations can be found in Armitage (1971), Draper and Smith (1998), and Kutner et al. (2004). In addition, Box and Cox (1964) describe an approach to making an exploratory search for one of a family of transformations (see also Box and Cox [1984] and Carroll and Ruppert [1984]). Nevertheless, we consider it useful to describe a few of the more commonly used transformations:

1. The *log transformation* ($Y' = \log Y$), where $Y > 0$, is used to stabilize the variance of Y if the variance increases markedly with increasing Y ; to normalize the dependent variable if the distribution of the residuals for Y is positively skewed; and to linearize the regression model if the relationship of Y to some independent variable suggests a model with consistently increasing slope (e.g., an exponential relationship). While any choice of logarithmic base will similarly achieve these objectives, \log_{10} and \log_e (i.e., \ln , the natural log) are most commonly used.
2. The *square root transformation* ($Y' = \sqrt{Y}$), where $Y \geq 0$, is used to stabilize the variance if the variance is proportional to the mean of Y . This is particularly appropriate if the dependent variable has the Poisson distribution.
3. The *square transformation* ($Y' = Y^2$) is used to stabilize the variance if the variance decreases with the mean of Y ; to normalize the dependent variable if the distribution of the residuals for Y is negatively skewed; and to linearize the model if the original relationship with some independent variable is curvilinear downward (i.e., if the slope consistently decreases as the independent variable increases).
4. The *arcsin transformation* ($Y' = \arcsin \sqrt{Y} = \sin^{-1} \sqrt{Y}$), where $Y \geq 0$, is used to stabilize the variance if Y is a proportion or rate.

Sometimes the data are poorly described by a multiple linear regression model, and model assumptions may be violated even after applying transformations. Other types of regression models are available that allow nonlinear relationships between predictors and the response and permit the consideration of responses that are not continuous. Logistic and Poisson regression are two such options, and their use will be discussed in Chapters 22–24.

14.4.2 Weighted Least-Squares Analysis

The *weighted least-squares* method of analysis is a modification of standard regression analysis procedures that is used when a regression model is to be fit to a set of data for which the assumptions of variance homogeneity and/or independence do not hold. We shall briefly describe here the weighted least-squares approach for dealing with variance heterogeneity. For discussions of the general method of weighted regression (which incorporate a discussion of treating nonindependence), see Draper and Smith (1998) and Kutner et al. (2004).

Weighted least-squares analysis can be used when the variance of Y varies for different values of the independent variable(s), provided that these variances (i.e., σ_i^2 for the i th observation on Y) are known or can be assumed to be of the form $\sigma_i^2 = \sigma^2/W_i$, where the weights $\{W_i\}$ are known. The methodology then involves determining the regression coefficients $\hat{\beta}'_0, \hat{\beta}'_1, \dots, \hat{\beta}'_k$ that minimize the expression

$$\sum_{i=1}^n W_i(Y_i - \hat{\beta}'_0 - \hat{\beta}'_1 X_{i1} - \hat{\beta}'_2 X_{i2} - \cdots - \hat{\beta}'_k X_{ik})^2$$

where the weight W_i is given by $1/\sigma_i^2$ (when the $\{\sigma_i^2\}$ are known) or is exactly the W_i in the expression σ^2/W_i (when this form applies).

The specific weighted least-squares solution for the straight-line regression case (i.e., $Y = \beta_0 + \beta_1 X + E$) is given by the formulas

$$\hat{\beta}'_1 = \frac{\sum_{i=1}^n W_i(X_i - \bar{X}')(Y_i - \bar{Y}')}{\sum_{i=1}^n W_i(X_i - \bar{X}')^2}$$

and

$$\hat{\beta}'_0 = \bar{Y}' - \hat{\beta}'_1 \bar{X}'$$

in which

$$\bar{Y}' = \frac{\sum_{i=1}^n W_i Y_i}{\sum_{i=1}^n W_i} \quad \text{and} \quad \bar{X}' = \frac{\sum_{i=1}^n W_i X_i}{\sum_{i=1}^n W_i}$$

Under the usual normality and mutual independence assumptions for the Y_i 's, the same general procedures are applicable as are used in the unweighted least-squares case regarding t tests, F tests, and confidence intervals about the various regression parameters. For example, to test $H_0: \beta_1 = 0$, we may use the following test statistic:

$$T = \frac{\hat{\beta}'_1 - 0}{S'_{Y|X}/S'_X \sqrt{n-1}} \sim t_{n-2} \text{ under } H_0$$

in which

$$S'_{Y|X}^2 = \frac{1}{n-2} \sum_{i=1}^n W_i(Y_i - \hat{\beta}'_0 - \hat{\beta}'_1 X_i)^2$$

and

$$S'_X^2 = \frac{1}{n-1} \sum_{i=1}^n W_i(X_i - \bar{X}')^2$$

14.5 Collinearity

Detection of collinearity is the last component of the regression diagnostic methods that we will discuss, but it is by no means the least important. As mentioned earlier, collinearity exists when there are strong linear relationships among *independent variables* that produce unreliable and unstable parameter estimates and standard errors. We begin with simple examples illustrating the symptoms and effects of collinearity, after which we present in more detail the mathematics behind the collinearity problem. We then present three approaches for diagnosing the presence of collinearity and some simple remedies for the problem.

14.5.1 Simple Examples Illustrating Collinearity

For the data of Lewis and Taylor (1967), consider the following three competing models:

$$\text{WEIGHT} = \beta_0 + \beta_1 \text{HEIGHT} + \beta_2 \text{AGE} + E \quad (\text{Model 1})$$

$$\text{WEIGHT} = \beta'_0 + \beta'_1 \text{HEIGHT} + \beta'_2 \text{AGE} + \beta'_3 \text{AGE}^2 + E \quad (\text{Model 2})$$

$$\text{WEIGHT} = \beta''_0 + \beta''_1 \text{HEIGHT} + \beta''_2 \text{AGE}_p + \beta''_3 (\text{AGE}_p)^2 + E \quad (\text{Model 3})$$

We have already conducted several diagnostics analyses for the first model. Users who perceived a slight curvilinear relationship in Figure 14.1 may be interested in considering model 2. The third model is similar to the second, except that we will fit it using data for AGE that have been perturbed slightly by adding a randomly generated value between -0.25 and 0.25 to each AGE value. Note that this perturbation changes individual AGE values to AGE_p values by no more than 2%, an amount that is unimportant, practically speaking, and that causes virtually no change in summary statistics for AGE and AGE_p. Therefore, such a perturbation should not yield regression results that are very different from the original data if the regression model is “good” (i.e., is valid and precise). The parameter estimates and their standard errors for each model are shown in Table 14.2.

Comparing models 1 and 2, we see that the estimates of the regression coefficient for AGE differ in several important ways: the signs of the estimates are different, the magnitudes are considerably different, and the standard error of the estimate is very large in model 2. As a result, the regression coefficient is insignificant (i.e., is not significantly different from 0) in model 2, even though the corresponding coefficient in model 1 is significant. Indeed, a multiple partial test for the joint significance of the AGE and AGE² coefficients, controlling

TABLE 14.2 Parameter estimates for three regression models based on Lewis and Taylor data (1967) ($n = 127$)

Parameter	Parameter Estimates (Std. Errors)					
	Model 1 (Without AGE ²)		Model 2 (With AGE ²)		Model 3 (AGE Values Perturbed)	
Intercept	−39.0*	(9.0)	52.9	(54.6)	66.5	(54.0)
Regression Coefficient						
For HEIGHT	31.6*	(8.6)	34.0*	(8.7)	34.6*	(8.4)
For AGE	2.7*	(0.6)	−11.0	(8.0)	−13.1*	(7.9)
For AGE ²	—		0.5*	(0.3)	0.6*	(0.3)

*Statistically significant at the 10% significance level (two-tailed test).

© Cengage Learning

for HEIGHT, in model 2 turns out to be nonsignificant, indicating that a child's age is not important, even though it was important in model 1. It is impossible to discern the true nature of the association between AGE and WEIGHT because of these inconsistencies. Such inconsistencies often occur in the presence of collinearity and are an indication of the instability caused by this undesirable condition.

In this example, it is easily concluded that collinearity exists because of the obvious association between the predictors AGE and AGE² (which can be regarded as having a linear component). However, it is not always this easy to diagnose the presence of collinearity or to identify predictors involved in collinearity problems; in Section 14.5.3, we present techniques that can be used for this purpose.

A comparison of models 2 and 3 further highlights the instability that collinearity can cause. The models are the same and are fitted with data that are similar except for a slight, practically meaningless perturbation of the children's ages. However, the coefficient estimate for AGE in model 3 is almost 20% more negative than in model 2 (a considerable difference) and is now statistically significant. The fact that the regression results for AGE can be so different for data that are not very different should be of concern. In actual analyses, models are usually not refitted with perturbed data; therefore, this aspect of the serious instability due to collinearity may go undetected.

14.5.2 Mathematical Concepts in Collinearity

In this section, we explore some of the mathematical concepts related to the collinearity problem. This will be helpful in understanding the statistics used to diagnose the presence of collinearity; these statistics are presented in Section 14.5.3. We recognize that some applied regression analysts may not require the detail presented here and that others may be very interested in delving more deeply into the mathematical roots of the collinearity problem. In the former case, readers may simply skip this subsection; in the latter, we refer readers to the detailed discussion of the topic in the book by Belsley, Kuh, and Welsch (1980).

Collinearity with Two Predictors

The problems emanating from collinearity can be illustrated with simple two-variable regression examples. Consider fitting the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + E_i$$

to produce $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$. In general, we can show that

$$\hat{\beta}_j = c_j \left[\frac{1}{1 - r_{X_1 X_2}^2} \right]$$

for $j = 1$ or $j = 2$. Here c_j is a value that depends on the data, and $r_{X_1 X_2}^2$ is the squared correlation between X_1 and X_2 . In turn,

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

so

$$\bar{Y} - \hat{\beta}_0 = \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2 = \left[\frac{1}{1 - r_{X_1 X_2}^2} \right] (c_1 \bar{X}_1 + c_2 \bar{X}_2)$$

Here \bar{X}_1 and \bar{X}_2 are the means of X_1 and X_2 , respectively. These expressions tell us that $\hat{\beta}_1$, $\hat{\beta}_2$, and $(\bar{Y} - \hat{\beta}_0)$ are all proportional to $1/[1 - r_{X_1 X_2}^2]$. To appreciate the implications of this, it is informative to consider fitting the model

$$Y_i = \theta_0 + \theta_1 X_{i1} + \theta_2 X_{i1} + E_i$$

In this case, a single variable, X_1 , is included in the model twice. What are the estimates of the regression coefficients? Since $r_{X_1 X_2}^2 = r_{X_1 X_1}^2 = 1$, it follows that $1 - r_{X_1 X_2}^2 = 0$ and

$$\hat{\theta}_j = c'_j \left(\begin{matrix} 1 \\ 0 \end{matrix} \right) = ?$$

From this, we conclude that the estimates of the regression coefficients are indeterminate. Since the estimates of the variances of the regression coefficients are proportional to the “inflation factor” $1/(1 - r_{X_1 X_2}^2)$, they are indeterminate as well. In turn, the P -values for tests about the coefficients are also indeterminate, since they involve the estimates just discussed. The preceding model can be rewritten in the form

$$Y_i = \theta_0 + (\theta_1 + \theta_2) X_{i1} + E_i$$

which establishes that an infinite number of pairs of θ_1 and θ_2 values add up to the same coefficient value; thus, an estimate of the coefficient of X_1 —say, $\widehat{\theta_1 + \theta_2}$ —does not permit a unique determination of the individual estimates $\hat{\theta}_1$ and $\hat{\theta}_2$. Consequently, θ_1 and θ_2 cannot be estimated separately. For example, if the actual estimate for the slope of X_1 is 12.5, then $\hat{\theta}_1 = 12.5$, $\hat{\theta}_2 = 0$ are possible components, as are $\hat{\theta}_1 = 0$, $\hat{\theta}_2 = 12.5$ and $\hat{\theta}_1 = 6$, $\hat{\theta}_2 = 6.5$,

and so on. In this extreme example, one of the predictor variables is a *perfect linear combination* of the other—namely,

$$X_1 = \alpha + \beta X_2 = 0 + (1)X_2 = X_2$$

Geometrically, the points (X_{i1}, X_{i2}) all fall on the straight line $X_1 = X_2$; hence, the term *collinear* is applicable.

The following data provide a slightly more general example:

X_1	X_2
0	3
1	5
3	9
4	11
7	17

Plotting X_2 against X_1 yields a very simple picture, as shown in Figure 14.9. The plot illustrates that all of the data points (X_{i1}, X_{i2}) fall exactly on the straight line $X_2 = 3 + 2X_1$, which demonstrates that X_1 and X_2 are exactly collinear. Moreover, $r^2_{X_1X_2} = 1$ directly measures this perfect collinearity. The collinearity issue involves only the predictor variables and does not depend on the relationship between the response and any of the predictors. As $r^2_{X_1X_2}$ decreases, the collinearity problem between X_1 and X_2 becomes less severe, with the ideal situation occurring when X_1 and X_2 are uncorrelated.

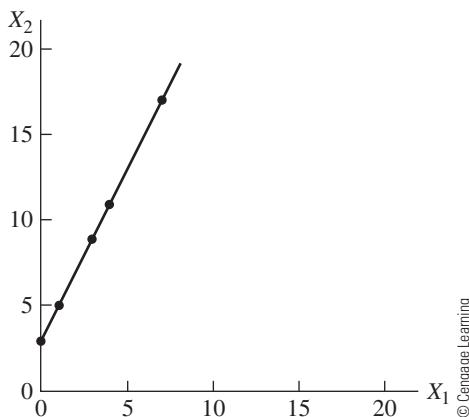


FIGURE 14.9 Perfectly collinear set of pairs of variable values: $X_2 = 3 + 2X_1$

In general, for a two-variable model, if $r^2_{X_1X_2}$ is nearly 1.0, then collinearity is likely present. Although the regression coefficient estimates can be computed, they are highly unstable. In particular, this instability is reflected in large estimates of the coefficient variances, since such variance estimates are proportional to $1/(1 - r^2_{X_1X_2})$. As $r^2_{X_1X_2}$ gets closer to 1.0, this factor

becomes large, thereby inflating the estimated variances of the regression coefficients. Later we will see that this factor is a special case of a variance inflation factor.

Collinearity Concepts

We now generalize the discussion to treat any number of predictors. In addition, we describe methods for quantifying the degree of collinearity in fitting a regression model to a particular set of data.

As we observed in the examples just discussed, collinearity involves relationships among the predictor variables and does not directly involve the response variable. As such, one informative way to examine collinearity is to consider what happens if each predictor variable is treated as the response variable in a multiple regression model in which the independent variables are all of the remaining predictors. For k predictors, then, we have k such models. For example, with the four-predictor model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + E_i$$

four models would be fitted as follows:

$$\begin{aligned} X_{i1} &= \alpha_{01} + \alpha_{21} X_{i2} + \alpha_{31} X_{i3} + \alpha_{41} X_{i4} + E_i \\ X_{i2} &= \alpha_{02} + \alpha_{12} X_{i1} + \alpha_{32} X_{i3} + \alpha_{42} X_{i4} + E_i \\ X_{i3} &= \alpha_{03} + \alpha_{13} X_{i1} + \alpha_{23} X_{i2} + \alpha_{43} X_{i4} + E_i \\ X_{i4} &= \alpha_{04} + \alpha_{14} X_{i1} + \alpha_{24} X_{i2} + \alpha_{34} X_{i3} + E_i \end{aligned}$$

To assess collinearity, we need to know the associated R^2 -values based on fitting these four models—namely, $R^2_{X_1|X_2, X_3, X_4}$, $R^2_{X_2|X_1, X_3, X_4}$, and $R^2_{X_4|X_1, X_2, X_3}$. If any of these multiple R^2 -values equals 1.0, a perfect collinearity is said to exist among the set of predictors. The term *collinearity* is used to indicate that one of the predictors is nearly an exact linear combination of the others. As described in earlier examples, in the special case $k = 2$, perfect collinearity means that X_1 is a straight-line function of X_2 —say, $X_1 = \alpha + \beta X_2$ —so that the points (X_{i1}, X_{i2}) lie on a straight line.

Consider, for example, predicting college grade-point average (CGPA) on the basis of high school grade-point average (HGPA) and College Board SAT scores on the mathematics (MATH) and verbal (VERB) sections. The model is

$$\text{CGPA}_i = \beta_0 + \beta_1 \text{HGPA}_i + \beta_2 \text{MATH}_i + \beta_3 \text{VERB}_i + E_i$$

Now imagine trying to improve prediction by adding the total (combined) board scores ($\text{TOT} = \text{MATH} + \text{VERB}$) to create a new model:

$$\text{CGPA}_i = \alpha_0 + \alpha_1 \text{HGPA}_i + \alpha_2 \text{MATH}_i + \alpha_3 \text{VERB}_i + \alpha_4 \text{TOT}_i + E_i$$

This model has a perfect collinearity, which means that the parameters in the model cannot be estimated uniquely. To see this, begin by rewriting the model as

$$\text{CGPA}_i = \alpha_0 + \alpha_1 \text{HGPA}_i + \alpha_2 \text{MATH}_i + \alpha_3 \text{VERB}_i + \alpha_4 (\text{MATH}_i + \text{VERB}_i) + E_i$$

It follows that

$$\text{CGPA}_i = \alpha_0 + \alpha_1 \text{HGPA}_i + (\alpha_2 + \alpha_4) \text{MATH}_i + (\alpha_3 + \alpha_4) \text{VERB}_i + E_i$$

With this version of the model, consider choosing $\alpha_4 = 0$. Then $\alpha_2 = \beta_2$ and $\alpha_3 = \beta_3$ give the correct original model. Next, choose $\alpha_4 = 3$. Then $\alpha_2 = \beta_2 - 3$ and $\alpha_3 = \beta_3 - 3$ also give the correct original model. In fact, for any choice of α_4 , we can choose values for α_2 and α_3 that provide the correct model. Since the α_4 parameter is irrelevant under the circumstances (it could best be set equal to 0), a model containing a perfect collinearity is sometimes said to be *overparameterized*.

Near collinearity arises if the multiple R^2 -value relating one predictor with the remaining predictors is nearly 1. For a general model involving k predictors such as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + E_i$$

the multiple R^2 -value of interest for the first predictor is $R_{X_1|X_2, X_3, \dots, X_k}^2$, the multiple R^2 -value of interest for the second predictor is $R_{X_2|X_1, X_3, \dots, X_k}^2$, and so on. These quantities are generalizations of the statistic $r_{X_1 X_2}^2$ for a $k = 2$ variable model. For convenience, we denote by R_j^2 the squared multiple correlation based on regressing X_j on the remaining $(k - 1)$ predictors.

The *variance inflation factor* (VIF) is often used to measure collinearity in a multiple regression analysis. It may be computed as

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \quad j = 1, 2, \dots, k$$

The quantity VIF_j generalizes the variance inflation factor for a two-predictor model, $1/(1 - r_{X_1 X_2}^2)$. Since $0 \leq r_{X_1 X_2}^2 \leq 1$, $\text{VIF}_j \geq 0$. As for the two-variable case, the estimates of the variances for the regression coefficients are proportional to the VIFs—namely,

$$S_{\hat{\beta}_j}^2 = c_j^*(\text{VIF}_j) \quad j = 1, 2, \dots, k$$

This expression suggests that the larger the value of VIF_j , the more troublesome the variable X_j is. A rule of thumb for evaluating VIFs is to be concerned with any value larger than 10.0; this rule is equivalent to $R_j^2 > 0.90$ or, approximately, to $R_j > 0.95$. Some people prefer to consider

$$\text{Tolerance}_j = \frac{1}{\text{VIF}_j} = 1 - R_j^2$$

The choice among R_j^2 , $1 - R_j^2$, and VIF_j is a matter of personal preference, since they all contain exactly the same information. As R_j^2 goes to 1.0, the tolerance $(1 - R_j^2)$ goes to 0, and VIF_j goes to infinity.

Because of its special nature, the intercept requires separate treatment in evaluating collinearity. For the general model involving k predictors,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + E_i$$

we find regression coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$. The intercept estimate can be expressed simply as

$$\hat{\beta}_0 = \bar{Y} - (\hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2 + \dots + \hat{\beta}_k \bar{X}_k) = \bar{Y} - \sum_{j=1}^k \hat{\beta}_j \bar{X}_j$$

From this, we can deduce that the estimated intercept is affected by the VIF_j 's, $j = 1, 2, \dots, k$, since it is a function of the $\hat{\beta}_j$'s. The problem disappears if the means of all X_j 's are 0 (e.g., if the predictor data are centered). In that case, \bar{Y} is the estimated intercept.

In general, even if the predictor data are not centered, a variance inflation factor VIF_0 for $\hat{\beta}_0$ can be defined and interpreted in the same way as VIF_j . First, define

$$VIF_0 = \frac{1}{1 - R_0^2}$$

Here R_0^2 is the generalized squared multiple correlation for the regression model

$$I_i = \alpha_1 X_{i1} + \alpha_2 X_{i2} + \dots + \alpha_k X_{ik} + E_i$$

in which I_i is identically 1 (which may be thought of as the value for the intercept variable). No intercept is included in this model (as a predictor), which is why R_0^2 is called a *generalized* squared correlation. As with the VIF_j 's, $VIF_0 \geq 0$, and

$$S_{\hat{\beta}_0}^2 = c_0^*(VIF_0)$$

Hence, the interpretation for VIF_0 is the same as for VIF_j , $j = 1, 2, \dots, k$.

Controversy surrounds the treatment of the intercept in regression diagnostics. For some, it is simply another predictor; for others, it should be eliminated from discussion. We take a middle position, arguing that the model and the data at hand determine the role of the intercept. (See, for example, formal commentaries following the paper by Belsley 1984.) This leads us to discuss diagnostics both with and without the intercept included in the regression model, corresponding to the cases with and without centering of the predictors and response variables.

The presence of collinearity or (more typically) near collinearity causes computational difficulties in calculating numerically reliable estimates of the R_j^2 -values, tolerances, and VIF_j 's using standard regression procedures. This apparent impasse can be solved in at least three ways. The first way is to use computational algorithms that detect collinearity problems as they arise in the midst of the calculations. A discussion of such algorithms is beyond the scope of this text.

A second way to avoid the impasse is to scale the data appropriately. By *scaling*, we mean the choice of measurement unit (e.g., degrees Celsius versus degrees Fahrenheit) and the choice of measurement origin (e.g., degrees Celsius versus degrees Kelvin). We shall consider only linear changes in scale, such as

$$X_1 = \alpha + \beta X_2$$

$$C = \frac{5}{9}(F - 32) = -32\left(\frac{5}{9}\right) + \left(\frac{5}{9}\right)(F)$$

or

$$K = (-273) + (1)(C)$$

where C , F , and K are temperatures in degrees Celsius, degrees Fahrenheit, and degrees Kelvin, respectively.

Often scaling refers just to multiplying by a constant rather than also to adding or subtracting a constant. An example of scaling is the conversion from feet to inches. One important case of subtracting a constant, a form of scaling, is *centering*. A set of values $\{X_{ij}\}$ for predictor X_j is centered by subtracting the mean \bar{X}_j of the values for predictor X_j from each individual value for that predictor, giving

$$X_{ij}^* = X_{ij} - \bar{X}_j$$

in which $\bar{X}_j = \sum_{i=1}^n X_{ij}/n$ for predictor X_j .

Computing standardized scores (z scores) is a closely related method of scaling. In particular, the standardized score corresponding to X_{ij} is

$$z_{ij} = \frac{X_{ij} - \bar{X}_j}{S_j}$$

in which $S_j^2 = \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2/(n - 1)$. Centered and standardized scores have mean 0, since $\sum_{i=1}^n X_{ij}^* = \sum_{i=1}^n z_{ij} = 0$. Also, $\sum_{i=1}^n z_{ij}^2/(n - 1) = 1$, so the set $\{z_{ij}\}$ of standardized scores has a variance equal to 1. We shall delay the discussion of detecting and fixing scaling problems until later in this chapter. For now, we need the concept of scaling in order to understand the following discussion of other diagnostic statistics.

A third way to avoid the impasse created by collinearity and near collinearity is to use alternate computational methods to diagnose collinearity. One especially popular method for characterizing near and/or exact collinearities among the predictors involves computing the *eigenvalues* of the predictor variable *correlation matrix*. The eigenvalues are connected with the *principal component analysis* of the predictors. The principal components of the predictors are a set of new variables that are linear combinations of the original predictors. These components have two special properties: they are not correlated with each other; and each, in turn, has maximum variance, given that all components are mutually uncorrelated. The principal components provide idealized predictor variables that still retain all of the same information as the original variables. The variances of these components (the new variables) are called *eigenvalues*. The larger the eigenvalue, the more important the associated principal component is in representing the variation in the predictors. As an eigenvalue approaches zero, the presence of a near collinearity among the original predictors is indicated. The presence of an eigenvalue of exactly 0 means that a perfect linear dependency (i.e., an exact collinearity) exists among the predictors.

If a set of k predictor variables does *not* involve an *exact* collinearity, then k principal components are needed to partition all the variability contained in the original set of predictor variables. If one of the predictors is a perfect linear combination of the others, then only $(k - 1)$ principal components are needed to describe all of the variation in the predictors. The number of zero (or near-zero) eigenvalues is the number of collinearities (or near collinearities) among the predictors. Even a single eigenvalue near 0 presents a serious problem that must be resolved.

In using the eigenvalues to determine the presence of near collinearity, researchers usually employ three kinds of statistics: the *condition index* (CI), the *condition number* (CN), and the *variance proportions*.

Consider again the general linear model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + E_i$$

Often collinearity is assessed by considering only the k parameters β_1 through β_k and ignoring the intercept β_0 . This is accomplished by centering the response and predictor variables, which corresponds to fitting the model

$$Y_i - \bar{Y} = \beta_1(X_{i1} - \bar{X}_1) + \beta_2(X_{i2} - \bar{X}_2) + \cdots + \beta_k(X_{ik} - \bar{X}_k) + E_i$$

Noting that $\hat{\beta}_0 = \bar{Y} - \sum_{j=1}^k \hat{\beta}_j \bar{X}_j$ in the original model, we observe that centering the predictors and the response forces the estimated intercept in that centered model to be 0. Hence, it can be dropped from that model.

It is also common to assess collinearity after centering and standardizing both the predictors and the response. This leads to the standardized model

$$\frac{Y_i - \bar{Y}}{S_Y} = \beta_1^* \frac{(X_{i1} - \bar{X}_1)}{S_1} + \beta_2^* \frac{(X_{i2} - \bar{X}_2)}{S_2} + \cdots + \beta_k^* \frac{(X_{ik} - \bar{X}_k)}{S_k} + E_i^*$$

The coefficients for this standardized model are often called the *standardized regression coefficients*; in particular,

$$\beta_j^* = \beta_j \left(\frac{S_j}{S_Y} \right) \quad j = 1, 2, \dots, k$$

Table 14.3 summarizes an eigenanalysis of the predictor correlation matrix R_{xx} for a hypothetical four-predictor ($k = 4$) standardized regression model. With $k = 4$ predictors, four eigenvalues (denoted by λ 's) exist. Later we will include the intercept in the analysis and have $(k + 1)$ eigenvalues. In the present case, $\lambda_1 = 2.0$, $\lambda_2 = 1.0$, $\lambda_3 = 0.6$, and $\lambda_4 = 0.4$. (The sum of the eigenvalues for a correlation matrix involving k predictors is *always* equal to k .)

TABLE 14.3 Eigenanalysis of the predictor correlation matrix for a hypothetical four-predictor standardized regression model

Principal Component	Eigenvalue	Condition Index	Variance Proportions			
			X_1	X_2	X_3	X_4
1	2.0	1.00	0.09	0.11	0.08	0.15
2	1.0	1.41	0.32	0.10	0.25	0.07
3	0.6	1.87	0.40	0.52	0.12	0.13
4	0.4	2.24	0.19	0.27	0.55	0.65

It is customary to list the eigenvalues from largest (λ_1) to smallest (λ_k). A *condition index* can be computed for each eigenvalue as

$$CI_j = \sqrt{\lambda_1/\lambda_j}$$

In particular, CI_3 for this example is $\sqrt{2.0/0.6} = 1.87$. The first (largest) eigenvalue always has an associated condition index (CI_1) of 1.0. The largest CI_j , called the *condition number* (CN), always involves the largest (λ_1) and smallest (λ_k) eigenvalues. It is given by the formula

$$CN = \sqrt{\lambda_1/\lambda_k}$$

For this example, $CN = \sqrt{2.0/0.4} = 2.24$. Since eigenvalues are variances, CI_j and CN are ratios of standard deviations (of principal components, the idealized predictors). Like VIFs, values of CI_j and CN are nonnegative, and larger values suggest potential near collinearity. Belsley, Kuh, and Welsch (1980) recommended interpreting a CN of 30 or more as reflecting moderate to severe collinearity, worthy of further investigation. Of course, such a CN may be associated with two or more CI_j 's that are greater than or equal to 30.

A *variance proportion* indicates, for each predictor, the proportion of the estimated variance of its estimated regression coefficient that is associated with a particular principal component. The variance proportions suggest collinearity problems if more than one predictor has a high variance proportion (loads highly) on a principal component having a high condition index. The presence of two or more proportions of at least 0.5 for such a component suggests a problem. One should definitely be concerned when two or more loadings greater than 0.9 appear on a component with a large condition index.

Table 14.3 includes variance proportions for the hypothetical four-predictor model under consideration. The entries represent a typical pattern of condition indices and proportions for a regression analysis with no major collinearity problems. Each column sums to a total proportion of 1.00 because each estimated regression coefficient has its own estimated variance partitioned among the four components. The last row involves two loadings greater than 0.5, corresponding to the smallest eigenvalue; but since the smallest eigenvalue has a CI of only 2.24 (far from the suggested warning level of 30.0 for moderate to severe collinearity), the proportion pattern does not indicate a major problem.

The intercept can play an important role in collinearity. In defining regression models, researchers follow the standard practice of including the intercept term β_0 . The intercept is a regression coefficient for a variable that has a constant value of 1. Consequently, any variable with near-zero variance will be nearly a constant multiple of the intercept and hence will be nearly collinear with it. This problem can arise spuriously when variables are improperly scaled. To evaluate this possibility, the eigenanalysis discussed earlier can be modified to include the intercept by basing it on the scaled cross-products matrix.⁹ The eigenanalysis including the intercept may suggest that this constant is nearly collinear with one or more predictors.

Centering may help decrease collinearity. From a purely theoretical perspective, statisticians disagree as to when centering helps regression calculations (for example, see Belsley

⁹ The eigenanalysis would then be based on the $(k + 1) \times (k + 1)$ cross-products matrix $\mathbf{X}'\mathbf{X}$ suitably scaled to have 1's on the diagonal rather than on the $k \times k$ correlation matrix \mathbf{R}_{XX} . The sum of the eigenvalues for this scaled cross-products matrix is equal to $(k + 1)$.

1984, 1991; Smith and Campbell 1980; and formal commentaries on the papers in the same issues). For actual computations, centering can increase numerical accuracy in many situations. Polynomial regression (Chapter 15) is one situation where centering is recommended. In the example that follows, we numerically illustrate the effects of centering on procedures for diagnosing collinearity.

14.5.3 Collinearity Diagnostics

There are three steps that should be taken to diagnose the presence of collinearity. These steps will also identify the particular predictor variables involved in the collinearity problem—information that is needed before the appropriate corrective action can be taken.

The first step is taken during the simple descriptive analyses mentioned in Section 14.1 and serves to bring the most obvious possible collinearity problems to our attention. Scatterplots of each continuous predictor versus each of the other continuous predictors, and correlations between predictors, can quickly reveal potentially strong and problematic linear associations. For example, Figure 14.5 and the SAS output (PROC CORR) for Example 14.1 suggested that there may be a collinearity problem involving HEIGHT and AGE.

Next, the VIF values should be examined for each predictor in the model. We demonstrated, in Section 14.5.1, that collinearity can result in inflated standard errors for the fitted model. As discussed in Section 14.5.3, the VIF values provide an indication of the degree to which the estimated standard errors of regression parameter estimates are affected by linear associations among predictor variables. Generally, VIF values greater than 10 indicate collinearity issues among predictor variables that are serious enough to warrant corrective action.

In the case of model 1, described in Section 14.5.1, the VIF values for HEIGHT and AGE are each equal to 2.3. These are not large enough to conclude that there is a collinearity problem needing correction in the model with two predictors, HEIGHT and AGE.

It should be noted that, in models with just two predictors, the two VIF values will be identical (see Section 14.5.2 for details on VIF calculations). In larger models, the VIF values will generally all be different, so identifying which sets of predictors are strongly linearly associated may not be easy. Indeed, it may be the case that three or more predictors are simultaneously collinear with each other. In such a situation, the analyst must use good scientific judgment to determine which predictors are collinear with each other and should also proceed to the third step of the collinearity investigation.

In the third step of the collinearity investigation, the condition index and variance proportion statistics should be examined. As described in Section 14.5.2, the condition indices for a data set describe the degree to which the data are “ill-conditioned”—that is, the degree to which small changes in data values can result in large changes in the regression parameter estimates and their standard errors. There are as many condition indices as there are parameters in the regression model. Condition indices greater than 30 reflect moderate collinearity or worse. For the largest condition index greater than 30, the variance proportion statistics should be examined in order to determine which predictor variables are primarily responsible for the large condition index. Predictor variables with variance proportions higher than 0.5 can be considered to be involved in the collinearity problem. Additionally, if the intercept is involved in the collinearity (i.e., it has a high variance proportion), it is recommended that intercept-adjusted collinearity diagnostic statistics (i.e., diagnostic statistics computed after centering all predictor and response variables) be examined rather than unadjusted statistics

using uncentered variables. Corrective action is taken (see Section 14.5.4) and the statistics are then re-examined for additional collinearity problems.

Variance inflation factors, condition indices, and variance proportions are all easily produced using SAS and other packages.

■ **Example 14.5** Consider model 2 from Section 14.5.1:

$$\text{WEIGHT} = \beta_0 + \beta_1 \text{HEIGHT} + \beta_2 \text{AGE} + \beta_3 \text{AGE}^2 + E$$

The relevant SAS output is shown below.

Edited SAS Output (PROC CORR) for Example 14.5

PEARSON CORRELATION COEFFICIENTS, N = 127			
	HEIGHT	AGE	AGE_SQUARED
HEIGHT	1.00000	0.75328	0.74648
AGE	0.75328	1.00000	0.99853
AGE_SQUARED	0.74648	0.99853	1.00000

© Cengage Learning

Edited SAS Output (PROC REG) for Example 14.5

The REG Procedure

Model: MODEL1

Dependent Variable: WEIGHT IN POUNDS

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	6103.48126	2034.49375	42.42	<.0001
Error	123	5899.48413	47.96329		
Corrected Total	126	12003			

PARAMETER ESTIMATES

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	Intercept	1	52.87854	54.59020	0.97	0.3346	0
HEIGHT	HEIGHT IN INCHES	1	33.96409	8.65471	3.92	0.0001	2.37242
AGE	AGE IN MONTHS	1	-10.98523	8.02559	-1.37	0.1736	357.96034
AGE_SQUARED		1	0.48156	0.28222	1.71	0.0905	349.71180

↑
VIF values
(continued)

COLLINEARITY DIAGNOSTICS						
Number	Eigenvalue	Condition Index	Proportion of Variation			
			Intercept	HEIGHT	AGE	AGE_SQUARED
1	3.97387	1.00000	0.00000796	0.00012761	0.00000196	0.00000790
2	0.02469	12.68626	0.00171	0.00610	0.00001001	0.00299
3	0.00142	52.93716	0.02287	0.96076	0.00145	0.00030963
4	0.00002091	435.94470	0.97541	0.03301	0.99854	0.99669

COLLINEARITY DIAGNOSTICS (INTERCEPT ADJUSTED)						
Number	Eigenvalue	Condition Index	Proportion of Variation			AGE_SQUARED
			HEIGHT	AGE	AGE_SQUARED	
1	2.67140	1.00000	0.04529	0.00037365	0.00038071	
2	0.32718	2.85743	0.91837	0.00117	0.00131	
3	0.00142	43.44709	0.03634	0.99846	0.99831	

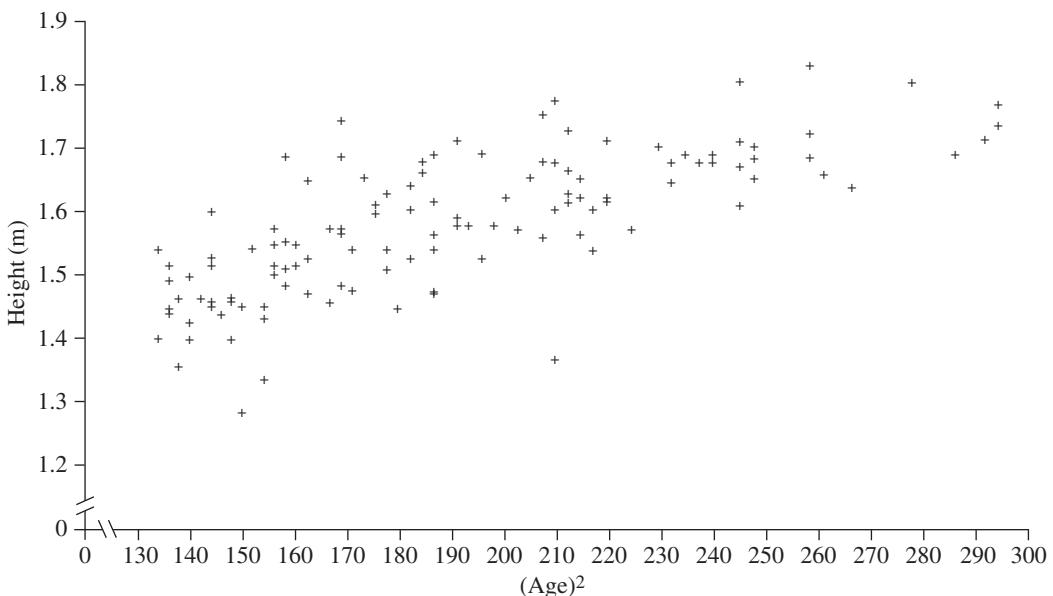
Largest condition index

Variance proportion for the intercept is high

Intercept-adjusted variance proportion

© Cengage Learning

From the scatterplot in Figure 14.10, we see that there is an approximate linear relationship between HEIGHT and AGE^2 . This is not surprising, since there was such a relationship between HEIGHT and AGE. A scatterplot of AGE versus AGE^2 is not necessary; clearly, there is a relationship, and all that remains to be determined is whether that relationship is the source of moderate or strong collinearity. The estimated correlation coefficient output (from SAS PROC CORR) confirms the strong associations.



© Cengage Learning

FIGURE 14.10 SAS scatterplot of HEIGHT versus AGE^2 for Example 14.5

The collinearity diagnostics output (from SAS PROC REG) shows VIF values of 358 and 350 for AGE and AGE², respectively. Clearly, these terms are involved in a severe collinearity problem. The VIF for HEIGHT is small, so it is not involved (even though early indications were that the HEIGHT–AGE relationship might be a problem).

For illustration purposes, we discuss the use of the condition indices, although, in this example, it is clear that the collinearity is due to the relationship between AGE and AGE². The largest condition index in the SAS output is 435.9, indicating the presence of strong collinearity. The variance proportions in the same row indicate that the intercept is involved in the collinearity (variance proportion for the intercept = 0.97541). As a result, we consider the intercept-adjusted condition indices. Here the variance proportions corresponding to the largest condition index are examined (the condition indices are ordered, so the largest index appears in the last row). The variance proportions for AGE and AGE² are high (0.99846 and 0.99831, respectively), indicating that they are involved in the collinearity problem. Corrective action should be taken as described in Section 14.5.4.

14.5.4 Treating Collinearity Problems

Collinearity problems are often easily remedied by eliminating one or more of the predictors in the collinear set. At first, this may seem to be a drastic step; however, the strategy does make sense from the point of view that, since strongly collinear variables are closely related, they are very similar and can be regarded as measuring much the same thing. In that case, there may not be a great loss of information if one variable is dropped. In Example 14.5, AGE² would probably be removed. The collinearity diagnostics should then be re-examined to ensure that the problem has been adequately addressed. If not (i.e., if VIF values or condition indices are still high), another variable may need to be removed. The process repeats until collinearity is no longer judged to be a problem.¹⁰

Decisions about which variables to remove from a collinear set can be difficult to make. Judgments should not be made on the basis of *P*-values, since hypothesis test results may be misleading in the presence of collinearity. Instead, researchers should use their scientific expertise and previous experience to identify the variables that, from a scientific point of view, would be acceptable to drop. Practical concerns, such as cost and difficulty of collecting data for each variable, may also be important to take into account. Throughout, a conservative approach should be taken that ensures that only the least scientifically interesting predictors are dropped and that the original objectives of the study can still be met.

In regression models in which *natural polynomials* or powers of a continuous predictor are included (e.g., AGE² is included along with AGE), an alternative to dropping higher-order terms to reduce the naturally occurring collinearity is to use the method of orthogonal polynomials. This topic is discussed in detail in Chapter 15.

The improper use of dummy variables may inadvertently introduce severe collinearity problems. Discussion of how to avoid this problem is found in Section 12.3.

Interaction terms generally create the atmosphere for collinearity problems, especially if such terms are overused. For example, if the predictors are AGE, HEIGHT, and AGE × HEIGHT, then collinearity may be a problem due to the close functional relationship

¹⁰ Note that it is usually best to remove just one variable from a collinear set, re-examine the collinearity diagnostics, and then remove another variable if necessary, and so on. Often, removing just one variable will greatly reduce or eliminate collinearity problems.

between the product term and the two basic predictors. The potential for collinearity grows with the inclusion of multiple higher-order interaction terms in one's model, such as second-order and third-order interactions. In general, we recommend against including three- and four-variable product terms, due to the potential for collinearity and problems in interpreting their meanings, unless there is a clear scientific basis for their inclusion.

Researchers designing studies can attempt to head off potential collinearity problems by collecting data that break the pattern of collinearity. For example, suppose that family income and years of education are to be used as predictors in a regression analysis. Since these variables are likely to be highly positively correlated, researchers can use stratified sampling schemes that deliberately sample from subpopulations of highly educated subjects who earn low incomes and higher-income subjects who have few years of education. If enough such subjects are sampled, collinearity problems between the two predictors might be lessened or eliminated. However, this approach may not be possible (for example, in observational studies with simple random sampling) and may not be successful in reducing collinearity problems, regardless of the researchers' efforts.

There are some other practical techniques that may be helpful in eliminating collinearity. In some cases, using centered data (data in which the predictor variables have been transformed by subtracting the mean value of the variable from each value) can alleviate collinearity problems. A discussion about centering was presented in Section 14.5.2. However, analysts should proceed with caution. Belsley, Kuh, and Welsch (1980) have shown that centering can, in some situations, render the usual collinearity diagnostics, VIF values and condition indices, ineffective. That is, the instability characteristic of collinearity may remain even though the VIF values and condition indices appear normal when data are centered.

Other advanced techniques, such as regression on *principal components* and the technique known as *ridge regression*, may also be considered in the treatment of collinearity, although we do not favor the use of these methods. In the analysis of principal components, the original predictor variables are replaced by a set of mutually uncorrelated variables, the principal components. If necessary, components associated with eigenvalues near zero are dropped from the analysis, thus eliminating the attendant collinearity problem. Ridge regression involves perturbing the eigenvalues of the original predictor variable cross-products matrix to push them away from zero, thus reducing the amount of collinearity. A detailed discussion of these methods is beyond the scope of this book (e.g., see Gunst and Mason 1980). Both methods lead to biased estimators of the regression coefficients under the assumed model. In addition, *P*-values for statistical tests may be optimistically small when based on such biased estimation methods.

14.6 Diagnostics Example

We now present a detailed example illustrating the implementation of the regression diagnostics methods discussed and recommended in this chapter.

■ **Example 14.6** The data for this example arise from a hypothetical experiment concerned with environmental monitoring of surface water pollutants. The concentrations (X_1 and X_2) of two pollutants, A and B, and the instrument reading (Y) for a third pollutant, C, were recorded (see Table 14.4). Instrument readings for pollutant C are

TABLE 14.4 Data for hypothetical pollutant concentration example

Observation	Pollutant A Concentration (X_1)	Pollutant B Concentration (X_2)	Instrument Reading (Y)
1	0.1	0.5	10.7
2	0.1	1.1	14.2
3	0.1	2.3	16.7
4	0.5	0.4	19.1
5	0.5	1.2	24.9
6	0.5	2.0	25.4
7	1.0	0.1	32.3
8	1.0	1.2	33.8
9	1.0	2.1	39.6
10	1.5	0.8	33.3
11	1.5	1.5	37.2
12	1.5	2.9	37.8
13	2.0	0.2	37.5
14	2.0	1.5	38.6
15	2.0	2.5	42.6
16	2.5	0.7	44.3
17	2.5	1.6	45.2
18	2.5	2.3	47.2
19	3.0	0.3	45.2
20	3.0	1.3	46.4
21	3.0	2.3	65.3
22	3.5	0.9	47.5
23	3.5	1.7	48.2
24	3.5	2.6	48.2
25	4.0	0.5	47.4
26	4.0	1.9	48.6
27	4.0	2.2	48.9
28	4.5	0.4	48.0
29	4.5	1.7	48.9
30	4.5	2.4	50.1
31	5.0	0.5	48.5
32	5.0	1.7	48.9
33	5.0	2.7	50.2

difficult and expensive to obtain; therefore, it would be useful to develop a predictive model based on the more easily measured concentrations of pollutants A and B. The regression model under consideration is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$$

The experiment was designed so that several specific levels of pollutant A, between 0.1 and 5.0 parts per million (ppm), were studied. Based on past experience, it is thought that plausible ranges of concentrations would be 0.1 to 3.0 ppm for pollutant B, and 0 to 60 ppm for pollutant C. It is believed that there will be a positive association between the pollutant levels of pollutants A and B.

As a first step in the diagnostics analysis, we examine simple descriptive statistics for each variable. The statistics (mean, median, standard deviation, range, interquartile range) in the SAS output (PROC UNIVARIATE) below do not, for the most part, look unusual given the prior knowledge about plausible values for X_1 , X_2 , and Y . However, the largest value of Y is 65.3 (see the Extreme Observations section of the output for Y), which is quite a bit larger than the next largest value and is larger than expected based on the plausible range for Y . Efforts should be made to ensure that this value was not recorded in error; if it was, the error should now be corrected. If not recorded in error, this potentially influential observation should be noted and further attention paid to it during the formal check for outliers in the next step.

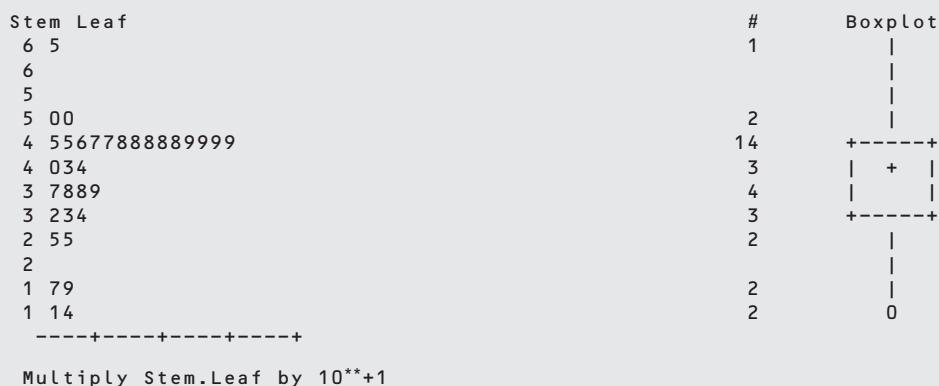
Edited SAS Output (PROC UNIVARIATE) for Example 14.6

VARIABLE: Y			
Moments			
N	33	Sum Weights	33
Mean	40.0212121	Sum Observations	1320.7
Std Deviation	12.2732625	Variance	150.632973
Skewness	-0.8440522	Kurtosis	0.45212782

BASIC STATISTICAL MEASURES			
Location		Variability	
Mean	40.02121	Std Deviation	12.27326
Median	45.20000	Variance	150.63297
Mode	48.90000	Range	54.60000
		Interquartile Range	14.40000

EXTREME OBSERVATIONS			
Lowest		Highest	
Value	Obs	Value	Obs
10.7	1	48.9	29
14.2	2	48.9	32
16.7	3	50.1	30
19.1	4	50.2	33
24.9	5	65.3	21

(continued)



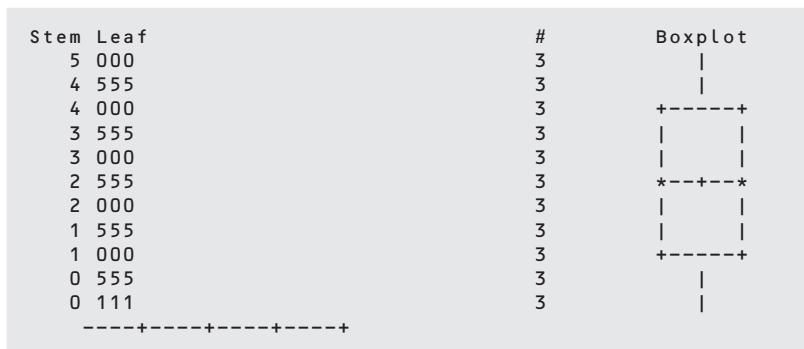
VARIABLE: X1			
Moments			
N	33	Sum Weights	33
Mean	2.50909091	Sum Observations	82.8
Std Deviation	1.59125808	Variance	2.53210227
Skewness	0.02636322	Kurtosis	-1.2491643
Uncorrected SS	288.78	Corrected SS	81.0272727
Coeff Variation	63.419706	Std Error Mean	0.27700248

BASIC STATISTICAL MEASURES			
Location		Variability	
Mean	2.509091	Std Deviation	1.59126
Median	2.500000	Variance	2.53210
Mode	0.100000	Range	4.90000
		Interquartile Range	3.00000

Note: The mode displayed is the smallest of 11 modes with a count of 3.

EXTREME OBSERVATIONS			
Lowest		Highest	
Value	Obs	Value	Obs
0.1	3	4.5	29
0.1	2	4.5	30
0.1	1	5.0	31
0.5	6	5.0	32
0.5	5	5.0	33

(continued)



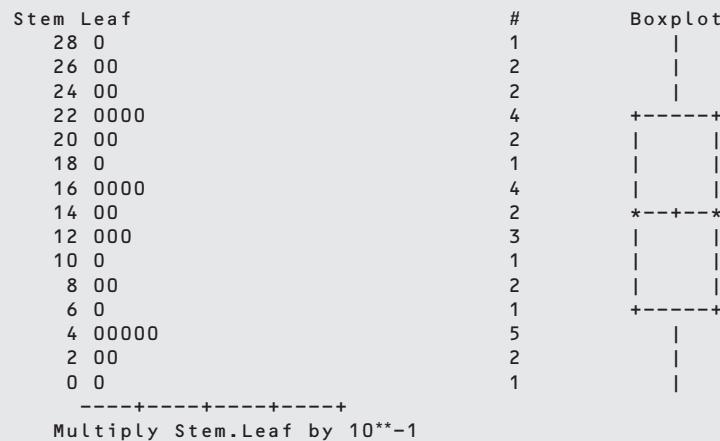
VARIABLE: X2			
Moments			
N	33	Sum Weights	33
Mean	1.45454545	Sum Observations	48
Std Deviation	0.82994386	Variance	0.68880682
Skewness	-0.0394651	Kurtosis	-1.2489157
Uncorrected SS	91.86	Corrected SS	22.0418182
Coeff Variation	57.0586407	Std Error Mean	0.14447468

BASIC STATISTICAL MEASURES			
Location		Variability	
Mean	1.454545	Std Deviation	0.82994
Median	1.500000	Variance	0.68881
Mode	0.500000	Range	2.80000
		Interquartile Range	1.50000

Note: The mode displayed is the smallest of 3 modes with a count of 3.

EXTREME OBSERVATIONS			
Lowest		Highest	
Value	Obs	Value	Obs
0.1	7	2.4	30
0.2	13	2.5	15
0.3	19	2.6	24
0.4	28	2.7	33
0.4	4	2.9	12

(continued)



© Cengage Learning

Edited SAS Output (PROC CORR) for Example 14.6

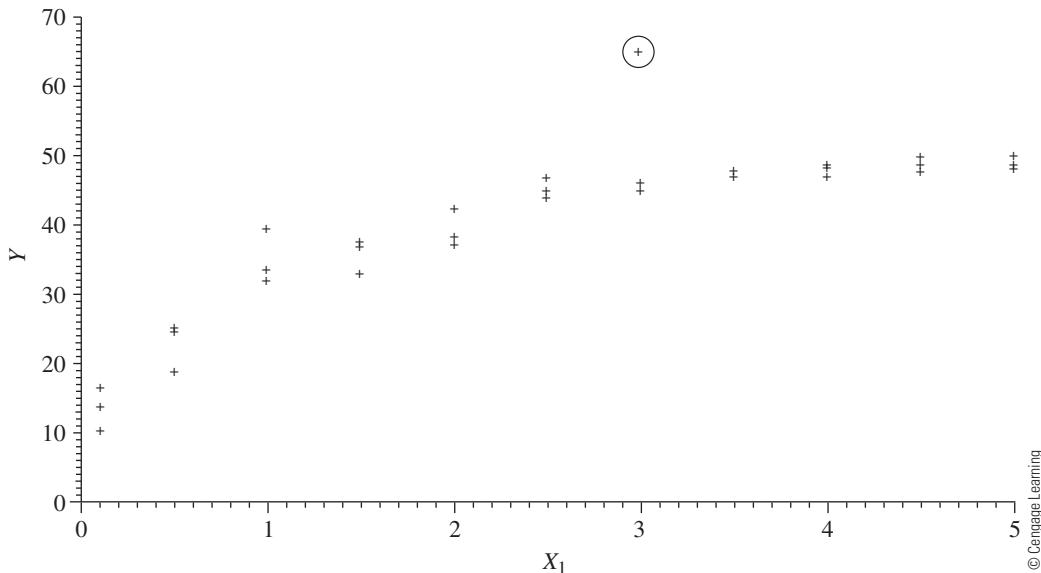
PEARSON CORRELATION COEFFICIENTS, N = 33			
	Y	X1	X2
Y	1.00000	0.84142	0.28431
X1	0.84142	1.00000	0.13969
X2	0.28431	0.13969	1.00000

© Cengage Learning

The scatterplot of Y versus X_1 (Figure 14.11) shows an obvious relationship, although perhaps a curvilinear one. The scatterplot of Y versus X_2 (Figure 14.12) shows no obvious relationship. The estimated correlation output shows that the association between Y and X_1 has a strong linear component ($r_{YX_1} = 0.84$), although the association may be better described as curvilinear. The correlation between Y and X_2 is weak ($r_{YX_2} = 0.28$).

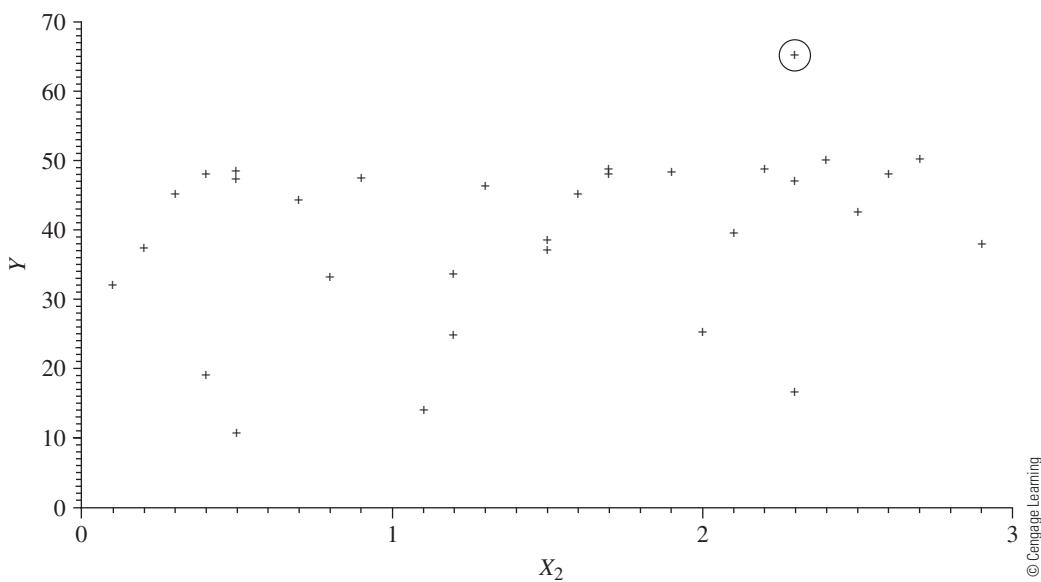
An outlier is circled in each figure. This is likely to be the same value identified in the univariate descriptive output for Y . No other outliers are immediately apparent, but a more formal check using outlier diagnostics may reveal others.

Since a multiple regression is to be performed, partial regression plots are also inspected (Figures 14.13 and 14.14). Since there are no strong linear relationships between Y and X_2 or between X_1 and X_2 ($r_{X_1X_2} = 0.14$ from the estimated correlation output), the partial regression plot of Y versus X_1 adjusted for X_2 is very similar in appearance to the unadjusted plot in Figure 14.11. The partial regression plot of Y versus X_2 adjusted for X_1 shows a slightly more distinct (though still fairly weak) positive linear association than does the unadjusted scatterplot in Figure 14.12. Partial correlation calculations confirm these findings ($r_{YX_1|X_2} = 0.84$, $r_{YX_2|X_1} = 0.31$; output not shown).



© Cengage Learning

FIGURE 14.11 Plot of pollutant C concentration level (Y) versus pollutant A concentration level (X_1) for hypothetical environmental monitoring data



© Cengage Learning

FIGURE 14.12 Plot of pollutant C concentration level (Y) versus pollutant B concentration level (X_2) for hypothetical environmental monitoring data

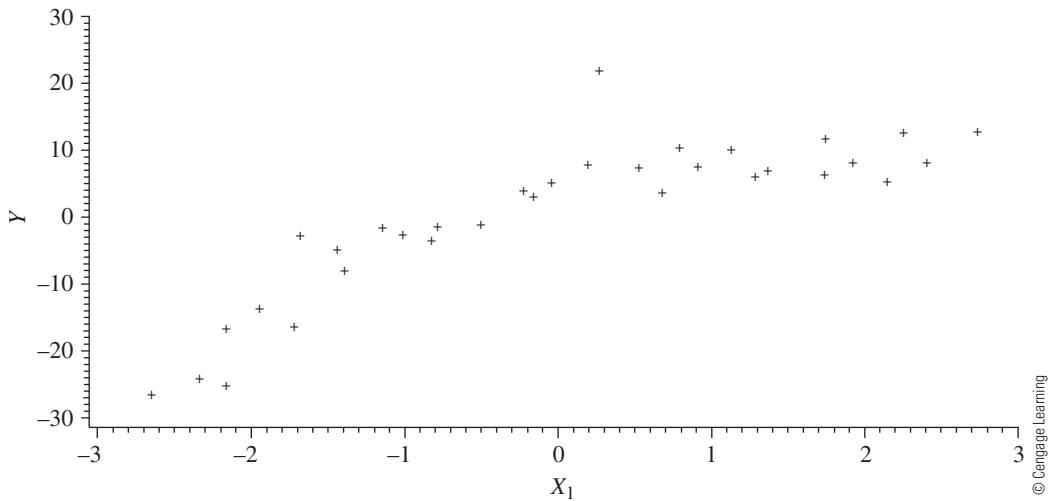


FIGURE 14.13 Partial regression plot (adjusted for pollutant B) of pollutant C concentration level (Y) versus pollutant A concentration level (X_1) for hypothetical environmental monitoring data

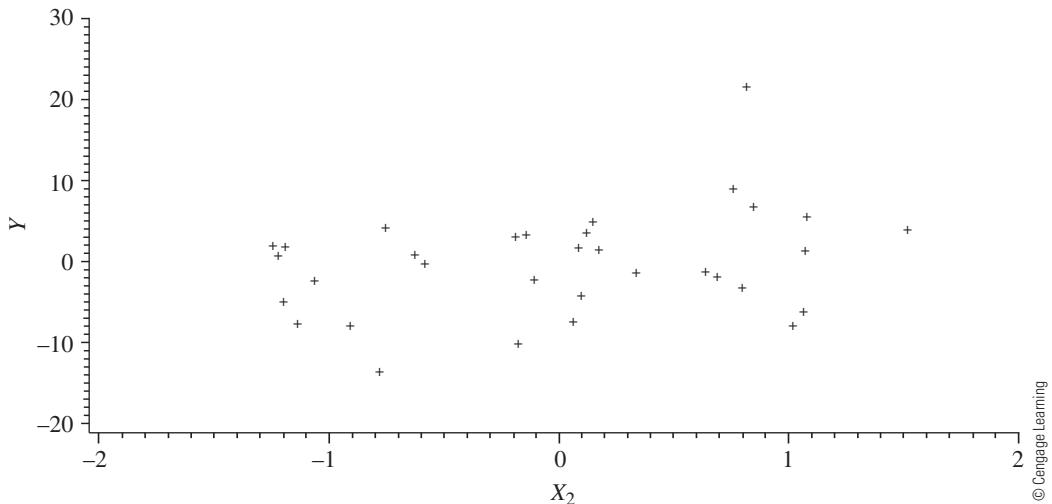


FIGURE 14.14 Partial regression plot (adjusted for pollutant A) of pollutant C concentration level (Y) versus pollutant B concentration level (X_2) for hypothetical environmental monitoring data

In the next stage of the diagnostics analysis, we inspect outlier diagnostic statistics. The observations for which at least one of the diagnostic statistics exceeded the corresponding critical value (i.e., $d_i > 1$, $b_i > 2(2 + 1)/33 = 0.18$, and $|r_{(-i)}| > t_{33-2-2, 0.95} \approx 1.7$) are shown in Table 14.5. For the three possible outliers, only the jackknife residual cutoff value

TABLE 14.5 Outliers in environmental monitoring example

Observation	X_1	X_2	Y	d_i	h_i	$r_{(i)}$
1	0.1	0.5	10.7	0.186	0.130	2.030
3	0.1	2.3	16.7	0.172	0.150	1.769
21	3.0	2.3	65.3	0.230	0.064	3.848

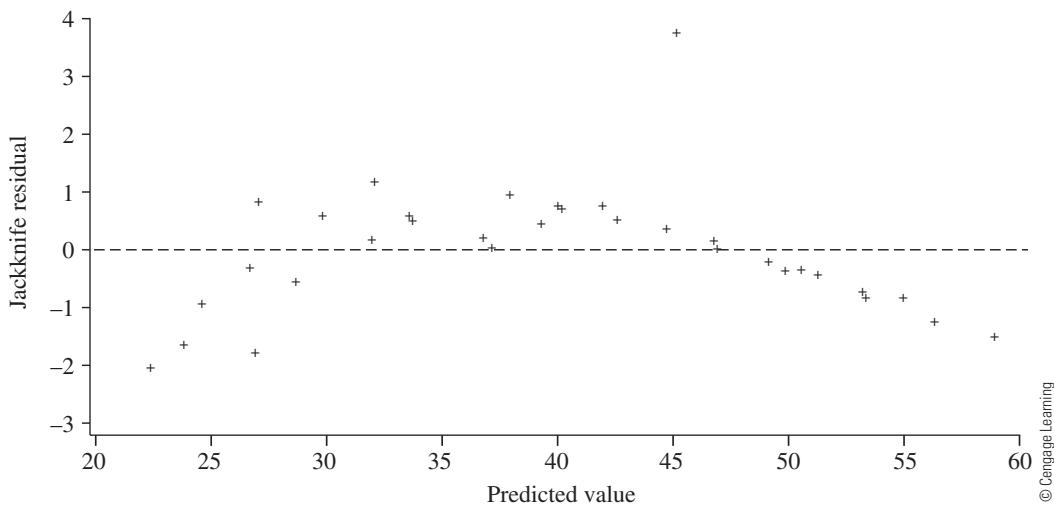
© Cengage Learning

is exceeded. For each of these observations, we first determine whether any data values were recorded in error. If they were, then every effort should be made to recover the correct value and repeat the analysis. If not, we examine each diagnostic statistic value for each observation and proceed accordingly.

Suppose we judge observations 1 and 3 to be plausible, since they are within the ranges identified prior to the study. Why then do they have large jackknife residual values and appear to be outliers? It is probably because they are at the end of the X_1 scale, where the curvilinear relationship between X_1 and Y departs to the greatest extent from the fitted regression function $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$. Once the curvature issue is dealt with, it is likely that the residuals for these observations will not be overly large in magnitude.

The value of Y for observation 21 is beyond the plausible range for Y (0–60 ppm) identified prior to the study. If the correct value of Y cannot be recovered for this observation, then the value of Y may be set to missing (effectively excluding observation 21 from the analysis). This is the course of action we recommend.

Next we examine the validity of regression assumptions. Figure 14.15 presents a plot of jackknife residuals versus predicted values. Instead of the desired random scatter of points, a



© Cengage Learning

FIGURE 14.15 SAS plot of jackknife residuals versus predicted values for regression of pollutant C concentration level (Y) on pollutant A concentration level (X_1) and pollutant B concentration level (X_2)

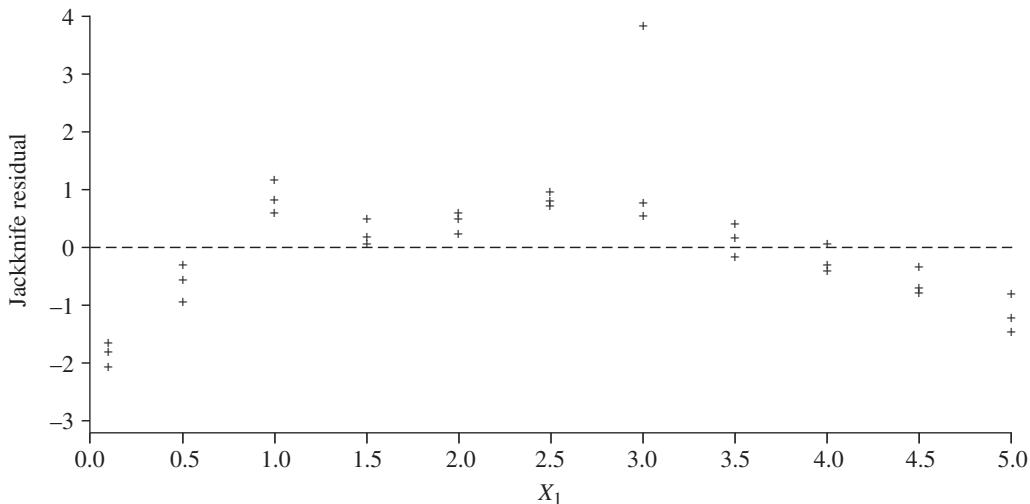


FIGURE 14.16 SAS plot of jackknife residuals versus pollutant A concentration level (X_1) for regression of pollutant C concentration level (Y) on pollutant A concentration level (X_1) and pollutant B concentration level (X_2)

© Cengage Learning

distinctly quadratic pattern is apparent. The same pattern is apparent in the plot of jackknife residuals versus X_1 (Figure 14.16), whereas the analogous plot involving X_2 (Figure 14.17) is more consistent with the linearity assumption. This suggests that the quadratic relationship primarily involves X_1 and not X_2 . A modification of the model should be attempted in order

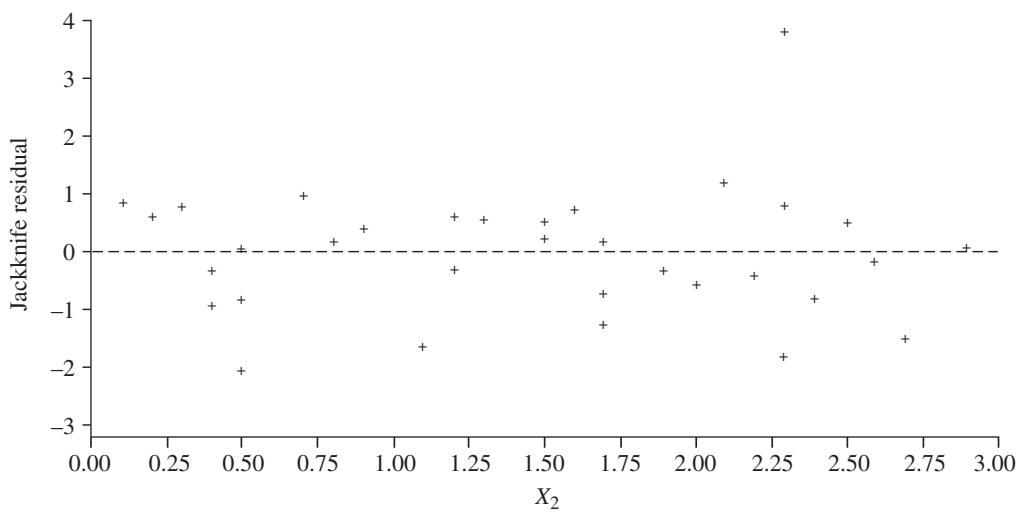
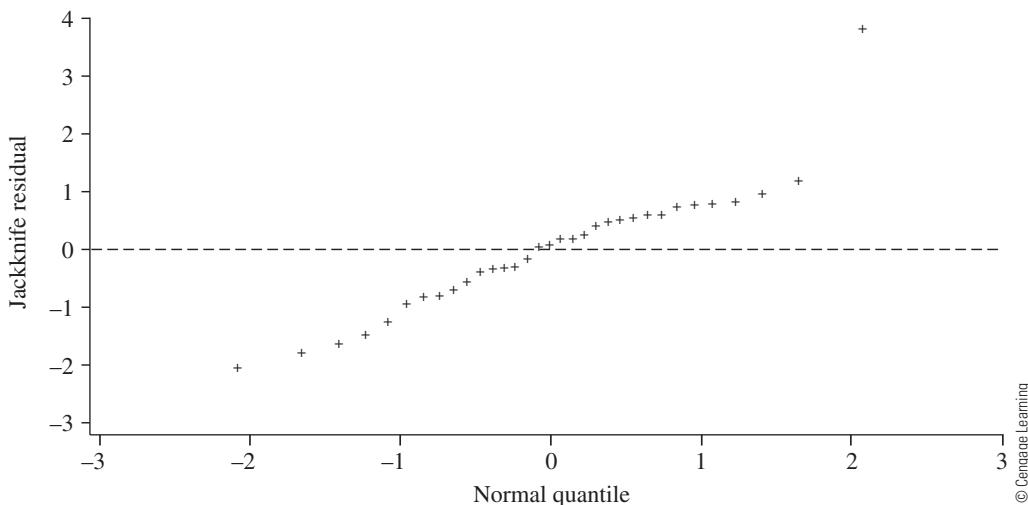


FIGURE 14.17 SAS plot of jackknife residuals versus pollutant B concentration level (X_2) for regression of pollutant C concentration level (Y) on pollutant A concentration level (X_1) and pollutant B concentration level (X_2)

© Cengage Learning



© Cengage Learning

FIGURE 14.18 SAS normal quantile-quantile plot of jackknife residuals from the regression of pollutant C concentration level (Y) on pollutant A concentration level (X_1) and pollutant B concentration level (X_2)

to eliminate the systematic pattern in the jackknife residuals reflected in Figures 14.15 and 14.16. Based on these residual plots, there does not appear to be any gross violation of the homoscedasticity assumption.

The normal quantile-quantile plot of jackknife residuals shown in Figure 14.18 exhibits a fairly linear pattern, indicating that the normality assumption is not badly violated.

Finally, we conduct a check for collinearity problems. The correlation between X_1 and X_2 shown in the SAS output is not high, indicating that there is no moderate or strong relationship between the two predictors and, therefore, probably no collinearity problem. The variance inflation factors confirm this; they are small (both equal to 1.02; output not shown). Hence, there is no real need to inspect any more diagnostic statistics such as condition indices and variance proportions.

In summary, in this example, it appears that there are two issues that need to be addressed: the effect of the outlier (observation 21); and the nonlinear relationship between Y and X_1 . A reasonable strategy would be to eliminate observation 21 from the data set and then fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + E$ to the revised data set. ■

Problems

The reader may use the provided computer output to answer the questions posed below; alternatively, the reader may use his or her own statistical software to create the necessary output. For some of the questions, be aware that the data sets have hidden problems: outliers, collinearities, and variables in need of transformation. In response to certain requests, some computer programs may either balk or produce invalid results. This state of affairs is realistic.

Other problems use “nice” data that, while not realistic, do nevertheless allow us to focus on the particular statistical concepts discussed in this chapter.

1. Consider the data of Problem 1, Chapter 5. For the model with dry weight as the response and age as the predictor,
 - a. Examine a plot of the studentized or jackknife residuals versus the predicted values. Are any regression assumption violations apparent? If so, suggest possible remedies.
 - b. Examine numerical descriptive statistics, histograms, box-and-whisker plots, and normal probability plots of jackknife residuals. Is the normality assumption violated? If so, suggest possible remedies.
 - c. Examine outlier diagnostics, including Cook’s distance, leverage statistics, and jackknife residuals, and identify any potential outliers. What course of action, if any, should be taken when outliers are identified?
2. **a.–c.** Repeat Problem 1 using $\log_{10}(\text{dry weight})$ as the response.
- 3–14. **a.–c.** Repeat parts (a) through (c) of Problem 1 using the data of Problems 3 through 14 of Chapter 5.
15. **a.–c.** Consider the data of Problem 15, Chapter 5. For the model with BLOOD-TOL as the response and PPM_TOL as the predictor, repeat (a) through (c) of Problem 1.
16. **a.–c.** Repeat Problem 15 using LN_BLDTL as the response and LN_PPMLT as the predictor.
 - d. Which approach—using the original variables or using the logged variables—leads to fewer diagnostic problems?
17. **a.–c.** Repeat Problem 15 using BLOODTOL as the response and BRAINTOL as the predictor.
18. **a.–d.** Repeat Problem 16 using LN_BLDTL as the response and LN_BRNTL as the predictor.
19. The data in the following table come from an article by Bethel et al. (1985). All subjects are asthmatics. For the model with FEV_1 as the response and HEIGHT, WEIGHT, and AGE as the predictors,

Subject	AGE (yr)	Sex	Height (cm)	Weight (kg)	FEV_1^* (L)	Subject	AGE (yr)	Sex	Height (cm)	Weight (kg)	FEV_1^* (L)
1	24	M	175	78.0	4.7	11	26	M	180	70.5	3.5
2	36	M	172	67.6	4.3	12	29	M	163	75.0	3.2
3	28	F	171	98.0	3.5	13	33	F	180	68.0	2.6
4	25	M	166	65.5	4.0	14	31	M	180	65.0	2.0
5	26	F	166	65.0	3.2	15	30	M	180	70.4	4.0
6	22	M	176	65.5	4.7	16	22	M	168	63.0	3.9
7	27	M	185	85.5	4.3	17	27	M	168	91.2	3.0
8	27	M	171	76.3	4.7	18	46	M	178	67.0	4.5
9	36	M	185	79.0	5.2	19	36	M	173	62.0	2.4
10	24	M	182	88.2	4.2						

*Forced expiratory volume in 1 second.

- a.-c.** Repeat (a) through (c) in Problem 1.
- d.** Examine variance inflation factors, condition indices (unadjusted and adjusted for the intercept), and variance proportions. Are there any important collinearity problems? If so, suggest possible remedies.
- 20.** Repeat Problem 19(d), this time including SEX as a predictor (coded SEX = 1 if female, SEX = 0 if male).
- 21.** Repeat Problem 20, this time including three interactions: SEX and AGE, SEX and HEIGHT, and SEX and WEIGHT.
- 22.** In an analysis of daily soil evaporation (EVAP), Freund (1979) identified the following predictor variables:

MAXAT = Maximum daily air temperature

MINAT = Minimum daily air temperature

AVAT = Integrated area under the daily air temperature curve (i.e., a measure of average air temperature)

MAXST = Maximum daily soil temperature

MINST = Minimum daily soil temperature

AVST = Integrated area under the daily soil temperature curve

MAXH = Maximum daily relative humidity

MINH = Minimum daily relative humidity

AVH = Integrated area under the daily relative humidity curve

WIND = Total wind, measured in miles per day

In addition, Freund provided the following overall ANOVA table and significance tests about the regression coefficients.

Source	d.f.	SS	MS	F
Regression	10	8,159.35	815.94	19.27
Residual	35	1,482.76	42.36	
Total	45	9,642.11		

Variable	$\hat{\beta}$	t	VIF
MAXAT	0.5011	0.88	8.828
MINAT	0.3041	0.39	8.887
AVAT	0.09219	0.42	22.21
MAXST	2.232	2.22	39.29
MINST	0.2049	0.19	14.08
AVST	0.7426	-2.12	52.36
MAXH	1.110	0.98	1.981
MINH	0.7514	1.54	25.38
AVH	-0.5563	-3.44	24.12
WIND	0.00892	0.97	1.985

- a. Compute R^2 , and provide a test of significance. Use $\alpha = .01$.
 - b. Compute R_j^2 for each predictor.
 - c. Which, if any, variables are implicated as possibly inducing collinearity?
 - d. Freund (1979) noted that “some of the coefficients have ‘wrong’ signs.” Based on what you know about evaporation, explain which coefficients have wrong signs, paying particular attention to those with extreme t -values.
23. The raw data for Problem 22, from Freund (1979), appear below. For the model of Problem 22,
- a. Fit the model.
 - b. What discrepancies do you note between the results of (a) and the data summary presented in Problem 22?
 - c. Examine the correlation matrix for all predictor variables in this problem. Are any collinearity problems apparent?
 - d. Determine the variance inflation factors, condition indices (unadjusted and adjusted for the intercept), and variance proportions. Are any collinearity problems apparent? If so, suggest a remedy.
 - e. Examine outlier diagnostics. Are any outliers apparent? If so, suggest a course of action.
 - f. Determine eigenvalues, condition indices, and condition numbers for the correlation matrix (excluding the intercept).
 - g. Determine eigenvalues, condition indices, and condition numbers for the scaled cross-products matrix (including the intercept).
 - h. Determine residuals (preferably studentized) and leverage values. Do any observations seem bothersome? Explain.
 - i. Does there appear to be any problem with collinearity? Explain.

Observation	Month	Day	MAXST	MINST	AVST	MAXAT	MINAT	AVAT	MAXH	MINH	AVH	WIND	EVAP
1	6	6	84	65	147	85	59	151	95	40	398	273	30
2	6	7	84	65	149	86	61	159	94	28	345	140	34
3	6	8	79	66	142	83	64	152	94	41	388	318	33
4	6	9	81	67	147	83	65	158	94	50	406	282	26
5	6	10	84	68	167	88	69	180	93	46	379	311	41
6	6	11	74	66	131	77	67	147	96	73	478	446	4
7	6	12	73	66	131	78	69	159	96	72	462	294	5
8	6	13	75	67	134	84	68	159	95	70	464	313	20
9	6	14	84	68	161	89	71	195	95	63	430	455	31
10	6	15	86	72	169	91	76	206	93	56	406	604	38
11	6	16	88	73	178	91	76	208	94	55	393	610	43
12	6	17	90	74	187	94	76	211	94	51	385	520	47
13	6	18	88	72	171	94	75	211	96	54	405	663	45
14	6	19	88	72	171	92	70	201	95	51	392	467	45
15	6	20	81	69	154	87	68	167	95	61	448	184	11
16	6	21	79	68	149	83	68	162	95	59	436	177	10
17	6	22	84	69	160	87	66	173	95	42	392	173	30

(continued)

Observation	Month	Day	MAXST	MINST	AVST	MAXAT	MINAT	AVAT	MAXH	MINH	AVH	WIND	EVAP
18	6	23	84	70	160	87	68	177	94	44	392	76	29
19	6	24	84	70	168	88	70	169	95	48	398	72	23
20	6	25	77	67	147	83	66	170	97	60	431	183	16
21	6	26	87	67	166	92	67	196	96	44	379	76	37
22	6	27	89	69	171	92	72	199	94	48	393	230	50
23	6	28	89	72	180	94	72	204	95	48	394	193	36
24	6	29	93	72	186	92	73	201	94	47	386	400	54
25	6	30	93	74	188	93	72	206	95	47	385	339	44
26	7	1	94	75	199	94	72	208	96	45	370	172	41
27	7	2	93	74	193	95	73	214	95	50	396	238	45
28	7	3	93	74	196	95	70	210	96	45	380	118	42
29	7	4	96	75	198	95	71	207	93	40	365	93	50
30	7	5	95	76	202	95	69	202	93	39	357	269	48
31	7	6	84	73	173	96	69	173	94	58	418	128	17
32	7	7	91	71	170	91	69	168	94	44	420	423	20
33	7	8	88	72	179	89	70	189	93	50	399	415	15
34	7	9	89	72	179	95	71	210	98	46	389	300	42
35	7	10	91	72	182	96	73	208	95	43	384	193	44
36	7	11	92	74	196	97	75	215	96	46	389	195	41
37	7	12	94	75	192	96	69	198	95	36	380	215	49
38	7	13	96	75	195	95	67	196	97	24	354	185	53
39	7	14	93	76	198	94	75	211	93	43	364	466	53
40	7	15	88	74	188	92	73	198	95	52	405	399	21
41	7	16	88	74	178	90	74	197	95	61	447	232	1
42	7	17	91	72	175	94	70	205	94	42	380	275	44
43	7	18	92	72	190	95	71	209	96	44	379	166	44
44	7	19	92	73	189	96	72	208	93	42	372	189	46
45	7	20	94	75	194	95	71	208	93	43	373	164	47
46	7	21	96	76	202	96	71	208	94	40	368	139	50

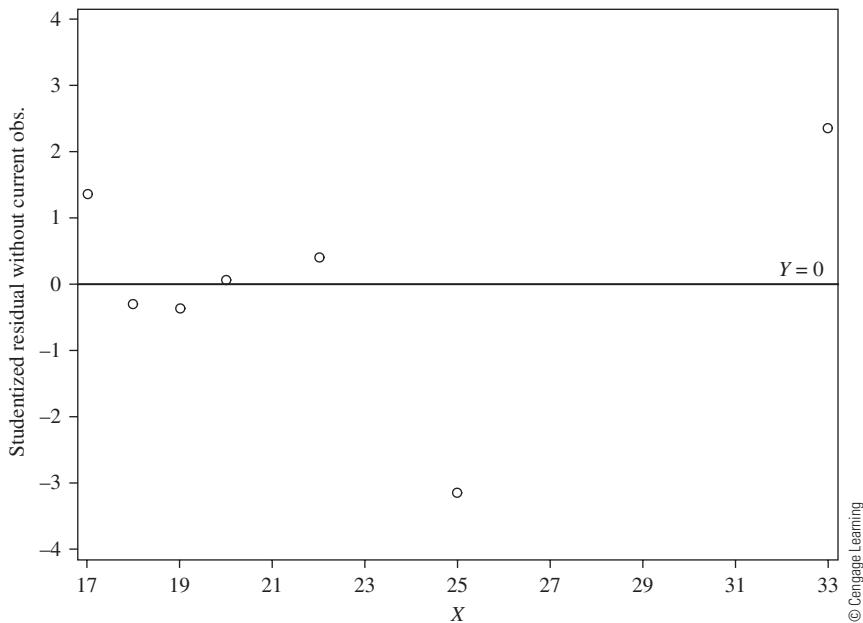
24. Real estate prices depend, in part, on property size. The house size X (in hundreds of square feet) and house price Y (in thousands of dollars) of a random sample of houses in a certain county were observed. The data (which first appeared in Problem 16 of Chapter 5) are as follows:

X	18	20	25	22	33	19	17
Y	80	95	104	110	175	85	89

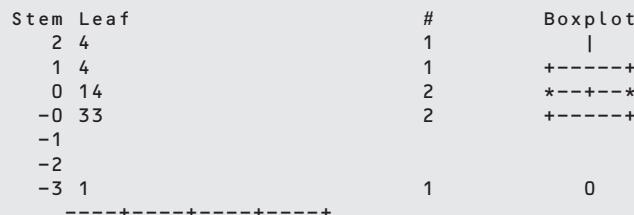
- a. Fit a straight-line model with house price Y as the response and house size X as the predictor.
- b. Compared to the appropriate t -distribution value, do any residuals appear to have extreme values?
- c. Do these analyses highlight any potentially troublesome observations? Why or why not?

Edited SAS Output for Problem 24

Plot of JACKKNIFE by X. Legend: A = 1 obs, B = 2 observation, etc.



© Cengage Learning



25. Data on sales revenues (Y) and advertising expenditures (X) for a large retailer for the period 1988–1993 are given in the following table:

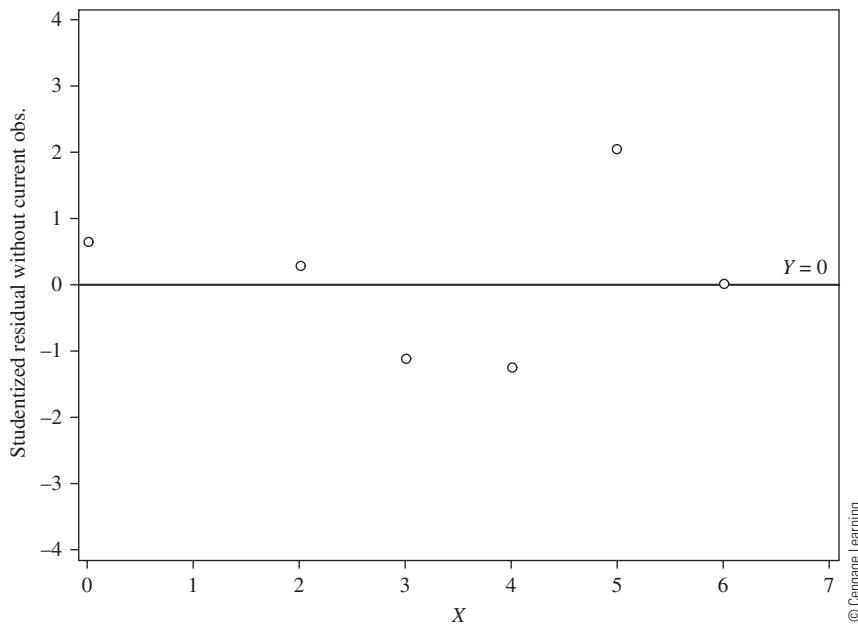
Year	1988	1989	1990	1991	1992	1993
Sales Y (\$millions)	4	8	2	8	5	4
Advertising X (\$millions)	2	5	0	6	4	3

These data first appeared in Problem 17 of Chapter 5.

- a.–c. Repeat parts (a) through (c) of Problem 24 for the current data set, using sales revenues Y as the response and advertising expenditures X as the predictor.

Edited SAS Output for Problem 25

Plot of JACKKNIFE by ADV₁. Legend: A = 1 obs, B = 2 obs, etc.



© Cengage Learning



26. The production manager of a plant that manufactures syringes has recorded the marginal cost Y at various levels of output X for 14 randomly selected months. The data are shown in the following table:

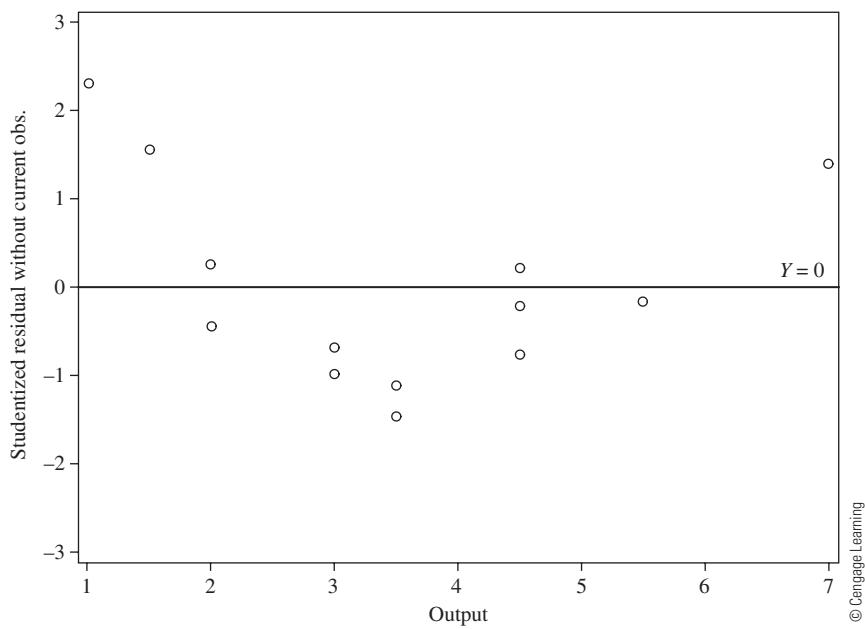
Marginal Cost Y (per 100 units)	Output X (thousands of units)	Marginal Cost Y (per 100 units)	Output X (thousands of units)
31.00	3.0	27.00	3.5
30.00	3.0	25.00	5.5
28.00	3.5	24.00	7.0
46.00	1.0	25.00	7.0
43.00	1.5	29.50	4.5
35.00	2.0	26.00	4.5
37.50	2.0	28.00	4.5

These data first appeared in Problem 18 of Chapter 5.

a.-c. Repeat parts (a) through (c) of Problem 24 for the current data set, using marginal cost Y as the response and output X as the predictor.

Edited SAS Output for Problem 26

Plot of JACKKNIFE by OUTPUT. Legend: A = 1 obs, B = 2 obs, etc.



© Cengage Learning

STEM LEAF	#	Boxplot
2 3	1	
1 6	1	
1 04	2	+-----+
0		+
0 23	2	+
-0 422	3	*-----*
-0 87	2	+-----+
-1 10	2	
-1 5	1	
		-----+-----+-----+

27. Radial keratotomy is a type of refractive surgery in which radial incisions are made in the cornea of myopic (nearsighted) patients in an effort to reduce their myopia. Theoretically, the incisions allow the curvature of the cornea to become less steep, reducing the refractive error of the patient's vision. This and other types of vision correction surgery were growing in popularity in the 1980s and 1990s among the public and among ophthalmologists.

The Prospective Evaluation of Radial Keratotomy (PERK) study was begun in 1983 to investigate the effects of radial keratotomy. Lynn et al. (1987) examined the factors associated with the five-year postsurgical change in refractive error (Y), measured in diopters. Two of the independent variables under consideration were baseline refractive error (X_1 , in diopters) and baseline curvature of the cornea (X_2 , in diopters). (Note: Myopic patients have negative refractive errors. Patients who are farsighted have positive refractive errors. Patients who are neither near- nor farsighted have zero refractive error.)

The computer output below is based on data adapted from the PERK study. These data first appeared in Chapter 8, Problem 12. Use the output to answer the following questions.

- Fit a model relating change in refractive error to baseline refractive error and baseline curvature.
- Conduct variable-added-last tests for both predictors, and perform a test for significance of the estimated intercept.
- Determine the variance inflation factors for each predictor.
- Determine a correlation matrix that includes both predictors and the response.
- Determine eigenvalues, condition indices, and condition numbers for the correlation matrix (excluding the intercept).
- Determine eigenvalues, condition indices, and condition numbers for the scaled cross-products matrix (including the intercept).
- Determine residuals and leverage values. Do any observations seem bothersome? Explain.
- Does there appear to be any problem with collinearity? Explain.

Edited SAS Output (PROC REG) for Problem 27

CORRELATION			
Variable	X1	X2	Y
X1	1.0000	0.0424	-0.4040
X2	0.0424	1.0000	-0.2544
Y	-0.4040	-0.2544	1.0000

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	17.62277	8.81139	7.18	0.0018
Error	51	62.63017	1.22804		
Corrected Total	53	80.25294			

Root MSE	1.10817	R-Square	0.2196
Dependent Mean	3.83343	Adj R-Sq	0.1890
Coeff Var	28.90811		

(continued)

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	12.36002	5.08621	2.43	0.0187	0
X1	1	-0.29160	0.09165	-3.18	0.0025	1.00180
X2	1	-0.22040	0.11482	-1.92	0.0605	1.00180

COLLINEARITY DIAGNOSTICS						
Number	Eigenvalue	Condition Index	Proportion of Variation			
			Intercept	X1	X2	
1	2.90058	1.00000	0.00010283	0.01582	0.00010396	
2	0.09898	5.41336	0.00140	0.97792	0.00147	
3	0.00044274	80.94123	0.99850	0.00626	0.99842	

COLLINEARITY DIAGNOSTICS (INTERCEPT ADJUSTED)				
Number	Eigenvalue	Condition Index	Proportion of Variation	
			X1	X2
1	1.04240	1.00000	0.47880	0.47880
2	0.95760	1.04334	0.52120	0.52120

OUTPUT STATISTICS											
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2 -1 0 1 2	Cook's D	RStudent	Hat Diag H	
1	5.3750	4.1350	0.3944	1.2400	1.036	1.197	**	0.069	1.2026	0.1267	
2	2.6250	3.7073	0.1762	-1.0823	1.094	-0.989	*	0.008	-0.9890	0.0253	
3	1.5000	3.2309	0.2226	-1.7309	1.086	-1.594	***	0.036	-1.6196	0.0404	
4	3.7500	3.9667	0.1682	-0.2167	1.095	-0.198		0.000	-0.1960	0.0230	
5	4.1250	3.7652	0.1672	0.3598	1.095	0.328		0.001	0.3255	0.0228	
6	4.6250	4.4647	0.2284	0.1603	1.084	0.148		0.000	0.1464	0.0425	
7	4.2500	5.3056	0.4292	-1.0556	1.022	-1.033	**	0.063	-1.0339	0.1500	
8	3.7500	1.9284	0.5750	1.8216	0.947	1.923	***	0.454	1.9770	0.2693	
9	3.2500	3.5466	0.3875	-0.2966	1.038	-0.286		0.004	-0.2831	0.1223	
10	2.6250	3.3669	0.2253	-0.7419	1.085	-0.684	*	0.007	-0.6802	0.0413	
11	4.0000	3.5386	0.3122	0.4614	1.063	0.434		0.005	0.4305	0.0794	
12	4.1250	4.2265	0.1842	-0.1015	1.093	-0.0929		0.000	-0.0919	0.0276	
13	3.0000	3.0159	0.2864	-0.0159	1.071	-0.0148		0.000	-0.0147	0.0668	
14	3.1250	3.5499	0.2931	-0.4249	1.069	-0.398		0.004	-0.3943	0.0700	
15	4.7500	4.6457	0.4180	0.1043	1.026	0.102		0.001	0.1006	0.1423	
16	3.5000	3.8370	0.1721	-0.3370	1.095	-0.308		0.001	-0.3051	0.0241	
17	4.5000	4.0543	0.3622	0.4457	1.047	0.426		0.007	0.4221	0.1068	
18	4.7500	3.9044	0.2479	0.8456	1.080	0.783	*	0.011	0.7799	0.0501	
19	4.8750	4.6426	0.2619	0.2324	1.077	0.216		0.001	0.2138	0.0558	
20	3.0000	3.6310	0.2374	-0.6310	1.082	-0.583	*	0.005	-0.5791	0.0459	
21	3.6250	4.0459	0.2176	-0.4209	1.087	-0.387		0.002	-0.3841	0.0386	

(continued)

OUTPUT STATISTICS											
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2 -1 0 1 2	Cook's D	RStudent	Hat Diag H	
22	3.2500	3.4788	0.1917	-0.2288	1.091	-0.210		0.000	-0.2077	0.0299	
23	6.2500	3.9794	0.1683	2.2706	1.095	2.073	* ***	0.034	2.1450	0.0231	
24	5.1250	4.0514	0.2168	1.0736	1.087	0.988	*	0.013	0.9876	0.0383	
25	2.0000	3.1571	0.2521	-1.1571	1.079	-1.072	**	0.021	-1.0739	0.0518	
26	4.1300	3.0555	0.2553	1.0745	1.078	0.996	*	0.019	0.9964	0.0531	
27	4.1250	3.3367	0.2003	0.7883	1.090	0.723	*	0.006	0.7198	0.0327	
28	4.2500	3.9357	0.1813	0.3143	1.093	0.287		0.001	0.2849	0.0268	
29	6.1250	4.9313	0.3463	1.1937	1.053	1.134	**	0.046	1.1372	0.0977	
30	6.1250	3.9038	0.1634	2.2212	1.096	2.026	* ***	0.030	2.0925	0.0217	
31	2.8750	3.4664	0.2242	-0.5914	1.085	-0.545	*	0.004	-0.5412	0.0409	
32	3.2500	4.8782	0.3515	-1.6282	1.051	-1.549	***	0.090	-1.5715	0.1006	
33	5.0000	4.0832	0.2656	0.9168	1.076	0.852	*	0.015	0.8498	0.0574	
34	3.7500	3.2816	0.2132	0.4684	1.087	0.431		0.002	0.4272	0.0370	
35	6.8750	4.4024	0.2576	2.4726	1.078	2.294	* ***	0.100	2.3986	0.0541	
36	2.3750	3.5047	0.1814	-1.1297	1.093	-1.033	**	0.010	-1.0340	0.0268	
37	5.1250	4.3261	0.2182	0.7989	1.086	0.735	*	0.007	0.7320	0.0388	
38	4.3750	4.3028	0.2098	0.0722	1.088	0.0663		0.000	0.0657	0.0358	
39	5.5000	3.8129	0.2329	1.6871	1.083	1.557	***	0.037	1.5799	0.0442	
40	2.2500	3.5445	0.2734	-1.2945	1.074	-1.205	**	0.031	-1.2109	0.0609	
41	4.1250	3.4986	0.2750	0.6264	1.074	0.583	*	0.007	0.5796	0.0616	
42	3.8750	4.7244	0.2831	-0.8494	1.071	-0.793	*	0.015	-0.7899	0.0653	
43	3.2500	4.5046	0.2328	-1.2546	1.083	-1.158	**	0.021	-1.1619	0.0441	
44	2.7500	3.5428	0.1866	-0.7928	1.092	-0.726	*	0.005	-0.7224	0.0283	
45	3.1250	3.3723	0.2130	-0.2473	1.088	-0.227		0.001	-0.2253	0.0369	
46	2.6250	4.2043	0.1800	-1.5793	1.093	-1.444	**	0.019	-1.4603	0.0264	
47	2.8750	3.7925	0.1515	-0.9175	1.098	-0.836	*	0.004	-0.8333	0.0187	
48	5.2500	4.2388	0.2975	1.0112	1.067	0.947	*	0.023	0.9463	0.0721	
49	2.8750	3.7997	0.2285	-0.9247	1.084	-0.853	*	0.011	-0.8504	0.0425	
50	2.1250	3.2733	0.2408	-1.1483	1.082	-1.062	**	0.019	-1.0630	0.0472	
51	0.8750	3.4903	0.1852	-2.6153	1.093	-2.394	***	0.055	-2.5157	0.0279	
52	3.5000	3.4745	0.1844	0.0255	1.093	0.0233		0.000	0.0231	0.0277	
53	3.5000	3.2263	0.2201	0.2737	1.086	0.252		0.001	0.2497	0.0395	
54	4.3750	3.9205	0.1542	0.4545	1.097	0.414		0.001	0.4107	0.0194	

28. In 1990, *Business Week* magazine compiled financial data on the 1,000 companies that had the biggest impact on the U.S. economy. Data from this compilation were presented in Problem 13 of Chapter 8. In addition to the company name, data on the following variables were shown in that problem:

1990 Rank: Based on company's market value (share price on March 16, 1990, multiplied by available common shares outstanding)

1989 Rank: Rank in 1989 compilation

P-E Ratio: Price-to-earnings ratio, based on 1989 earnings and March 16, 1990, share price

Yield: Annual dividend rate as a percentage of March 16, 1990, share price

The computer output given next is based on the *Business Week* magazine data. Use the output to answer the following questions:

- Fit a model with yield (Y) as the response and 1990 rank (X_1), 1989 rank (X_2), and P-E ratio (X_3) as the predictors. State the estimated model.
- Perform variables-added-last tests for each predictor.
- Determine the variance inflation factors for each predictor.
- Does there appear to be any problem with collinearity? Explain.
- Examine each plot of studentized residuals versus the predictor. Are any violations of linear regression assumptions obvious?
- Determine residuals and leverage values. Do any observations seem bothersome? Explain.

Edited SAS Output (PROC REG) for Problem 28

Yield Regressed on 1990 Rank (X1), 1989 Rank (X2), and P-E Ratio (X3)

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	29.97726	9.99242	8.83	0.0011
Error	16	18.10304	1.13144		
Corrected Total	19	48.08030			

Root MSE	1.06369	R-Square	0.6235
Dependent Mean	2.48550	Adj R-Sq	0.5529
Coeff Var	42.79588		

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	7.42119	0.98839	7.51	<.0001	0
X1	1	-0.01335	0.00771	-1.73	0.1025	23.11906
X2	1	0.00762	0.00697	1.09	0.2901	22.15954
X3	1	-0.26185	0.05250	-4.99	0.0001	1.21991

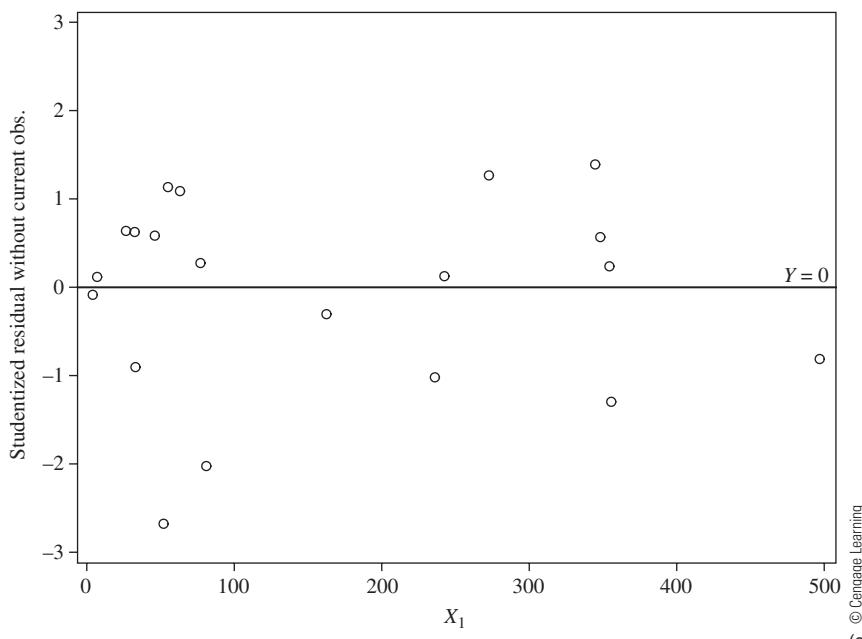
OUTPUT STATISTICS											
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2 -1 0 1 2	Cook's D	RStudent	Hat Diag H	
1	2.8700	2.9469	0.3519	-0.0769	1.004	-0.0766		0.000	-0.0742	0.1095	
2	2.5400	2.4060	0.3567	0.1340	1.002	0.134		0.001	0.1296	0.1124	
3	1.7200	1.0888	0.4715	0.6312	0.953	0.662	*	0.027	0.6499	0.1965	
4	4.2600	3.6101	0.3620	0.6499	1.000	0.650	*	0.014	0.6377	0.1158	

(continued)

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	OUTPUT STATISTICS				Cook's D	RStudent	Hat Diag H
							-2	-1	0	1			
5	0.4100	1.2782	0.4351	-0.8682	0.971	-0.894		*			0.040	-0.8886	0.1673
6	4.1000	3.4917	0.3426	0.6083	1.007	0.604		*			0.011	0.5916	0.1037
7	0	2.3005	0.3247	-2.3005	1.013	-2.271		****			0.133	-2.6714	0.0932
8	4.2700	3.1250	0.3190	1.1450	1.015	1.128		**			0.031	1.1388	0.0899
9	2.6100	1.5282	0.4037	1.0818	0.984	1.099		**			0.051	1.1070	0.1440
10	2.9400	2.6352	0.2802	0.3048	1.026	0.297					0.002	0.2884	0.0694
11	2.9800	4.6793	0.5318	-1.6993	0.921	-1.845		***			0.283	-2.0129	0.2499
12	2.6200	2.9233	0.3480	-0.3033	1.005	-0.302					0.003	-0.2930	0.1070
13	2.9100	3.7481	0.6654	-0.8381	0.830	-1.010		**			0.164	-1.0106	0.3914
14	1.8700	1.7354	0.2943	0.1346	1.022	0.132					0.000	0.1276	0.0766
15	4.7300	3.5077	0.4028	1.2223	0.984	1.242		**			0.065	1.2646	0.1434
16	0	-0.4225	0.7882	0.4225	0.714	0.591		*			0.106	0.5790	0.5491
17	5.4600	4.2769	0.6101	1.1831	0.871	1.358		**			0.226	1.3976	0.3289
18	1.9900	1.7468	0.4036	0.2432	0.984	0.247					0.003	0.2398	0.1440
19	1.1000	2.0679	0.7349	-0.9679	0.769	-1.259		**			0.362	-1.2838	0.4773
20	0.3300	1.0367	0.6116	-0.7067	0.870	-0.812		*			0.081	-0.8029	0.3306

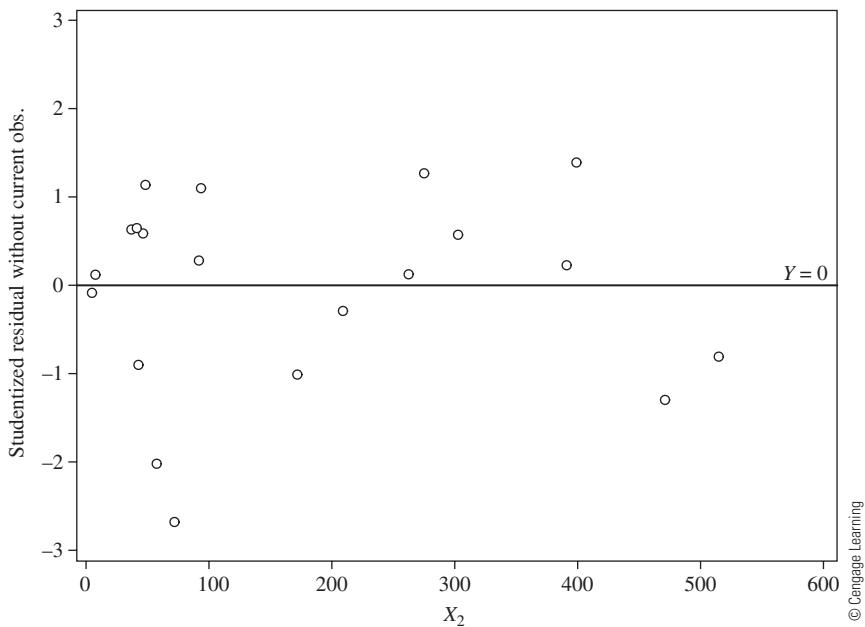
Sum of Residuals	0
Sum of Squared Residuals	18.10304
Predicted Residual SS (PRESS)	30.27079

Plot of JACKKNIFE*X1. Legend: A = 1 obs, B = 2 obs, etc.



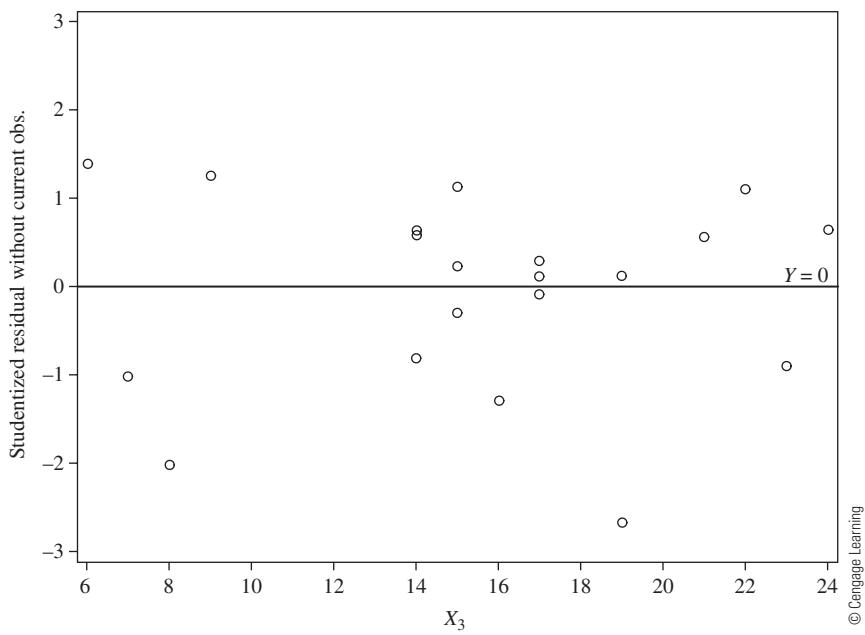
(continued)

Plot of JACKKNIFE*X2. Legend: A = 1 obs, B = 2 obs, etc.



© Cengage Learning

Plot of JACKKNIFE*X3. Legend: A = 1 obs, B = 2 obs, etc.



© Cengage Learning

- 29.** In Problem 19 of Chapter 5, data from the 1990 Census were shown for 26 randomly selected Metropolitan Statistical Areas (MSAs). Of interest are factors potentially associated with the rate of owner occupancy of housing units. Three variables included in the data set were as follows:

OWNEROCC: Proportion of housing units that are owner-occupied
 (as opposed to renter-occupied)

OWNCOST: Median selected monthly ownership costs, in \$

URBAN: Proportion of population living in urban areas

Use the output given next to answer the following questions:

- Fit a model with OWNEROCC as the response and OWCOST and URBAN as the predictors. State the estimated model.
- Perform variables-added-last tests for each predictor.
- Determine the variance inflation factors for each predictor.
- Does there appear to be any problem with collinearity? Explain.
- Examine each plot of studentized residuals versus the predictor. Are any violations of the assumptions of linear regression evident from the plot?
- Determine residuals and leverage values. Do any observations seem bothersome? Explain.

Edited SAS Output (PROC REG) for Problem 29

OWNEROCC Regressed on OWCOST and URBAN

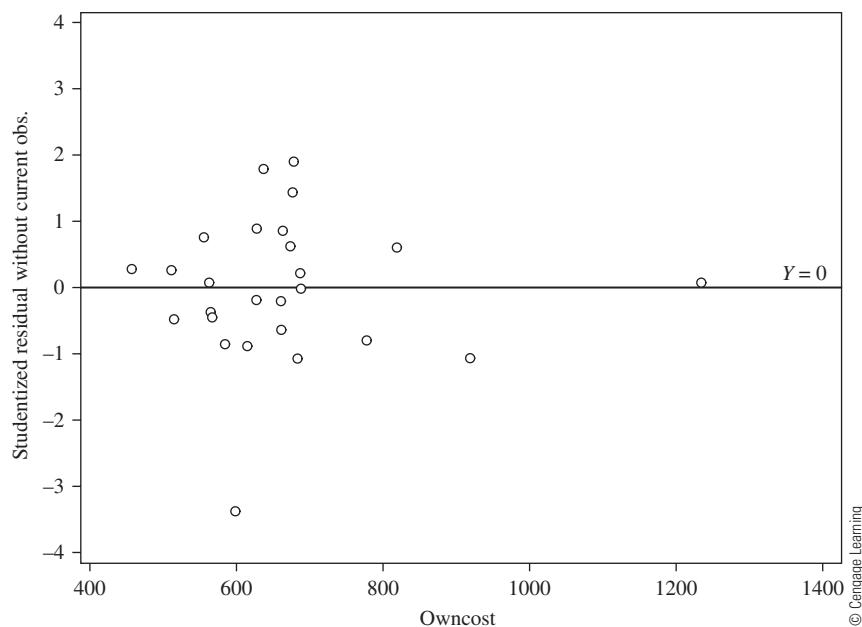
ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	255.77851	127.88925	8.52	0.0017
Error	23	345.18303	15.00796		
Corrected Total	25	600.96154			

Root MSE	3.87401	R-Square	0.4256
Dependent Mean	65.96154	Adj R-Sq	0.3757
Coeff Var	5.87314		

(continued)

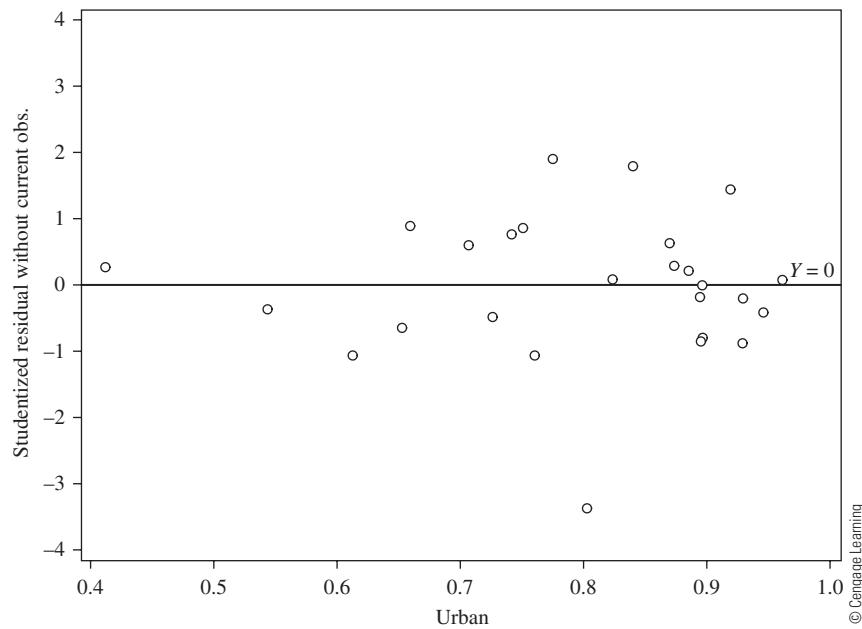
PARAMETER ESTIMATES										
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation				
Intercept	1	86.75877	5.10721	16.99	<.0001	0				
OWNCOST	1	-0.01113	0.00529	-2.10	0.0467	1.07635				
URBAN	1	-16.87557	5.89098	-2.86	0.0088	1.07635				
OUTPUT STATISTICS										
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2 -1 0 1 2	Cook's D	RStudent	Hat Diag H
1	62.0000	65.1719	1.1065	-3.1719	3.713	-0.854	*	0.022	-0.8492	0.0816
2	72.0000	68.6630	1.0870	3.3370	3.718	0.897	*	0.023	0.8935	0.0787
3	72.0000	65.5198	0.8254	6.4802	3.785	1.712	***	0.046	1.7925	0.0454
4	73.0000	66.1494	0.7769	6.8506	3.795	1.805	***	0.046	1.9054	0.0402
5	61.0000	64.2599	1.1661	-3.2599	3.694	-0.882	*	0.026	-0.8780	0.0906
6	67.0000	66.6296	0.9675	0.3704	3.751	0.0987		0.000	0.0966	0.0624
7	64.0000	64.6950	1.0026	-0.6950	3.742	-0.186		0.001	-0.1818	0.0670
8	67.0000	68.8092	1.0949	-1.8092	3.716	-0.487		0.007	-0.4786	0.0799
9	57.0000	56.8103	3.0194	0.1897	2.427	0.0782		0.003	0.0765	0.6075
10	63.0000	64.5295	1.3615	-1.5295	3.627	-0.422		0.008	-0.4140	0.1235
11	56.0000	66.5931	0.8389	-10.5931	3.782	-2.801	*****	0.129	-3.3746	0.0469
12	71.0000	68.0832	0.9522	2.9168	3.755	0.777	*	0.013	0.7698	0.0604
13	67.0000	64.5928	0.8703	2.4072	3.775	0.638	*	0.007	0.6293	0.0505
14	65.0000	64.1798	0.9134	0.8202	3.765	0.218		0.001	0.2133	0.0556
15	75.0000	74.1348	2.3174	0.8652	3.104	0.279		0.014	0.2730	0.3578
16	66.0000	68.4098	1.1346	-2.4098	3.704	-0.651	*	0.013	-0.6422	0.0858
17	64.0000	63.9992	0.9488	0.000805	3.756	0.00021		0.000	0.000210	0.0600
18	63.0000	63.7494	1.0958	-0.7494	3.716	-0.202		0.001	-0.1974	0.0800
19	70.0000	71.3160	1.6272	-1.3160	3.516	-0.374		0.010	-0.3672	0.1764
20	70.0000	66.7188	0.8057	3.2812	3.789	0.866	*	0.011	0.8610	0.0433
21	69.0000	63.7458	1.0392	5.2542	3.732	1.408	**	0.051	1.4404	0.0720
22	65.0000	68.8408	1.3485	-3.8408	3.632	-1.058	**	0.051	-1.0604	0.1212
23	60.0000	63.7172	1.6143	-3.7172	3.522	-1.056	**	0.078	-1.0583	0.1736
24	60.0000	62.9960	1.0502	-2.9960	3.729	-0.803	*	0.017	-0.7970	0.0735
25	68.0000	65.7428	1.3299	2.2572	3.639	0.620	*	0.017	0.6119	0.1178
26	68.0000	66.9432	1.4928	1.0568	3.575	0.296		0.005	0.2897	0.1485
Sum of Residuals							0			
Sum of Squared Residuals							345.18303			
Predicted Residual SS (PRESS)							396.79994			

Plot of JACKKNIFE*OWNCOST. Legend: A = 1 obs, B = 2 obs, etc.



© Cengage Learning

Plot of JACKKNIFE*URBAN. Legend: A = 1 obs, B = 2 obs, etc.



© Cengage Learning

References

The following list includes sources not cited in Chapter 14 that nonetheless offer helpful discussions of topics covered here.

- Agresti, A. 1990. *Categorical Data Analysis*. New York: John Wiley & Sons.
- Anscombe, F. J. 1960. "Rejection of Outliers." *Technometrics* 2: 123–47.
- Anscombe, F. J., and Tukey, J. W. 1963. "The Examination and Analysis of Residuals." *Technometrics* 5: 141–60.
- Armitage, P. 1971. *Statistical Methods in Medical Research*. Oxford: Blackwell Scientific.
- Barnett, V., and Lewis, T. 1978. *Outliers in Statistical Data*. New York: John Wiley & Sons.
- Bartlett, M. S. 1947. "The Use of Transformations." *Biometrics* 3: 39–52.
- Belsley, D. A. 1984. "Demeaning Conditioning Diagnostics through Centering." *American Statistician* 38: 73–77.
- . 1991. *Conditioning Diagnostics, Collinearity and Weak Data in Regression*. New York: John Wiley & Sons.
- Belsley, D. A.; Kuh, E.; and Welsch, R. E. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.
- Bethel, R. A.; Sheppard, D.; Geffroy, B.; Tam, E.; Nadel, J. A.; and Boushey, J. A. 1985. "Effect of 0.25 ppm Sulfur Dioxide on Airway Resistance in Freely Breathing, Heavily Exercising, Asthmatic Subjects." *American Review of Respiratory Diseases* 131: 659–61.
- Box, G. E. P., and Cox, D. R. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society B* 26: 211–43 (with discussion, 244–52).
- . 1984. "An Analysis of Transformations Revisited, Rebuttal." *Journal of the American Statistical Association* 79: 209–10.
- Carroll, R. J., and Ruppert, D. 1984. "Power Transformations When Fitting Theoretical Models to Data." *Journal of the American Statistical Association* 79: 321–28.
- Conover, W. J., and Iman, R. L. 1981. "Rank Transformations as a Bridge between Parametric and Nonparametric Statistics." *American Statistician* 35: 124–28.
- Cook, R. D., and Weisberg, S. 1982. *Residuals and Influence in Regression*. New York: Chapman & Hall.
- Draper, N. R., and Smith, H. 1998. *Applied Regression Analysis* (3rd ed.). New York: John Wiley & Sons.
- Durbin, J., and Watson, G. S. 1951. "Testing for Serial Correlation in Least Squares Regression." *Biometrika* 37: 409–28.
- Freund, R. J. 1979. "Multicollinearity etc., Some 'New' Examples." *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 111–12.
- Gallant, A. R. 1975. "Nonlinear Regression." *American Statistician* 29: 73–81.
- Gunst, R. F., and Mason, R. L. 1980. *Regression Analysis and Its Application*. New York: Marcel Dekker.
- Hackney, O. J., and Hocking, R. R. 1979. "Diagnostic Techniques for Identifying Data Problems in Multiple Linear Regression." *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 94–98.
- Hinkley, D. V., and Rungger, G. 1984. "The Analysis of Transformed Data." *Journal of the American Statistical Association* 79: 302–20.

- Hoaglin, D. C., and Welsch, R. E. 1978. "The Hat Matrix in Regression and ANOVA." *American Statistician* 32: 17–22.
- Hocking, R. R. 1983. "Developments in Linear Regression Methodology: 1959–1982." *Technometrics* 25: 219–48.
- Huber, P. J. 1981. *Robust Statistics*. New York: John Wiley & Sons.
- Jensen, D. R., and Ramirez, D. E. 1996. "Computing the CDF of Cook's D_I Statistic." In A. Prat and E. Ripoll, eds., *Proceedings of the 12th Symposium in Computational Statistics*, pp. 65–66. Barcelona, Spain: Institut d'Estadística de Catalunya.
- . 1998. "Some Exact Properties of Cook's D_I ." In C. R. Rao and N. Balakrishnan, eds., *Handbook of Statistics-16: Order Statistics and Their Applications*. Amsterdam: North Holland.
- Kutner, M. H.; Neter, J.; Nachtsheim, C. J.; and Li, W. 2004. *Applied Linear Statistical Models* (5th ed.). New York: McGraw-Hill/Irwin.
- Lewis, T., and Taylor, L. R. 1967. *Introduction to Experimental Ecology*. New York: Academic Press.
- Lynn, M. J., Waring, G. O., III, Sperduto, R. D. 1987. "Factors Affecting Outcome and Predictability of Radial Keratotomy in the PERK Study." *Archives of Ophthalmology* 105: 42–51.
- McCabe, G. P. 1984. "Principal Variables." *Technometrics* 26: 137–44.
- Morrison, D. F. 1976. *Multivariate Statistical Methods*. New York: McGraw-Hill.
- Mosteller, F., and Tukey, J. W. 1977. *Data Analysis and Regression*. Reading, Mass.: Addison-Wesley.
- Muller, K. E., and Chen Mok, M. 1997. "The Distribution of Cook's Statistic." *Communications in Statistics: Theory and Methods* 26(3): 525–46.
- Obenchain, R. L. 1977. Letter to the Editor. *Technometrics* 19: 348–49.
- Picard, R. R., and Cook, R. D. 1984. "Cross-validation of Regression Models." *Journal of the American Statistical Association* 79: 575–83.
- Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. New York: Chapman and Hall.
- Shapiro, S. S., and Wilks, M. B. 1965. "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika* 52: 591–611.
- Siegel, S. 1956. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.
- Smith, G., and Campbell, F. 1980. "A Critique of Some Ridge Regression Methods." *Journal of the American Statistical Association* 75: 74–81.
- Stephens, M. A. 1974. "EDF Statistics for Goodness of Fit and Some Comparisons." *Journal of the American Statistical Association* 69: 730–37.
- Stevens, J. P. 1984. "Outliers and Influential Data Points in Regression Analysis." *Psychology Bulletin* 95: 334–44.
- Timm, N. H. 1975. *Multivariate Analysis with Applications in Education and Psychology*. Belmont, Calif.: Wadsworth.
- Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley.
- Weisberg, S. 1980. *Applied Linear Regressions*. New York: John Wiley & Sons.

15

Polynomial Regression

15.1 Preview

In this chapter, we focus on a special case of the multiple regression model, the *polynomial model*, which is often of interest when only *one basic* independent variable—say, X —is to be considered. We initially considered a straight-line model (Chapter 5) for this situation; however, we may want to determine whether prediction can be improved significantly by increasing the complexity of the fitted straight-line model. The simplest extension of the straight-line model is the second-order polynomial, or *parabola*, which involves a second term, X^2 , in addition to X . Adding high-order terms like X^2 and X^3 , which are simple functions of a single basic variable, can be considered equivalent to adding new independent variables. Thus, if we rename X as X_1 and X^2 as X_2 , the second-order model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + E$$

becomes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$$

In general, polynomial models are special cases of the general multiple regression model. Since only one basic independent variable is being considered, however, any polynomial model can be represented by a curvilinear plot on a two-dimensional graph (rather than as a surface in higher-dimensional space). As mentioned in Chapter 5, when only one basic independent variable X is being considered, the fundamental goal is to find the curve that best fits the data so that the relationship between X and Y is appropriately described. Because a higher-order curve may be more appropriate than a straight line, a discussion of how to fit and evaluate such (polynomial) curves is important.

We first consider methods for fitting and evaluating the second-order (parabolic) model, after which we consider higher-order polynomial models. Because these models are special

cases of the general multiple regression model, the procedure for fitting these models and the methods for inference are essentially the same as those described more generally in Chapters 8 and 9. Since the independent variables in a polynomial model are functions of the same basic variable (X), they are inherently correlated. This, in turn, can lead to computational difficulties due to collinearity (Chapter 14). Fortunately, techniques such as centering and the use of orthogonal polynomials are available to help remedy such problems; these procedures are discussed later in this chapter. We shall also see that the use of orthogonal polynomials helps simplify hypothesis testing.

15.2 Polynomial Models

The most general kind of curve usually considered for describing the relationship between a single independent variable X and a response Y is called a *polynomial*. Mathematically, a polynomial of order k in x is an expression of the form

$$y = c_0 + c_1x + c_2x^2 + \cdots + c_kx^k$$

in which the c 's and k (which must be a nonnegative integer) are constants. We have already considered the simple polynomial corresponding to $k = 1$ (namely, the straight line having the form $y = c_0 + c_1x$). The second-order polynomial corresponding to $k = 2$ (namely, the parabola) has the general form $y = c_0 + c_1x + c_2x^2$.

In going from a *mathematical* model to a *statistical* model, as we did in the straight-line case, we may write a parabolic model in either of the following forms:

$$\mu_{Y|X} = \beta_0 + \beta_1X + \beta_2X^2 \quad (15.1)$$

or

$$Y = \beta_0 + \beta_1X + \beta_2X^2 + E \quad (15.2)$$

If we tentatively assume that a parabolic model—as given by either (15.1) or (15.2)—is appropriate for describing the relationship between X and Y , we must then determine a specific estimated parabola that best fits the data. As in the straight-line case, this best-fitting parabola may be determined by employing the least-squares method.

15.3 Least-squares Procedure for Fitting a Parabola

The least-squares estimates of the parameters β_0 , β_1 , and β_2 in a parabolic model are chosen so as to minimize the sum of squares of deviations of observed points from corresponding points on the fitted parabola (Figure 15.1). Letting $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ denote the least-squares estimates of the unknown regression coefficients in the parabolic model

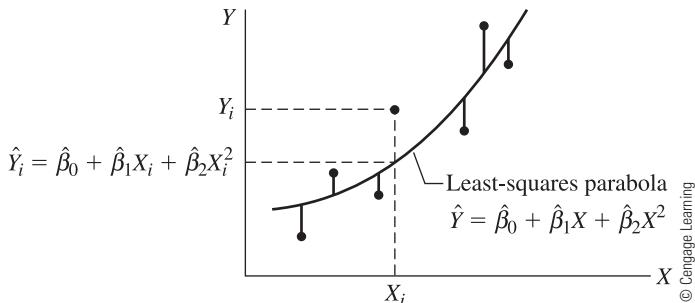


FIGURE 15.1 Deviations of observed points from the least-squares parabola

(15.1) and letting \hat{Y} denote the value of the predicted response at X , we can write the estimated parabola as

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 \quad (15.3)$$

The minimum sum of squares obtained by using this least-squares parabola is

$$\text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i - \hat{\beta}_2 X_i^2)^2 \quad (15.4)$$

As with the general regression model, we do not find it necessary to present the precise formulas for calculating the least-squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$. These formulas are quite complex and become even more so for polynomials of orders higher than two. The researcher is not likely to employ such polynomial regression methods without using a packaged computer program, which can perform the necessary calculations and print the numerical results. (Appendix B contains a discussion of matrices and their relationship to regression analysis; by using matrix mathematics, we can compactly represent the general regression model and the associated least-squares methodology.)

■ **Example 15.1** For the age–systolic blood pressure data of Table 5.1, with the outlier removed,¹ the least-squares estimates for the parabolic regression coefficients are computed to be

$$\hat{\beta}_0 = 113.41, \quad \hat{\beta}_1 = 0.088, \quad \hat{\beta}_2 = 0.010$$

¹ As described in Section 5.5.3, the outlier corresponds to the data point ($X = 47$, $Y = 220$) for the second individual listed in Table 5.1.

The fitted model given by (15.3) then becomes

$$\hat{Y} = 113.41 + 0.088X + 0.010X^2 \quad (15.5)$$

This equation can be compared with the straight-line equation obtained in Section 5.5 for these data with the outlier removed—namely,

$$\hat{Y} = 97.08 + 0.95X \quad (15.6)$$

A comparison between (15.5) and (15.6) reveals that the estimates of β_0 and β_1 differ in the two models, indicating that the estimation of β_2 affects the estimation of β_0 and β_1 in the quadratic model. ■

15.4 ANOVA Table for Second-order Polynomial Regression

As in the straight-line case, the essential results gathered from a second- or higher-order polynomial model can be summarized in an ANOVA table. The ANOVA table for a parabolic fit to the age–systolic blood pressure data of Table 5.1 (with the outlier removed) is given in Table 15.1.

The contents of Table 15.1 deserve comment. First, only variables-added-in-order tests are described. Natural variable orderings suggest themselves, either from the largest to the smallest power of the predictor or vice versa. Consequently, a variables-added-last test for each term in the final model should be avoided with polynomial models. Using variables-added-in-order tests aids in choosing the most parsimonious yet relevant model possible. Such tests might utilize the residual mean square from the largest model considered or from the sequential “full” models at every step of variable addition (as done in Section 9.3.2). This is discussed more fully in Section 15.10.

TABLE 15.1 ANOVA table for a parabola fit to the age–systolic blood pressure data of Table 5.1, with the outlier removed

Source	d.f.	SS	MS	F
Regression	X	6,110.10	6,110.10	68.89
	$X^2 X$	163.30	163.30	1.84
Residual	26	2,306.05	88.69	
Total (corrected)*	28	8,579.45		

Note: The residual from the largest model was used for all tests.

* $R^2 = 0.731$

15.5 Inferences Associated with Second-order Polynomial Regression

Three basic inferential questions are associated with second-order polynomial regression:

1. Is the overall regression significant? That is, is more of the variation in Y explained by using the second-order model than by ignoring X completely (and just using \bar{Y})?
2. Does the second-order model provide significantly more predictive power than the straight-line model does?
3. Given that a second-order model is more appropriate than a straight-line model, should we add higher-order terms (X^3 , X^4 , etc.) to the second-order model?

15.5.1 Test for Overall Regression and Strength of the Overall Parabolic Relationship

Determining whether the overall regression is significant involves testing the null hypothesis H_0 : “There is no significant overall regression using X and X^2 ” (i.e., $\beta_1 = \beta_2 = 0$). The testing procedure used for this null hypothesis involves the overall F test described in Chapter 9—namely, computing

$$F = \frac{\text{MS Regression}}{\text{MS Residual}}$$

and then comparing the value of this F statistic with an appropriate critical point of the F distribution, which (in our example) has 2 and 26 degrees of freedom in the numerator and denominator, respectively. For $\alpha = .001$, we find that $F = 35.37 > F_{2, 26, 0.999} = 9.12$, so we reject the null hypothesis of nonsignificant overall regression ($P < .001$).

To obtain a quantitative measure of how well the second-order model predicts the dependent variable, we can use the squared multiple correlation coefficient (the multiple R^2). As with r^2 in straight-line regression, R^2 represents the proportionate reduction in the error sum of squares obtained by using X and X^2 instead of the naive predictor \bar{Y} . The formula for calculating R^2 is given by

$$R^2(\text{second-order model}) = \frac{\text{SSY} - \text{SSE}(\text{second-order model})}{\text{SSY}} \quad (15.7)$$

For this example, $R^2 = 0.731$. The preceding F test (with $P < .001$) tells us that this R^2 is significantly different from 0.

15.5.2 Test for the Addition of the X^2 Term to the Model

To answer the second question, about increased predictive power, we must perform a partial F test of the null hypothesis H_0 : “The addition of the X^2 term to the straight-line model does not significantly improve the prediction of Y over and above that achieved by

the straight-line model itself" (i.e., $\beta_2 = 0$). To test this null hypothesis, we compute the partial F statistic

$$F(X^2|X) = \frac{(\text{Extra SS due to adding } X^2)/1}{\text{MS Residual(second-order model)}} \quad (15.8)$$

and then compare this F value to an appropriate F percentage point (which is an $F_{1, 26}$ value in our example). Since X^2 is the last variable added, this is a variables-added-last test. Alternatively, we could divide the estimated coefficient $\hat{\beta}_2$ by its estimated standard error to form a statistic that has a t distribution under H_0 with 26 degrees of freedom.

The ANOVA information needed to compute the F test for our example is given in Table 15.1. The extra sum of squares for $X^2|X$ is 163.30 and is computed as the difference between the sum-of-squares regression values for the first- and second-order models. The partial F statistic (15.8) is then computed to be

$$F = \frac{163.30}{88.69} = 1.84$$

Since $F_{1, 26, 0.90} = 2.91$, we would not reject H_0 at the $\alpha = .10$ level. Specifically, the P -value for this test is .19. Thus, for this example, we conclude that adding a quadratic term to the straight-line model does not significantly improve prediction. Corroboration of the results of this partial F test is provided by a scatter diagram of the data (Figure 5.8 in Chapter 5), which offers no evidence of a parabolic relationship between X and Y . Finally, the conclusion is also supported by the small increase in R^2 when the X^2 term is added to the straight-line model. Since $r^2 = R^2 = 0.712$ for the straight-line model and $R^2 = 0.731$ for the parabolic model, the increase in R^2 is $0.731 - 0.712 = 0.019$.

15.5.3 Testing for Adequacy of the Second-order Model

The preceding analysis of the Table 5.1 data has shown that a straight-line model fits the data adequately, is significantly predictive of the response, and is preferable to a parabolic model. Consequently, it would seem superfluous in this case to evaluate whether a model of an order higher than two would be significantly better than a straight-line model. Nevertheless, in general, any question of model adequacy can be addressed (with a *lack-of-fit* test) for any model (of any order) that is being considered at a given stage of an analysis. Any such lack-of-fit test can be characterized by a partial or multiple partial F test for the addition of one or more terms to the model under study. More detailed discussion of lack-of-fit tests is provided in the subsequent sections.

15.6 Example Requiring a Second-order Model

We now turn to another hypothetical example to illustrate the methods of polynomial regression. This example will lead us to a different conclusion regarding the appropriateness of a second-order model.

TABLE 15.2 Weight gain after two weeks as a function of dosage level

Dosage level (X)	1	2	3	4	5	6	7	8
Weight gain (Y) (dag)	1	1.2	1.8	2.5	3.6	4.7	6.6	9.1

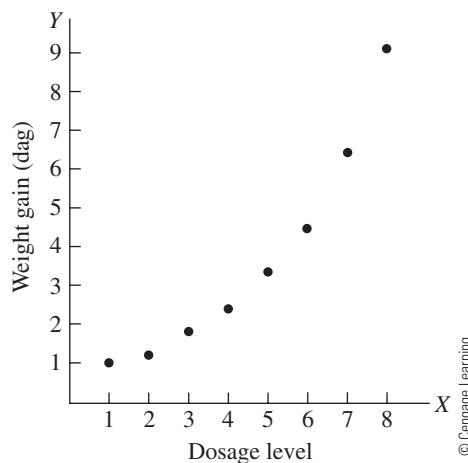
© Cengage Learning

Suppose that a laboratory study is undertaken to determine the relationship between the dosage (X) of a certain drug and weight gain (Y). Eight laboratory animals of the same sex, age, and size are selected and randomly assigned to one of eight dosage levels.²

The gain in weight (in dekagrams [10 grams]) is measured for each animal after a two-week time period during which all animals were subject to the same dietary regimen and general laboratory conditions. The data are given in Table 15.2 and a scatter diagram for these data in Figure 15.2. By simply eyeballing this diagram, one can see that a parabolic curve is a more appropriate model than a straight line. We shall now proceed to quantify this visual impression.

The complete ANOVA table, based on fitting a parabola to the data in Table 15.2, is given in Table 15.3. The equation for the least-squares parabola on which the ANOVA table is based has the form

$$Y = 1.35 - 0.41X + 0.17X^2$$



© Cengage Learning

FIGURE 15.2 Scatter diagram of hypothetical data for the animal weight gain study

² The study design here can certainly be criticized for not having more than one animal receive a particular dosage, as well as for involving such a small total sample size. Replication at each dosage would provide a reliable estimate of animal-to-animal variation in the data. However, for some laboratory studies, sufficient numbers of animals are not easily obtainable; cost and time are often limiting factors as well. Finally, the data for this example have been constructed to simplify the analysis and to present a relationship that is clearly second-order in nature.

TABLE 15.3 Regression ANOVA table for the straight-line model fit to the weight gain data

Source	d.f.	SS	MS	F
Regression	X	1	52.04	52.04
	$X^2 X$	1	4.83	4.83
Residual	5	0.20	0.04	
Total (corrected)*	7	57.07		

* $R^2 = 0.997$.

© Cengage Learning

Let us investigate the information contained in this ANOVA table. First, by dividing the regression X sum of squares by the residual mean square for the straight-line model, we can test whether there is a significant straight-line regression effect before we add the X^2 term to the model; in particular, the ANOVA table for straight-line regression derived from Table 15.3 is given in Table 15.4. The least-squares line is $\hat{Y} = -1.20 + 1.11X$. The null hypothesis of no significant linear regression is clearly rejected, since an F of 62.05 exceeds $F_{1, 6, 0.999} = 35.51$ ($P < .001$). Our next step is to examine the complete ANOVA table and decide whether adding the X^2 term significantly improves the prediction of Y over and above what is obtained from a simple straight-line model. In doing so, we are asking whether the increase in R^2 of 0.085 (0.997 – 0.912), obtained by including the X^2 term in the model, significantly improves the fit. The appropriate test statistic to use in answering this question is the partial F statistic

$$F = \frac{\text{(Extra sum of squares due to adding } X^2\text{)}/1}{\text{MS Residual (second-order model)}} \\ = \frac{4.83}{0.04} = 120.75$$

which exceeds $F_{1, 5, 0.999} = 47.18$ ($P < .001$). Therefore, adding the X^2 term to the model significantly improves prediction. As might be expected, a test for overall significant regression of the second-order model yields a highly significant F —namely,

TABLE 15.4 Regression ANOVA table for the linear model fit to the weight gain data

Source	d.f.	SS	MS	F
Regression (X)	1	52.04	52.04	62.05
Residual	6	5.03	0.84	
Total (corrected)*	7	57.07		

* $R^2 = 0.9118$.

© Cengage Learning

$$F = \frac{\text{MS Regression (second-order model)}}{\text{MS Residual (second-order model)}}$$

$$= \frac{(52.04 + 4.83)/2}{0.04} = 710.88$$

Up to this point, we have concluded that a first-order (straight-line) model is not as good as a second-order model. We now need to determine whether adding higher-order terms to the second-order model is warranted. For example, we can add an X^3 term (with associated regression coefficient β_3) to the second-order model and then test whether prediction is significantly improved. Fitting this third-order model by least squares results in the ANOVA table given in Table 15.5. To test whether adding the third-order term significantly improves the fit, we calculate the following statistic:

$$F = \frac{(\text{Extra sum of squares due to adding } X^3)/1}{\text{MS Residual (third-order model)}}$$

$$= \frac{0.14}{0.014} = 10.00$$

This F statistic has an F distribution with 1 and 4 degrees of freedom under H_0 : “The addition of the X^3 term is not worthwhile” (i.e., $\beta_3 = 0$). Since $F_{1, 4, 0.95} = 7.71$ and $F_{1, 4, 0.975} = 12.22$, we have $.025 < P < .05$. This P -value would thus reject H_0 for $\alpha = .05$ but not for $\alpha = .025$. This makes the decision of whether to include the X^3 term in the model somewhat difficult. However, several other factors should be taken into consideration: (1) the R^2 -value for the parabolic fit is very high (namely, 0.997); (2) the R^2 -value only increases from 0.997 to 0.999 in going from a second-order model to a third-order model; (3) the scatter diagram clearly suggests a second-order curve; (4) when in doubt, the simpler model is preferable because it is easier to interpret. All things considered, then, it is most sensible to conclude that the second-order model is most appropriate.

TABLE 15.5 Regression ANOVA table for the cubic model fit to the weight gain data

Source	d.f.	SS	MS	F
Regression	X	1	52.037	52.04
	$X^2 X$	1	4.835	4.84
	$X^3 X, X^2$	1	0.141	0.14
Residual	4	0.056	0.014	10.00*
Total (corrected) [†]	7	57.066		

* $.025 < P < .05$

[†] $R^2 = 0.999$

In summary, for the data in Table 15.2, the best-fitting model is

$$\hat{Y} = 1.35 - 0.41X + 0.17X^2$$

with an R^2 of 0.997.

15.7 Fitting and Testing Higher-order Models

So far we have seen how the basic ideas of multiple regression may be applied to fitting and testing quadratic and cubic polynomial models. These same methods generalize to all higher-order polynomial models.

How large an order of polynomial model to consider depends on the problem being studied and the amount and type of data being collected. For studies in the biological and social sciences, one important consideration is whether the regression relationship can be described by a monotonic function (i.e., one that is always increasing or decreasing). If only monotonic functions are of interest, a second- or third-order model usually suffices (although monotonicity is not guaranteed, since, for example, some parabolas increase and then decrease). A large number of well-placed predictor values and a small error variance are needed to obtain reliable fits for models of higher order than cubic.

A more general consideration is the number of *bends* (more technically, *relative extrema*) in the polynomial curve that one wishes to fit. For example, a first-order model has no bends; a second-order model has no more than one bend; and each higher-order term adds another potential bend. In practice, fitting polynomial models of orders higher than three usually leads to models that are neither always decreasing nor always increasing. Substantial theoretical and/or empirical evidence should exist to support the employment of such complicated nonmonotonic models.

The quantity of data directly limits the maximum order of a polynomial that may be fit. Consider the weight gain data (Table 15.2). Given those eight distinct values, a polynomial of order seven would fit the eight points perfectly, giving an SSE value of 0 and an R^2 -value of 1. (Nevertheless, because the fitted equation would have eight estimated parameters, no gain in parsimony is made over simply listing the eight data points.) *Generally, the maximum-order polynomial that may be fit is one less than the number of distinct X-values.* Thus, observations that are *replicates* (i.e., observations sharing the same X -value) contribute only once to the number of distinct X -values observed. For example, consider the age–systolic blood pressure data in Table 5.1 (with the outlier removed). Of the 29 observations, 5 are replicates, which implies that 24 distinct X -values exist. Hence, a polynomial of order 23 could be fit to these data, although it would be highly unlikely that a model of such high order would be necessary to provide good prediction.

15.8 Lack-of-fit Tests

Given that a polynomial model has been fitted and the estimated regression coefficients have been tested for their significance, how can one be confident that a model of order higher than the highest order tested is probably not needed? A lack-of-fit (LOF) test can be used to

TABLE 15.6 Replicates and pure-error estimates for the age–systolic blood pressure data of Table 5.1

X	Y	SS*	d.f.
39	144, 120	288.0	1
42	124, 128	8.0	1
45	138, 135	4.5	1
56	154, 150	8.0	1
67	170, 158	72.0	1
		380.5 (SS _{PE})	5

*The sum of squares for a given X is calculated by using the formula $\sum_m (Y_{mx} - \bar{Y}_x)^2$, where Y_{mx} is the mth observation of Y at $X = x$ and \bar{Y}_x is the mean of all replicates at $X = x$.

© Cengage Learning

address this question. Conceptually, an LOF test evaluates a model more complex than the one under primary consideration.

The classical LOF test can be applied only if there are replicate observations. With n total observations, if d X -values are distinct, the number of replicates is $r = n - d$. Recall that a polynomial curve of order $d - 1$ can pass through exactly d distinct points. A classical LOF test compares the fit of a polynomial of order $d - 1$ with the fit of the polynomial model currently under consideration.

For the age–systolic blood pressure example (with the outlier removed), 5 X -values out of the total of 29 involve replicates (i.e., $r = 5$). These are listed in Table 15.6. (Notice that these replicate data in Table 15.6 call into question the validity of the variance homogeneity assumption.) In a classical LOF test of these data, we consider a polynomial of order $d - 1 = n - r - 1 = 29 - 5 - 1 = 23$, which is the higher order of the two polynomial models being compared. The lower-order polynomial is the model of primary interest, such as the second-order model fit to the age–systolic blood pressure data. For our example, the two models would be

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + E$$

and

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_{23} X^{23} + E$$

Table 15.7 contains the ANOVA information needed to compare these two models. The F test for such a comparison is a multiple partial F test of the null hypothesis $H_0: \beta_j = 0$, $j = 3, 4, \dots, 23$. (Trying to fit the larger 23rd-degree polynomial directly would lead to serious collinearity problems; a possible alternative method of fitting, described in Section 15.9, involves the use of orthogonal polynomials.)

The ANOVA framework for the classical LOF test (see Table 15.7) partitions the residual sum of squares (SSE) for the model whose fit is being questioned into two components:

TABLE 15.7 Regression ANOVA table for the classical LOF test of the second-order polynomial fit to the age–systolic blood pressure data of Table 5.1

Source	d.f.
X	1
$X^2 X$	1
Residual { Lack of fit ($X^3, X^4, \dots, X^{23} X, X^2$)	21
Pure error	5
Total	28

Note: The LOF test statistic is given by

$$\begin{aligned} F &= F(X^3, X^4, \dots, X^{23}|X, X^2) \\ &= \frac{[\text{Regression SS}(X, X^2, \dots, X^{23}) - \text{Regression SS}(X, X^2)]/21}{\text{Residual SS}(X, X^2, \dots, X^{23})/5} \\ &= \frac{\text{MS}_{\text{LOF}}}{\text{MS}_{\text{PE}}} \end{aligned}$$

Under H_0 : "No lack of fit of second-order model," this F statistic should have an F distribution with 21 and 5 degrees of freedom.

© Cengage Learning

a pure-error sum of squares SS_{PE} (with degrees of freedom df_{PE}); and an LOF sum of squares SS_{LOF} . The test statistic, which is equivalent to the multiple partial F test just described, can be written as $F = \text{MS}_{\text{LOF}}/\text{MS}_{\text{PE}}$. In actuality, SS_{PE} is the error sum of squares for the higher-degree polynomial model being compared, of order $d - 1 = n - r - 1$. The quantity SS_{LOF} is the extra sum of squares due to the addition to the lower-order model of all higher-order terms needed to construct the higher-order model. In the example under consideration,

$$\text{SS}_{\text{LOF}} = \text{Regression SS}(X, X^2, \dots, X^{23}) - \text{Regression SS}(X, X^2)$$

and

$$\text{SS}_{\text{PE}} = \text{Residual SS}(X, X^2, \dots, X^{23})$$

With most computer regression packages, using a multiple partial F test for lack of fit can be less computationally cumbersome than identifying replicate observations in order to compute SS_{PE} directly. Orthogonal polynomials *must* be used to avoid serious inaccuracy in the multiple regression computations leading to SS_{LOF} .

15.9 Orthogonal Polynomials

In Chapter 14, we illustrated collinearity problems that can arise in work with polynomial models, and we demonstrated that centering the predictor helped remedy such problems for a second-order polynomial model. A more sophisticated approach is needed for higher-order

models. The polynomials we have discussed so far have all been *natural polynomials*. This terminology derives from the fact that each of the independent variables (X, X^2, X^3 , etc.) in a polynomial of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + E$$

is a simple polynomial by itself. Another method of fitting polynomial models involves *orthogonal polynomials*, which constitute a new set of independent variables that are defined in terms of the simple polynomials but have more complicated structures. In this section, we explain the method and describe its advantages and disadvantages. The basic motivation for using orthogonal polynomials is to avoid the serious collinearity inherent in using natural polynomials.

In Section 15.2, we alluded to the natural polynomial model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + E$$

The simple polynomials X, X^2, \dots, X^k are the predictors. Orthogonal polynomial variables are new predictor variables that consist of linear combinations of these simple polynomials. Denoting orthogonal polynomials as $X_1^*, X_2^*, \dots, X_k^*$, we can write them as linear combinations of the form

$$\begin{aligned} X_1^* &= a_{01} + a_{11}X \\ X_2^* &= a_{02} + a_{12}X + a_{22}X^2 \\ &\vdots \\ X_k^* &= a_{0k} + a_{1k}X + a_{2k}X^2 + \cdots + a_{kk}X^k \end{aligned}$$

where the a 's are constants that relate the X^* 's to the original predictors. Each linear combination has the form of a polynomial. It can be shown that each simple polynomial can be written as a linear combination of the X^* 's as follows:

$$\begin{aligned} X &= b_{01} + b_{11}X_1^* \\ X^2 &= b_{02} + b_{12}X_1^* + b_{22}X_2^* \\ &\vdots \\ X^k &= b_{0k} + b_{1k}X_1^* + b_{2k}X_2^* + \cdots + b_{kk}X_k^* \end{aligned}$$

where the b 's are constants. With no loss of information, we can write either

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + E$$

or

$$Y = \beta_0^* + \beta_1^* X_1^* + \beta_2^* X_2^* + \cdots + \beta_k^* X_k^* + E$$

The parameters $\{\beta_j^*\}$ in the latter model differ numerically from those in the former. Moreover, whereas the simple polynomials are highly correlated with one another, the orthogonal polynomials (via appropriate specification of the a 's) are pairwise uncorrelated.³ However,

³ $r_{\mathbf{x}_j^*, \mathbf{x}_{j'}^*} = 0$ for all $j \neq j'$.

although the parameters and predictors in the two models have very different interpretations, it can be shown that the multiple R^2 -values and the overall regression F tests obtained by fitting these two models are exactly the same.

In summary, the orthogonal polynomial values represent a recoding of the original predictors with two desirable basic properties: *the orthogonal polynomial variables contain exactly the same information as the simple polynomial variables*; and *the orthogonal polynomial variables are uncorrelated with each other*. The first property means that all questions about the natural polynomial model can be answered by using the orthogonal polynomial model. For example, one can assess the overall strength of the regression relationship, conduct the overall regression F test, and even compute partial F tests while using the orthogonal polynomial model. The second property, zero pairwise correlation, completely eliminates any collinearity.

With regard to partial F tests, it can be shown that the partial F test of $H_0: \beta_j^* = 0$ for the orthogonal polynomial model of order k is equivalent to the partial F test of $H_0: \beta_j = 0$ in the reduced natural polynomial model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_j X^j + E$$

for any $j \leq k$. Thus, only one orthogonal polynomial model has to be fit (of the highest order—say, k —of interest) in order to test sequentially for the significance of each simple polynomial term in the original natural polynomial model. This may be summarized by saying that the k tests of $H_0: \beta_j^* = 0$, for $j = 1, 2, \dots, k$, using the orthogonal polynomial model are equivalent to the k variables-added-in-order tests of $H_0: \beta_j = 0$, for $j = 1, 2, \dots, k$, using an appropriate sequence of k natural polynomial models.

As an illustration, consider a third-order natural polynomial model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + E$$

Suppose that it is of interest to determine a best model by proceeding backward, starting with a test of $H_0: \beta_3 = 0$ in the preceding cubic model. If this test is nonsignificant, one would proceed to a test of $H_0: \beta_2 = 0$ in a (reduced) parabolic model; in turn, if $H_0: \beta_2 = 0$ is not rejected, one would perform a test of $H_0: \beta_1 = 0$ in a (reduced) straight-line model. One way to conduct these three tests would be to fit three separate natural polynomial models (cubic, parabolic, and straight-line models) and then to perform partial F (or t) tests—starting with the cubic model—to assess the significance of the highest-order term in each model, stopping at any point if significance is obtained. Unfortunately, collinearity will generally compromise reliable fitting of the cubic and (without centering) quadratic models. To avoid this problem, one can fit a *single* third-order orthogonal polynomial model and then test sequentially $H_0: \beta_3^* = 0$, $H_0: \beta_2^* = 0$, and finally $H_0: \beta_1^* = 0$, stopping at any point if significance is obtained. This approach generally ensures computational accuracy.

15.9.1 Transforming from Natural to Orthogonal Polynomials

In this section, we describe a method for expressing orthogonal polynomials in terms of the original simple polynomials. This procedure involves specifying the a 's in the earlier equations relating the X 's to powers of X . To illustrate the method, assume that the scientist who collected the data in Table 15.8 decided to try to confirm the conclusions made in the first study by collecting new data. The same eight dosage values, from 1 to 8, were used. Three

TABLE 15.8 Observed weight gains and simple polynomial values

Observation	WGTGAIN (Y)	DOSE1 (X)	DOSE2 (X^2)	DOSE3 (X^3)	DOSE4 (X^4)	DOSE5 (X^5)	DOSE6 (X^6)	DOSE7 (X^7)
1	0.9	1	1	1	1	1	1	1
2	1.1	2	4	8	16	32	64	128
3	1.6	3	9	27	81	243	729	2187
4	2.3	4	16	64	256	1024	4096	16384
5	3.5	5	25	125	625	3125	15625	78125
6	5.0	6	36	216	1296	7776	46656	279936
7	6.6	7	49	343	2401	16807	117649	823543
8	8.7	8	64	512	4096	32768	262144	2097152
9	0.9	1	1	1	1	1	1	1
10	1.1	2	4	8	16	32	64	128
11	1.6	3	9	27	81	243	729	2187
12	2.1	4	16	64	256	1024	4096	16384
13	3.4	5	25	125	625	3125	15625	78125
14	4.5	6	36	216	1296	7776	46656	279936
15	6.7	7	49	343	2401	16807	117649	823543
16	8.6	8	64	512	4096	32768	262144	2097152
17	0.8	1	1	1	1	1	1	1
18	1.2	2	4	8	16	32	64	128
19	1.4	3	9	27	81	243	729	2187
20	2.2	4	16	64	256	1024	4096	16384
21	3.2	5	25	125	625	3125	15625	78125
22	4.8	6	36	216	1296	7776	46656	279936
23	6.7	7	49	343	2401	16807	117649	823543
24	8.8	8	64	512	4096	32768	262144	2097152

Note: The linear predictor variable X is called DOSE1. The quadratic predictor variable X^2 is called DOSE2, etc.; in general, the variable X^j is called DOSE j .

© Cengage Learning

new observations were taken at each dosage, giving a total of 24 observations. Since eight distinct X -values are available, the highest-order polynomial that can be fitted is of the seventh order. To demonstrate the use of orthogonal polynomials and an associated LOF test, let us consider fitting third- and seventh-order polynomials to these data.

Table 15.8 provides the data from the new study and also gives the values of the X^2, X^3, \dots, X^7 terms (which are labeled DOSE2, DOSE3, ..., DOSE7 in the table). These are the natural polynomial values. Table 15.9 includes the same response (Y) values, but the predictor (X) values have been replaced by orthogonal polynomial values (labeled ODOSE1, ODOSE2, ..., ODOSE7). These values were obtained by using Table A.7 in Appendix A as follows.

Table A.7 can be used *only if* the predictor (X) values are equally spaced and if the same number of observations (i.e., replicates) occurs at each value. If either of these conditions is not satisfied, Table A.7 cannot be used. In that case, a reasonable alternative is to use a computer program (such as the ORPOL function in SAS PROC IML) to calculate the appropriate orthogonal polynomial values.

In describing the transition from Table 15.8 to Table 15.9, we begin by noting that the variable “dosage” has eight distinct, equally spaced values and that three replicates are taken

TABLE 15.9 Observed weight gains and orthogonal polynomial values

Observation	WGTGAIN	ODOSE1	ODOSE2	ODOSE3	ODOSE4	ODOSE5	ODOSE6	ODOSE7
1	0.9	-7	7	-7	7	-7	1	-1
2	1.1	-5	1	5	-13	23	-5	7
3	1.6	-3	-3	7	-3	-17	9	-21
4	2.3	-1	-5	3	9	-15	-5	35
5	3.5	1	-5	-3	9	15	-5	-35
6	5.0	3	-3	-7	-3	17	9	21
7	6.6	5	1	-5	-13	-23	-5	-7
8	8.7	7	7	7	7	7	1	1
9	0.9	-7	7	-7	7	-7	1	-1
10	1.1	-5	1	5	-13	23	-5	7
11	1.6	-3	-3	7	-3	-17	9	-21
12	2.1	-1	-5	3	9	-15	-5	35
13	3.4	1	-5	-3	9	15	-5	-35
14	4.5	3	-3	-7	-3	17	9	21
15	6.7	5	1	-5	-13	-23	-5	-7
16	8.6	7	7	7	7	7	1	1
17	0.8	-7	7	-7	7	-7	1	-1
18	1.2	-5	1	5	-13	23	-5	7
19	1.4	-3	-3	7	-3	-17	9	-21
20	2.2	-1	-5	3	9	-15	-5	35
21	3.2	1	-5	-3	9	15	-5	-35
22	4.8	3	-3	-7	-3	17	9	21
23	6.7	5	1	-5	-13	-23	-5	-7
24	8.8	7	7	7	7	7	1	1

Note: The linear orthogonal polynomial predictor variable is called ODOSE1, the quadratic orthogonal polynomial variable is called ODOSE2, etc.; in general, the j th-order variable is called ODOSE j .

© Cengage Learning

at each of these values. Hence, Table A.7 can be used. In Table A.7, k indicates the number of distinct values of X —in our case, 8. The row in Table A.7 labeled “Linear” for $k = 8$ consists of the entries $(-7, -5, -3, -1, 1, 3, 5, 7)$. These eight entries will be used to replace the eight original values for the linear term X . Thus, a subject who received a dosage of 1 (as did the first subject) is given a linear orthogonal polynomial score of -7 ; a subject who received a dosage of 2 is given a linear orthogonal polynomial score of -5 ; and so on, with a dosage of 8 corresponding to a score of 7. This pattern is repeated in Table 15.9 twice more for the replicate observations. Analogous operations yield the remaining columns appearing in Table 15.9. For example, the eight quadratic entries for $k = 8$ $(7, 1, -3, -5, -5, -3, 1, 7)$ give the orthogonal polynomial values for the variable ODOSE2 in Table 15.9, which replace the original eight values for the quadratic term X^2 (given in the corresponding column of Table 15.8).

Finally, the rightmost column of Table A.7 gives the sum of squared values for each row in the table. For $k = 8$, dividing each linear orthogonal polynomial score by $\sqrt{168}$, each quadratic score by $\sqrt{168}$, each cubic score by $\sqrt{264}$, and so on makes the variance of each set of orthogonal polynomial scores (i.e., each set of row entries) equal to 1. This helps in two ways. First, numerical accuracy is improved by the avoidance of scaling problems. Second,

the estimated standard errors of the resulting estimated regression coefficients are all equal, which simplifies the task of comparing and interpreting such regression coefficients.

15.9.2 Regression Analysis with Orthogonal Polynomials

Let us now consider the regression ANOVA tables based on fitting third- and seventh-order polynomial models to the weight gain data in Tables 15.8 and 15.9. Table 15.10 summarizes the results for the third-order natural polynomial model. Clearly, the model fits extremely well ($R^2 = 0.998$). Furthermore, the variables-added-in-order tests give P -values of .0001, .0001, and .3995 for linear, quadratic, and cubic tests, respectively. These results argue persuasively for a second-order model. Since the simple polynomials X , X^2 , and X^3 are correlated, the variables-added-last tests are more difficult to interpret. Nevertheless, the cubic test again gives $P = .3995$, since it is the last variable to enter; the linear effect is not significant ($P = .1359$), but the quadratic effect is ($P = .0046$). When using variables-added-last tests, analysts should proceed backward, starting with the cubic model and including in the final

TABLE 15.10 Third-order natural polynomial model for the weight gain data ($n = 24$)

Source	d.f.	SS	MS
Model*	3	169.7122	56.5707
Error	20	0.3274	0.0164
Total (corrected) [†]	23	170.0396	
Source	d.f.	Variables-Added-in-Order SS	F
DOSE1 (X)	1	155.6667	9508.98 ($P < .0001$)
DOSE2 ($X^2 X$)	1	14.0334	857.23 ($P < .0001$)
DOSE3 ($X^3 X, X^2$)	1	0.0121	0.74 ($P = .3995$)
Source	d.f.	Variables-Added-Last SS	F
DOSE1 ($X X^2, X^3$)	1	0.0395	2.41 ($P = .1359$)
DOSE2 ($X^2 X, X^3$)	1	0.1662	10.15 ($P = .0046$)
DOSE3 ($X^3 X, X^2$)	1	0.0121	0.74 ($P = .3995$)
Parameter	Estimate	S_{β_j}	T for $H_0: \text{Parameter} = 0$
Intercept	1.0261	0.182	5.64 ($P = .0001$)
DOSE1	-0.2559	0.165	-1.55 ($P = .1359$)
DOSE2	0.1316	0.041	3.19 ($P = .0046$)
DOSE3	0.0026	0.003	0.86 ($P = .3995$)

* $F = 3,455.65$ ($P < .0001$).

[†] $R^2 = 0.998075$.

model all lower-order terms below the highest term deemed significant. Then, the series of variables-added-last tests also supports the choice of the second-order model.

The preference for a second-order model is further supported by comparing R^2 -values for the second- and third-order models. It can be shown, using the variables-added-in-order SS entries in Table 15.10, that the R^2 -value for the second-order model is

$$R^2 = \frac{155.6667 + 14.0334}{170.0396} = 0.998$$

which is the same (within round-off error) as the R^2 -value for the third-order model.

For comparison, consider the results of fitting the third-order orthogonal polynomial model, as summarized in Table 15.11. The most important difference between Tables 15.10 and 15.11 is that the two types of sum of squares (namely, variables-added-in-order and variables-added-last) are equal in the latter. Consequently, the corresponding partial F -test results are the same and coincide with the t -test results.

TABLE 15.11 Third-order orthogonal polynomial model for the weight gain data ($n = 24$)

Source	d.f.	SS	MS
Model*	3	169.7122	56.5707
Error	20	0.3274	0.0164
Total (corrected) [†]	23	170.0396	
Source	d.f.	Variables-Added-in-Order SS	<i>F</i>
ODOSE1	1	155.6667	9,508.98 ($P < .0001$)
ODOSE2	1	14.0334	857.23 ($P < .0001$)
ODOSE3	1	0.0121	0.74 ($P = .3995$)
Source	d.f.	Variables-Added-Last SS	<i>F</i>
ODOSE1	1	155.6667	9,508.98 ($P < .0001$)
ODOSE2	1	14.0334	857.23 ($P < .0001$)
ODOSE3	1	0.0121	0.74 ($P = .3995$)
Parameter	Estimate	$S_{\hat{\beta}_j}$	<i>T</i> for $H_0:$ Parameter = 0
Intercept	3.6542	0.0261	139.91 ($P < .0001$)
ODOSE1	0.5558	0.0057	97.51 ($P < .0001$)
ODOSE2	0.1669	0.0057	29.27 ($P < .0001$)
ODOSE3	0.0039	0.0045	0.86 ($P = .3995$)

* $F = 3,455.65$ ($P < .0001$).

[†] $R^2 = 0.998075$.

The preceding results, whether from analysis using natural or orthogonal polynomials, indicate that we need not consider any polynomial model higher than second order. Nevertheless, it is instructive to use these data to illustrate particular problems that arise if the model order does exceed two. Specifically, we address the (hidden) problem of collinearity. The existence of this collinearity problem can be demonstrated by considering the predictor correlations given in Table 15.12 for both the simple polynomials and their centered counterparts—that is, for $(X - \bar{X})$, $(X - \bar{X})^2$, ..., as well as for X, X^2, \dots . We first focus on the correlations among the linear, quadratic, and cubic terms in a third-order model; this information is contained in the upper left corner of each array (above and to the left of the dashed lines). The existence of a collinearity problem is suggested by the presence of three very high correlations (all above 0.93) for the noncentered predictor data. Centering X helps reduce collinearity: two of the three correlations become 0 after centering. Turning to the complete array, we again see, for the noncentered polynomials, that the smallest off-diagonal correlation is 0.779 and that four are greater than 0.990. As before, centering X helps reduce collinearity substantially, since correlations between odd and even powers of centered X values are then all 0 for these equally spaced X values. Nevertheless, the remaining correlations are high, with two correlations greater than 0.99. The analogous correlation array using orthogonal polynomials would have all zeros off the diagonal.

In Table 15.13, the collinearity problem suggested by Table 15.12 can be examined further. Predictor correlation matrix eigenvalues (see Chapter 14) are reported here for third- and seventh-order natural, centered, and orthogonal polynomial models. For the third-order natural polynomial model, the condition number (CN; see Chapter 14) is 70.8, suggesting a

TABLE 15.12 Predictor correlations for the weight gain data ($n = 24$)

Simple Polynomials							
	X	X^2	X^3	X^4	X^5	X^6	X^7
X	1	0.976	0.932	0.887	0.846	0.810	0.779
X^2		1	0.988	0.963	0.935	0.908	0.882
X^3			1	0.993	0.978	0.960	0.941
X^4				1	0.996	0.986	0.973
X^5					1	0.997	0.990
X^6						1	0.998
X^7							1

Centered Simple Polynomials							
	$(X - \bar{X})$	$(X - \bar{X})^2$	$(X - \bar{X})^3$	$(X - \bar{X})^4$	$(X - \bar{X})^5$	$(X - \bar{X})^6$	$(X - \bar{X})^7$
$(X - \bar{X})$	1	0	0.926	0	0.855	0	0.813
$(X - \bar{X})^2$		1	0	0.969	0	0.932	0
$(X - \bar{X})^3$			1	0	0.985	0	0.965
$(X - \bar{X})^4$				1	0	0.992	0
$(X - \bar{X})^5$					1	0	0.996
$(X - \bar{X})^6$						1	0
$(X - \bar{X})^7$							1

TABLE 15.13 Eigenvalues of predictor correlation matrices for polynomial models for the weight gain data ($n = 24$)

Third-order Model			
Eigenvalue	Natural Polynomial		Orthogonal Polynomial
	Uncentered	Centered	
1	2.931	1.926	1.000
2	0.069	1.000	1.000
3	6×10^{-4}	0.074	1.000
CN	70.8	5.1	1.0

Seventh-order Model			
Eigenvalue	Natural Polynomial		Orthogonal Polynomial
	Uncentered	Centered	
1	6.638	3.773	1.000
2	0.348	2.929	1.000
3	0.013	0.221	1.000
4	3×10^{-4}	0.071	1.000
5	4×10^{-6}	0.006	1.000
6	2×10^{-8}	5×10^{-4}	1.000
7	2×10^{-11}	2×10^{-5}	1.000
CN	569,664.0	430.0	1.0

© Cengage Learning

collinearity problem. If the X -variable is centered, the condition number is reduced to 5.1, indicating no severe collinearity problem. Unfortunately, as shown by the correlation arrays in Table 15.12, such centering does not solve the collinearity problem for higher-order models. For a seventh-order model based on centered dosage data, the condition number is a disturbing 430.0. And the condition number for an uncentered, seventh-order natural polynomial model is an extremely alarming 569,664. In contrast, the condition number for both the third-order and the seventh-order orthogonal polynomial models is 1.0.

The preceding condition numbers lead us to recommend using only orthogonal polynomials (i.e., we rule out using natural polynomials) to conduct LOF tests. Table 15.14 summarizes the ANOVA results based on fitting a seventh-order orthogonal polynomial model to the $n = 24$ dosage-weight gain observations. The lower-order polynomial results (up to, say, order three) remain virtually unchanged. This follows from the fact that the quadratic model fits so well and from the properties of orthogonal polynomials. After the .0001 P -values for the linear and quadratic terms, the next smallest P -value is .0958 for the fourth-order term. As before, a second-order polynomial model seems most appropriate.

TABLE 15.14 Seventh-order orthogonal polynomial model for the weight gain data ($n = 24$)

Source	d.f.	SS	MS
Model*	7	169.7796	24.2542
Error	16	0.2600	0.0163
Total (corrected) [†]	23	170.0396	
Source	d.f.	Variables-Added-in-Order SS	F
ODOSE1	1	155.6667	9,579.49 ($P < .0001$)
ODOSE2	1	14.0334	863.59 ($P < .0001$)
ODOSE3	1	0.0121	0.75 ($P = .4003$)
ODOSE4	1	0.0509	3.13 ($P = .0958$)
ODOSE5	1	0.0000	0.00 ($P = .9924$)
ODOSE6	1	0.0036	0.22 ($P = .6420$)
ODOSE7	1	0.0128	0.79 ($P = .3871$)
Source	d.f.	Variables-Added-Last SS	F
ODOSE1	1	155.6667	9,579.49 ($P < .0001$)
ODOSE2	1	14.0334	863.59 ($P < .0001$)
ODOSE3	1	0.0121	0.75 ($P = .4003$)
ODOSE4	1	0.0509	3.13 ($P = .0958$)
ODOSE5	1	0.0000	0.00 ($P = .9924$)
ODOSE6	1	0.0036	0.22 ($P = .6420$)
ODOSE7	1	0.0128	0.79 ($P = .3871$)
Parameter	Estimate	S_{β_j}	T for H_0 : Parameter = 0
Intercept	3.6541	0.0260	140.54 ($P = .0001$)
ODOSE1	0.5558	0.0057	98.00 ($P = .0001$)
ODOSE2	0.1669	0.0057	29.43 ($P = .0001$)
ODOSE3	0.0039	0.0045	0.86 ($P = .4003$)
ODOSE4	-0.0052	0.0030	-1.77 ($P = .0958$)
ODOSE5	0.0000	0.0016	0.01 ($P = .9924$)
ODOSE6	-0.0021	0.0045	0.47 ($P = .6420$)
ODOSE7	-0.0011	0.0013	-0.89 ($P = .3871$)

* $F = 1,492.57$ ($P = .0001$).

[†] $R^2 = 0.998075$.

© Cengage Learning

As discussed in Section 15.8, the LOF test for the second-order model may be performed by using a multiple partial F statistic of the form

$$F = \frac{[\text{Regression SS (seventh order)} - \text{Regression SS (second order)}]/(7 - 2)}{\text{Residual SS (seventh order)}/(24 - 1 - 7)}$$

The actual value of the statistic in our example is

$$F = \frac{(169.7795 - 169.7001)/5}{0.2600/16} = 0.98$$

With 5 and 16 degrees of freedom for $\alpha = .25$, the critical value is 1.48, so the P -value must be $> .25$. Hence, we fail to reject the null hypothesis of no lack of fit; that is, a second-order model provides a good description of these data.

15.10 Strategies for Choosing a Polynomial Model

In our discussion of polynomial models, we have sometimes started with the smallest model, involving only a linear term, and sequentially added increasing powers of X . This is a forward-selection model-building strategy. Although it is a natural approach, it can produce misleading results for inference-making procedures.

With a forward-selection strategy, one usually tests for the importance of a candidate predictor by comparing the extra regression sum of squares for the addition of that predictor to the residual mean square. This residual mean square is based on fitting a model containing the candidate (predictor) variable and the variables already in the model. The corresponding partial F statistic is of the form

$$F(X_j|X, X^2, \dots, X^{j-1}) = \frac{\text{SS}(X^j|X, X^2, \dots, X^{j-1})/1}{\text{MS Residual}(X, X^2, \dots, X^j)}$$

when one is testing for the importance of X^j in a polynomial model already containing terms through X^{j-1} . The mean-square residual in the preceding expression is not based on terms of order higher than X^j , even though such terms may actually belong in the final model.

The forward-selection testing approach just described can lead to underfitting the data (i.e., the forward-selection algorithm is likely to quit too soon, thereby choosing a model of an order lower than is actually required). The reason for this problem is that, when the computations proceed forward, the residual mean-square error estimate of σ^2 at any step will be biased upward if the polynomial model at that step is of too low a degree. Since this can cause the denominator in the above partial F expression to be too large, the F statistic itself may be too small and hence nonsignificant, thus stopping the forward-selection algorithm prematurely.

The underfitting bias can be avoided by using a backward-elimination strategy (see Chapter 16), where the denominator in the partial F statistic at each backward step involves the residual mean square for the full (or largest) model fitted. When using this backward-elimination approach, however, one may overfit the data (i.e., choose a final model of an order slightly higher than required). Fortunately, the MS Residual computed using the full model is still a valid (unbiased) estimator of σ^2 . Consequently, using this estimator in the denominator

of the partial F test at any backward step is a statistically valid procedure. By slightly overfitting the data, one loses some statistical power, but usually this loss is negligible.

Throughout the model selection process, it is important to iteratively conduct the residual analysis methods of Chapter 14 in order to assess model appropriateness. Of particular use for polynomial regression is a plot of jackknife residuals against X . The need for a higher-order model often appears as a nonlinear trend in the residuals. In the appropriate higher-order model, the nonlinear trend in the residuals should be eliminated.

Problems

- 1.** In an environmental engineering study of a certain chemical reaction, the concentrations of 18 separately prepared solutions were recorded at different times (three measurements at each of six times). The natural logarithms of the concentrations were also computed. The data recorded are reproduced in the following table:

Solution Number (<i>i</i>)	TIME (X_i) (hr)	Concentration (Y_i) (mg/ml)	ln Concentration ($\ln Y_i$)
1	6	0.029	-3.540
2	6	0.032	-3.442
3	6	0.027	-3.612
4	8	0.079	-2.538
5	8	0.072	-2.631
6	8	0.088	-2.430
7	10	0.181	-1.709
8	10	0.165	-1.802
9	10	0.201	-1.604
10	12	0.425	-0.856
11	12	0.384	-0.957
12	12	0.472	-0.751
13	14	1.130	0.122
14	14	1.020	0.020
15	14	1.249	0.222
16	16	2.812	1.034
17	16	2.465	0.902
18	16	3.099	1.131

- a. Plot on separate sheets of graph paper:
 - (1) Concentration (Y) versus time (X).
 - (2) Natural logarithm of concentration ($\ln Y$) versus time (X).
- b. Using the accompanying computer output, obtain the following:
 - (1) The estimated equation of the straight-line (degree 1) regression of Y on X .
 - (2) The estimated equation of the quadratic (degree 2) regression of Y on X .
 - (3) The estimated equation of the straight-line (degree 1) regression of $\ln Y$ on X .
 - (4) Graphs of each of these fitted equations on their respective scatter diagrams.
- c. Based on the accompanying computer output, complete the following table for the straight-line regression of Y on X .

Source	d.f.	SS	MS	F
Regression	1			
Residual	4			
Total	12			
	17			

- d. Based on the accompanying computer output, complete the following ANOVA table.

Source	d.f.	SS	MS	F
Regression	1			
Residual	3			
Total	12			
	17			

- e. Determine and compare the proportions of the total variation in Y that are explained by the straight-line regression on X and by the quadratic regression on X .
- f. Carry out F tests for the significance of the straight-line regression of Y on X and for the adequacy of fit of the estimated regression line.
- g. Carry out an overall F test for the significance of the quadratic regression of Y on X , a test for the significance of the addition of X^2 to the model, and an F test for the adequacy of fit of the estimated quadratic model.
- h. For the straight-line regression of $\ln Y$ on X , carry out F tests for the significance of the overall regression and for the adequacy of fit of the straight-line model.
- i. What proportion of the variation in $\ln Y$ is explained by the straight-line regression of $\ln Y$ on X ? Compare this result with the result you obtained in part (e) for the quadratic regression of Y on X .
- j. A fundamental assumption in regression analysis is variance homoscedasticity.
- Examine the scatter diagrams constructed in part (a), and state why taking natural logarithms of the concentrations helps with regard to the assumption of variance homogeneity.
 - Is the straight-line regression of $\ln Y$ on X better for describing this set of data than the quadratic regression of Y on X ? Explain.

Edited SAS Output (PROC REG and PROC RSREG) for Problem 1

Straight-line regression of Y on X

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-1.93180	0.42858	-4.51	0.0004
X	1	0.24597	0.03721	6.61	<.0001

(continued)

Quadratic regression of Y on X

Regression	DF	Type I Sum of Squares	R-Square	F Value	Pr > F
Linear	1	12.705408	0.7320	255.15	<.0001
Quadratic	1	3.905067	0.2250	78.42	<.0001
Total Model	2	16.610475	0.9570	166.79	<.0001

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	3	0.514457	0.171486	8.85	0.0023
Pure Error	12	0.232482	0.019374		
Total Error	15	0.746939	0.049796		

Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	3.172052	0.603016	5.26	<.0001
X	1	-0.781023	0.116989	-6.68	<.0001
X*X	1	0.046682	0.005271	8.86	<.0001

Straight-line regression of ln Y on X

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-6.20956	0.07704	-80.60	<.0001
X	1	0.45117	0.00669	67.45	<.0001

Quadratic regression of ln Y on X

Regression	DF	Type I Sum of Squares	R-Square	F Value	Pr > F
Linear	1	42.745786	0.9965	4277.04	<.0001
Quadratic	1	0.000396	0.0000	0.04	0.8448
Total Model	2	42.746182	0.9965	2138.54	<.0001

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	3	0.027439	0.009146	0.90	0.4714
Pure Error	12	0.122475	0.010206		
Total Error	15	0.149914	0.009994		

- k. What key assumption about the data would be in question if, instead of 18 different solutions, there were only 3 different solutions, each of which was analyzed at the six different time points?

2. With the addition of five pairs of observations—(18,000, 39.2), (22,400, 27.9), (24,210, 22.3), (5,400, 11.7), and (9,340, 32.5)—to the data in Problem 3 in Chapter 5, the accompanying computer output is obtained for the regression of TIME (Y) on INC (X).
- a. Using the accompanying computer output, complete the following ANOVA table for the straight-line regression of TIME (Y) on INC (X).

Source	d.f.	SS	MS	F
Regression	1			
Residual	18			
	5			
Total	24			

- b. Using the accompanying computer output, complete the following ANOVA table for the quadratic regression of TIME (Y) on INC (X).

Source	d.f.	SS	MS	F
Regression	1			
	1			
Residual	17			
	5			
Total	24			

- c. On the scatter diagram of the data for this problem, plot the fitted straight-line (degree 1) equation and the fitted quadratic (degree 2) equation.
- d. Calculate and compare the R^2 -values obtained for the straight-line, quadratic, and cubic fits.
- e. Carry out F tests for the significance of the straight-line regression and for the adequacy of fit of the straight-line model.
- f. Carry out F tests for the significance of the quadratic regression, of the addition of the quadratic term to the model, and of the adequacy of fit of the quadratic model.
- g. Which model is most appropriate: straight-line, quadratic, or cubic?

Edited SAS Output (PROC REG and PROC RSREG) for Problem 2

Linear regression of TIME (Y) on INC (X)

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	20.17655	4.60655	4.38	0.0002
X	1	0.00061197	0.00030000	2.04	0.0530

(continued)

Quadratic regression of TIME (Y) on INC (X)

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-19.86602	3.90293672	-5.09	<.0001
X	1	0.00787	0.00064060	12.29	<.0001
X2	1	-0.00000025	0.00000002	-11.52	<.0001

Cubic regression of TIME (Y) on INC (X)

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-35.292777	8.14734679	-4.33	0.0003
X	1	0.012227	0.00214424	5.70	<.0001
X2	1	-0.00000059	0.00000016	-3.66	0.0015
X3	1	7.80716E-12	0.00000000	2.11	0.0466

Regression of TIME (Y) on INC (X)

[Portion of output omitted]

Regression	DF	Type I Sum of Squares
Linear	1	442.914529
Quadratic	1	2100.080668
Cubic	1	61.100193

Residual	DF	Sum of Squares	Mean Square
Lack of Fit	16	271.748723	16.984300
Pure Error	5	15.201250	3.040250

3. a.-g. For the data on DIST (Y) and MPH (X) in Problem 7 in Chapter 5, use the following information to answer the same questions as in parts (a) through (g) of Problem 2.

$$\text{Degree 1 fit: } \hat{Y} = -122.345 + 6.227X$$

$$\text{Degree 2 fit: } \hat{Y} = 32.901 - 3.051X + 0.1176X^2$$

$$\text{Degree 3 fit: } \hat{Y} = 114.621 - 10.620X + 0.3247X^2 - 0.00173X^3$$

Source	d.f.	SS	MS
Regression	1	173,473.96	173,473.96
	1	10,515.44	10,515.44
	1	415.19	415.19
Residual	11	2,664.15	242.20
	4	3,433.93	858.48
Total	18	190,502.67	

4. For the data on VOTE (Y) and TVEXP (X) in Problem 5 in Chapter 5, it was found that the straight-line model was adequate. Using the accompanying computer output for quadratic regression, do the following:
- Plot the fitted straight-line model and the fitted quadratic model on the scatter diagram for the data of this problem.
 - Determine the change in R^2 in going from a degree 1 to a degree 2 model.
 - Test for the significance of the addition of the X^2 term to the model.
 - Assess whether the results in parts (a) through (c) contradict your earlier conclusion about the adequacy of fit of the straight-line model.

Edited SAS Output (PROC REG and PROC RSREG) for Problem 4

Straight-line regression of Y on X

Root MSE	3.33177	R-Square	0.9101
Dependent Mean	45.71000	Adj R-Sq	0.9051
Coeff Var	7.28894		

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	2.17407	3.30974	0.66	0.5196
X	1	1.17696	0.08718	13.50	<.0001

Quadratic regression of Y on X

Regression	DF	Type I Sum of Squares	R-Square	F Value	Pr > F
Linear	1	2023.205063	0.9101	174.27	<.0001
Quadratic	1	2.450153	0.0011	0.21	0.6518
Total Model	2	2025.655215	0.9112	87.24	<.0001

(continued)

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	12	166.802785	13.900232	2.27	0.1874
Pure Error	5	30.560000	6.112000		
Total Error	17	197.362785	11.609576		

Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	9.656782	16.636082	0.58	0.5692
X	1	0.750772	0.931996	0.81	0.4316
X*X	1	0.005746	0.012508	0.46	0.6518

5. For the regression of PCI (Y) on YNG (X) for the African countries considered in Problem 2 in Chapter 12, use the accompanying information to do the following:
- Plot the estimated straight-line and quadratic models on the scatter diagram for the data on the African countries.
 - Test for the significance of the straight-line regression and for the adequacy of fit of the straight-line model.
 - Test for the significance of the addition of the X^2 term to the model.
 - Which model is more appropriate, the straight-line model or the quadratic model?

$$\text{Degree 1 fit: } \hat{Y} = 893.57 - 17.276X$$

$$\text{Degree 2 fit: } \hat{Y} = 732.05 - 9.203X - 0.0996X^2$$

Source	d.f.	SS	MS
Regression $\begin{cases} X \\ X^2 X \end{cases}$	1	153,784.8	153,784.8
	1	88.3	88.3
Residual $\begin{cases} \text{Lack of fit} \\ \text{Pure error} \end{cases}$	15	2,773.9	184.9
	8	911.5	113.9
Total	25	157,558.5	

6. For the data given in Problem 11 in Chapter 7, which concerns the relationship between the temperature (X) of a certain medium and the growth (Y) of human amniotic cells in a tissue culture, researchers wish to evaluate whether a parabolic model is more appropriate than a straight-line model. Use the accompanying computer output to answer the following questions:
- Plot the fitted straight-line model and the fitted quadratic model on the same scatter diagram.
 - Test for the significance of adding the X^2 term to the model.
 - Determine the change in R^2 in going from a straight-line to a parabolic model.
 - How do your results in parts (b) and (c) compare with the results in Problem 11 in Chapter 7 for the test of adequacy of fit of the straight-line model?
 - Which model is more appropriate—straight line or parabolic?

Edited SPSS Output for Problem 6

DEPENDENT VARIABLE.. Y		ANALYSIS OF VARIANCE			MEAN SQUARE		
VARIABLE(S) ENTERED ON STEP NUMBER 1.. X		REGRESSION	RESIDUAL	DF	SUM OF SQUARES	F	
MULTIPLE R	0.98603			1.	12.83072		
R SQUARE	0.97225				0.36618		
ADJUSTED R SQUARE	0.97071						
STANDARD ERROR	0.14263						
 -----VARIABLES IN THE EQUATION-----		 -----VARIABLES NOT IN THE EQUATION-----			 -----VARIABLES NOT IN THE EQUATION-----		
VARIABLE	B	BETA	STD ERROR B	DF	BETA IN	PARTIAL	MEAN SQUARE
X	0.03582	0.98603	0.00143	630.716	0.84602	0.64297	12.83072
(CONSTANT)	-0.46240						
VARIABLE(S) ENTERED ON STEP NUMBER 2.. XX		ANALYSIS OF VARIANCE			MEAN SQUARE		
MULTIPLE R	0.99183			2.	12.98210		
R SQUARE	0.98372						
ADJUSTED R SQUARE	0.98181						
STANDARD ERROR	0.11241						
 -----VARIABLES IN THE EQUATION-----		 -----VARIABLES NOT IN THE EQUATION-----			 -----VARIABLES NOT IN THE EQUATION-----		
VARIABLE	B	BETA	STD ERROR B	DF	VARIABLE	BETA IN	MEAN SQUARE
X	0.00537	0.14782	0.00887	0.367			6.49105
XX	0.00022	0.84502	0.00006	11.981			513.73362
(CONSTANT)	0.49460						
MAXIMUM STEP REACHED							

Note: B stands for the regression coefficients $\hat{\beta}$, XX stands for X^2 , and you can ignore for now the terms "BETA," "PARTIAL," and "TOLERANCE." Also,

$$\text{adjusted } R^2 = R^2 - \left(\frac{k}{n-k-1} \right) (1 - R^2).$$

From *Statistical Package for the Social Sciences* by Nie et al. Copyright © 1975 by McGraw-Hill, Inc. Used with permission of McGraw-Hill Book Company and Dr. Norman Nie, President, SPSS Inc.

7. The skin response in rats to different concentrations of a newly developed vaccine was measured in an experiment, resulting in the data, models, and computer output that follow.

Concentration (X) (ml/l)	0.5	0.5	1.0	1.0	1.5	1.5	2.0	2.0	2.5	2.5	3.0	3.0
Skin response (Y) (mm)	13.90	13.81	14.08	13.99	13.75	13.60	13.32	13.39	13.45	13.53	13.59	13.64

- a. Plot the straight-line, quadratic, and cubic equations on the scatter diagram for this data set.
- b. Test sequentially for significant straight-line fit, for significant addition of X^2 , and for significant addition of X^3 to the model.
- c. Which of the three models do you recommend, and why? (Note: You might also want to consider R^2 for each model.)

Edited SAS Output (PROC REG) for Problem 7

Linear regression of Y on X

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.28440	0.28440	8.22	0.0168
Error	10	0.34609	0.03461		
Corrected Total	11	0.63049			

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	13.98633	0.12246	114.21	<.0001
X	1	-0.18029	0.06289	-2.87	0.0168

Quadratic regression of Y on X

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.35362	0.17681	5.75	0.0246
Error	9	0.27688	0.03076		
Corrected Total	11	0.63049			

(continued)

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	14.27050	0.22186	64.32	<.0001
X	1	-0.60654	0.29030	-2.09	0.0663
X2	1	0.12179	0.08119	1.50	0.1679

Cubic regression of Y on X

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.52218	0.17406	12.86	0.0020
Error	8	0.10831	0.01354		
Corrected Total	11	0.63049			

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	13.36167	0.29665	45.04	<.0001
X	1	1.67997	0.67601	2.49	0.0378
X2	1	-1.39294	0.43264	-3.22	0.0122
X3	1	0.28852	0.08177	3.53	0.0077

8. This problem uses the data presented in the table for Problem 12 in Chapter 5. Use $\alpha = .05$.
 - a. Use a computer program to fit a natural polynomial cubic model for predicting vocabulary size as a function of age in years. Provide estimated regression coefficients.
 - b. Using variables-added-in-order tests, determine the best model.
 - c. Report variables-added-last tests. Explain any differences between your results here and those you obtained in part (b).
 - d. Report appropriate collinearity diagnostics for the model, and evaluate them. Include predictor correlations.
 - e. Plot jackknife (or studentized) residuals against predicted values for the best model based on your results in part (b). Provide a frequency histogram or schematic plot of the residuals, and comment on it.
 - f. Compare your results here with those you obtained in Problem 12 in Chapter 5.
9. a.–e. Repeat Problem 8, parts (a) through (e), after centering the predictor (age).
 - f. Compare the results obtained here to those obtained in Problem 8.
10. a.–e. Repeat Problem 8, parts (a) through (e), but use orthogonal polynomials.
 - f. Compare the results obtained here to those obtained in Problems 8 and 9.

(Hint: Table A.7 cannot be used.)

- 11.** This problem uses the data from Problem 13 in Chapter 5. Use $\alpha = .10$.
- Use a computer program to fit a quadratic natural polynomial model for predicting latency as a function of weight minus average weight.
 - Repeat parts (b) through (e) from Problem 8 for this analysis.
 - Compare the results obtained here to those obtained in Chapter 5.
- 12.** This problem uses the data from Problem 15 in Chapter 5.
- Using a computer program, fit a cubic polynomial model with BLOODTOL as the response and PPM_TOLU as the predictor. Center PPM_TOLU. Provide the prediction equation.
 - Repeat parts (b) through (e) from Problem 8 for this analysis.
 - Compute the estimated variance for each estimated regression coefficient. Are these estimates approximately equal?
 - Compare the results obtained here to those obtained in Chapter 5.
- 13.** **a.** For the data from Problem 15 in Chapter 5, specify the orthogonal polynomial codings needed for coding PPM_TOLU linear, quadratic, and cubic terms. Repeat part (a) of Problem 12, but use the orthogonal coding. (*Hint:* Table A.7 cannot be used.)
- b.–e.** Repeat parts (b) through (e) from Problem 8 for this analysis.
- 14.** To promote safe driving habits and to better protect its customers, an insurance company offers a discount of between 5% and 20% on renewal insurance premiums to customers who have completed a defensive driving course. The following table of data shows the number of customers who have applied for the discount at various discount levels, over a period of 12 months.

Month	Discount (X , %)	Number of Renewing Customers Applying for Discount
1	5	485
2	5	1,025
3	5	1,056
4	10	1,020
5	10	1,149
6	10	1,100
7	15	1,800
8	15	1,805
9	15	1,725
10	20	2,225
11	20	2,325
12	20	2,650

Use the accompanying computer output to answer the following questions.

- Determine the estimated equation of the straight-line regression of the number of customers applying for the discount (Y) on the discount level (X).
- Determine the estimated equation of the quadratic regression of the number of customers applying for the discount (Y) on the discount level (X).
- Plot both estimated models, along with a scatterplot of the data. Which model appears to fit the data better?
- Conduct variables-added-in-order tests for the model in part (b).
- Carry out tests for the significance of the straight-line regression in part (a) and for the adequacy of fit of the estimated regression line.

- f. Carry out tests for the significance of the quadratic regression in part (b) and for the adequacy of fit of the second-order model.
- g. Based on the results from parts (a) through (f), which of the two regressions appears to be more appropriate for predicting the number of customers who apply for the discount?

Edited SAS Output (PROC REG and PROC RSREG) for Problem 14

Straight-line regression of Y on X

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4246956	4246956	90.18	<.0001
Error	10	470929	47093		
Corrected Total	11	4717885			

Root MSE	217.00893	R-Square	0.9002
Dependent Mean	1530.41667	Adj R-Sq	0.8902
Coeff Var	14.17973		

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	200.16667	153.44849	1.30	0.2213
X	1	106.42000	11.20629	9.50	<.0001

Quadratic regression of Y on X

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4360447	2180223	54.90	<.0001
Error	9	357438	39715		
Corrected Total	11	4717885			

Root MSE	199.28707	R-Square	0.9242
Dependent Mean	1530.41667	Adj R-Sq	0.9074
Coeff Var	13.02175		

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	686.41667	320.30915	2.14	0.0607	0
X	1	9.17000	58.44244	0.16	0.8788	32.25000
X2	1	3.89000	2.30117	1.69	0.1252	32.25000

(continued)

[Portion of output omitted]

Regression	DF	Type I Sum of Squares	R-Square	F Value	Pr > F
Linear	1	4246956	0.9002	106.93	<.0001
Quadratic	1	113491	0.0241	2.86	0.1252
Total Model	2	4360447	0.9242	54.90	<.0001

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	1	39990	39990	1.01	0.3448
Pure Error	8	317448	39681		
Total Error	9	357438	39715		

15. Refer to Problem 14. Using the computer output for that problem, and the accompanying output here, answer the following questions.
- Determine the variance inflation factors for the estimated model in part (b) of Problem 14. Does collinearity appear to be a problem?
 - Determine the estimated equation of the quadratic regression of the number of customers applying for the discount (Y) on the centered discount level (Z).
 - Determine the variance inflation factors for the estimated model in part (b) of this problem. Does collinearity appear to be a problem?
 - Conduct variables-added-in-order tests for the model in part (b).
 - Carry out tests for the significance of the quadratic regression in part (b) and for the adequacy of fit of the second-order model.
 - Based on the results from Problem 14 and parts (a) through (e) of this problem, which of the two regressions appears to be more appropriate for predicting the number of customers who apply for the discount?

Edited SAS Output (PROC REG and PROC RSREG) for Problem 15

Quadratic regression of Y on Z

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4360447	2180223	54.90	<.0001
Error	9	357438	39715		
Corrected Total	11	4717885			

(continued)

Root MSE	199.28707	R-Square	0.9242
Dependent Mean	1530.41667	Adj R-Sq	0.9074
Coeff Var	13.02175		

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t 	Variance Inflation
Intercept	1	1408.85417	92.09169	15.30	<.0001	0
Z	1	106.42000	10.29114	10.34	<.0001	1.00000
Z2	1	3.89000	2.30117	1.69	0.1252	1.00000

Quadratic regression of Y on X

Regression	DF	Type I Sum of Squares	R-Square	F Value	Pr > F
Linear	1	4246956	0.9002	106.93	<.0001
Quadratic	1	113491	0.0241	2.86	0.1252
Total Model	2	4360447	0.9242	54.90	<.0001

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	1	39990	39990	1.01	0.3448
Pure Error	8	317448	39681		
Total Error	9	357438	39715		

16. Columbus Airlines introduced a new, specially discounted line of air fares in 1995. Annual ticket revenues Y (in \$1,000s) are shown in the following table, along with the time period X (in months, with $X = 1$ for January 1995).

Month	Ticket Revenues
1	34.9
2	38.8
3	41.5
4	45.1
5	48.3
6	51.2
7	56.6
8	59.9
9	65.4

Use the accompanying computer output to answer the following questions.

- Determine the estimated equation for the straight-line regression of ticket revenues (Y) on month (X).
- Determine the estimated equation for the quadratic regression of ticket revenues (Y) on month (X).
- Plot both estimated models, along with a scatterplot of the data. Which model appears to fit the data better?

- d. Conduct variables-added-in-order tests for the model in part (b).
- e. Carry out tests for the significance of the straight-line regression in part (a).
- f. Carry out tests for the significance of the quadratic regression in part (b).
- g. Examine the variance inflation factors for the estimated model in part (b). Does there appear to be a problem with collinearity?
- h. Based on the results from parts (a) through (g), which of the two regression models appears to be more appropriate for predicting ticket revenues?

Edited SAS Output (PROC REG) for Problem 16

Straight-line regression of Y on X

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	818.44267	818.44267	858.56	<.0001
Error	7	6.67289	0.95327		
Corrected Total	8	825.11556			

Root MSE	0.97636	R-Square	0.9919
Dependent Mean	49.07778	Adj R-Sq	0.9908
Coeff Var	1.98940		

PARAMETER ESTIMATES					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	30.61111	0.70931	43.16	<.0001
X	1	3.69333	0.12605	29.30	<.0001

Quadratic regression of Y on X

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	822.91956	411.45978	1124.21	<.0001
Error	6	2.19599	0.36600		
Corrected Total	8	825.11556			

Root MSE	0.60498	R-Square	0.9973
Dependent Mean	49.07778	Adj R-Sq	0.9965
Coeff Var	1.23269		

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	32.82143	0.76979	42.64	<.0001	0
X	1	2.48771	0.35346	7.04	0.0004	20.48052
X2	1	0.12056	0.03447	3.50	0.0129	20.48052

16

Selecting the Best Regression Equation

16.1 Preview

The general problem to be discussed in this chapter is as follows: we have one response variable Y and a set of k predictor variables X_1, X_2, \dots, X_k ; and we want to determine the *best* subset of the k predictors and the corresponding *best-fitting* regression model for describing the relationship between Y and the X 's.¹ What exactly we mean by “best” depends in part on our overall goal for modeling. Two such goals were described in Chapter 11 and are briefly reviewed now.

One goal is to find a model that provides the best prediction of Y , given X_1, X_2, \dots, X_k , for some new observation or for a batch of new observations. In practice, we emphasize estimating the regression of Y on the X 's (see Chapter 8)— $E(Y|X_1, X_2, \dots, X_k)$ —which expresses the mean of Y as a function of the predictors. Using this goal, we may say that our best model is *reliable* if it predicts well for a new random sample from the population under study. The details of the model may be of little or no consequence, such as whether we include any particular variable or what magnitude or sign we use for its regression coefficient. For example, in considering a sample of systolic blood pressures, we may simply wish to predict systolic blood pressure (SBP) as a function of demographic variables like AGE, RACE, and GENDER. We may not care which variables are in the model or how they are defined as long as the final model obtained gives the best prediction possible.

Alongside the question of prediction is the question of validity—that is, of obtaining accurate estimates for one or more regression coefficient parameters in a model and then making inferences about these parameters of interest. The goal here is to quantify the relationship between one or more independent variables of interest and the dependent variable, controlling when necessary for other variables. For example, we might wish to describe the relationship of

¹ Hocking (1976) provided an exceptionally thorough and well-written review of this topic. His presentation is at a technical level somewhat higher than that of this text.

SBP to AGE, controlling for RACE and GENDER. In this case, we are focusing on regression coefficients involving AGE (including functions of AGE); although RACE and GENDER may remain in the model for control purposes, these variables are not of primary interest.

In this chapter, we focus on strategies for selecting the best model when the primary goal of analysis is prediction. In the last section, we briefly mention a strategy for modeling in situations where the validity of the estimates of one or more regression coefficients is of primary importance (see also Chapter 11).

16.2 Steps in Selecting the Best Regression Equation: Prediction Goal

To select the best regression equation, carry out the following steps:

1. Specify the maximum model (defined in Section 16.3) to be considered.
2. Specify a criterion for selecting a model.
3. Specify a strategy for selecting variables.
4. Conduct the specified analysis.
5. Evaluate the reliability of the model chosen.

By following these steps, one can convert the global goal of finding the best predictors of Y into simple, concrete actions. Each step helps to ensure reliability and to reduce the work required. Specifying the maximum model forces the analyst to state the analysis goal clearly, recognize the limitations of the data at hand, and describe the range of plausible models explicitly. In doing so, the analyst should consider all available scientific knowledge. In turn, specifying the criterion for selecting a model and the strategy for applying that criterion (i.e., for selecting variables) simplifies and speeds the analysis process. Finally, whether the primary goal is prediction or validity, the reliability of the chosen model must be demonstrated.

We illustrate the recommended prediction-oriented process of analysis via two examples. The first of these was introduced in Chapter 8. The data (given in Table 8.1) are the hypothetical results of measuring a response variable weight (WGT) and predictor variables height (HGT) and age (AGE) on 12 children. The second example uses real data from a study involving more than 200 subjects (Lewis and Taylor 1967) discussed in Chapter 14.

16.3 Step 1: Specifying the Maximum Model: Prediction Goal

The *maximum model* is defined to be the largest model (the one having the most predictor variables) considered at any point in the process of model selection. All other possible models can be created by deleting predictor variables from the maximum model. A model created by deleting predictors from the maximum model is called a *restriction* of the maximum model.

Throughout this chapter we assume that the maximum model with k variables or some restriction of it with $p \leq k$ variables is the “correct” model in the population for the k variables selected. An important implication of this assumption is that the *population* squared multiple correlation for the maximum model—namely, $\rho_{Y|X_1, X_2, \dots, X_k}^2$ —is no larger than that for the correct model (which may have fewer variables); as a result, adding more predictors to the correct model does not increase the *population* squared multiple correlation for the correct model. In turn, for the sample at hand, $R_{Y|X_1, X_2, \dots, X_k}^2$ for the maximum model is always at least as large as the corresponding R^2 for any subset model. However, other sample-based criteria may not necessarily suggest that the largest model is best.

To illustrate, consider the data from Table 8.1, reproduced in the following table.

Child	1	2	3	4	5	6	7	8	9	10	11	12
$Y(\text{WGT})$	64	71	53	67	55	58	77	57	56	51	76	68
$X_1(\text{HGT})$	57	59	49	62	51	50	55	48	42	42	61	57
$X_2(\text{AGE})$	8	10	6	11	8	7	10	9	10	6	12	9

One possible model (not necessarily the maximum one) to consider is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$$

This model allows for only a linear (planar) relationship between the response (WGT) and the two predictors (HGT and AGE). Nevertheless, the nature of growth suggests that the relationship, although monotonic in both HGT and AGE, may well be nonlinear. This implies that at least one quadratic (squared) term may have to be included in the model. Since HGT and AGE are highly correlated, and since the sample size is very small,² only $(\text{AGE})^2$ will be considered. The limitations of the data lead us to define the maximum model as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$$

with $X_1 = \text{HGT}$, $X_2 = \text{AGE}$, and $X_3 = (\text{AGE})^2$.

The most important reason for choosing a large maximum model is to minimize the chance of making a Type II (false negative) error. In a regression analysis, a Type II error corresponds to *omitting* a predictor that has a truly nonzero regression coefficient in the population. Other reasons for considering a large maximum model are the wishes to include the following elements:

1. All conceivable basic predictors (AGE, HGT)
2. Higher-order powers of primary predictors [e.g., $(\text{AGE})^2$, $(\text{HGT})^3$]
3. Other transformations of primary predictors [e.g., $\log(\text{AGE})$, $1/\text{HGT}$]
4. Interactions among primary predictors, including two-way (e.g., $\text{AGE} \times \text{HGT}$) and higher-order interactions
5. All possible “control” variables, as well as functions of these control variables

² Because the sample size here is so small, it almost precludes making reliable conclusions from any regression analysis. This small sample size is used to simplify our discussion.

Recall that overfitting a model (including variables in the model with truly zero regression coefficients in the population) will not introduce bias when population regression coefficients are estimated, if the usual regression assumptions are met. We must be careful, however, to ensure that overfitting does not introduce harmful collinearity (Chapter 14). Underfitting (i.e., leaving important predictors out of the final model), however, will introduce bias in the estimated regression coefficients.

There are also important reasons for working with a small maximum model. With a prediction goal, the need for reliability strongly argues for a small maximum model; and with a validity goal, we want to focus on a few important variables. In either case, we wish to avoid a Type I (false positive) error. In a regression analysis, a Type I error corresponds to *including* a predictor that has a zero population regression coefficient. The desire for parsimony is another important reason for choosing a small maximum model: practically unimportant but statistically significant predictors can greatly confuse the interpretation of regression results, and complex interaction terms are particularly troublesome in this regard.

The particular sample of data to be analyzed imposes certain constraints on the choice of the maximum model. In general, the smaller the sample size, the smaller the maximum model should be. The idea here is that a larger number of independent observations are needed to estimate reliably a larger number of regression coefficients. This notion has led to various guidelines about the size of a maximum model. The most basic constraint is that the error degrees of freedom must be positive. Symbolically, we require

$$\text{d.f. error} = n - k - 1 > 0$$

which is equivalent to the constraint

$$n > k + 1$$

As always, n is the number of observations and k is the number of predictors, giving $k + 1$ regression coefficients (including the intercept). With negative error degrees of freedom, the model has at least one perfect collinearity; consequently, unique estimates of regression coefficients and variances cannot be computed. With zero error degrees of freedom ($n = k + 1$), unique estimates of the regression coefficients can be obtained, but unique estimates of variances cannot since $\text{SSE} = 0$. Furthermore, even if the population squared multiple correlation is 0, $R^2 = 1.00$ when $\text{SSE} = 0$, reflecting the fact that the model exactly fits the observed data. In such a situation, we have exchanged n values of Y for n estimated regression coefficients. Hence, we have gained nothing, since the dimensionality of the problem remains the same.

The question then arises as to how many error degrees of freedom are needed. The weakest requirement is for a minimum of approximately 10 error degrees of freedom—namely,

$$n - k - 1 \geq 10$$

or

$$n \geq 10 + k + 1$$

Another suggested rule of thumb for multiple linear regression is to have at least 5 (or 10) observations per predictor—namely, to require that

$$n \geq 5k \quad (\text{or } n \geq 10k)$$

Assume, for example, that we wish to consider a maximum model involving 30 predictors. To have at least 10 error degrees of freedom, we need a minimum sample of size 41, and $n \geq 5k$ demands a sample size of at least 150. Split-sample approaches, discussed later in this chapter, may reduce the required sample size substantially.

Another constraint on the maximum model involves the amount of variability present in the predictor values, considered either individually or jointly. If a predictor has the same value for all subjects, it obviously cannot be used in any model. For example, consider the variable SEX of child (male = 1, female = 0) as a candidate predictor for the weight example. If all the subjects are male, then SEX = 1 for all subjects, the sample variance for this predictor is zero, and it is perfectly collinear with the intercept variable. Clearly, if all subjects are of one gender, comparisons involving gender cannot be made.

Similarly, consider the variables SEX and RACE (white = 0, black = 1) and their interaction (SEXRACE = SEX \times RACE). If no black females are represented in the data, the 1 degree of freedom interaction effect cannot be estimated. If a race–sex cell is nearly empty, the estimated interaction coefficient may be very unstable (i.e., have a large estimated variance).

Polynomial terms (e.g., X^2 and X^3) and other transformations merit particular consideration when the maximum model is being specified. For the weight example, we might wish to consider AGE, (AGE)², and exp(AGE) as possible predictors. Using complicated functions of predictors can lead to near collinearities (see Chapter 14) and to very unstable and often uninterpretable results when trying to find the best model (see Marquardt and Snee 1975, for further discussion on this topic). Consequently, we should attempt to reduce such collinearity if possible. Centering, if applicable, almost always helps increase numerical accuracy, as does multiplying variables by various constants (a special form of scaling; see Chapter 14) to produce nearly equal variances for all predictors. If we do nothing about severe collinearity, the estimated regression coefficients in the best model may be highly unstable (i.e., have large estimated variances) and may have values quite different from the true parameter values.

16.4 Step 2: Specifying a Criterion for Selecting a Model: Prediction Goal

The second step in selecting the best model is to specify the selection criterion. A *selection criterion* is an index that can be computed for each candidate model and used to compare models. Thus, given one particular selection criterion, candidate models can be ordered from best to worst. This helps automate the process of choosing the best model. As we shall see, this selection-criterion-specific process may not find the *best* model in a global sense. Nonetheless, using a specific selection criterion can substantially reduce the work involved in finding a *good* model.

Obviously, the selection criterion should be related to the goal of the analysis. For example, if the goal is reliable prediction of future observations, the selection criterion should be somewhat liberal to avoid missing useful predictors. Distinctions can be drawn among *numerical* differences, *statistically significant* differences, and *scientifically important* differences. Numerical differences may or may not correspond to significant or important differences. With sensitive tests (typical of analyses of large samples), significant differences

may or may not correspond to important differences. And, with insensitive tests (typical of analyses of very small samples), important differences may not be significant.

Many selection criteria for choosing the best model have been suggested. Hocking (1976), for example, reviewed eight candidates. We consider three criteria: R_p^2 , F_p , and C_p . But before discussing these, we must define the notation needed to understand them. All three criteria attempt to compare two model equations: the maximum model with k predictors and a restricted model with p predictors ($p \leq k$). The maximum model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \beta_{p+1} X_{p+1} + \cdots + \beta_k X_k + E \quad (16.1)$$

and the reduced model (a restriction of the maximum model) is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + E \quad (16.2)$$

Let $\text{SSE}(k)$ be the error sum of squares for the k -variable model, let $\text{SSE}(p)$ be the error sum of squares for the p -variable model, and so on. Also, $\text{SSY} = \sum_{i=1}^n (Y_i - \bar{Y})^2$ is the total (corrected) sum of squares for the response Y . Often $p = k - 1$, which is the case when we evaluate the addition or deletion of a single variable. We assume that the $(k - p)$ variables under consideration for addition or deletion are denoted $X_{p+1}, X_{p+2}, \dots, X_k$, for notational convenience.

The sample squared multiple correlation R^2 is a natural candidate for deciding which model is best, so we will discuss this criterion first. The multiple R^2 -value for the p -variable model is

$$R_p^2 = R_{Y|X_1, X_2, \dots, X_p}^2 = 1 - \frac{\text{SSE}(p)}{\text{SSY}} \quad (16.3)$$

Unfortunately, R_p^2 has two potentially misleading characteristics. First, adding predictors, even useless ones, can never decrease R_p^2 . In fact, adding variables invariably increases R_p^2 at least slightly because Regression SS (i.e., $\text{SSY} - \text{SSE}$) always increases with the addition of predictors, thereby decreasing SSE. Second, R_p^2 is always largest for the maximum model, even though a better model may be obtained by deleting some (or even many) variables. The reduced (or restricted) model may be better because it may sacrifice only a negligible amount of predictive strength, while substantially simplifying the model.

Another reasonable criterion for selecting the best model is the F -test statistic for comparing the full and restricted models. This statistic, F_p , can be expressed in terms of sums of squares for error (SSEs) as

$$F_p = \frac{[\text{SSE}(p) - \text{SSE}(k)]/(k - p)}{\text{SSE}(k)/(n - k - 1)} = \frac{[\text{SSE}(p) - \text{SSE}(k)]/(k - p)}{\text{MSE}(k)} \quad (16.4)$$

This statistic may be compared to an F distribution with $k - p$ and $n - k - 1$ degrees of freedom. The criterion F_p tests whether $\text{SSE}(p) - \text{SSE}(k)$, the difference between the residual

sum of squares for the p -variable model and the residual sum of squares for the maximum model, differs significantly from 0. If F_p is *not* significant, we can use the smaller (p -variable) model and achieve roughly the same predictive ability as that yielded by the full model. Hence, a reasonable rule for selecting variables is to retain p variables if F_p is not significant and if p is as small as possible. An often-used special case of F_p occurs when $p = k - 1$, in which case F_p is a test of $H_0: \beta_k = 0$ in the full model.

A less obvious candidate for a selection criterion involving $\text{SSE}(p)$ is Mallow's C_p :

$$C_p = \frac{\text{SSE}(p)}{\text{MSE}(k)} - [n - 2(p + 1)] \quad (16.5)$$

The C_p criterion helps us to decide how many variables to put in the best model, since it achieves a value of approximately $p + 1$ if $\text{MSE}(p)$, the MSE for the p -variable model, is roughly equal to $\text{MSE}(k)$ (i.e., if the correct model is of size p). Knowing the correct model size greatly aids in choosing the best model.

The criteria F_p , R_p^2 , and C_p are intimately related. For example, the F_p test can be expressed in terms of multiple squared correlations as follows:

$$F_p = \frac{(R_k^2 - R_p^2)/(k - p)}{(1 - R_k^2)/(n - k - 1)} \quad (16.6)$$

and the C_p statistic is the following simple function of the F_p statistic:

$$C_p = (k - p)F_p + (2p - k + 1) \quad (16.7)$$

The reason to consider more than one criterion is that no single criterion is always best. In practice, the alternatives can lead to different model choices. In the remainder of the chapter, we shall consider all of the criteria mentioned, at least to some extent. An important aspect of our discussion will be a demonstration of the limitations of R_p^2 as the sole criterion for selecting a model. We favor C_p because it tends to simplify the decision about how many variables to retain in the final model.

16.5 Step 3: Specifying a Strategy for Selecting Variables: Prediction Goal

The third step in choosing the best model is to specify the strategy for selecting variables. Such a strategy is concerned with determining how many variables and also which particular variables should be in the final model. Traditionally, such strategies have focused on deciding whether a single variable should be added to a model (a forward-selection method) or whether a single variable should be deleted from a model (a backward-elimination method). As computers became more powerful, methods for considering more than one variable per step became practical (by generalizing single-variable methods to deal with sets, or *chunks*, of

variables). Before discussing these strategies in detail, we consider an algorithm for evaluating models that, if feasible to conduct, should be the method of choice.

16.5.1 All Possible Regressions Procedure

Whenever practical, the *all possible regressions procedure* (or all subsets regression) is to be preferred over any other variable selection strategy. It is the only method guaranteed to find the model having the largest R_p^2 , the most ideal value of C_p , and so on. This strategy is not always used because the amount of calculation and interpretation of results becomes impractical when the number of variables k in the maximum model is large. The all possible regressions procedure requires that we fit each possible regression equation associated with all possible sets of the k independent variables. For our example, we must fit the seven models corresponding to the following seven sets of independent variables: HGT; AGE; (AGE)²; HGT and AGE; HGT and (AGE)²; AGE and (AGE)²; and HGT, AGE, and (AGE)². For k independent variables, the number of models to be fitted would be $(2^k - 1)$; for example, if $k = 10$, then $(2^{10} - 1) = 1,023$.

Once all $(2^k - 1)$ models have been fitted, we assemble the fitted models into sets involving from 1 to k variables and then order the models within each set according to some criterion (e.g., R_p^2 , F_p , or C_p).

For our data, a summary of the results of the all possible regressions procedure appears in Table 16.1. From the table, the leaders (in terms of R_p^2 -values) in each of the sets involving one, two, and three variables are

One-variable set:	HGT with $R_1^2 = 0.6630$
Two-variable set:	HGT, AGE with $R_2^2 = 0.7800$
Three-variable set:	HGT, AGE, (AGE) ² with $R_3^2 = 0.7802$

Having arranged the top-ranking models of each size in this way, it is helpful to apply a decision rule to the values of R_p^2 to aid in selecting the best model. Due to the properties of R_p^2 discussed earlier, the R_p^2 for the top-ranking models with more predictors will necessarily be greater than that for models with fewer predictors. Some researchers have suggested that more complex models that fail to increase R_p^2 by 0.02 should not be adopted. This decision rule is not necessarily appropriate in all situations; in particular, the choice of an appropriate decision rule depends on the subject matter under consideration and on what the investigators consider to be an adequate level of prediction for their particular situation. Of the three models (models 1, 4, and 7, respectively, in Table 16.1), model 4, involving HGT and AGE, should clearly be our choice, since its R^2 -value is essentially the same as that for model 7 and is much higher than the value for model 1. Thus, our choice of the best regression equation based on the all possible regressions procedure with R_p^2 as the criterion is

$$\text{WGT} = 6.553 + 0.722\text{HGT} + 2.050\text{AGE}$$

Now consider the partial F statistics in Table 16.1. For a given variable in a given model, the associated partial F statistic assesses the contribution made by that variable to the prediction of Y (i.e., WGT) over and above the contributions made by other variables already in the given model. For example, for model 4, which involves only HGT and AGE, the partial

TABLE 16.1 Summary of results of all possible regressions procedure

Model	No. of Variables (p)	Variables Used	Estimated Coefficients				Partial F Statistics			MSE(p)	Overall F Statistic	F_p	R_p^2	C_p
			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	HGT	AGE	(AGE) 2					
1	1	HGT (X_1)	6.190	1.073			19.67**			29.93	19.67**	2.133	0.6630	4.27
2	1	AGE (X_2)	30.571		3.643			14.55**		36.18	14.55**	3.414	0.5926	6.83
3	1	(AGE) 2 (X_3)	45.998		0.206			14.25**		36.63	14.25**	3.505	0.5876	7.01
4	2	HGT, AGE	6.553	0.722	2.050		7.665*	4.785		21.71	15.95**	0.007	0.7800	2.01
5	2	HGT, (AGE) 2	15.118	0.726	0.115		7.601*		4.565	22.07	15.63**	0.138	0.7764	2.14
6	2	AGE, (AGE) 2	32.404		3.205	0.025		0.113	0.002	40.20	6.55*	6.824	0.5927	8.83
7	3	HGT, AGE, (AGE) 2	3.438	0.724	2.777	-0.042	6.827*	0.140	0.010	24.40	9.47**	—	0.7802	4.00

© Cengage Learning

F for AGE is $F(\text{AGE} | \text{HGT}) = 4.785$ but for model 7, which includes HGT, AGE, and (AGE) 2 , the partial F for AGE is $F(\text{AGE} | \text{HGT} (\text{AGE})^2) = 0.140$.

These partial F 's pertain to variables-added-last tests, each based on the MSE for the corresponding model being fit for that row. Such test statistics must be treated with some caution, since any test statistic based on a model with fewer terms than needed can be misleading, perhaps substantially so, because biased variance estimators are being used. Furthermore, many hypothesis tests are being conducted, so that the overall Type I error rate should be higher than the nominal rate α .

Overall F tests are also affected by the estimate of σ^2 being used. With a forward selection algorithm, hypothesis tests conducted early on are often based on MSE values that are too large, thus leading to a premature stopping of the algorithm and hence to the possible omission of important predictors. As calculated in Table 16.1, each overall F statistic is based on the corresponding MSE listed in the same row. If the MSE for the largest model, number 7, had been used in the denominator of each overall F test instead, each resulting test statistic would have been an F_p statistic. Such a statistic involves a comparison of the R_p^2 -value for a reduced model (models 1 through 6) and the R_k^2 -value for the largest model (model 7). For example, for model 4,

$$F_p = \frac{(R_k^2 - R_p^2)/(k - p)}{(1 - R_k^2)/(n - k - 1)} = \frac{(0.7802 - 0.7800)/(3 - 2)}{(1 - 0.7802)/(12 - 3 - 1)} = 0.007$$

In general, such a test is a multiple partial F test. The small F -value here indicates that the predictive abilities of the maximum model and of model 4 do not differ significantly.

Table 16.1 also provides C_p -values. The value of C_p is expected to be close to $(p + 1)$ if the correct model—or a larger model that contains the correct model—is considered. If important

predictors are omitted, C_p should be larger than $(p + 1)$. Also, if $F_p < 1$, then $C_p < (p + 1)$; this can occur when R_p^2 is close enough to R_k^2 in value. Models with C_p not too far from $(p + 1)$ are preferred. Therefore, for a model with one variable, C_p is compared with the value 2.0; for two variables, with the value 3.0; and so on. For this example, no one-variable model has a C_p -value near 2.0. For the only three-variable model, C_p is exactly 4.00. The full model with k predictors is guaranteed to have C_p exactly equal to $(k + 1)$; to see this, examine equation (16.6), the formula for C_p . The two-variable model with AGE and $(AGE)^2$ has a C_p -value much greater than 3, while models 4 and 5 have C_p -values near the minimum possible C_p -value of

$$\frac{(n - k - 1)\text{MSE}(k)}{\text{MSE}(k)} - [n - 2(p + 1)] = (2p - k + 1)$$

which equals 2 when $k = 3$ and $p = 2$. We can see from (16.7) that this lower bound for C_p is attained when $F_p = 0$ and that it can be negative.

The all possible regressions procedure was presented first, since it is preferred whenever practical. It alone is guaranteed to find the best model, in the sense that any selection criterion will be numerically optimized for the particular sample under study. Naturally, using it does not guarantee finding the correct (or population) model. In fact, in many situations, several reasonable candidates for the best model can be found, with different selection criteria suggesting different best models. Furthermore, such findings may vary from sample to sample, even though all the samples are chosen from the same population. Consequently, the choice of the best model may vary from sample to sample. These considerations motivate the discussion in Section 16.7 of evaluating the reliability of the chosen regression equation.

As mentioned earlier, the all possible regressions algorithm is often impractical, since $(2^k - 1)$ models must be fitted when k candidate predictors are being evaluated. Many methods have been suggested as computationally feasible alternatives for approximating the all possible regressions procedure. Although these methods are not guaranteed to find the best model, they can (with careful use) glean essentially all the information from the data needed to choose the best model.

16.5.2 Backward-Elimination Procedure

In the *backward-elimination procedure*, we proceed as follows:

1. Determine the fitted regression equation containing all independent variables (i.e., the estimated maximum model). From the accompanying SAS computer output, we obtain

$$\widehat{\text{WGT}} = 3.438 + 0.724\text{HGT} + 2.777\text{AGE} - 0.042(\text{AGE})^2$$

2. Determine the partial F statistic for every variable in the model as though it were the last variable to enter, and determine the P -values associated with the test statistics. The partial F statistics and P -values are indicated in the Step 0 portion of the SAS output. (Recall that the partial F statistics test whether adding the last variable to

the model significantly helps predict the dependent variable given that the other variables are already in the model.)

3. *Focus on the lowest observed partial F statistic (or, equivalently, on the highest P-value).* From the Step 0 portion of the SAS output, we see that we should focus on (AGE)², represented by “AGE2.”
4. *Compare the P-value with a preselected significance level (say, 10%), and decide whether to remove the variable under consideration. If the variable is dropped, recompute the regression equation for the remaining variables, and repeat backward-elimination procedure Steps 2, 3, and 4. If the variable is not dropped, the backward-elimination procedure ends, and the selected model consists of variables remaining in the model.* From the Step 1 portion of the SAS output, we see that (AGE)² is dropped and the regression model with HGT and AGE as predictors is estimated. With (AGE)² out of the picture, the smallest partial F statistic is 4.78, with an associated P-value of .0565. The P-value is less than .10, so we stop here with this model, which is the same model that we arrived at using the all possible regressions procedure:

$$\widehat{\text{WGT}} = 6.553 + 0.722\text{HGT} + 2.050\text{AGE}$$

Edited SAS Output (PROC REG) for Backward-Elimination Procedure

Backward-Elimination Procedure for Dependent Variable WGT

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.7803 and C(p) = 4.0000

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	693.06046340	231.02015447	9.47	0.0052
Error	8	195.18953660	24.39869208		
Corrected Total	11	888.25000000			

Variable	Parameter Estimate	Standard Error	Type II SS	Partial F statistics	P-value
				F Value	Pr > F
Intercept	3.43842600	33.61081984	0.25534520	0.01	0.9210
HGT	0.72369024	0.27696316	166.58195495	6.83	0.0310
AGE	2.77687456	7.42727877	3.41051231	0.14	0.7182
AGE2	-0.04170670	0.42240715	0.23785686	0.01	0.9238

Bounds on condition number: 89.96948, 543.7942

(continued)

Backward Elimination: Step 1

Variable AGE2 Removed: R-Square = 0.77998605 and C(p) = 2.00974875

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	692.82260654	346.41130327	15.95	0.0011
Error	9	195.42739346	21.7141483		
Corrected Total	11	888.25000000			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	6.55304825	10.94482708	7.78416178	0.36	0.5641
HGT	0.72203796	0.26080506	166.42974940	7.66	0.0218
AGE	2.05012635	0.93722561	103.90008336	4.78	0.0565

Bounds on condition number: 1.604616, 6.418463

Backward Elimination: Step 2

All variables left in the model are significant at the 0.1000 level.

SUMMARY OF BACKWARD ELIMINATION							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	AGE2	2	0.0003	0.7800	2.0097	0.0097	0.923

© Cengage Learning

16.5.3 Forward-Selection Procedure

In the *forward-selection procedure*, we proceed as follows:

1. *Select as the first variable to enter the model the variable most highly correlated with the dependent variable. (Note: The model containing only this particular predictor will have the smallest overall F test P-value among all possible single-variable models.) Then fit the associated straight-line regression equation. If the overall F test for this regression is not significant, stop and conclude that no independent variables are important predictors. If the test is significant, include this variable in the model, and proceed to Step 2 of the procedure.* From the accompanying SAS output, we see that the highest squared correlation is for X_1 , HGT. The straight-line regression equation relating WGT and HGT is

$$\widehat{WGT} = 6.190 + 1.072HGT$$

2. *Determine the partial F statistic and P-value associated with each remaining variable, based on a regression equation containing that variable and the variable*

initially selected. For our data, the statistics are shown in the Step 2 portion of the SAS output.

3. *Focus on the variable with the smallest partial F statistic (i.e., the variable with the largest P-value).* From the Step 2 portion of the SAS output, we see that AGE has the smallest partial F statistic, with a P-value of .0565.
4. *Test for the significance of the partial F statistic associated with the variable from Step 3 of the forward-selection procedure. If this test statistic is significant, add the new variable to the regression equation. If it is not significant, stop the model selection process, and use in the model only the variable added in Step 1.* For our data, since the P-value for AGE is less than .10, we add AGE to get the following two-variable model:

$$\widehat{\text{WGT}} = 6.553 + 0.722\text{HGT} + 2.050\text{AGE}$$

5. *At each subsequent step, determine the partial F statistics for the variables not yet in the model, and then add to the model the variable with the largest partial F-value (if it is statistically significant). At any step, if the largest partial F is not significant, no more variables are included in the model, and the process is terminated.* For our example, we have already added HGT and AGE to the model. We now check to see whether we should add (AGE)², represented by “AGE2.” The partial F for (AGE)², controlling for HGT and AGE, is 0.01, with a P-value of .9238. This value is not statistically significant at $\alpha = .10$. Again, we have arrived at the same two-variable model chosen via the previously discussed methods.

Edited SAS Output (PROC REG) for Forward-Selection Procedure

Forward Selection: Step 0

STATISTICS FOR ENTRY DF = 1,10				
Variable	Tolerance	Model R-Square	F Value	Pr > F
HGT	1.000000	0.6630	19.67	0.0013
AGE	1.000000	0.5926	14.55	0.0034
AGE2	1.000000	0.5876	14.25	0.0036

Forward Selection: Step 1

Variable HGT Entered: R-Square = 0.66301438 and C(p) = 4.26817716

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	588.92252318	588.92252318	19.67	0.0013
Error	10	299.32747682	29.93274768		
Corrected Total	11	888.25000000			

(continued)

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	6.18984871	12.84874620	6.94680569	0.23	0.6404
HGT	1.07223036	0.24173098	588.92252318	19.67	0.0013

Bounds on condition number: 1, 1

Forward Selection: Step 2

STATISTICS FOR ENTRY DF = 1,9				
Variable	Tolerance	Model R-Square	F Value	Pr > F
AGE	0.623202	0.7800	4.7849	0.0565
AGE2	0.621230	0.7764	(4.5647)	(0.0614)

Variable AGE Entered: R-Square = 0.77998605 and C(p) = 2.00974875

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	692.82260654	346.41130327	15.95	0.0011
Error	9	195.42739346	21.71415483		
Corrected Total	11	888.25000000			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	6.55304825	10.94482708	7.78416178	0.36	0.5641
HGT	0.72203796	0.26080506	166.42974940	7.66	0.0218
AGE	2.05012635	0.93722561	103.90008336	4.78	0.0565

Bounds on condition number: 1.6046, 6.4185

Forward Selection: Step 3

STATISTICS FOR ENTRY DF = 1,8				
Variable	Tolerance	Model R-Square	F Value	Pr > F
AGE2	0.011115	0.7803	(0.0097)	(0.9238)

No other variable met the 0.1000 significance level for entry into the model.

SUMMARY OF FORWARD SELECTION							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	HGT	1	0.6630	0.6630	4.2682	19.6749	0.0013
2	AGE	2	0.1170	0.7800	2.0097	4.7849	0.0565

16.5.4 Stepwise Regression Procedure

Stepwise regression is a modified version of forward regression that permits re-examination, at every step, of the variables incorporated in the model in previous steps. A variable that entered at an early stage may become superfluous at a later stage because of its relationship with other variables subsequently added to the model. To check this possibility, at each step we conduct a partial F test for each variable currently in the model, as though it were the most recent variable entered, irrespective of its actual entry point into the model. The variable with the smallest nonsignificant partial F statistic (if there is such a variable) is removed; the model is refitted with the remaining variables; the partial F 's are obtained and similarly examined; and so on. The whole process continues until no more variables can be entered or removed.

For our example, the first step, as in the forward-selection procedure, is to add the variable HGT to the model, since it has the highest significant correlation with WGT. Next, as before, we add AGE to the model, since it has a higher significant partial correlation with WGT than does $(AGE)^2$, controlling for HGT. Now, before testing to see whether $(AGE)^2$ should also be added to the model, we look at the partial F of HGT, given that AGE is already in the model, to see whether HGT should now be removed. Since $F(HGT | AGE) = 7.665$ (see model 4, Table 16.1) exceeds $F_{1,9,0.90} = 3.36$, we do not remove HGT from the model. Next, we check to see whether we should add $(AGE)^2$; of course, the answer is no, since we have dealt with this situation before.

The analysis-of-variance table that best summarizes the results obtained for our example is as follows:

Source	d.f.	SS	MS	F	R ²
Regression { HGT (AGE HGT)	1	588.92	588.92	19.67**	0.7800
	1	103.90	103.90	4.79 ($P < .10$)	
Residual	9	195.43	21.71		
Total	11	888.25			

The ANOVA table that considers all variables is

Source	d.f.	SS	MS	F	R ²
Regression { HGT (AGE HGT $(AGE)^2 HGT, AGE$	1	588.92	588.92	19.67**	0.7802
	1	103.90	103.90	4.79 ($P < .10$)	
	1	0.24	0.24	0.01	
Error	8	195.19	24.40		
Total	11	888.25			

16.5.5 Using Computer Programs

So far we have discussed decision-making algorithms in stepwise variable selection in terms of F_p and its P -value, R_p^2 , C_p , and MSE_p . In backward elimination, the F statistic

is often called an “ F -to-leave,” while in forward selection, the F statistic is often called an “ F -to-enter.” Stepwise computer programs require specifying comparison or critical values for these F 's or specifying their associated significance levels. These values must not be interpreted as when conducting just one hypothesis test, since the probability of finding at least one significant independent variable when there are actually none increases rapidly as the number k of candidate independent variables increases—an approximate upper bound on this overall significance level being $1 - (1 - \alpha)^k$, where α is the significance level of any one test.

To prevent this overall significance level from exceeding α in value, a conservative but easily implemented approach is to conduct any one test at level α/k (see Pope and Webster 1972, and Kupper, Stewart, and Williams 1976, for further discussion). In Section 16.7, we discuss other techniques for helping to ensure the reliability of our conclusions.

16.5.6 Chunkwise Methods

The methods just described for selecting variables one-at-a-time can be generalized in a very useful way. The basic idea is that any selection method in which a single variable is added or deleted can be generalized to apply to adding or deleting a group of variables. Consider, for example, using backward elimination to build a model for which the response variable is blood cholesterol level. Assume that three groups of predictors are available: demographic (gender, race, age, and their pairwise interactions), anthropometric (height, weight, and their interactions), and diet recall (amounts of five food types). The three groups of variables constitute *chunks*—sets of predictors that are logically related and equally important (within a chunk) as candidate predictors.

Several possible chunkwise testing methods are available. Choosing among them depends on two factors: the analyst's preference for backward, forward, or other selection strategies; and the extent to which the analyst can logically group variables (i.e., form chunks) and then order the groups in importance. In many applications, an *a priori* order exists among the chunks. For this example, the researcher may wish to consider diet variables *only* after controlling for important demographic and anthropometric variables. Imposing order in this fashion simplifies the analysis, typically increases reliability, and increases the chance of finding a scientifically plausible model. Hence, whenever possible, we recommend imposing an order on chunk selection.

We illustrate the use of chunkwise testing methods by describing a backward-elimination strategy. Other strategies may be preferred, depending on the situation. One approach to a backward-elimination method for chunkwise testing involves requiring all but one specified chunk of variables to stay in the model, the variables in that specified chunk being candidates for deletion.³ For our example, assume that the set of diet variables is the first chunk considered for deletion. (If an *a priori* order among chunks exists, that order determines which chunk is considered first for deletion. Otherwise, we choose first the chunk that makes the least important predictive contribution [e.g., because it has the smallest multiple partial F statistic].)

³ Stepwise regression computer packages permit this approach by letting the user specify variables to be forced into the model. The same feature can be used for subsequent chunkwise steps.

Thus, in our example, all of the demographic and anthropometric variables are forced to remain in the model. If, for example, the chunkwise multiple partial F test for the set of diet variables is not significant, the entire chunk can be deleted. If this test shows that the chunk is significant, at least one of the variables in the diet chunk should be retained.

Of course, the simplest chunkwise method adds or deletes all variables in a chunk together. A more sensitive approach is to manipulate single variables within a significant chunk while keeping the other chunks in place. If we assume that the diet chunk is important, we must decide which of the diet variables to retain as important predictors. A reasonable second step here is then to require all demographic variables and the important individual diet variables to be retained, while considering the (second) chunk of anthropometric variables for deletion. The final step in this three-chunk example requires that the individual variables selected from the first two chunks remain in the model, while variables in the third (demographic) chunk become candidates for deletion. Forward and stepwise single-variable selection methods can also be generalized for use in chunkwise testing.

Chunkwise methods for selecting variables can have substantial advantages over single-variable selection methods. First, chunkwise methods effectively incorporate into the analysis prior scientific knowledge and preferences about sets of variables. Second, the number of possible models to be evaluated is reduced. If a chunk test is not significant and the entire chunk of variables is deleted, then no tests about individual variables in that chunk are carried out. In many situations, such testing for group (or chunk) significance is more effective and reliable than evaluating variables one at a time.

16.6 Step 4: Conducting the Analysis: Prediction Goal

Having specified the maximum model, the criterion for selecting a model, and the strategy for selecting variables, we must conduct the analysis as planned. Obviously, this requires some type of computer program. The goodness of fit of the model chosen should certainly be examined. Also, the regression diagnostic methods of Chapter 14, such as residual analysis, are needed to demonstrate that the model chosen is reasonable for the data at hand. In the next section, we discuss methods for evaluating whether the model chosen has a good chance of fitting new samples of data from the same (or similar) populations.

16.7 Step 5: Evaluating Reliability with Split Samples: Prediction Goal

Having chosen a model that is best for a particular sample of data, we have no assurance that the model can reliably be applied to other samples. If the chosen model predicts well for subsequent samples from the population of interest, we say that the model is *reliable*. In this section, we discuss methods for evaluating the reliability of a model. Most

generally accepted methods for assessing model reliability involve some form of a *split-sample* approach.

Clearly, the most rigorous way to assess the reliability of a chosen model is to assess the fit of the chosen model to a new set of data randomly selected from the population under study. However, this approach is usually expensive and sometimes not feasible, leading to the use of split-sample methods (involving training and holdout groups).

For the cholesterol example in Section 16.5.6, the simplest split-sample analysis would proceed as follows. First, we randomly assign each observation to one or the other of two groups, the training group or the holdout group. This must be done before any analysis is conducted. Subjects may be grouped in strata based on one or more important categorical variables. For the cholesterol example, appropriate strata might be defined by various gender-race combinations. If such strata are important, we may use a stratum-specific split-sample scheme. With this method, we randomly assign each subject within a stratum to either the training group or the holdout group; such random assignment is done separately for each stratum. The goal of stratum-specific random assignment is to ensure that the two groups of observations (training and holdout) are equally representative of the parent population.

An alternative to stratified random splitting is a pair-matching assignment scheme. With pair-matching assignment, we find pairs of subjects that are as similar as possible and then randomly assign one member of the pair to the training sample and the other to the holdout sample. Unfortunately, the differences between the resulting training and holdout groups tend to be fewer than corresponding differences among subsequent randomly chosen samples. This tends to create an unrealistically optimistic opinion about model reliability, which we wish to avoid.

Either of two approaches can be recommended for assessing reliability once the training and holdout groups have been determined. The first approach is to apply the same variable selection algorithm to the training and holdout groups, and any differences in the sets of predictor variables chosen, and/or their estimated regression coefficients, for these two groups is taken as an indication of unreliability. This approach for assessing reliability is far too stringent when prediction is the goal. A second approach is based on the idea that a very good predictive model should predict about as well for the holdout group as it does for the training group, and we will describe and illustrate this second approach using the children's weight example.

With this second approach, we begin by conducting a model-building process using the data for the training group. Suppose that the fitted prediction equation obtained for the training group data is denoted as

$$\hat{Y}_1 = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

For the children's weight example, this equation takes the form

$$\widehat{\text{WGT}} = 6.553 + 0.722\text{HGT} + 2.050\text{AGE}$$

Next, we use the estimated prediction equation, which is based only on the training group data, to compute predicted values for this group. We denote this set of predicted values as $\{\hat{Y}_{i1}\}$,

$i = 1, 2, \dots, n_1$, with the subscript i indexing the subject and the subscript 1 denoting the training group. For this training sample, we let

$$R^2(1) = R_{Y_1|X_1, X_2, \dots, X_p}^2 = r_{Y_1, \hat{Y}_1}^2$$

denote the sample squared *multiple* correlation, which equals the sample squared *univariate* correlation between the observed and predicted response values.

Next, we use the prediction equation for the training group to compute predicted values for the holdout (or “validation”) sample. We denote this set of predicted values as

$$\{\hat{Y}_{i2}^*\}, i = n_1 + 1, n_1 + 2, \dots, n_1 + n_2$$

where n_2 is the number of observations in the holdout group and $n = n_1 + n_2$.

Finally, we compute the univariate correlation between these predicted values and the observed responses in the holdout sample (group 2)—namely,

$$R_*^2(2) = r_{Y_2, \hat{Y}_2}^2$$

The quantity $R_*^2(2)$ is called the *cross-validation correlation*, and the quantity

$$R^2(1) - R_*^2(2)$$

is called the *shrinkage on cross-validation*. The shrinkage statistic is almost always positive. How large must shrinkage be to cast doubt on model reliability? No firm rules can be given. Certainly, the fitted model is unreliable if shrinkage is 0.90 or more. In contrast, a shrinkage value less than 0.10 indicates an unreliable model. If the shrinkage is small enough to indicate reliability of the model, it seems reasonable to pool the data and calculate pooled estimates of the regression coefficients.

Depending on the situation, using only half of the data for the training sample analysis may be inadvisable. A useful rule of thumb is to increase the relative size of the training sample as the total sample size decreases and to decrease the relative size of the training sample as the total sample size increases. To illustrate this, consider two situations. In the first situation, data from over 3,000 subjects were available for regression analysis. The maximum model included fewer than 20 variables, consisting of a chunk of demographic control variables and a chunk of pollution exposure variables. The primary goal of the study was to test for the presence of a relationship between the response variable (a measure of pulmonary function) and the pollution variables, controlling for demographic effects. As a framework for choosing the best set of demographic control variables, a 10% stratified random sample ($n_1 \approx 300$) was used as a training sample.

The second situation involved a study of the relationship between measurements of amount of body fat by X-ray methods (CAT scans) and by traditional measures such as skinfold thickness. The primary goal was to provide an equation to predict the X-ray-measured body fat from demographic information (gender and age) and simple anthropometric measures (body weight, height, and three readings of skinfold thickness). The maximum model contained fewer than 20 predictors. Data from approximately 200 subjects were available. These data constituted nearly one year’s collection from the hospital X-ray laboratory and demanded substantial human effort and computer processing to extract. Consequently,

approximately 75% of the data were used as a training sample. These two examples illustrate the general concept that the splitting proportion should be tailored to the problem under study.

16.8 Example Analysis of Actual Data

We are now ready to apply the methods discussed in this chapter to a set of actual data. In Chapter 14 (on regression diagnostics), we introduced data on more than 200 children, reported by Lewis and Taylor (1967). The categories reported for all subjects include body weight (WGT), standing height (HGT), AGE, and SEX. Our goal here is to provide a reliable equation for predicting weight.

The first step is to choose a maximum model. The number of subjects is large enough that we can consider a fairly wide range of models. It is natural to want to include the *linear* effects of AGE, HGT, and SEX in the model. Quadratic and cubic terms for both AGE and HGT will also be included, since WGT is expected to increase in a nonlinear way as a function of AGE and HGT. Because the same model is unlikely to hold for both males and females, the interaction of the dichotomous variable SEX with each power of AGE and HGT will be included. Terms involving cross-products between AGE and HGT powers will not be included, since no biological fact suggests that such interactions are important. Hence, $k = 13$ variables will be included in the maximum model as predictors of WGT: AGE; AGE2 = $(AGE)^2$; AGE3 = $(AGE)^3$; HGT; HT2 = $(HGT)^2$; HT3 = $(HGT)^3$; SEX = 1 for a boy and 0 for a girl; SEX \times AGE; SEX \times AGE2; SEX \times AGE3; SEX \times HGT; SEX \times HT2; and SEX \times HT3. (In the analysis to follow, AGE and HGT denote centered variables.)

The second step is to choose a selection criterion. We choose to emphasize C_p , while also looking at R_p^2 . The former criterion is helpful in deciding the size of the best model, while the latter provides an easily interpretable measure of predictive ability.

The third step is to choose a strategy for selecting variables. We prefer backward-elimination methods, with as much structure imposed on the process as possible. As mentioned earlier, one procedure is to group the variables into chunks and then to order the chunks by degree of importance. However, we shall not consider a chunkwise strategy here.

After the (seemingly) best model is chosen, the fourth step is to evaluate the reliability of the model. Since the goal here is to produce a reliable prediction equation, this step is especially important. To implement a split-sample approach to assessing reliability, we first randomly assign each subject to either data set ONE (the training group) or data set TWO (the holdout group). Random splitting is sex-specific. The following table summarizes the observed split-sample subject distribution.

Sample	Female	Male	Total
ONE	55	63	118
TWO	56	63	119
Total	111	126	237

Data set ONE will be used to build the best model. Then data set TWO will be used to compute cross-validation correlation and shrinkage statistics.

16.8.1 Analysis of Data Set ONE

Table 16.2 provides descriptive statistics for data set ONE, and Table 16.3 provides the matrix of zero-order correlations. To avoid computational problems, the predictor variables, HGT and AGE, have been centered (see Table 16.2 and Chapter 14).

Table 16.4 summarizes the backward-elimination analysis based on data set ONE. For data set ONE, $\text{SSY} = 35,700.70$ and $\text{MSE} = 120.00$ for the full model. Table 16.4 specifies the order of variable elimination, the variables in the best model for each possible model size, and C_p and R_p^2 for each such model. Since the maximum model includes $k = 13$ variables, the top row corresponds to $p = 13$, for which $R_p^2 = 0.65042$. Since the maximum model always has $C_p = k + 1$, the C_p -value of 14 provides no information regarding the best model size. The second row corresponds to a $p = 12$ variable model. The variable HT3 (listed in the preceding row) has been deleted, leaving $C_p = 12$. Similarly, the bottom row tells us that the best single-variable model includes HGT ($C_p = 6.72$, $R_p^2 = 0.59422$), the best two-variable model includes HGT and AGE3 ($C_p = 0.52$, $R_p^2 = 0.62177$), and so on. In general, the variable identified in row $(14 - p)$ in the table and all variables listed in rows below it are included in the best p -variable model, while all variables listed in rows above row $(14 - p)$ are excluded.

TABLE 16.2 Descriptive statistics for data set ONE ($n = 118$)

Variable	Mean	S.D.	Minimum	Maximum
WGT (pounds)	100.05	17.47	63.5	150.0
AGE* (years)	0	1.50	-2.00	3.91
HGT† (inches)	0	3.7	-8.47	10.73

* Equals $(\text{AGE} - \bar{\text{AGE}}_1)$, where $\bar{\text{AGE}}_1 \doteq 13.59$

† Equals $(\text{HGT} - \bar{\text{HGT}}_1)$, where $\bar{\text{HGT}}_1 \doteq 61.27$

© Cengage Learning

TABLE 16.3 Correlations ($\times 100$) for data set ONE with centered age and height variables ($n = 118$)

WGT													
AGE	59												
AGE2	25	44											
AGE3	48	78	80										
HGT	77	67	16	43									
HT2	05	10	23	11	15								
HT3	63	46	14	32	79	35							
SEX	03	-01	-04	-03	18	18	24						
SEX × AGE	46	72	28	50	62	29	46	00					
SEX × AGE2	18	26	55	38	27	35	32	51	37				
SEX × AGE3	38	58	49	60	46	30	40	12	81	70			
SEX × HGT	57	55	16	34	81	41	68	10	77	27	57		
SEX × HT2	26	22	19	17	43	78	70	50	31	53	36	47	
SEX × HT3	48	38	17	27	64	65	88	16	54	32	46	78	75

© Cengage Learning

TABLE 16.4 Backward elimination based on data set ONE ($n = 118$)

p	C_p	R_p^2	Variable to Be Deleted
13	14.00	0.65042	HT3
12	12.00	0.65042	SEX × AGE
11	10.00	0.65042	AGE2
10	8.00	0.65040	SEX × AGE2
9	6.00	0.65040	SEX × AGE3
8	4.03	0.65032	HT2
7	2.08	0.65015	AGE
6	0.38	0.64914	SEX
5	-0.82	0.64644	SEX × HT3
4	1.15	0.63312	SEX × HT2
3	-0.12	0.63066	SEX × HGT
2	0.52	0.62177	AGE3
1	6.72	0.59422	HGT

© Cengage Learning

Table 16.4 does not suggest any obvious choice for a best model. The maximum R^2 is about 0.650, and the minimum is around 0.594, which is a small operating range, indicating that many different-size models predict about equally well. This is not unusual with moderate-to-strong predictor intercorrelations—say, above 0.50 in absolute value (see Table 16.3). Since $R^2 \approx 0.650$ and $R^2_{13} \approx 0.650$, it seems unreasonable to use more than seven predictors. However, on the basis of R_p^2 alone, it is difficult to decide whether a model containing fewer than seven predictors is appropriate.

In contrast, the C_p statistic suggests that only two variables are needed. (Recall that C_p should be compared with the value $p + 1$.) C_p is greater than $p + 1$ only when $p = 1$, which calls into question only the one-variable model in Table 16.4. Unfortunately, the $p = 2$ model with HGT and AGE3 as predictors is not appealing, since the linear and quadratic terms AGE and AGE2 are not included. As discussed in detail in Chapter 15, we strongly recommend including such lower-order terms in polynomial-type regression models.

An apparently nonlinear relationship may indicate the need to transform the response and/or predictor variables. Since quadratic and cubic terms are included as candidate predictors, $\log(\text{WGT})$ and $(\text{WGT})^{1/2}$ can be tried as alternative response variables. Such transformations can sometimes linearize a relationship (see Chapter 14). Based on all 13 predictors, $R^2 = 0.668$ for the log transformation and 0.659 for the square root transformation. Because neither transformation produces a substantial gain in R^2 , they will not be considered further.

In additional exploratory analyses of the data in data set ONE, models were fitted in two fixed orders: an interaction ordering (Table 16.5); and a power ordering (Table 16.6). We can interpret Tables 16.5 and 16.6 similarly to Table 16.4; each row gives p , the number of variables in the model, and the associated C_p - and R_p^2 -values. All variables on or below row $(14 - p)$ are included in the fitted model, and all those above row $(14 - p)$ are excluded.

TABLE 16.5 Interaction-ordered fitting based on data set ONE ($n = 118$)

p	C_p	R_p^2	Variable to Be Deleted
13	14.00	0.65042	SEX × AGE3
12	12.01	0.65040	SEX × HT3
11	10.39	0.64909	SEX × AGE2
10	8.39	0.64909	SEX × HT2
9	6.94	0.64727	AGE3
8	6.97	0.64045	HT3
7	6.37	0.63575	AGE2
6	7.35	0.62571	HT2
5	5.36	0.62567	SEX × AGE
4	3.60	0.62486	SEX × HGT
3	4.98	0.61352	SEX
2	5.86	0.60383	AGE
1	6.72	0.59422	HGT

© Cengage Learning

TABLE 16.6 Power-ordered fitting based on data set ONE ($n = 118$)

p	C_p	R_p^2	Variable to Be Deleted
13	14.00	0.65042	SEX × AGE3
12	12.01	0.65040	SEX × HT3
11	10.39	0.64909	SEX × AGE2
10	8.39	0.64909	SEX × HT2
9	6.94	0.64727	SEX × AGE
8	8.00	0.64698	SEX × HGT
7	4.77	0.64112	AGE3
6	5.05	0.63343	HT3
5	4.83	0.62747	AGE2
4	6.27	0.61589	HT2
3	4.98	0.61352	SEX
2	5.86	0.60383	AGE
1	6.72	0.59422	HGT

© Cengage Learning

In both tables, certain values of p are natural break points for model evaluation. For example, in Table 16.5, $p = 3$ corresponds to including just the linear terms of the predictors HGT, AGE, and SEX. The fact that $4.98 > (3 + 1)$ leads us to consider larger models. Including the

pairwise interactions SEX \times HGT and SEX \times AGE gives $p = 5$ and $C_5 = 5.36 < (5 + 1)$, which suggests that this five-variable model is reasonable. Notice that $R_p^2 = 0.626$ and that each subsequent variable addition increases R^2 very little. In fact, the best model might be of size $p = 4$, since $C_4 = 3.60$ and R^2 is not appreciably reduced. Taken together, these comments suggest choosing a four-variable model containing HGT, AGE, SEX, and SEX \times HGT, with $R_p^2 = 0.625$ and $C_4 = 3.60$.

Table 16.6 summarizes a similar analysis in which the powers of the continuous predictors are entered into the model in a logical ordering. Here, as in Table 16.5, $p = 3$ gives $C_3 = 4.98$, encouraging us to consider larger models. The model with all three original variables and the two quadratic terms gives $p = 5$, $C_p = 4.83$, and $R_p^2 = 0.62747$. Adding higher-order powers or any interactions does not improve R^2 appreciably, nor does it lead to C_p -values noticeably greater than $p + 1$. All models smaller than $p = 5$, on the other hand, do have C_p -values noticeably greater than $p + 1$. The results in Table 16.6 lead to choosing a five-variable model containing HGT, AGE, SEX, HT2, and AGE2, with $R_p^2 = 0.627$ and $C_5 = 4.83$.

In this example, then, two possible models have been suggested, both with roughly the same R^2 -value. Personal preference may be exercised within the constraints of parsimony and scientific knowledge. In this case, we prefer the five-variable model with HGT, AGE, SEX, HT2, and AGE2 as predictors. We choose this model as best because we expect growth to be nonlinear and because we prefer to avoid using interaction terms. The chosen model is thus of the general form

$$\widehat{\text{WGT}} = \hat{\beta}_0 + \hat{\beta}_1 \text{SEX} + \hat{\beta}_2 \text{HGT} + \hat{\beta}_3 \text{AGE} + \hat{\beta}_4 \text{HT2} + \hat{\beta}_5 \text{AGE2}$$

A predicted weight for the i th subject may be computed with the formula

$$\begin{aligned}\widehat{\text{WGT}}_i = & 100.9 - 3.153(\text{SEX}_i) + 3.53(\text{HGT}_i - 61.27) + 0.4199(\text{AGE}_i - 13.59) \\ & - 0.0731(\text{HGT}_i - 61.27)^2 + 0.8067(\text{AGE}_i - 13.59)^2\end{aligned}$$

Recall that SEX has a value of 1 for boys and 0 for girls. For males, the estimated intercept is $100.9 - 3.153(1) = 97.747$, while for females, it is $100.9 - 3.153(0) = 100.9$. For these particular data, then, the fitted model predicts that, on average, girls outweigh boys of the same height and age by about 3 pounds. After selecting a model, we must consider residual analysis and other regression diagnostic procedures (see Chapter 14). Some large residuals are present, but these do not justify a more complex model. Since they turn out not to be influential, they need not be deleted to improve estimation accuracy.

16.8.2 Analysis of Sample TWO

Since we used exploratory analysis of data set ONE to choose the best (most appealing) model, we will use analysis of the holdout data set TWO to evaluate the reliability of the chosen prediction equation. Table 16.7 provides descriptive statistics for data set TWO. These appear very similar to those of data set ONE, as one would hope. In particular, the AGE and HGT variables are centered using the sample ONE means. This continuity is necessary to maintain the definitions of these variables as used in the fitted model. Cross-validation analysis would look spuriously bad if this were not done.

TABLE 16.7 Descriptive statistics for data set TWO ($n = 119$)

Variable	Mean	S.D.	Minimum	Maximum
WGT (pounds)	102.6	21.22	50.5	171.5
AGE* (years)	0.20	1.46	-1.92	3.58
HGT [†] (inches)	0.18	4.16	-10.77	9.73

* Equals $(AGE - \bar{AGE})$, where $\bar{AGE}_1 \doteq 13.59$

[†] Equals $(HGT - \bar{HGT})$, where $\bar{HGT}_1 \doteq 61.27$

© Cengage Learning

In comparison to Table 16.4, Table 16.8 illustrates the instability of stepwise methods for selecting variables. Despite using the same backward-elimination strategy, we find that almost no variable appears in the same place as it did in the analysis for data set ONE.

Our recommended cross-validation analysis begins by using the regression equation estimated from data set ONE (with predictors HGT, AGE, SEX, HT2, and AGE2) to predict WGT values for data set TWO. The squared multiple correlation between these predicted values and the observed WGT values in data set TWO is $R^2(2) = 0.621$, and this is the cross-validation squared correlation. Since $R^2(1) = 0.627$ for sample ONE, shrinkage is $0.627 - 0.621 = 0.006$, which is quite small and indicates excellent reliability of prediction.

TABLE 16.8 Backward elimination based on data set TWO ($n = 119$)

p	C_p	R_p^2	Variable to Be Deleted
13	14.00	0.72328	AGE2
12	12.00	0.72328	SEX
11	10.02	0.72323	AGE3
10	8.31	0.72245	SEX × AGE2
9	7.22	0.72006	HT2
8	6.06	0.71781	SEX × HT2
7	4.15	0.71763	SEX × HGT
6	3.16	0.71496	HT3
5	1.52	0.71402	SEX × HT3
4	2.58	0.70595	SEX × AGE
3	7.99	0.68642	AGE
2	9.60	0.67690	SEX × AGE3
1	33.90	0.60762	HGT

© Cengage Learning

An important aspect of assessing the reliability of a model involves considering difference scores of the form

$$Y_i - \hat{Y}_i$$

or (in our example)

$$\text{WGT}_i - \widehat{\text{WGT}}_i$$

where only data set TWO subjects are used and where the data set ONE equation is used to compute the predicted values. These “unstandardized residuals” can be subjected to various residual analyses. The most helpful such analysis entails univariate descriptive statistics, a box-and-whisker plot, and a plot of the difference scores ($Y_i - \hat{Y}_i$) versus the predicted values (\hat{Y}_i) (such plots are described in Chapter 14). In our example, a few large positive residuals are present, but they are not sufficiently implausible or influential to require further investigation. Their presence does hint at why cubic terms and interactions keep trying to creep into the model. These residuals could be reduced in size but only by adding many more predictors.

For prediction purposes, if the model is finally deemed acceptable, we should pool all the data and report the coefficient estimates based on the combined data. If the model is not acceptable, we must review the model-building process, paying particular attention to variable selection and large differences between training and holdout groups of data.

16.9 Selecting the Most Valid Model

In this chapter, we have considered variable selection strategies for situations where the primary study goal is prediction. In contrast, validity-oriented strategies are directed toward providing valid estimates of one or more regression coefficients in a model. A valid estimate reflects the value of the parameter it is estimating in the target population being studied. Validity-oriented analysis strategies are often of interest in public health research studies, where quantifying the true strength of associative or causal relationships in the target population, while appropriately accounting for potentially complex social and/or biological confounders and interactions, is often the primary objective.

In a validity-based strategy for selecting variables to be included in a final model, both confounding and interaction must be considered (see Chapter 11 for details).⁴ The specific model selection strategy that we recommend is a variant of the backward-elimination process. The first step involves a careful selection of the variables to be included in the initial “full” model, followed by the fitting of this initial model and a determination of important interaction effects. This first step is followed by an evaluation of confounding, an evaluation that is contingent on the results of the interaction assessment. The final step is a determination as

⁴ In this consideration of validity, other biases are assumed not to affect model estimates. For a truly valid estimate, the sample of data under consideration should accurately reflect the population of interest (i.e., no selection bias), the data should be accurately measured (i.e., no information bias), and the assumptions underlying the model and the selected analysis methods must be reasonably well satisfied.

to which subsets of control variables yield the most precise estimated measures of effect (e.g., estimated regression coefficients) for the main variables of interest.

In the variable selection step, one generally considers not only which measured independent variables might be related to the dependent variable but also the relative roles that these predictors might play with respect to the associations of research interest. Here it is helpful to consider an alternative notation and formulation of the multiple linear regression model—namely, the *EVW* model. In this model, not all independent variables are thought of in the same way, and the letters *E*, *V*, and *W* are used (instead of *X*) to delineate these variables by their conceptualized underlying relationships with each other and with the outcome *Y*. We specifically define these variables as follows: the *E* variables are the main independent variables whose associations with *Y* in the target population are of primary interest; the *V* variables are the potential confounders of the associations between the *E* variables and *Y*; and the *W* variables are the potential effect-modifiers of the associations between the *E* variables and *Y*. Typically, there are only a few *E* variables under study. Here we provide an illustration of a multiple linear regression model with one *E* variable, p_1 confounder variables *V*, and p_2 effect-modifiers *W*:

$$Y = \alpha + \beta E + (\gamma_1 V_1 + \gamma_2 V_2 + \cdots + \gamma_{p_1} V_{p_1}) + (\delta_1 EW_1 + \delta_2 EW_2 + \cdots + \delta_{p_2} EW_{p_2}) + \text{Error}$$

which can be written, using summation notation, in the form

$$Y = \alpha + \beta E + \sum_{i=1}^{p_1} \gamma_i V_i + \sum_{j=1}^{p_2} \delta_j EW_j + \text{Error}$$

Note that different sets of Greek letters are assigned to the different types of variables: α for the intercept and β 's, γ 's, and δ 's for the *E*, *V*, and *EW* product terms, respectively. This model formulation also requires a more nuanced specification of product terms. Because the emphasis in this framework is on the *E-Y* relationship, only variables involved in product terms with the main predictor of interest (*E*) are considered to be *W* variables; all other product terms are considered to be potential confounders (of the *E-Y* relationship) and are represented as *V* variables. As described in Chapter 11, the model must be hierarchically well-formulated (HWF), so all *W* variables in *EW* product terms must also be included as lower-order *V* variables.

As an example, returning to the BRFSS body-mass index (BMI) and drinking frequency example, suppose a model with drinking frequency (*drink_days*) is to be fit, controlling for exercise (*exercise*) and tobacco use (*tobacco_now*) and including all possible product terms. For this model, *E* = *drink_days*, *V*₁ = *exercise*, *V*₂ = *tobacco_now*, and *V*₃ = *exercise* × *tobacco_now*, with *W*₁ = *exercise* and *W*₂ = *tobacco_now* appearing in product terms with the *E* variable *drink_days*:

$$Y = \alpha + \beta (\text{drink_days}) + \gamma_1 (\text{exercise}) + \gamma_2 (\text{tobacco_now}) + \gamma_3 (\text{exercise} \times \text{tobacco_now}) + \delta_1 (\text{drink days} \times \text{exercise}) + \delta_2 (\text{drink days} \times \text{tobacco_now}) + \text{Error}$$

After variable and model specification, the next step is to assess the statistical significance of *EW* interaction terms in the fitted model (i.e., to assess statistical evidence for

effect modification). Here we recommend a “global” test for the presence of any interaction—namely, the multiple partial F test of $H_0: \delta_1 = \delta_2 = \cdots = \delta_{p_2} = 0$. If this test is significant, thus providing statistical evidence that at least one of the δ ’s is not equal to 0, the next step is to perform backward elimination (see Section 16.5.2), considering only the EW terms for possible elimination and retaining all lower-order terms, to identify which of the δ ’s are nonzero. If the global test is not significant, suggesting no interaction, then one typically drops all EW terms from the model. At the end of this interaction assessment step, all statistically significant EW terms should be retained in the “final” model.⁵

After assessing interaction, confounding is subjectively assessed using the methods of Section 11.4. In a model with no significant EW terms, this assessment is done by looking for meaningful changes in $\hat{\beta}$ as a result of the exclusion of one or more potential confounders V from the model. The simplest method for organizing this process is an “all-subsets” approach, where $\hat{\beta}$ is estimated in separate models that consider every possible subset of V terms. This is analogous to the *all possible regressions procedure* for prediction-oriented model selection, with changes in $\hat{\beta}$ serving as the *selection criterion*. To again maintain an HWF model, only V variables not appearing in significant EW terms are considered for exclusion.

Using the results from this all-subsets approach, one can construct a table containing all the $\hat{\beta}$ values. Those models in which $\hat{\beta}$ does not meaningfully change by a predetermined criterion (e.g., 10%) from the fully specified *maximum* (i.e., gold-standard) *model*, obtained after interaction assessment, may be considered valid and thus are candidates for being chosen as the final model. For models that contain a large number of V variables, an all-subsets approach may be impractical, in which case forward-, backward-, and stepwise-selection methods can be used to identify sets of confounders that produce valid models.

The presence of significant EW terms in the model complicates confounding assessment, since β alone does not fully represent the change in Y that results from changes in E . A detailed discussion of evaluating confounding in this situation is described elsewhere (e.g., see Kleinbaum and Klein 2010).

After a candidate set of valid models is identified, how should one select the “final” model? In this regard, two issues need to be considered. The first issue is that it may be prudent to force certain control variables to be in the final model for so-called political reasons (e.g., historically, all published research has considered these control variables because they are well-established risk factors for the outcome under study). A second issue concerns the evaluation of the precision of the estimated association between E and Y . For models without EW terms, this evaluation corresponds to examining the width of the confidence interval for $\hat{\beta}$ and favoring models with narrower confidence intervals. While precision will generally be greater in multiple linear regression models that contain more variables, certain V terms may negligibly increase precision and thus may be deemed unnecessary for inclusion.

For further details and examples of this validity-oriented strategy, see the epidemiologic research-oriented texts of Kleinbaum, Kupper, and Morgenstern (1982, chaps. 21–24) and Kleinbaum and Klein (2010, chaps. 6–7).

⁵ The inclusion of more than one E variable requires more complicated decisions to be made during the assessment of interaction; a discussion of these decisions is beyond the scope of this text. See Chapter 8 in Kleinbaum and Klein (2010) for more information.

Problems

1. Add the variables $(HGT)^2$ and $(AGE \times HGT)$ to the data set for WGT, HGT, AGE, and $(AGE)^2$ given in Section 16.3. Then, using an available regression program and the accompanying computer output, determine the best regression model relating WGT to the five independent variables as follows.
 - a. Use the forward approach.
 - b. Use the backward approach.
 - c. Use an approach that first determines the best model, employing HGT and AGE as the only candidate independent variables, and then determines whether any second-order terms should be added to the model. (Use $\alpha = .10$.)
 - d. Compare and discuss the models obtained for each of the three approaches.

Edited SAS Output (PROC REG) for Problem 1

Regression Models for Dependent Variable: WGT

SSY = 888.25000, N = 12

Number in Model	R-Square	C(p)	MSE	Variables in Model
1	0.7535	-0.4453	21.89185	AGEHGT
1	0.6675	2.1916	29.53302	HGT2
1	0.6630	2.3295	29.93275	HGT
1	0.5926	4.4874	36.18571	AGE
1	0.5876	4.6413	36.63180	AGE2
2	0.7800	0.7440	21.71415	HGT AGE
2	0.7764	0.8535	22.06667	HGT AGE2
2	0.7755	0.8809	22.15495	HGT AGEHGT
2	0.7731	0.9539	22.38984	AGE2 AGEHGT
2	0.7724	0.9765	22.46250	AGE HGT2
2	0.7711	1.0152	22.58723	HGT2 AGEHGT
2	0.7677	1.1206	22.92643	AGE2 HGT2
2	0.7669	1.1459	23.00799	AGE AGEHGT
2	0.6681	4.1729	32.75423	HGT HGT2
2	0.5927	6.4844	40.19683	AGE AGE2
3	0.8028	2.0454	21.89768	HGT HGT2 AGEHGT
3	0.8002	2.1249	22.18565	HGT AGE2 HGT2
3	0.7959	2.2574	22.66559	HGT AGE HGT2
3	0.7803	2.7358	24.39869	HGT AGE AGE2
3	0.7802	2.7367	24.40178	HGT AGE AGEHGT

(continued)

3	0.7769	2.8376	24.76744	AGE AGE2 AGEHGT
3	0.7764	2.8534	24.82438	HGT AGE2 AGEHGT
3	0.7739	2.9311	25.10614	AGE2 HGT2 AGEHGT
3	0.7734	2.9464	25.16136	AGE AGE2 HGT2
3	0.7725	2.9742	25.26208	AGE HGT2 AGEHGT
4	0.8043	4.0001	24.83873	HGT AGE HGT2 AGEHGT
4	0.8031	4.0344	24.98033	HGT AGE2 HGT2 AGEHGT
4	0.8006	4.1127	25.30469	HGT AGE AGE2 HGT2
4	0.7803	4.7356	27.88308	HGT AGE AGE2 AGEHGT
4	0.7771	4.8338	28.28980	AGE AGE2 HGT2 AGEHGT
5	0.8043	6.0000	28.97780	HGT AGE AGE2 HGT2 AGEHGT

© Cengage Learning

2. Using the data given in Problem 2 in Chapter 5 (with SBP as the dependent variable) and the accompanying computer output, find the best regression model, using $\alpha = .05$ and the independent variables AGE, QUET, and SMK as follows.
- Use the forward approach.
 - Use the backward approach.
 - Use the all possible regressions approach.
 - Based on your results in parts (a) through (c), select a model for further analysis to determine which, if any, of the following interaction (i.e., product) terms should be added to the model: AQ = AGE \times QUET, AS = AGE \times SMK, QS = QUET \times SMK, and AQS = AGE \times QUET \times SMK.

Edited SAS Output (PROC REG) for Problem 2

Regression Models for Dependent Variable: SBP

SSY = 6425.97, N = 32

Number in Model	R-Square	C(p)	MSE	Variables in Model
1	0.6009	18.7539	85.47795	AGE
1	0.5506	24.6554	96.26743	QUET
1	0.0612	81.9934	201.09569	SMK
2	0.7298	5.6565	59.87188	AGE SMK
2	0.6412	16.0324	79.49574	AGE QUET
2	0.6412	16.0365	79.50353	SMK QUET

(continued)

3	0.7609	4.0075	54.86225	AGE SMK QUET
3	0.7405	6.4033	59.55518	AGE SMK AS
3	0.6776	13.7682	73.98197	AGE QUET AQ
3	0.6511	16.8744	80.06647	SMK QUET QS
4	0.7901	2.5924	49.95688	AGE SMK QUET AQ
4	0.7648	5.5544	55.97387	AGE SMK QUET AQS
4	0.7639	5.6625	56.19343	AGE SMK QUET AS
4	0.7616	5.9357	56.74834	AGE SMK QUET QS
5	0.7922	4.3428	51.35172	AGE SMK QUET AQ AS
5	0.7918	4.3875	51.44602	AGE SMK QUET AQ AQS
5	0.7908	4.5061	51.69621	AGE SMK QUET AQ QS
5	0.7750	6.3632	55.61377	AGE SMK QUET QS AQS
5	0.7650	7.5314	58.07815	AGE SMK QUET AS AQS
5	0.7650	7.5370	58.08994	AGE SMK QUET AS QS
6	0.7935	6.1929	53.07696	AGE SMK QUET AQ QS AQS
6	0.7927	6.2869	53.28300	AGE SMK QUET AS QS AQS
6	0.7925	6.3063	53.32556	AGE SMK QUET AQ AS QS
6	0.7923	6.3385	53.39624	AGE SMK QUET AQ AS AQS
7	0.7952	8.0000	54.84757	AGE SMK QUET AQ AS QS AQS

© Cengage Learning

3. For the same data set considered in Problem 2 (and employing the accompanying computer output), use the stepwise regression approach to find the best regression models of SBP on QUET, AGE, and QUET \times AGE for smokers and nonsmokers separately. (Use $\alpha = .05$.) Compare the results for each group here with those obtained for Problem 2(d).

Edited SAS Output (PROC REG) for Problem 3

----- SMK = 0 -----

Dependent Variable SBP
Stepwise Selection: Step 1

STATISTICS FOR ENTRY DF = 1,13				
Variable	Tolerance	Model R-Square	F Value	Pr > F
AGE	1.00000	0.8029	52.94	<.0001
QUET	1.00000	0.7307	35.27	<.0001
AQ	1.00000	0.8381	67.31	<.0001

(continued)

Variable AQ Entered: R-Square = 0.8381 and C(p) = 0.2892

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1953.17097	1953.17097	67.31	<.0001
Error	13	377.22903	29.01762		
Corrected Total	14	2330.40000			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	93.07315	5.98129	7026.21152	242.14	<.0001
AQ	0.24951	0.03041	1953.17097	67.31	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

STATISTICS FOR ENTRY DF = 1,12				
Variable	Tolerance	Model R-Square	F Value	Pr > F
AGE	0.070103	0.8406	0.19	0.6729
QUET	0.089970	0.8419	0.29	0.6003

All variables left in the model are significant at the 0.0500 level.

No other variable met the 0.0500 significance level for entry into the model.

SUMMARY OF STEPWISE SELECTION								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	AQ		1	0.8381	0.8381	0.2892	67.31	<.0001

----- SMK = 1 -----

Dependent Variable SBP

Stepwise Selection: Step 1

STATISTICS FOR ENTRY DF = 1,15				
Variable	Tolerance	Model R-Square	F Value	Pr > F
AGE	1.000000	0.6737	30.97	<.0001
QUET	1.000000	0.5640	19.40	0.0005
AQ	1.000000	0.6978	34.63	<.0001

(continued)

Variable AQ Entered: R-Square = 0.6978 and C(p) = 2.3328

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2583.44049	2583.44049	34.63	<.0001
Error	15	1119.03010	74.60201		
Corrected Total	16	3702.47059			
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	102.21983	8.02769	12096	162.14	<.0001
AQ	0.25167	0.04277	2583.44049	34.63	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

STATISTICS FOR ENTRY DF = 1,14				
Variable	Tolerance	Model R-Square	F Value	Pr > F
AGE	0.126133	0.7104	0.61	0.4475
QUET	0.078094	0.7311	1.74	0.2086

All variables left in the model are significant at the 0.0500 level.

No other variable met the 0.0500 significance level for entry into the model.

SUMMARY OF STEPWISE SELECTION								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	AQ		1	0.6978	0.6978	2.3328	34.63	<.0001

4. For the data given in Problem 4 in Chapter 8 (plus the accompanying computer output), find (using $\alpha = .10$) the best regression model relating homicide rate (Y) to population size (X_1), percentage of families with income less than \$5,000 (X_2), and unemployment rate (X_3).
 - a. Use the stepwise approach.
 - b. Use the backward approach.
 - c. Use the all possible regressions approach.

Edited SAS Output (PROC REG) for Problem 4

Regression Models for Dependent Variable: Y

SSY 1855.20200, N = 20

Number in Model	R-Square	C(p)	MSE	Variables in Model
1	0.7480	6.1969	25.97790	X3
1	0.7052	9.9594	30.38125	X2
1	0.0045	71.6694	102.60275	X1
2	0.8020	3.4376	21.60839	X2 X3
2	0.7672	6.5055	25.41000	X1 X3
2	0.7103	11.5104	31.61198	X1 X2
3	0.8183	4.0000	21.06607	X1 X2 X3

5. The data set listed in the following table contains information on AGE, SEX (1 = male, 2 = female), work problems index (WP), marital conflict index (MC), and depression index (DEP) for a sample of 39 new admissions to a psychiatric clinic at a large university hospital. For each sex *separately*, determine (using $\alpha = .10$) the best regression model relating DEP to MC and WP, controlling for AGE, using the

ID No.	AGE	SEX	WP	MC	DEP
1	45	2	90	70	69
2	35	1	90	75	75
3	32	2	70	32	35
4	32	2	80	30	73
5	39	2	85	55	86
6	25	2	85	6	161
7	22	1	75	20	202
8	30	2	70	63	91
9	49	2	75	4	113
10	47	1	84	12	68
11	48	1	64	11	109
12	49	2	85	7	92
13	45	2	80	8	80
14	41	2	80	15	82
15	45	2	82	6	156
16	59	2	72	5	198
17	42	2	70	17	170
18	35	1	70	29	188
19	31	2	70	80	82
20	45	1	70	126	37
21	28	1	85	30	194
22	37	1	90	9	294
23	29	1	80	14	94
24	29	1	70	24	126
25	31	1	80	21	192
26	29	1	60	11	232
27	29	1	70	10	184

(continued)

ID No.	AGE	SEX	WP	MC	DEP
28	23	2	80	10	238
29	44	2	78	19	112
30	28	1	70	22	141
31	32	2	70	21	108
32	36	2	74	77	87
33	22	2	78	67	33
34	46	2	70	25	73
35	21	1	70	14	168
36	34	1	80	17	218
37	27	2	80	18	175
38	31	2	80	42	126
39	19	2	75	36	135

following sequential procedure: (1) force AGE into the model first; (2) use the all possible regressions approach on the remaining two independent variables WP and MC; (3) determine whether the interaction term ($MC \times WP$) should be added to the model. Compare and discuss the results obtained for each sex.

6. For the data in Problem 15 in Chapter 5, use LN_BRNTL as the response variable and LN_PPML, LN_BLDL, AGE, and WEIGHT as predictors. (Use $\alpha = .05$.)
 - a. Indicate a plausible fixed order for testing predictors, based on the nature of the data. Briefly defend your choice.
 - b. Use a computer program to fit the full model. Provide a test of whether or not the population multiple correlation is 0.
 - c. Using the *fixed* order, choose a best model, adding variables in a forward fashion. Do not adhere to a particular α , but use your judgment.
 - d. Repeat part (c) but delete the variables in the fixed order (a backward method).
 - e. Using a computer program employing a stepwise procedure, use a backward algorithm to find the best model.
 - f. Repeat part (e), using a forward algorithm.
 - g. Compare and contrast your conclusions for parts (c), (d), (e), and (f). Indicate your preferred model.
 - h. Using the ideas presented in Chapter 14, indicate how you could evaluate the adequacy of the best model and the validity of the underlying assumptions.
 - i. Suggest a practical split-sample approach for this particular set of data. Include recommended sample sizes, any stratification variables, and a variable selection strategy.
7. Consider the data of Bethel et al. (1985), discussed in Problem 19 in Chapter 14. Delete the three female subjects, leaving 16 observations. Use FEV_1 as the response and AGE, WEIGHT, and HEIGHT as predictors.
 - a. Use the all possible regressions procedure to suggest a best model.
 - b. Consider a model with centered AGE, WEIGHT, HEIGHT, and their squares as predictors. Suggest a plausible forward chunkwise strategy for choosing a model, and implement it.
 - c. Use the all possible regressions procedure for the expanded model to choose a best model.

- d. Compare results from parts (a), (b), and (c). What model seems most plausible? How do the data limit your conclusions?
8. Use the data from Freund (1979), presented in Problem 22 in Chapter 14. Taking the model discussed there as the maximum model, repeat parts (a) through (h) of Problem 6. In part (h), note the possible role of collinearity.
9. A random sample of data was collected on residential sales in a large city. The accompanying table shows the selling price (Y , in \$1,000s), area (X_1 , in hundreds of square feet), number of bedrooms (X_2), total number of rooms (X_3), house age (X_4 , in years), and location ($Z = 0$ for in-town and inner suburbs, $Z = 1$ for outer suburbs).
- In parts (a) through (c), use variables X_1 , X_2 , X_3 , X_4 , and Z as the predictor variables. Use the accompanying computer output to answer parts (a)–(d).
- Use the all possible regressions procedure to suggest a best model.
 - Use the stepwise regression algorithm to suggest a best model.
 - Use the backward-elimination algorithm to suggest a best model.
 - Which of the models selected in parts (a), (b), and (c) seems to be the best model, and why?

House	Y	X_1	X_2	X_3	X_4	Z
1	84.0	13.8	3	7	10	0
2	93.0	19.0	2	7	22	1
3	83.1	10.0	2	7	15	1
4	85.2	15.0	3	7	12	1
5	85.2	12.0	3	7	8	1
6	85.2	15.0	3	7	12	1
7	85.2	12.0	3	7	8	1
8	63.3	9.1	3	6	2	1
9	84.3	12.5	3	7	11	1
10	84.3	12.5	3	7	11	1
11	77.4	12.0	3	7	5	0
12	92.4	17.9	3	7	18	0
13	92.4	17.9	3	7	18	0
14	61.5	9.5	2	5	8	0
15	88.5	16.0	3	7	11	0
16	88.5	16.0	3	7	11	0
17	40.6	8.0	2	5	5	0
18	81.6	11.8	3	7	8	1
19	86.7	16.0	3	7	9	0
20	89.7	16.8	2	7	12	0
21	86.7	16.0	3	7	9	0
22	89.7	16.8	2	7	12	0
23	75.9	9.5	3	6	6	1
24	78.9	10.0	3	6	11	0
25	87.9	16.5	3	7	15	0
26	91.0	15.1	3	7	8	1
27	92.0	17.9	3	8	13	1
28	87.9	16.5	3	7	15	0
29	90.9	15.0	3	7	8	1
30	91.9	17.8	3	8	13	1

Edited SAS Output (PROC REG) for Problem 9

Regression Models for Dependent Variable: Y

N = 30

Number in Model	R-Square	C(p)	MSE	Variables in Model
1	0.7417	11.5912	32.78751	X3
1	0.6392	26.5059	45.79629	X1
1	0.3728	65.2771	79.61294	X4
1	0.1103	103.4723	112.92725	X2
1	0.0145	117.4173	125.09024	Z
2	0.8069	4.0972	25.41440	X1 X3
2	0.8063	4.1814	25.49052	X3 X4
2	0.7535	11.8678	32.44298	X3 Z
2	0.7417	13.5851	33.99634	X2 X3
2	0.7055	18.8517	38.76009	X1 Z
2	0.6936	20.5949	40.33682	X1 X2
2	0.6445	27.7395	46.79920	X1 X4
2	0.5723	38.2360	56.29347	X2 X4
2	0.4051	62.5683	78.30242	X4 Z
2	0.1147	104.8398	116.53767	X2 Z
3	0.8224	3.8492	24.28032	X1 X3 X4
3	0.8190	4.3407	24.74191	X2 X3 X4
3	0.8104	5.5917	25.91700	X1 X2 X3
3	0.8086	5.8488	26.15849	X3 X4 Z
3	0.8080	5.9389	26.24314	X1 X3 Z
3	0.7535	13.8675	33.69053	X2 X3 Z
3	0.7406	15.7521	35.46078	X1 X2 Z
3	0.7231	18.3003	37.85424	X1 X2 X4
3	0.7098	20.2356	39.67213	X1 X4 Z
3	0.5851	38.3797	56.71498	X2 X4 Z
4	0.8346	4.0700	23.51341	X1 X2 X3 X4
4	0.8227	5.8014	25.20478	X1 X3 X4 Z
4	0.8209	6.0632	25.46058	X2 X3 X4 Z
4	0.8118	7.3871	26.75380	X1 X2 X3 Z
4	0.7627	14.5337	33.73516	X1 X2 X4 Z
5	0.8351	6.0000	24.42193	X1 X2 X3 X4 Z

(continued)

Dependent Variable Y

Stepwise Selection: Step 1

Variable X3 Entered: R-Square = 0.7417 and C(p) = 11.5912

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2635.95947	2635.95947	80.40	<.0001
Error	28	918.05019	32.78751		
Corrected Total	29	3554.00967			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-17.08438	11.26624	75.39618	2.30	0.1406
X3	14.71918	1.64160	2635.95947	80.40	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable X1 Entered: R-Square = 0.8069 and C(p) = 4.0972

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2867.82088	1433.91044	56.42	<.0001
Error	27	686.18878	25.41440		
Corrected Total	29	3554.00967			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-4.21356	10.79550	3.87161	0.15	0.6994
X1	1.30268	0.43128	231.86141	9.12	0.0055
X3	10.14196	2.09411	596.10738	23.46	<.0001

Bounds on condition number: 2.0994, 8.3975

Stepwise Selection: Step 3

Variable X4 Entered: R-Square = 0.8224 and C(p) = 3.8492

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2922.72134	974.24045	40.12	<.0001
Error	26	631.28833	24.28032		
Corrected Total	29	3554.00967			

(continued)

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-4.92269	10.56242	5.27391	0.22	0.6451
X1	0.81476	0.53197	56.95559	2.35	0.1377
X3	10.52638	2.06276	632.29012	26.04	<.0001
X4	0.45797	0.30456	54.90046	2.26	0.1447

Bounds on condition number: 3.3432, 22.38

Stepwise Selection: Step 4

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

SUMMARY OF STEPWISE SELECTION								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	X3		1	0.7417	0.7417	11.5912	80.40	<.0001
2	X1		2	0.0652	0.8069	4.0972	9.12	0.0055
3	X4		3	0.0154	0.8224	3.8492	2.26	0.1447

10. In Problem 9, the first-order interactions between Z and the predictor variables X_1 , X_3 , and X_4 were not included in the model selection process. Investigate whether this was appropriate, as follows (using the accompanying computer output).
 - a. Conduct a test to investigate the importance of the interactions, given that X_1 , X_2 , X_3 , X_4 , and Z are in the model.
 - b. In view of your answer in part (a) and in light of the model selected in Problem 9, was it appropriate to exclude the interactions?
 - c. Why is it necessary to ignore the interaction term $X_2 \times Z$?

Edited SAS Output (PROC REG) for Problem 10

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	3100.73976	387.59247	17.96	<.0001
Error	21	453.26991	21.58428		
Corrected Total	29	3554.00967			

Root MSE	4.64589	R-Square	0.8725
Dependent Mean	83.49667	Adj R-Sq	0.8239
Coeff Var	5.56416		

(continued)

PARAMETER ESTIMATES						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	-24.70798	16.10563	-1.53	0.1399	209151
X1	1	0.27720	1.30212	0.21	0.8335	2271.71350
X2	1	0.28357	3.48718	0.08	0.9360	193.20196
X3	1	13.60786	4.98736	2.73	0.0126	415.25208
X4	1	1.00655	0.57143	1.76	0.0927	86.00677
Z	1	53.46722	26.89804	1.99	0.0600	1.70905
X1Z	1	0.70683	1.55573	0.45	0.6542	89.28986
X3Z	1	-8.12422	5.76510	-1.41	0.1734	29.49488
X4Z	1	-0.67663	0.83801	-0.81	0.4285	14.07165

11. In 1990, *Business Week* magazine compiled financial data on the 1,000 companies that had the biggest impact on the U.S. economy.⁶ Data from a sample of the top 500 companies in *Business Week's* report were presented in Problem 13

Edited SAS Output (PROC REG) for Problem 11

All Possible Regressions Output
N = 20

Number in Model	R-Square	C(p)	MSE	Variables in Model
1	0.3870	10.0513	1.63753	X3
1	0.0380	24.8783	2.56952	X1
1	0.0367	24.9348	2.57307	X2
2	0.5953	3.1971	1.14456	X1 X3
2	0.5529	5.0002	1.26456	X2 X3
2	0.0381	26.8769	2.72057	X1 X2
3	0.6235	4.0000	1.13144	X1 X2 X3

Stepwise Regression Output
Dependent Variable Y

(continued)

⁶ "The Business Week 1000" (1990).

Stepwise Selection: Step 1

Variable X3 Entered: R-Square = 0.3870 and C(p) = 10.0513

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	18.60477	18.60477	11.36	0.0034
Error	18	29.47553	1.63753		
Corrected Total	19	48.08030			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	5.49230	0.93681	56.28470	34.37	<.0001
X3	-0.19274	0.05718	18.60477	11.36	0.0034

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable X1 Entered: R-Square = 0.5953 and C(p) = 3.1971

ANALYSIS OF VARIANCE					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	28.62285	14.31142	12.50	0.0005
Error	17	19.45745	1.14456		
Corrected Total	19	48.08030			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	7.21135	0.97521	62.58539	54.68	<.0001
X1	-0.00513	0.00173	10.01808	8.75	0.0088
X3	-0.24883	0.05143	26.79390	23.41	0.0002

Bounds on condition number: 1.1572, 4.6289

Stepwise Selection: Step 3

All variables left in the model are significant at the 0.0500 level.

No other variable met the 0.0500 significance level for entry into the model.

SUMMARY OF STEPWISE SELECTION								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	X3		1	0.3870	0.3870	10.0513	11.36	0.0034
2	X1		2	0.2084	0.5953	3.1971	8.75	0.0088

in Chapter 8. In addition to the company name, the following variables were shown:

1990 Rank (X_1):	Based on company's market value (share price on March 16, 1990, multiplied by available common shares outstanding).
1989 Rank (X_2):	Rank in 1989 compilation.
P-E Ratio (X_3):	Price-to-earning ratio based on 1989 earnings and March 16, 1990, share price.
Yield (Y):	Annual dividend rate as a percentage of March 16, 1990, share price.

In parts (a) and (b), use variables X_1 , X_2 , and X_3 as the predictor variables.

- a. Use the all possible regressions procedure to suggest a best model.
- b. Use the stepwise regression algorithm to suggest a best model.
- c. Which model seems to be the better one? Why?

12. Radial keratotomy is a type of refractive surgery in which radial incisions are made in a myopic (nearsighted) patient's cornea to reduce the person's myopia. Theoretically, the incisions allow the curvature of the cornea to become less steep, thereby reducing the patient's refractive error. This and other vision-correction surgery techniques grew in popularity in the 1980s and 1990s, both among the public and among ophthalmologists.

The Prospective Evaluation of Radial Keratotomy (PERK) clinical trial was begun in 1983 to evaluate the effects of radial keratotomy. As part of the study, Lynn et al. (1987) examined the variables associated with the sample patients' five-year postsurgical change in refractive error (Y , measured in diopters, D). Several independent variables were under consideration:

Baseline refractive error (X_1 , diopters)

Patient age (X_2 , in years)

Patient's sex (X_3)

Baseline average central keratometric power (X_4 , a measure of corneal curvature, in diopters)

Depth of incision scars (X_5 , in mm)

Baseline horizontal corneal diameter (X_6 , in mm)

Baseline intraocular pressure X_7 , in mm Hg

Baseline central corneal thickness X_8 , in mm)

Diameter of clear zone (X_9 , in mm) (The clear zone is the circular central portion of the cornea that is left uncut during the surgery; the surgical incisions are made radially from the periphery of the cornea to the edge of the clear zone. Smaller clear zones are used for more myopic patients, the thinking being that "more" surgery, in the form of longer incisions, is probably needed for such patients.)

Some of the PERK study results from the all possible subsets analysis that was performed are shown in the following table.

Variable Added to Model	Model R^2 as Variables Are Added
Diameter of clear zone (X_9)	0.28
Patient age (X_2)	0.40
Depth of incision scars (X_3)	0.44
Baseline refractive error (X_1)	0.45
Baseline horizontal corneal diameter (X_6)	0.47
Baseline avg. central keratometric power (X_4)	0.48
Baseline intraocular pressure (X_7)	0.49
Patient's sex (X_3)	0.49
Baseline central corneal thickness (X_8)	0.49

- a. Using this table of R^2 values, perform an all possible regressions analysis to suggest a best model.
- b. The PERK study researchers concluded: "The regression analysis of the factors affecting the outcome of radial keratotomy [i.e., the change in refractive error] showed that the diameter of the clear zone, patient age, and the average depth of the incision scars were the most important factors." Do you agree with this assessment? Explain.

References

- Bethel, R. A.; Sheppard, D.; Geffroy, B.; Tam, E.; Nadel, J. A.; and Boushey, J. A. 1985. "Effect of 0.25 ppm Sulfur Dioxide on Airway Resistance in Freely Breathing, Heavily Exercising, Asthmatic Subjects." *American Review of Respiratory Diseases* 131: 659–61.
- "The Business Week 1000, America's Most Valuable Companies." 1990. *Business Week*, special issue, April 13.
- Freund, R. J. 1979. "Multicollinearity etc., Some 'New' Examples." *Proceedings of the Statistical Computing Section*, American Statistical Association, pp. 111–12.
- Hocking, R. R. 1976. "The Analysis and Selection of Variables in Linear Regression." *Biometrics* 32: 1–49.
- Kleinbaum, D. G., and Klein, M. 2010. *Logistic Regression: A Self-Learning Text, Third Edition*. New York: Springer.
- Kleinbaum, D. G.; Kupper, L. L.; and Morgenstern, H. 1982. *Epidemiologic Research: Principles and Quantitative Methods*. Belmont, Calif.: Lifetime Learning Publications.
- Kupper, L. L.; Stewart, J. R.; and Williams, K. A. 1976. "A Note on Controlling Significance Levels in Stepwise Regression." *American Journal of Epidemiology* 103(1): 13–15.
- Lewis, T., and Taylor, L. R. 1967. *Introduction to Experimental Ecology*. New York: Academic Press.
- Lynn, M. J.; Waring, G. O. III; Sperduto, R. D. 1987. "Factors Affecting Outcome and Predict Ability of Radial Keratotomy in the PERK Study." *Archives of Ophthalmology* 105: 42–51.
- Marquardt, D. W., and Snee, R. D. 1975. "Ridge Regression in Practice." *The American Statistician* 29(1): 3–20.
- Pope, P. T., and Webster, J. T. 1972. "The Use of an F-statistic in Stepwise Regression Procedures." *Technometrics* 14(2): 327–40.

17

One-way Analysis of Variance

17.1 Preview

This chapter is the first of four that focus on *analysis of variance* (ANOVA). Earlier we described ANOVA as a technique for assessing how several *nominal* independent variables affect a *continuous* dependent variable. An example given in Table 2.2 (the Ponape study) involved describing the effects on blood pressure of two cultural incongruity indices, each dichotomized into “high” and “low” categories (see Example 1.3). In this case, the dependent variable (blood pressure) was continuous, and the two independent variables (the cultural incongruity indices) were both nominal.

The fact that ANOVA is generally restricted to use with nominal independent variables suggests an interesting interpretation of the purpose of the technique. Loosely speaking, *ANOVA is usually employed in comparisons involving several population means*. In fact, in the simplest case, involving a comparison of two population means, the ANOVA comparison procedure is equivalent to the usual two-sample *t* test, which requires the assumption of equal population variances.

The population means to be compared can generally be easily specified by cross-classifying the nominal independent variables under consideration to form different combinations of categories.¹ In the example dealing with the Ponape study, we need only cross-classify the HI and LO categories of incongruity index 1 with the HI and LO categories of index 2. This yields the four population means corresponding to the four combinations HI–HI, HI–LO, LO–HI, and LO–LO, as indicated in the following configuration:

¹ Such specification is not possible if the categories of any nominal variable are viewed as being only a sample from a much larger population of categories of interest. We consider such situations later when discussing *random-effects models*.

		INDEX 2	
		HI	LO
INDEX 1	HI	μ_1	μ_2
	LO	μ_3	μ_4

Assessing whether the two indices have some effect on the dependent variable “blood pressure” is equivalent to determining what kind of differences, if any, exist among the four population means.

17.1.1 Why the Name ANOVA?

If the ANOVA technique usually involves comparing means, it seems somewhat inappropriate to call it analysis of *variance*. Why not instead use the acronym ANOME, where *ME* stands for *means*? Actually, the designation *ANOVA* is quite justifiable: although typically means are compared, the comparisons are made using estimates of variance. As with regression analysis, the ANOVA test statistics are *F* statistics and are actually ratios of estimates of variance. In fact, it is even possible, and in some cases appropriate, to specify the null hypotheses of interest in terms of population variances.

17.1.2 ANOVA versus Regression

Another general distinction relates to the difference between an “ANOVA problem” and a “regression problem.” For ANOVA, *all* independent variables must be treated as nominal; for regression analysis, any mixture of measurement scales (nominal, ordinal, or interval) is permitted for the independent variables. In fact, ANOVA is often viewed as a special case of regression analysis, and almost any ANOVA model can be represented by a regression model whose parameters can be estimated and inferred about in the usual manner. The same may be said for certain other multivariable techniques, such as analysis of covariance. Hence, we may view the various names given to these techniques as indicators of different (linear) models having the same general form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + E$$

yet involving different types of variables and perhaps different assumptions about these variables. The choice of method can thus be regarded as equivalent to the choice of an appropriate linear model.

17.1.3 Factors and Levels

Some additional terms must be introduced at this point. In Chapter 12, in connection with using dummy variables in regression, we saw that a nominal variable with *k* categories can generally be incorporated into a regression model with an intercept if we define $(k - 1)$ dummy variables. These $(k - 1)$ variables collectively describe the *basic* nominal variable under consideration. To refer to a basic variable without having to identify the specific dummy variables used to define it in the regression model, we can follow the approach of calling the basic nominal variable a *factor*. The different categories of the factor are often referred to as its *levels*.

For example, if we wanted to compare the effects of several drugs on some human health response, we would consider the nominal variable “drugs” as a single factor and the specific drug categories as the levels. If we were comparing k drugs, we would incorporate them into a regression model by defining $(k - 1)$ dummy variables. If, in addition to comparing the drugs, we wanted to consider whether males and females responded differently, we would consider the nominal variable “sex” as a second factor and the specific categories (male and female) as the levels of this dichotomous factor.

17.1.4 Fixed versus Random Factors

A *random factor* is a factor whose levels may be regarded as a sample from some large population of levels.² A *fixed factor* is a factor whose levels are the only ones of interest. The distinction is important in any ANOVA, since different tests of significance are required for different configurations of random and fixed factors. We will see this more specifically when considering two-way ANOVA. For now, we can simply look at some examples of random and fixed factors (summarized in Table 17.1):

1. “Subjects” is usually considered a random factor, since we ordinarily want to make inferences about a large population of potential subjects on the basis of the subjects sampled and inference about specific subjects is not of interest.
2. “Observers” is a random factor we often consider when examining the effect of different observers on the response variable of interest.
3. “Days,” “weeks,” and so on are usually considered random factors in investigations of the effect of time on a response variable observed during different time periods. We normally use many levels for such temporal factors to represent a large number of time periods.
4. “Sex” is always a fixed factor, since its typically two levels include all possible levels of interest.
5. “Locations” (e.g., cities, plants, or states) may be fixed or random, depending on whether a set of specific sites or a larger geographical universe is to be considered.

TABLE 17.1 Examples of random and fixed factors

Random	Fixed	Random or Fixed
Subjects	Sex	Locations
Observers	Age	Treatments
Days	Marital status	Drugs
Weeks	Education	Exposures

© Cengage Learning

² In practice, the experimental levels of a random factor need not be selected at random as long as they are reasonably representative of the larger population of levels of interest.

6. “Age” is usually treated as a fixed factor, regardless of how the different age groups are defined.
7. “Treatments,” “drugs,” “exposures,” and so on are usually considered fixed factors, but they may be considered random if their levels represent a much larger group of possible levels. For example, consider a research project designed to investigate whether antibiotic drug usage is associated with elevated numbers of antibiotic-resistant bacterial infections among hospital in-patients. In this situation, one might consider the particular types of antibiotic drugs administered to these in-patients as representing just a small subset of a large number of possible types of antibiotic drugs that could have been administered to these inpatients. Then, it would be reasonable to use a random effect to reflect the variation in risk of antibiotic-resistant bacterial infection as a function of variation in the type of antibiotic drug being administered.
8. “Marital status” is treated as a fixed factor.
9. “Education” is treated as a fixed factor.

17.2 One-way ANOVA: The Problem, Assumptions, and Data Configuration

One-way ANOVA deals with the effect of a single nominal factor on a single continuous response variable. When that one factor is a fixed factor, one-way ANOVA (often referred to as *fixed-effects one-way ANOVA*) involves a comparison of several (two or more) population means.³ The different populations effectively correspond to the different subgroups defined by the categories of the nominal factor under study.

17.2.1 The Problem

The main analysis problem in fixed-effects one-way ANOVA is to determine whether the population means are all equal or not. Thus, given k means (denoted as $\mu_1, \mu_2, \dots, \mu_k$), the basic null hypothesis of interest is

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k \quad (17.1)$$

The alternative hypothesis is given by

H_A : “The k population means are not all equal”

If the null hypothesis (17.1) is rejected, the next problem is to find out where the differences are. For example, if $k = 3$ and $H_0: \mu_1 = \mu_2 = \mu_3$ is rejected, we might wish to determine whether the main differences are between μ_1 and μ_2 , between μ_1 and μ_3 , between μ_1 and the average of the other two means, and so on. Such questions fall under the general statistical subject of multiple-comparison procedures, discussed in Section 17.7.

³ In this section, we focus on situations involving fixed factors. Random factors are discussed in Section 17.6.

17.2.2 The Assumptions

Four assumptions must be made for fixed-effects one-way ANOVA:

1. Independent random samples (individuals, animals, etc.) have been selected from each of k populations or groups.
2. A value of a specified dependent variable has been recorded for each experimental unit (individual, animal, etc.) sampled.
3. The dependent variable is normally distributed in each population.
4. The variance of the dependent variable is the same in each population (this common variance is denoted as σ^2).

Although these assumptions provide the theoretical justification for applying this method, it is sometimes necessary to use fixed-effects one-way ANOVA to compare several means even when the necessary assumptions are not clearly satisfied. Indeed, these assumptions rarely hold exactly. It is, therefore, important to consider the consequences of applying fixed-effects one-way ANOVA when the assumptions are in question.

In general, fixed-effects one-way ANOVA can be applied as long as none of the assumptions is badly violated. This is true for more complex ANOVA situations, as well as for fixed-effects one-way ANOVA. The term generally used to denote this property of broad applicability is *robustness*: a procedure is robust if moderate departures from the basic assumptions do not adversely affect its performance in any meaningful way.

We must nevertheless avoid asserting robustness as an automatic justification for carelessly applying the ANOVA method. Certain facts should be kept in mind when considering the use of ANOVA in a given situation. It is true that the normality assumption does not have to be exactly satisfied as long as we are dealing with relatively large samples (e.g., 20 or more observations from each population), although the consequences of large deviations from normality are somewhat more severe for random factors than for fixed factors. Similarly, the assumption of variance homogeneity can be mildly violated without serious risk, provided that the numbers of observations selected from each population are more or less the same (again, the consequences are more severe for random factors).

On the other hand, an inappropriate assumption of independence of the observations can lead to serious errors in inference for both fixed and random cases. In general, great care should be taken to ensure that the observations are independent. This concern arises primarily in studies where repeated observations are recorded on the same experimental subjects: the level of response of a subject on one occasion commonly has a decided effect on subsequent responses.

What should we do when one or more of these assumptions are in serious question? One option is to transform the data (e.g., by means of a log, square root, or other transformation) so that they more closely satisfy the assumptions (see Section 14.4). Another strategy is to select a more appropriate method of analysis (e.g., nonparametric ANOVA methods or longitudinal/correlated data analysis procedures).⁴ A detailed discussion concerning the analysis of longitudinal data is provided in Chapters 25 and 26.

⁴ Descriptions of nonparametric methods that can be used when these assumptions are clearly and strongly violated can be found in Siegel (1956), Lehmann (1975), and Hollander and Wolfe (1973). Recent textbooks on longitudinal/correlated data analysis methods include Diggle et al. (2013) and Fitzmaurice et al. (2008); also see Kleinbaum and Klein (chaps. 13–16, 2010).

TABLE 17.2 General data configuration for one-way ANOVA

Population	Sample Size	Observations	Total	Sample Mean
1	n_1	$Y_{11}, Y_{12}, Y_{13}, \dots, Y_{1n_1}$	$Y_{1\cdot} = T_1$	$\bar{Y}_{1\cdot} = T_1/n_1$
2	n_2	$Y_{21}, Y_{22}, Y_{23}, \dots, Y_{2n_2}$	$Y_{2\cdot} = T_2$	$\bar{Y}_{2\cdot} = T_2/n_2$
3	n_3	$Y_{31}, Y_{32}, Y_{33}, \dots, Y_{3n_3}$	$Y_{3\cdot} = T_3$	$\bar{Y}_{3\cdot} = T_3/n_3$
\vdots	\vdots	\vdots	\vdots	\vdots
k	n_k	$Y_{k1}, Y_{k2}, Y_{k3}, \dots, Y_{kn_k}$	$Y_{k\cdot} = T_k$	$\bar{Y}_{k\cdot} = T_k/n_k$
$n = \sum_{i=1}^k n_i$			$Y_{\cdot\cdot} = G$	$\bar{Y}_{\cdot\cdot} = \bar{Y} = G/n$

© Cengage Learning

17.2.3 Data Presentation for One-way ANOVA

Computations necessary for one-way ANOVA are simple enough that they can be performed with an ordinary calculator when the data are conveniently arranged. Table 17.2 illustrates a useful way of presenting the data for the general one-way situation. Each row contains the set of observations, and sample total and sample mean, pertaining to a random sample from a particular population. Clearly, the number of observations selected from each population does *not* have to be the same; that is, there are n_i observations from the i th population, and n_i need not equal n_j if $i \neq j$. Double-subscript notation (Y_{ij}) is used to distinguish one observation from another. The first subscript for a given observation denotes the population number, and the second distinguishes that observation from the others in the sample from that particular population. Thus, Y_{23} denotes the third observation from the second population, Y_{62} denotes the second observation from the sixth population, and Y_{kn_k} denotes the last observation from the k th population. The totals for each sample (from each population) are denoted alternatively by T_i or $Y_{i\cdot}$ for the i th sample, where the \cdot denotes that we are summing over all values of j (i.e., we are adding together all observations making up the given sample). The grand total over all samples is denoted as $Y_{\cdot\cdot} = G$. The sample means are alternatively denoted by $\bar{Y}_{i\cdot}$ or T_i/n_i for the i th sample; these statistics are particularly important because they represent the estimates of the population means of interest. Finally, the grand mean over all samples is $\bar{Y}_{\cdot\cdot} = G/n$.

■ **Example 17.1** In a study by Daly (1973) of the effects of neighborhood characteristics on health, a stratified random sample of 100 households was selected—25 from each of four turnkey neighborhoods included in the study. The data presentation of Cornell Medical Index (CMI) scores for female heads of household is given in Table 17.3. Such scores are measures (derived from questionnaires) of the overall (self-perceived) health of individuals; the *higher* the score, the *poorer* the health. Each of the turnkey neighborhoods differed in total number of households and in percentage of blacks in the surrounding neighborhoods. The racial composition of the turnkey neighborhoods

TABLE 17.3 Cornell Medical Index scores for a sample of women from different households in four turnkey housing neighborhoods

Neighborhood	No. of Households	% Blacks in Surrounding Neighborhoods	Sample Size (n_i)	Observations (Y_{ij})	Total (T_i)	Sample Mean (\bar{Y}_i)
Cherryview	98	17	25	49, 12, 28, 24, 16, 28, 21, 48, 30, 18, 10, 10, 15, 7, 6, 11, 13, 17, 43, 18, 6, 10, 9, 12, 12	$T_1 = 473$	$\bar{Y}_{1\cdot} = 18.92$
Morningside	211	100	25	5, 1, 44, 11, 4, 3, 14, 2, 13, 68, 34, 40, 36, 40, 22, 25, 14, 23, 26, 11, 20, 4, 16, 25, 17	$T_2 = 518$	$\bar{Y}_{2\cdot} = 20.72$
Northhills	212	36	25	20, 31, 19, 9, 7, 16, 11, 17, 9, 14, 10, 5, 15, 19, 29, 23, 70, 25, 6, 62, 2, 14, 26, 7, 55	$T_3 = 521$	$\bar{Y}_{3\cdot} = 20.84$
Easton	40	65	25	13, 10, 20, 20, 22, 14, 10, 8, 21, 35, 17, 23, 17, 23, 83, 21, 17, 41, 20, 25, 49, 41, 27, 37, 57	$T_4 = 671$	$\bar{Y}_{4\cdot} = 26.84$
$\sum_{i=1}^4 n_i = 100$					$G = 2,183$	$\bar{Y} = 21.83$

© Cengage Learning

themselves was over 95% black. Daly's main thesis was that the health of persons living in similar federal or state housing projects varied according to the racial composition of the surrounding neighborhoods: the more similar the racial composition of the surrounding neighborhood was, the better the health of the residents in the project would be. According to Daly, federal housing planners had never considered information about overall neighborhood racial composition and its relationship to health as criteria for selecting areas for such projects. This study, it was hoped, might provide some concrete recommendations for improved federal planning.

The sample means of the data in Table 17.3 vary. To determine whether the observed differences in these sample means are attributable solely to chance or whether there is evidence of true underlying differences in these means, we can perform a one-way fixed-effects ANOVA. The possibility of violations of the assumptions underlying this methodology are not of great concern for this data set, since the sample sizes are equal and reasonably large and since observations on women from different households may be treated as independent. ■

17.3 Methodology for One-way Fixed-effects ANOVA

The null hypothesis of equal population means ($H_0: \mu_1 = \mu_2 = \dots = \mu_k$) is tested by using an F test. The test statistic is calculated as follows:⁵

$$F = \frac{\text{MST}}{\text{MSE}} \quad (17.2)$$

where

$$\text{MST} = \frac{\sum_{i=1}^k (T_i^2/n_i) - G^2/n}{k - 1} \quad (17.3)$$

and

$$\text{MSE} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^k (T_i^2/n_i)}{n - k} \quad (17.4)$$

When H_0 is true (i.e., when the population means are all equal), the F statistic of (17.2) has an F distribution with $(k - 1)$ numerator and $(n - k)$ denominator degrees of freedom. Thus, for a given α , we would reject H_0 and conclude that some (i.e., at least two) of the population means differ from one another if

$$F \geq F_{k-1, n-k, 1-\alpha}$$

where $F_{k-1, n-k, 1-\alpha}$ is the $100(1 - \alpha)\%$ point of the F distribution with $(k - 1)$ and $(n - k)$ degrees of freedom. The critical region for this test involves only upper percentage points of the F distribution, since only large values of the F statistic (usually values much greater than 1) will provide significant evidence for rejecting H_0 .

17.3.1 Numerical Illustration

For the data given in Table 17.3, the calculations needed to perform the F test proceed as follows:

$$\sum_{i=1}^4 \sum_{j=1}^{25} Y_{ij}^2 = \underbrace{(49)^2 + (12)^2 + \dots + (37)^2 + (57)^2}_{\text{Sum of 100 squared observations}} = 72,851.00$$

⁵ The designation MST is read "mean square due to treatments," since the populations being compared often represent treatment groups. You might notice that (17.2) resembles the F statistic MSR/MSE used in multiple linear regression, despite differences in their calculation. The relationship between ANOVA and regression is explored more in Section 17.4. Some texts also use the terms *MS-Between* and *MS-Within* to describe the MST and MSE, respectively, as explained in Section 17.3.2.

$$\sum_{i=1}^4 \frac{T_i^2}{n_i} = \frac{(473)^2}{25} + \frac{(518)^2}{25} + \frac{(521)^2}{25} + \frac{(671)^2}{25} = 48,549.40$$

$$\frac{G^2}{n} = \frac{(2,183)^2}{100} = 47,654.89$$

$$\begin{aligned} \text{MST} &= \frac{\sum_{i=1}^4 (T_i^2/n_i) - G^2/n}{4 - 1} \\ &= \frac{48,549.40 - 47,654.89}{3} \\ &= 298.17 \end{aligned}$$

$$\begin{aligned} \text{MSE} &= \frac{\sum_{i=1}^4 \sum_{j=1}^{25} Y_{ij}^2 - \sum_{i=1}^4 (T_i^2/n_i)}{100 - 4} \\ &= \frac{72,851.00 - 48,549.40}{96} \\ &= 253.14 \end{aligned}$$

$$\begin{aligned} F &= \frac{\text{MST}}{\text{MSE}} \\ &= \frac{298.17}{253.14} \\ &= 1.178 \end{aligned}$$

The preceding calculations can be conveniently performed by a computer program. An example of computer output for these data (using SAS's GLM procedure) is shown next.

Edited SAS Output (PROC GLM) for ANOVA of CMI Scores

CLASS LEVEL INFORMATION					
Class	Levels	Values			
NBRHOOD	4	1 2 3 4			
Number of Observations Read		100			
Number of Observations Used		100			
Dependent Variable: CMI					
		F statistic			
		P-value for F test			
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	894.51000	298.17000	1.18	0.3223
Error	96	24301.60000	253.14167	MST	MSE
Corrected Total	99	25196.11000			

Using these calculations, we may test $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ (i.e., the hypothesis that there are *no* differences among the true mean CMI scores for the four neighborhoods) against H_A : “There are differences among the true mean CMI scores.” For example, if $\alpha = .10$, we would find from the F tables that $F_{3, 96, 0.90} = 2.15$, which is greater than the computed F of 1.178. Thus, we would not reject H_0 at $\alpha = .10$.

To find the P -value for this test, we first note that $F_{3, 96, 0.75} = 1.41$, which also exceeds the computed F . Thus, we know that $P > .25$. From the preceding SAS output, we see that the P -value is in fact .3223. Therefore, we conclude (as did Daly) that the observed mean CMI scores for the four neighborhoods do not significantly differ.

If a significant difference among the sample means had been found, it would still have been up to the investigator to determine whether the actual magnitude of the difference(s) was meaningful in a practical sense and whether the pattern of the difference(s) was as hypothesized. In our example, the distribution of percentages of blacks in the surrounding neighborhoods (see Table 17.3) indicates that the observed differences among the sample means clearly do not match the pattern hypothesized. Under Daly’s conjecture, Cherryview (with a surrounding neighborhood that was 17% black) would have been expected to register the highest observed mean CMI score (i.e., poorer health status), followed by Northhills (36%), Easton (65%), and finally Morningside (100%). This was not the order actually obtained. Daly also examined whether her conjecture was supported when she controlled for other possibly relevant factors, such as “months lived in the neighborhood,” “number of children,” and “marital status.” However, no significant results were obtained from these analyses either.

17.3.2 Rationale for the F Test in One-way ANOVA

The use of the F test described in the preceding subsection may be motivated by various considerations. We, therefore, offer an intuitive theoretical appreciation of its purpose and also provide some insight into the rationale behind more complex ANOVA testing procedures.

1. The F test in one-way fixed-effects ANOVA is a generalization of the two-sample t test. We can easily show with a little algebra that the numerator and denominator components in the F statistic (17.2) for one-way ANOVA are simple generalizations of the corresponding components in the square of the ordinary two-sample t -test statistic. In fact, when $k = 2$, the F statistic for one-way ANOVA is exactly equal to the square of the corresponding t statistic. Such a result is intuitively reasonable, since the numerator degrees of freedom of F when $k = 2$ is 1, and we have previously noted that the square of a t statistic with v degrees of freedom has the F distribution with 1 and v degrees of freedom in the numerator and denominator, respectively.

In particular, recall that the two-sample t -test statistic is given by the formula

$$T = \frac{(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) / \sqrt{1/n_1 + 1/n_2}}{S_p}$$

where the pooled sample variance S_p^2 is given by

$$S_p^2 = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^2 \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2$$

or, equivalently, by

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

where S_1^2 and S_2^2 are the sample variances for groups 1 and 2, respectively. Focusing first on the denominator (MSE) of the F statistic (17.2), we can show with some algebra that

$$\begin{aligned} \text{MSE} &= \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^k (T_i^2/n_i)}{n - k} \\ &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \cdots + (n_k - 1)S_k^2}{n_1 + n_2 + \cdots + n_k - k} \end{aligned}$$

Thus, MSE is a pooled estimate of the common population variance σ^2 , since it is a weighted sum of the k estimates of σ^2 obtained by using the k different sets of observations. Furthermore, when $k = 2$, MSE is equal to S_p^2 .

Looking at the numerator (MST) of the F statistic, we can show that

$$\begin{aligned} \text{MST} &= \frac{\sum_{i=1}^k (T_i^2/n_i) - G^2/n}{k - 1} \\ &= \frac{1}{k - 1} \sum_{i=1}^k n_i(\bar{Y}_{i\cdot} - \bar{Y})^2 \end{aligned}$$

which simplifies to $(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot})^2/(1/n_1 + 1/n_2)$ when $k = 2$. Thus, the equivalence is established.

2. The F statistic is the ratio of two variance estimates. We have already seen that MSE is a pooled estimate of the common population variance σ^2 ; that is, the true average (or mean) value (μ_{MSE} , say) of MSE is σ^2 . It turns out, however, that MST estimates σ^2 only when H_0 is true—that is, only when the population means $\mu_1, \mu_2, \dots, \mu_k$ are all equal. In fact, the general formula for the expected value of MST (μ_{MST} , say) is given by

$$\mu_{\text{MST}} = \sigma^2 + \frac{1}{k - 1} \sum_{i=1}^k n_i(\mu_i - \bar{\mu})^2 \quad (17.5)$$

where $\bar{\mu} = \sum_{i=1}^k n_i \mu_i / n$. By inspection of expression (17.5), we can see that MST estimates σ^2 only when all the μ_i are equal, in which case $\mu_i = \bar{\mu}$ for every i , and so $\sum_{i=1}^k n_i(\mu_i - \bar{\mu})^2 = 0$. Otherwise, both terms on the right-hand side of (17.5) are positive, and MST estimates something greater in value than σ^2 . In other words,

$$\mu_{\text{MST}} = \sigma^2 \quad \text{when } H_0 \text{ is true}$$

and

$$\mu_{\text{MST}} > \sigma^2 \quad \text{when } H_0 \text{ is not true}$$

Loosely speaking, then, the F statistic MST/MSE may be viewed as approximating in some sense the ratio of population means

$$\frac{\mu_{\text{MST}}}{\mu_{\text{MSE}}} = \frac{\sigma^2 + \frac{1}{(k-1)} \sum_{i=1}^k n_i(\mu_i - \bar{\mu})^2}{\sigma^2} \quad (17.6)$$

When H_0 is true, the numerator and denominator of (17.6) both equal σ^2 , and the F statistic is the ratio of two estimates of the same variance. Furthermore, F can be expected to give different values depending on whether H_0 is true or not; that is, F should take a value close to 1 if H_0 is true (since in that case it approximates $\sigma^2/\sigma^2 = 1$), whereas F should be larger than 1 if H_0 is false (since the numerator of (17.6) is greater than the denominator).

3. *The F statistic compares the variability between groups to the variability within groups.* As with regression analysis, the total variability in the observations in a one-way ANOVA situation is measured by a total sum of squares:

$$\text{SSY} = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2 \quad (17.7)$$

Furthermore, it can be shown that

$$\text{SSY} = \text{SST} + \text{SSE} \quad (17.8)$$

where

$$\text{SST} = (k-1)\text{MST} = \sum_{i=1}^k n_i(\bar{Y}_i - \bar{Y})^2$$

and

$$\text{SSE} = (n-k)\text{MSE} = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2$$

SST can be considered to be a measure of the variability *between* (or *across*) populations. It involves components of the general form $(\bar{Y}_i - \bar{Y})$, which is the difference between the i th group mean and the overall mean.

SSE is a measure of the variability *within* populations and gives no information about variability between populations. It involves components of the general form $(Y_{ij} - \bar{Y}_{i\cdot})$, which is the difference between the j th observation in the i th group and the mean for the i th group.

If SST is quite large in comparison to SSE, we know that most of the total variability is due to differences *between* populations rather than to differences *within* populations. Thus, it is natural in such a case to suspect that the population means are not all equal.

By writing the F statistic (17.2) in the form

$$F = \frac{\text{SST} \left(\frac{n - k}{k - 1} \right)}{\text{SSE}}$$

we can see that F will be large whenever SST accounts for a much larger proportion of the total sum of squares than does SSE.

17.3.3 ANOVA Table for One-way ANOVA

As in regression analysis, the results of any ANOVA procedure can be summarized in an ANOVA table. The ANOVA table for one-way ANOVA is given in general form in Table 17.4. Table 17.5 is the ANOVA table for our example involving the CMI data; this ANOVA table was also previously shown in the SAS output on page 489.

The “Source” and “SS” columns in Table 17.4 display the components of the fundamental equation of one-way ANOVA:

$$\text{SSY} = \text{SST} + \text{SSE}$$

The “MS” column contains the sums of squares divided by their corresponding degrees of freedom. The two mean squares are then used to form the numerator and denominator for the F test.

TABLE 17.4 General ANOVA table for one-way ANOVA (k populations)

Source	d.f.	SS	MS	F
Between	$k - 1$	SST	$\text{MST} = \frac{\text{SST}}{k - 1}$	$\frac{\text{MST}}{\text{MSE}}$
Within	$n - k$	SSE	$\text{MSE} = \frac{\text{SSE}}{n - k}$	
Total	$n - 1$	SSY		

© Cengage Learning

TABLE 17.5 ANOVA table for CMI data ($k = 4$)

Source	d.f.	SS	MS	F
Between (neighborhoods)	3	894.51	298.17	1.18
Within (error)	96	24,301.60	253.14	
Total	99	25,196.11		

© Cengage Learning

17.4 Regression Model for Fixed-effects One-way ANOVA

We observed earlier that most ANOVA procedures can also be considered in a regression analysis setting; this can be done by defining appropriate dummy variables in a regression model.⁶ The ANOVA F tests are then formulated in terms of hypotheses concerning the coefficients of the dummy variables in the regression model.⁷

- **Example 17.2** For the example involving the CMI data of Daly's (1973) study (see Table 17.3), a number of alternative regression models could be used to describe the situation, depending on the coding schemes used for the dummy variables. One such model is

$$Y = \mu + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + E \quad (17.9)$$

where the regression coefficients are denoted as μ , α_1 , α_2 , and α_3 and the independent variables are defined as

$$X_1 = \begin{cases} 1 & \text{if neighborhood 1} \\ -1 & \text{if neighborhood 4} \\ 0 & \text{if otherwise} \end{cases} \quad X_2 = \begin{cases} 1 & \text{if neighborhood 2} \\ -1 & \text{if neighborhood 4} \\ 0 & \text{if otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if neighborhood 3} \\ -1 & \text{if neighborhood 4} \\ 0 & \text{if otherwise} \end{cases}$$

Although we previously used the Greek letter β with subscripts to denote regression coefficients, we have changed the notation for our ANOVA regression model so that these coefficients correspond to the parameters in the classical fixed-effects ANOVA model described in Section 17.5.

The coding scheme used here to define the dummy variables is called an *effect* coding scheme. The coefficients μ , α_1 , α_2 , and α_3 for this (dummy variable) model can each be expressed in terms of the underlying population means (μ_1 , μ_2 , μ_3 , and μ_4), as follows:

$$\mu = \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4}$$

$$\alpha_1 = \mu_1 - \mu$$

$$\alpha_2 = \mu_2 - \mu$$

$$\alpha_3 = \mu_3 - \mu \quad (17.10)$$

⁶ As mentioned earlier, we are restricting our attention here entirely to models with fixed factors. Models involving random factors will be treated in Section 17.6.

⁷ We will see later that a regression formulation is often desirable, if not mandatory, for dealing with certain nonorthogonal ANOVA problems involving two or more factors. We will discuss such problems in Chapter 20.

The coefficients in (17.10) are related as follows: $\mu_{Y|X_1, X_2, X_3} = \mu + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3$. Thus,

$$\begin{aligned}\mu_1 &= \mu_{Y|1, 0, 0} = \mu + \alpha_1 && \text{since } X_1 = 1, X_2 = 0, X_3 = 0 \text{ for neighborhood 1} \\ \mu_2 &= \mu_{Y|0, 1, 0} = \mu + \alpha_2 && \text{since } X_1 = 0, X_2 = 1, X_3 = 0 \text{ for neighborhood 2} \\ \mu_3 &= \mu_{Y|0, 0, 1} = \mu + \alpha_3 && \text{since } X_1 = 0, X_2 = 0, X_3 = 1 \text{ for neighborhood 3} \\ \mu_4 &= \mu_{Y|-1, -1, -1} \\ &= \mu - \alpha_1 - \alpha_2 - \alpha_3 && \text{since } X_1 = X_2 = X_3 = -1 \text{ for neighborhood 4}\end{aligned}$$

Adding the left-hand sides and right-hand sides of these equations yields

$$\mu_1 + \mu_2 + \mu_3 + \mu_4 = 4\mu$$

or

$$\mu = \frac{1}{4} \sum_{i=1}^4 \mu_i$$

Solution (17.10) is obtained by solving for the regression coefficients α_i in terms of μ_1 , μ_2 , μ_3 , and μ_4 .

Model (17.9) involves coefficients that describe separate comparisons of the first three group means with the overall unweighted mean μ .⁸ In this model, $(\mu_4 - \mu)$ can be expressed as the negative sum of α_1 , α_2 , and α_3 . Moreover, model (17.9) can be fitted to provide *exactly* the same F statistic as is required in one-way ANOVA for the test of $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$. The equivalent regression null hypothesis is $H_0: \alpha_1 = \alpha_2 = \alpha_3 = 0$,⁹ the regression F statistic will have the same degrees of freedom (i.e., $k - 1 = 3$ and $n - k = 96$) as given previously, and the ANOVA table will be exactly the same as the one given in the last section (where we pooled the dummy variable effects into one source of variation with 3 degrees of freedom).

Other coding schemes for the independent variables yield exactly the same ANOVA table and F test as model (17.9), although the regression coefficients themselves represent different parameters and have different least-squares estimators. One frequently used coding scheme defines the independent variables as

$$X_i = \begin{cases} 1 & \text{if neighborhood } i \\ 0 & \text{otherwise} \end{cases} \quad i = 1, 2, 3$$

⁸ It is worth distinguishing μ , which is the unweighted mean of k group means, from the overall mean μ^* for the entire population. The mean μ^* is a weighted sum of the individual population means, with the weight w_i for i th mean being $\frac{n_i}{\sum_{j=1}^k n_j}$. In the Daly example, $w_i = \frac{n_i}{\sum_{j=1}^4 n_j} = \frac{1}{4}$, the proportion of the total residents that each neighborhood contributes.

⁹ When $\alpha_1 = \alpha_2 = \alpha_3 = 0$, it follows from simple algebra based on (17.10) that $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (e.g., $\alpha_1 = \mu_1 - \mu = 0$ implies that $\mu_1 = \mu$; $\alpha_2 = \mu_2 - \mu = 0$ implies that $\mu_2 = \mu = \mu_1$, etc.).

This coding scheme is an example of *reference cell* coding. The referent group in this case is group 4, and the regression coefficients describe separate comparisons of the first three population means with μ_4 :

$$\mu = \mu_4$$

$$\alpha_1 = \mu_1 - \mu_4$$

$$\alpha_2 = \mu_2 - \mu_4$$

$$\alpha_3 = \mu_3 - \mu_4$$



17.4.1 Effect Coding Model

For the general situation involving k populations, the following model using effect coding is analogous to that of the previous section:

$$Y = \mu + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_{k-1} X_{k-1} + E \quad (17.11)$$

in which

$$X_i = \begin{cases} 1 & \text{for population } i \\ -1 & \text{for population } k \\ 0 & \text{otherwise} \end{cases} \quad i = 1, 2, \dots, k-1$$

The coefficients of this model can be expressed in terms of the k population means $\mu_1, \mu_2, \dots, \mu_k$ as

$$\begin{aligned} \mu &= \frac{\mu_1 + \mu_2 + \cdots + \mu_k}{k} \\ \alpha_1 &= \mu_1 - \mu \\ \alpha_2 &= \mu_2 - \mu \\ &\vdots \\ \alpha_{k-1} &= \mu_{k-1} - \mu \\ -(\alpha_1 + \alpha_2 + \cdots + \alpha_{k-1}) &= \mu_k - \mu \end{aligned} \quad (17.12)$$

for model (17.11). The F statistic for one-way ANOVA with k populations can be obtained equivalently by testing the null hypothesis $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_{k-1} = 0$ in model (17.11).

17.4.2 Reference Cell Coding Model

Another coding scheme for one-way ANOVA of k populations uses the following reference cell coding:

$$X_i = \begin{cases} 1 & \text{for population } i \\ 0 & \text{otherwise} \end{cases} \quad i = 1, 2, \dots, k-1$$

Again, only $k-1$ such dummy variables are needed, and the population “left out” becomes the reference population (group or cell). Thus, given X_i as defined here, group k is the

reference group. (If X_1 had been left out instead of X_k , then group 1 would have been the reference group.)

With the specified reference cell coding, the responses in each group under the model $Y = \mu_k + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_{k-1} X_{k-1} + E$ are as follows:

$$\text{Group 1: } Y = \mu_k + \alpha_1 + E$$

$$\text{Group 2: } Y = \mu_k + \alpha_2 + E$$

$$\vdots$$

$$\text{Group } k-1: Y = \mu_k + \alpha_{k-1} + E$$

$$\text{Group } k: Y = \mu_k + E$$

In turn, the regression coefficients can be written in terms of group means as follows:

$$\alpha_1 = \mu_1 - \mu_k$$

$$\alpha_2 = \mu_2 - \mu_k$$

$$\vdots$$

$$\alpha_{k-1} = \mu_{k-1} - \mu_k$$

Thus, different coding schemes (e.g., an effect coding or reference cell coding) yield regression coefficients representing different parameters (e.g., $\alpha_1 = \mu_1 - \mu$ for the effect coding but $\alpha_1 = \mu_1 - \mu_k$ for the reference cell coding). Regardless of the coding scheme used, the test of the hypothesis $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ can be obtained equivalently by testing $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_{k-1} = 0$ in the regression model (17.11). In other words, the correct SST, SSE, and $F_{k-1, n-k}$ values are obtained from the regression analysis, regardless of the (legitimate) coding scheme chosen.

17.5 Fixed-effects Model for One-way ANOVA

Many textbooks and articles that deal strictly with ANOVA procedures use a more classical type of model than the regression model given earlier to describe the fixed-effects one-way ANOVA situation. The more classical type of model is often referred to as a *fixed-effects ANOVA model*; in it, all factors under consideration are fixed (i.e., the levels of each factor are the only levels of interest). The *effects* referred to in this type of model represent measures of the influence (i.e., the effect) that different levels of the factor have on the dependent variable.¹⁰ Such measures are often expressed in the form of differences between a given mean and an overall mean; that is, the effect of the i th population is often measured as the amount by which the i th population mean differs from an overall mean.

¹⁰ In situations involving models with two or more factors, *effects* can also refer to measures of the influence of combinations of levels of the different factors on the dependent variable.

■ **Example 17.3** For the CMI data ($k = 4$), the fixed-effects ANOVA model is

$$Y_{ij} = \mu + \alpha_i + E_{ij} \quad i = 1, 2, 3, 4; \quad j = 1, 2, \dots, 25 \quad (17.13)$$

where

Y_{ij} = j th observation from the i th population

$$\mu = \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4}$$

$\alpha_1 = \mu_1 - \mu$ = Differential effect of neighborhood 1

$\alpha_2 = \mu_2 - \mu$ = Differential effect of neighborhood 2

$\alpha_3 = \mu_3 - \mu$ = Differential effect of neighborhood 3

$\alpha_4 = \mu_4 - \mu$ = Differential effect of neighborhood 4

$E_{ij} = Y_{ij} - \mu - \alpha_i$ = Error component associated with the j th observation from the i th population

One important property of this model is that the sum of the four α effects is 0; that is, $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0$. Thus, these effects represent differentials from the overall population mean μ that average out to 0. Nevertheless, the effect of one level (i.e., a neighborhood) may differ considerably from the effect of another. If this proved to be the case, we would probably find that our F test leads to rejection of the null hypothesis of equal population mean CMI scores for the four neighborhoods.

Another important property of this model is that the effects $\alpha_1, \alpha_2, \alpha_3$, and α_4 , which are population parameters defined in terms of population means, can each be estimated from the data by appropriately substituting the usual estimates of the means into the expressions for the effects. For our example, the estimated effects are given by

$$\hat{\alpha}_1 = \bar{Y}_{1\cdot} - \bar{Y} = \text{Sample mean CMI score for neighborhood 1} - \text{Overall sample mean CMI score for all neighborhoods}$$

$$\hat{\alpha}_2 = \bar{Y}_{2\cdot} - \bar{Y} = \text{Sample mean CMI score for neighborhood 2} - \text{Overall sample mean CMI score for all neighborhoods}$$

$$\hat{\alpha}_3 = \bar{Y}_{3\cdot} - \bar{Y} = \text{Sample mean CMI score for neighborhood 3} - \text{Overall sample mean CMI score for all neighborhoods}$$

$$\hat{\alpha}_4 = \bar{Y}_{4\cdot} - \bar{Y} = \text{Sample mean CMI score for neighborhood 4} - \text{Overall sample mean CMI score for all neighborhoods}$$

The actual numerical values obtained from these formulas are as follows:

$$\hat{\alpha}_1 = 18.92 - 21.83 = -2.91$$

$$\hat{\alpha}_2 = 20.72 - 21.83 = -1.11$$

$$\hat{\alpha}_3 = 20.84 - 21.83 = -0.99$$

$$\hat{\alpha}_4 = 26.84 - 21.83 = 5.01$$

Like the population effects, the estimated effects sum to 0; that is, $\sum_{i=1}^4 \hat{\alpha}_i = 0$.

If we consider the general one-way ANOVA situation (with k populations and n_i observations from the i th population), the fixed-effects one-way ANOVA model may be written as follows:

$$Y_{ij} = \mu + \alpha_i + E_{ij} \quad i = 1, 2, \dots, k; \quad j = 1, 2, \dots, n_i \quad (17.14)$$

where

Y_{ij} = j th observation from the i th population

$$\mu = \frac{\mu_1 + \mu_2 + \dots + \mu_k}{k}$$

$\alpha_i = \mu_i - \mu$ = Differential effect of population i

$E_{ij} = Y_{ij} - \mu - \alpha_i$ = Error component associated with the j th observation from the i th population

Here it is easy to show that the sum of the α effects is 0; that is, $\sum_{i=1}^k \alpha_i = 0$. Similarly, the estimated effects, $\hat{\alpha}_i^* = (\bar{Y}_i - \bar{Y}^*)$, where $\bar{Y}^* = \sum_{i=1}^k \bar{Y}_i / k$, satisfy the constraint $\sum_{i=1}^k \hat{\alpha}_i^* = 0$.

An alternative definition of μ is $\mu^* = \sum_{i=1}^k n_i \mu_i / n$, the overall weighted mean of the means. In this case, the weighted sum $\sum_{i=1}^k n_i \alpha_i = 0$, and the weighted sum of the estimated effects, $\hat{\alpha}_i = (\bar{Y}_i - \bar{Y})$, where $\bar{Y} = \sum_{i=1}^k n_i \bar{Y}_i / n$, satisfies $\sum_{i=1}^k n_i \hat{\alpha}_i = 0$.

Model (17.14) corresponds in structure to the regression model given by (17.9): the regression coefficients $\alpha_1, \alpha_2, \dots, \alpha_{k-1}$ are precisely the effects $\alpha_1 = \mu_1 - \mu, \alpha_2 = \mu_2 - \mu, \dots, \alpha_{k-1} = \mu_{k-1} - \mu$; the regression constant μ represents the overall (unweighted) mean μ ; and the negative sum of the regression coefficients $(-\sum_{i=1}^{k-1} \alpha_i)$ represents the effect $\alpha_k = \mu_k - \mu$. This is why we have defined each of these models using the same notation for the unknown parameters:

$$Y = \mu + \sum_{i=1}^{k-1} \alpha_i X_i + E \quad (\text{dummy variable regression model})$$

$Y_{ij} = \mu + \alpha_i + E_{ij} \quad (\text{fixed-effects ANOVA model})$

Notice that μ represents the unweighted average of the k population means, μ , rather than the weighted average, μ^* , even though the sample sizes can be different in the different populations. Correspondingly, the least-squares estimate of μ is $\sum_{i=1}^k \bar{Y}_i / k$, the unweighted average of the k sample means, rather than \bar{Y} . Nevertheless, the dummy variables in the regression model can be redefined to obtain a least-squares solution yielding \bar{Y} as the estimate of μ . The following dummy variable definitions

$$X_i = \begin{cases} -n_i & \text{if population } k \\ n_k & \text{if population } i \\ 0 & \text{otherwise} \end{cases} \quad i = 1, 2, \dots, k - 1$$

are required.

17.6 Random-effects Model for One-way ANOVA

In Section 17.1, we distinguished between fixed and random factors. In this section, we discuss how the one-way ANOVA model is conceptualized differently when the independent variable is a random factor.

■ **Example 17.4** To get some insight into the structure of random-effects models, reconsider Daly's (1973) study (see Table 17.3). It might be argued that the four different turnkey neighborhoods form a representative sample of a larger population of similar types of neighborhoods (some of which might even be predominantly white with differing percentages of blacks in the surrounding neighborhoods). If so, the neighborhood factor would have to be considered random, and the appropriate ANOVA model would be a *random-effects* one-way ANOVA model.¹¹ Its form would be essentially the same as that given in (17.13), except that the α components would be treated differently; that is, the random-effects model would be of the form¹²

$$Y_{ij} = \mu + A_i + E_{ij} \quad i = 1, 2, 3, 4; \quad j = 1, 2, \dots, 25 \quad (17.15)$$

In this model, the A_i 's can be viewed as constituting a set of random variables that have a common distribution—one that represents the entire population of possible effects (in our example, neighborhoods).

To perform the appropriate analysis, we must assume that the distribution of A_i is normal with zero mean:

$$A_i \sim N(0, \sigma_A^2) \quad i = 1, 2, 3, 4 \quad (17.16)$$

where σ_A^2 denotes the variance of A_i . We must also assume that the A_i 's are independent of the E_{ij} 's and of each other.¹³

The requirement of zero mean in (17.16) is similar in philosophy to the requirement that $\sum_{i=1}^k \alpha_i = 0$ for the fixed-effects model. When the random-effects model (17.15) applies, we assume that the average (i.e., mean) effect of neighborhoods is 0 over the entire collection of neighborhoods; that is, we assume that $\mu_{A_i} = 0$, $i = 1, 2, 3, 4$.

How do we state our null hypothesis? Because we have specified that the neighborhood effects average out to 0 over the entire collection of possible effects, the only way to

¹¹ This type of model is also referred to as a *variance-components model*.

¹² The historical convention has been to use Latin letters (X, Y, Z , etc.) to denote random variables and to use Greek letters (β, μ, σ, τ) to denote parameters. This requires using A 's rather than α 's to denote random effects.

¹³ Since the *same* random variable A_i defined by (17.16) appears in (17.15) for each of the observations from population i , it follows that $\text{corr}(Y_{ij}, Y_{ij}) = \sigma_A^2 / (\sigma_A^2 + \sigma^2)$, the so-called intraclass (or intrapopulation) correlation. Thus, in contrast to the one-way *fixed-effects* ANOVA model, the one-way *random-effects* ANOVA model introduces a dependency among the set of observations from the same population.

assess whether any significant neighborhood effects are present at all involves considering σ_A^2 . If there is no variability (i.e., $\sigma_A^2 = 0$), all neighborhood effects must be 0. If there is variability (i.e., $\sigma_A^2 > 0$), some nonzero effects must exist in the collection of neighborhood effects.

Thus, our null hypothesis of no neighborhood effects should be stated as follows:

$$H_0: \sigma_A^2 = 0 \quad (17.17)$$

This hypothesis is analogous to the null hypothesis (17.1) used in the fixed-effects case, although it happens to be stated in terms of a population variance rather than in terms of population means. In a random-effects model, it is the variability among the levels of the random factor that is of interest rather than particular mean values or comparisons among them.

We must still explain why the F test given by (17.2) for the fixed-effects model is exactly the same as that used for the random-effects model.¹⁴ Such an explanation is best made by considering the properties of the mean squares MST and MSE. In Section 17.3.2, in connection with the fixed-effects model, we saw that the F statistic, MST/MSE , could be considered a rough approximation to the ratio of the means of these mean squares,

$$\frac{\mu_{MST}}{\mu_{MSE}} = \frac{\sigma^2 + \frac{1}{(k-1)} \sum_{i=1}^k n_i(\mu_i - \bar{\mu})^2}{\sigma^2}$$

The parameters μ_{MST} and μ_{MSE} are often called *expected mean squares*.

A similar argument can be made with regard to the F statistic for the random-effects model. In particular, for the random-effects model, as well as for the fixed-effects model, the denominator MSE estimates σ^2 ; that is,

$$\mu_{MSE} = \sigma^2$$

Furthermore, for the random-effects model applied to our example ($k = 4$, $n_i = 25$), it can be shown that MST estimates

$$\mu_{MST(\text{random})} = \sigma^2 + 25\sigma_A^2 \quad (17.18)$$

Thus, for the random-effects model, F approximates the ratio

$$\frac{\mu_{MST(\text{random})}}{\mu_{MSE}} = \frac{\sigma^2 + 25\sigma_A^2}{\sigma^2} \quad (17.19)$$

Since the null hypothesis in this case is $H_0: \sigma_A^2 = 0$, the ratio (17.19) simplifies to $\sigma^2/\sigma^2 = 1$ when H_0 is true. Thus, the F statistic under H_0 again consists of the ratio of two estimates of the same variance σ^2 . Furthermore, because $\sigma_A^2 > 0$ when H_0 is not true, the greater the variability among neighborhood effects is, the larger the observed value of F should be.

¹⁴ Again, the F tests are computationally equivalent for fixed-effects and random-effects models only in one-way ANOVA. When dealing with two-way or higher-way ANOVA, the testing procedures may be different.

In general, the random-effects model for one-way ANOVA is given by

$$Y_{ij} = \mu + A_i + E_{ij} \quad i = 1, 2, \dots, k; \quad j = 1, 2, \dots, n_i \quad (17.20)$$

where A_i and E_{ij} are independent random variables satisfying $A_i \sim N(0, \sigma_A^2)$ and $E_{ij} \sim N(0, \sigma^2)$.¹⁵

For this model, F approximates the following ratio of expected mean squares:

$$\frac{\mu_{\text{MST}(\text{random})}}{\mu_{\text{MSE}}} = \frac{\sigma^2 + n_0 \sigma_A^2}{\sigma^2}$$

where

$$n_0 = \frac{\sum_{i=1}^k n_i - \left(\sum_{i=1}^k n_i^2 / \sum_{i=1}^k n_i \right)}{k - 1}$$

functions as an average of the n_i observations selected from each population.¹⁶ The F statistic for the random-effects model is, therefore, the ratio of two estimates of σ^2 when $H_0: \sigma_A^2 = 0$ is true.

Table 17.6 summarizes the similarities and the differences between the fixed- and random-effects models. Tables with similar formats will be used in subsequent chapters to highlight distinctions for ANOVA situations with more than two factors.

TABLE 17.6 Combined one-way ANOVA table for fixed- and random-effects models

Source	d.f.	MS	F	Expected Mean Square (EMS)	
				Fixed Effects	Random Effects
Between	$k - 1$	MST	$\frac{\text{MST}}{\text{MSE}}$	$\sigma^2 + \frac{1}{k-1} \sum_{i=1}^k n_i (\mu_i - \bar{\mu})^2$	$\sigma^2 + n_0 \sigma_A^2$
Within	$n - k$	MSE		σ^2	σ^2
Total	$n - 1$				
				$H_0: \mu_1 = \mu_2 = \dots = \mu_k$	$H_0: \sigma_A^2 = 0$

© Cengage Learning

¹⁵ Under these assumptions, $Y_{ij} \sim N(\mu, \sigma_A^2 + \sigma^2)$ for all (i, j) . And $\text{corr}(Y_{ij}, Y_{ij'}) = \sigma_A^2 / (\sigma_A^2 + \sigma^2)$ for all $j \neq j'$ with fixed i . In other words, observations from *different* populations are independent, but observations from the *same* population are correlated.

¹⁶ When all the n_i are equal, as in the Daly (1973) example (i.e., $n_i = n^*$), then n_0 is equal to n^* , since

$$n_0 = \frac{kn^* - (kn^*/kn^*)}{k - 1} = n^*$$

In the Daly (1973) example, $n^* = 25$.

17.7 Multiple-comparison Procedures for Fixed-effects One-way ANOVA

When we find that an ANOVA F test for simultaneously comparing several population means in a fixed-effects ANOVA model is statistically significant, our next step customarily is to determine which *specific* differences exist among the population means. For example, if we are comparing four means (fixed-effects case) and the null hypothesis $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ is rejected,¹⁷ we next try to determine which subgroups of means are different by considering more specific hypotheses, such as $H_{01}: \mu_1 = \mu_2$, $H_{02}: \mu_2 = \mu_3$, $H_{03}: \mu_3 = \mu_4$, or even $H_{04}: (\mu_1 + \mu_2)/2 = (\mu_3 + \mu_4)/2$, which compares the average effect of populations 1 and 2 with the average effect of populations 3 and 4. Such specific comparisons may have been of interest to us before (a priori) the data were collected, or they may arise in completely exploratory studies only after (a posteriori) the data have been examined. In either event, a seemingly reasonable first approach to drawing inferences about differences among the population means would be to conduct several t tests and to focus on all the tests found to be significant. For example, if all pairwise comparisons among the means are desired, then ${}_4C_2 = 6$ such tests must be performed with regard to 4 means (or in general, ${}_kC_2 = k(k - 1)/2$ tests with regard to k means). Thus, in testing $H_0: \mu_i = \mu_j$ at the α level of significance, we could reject this H_0 when

$$|T| \geq t_{n-k, 1-\alpha/2}$$

where

$$T = \frac{(\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}) - 0}{\sqrt{\text{MSE}(1/n_i + 1/n_j)}}$$

and where n is the total number of observations; k is the number of means under consideration; n_i and n_j are the sizes of the samples selected from the i th and j th populations, respectively; $\bar{Y}_{i\cdot}$ and $\bar{Y}_{j\cdot}$ are the corresponding sample means; and MSE is the mean-square-error term, with $(n - k)$ degrees of freedom, that estimates the (homoscedastic) variance σ^2 . MSE is used instead of a simple two-sample estimate of σ^2 based entirely on data from groups i and j ; this is because MSE is a better estimate of σ^2 (in terms of degrees of freedom) under the assumption of variance homogeneity over all k populations.

Equivalently, one could reject H_0 if the $100(1 - \alpha)\%$ confidence interval

$$(\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}) \pm t_{n-k, 1-\alpha/2} \sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

does not include 0.

¹⁷ This section deals only with fixed-effects ANOVA problems. The random-effects model treats the observed factor levels as a sample from a larger population of levels of interest and, therefore, is not directed exclusively at comparisons of the sampled levels.

Unfortunately, performing several such t tests has a serious drawback: the more null hypotheses there are to be tested, the more likely it is that one of them will be rejected even if all the null hypotheses are actually true. In other words, if several such tests are made, each at the α level, the probability of incorrectly rejecting *at least one* null hypothesis (i.e., making a Type I error) is much larger than α and continues to increase with each additional test made. Moreover, if in an exploratory study the investigator decides to compare only the sample means that are most discrepant (e.g., the largest versus the smallest), the testing procedure becomes biased in favor of rejecting H_0 because only the comparisons most likely to be significant are made. This bias will be reflected in the fact that the actual probability of incorrectly rejecting a given true null hypothesis exceeds the α level specified for the test.

Many different hypothesis-testing procedures have been devised to provide an overall significance level of α even when several tests are performed. All of these procedures are grouped under the heading “multiple-comparison procedures.” We shall focus here on three such approaches—the first attributed to Bonferroni, the second attributed to Tukey and Kramer, and the third attributed to Scheffé. Detailed mathematical discussions of these and other multiple-comparison methods can be found in Miller (1966), Guenther (1964), Lindman (1974) and Kutner et al. (2004).

17.7.1 The Bonferroni Approach

A conservative way to circumvent the problem of distorted significance levels when performing several tests involves reducing the significance level used for each individual test sufficiently to fix the *overall significance level* (i.e., the probability of incorrectly rejecting at least one of the null hypotheses being tested) at some desired level (say, α). If we perform l such tests, the maximum possible value for this overall significance level is $l\alpha$. Thus, one simple way to ensure an overall significance level of at most α is to use α/l as the significance level for each separate test. This approach is often referred to as the Bonferroni method. For example, if five contrasts have been identified a priori as being of interest, then performing each of the five relevant hypothesis tests at the $.05/5 = .01$ level of significance will ensure an overall significance level of no more than $\alpha = .05$.

One disadvantage of the Bonferroni method is that the *true* overall significance level may be considerably lower than α , and, in extreme situations, it may be so low that none of the individual tests will be rejected (i.e., the overall power of the method will be low). However, the true overall significance level using the Bonferroni approach will generally not be too low when the number of tests performed is close to the number of levels of the factor in question.

■ Example 17.5 Consider the set of data given in Table 17.7, which was collected from an experiment designed to compare the relative potencies of four cardiac substances. In the experiment, a suitable dilution of one of the substances was slowly infused into an anesthetized guinea pig, and the dosage at which the pig died was recorded. Ten guinea pigs were used for each substance, and the laboratory environment and the measurement procedures were assumed to be identical for each guinea pig. The main research goal was to determine whether any differences existed among the potencies of the four substances and, if so, to quantify those differences. The overall ANOVA table for comparing the mean potencies of the four cardiac substances is given in Table 17.8.

TABLE 17.7 Potencies (dosages at death) of four cardiac substances

Substance	Sample Size (n_i)	Dosage at Death (Y_{ij})	Total	Sample Mean (\bar{Y}_i)	Sample Variance (S_i^2)
1	10	29, 28, 23, 26, 26, 19, 25, 29, 26, 28	259	25.9	9.4333
2	10	17, 25, 24, 19, 28 21, 20, 25, 19, 24	222	22.2	12.1778
3	10	17, 16, 21, 22, 23 18, 20, 17, 25, 21	200	20.0	8.6667
4	10	18, 20, 25, 24, 16 20, 20, 17, 19, 17	196	19.6	8.7111

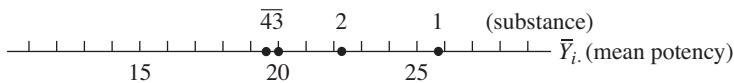
© Cengage Learning

TABLE 17.8 ANOVA table for data of Table 17.7

Source	d.f.	SS	MS	F
Substances	3	249.875	83.292	8.545 ($P < .001$)
Error	36	350.900	9.747	
Total	39	600.775		

© Cengage Learning

The global F test strongly rejects ($P < .001$) the null hypothesis of equality of the four population means. Therefore, the multiple-comparison question arises: What is the best way to account for the differences found? As a crude first step, we can examine the nature of the differences with the help of a schematic diagram of ordered sample means (Figure 17.1). In the diagram, an overbar has been drawn over the labels for substances 3 and 4 to indicate that the sample mean potencies for these two substances are quite similar. On the other hand, no overbar has been drawn connecting 1 and 2 with each other or with 3 and 4, suggesting that substances 1 and 2 differ from each other as well as from both 3 and 4.

**FIGURE 17.1** Crude comparison of sample means for potency data (© Cengage Learning)

Such an overall quantification of the differences among the population means is desired from a multiple-comparison analysis. Nevertheless, the purely descriptive approach taken does not account for the sampling variability associated with any estimated comparison of interest. As a result, two sample means that seem practically different may not, in fact, be statistically different. Let us consider how to evaluate the data in Table 17.7 using the Bonferroni method and an overall significance level of $\alpha = .05$ for all pairwise comparisons of the mean potencies of the four cardiac substances. This approach requires

computing ${}_4C_2 = 6$ confidence intervals, each associated with a significance level of $\alpha/6 = .05/6 = .0083$, utilizing the formula

$$(\bar{Y}_i - \bar{Y}_j) \pm t_{36,1-0.0083/2} \sqrt{\text{MSE} \left(\frac{1}{10} + \frac{1}{10} \right)}$$

The right-hand side of this expression is calculated as

$$t_{36,0.99585} \sqrt{9.747 \left(\frac{1}{5} \right)} = 2.79(1.396) = 3.895$$

Thus, the confidence intervals for the six population mean differences are given as

- $\mu_1 - \mu_4: 6.3 \pm 3.895$; i.e., $(2.405, 10.195)^*$
- $\mu_1 - \mu_3: 5.9 \pm 3.895$; i.e., $(2.005, 9.795)^*$
- $\mu_1 - \mu_2: 3.7 \pm 3.895$; i.e., $(-0.195, 7.595)$
- $\mu_2 - \mu_4: 2.6 \pm 3.895$; i.e., $(-1.295, 6.495)$
- $\mu_2 - \mu_3: 2.2 \pm 3.895$; i.e., $(-1.695, 6.095)$
- $\mu_3 - \mu_4: 0.4 \pm 3.895$; i.e., $(-3.495, 4.295)$

The results are ordered by the value of the sample mean difference (largest to smallest). The preceding intervals reveal only two significant comparisons (the ones starred) and translate into the diagrammatic overall ranking shown in Figure 17.2. These results are somewhat ambiguous due to overlapping “sets of similarities,” which indicate that substances 2, 3, and 4 have essentially the same potency; that 1 and 2 have about the same potency; but also that 1 differs from both 3 and 4. In other words, one possible conclusion is that 2, 3, and 4 are to be grouped together and that 1 and 2 are to be grouped together—which is difficult to reconcile because substance 2 is common to both groups. Having to confront this ambiguity is quite fortuitous from a pedagogical standpoint, since such ambiguous results are not infrequently encountered in connection with multiple-comparison procedures. In our case, the results indicate that the procedure used was not sensitive enough to permit an adequate evaluation of substance 2. Repeating the analysis with a larger data set would help clear up the ambiguity. Alternatively, since the Bonferroni approach tends to be conservative (i.e., the confidence intervals tend to be wider than necessary to achieve the overall significance level desired), other multiple-comparison methods—such as that of Tukey and Kramer—may provide more precise results (i.e., narrower confidence intervals) and so may reduce or eliminate any ambiguity.

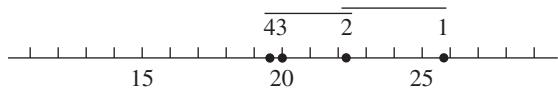


FIGURE 17.2 Bonferroni comparison of sample means for potency data (© Cengage Learning)

In the SAS output shown below, a representation similar to that of Figure 17.2 is displayed. In the output, sample means that are similar (i.e., not significantly different) at $\alpha = .05$ are labeled with the same letter.

Edited SAS Output (PROC GLM) for Data of Table 17.7 Using the Bonferroni Approach

Bonferroni (Dunn) *t* Tests for DOSAGE

Alpha	0.05		
Error Degrees of Freedom	36		
Error Mean Square	9.747222		
Critical Value of <i>t</i>	2.79197		
Minimum Significant Difference	3.8982		
MEANS WITH THE SAME LETTER ARE NOT SIGNIFICANTLY DIFFERENT.			
Bon Grouping	Mean	N	SUBSTANCE
A	25.900	10	1
A			
B	22.200	10	2
B			
B	20.000	10	3
B			
B	19.600	10	4

17.7.2 The Tukey–Kramer Method

The Tukey–Kramer method¹⁸ is applicable when pairwise comparisons of population means are of primary interest; that is, null hypotheses of the form $H_0: \mu_i = \mu_j$ are to be considered. To use the Tukey–Kramer method, we compute the following confidence interval for the population mean difference ($\mu_i - \mu_j$):

$$(\bar{Y}_i - \bar{Y}_j) \pm \frac{q_{k, n-k, l-\alpha}}{\sqrt{2}} \sqrt{MSE \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \quad (17.21)$$

where $q_{k, n-k, l-\alpha}$ is the $100(1 - \alpha)\%$ point of the studentized range distribution,¹⁹ with k and $(n - k)$ degrees of freedom (see Table A.6 in Appendix A); k is the number of populations or groups; and n is the total number of observations.

¹⁸ The Tukey–Kramer method allows for unequal group sample sizes; it is the more general version of the method proposed by Tukey (1953), in which equal group sample sizes are assumed, and later extended by Kramer (1956).

¹⁹ The studentized range distribution with k and r degrees of freedom is the distribution of a statistic of the form R/S , where $R = \{\max_i(Y_i) - \min_i(Y_i)\}$ is the range of a set of k independent observations Y_1, Y_2, \dots, Y_k from a normal distribution with mean μ and variance σ^2 and where S^2 is an estimate of σ^2 based on r degrees of freedom (which is independent of the Y 's). In particular, when k means are being compared in fixed-effects one-way ANOVA, the statistic $\{\max_i(Y_i) - \min_i(Y_i)\}/\sqrt{MSE/n^*}$ (where n^* is the common sample size for each group) has the studentized range distribution with k and $(n - k)$ degrees of freedom under $H_0: \mu_1 = \mu_2 = \dots = \mu_k$, where $n_i = n^*$ for each i and where $n = kn^*$.

In the set of all kC_2 Tukey–Kramer pairwise confidence intervals of the form (17.21), when all of the group sample sizes are equal (that is, $n_i \equiv n^*$ for all $i = 1, 2, \dots, k$), the probability is $(1 - \alpha)$ that these intervals simultaneously contain the associated population mean differences that are being estimated; that is, $100(1 - \alpha)$ is the *overall confidence level* for all pairwise confidence intervals taken together. In particular, if each confidence interval is used to evaluate the corresponding pairwise null hypothesis of the general form $H_0: \mu_i = \mu_j$ by determining whether the value 0 is contained in the calculated interval, the probability of falsely rejecting the null hypothesis for *at least one* of the kC_2 comparisons is equal to α . When the group sample sizes are unequal or when fewer than all kC_2 pairwise comparisons are considered, the Tukey–Kramer method is conservative; that is, the overall significance level for the pairwise comparisons performed is less than α , and the overall confidence level is greater than $(1 - \alpha)$.

Let us now perform all pairwise comparisons of the group means for the potency data of Table 17.7 using the Tukey–Kramer method with an overall significance level of $\alpha = .05$. Since the value of $(q_{k,n-k,1-\alpha}/\sqrt{2})\sqrt{\text{MSE}(\frac{1}{n_i} + \frac{1}{n_j})}$ is needed for all of the six possible confidence intervals, we compute it first. We have $n_i = 10$ for all $i (= 1, 2, 3, 4)$, $k = 4$, $n = 40$, $\text{MSE} = 9.747$, and $q_{k,n-k,1-\alpha} = 3.81$. (This value was obtained from Table A.6 by interpolation.) Therefore,

$$\frac{q_{k,n-k,1-\alpha}}{\sqrt{2}} \sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} = \frac{3.81}{\sqrt{2}} \sqrt{9.747\left(\frac{1}{10} + \frac{1}{10}\right)} = 3.762$$

Thus, the pairwise Tukey–Kramer confidence intervals are

$$\mu_1 - \mu_4: 6.3 \pm 3.762; \text{i.e., } (2.538, 10.062)^*$$

$$\mu_1 - \mu_3: 5.9 \pm 3.762; \text{i.e., } (2.138, 9.662)^*$$

$$\mu_1 - \mu_2: 3.7 \pm 3.762; \text{i.e., } (-0.062, 7.462)$$

$$\mu_2 - \mu_4: 2.6 \pm 3.762; \text{i.e., } (-1.162, 6.362)$$

$$\mu_2 - \mu_3: 2.2 \pm 3.762; \text{i.e., } (-1.562, 5.962)$$

$$\mu_3 - \mu_4: 0.4 \pm 3.762; \text{i.e., } (-3.362, 4.162)$$

Because the first confidence interval, comparing sample means for groups 1 and 4, does not contain 0, we have statistical evidence (based on an overall significance level of $\alpha = .05$) that $\mu_1 \neq \mu_4$. Similarly, based on the second confidence interval, we conclude $\mu_1 \neq \mu_3$. All of the other intervals contain 0; therefore, all other pairwise mean difference comparisons are nonsignificant. (Note that, once again, the intervals were sorted in descending order based on the values of the sample mean differences; therefore, once the third pairwise mean difference in the list was determined to be nonsignificant, all remaining differences were also nonsignificant.)

Like the Bonferroni method, the Tukey–Kramer method leaves some ambiguity as to results: substance 2 has again been associated with substance 1 and also with substances 3 and 4. Such ambiguity is not uncommon. In this instance, it suggests that the amount of data being collected was insufficient to permit a clear characterization of substance 2.

Note that the Tukey–Kramer confidence intervals are narrower than the Bonferroni confidence intervals for the potency data example. Indeed, when only all possible pairwise comparisons are being investigated and group sample sizes are equal, the Tukey–Kramer approach guarantees an overall significance level of α , whereas the Bonferroni approach guarantees an overall significance level $\leq \alpha$; in other words, for this situation, the Bonferroni approach usually will be less powerful than the Tukey–Kramer approach and will, at best, be equally as powerful. This usually will also be true when the group sample sizes are unequal. Therefore, it is recommended that, when only all possible pairwise comparisons are being made, the Tukey–Kramer approach be used rather than the Bonferroni approach.

In the SAS output below, sample means that are similar (i.e., not significantly different) using the Tukey–Kramer method at $\alpha = .05$ are labeled with the same letter.

Edited SAS Output (PROC GLM) for Data of Table 17.7 Using the Tukey–Kramer Approach

Tukey's Studentized Range (HSD) Test for DOSAGE

Alpha				0.05
Error Degrees of Freedom				36
Error Mean Square				9.747222
Critical Value of Studentized Range				3.80880
Minimum Significant Difference				3.7604
MEANS WITH THE SAME LETTER ARE NOT SIGNIFICANTLY DIFFERENT.				
Tukey Grouping	Mean	N	SUBSTANCE	
	A	25.900	10	1
	A			
B	A	22.200	10	2
B				
B		20.000	10	3
B				
B		19.600	10	4

17.7.3 Scheffé's Method and Comparisons of Means Using Contrasts

Scheffé's method is generally recommended when comparisons other than simple pairwise differences between means are of interest and when these more general comparisons are not planned *a priori* (i.e., are, instead, suggested by observed data patterns during the course of the analysis).

These more general comparisons are referred to as *contrasts*, first introduced in Section 9.6. To illustrate contrasts in an ANOVA context, suppose that the investigator who collected the potency data of Table 17.7 suspected that substances 1 and 3 had similar potencies, that substances 2 and 4 had similar potencies, and that the

potencies of 1 and 3, on average, differed significantly from those of 2 and 4. Then it would be of interest to compare the average results obtained for 1 and 3 with the average results for 2 and 4—namely, to assess whether $(\mu_1 + \mu_3)/2$ really differs from $(\mu_2 + \mu_4)/2$. In other words, we could consider the contrast

$$L_1 = \frac{\mu_1 + \mu_3}{2} - \frac{\mu_2 + \mu_4}{2}$$

which would be 0 if the null hypothesis $H_0: (\mu_1 + \mu_3)/2 = (\mu_2 + \mu_4)/2$ were true. We can rewrite L_1 as follows:

$$L_1 = \frac{\mu_1 + \mu_3}{2} - \frac{\mu_2 + \mu_4}{2} = \frac{1}{2}\mu_1 - \frac{1}{2}\mu_2 + \frac{1}{2}\mu_3 - \frac{1}{2}\mu_4$$

or

$$L_1 = \sum_{i=1}^4 c_{1i}\mu_i$$

so that L_1 is a *linear* function of the population means, with $c_{11} = \frac{1}{2}$, $c_{12} = -\frac{1}{2}$, $c_{13} = \frac{1}{2}$, and $c_{14} = -\frac{1}{2}$. Further,

$$c_{11} + c_{12} + c_{13} + c_{14} = \frac{1}{2} - \frac{1}{2} + \frac{1}{2} - \frac{1}{2} = 0$$

In general, a *contrast* is defined as any linear function of the population means—say,

$$L = \sum_{i=1}^k c_i\mu_i$$

such that

$$\sum_{i=1}^k c_i = 0$$

The associated null hypothesis is

$$H_0: \sum_{i=1}^k c_i\mu_i = 0$$

and the two-sided alternative hypothesis is

$$H_A: \sum_{i=1}^k c_i\mu_i \neq 0$$

The data in Table 17.7 suggest that the mean potency of substance 1 is definitely higher than the mean potencies of the other three substances. Such an observation invites a comparison of the mean potency of substance 1 with the average potency of substances 2, 3, and 4. In this case, the appropriate contrast to consider is

$$L_2 = \mu_1 - \frac{\mu_2 + \mu_3 + \mu_4}{3}$$

or

$$L_2 = \sum_{i=1}^4 c_{2i} \mu_i$$

where $c_{21} = 1$, $c_{22} = c_{23} = c_{24} = -\frac{1}{3}$. (Again, $\sum_{i=1}^4 c_{2i} = 0$.)

A third possible contrast of interest involves a comparison of the average of the means for substances 1 and 2 with those for 3 and 4. The contrast here is

$$L_3 = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2} = \sum_{i=1}^4 c_{3i} \mu_i$$

where $c_{31} = c_{32} = \frac{1}{2}$ and $c_{33} = c_{34} = -\frac{1}{2}$.

Finally, any pairwise comparison is also a contrast. For example, a comparison of μ_1 with μ_4 takes the form

$$L_4 = \mu_1 - \mu_4 = \sum_{i=1}^4 c_{4i} \mu_i$$

where $c_{41} = 1$, $c_{42} = c_{43} = 0$, and $c_{44} = -1$.

Scheffé's method provides a family of confidence intervals for evaluating *all possible contrasts* that can be defined, given a fixed number k of population means, such that the overall confidence coefficient associated with the entire family is $(1 - \alpha)$, where α is specified by the investigator. In other words, the probability is $(1 - \alpha)$ that these confidence intervals simultaneously contain the true values of all the contrasts being considered. Equivalently, the overall significance level is α for testing hypotheses of the general form $H_0: L = \sum_{i=1}^k c_i \mu_i = 0$ concerning all possible contrasts; that is, the probability is α that at least one such null hypothesis will falsely be rejected.

The general form of a Scheffé-type confidence interval is as follows. Let $L = \sum_{i=1}^k c_i \mu_i$ be some contrast of interest. Then the appropriate confidence interval concerning L is given by

$$\sum_{i=1}^k c_i \bar{Y}_i \pm S \sqrt{\text{MSE} \left(\sum_{i=1}^k \frac{c_i^2}{n_i} \right)} \quad (17.22)$$

where $\hat{L} = \sum_{i=1}^k c_i \bar{Y}_i$ is the unbiased point estimator of L and where $S^2 = (k-1)F_{k-1, n-k, 1-\alpha}$, with $n = \sum_{i=1}^k n_i$.

As a special case, when only pairwise comparisons are of interest, this formula simplifies to

$$(\bar{Y}_i - \bar{Y}_j) \pm S \sqrt{\text{MSE} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \quad (17.23)$$

when $(\mu_i - \mu_j)$ is the parameter of interest.

If the investigator is interested only in pairwise comparisons, Scheffé's method using (17.23) is not recommended; in such cases, the Tukey–Kramer method will always provide narrower confidence intervals (i.e., will give more precise estimates of the true pairwise

differences). However, Scheffé's method does have a desirable property: whenever the overall F test of the null hypothesis that all k population means are equal is rejected, at least one estimated contrast will be found that differs significantly from 0. In contrast, the Tukey–Kramer method may not turn up any significant pairwise comparisons, even when the overall F statistic is significant.

To illustrate Scheffé's method, let us first consider all pairwise comparisons for the data of Table 17.7. We begin by first computing the quantity

$$S\sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

which will have the same value for all pairwise comparisons, since all the sample sizes are equal to 10. So

$$\sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} = \sqrt{9.747\left(\frac{1}{10} + \frac{1}{10}\right)} = 1.3962$$

And with $k = 4$, $n = 40$, and $\alpha = .05$, we have

$$S = \sqrt{(k - 1)F_{k-1, n-k, 1-\alpha}} = \sqrt{3F_{3, 36, 0.95}} = \sqrt{3(2.866)} = 2.9322$$

Therefore, $S\sqrt{\text{MSE}(1/n_i + 1/n_j)} = 2.9322(1.3962) = 4.094$
Thus, the pairwise Scheffé's method confidence intervals are

- $\mu_1 - \mu_4: 6.3 \pm 4.094$; i.e., $(2.206, 10.394)^*$
- $\mu_1 - \mu_3: 5.9 \pm 4.094$; i.e., $(1.806, 9.994)^*$
- $\mu_1 - \mu_2: 3.7 \pm 4.094$; i.e., $(-0.394, 7.794)$
- $\mu_2 - \mu_4: 2.6 \pm 4.094$; i.e., $(-1.494, 6.694)$
- $\mu_2 - \mu_3: 2.2 \pm 4.094$; i.e., $(-1.894, 6.294)$
- $\mu_3 - \mu_4: 0.4 \pm 4.094$; i.e., $(-3.694, 4.494)$

The conclusions using these Scheffé confidence intervals are similar to those using the Bonferroni and Tukey–Kramer methods (see Table 17.9). However, the Scheffé intervals are widest, underscoring the point that this method should not be used when performing only all possible pairwise comparisons.

The results based on Scheffé's method are illustrated in the SAS output below.

Edited SAS Output (PROC GLM) for Data of Table 17.7 Using Scheffé's Approach

Scheffé's Test for DOSAGE

Alpha	0.05
Error Degrees of Freedom	36
Error Mean Square	9.747222
Critical Value of F	2.86627
Minimum Significant Difference	4.0942

(continued)

MEANS WITH THE SAME LETTER ARE NOT SIGNIFICANTLY DIFFERENT.				
Scheffe Grouping		Mean	N	Substance
	A	25.900	10	1
	A			
B	A	22.200	10	2
B				
B		20.000	10	3
B				
B		19.600	10	4

TABLE 17.9 Comparison of some Tukey–Kramer and Scheffé confidence intervals for the potency data of Table 17.7

Pairwise Comparison	Tukey–Kramer		Scheffé	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit
$\mu_1 - \mu_4$	2.538	10.062	2.206	10.394
$\mu_1 - \mu_3$	2.138	9.662	1.806	9.999
$\mu_1 - \mu_2$	-0.062	7.462	-0.394	7.774
$\mu_2 - \mu_4$	-1.162	6.362	-1.494	6.694
$\mu_2 - \mu_3$	-1.562	5.962	-1.894	6.294
$\mu_3 - \mu_4$	-3.362	4.162	-3.694	4.494

© Cengage Learning

Let us now use Scheffé's method to make inferences regarding two contrasts of interest mentioned earlier. In particular, let us consider

$$L_2 = \mu_1 - \frac{\mu_2 + \mu_3 + \mu_4}{3} = \sum_{i=1}^4 c_{2i} \mu_i$$

where $c_{21} = 1$ and $c_{22} = c_{23} = c_{24} = -\frac{1}{3}$, and

$$L_3 = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2} = \sum_{i=1}^4 c_{3i} \mu_i$$

where $c_{31} = c_{32} = \frac{1}{2}$ and $c_{33} = c_{34} = -\frac{1}{2}$.

Using our previously computed value $S = \sqrt{(k - 1)F_{k-1, n-k, 1-\alpha}} = 2.9322$, we calculate using (17.22) as follows:

$$\begin{aligned} & \left(\bar{Y}_{1.} - \frac{\bar{Y}_{2.} + \bar{Y}_{3.} + \bar{Y}_{4.}}{3} \right) \pm S \sqrt{\text{MSE} \left[\frac{(1)^2}{10} + \frac{(-1/3)^2}{10} + \frac{(-1/3)^2}{10} + \frac{(-1/3)^2}{10} \right]} \\ & \left(25.9 - \frac{22.2 + 20.0 + 19.6}{3} \right) \pm 2.9322 \sqrt{9.747 \left(\frac{12}{90} \right)} \\ & (25.9 - 20.6) \pm 2.9322(1.1400) \\ & 5.3 \pm 3.3427 \end{aligned}$$

or

$$(1.9573, 8.6427)$$

Since this interval does not contain the value 0, we have evidence that the average potency of substance 1 differs from the average potency of substances 2, 3, and 4.

Next, we calculate the following Schéffé interval regarding L_3 :

$$\begin{aligned} & \left(\frac{\bar{Y}_{1.} + \bar{Y}_{2.}}{2} - \frac{\bar{Y}_{3.} + \bar{Y}_{4.}}{2} \right) \pm S \sqrt{\text{MSE} \left[\frac{(1/2)^2}{10} + \frac{(1/2)^2}{10} + \frac{(-1/2)^2}{10} + \frac{(-1/2)^2}{10} \right]} \\ & \left(\frac{25.9 + 22.2}{2} - \frac{20.0 + 19.6}{2} \right) \pm 2.9322 \sqrt{9.747 \left(\frac{1}{10} \right)} \\ & (24.05 - 19.80) \pm 2.9322(0.9873) \\ & 4.25 \pm 2.8949 \end{aligned}$$

or

$$(1.3551, 7.1449)$$

Because this interval also does not contain the value 0, we conclude that the average potency of substances 1 and 2 differs from the average potency of substances 3 and 4.

How do these results about the contrasts L_2 and L_3 help clear up the ambiguity created by considering all pairwise comparisons? First, uncertainty regarding substance 2 remains. Nevertheless, the confidence interval for comparing 1 with the average of 2, 3, and 4 (i.e., 1.9573 to 8.6427) is farther away from 0 than is the confidence interval for comparing the average of 1 and 2 with the average of 3 and 4 (i.e., 1.3551 to 7.1449); this suggests that substance 2 is closer to 3 and 4 in potency than it is to substance 1. The pairwise comparisons also support this contention, since the confidence interval for $(\mu_1 - \mu_2)$ is farther from 0 than is the interval for $(\mu_2 - \mu_3)$, again indicating that 2 is closer to 3 than to 1. Thus, if a definite decision regarding substance 2 had to be made on the basis of this set of data, the most logical thing to do would be to consider the potency of 1 as distinct from the potencies of 2, 3, and 4, which, as a threesome, are too similar to separate. Schematically, this conclusion is represented in Figure 17.3, which differs from Figure 17.2.

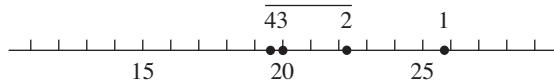


FIGURE 17.3 Conclusion regarding sample means for potency data (© Cengage Learning)

17.8 Choosing a Multiple-comparison Technique

Figure 17.4 summarizes a strategy for choosing a multiple-comparison technique for ANOVA. The first choice is between pairwise and nonpairwise comparisons. If only pairwise comparisons are being considered, the Tukey–Kramer method is preferable.

If any nonpairwise comparisons are to be considered, a choice must be made between planned (a priori) and unplanned (a posteriori) comparisons. The Bonferroni method should be used for planned comparisons; unplanned comparisons should be evaluated using the Scheffé method.

Although we have discussed only the Bonferroni, Tukey–Kramer, and Scheffé methods, a number of other multiple-comparison techniques have been suggested in the statistical literature. Miller (1981) provides a comprehensive review of these procedures. The methods differ with regard to the target set of contrasts (e.g., pairwise versus nonpairwise), the cell-specific sample sizes, and other properties. The developers of these methods have sought to minimize the Type II error rate (i.e., to maximize power; see Chapter 3) for a particular set of comparisons, while controlling the Type I error rate.

The reasons for considering other multiple-comparison methods are to improve power and to achieve robustness. A multiple-comparison procedure is robust if a violation of its underlying assumptions (such as homogeneity of variance) does not seriously affect the validity of an analysis. If a robust procedure is used, even when some assumptions are violated,

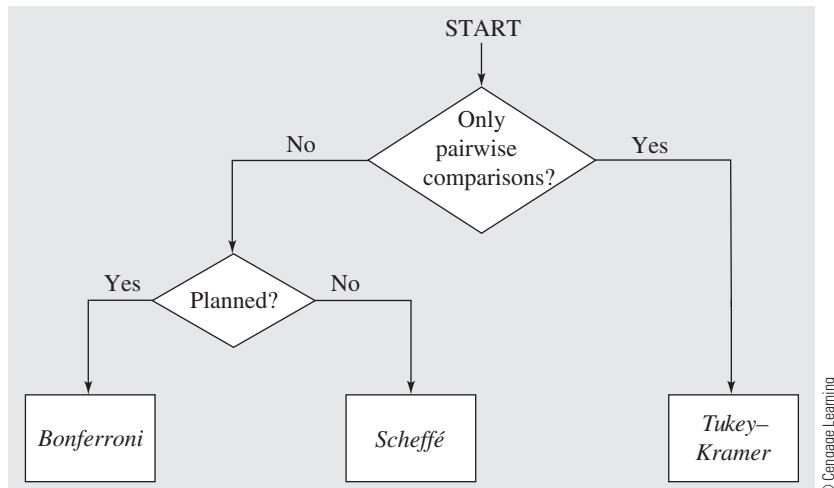


FIGURE 17.4 Recommended procedure for choosing a multiple-comparison technique in ANOVA

Type I and Type II error rates are unlikely to be seriously compromised. Both the Bonferroni and Scheffé methods tend to be fairly robust; the Tukey–Kramer method is less so. Scheffé's method is always the least powerful, since all possible comparisons are considered. The Bonferroni method is generally presumed to be insensitive (i.e., to have low power), but this presumption can be somewhat discounted if the sets of comparisons are well planned.

Although many acceptable multiple-comparison procedures are available, most are limited to a rather narrow range of application (e.g., the Tukey–Kramer method is limited to pairwise contrasts). Both the Bonferroni and the Scheffé procedures are completely general methods—the former for planned (a priori) and the latter for unplanned (a posteriori) multiple comparisons.

17.9 Orthogonal Contrasts and Partitioning an ANOVA Sum of Squares

Our previous discussions of multiple regression have touched on the notion of a partitioned sum of squares in regression analysis, where the sum of squares due to regression (SSR) is broken down into various components reflecting the relative contributions of various terms in the fitted model. In an ANOVA framework, it is possible (via the use of orthogonal contrasts) to additionally partition the treatment sum of squares SST into meaningful components associated with certain specific comparisons of particular interest. To illustrate how such a partitioning can be accomplished, we must discuss two new concepts: *orthogonal contrasts* and the *sum of squares associated with a contrast*.

In the notation of Section 17.7, two estimated contrasts

$$\hat{L}_A = \sum_{i=1}^k c_{Ai} \bar{Y}_i \quad \text{and} \quad \hat{L}_B = \sum_{i=1}^k c_{Bi} \bar{Y}_i$$

are *orthogonal* to each other (i.e., are orthogonal contrasts) if

$$\sum_{i=1}^k \frac{c_{Ai} c_{Bi}}{n_i} = 0 \quad (17.24)$$

In the special case where the n_i 's are equal, then equation (17.24) reduces to the condition

$$\sum_{i=1}^k c_{Ai} c_{Bi} = 0 \quad (17.25)$$

Consider the three estimated contrasts discussed earlier with regard to the potency data of Table 17.7:

$$\hat{L}_1 = \frac{\bar{Y}_{1\cdot} + \bar{Y}_{3\cdot}}{2} - \frac{\bar{Y}_{2\cdot} + \bar{Y}_{4\cdot}}{2} = \frac{1}{2} \bar{Y}_{1\cdot} - \frac{1}{2} \bar{Y}_{2\cdot} + \frac{1}{2} \bar{Y}_{3\cdot} - \frac{1}{2} \bar{Y}_{4\cdot}$$

where $c_{11} = c_{13} = \frac{1}{2}$ and $c_{12} = c_{14} = -\frac{1}{2}$;

$$\hat{L}_2 = \bar{Y}_{1\cdot} - \frac{1}{3}(\bar{Y}_{2\cdot} + \bar{Y}_{3\cdot} + \bar{Y}_{4\cdot}) = \bar{Y}_{1\cdot} - \frac{1}{3}\bar{Y}_{2\cdot} - \frac{1}{3}\bar{Y}_{3\cdot} - \frac{1}{3}\bar{Y}_{4\cdot}$$

where $c_{21} = 1$ and $c_{22} = c_{23} = c_{24} = -\frac{1}{3}$; and

$$\hat{L}_3 = \frac{\bar{Y}_{1\cdot} + \bar{Y}_{2\cdot}}{2} - \frac{\bar{Y}_{3\cdot} + \bar{Y}_{4\cdot}}{2} = \frac{1}{2}\bar{Y}_{1\cdot} + \frac{1}{2}\bar{Y}_{2\cdot} - \frac{1}{2}\bar{Y}_{3\cdot} - \frac{1}{2}\bar{Y}_{4\cdot}$$

where $c_{31} = c_{32} = \frac{1}{2}$ and $c_{33} = c_{34} = -\frac{1}{2}$. Since $n_i = 10$ for every i , we need only verify that condition (17.25) holds to demonstrate orthogonality. In particular, for \hat{L}_1 and \hat{L}_2 , we have

$$\sum_{i=1}^4 c_{1i}c_{2i} = \left(\frac{1}{2}\right)(1) + \left(-\frac{1}{2}\right)\left(-\frac{1}{3}\right) + \left(\frac{1}{2}\right)\left(-\frac{1}{3}\right) + \left(-\frac{1}{2}\right)\left(-\frac{1}{3}\right) = \frac{2}{3} \neq 0$$

For \hat{L}_1 and \hat{L}_3 , we have

$$\sum_{i=1}^4 c_{1i}c_{3i} = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) + \left(-\frac{1}{2}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)\left(-\frac{1}{2}\right) + \left(-\frac{1}{2}\right)\left(-\frac{1}{2}\right) = 0$$

And, for \hat{L}_2 and \hat{L}_3 , we have

$$\sum_{i=1}^4 c_{2i}c_{3i} = (1)\left(\frac{1}{2}\right) + \left(-\frac{1}{3}\right)\left(\frac{1}{2}\right) + \left(-\frac{1}{3}\right)\left(-\frac{1}{2}\right) + \left(-\frac{1}{3}\right)\left(-\frac{1}{2}\right) = \frac{2}{3} \neq 0$$

Thus, we conclude that \hat{L}_1 and \hat{L}_3 are orthogonal to one another but that neither is orthogonal to \hat{L}_2 .

Orthogonality is a desirable property for several reasons. Suppose that SST denotes the treatment sum of squares with $(k - 1)$ degrees of freedom for a fixed-effects one-way ANOVA, and suppose that $\hat{L}_1, \hat{L}_2, \dots, \hat{L}_t$ are a set of t ($\leq k - 1$) mutually orthogonal contrasts of the k sample means (*mutually orthogonal* means that any two contrasts selected from the set of t contrasts are orthogonal to one another). Then SST can be partitioned into $(t + 1)$ statistically independent sums of squares, with t of these sums of squares having 1 degree of freedom each and being associated with the t orthogonal contrasts and with the remaining sum of squares having $(k - t - 1)$ degrees of freedom and being associated with what remains after the t orthogonal contrast sums of squares have been accounted for. In other words, we can write

$$\text{SST} = \text{SS}(\hat{L}_1) + \text{SS}(\hat{L}_2) + \cdots + \text{SS}(\hat{L}_t) + \text{SS}(\text{remainder})$$

where $\text{SS}(\hat{L})$ is the notation for the sum of squares (with 1 degree of freedom) associated with the contrast \hat{L} . In particular, it can be shown that

$$\text{SS}(\hat{L}) = \frac{(\hat{L})^2}{\sum_{i=1}^k (c_i^2/n_i)} \quad \text{when} \quad \hat{L} = \sum_{i=1}^k c_i \bar{Y}_{i\cdot} \quad (17.26)$$

that

$$\frac{SS(\hat{L})}{MSE} \sim F_{1, n-k}$$

under $H_0: L = \sum_{i=1}^k c_i \mu_i = 0$, and, in general, that

$$\frac{[SS(\hat{L}_1) + SS(\hat{L}_2) + \dots + SS(\hat{L}_t)]/t}{MSE} \sim F_{t, n-k}$$

under $H_0: L_1 = L_2 = \dots = L_t = 0$ when $\hat{L}_1, \hat{L}_2, \dots, \hat{L}_t$ are mutually orthogonal. Thus, by partitioning SST as described, we can test hypotheses concerning sets of orthogonal contrasts that are of more specific interest than the global hypothesis $H_0: \mu_1 = \mu_2 = \dots = \mu_k$.

For example, to test $H_0: L_2 = \mu_1 - \frac{1}{3}(\mu_2 + \mu_3 + \mu_4) = 0$ for the potency data in Table 17.7, we first use (17.26) to calculate

$$SS(\hat{L}_2) = \frac{(5.30)^2}{[(1)^2 + (-1/3)^2 + (-1/3)^2 + (-1/3)^2]/10} = 210.675$$

and then we form the ratio

$$\frac{SS(\hat{L}_2)}{MSE} = \frac{210.675}{9.747} = 21.614$$

which is highly significant ($P < .001$ based on the $F_{1, 36}$ distribution). Modifying Table 17.8 to reflect this partitioning of SST (the sum of squares for “substances”) yields

Source	d.f.	SS	MS	F
Substances 1 vs. (2, 3, 4)	1	210.675	210.675	21.614 ($P < .001$)
	2	39.200	irrelevant	
Error	36	350.900	9.747	
Total	39	600.775		

Similarly, the partitioned ANOVA table for testing

$$H_0: L_3 = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2} = 0$$

is as follows:

Source	d.f.	SS	MS	F
Substances (1, 2) vs. (3, 4)	1	180.625	180.625	18.531 ($P < .001$)
	2	69.25	irrelevant	
Error	36	350.900	9.747	
Total	39	600.775		

This partition follows from the fact that

$$\text{SS}(\hat{L}_3) = \frac{(4.25)^2}{[(1/2)^2 + (1/2)^2 + (-1/2)^2 + (-1/2)^2]/10} = 180.625$$

Finally, since \hat{L}_1 and \hat{L}_3 are orthogonal, we can represent the independent contributions of these two contrasts to SST in one ANOVA table:

Source	d.f.	SS	MS	F
Substances	(1, 3) vs. (2, 4)	1	42.025	42.025
	(1, 2) vs. (3, 4)	1	180.625	180.625
	Remainder	1	27.225	irrelevant
Error	36	350.900	9.747	—
Total	39	600.775		

Here

$$\text{SS}(\hat{L}_1) = \frac{\left(\frac{25.9 + 20.0}{2} - \frac{22.2 + 19.6}{2} \right)^2}{[(1/2)^2 + (-1/2)^2 + (1/2)^2 + (-1/2)^2]/10} = 42.025$$

It would not be valid to present a partitioned ANOVA table that simultaneously included partitions due to \hat{L}_2 , as well as \hat{L}_1 and/or \hat{L}_3 . This is because \hat{L}_2 is not orthogonal to these contrasts, and so its sum of squares does not represent a separate and independent contribution to SST; this can easily be confirmed by observing from the preceding tables that

$$\text{SST} = 249.875 \neq \text{SS}(\hat{L}_1) + \text{SS}(\hat{L}_2) + \text{SS}(\hat{L}_3) = 42.025 + 210.675 + 180.625$$

A final example of using orthogonal contrasts involves the need to assess whether the sample means exhibit a trend of some sort; this need often arises when the treatments (or populations) being studied represent, for example, different levels of the same factor (e.g., different concentrations of the same material or different temperature or pressure settings). In such situations, it is of interest to quantify how the sample means vary with changes in the level of the factor—that is, to clarify whether the change in mean response takes place in a linear, quadratic, or other way as the level of the factor increases or decreases.

A qualitative first step in assessing such a trend is to plot the observed treatment means as a function of the factor levels. This may yield some general idea of the pattern (if any) present. Standard regression techniques can then be used to quantify any trends suggested by such plots, but our goal here is not to fit a regression model; rather, it is to evaluate a possible general trend in the sample means statistically, instead of forming an opinion on the basis of a simple examination of a plot of these means.

In a standard regression approach, the independent variable (“treatments”) could possibly be considered as an interval variable, and the actual value of the variable at each treatment setting could be used. To test for a linear trend by using regression, we would apply the model $Y = \beta_0 + \beta_1 X + E$, where X denotes the (interval) treatment variable. With this model, the test for linear trend is the usual test for zero slope. In general, this test is not

equivalent to the test for a linear trend in *mean* response using the orthogonal polynomials described in this section, except when the pure error mean square is used in place of the residual mean square in the denominator of the *F* statistic to test for zero slope (because then the regression pure error mean square and the one-way ANOVA error mean square are identical). The usual test for lack of fit of a straight-line regression model is equivalent to the test for a nonlinear trend in mean response discussed in this section. (See Problem 10 at the end of this chapter for more discussion about this issue.)

A statistical trend analysis may be carried out by determining how much of the sum of squares due to treatments (SST) is associated with each of the terms (linear, quadratic, cubic, etc.) in a polynomial regression. If the various levels of the treatment or factor being studied are equally spaced, this determination is best carried out by using the method of orthogonal polynomials. (For a discussion, see Armitage [1971] and also Chapter 15.)

To illustrate the use of orthogonal polynomials, let us again turn to the potency data of Table 17.7. Further, let us suppose that the four substances actually represent four equally spaced concentrations of some toxic material, with substance 1 the least concentrated solution and substance 4 the most concentrated. For example, substance 1 might represent a 10% solution of the toxic material, substance 2 a 20% solution, substance 3 a 30% solution, and substance 4 a 40% solution. In this case, a plot of the four sample means versus concentration takes the form shown in Figure 17.5. This plot suggests at least a linear and possibly a quadratic relationship between concentration and response, and this general impression can be quantified via the use of orthogonal polynomials. In particular, given $k = 4$ sample means, it is possible to fit up to a third-order ($k - 1 = 3$) polynomial to these means; the cubic model (with four terms) would pass through all four points on the Figure 17.5 graph, thus explaining all the variation in the four sample means (or, equivalently, in SST). Because the concentrations are equally spaced, it is possible via the use of orthogonal polynomials to define three orthogonal contrasts of the four sample means—one (say, \hat{L}_l) measuring the strength of the linear component of the third-degree polynomial, one (say, \hat{L}_q) the quadratic component contribution, and one (say, \hat{L}_c) the cubic component effect. The sums of squares associated with these three orthogonal contrasts each have 1 degree of freedom, are statistically independent, and satisfy the relationship

$$\text{SST} = \text{SS}(\hat{L}_l) + \text{SS}(\hat{L}_q) + \text{SS}(\hat{L}_c)$$

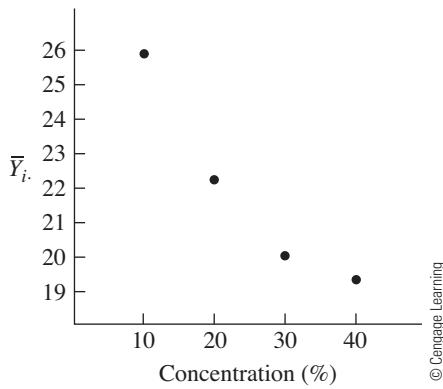


FIGURE 17.5 Plot of the sample means versus concentration

It can be shown that these three orthogonal contrasts have the following coefficients:

Contrast	Coefficient of				Calculated Value of Contrast
	$\bar{Y}_1 = 25.9$	$\bar{Y}_2 = 22.2$	$\bar{Y}_3 = 20.0$	$\bar{Y}_4 = 19.6$	
\hat{L}_l	-3	-1	1	3	-21.1
\hat{L}_q	1	-1	-1	1	3.3
\hat{L}_c	-1	3	-3	1	0.3

Condition (17.25) clearly holds for these three sets of coefficients, which is sufficient to establish mutual orthogonality here, since the n_i 's are all equal. These coefficients were taken from Table A.7 in Appendix A. The table may be used only for equally spaced treatment values and equal cell-specific sample sizes. Kirk (1969) summarizes an algorithm for the general case that does not require these two restrictions. More conveniently, computer software can be used to obtain orthogonal contrast coefficients.

Now, from (17.26), the sums of squares for these three particular contrasts are

$$SS(\hat{L}_l) = \frac{(-21.1)^2}{[(-3)^2 + (-1)^2 + (1)^2 + (3)^2]/10} = 222.605$$

$$SS(\hat{L}_q) = \frac{(3.3)^2}{[(1)^2 + (-1)^2 + (-1)^2 + (1)^2]/10} = 27.225$$

$$SS(\hat{L}_c) = \frac{(0.3)^2}{[(-1)^2 + (3)^2 + (-3)^2 + (1)^2]/10} = 0.045$$

Notice that

$$SST = 249.875 = 222.605 + 27.225 + 0.045$$

Finally, to assess the significance of these sums of squares, we form the following partitioned ANOVA table:

Source	d.f.	SS	MS	F
Substances	\hat{L}_l	222.605	222.605	22.838 ($P < .001$)
	\hat{L}_q	27.225	27.225	2.793 ($.1 < P < .25$)
	\hat{L}_c	0.045	0.045	< 1 (n.s.)
Error	36	350.900	9.747	
Total	39	600.775		

It is clear from the preceding F tests that the relationship between potency (as measured by the amount of injected material needed to cause death) and concentration is strongly linear, with no evidence of higher-order effects.

Problems

1. Five treatments for fever blisters, including a placebo, were randomly assigned to 30 patients. For each of the five treatments, the data in the accompanying table identify the number of days from initial appearance of the blisters until healing is complete.

Treatment	No. of Days
Placebo (1)	5, 8, 7, 7, 10, 8
2	4, 6, 6, 3, 5, 6
3	6, 4, 4, 5, 4, 3
4	7, 4, 6, 6, 3, 5
5	9, 3, 5, 7, 7, 6

- a. Compute the sample means and the sample standard deviations for each treatment.
- b. Complete the ANOVA table in the accompanying computer output for the data given.
- c. Do the effects of the five treatments differ significantly with regard to healing fever blisters? In other words, test $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ against H_A : “At least two treatments have different population means.”
- d. What are the estimates of the true effects ($\mu_i - \mu$) of the treatments? Verify that the sum of these estimated effects is 0. (Note: μ_i is the population mean for the i th treatment, and $\mu = \frac{1}{5} \sum_{i=1}^5 \mu_i$ is the overall population mean.)
- e. Using dummy variables, create an appropriate regression model that describes this experiment. Give two possible ways of defining these dummy variables (one using 0's and 1's and the other using 1's and -1's), and describe for each coding scheme how the regression coefficients are related to the population means $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$, and μ .
- f. For the data of this problem, carry out the Scheffé, Tukey–Kramer, and Bonferroni multiple-comparison procedures for examining pairwise differences between means, as described in Section 17.7. Also, compare the widths of the confidence intervals obtained by the three procedures.

Edited SAS Output (PROC GLM) for Problem 1

Dependent Variable: days

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	—	—	—	—	0.0136
Error	25	58.50000000	2.34000000		
Corrected Total	29	—			
<hr/>					
R-Square	Coeff Var	Root MSE	days Mean		
0.383994	27.15454	1.529706	5.633333		

(continued)

Source	DF	Type I SS	Mean Square	F Value	Pr > F
treat	4	36.46666667	9.11666667	3.90	0.0136

Source	DF	Type III SS	Mean Square	F Value	Pr > F
treat	4	36.46666667	9.11666667	3.90	0.0136

Tukey's Studentized Range (HSD) Test for Days

Error Degrees of Freedom	25
Error Mean Square	2.34
Critical Value of Studentized Range	4.15336
Minimum Significant Difference	2.5938

MEANS WITH THE SAME LETTER ARE NOT SIGNIFICANTLY DIFFERENT.				
Tukey Grouping	Mean	N	treat	
A	7.5000	6	1	
A				
B	A	6.1667	6	5
B	A			
B	A	5.1667	6	4
B	A			
B	A	5.0000	6	2
B				
B		4.3333	6	3

Bonferroni (Dunn) t Tests for Days

Error Degrees of Freedom	25
Error Mean Square	2.34
Critical Value of t	3.07820
Minimum Significant Difference	2.7186

MEANS WITH THE SAME LETTER ARE NOT SIGNIFICANTLY DIFFERENT.				
Bon Grouping	Mean	N	treat	
A	7.5000	6	1	
A				
B	A	6.1667	6	5
B	A			
B	A	5.1667	6	4
B	A			
B	A	5.0000	6	2
B				
B		4.3333	6	3

(continued)

Scheffe's Test for Days

Error Degrees of Freedom	25
Error Mean Square	2.34
Critical Value of F	2.75871
Minimum Significant Difference	2.9338

MEANS WITH THE SAME LETTER ARE NOT SIGNIFICANTLY DIFFERENT.

Scheffe Grouping		Mean	N	treat
	A	7.5000	6	1
	A			
B	A	6.1667	6	5
B	A			
B	A	5.1667	6	4
B	A			
B	A	5.0000	6	2
B				
B		4.3333	6	3

2. The following data are replicate measurements of the sulfur dioxide concentration in each of three cities:

City I: 2, 1, 3

City II: 4, 6, 8

City III: 2, 5, 2

- a. Complete the ANOVA table in the accompanying computer output for simultaneously comparing the mean sulfur dioxide concentrations in the three cities.
- b. Test whether the three cities differ significantly in mean sulfur dioxide concentration levels.
- c. What is the estimated effect associated with each city?
- d. State precisely the appropriate ANOVA fixed-effects model for these data.
- e. Using dummy variables, state precisely the regression model that corresponds to the fixed-effects model in part (d). What is the relationship between the coefficients in the regression model and the effects in the ANOVA model?
- f. Using the t distribution, find a 90% confidence interval for the true difference between the effects of cities I and II (making sure to use the best estimate of σ^2 provided by the data).

Edited SAS Output (PROC GLM) for Problem 2

Dependent Variable: S02

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	26.00000000	13.00000000	_____	0.0553
Error	—	_____	_____	_____	_____
Corrected Total	8	_____	_____	_____	_____

R-Square	Coeff Var	Root MSE	s02 Mean
0.619048	44.53618	1.632993	3.666667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
city	2	26.00000000	13.00000000	4.88	0.0553

Source	DF	Type III SS	Mean Square	F Value	Pr > F
city	2	26.00000000	13.00000000	4.88	0.0553

3. Each of three chemical laboratories performed four replicate determinations of the concentration of suspended particulate matter in a certain area using the “Hi-Vol” method of analysis. The resulting data are presented next.

Lab 1	Lab II	Lab III
4	2	5
4	2	2
6	5	5
10	3	8

Edited SAS Output (PROC GLM) for Problem 3

Dependent Variable: conc

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	18.66666667	9.33333333	1.75	0.2280
Error	9	48.00000000	5.33333333	_____	_____
Corrected Total	11	66.66666667	_____	_____	_____

R-Square	Coeff Var	Root MSE	conc Mean
0.280000	49.48717	2.309401	4.666667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
lab	2	18.66666667	9.33333333	1.75	0.2280

Source	DF	Type III SS	Mean Square	F Value	Pr > F
lab	2	18.66666667	9.33333333	1.75	0.2280

- a. Identify the appropriate ANOVA table for these data.
 - b. Test the null hypothesis of no differences among the laboratories.
 - c. With large-scale interlaboratory studies, analysts usually make inferences about a large population of laboratories of which only a random sample (e.g., laboratories I, II, and III) can be investigated. In such a case, describe the appropriate random-effects model for the data.
 - d. Two quantities of particular interest in a large-scale interlaboratory study are repeatability (i.e., a measure of the variability among replicate measurements within a single laboratory) and reproducibility (i.e., a measure of the variability between results from different laboratories). Using the random-effects model defined in part (c), define what you think are reasonable measures of repeatability and reproducibility, and obtain estimates of the quantities you have defined using the data in the accompanying computer output.
4. Ten randomly selected mental institutions were examined to determine the effects of three different antipsychotic drugs on patients with the same types of symptoms. Each institution used one and only one of the three drugs exclusively for a one-year period. The proportion of treated patients in each institution who were discharged after one year of treatment is as follows for each drug used:
- | | |
|--------------------------------|------------------------------------|
| Drug 1: 0.10, 0.12, 0.08, 0.14 | $(\bar{Y}_1 = 0.11, S_1 = 0.0192)$ |
| Drug 2: 0.12, 0.14, 0.19 | $(\bar{Y}_2 = 0.15, S_2 = 0.0361)$ |
| Drug 3: 0.20, 0.25, 0.15 | $(\bar{Y}_3 = 0.20, S_3 = 0.0500)$ |
- a. Determine the appropriate ANOVA table for this data set using the accompanying computer output.

Edited SAS Output (PROC GLM) for Problem 4

Dependent Variable: *disch*

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.01389000	0.00694500	5.06	0.0436
Error	7	0.00960000	0.00137143		
Corrected Total	9	0.02349000			
R-Square		Coeff Var	Root MSE	disch Mean	
0.591315		24.85423	0.037033	0.149000	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
drug	2	0.01389000	0.00694500	5.06	0.0436
Source	DF	Type III SS	Mean Square	F Value	Pr > F
drug	2	0.01389000	0.00694500	5.06	0.0436

- b. Test to see whether significant differences exist among drugs with regard to the average proportion of patients discharged.
 - c. What other factors should be considered in comparing the effects of the three drugs?
 - d. What basic ANOVA assumptions might be violated here?
5. Suppose that a random sample of five active members in each of four political parties in a certain western European country was given a questionnaire purported to measure (on a 100-point scale) the extent of “general authoritarian attitude toward interpersonal relationships.” The means and standard deviations of the authoritarianism scores for each party are given in the following table.

	Party 1	Party 2	Party 3	Party 4
\bar{Y}_i	85	80	95	50
S_i	6	7	4	10
n_i	5	5	5	5

- a. Determine the appropriate ANOVA table for this data set.
 - b. Test to see whether significant differences exist among parties with respect to mean authoritarianism scores.
 - c. Using dummy variables, state an appropriate regression model for this experimental situation.
 - d. Apply the Tukey–Kramer method of multiple comparisons to identify the pairs in which the means significantly differ from one another. (Use $\alpha = .05$.)
6. A psychosociological questionnaire was administered to a random sample of 200 persons on an island in the South Pacific that has become increasingly westernized over the past 30 years. From the questionnaire data, each of the 200 persons was classified into one of three groups—HI-POS, NO-DIF, and HI-NEG—according to the discrepancy between the amount of prestige in that person’s traditional culture and the amount of prestige in the modern (westernized) culture. On the basis of the questionnaire data, a measure of anomie (i.e., social disorientation), denoted as Y , was determined on a 100-point scale, with the results summarized in the following table.

Group	n_i	\bar{Y}_i	S_i
HI-POS	50	65	9
NO-DIF	75	50	11
HI-NEG	75	55	10

- a. Determine the appropriate ANOVA table.
 - b. Test whether the three different categories of prestige discrepancy have significantly different sample mean anomie scores.
 - c. How would you test whether a significant difference exists between the NO-DIF category and the other two categories combined?
7. To examine whether mathematics skills of the average high school graduate in urban areas of the United States have declined over the past 10 years, the average scores on the math section of the SAT Reasoning Test (MSAT) were compared for a random sample of five big-city high schools for the years 2002, 2007, and 2012. The results are shown in the following table and accompanying computer output.

Year	School 1	School 2	School 3	School 4	School 5
2002	550	560	535	545	555
2007	545	560	528	532	541
2012	536	552	526	527	530

- Determine the sample means for each year.
- Comment on the independence assumption required for using one-way ANOVA on the data given.
- Determine the one-way ANOVA table for these data.
- Test by means of a one-way ANOVA whether any significant differences exist among the three MSAT average scores for the years 2002, 2007, and 2012.
- Use Scheffé's method to locate any significant differences between pairs of means. (Use $\alpha = .05$.)

Edited SAS Output (PROC GLM) for Problem 7

Dependent Variable: vsat

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	548.133333	274.066667	2.26	0.1466
Error	12	1453.600000	121.133333		
Corrected Total	14	2001.733333			

R-Square	Coeff Var	Root MSE	vsat Mean
0.273829	2.032638	11.00606	541.4667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
year	2	548.133333	274.066667	2.26	0.1466

Source	DF	Type III SS	Mean Square	F Value	Pr > F
year	2	548.133333	274.066667	2.26	0.1466

- Three persons (denoted A, B, and C) claiming to have unusual psychic ability underwent ESP tests at an eastern U.S. psychic research institute. On each of five randomly selected days, each person was asked to specify for 26 pairs of cards whether both cards in a given pair were of the same color or not. The numbers of correct answers are given in the accompanying table.

Person	Day 1	Day 2	Day 3	Day 4	Day 5
A	20	22	20	21	18
B	24	21	18	22	20
C	16	18	14	13	16

- Determine the mean score for each person, and interpret the results.
- Test whether the three persons have significantly different ESP ability.

- c. Carry out Scheffé's multiple-comparison procedure to determine which pairs of persons, if any, significantly differ in ESP ability.
 - d. On the basis of the results in parts (b) and (c), can one conclude that any of these persons has statistically significant ESP ability? Explain.
 - e. What basic ANOVA assumptions might be violated here?
9. The average generation times for four different strains of influenza virus were determined using six cultures for each strain. The data are summarized in the following table.

Statistic	Strain A	Strain B	Strain C	Strain D
\bar{Y}	420.3	330.7	540.4	450.8
S	30.22	28.90	31.08	33.29

- a. Test whether the true mean generation time differs among the four strains.
 - b. What is the appropriate ANOVA table for these data?
 - c. Use the Tukey-Kramer multiple-comparison procedure to identify where any differences occur among the means. (Use $\alpha = .05$.)
10. Three replicate water samples were taken at each of four locations in a river to determine whether the quantity of dissolved oxygen—a measure of water pollution—varied from one location to another (the higher the level of pollution, the lower the dissolved oxygen reading). Location 1 was adjacent to the wastewater discharge point for a certain industrial plant, and locations 2, 3, and 4 were selected at points 10, 20, and 30 miles downstream from this discharge point. The resulting data appear in the accompanying table. The quantity Y_{ij} denotes the value of the dissolved oxygen content for the j th replicate at location i ($j = 1, 2, 3$ and $i = 1, 2, 3, 4$), and \bar{Y}_i denotes the mean of the three replicates taken at location i .

Location	Dissolved Oxygen	
	Content (Y_{ij})	Mean (\bar{Y}_i)
1	4, 5, 6	5
2	6, 6, 6	6
3	7, 8, 9	8
4	8, 9, 10	9

- a. Do the data provide sufficient evidence to suggest that values for mean dissolved oxygen content differ significantly among the four locations? (Use $\alpha = .05$.) Make sure to construct the appropriate ANOVA table.
- b. Given that μ_i represents the true mean level of dissolved oxygen at location i ($i = 1, 2, 3, 4$), test the null hypothesis

$$H_0: -3\mu_1 - \mu_2 + \mu_3 + 3\mu_4 = 0$$

versus

$$H_A: -3\mu_1 - \mu_2 + \mu_3 + 3\mu_4 \neq 0$$

at the 2% level. The quantity $(-3\mu_1 - \mu_2 + \mu_3 + 3\mu_4)$ is a contrast based on orthogonal polynomials (see Section 17.9), which can be shown to be a measure of the linear relationship between "location" (the four equally spaced distances 0, 10, 20, and 30 miles downstream from the plant) and "dissolved oxygen content."

- c. Another way to quantify the strength of this linear relationship is to fit by least squares the model

$$Y = \beta_0 + \beta_1 X + E$$

where

$$X = \begin{cases} 0 & \text{for location 1} \\ 10 & \text{for location 2} \\ 20 & \text{for location 3} \\ 30 & \text{for location 4} \end{cases}$$

Fitting such a regression model to the $n = 12$ data points yields the accompanying ANOVA table. Use this table to perform a test of $H_0: \beta_1 = 0$ at the 2% significance level.

Source	d.f.	SS
Regression	1	29.40
Residual Lack of fit Pure error	10 2 8	6.60 0.60 6.00

- d. The regression model in part (c) amounts to saying that $\mu_i = \beta_0 + \beta_1 X_i$. Show that the hypothesis tested in part (b) is equivalent to the hypothesis tested in part (c).
- e. Why do the two test statistics calculated in parts (b) and (c) *not* have the same numerical value? What reasonable modification of the test in part (c) would yield the same F -value as that obtained in part (b)?
- f. Given the results of part (b), a test for a nonlinear trend in mean response can be obtained by subtracting the sum of squares for the contrast

$$\hat{L} = -3\bar{Y}_1 + \bar{Y}_2 + \bar{Y}_3 + 3\bar{Y}_4$$

from the sum of squares for treatments and then dividing this difference by the appropriate degrees of freedom to yield an F statistic of the form

$$F(\text{Nonlinear trend}) = \frac{[\text{SST} - \text{SS}(\hat{L})]/\text{df}}{\text{MSE}}$$

Carry out this test based on the results obtained in parts (a) and (b). (Use $\alpha = .05$.)

- g. Carry out the usual regression lack-of-fit test for adequacy of the straight-line model fit in part (c), using $\alpha = .05$. Does the value of the F statistic equal the value obtained in part (f)?
11. Consider the data from Problem 15 in Chapter 5. The dependent variable of interest is LN_BRNTL. Use $\alpha = .05$.
- a. Conduct a one-way ANOVA, with dosage level (PPM_TOLU) as a categorical predictor.

- b. Using the Bonferroni technique, compute all pairwise comparisons.
 - c. Repeat part (b) using the Tukey–Kramer method.
 - d. Repeat part (b) using Scheffé’s method.
 - e. The ANOVA overall test corresponds to fitting a polynomial model of order k . What is k ?
 - f. Examine the ANOVA assumptions by computing the estimates of the within-cell variances. Provide frequency histograms for each cell to aid in this examination.
12. Repeat Problem 11, skipping part (e), but use LN_BLDTL as the dependent variable.
13. The accompanying source table relates to a study involving the effects of trimethylin doses of 0, 3, 6, and 9 mg/kg on a sample of 48 rats. Each rat received only one dose. The response variable was the log of the activity counts after one hour spent in a residential maze. Compute the unknown values for a , b , c , d , e , f , and g .

Source	SS	d.f.	MS	F
Dosage	a	d	g	14.71
Within dosage	b	e	12.84	
Total	c	f		

14. For each of the following contrasts, indicate the null hypothesis that is being tested.

	$\bar{Y}_1.$	$\bar{Y}_2.$	$\bar{Y}_3.$	$\bar{Y}_4.$	$\bar{Y}_5.$
c_1	1	0	-1	0	0
c_2	0	$\frac{1}{2}$	0	$\frac{1}{2}$	-1
c_3	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	-1

15. Persistent patent ductus arteriosus (PDA) is a medical condition that affects more than 40% of very low birth weight infants (see Cotton et al. 1979). Infants with PDA have an increased risk of ailments such as cerebral hemorrhage and bronchopulmonary dysplasia, as well as of death. In a study, Varvarigou et al. (1996) determined that early ibuprofen administration is effective in preventing PDA in preterm neonates. In a sample of 34 such infants, 11 were treated with one dose of ibuprofen, 12 with three doses, and 11 with a saline solution, all treatments beginning within three hours of birth.

As part of the study, clinical factors such as birth weight, gestational age, and one-minute Apgar scores were compared for the three groups. Summary statistics for these variables are presented in the following table.

Treatment Group	Mean Birth Weight (grams) (SD in parentheses)	Mean Gestational Age (weeks) (SD in parentheses)	Mean One-minute Apgar Score (SD in parentheses)
Saline ($n = 11$)	843.0 (290.5)	27.3 (2.4)	5.5 (1.9)
1 dose of ibuprofen ($n = 11$)	908.6 (334.0)	26.6 (2.9)	3.7 (2.3)
3 doses of ibuprofen ($n = 12$)	947.5 (245.3)	26.5 (2.4)	5.0 (2.1)

- a. Suppose that an ANOVA is to be performed to compare the average birth weight for the different treatment groups. State precisely the ANOVA model. Is treatment group a fixed-effects factor or a random-effects factor?
- b. Using the summary statistics given in the table, determine the ANOVA table for the model in (a).
- c. Test whether the three treatment groups differ significantly in mean birth weights.
- d–f. Repeat parts (a)–(c) for gestational age.
- g–i. Repeat parts (a)–(c) for Apgar score.
16. In the PDA research described in Problem 15, one of the outcome variables was mean airway pressure (measured in cm H₂O) seven days after birth—a measure of respiratory status. The data on this outcome are summarized in the accompanying table.

Treatment Group	Mean Airway Pressure (cm H ₂ O) (SD in parentheses)
Saline (n = 11)	4.7 (1.8)
1 dose of ibuprofen (n = 11)	3.4 (0.7)
3 doses of ibuprofen (n = 12)	2.6 (0.6)

- a.–c. Repeat parts (a)–(c) of Problem 15, using mean airway pressure as the dependent variable.
- d. Use the Tukey–Kramer method to locate any significant differences between pairs of means. (Use $\alpha = .05$.)
17. A survey was conducted to examine the effects of pharmaceutical drug advertising on physician practices. Self-administered questionnaires were delivered to randomly selected physicians in various practice types: general practice, family practice, and internal medicine. Physicians reported on their attitude toward drug advertising and on the influence of drug advertising on their prescription-writing habits. Likert Scale measurements were used in the survey (1 = Strongly disagree, 5 = Strongly agree; the higher the score, the more favorably disposed the respondent is to advertising). The sample data are presented in the following table and in the accompanying computer output.

Practice Type	Attitude toward Advertising	Influence on Prescription-writing Habits
General practice	3, 4, 4, 4, 4, 3	4, 5, 5, 3, 5, 4
Family practice	3, 4, 2, 2, 2, 1	3, 2, 2, 2, 1, 3
Internal medicine	2, 4, 2, 1, 1, 2	1, 2, 3, 1, 1, 1

- a. Suppose that an ANOVA is to be performed to compare the average attitude toward advertising for the different practice types. State precisely the ANOVA model for these data.

Edited SAS Output (PROC GLM) for Problem 17

CLASS LEVEL INFORMATION		
Class	Levels	Values
practice	3	Fam Gen Int

Dependent Variable: attitude

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	9.33333333	_____	_____	0.0159
Error	15	12.66666667	_____	_____	_____
Corrected Total	17	22.00000000	_____	_____	_____

R-Square	Coeff Var	Root MSE	attitude Mean
0.424242	34.46012	0.918937	2.666667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
practice	2	9.33333333	4.66666667	5.53	0.0159

Source	DF	Type III SS	Mean Square	F Value	Pr > F
practice	2	9.33333333	4.66666667	5.53	0.0159

Scheffe's Test for Attitude

Alpha	0.05
Error Degrees of Freedom	15
Error Mean Square	0.844444
Critical Value of F	3.68232
Minimum Significant Difference	1.4398

MEANS WITH THE SAME LETTER ARE NOT SIGNIFICANTLY DIFFERENT.

Scheffe Grouping	Mean	N	practice
A	3.6667	6	Gen
A			
B	2.3333	6	Fam
B			
B	2.0000	6	Int

CLASS LEVEL INFORMATION

Class	Levels	Values
practice	3	Fam Gen Int

Dependent Variable: habits

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	_____	_____	_____	<.0001
Error	—	9.66666667	0.64444444	_____	_____
Corrected Total	17	_____	_____	_____	_____

(continued)

R-Square	Coeff Var	Root MSE	habits Mean
0.731481	30.10399	0.802773	2.666667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
practice	2	26.33333333	13.16666667	20.43	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
practice	2	26.33333333	13.16666667	20.43	<.0001

Scheffe's Test for Habits

Alpha	0.05
Error Degrees of Freedom	15
Error Mean Square	0.644444
Critical Value of F	3.68232
Minimum Significant Difference	1.2578

MEANS WITH THE SAME LETTER ARE NOT SIGNIFICANTLY DIFFERENT.			
Scheffe Grouping	Mean	N	practice
A	4.3333	6	Gen
B	2.1667	6	Fam
B	1.5000	6	Int

- b. Complete the ANOVA table in the computer output for the model in part (a).
- c. Test whether the physician types differ significantly in average attitude toward advertising.
- d. Use Scheffe's method to locate differences between pairs of means.
- e.–h. Repeat parts (a)–(d) with regard to influence on prescription-writing habits.
18. Customer portfolio analyses provide retailers with data on customer characteristics, including expenditure patterns. Based on this information, retailers can make marketing and operational adjustments to improve their market position. ABC Foods, a large supermarket chain, conducted a survey of the average weekly expenditures of 542 randomly selected residents in a large metropolitan market. The customers were classified as Loyal to ABC, New to ABC, Defectors from ABC, Loyal to competitors, and Unaffiliated. The average weekly expenditure data are summarized in the following table.

Customer Type	Sample Size	Avg. Weekly Expenditure (\$D)
Loyal to ABC	84	\$75 (\$13)
New to ABC	25	\$65 (\$10)
Defectors from ABC	27	\$82 (\$14)
Loyal to competitors	173	\$93 (\$17)
Unaffiliated	233	\$82 (\$15)

- a. Suppose that an ANOVA is to be performed to compare the average weekly expenditures for the different customer types. State precisely the ANOVA model. Is customer type a fixed-effects factor or a random-effects factor?
- b. Using the summary statistics given in the table, determine the ANOVA table for the model in part (a).
- c. Test whether the five customer types differ significantly in average weekly expenditures.
- d. Use Scheffé's method to locate any significant differences between pairs of means. (Use $\alpha = .05$.)
19. The data in the accompanying table were sampled from *U.S. News & World Report's* 1996 story on the "Best Mutual Funds."²⁰ The following variables are shown for each fund:
- CAT** (fund category): 1 = Aggressive growth; 2 = Long-term growth;
 3 = Growth and income; 4 = Income.
- LOAD** (load status): N = No load; L = Load.
- VOL** (volatility): A letter grade from A+ to F indicating how much the month-to-month return varied from the fund's three-year total return: A+ = Least variability; F = Most variability.
- OPI** (Overall Performance Index): An overall measure of the relative performance of each fund over the past 1, 3, 5, and 10 years. The higher the OPI, the better the performance.

Fund	CAT	LOAD	VOL	OPI
20th Century Giftrust Investors	1	N	F	89.9
AIM Aggressive Growth	1	L	F+	92.4
Stein Roe Capital Opportunity	1	N	F+	88.1
FPA Capital	1	L	D	92.2
Third Avenue Value	1	N	B	88.2
Phoenix U.S. Govt. A	1	L	D	88.0
Vanguard Primecap	2	N	D+	93.9
MFS Research A	2	L	D+	94.5
Fidelity Value	2	N	B	91.4
Mairs & Power Growth	2	N	C	99.1
Guardian Park Avenue	2	L	C	92.2
AIM Value A	2	L	D+	90.6
Lexington Corp. Leaders	3	N	C	91.3
Fundamental Investors	3	L	B	92.7
Oppenheimer Main St. Inc. & Growth A	3	L	D+	90.8
MAS Value Portfolio	3	N	B	99.8
Vanguard Index Value	3	N	B	88.9
Putnam Growth and Income	3	L	B+	90.8
Federated Liberty Equity Income A	4	L	B	89.5
United Income A	4	L	C+	76.8
Benham Inc. & Growth	4	N	B+	89.8
Manager's Income Equity	4	N	B	84.7
Pioneer Equity Income	4	L	B+	84.0
One Group Income Equity A	4	N	B+	83.3

²⁰ "1996 Mutual Funds Guide" 1996.

Edited SAS Output (PROC GLM) for Problem 19

CLASS LEVEL INFORMATION		
Class	Levels	Values
CAT	4	1 2 3 4

Number of Observations Read	24
Number of Observations Used	24

Dependent Variable: OPI

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	—	282.0545833	94.0181944	—	—
Error	20	255.5050000	12.7752500		
Corrected Total	23	537.5595833			

R-Square	Coeff Var	Root MSE	OPI Mean
0.524695	3.966062	3.574248	90.12083

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CAT	3	282.0545833	94.0181944	7.36	0.0016

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CAT	3	282.0545833	94.0181944	7.36	0.0016

Tukey's Studentized Range (HSD) Test for OPI

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	12.77525
Critical Value of Studentized Range	3.95825
Minimum Significant Difference	5.7758

MEANS WITH THE SAME LETTER ARE NOT SIGNIFICANTLY DIFFERENT.			
Tukey Grouping	Mean	N	CAT
A	93.617	6	2
A			
A	92.383	6	3
A			
B	A	89.800	6
B			
B		84.683	6
			4

- a. Suppose that an ANOVA is to be performed to compare the average OPI values for the different fund categories. State precisely the ANOVA model. Is fund category a fixed- or random-effects factor?
 - b. Using the SAS output that follows, complete the ANOVA table.
 - c. Test whether the average overall performance indices (OPI) differ significantly across the four fund categories (CAT).
 - d. Use the Tukey–Kramer method to locate any significant differences between pairs of means. (Use $\alpha = .05$.)
20. This problem refers to the data of Problem 19.
- a. Suppose that an ANOVA is to be performed to compare the average OPI values for funds with different volatilities. State precisely the ANOVA model. Is fund volatility a fixed- or random-effects factor?
 - b. In the SAS output that follows, complete the ANOVA table.
 - c. Test whether the average overall performance indices differ significantly by volatility rating. Interpret your results.

Edited SAS Output (PROC GLM) for Problem 20

Dependent Variable: OPI

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	—	291.4399405	41.6342772	—	—
Error	16	246.1196429	15.3824777		
Corrected Total	23	537.5595833			

R-Square	Coeff Var	Root MSE	OPI Mean
0.542154	4.351991	3.922050	90.12083

Source	DF	Type I SS	Mean Square	F Value	Pr > F
VOL	7	291.4399405	41.6342772	2.71	0.0470

Source	DF	Type III SS	Mean Square	F Value	Pr > F
VOL	7	291.4399405	41.6342772	2.71	0.0470

21. Radial keratotomy is a type of refractive surgery in which radial incisions are made in a myopic (nearsighted) patient's cornea to reduce the person's myopia. Theoretically, the incisions allow the curvature of the cornea to become less steep, thereby reducing the patient's refractive error. (*Note:* Myopic patients have negative refractive errors. Patients who are farsighted have positive refractive errors. Patients who are neither near- nor farsighted have zero refractive error.)

The incisions extend radially from the periphery toward the center of the cornea. A circular central portion of the cornea, known as the clear zone, remains uncut. The diameter of the clear zone is determined by the baseline refraction of the patient. Patients with a greater degree of myopia may receive longer incisions, leaving them with smaller clear zones. The thinking here is that "more surgery" is needed to correct the worse initial vision of these patients.

Edited SAS Output (PROC GLM) for Problem 21

CLASS LEVEL INFORMATION		
Class	Levels	Values
CLRZONE	3	3.0 3.5 4.0

Number of Observations Read	54
Number of Observations Used	51

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	—	14.70441590	—	—	—
Error	48	64.59163802	1.34565913	—	—
Corrected Total	50	79.29605392	—	—	—

R-Square	Coeff Var	Root MSE	Y Mean
0.185437	30.43590	1.160025	3.811373

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CLRZONE	2	14.70441590	7.35220795	5.46	0.0073

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CLRZONE	2	14.70441590	7.35220795	5.46	0.0073

Tukey's Studentized Range (HSD) Test for Y

Alpha	0.05
Error Degrees of Freedom	48
Error Mean Square	1.345659
Critical Value of Studentized Range	3.42021

COMPARISONS SIGNIFICANT AT THE 0.05 LEVEL ARE INDICATED BY ***.				
CLRZONE Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
3.0 - 3.5	0.7021	-0.2562	1.6603	
3.0 - 4.0	1.2778	0.3368	2.2188	***
3.5 - 3.0	-0.7021	-1.6603	0.2562	
3.5 - 4.0	0.5757	-0.4325	1.5840	
4.0 - 3.0	-1.2778	-2.2188	-0.3368	***
4.0 - 3.5	-0.5757	-1.5840	0.4325	

Radial keratotomy and other vision-correction surgery techniques grew in popularity in the 1980s and 1990s, both among the public and among ophthalmologists. The Prospective Evaluation of Radial Keratotomy (PERK) study was begun in 1983 to evaluate the effects of radial keratotomy. Lynn et al. (1987) examined the variables associated with the five-year postsurgical change in refractive error (Y , measured in diopters, D). One of the independent variables under consideration was diameter of the clear zone (X). In the PERK study, three clear zone sizes were used: 3.0 mm, 3.5 mm, and 4.0 mm.

The above computer output is based on data adapted from the PERK study.

- a. Suppose that an ANOVA is to be performed to compare the average five-year change in refraction for patients with different clear zones. State precisely the ANOVA model. Is clear zone size a fixed- or random-effects factor?
 - b. In the above SAS output, complete the ANOVA table.
 - c. Test whether the average five-year change in refraction differs significantly by clear zone size. Interpret your results.
 - d. Use the Tukey–Kramer method to locate any significant pairwise differences between clear zones. (Use $\alpha = .05$.) Interpret your results.
22. Data on law and business schools were sampled from *U.S. News & World Report's* 1996 report on the “America’s Best Graduate Schools 1996 Annual Guide.”²¹ The school’s reputation rank among academics and the 1995 median starting salary for graduates are shown in the accompanying table for 12 randomly sampled law schools and 12 randomly sampled business schools.

University	School	Reputation Rank by Academics	1995 Median Starting Salaries (in \$1000s)
Vanderbilt University	Law	17	62
University of Chicago	Law	2	70
Brigham Young University	Law	45	45.5
George Washington University	Law	24	60
Cornell University	Law	11	70
Rutgers University	Law	45	62
University of Pennsylvania	Law	6	70
University of Illinois (Urbana–Champaign)	Law	23	53
Villanova University	Law	65	62.2
University of Florida	Law	37	43.4
Tulane University	Law	49	53
Case Western Reserve University	Law	41	47
University of Georgia	Business	45	45
Tulane University	Business	40	50
University of Southern California	Business	24	55
Brigham Young University	Business	57	68
Emory University	Business	33	58
University of Illinois (Urbana–Champaign)	Business	24	45.6
Cornell University	Business	10	60
University of North Carolina (Chapel Hill)	Business	16	59
Massachusetts Institute of Technology	Business	1	75
University of Texas (Austin)	Business	19	55
College of William and Mary	Business	67	48
Penn State University	Business	33	49.7

²¹ “America’s Best Graduate Schools” 1996.

Edited SAS Output (PROC GLM) for Problem 22

CLASS LEVEL INFORMATION		
Class	Levels	Values
SCHOOL	2	Bus Law

Number of Observations Read	24
Number of Observations Used	24

Dependent Variable: SAL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	—	37.001667	37.001667	—	—
Error	22	1929.391667	87.699621		
Corrected Total	23	1966.393333			

R-Square	Coeff Var	Root MSE	SAL Mean
0.018817	16.44873	9.364808	56.93333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
SCHOOL	1	37.00166667	37.00166667	0.42	0.5227
Source	DF	Type III SS	Mean Square	F Value	Pr > F
SCHOOL	1	37.00166667	37.00166667	0.42	0.5227

CLASS LEVEL INFORMATION		
Class	Levels	Values
REP	2	1 2

Number of Observations Read	24
Number of Observations Used	24

Dependent Variable: SAL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	—	440.326667	440.326667	—	—
Error	22	1526.066667	69.366667		
Corrected Total	23	1966.393333			

R-Square	Coeff Var	Root MSE	SAL Mean
0.223926	14.62880	8.328665	56.93333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
REP	1	440.3266667	440.3266667	6.35	0.0195

Source	DF	Type III SS	Mean Square	F Value	Pr > F
REP	1	440.3266667	440.3266667	6.35	0.0195

- a. Suppose that an ANOVA is to be performed to compare the average 1995 starting salaries (SAL) for business and law schools (let the variable SCHOOL = Law or Business, depending on the type of school). State precisely the ANOVA model. Is SCHOOL a fixed- or random-effects factor?
 - b. In the above SAS output, complete the ANOVA table.
 - c. Test whether the average 1995 starting salaries differ significantly for graduates of law schools and of business schools.
 - d. Suppose that an ANOVA is to be performed to compare the average 1995 starting salaries (SAL) for the top 25 schools and schools not in the top 25 in terms of reputation rank (let the variable REP = 1 if a school's reputation rank is 25 or less, REP = 2 if the rank is 26 or more). State precisely the ANOVA model.
 - e. In the second part of the above SAS output, complete the ANOVA table.
 - f. Test whether the average 1995 starting salaries differ significantly between top 25 schools and schools not in the top 25 in terms of reputation rank.
23. In September 1996, *U.S. News & World Report* published a report on America's health maintenance organizations (HMOs).²² The report was intended to serve as a consumer guide to HMO quality. For each HMO included in the report, data were provided on several variables, including the following:

PHYSTURN: Physician turnover rate (%).

PHYSCERT: Percentage of doctors who were board certified.

PREV: Prevention score, indicating how well the HMO meets Public Health Service goals in various measures of preventive care (including immunizations and prenatal care). The results can be negative (indicating that the HMO falls short of the goals) or positive (indicating that it exceeds the goals).

HMO	PREV	PHYSTURN	PHYSCERT
HMO Kentucky	-114	1	64
Kaiser Foundation (HI Region)	-6	7	92
CIGNA HealthCare of LA	-97	6	80
CIGNA HealthCare of S. California	-82	8	81
CIGNA HealthCare of N. California	-42	21	82
HIP Health Plan of Florida	-8	3	80
CIGNA HealthCare of San Diego	-36	5	76
NYLCare of the Mid-Atlantic	-9	6	80
Personalcare Insurance of Illinois	-26	13	74
Prudential Healthcare Tri-state	-78	9	71
CIGNA HealthCare of S. Florida	-26	4	77
Health New England	11	5	79
Group Health Northwest	4	3	90
Kaiser Foundation—Mid-Atlantic	29	4	92
Health Alliance Medical Plans	-28	6	82
Pilgrim Health Care	22	2	83
Partners National Health, NC	-1	6	84
Healthsource of New Hampshire	11	5	81

²² "Rating the HMOs" 1996.

Edited SAS Output (PROC GLM) for Problem 23

CLASS LEVEL INFORMATION		
Class	Levels	Values
TURN	2	1 2

Number of Observations Read	18
Number of Observations Used	18

Dependent Variable: PREV

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2868.56752	2868.56752	1.72	0.2085
Error	16	26717.87692	1669.86731		
Corrected Total	17	29586.44444			

R-Square	Coeff Var	Root MSE	PREV Mean
0.096955	-154.5278	40.86401	-26.44444

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TURN	1	2868.567521	2868.567521	1.72	0.2085

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TURN	1	2868.567521	2868.567521	1.72	0.2085

CLASS LEVEL INFORMATION		
Class	Levels	Values
BORDCERT	2	1 2

Number of Observations Read	18
Number of Observations Used	18

Dependent Variable: PREV

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7691.37778	7691.37778	5.62	0.0306
Error	16	21895.06667	1368.44167		
Corrected Total	17	29586.44444			

R-Square	Coeff Var	Root MSE	PREV Mean
0.259963	-139.8874	36.99245	-26.44444

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BORDCERT	1	7691.377778	7691.377778	5.62	0.0306

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BORDCERT	1	7691.377778	7691.377778	5.62	0.0306

- a. In the *U.S. News & World Report* story, the average physician turnover rate for HMOs is reported to be 6%. Suppose that an ANOVA is to be performed to compare the average prevention scores for HMOs with low physician turnover rates ($\leq 6\%$) and HMOs with higher rates ($> 6\%$). State precisely the ANOVA model.
- b. In the above SAS output, complete the ANOVA table. To produce the output, the following coding scheme was followed: **TURN** = 1 if **PHYSTURN** $\leq 6\%$; 2 otherwise.
- c. Test whether the average prevention scores differ significantly for the two types of HMOs mentioned in part (a).
- d. Suppose that an ANOVA is to be performed to compare the average prevention scores for HMOs with low percentages of board-certified primary care physicians ($\leq 75\%$) and HMOs with higher percentages of board-certified physicians ($> 75\%$). State precisely the ANOVA model.
- e. In the second part of the above SAS output, complete the ANOVA table. To produce the output, the following coding scheme was followed: **CERT** = 1 if **PHYSCERT** $\leq 75\%$; 2 otherwise.
- f. Test whether the average prevention scores differ significantly for the two types of HMOs mentioned in part (d).

References

- “America’s Best Graduate Schools.” 1996. *U.S. News & World Report*, March 18, 79–91.
- Armitage, P. 1971. *Statistical Methods in Medical Research*. Oxford: Blackwell Scientific.
- Cotton, R. B.; Stahlman, M. T.; Kovar, I.; and Catterton, W. Z. 1979. “Medical Management of Small Preterm Infants with Symptomatic Patent Ductus Arteriosus.” *Journal of Pediatrics* 2: 467–73.
- Daly, M. B. 1973. “The Effect of Neighborhood Racial Characteristics on the Attitudes, Social Behavior, and Health of Low Income Housing Residents.” Ph.D. dissertation, Department of Epidemiology, University of North Carolina, Chapel Hill, N.C.
- Diggle, P. J.; Heagerty, P. J.; Liang, K. Y.; and Zeger, S. L. 2013. *Analysis of Longitudinal Data*, Second Edition (paperback). Oxford: Oxford University Press.
- Fitzmaurice G.; Davidian, M.; Verbeke, G.; and Molenberghs, G. 2009. *Longitudinal Data Analysis*. London: Chapman and Hall/CRC.
- Guenther, W. C. 1964. *Analysis of Variance*. Englewood Cliffs, N.J.: Prentice-Hall.
- Hollander, M., and Wolfe, D. A. 1973. *Nonparametric Statistical Methods*. New York: John Wiley & Sons.
- Kirk, R. E. 1969. *Experimental Design: Procedures for the Behavioral Sciences*. Belmont, Calif.: Wadsworth.
- Kleinbaum, D.G., and Klein, M. 2010. *Logistic Regression—A Self-Learning Text*, Third Edition (Chapters 13–16). New York: Springer Publishers.
- Kramer, C. Y. 1956. “Extension of the Multiple Range Test to Group Means with Unequal Numbers of Replications.” *Biometrics* 12: 307–10.
- Kutner, M. H.; Neter, J.; Nachtsheim, C. J.; and Li, W. 2004. *Applied Linear Statistical Models*, Fifth Edition. New York: McGraw-Hill/Irwin.
- Lehmann, E. L. 1975. *Non-parametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day.
- Lindman, H. R. 1974. *Analysis of Variance in Complex Experimental Designs*. San Francisco: W. H. Freeman.

- Lynn, M. J.; Waring, G. O. III; and Sperduto, R. D. 1987. "Factors Affecting Outcome and Predictability of Radial Keratotomy in the PERK Study." *Archives of Ophthalmology* 105: 42–51.
- Miller, R. G., Jr. 1966. *Simultaneous Statistical Inference*. New York: McGraw-Hill.
- . 1981. *Simultaneous Statistical Inference*, Second Edition. New York: Springer-Verlag.
- "1996 Mutual Funds Guide, Best Mutual Funds." 1996. *U.S. News & World Report*, January 29, 88–100.
- "Rating the HMOs." 1996. *U.S. News & World Report*, September 2, 52–63.
- Siegel, S. 1956. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.
- Tukey, J. W. 1953. "The Problem of Multiple Comparisons" (mimeographed notes, Princeton University, Princeton, N.J.).
- Varvarigou, A.; Bardin, C. L.; Beharry, K.; Chemtob, S.; Papageorgiou, A.; and Aranda, J. 1996. "Early Ibuprofen Administration to Prevent Patent Ductus Arteriosus in Premature Newborn Infants." *Journal of American Medical Association* 275: 539–4.

18

Randomized Blocks: Special Case of Two-way ANOVA

18.1 Preview

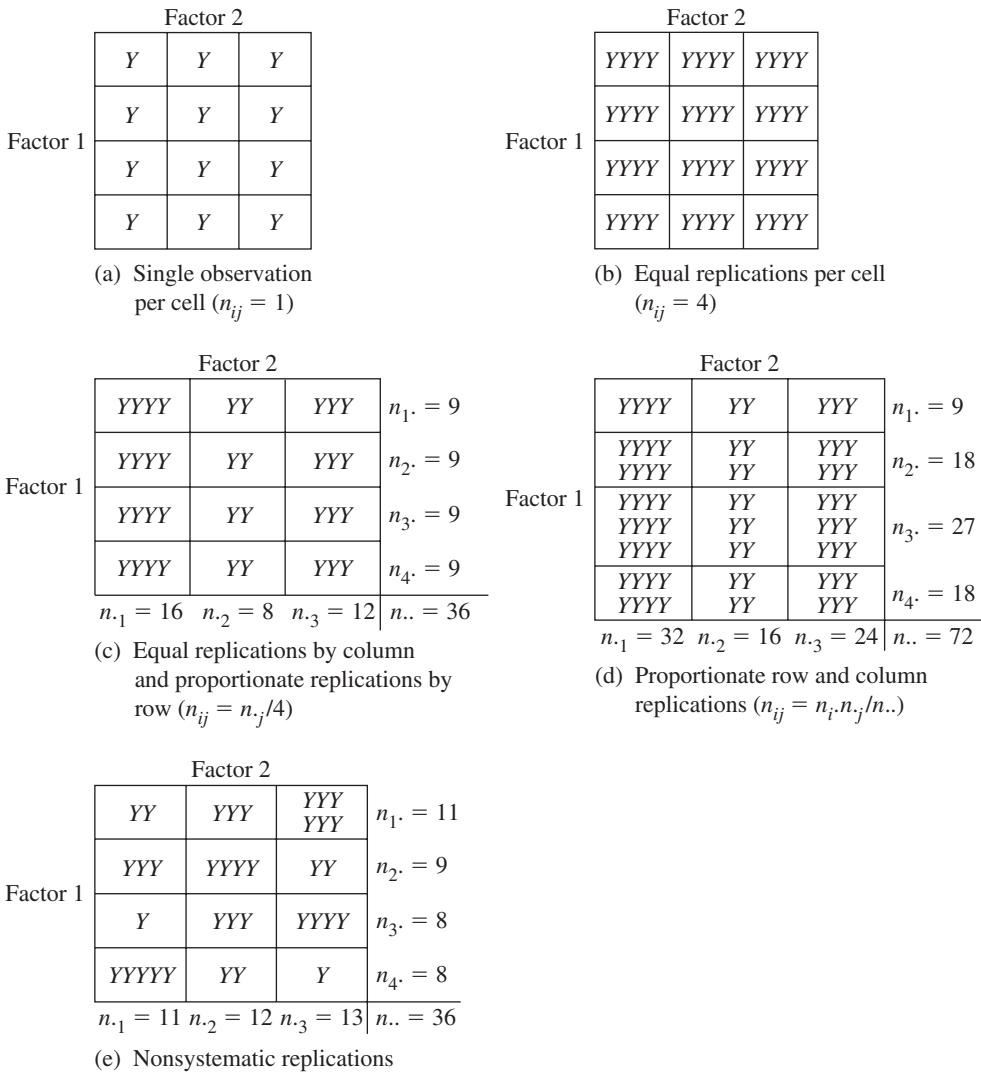
In Chapter 17, we considered the simplest kind of ANOVA problem—one involving a single factor (or independent variable). We now focus on the two-factor case, which is generally referred to as two-way ANOVA. This extension is by no means trivial. In fact, we will devote three chapters (Chapters 18, 19, and 20) to different aspects of the two-factor case. In this chapter, we first examine how a two-factor situation may be classified according to the data pattern. We then restrict our attention to a specific type of pattern, which will lead us to consider the randomized-blocks design, the main topic of this chapter.

18.1.1 Two-way Data Patterns

Several different types of data patterns for two-way ANOVA are illustrated in Figure 18.1. Each of these tables describes a two-factor study with four levels of factor 1 (the “row” factor) and three levels of factor 2 (the “column” factor). The Y ’s in each table correspond to individual observations on a single dependent variable Y . The number of Y ’s in a given cell is denoted by n_{ij} for the i th level of factor 1 and the j th level of factor 2. The marginal total number of Y ’s in the i th row is denoted by $n_{i\cdot}$, and the marginal total for the j th column is denoted by $n_{\cdot j}$. The total number of observations is denoted by $n\ldots$.

The simplest two-factor pattern, which is illustrated in Figure 18.1(a), arises when each cell holds a single observation (i.e., when $n_{ij} = 1$ for all i and j). The randomized-blocks design discussed in this chapter is a special case of this pattern, although such single-observation-per-cell data may arise in other ways.

A second type of pattern, illustrated in Figure 18.1(b), occurs when equal numbers of observations occur in each cell. Here, $n_{ij} = 4$ for all i and j . Chapter 19 will focus on this equal-replications situation.

**FIGURE 18.1** Some two-way data patterns for a 4×3 table

The patterns in parts (c) through (e) of Figure 18.1 present different problems in statistical analysis,¹ which we will discuss in Chapter 20. The common property of these three patterns is that not all cells have the same number of observations. Unequal cell numbers often arise in observational studies in which the number of observations at each level of a factor is not necessarily under the control of the investigators. Alternatively, an experiment designed for equal cell numbers may ultimately produce unequal cell numbers due to either

¹ In Chapters 18, 19, and 20, we assume that each cell in a table contains *at least* one observation. When one or more cells contain no observations, the analysis required is considerably more complicated. Further discussion of this situation is provided in texts by Ostle (1963), Peng (1967), and Armitage (1971).

the occurrence of randomly missing data or the purposeful elimination of clearly erroneously measured observations.

For the pattern in Figure 18.1(c), cells in the same column have the same number of observations, whereas cells in the same row are in the ratio 4:2:3. For this table, each of the four cell frequencies in the j th column is equal to the same fraction of the corresponding total column frequency (i.e., $n_{ij} = n_{.j}/4$ in this case). Note, for example, that $n_{.1}/4 = 16/4 = 4$, which is the number of observations in any cell in column 1.

For Figure 18.1(d), the cells in a given column are in the ratio 1:2:3:2, whereas the cells in a given row are in the ratio 4:2:3. This pattern results because n_{ij} is determined as

$$n_{ij} = \frac{n_i \cdot n_{.j}}{n_{..}}$$

which means that any cell frequency can be obtained by multiplying the corresponding row and column marginal frequencies together and then dividing by the total number of observations. Thus, for cell (1, 2) in Figure 18.1(d), we have $n_{1.} \cdot n_{.2} / n_{..} = 9(16)/72 = 2$, which equals n_{12} . Similarly, for cell (4, 3), $n_{4.} \cdot n_{.3} / n_{..} = 18(24)/72 = 6$, which equals n_{43} .

There is no mathematical rule for describing the pattern of cell frequencies in Figure 18.1(e), so we say that such a pattern is nonsystematic. As we will see in Chapter 20, the ANOVA procedures required for the patterns in Figure 18.1(c) and 18.1(d) differ from those required for the irregular pattern in Figure 18.1(e). For the former two patterns, the same computational procedure may be used as when an equal number of observations exists per cell. For the nonsystematic case, a different procedure is required.

18.1.2 The Case of a Single Observation per Cell

A two-way table with a single observation in each cell can arise in a number of different experimental situations. Consider the following three examples:

1. Six hypertensive individuals, matched pairwise by age and sex, are randomly assigned (within each pair) to either a treatment or a control group. For each individual, a measure of change in self-perception of health is determined after one year. The main question of interest is whether the true mean change in self-perception for the treatment group differs from that for the control group.
2. Six growth-inducing treatment combinations are randomly assigned to six mice from the same litter. The treatment combinations are defined by the cross-classification of the levels of two factors: factor A (drug A1 or placebo A0) and factor B (drug B2 [high dose], drug B1 [low dose], or placebo B0). The dependent variable of interest is weight gain measured one week after treatment is initiated. The questions to be considered include (a) whether the effect of drug A1 differs from that of placebo A0; (b) whether differences exist among the effects of drugs B1 and B2 and placebo B0; and (c) whether the drug A1 effect differs from that of placebo A0 in the same way at each level of factor B.
3. Scores of satisfaction with medical care are recorded for six hypertensive patients assigned to one of six categories, depending on whether the nurse practitioner assigned to the patient was measured to have high or low autonomy (factor A)

and high, medium, or low knowledge of hypertension (factor B). The main questions of interest are (a) whether mean satisfaction scores differ between patients with high-autonomy nurse practitioners and those with low-autonomy nurse practitioners and (b) whether these differences in mean satisfaction scores differ in the same way at each level of knowledge grouping.

Each of these experiments may be represented by a 2×3 two-way table, as shown in Figure 18.2.

In the figure, the third example, Figure 18.2(c), involves two factors whose levels (or categories) were determined after the data were gathered. Such a study is often referred to as an *observational study*, rather than as an *experiment*, since the latter term is usually reserved for studies involving factors whose levels are specified beforehand. Epidemiological, sociological, and psychological studies are often observational rather than experimental. For this example, the autonomy groupings were determined after the frequency distribution of autonomy scores on nurse practitioners was considered. Similarly, the knowledge categories were determined using the observed knowledge scores. Thus, one patient was in each of the six groups because of a posteriori (rather than a priori) considerations. In actual practice, it is often impossible to arrange things so nicely, especially when large samples are involved. That is why most such observational studies have unequal numbers of observations in various cells.²

		Pair 1	Pair 2	Pair 3	
		<i>Y</i>	<i>Y</i>	<i>Y</i>	<i>Y</i> = Change in self-perception of health after 1 year
		<i>Y</i>	<i>Y</i>	<i>Y</i>	
Treatment					
Control					

(a)

		Placebo B0	Drug B1	Drug B2	
		<i>Y</i>	<i>Y</i>	<i>Y</i>	<i>Y</i> = Weight gain after 1 week
		<i>Y</i>	<i>Y</i>	<i>Y</i>	
Placebo A0					
Drug A1					

(b)

		Low Knowledge	Medium Knowledge	High Knowledge	
		<i>Y</i>	<i>Y</i>	<i>Y</i>	<i>Y</i> = Patient satisfaction with medical care
		<i>Y</i>	<i>Y</i>	<i>Y</i>	
Low Autonomy					
High Autonomy					

(c)

FIGURE 18.2 Different experimental situations resulting in two-way tables with a single observation per cell

² In this chapter, we focus on the case where the levels of each factor are considered fixed; nevertheless, even if one or both factors were considered random, the tests of hypotheses of interest, as with one-way ANOVA, would be computed in exactly the same way as in the fixed-factor case. This is not so when more than one observation occurs per cell, as discussed in Chapter 19.

The second example, Figure 18.2(b), involves two factors whose levels were determined before the data were collected; the resulting six *treatment combinations* can be viewed, in one sense, as representing the different levels of a single factor that have been randomly assigned to the six individuals. Although it may be necessary because of limited experimental material or prohibitive cost to apply or assign each treatment combination to only a single experimental unit, considerably more information is obtained if each treatment combination has several replications.

This brings us to Figure 18.2(a), which is of the general type to be considered in this chapter. Like Figure 18.2(b), this example represents a randomized experiment. It differs from the other tables in Figure 18.2, however, in its allocation of individuals to cells (i.e., treatment combinations). The allocation here was done by randomization within each pair rather than, say, by randomization across all six individuals, ignoring any pairing. Another feature unique to this table is that the effect of only one of the two factors involved (in this case, “treatment versus control”) is of primary interest. The other factor, “pair” (with three levels corresponding to the three pairs), is used only to help increase the precision of the comparison between the treatment and control groups. Thus, if pair matching (on age and sex) is used and significant differences are found between the Y values for paired treatment and control subjects, such differences cannot solely be attributed to one group being older, say, or having a different sex composition than the other. The pairing, therefore, serves to eliminate or block out noise that otherwise would affect the comparison between treatment and control groups due to the confounding effects of age and sex. Such pairs are often referred to as *blocks*, and the associated experimental design is called a *randomized-blocks design*.

The analysis required for data arranged as in Figure 18.2(a) is described in most introductory statistics texts. Since two groups are involved (treatment and control) and since matching has been done, the generally recommended method of analysis involves using the paired-difference t test, which focuses on differences in Y values within pairs. Hence, the key test statistic involved is of the form $T = \bar{d} \sqrt{n}/S_d$, where \bar{d} is the difference between the treatment group mean and the control group mean, S_d is the standard deviation of the difference scores for all pairs, and n is the number of pairs (3, in Figure 18.2(a)). This statistic has the t distribution with $(n - 1)$ degrees of freedom under H_0 , so the critical region is of the form $|T| \geq t_{n-1, 1-\alpha/2}$ for a two-sided test of the null hypothesis of a true average difference score of zero.

The paired-difference t test can be viewed as a special case of the general F test applied to a randomized-blocks ANOVA model in which there are only two treatments per block. In fact, this randomized-blocks F test represents a generalization of the paired-difference t test just as the one-way ANOVA F test represents a generalization of the two-sample t test.

18.2 Equivalent Analysis of a Matched-pairs Experiment

Let us make the matched-pairs example of the previous section more realistic by considering the data given in Table 18.1, which involves 15 pairs of individuals matched on age and sex. The main inferential question for these data involves whether the mean change in

TABLE 18.1 Matched-pairs design involving change scores in self-perception of health (Y) among hypertensives

Group	Pair															Total	Mean
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
Treatment	10	12	8	8	13	11	15	16	4	13	2	15	5	6	8	146	9.73
Control	6	5	7	9	10	12	9	8	3	14	6	10	1	2	1	103	6.87
Total	16	17	15	17	23	23	24	24	7	27	8	25	6	8	9	249	8.30
Difference	4	7	1	-1	3	-1	6	8	1	-1	-4	5	4	4	7	43	2.86

© Cengage Learning

self-perception score for the treatment (T) group significantly differs from that for the control (C) group. Stated in terms of population means, the null hypothesis is

$$H_0: \mu_T = \mu_C$$

where μ_T and μ_C denote the population mean change scores for the treatment and control groups, respectively. The alternative hypothesis is then

$$H_A: \mu_T \neq \mu_C$$

assuming that the analyst did not theorize in advance that one particular group would have a higher or lower population mean change score than the other. (If, however, the analyst thought a priori that the treatment group would have a significantly higher mean change score than the control group, the alternative would be one-sided: $H_A: \mu_T > \mu_C$.)

18.2.1 Paired-difference t Test

One method for testing the null hypothesis H_0 was described in the previous section—the paired-difference t test. To perform this test, we first determine from Table 18.1 that $\bar{d} = \bar{Y}_T - \bar{Y}_C = 2.86$ and

$$S_d^2 = \frac{1}{14} \left[\sum_{i=1}^{15} d_i^2 - \frac{\left(\sum_{i=1}^{15} d_i \right)^2}{15} \right] = 12.695$$

where d_i is the observed difference between the scores for the treatment and control groups for the i th pair. Then we compute the test statistic as

$$T = \frac{\bar{d} \sqrt{n}}{S_d} = \frac{2.86 \sqrt{15}}{\sqrt{12.695}} = 3.109$$

Since $t_{14, 0.995} = 2.976$ and $t_{14, 0.9995} = 4.140$, the P -value for this two-sided test is given by

$$.001 < P < .01$$

We, therefore, reject H_0 and conclude that the observed mean change score for the treatment group differs significantly from that for the control group.

18.2.2 Randomized-blocks F Test

Another way to test the null hypothesis $H_0: \mu_T = \mu_C$ for the data of Table 18.1 is to use an F test, based on the ANOVA table given in Table 18.2.

This ANOVA table differs in form from the one used for one-way ANOVA in that the total sum of squares is partitioned into three components (treatments, pairs, and error) instead of just two (between and within). The treatment component in Table 18.2 is similar to the between (i.e., treatment) source in the one-way ANOVA case, and the error component here is analogous to the within component in the one-way ANOVA case because this source is used as an estimate of the population variance σ^2 . Finally, we have a pairs (or blocks) component in Table 18.2 that has no corresponding component in the one-way ANOVA case.

To recap, whereas the total sum of squares in one-way ANOVA may be split up into two components, as indicated by the equation

$$\text{SS(Total)} = \text{SS(Treatments)} + \text{SS(Error)}$$

expressed in Chapter 17 as $\text{SSY} = \text{SST} + \text{SSE}$, the total sum of squares in a randomized-blocks ANOVA can be partitioned into three meaningful components:

$$\text{SS(Total)} = \text{SS(Treatments)} + \text{SS(Blocks)} + \text{SS(Error)} \quad (18.1)$$

The last two components on the right-hand side of (18.1) can be seen as representing a partition of the experimental error (or within) sum of squares associated with one-way ANOVA (i.e., with the blocking ignored). By separating out the blocking effect, we obtain a more precise estimate of experimental error—one not contaminated by any noise due to the effects of the blocking variables (in our case, age and sex).

The computed sums of squares in expression (18.1) are shown in the ANOVA table in Table 18.2. They are

SS(Treatments)	= 61.63
SS(Blocks)	= 391.80
SS/Error)	= 88.87
SS(Total)	= 542.30

TABLE 18.2 ANOVA table for matched-pairs data of Table 18.1

Source	SS	d.f.	MS	F
Treatment	61.63	1	61.63	9.71 (.005 < P < .01)
Pairs (blocks)	391.80	14	27.99	4.41 (.001 < P < .005)
Error	88.87	14	6.35	
Total	542.30	29		

The degrees of freedom corresponding to the sums of squares are shown in the second column of Table 18.2. The d.f. for “treatments” is always equal to 1 less than the number (k) of treatments (in our example, $k - 1 = 2 - 1 = 1$). The d.f. for “blocks” is equal to 1 less than the number (b) of blocks (in our example, $b - 1 = 15 - 1 = 14$). The d.f. for “error” is the product of the treatment degrees of freedom and the block degrees of freedom:

$$\text{Error d.f.} = (\text{Treatment d.f.})(\text{Block d.f.}) = (k - 1)(b - 1) = (1)(14) = 14$$

Finally, the d.f. for the total sum of squares is equal to 1 less than the number (kb) of observations (in our example, $30 - 1 = 29$).

The mean squares, as usual, are obtained by dividing the sums of squares by their corresponding degrees of freedom. Finally, the F statistic for the test of the null hypothesis that no differences exist among the treatments is given by the formula

$$F = \frac{\text{MS(Treatments)}}{\text{MS(Error)}} \quad (18.2)$$

In particular, to test $H_0: \mu_T = \mu_C$ for the data of Table 18.1, we calculate

$$F = \frac{61.63/1}{88.87/14} = \frac{61.63}{6.35} = 9.71$$

which is the observed value for an F random variable with 1 and 14 degrees of freedom under H_0 . Thus, H_0 is rejected at significance level α if the observed value of F is greater than $F_{1, 14, 1-\alpha}$. Since $F_{1, 14, 0.99} = 8.86$ and $F_{1, 14, 0.995} = 11.06$, the P -value for this test satisfies the inequality $.005 < P < .01$. This is quite small, so it is reasonable to reject H_0 and to conclude that the treatment group has a significantly different observed mean change score compared to that of the control group.

Thus, the conclusions reached via the paired-difference t test and the randomized-blocks F test are exactly the same: reject H_0 . This agreement occurs because the two tests are completely equivalent. It can be shown mathematically that the square of a paired-difference T statistic with v degrees of freedom is exactly equal to a randomized-blocks F statistic with 1 and v degrees of freedom (i.e., $F_{1, v} = T_v^2$) and that $F \geq F_{1, v, 1-\alpha}$ is equivalent to $|T| \geq t_{v, 1-\alpha/2}$.

In our example,

$$T_{14}^2 = (3.109)^2 = 9.67 = F_{1, 14}$$

and

$$t_{14, 0.995}^2 = (2.977)^2 = 8.86 = F_{1, 14, 0.99}$$

An F test of the null hypothesis H_0 : “No significant differences exist among the blocks” may also be performed. This test is not of primary interest, since the very use of a randomized-blocks design is based on the a priori assumption that significant block-to-block

variation exists. Nevertheless, an *a posteriori* F test may be used to check the reasonableness of this assumption. The test statistic to be used in this case is

$$F = \frac{\text{MS(Blocks)}}{\text{MS(Error)}}$$

which, for the example of Table 18.1, has the F distribution under H_0 with 14 degrees of freedom in the numerator and 14 degrees of freedom in the denominator. From Table 18.2, we find that this F statistic is computed to be 4.41, with a P -value satisfying $.001 < P < .005$. The conclusion for this test, as expected, is to reject the null hypothesis that no block differences exist.

18.3 Principle of Blocking

For the matched-pairs example, we indicated that the primary reason for “pairing up” the data was to prevent the confounding factors age and sex from blurring the comparison between the treatment and control groups. In other words, using the matched-pairs design represented an attempt to account for the fact that the experimental units (i.e., the subjects) were not homogeneous with regard to factors (other than experimental group membership status) that were likely to affect the response variable. In another experimental situation, age and sex might not be important covariates and so would not have to be controlled or adjusted for.

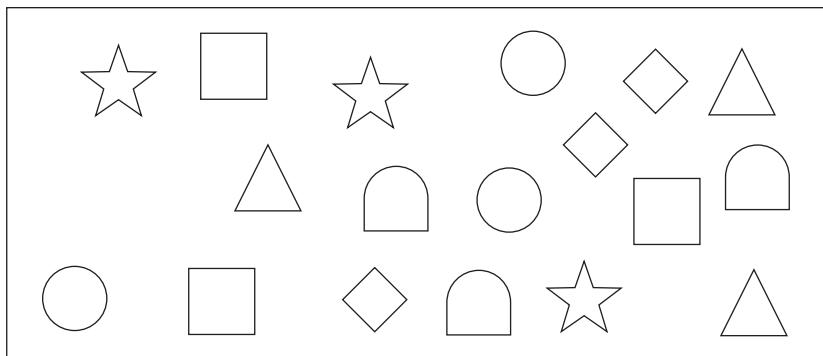
In a situation where the experimental units under study are heterogeneous relative to certain concomitant variables that affect the response variable (but are not of primary interest), using a randomized-blocks design requires the following two steps:

1. The experimental units (e.g., people, animals) that are homogeneous (with respect to levels of the concomitant variables) are collected together to form a block.
2. The various treatments are assigned at random to the experimental units within each block.

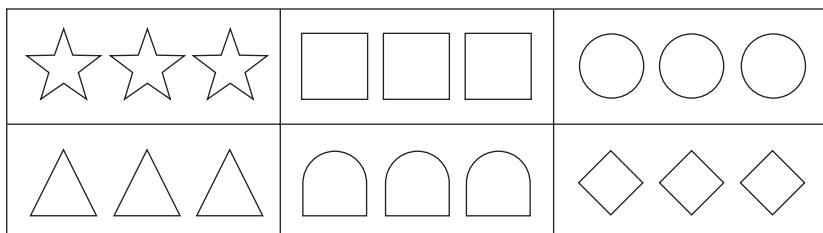
These steps are illustrated in Figure 18.3, where six blocks are formed, each consisting of three homogeneous experimental units. Three treatments (labeled A, B, and C) are then assigned at random to the three units within each block.

Figure 18.4, on the other hand, provides an example of incorrect blocking. In this case, one type of experimental unit might predominantly receive one kind of treatment, and another type might not get that treatment at all. For example, the experimental unit type \star was assigned treatment B twice, whereas experimental unit types \circ and \triangle were not assigned treatment B at all. If the blocking had been done correctly, every distinct type of experimental unit would have been assigned each treatment exactly once.

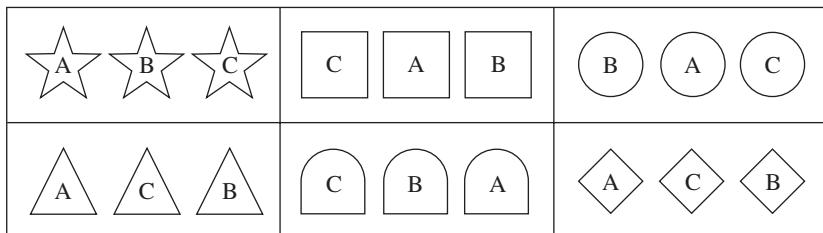
With regard to our matched-pairs design in Table 18.1, if we had not blocked on age, the treatment might have been assigned mostly to older subjects, and the control group might have consisted mostly of younger subjects. Any treatment-control differences found subsequently might very well have been due entirely to age differences between the two groups and not to a real differential effect of the treatment relative to control.



(a) Heterogeneous experimental units

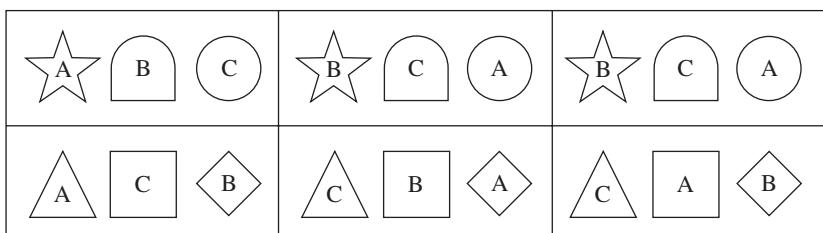


(b) Formation of blocks



(c) Randomization of treatments A, B, and C within each block

© Cengage Learning

FIGURE 18.3 Steps used in forming randomized blocks

© Cengage Learning

FIGURE 18.4 Example of incorrect blocking

18.4 Analysis of a Randomized-blocks Study

In this section, we discuss the appropriate analysis for a randomized-blocks study, focusing on the case where both factors (treatments and blocks) are considered fixed, although the tests of hypotheses of interest are computed in exactly the same way even if one or both factors are considered random; that is, practically speaking, it does not matter how the factors are defined.³

18.4.1 Data Presentation

Table 18.3 gives the general layout of the data for a randomized-blocks design involving k treatments and b blocks. In this table, Y_{ij} denotes the value of the observation on the dependent variable Y corresponding to the i th treatment in the j th block (e.g., Y_{23} denotes the value of the observation associated with treatment 2 in block 3). The (row) total for treatment i is denoted by T_i ; that is, $T_i = \sum_{j=1}^b Y_{ij}$. The (column) total for block j is denoted by B_j ; that is, $B_j = \sum_{i=1}^k Y_{ij}$. The grand total of all bk observations is $G = \sum_{i=1}^k \sum_{j=1}^b Y_{ij}$. Finally, the treatment (row) means are denoted by \bar{Y}_i , for the i th treatment; the block (column) means are denoted by $\bar{Y}_{\cdot j}$ for the j th block; and $\bar{Y}_{\cdot \cdot}$ denotes the grand mean.

A special case of this general format was given in Table 18.1, where $k = 2$ and $b = 15$. Another example (with $k = 4$ and $b = 8$) appears in Table 18.4. Although based on artificial data, this example illustrates the type of information being considered in several ongoing, large-scale U.S. intervention studies dealing with risk factors associated with heart disease. Table 18.4 presents the data for a small experiment designed to assess the effects of four different cholesterol-reducing diets on persons who have hypercholesterolemia. In such a study, it would seem logical to try to take into account (or adjust for) the effects of sex, age, and body size on the dependent variable Y , which represents reduction in cholesterol level after one year. One way to do this is by using a randomized-blocks design, where the blocks

TABLE 18.3 Data layout for a randomized-block design

Treatment	Block					Total	Mean
	1	2	3	...	b		
1	Y_{11}	Y_{12}	Y_{13}	...	Y_{1b}	T_1	$\bar{Y}_{1\cdot}$
2	Y_{21}	Y_{22}	Y_{23}	...	Y_{2b}	T_2	$\bar{Y}_{2\cdot}$
3	Y_{31}	Y_{32}	Y_{33}	...	Y_{3b}	T_3	$\bar{Y}_{3\cdot}$
...
k	Y_{k1}	Y_{k2}	Y_{k3}	...	Y_{kb}	T_k	$\bar{Y}_{k\cdot}$
Total	B_1	B_2	B_3	...	B_b	G	—
Mean	$\bar{Y}_{\cdot 1}$	$\bar{Y}_{\cdot 2}$	$\bar{Y}_{\cdot 3}$...	$\bar{Y}_{\cdot b}$	—	$\bar{Y}_{\cdot \cdot}$

© Cengage Learning

³ In Chapter 19, which discusses the situation of equal replications (at least two) per cell, the required F tests differ, depending on how the factors are defined.

TABLE 18.4 Randomized-blocks study for comparing the effects of four cholesterol-reducing diets on persons with hypercholesterolemia (Y = reduction in cholesterol level after one year)

Treatment (Diet)	Block								Total	Mean
	1 (Male, Age \geq 50, QUET > 3.5)	2 (Male, Age \geq 50, QUET ≤ 3.5)	3 (Male, Age $<$ 50, QUET > 3.5)	4 (Male, Age $<$ 50, QUET ≤ 3.5)	5 (Female, Age \geq 50, QUET > 3.5)	6 (Female, Age \geq 50, QUET ≤ 3.5)	7 (Female, Age $<$ 50, QUET > 3.5)	8 (Female, Age $<$ 50, QUET ≤ 3.5)		
1	11.2	6.2	16.5	8.4	14.1	9.5	21.5	13.2	100.6	12.58
2	9.3	4.1	14.2	6.9	14.2	8.9	15.2	10.1	82.9	10.36
3	10.4	5.1	14.0	6.2	11.1	8.4	17.3	11.2	83.7	10.46
4	9.0	4.9	13.7	6.1	11.8	8.4	15.9	9.7	79.5	9.94
Total	39.9	20.3	58.4	27.6	51.2	35.2	69.9	44.2	346.7	—
Mean	9.98	5.08	14.60	6.90	12.80	8.80	17.48	11.05	—	10.83

© Cengage Learning

are chosen to represent combinations of the various sex–age–body size categories of interest. Thus, in Table 18.4, block 1 consists of males at least 50 years of age with a quetelet index ($= 100[\text{weight}]/[\text{height}]^2$) above 3.5; block 3 consists of males under 50 with a quetelet index above 3.5; and so on. For each block of subjects, the four diets are randomly assigned to the sample of four persons in the block. Each subject is then followed for one year, after which the change in cholesterol level (Y) is recorded.

The primary research question of interest in this study concerns whether any of the observed average reductions in cholesterol level achieved by the four diets significantly differ. By inspecting Table 18.4, we can see that diet 1 appears to be the best, since it is associated with the largest average reduction (12.58 units). Nevertheless, this assessment needs to be evaluated statistically, which involves determining whether the observed differences among the mean reductions for the four diets can be attributed solely to chance. The randomized-blocks F test provides a method for making such a statistical evaluation.

Further inspection of Table 18.4 indicates that considerable differences exist among the block means. This is to be expected, since the reason for blocking was that different blocks (or, equivalently, different categories of the covariates age, sex, and body size) were expected to have different effects on the response. We can nevertheless perform a statistical test to satisfy ourselves that such block differences are statistically significant.

18.4.2 Hypotheses Tested in a Randomized-blocks Analysis

The primary null hypothesis of interest in a randomized-blocks analysis is the equality of treatment means:⁴

$$H_0: \mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{k\cdot} \quad (18.3)$$

⁴ Again, we are considering the null hypothesis for the fixed-effects case. In two-way ANOVA, this hypothesis would be stated differently if the treatment effects were considered random, but the test of hypothesis would be computed in exactly the same way.

where μ_i denotes the population mean response associated with the i th treatment. The alternative hypothesis may be stated as

H_A : "Not all the μ_i 's are equal."

If performing this test leads to rejection of H_0 , it becomes of interest to determine where the important differences are among the treatment effects. One qualitative way to make such a determination is simply to look at the observed treatment means and to make visual comparisons. Thus, from Table 18.4, rejecting the null hypothesis that no differences exist among the diets leads to the conclusion that diet 1 is best in reducing cholesterol level, that diets 2 and 3 are next in line and are about equally effective, and that diet 4 is the worst of the four. To verify such conclusions statistically, we can use multiple-comparison techniques to conduct appropriate tests, as described in Chapter 17.

As mentioned earlier, another null hypothesis sometimes of interest is that no significant differences exist among the block means. Since use of a randomized-blocks design is based on the a priori opinion that the block means are different, such a hypothesis is usually tested only to check whether the blocking was justified.

18.5 ANOVA Table for a Randomized-blocks Study

Table 18.5 summarizes the procedures necessary to test the null hypotheses of equality among the treatment means and among the block means. The computation of the ANOVA table sums of squares is routinely carried out by a computer program (e.g., SAS's GLM procedure). The ANOVA table values for the data given in Table 18.4 are shown in the accompanying SAS computer output.

From Table 18.5 and the SAS output, it is clear that the total (corrected) sum of squares, SSY, is split up into the three components—SST (treatments), SSB (blocks), and SSE (error)—using the fundamental equation

$$\text{SSY} = \text{SST}(\text{Treatments}) + \text{SSB}(\text{Blocks}) + \text{SSE}(\text{Error}) \quad (18.4)$$

TABLE 18.5 ANOVA table for a randomized-blocks study with k treatments and b blocks

Source	SS	d.f.	MS	F
Treatment	$k - 1$	SST	$MST = \frac{\text{SST}}{k - 1}$	$\frac{MST}{MSE}$
Blocks	$b - 1$	SSB	$MSB = \frac{\text{SSB}}{b - 1}$	$\frac{MSB}{MSE}$
Error	$(k - 1)(b - 1)$	SSE	$MSE = \frac{\text{SSE}}{(k - 1)(b - 1)}$	
Total	$kb - 1$	SSY		

Edited SAS Output (PROC GLM) for Data of Table 18.4

CLASS LEVEL INFORMATION		
Class	Levels	Values
trt	4	1 2 3 4
blk	8	1 2 3 4 5 6 7 8

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	496.4206250	49.6420625	52.89	<.0001
Error	21	19.7115625	0.9386458		MSE
Corrected Total	31	516.1321875		SSY	

R-Square	Coeff Var	Root MSE	y Mean
0.961809	8.942254	0.968837	10.83438

Source	DF	SST	MST	F Statistics
trt	3	33.5609375	11.1869792	11.92 <.0001
blk	7	462.8596875	66.1228125	70.44 <.0001

Source	DF	SSB	MSB	SSB	MSB
trt	3	33.5609375	11.1869792	11.92 <.0001	
blk	7	462.8596875	66.1228125	70.44 <.0001	

The formula for the total sum of squares is

$$SSY = \sum_{i=1}^k \sum_{j=1}^b (Y_{ij} - \bar{Y}_{..})^2 \quad (18.5)$$

As is the case in any ANOVA situation, SSY measures the total unexplained variation in the data, corrected for the grand mean. The degrees of freedom associated with SSY are $(kb - 1)$. From the SAS output, we see that SSY is 516.132.

The treatment sum of squares SST is defined as

$$SST = b \sum_{i=1}^k (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2 \quad (18.6)$$

SST reflects the variation among the observed treatment means. Its basic components are of the form $(\bar{Y}_{i\cdot} - \bar{Y}_{..})$, which is the difference between the i th treatment mean and the grand mean. SST has $(k - 1)$ degrees of freedom. For the SAS output, SST is computed as 33.561.

The block sum of squares SSB is defined as

$$\text{SSB} = k \sum_{j=1}^b (\bar{Y}_{.j} - \bar{Y}_{..})^2 \quad (18.7)$$

with $(b - 1)$ degrees of freedom. The basic components of SSB take the form $(\bar{Y}_{.j} - \bar{Y}_{..})$, which is the difference between the j th block mean and the grand mean. From the SAS output, we see that SSB is computed as 462.860.

Finally, the residual sum of squares SSE is given by

$$\text{SSE} = \sum_{i=1}^k \sum_{j=1}^b (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2 \quad (18.8)$$

with degrees of freedom equal to $(k - 1)(b - 1)$. For the SAS output, SSE is computed as 19.712.

The complexity of expression (18.8) indicates that SSE is not easily recognizable as an estimate of σ^2 . In fact, its basic component, $(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})$, can be written in the form

$$(Y_{ij} - \bar{Y}_{.j}) - (\bar{Y}_{i.} - \bar{Y}_{..})$$

which is an estimate of the difference between the effect of the i th treatment relative to the average effect of all treatments in the j th block (i.e., $\underline{Y}_{ij} - \bar{Y}_{.j}$) and the overall effect of the i th treatment relative to the overall mean (i.e. $\bar{Y}_{i.} - \bar{Y}_{..}$). Actually, SSE measures block-treatment interaction here, in the sense that SSE is large when the treatment effects vary from block to block. Although we discuss the concept of interaction more thoroughly in Chapter 19, using a randomized-blocks design requires us to assume that no such block-treatment interaction exists; under this assumption, the residual variation reflected in SSE can then be attributed solely to experimental error.⁵ The expected effect of violating the no-interaction assumption is a reduction in the power of the tests of treatment and block effects. This assumption must not be violated; otherwise, we would have no way of obtaining an unbiased estimator of σ^2 to use in the denominator of the F statistics. After all, for each block in a randomized-blocks design, no treatment is applied more than once, which makes it impossible to obtain a pure error estimate of σ^2 associated with a particular treatment in a given block. In the case of two-way ANOVA with more than one observation per cell, a pure error estimate of σ^2 can be developed by utilizing the available information on within-cell variability.

Each mean-square term in Table 18.5 is, as usual, obtained by dividing the corresponding sum-of-squares term by its degrees of freedom. The F statistics are formed as the ratios of mean-square terms, with MSE in the denominator in each case (regardless of whether each factor is treated as fixed or random).

⁵ A method for testing the validity of the assumption of no block-treatment interaction, developed by Tukey (1949), is described in Problem 3(f) at the end of this chapter.

18.5.1 *F* Test for the Equality of Treatment Means

The test of $H_0: \mu_{1\cdot} = \mu_{2\cdot} = \dots = \mu_{k\cdot}$ is performed by using the following *F* statistic (defined in terms of the notation of Table 18.3):

$$F = \frac{\text{MST}}{\text{MSE}} \quad (18.9)$$

where

$$\text{MST} = \frac{1}{k-1}(\text{SST}) = \frac{1}{k-1}\left(\frac{1}{b} \sum_{i=1}^k T_i^2 - \frac{G^2}{bk}\right)$$

and

$$\text{MSE} = \frac{1}{(k-1)(b-1)}(\text{SSE})$$

The assumptions required for the validity of this test are as follows:

1. The observations are statistically independent of one another.
2. Each observation is selected from a normally distributed population.
3. Each observation is selected from a population with variance σ^2 (i.e., variance homogeneity is assumed).
4. There is no block-treatment interaction effect (i.e., the true extent to which treatments differ is the same, regardless of the block considered).

If H_0 is true, the *F* statistic in (18.9) has the *F* distribution with $(k-1)$ degrees of freedom in the numerator and $(k-1)(b-1)$ degrees of freedom in the denominator. Thus, for a given α , we would reject H_0 and conclude that not all treatments have the same effect on the response when

$$F \geq F_{k-1, (k-1)(b-1), 1-\alpha}$$

From the SAS output shown in Section 18.5, we see that MST is equal to 11.187 and MSE is equal to 0.939, so

$$F = \frac{11.187}{0.939} = 11.914$$

Since $F_{3, 21, 0.999} = 7.94$, the *P*-value for this test satisfies $P < .001$. Thus, we reject the null hypothesis H_0 and conclude that significant differences exist among the four diets.

One might ask, “How is the statistic in (18.9) different from that found in (17.2) for one-way ANOVA?” Though they are functionally the same, the difference lies in the MSE used. In a randomized-blocks study that controls for one or more important factors, the MSE is reduced because the blocks factor has “absorbed” some of the variability in the observed responses, thereby increasing statistical precision.

18.5.2 *F* Test for the Equality of Block Means

As previously mentioned, this test is rarely used except as an a posteriori check to confirm that blocking was effective. To perform this test, we calculate the following *F* statistic:

$$F = \frac{\text{MSB}}{\text{MSE}} \quad (18.10)$$

where

$$\text{MSB} = \frac{1}{b-1}(\text{SSB}) = \frac{1}{b-1}\left(\frac{1}{k} \sum_{j=1}^b B_j^2 - \frac{G^2}{bk}\right)$$

and where MSE is calculated as before.

Under the null hypothesis H_0 : “There are no differences among the true block means,” the *F* statistic (18.10) has the *F* distribution with $(b-1)$ and $(k-1)(b-1)$ degrees of freedom. Thus, H_0 is rejected at significance level α when F exceeds $F_{b-1, (k-1)(b-1), 1-\alpha}$. As shown in the SAS output in Section 18.5, this test yields the following *F* statistic:

$$F = \frac{66.123}{0.939} = 70.419$$

The *P*-value for this test satisfies $P < .001$, so H_0 is rejected, as expected.

18.6 Regression Models for a Randomized-blocks Study

The randomized-blocks study, as for a one-way ANOVA study, can be described by either a regression model or a classical ANOVA effects model. The regression formulation, which we consider in this section, is technically equivalent to the fixed-effects ANOVA approach in terms of estimating the unknown parameters in each model. Nevertheless, for testing purposes, mean-square terms in the ANOVA table obtained from the fit of a proper regression model can be used to compute appropriate *F* statistics, regardless of whether the actual ANOVA model involves fixed and/or random effects.

An appropriate regression model for the randomized-blocks study should contain $(k-1)$ dummy variables for the k treatments and $(b-1)$ dummy variables for the b blocks. One such model formulation (using effect coding) is

$$Y = \mu + \sum_{i=1}^{k-1} \alpha_i X_i + \sum_{j=1}^{b-1} \beta_j Z_j + E \quad (18.11)$$

where

$$X_i = \begin{cases} -1 & \text{if treatment } k \\ 1 & \text{if treatment } i \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad Z_j = \begin{cases} -1 & \text{if block } b \\ 1 & \text{if block } j \\ 0 & \text{otherwise} \end{cases}$$

(for $i = 1, 2, \dots, k - 1$; and $j = 1, 2, \dots, b - 1$).

As in one-way ANOVA, other coding schemes for the independent variables are possible. For example, we may let

$$X_i = \begin{cases} 1 & \text{if treatment } i \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad Z_j = \begin{cases} 1 & \text{if block } j \\ 0 & \text{otherwise} \end{cases}$$

(for $i = 1, 2, \dots, k - 1$; and $j = 1, 2, \dots, b - 1$). For any such coding scheme, the F tests obtained from fitting the regression model are exactly equivalent to the randomized-blocks ANOVA F tests described previously. The regression coefficients, however, represent different functions of the cell population means.

As in the one-way ANOVA situation described in Chapter 17, the regression coefficients in model (18.11) may be expressed in terms of underlying cell (i.e., block-treatment combination) means. To see this, consider the matrix of population cell means associated with the general randomized-blocks layout presented in Table 18.6. In this table, the (cell) mean for the i th treatment in the j th block is denoted by μ_{ij} ; the mean for the i th treatment (averaged over the b blocks) is denoted by $\mu_{i\cdot}$; the mean for the j th block (averaged over the k treatments) is denoted by $\mu_{\cdot j}$; and $\mu_{\cdot\cdot}$ denotes the overall mean. Hence, $\mu_{i\cdot}$, $\mu_{\cdot j}$, and $\mu_{\cdot\cdot}$ satisfy

$$\mu_{i\cdot} = \frac{1}{b} \sum_{j=1}^b \mu_{ij}, \quad \mu_{\cdot j} = \frac{1}{k} \sum_{i=1}^k \mu_{ij}, \quad \mu_{\cdot\cdot} = \frac{1}{bk} \sum_{i=1}^k \sum_{j=1}^b \mu_{ij}$$

For model (18.11), the coefficients α_i and β_j can be expressed as follows:

$$\begin{aligned} \mu &= \mu_{\cdot\cdot} \\ \alpha_i &= \mu_{i\cdot} - \mu_{\cdot\cdot} \quad i = 1, 2, \dots, k - 1 \\ \beta_j &= \mu_{\cdot j} - \mu_{\cdot\cdot} \quad j = 1, 2, \dots, b - 1 \\ - \sum_{i=1}^{k-1} \alpha_i &= \mu_{k\cdot} - \mu_{\cdot\cdot} \\ - \sum_{j=1}^{b-1} \beta_j &= \mu_{\cdot b} - \mu_{\cdot\cdot} \end{aligned} \tag{18.12}$$

Thus, the coefficient μ represents the overall mean, the coefficient α_i represents the difference between the i th treatment mean and the overall mean, and the coefficient β_j represents the difference between the j th block mean and the overall mean. Furthermore, the negative sum of the α_i (i.e., $-\sum_{i=1}^{k-1} \alpha_i$) gives the difference between the k th treatment mean and the overall mean, and the negative sum of the β_j (i.e., $-\sum_{j=1}^{b-1} \beta_j$) gives the difference between the b th block mean and the overall mean.

TABLE 18.6 Matrix of cell means for a randomized-blocks layout

Treatment (i)	Block (j)					Mean
	1	2	3	...	b	
1	μ_{11}	μ_{12}	μ_{13}	...	μ_{1b}	$\mu_{1\cdot}$
2	μ_{21}	μ_{22}	μ_{23}	...	μ_{2b}	$\mu_{2\cdot}$
3	μ_{31}	μ_{32}	μ_{33}	...	μ_{3b}	$\mu_{3\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	μ_{k1}	μ_{k2}	μ_{k3}	...	μ_{kb}	$\mu_{k\cdot}$
Mean	$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	$\mu_{\cdot 3}$...	$\mu_{\cdot b}$	$\mu_{\cdot \cdot}$

© Cengage Learning

These results can also be used to express the cell, treatment, and block means as a function of the regression parameters. The cell means are

$$\mu_{ij} = \mu + \alpha_i + \beta_j \quad i = 1, 2, \dots, k - 1; j = 1, 2, \dots, b - 1$$

$$\mu_{kj} = \mu - \sum_{i=1}^{k-1} \alpha_i + \beta_j \quad j = 1, 2, \dots, b - 1$$

$$\mu_{ib} = \mu + \alpha_i - \sum_{j=1}^{b-1} \beta_j \quad i = 1, 2, \dots, k - 1$$

$$\mu_{kb} = \mu - \sum_{i=1}^{k-1} \alpha_i - \sum_{j=1}^{b-1} \beta_j$$

In turn, the treatment means are

$$\mu_{i\cdot} = \mu + \alpha_i \quad i = 1, 2, \dots, k - 1$$

$$\mu_{k\cdot} = \mu - \sum_{i=1}^{k-1} \alpha_i$$

and the block means are

$$\mu_{\cdot j} = \mu + \beta_j \quad j = 1, 2, \dots, b - 1$$

$$\mu_{\cdot b} = \mu - \sum_{j=1}^{b-1} \beta_j$$

Similar results can be obtained by using reference cell coding. The model (18.11) remains the same, but the dummy variables change as follows:

$$X_i = \begin{cases} 1 & \text{if treatment } i \\ 0 & \text{otherwise} \end{cases} \quad i = 1, 2, \dots, k - 1$$

$$Z_j = \begin{cases} 1 & \text{if block } j \\ 0 & \text{otherwise} \end{cases} \quad j = 1, 2, \dots, b - 1$$

Although the same parameter symbols are used, their meanings change. In particular, the cell means (for reference cell coding) are

$$\mu_{ij} = \mu + \alpha_i + \beta_j \quad i = 1, 2, \dots, k - 1; j = 1, 2, \dots, b - 1$$

$$\mu_{kj} = \mu + \beta_j \quad j = 1, 2, \dots, b - 1$$

$$\mu_{ib} = \mu + \alpha_i \quad i = 1, 2, \dots, k - 1$$

$$\mu_{kb} = \mu$$

The treatment means are

$$\mu_{i\cdot} = \mu + \alpha_i + \sum_{j=1}^{b-1} \frac{\beta_j}{b} \quad i = 1, 2, \dots, k - 1$$

$$\mu_{k\cdot} = \mu + \sum_{j=1}^{b-1} \frac{\beta_j}{b}$$

and the block means are

$$\mu_{\cdot j} = \mu + \beta_j + \sum_{i=1}^{k-1} \frac{\alpha_i}{k} \quad j = 1, 2, \dots, b - 1$$

$$\mu_{\cdot b} = \mu + \sum_{i=1}^{k-1} \frac{\alpha_i}{k}$$

These results demonstrate that estimates of means, and contrasts between means, can be expressed as linear combinations of estimated regression coefficients. (*Caution:* When using a computer package to estimate effects in ANOVA, make sure you understand exactly what parameters the program is estimating.)

As with one-way ANOVA, the F tests resulting from the fit of regression model (18.11), or from any other properly coded regression model used for testing the hypotheses of equality of treatment means and of equality of block means, are exactly the same as those obtained for the ANOVA procedures presented earlier. The multiple partial F test of $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_{k-1} = 0$ under model (18.11) provides exactly the same F -value as that given by

$$F = \frac{MST}{MSE}$$

under (18.9). Similarly, the multiple partial F test of $H_0: \beta_1 = \beta_2 = \dots = \beta_{b-1} = 0$ in model (18.11) yields exactly the same F -value as is given by

$$F = \frac{MSB}{MSE}$$

under (18.10).

Thus, it does not matter which, if any, regression model (i.e., coding scheme) we use if we are interested only in performing the preceding global F tests. Furthermore, if we want to make certain specific comparisons of means, we can always calculate these comparisons directly without using a regression model.

18.7 Fixed-effects ANOVA Model for a Randomized-blocks Study

If both the block effects and the treatment effects are considered fixed,⁶ a classical ANOVA model may be written in terms of these effects, as was done in the one-way ANOVA case. The effects for a randomized-blocks study are defined as differences between a given treatment mean and the overall mean (i.e., $\mu_{i\cdot} - \mu_{..}$) and as differences between a given block mean and the overall mean (i.e., $\mu_{\cdot j} - \mu_{..}$). The fixed-effects model may be written in the form

$$Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij} \quad i = 1, 2, \dots, k; j = 1, 2, \dots, b \quad (18.13)$$

where

Y_{ij} = Observed response associated with the i th treatment in the j th block

$\mu = \mu_{..}$ = Grand (overall) mean

$\alpha_i = \mu_{i\cdot} - \mu_{..}$ = Effect of treatment i

$\beta_j = \mu_{\cdot j} - \mu_{..}$ = Effect of block j

$E_{ij} = Y_{ij} - \mu - \alpha_i - \beta_j$ = Error component associated with the i th treatment in the j th block

In model (18.13), the effect of any given treatment is the same, regardless of the block; and similarly, the effect of any given block is the same, regardless of the treatment. In other words, there is no block-treatment interaction. Thus, to determine the mean response for a given cell, we need only know the treatment effect and block effect associated with that cell—and not the particular contribution of the cell itself. A model incorporating block-treatment interaction would be of the form

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ij}$$

which contains the (interaction) term γ_{ij} specific to cell (i, j) .

A few additional properties of the fixed-effects model (18.13) are worth noting. First, it can be shown that

$$\sum_{i=1}^k \alpha_i = 0 \quad \text{and} \quad \sum_{j=1}^b \beta_j = 0$$

since $\alpha_i = (\mu_{i\cdot} - \mu_{..})$ and $\beta_j = (\mu_{\cdot j} - \mu_{..})$.

Another property of interest is that the various treatment and block effects can be estimated from the data, as follows:

$$\hat{\alpha}_i = (\bar{Y}_{i\cdot} - \bar{Y}_{..}) \quad \text{and} \quad \hat{\beta}_j = (\bar{Y}_{\cdot j} - \bar{Y}_{..})$$

⁶ Random-effects models for two-way ANOVA are discussed in Chapter 19.

A final property of the fixed-effects model (18.13) is that it corresponds in structure to the regression model given by (18.11); that is, the coefficients α_1 through α_{k-1} of model (18.11) correspond to the first $(k - 1)$ treatment effects in model (18.13). Similarly, the coefficients β_1 through β_{b-1} correspond to the first $(b - 1)$ block effects in model (18.13). Finally, the negative sums $-\sum_{i=1}^{k-1} \alpha_i$ and $-\sum_{j=1}^{b-1} \beta_j$ represent the effects of the k th treatment and the b th block, respectively.

Problems

- 1.** A private research corporation conducted an experiment to investigate the toxic effects of three chemicals (I, II, and III) used in the tire-manufacturing industry. In this experiment, three adjacent 1-inch squares were marked on the back of each of eight rats, and the three chemicals were applied separately to the three squares on each rat. The squares were then rated from 0 to 10, depending on the degree of irritation. The data are as shown in the following table.

Chemical	Rat								Total
	1	2	3	4	5	6	7	8	
I	6	9	6	5	7	5	6	6	50
II	5	9	9	8	8	7	7	7	60
III	3	4	3	6	8	5	5	6	40
Total	14	22	18	19	23	17	18	19	150

- a. What are the blocks and what are the treatments in this randomized-blocks design?
- b. Complete the ANOVA table in the accompanying computer output for the data set given.
- c. Do the data provide sufficient evidence to indicate a significant difference in the toxic effects of the three chemicals?
- d. Using a confidence interval of the form

$$(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) \pm t_{v, 1-\alpha/2} \sqrt{\text{MSE} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

where v is the degrees of freedom, find a 98% confidence interval for the true difference in the toxic effects of chemicals I and II.

- e. Provide a reasonable measure of the proportion of total variation that is explained by the particular statistical model used in analyzing this data set.
- f. State the fixed-effects ANOVA model and the corresponding regression model for this analysis.
- g. State the assumptions on which the validity of the analysis depends.

Edited SAS Output (PROC GLM) for Problem 1

Dependent Variable: score

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	43.50000000	4.83333333	2.71	0.0463
Error	14	_____	_____		
Corrected Total	23	68.50000000			
R-Square					
		Coeff Var	Root MSE	score Mean	
		0.635036	21.38090	1.336306	6.250000
Source	DF	Type I SS	Mean Square	F Value	Pr > F
chem	2	_____	_____	7.00	0.0078
rat	7	18.50000000	2.64285714	1.48	0.2518
Source	DF	Type III SS	Mean Square	F Value	Pr > F
chem	2	25.00000000	12.50000000	7.00	0.0078
rat	7	18.50000000	2.64285714	1.48	0.2518

2. In a study of the psychosocial changes in individuals who participated in a community-based intervention program, individuals who were clinically identified as hypertensives were randomly assigned to one of two treatment groups (with n individuals in each group). Group 1 was given the usual care that existing community facilities provide for hypertensives; group 2 was given the special care that the intervention study team provided. Among the variables measured on each individual were SP1, an index of the individual's self-perception of health immediately after being identified as hypertensive but before being assigned to one of the two groups; SP2, an index of the individual's self-perception of health one year after assignment to one of the two groups; AGE; and SEX. One main research question of interest was whether the change in self-perception of health after one year would be greater for individuals in group 2 than for those in group 1. Several different analytical approaches to answering this question are possible, depending on the dependent variable chosen and on the treatment of the variables SP1, AGE, and SEX in the analysis. Among these approaches are the following six:
- Matching pairwise on AGE and SEX (which we assume is possible) and then performing a paired-difference t test to determine whether the mean of group 1 change scores significantly differs from the mean of group 2 change scores. (*Note:* This is equivalent to performing a randomized-blocks analysis, where the blocks are the pairs of individuals.)
 - Matching pairwise on AGE and SEX and then performing a regression analysis, with the change score $Y = (SP2 - SP1)$ as the dependent variable and with SP1 as one of the independent variables.

- Matching pairwise on AGE and SEX and then performing a regression analysis, with SP2 as the dependent variable and with SP1 as one of the independent variables.
 - Controlling for AGE and SEX (without any prior matching) via analysis of covariance, with the change score $Y = (SP2 - SP1)$ as the dependent variable.
 - Controlling for AGE and SEX via analysis of covariance, with the change score $Y = (SP2 - SP1)$ as the dependent variable and with SP1 as one of the independent variables.
 - Controlling for AGE and SEX via analysis of covariance, with SP2 as the dependent variable and with SP1 as one of the independent variables.
- a. What regression model is associated with each of the preceding six approaches? Make sure to define your variables carefully.
- b. For each of the preceding regression models, state the appropriate null hypothesis (in terms of regression coefficients) for testing for group differences with respect to self-perception scores. For each regression model, indicate how to set up the appropriate ANOVA table to carry out the desired test.
- c. Assuming that you have decided to match pairwise on AGE and SEX, which of the preceding regression models would you prefer to use, and why?
3. An experiment was conducted at the University of North Carolina to see whether the *biochemical oxygen demand* (BOD) test for water pollution is biased by the presence of copper. In this test, the amount of dissolved oxygen in a sample of water is measured at the beginning and at the end of a five-day period; the difference in dissolved oxygen content is ascribed to the action of bacteria on the impurities in the sample and is called BOD. The question is whether dissolved copper retards the bacterial action and produces artificially low responses for the test.

The data in the following table are partial results from this experiment. The three samples (which are from different sources) are split into five subsamples, and the concentration of copper ions in each subsample is given. The BOD measurements are given for each subsample–copper ion concentration combination.

Sample	Copper Ion Concentration (ppm)					Mean
	0	0.1	0.3	0.5	0.75	
1	210	195	150	148	140	168.60
2	194	183	135	125	130	153.40
3	138	98	89	90	85	100.00
Mean	180.67	158.67	124.67	121.00	118.33	140.67

- a. Using dummy variables and treating the copper ion concentration as a nominal variable, provide an appropriate regression model for this experiment. Is this a randomized-blocks design?
- b. If copper ion concentration is treated as an interval (continuous) variable, one appropriate regression model would be

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 X + \beta_4 Z_1 X + \beta_5 Z_2 X + E$$

where

$$Z_1 = \begin{cases} 1 & \text{for sample 1} \\ 0 & \text{for sample 2} \\ -1 & \text{for sample 3} \end{cases} \quad Z_2 = \begin{cases} 0 & \text{for sample 1} \\ 1 & \text{for sample 2} \\ -1 & \text{for sample 3} \end{cases}$$

and X = copper ion concentration. What are the advantages and disadvantages of using the models in parts (a) and (b)? Which model would you prefer to use, and why?

- c. Compare (without doing any statistical tests) the average BOD responses at the various levels of copper ion concentration.
- d. Use the following table, which is based on a randomized-blocks analysis, to test (at $\alpha = .05$) the null hypothesis that copper ion concentration has no effect on the BOD test.

Source	d.f.	SS	MS	F
Samples	2	12,980.9333	6,490.4667	56.83
Concentrations	4	9,196.6667	2,299.1667	20.13
Error	8	913.7333	114.2167	

- e. Judging from the ANOVA table and the observed block means, does blocking appear to be justified?
- f. The randomized-blocks analysis assumes that the relative differences in BOD responses at different levels of copper ion concentration are the same, regardless of the sample used; in other words, there is no copper ion concentration–sample interaction. One method (see Tukey 1949) for testing whether such an interaction effect actually exists is *Tukey's test for additivity*. It addresses the null hypothesis

H_0 : No interaction exists (i.e., the model is additive in the block and treatment effects).

versus the alternative hypothesis

H_A : The model is not additive, and a transformation $f(Y)$ exists that removes the nonadditivity in the model for Y .

Tukey's test statistic is given by

$$F = \frac{\text{SSN}}{(\text{SSE} - \text{SSN})/[(k-1)(b-1)-1]}$$

(which is distributed as $F_{1, (k-1)(b-1)-1}$ under H_0), where

$$\text{SSN} = \frac{\left[\sum_{i=1}^k \sum_{j=1}^b Y_{ij} (\bar{Y}_{i\cdot} - \bar{Y}_{..})(\bar{Y}_{\cdot j} - \bar{Y}_{..}) \right]^2}{\sum_{i=1}^k (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2 \sum_{j=1}^b (\bar{Y}_{\cdot j} - \bar{Y}_{..})^2}$$

using the notation in this chapter.

Given that the computed F statistic equals 4.37 in Tukey's test for additivity, is there significant evidence of nonadditivity? (Use $\alpha = .05$.)

- g. The following table gives results from fitting the multiple regression model given in part (b). Use this table to test whether evidence exists of a significant effect due to copper ion concentration (i.e., test $H_0: \beta_3 = \beta_4 = \beta_5 = 0$). Use the result that $R^2 = 0.888$.

Source	d.f.	MS
Z_1	1	11,764.90
$Z_2 Z_1$	1	1,216.03
$X Z_1, Z_2$	1	7,134.57
$Z_1 X Z_1, Z_2, X$	1	303.27
$Z_2 X Z_1, Z_2, X, Z_1 X$	1	91.07
Error	9	286.83

4. Using the accompanying computer output based on the data in Problem 7 in Chapter 17, conduct a randomized-blocks analysis, treating the high schools as blocks, to test whether significant differences exist among the average math SAT (MSAT) scores for the years 2002, 2007, and 2012. Do the results obtained from this randomized-blocks analysis differ from those obtained earlier from the one-way ANOVA?

Edited SAS Output (PROC GLM) for Problem 4

Dependent Variable: MSAT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1875.200000	312.533333	19.76	0.0002
Error	8	126.533333	15.816667		
Corrected Total	14	2001.733333			
R-Square	Coeff Var	Root MSE	vsat Mean		
0.936788	0.734490	3.977017	541.4667		
Source	DF	Type I SS	Mean Square	F Value	Pr > F
year	2	548.133333	274.066667	17.33	0.0012
school	4	1327.066667	331.766667	20.98	0.0003
Source	DF	Type III SS	Mean Square	F Value	Pr > F
year	2	548.133333	274.066667	17.33	0.0012
school	4	1327.066667	331.766667	20.98	0.0003

5. Using the accompanying computer output based on the data in Problem 8 in Chapter 17, conduct a randomized-blocks analysis, treating the days as blocks, to test whether the three persons have significantly different ESP ability. Does blocking on days seem appropriate?

Edited SAS Output (PROC GLM) for Problem 5

Dependent Variable: SCORE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	111.46666667	18.57777778	6.12	0.0113
Error	8	24.26666667	3.0333333		
Corrected Total	14	135.7333333			
R-Square		Coeff Var	Root MSE	score Mean	
0.821218		9.231343	1.741647	18.86667	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
person	2	91.73333333	45.86666667	15.12	0.0019
day	4	19.73333333	4.93333333	1.63	0.2585
Source	DF	Type III SS	Mean Square	F Value	Pr > F
person	2	91.73333333	45.86666667	15.12	0.0019
day	4	19.73333333	4.93333333	1.63	0.2585

6. The promotional policies of four companies in a certain industry were compared to determine whether their rates for promoting blacks and whites differed. Data on the variable “rate discrepancy” were obtained for the four companies in each of three different two-year periods. This variable was defined as

$$d = \hat{p}_W - \hat{p}_B$$

where

$$\hat{p}_W = \frac{\text{(Number of whites promoted)}(100)}{\text{Number of whites eligible for promotion}}$$

and

$$\hat{p}_B = \frac{\text{(Number of blacks promoted)}(100)}{\text{Number of blacks eligible for promotion}}$$

The resulting data are reproduced in the following table.

Period	Company			
	1	2	3	4
1	3	5	5	4
2	4	4	3	5
3	8	12	10	9

- a. Using dummy variables, create an appropriate regression model for this data set. Is this a randomized-blocks design?
- b. Use the following table to test whether any significant differences exist among the average rate discrepancies for the four companies. Tukey's F for testing additivity equals 3.583 with 1 and 5 degrees of freedom.
- | Source | d.f. | SS | MS | F |
|-----------|------|---------|---------|--------|
| Periods | 2 | 84.5000 | 42.2500 | 33.800 |
| Companies | 3 | 6.0000 | 2.0000 | 1.600 |
| Error | 6 | 7.5000 | 1.2500 | |
- c. Does Tukey's test for these data indicate that a removable interaction effect is present?
- d. If no significant differences are found among the rate discrepancies for the four companies, would this support the contention that none of the companies has a discriminatory promotional policy?
- e. Comment on the suitability of this analysis in view of the fact that the response variable is a difference in proportions.
7. Suppose that, in a study to compare body sizes of three genotypes of fourth-instar silkworms, the mean lengths (in millimeters) for separately reared cocoons of heterozygous (HET), homozygous (HOM), and wild (WLD) silkworms were determined at five laboratory sites, as detailed in the following table and in the accompanying computer output.⁷

Variable	Site				
	1	2	3	4	5
HOM	29.87	28.16	32.08	30.84	29.44
HET	32.51	30.82	34.17	33.46	32.99
WLD	35.76	33.14	36.29	34.95	35.89

- a. Assuming that this is a randomized-blocks design, what are the blocks and what are the treatments?
- b. Why do you think a randomized-blocks analysis is appropriate (or inappropriate) for this experiment?
- c. Carry out an appropriate analysis of the data for this experiment, and state your conclusions. Be sure to state the null hypothesis for each test performed.

Edited SAS Output (PROC GLM) for Problem 7

Dependent Variable: LENGTH

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	84.76168000	14.12694667	46.97	<.0001
Error	8	2.40629333	0.30078667		
Corrected Total	14	87.16797333			

(continued)

⁷ Adapted from a study by Sokal and Karlen (1964).

R-Square	Coeff Var	Root MSE	length Mean		
0.972395	1.677632	0.548440	32.69133		
Source	DF	Type I SS	Mean Square	F Value	Pr > F
variable	2	65.81397333	32.90698667	109.40	<.0001
site	4	18.94770667	4.73692667	15.75	0.0007
Source	DF	Type III SS	Mean Square	F Value	Pr > F
variable	2	65.81397333	32.90698667	109.40	<.0001
site	4	18.94770667	4.73692667	15.75	0.0007

8. In Problem 3, the independent variable of interest, copper ion concentration, is an interval-scaled variable.
- Treating copper ion concentration as a five-level categorical predictor (as in the randomized-blocks analysis) corresponds to including a collection of polynomial terms for copper ion concentration (e.g., X, X^2, \dots) in the model. What is the highest order power of X allowed?
 - State, in terms of k levels of a predictor, a general principle based on the example in part (a).
 - State the expansion of the regression model given in Problem 3(b) that encompasses the polynomial terms discussed in part (a).
 - Use a computer program to fit the model stated in part (c). Center X by subtracting 0.330. (Why?) Provide estimated regression coefficients.
 - Provide a multiple partial F test comparing the model in parts (b) and (g) of Problem 3 to the one fitted here; the latter is equivalent to the model considered in part (d) of Problem 3.
9. Assume that the following table came from an analysis of a randomized-blocks study.

Source	d.f.	SS	MS	F
Treatments	4	b	e	5.00
Blocks	a	c	48.00	6.00
Error	20	d	f	

- a–f. Complete the ANOVA table by determining the numerical values for the entries a through f . Show your work.
- g. Provide tests concerning treatment and block effects using $\alpha = .05$. What are your conclusions?
10. An educator has obtained class average scores on a standardized test for classes from each grade from first through seventh at each of four schools in a city. These are shown in the following table. To assess differences among schools, the educator decides to conduct a randomized-blocks ANOVA, treating grade as a blocking factor.

Observation	School	Grade	Score
1	1	1	70.4
2	1	2	74.5
3	1	3	49.6
4	1	4	72.9
5	1	5	60.3
6	1	6	59.5
7	1	7	77.0
8	2	1	73.8
9	2	2	71.0
10	2	3	65.0
11	2	4	73.2
12	2	5	79.7
13	2	6	82.5
14	2	7	60.9
15	3	1	71.3
16	3	2	62.7
17	3	3	71.9
18	3	4	99.4
19	3	5	74.4
20	3	6	82.8
21	3	7	74.4
22	4	1	68.5
23	4	2	79.5
24	4	3	72.1
25	4	4	77.0
26	4	5	86.5
27	4	6	101.8
28	4	7	100.7

- Conduct the appropriate analysis, using the accompanying computer output.
- Provide F tests about school and grade differences, using $\alpha = .05$. What do you conclude?
- Why might using the children's individual scores be better than using mean scores?

Edited SAS Output (PROC GLM) for Problem 10

Dependent Variable: SCORE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	2005.678214	222.853135	2.07	0.0905
Error	18	1939.268571	107.737143		
Corrected Total	27	3944.946786			

R-Square	Coeff Var	Root MSE	score Mean
0.508417	13.88383	10.37965	74.76071

Source	DF	Type I SS	Mean Square	F Value	Pr > F
school	3	1131.063929	377.021310	3.50	0.0370
grade	6	874.614286	145.769048	1.35	0.2857

(continued)

Source	DF	Type III SS	Mean Square	F Value	Pr > F
school	3	1131.063929	377.021310	3.50	0.0370
grade	6	874.614286	145.769048	1.35	0.2857

11. A company evaluated three new production-line technologies, one involving high automation, the second involving moderate automation, and the third involving low automation. All three technologies were intended to improve productivity through increased automation, but they continued to require operator intervention. Three groups were identified, each consisting of three line operators with similar years of production-line experience. Within each group, the operators were randomly assigned to one of the three new production-line technologies. The output (in units per hour) was recorded for each operator and is shown in the accompanying table.

Operator Experience	Technology		
	High Automation (H)	Moderate Automation (M)	Low Automation (L)
<1 Year	10	8	5
1–2 Years	18	12	10
>2 Years	19	13	12

- a. What are the blocks and what are the treatments in this randomized-blocks design?
- b. Complete the ANOVA table in the accompanying computer output for the data set given.
- c. Do the data provide sufficient evidence to indicate that any significant differences exist among the outputs for the three technologies?
- d. Using a confidence interval of the form

$$(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) \pm t_{v, 1-\alpha/2} \sqrt{\text{MSE} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

where v represents the degrees of freedom, find a 99% confidence interval for the true difference in average output between the high and moderate automation lines.

Edited SAS Output (PROC GLM) for Problem 11

CLASS LEVEL INFORMATION		
Class	Levels	Values
LINE	3	H L M
EXPER	3	1-2 <1 >2

Number of Observations Read	9
Number of Observations Used	9

(continued)

Dependent Variable: OUTPUT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	153.1111111	38.2777778	_____	0.0039
Error	4	_____	_____	_____	_____
Corrected Total	8	158.8888889	_____	_____	_____
R-Square		Coeff Var	Root MSE	OUTPUT Mean	
0.963636		10.10902	1.201850	11.88889	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
LINE	2	70.22222222	35.11111111	_____	_____
EXPER	2	82.88888889	41.44444444	_____	_____
Source	DF	Type III SS	Mean Square	F Value	Pr > F
LINE	2	70.22222222	35.11111111	_____	_____
EXPER	2	82.88888889	41.44444444	_____	_____
Level of LINE	N	OUTPUT			
		Mean	Std Dev		
H	3	15.6666667	4.93288286	_____	_____
L	3	9.0000000	3.60555128	_____	_____
M	3	11.0000000	2.64575131	_____	_____
Level of EXPER	N	OUTPUT			
		Mean	Std Dev		
1-2	3	13.3333333	4.16333200	_____	_____
<1	3	7.6666667	2.51661148	_____	_____
>2	3	14.6666667	3.78593890	_____	_____

12. An advertising company evaluated three types of television ads for a new low-cost subcompact automobile: visual-appeal ads, budget-appeal ads, and feature-appeal ads. To control for age differences, the company randomly selected viewers from four age groups to evaluate the persuasiveness of the ads (measured on a scale ranging from 1 to 10, with 1 being the lowest level of persuasion and 10 the highest). Within each age group, each of three viewers was randomly assigned to view one of the three types of ads. The sample persuasion scores are reproduced in the following table and processed as shown in the accompanying computer output.

Viewer Age	Type of Ad		
	Visual Appeal (V)	Budget Appeal (B)	Feature Appeal (F)
18–25 Years	7	8	5
26–35 Years	6	9	4
36–45 Years	8	9	4
46 and Older	10	10	5

Edited SAS Output (PROC GLM) for Problem 12

CLASS LEVEL INFORMATION			
Class	Levels	Values	
AD	3	B F V	
AGE	4	18-25 26-35 36-45 >45	

Number of Observations Read	12
Number of Observations Used	12

Dependent Variable: SCORE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	50.08333333	10.01666667	_____	0.0040
Error	6	_____	_____	_____	_____
Corrected Total	11	54.91666667	_____	_____	_____

R-Square	Coeff Var	Root MSE	SCORE Mean
0.911988	12.67098	0.897527	7.083333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
AD	2	43.16666667	21.58333333	_____	_____
AGE	3	6.91666667	2.30555556	_____	_____

Source	DF	Type III SS	Mean Square	F Value	Pr > F
AD	2	43.16666667	21.58333333	_____	_____
AGE	3	6.91666667	2.30555556	_____	_____

Level of AD	N	SCORE	
		Mean	Std Dev
B	4	9.00000000	0.81649658
F	4	4.50000000	0.57735027
V	4	7.75000000	1.70782513

Level of AGE	N	SCORE	
		Mean	Std Dev
18-25	3	6.66666667	1.52752523
26-35	3	6.33333333	2.51661148
36-45	3	7.00000000	2.64575131
>45	3	8.33333333	2.88675135

- a. Identify the blocks and treatments in this randomized-blocks design.
- b. Complete the ANOVA table for the data set given.
- c. Do the data provide sufficient evidence to indicate significant differences among the persuasion scores for the three types of ads?
- d. Using a confidence interval of the form

$$(\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}) \pm t_{v, 1-\alpha/2} \sqrt{\text{MSE} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

where v is the error degrees of freedom, find a 99% confidence interval for the true difference in average output between the visual-appeal and budget-appeal ads.

- e. Repeat part (d) for visual-appeal and feature-appeal ads.
- f. Repeat part (d) for budget-appeal and feature-appeal ads.
- g. Based on your results from parts (b)–(f), what do you conclude about the average persuasion scores for the different types of ads?

References

- Armitage, P. 1971. *Statistical Methods in Medical Research*. Oxford: Blackwell Scientific.
- Ostle, B. 1963. *Statistics in Research*, Second Edition. Ames, Iowa: Iowa State University Press.
- Peng, K. C. 1967. *Design and Analysis of Scientific Experiments*. Reading, Mass.: Addison-Wesley.
- Sokal, R. R., and Karlen, I. 1964. "Competition among Genotypes in *Tibolium castaneum* at Varying Densities and Gene Frequencies (the Black Locus)." *Genetics* 49: 195–211.
- Tukey, J. W. 1949. "One Degree of Freedom for Nonadditivity." *Biometrics* 5: 232.

19

Two-way ANOVA with Equal Cell Numbers

19.1 Preview

In this chapter, we shall consider the analysis of the simplest two-way data layout involving more than one observation per cell—namely, the layout for which the number of observations in any given cell is at least two and is exactly the same as in any other cell (i.e., the data are considered “balanced”; see Figure 18.1(b)). As we will see in this chapter, having equal cell numbers makes for a straightforward analysis requiring only slightly more involved calculations than those used for a randomized-blocks experiment. On the other hand, having unequal cell numbers leads to a much more complicated analysis (see Chapter 20).

Two-way layouts with equal cell numbers are rarely seen in observational studies, but they are often generated intentionally in experimental studies where the investigator chooses the levels of the factors and the allocation of subjects to the various factor combinations. Such two-way layouts with equal cell numbers can be obtained in any of three ways:

1. *Blocking* whereby several (but equal numbers of) observations on each treatment occur in each block.
2. *Stratifying* according to the levels of the two factors of interest and then sampling within each stratum.
3. *Forming treatment combinations* (i.e., cells) and then allocating these combinations to individuals.

The design chosen depends, of course, on the study characteristics. For example, if we want to help eliminate the effects of a confounding factor, such as ‘households’ in a study of residential radon exposure and the risk of lung cancer, we can use blocking. On the other hand, if we are interested in measuring the respiratory function of industrial workers in different

plants (factor 1) subject to different levels of exposure to some substance (factor 2), a stratified sampling procedure is appropriate. Or if we are interested in the effects of combinations of different dosages of two different drugs, we may randomly assign the different drug combinations to different subjects in a way that ensures that each drug combination is given to the same number of subjects.

Regardless of the experimental design, having more than one observation at each combination of factor levels is essential. In the respiratory function example, if only one person experiencing a certain level of exposure was examined from a given plant, we would have no direct way to determine how the responses of other persons in the same circumstances differed from the response of that individual. Similarly, for the drug example, if no more than one individual received a specific treatment combination of drugs, we could not assess the variation in response among persons receiving that same treatment combination. Thus, a major reason for having more than one observation at each combination of factor levels (i.e., in each cell) is to be able to compute a pure estimate of experimental error (σ^2).

Pure in this context means within-cell. Using a randomized-blocks design (with a single observation per cell) precludes the possibility of obtaining a within-cell estimate of σ^2 . However, if the blocking does *only* what it is assumed to do (i.e., eliminate the effects of confounding factors), we can still obtain an estimate of σ^2 (although not a pure one) from a statistic that would ordinarily measure block-treatment interaction (which is assumed not to exist).

The detection of an interaction effect between two factors, although not of interest for the randomized-blocks design (in which the blocks are not considered as the levels of an important factor or independent variable), is an important reason to have multiple observations in each cell in two-way layouts. We will elaborate on this notion of interaction later in the chapter.

■ **Example 19.1** In Table 19.1 the Cornell Medical Index (CMI) data of Table 17.3 from Daly's (1973) study have been categorized according to the levels of two factors:

Factor 1: Percentage black in surrounding neighborhoods (PSN);
Level 1 = Low ($\leq 50\%$), Level 2 = High ($> 50\%$)

Factor 2: Number of households in turnkey neighborhood (NHT);
Level 1 = Low (≤ 100), Level 2 = High (> 100)

Each of the four cells in Table 19.1 represents a combination of a level of factor 1 and a level of factor 2. The 25 observations in each cell constitute random samples of 25 women who were heads of household selected from the four turnkey neighborhoods, as defined by the stratification of the two factors NHT and PSN.

We have already seen that the F test for one-way ANOVA, which treats the four cells of Table 19.1 as the levels of a single variable "neighborhood," was nonsignificant for these data. Thus, we concluded that the mean CMI scores for the four neighborhoods, when compared simultaneously, do not significantly differ from one another.

If the one-way ANOVA F test had led to the conclusion that the four neighborhoods have significantly different mean CMI scores, the nature of these differences would have been of considerable interest. For example, do neighborhoods with a high percentage of blacks

TABLE 19.1 Cornell Medical Index scores for a sample of women from four turnkey neighborhoods*

Number of Households (NHT)	Percentage Black in Surrounding Neighborhoods (PSN)		Total
	Low ($\leq 50\%$)	High ($> 50\%$)	
Low (≤ 100)	Cherryview: 49, 12, 28, 24, 16, 28, 21, 48, 30, 18, 10, 10, 15, 7, 6, 11, 13, 17, 43, 18, 6, 10, 9, 12, 12 ($n_{11} = 25$, $\bar{Y}_{11..} = 18.92$)	Easton: 13, 10, 20, 20, 22, 14, 10, 8, 21, 35, 17, 23, 17, 23, 83, 21, 17, 41, 20, 25, 49, 41, 27, 37, 57 ($n_{12} = 25$, $\bar{Y}_{12..} = 26.84$)	$n_{1..} = 50$ $\bar{Y}_{1..} = 22.88$
High (> 100)	Northhills: 20, 31, 19, 9, 7, 16, 11, 17, 9, 14, 10, 5, 15, 19, 29, 23, 70, 25, 6, 62, 2, 14, 26, 7, 55 ($n_{21} = 25$, $\bar{Y}_{21..} = 20.84$)	Morningside: 5, 1, 44, 11, 4, 3, 14, 2, 13, 68, 34, 40, 36, 40, 22, 25, 14, 23, 26, 11, 20, 4, 16, 25, 17 ($n_{22} = 25$, $\bar{Y}_{22..} = 20.72$)	$n_{2..} = 50$ $\bar{Y}_{2..} = 20.78$
Total	$n_{1..} = 50$, $\bar{Y}_{1..} = 19.88$	$n_{2..} = 50$, $\bar{Y}_{2..} = 23.78$	$n_{..} = 100$ $\bar{Y}_{..} = 21.83$

*In this table, we use the notation $\bar{Y}_{ij..}$ to denote the sample mean for the cell (i, j) , $\bar{Y}_{i..}$ to denote the sample mean for row i , $\bar{Y}_{.j..}$ to denote the sample mean for column j , and $\bar{Y}_{..}$ to denote the overall sample mean.

© Cengage Learning

in the surrounding environs have significantly smaller mean CMI scores than those with a low percentage of blacks in surrounding environs? Do neighborhoods with a large number of households have significantly smaller mean CMI scores than neighborhoods with a small number of households? Do the amount and direction of the differences in CMI scores between neighborhoods of different size depend significantly on the racial makeup of the surrounding environs (i.e., is there an interaction effect)?

These questions can be answered by using two-way ANOVA. In fact, in spite of the nonsignificance of the one-way ANOVA F test, we will perform a two-way ANOVA to quantify the separate effects of the factors PSN and NHT and (even more importantly) to examine the possibility of an interaction between these two factors. ■

19.2 Using a Table of Cell Means

Our first step in examining a two-way layout should always be to construct a table of cell means. For our CMI data, a table of cell means is presented in Table 19.2. From this table, we can make three important observations:

1. The mean CMI score for low NHT is larger than that for high NHT:

$$\hat{\mu}_{1..} - \hat{\mu}_{2..} = \bar{Y}_{1..} - \bar{Y}_{2..} = 22.88 - 20.78 = 2.10$$

This comparison measures the *main effect* of NHT.

TABLE 19.2 Cell means for CMI data

NHT	PSN		Row Mean
	Low	High	
Low	$\hat{\mu}_{11} = \bar{Y}_{11\cdot} = 18.92$	$\hat{\mu}_{12} = \bar{Y}_{12\cdot} = 26.84$	$\hat{\mu}_{1\cdot} = \bar{Y}_{1\cdot\cdot} = 22.88$
High	$\hat{\mu}_{21} = \bar{Y}_{21\cdot} = 20.84$	$\hat{\mu}_{22} = \bar{Y}_{22\cdot} = 20.72$	$\hat{\mu}_{2\cdot} = \bar{Y}_{2\cdot\cdot} = 20.78$
Column Mean	$\hat{\mu}_{\cdot 1} = \bar{Y}_{\cdot 1\cdot} = 19.88$	$\hat{\mu}_{\cdot 2} = \bar{Y}_{\cdot 2\cdot} = 23.78$	$\hat{\mu}_{\cdot\cdot} = \bar{Y}_{\cdot\cdot\cdot} = 21.83$

© Cengage Learning

2. The mean CMI score for low PSN is smaller than that for high PSN:

$$\hat{\mu}_{\cdot 1} - \hat{\mu}_{\cdot 2} = \bar{Y}_{\cdot 1\cdot} - \bar{Y}_{\cdot 2\cdot} = 19.88 - 23.78 = -3.90$$

This comparison measures the main effect of PSN.

3. There is little difference between high PSN and low PSN when NHT is high:

$$\hat{\mu}_{22} - \hat{\mu}_{21} = \bar{Y}_{22\cdot} - \bar{Y}_{21\cdot} = 20.72 - 20.84 = -0.12$$

but there is considerable difference between high PSN and low PSN when NHT is low:

$$\hat{\mu}_{12} - \hat{\mu}_{11} = \bar{Y}_{12\cdot} - \bar{Y}_{11\cdot} = 26.84 - 18.92 = 7.92$$

These two comparisons measure the *interaction* between NHT and PSN.

Observation 1 suggests that persons from small turnkey neighborhoods might not be as healthy as persons from large turnkey neighborhoods (the lower the CMI score, the healthier), which is consistent with what Daly theorized. Observation 2 suggests that persons from turnkey neighborhoods containing a high percentage of blacks in the surrounding neighborhoods might not be as healthy as persons from turnkey neighborhoods with a low percentage of blacks in the surrounding neighborhoods. This observation runs counter to Daly's theory. Finally, observation 3 suggests that there is little difference between neighborhoods containing high and low black percentages in the surroundings when the turnkey neighborhood size is large, whereas there is considerable difference between neighborhoods with high and low black percentages in the surroundings when the size of the turnkey neighborhood is small.

Another way of describing the interaction effect pointed out in observation 3 is to say that, when PSN is low, persons from neighborhoods with low NHT seem to be healthier than persons from neighborhoods with high NHT, but that, when PSN is high, persons from neighborhoods with high NHT seem to be healthier than persons from neighborhoods with low NHT.

Clearly, the most important of these observations is observation 3, which suggests some kind of interaction between NHT and PSN; that is, any difference in the health of persons from different PSN categories seems to depend on what NHT category is being considered. Equivalently, any difference between persons from different NHT categories appears to depend on the PSN category.

Nevertheless, we must remember that the differences found in this study were obtained from a sample, so the differences could have occurred solely by chance. In other words, we must determine whether the differences found are statistically significant. This can be done by using two-way ANOVA.

19.2.1 Fixed, Random, or Mixed Model

To determine the appropriate significance tests for a two-way ANOVA, we first have to specify whether each of the two factors is fixed or random. Although such a specification in the one-way ANOVA situation alters only the statement of the null and alternative hypotheses (and not the form of the F test used), the way the factors are classified in the two-way case affects the F test as well.

In fact, three different cases must be considered, depending on the classification of the two factors:

1. The fixed-effects case, where both factors are fixed
2. The random-effects case, where both factors are random
3. The mixed-effects case, where one factor is fixed and the other is random

In our example (i.e., Table 19.1), the classification of the factors NHT and PSN depends on the researcher's point of view. The fixed-effects case would apply if the researcher were interested only in the particular turnkey neighborhoods selected for study or (in terms of the two factors NHT and PSN) did not wish to make inferences to neighborhoods of different sizes or to different black percentages for surrounding neighborhoods. The random-effects case would apply if the turnkey neighborhood sizes chosen were considered representative of a larger population of sizes of interest and if the black percentages were considered representative of a larger population of black percentages of interest. The mixed-effects case would be applicable if one of the factors was considered fixed and the other random. Of these three cases, the random-effects case probably best represents the true situation. Nevertheless, we will discuss the appropriate analysis for each case.

19.2.2 Results of Two-way ANOVA, with Interaction, for the Data of Table 19.1

The computer output that follows gives the two-way ANOVA results for the CMI data of Table 19.1. The four sources of variation shown in the output correspond to the two main effects (for NHT and PSN), the interaction effect ($NHT \times PSN$), and the error variation. Since all possible interactions among the main effects are included here, this type of model is called a *fully specified model*. Corresponding to these four sources are three null hypotheses that may be tested:

1. H_0 : No main effect of NHT
2. H_0 : No main effect of PSN
3. H_0 : No interaction effect between NHT and PSN

Edited SAS Output (PROC GLM) for the Data of Table 19.1

CLASS LEVEL INFORMATION		
Class	Levels	Values
NHT	2	1 2
PSN	2	1 2

Dependent Variable: CMI

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	894.51000	298.17000	1.18	0.3223
Error	96	24301.60000	253.14167		
Corrected Total	99	25196.11000			

R-Square	Coeff Var	Root MSE	cmi Mean
0.035502	72.88331	15.91043	21.83000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
NHT	1	110.2500000	110.2500000	0.44	0.5109
PSN	1	380.2500000	380.2500000	1.50	0.2233
NHT*PSN	1	404.0100000	404.0100000	1.60	0.2095

Source	DF	Type III SS	Mean Square	F Value	Pr > F
NHT	1	110.2500000	110.2500000	0.44	0.5109
PSN	1	380.2500000	380.2500000	1.50	0.2233
NHT*PSN	1	404.0100000	404.0100000	1.60	0.2095

For random-effects model

Fixed-effects tests

Source	Type III Expected Mean Square
NHT	Var(Error) + 25 Var(NHT*PSN) + 50 Var(NHT)
PSN	Var(Error) + 25 Var(NHT*PSN) + 50 Var(PSN)
NHT*PSN	Var(Error) + 25 Var(NHT*PSN)

Source: NHT
Error: MS(NHT*PSN)

Random- and mixed-effects tests

DF	Type III MS	Denominator DF	Denominator MS	F Value	Pr > F
1	110.25	1	404.01	0.2729	0.6935

(continued)

Source: PSN

Error: MS(NHT*PSN)

DF	Type III MS	Denominator DF	Denominator MS	F Value	Pr > F
1	380.25	1	404.01	0.9412	0.5096

Source: NHT*PSN

Error: MS(Error)

DF	Type III MS	Denominator DF	Denominator MS	F Value	Pr > F
1	404.01	96	253.14166667	1.5960	0.2095

Each of these hypotheses can be stated more precisely in terms of population cell means and/or variances, depending on whether the fixed-, random-, or mixed-effects case applies. For example, in the fixed-effects case, the null hypotheses may be given in terms of cell means (see Table 19.2) as follows:

1. $H_0: \mu_{1.} = \mu_{2.}$ (no main effect of NHT)
2. $H_0: \mu_{.1} = \mu_{.2}$ (no main effect of PSN)
3. $H_0: \mu_{22} - \mu_{21} - \mu_{12} + \mu_{11} = 0$ (no interaction effect between NHT and PSN)

More will be said in Sections 19.4 and 19.7 about testing null hypotheses in the fixed-, random-, and mixed-effects cases. We shall describe in Section 19.3 how the sum-of-squares and degrees-of-freedom terms are determined for the general two-way ANOVA case. At present, we focus entirely on the F statistics given in the output, which differ according to how the factors are classified. The two numbers in parentheses next to each F statistic below indicate the appropriate degrees of freedom to be used for that F test. None of the tests is significant, as we might have expected from the previous results for one-way ANOVA.

Tests for Fixed Effects

F tests for the fixed-effects case *always* involve dividing the mean square for the particular factor of interest by the error mean square. The degrees of freedom correspond to the particular mean squares used. Thus,

$$F(\text{NHT}) = \frac{\text{MS}(\text{NHT})}{\text{MS}(\text{Error})} = \frac{110.25}{253.14} = 0.44_{(1, 96)}$$

$$F(\text{PSN}) = \frac{\text{MS}(\text{PSN})}{\text{MS}(\text{Error})} = \frac{380.25}{253.14} = 1.50_{(1, 96)}$$

$$F(\text{NHT} \times \text{PSN}) = \frac{\text{MS}(\text{NHT} \times \text{PSN})}{\text{MS}(\text{Error})} = \frac{404.01}{253.14} = 1.60_{(1, 96)}$$

Tests for Random or Mixed Effects¹

In either the random- or the mixed-effects case, the F test for each main effect consists of dividing the mean square for the particular main effect under consideration by the interaction mean square. Again, the degrees of freedom correspond to the particular mean squares used. Thus,

$$F(NHT) = \frac{MS(NHT)}{MS(NHT \times PSN)} = \frac{110.25}{404.01} = 0.27_{(1, 1)}$$

$$F(PSN) = \frac{MS(PSN)}{MS(NHT \times PSN)} = \frac{380.25}{404.01} = 0.94_{(1, 1)}$$

The F test for interaction is the same for the random- and mixed-effects cases as for the fixed-effects case.

19.3 General Methodology

In this section, we describe the data configuration, computational formulas, and ANOVA table for the general balanced (i.e., equal cell numbers) two-way situation consisting of r levels of one factor (which we call the *row factor*), c levels of the other factor (which we call the *column factor*), and n observations in each of the rc cells.

19.3.1 Presentation of Data for Two-way ANOVA: Equal Cell Numbers

Table 19.3 offers a convenient format for presenting the data for the general two-way situation when all cells contain an equal number of observations. Table 19.4 gives the corresponding table of (sample) cell means.

Table 19.3 uses three subscripts to differentiate the individual observations. The first two subscripts index the row and column (i.e., the cell), and the third subscript denotes the observation number within that cell. For example, Y_{122} denotes the second observation in cell (1, 2), which corresponds to row 1 and column 2. In general, Y_{ijk} denotes the k th observation in the (i, j) th cell of the table. The cell total for the (i, j) th cell is denoted by T_{ij} ; the i th row total is R_i ; the j th column total is C_j ; and the grand total is G . In other words,

$$R_i = \sum_{j=1}^r \sum_{k=1}^n Y_{ijk}, \quad C_j = \sum_{i=1}^r \sum_{k=1}^n Y_{ijk}, \quad G = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^n Y_{ijk}$$

In Table 19.4, the mean of the n observations in cell (i, j) is denoted by $\bar{Y}_{ij\cdot}$; this sample mean estimates the population cell mean $\mu_{ij\cdot}$. The i th row mean is $\bar{Y}_{i\cdot\cdot\cdot}$; the j th column mean is $\bar{Y}_{\cdot j\cdot\cdot}$; and $\bar{Y}_{\dots\dots\dots}$ is the grand (overall) mean. Thus, we have

$$\bar{Y}_{ij\cdot} = \frac{1}{n} \sum_{k=1}^n Y_{ijk}, \quad \bar{Y}_{i\cdot\cdot\cdot} = \frac{R_i}{cn}, \quad \bar{Y}_{\cdot j\cdot\cdot} = \frac{C_j}{rn}, \quad \bar{Y}_{\dots\dots\dots} = \frac{G}{rcn}$$

¹ See Section 19.7 (and footnote 8 on page 607) for a discussion of the rationale for these F tests in terms of "expected mean squares."

TABLE 19.3 General layout of data for two-way ANOVA with equal numbers of observations per cell

Row Factor	Column Factor			Row Total	
	1	2	...		
1	$(Y_{111}, Y_{112}, \dots, Y_{11n})$ T_{11}	$(Y_{121}, Y_{122}, \dots, Y_{12n})$ T_{12}	...	$(Y_{1c1}, Y_{1c2}, \dots, Y_{1cn})$ T_{1c}	R_1
2	$(Y_{211}, Y_{212}, \dots, Y_{21n})$ T_{21}	$(Y_{221}, Y_{222}, \dots, Y_{22n})$ T_{22}	...	$(Y_{2c1}, Y_{2c2}, \dots, Y_{2cn})$ T_{2c}	R_2
⋮	⋮	⋮	⋮	⋮	⋮
r	$(Y_{r11}, Y_{r12}, \dots, Y_{r1n})$ T_{r1}	$(Y_{r21}, Y_{r22}, \dots, Y_{r2n})$ T_{r2}	...	$(Y_{rc1}, Y_{rc2}, \dots, Y_{rcn})$ T_{rc}	R_r
Column Total	C_1	C_2	...	C_c	G

© Cengage Learning

TABLE 19.4 Sample cell means for two-way ANOVA

Row Factor	Column Factor			Row Mean	
	1	2	...		
1	$\bar{Y}_{11\cdot}$	$\bar{Y}_{12\cdot}$...	$\bar{Y}_{1c\cdot}$	$\bar{Y}_{1\cdot\cdot}$
2	$\bar{Y}_{21\cdot}$	$\bar{Y}_{22\cdot}$...	$\bar{Y}_{2c\cdot}$	$\bar{Y}_{2\cdot\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮
r	$\bar{Y}_{r1\cdot}$	$\bar{Y}_{r2\cdot}$...	$\bar{Y}_{rc\cdot}$	$\bar{Y}_{r\cdot\cdot}$
Column Mean	$\bar{Y}_{\cdot 1\cdot}$	$\bar{Y}_{\cdot 2\cdot}$...	$\bar{Y}_{\cdot c\cdot}$	$\bar{Y}_{\cdot\cdot\cdot}$

© Cengage Learning

■ **Example 19.2** In our earlier example (Table 19.1), $r = c = 2$ and $n = 25$. Tables 19.5 and 19.6 give the data layout and the table of cell means for an example in which $r = 3$, $c = 3$, and $n = 12$. This example deals with one kind of data set used in occupational health studies for evaluating the health status of industrial workers. The dependent variable here is forced expiratory volume (FEV), which is a measure of respiratory function. Very low FEV indicates possible respiratory dysfunction, whereas high FEV indicates good respiratory function. In this example, observations are taken on $n = 12$ individuals in each of three plants in a given industry where workers are exposed to one of three toxic substances. Thus, we have two factors, each with three levels. The categories of the row factor (Plant) are labeled 1, 2, and 3 in Table 19.5. The categories

TABLE 19.5 Forced expiratory volume classified by plant and toxic exposure

Plant	Toxic Substance			Row Total
	A	B	C	
1	4.64, 5.92, 5.25, 6.17, 4.20, 5.90, 5.07, 4.13, 4.07, 5.30, 4.37, 3.76 ($T_{11} = 58.78$)	3.21, 3.17, 3.88, 3.50, 2.47, 4.12, 3.51, 3.85, 4.22, 3.07, 3.62, 2.95 ($T_{12} = 41.57$)	3.75, 2.50, 2.65, 2.84, 3.09, 2.90, 2.62, 2.75, 3.10, 1.99, 2.42, 2.37 ($T_{13} = 32.98$)	$R_1 = 133.33$
2	5.12, 6.10, 4.85, 4.72, 5.36, 5.41, 5.31, 4.78, 5.08, 4.97, 5.85, 5.26 ($T_{21} = 62.81$)	3.92, 3.75, 4.01, 4.64, 3.63, 3.46, 4.01, 3.39, 3.78, 3.51, 3.19, 4.04 ($T_{22} = 45.33$)	2.95, 3.21, 3.15, 3.25, 2.30, 2.76, 3.01, 2.31, 2.50, 2.02, 2.64, 2.27 ($T_{23} = 32.37$)	$R_2 = 140.51$
3	4.64, 4.32, 4.13, 5.17, 3.77, 3.85, 4.12, 5.07, 3.25, 3.49, 3.65, 4.10 ($T_{31} = 49.56$)	4.95, 5.22, 5.16, 5.35, 4.35, 4.89, 5.61, 4.98, 5.77, 5.23, 4.86, 5.15 ($T_{32} = 61.52$)	2.95, 2.80, 3.63, 3.85, 2.19, 3.32, 2.68, 3.35, 3.12, 4.11, 2.90, 2.75 ($T_{33} = 37.65$)	$R_3 = 148.73$
Column Total	$C_1 = 171.15$	$C_2 = 148.42$	$C_3 = 103.00$	$G = 422.57$

© Cengage Learning

of the column factor (Toxsub) are labeled A, B, and C. Three questions are of particular interest here:

1. Does the mean FEV level differ among plants (i.e., is there a main effect due to Plant)?
2. Does the mean FEV level differ among exposure categories (i.e., is there a main effect due to Toxsub)?
3. Do the differences in mean FEV levels among plants depend on the exposure category, and vice versa (i.e., is there an interaction effect between Plant and Toxsub)?

We can evaluate these questions initially by examining the cell means in Table 19.6. Of the three plants, plant 1 has the lowest mean FEV (3.70), followed by plant 2 (3.90), and then plant 3 (4.13). This suggests that the workers in plant 1 have poorer respiratory health than those in plant 2, and so on. Nevertheless, if the 3.70 value for plant 1 is considered clinically normal, then—despite the differences observed—all plants will be given a “clean bill of health.” Furthermore, these differences among plants might have occurred solely by chance (i.e., might not be statistically significant).

In addition, it can be seen from Table 19.6 that exposure to substance C is associated with the poorest respiratory health (2.86), whereas exposure to substance A (4.75) and exposure to substance B (4.12) are associated with considerably better respiratory health. Again, we must decide whether to view the 2.86 value as meaningfully low and determine whether the differences among substances A, B, and C are statistically significant.

TABLE 19.6 Cell means for data of Table 19.5

Plant	Toxic Substance			Row Mean
	A	B	C	
1	4.90	3.46	2.75	3.70
2	5.23	3.78	2.70	3.90
3	4.13	5.13	3.14	4.13
Column Total	4.75	4.12	2.86	3.91

© Cengage Learning

Finally, Table 19.6 indicates that the differences among plants depend somewhat on the toxic exposure being considered. For example, with respect to toxic substance A, plant 3 has the lowest mean FEV (4.13). With respect to toxic substance B, however, plant 1 has the lowest mean (3.46); and with respect to toxic substance C, plant 2 has the lowest mean (2.70). Furthermore, the magnitude of the differences among plants varies with the toxic substance. For toxic substance B, the difference between the highest and lowest plant means is $5.13 - 3.46 = 1.67$, whereas for toxic substances A and C the maximum differences are smaller ($5.23 - 4.13 = 1.10$ and $3.14 - 2.70 = 0.44$, respectively). Such fluctuations suggest that a significant interaction effect may be present, although this must be verified statistically. ■

19.3.2 ANOVA Table for Two-way ANOVA

Table 19.7 gives the general form of the two-way ANOVA table with r levels of the row factor, c levels of the column factor, and n observations per cell. The computer output, given right after Table 19.7, shows the corresponding ANOVA results associated with the FEV data of Table 19.5. From Table 19.7, we see that the total (corrected) sum of squares (SSY) has been divided into the four components—SSR (rows), SSC (columns), SSRC (row \times column interaction), and SSE (error)—based on the following fundamental equation:

$$\text{SSY} = \text{SSR} + \text{SSC} + \text{SSRC} + \text{SSE}, \quad (19.1)$$

or, equivalently,

$$\begin{aligned} \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{...})^2 &= cn \sum_{i=1}^r (\bar{Y}_{i..} - \bar{Y}_{...})^2 + rn \sum_{j=1}^c (\bar{Y}_{.j} - \bar{Y}_{...})^2 \\ &+ n \sum_{i=1}^r \sum_{j=1}^c (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j} + \bar{Y}_{...})^2 + \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{ij.})^2 \end{aligned}$$

TABLE 19.7 General (balanced) two-way ANOVA table²

Source	d.f.	SS	MS	F	
				Fixed	Mixed or Random
Row (main effect)	$r - 1$	SSR	$MSR = \frac{SSR}{r - 1}$	$\frac{MSR}{MSE}$	$\frac{MSR}{MSRC}$
Column (main effect)	$c - 1$	SSC	$MSC = \frac{SSC}{c - 1}$	$\frac{MSC}{MSE}$	$\frac{MSC}{MSRC}$
Row \times column (interaction)	$(r - 1)(c - 1)$	SSRC	$MSRC = \frac{SSRC}{(r - 1)(c - 1)}$	$\frac{MSRC}{MSE}$	$\frac{MSRC}{MSE}$
Error	$rc(n - 1)$	SSE	$MSE = \frac{SSE}{rc(n - 1)}$		
Total	$rcn - 1$	SSY			

© Cengage Learning

Edited SAS Output (PROC GLM) for the Data of Table 19.5

CLASS LEVEL INFORMATION		
Class	Levels	Values
PLANT	3	1 2 3
SUBSTANCE	3	A B C

Dependent Variable: FEV

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	94.6984630	11.8373079	44.10	<.0001
Error	99	26.5758583	0.2684430		
Corrected Total	107	121.2743213			

R-Square	Coeff Var	Root MSE	FEV Mean
0.780862	13.24193	0.518115	3.912685

Source	DF	Type I SS	Mean Square	F Value	Pr > F
PLANT	2	3.29889630	1.64944815	6.14	0.0031
SUBSTANCE	2	66.88936852	33.44468426	124.59	<.0001
PLANT*SUBSTANCE	4	24.51019815	6.12754954	22.83	<.0001

(continued)

² See Section 19.7 for a discussion of the rationale for these *F* tests in terms of "expected mean squares."

The flowchart illustrates the derivation of Type III SS and Mean Square terms from SAS output. It branches into random-effects and fixed-effects tests, then into random- and mixed-effects tests.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PLANT	2	3.29889630	1.64944815	6.14	0.0031
SUBSTANCE	2	66.88936852	33.44468426	124.59	<.0001
PLANT*SUBSTANCE	4	24.51019815	6.12754954	22.83	<.0001

Source	Type III Expected Mean Square				
PLANT	Var(Error) + 12 Var(PLANT*SUBSTANCE) + 36 Var(PLANT)				
SUBSTANCE	Var(Error) + 12 Var(PLANT*SUBSTANCE) + 36 Var(SUBSTANCE)				
PLANT*SUBSTANCE	Var(Error) + 12 Var(PLANT*SUBSTANCE)				

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PLANT	2	3.298896	1.649448	0.27	0.7768
SUBSTANCE	2	66.889369	33.444684	5.46	0.0719
Error	4	24.510198	6.127550		
Error: MS(PLANT*SUBSTANCE)					

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PLANT*SUBSTANCE	4	24.510198	6.127550	22.83	<.0001
Error: MS(Error)	99	26.575858	0.268443		

For the FEV data of Table 19.5, the sums of squares obtained (see the above SAS output) are

$$\text{SSR}(\text{Plant}) = 3.299$$

$$\text{SSC}(\text{Toxsub}) = 66.889$$

$$\text{SSRC}(\text{Plant} \times \text{Toxsub}) = 24.510$$

$$\text{SSE} = 26.576$$

$$\text{SSY} = 121.274$$

The degrees of freedom associated with these sums of squares are as follows:

SSR has $(r - 1)$ d.f.

SSC has $(c - 1)$ d.f.

SSRC has $(r - 1)(c - 1)$ d.f.

(19.2)

SSE has $rc(n - 1)$ d.f.

SSY has $(rcn - 1)$ d.f.

Each mean-square term is obtained (as usual) by dividing the corresponding sum of squares by its associated degrees of freedom. Again, the appropriate F statistics to use depend on whether the row and column factors are classified as fixed or as random. These F tests are described in Section 19.4.

19.4 F Tests for Two-way ANOVA

The null hypotheses of interest for two-way ANOVA—as well as the basic statistical assumptions required for validly testing them—can be stated quite generally to encompass the three possible schemes of factor classification that depend on whether or not the row factor (R) and column factor (C) are classified as both fixed, both random, or mixed (i.e., one fixed and the other random).³

Null Hypotheses

1. $H_0(R)$: *There is no row factor (main) effect* (i.e., there are no differences among the effects of the levels of the row factor).
2. $H_0(C)$: *There is no column factor (main) effect* (i.e., there are no differences among the effects of the levels of the column factor).
3. $H_0(RC)$: *There is no interaction effect between rows and columns* (i.e., the row-level effects within any one column are the same as those within any other column; equivalently, the column-level effects within any one row are the same as those within any other row).

Assumptions

1. For fixed-effects models, all observations are *statistically independent* of one another.⁴
2. Each observation comes from a *normally distributed population*.
3. Each observation has the same population variance (i.e., the usual assumption of *variance homogeneity* applies).

As previously stated, the appropriate F statistics depend on whether the row and column factors are classified as fixed or random. This is because the mean-square term associated with a given source of variation will estimate different quantities, depending on whether the row and column factors are fixed or random.

Regardless of how the factors are classified, however, the F statistic used to test $H_0(RC)$ of no row–column interaction always has the form

$$F(RC) = \frac{MS_{RC}}{MSE}$$

with $(r - 1)(c - 1)$ and $rc(n - 1)$ degrees of freedom.⁵

³ However, the null hypotheses are quite different when stated more precisely in terms of population cell means and/or variances.

⁴ The responses are *not* mutually independent when random effects are included in the regression model for two-way ANOVA with equal cell numbers.

⁵ $F(RC)$ denotes the F test of $H_0(RC)$. Similarly, $F(R)$ and $F(C)$ denote the F tests of $H_0(R)$ and $H_0(C)$, respectively.

The F statistics for evaluating main effects differ as follows with respect to the factor classification scheme:

1. *Rows and columns fixed.* Divide the mean squares for rows and for columns by the mean square for error:

$$F(R) = \frac{MSR}{MSE}$$

with $(r - 1)$ and $rc(n - 1)$ degrees of freedom or

$$F(C) = \frac{MSC}{MSE}$$

with $(c - 1)$ and $rc(n - 1)$ degrees of freedom.

2. *Rows and columns random, or one factor fixed and the other factor random.* Divide the mean squares for rows and columns by the mean square for interaction:

$$F(R) = \frac{MSR}{MSRC}$$

with $(r - 1)$ and $(r - 1)(c - 1)$ degrees of freedom or

$$F(C) = \frac{MSC}{MSRC}$$

with $(c - 1)$ and $(r - 1)(c - 1)$ degrees of freedom.

For the FEV data, the classification of the factors depends on the point of view of the researcher. If, for example, the plants and toxic substances were the only ones of interest, both factors would be considered fixed. However, if the plants were considered to represent a sample from a large population of plants of interest and if the toxic substances likewise were viewed as representing a population of toxic agents of interest, both factors would be treated as random. Of course, the classification would be mixed if one of these factors were considered fixed and the other random.

The decisions regarding certain null hypotheses differ, however, depending on how the factors are classified. This can be seen from the earlier SAS output, as follows:

1. *Both factors fixed.* Both main effects are significant, since $F(\text{Plant}) = 6.14^{**}$ (with 2 and 99 degrees of freedom) and $F(\text{Toxsub}) = 124.59^{**}$ (with 2 and 99 degrees of freedom).
2. *Both factors random, or one factor fixed and the other factor random.* Neither main effect is significant, since $F(\text{Plant}) = 0.27$ (with 2 and 4 degrees of freedom) and $F(\text{Toxsub}) = 5.46$ (with 2 and 4 degrees of freedom).

Despite these differences among the main-effect test results, the most important finding from this analysis is that the interaction effect is significant:

$$F(\text{Plant} \times \text{Toxsub}) = 22.83^{**} \text{ (with 4 and 99 d.f.)}$$

In Section 19.6, we discuss the interpretation of such interaction effects. Certainly, it does not make much sense to talk about the separate or independent effects (i.e., main effects) of Plant and Toxsub on FEV, since there is strong evidence that these factors do not affect FEV independently of one another.

Regardless of whether the factors are fixed or random, eight distinct patterns of significance may result, depending on the significance (or nonsignificance) of the three tests involved (two main-effect tests and the interaction test). Suppose that one factor is labeled A and the second factor B. Then Table 19.8 summarizes the possible outcomes (with significant results indicated by asterisks). In pattern I, the F tests for the A effect, the B effect, and the $A \times B$ interaction are all nonsignificant. In pattern VIII, all three are significant.

TABLE 19.8 Patterns of significance possible in two-way ANOVA with equal cell numbers

Source	Pattern							
	I	II	III	IV	V	VI	VII	VIII
A	*			*		*		*
B		*		*			*	*
$A \times B$				*	*	*	*	*

Asterisk denotes significance for the effect in that row.

© Cengage Learning

Each pattern deserves some comment. Pattern I leads us to conclude that there is no evidence of any relationship between the response variable and the predictors (considered either separately or together). Pattern II implies that only factor A is related to the response variable. Similarly, Pattern III implies that only factor B is related to the response variable. In Pattern IV, we conclude that both A and B affect the response. Furthermore, the nature of the relationship between factor A and the response remains the same at all levels of factor B studied, and, conversely. Patterns V, VI, VII, and VIII all involve significant interaction. Most statisticians recommend focusing only on this interaction in such cases. A significant interaction implies that both factors are important and that the level of one of the factors must be known to characterize the effect of the other factor on the response. Section 19.6 discusses interpreting interactions.

19.5 Regression Model for Fixed-effects Two-way ANOVA

In this section, we describe a particular regression model⁶ and a related classical fixed-effects ANOVA model for two-way ANOVA when both factors are considered fixed.

⁶ Several other regression models can be defined, of course, depending on the coding scheme for the dummy variables. The regression model given here is the one most commonly used, due to its natural connection with the classical fixed-effects two-way ANOVA model.

TABLE 19.9 Table of population cell means for two-way layout

Row	Column					
	1	2	c			
1	μ_{11}	μ_{12}	...	μ_{1c}	$\mu_{1\cdot}$	
2	μ_{21}	μ_{22}	...	μ_{2c}	$\mu_{2\cdot}$	
:	:	:		:	:	
r	μ_{r1}	μ_{r2}	...	μ_{rc}	$\mu_{r\cdot}$	
Column Mean	$\mu_{\cdot 1}$	$\mu_{\cdot 2}$...	$\mu_{\cdot c}$	$\mu_{\cdot \cdot}$	

© Cengage Learning

Random-effects and mixed-effects models are discussed in detail in Section 19.7. As in the cases of one-way ANOVA and randomized-blocks ANOVA, a regression model for two-way ANOVA can be interpreted in terms of the cell, marginal, and overall means associated with the two-way layout (see Table 19.9). In the table,

$$\mu_{i\cdot} = \frac{1}{c} \sum_{j=1}^c \mu_{ij} \quad i = 1, 2, \dots, r$$

$$\mu_{\cdot j} = \frac{1}{r} \sum_{i=1}^r \mu_{ij} \quad j = 1, 2, \dots, c$$

$$\mu_{\cdot \cdot} = \frac{1}{rc} \sum_{i=1}^r \sum_{j=1}^c \mu_{ij}$$

19.5.1 Regression Model

When there are r rows and c columns, a regression model can be formulated involving $(r - 1)$ dummy variables for the row factor, $(c - 1)$ dummy variables for the column factor, and $(r - 1)(c - 1)$ interaction dummy variables, which are constructed by forming products of each of the row dummy variables with each of the column dummy variables. Such a model can be expressed as

$$Y = \mu + \sum_{i=1}^{r-1} \alpha_i X_i + \sum_{j=1}^{c-1} \beta_j Z_j + \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} \gamma_{ij} X_i Z_j + E \quad (19.3)$$

in which

$$X_i = \begin{cases} -1 & \text{for level } r \text{ of the row factor} \\ 1 & \text{for level } i \text{ of the row factor} \\ 0 & \text{otherwise} \end{cases}$$

and

$$Z_j = \begin{cases} -1 & \text{for level } c \text{ of the column factor} \\ 1 & \text{for level } j \text{ of the column factor} \\ 0 & \text{otherwise} \end{cases}$$

(for $i = 1, 2, \dots, r - 1$; and $j = 1, 2, \dots, c - 1$).

The formulas relating the coefficients α_i , β_j , and γ_{ij} to the various means of Table 19.9 are

$$\begin{aligned} \mu &= \mu.. \\ \alpha_i &= \mu_{i..} - \mu.. & i = 1, 2, \dots, r - 1 \\ \beta_j &= \mu_{.j} - \mu.. & j = 1, 2, \dots, c - 1 \\ \gamma_{ij} &= \mu_{ij} - \mu_{i..} - \mu_{.j} + \mu.. & i = 1, 2, \dots, r - 1; j = 1, 2, \dots, c - 1 \\ -\sum_{i=1}^{r-1} \alpha_i &= \mu_{r..} - \mu.. \\ -\sum_{j=1}^{c-1} \beta_j &= \mu_{.c} - \mu.. \\ -\sum_{i=1}^{r-1} \gamma_{ij} &= \mu_{rj} - \mu_{r..} - \mu_{.j} + \mu.. & j = 1, 2, \dots, c - 1 \\ -\sum_{j=1}^{c-1} \gamma_{ij} &= \mu_{ic} - \mu_{i..} - \mu_{.c} + \mu.. & i = 1, 2, \dots, r - 1 \end{aligned} \tag{19.4}$$

As with the ANOVA regression analogies made in earlier chapters, the same F tests given in Table 19.7 (for the case where both factors are fixed) can be obtained by using the appropriate multiple partial F tests concerning subsets of the coefficients in the regression model (19.3). Specifically, the multiple partial F test of $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_{r-1} = 0$ for model (19.3) yields exactly the same F statistic as that used in standard (balanced) two-way fixed-effects ANOVA for testing the significance of the main effect of the row factor (i.e., $F = \text{MSR}/\text{MSE}$). Similarly, the multiple partial F test of $H_0: \beta_1 = \beta_2 = \dots = \beta_{c-1} = 0$ in model (19.3) yields exactly the same F statistic as is used in standard (balanced) two-way fixed-effects ANOVA to test the significance of the main effect of the column factor (i.e., $F = \text{MSC}/\text{MSE}$). Finally, the multiple partial F test of $H_0: \gamma_{ij} = 0$ (for $i = 1, 2, \dots, r - 1$; and $j = 1, 2, \dots, c - 1$) is identical to the (balanced) two-way fixed-effects ANOVA F test for interaction (i.e., $F = \text{MSRC}/\text{MSE}$).

The preceding formulas may also be used to express the cell, row marginal, and column marginal means as functions of the regression coefficients, as follows:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \quad i = 1, 2, \dots, r - 1; j = 1, 2, \dots, c - 1$$

$$\mu_{rj} = \mu - \sum_{i=1}^{r-1} \gamma_{ij} - \sum_{i=1}^{r-1} \alpha_i + \beta_j \quad j = 1, 2, \dots, c - 1$$

$$\mu_{ic} = \mu - \sum_{j=1}^{c-1} \gamma_{ij} - \sum_{j=1}^{c-1} \beta_j + \alpha_i \quad i = 1, 2, \dots, r-1$$

$$\mu_{rc} = \mu - \sum_{i=1}^{r-1} \alpha_i - \sum_{j=1}^{c-1} \beta_j + \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} \gamma_{ij}$$

The row marginal means are

$$\mu_{i\cdot} = \mu + \alpha_i \quad i = 1, 2, \dots, r-1$$

$$\mu_{r\cdot} = \mu - \sum_{i=1}^{r-1} \alpha_i$$

and the column marginal means are

$$\mu_{\cdot j} = \mu + \beta_j \quad j = 1, 2, \dots, c-1$$

$$\mu_{\cdot c} = \mu - \sum_{j=1}^{c-1} \beta_j$$

If reference cell coding is used, the general form of model (19.3) stays the same, but the definitions of X_i and Z_j change. For example, we can define

$$X_i = \begin{cases} 1 & \text{for level } i \text{ of the row factor} \\ 0 & \text{otherwise} \end{cases} \quad i = 1, 2, \dots, r-1$$

and

$$Z_j = \begin{cases} 1 & \text{for level } j \text{ of the column factor} \\ 0 & \text{otherwise} \end{cases} \quad j = 1, 2, \dots, c-1$$

With these definitions in place, the parameters in model (19.3) have different interpretations from those given earlier. In particular, the cell means (for the above reference cell coding) can be expressed as the following functions of the parameters in model (19.3):

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \quad i = 1, 2, \dots, r-1; j = 1, 2, \dots, c-1$$

$$\mu_{rj} = \mu + \beta_j \quad j = 1, 2, \dots, c-1$$

$$\mu_{ic} = \mu + \alpha_i \quad i = 1, 2, \dots, r-1$$

$$\mu_{rc} = \mu$$

The row marginal means are

$$\mu_{i\cdot} = \mu + \alpha_i + \sum_{j=1}^{c-1} \frac{(\beta_j + \gamma_{ij})}{c} \quad i = 1, 2, \dots, r-1$$

$$\mu_{r\cdot} = \mu + \sum_{j=1}^{c-1} \frac{\beta_j}{c}$$

and the column marginal means are

$$\begin{aligned}\mu_{\cdot j} &= \mu + \beta_j + \sum_{i=1}^{r-1} \frac{(\alpha_i + \gamma_{ij})}{r} \quad j = 1, 2, \dots, c-1 \\ \mu_{\cdot c} &= \mu + \sum_{i=1}^{r-1} \frac{\alpha_i}{r}\end{aligned}$$

These expressions must be modified when cell-specific sample sizes are not all equal. Chapter 20 addresses two-way ANOVA with unequal cell-specific sample sizes.

19.5.2 Classical Two-way Fixed-effects ANOVA Model

When both factors are considered fixed, there are three types of effects to consider:

1. *Main effects of the row factor:* The differences between the various row means and the overall mean (i.e., $\mu_{\cdot i} - \mu_{\cdot \cdot}$, $i = 1, 2, \dots, r$).
2. *Main effects of the column factor:* The differences between the various column means and the overall mean (i.e., $\mu_{\cdot j} - \mu_{\cdot \cdot}$, $j = 1, 2, \dots, c$).
3. *Interaction effects:* The differences between differences of the form $(\mu_{ij} - \mu_{\cdot i}) - (\mu_{\cdot j} - \mu_{\cdot \cdot})$ or $(\mu_{ij} - \mu_{\cdot j}) - (\mu_{\cdot i} - \mu_{\cdot \cdot})$, for $i = 1, 2, \dots, r$ and $j = 1, 2, \dots, c$.

The classical two-way fixed-effects ANOVA model involving such effects is of the following form:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk} \quad (19.5)$$

where

$\mu = \mu_{\cdot \cdot}$ = Overall mean

$\alpha_i = \mu_{\cdot i} - \mu_{\cdot \cdot}$ = Effect of row i

$\beta_j = \mu_{\cdot j} - \mu_{\cdot \cdot}$ = Effect of column j

$\gamma_{ij} = \mu_{ij} - \mu_{\cdot i} - \mu_{\cdot j} + \mu_{\cdot \cdot}$ = Interaction effect associated with cell (i, j)

$E_{ijk} = Y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij}$ = Error (or residual) associated with the k th observation in cell (i, j)

(for $i = 1, 2, \dots, r$; $j = 1, 2, \dots, c$; and $k = 1, 2, \dots, n$).

The following relationships are clearly satisfied by the effects in the preceding model:

$$\sum_{i=1}^r \alpha_i = 0, \quad \sum_{j=1}^c \beta_j = 0, \quad \sum_{i=1}^r \gamma_{ij} = 0, \quad \sum_{j=1}^c \gamma_{ij} = 0 \quad (19.6)$$

A comparison of the regression coefficients in model (19.3) with the ANOVA effects in model (19.5) makes clear that the models are completely equivalent.

Finally, each of the effects in model (19.5) can be simply estimated by using sample means, as follows:

$$\begin{aligned}\hat{\mu} &= \bar{Y}... \\ \hat{\alpha}_i &= \bar{Y}_{i..} - \bar{Y}... & i = 1, 2, \dots, r \\ \hat{\beta}_j &= \bar{Y}_{.j} - \bar{Y}... & j = 1, 2, \dots, c \\ \hat{\gamma}_{ij} &= \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j} + \bar{Y}... & i = 1, 2, \dots, r; j = 1, 2, \dots, c\end{aligned}$$

19.6 Interactions in Two-way ANOVA

In this section, we investigate several ways to evaluate interaction in two-way ANOVA. For convenience, we focus on the fixed-effects case. Nevertheless, even though the parameters involved and the test statistics used for making inferences in the fixed-effects case differ from those used in the random- and mixed-effects cases, the interpretations of interactions are generally the same.

19.6.1 Concept of Interaction

An interaction exists between two factors if the relationship among the effects associated with the levels of one factor differs according to the levels of the second factor. In other words, an interaction represents an effect due to the joint influence of two factors, over and above the effects of each factor considered separately.

More specifically, if we consider the two-way table of cell means and the various ways of writing the statistical model in the fixed-effects case, there are three equivalent ways of describing or representing an interaction in statistical terms.

Method 1: Interaction as a Difference in Differences of Means In two-way ANOVA, an interaction exists between the row and column factors if any of the following equivalent statements are true:

- For some pair of columns, the difference between the means in these columns for a given row is not equal to the difference between the corresponding means for some other row. For example, for rows 1 and 2 and columns 1 and 2, $\mu_{11} - \mu_{12} \neq \mu_{21} - \mu_{22}$.
- For some pair of rows, the difference between the means in these rows for a given column is not equal to the difference between the corresponding means for some other column. For example, for rows 1 and 2 and columns 1 and 2, $\mu_{11} - \mu_{21} \neq \mu_{12} - \mu_{22}$.
- For some cell in the table, the difference between that cell's mean and its associated marginal row mean is not equal to the difference between its associated marginal column mean and the overall mean. For example, for the (i, j)th cell, $\mu_{ij} - \mu_{i..} \neq \mu_{.j} - \mu_{...}$ or $\mu_{ij} - \mu_{i..} - \mu_{.j} + \mu_{...} \neq 0$.
- For some cell in the table, the difference between that cell's mean and its associated marginal column mean is not equal to the difference between its associated

marginal row mean and the overall mean. For example, $\mu_{ij} - \mu_{\cdot j} \neq \mu_{i\cdot} - \mu_{\cdot\cdot}$ or $\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot} \neq 0$.

Thus, when there is no interaction, the relationship among the column effects (β_j 's) is the same, regardless of the row being considered, and vice versa. And, since, from statements 3 and 4,

$$\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot} = 0$$

when there is no interaction, it follows that

$$\text{MSRC} = \frac{n}{(r-1)(c-1)} \sum_{i=1}^r \sum_{j=1}^c (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2$$

(which estimates

$$\frac{n}{(r-1)(c-1)} \sum_{i=1}^r \sum_{j=1}^c (\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot})^2$$

for the population) is small when there is no interaction and large when there is interaction.

Method 2: Interaction as an Effect in the Fixed-effects Model An interaction exists if the appropriate fixed-effects ANOVA model has the form

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk}$$

where $\gamma_{ij} \neq 0$ for at least one (i, j) pair.

Methods 1 and 2 are completely equivalent. For example, if there is no interaction, then $\mu_{ij} = \mu + \alpha_i + \beta_j$, so $\mu_{1j} - \mu_{2j} = (\mu + \alpha_1 + \beta_j) - (\mu + \alpha_2 + \beta_j) = \alpha_1 - \alpha_2$, which is independent of j . In general, $\mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = \dots = \mu_{1c} - \mu_{2c} = \alpha_1 - \alpha_2$.

Method 3: Interaction as a Term in a Regression Model An interaction exists if the appropriate regression model (using dummy variables) contains a term that involves the product (or in general, any function) of variables from different factors—for example, if the appropriate model is of the form

$$Y = \mu + \sum_{i=1}^{r-1} \alpha_i X_i + \sum_{j=1}^{c-1} \beta_j Z_j + \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} \gamma_{ij} X_i Z_j + E$$

where at least one of the γ_{ij} is not 0.

When $r = c = 2$, the model simplifies to

$$Y = \mu + \alpha_1 X_1 + \beta_1 Z_1 + \gamma_{11} X_1 Z_1 + E$$

where

$$X_1 = \begin{cases} -1 & \text{if level 2 of the row factor} \\ 1 & \text{if level 1 of the row factor} \end{cases}$$

$$Z_1 = \begin{cases} -1 & \text{if level 2 of the column factor} \\ 1 & \text{if level 1 of the column factor} \end{cases}$$

Then, $\mu = \mu_{..}$, $\alpha_1 = \mu_{1.} - \mu_{..}$, $\beta_1 = \mu_{.1} - \mu_{..}$, and $\gamma_{11} = \mu_{11} - \mu_{1.} - \mu_{.1} + \mu_{..}$.

Method 3 is equivalent to the other two methods, provided that both independent variables (i.e., factors) are considered nominal and so are represented by dummy variables. If, however, both independent variables are continuous, a regression model with any product term will exhibit a somewhat different interaction effect, not necessarily characterized by a nonzero difference of mean differences.

19.6.2 Some Hypothetical Examples

We now consider some hypothetical two-way tables of population cell means that illustrate different patterns of interaction. These tables pertain to the example in Section 19.2, for which the factors are NHT and PSN and the dependent variable is CMI score. Subsequently, we will examine the table of sample cell means actually obtained (Table 19.2), keeping in mind that the statistical test for interaction may render as statistically unimportant any tentative trends suggested by the sample means. We will also examine the example of Section 19.3 in this light.

Row and Column Main Effects but No Interaction Effect

Table 19.10 presents three alternative layouts, each representing the general situation involving *both* a row main effect and a column main effect but no interaction effect. Each of these tables gives *population* (and not sample) mean values, so there is no sampling variation to consider.

The main effects are reflected in the differences between marginal row means and between marginal column means in each table. The lack of an interaction effect can be established by comparing the differences among the cell means, as discussed in Section 19.6.1. From Table 19.10(a), for example, we have $\mu_{11} - \mu_{21} = \mu_{21} - \mu_{22}$, since $26 - 23 = 3 = 20 - 17$. For this same table, $\mu_{11} - \mu_{1.} - \mu_{.1} + \mu_{..} = 26 - 24.5 - 23 + 21.5 = 0$, and similar terms associated with the other three cells in the table are also 0. Furthermore, for this table, the model (19.5) can be shown to have the following specific structure:

$$\mu_{ij} = 21.5 + \alpha_i + \beta_j$$

where

$$\alpha_i = \begin{cases} 3 & \text{if } i = 1 \\ -3 & \text{if } i = 2 \end{cases} \quad \text{and} \quad \beta_j = \begin{cases} 1.5 & \text{if } j = 1 \\ -1.5 & \text{if } j = 2 \end{cases}$$

TABLE 19.10 Alternative layouts for main effects but no interaction

(a)		PSN			(b)		PSN			(c)		PSN		
NHT		Low	High	Row Mean	NHT		Low	High	Row Mean	NHT		Low	High	Row Mean
Low		26	23	24.5	Low		18	26	22	Low		18	26	22
High		20	17	18.5	High		20	28	24	High		16	24	20
Column Mean		23	20	21.5	Column Mean		19	27	23	Column Mean		17	25	21

This model does not involve any γ_{ij} term (i.e., there is no interaction term in the model). Thus, we have

$$\mu_{11} = 21.5 + 3 + 1.5 = 26$$

$$\mu_{12} = 21.5 + 3 - 1.5 = 23$$

$$\mu_{21} = 21.5 - 3 + 1.5 = 20$$

$$\mu_{22} = 21.5 - 3 - 1.5 = 17$$

The layouts for Tables 19.10(b) and (c) also represent no-interaction models; they have the particular forms, respectively, of

$$\mu_{ij} = 23 + \alpha_i + \beta_j$$

where

$$\alpha_i = \begin{cases} -1 & \text{if } i = 1 \\ 1 & \text{if } i = 2 \end{cases} \quad \text{and} \quad \beta_j = \begin{cases} -4 & \text{if } j = 1 \\ 4 & \text{if } j = 2 \end{cases}$$

and

$$\mu_{ij} = 21 + \alpha_i + \beta_j$$

where

$$\alpha_i = \begin{cases} 1 & \text{if } i = 1 \\ -1 & \text{if } i = 2 \end{cases} \quad \text{and} \quad \beta_j = \begin{cases} -4 & \text{if } j = 1 \\ 4 & \text{if } j = 2 \end{cases}$$

Exactly One Main Effect and No Interaction Effect

This situation is depicted in Table 19.11. Table 19.11(a) contains a main effect due to PSN but no main effect due to NHT. Table 19.11(b) contains a main effect due to NHT but no main effect due to PSN. There is no PSN \times NHT interaction.

Same-direction Interaction

Three examples of same-direction interaction are given in Table 19.12. In Table 19.12(a), we see that $\mu_{11} - \mu_{12} = 26 - 23 = 3$, whereas $\mu_{21} - \mu_{22} = 20 - 13 = 7$. Also, $\mu_{11} - \mu_{12} = \mu_{21} - \mu_{22}$.

TABLE 19.11 Layouts for one main effect and no interaction

(a) Main Effect Due to PSN

(b) Main Effect Due to NHT

PSN			Row Mean	PSN			Row Mean
NHT	Low	High		NHT	Low	High	
Low	18	26	22	Low	19	19	19
High	18	26	22	High	24	24	24
Column Mean	18	26	22	Column Mean	21.5	21.5	21.5

TABLE 19.12 Layouts for same-direction interaction

(a)		PSN				(b)		PSN				(c)		PSN			
NHT		Low	High	Row Mean	NHT	Low	High	Row Mean	NHT	Low	High	Row Mean	NHT	Low	High	Row Mean	
Low		26	23	24.5	Low	18	26	22	Low	18	26	22	Low	18	26	22	
High		20	13	16.5	High	20	36	28	High	12	24	18	High	12	24	18	
Column Mean		23	18	20.5	Column Mean	19	31	25	Column Mean	15	25	20	Column Mean	15	25	20	

© Cengage Learning

$= 26 - 24.5 = 1.5$, and $\mu_{11} - \mu_{..} = 23 - 20.5 = 2.5$, so $\mu_{11} - \mu_{1.} - \mu_{.1} + \mu_{..} = -1.0$. The other interactions of the general form ($\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..}$) are similarly determined to be either 1.0 (for $i = 1, j = 2$ or for $i = 2, j = 1$) or -1.0 (for $i = 2, j = 2$).

These hypothetical results in Table 19.12(a) indicate that in *both* low- and high-NHT neighborhoods, persons in friendly surroundings (i.e., high PSN) are healthier (i.e., have a lower CMI score) than persons in unfriendly surroundings (i.e., low PSN) but that the extent of this difference is greater when there are a large number of households (high NHT) than when there are a small number (low NHT). In other words, at each level of NHT, the difference between the PSN-level effects lies in the same direction (i.e., high PSN is associated with a lower mean CMI score than is low PSN), but the magnitude of the difference depends on the NHT level. This is the meaning of *same-direction interaction*.

The model for Table 19.12(a) may be expressed as

$$\mu_{ij} = 20.5 + \alpha_i + \beta_j + \gamma_{ij}$$

where

$$\alpha_i = \begin{cases} 4.0 & \text{if } i = 1 \\ -4.0 & \text{if } i = 2 \end{cases} \quad \beta_j = \begin{cases} 2.5 & \text{if } j = 1 \\ -2.5 & \text{if } j = 2 \end{cases}$$

$$\gamma_{ij} = \begin{cases} -1.0 & \text{if } i = 1, j = 1 \\ 1.0 & \text{if } i = 1, j = 2 \\ 1.0 & \text{if } i = 2, j = 1 \\ -1.0 & \text{if } i = 2, j = 2 \end{cases}$$

Thus,

$$\mu_{11} = 20.5 + 4.0 + 2.5 - 1.0 = 26$$

$$\mu_{12} = 20.5 + 4.0 - 2.5 + 1.0 = 23$$

$$\mu_{21} = 20.5 - 4.0 + 2.5 + 1.0 = 20$$

$$\mu_{22} = 20.5 - 4.0 - 2.5 - 1.0 = 13$$

The same general type of model holds for Tables 19.12(b) and (c).

TABLE 19.13 Layouts for reverse interaction

(a)		PSN			(b)		PSN		
NHT		Low	High	Row Mean	NHT		Low	High	Row Mean
Low		18	26	22	Low		26	22	24
High		22	20	21	High		18	24	21
Column Mean		20	23	21.5	Column Mean		22	23	22.5

© Cengage Learning

Reverse Interaction

Two examples of reverse interaction are given in Table 19.13. In these instances, the direction of the difference between two cell means for one row (column) is opposite to, or reversed from, the direction of the difference between the corresponding cell means for some other row (column).

In Table 19.13(a), we see that $\mu_{11} - \mu_{12} = 18 - 26 = -8$, whereas $\mu_{21} - \mu_{22} = 22 - 20 = 2$. Also, $\mu_{21} - \mu_{2.} = 22 - 21 = 1$, and $\mu_{.1} - \mu_{..} = 20 - 21.5 = -1.5$, so $\mu_{21} - \mu_{2.} - \mu_{.1} + \mu_{..} = 2.5$.

These hypothetical results for this table indicate that, for neighborhoods with a small number of households (low NHT), persons in unfriendly surroundings (low PSN) are healthier than are persons in friendly surroundings; but for neighborhoods with a large number of households, the situation is reversed. In other words, the difference between the effects of the high and low PSN levels is positive for low NHT but negative for high NHT (i.e., there is a reversal in sign). This is the meaning of *reverse interaction*.

The model in this case is given as

$$\mu_{ij} = 21.5 + \alpha_i + \beta_j + \gamma_{ij}$$

where

$$\alpha_i = \begin{cases} 0.5 & \text{if } i = 1 \\ -0.5 & \text{if } i = 2 \end{cases} \quad \beta_j = \begin{cases} -1.5 & \text{if } j = 1 \\ 1.5 & \text{if } j = 2 \end{cases}$$

$$\gamma_{ij} = \begin{cases} -2.5 & \text{if } i = 1, j = 1 \\ 2.5 & \text{if } i = 1, j = 2 \\ 2.5 & \text{if } i = 2, j = 1 \\ -2.5 & \text{if } i = 2, j = 2 \end{cases}$$

Thus,

$$\mu_{11} = 21.5 + 0.5 - 1.5 - 2.5 = 18$$

$$\mu_{12} = 21.5 + 0.5 + 1.5 + 2.5 = 26$$

$$\mu_{21} = 21.5 - 0.5 - 1.5 + 2.5 = 22$$

$$\mu_{22} = 21.5 - 0.5 + 1.5 - 2.5 = 20$$

It can be shown that a reverse interaction is indicated in Table 19.13(b), as well.

19.6.3 Interaction Effects for Data of Table 19.1

The table of sample cell means actually obtained for the CMI example of Section 19.2 is given in Table 19.14. From this table, the following differences of means can be determined:

$$\bar{Y}_{11\cdot} - \bar{Y}_{12\cdot} = 18.92 - 26.84 = -7.92 \text{ whereas } \bar{Y}_{21\cdot} - \bar{Y}_{22\cdot} = 20.84 - 20.72 = 0.12$$

$$\bar{Y}_{11\cdot} - \bar{Y}_{21\cdot} = 18.92 - 20.84 = -1.92 \text{ whereas } \bar{Y}_{12\cdot} - \bar{Y}_{22\cdot} = 26.84 - 20.72 = 6.12$$

$$\bar{Y}_{11\cdot} - \bar{Y}_{1\cdot\cdot} - \bar{Y}_{\cdot 1\cdot} + \bar{Y}_{\cdot\cdot\cdot} = 18.92 - 22.88 - 19.88 + 21.83 = -2.01$$

$$\bar{Y}_{12\cdot} - \bar{Y}_{1\cdot\cdot} - \bar{Y}_{\cdot 2\cdot} + \bar{Y}_{\cdot\cdot\cdot} = 26.84 - 22.88 - 23.78 + 21.83 = 2.01$$

$$\bar{Y}_{21\cdot} - \bar{Y}_{2\cdot\cdot} - \bar{Y}_{\cdot 1\cdot} + \bar{Y}_{\cdot\cdot\cdot} = 20.84 - 20.78 - 19.88 + 21.83 = 2.01$$

$$\bar{Y}_{22\cdot} - \bar{Y}_{2\cdot\cdot} - \bar{Y}_{\cdot 2\cdot} + \bar{Y}_{\cdot\cdot\cdot} = 20.72 - 20.78 - 23.78 + 21.83 = -2.01$$

TABLE 19.14 Cell means for data of Table 19.1

NHT	PSN		Row Mean
	Low	High	
Low	18.92	26.84	22.88
High	20.84	20.72	20.78
Column Mean	19.88	23.78	21.83

© Cengage Learning

These comparisons suggest a possible reverse interaction. Specifically, for small turnkey neighborhoods, persons from friendly surroundings (high PSN) appear to have worse health (higher mean CMI scores) than do persons from unfriendly surroundings; but little difference in mean CMI scores is found on the friendliness variable for large turnkey neighborhoods. As was mentioned in Section 19.2, this pattern runs counter to what Daly (1973) expected to find. However, these observed differences are subject to sampling variation (i.e., they are sample values and not population values), and the test for interaction for these data does not reveal significant interaction.

19.6.4 Interaction Effects for Data of Table 19.5

The table of sample cell means for the FEV example (Table 19.5) is as shown in Table 19.15. This table of means is slightly more difficult to interpret than the one for the CMI data because it contains three rows and columns instead of two. Nevertheless, we can immediately observe that the relative magnitudes of the means vary from column to column. For example, for Toxsub A, the order of Plants by increasing mean FEV level is 3, 1, 2. For Toxsub B, on

TABLE 19.15 Cell means for data of Table 19.5

Plant	Toxsub			Row Mean
	A	B	C	
1	4.90	3.46	2.75	3.70
2	5.23	3.78	2.70	3.90
3	4.13	5.13	3.14	4.13
Column Mean	4.75	4.12	2.86	3.91

© Cengage Learning

the other hand, the order is 1, 2, 3, and for Toxsub C the order is 2, 1, 3. These differences in ordering indicate some interaction, the significance of which was established earlier. The following comparisons of cell means should help in interpreting the nature of this significant interaction effect:

$$\begin{aligned}\bar{Y}_{11\cdot} - \bar{Y}_{12\cdot} &= 4.90 - 3.46 = 1.44 \quad \text{whereas} \quad \bar{Y}_{31\cdot} - \bar{Y}_{32\cdot} = 4.13 - 5.13 = -1.00 \\ \bar{Y}_{21\cdot} - \bar{Y}_{31\cdot} &= 5.23 - 4.13 = 1.10 \quad \text{whereas} \quad \bar{Y}_{22\cdot} - \bar{Y}_{32\cdot} = 3.78 - 5.13 = -1.35 \\ \bar{Y}_{21\cdot} - \bar{Y}_{1\cdot} &= 5.23 - 4.75 = 0.48 \quad \text{whereas} \quad \bar{Y}_{2\cdot} - \bar{Y}_{\dots} = 3.90 - 3.91 = -0.01\end{aligned}$$

The set of interaction effects of the form $\hat{\gamma}_{ij} = \bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\dots} - \bar{Y}_{\cdot j\dots} + \bar{Y}_{\dots\dots}$ is given in Table 19.16. These patterns demonstrate that some plants are associated with better respiratory health than others for one kind of toxic exposure but are worse for other kinds of such exposure. Therefore, we cannot conclude that one plant was better overall than another. Rather, the differences in respiratory health among plants depend on which toxic substance is being considered.

TABLE 19.16 Interaction effects ($\hat{\gamma}_{ij}$ values) for data of Table 19.5

Plant	Toxsub			Row Mean
	A	B	C	
1	0.35	-0.45	0.10	0.00
2	0.49	-0.34	-0.15	0.00
3	-0.84	0.78	0.06	0.00
Column Mean	0.00	0.00	0.00	0.00

© Cengage Learning

19.7 Random- and Mixed-effects Two-way ANOVA Models

In this section, we examine the classical two-way ANOVA statistical models appropriate when both factors are random or when one factor is fixed and the other is random. We will specify the appropriate null hypotheses of interest and the expected mean squares associated with each model. The expected mean square for a particular source is the true average (population) value of the mean-square term in the ANOVA table.

19.7.1 Random-effects Model⁷

When both factors are random, the two-way ANOVA model is given as

$$Y_{ijk} = \mu + A_i + B_j + C_{ij} + E_{ijk} \quad (19.7)$$

where A_i , B_j , C_{ij} , and E_{ijk} are mutually independent random variables satisfying

$$A_i \sim N(0, \sigma_R^2)$$

$$B_j \sim N(0, \sigma_C^2)$$

$$C_{ij} \sim N(0, \sigma_{RC}^2)$$

$$E_{ijk} \sim N(0, \sigma^2) \quad i = 1, 2, \dots, r; \quad j = 1, 2, \dots, c; \quad k = 1, 2, \dots, n$$

19.7.2 Mixed-effects Model with Fixed Row Factor and Random Column Factor

One particular model⁸ is

$$Y_{ijk} = \mu + \alpha_i + B_j + C_{ij} + E_{ijk} \quad (19.8)$$

⁷ As previously footnoted in Chapter 17 for one-way ANOVA models, whenever random effects occur in a two-way ANOVA model, we can no longer assume that the Y_{ijk} are mutually independent.

⁸ There are several ways to define the assumptions for a two-way mixed model. In particular, one must choose between two alternative forms of summation restrictions for the random components of a two-way mixed model (see Searle, Casella, and McCulloch 1992). The assumptions we have made (in the body of the text) for mixed models (19.8) and (19.9) include summation restrictions for the fixed effects—e.g., $\sum_{i=1}^r \alpha_i = 0$ for model (19.8)—but no such summation restrictions for random effects, particularly the (random) interaction effects C_{ij} . Alternatively, we can choose to assume (for either model 19.8 or model 19.9) that the sum of the random interaction factors C_{ij} over the levels of the fixed factor are zero; i.e., $\sum_{i=1}^r C_{ij} = 0$ for each j in model (19.8). The latter summation restriction is what we assumed in previous editions of this text. Moreover, this restriction implied that C_{ij} and $C_{i'j}$, $i \neq i'$, are correlated, which leads to a different formula for the F statistic for the test of hypothesis about the random main effects in the mixed model; specifically, the denominator mean square is MSE instead of MSRC. The question of which form of mixed model to use—one without the restrictions on the C_{ij} or one with them—remains open (Searle et al. 1992). Nevertheless, we have chosen to assume (in the main body of the text) no restrictions on the C_{ij} because the F statistics based on no restrictions correspond to those calculated by standard computer programs for ANOVA models with mixed effects (including SAS's GLM, whose output we illustrate in the text). Also, assumptions not involving summation restrictions correspond to assumptions statisticians typically make when considering the analysis of repeated measures data, which we discuss in Chapters 25 and 26.

where each α_i is a constant such that $\sum_{i=1}^r \alpha_i = 0$ and where B_j , C_{ij} , and E_{ijk} are mutually independent random variables satisfying

$$B_j \sim N(0, \sigma_C^2)$$

$$C_{ij} \sim N(0, \sigma_{RC}^2) \quad i = 1, 2, \dots, r; j = 1, 2, \dots, c; k = 1, 2, \dots, n$$

$$E_{ijk} \sim N(0, \sigma^2)$$

19.7.3 Mixed-effects Model with Random Row Factor and Fixed Column Factor

One particular model is

$$Y_{ijk} = \mu + A_i + \beta_j + C_{ij} + E_{ijk} \quad (19.9)$$

where each β_j is a constant such that $\sum_{j=1}^c \beta_j = 0$ and where A_i , C_{ij} , and E_{ijk} are mutually independent random variables satisfying

$$A_i \sim N(0, \sigma_R^2)$$

$$C_{ij} \sim N(0, \sigma_{RC}^2) \quad i = 1, 2, \dots, r; j = 1, 2, \dots, c; k = 1, 2, \dots, n$$

$$E_{ijk} \sim N(0, \sigma^2)$$

19.7.4 Null Hypotheses and Expected Mean Squares for Two-way ANOVA Models

For fixed-, random-, and mixed-effects models, Table 19.17 gives the specific null hypotheses being tested with regard to row main effects, column main effects, and interaction effects. Table 19.18 gives the expected mean square for each factor in each of the models. These two tables demonstrate why different F statistics are required for testing the various hypotheses of interest. In this regard, the primary consideration is the choice of the appropriate denominator mean squares to use in the various F statistics. The numerator mean square always corresponds to the factor being considered; for example, if the factor is “rows,” the numerator mean square is MSR, regardless of the type of model. Similarly, if the factor is “columns” or “interaction,” the numerator mean square is MSC or MSRC, respectively. The denominator mean square, however, is chosen to correspond to the expected mean square to which the numerator expected mean square reduces under the null hypothesis of interest. For example, in a test for significant row effects in a random-effects model, the numerator expected mean square, $(\sigma^2 + n\sigma_{RC}^2 + cn\sigma_R^2)$ from Table 19.18, reduces to $(\sigma^2 + n\sigma_{RC}^2)$ under $H_0: \sigma_R^2 = 0$. This requires that the denominator mean square be MSRC, since the expected mean square of MSRC under the random-effects model is exactly $(\sigma^2 + n\sigma_{RC}^2)$.

Thus, the ratio of expected mean squares

$$\frac{\text{EMS}(R)}{\text{EMS}(RC)} = \frac{\sigma^2 + n\sigma_{RC}^2 + cn\sigma_R^2}{\sigma^2 + n\sigma_{RC}^2}$$

TABLE 19.17 Null hypotheses for two-way ANOVA

Source	Model Type			
	Mixed Effects			
	Fixed Effects	Random Effects	Rows Fixed, Columns Random	Rows Random, Columns Fixed
Rows	$\alpha_1 = \alpha_2 = \dots = \alpha_r = 0$	$\sigma_R^2 = 0$	$\alpha_1 = \alpha_2 = \dots = \alpha_r = 0$	$\sigma_R^2 = 0$
Columns	$\beta_1 = \beta_2 = \dots = \beta_c = 0$	$\sigma_C^2 = 0$	$\sigma_C^2 = 0$	$\beta_1 = \beta_2 = \dots = \beta_c = 0$
Interactions	$\gamma_{ij} = 0$ for all i, j	$\sigma_{RC}^2 = 0$	$\sigma_{RC}^2 = 0$	$\sigma_{RC}^2 = 0$

© Cengage Learning

reduces to $(\sigma^2 + n\sigma_{RC}^2)/(\sigma^2 + n\sigma_{RC}^2) = 1$ under $H_0: \sigma_R^2 = 0$, so the F statistic MSR/MSRC is the ratio of two estimators of the same variance under H_0 .

As another example, let us consider the F test for significant row effects based on the mixed-effects model with the row factor fixed and column factor random. The test statistic in this case, $F = \text{MSR}/\text{MSRC}$, involves the following ratio of expected mean squares (see Table 19.18):

$$\frac{\text{EMS}(R)}{\text{EMS}(RC)} = \frac{\sigma^2 + n\sigma_{RC}^2 + cn \sum_{i=1}^r \alpha_i^2 / (r-1)}{\sigma^2 + n\sigma_{RC}^2}$$

Under $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_r = 0$, this ratio simplifies to $(\sigma^2 + n\sigma_{RC}^2)/(\sigma^2 + n\sigma_{RC}^2) = 1$. Thus, the F statistic is the ratio of two estimators of the same variance under H_0 .

TABLE 19.18 Expected mean squares for two-way ANOVA (r rows, c columns, n observations per cell)

Source	Model Type			
	Mixed Effects			
	Fixed Effects	Random Effects	Rows Fixed, Columns Random	Rows Random, Columns Fixed
Rows	$\sigma^2 + cn \sum_{i=1}^r \frac{\alpha_i^2}{r-1}$	$\sigma^2 + n\sigma_{RC}^2 + cn\sigma_R^2$	$\sigma^2 + n\sigma_{RC}^2 + cn \sum_{i=1}^r \frac{\alpha_i^2}{r-1}$	$\sigma^2 + n\sigma_{RC}^2 + cn\sigma_R^2$
Columns	$\sigma^2 + rn \sum_{j=1}^c \frac{\beta_j^2}{c-1}$	$\sigma^2 + n\sigma_{RC}^2 + rn\sigma_C^2$	$\sigma^2 + n\sigma_{RC}^2 + rn\sigma_C^2$	$\sigma^2 + n\sigma_{RC}^2 + rn \sum_{j=1}^c \frac{\beta_j^2}{c-1}$
Interactions	$\sigma^2 + n \sum_{i=1}^r \sum_{j=1}^c \frac{\gamma_{ij}^2}{(r-1)(c-1)}$	$\sigma^2 + n\sigma_{RC}^2$	$\sigma^2 + n\sigma_{RC}^2$	$\sigma^2 + n\sigma_{RC}^2$
Error	σ^2	σ^2	σ^2	σ^2

© Cengage Learning

As a final example, we consider the F test for significant row effects based on the mixed-effects model with the row factor random and the column factor fixed. The test statistic is $F = \text{MSR}/\text{MSRC}$, which involves

$$\frac{\text{EMS}(\text{R})}{\text{EMS}(\text{RC})} = \frac{\sigma^2 + n\sigma_{\text{RC}}^2 + cn\sigma_{\text{R}}^2}{\sigma^2 + n\sigma_{\text{RC}}^2}$$

Under $H_0: \sigma_{\text{R}}^2 = 0$, this ratio simplifies to

$$\frac{\sigma^2 + n\sigma_{\text{RC}}^2}{\sigma^2 + n\sigma_{\text{RC}}^2} = 1$$

as desired.

Problems

- The data in the following table and accompanying computer output come from an animal experiment designed to investigate whether levorphanol reduces stress as reflected in the cortical sterone level. The four treatment groups contained five animals each.

Control	Levorphanol Only	Epinephrine Only	Levorphanol and Epinephrine
1.90	0.82	5.33	3.08
1.80	3.36	4.84	1.42
1.54	1.64	5.26	4.54
4.10	1.74	4.92	1.25
1.89	1.21	6.07	2.57

- These data may be analyzed by means of two-way ANOVA. What are the two factors?
- Classify each factor as either fixed or random.
- Rearrange the data into a two-way table appropriate for two-way ANOVA.
- Form the table of sample means, and comment on the patterns observed.
- Complete the ANOVA table shown below.
- Analyze the data to determine whether significant main effects exist due to levorphanol and epinephrine and whether a significant interaction effect exists between epinephrine and levorphanol.

Edited SAS Output (PROC GLM) for Problem 1

Dependent Variable: LEVEL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	37.57844000	12.52614667	_____	0.0002
Error	16	_____	_____	_____	_____
Corrected Total	19	_____	_____	_____	_____

(continued)

R-Square	Coeff Var	Root MSE	Y Mean
0.697495	34.05076	1.009265	2.964000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
L	1	12.83202000	12.83202000	12.60	0.0027
E	1	18.58592000	18.58592000	18.25	0.0006
L*E	1	6.16050000	6.16050000	6.05	0.0257

Source	DF	Type III SS	Mean Square	F Value	Pr > F
L	1	12.83202000	12.83202000	12.60	0.0027
E	1	18.58592000	18.58592000	18.25	0.0006
L*E	1	6.16050000	6.16050000	6.05	0.0257

2. The following table gives the performance competency scores for a random sample of family nurse practitioners (FNPs) with different specialties, from hospitals in three cities.

Specialty	City 1	City 2	City 3
Pediatrics	91.7, 74.9, 88.2, 79.5	86.3, 88.1, 92.0, 69.5	82.3, 78.7, 89.8, 84.5
Obstetrics and gynecology	80.1, 76.2, 70.3, 89.5	71.3, 73.4, 76.9, 87.2	90.1, 65.6, 74.6, 79.1
Diabetes and hypertension	71.5, 49.8, 55.1, 75.4	80.2, 76.1, 44.2, 50.5	48.7, 54.4, 60.1, 70.8

- a. Classify each factor as either fixed or random, and justify your classification.
- b. Form the table of sample means (use the accompanying computer output to do so), and then comment on the patterns observed.
- c. Using the computer output, compute the appropriate F statistic for each of the four possible factor classification schemes (i.e., both factors fixed, both random, and one factor of each type).
- d. Analyze the data based on each possible factor classification scheme. How do the results compare?
- e. Using Scheffé's method, as described in Chapter 17, find a 95% confidence interval for the true difference in mean scores between pediatric FNPs and ob-gyn FNPs.
- f. Using the sample means obtained in part (b) and assuming each factor to be fixed, state a regression model appropriate for the two-way ANOVA table, and provide estimates of the regression coefficients associated with the main effects of each factor.

Edited SAS Output (PROC GLM) for Problem 2

Class Level Information		
Class	Levels	Values
SPEC	3	D&H OBGYN PED
CITY	3	1 2 3

Dependent Variable: SCORE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	3288.950000	411.118750	3.82	0.0040
Error	27	2902.180000	107.488148		
Corrected Total	35	6191.130000			

R-Square	Coeff Var	Root MSE	SCORE Mean
0.531236	13.94438	10.36765	74.35000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
SPEC	2	3229.871667	1614.935833	15.02	<.0001
CITY	2	24.541667	12.270833	0.11	0.8925
SPEC*CITY	4	34.536667	8.634167	0.08	0.9877

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SPEC	2	3229.871667	1614.935833	15.02	<.0001
CITY	2	24.541667	12.270833	0.11	0.8925
SPEC*CITY	4	34.536667	8.634167	0.08	0.9877

Source	Type III Expected Mean Square
SPEC	Var(Error) + 4 Var(SPEC*CITY) + 12 Var(SPEC)
CITY	Var(Error) + 4 Var(SPEC*CITY) + 12 Var(CITY)
SPEC*CITY	Var(Error) + 4 Var(SPEC*CITY)

Test of Hypotheses for Random Model Analysis of Variance

Dependent Variable: Score

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SPEC	2	3229.871667	1614.935833	187.04	0.0001
CITY	2	24.541667	12.270833	1.42	0.3417
Error: MS(SPEC*CITY)	4	34.536667	8.634167		

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SPEC*CITY	4	34.536667	8.634167	0.08	0.9877
Error: MS(Error)	27	2902.180000	107.488148		

(continued)

Level of SPEC	Level of CITY	N	SCORE	
			Mean	Std Dev
D&H	1	4	62.9500000	12.4184003
D&H	2	4	62.7500000	18.0452210
D&H	3	4	58.5000000	9.4286797
OBGYN	1	4	79.0250000	8.0619993
OBGYN	2	4	77.2000000	7.0554943
OBGYN	3	4	77.3500000	10.1857744
PED	1	4	83.5750000	7.7301897
PED	2	4	83.9750000	9.9389386
PED	3	4	83.8250000	4.6456969

3. The following table gives the average patient waiting time in minutes for patients from a random sample of 16 physicians, classified by type of practice and type of physician.

Physician Type	Type of Practice	
	Group	Solo
General practitioner	15, 20, 25, 20	20, 25, 30, 25
Specialist	30, 25, 30, 35	25, 20, 30, 30

- a. Classify each factor as either fixed or random, and justify your classification scheme.
- b. Using the accompanying computer output, compute the F statistic corresponding to each of the four possible factor classification schemes.
- c. Discuss the analysis of the data when both factors are considered fixed.
- d. What are the estimates of the (fixed) effect due to “general practitioner,” the (fixed) effect due to “group practice,” and the interaction effect ($\mu_{11} - \mu_{12} - \mu_{21} + \mu_{22}$), where μ_{ij} denotes the cell mean in the i th row and j th column of the table of cell means?
- e. Interpret the interaction effect observed.
- f. What is an appropriate regression model for this two-way ANOVA?
- g. How might you modify the model in part (f) to reflect the conclusions made in part (c)?

Edited SAS Output (PROC GLM) for Problem 3

CLASS LEVEL INFORMATION		
Class	Levels	Values
PHYS	2	GEN SPEC
PRACTICE	2	GROUP SOLO

(continued)

Dependent Variable: TIME

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	204.6875000	68.2291667	3.74	0.0415
Error	12	218.7500000	18.2291667		
Corrected Total	15	423.4375000			

R-Square	Coeff Var	Root MSE	TIME Mean
0.483395	16.86741	4.269563	25.31250

Source	DF	Type I SS	Mean Square	F Value	Pr > F
PHYS	1	126.5625000	126.5625000	6.94	0.0218
PRACTICE	1	1.5625000	1.5625000	0.09	0.7747
PHYS*PRACTICE	1	76.5625000	76.5625000	4.20	0.0629

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PHYS	1	126.5625000	126.5625000	6.94	0.0218
PRACTICE	1	1.5625000	1.5625000	0.09	0.7747
PHYS*PRACTICE	1	76.5625000	76.5625000	4.20	0.0629

Source	Type III Expected Mean Square
PHYS	Var(Error) + 4 Var(PHYS*PRACTICE) + 8 Var(PHYS)
PRACTICE	Var(Error) + 4 Var(PHYS*PRACTICE) + 8 Var(PRACTICE)
PHYS*PRACTICE	Var(Error) + 4 Var(PHYS*PRACTICE)

Test of Hypotheses for Random Model Analysis of Variance

Dependent Variable: TIME

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PHYS	1	126.562500	126.562500	1.65	0.4208
PRACTICE	1	1.562500	1.562500	0.02	0.9097
Error	1	76.562500	76.562500		
Error: MS(PHYS*PRACTICE)					

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PHYS*PRACTICE	1	76.562500	76.562500	4.20	0.0629
Error: MS(Error)	12	218.750000	18.229167		

Level of PHYS	Level of PRACTICE	N	TIME	
			Mean	Std Dev
GEN	GROUP	4	20.0000000	4.08248290
GEN	SOLO	4	25.0000000	4.08248290
SPEC	GROUP	4	30.0000000	4.08248290
SPEC	SOLO	4	26.2500000	4.78713554

4. A study was undertaken to measure and compare sexist attitudes of students at various types of colleges. Random samples of 10 undergraduate seniors of each sex were selected from each of three types of colleges. A questionnaire was then administered to each student, from which a score for “degree of sexism”—defined as the extent to which a student considered males and females to have different life roles—was determined (the higher the score, the more sexist the attitude). The resulting data are given in the following table.

College Type	Male	Female
Coed with 75% or more males	50, 35, 37, 32, 46, 38, 36, 40, 38, 41	38, 27, 34, 30, 22, 32, 26, 24, 31, 33
Coed with less than 75% males	30, 29, 31, 27, 22, 20, 31, 22, 25, 30	28, 31, 28, 26, 20, 24, 31, 24, 31, 26
Not coed	45, 40, 32, 31, 26, 28, 39, 27, 37, 35	40, 35, 32, 29, 24, 26, 36, 25, 35, 35

- a. Form the table of cell means, and interpret the results obtained (see the accompanying computer printout).
- b. Using the computer output, calculate the F statistics corresponding to a model with both factors fixed.
- c. Discuss the analysis of the data for this fixed-effects model case.

Edited SAS Output (PROC GLM) for Problem 4

Dependent Variable: SCORE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1144.883333	228.976667	9.06	<.0001
Error	54	1365.300000	25.283333		
Corrected Total	59	2510.183333			

R-Square	Coeff Var	Root MSE	SCORE Mean
0.456096	16.02205	5.028254	31.38333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
COLLEGE	2	657.4333333	328.7166667	13.00	<.0001
SEX	1	228.1500000	228.1500000	9.02	0.0040
COLLEGE*SEX	2	259.3000000	129.6500000	5.13	0.0091

Source	DF	Type III SS	Mean Square	F Value	Pr > F
COLLEGE	2	657.4333333	328.7166667	13.00	<.0001
SEX	1	228.1500000	228.1500000	9.02	0.0040
COLLEGE*SEX	2	259.3000000	129.6500000	5.13	0.0091

(continued)

Level of COLLEGE	Level of SEX	N	SCORE	
			Mean	Std Dev
<75%	FEMALE	10	26.9000000	3.63470922
<75%	MALE	10	26.7000000	4.16466619
>75%	FEMALE	10	29.7000000	4.92273637
>75%	MALE	10	39.3000000	5.31350481
NOT	FEMALE	10	31.7000000	5.41705127
NOT	MALE	10	34.0000000	6.27162924

5. Random samples of 100 persons awaiting trial on felony charges were selected from rural, urban, and suburban court locations in each of two states, one (state 1) in the Northeast and the other (state 2) in the South. The following table summarizes the data on the time \bar{Y} (in months) between arrest and beginning of trial for these random samples.

State	Court Location		
	Rural	Suburban	Urban
1	$\bar{Y} = 3.4, S = 1.3$	$\bar{Y} = 5.8, S = 1.2$	$\bar{Y} = 6.8, S = 1.5$
2	$\bar{Y} = 2.4, S = 1.5$	$\bar{Y} = 3.5, S = 1.7$	$\bar{Y} = 4.7, S = 1.7$

- a. Do the sample means in the table suggest that the average waiting times for state 1 vary by court location differently from how they vary for state 2? Is there an interaction effect?
- b. Analyze these data. Use the following ANOVA table. Assume that both factors are fixed.

Source	d.f.	SS	MS
States	1	486.00	486.00
Court locations	2	826.33	413.17
Interaction	2	49.00	24.50
Error	594	1,327.591	2.235

- c. Define an appropriate regression model for this two-way ANOVA.
- d. How might one revise the model in part (c) and the associated ANOVA table in order to investigate whether a linear trend exists between waiting time and degree of urbanization (as determined by treating the categories rural, suburban, and urban on an ordinal scale)? What difficulty does one encounter when considering such a model?
- 6. An experiment was conducted at a large state university to determine whether two different instructional methods for teaching a beginning statistics course would yield different levels of cognitive achievement. One instructional method involved using a self-instructional format, including a sequence of slide-tape presentations; the other

method utilized the standard lecture format. The 100 students who registered for the course were randomly assigned to one of four sections, 25 per section, corresponding to the combinations of one of the two methods with one of two instructors. The results obtained from identical final exams given to each section are summarized in the following table.

Instructor	Method	
	Lecture	Self-instruction
A	$\bar{Y} = 71.2, S = 13.8$	$\bar{Y} = 80.2, S = 12.1$
B	$\bar{Y} = 73.8, S = 11.7$	$\bar{Y} = 77.5, S = 14.1$

- a. What do the results suggest about the comparative effects of the two instructional methods?
 - b. Classify each factor as either fixed or random, and explain your classification.
-
- | Source | d.f. | SS | MS |
|-------------------------|------|--------------|--------------|
| INSTRUC | 1 | 6.2500E - 02 | 6.2500E - 02 |
| METHOD | 1 | 6.0081E + 03 | 1.0081E + 03 |
| INSTRUC \times METHOD | 1 | 1.7556E + 02 | 1.7556E + 02 |
| Error | 96 | 1.6141E + 04 | 1.6814E + 02 |
-
- c. Using the ANOVA table, perform the appropriate F tests for each of the four types of factor classification schemes. Compare the conclusions reached under each scheme.
 - d. What factors should be controlled for in this experiment?
 - e. Given a continuous variable C to be controlled for, write an appropriate regression model for this data set that takes C into account. What general method of analysis is characterized by such a model?
7. The following table and accompanying computer output present data on the uric acid level found in the bloodstreams of persons with Down's syndrome and in the bloodstreams of non-Down's syndrome subjects. All subjects were between the ages of 21 and 25. Analyze these data, using the ANOVA table to determine whether evidence of a higher uric acid level exists in the group with Down's syndrome, making sure to characterize any sex relationships that exist.

Group	Sex	
	Male	Female
Down's syndrome	5.84, 6.30, 6.95, 5.92, 7.94	4.90, 6.95, 6.73, 5.32, 4.81
Others	5.50, 6.08, 5.12, 7.58, 6.78	4.94, 7.20, 5.22, 4.60, 3.88

Edited SAS Output (PROC GLM) for Problem 7

Dependent Variable: ACID

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5.65548000	1.88516000	1.74	0.1987
Error	16	17.31484000	1.08217750		
Corrected Total	19	22.97032000			
R-Square		Coeff Var		Root MSE	
0.246208		17.54854		1.040278	
Source		DF		Type I SS	
GROUP		1		1.13288000	
SEX		1		4.47458000	
GROUP*SEX		1		0.04802000	
Source		DF		Type III SS	
GROUP		1		1.13288000	
SEX		1		4.47458000	
GROUP*SEX		1		0.04802000	
Level of GROUP		Level of SEX		ACID	
		N		Mean	
MONGOLOID		FEMALE		5.74200000	
MONGOLOID		MALE		6.59000000	
OTHERS		FEMALE		5.16800000	
OTHERS		MALE		6.21200000	
				Std Dev	
				1.02360637	
				0.87286883	
				1.24149909	
				0.98879725	

8. An experiment was conducted to investigate the survival of diplococcus pneumonia bacteria in chick embryos under relative humidities (RH) of 0%, 25%, 50%, and 100% and under temperatures (Temp) of 10°C, 20°C, 30°C, and 40°C, using 10 chicks for each RH-Temp combination.⁹ The partially completed ANOVA table is as given next.

Source	d.f.	MS
RH		2.010
Temp		7.816
Interaction		1.642
Error		0.775
Total		

⁹ Adapted from a study by Price (1954).

- a. Should the two factors RH and Temp be considered as fixed or random? Explain.
 - b. Carry out the analysis of variance for both the fixed-effects case and the random-effects case. Do your conclusions differ in the two cases?
 - c. Write the fixed-effects and the random-effects models that could describe this experiment.
 - d. Using dummy variables, provide a regression model that can be used to obtain the results in the ANOVA table.
 - e. What regression model would be appropriate for describing the relationship of RH and Temp to survival time (Y) if the data for the independent variables are to be treated as interval rather than as nominal?
9. The diameters (Y) of three species of pine trees were compared at each of four locations, using samples of five trees per species at each location. The resulting data are given in the following table.

Species	Location			
	1	2	3	4
A	23	25	21	14
	15	20	17	17
	26	21	16	19
	13	16	24	20
	21	18	27	24
B	28	30	19	17
	22	26	24	21
	25	26	19	18
	19	20	25	26
	26	28	29	23
C	18	15	23	18
	10	21	25	12
	12	22	19	23
	22	14	13	22
	13	12	22	19

- a. Comment on whether each of the two factors should be considered fixed or random.
- b. Use the following partially completed ANOVA table to carry out your analysis, first considering both factors as fixed and then considering a mixed model with “locations” treated as random. Compare your conclusions.

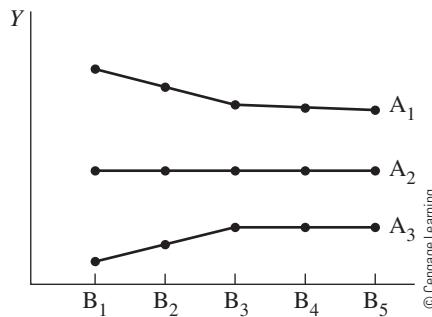
Source	d.f.	SS
Species		344.9333
Locations		46.0500
Interaction		113.6000
Error		875.6000

10. Consider an ANOVA table of the following form.

Source	d.f.	SS	MS	F	P
A					P_1
B					P_2
$A \times B$					P_3
Error					

Use $\alpha = .05$ for all parts of this problem. In each case, decide what effects (if any) are significant, and what conclusions to draw, based on a two-way ANOVA table with the following P -values:

- a. $P_1 = .03, P_2 = .51, P_3 = .31$
 - b. $P_1 = .001, P_2 = .63, P_3 = .007$
 - c. $P_1 = .093, P_2 = .79, P_3 = .02$
 - d. $P_1 = .56, P_2 = .38, P_3 = .24$
11. Assume that a total of 75 subjects were tested in a balanced two-way fixed-effects factorial experiment. A plot of the means from the study is shown next. The dependent variable is Y , and the factors are A and B.



- a. Give the left two columns (the source and degrees-of-freedom columns) of the ANOVA table for the data plotted.
 - b. Assume that the plot shows the population means. Indicate which significance tests should hopefully yield significant results in the ANOVA table.
12. Assume that the following ANOVA table came from a balanced two-way fixed-effects ANOVA. Show the formula used (with numbers filled in) and the numerical value of each letter in the table.

Source	d.f.	SS	MS	F
A	a	5.12	g	6.40*
B	b	e	0.76	3.80*
$A \times B$	c	4.32	0.36	i
Error	20	4.00	h	
Total	d	f		

Also, indicate which (if any) family or families of means should be evaluated with multiple comparisons.

13. The data in the following table and accompanying computer output are from a hypothetical study of human body temperature as affected by air temperature and a dietary supplement that is hoped to increase heat tolerance. Body temperatures (in degrees Celsius) were measured for 36 athletes immediately following a standard exercise routine in a room controlled to a fixed air temperature (in degrees Celsius). Each subject had been receiving a steady, fixed dose (in milligrams per kilogram of body weight) of the dietary supplement.
- Provide an appropriate two-way ANOVA table.
 - Provide F tests of the two main effects and the interaction. Use $\alpha = .05$. What do you conclude?
 - Define dummy variables, and specify an appropriate multiple regression model corresponding to the analysis done in part (a). (*Hint:* Use dummy variables that have the value -1 for Airtemp = 21 and for Dose = 0. Why is this a scientifically sensible choice?)
 - Since both factors are interval-scale variables, specify a corresponding natural-polynomial multiple regression model.

Edited SAS Output (PROC GLM) for Problem 13

Dependent Variable: BODYTEMP

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	0.13888889	0.01262626	0.43	0.9255
Error	24	0.70000000	0.02916667		
Corrected Total	35	0.83888889			

R-Square	Coeff Var	Root MSE	BODYTEMP Mean
0.165563	0.461644	0.170783	36.99444

Source	DF	Type I SS	Mean Square	F Value	Pr > F
AIRTEMP	3	0.02777778	0.00925926	0.32	0.8126
DOSE	2	0.06055556	0.03027778	1.04	0.3695
AIRTEMP*DOSE	6	0.05055556	0.00842593	0.29	0.9364

Source	DF	Type III SS	Mean Square	F Value	Pr > F
AIRTEMP	3	0.02777778	0.00925926	0.32	0.8126
DOSE	2	0.06055556	0.03027778	1.04	0.3695
AIRTEMP*DOSE	6	0.05055556	0.00842593	0.29	0.9364

Observation	Airtemp	Dose	Bodytemp
1	21	0.00	37.2
2	21	0.00	37.2
3	21	0.00	36.8
4	21	0.05	37.1
5	21	0.05	36.9
6	21	0.05	36.8
7	21	0.10	37.1
8	21	0.10	37.1
9	21	0.10	37.1
10	25	0.00	36.9
11	25	0.00	37.0
12	25	0.00	37.1
13	25	0.05	37.1
14	25	0.05	36.7
15	25	0.05	37.0
16	25	0.10	36.9
17	25	0.10	37.0
18	25	0.10	37.3
19	29	0.00	36.9
20	29	0.00	37.0
21	29	0.00	36.8
22	29	0.05	36.9
23	29	0.05	37.0
24	29	0.05	36.9
25	29	0.10	36.9
26	29	0.10	37.0
27	29	0.10	37.2
28	33	0.00	37.1
29	33	0.00	37.3
30	33	0.00	36.7
31	33	0.05	36.9
32	33	0.05	37.0
33	33	0.05	37.0
34	33	0.10	36.9
35	33	0.10	36.8
36	33	0.10	37.2

14. The experiment in Problem 11 in Chapter 18 was repeated—this time using three workers per technology/experience combination rather than just one. For each cell in the following table, the outputs (in number of units per hour) are listed for three randomly chosen workers at the given experience level.

Operator Experience	Technology		
	High Automation (H)	Moderate Automation (M)	Low Automation (L)
<1 Year	16, 14, 12	8, 12, 10	5, 4, 8
1–2 Years	15, 18, 17	10, 10, 12	8, 10, 10
>2 Years	20, 18, 19	13, 13, 14	12, 11, 13

- a. Are the two factors in this problem—Technology and Operator Experience—fixed or random factors?
 b. Form the table of sample means for these data, and comment on observed patterns.

- c. The data may be analyzed by using a two-way ANOVA. Determine the ANOVA table for these data.
- d. Analyze the accompanying computer output to determine whether significant main effects exist due to technology type and worker experience and whether these factors significantly interact.

Edited SAS Output (PROC GLM) for Problem 14

CLASS LEVEL INFORMATION		
Class	Levels	Values
LINE	3	H L M
EXPER	3	1-2 <1>2

Dependent Variable: OUTPUT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	386.2962963	48.2870370	22.10	<.0001
Error	18	39.3333333	2.1851852		
Corrected Total	26	425.6296296			

R-Square	Coeff Var	Root MSE	OUTPUT Mean
0.907588	12.02181	1.478237	12.29630

Source	DF	Type I SS	Mean Square	F Value	Pr > F
LINE	2	269.4074074	134.7037037	61.64	<.0001
EXPER	2	107.6296296	53.8148148	24.63	<.0001
LINE*EXPER	4	9.2592593	2.3148148	1.06	0.4050

Source	DF	Type III SS	Mean Square	F Value	Pr > F
LINE	2	269.4074074	134.7037037	61.64	<.0001
EXPER	2	107.6296296	53.8148148	24.63	<.0001
LINE*EXPER	4	9.2592593	2.3148148	1.06	0.4050

Level of LINE	N	OUTPUT	
		Mean	Std Dev
H	9	16.5555556	2.55495162
L	9	9.0000000	3.04138127
M	9	11.3333333	1.93649167

Level of EXPER	N	OUTPUT	
		Mean	Std Dev
1-2	9	12.2222222	3.56292639
<1	9	9.8888889	4.01386486
>2	9	14.7777778	3.30823887

(continued)

Level of LINE	Level of EXPER	N	OUTPUT	
			Mean	Std Dev
H	1-2	3	16.6666667	1.5275253
H	<1	3	14.0000000	2.00000000
H	>2	3	19.0000000	1.00000000
L	1-2	3	9.3333333	1.15470054
L	<1	3	5.6666667	2.08166600
L	>2	3	12.0000000	1.00000000
M	1-2	3	10.6666667	1.15470054
M	<1	3	10.0000000	2.00000000
M	>2	3	13.3333333	0.57735027

15. An advertising company evaluated three types of television ads for a new, low-cost, subcompact automobile: visual-appeal ads, budget-appeal ads, and feature-appeal ads. To control for age differences, viewers from four age groups were chosen to evaluate the persuasiveness of the ads (as measured on a scale from 1 to 10, where 1 represented the lowest level of persuasion and 10 the highest). Within each age group were six viewers; two each were randomly assigned to view one of the three types of ads. The sample persuasion scores are presented in the following table.

Viewer Age	Type of Ad		
	Visual Appeal (V)	Budget Appeal (B)	Feature Appeal (F)
18–25 years	6, 5	8, 7	5, 4
26–35 years	7, 6	9, 10	4, 8
36–45 years	8, 9	9, 8	4, 2
46 and older	10, 9	10, 8	5, 4

- a. Form the table of sample means for these data, and comment on observed patterns.
- b. The data may be analyzed by using a two-way ANOVA. Determine the ANOVA table for these data.
- c. Analyze the accompanying computer output to determine whether significant main effects exist due to ad type and viewer age and whether these factors significantly interact.

Edited SAS Output (PROC GLM) for Problem 15

CLASS LEVEL INFORMATION		
Class	Levels	Values
AD	3	B F V
AGE	4	18-25 26-35 36-45 >45

(continued)

Dependent Variable: SCORE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	106.1250000	9.6477273	7.02	0.0011
Error	12	16.5000000	1.3750000		
Corrected Total	23	122.6250000			

R-Square	Coeff Var	Root MSE	SCORE Mean
0.865443	17.05606	1.172604	6.875000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
AD	2	72.75000000	36.37500000	26.45	<.0001
AGE	3	11.79166667	3.93055556	2.86	0.0814
AD*AGE	6	21.58333333	3.59722222	2.62	0.0737

Level of AD	N	SCORE	
		Mean	Std Dev
B	8	8.62500000	1.06066017
F	8	4.50000000	1.69030851
V	8	7.50000000	1.77281052

Level of AGE	N	SCORE	
		Mean	Std Dev
18-25	6	5.83333333	1.47196014
26-35	6	7.33333333	2.16024690
36-45	6	6.66666667	2.94392029
>45	6	7.66666667	2.58198890

Level of AD	Level of AGE	N	SCORE	
			Mean	Std Dev
B	18-25	2	7.50000000	0.70710678
B	26-35	2	9.50000000	0.70710678
B	36-45	2	8.50000000	0.70710678
B	>45	2	9.00000000	1.41421356
F	18-25	2	4.50000000	0.70710678
F	26-35	2	6.00000000	2.82842712
F	36-45	2	3.00000000	1.41421356
F	>45	2	4.50000000	0.70710678
V	18-25	2	5.50000000	0.70710678
V	26-35	2	6.50000000	0.70710678
V	36-45	2	8.50000000	0.70710678
V	>45	2	9.50000000	0.70710678

16. This question refers to the *U.S. News & World Report* mutual fund data presented in Problem 19 in Chapter 17. The variables described in that question were:

CAT (fund category): 1 = Aggressive growth; 2 = Long-term growth;
3 = Growth and income; 4 = Income.

LOAD (load status): N = No load; L = Load.

VOL (volatility): A letter grade from A+ to F indicating how much the month-to-month return varied from the fund's three-year total return: A+ = Least variability; F = Most variability.

OPI (Overall Performance Index): A measure of the relative performance of each fund over the past 1, 3, 5, and 10 years.

- a. Suppose that a two-way ANOVA is to be performed, with OPI as the dependent variable and fund category (CAT) and load status (LOAD) as the two factors. State precisely the ANOVA model. Are the factors fixed or random?
 - b. In the SAS output that follows, complete the ANOVA table.

Edited SAS Output (PROC GLM) for Problem 16

CLASS LEVEL INFORMATION		
Class	Levels	Values
CAT	4	1 2 3 4
LOAD	2	L N

Dependent Variable: OPI

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	—	312.0729167	44.5818452	————	————
Error	16	225.4866667	14.0929167		
Corrected Total	23	537.5595833			

R-Square	Coeff Var	Root MSE	OPI Mean
0.580536	4.165578	3.754053	90.12083

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CAT	3	282.0545833	94.0181944	6.67	0.0039
LOAD	1	8.0504167	8.0504167	0.57	0.4608
CAT*LOAD	3	21.9679167	7.3226389	0.52	0.6748

(continued)

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CAT	3	282.0545833	94.0181944	6.67	0.0039
LOAD	1	8.0504167	8.0504167	0.57	0.4608
CAT*LOAD	3	21.9679167	7.3226389	0.52	0.6748

Level of CAT	N	OPI	
		Mean	Std Dev
1	6	89.8000000	2.06009709
2	6	93.6166667	3.06425630
3	6	92.3833333	3.83218823
4	6	84.6833333	4.77301442

Level of LOAD	N	OPI	
		Mean	Std Dev
L	12	89.5416667	4.82972959
N	12	90.7000000	4.98105502

Level of CAT	Level of LOAD	N	OPI	
			Mean	Std Dev
1	L	3	90.8666667	2.48461935
1	N	3	88.7333333	1.01159939
2	L	3	92.4333333	1.96044213
2	N	3	94.8000000	3.92810387
3	L	3	91.4333333	1.09696551
3	N	3	93.3333333	5.72741943
4	L	3	83.4333333	6.36893502
4	N	3	85.9333333	3.42101350

- c. Analyze the data to determine whether there are significant main effects due to fund category and load status and whether these factors significantly interact.
17. This question refers back to the *U.S. News & World Report* graduate school data presented in Problem 22 in Chapter 17.
- Suppose that a two-way ANOVA is to be performed, with 1995 starting salary as the dependent variable and school type (SCHOOL) and reputation rank (REP = 1 if in top 25; 2 if not) as the two factors. State precisely the ANOVA model. Are the factors fixed or random?
 - In the SAS output that follows, complete the ANOVA table.
 - Analyze the data to determine whether there are significant main effects due to school type and reputation rank and whether these factors significantly interact. Use $\alpha = .10$.

Edited SAS Output (PROC GLM) for Problem 17

CLASS LEVEL INFORMATION		
Class	Levels	Values
SCHOOL	2	Bus Law
REP	2	1 2

Dependent Variable: SAL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	—	547.370000	182.456667	_____	_____
Error	20	1419.023333	70.951167		
Corrected Total	23	1966.393333			

R-Square	Coeff Var	Root MSE	SAL Mean
0.278362	14.79494	8.423252	56.93333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
SCHOOL	1	37.0016667	37.0016667	0.52	0.4786
REP	1	440.3266667	440.3266667	6.21	0.0216
SCHOOL*REP	1	70.0416667	70.0416667	0.99	0.3323

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SCHOOL	1	37.0016667	37.0016667	0.52	0.4786
REP	1	440.3266667	440.3266667	6.21	0.0216
SCHOOL*REP	1	70.0416667	70.0416667	0.99	0.3323

Level of SCHOOL	N	SAL	
		Mean	Std Dev
Bus	12	55.6916667	9.06396044
Law	12	58.1750000	9.65628622

Level of REP	N	SAL	
		Mean	Std Dev
1	12	61.2166667	8.62657979
2	12	52.6500000	8.01969167

Level of SCHOOL	Level of REP	N	SAL	
			Mean	Std Dev
Bus	1	6	58.2666667	9.64710665
Bus	2	6	53.1166667	8.47122581
Law	1	6	64.1666667	7.05454936
Law	2	6	52.1833333	8.31923474

References

- Daly, M. B. 1973. "The Effect of Neighborhood Racial Characteristics on the Attitudes, Social Behavior, and Health of Low Income Housing Residents." Ph.D. dissertation, Department of Epidemiology, University of North Carolina, Chapel Hill, N.C.
- Hocking, R. R. 1973. "A Discussion of the Two-Way Mixed Model." *American Statistician* 27(4): 148–52.
- Price, R. D. 1954. "The Survival of *Bacterium Tularensis* in Lice and Louse Feces." *American Journal of Tropical Medicine Hygiene* 3: 179–86.
- Searle, S. R.; Casella, G.; and McCulloch, C. E. 1992. *Variance Components*. New York: John Wiley & Sons, pp. 118–27.

20

Two-way ANOVA with Unequal Cell Numbers

20.1 Preview

When we first began discussing two-way ANOVA in Chapter 18, we saw in Figure 18.1 several ways to classify a two-factor problem according to the observed data pattern. We have already covered methods for handling the single-observation-per-cell case (Chapter 18) and the equal-cell-number case (Chapter 19). In this chapter, we examine procedures for analyzing two-factor patterns that have unequal cell numbers. This situation presents special problems in statistical analysis—both in computation and in interpretation.

In treating these problems, we must make several distinctions among patterns of cell frequency. The first distinction is between *balanced* and *unbalanced* patterns. A balanced pattern has an equal number of observations in each cell, whereas an unbalanced pattern does not. Similarly, a *complete* pattern has at least one observation per cell, whereas an *incomplete* pattern has zero observations in one or more cells. All incomplete patterns are unbalanced. Finally, some unbalanced patterns exhibit *proportional cell frequencies*.

20.2 Presentation of Data for Two-way ANOVA: Unequal Cell Numbers

A general schematic for arranging data for the unequal-cell-number case in two-way ANOVA is presented in Table 20.1. A numerical example is given in Table 20.2, to which we will refer throughout this chapter. In the table, we have

r = Number of rows (i.e., number of levels of the row factor)

c = Number of columns (i.e., number of levels of the column factor)

Y_{ijk} = k th observation in the cell associated with the i th row and j th column

n_{ij} = Number of observations in the cell associated with the i th row and j th column

TABLE 20.1 Data layout for the unequal-cell-number case (two-way ANOVA)

Row Factor	Column Factor				Row Marginals
	1	2	...	c	
1	$Y_{111}, Y_{112}, \dots, Y_{11n_{11}}$ (Sample size = n_{11}) (Cell mean = $\bar{Y}_{11\cdot}$)	$Y_{121}, Y_{122}, \dots, Y_{12n_{12}}$ (Sample size = n_{12}) (Cell mean = $\bar{Y}_{12\cdot}$)	...	$Y_{1c1}, Y_{1c2}, \dots, Y_{1cn_{1c}}$ (Sample size = n_{1c}) (Cell mean = $\bar{Y}_{1c\cdot}$)	$n_{1\cdot}, \bar{Y}_{1\cdot}$
2	$Y_{211}, Y_{212}, \dots, Y_{21n_{21}}$ (Sample size = n_{21}) (Cell mean = $\bar{Y}_{21\cdot}$)	$Y_{221}, Y_{222}, \dots, Y_{22n_{22}}$ (Sample size = n_{22}) (Cell mean = $\bar{Y}_{22\cdot}$)	...	$Y_{2c1}, Y_{2c2}, \dots, Y_{2cn_{2c}}$ (Sample size = n_{2c}) (Cell mean = $\bar{Y}_{2c\cdot}$)	$n_{2\cdot}, \bar{Y}_{2\cdot}$
:	:	:			:
r	$Y_{r11}, Y_{r12}, \dots, Y_{r1n_{r1}}$ (Sample size = n_{r1}) (Cell mean = $\bar{Y}_{r1\cdot}$)	$Y_{r21}, Y_{r22}, \dots, Y_{r2n_{r2}}$ (Sample size = n_{r2}) (Cell mean = $\bar{Y}_{r2\cdot}$)	...	$Y_{rc1}, Y_{rc2}, \dots, Y_{rcn_{rc}}$ (Sample size = n_{rc}) (Cell mean = $\bar{Y}_{rc\cdot}$)	$n_{r\cdot}, \bar{Y}_{r\cdot}$
Column Marginals	$n_{\cdot 1}, \bar{Y}_{\cdot 1}$	$n_{\cdot 2}, \bar{Y}_{\cdot 2}$...	$n_{\cdot c}, \bar{Y}_{\cdot c}$	$n_{\cdot\cdot}, \bar{Y}_{\cdot\cdot}$

© Cengage Learning

TABLE 20.2 Satisfaction with medical care (Y), classified by patient worry and affective communication between patient and physician

Affective Communication	Worry		Row Marginals
	Negative	Positive	
High	2, 5, 8, 6, 2, 4, 3, 10 ($n_{11} = 8$) ($\bar{Y}_{11\cdot} = 5$)	7, 5, 8, 6, 3, 5, 6, 4, 5, 6, 8, 9 ($n_{12} = 12$) ($\bar{Y}_{12\cdot} = 6$)	$n_{1\cdot} = 20$ $\bar{Y}_{1\cdot\cdot} = 5.60$
Medium	4, 6, 3, 3 ($n_{21} = 4$) ($\bar{Y}_{21\cdot} = 4$)	7, 7, 8, 6, 4, 9, 8, 7 ($n_{22} = 8$) ($\bar{Y}_{22\cdot} = 7$)	$n_{2\cdot} = 12$ $\bar{Y}_{2\cdot\cdot} = 6.00$
Low	8, 7, 5, 9, 9, 10, 8, 6, 8, 10 ($n_{31} = 10$) ($\bar{Y}_{31\cdot} = 8$)	5, 8, 6, 6, 9, 7, 7, 8 ($n_{32} = 8$) ($\bar{Y}_{32\cdot} = 7$)	$n_{3\cdot} = 18$ $\bar{Y}_{3\cdot\cdot} = 7.56$
Column Marginals	$n_{\cdot 1} = 22$ $\bar{Y}_{\cdot 1} = 6.18$	$n_{\cdot 2} = 28$ $\bar{Y}_{\cdot 2} = 6.57$	$n_{\cdot\cdot} = 50$ $\bar{Y}_{\cdot\cdot\cdot} = 6.40$

© Cengage Learning

Also,

$$\bar{Y}_{ij\cdot} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk}$$

$$\bar{Y}_{i\cdot\cdot} = \frac{1}{n_{i\cdot\cdot}} \sum_{j=1}^c \sum_{k=1}^{n_{ij}} Y_{ijk} \quad \text{where} \quad n_{i\cdot\cdot} = \sum_{j=1}^c n_{ij}$$

$$\bar{Y}_{j\cdot} = \frac{1}{n_{\cdot j}} \sum_{i=1}^r \sum_{k=1}^{n_{ij}} Y_{ijk} \quad \text{where} \quad n_{\cdot j} = \sum_{i=1}^r n_{ij}$$

$$\bar{Y}_{\dots} = \frac{1}{n_{\dots}} \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n_{ij}} Y_{ijk} \quad \text{where} \quad n_{\dots} = \sum_{i=1}^r \sum_{j=1}^c n_{ij}$$

The unequal-cell-number case arises quite frequently in observational studies. In such studies, one or more of the following statements are typically true:

1. Some variables of interest are not categorized before the data are collected.
2. New variables are often considered after the data are collected.
3. When all the variables are separately categorized, it is often impractical or even impossible to control in advance how the various categories will combine to form strata of interest.

The unequal-cell-number case can also arise in experimental studies when a posteriori consideration is given to variables other than those of primary interest, even if the design based on the primary variables calls for equal cell numbers. Furthermore, unequal cell numbers generally result whenever data points are missing, which may occur (for example) because of study dropouts or incomplete records.

The example presented in Table 20.2 is derived from a study by Thompson (1972) of the relationship of two factors—patient perception of pregnancy and physician–patient communication—to patient satisfaction with medical care. Two main variables of interest were the patient’s degree of worry (WORRY) and a measure of affective communication (AFFCOM). These variables were developed from scales based on questionnaires administered to patients and their physicians. Based on the distribution of scores, the WORRY variable was grouped into the categories “positive” and “negative,” and the AFFCOM variable was grouped into the categories “high,” “medium,” and “low.” Table 20.2 presents data of this type, showing scores for satisfaction with medical care ($Y = \text{TOTSAT}$), classified according to these six combinations of levels of the factors WORRY and AFFCOM.

As the table indicates, the categorization scheme used leads to a two-way table with unequal cell numbers. For WORRY, there are 22 negative and 28 positive values; for AFFCOM, there are 20 high, 12 medium, and 18 low scores. When the separate categories for the two variables are considered together, the resulting six categories have different cell sample sizes, ranging from 4 (for medium AFFCOM, negative WORRY) to 12 (for high AFFCOM, positive WORRY).

20.3 Problem with Unequal Cell Numbers: Nonorthogonality

The key statistical concept associated with the special analytical problems encountered in the unequal-cell-number case pertains to the *nonorthogonality* of the sums of squares usually used to describe the sources of variation in a two-way ANOVA table. To clarify what *orthogonality*

means, we first state the general formulas for these sums of squares (given in Section 19.3 for the equal-cell-number case) in terms of unequal cell numbers:

$$\begin{aligned}
 \text{SSR} &= \sum_{i=1}^r n_i (\bar{Y}_{i..} - \bar{Y}...)^2 & \text{SSC} &= \sum_{j=1}^c n_j (\bar{Y}_{.j.} - \bar{Y}...)^2 \\
 \text{SSRC} &= \sum_{i=1}^r \sum_{j=1}^c n_{ij} (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}...)^2 \\
 \text{SSE} &= \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{ij.})^2, & \text{SSY} &= \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}...)^2
 \end{aligned} \tag{20.1}$$

These formulas for SSR, SSC, and SSRC are often referred to as the *unconditional* sums of squares for rows, columns, and interaction, respectively; here *unconditional* means that each sum of squares may be separately defined from basic principles to describe the variability associated with the estimated effects $(\bar{Y}_{i..} - \bar{Y}...)$, $(\bar{Y}_{.j.} - \bar{Y}...)$, and $(\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}...)$ for rows, columns, and interaction, respectively. (Later we will explore an equivalent way to illustrate the meaning of the term *unconditional* using regression analysis methodology.)

If the collection of sums of squares in (20.1) is orthogonal, the following fundamental equation holds:

$$\text{SSR} + \text{SSC} + \text{SSRC} + \text{SSE} = \text{SSY}$$

That is, the terms on the left-hand side must partition the total sum of squares into nonoverlapping sources of variation.

We have already seen that this fundamental equation holds true for the equal-cell-number case (Chapter 19). Unfortunately, when unequal cell numbers exist, the unconditional sums of squares no longer represent completely separate (i.e., orthogonal) sources of variation; thus,

$$\text{Unequal cell numbers} \Rightarrow \text{SSR} + \text{SSC} + \text{SSRC} + \text{SSE} \neq \text{SSY}$$

To see why this is so, consider the general regression formulation for two-way ANOVA, which accommodates the unequal-cell-number case, as well as the equal-cell-number case (also provided as equation (19.3)):

$$Y = \mu + \sum_{i=1}^{r-1} \alpha_i X_i + \sum_{j=1}^{c-1} \beta_j Z_j + \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} \gamma_{ij} X_i Z_j + E \tag{20.2}$$

where μ , α_i , β_j , and γ_{ij} are regression coefficients and X_i and Z_j are appropriately defined dummy variables. The general form of the fundamental regression equation for this model may be written

$$\text{SS(Total)} = \text{SS(Regression)} + \text{SS(Error)}$$

written here as

$$\text{SSY} = \text{SSReg} + \text{SSE}$$

where

$$\text{SSY} = \sum (Y_i - \bar{Y})^2$$

$$\text{SSReg} = \sum (\hat{Y}_i - \bar{Y})^2$$

= Regression SS($X_1, X_2, \dots, X_{r-1}; Z_1, Z_2, \dots, Z_{c-1}; X_1Z_1, X_1Z_2, \dots, X_{r-1}Z_{c-1}$)

$$\text{SSE} = \sum (Y_i - \hat{Y}_i)^2$$

and where the summation is over all $n..$ observations.

Now, using the *extra-sum-of-squares principle* (see Chapter 9), we can partition the regression sum of squares in various ways to emphasize the contribution due to adding sets of variables to a regression model that already contains other sets of variables. In particular, we can partition the fundamental regression equation as follows with regard to model (20.2):

$$\begin{aligned} \text{SSY} &= \text{Regression SS}(\overbrace{X_1, X_2, \dots, X_{r-1}}^R) \\ &\quad + \text{Regression SS}(\overbrace{Z_1, Z_2, \dots, Z_{c-1}}^C | \overbrace{X_1, X_2, \dots, X_{r-1}}^R) \\ &\quad + \text{Regression SS}(\overbrace{X_1Z_1, X_1Z_2, \dots, X_{r-1}Z_{c-1}}^{RC} | \overbrace{X_1, X_2, \dots, X_{r-1}, Z_1, Z_2, \dots, Z_{c-1}}^{R, C}) \\ &\quad + \text{SSE} \end{aligned} \tag{20.3}$$

On the other hand, if we wish to consider the contribution of the column factor first in the model, the appropriate decomposition of SSY becomes

$$\begin{aligned} \text{SSY} &= \text{Regression SS}(\overbrace{Z_1, Z_2, \dots, Z_{c-1}}^C) \\ &\quad + \text{Regression SS}(\overbrace{X_1, X_2, \dots, X_{r-1}}^R | \overbrace{Z_1, Z_2, \dots, Z_{c-1}}^C) \\ &\quad + \text{Regression SS}(\overbrace{X_1Z_1, X_1Z_2, \dots, X_{r-1}Z_{c-1}}^{RC} | \overbrace{X_1, X_2, \dots, X_{r-1}, Z_1, Z_2, \dots, Z_{c-1}}^{R, C}) \\ &\quad + \text{SSE} \end{aligned} \tag{20.4}$$

As suggested by (20.3) and (20.4), it can be shown that

$$\begin{aligned}\text{Regression SS}(X_1, X_2, \dots, X_{r-1}) &\equiv \text{SSR} \\ \text{Regression SS}(Z_1, Z_2, \dots, Z_{c-1}) &\equiv \text{SSC} \\ \text{Regression SS}(X_1Z_1, X_1Z_2, \dots, X_{r-1}Z_{c-1}) &\equiv \text{SSRC}\end{aligned}\tag{20.5}$$

where SSR, SSC, and SSRC are the unconditional sums of squares given by (20.1). For example, we can express (20.3) and (20.4) as

$$\text{SSY} = \text{SSR} + \text{SS}(C|R) + \text{SS}(RC|R, C) + \text{SSE}$$

and

$$\text{SSY} = \text{SSC} + \text{SS}(R|C) + \text{SS}(RC|R, C) + \text{SSE}$$

respectively. Both of these equations involve conditional sums of squares and are always true, regardless of whether the cell sample sizes are equal or not.

When all the cell sample sizes are equal, however, it is also true that

$$\text{Equal cell numbers} \Rightarrow \begin{cases} \text{SSR} = \text{SS}(R|C) \\ \text{SSC} = \text{SS}(C|R) \\ \text{SSRC} = \text{SS}(RC|R, C) \end{cases}$$

Consequently, when all the cell sample sizes are equal, the extra sums of squares are not affected by variables already in the model, and the following relationship holds:

$$\text{Equal cell numbers} \Rightarrow \text{SSR} + \text{SSC} + \text{SSRC} + \text{SSE} = \text{SSY}\tag{20.6}$$

In the unequal-cell-number case, we have

$$\text{Unequal cell numbers} \Rightarrow \begin{cases} \text{SSR} \neq \text{SS}(R|C) \\ \text{SSC} \neq \text{SS}(C|R) \\ \text{SSRC} \neq \text{SS}(RC|R, C) \end{cases}$$

Therefore, in the unequal-cell-number case, (20.6) does not hold, and we must consider such expressions as (20.3) and (20.4), which reflect the importance of the order in which the effects are entered into the model. As discussed further in Section 20.4, the unequal-cell-number case is best handled by using regression analysis to carry out the two-way ANOVA calculations.

One exception to this nonorthogonality occurs, when the cell frequencies satisfy the proportional relationship

$$n_{ij} = \frac{n_i \cdot n_j}{n_{..}}\tag{20.7}$$

As an example of proportional allocation, consider a clinical trial where each of $n_{..} = 100$ patients is to receive two of four drugs (A, B, C, D) for hypertension. Suppose $n_{1.} = 60$ patients receive drug A, $n_{2.} = 40$ receive drug B, $n_{11} = 60$ receive drug C, and $n_{22} = 40$ receive drug D. Then proportional allocation would result in $n_{11} = 36$ patients receiving drugs A and C, $n_{12} = 24$ receiving A and D, $n_{21} = 24$ receiving B and C, and $n_{22} = 16$ receiving B and D. When (20.7) holds, the following statement can be made:

$$n_{ij} = \frac{n_i \cdot n_j}{n_{..}} \Rightarrow \begin{cases} \text{SSR} = \text{SS}(R|C) \\ \text{SSC} = \text{SS}(C|R) \\ \text{SSRC} \neq \text{SS}(RC|R, C) \end{cases}$$

Thus, although (20.6) still does not hold in this case, (20.3) and (20.4) simplify to the single equation

$$\text{SSR} + \text{SSC} + \text{SS}(RC|R, C) + \text{SSE} = \text{SSY} \quad (20.8)$$

Hence, (20.8) contains only one term, $\text{SS}(RC|R, C)$, that differs from the terms in (20.6). This sum of squares, however, can easily be obtained by subtraction, as is clear from (20.8). Thus, *when the proportional cell frequency allocation of (20.7) is used, the standard equal-cell-number ANOVA calculations can be performed*, without any need to resort to regression analysis methods.

To summarize, we have the flow diagram for two-way ANOVA shown in Figure 20.1.

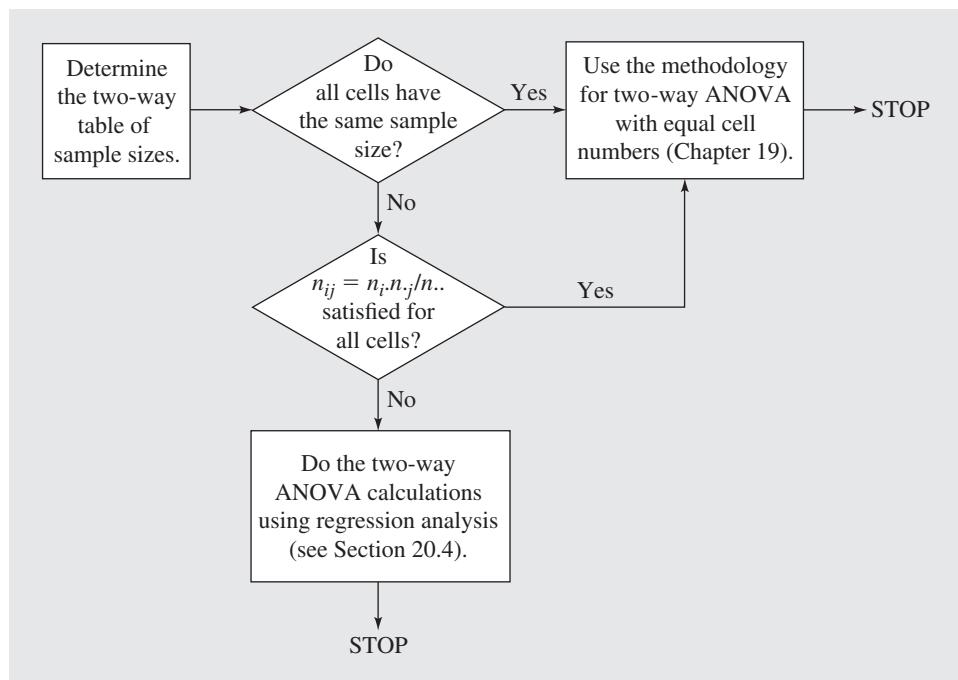


FIGURE 20.1 Flow diagram for two-way ANOVA

20.4 Regression Approach for Unequal Cell Sample Sizes

As described in Section 20.3, the general regression model applicable to both the equal- and unequal-cell-number cases is

$$Y = \mu + \sum_{i=1}^{r-1} \alpha_i X_i + \sum_{j=1}^{c-1} \beta_j Z_j + \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} \gamma_{ij} X_i Z_j + E \quad (20.9)$$

where the $\{X_i\}$ and $\{Z_j\}$ are sets of dummy variables representing the r levels of the row factor and the c levels of the column factor, respectively. In general, *any* two-way ANOVA problem can be analyzed via a regression approach utilizing such a model. When there are unequal cell numbers, however, the order in which effects (row, column, or interaction) are tested becomes important, and a careless decision can yield inappropriate conclusions. The procedure we recommend is a backward-type algorithm in which interaction is considered before main effects (see Appelbaum and Cramer 1974). This algorithm involves the following steps:

Step 1. After fitting the full model (20.9), perform a chunk test for interaction (i.e., test $H_0: \gamma_{ij} = 0$ for all i and j). The (multiple partial) F statistic is given by

$$F(X_1 Z_1, X_1 Z_2, \dots, X_{r-1} Z_{c-1} | X_1, X_2, \dots, X_{r-1}, Z_1, Z_2, \dots, Z_{c-1})$$

and its numerator and denominator degrees of freedom are $(r - 1)(c - 1)$ and $(n.. - rc)$, respectively.

Step 2.

- a. If the Step 1 test is significant, two primary options are available:
 - i. Do no further testing and use the above full model as the final model.
 - ii. Do individual testing to eliminate any nonsignificant interaction terms. The final model will then contain all main effects and all significant product terms.¹
- b. If the Step 1 test is not significant, reduce the model by eliminating all interaction terms. This reduced model is of the form

$$Y = \mu + \sum_{i=1}^{r-1} \alpha_i X_i + \sum_{j=1}^{c-1} \beta_j Z_j + E \quad (20.10)$$

Then conduct two main-effect chunk tests of $H_0: \alpha_i = 0$ for all i and of $H_0: \beta_j = 0$ for all j , using the F statistics $F(X_1, X_2, \dots, X_{r-1} | Z_1, Z_2, \dots, Z_{c-1})$ and

¹ Another possible strategy within option ii is to allow for the possible removal of main-effect terms that are not components of significant interaction terms.

$F(Z_1, Z_2, \dots, Z_{c-1}|X_1, X_2, \dots, X_{r-1})$, respectively.² These tests consider the significance of the row effects, given the column effects, and the significance of the column effects, given the row effects, respectively.

Step 3.

- a. If Step 2(b) yields nonsignificant results for both tests, reduce the model further by eliminating the chunk of variables (the set of row or column main effects) corresponding to the least significant chunk test (the test having the larger P -value). Thus, if the test of $H_0: \alpha_i = 0$ for all i (the test for “rows” given “columns”) has the larger P -value, the new reduced model is

$$Y = \mu + \sum_{j=1}^{c-1} \beta_j Z_j + E \quad (20.11)$$

Alternatively, if the test of $H_0: \beta_j = 0$ for all j (the test for “columns” given “rows”) has the larger P -value, the new reduced model is

$$Y = \mu + \sum_{i=1}^{r-1} \alpha_i X_i + E \quad (20.12)$$

After reducing the model, conduct a chunk test for the main effects in this new reduced model, using either $F(X_1, X_2, \dots, X_{r-1})$ or $F(Z_1, Z_2, \dots, Z_{c-1})$, depending on which set of main effects remains.³ If this final test is nonsignificant, the overall conclusion is that the row, column, and interaction effects are all unimportant. If the test is significant, the final model contains only the significant (row or column) main effects, and the conclusion is that only these effects are important.

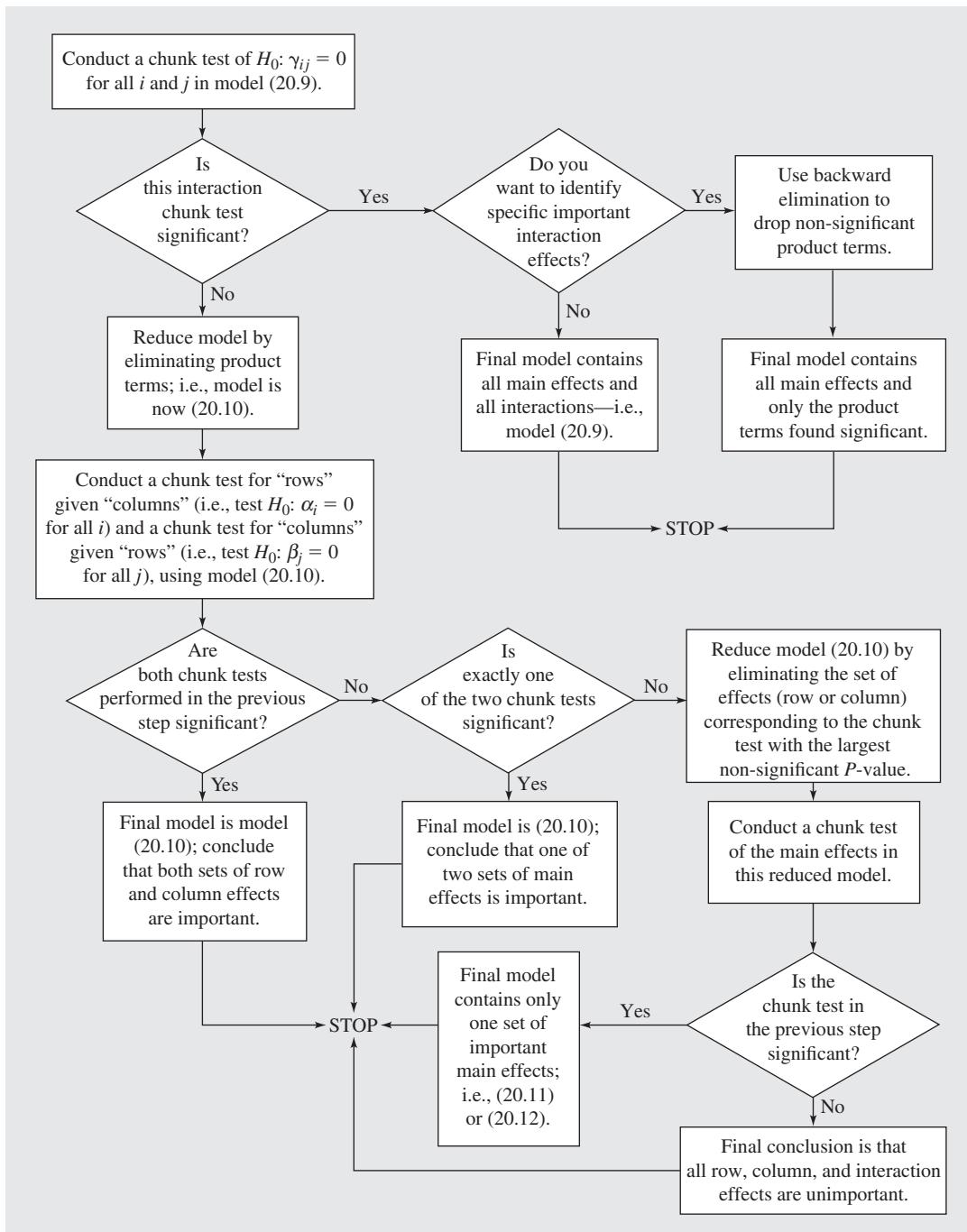
- b. If Step 2(b) yields significant results for both tests, the reduced model (20.10) is the final model, and the conclusion is that both row and column effects are important but that there are no important interaction terms.
- c. If Step 2(b) produces exactly one significant test, there is no need to reduce the model further; the conclusion is that one of the two sets of main effects is important and that there is no significant interaction.⁴

Figure 20.2 provides a flow diagram for the preceding strategy.

² As mentioned elsewhere (e.g., Chapter 9), some statisticians prefer to use the mean-square residual for the full model (20.9) in these main-effect F tests rather than using the mean-square residual for the reduced model (20.10). The default partial F -test results presented by most computer packages use the mean-square residual for the full model.

³ As in Step 2, some statisticians prefer to use the mean-square residual for the full model, rather than the mean-square residual for the reduced model, for the denominator of these tests.

⁴ An alternative here is to reduce the model further by eliminating the nonsignificant set of main-effect variables. However, the unconditional test for the remaining set of main-effect variables may then be nonsignificant, even though the corresponding conditional test under model (20.10) is significant. In this situation, we believe that the conclusions based on model (20.10) are more appropriate.

**FIGURE 20.2** Flow diagram for regression analysis of unbalanced two-way ANOVA data

20.4.1 Example of Regression Approach to Analyzing Unbalanced Two-way ANOVA Data

For the data on satisfaction with medical care from Table 20.2, two regression-model-based ANOVA tables (Tables 20.3 and 20.4) can be produced, depending on whether rows precede columns or columns precede rows into the model. If we follow the strategy for regression analysis previously outlined, our first step is to conduct a chunk test for interaction. The regression model we are using for this test is

$$Y = \mu + \alpha_1 X_1 + \alpha_2 X_2 + \beta_1 Z_1 + \gamma_{11} X_1 Z_1 + \gamma_{21} X_2 Z_1 + E$$

where X_1 and X_2 are dummy variables for the row effects corresponding to the variable AFFCOM and where Z_1 is a dummy variable for the column effects corresponding to the variable WORRY. The null hypothesis of no interaction is $H_0: \gamma_{11} = \gamma_{21} = 0$. As expected, Tables 20.3 and 20.4 provide the same numerical results for this conditional test. The P -value for this interaction test indicates significance at the .05 level but not at the .01 level. Under our strategy, if the investigator is using a .05 significance level, the analysis would stop at this point; we would conclude that significant interaction exists and that main-effect interpretations are not relevant. If the significance level is .01, however,

TABLE 20.3 ANOVA table when row effects enter before column effects for regression analysis of data in Table 20.2

Source	d.f.	SS	MS	F	P
X_1, X_2	2	38.756	19.378	4.97 _{2, 47}	.01 < P < .025
$Z_1 X_1, X_2$	1	5.861	5.861	1.52 _{1, 46}	.10 < P < .25
$X_1 Z_1, X_2 Z_1 X_1, X_2, Z_1$	2	27.383	13.692	4.02 _{2, 44}	.01 < P < .025
Residual	44	150.000	3.409		

© Cengage Learning

TABLE 20.4 ANOVA table when column effects enter before row effects for regression analysis of data in Table 20.2

Source	d.f.	SS	MS	F	P
Z_1	1	1.870	1.870	0.41 _{1, 48}	P > .25
$X_1, X_2 Z_1$	2	42.747	21.373	5.54 _{2, 46}	.005 < P < .01
$X_1 Z_1, X_2 Z_1 X_1, X_2, Z_1$	2	27.383	13.692	4.02 _{2, 44}	.01 < P < .025
Residual	44	150.000	3.409		

© Cengage Learning

we would conclude that no significant interaction effects exist, and then we would consider the reduced model

$$Y = \mu + \alpha_1 X_1 + \alpha_2 X_2 + \beta_1 Z_1 + E$$

The conditional (multiple partial F) test for “rows” given “columns” (i.e., $H_0: \alpha_1 = \alpha_2 = 0$) based on fitting the preceding model leads to an F ratio of 5.54, as given in the “ $X_1, X_2|Z_1$ ” row of Table 20.4. The corresponding P -value indicates significance at the .01 level. The conditional test for “columns” given “rows” (i.e., $H_0: \beta_1 = 0$) in the reduced model results in an F -value of 1.52 (see the “ $Z_1|X_1, X_2$ ” row of Table 20.3). The corresponding P -value ($>.10$) clearly indicates nonsignificance. Thus, based on our strategy, we would conclude (using $\alpha = .01$) that a significant AFFCOM main effect and a nonsignificant WORRY main effect are present.

If the model is further reduced to contain only the AFFCOM main effects (X_1 and X_2), the corresponding unconditional main-effect F test, as found in the “ X_1, X_2 ” row of Table 20.3, is no longer significant at the .01 level. Nevertheless, since the model containing both sets of main effects indicates the existence of a significant AFFCOM main effect, we would argue that AFFCOM is an important variable because this model is taking the WORRY variable into account (i.e., the AFFCOM main-effect F test is conditional on WORRY in Table 20.4).

20.4.2 Using Computer Programs

Nearly all complex data analysis is now conducted with the aid of computer programs. Statistical packages usually contain ANOVA programs or ANOVA options within regression programs. With unbalanced data, and especially with incomplete data, the user of such packages must make an extra effort to understand the dummy variable coding schemes being used and the partial F tests being computed (e.g., which sets of parameters are being considered in which model). Without this understanding, the user may inadvertently conduct tests quite different from those desired.

20.5 Higher-way ANOVA

As a special case of regression analysis, ANOVA can be generalized to any number of factors (i.e., independent variables). Nevertheless, extreme emphasis on complex ANOVA models and the associated testing procedures is likely to be unwarranted, especially for researchers in the health, medical, social, and behavioral sciences. Two reasons underlie this contention:

1. As more independent variables are considered, the researcher’s wish to treat them all as nominal variables becomes less likely.
2. Even if all independent variables can be treated as nominal, sufficient numbers of observations may not be available in all cells (e.g., there may even be several empty cells) to guarantee valid and precise statistical analyses.

TABLE 20.5 General three-way ANOVA table for equal cell sample sizes (all factors fixed)

Source	d.f.	MS	F
A (a levels)	$a - 1$	MS	MSA/MSE
B (b levels)	$b - 1$	MSA	MSB/MSE
C (c levels)	$c - 1$	MSB	MSC/MSE
AB	$(a - 1)(b - 1)$	MSC	MSAB/MSE
AC	$(a - 1)(c - 1)$	MSAB	MSAC/MSE
BC	$(b - 1)(c - 1)$	MSAC	MSBC/MSE
ABC	$(a - 1)(b - 1)(c - 1)$	MSBC	MSABC/MSE
Error	$n - abc$	MSABC	
Total	$n - 1$	MSE	

Note: See Ostle (1963) for information about F statistics when one or more factors are assumed to be random.

© Cengage Learning

Methods are available, however, for designing and analyzing experimental studies in which only a *fraction* of the total number of possible cells need be used; these methods permit the researcher to estimate the effects of primary interest. For instance, if there are some empty cells, it may be possible to estimate specific interactions of interest despite not being able to estimate all possible interactions. Refer to texts by Ostle (1963), Snedecor and Cochran (1967), and Peng (1967) for applications of such methods.

In general, however, regression analysis should predominate in higher-way ANOVA situations, especially since so much research in the health, medical, social, and behavioral sciences is observational in nature, thus typically leading to data sets with highly-varying cell-specific sample sizes. Consequently, we will not extend our discussion to three-way or higher ANOVA situations. For reference purposes, however, we present Table 20.5, the general three-way ANOVA table for the equal-cell-number situation when all factors are assumed to be fixed. An exercise at the end of the chapter deals with the three-way case.

Problems

1. Consider hypothetical data based on a study concerning the effects of rapid cultural change on blood pressure levels for native citizens of an island in Micronesia. Blood pressures were taken on a random sample of 30 males over age 40 from a certain province. These persons, who commuted to work in the nearby westernized capital city, were also given a sociological questionnaire from which their social rankings in both their traditional and their modern (i.e., westernized) cultures were determined. The results are summarized in the following table.

Modern Rank (Factor A)	Traditional Rank (Factor B)		
	HI	MED	LO
HI	130, 140, 135	150, 145	175, 160, 170, 165, 155
MED	145, 140, 150	150, 160, 155	165, 155, 165, 170, 160
LO	180, 160, 145	155, 140, 135	125, 130, 110

- a. Discuss the table of sample means for this data set.
- b. Give an appropriate regression model for this data set, treating the two factors as nominal variables.
- c. Using the regression ANOVA tables that follow (where X pertains to factor A and Z pertains to factor B), carry out two different main-effect tests for each factor, and also test for interaction. Compare the results of the two main-effect tests for each factor.

Source	d.f.	MS
X_1	1	469.17985
$X_2 X_1$	1	508.52217
$Z_1 X_1, X_2$	1	187.97673
$Z_2 X_1, X_2, Z_1$	1	7.54570
$X_1Z_1 X_1, X_2, Z_1, Z_2$	1	3,925.29395
$X_1Z_2 X_1, X_2, Z_1, Z_2, X_1Z_1$	1	9.70621
$X_2Z_1 X_1, X_2, Z_1, Z_2, X_1Z_1, X_1Z_2$	1	633.17613
$X_2Z_2 X_1, X_2, Z_1, Z_2, X_1Z_1, X_1Z_2, X_2Z_1$	1	2.67593
Residual	21	75.83333

Source	d.f.	MS
Z_1	1	278.59213
$Z_2 Z_1$	1	22.71129
$X_1 Z_1, Z_2$	1	391.77041
$X_2 Z_1, Z_2, X_1$	1	480.15062
$X_1Z_1 Z_1, Z_2, X_1, X_2$	1	3,925.29395
$X_1Z_2 Z_1, Z_2, X_1, X_2, X_1Z_1$	1	9.70621
$X_2Z_1 Z_1, Z_2, X_1, X_2, X_1Z_1, X_1Z_2$	1	633.17613
$X_2Z_2 Z_1, Z_2, X_1, X_2, X_1Z_1, X_1Z_2, X_2Z_1$	1	2.67593
Residual	21	75.83333

- d. How might you modify the regression model given in part (b) so that any trends in blood pressure levels could be quantified in terms of increasing social rankings for the two factors? (This requires assigning numerical values to the categories of each factor.) What difficulty do you encounter in defining such a model?

2. A study was conducted to assess the combined effects of patient attitude and patient–physician communication on patient satisfaction with medical care during pregnancy. A random sample of 110 pregnant women under the care of private physicians was followed from the first visit with the physician until delivery. On the basis of specially devised questionnaires, the following variables were measured for each patient: Y = Satisfaction score; X_1 = Attitude score; and X_2 = Communication score. Each score was developed as an interval variable, but some question remains as to whether the analysis should treat the attitude and/or communication scores as nominal variables.
- What would be an appropriate regression model for describing the joint effect of X_1 and X_2 on Y if an interaction between communication and attitude is possible and if all variables are treated as interval variables?
 - What would be an appropriate regression model (using dummy variables) if the analyst wished to allow for an interaction effect but desired only to compare high values versus low values (i.e., to make group comparisons) for both the communication and attitude variables? What kind of ANOVA model would this regression model correspond to?
 - When would the model in part (a) be preferable to that in part (b), and vice versa?
 - If both independent variables are treated nominally, as in part (b), would you expect the associated 2×2 table to have equal numbers in each of the four cells?
3. The data listed in the table on the following page are from an unpublished study by Harbin and others (1985). Subjects were young (ages 18 through 29) or old (ages 60 through 86) men. Each subject was exposed to 0 or 100 parts per million (ppm) of carbon monoxide (CO) for a period before and during testing. Median reaction times for 30 trials are reported for two different tasks: (1) simple reaction time (no choice), REACTIM1; (2) two-choice reaction time, REACTIM2. Pilot data analysis from an earlier study established that this dependent variable followed an appropriate Gaussian distribution. For this problem, consider only REACTIM1 as a dependent variable. Use $\alpha = .01$.
- Observe the cross-tabulation of AGEGROUP and PPM_CO in the accompanying SAS PROC FREQ computer output. Why might this be called a nearly orthogonal two-way design? Any missing data were due to technical problems unrelated to the response variable or treatments.
 - Using dummy variables coded -1 and 1 for AGEGROUP = young and old, respectively, and coded -1 and 1 for PPM_CO = 0 and 100 , respectively, define dummy variables and a corresponding multiple regression model for a two-way ANOVA.
 - Use an appropriate computer program to fit the regression model in part (b). Report appropriate tests of the AGEGROUP \times PPM_CO interaction and tests of their main effects in an appropriate summary source table.
 - SAS's PROC GLM procedure automatically codes dummy variables (and properly treats unequal sample sizes). The accompanying PROC GLM computer output shows the two-way ANOVA results for this problem. Complete the F statistics in the ANOVA table, and report the results of the same tests as in part (c). Compare the results to those in part (c), and explain any differences.

Observation	AGEGROUP	PPM_CO	REACTIM1	REACTIM2
1	Young	0	291.5	632.0
2	Young	0	471.0	607.5
3	Young	0	692.0	859.0
4	Young	0	376.0	484.0
5	Young	0	372.5	501.0
6	Young	0	307.0	381.0
7	Young	0	501.0	559.0
8	Young	0	466.0	632.0
9	Young	0	375.0	434.0
10	Young	0	425.0	454.0
11	Young	0	343.0	542.0
12	Young	0	348.0	471.0
13	Young	0	503.0	521.0
14	Young	0	382.5	519.0
15	Young	100	472.5	515.0
16	Young	100	354.0	521.0
17	Young	100	350.0	456.0
18	Young	100	486.0	522.0
19	Young	100	402.0	472.0
20	Young	100	347.0	414.0
21	Young	100	320.0	363.0
22	Young	100	446.0	591.0
23	Young	100	410.0	539.5
24	Young	100	302.0	472.5
25	Young	100	692.5	656.0
26	Young	100	447.5	548.0
27	Young	100	525.5	527.0
28	Young	100	322.5	574.0
29	Young	100	468.5	559.5
30	Young	100	378.0	499.5
31	Young	100	497.5	529.5
32	Old	0	542.0	595.0
33	Old	0	599.0	606.0
34	Old	0	562.0	598.0
35	Old	0	586.0	744.0
36	Old	0	674.0	724.0
37	Old	0	762.0	836.5
38	Old	0	697.0	834.0
39	Old	0	583.0	698.5
40	Old	0	533.5	668.0
41	Old	0	524.5	670.0
42	Old	0	500.0	587.0
43	Old	0	680.0	912.5
44	Old	0	563.5	619.0
45	Old	100	523.5	646.5
46	Old	100	770.0	862.5
47	Old	100	712.0	829.0
48	Old	100	653.0	697.0
49	Old	100	699.5	818.0
50	Old	100	561.0	819.5
51	Old	100	751.0	872.0
52	Old	100	520.5	889.0
53	Old	100	523.0	601.0

Edited SAS Output (PROC FREQ and PROC GLM) for Problem 3

TABLE OF AGEGROUP BY PPM_CO			
AGEGROUP	PPM_CO		
Frequency			
Percent	0	100	Total
Row Pct			
Col Pct	0	100	Total
Old	13 24.53 59.09 48.15	9 16.98 40.91 34.62	22 41.51
Young	14 26.42 45.16 51.85	17 32.08 54.84 65.38	31 58.49
Total	27 50.94	26 49.06	53 100.00

CLASS LEVEL INFORMATION		
Class	Levels	Values
AGEGROUP	2	Old Young
PPM_CO	2	0 100

Dependent Variable: REACTIM1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	484794.0983	161598.0328	17.51	<.0001
Error	49	452176.6187	9228.0943		
Corrected Total	52	936970.7170			

R-Square	Coeff Var	Root MSE	reactim1 Mean
0.517406	19.14396	96.06297	501.7925

SS(AGEGROUP)					
Source	DF	Type I SS	Mean Square	F Value	Pr > F
AGEGROUP	1	478181.8431	478181.8431	_____	_____
PPM_CO	1	4210.8915	4210.8915	_____	_____
AGEGROUP*PPM_CO	1	2401.3637	2401.3637	_____	_____

SS(PPM_CO AGEGROUP)					
Source	DF	Type II SS	Mean Square	F Value	Pr > F
AGEGROUP	1	481452.3695	481452.3695	_____	_____
PPM_CO	1	4210.8915	4210.8915	_____	_____
AGEGROUP*PPM_CO	1	2401.3637	2401.3637	_____	_____

SS(AGEGROUP | PPM_CO) SS(AGEGROUP*PPM_CO | PPM_CO, AGEGROUP)

4. a.–d. Repeat Problem 3, using REACTIM2 as the dependent variable and referring to the accompanying computer output.

Edited SAS Output (PROC GLM) for Problem 4

CLASS LEVEL INFORMATION		
Class	Levels	Values
AGEGROUP	2	Old Young
PPM_CO	2	0 100

Dependent Variable: REACTIM2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	584722.361	194907.454	20.04	<.0001
Error	49	476633.960	9727.224		
Corrected Total	52	1061356.321			

R-Square	Coeff Var	Root MSE	REACTIM2 Mean
0.550920	16.09215	98.62669	612.8868

Source	DF	Type I SS	Mean Square	F Value	Pr > F
AGEGROUP	1	543059.0114	543059.0114	_____	_____
PPM_CO	1	3971.1064	3971.1064	_____	_____
AGEGROUP * PPM_CO	1	37692.2429	37692.2429	_____	_____

Source	DF	Type II SS	Mean Square	F Value	Pr > F
AGEGROUP	1	545527.9252	545527.9252	_____	_____
PPM_CO	1	3971.1064	3971.1064	_____	_____
AGEGROUP * PPM_CO	1	37692.2429	37692.2429	_____	_____

5. A crime victimization study was undertaken in a medium-size southern city. The main purpose was to determine the effects of being a crime victim on confidence in law enforcement authority and in the legal system itself. A questionnaire was administered to a stratified random sample of 40 city residents; among the information elicited were data on the number of times victimized, a measure of social class status (SCLS), and a measure of the respondent's confidence in law enforcement and in the legal system. The data are reproduced in the following table.

No. of Times Victimized	Social Class Status		
	LO MED HI		
0	4, 14, 15, 19, 17, 17, 16	7, 10, 12, 15, 16	8, 19, 10, 17
1	2, 7, 18	6, 19, 12, 12	7, 6, 5, 3, 16
2+	7, 8, 2, 11, 12	1, 2, 4	4, 2, 8, 9

- a. Determine the table of sample means, and comment on any patterns noted.
 b. Analyze this data set using the following ANOVA table.

- c. How would you analyze this data set by using the two tables of regression results that follow?
- d. What ANOVA assumption(s) might not hold for these data?

Source	d.f.	SS	MS
VICTIM	2	400.00	200.00
SCLS	2	22.739	11.370
VICTIM \times SCLS	4	109.93	27.483
Error	31	704.08	22.712

Source	d.f.	MS
Z_1	1	44.03235
$Z_2 Z_1$	1	1.03496
$X_1 Z_1, Z_2$	1	395.75734
$X_2 Z_1, Z_2, X_1$	1	0.06778
$X_1Z_1 Z_1, Z_2, X_1, X_2$	1	1.68985
$X_1Z_2 Z_1, Z_2, X_1, X_2, X_1Z_1$	1	3.31635
$X_2Z_1 Z_1, Z_2, X_1, X_2, X_1Z_1, X_1Z_2$	1	0.40190
$X_2Z_2 Z_1, Z_2, X_1, X_2, X_1Z_1, X_1Z_2, X_2Z_1$	1	94.59353
Residual	31	22.71229

Source	d.f.	MS
X_1	1	407.86993
$X_2 X_1$	1	0.52174
$Z_1 X_1, X_2$	1	27.98766
$Z_2 X_1, X_2, Z_1$	1	4.51309
$X_1Z_1 X_1, X_2, Z_1, Z_2$	1	1.68985
$X_1Z_2 X_1, X_2, Z_1, Z_2, X_1Z_1, X_1Z_2$	1	3.31635
$X_2Z_1 X_1, X_2, Z_1, Z_2, X_1Z_1, X_1Z_2$	1	0.40190
$X_2Z_2 X_1, X_2, Z_1, Z_2, X_1Z_1, X_1Z_2, X_2Z_1$	1	94.59353
Residual	31	22.71229

Note: X pertains to number of times victimized; Z pertains to social class status.

- 6. The effect of a new antidepressant drug on reducing the severity of depression was studied in manic-depressive patients at two state mental hospitals. In each hospital all such patients were randomly assigned to either a treatment (new drug) or a control (old drug) group. The results of this experiment are summarized in the following table; a high mean score indicates more of a lowering in depression level than does a low mean score.

Hospital	Group	
	Treatment	Control
A	$n = 25, \bar{Y} = 8.5, S = 1.3$	$n = 31, \bar{Y} = 4.6, S = 1.8$
B	$n = 25, \bar{Y} = 2.3, S = 0.9$	$n = 31, \bar{Y} = -1.7, S = 1.1$

- a. Without performing any statistical tests, interpret the means in the table.
- b. What regression model is appropriate for analyzing the data? For this model, describe how to test whether the effect of the new drug is significantly different from the effect of the old drug.

7. A study was conducted by a television network in a certain state to evaluate the viewing characteristics of adult females. Each individual in a stratified random sample of 480 women was sent a questionnaire; the strata were formed on the basis of the following three factors: season (winter or summer), region (eastern, central, or western), and residence (rural or urban). The averages of the total time reported watching TV (hours per day) are summarized in the accompanying table of sample means and standard deviations.

Residence and Region	Summer			Winter			Marginals	
	n	\bar{Y}	S	n	\bar{Y}	S	n	\bar{Y}
Rural								
East	40	2.75	1.340	40	4.80	0.851	80	3.78
Central	40	2.75	1.380	40	4.85	0.935	80	3.80
West	40	2.65	1.180	40	4.78	0.843	80	3.71
Marginals	120	2.72		120	4.81		240	3.76
Urban								
East	40	3.38	0.958	40	3.65	0.947	80	3.52
Central	40	3.15	1.130	40	4.50	0.743	80	3.83
West	40	3.65	0.779	40	4.05	0.781	80	3.85
Marginals	120	3.39		120	4.07		240	3.73
Marginals	240	3.06		240	4.44		480	3.75

- a. Suppose that the questionnaire contained items concerning additional factors such as occupation (categorized as housewife, blue-collar worker, white-collar worker, or professional), age (categorized as 20 to 34, 35 to 50, and over 50), and number of children (categorized as 0, 1–2, and 3+). What is the likelihood of obtaining equal cell numbers when carrying out an ANOVA to consider these additional variables?
- b. Examine the table of sample means for main effects and interactions. (You may want to form two-factor summary tables for assessing two-factor interactions.)
- c. Using the following table of ANOVA results, carry out appropriate *F* tests, and discuss your results.
- d. State a regression model appropriate for obtaining information equivalent to the ANOVA results presented next.

Source	d.f.	SS	MS
RESID	1	0.1333	0.1333
REGION	2	2.5527	1.2763
SEASON	1	229.63	229.63
RESID × REGION	2	3.3247	1.6623
RESID × SEASON	1	60.492	60.492
REGION × SEASON	2	7.2247	3.6123
RESID × REGION × SEASON	2	6.7460	3.3730
Error	468	478.21	1.0218

8. Suppose that the following data were obtained by an investigator studying the influence of estrogen injections on change in the pulse rate of adolescent chimpanzees:

$$\text{Male} \left\{ \begin{array}{l} \text{Control: } 5.1, -2.3, 4.2, 3.8, 3.2, -1.5, 6.1, -2.5, 1.9, -3.0, -2.8, 1.7 \\ \text{Estrogen: } 15.0, 6.2, 4.1, 2.3, 7.6, 14.8, 12.3, 13.1, 3.4, 8.5, 11.2, 6.9 \end{array} \right.$$

Female $\begin{cases} \text{Control: } -2.3, -5.8, -1.5, 3.8, 5.5, 1.6, -2.4, 1.9 \\ \text{Estrogen: } 7.3, 2.4, 6.5, 8.1, 10.3, 2.2, 12.7, 6.3 \end{cases}$

- What are the factors in this experiment? Should they be designated as fixed or random?
- Demonstrate that the cell frequencies for two-way ANOVA in this problem are proportional; that is, $n_{ij} = n_i n_j / n..$ for each of the four cells.
- Use the following table of sample means

Sex	Control	Estrogen
Male	1.158333	8.783333
Female	0.100000	6.975000

and the general formulas

$$\begin{aligned} \text{SS(rows)} &= \text{SSR} = \sum_{i=1}^r n_i (\bar{Y}_{i..} - \bar{Y}_{...})^2 \\ \text{SS(columns)} &= \sum_{j=1}^c n_j (\bar{Y}_{.j} - \bar{Y}_{...})^2 \\ \text{SS(cells)} &= \sum_{i=1}^r \sum_{j=1}^c n_{ij} (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j} + \bar{Y}_{...})^2 \end{aligned}$$

to analyze the data for this problem, employing the usual methodology for equal-cell-number two-way ANOVA and the fact that SSE = 530.24078 (i.e., compute the sums of squares for rows, columns, and cells directly, and then obtain the sum of squares for interaction by subtraction).

- What regression model is appropriate for analyzing this data set?
- Using the regression analysis results given in the following table, check whether SS(Rows) = Regression SS(SEX) = Regression SS(SEX|TREATMENT) and SS(COLUMNS) = Regression SS(TREATMENT) = Regression SS(TREATMENT|SEX), where SS(Rows) and SS(COLUMNS) are as obtained in part (c). What has been demonstrated here?

Source	d.f.	SS
SEX	1	19.72667
TREATMENT SEX	1	536.55619
SEX \times TREATMENT SEX, TREATMENT	1	1.35000
Residual	36	530.24078

Source	d.f.	SS
TREATMENT	1	536.55619
SEX TREATMENT	1	19.72667
SEX \times TREATMENT SEX, TREATMENT	1	1.35000
Residual	36	530.24078

Note: SEX $\begin{cases} -1 & \text{if male} \\ 1 & \text{if female} \end{cases}$ and TREATMENT = $\begin{cases} -1 & \text{if control} \\ 1 & \text{if estrogen} \end{cases}$

9. The data listed in the following table relate to a study by Reiter and others (1981) concerning the effects of injecting triethyl-tin (TET) into rats once at age 5 days. The animals were injected with 0, 3, or 6 mg per kilogram of body weight. The response was the log of the activity count for 1 hour, recorded at 21 days of age. The rat was left to move about freely in a figure 8 maze. Analysis of other studies with this type of activity count confirms that log counts should yield Gaussian errors if the model is correct.
- a. Tabulate the DOSAGE \times SEX cell sample sizes. Explain why this might be called a nearly orthogonal design.
 - b. Using the accompanying SAS computer output, conduct a two-way ANOVA with SEX and DOSAGE as factors. (First provide the F statistic values and P -values in the accompanying ANOVA table.)
 - c. Using $\alpha = .05$, report your conclusions based on the ANOVA.
 - d. Which, if any, families of means should be followed up with multiple-comparison tests? What type of comparisons would you recommend?

Observation	LOGACT21	DOSAGE	SEX	CAGE
1	2.636	0	Male	5
2	2.736	0	Male	6
3	2.775	0	Male	7
4	2.672	0	Male	9
5	2.653	0	Male	11
6	2.569	0	Male	12
7	2.737	0	Male	15
8	2.588	0	Male	16
9	2.735	0	Male	17
10	2.444	3	Male	3
11	2.744	3	Male	5
12	2.207	3	Male	6
13	2.851	3	Male	7
14	2.533	3	Male	9
15	2.630	3	Male	11
16	2.688	3	Male	12
17	2.665	3	Male	15
18	2.517	3	Male	16
19	2.769	3	Male	17
20	2.694	6	Male	3
21	2.845	6	Male	5
22	2.865	6	Male	6
23	3.001	6	Male	7
24	3.043	6	Male	9
25	3.066	6	Male	11
26	2.747	6	Male	12
27	2.894	6	Male	15
28	1.851	6	Male	16
29	2.489	6	Male	17
30	2.494	0	Female	3
31	2.723	0	Female	5
32	2.841	0	Female	6
33	2.620	0	Female	7
34	2.682	0	Female	9
35	2.644	0	Female	11
36	2.684	0	Female	12
37	2.607	0	Female	15

(continued)

Observation	LOGACT21	DOSAGE	SEX	CAGE
38	2.591	0	Female	16
39	2.737	0	Female	17
40	2.220	3	Female	3
41	2.371	3	Female	5
42	2.679	3	Female	6
43	2.591	3	Female	7
44	2.942	3	Female	9
45	2.473	3	Female	11
46	2.814	3	Female	12
47	2.622	3	Female	15
48	2.730	3	Female	16
49	2.955	3	Female	17
50	2.540	6	Female	3
51	3.113	6	Female	5
52	2.468	6	Female	6
53	2.606	6	Female	7
54	2.764	6	Female	9
55	2.859	6	Female	11
56	2.763	6	Female	12
57	3.000	6	Female	15
58	3.111	6	Female	16
59	2.858	6	Female	17

Edited SAS Output (PROC FREQ and PROC GLM) for Problem 9

TABLE OF DOSAGE BY SEX			
Dosage	Sex		
	Female	Male	Total
0	10 16.95 52.63 33.33	9 15.25 47.37 31.03	19 32.20
3	10 16.95 50.00 33.33	10 16.95 50.00 34.48	20 33.90
6	10 16.95 50.00 33.33	10 16.95 50.00 34.48	20 33.90
Total	30 50.85	29 49.15	59 100.00

CLASS LEVEL INFORMATION		
Class	Levels	Values
Dosage	3	0 3 6
Sex	2	Female Male

(continued)

Dependent Variable: LOGACT21

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	0.28197725	0.05639545	1.16	0.3412
Error	53	2.57750079	0.04863209		
Corrected Total	58	2.85947803			
R-Square		Coeff Var	Root MSE	logact21 Mean	
0.098611		8.196165	0.220527	2.690610	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
Dosage	2	0.25750763	0.12875381	_____	_____
Sex	1	0.01054232	0.01054232	_____	_____
Dosage*sex	2	0.01392730	0.00696365	_____	_____
Source	DF	Type II SS	Mean Square	F Value	Pr > F
Dosage	2	0.25806594	0.12903297	_____	_____
Sex	1	0.01054232	0.01054232	_____	_____
Dosage*sex	2	0.01392730	0.00696365	_____	_____

10. The experimenters described in Problem 9 hoped that home cage would not affect activity level in any systematic fashion. Explore this question by repeating Problem 9, but replacing SEX with CAGE in your analysis.

Edited SAS Output (PROC FREQ and PROC GLM) for Problem 10

TABLE OF DOSAGE BY CAGE											
dosage	Cage										
	3	5	6	7	9	11	12	15	16	17	Total
0	1 1.69 5.26 20.00	2 3.39 10.53 33.33	19 32.20								
3	2 3.39 10.00 40.00	2 3.39 10.00 33.33	20 33.90								
6	2 3.39 10.00 40.00	2 3.39 10.00 33.33	20 33.90								
Total	5 8.47	6 10.17	59 100.00								

(continued)

CLASS LEVEL INFORMATION		
Class	Levels	Values
Dosage	3	0 3 6
Cage	10	3 5 6 7 9 11 12 15 16 17

Dependent Variable: LOGACT21

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	29	1.30579603	0.04502745	0.84	0.6786
Error	29	1.55368200	0.05357524		
Corrected Total	58	2.85947803			

R-Square	Coeff Var	Root MSE	logact21 Mean
0.456655	8.602631	0.231463	2.690610

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Dosage	2	0.25750763	0.12875381	_____	_____
Cage	9	0.47895137	0.05321682	_____	_____
Dosage*cage	18	0.56933703	0.03162984	_____	_____

Source	DF	Type II SS	Mean Square	F Value	Pr > F
Dosage	2	0.26793817	0.13396908	_____	_____
Cage	9	0.47895137	0.05321682	_____	_____
Dosage*cage	18	0.56933703	0.03162984	_____	_____

11. A manufacturer conducted a pricing experiment to explore the effects of price decreases on sales of one of its breakfast cereals. The two largest supermarket chains in a particular area participated in the experiment. Ten stores from each chain were randomly selected, and each store was assigned a price level for the cereal (either the original price or a 10% reduced price). If the competing chain had a store in the same vicinity, the two stores both were assigned the same price level. Some stores failed to complete the experiment due to competition from other supermarkets chains. Sales volumes (in hundreds of units) over the period of the study were noted for each of the remaining 17 stores and are shown in the following table.

Supermarket Chain	Price Level	
	Original Price (Price = 0)	10% Reduced Price (Price = R)
1	14, 14, 15	14, 14, 17, 19, 13
2	9, 7, 12, 8	10, 12, 14, 15, 13

- Using the accompanying SAS computer output, conduct a two-way ANOVA with Chain and Price as factors. (Provide the F statistic values and P -values in the accompanying ANOVA table as part of your analysis.)
- Do the effects of a price decrease on sales volume significantly differ between the different chains? If not, does it appear that a price decrease will significantly increase sales volume? How do the sales volumes compare at the two chains? Report your conclusions using the $\alpha = .05$ level. Also, do the conclusions differ for the different approaches taken in parts (a) and (b)?

Edited SAS Output (PROC GLM) for Problem 11

CLASS LEVEL INFORMATION		
Class	Levels	Values
CHAIN	2	1 2
PRICE	2	O R

Dependent Variable: SALES

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	98.2745098	32.7581699	7.79	0.0031
Error	13	54.6666667	4.2051282		
Corrected Total	16	152.9411765			

R-Square	Coeff Var	Root MSE	SALES Mean
0.642564	15.84586	2.050641	12.94118

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CHAIN	1	64.05228758	64.05228758	_____	_____
PRICE	1	26.62448211	26.62448211	_____	_____
CHAIN*PRICE	1	7.59774011	7.59774011	_____	_____

Source	DF	Type II SS	Mean Square	F Value	Pr > F
CHAIN	1	58.06416465	58.06416465	_____	_____
PRICE	1	26.62448211	26.62448211	_____	_____
CHAIN*PRICE	1	7.59774011	7.59774011	_____	_____

- The manager of a market research company conducted an experiment to investigate the productivity of three employees on each of two computerized data-entry systems. The employees conducted phone surveys, entering the survey data into the computer during the phone call. Productivity was measured as the time (in minutes) taken to complete a call in which the respondent agreed to complete the survey. Each

employee used each system for one hour, and the order of use was randomized. The productivity data are recorded in the following table.

Employee	System	
	1	2
1	8, 7, 8, 9, 8	6, 4, 7, 4, 3, 3, 4, 5, 6
2	5, 4, 6, 3, 8, 8, 7	6, 2, 3, 3, 4, 4, 5, 6, 7, 4
3	3, 4, 5, 4, 3, 5, 5, 6	3, 3, 4, 5, 4, 3, 2, 4, 3, 3, 3, 4

- Using the accompanying SAS computer output, conduct a two-way ANOVA. (Provide the F statistic values and P -values in the accompanying ANOVA table as part of your analysis.)
- Report your conclusions using the $\alpha = .01$ level. Is there an interaction between Employee and System? If not, does it appear that one system is significantly more productive than the other? Do the employees differ in terms of productivity?

Edited SAS Output (PROC GLM) for Problem 12

CLASS LEVEL INFORMATION		
Class	Levels	Values
EMPLOYEE	3	1 2 3
SYSTEM	2	1 2

Dependent Variable: PRODUCT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	85.1276611	17.0255322	9.82	<.0001
Error	45	78.0488095	1.7344180		
Corrected Total	50	163.1764706			

R-Square	Coeff Var	Root MSE	PRODUCT Mean
0.521691	27.64017	1.316973	4.764706

Source	DF	Type I SS	Mean Square	F Value	Pr > F
EMPLOYEE	2	36.26218487	18.13109244	_____	_____
SYSTEM	1	37.44652926	37.44652926	_____	_____
EMPLOYEE*SYSTEM	2	11.41894694	5.70947347	_____	_____

Source	DF	Type II SS	Mean Square	F Value	Pr > F
EMPLOYEE	2	38.44192096	19.22096048	_____	_____
SYSTEM	1	37.44652926	37.44652926	_____	_____
EMPLOYEE*SYSTEM	2	11.41894694	5.70947347	_____	_____

13. This question refers to the radial keratotomy data of Problem 21 in Chapter 17.
- Suppose that a two-way ANOVA is to be performed, with five-year postoperative change in refractive error as the dependent variable (Y) and diameter of clear zone (CLRZONE) and baseline average corneal curvature (BASECURV = 1 if curvature is < 43 diopters; = 2 if curvature is between 43 and 44 diopters; and = 3 if curvature is > 44 diopters) as the two factors. State precisely the ANOVA model. Are the factors fixed or random?
 - In the SAS output that follows, complete the ANOVA table.
 - Analyze the data to determine whether there are significant main effects due to clear zone and baseline curvature and whether these factors significantly interact. Use $\alpha = .10$.

Edited SAS Output (PROC GLM) for Problem 13

CLASS LEVEL INFORMATION		
Class	Levels	Values
CLRZONE	3	3.0 3.5 4.0
BASECURV	3	1 2 3

Dependent Variable: Y

Source	DF	Sum of Squares	F Value	Pr > F
Model	—	32.22787390	—	—
Error	42	47.06818002		
Corrected Total	50	79.29605392		

R-Square	Coeff Var	Y Mean
0.406425	27.77523	3.811373

Source	DF	Type I SS	F Value	Pr > F
CLRZONE	2	14.70441590	6.56	0.0033
BASECURV	2	5.37753783	2.40	0.1031
CLRZONE*BASECURV	4	12.14592017	2.71	0.0428

Source	DF	Type III SS	F Value	Pr > F
CLRZONE	2	12.92686249	5.77	0.0061
BASECURV	2	5.97019625	2.66	0.0814
CLRZONE*BASECURV	4	12.14592017	2.71	0.0428

Level of CLRZONE	N	Y	
		Mean	Std Dev
3.0	20	4.41875000	1.15228249
3.5	15	3.71666667	1.33836476
4.0	16	3.14093750	0.97595119

(continued)

Level of BASECURV	N	Y	
		Mean	Std Dev
1	11	4.30681818	1.20592759
2	15	3.60833333	1.03818466
3	25	3.71520000	1.38615361

Level of CLRZONE	Level of BASECURV	N	Y	
			Mean	Std Dev
3.0	1	3	5.00000000	0.99215674
3.0	2	6	3.66666667	0.79320027
3.0	3	11	4.67045455	1.22509276
3.5	1	4	4.65625000	0.71716310
3.5	2	5	4.22500000	1.26676261
3.5	3	6	2.66666667	1.06555932
4.0	1	4	3.43750000	1.42339090
4.0	2	4	2.75000000	0.46770717
4.0	3	8	3.18812500	0.96893881

14. This question refers to the *U.S. News & World Report* graduate school data presented in Problem 22 in Chapter 17.
- Suppose that a two-way ANOVA is to be performed, with 1995 starting salary as the dependent variable and with school type (SCHOOL) and reputation rank among academics (REP1 = 1 if reputation rank is in the top 10; = 2 if rank is 11 to 20; = 3 if rank is 21 or worse) as the two factors. State precisely the ANOVA model. Are the factors fixed or random?
 - In the SAS output that follows, complete the ANOVA table.
 - Analyze the data to determine whether there are significant main effects due to school type and reputation rank and whether these factors significantly interact.

Edited SAS Output (PROC GLM) for Problem 14

CLASS LEVEL INFORMATION		
Class	Levels	Values
SCHOOL	2	1 2
REP	3	1 2 3

Dependent Variable: SAL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	—	1000.505833	200.101167	—	—
Error	18	965.887500	53.660417		
Corrected Total	23	1966.393333			

(continued)

R-Square	Coeff Var	Root MSE	SAL Mean
0.508802	12.86650	7.325327	56.93333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
SCHOOL	1	37.0016667	37.0016667	0.69	0.4172
REP	2	910.3658333	455.1829167	8.48	0.0025
SCHOOL*REP	2	53.1383333	26.5691667	0.50	0.6176

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SCHOOL	1	67.7877778	67.7877778	1.26	0.2758
REP	2	910.3658333	455.1829167	8.48	0.0025
SCHOOL*REP	2	53.1383333	26.5691667	0.50	0.6176

Level of SCHOOL	N	SAL	
		Mean	Std Dev
1	12	58.1750000	9.65628622
2	12	55.6916667	9.06396044

Level of REP	N	SAL	
		Mean	Std Dev
1	4	68.7500000	6.29152870
2	4	61.5000000	6.35085296
3	16	52.8375000	7.37688959

Level of SCHOOL	Level of REP	N	SAL	
			Mean	Std Dev
1	1	2	70.0000000	0.0000000
1	2	2	66.0000000	5.6568542
1	3	8	53.2625000	7.5450906
2	1	2	67.5000000	10.6066017
2	2	2	57.0000000	2.8284271
2	3	8	52.4125000	7.6986896

References

- Appelbaum, M. I., and Cramer, E. M. 1974. "Some Problems in the Non-Orthogonal Analysis of Variance." *Psychology Bulletin* 81(6): 335–43.
- Harbin, T. J.; Benignus, V. A.; Muller, K. E.; and Barton, C. N. 1985. "Effects of Low-Level Carbon Monoxide Exposure upon Evoked Cortical Potentials in Young and Elderly Men." Manuscript in review, U.S. Environmental Protection Agency, Washington, D.C.
- Ostle, B. 1963. *Statistics in Research*, Second Edition. Ames: Iowa State University Press.
- Peng, K. C. 1967. *Design and Analysis of Scientific Experiments*. Reading, Mass.: Addison-Wesley.

- Reiter, L. W.; Heavner, G. G.; Dean, K. F; and Ruppert, P. H. 1981. "Developmental and Behavioral Effects of Early Postnatal Exposure to Triethyltin in Rats." *Neurobehavioral Toxicology and Teratology* 3: 285–93.
- Searle, S. R. 1971. *Linear Models*. New York: John Wiley & Sons.
- . 1987. *Linear Models for Unbalanced Data*. New York: John Wiley & Sons.
- Snedecor, G. W., and Cochran, W. G. 1967. *Statistical Methods*, Sixth Edition. Ames: Iowa State University Press.
- Thompson, S. J. 1972. "The Doctor–Patient Relationship and Outcomes of Pregnancy." Ph.D. dissertation, Department of Epidemiology, University of North Carolina, Chapel Hill, N.C.

21

The Method of Maximum Likelihood

21.1 Preview

This chapter describes the methodology of maximum likelihood estimation and its associated inference-making procedures. The term *maximum likelihood* (ML) refers to a very general algorithm for obtaining estimators of population parameters. Such estimators have large-sample statistical properties that are very useful for practical applications of regression modeling. Moreover, the ML method is applicable to a wide variety of statistical models. When a multiple linear regression model is fitted using normally distributed and mutually independent response variables, the least-squares estimators of the regression coefficients are identical to the ML estimators. Furthermore, ML estimation is the method of choice for estimating the parameters in nonlinear models such as the logistic regression model (discussed in Chapters 22 and 23) and the Poisson regression model (discussed in Chapter 24).

This chapter begins with a general discussion of the basic theoretical principles underlying the method of maximum likelihood. We then describe and illustrate inference-making procedures based on the use of ML estimators.

21.2 The Principle of Maximum Likelihood

In introducing the underlying principle of maximum likelihood, let us focus on a relatively simple problem of statistical estimation. Suppose that a large population contains a certain unknown proportion θ ($0 \leq \theta \leq 1$) of individuals with a certain genetic disorder. In epidemiology, the parameter θ is referred to as the prevalence of the disorder of interest in the population under study. Further, suppose that a random sample of n individuals is selected from this population and that Y represents the random variable denoting the number of individuals in the random sample of size n who have this genetic disorder. The possible values of Y are the $(n + 1)$ integer values $0, 1, 2, \dots, n$. The statistical estimation problem concerns

how to use n and Y to obtain an estimator of θ , denoted as $\hat{\theta}$, with good statistical properties. By “good” we mean that the chosen estimator (which is itself a random variable, since it is a function of the random variable Y) has little or no bias (i.e., its expected value is essentially equal to θ) and has a small variance.

Suppose, for example, that we sample $n = 100$ individuals and that $Y = 60$ out of these 100 individuals have the genetic disorder. Then, the intuitive choice for the estimate of the parameter θ , the true prevalence in the population, in this case would be $\hat{\theta} = 60/100 = 0.60$, the *sample proportion* of subjects with the genetic disorder of interest. It turns out that this value (0.60) is also the ML estimate of θ . If, on the other hand, we had sampled only $n = 5$ individuals and found that $Y = 4$ had the genetic disorder, then the sample proportion $\hat{\theta} = 4/5 = 0.80$ would be the ML estimate of θ . For a sample with a given n and Y , the sample proportion $\hat{\theta} = Y/n$ is the ML estimate for θ .

We now justify the above result mathematically. Given the above general scenario, it is reasonable to assume that the underlying probability distribution of the discrete random variable Y is binomial; in particular, we have

$$\text{pr}(Y; \theta) = {}_nC_Y \theta^Y (1 - \theta)^{n-Y} \quad Y = 0, 1, 2, \dots, n \quad (21.1)$$

where ${}_nC_Y = n! / Y!(n - Y)!$ denotes the number of combinations of n distinct objects chosen Y at a time.

For example, if $n = 5$, then the above binomial formula reduces to

$$\text{pr}(Y; \theta) = {}_5C_Y \theta^Y (1 - \theta)^{5-Y} \quad Y = 0, 1, 2, \dots, 5$$

If $Y = 4$, then ${}_5C_4 = 5! / 4!1! = 5$, so the binomial formula then simplifies further to

$$\text{pr}(Y = 4; \theta) = 5 \theta^4 (1 - \theta)$$

The method of maximum likelihood will then choose the value of θ , denoted as $\hat{\theta}$, that maximizes $\text{pr}(Y = 4; \theta)$; in other words, $\hat{\theta}$ satisfies the inequality

$$\text{pr}(Y = 4; \hat{\theta}) > \text{pr}(Y = 4; \theta^*)$$

where θ^* is any alternative value for θ other than $\hat{\theta}$. Notice that neither $\theta^* = 0$ nor $\theta^* = 1$ will maximize $\text{pr}(Y = 4; \theta)$, since

$$\text{pr}(Y = 4; 0) = \text{pr}(Y = 4; 1) = 0$$

For general n and Y , the probability formula

$$\text{pr}(Y; \theta) = {}_nC_Y \theta^Y (1 - \theta)^{n-Y}$$

is called a *likelihood function* and denoted as $L(\theta)$; that is, we can write the likelihood function as

$$L(\theta) = {}_nC_Y \theta^Y (1 - \theta)^{n-Y} \quad (21.2)$$

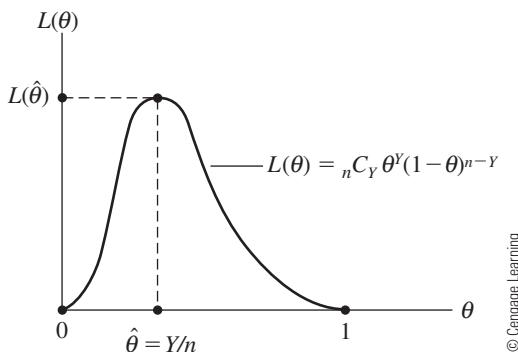
The likelihood function (21.2) gives the probability distribution of the observed data (i.e., Y) as a mathematical function of the unknown parameter θ . The mathematical problem addressed by ML estimation for this scenario is, therefore, to determine that value of θ , called $\hat{\theta}$, that maximizes $L(\theta)$. In other words, the ML estimator of θ is that numerical value that agrees most closely with the observed data in the sense of providing the largest possible value for the probability $L(\theta)$.

Mathematically, we can use calculus to maximize the function (21.2) by setting the derivative of $L(\theta)$ with respect to θ equal to 0 and then solving the resulting equation for θ to obtain $\hat{\theta}$.¹ The resulting ML estimator of θ is $\hat{\theta} = Y/n$, the *sample proportion*.

Thus, for general n and Y , the ML estimator of θ (i.e., $\hat{\theta} = Y/n$) has the property that

$$\text{pr}(Y; \hat{\theta}) > \text{pr}(Y; \theta^*) \quad (21.3)$$

where θ^* is any other value of θ satisfying $0 \leq \theta^* \leq 1$. Figure 21.1 illustrates graphically that the ML estimator $\hat{\theta} = Y/n$ maximizes the likelihood function $L(\theta)$ given by (21.2).



© Cengage Learning

FIGURE 21.1 Illustration of the principle of maximum likelihood, using the function (21.2)

In the example that we have been considering, the data are completely summarized by the single observed value of Y , the number of subjects out of n with the genetic disorder of interest. More generally, the data may consist of several values of Y —namely, Y_1, Y_2, \dots, Y_n obtained from a random sample of n subjects from some larger population.² For example,

$$\begin{aligned} {}^1 \frac{d}{d\theta} [L(\theta)] &= {}^n C_Y [\theta^{Y-1} (1-\theta)^{n-Y} - (n-Y)\theta^Y (1-\theta)^{n-Y-1}] \\ &= {}^n C_Y \theta^{Y-1} (1-\theta)^{n-Y-1} (Y-n\theta) \end{aligned}$$

Equating the above expression to 0 and then solving for θ yields $\hat{\theta} = Y/n$ as the value of θ that maximizes (21.2).

² In our earlier example, the outcome for the i th subject can be denoted by $Y_i = 1$ if the i th subject has a genetic disorder and by $Y_i = 0$ otherwise, $i = 1, 2, \dots, n$. The data on all subjects can then be summarized by the sum $\sum_{i=1}^n Y_i = Y$, the number of individuals in a sample of size n with the genetic disorder.

the Y_i may denote different values of systolic blood pressure (SBP) measured on a random sample of n subjects. For convenience, we denote the collection of n values $\{Y_i, i = 1, \dots, n\}$ using the bold letter \mathbf{Y} ; that is, the observations of Y on all study subjects are denoted by $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$. Also, more generally, there may be more than a single parameter being estimated; that is, instead of the single population proportion θ , there may be several parameters $\theta_1, \theta_2, \dots, \theta_q$ that are to be estimated. We denote the collection of q parameters using the bold Greek $\boldsymbol{\theta}$; that is, we define $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_q)$.

As an example of a situation involving several parameters, consider the multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + E$$

that is fitted using data on n subjects. The data set for the i th subject is given by $(Y_i, X_{i1}, \dots, X_{ik})$, $i = 1, 2, \dots, n$; and $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_k, \sigma^2)$ summarizes the set of parameters. (Note that q , the number of parameters in θ , equals $k + 2$ because of the constant term β_0 in the above model and the variance parameter σ^2 that needs to be estimated.) With the above notation for \mathbf{Y} and $\boldsymbol{\theta}$, we are now in a position to make a more general definition of the principle of maximum likelihood.

Let $L(\mathbf{Y}; \boldsymbol{\theta})$ denote the likelihood function for a data set \mathbf{Y} of n observations from some population characterized by the parameter set $\boldsymbol{\theta}$. The likelihood function $L(\mathbf{Y}; \boldsymbol{\theta})$ can informally be treated as the probability distribution of the multivariable \mathbf{Y} involving the random variables Y_1, Y_2, \dots, Y_n . By a straightforward extension of the general principles of maximum likelihood illustrated earlier when the parameter set consisted of only a single parameter θ , the ML estimator of $\boldsymbol{\theta}$ is the set of estimators $\hat{\boldsymbol{\theta}}$ for which

$$L(\mathbf{Y}; \hat{\boldsymbol{\theta}}) > L(\mathbf{Y}; \boldsymbol{\theta}^*) \quad (21.4)$$

where $\boldsymbol{\theta}^*$ denotes any other set of estimators of the elements of $\boldsymbol{\theta}$. The similarity between expressions (21.3) and (21.4) is clear.

In practice, finding the set $\hat{\boldsymbol{\theta}}$ of numerical functions of the observed data for which $L(\mathbf{Y}; \boldsymbol{\theta})$ is a maximum requires solving a system of q equations in q unknowns. Specifically, since maximizing $L(\mathbf{Y}; \boldsymbol{\theta})$ is equivalent to maximizing the natural logarithm of $L(\mathbf{Y}; \boldsymbol{\theta})$, the elements of $\hat{\boldsymbol{\theta}}$ are typically found as the solutions of the q equations obtained by setting the partial derivative of $\ln L(\mathbf{Y}; \boldsymbol{\theta})$ with respect to each $\theta_j, j = 1, 2, \dots, q$, equal to 0.

In particular, this set of ML equations can be written in the form

$$\frac{\partial}{\partial \theta_j} [\ln L(\mathbf{Y}; \boldsymbol{\theta})] = 0 \quad j = 1, 2, \dots, q \quad (21.5)$$

Except for some special cases, the system of ML equations (21.5) does not lead to explicit algebraic expressions for the ML estimators; consequently, these equations usually must be solved using sophisticated computer algorithms. This complexity results because the set of equations (21.5) typically involves nonlinear (as opposed to linear) functions of the elements of $\boldsymbol{\theta}$, thus requiring the use of so-called iterative computational procedures. Such calculations

are not a major problem, however, since sophisticated computer algorithms are available that have been designed specifically to perform such numerical manipulations.

Chapters 22 (on binary logistic regression), 23 (on polytomous and ordinal logistic regression), and 24 (on Poisson regression) introduce some particular forms of $L(\mathbf{Y}; \boldsymbol{\theta})$ that have important practical applications. For now, it is sufficient to state that ML estimation requires the specification of a likelihood function $L(\mathbf{Y}; \boldsymbol{\theta})$, which is then used to produce the estimator set $\hat{\boldsymbol{\theta}}$ of ML estimators of $\boldsymbol{\theta}$ via the system of equations (21.5).

21.3 Statistical Inference Using Maximum Likelihood

Once the ML estimates have been obtained, the next step is to use the elements of $\hat{\boldsymbol{\theta}}$ to make statistical inferences about the corresponding elements of $\boldsymbol{\theta}$. The computations involved in making such inferences are based on two quantities that are typically included in the output provided by standard ML computer programs.

The first of these two quantities is called the *maximized likelihood value*, which is the numerical value of the likelihood function $L(\mathbf{Y}; \boldsymbol{\theta})$ when the numerical values of the ML estimates (i.e., the elements of $\hat{\boldsymbol{\theta}}$) are substituted for the corresponding elements of $\boldsymbol{\theta}$ in the expression for $L(\mathbf{Y}; \boldsymbol{\theta})$. This corresponds to the Y -axis value of the peak of the curve shown in Figure 21.1. Mathematically, this ML value is denoted $L(\mathbf{Y}; \hat{\boldsymbol{\theta}})$.

The second quantity of interest is the *estimated covariance matrix* $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})$ of the ML estimators. This matrix has q rows and q columns (i.e., a $q \times q$ matrix) and has as its elements the q estimated variances of the q ML estimators (appearing on the diagonal) and the $q(q - 1)/2$ estimated covariances between pairs of estimators (appearing off the diagonal). If, for example, $q = 4$, then $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})$ becomes a 4×4 matrix of the following form:

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} \widehat{\text{Var}}(\hat{\theta}_1) & \widehat{\text{Cov}}(\hat{\theta}_1, \hat{\theta}_2) & \widehat{\text{Cov}}(\hat{\theta}_1, \hat{\theta}_3) & \widehat{\text{Cov}}(\hat{\theta}_1, \hat{\theta}_4) \\ \widehat{\text{Cov}}(\hat{\theta}_1, \hat{\theta}_2) & \widehat{\text{Var}}(\hat{\theta}_2) & \widehat{\text{Cov}}(\hat{\theta}_2, \hat{\theta}_3) & \widehat{\text{Cov}}(\hat{\theta}_2, \hat{\theta}_4) \\ \widehat{\text{Cov}}(\hat{\theta}_1, \hat{\theta}_3) & \widehat{\text{Cov}}(\hat{\theta}_2, \hat{\theta}_3) & \widehat{\text{Var}}(\hat{\theta}_3) & \widehat{\text{Cov}}(\hat{\theta}_3, \hat{\theta}_4) \\ \widehat{\text{Cov}}(\hat{\theta}_1, \hat{\theta}_4) & \widehat{\text{Cov}}(\hat{\theta}_2, \hat{\theta}_4) & \widehat{\text{Cov}}(\hat{\theta}_3, \hat{\theta}_4) & \widehat{\text{Var}}(\hat{\theta}_4) \end{bmatrix}$$

Diagonal

Notice that $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})$ is always a *symmetric matrix* since the collection of covariances below the diagonal of the matrix is the mirror image of the collection of covariances above the diagonal.

In general, the numerical value of the maximized likelihood function and the numerical values appearing in the estimated covariance matrix must be calculated by an ML computer program, since the form of the likelihood function typically under consideration does not lead (as we will later illustrate) to explicit algebraic expressions either for the ML estimators or for their variances and covariances. One special case in which the ML equations do have specific algebraic solutions is when a multiple linear regression model is fitted assuming normally distributed and mutually independent response variables; in this case, the least-squares estimators (or weighted least-squares estimators if it is assumed that $\text{Var}(Y_i) = \sigma_i^2$) of the regression coefficients are identical to the ML estimators.

For a numerical example using a linear regression model, we refer to the data in Table 15.2 of Chapter 15, which concern a laboratory study undertaken to determine the relationship between the dosage (X) of a certain drug and weight gain (Y) on a small sample of eight laboratory animals randomly assigned to one of eight dosage levels. For convenience, these data are reproduced in Table 21.1. (The sample size of $n = 8$ is actually too small to justify ML methods; see Section 21.3.1 for discussion of large-sample properties of ML estimates.)

TABLE 21.1 Weight gain after two weeks as a function of dosage level (repeat of Table 15.2)

Dosage level (X)	1	2	3	4	5	6	7	8
Weight gain (Y)	1.0	1.2	1.8	2.5	3.6	4.7	6.6	9.1

© Cengage Learning

For these data, we now consider fitting the following straight-line model,

$$\text{Model 1: } E(Y|X) = \beta_0 + \beta_1 X$$

We assume that the Y_i are normally distributed with variance $\text{Var}(Y_i) = \sigma^2$ not varying with i . We further assume that X_i is measured without error and that the n random variables Y_1, Y_2, \dots, Y_n are mutually independent.³

The expression for the distribution (density function) of the normally distributed random variable Y_i under model 1 is given by

$$f(Y_i; \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} [Y_i - (\beta_0 + \beta_1 X_i)]^2 \right\} \quad (21.6)$$

where $-\infty < Y_i < +\infty$. Note that (21.6) is a function of three parameters—namely, β_0 , β_1 , and σ^2 . Under the assumption that the $\{Y_i, i = 1, 2, \dots, n\}$ are mutually independent, it can be shown from statistical theory that the joint distribution of Y_1, Y_2, \dots, Y_n (i.e., the likelihood function) is, from (21.6),

$$L(\mathbf{Y}; \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n f(Y_i; \beta_0, \beta_1, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)]^2 \right\} \quad (21.7)$$

(In our example, $n = 8$.)

³ The assumption of *mutual independence* for a set of random variables is the strongest assumption that can be made about the joint behavior of this set. Such an assumption allows precise description of the joint distribution of the variables (i.e., the likelihood function) solely on the basis of knowledge of the separate behavior (i.e., the so-called marginal distribution) of each variable in the set. In particular, the joint distribution under mutual independence is simply the product of the marginal distributions.

The ML value obtained by maximizing the above likelihood formula is given by the algebraic form

$$L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2) = (2\pi\hat{\sigma}^2 e)^{-n/2} \quad (21.8)$$

For the particular set of $n = 8$ data points, the ML estimators obtained from maximizing the likelihood function (21.7) are⁴

$$\hat{\beta}_0 = -1.20, \quad \hat{\beta}_1 = 1.11, \quad \hat{\sigma}^2 = 0.6288$$

Hence, from (21.8),

$$\begin{aligned} L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2) &= L(\mathbf{Y}; -1.20, 1.11, 0.6288) \\ &= [(2(3.1416)(0.6288)(2.7183))]^{-8/2} = (10.7397)^{-4} \end{aligned}$$

The number that would typically appear in the computer output using an ML estimation program with the likelihood function (21.7) and the data set under consideration would be $-2 \ln L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2) = -2 \ln(10.7397)^{-4} = 18.9916$.

The estimated covariance matrix for the ML estimators $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$ has the following general form:

$$\hat{\mathbf{V}}(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2) = \begin{bmatrix} \widehat{\text{Var}}(\hat{\beta}_0) & \widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1) & \widehat{\text{Cov}}(\hat{\beta}_0, \hat{\sigma}^2) \\ \widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1) & \widehat{\text{Var}}(\hat{\beta}_1) & \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\sigma}^2) \\ \widehat{\text{Cov}}(\hat{\beta}_0, \hat{\sigma}^2) & \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\sigma}^2) & \widehat{\text{Var}}(\hat{\sigma}^2) \end{bmatrix}$$

The estimated variances of the ML estimators appear on the diagonal of this symmetric matrix, and the estimated covariances are given by the off-diagonal elements of the matrix.

For the data under consideration, the following estimated covariance matrix is obtained:

$$\hat{\mathbf{V}}(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2) = \begin{bmatrix} 0.3818 & -0.0674 & 0 \\ -0.0674 & 0.0150 & 0 \\ 0 & 0 & 0.0988 \end{bmatrix} \quad (21.9)$$

This is the matrix that a computer program would print out based on the use of the likelihood function (21.7) for these data.

⁴ The ML estimator $\hat{\sigma}^2$ of σ^2 is actually a biased estimator of σ^2 . The unbiased estimator of σ^2 is

$$\left(\frac{n}{n-2}\right)\hat{\sigma}^2 = \frac{\text{SSE}}{n-2}$$

where SSE is the sum of squares of residuals about the fitted straight line using the unweighted least-squares estimates for model 1. (In Table 15.4 of Chapter 15, we found SSE = 5.03). Thus, the ML method does not always produce unbiased estimators of the parameters of interest, although the extent of such bias generally decreases as the sample size increases when the likelihood function is correctly specified.

Later we will use the maximized likelihood value of $(10.7397)^{-4}$ to carry out a likelihood ratio test, which involves comparing the ratios of maximized likelihoods for different models. We now focus on the estimated covariance matrix (21.9) and use it to perform certain inference-making exercises.

21.3.1 Hypothesis Testing Using Wald Statistics

For the data we have been considering (Table 21.1), where $n = 8$, we have determined the maximized likelihood value $L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)$ and the estimated covariance matrix $\widehat{\mathbf{V}}(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)$ for the straight-line model (model 1)

$$E(Y|X) = \beta_0 + \beta_1 X$$

An important question of interest for this model is to determine whether the dosage (X) is a significant predictor of weight gain (Y). This can be formulated as a hypothesis-testing question, and the corresponding null hypothesis can be stated as $H_0: \beta_1 = 0$. We will assume that our alternative hypothesis is two-sided; that is, $H_A: \beta_1 \neq 0$.

Based on the large-sample properties⁵ of ML estimators, it can be shown for model 1 that the quantity

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_1)}} \quad (21.10)$$

will behave approximately as a standard normal random variable (i.e., a Z variate) when the sample size is large. Hence, a test of $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$ can be based on the Z statistic:

$$Z = \frac{\hat{\beta}_1}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_1)}}$$

which for large n has approximately a standard normal distribution under $H_0: \beta_1 = 0$. This test statistic is called a *Wald statistic*.

For our data, $\hat{\beta}_1 = 1.11$ and $\widehat{\text{Var}}(\hat{\beta}_1) = 0.0150$, so that

$$Z = \frac{1.11}{\sqrt{0.0150}} = 9.063$$

which is highly significant.

⁵ The ML method produces estimators whose properties are optimal for *large samples* when the assumed likelihood function is correct. ML estimators are said to be *asymptotically optimal* in the sense that desirable properties such as unbiasedness, minimum variance, and normality hold exactly in the limit only as the amount of data becomes infinitely large. Thus, it is typically reasonable to assume for large data sets that an ML estimator $\hat{\theta}$ of a parameter θ is essentially unbiased, has a small estimated variance $\widehat{\text{Var}}(\hat{\theta})$, and that the standardized random variable $(\hat{\theta} - \theta)/\sqrt{\widehat{\text{Var}}(\hat{\theta})}$ is approximately $N(0, 1)$ when the appropriate likelihood function is being used.

An equivalent test can be made using the chi-square distribution; in particular, since Z^2 has a χ_1^2 distribution when Z is $N(0,1)$, it follows that

$$Z^2 = \frac{(\hat{\beta}_1)^2}{\widehat{\text{Var}}(\hat{\beta}_1)} \quad (21.11)$$

will have approximately a χ_1^2 distribution under $H_0: \beta_1 = 0$ when the sample size is sufficiently large. In our numerical example, then, this chi-square statistic⁶ has the highly significant value of $(9.063)^2 = 82.14$.

ML-based test statistics such as Z^2 are usually assumed to have *large-sample* chi-square distributions under the null hypotheses of interest. (The degrees of freedom may be larger than 1 if the null hypothesis involves more than one parameter.) The large-sample requirement is crucial to the validity of such *asymptotic* procedures. In large-sample situations, the P -values associated with the use of ML-based chi-square statistics are comparable to those based on other methods of estimation and statistical inference (e.g., using least-squares methods in multiple linear regression models). In small-sample situations, discrepancies among different analysis procedures can be large. Since our particular data set consists of only $n = 8$ data points, the sample size here is admittedly too small to justify the use of asymptotic theory for Z^2 .

21.3.2 Interval Estimation

To obtain confidence intervals for parameters using ML estimation, we can again appeal to the approximate normality of standardized ML estimators (see footnote 5) when the sample size is large. In our example, since

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_1)}}$$

⁶ A reader who inspects Table 15.4 in Chapter 15 may wonder about the difference between the ML-based Z^2 statistic given by (21.11) and the $F (= t^2)$ statistic in Table 15.4 (with the value 62.05), both of which can be used to test $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$ under model 1. The primary reason for this numerical difference is that the very small sample size ($n = 8$) is totally inappropriate for the use of large-sample ML methods. More specifically, the F statistic given by

$$F = \frac{\text{MS Regression}(X)}{\text{MS Residual}(X)} = \frac{52.04}{0.84} = 62.05$$

uses a different (*unbiased*) estimator of $\text{Var}(\hat{\beta}_1)$ than the (*biased for small samples*) ML estimator of $\text{Var}(\hat{\beta}_1)$ that is used in the denominator of the Z^2 statistic. In particular, the numerical value of the *unbiased* estimator of $\text{Var}(\hat{\beta}_1)$ is 0.01986, whereas the numerical value of the *biased* estimator of $\text{Var}(\hat{\beta}_1)$ in (21.11) above is 0.0150. Note, also, that the t statistic corresponding to the F statistic is given by

$$t = \frac{1.11}{\sqrt{.01986}} = 7.88$$

which differs from the Wald statistic given by $Z = 9.063$. For large samples, there would be essentially no numerical differences between these two testing methods.

is approximately $N(0,1)$ for large samples under model 1, an approximate $100(1 - \alpha)\%$ large-sample ML-based confidence interval for β_1 has the general form

$$\hat{\beta}_1 \pm Z_{1-(\alpha/2)} \sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} \quad (21.12)$$

where $\text{pr}[Z > Z_{1-(\alpha/2)}] = \alpha/2$ when Z is $N(0, 1)$. For instance, a 95% ML-based confidence interval for β_1 using our data set of size $n = 8$ is

$$1.11 \pm 1.96 \sqrt{0.0150} = 1.11 \pm 0.24$$

which yields the interval $(0.87, 1.35)$.

As another example, let us obtain a 95% large-sample confidence interval for the true mean of Y when X is set to some value X_0 within the range of X -values for the data under consideration. If we denote this parameter as $E(Y|X = X_0)$, the ML estimator of $E(Y|X = X_0)$ based on fitting model 1 is

$$\hat{E}(Y|X = X_0) = \hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

Since \hat{Y}_0 is a linear combination of random variables (namely, $\hat{\beta}_0$ and $\hat{\beta}_1$), it follows that $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$ has estimated variance⁷

$$\widehat{\text{Var}}(\hat{Y}_0) = \widehat{\text{Var}}(\hat{\beta}_0) + X_0^2 \widehat{\text{Var}}(\hat{\beta}_1) + 2X_0 \widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1)$$

It follows that a 95% confidence interval for $E(Y|X = X_0)$ is given by

$$\hat{Y}_0 \pm 1.96 \sqrt{\widehat{\text{Var}}(\hat{Y}_0)}$$

For our particular numerical example, suppose that $X_0 = \bar{X} = 4.50$. Then using the values for $\widehat{\text{Var}}(\hat{\beta}_0)$, $\widehat{\text{Var}}(\hat{\beta}_1)$, and $\widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1)$ in the matrix (21.9), it follows that

$$\begin{aligned}\widehat{\text{Var}}(\hat{Y}_0) &= 0.3818 + (4.50)^2(0.0150) + 2(4.50)(-0.0674) \\ &= 0.3818 + 0.3038 - 0.6066 = 0.0790\end{aligned}$$

Since $\hat{Y}_0 = -1.20 + 1.11(4.50) = -1.20 + 5.00 = 3.80$, it follows that the 95% ML-based large-sample confidence interval for $E(Y|X = 4.50)$ is

$$3.80 \pm 1.96 \sqrt{0.0790}$$

which yields the interval $(3.25, 4.35)$.

⁷ In general, if $\hat{L} = \sum_{j=1}^k a_j X_j$ is a linear function of the k random variables X_1, X_2, \dots, X_k and if a_1, a_2, \dots, a_k are known constants, then

$$\text{Var } \hat{L} = \sum_{j=1}^k a_j^2 \text{Var } X_j + 2 \sum_{j=1}^{k-1} \sum_{j'=j+1}^k a_j a_{j'} \text{Cov}(X_j, X_{j'})$$

For example, when $k = 2$, we have

$$\text{Var } \hat{L} = a_1^2 \text{Var } X_1 + a_2^2 \text{Var } X_2 + 2a_1 a_2 \text{Cov}(X_1, X_2)$$

21.3.3 Hypothesis Testing Using Likelihood Ratio Tests

As its name suggests, a likelihood ratio (LR) test involves a ratio comparison of (maximized) likelihood values. To see how such a comparison can be used in making statistical inferences, let us now consider two other models in addition to model 1 that we might consider for the analysis of the previously described laboratory study to determine the relationship between the dosage (X) of a certain drug and weight gain (Y) on a small sample of eight laboratory animals. The three models are given as follows:

$$\text{Model 0: } E(Y) = \beta_0$$

$$\text{Model 1: } E(Y|X) = \beta_0 + \beta_1 X$$

$$\text{Model 2: } E(Y|X, X^2) = \beta_0 + \beta_1 X + \beta_2 X^2$$

Dispensing with the mathematical details, it can be shown that the maximized likelihood values $L(\mathbf{Y}; \hat{\beta}_0, \hat{\sigma}_0^2)$ and $L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}_2^2)$ for models 0 and 2, under the same assumptions used for the ML fitting of model 1, are equal to $(121.8426)^{-4}$ and $(0.4270)^{-4}$, respectively. Recall that the maximized likelihood value $L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_1^2)$ obtained for model 1 was $(10.7397)^{-4}$. (Note that we previously denoted $\hat{\sigma}_1^2$ as $\hat{\sigma}^2$ for model 1.)

From the above information, we can see that

$$L(\mathbf{Y}; \hat{\beta}_0, \hat{\sigma}_0^2) < L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_1^2) < L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}_2^2) \quad (21.13)$$

This result reflects the principle of multiple regression analysis that the squared multiple correlation coefficient (R^2) will always increase somewhat for any set of data whenever another regression parameter is included in the model under consideration. Analogously, since model 0 is a special case of model 1 when $\beta_1 = 0$ and since model 1, in turn, is a special case of model 2 when $\beta_2 = 0$, (21.13) must hold for any set of data.

The fact that models 0, 1, and 2 constitute a *hierarchical class* of models leads to an important application of LR tests.⁸ This is because the magnitude of the ratio of two maximized likelihood values reflects how much the maximized likelihood value for one specific model has changed based on the addition (or deletion) of one or more parameters to (or from) the given model. A decision to reject some null hypothesis based on an LR test depends on whether some appropriate function of the ratio of maximized likelihoods for the two models being compared (i.e., the test statistic) is large enough to indicate a statistically significant disparity between the two maximized likelihood values. This philosophy of assessing the significance of a change in maximized likelihood values is completely analogous to the philosophy in standard multiple regression analysis of assessing the significance of a change in R^2 based on the addition of one or more parameters to a given model.

⁸ For our purposes, a *hierarchical class* of models is a group of models each of which, except for the most complex (i.e., the one containing the largest number of regression coefficients), can be obtained as a special case of another more-complex model in the class by setting one or more regression coefficients equal to 0 in the more complex model. For example, the three models $E(Y|X_1) = \beta_0 + \beta_1 X_1$, $E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, and $E(Y|X_1, X_2, X_3, X_1 X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2$ constitute a hierarchical class, since $E(Y|X_1, X_2)$ becomes $E(Y|X_1)$ when $\beta_2 = 0$ and $E(Y|X_1, X_2, X_3, X_1 X_2)$ reduces to $E(Y|X_1, X_2)$ when $\beta_3 = \beta_{12} = 0$.

In earlier chapters, we discussed how to use a partial F test (or, equivalently, a t test) to assess whether an increase in R^2 is statistically significant when a parameter is added to a given regression model, and we illustrated how to use a multiple partial F test to make the same assessment when more than one new parameter is introduced into that model.

An *LR test* is performed analogously. Instead of having an F distribution, however, the LR test statistic to be used will have, *in large samples, approximately a chi-square distribution under the null hypothesis*. The degrees of freedom for this distribution will equal the number of parameters in the more complex model that must be set equal to 0 to obtain the less complex model as a special case. Also, as with a partial or multiple partial F test, *an LR test will always involve a comparison between two models that are members of a hierarchical class*, so that the terms *more complex* and *less complex* are well defined; alternatively, these more complex and less complex models are often conveniently referred to as the *full model* (F) and the *reduced model* (R), respectively.

The specific test statistic to be used in performing an LR test is of the general form

$$\text{LR} = -2 \ln(L_R/L_F)$$

where L_R is the maximized likelihood for the less complex model and L_F is the maximized likelihood for the more complex model. Thus, the LR statistic is a function of the ratio L_R/L_F of maximized likelihoods. Since L_F corresponds to the more complex model, it follows from algebra that

$$0 < L_R < L_F$$

so that

$$0 < L_R/L_F < 1$$

Thus,

$$-\infty < \ln(L_R/L_F) < 0$$

so that

$$0 < \text{LR} = -2 \ln(L_R/L_F) < +\infty$$

Clearly, the larger L_F is relative to L_R , the larger the value of the test statistic LR will be, and thus the more likely it will be that this value falls in the rejection region (i.e., the specified area in the upper tail of the appropriate chi-square distribution). If L_F is significantly larger than L_R , then the model corresponding to L_F agrees with (i.e., fits) the data significantly better than the model corresponding to L_R .

The LR statistic can then be written equivalently, using algebraic rules for logarithms,⁹ as follows:

$$\text{LR} = -2 \ln(L_R/L_F) = -2 \ln L_R - (-2 \ln L_F) \quad (21.14)$$

⁹ $\ln(a/b) = (\ln a - \ln b)$ for any two positive values a and b .

The expressions $-2 \ln L_R$ and $-2 \ln L_F$ are called *log-likelihood statistics*. When a computer program is used to carry out the ML estimation procedure, these two log-likelihood statistics are provided in separate outputs for the two models being compared. To carry out the LR test, the investigator simply finds $-2 \ln L_R$ and $-2 \ln L_F$ from the output and then subtracts $-2 \ln L_F$ from $-2 \ln L_R$ to obtain the value of the test statistic.

We now provide some numerical examples of LR tests. Again, we will consider models 0, 1, and 2 and the data used earlier. First, we will compare models 0 and 1 using an LR test. Since model 0 is a special case of model 1 when $\beta_1 = 0$, this LR test addresses $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$. The appropriate chi-square distribution for the LR statistic under this H_0 has 1 d.f. (since one parameter is restricted to be equal to 0 under H_0). Using (21.14) with $L_R = L(\mathbf{Y}; \hat{\beta}_0, \hat{\sigma}_0^2)$ and $L_F = L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_1^2)$, the LR statistic is given by the following formula:

$$\begin{aligned} \text{LR} &= -2 \ln L_R - (-2 \ln L_F) \\ &= -2 \ln L(\mathbf{Y}; \hat{\beta}_0, \hat{\sigma}_0^2) - (-2 \ln L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_1^2)) \end{aligned}$$

where the log-likelihood statistics¹⁰ produced by computer output are given by

$$-2 \ln L(\mathbf{Y}; \hat{\beta}_0, \hat{\sigma}_0^2) = 38.4218 \quad \text{and} \quad -2 \ln L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_1^2) = 18.9916$$

Substituting the above numerical values into the LR formula, we obtain

$$\text{LR} = 38.4218 - 18.9916 = 19.43$$

which corresponds to a P -value of less than .0005 and thus provides strong evidence in favor of $H_A: \beta_1 \neq 0$.

The discrepancy between the LR statistic of 19.43 and the Wald statistic of 82.14 is alarming. However, this discrepancy is entirely plausible because of the small sample size ($n = 8$). Because the two test statistics are only asymptotically equivalent, their numerical values are reasonably close only when n is large. Thus, although we have chosen this data set for pedagogical purposes, it is actually an inappropriate one for the application of large-sample statistical procedures. Nevertheless, the LR and Wald tests will generally not give identical numerical results even when n is large. Since the LR statistic is considered to have better statistical properties than the Wald statistic (e.g., the Wald statistic can be more influenced by collinearity problems), we recommend the use of the LR statistic over the Wald statistic when in doubt.

As another illustration, an LR test of $H_0: \beta_2 = 0$ versus $H_A: \beta_2 \neq 0$ involves a comparison of the maximized likelihood values for models 1 and 2. Using (21.14) again, this time with $L_R = L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_1^2)$ and $L_F = L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}_2^2)$, the LR statistic is given by the

¹⁰ Since we previously stated that $L(\mathbf{Y}; \hat{\beta}_0, \hat{\sigma}_0^2) = (121.8426)^{-4}$, it follows that

$$-2 \ln L(\mathbf{Y}; \hat{\beta}_0, \hat{\sigma}_0^2) = -2(-4) \ln(121.8426) = -2(-4)(4.8027) = 38.4218$$

Also, since we previously stated that $L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_1^2) = (10.7397)^{-4}$, it follows that

$$-2 \ln L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_1^2) = -2(-4) \ln(10.7397) = -2(-4)(2.3739) = 18.9916$$

following formula:

$$\begin{aligned} \text{LR} &= -2 \ln L_R - (-2 \ln L_F) \\ &= -2 \ln L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_1^2) - (-2 \ln L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}_2^2)) \end{aligned}$$

where the log-likelihood statistics¹¹ produced by computer output are given by

$$-2 \ln L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_1^2) = 18.9916 \quad \text{and} \quad -2 \ln L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}_2^2) = -6.8078$$

Substituting the above numerical values into the LR formula, we obtain

$$\text{LR} = 18.9916 - (-6.8078) = 25.80$$

which provides strong evidence in favor of $H_A: \beta_2 \neq 0$ (again, assuming that this LR statistic has approximately a chi-square distribution with 1 d.f. under H_0).

As a third illustration, we consider an LR test involving the maximized likelihood values for models 0 and 2, which tests $H_0: \beta_1 = \beta_2 = 0$ versus H_A : “At least one of the parameters β_1 and β_2 differs from 0.” This test is analogous to an overall F test in standard least-squares regression analysis because we are testing whether the regression coefficients of all predictors (other than the intercept) are simultaneously equal to 0. The appropriate LR-test statistic is

$$\begin{aligned} \text{LR} &= -2 \ln L_R - (-2 \ln L_F) \\ &= -2 \ln L(\mathbf{Y}; \hat{\beta}_0, \hat{\sigma}_0^2) - (-2 \ln L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}_2^2)) \end{aligned}$$

which for large samples has approximately a chi-square distribution with 2 d.f. under $H_0: \beta_1 = \beta_2 = 0$. The computed value of this test statistic is

$$\text{LR} = 38.4218 - (-6.8078) = 45.23$$

which is highly significant.

Even though a sample of size $n = 8$ is too small to justify using statistical inference-making procedures whose desirable properties hold only for large samples, the decisions made about the importance of the linear (β_1) and quadratic (β_2) effects in the data agree with the conclusions previously drawn based on the standard regression analysis given in Section 15.6 of Chapter 15.

One way to assess the goodness of fit of the second-degree model in X (i.e., model 2) is to employ an LR statistic to examine whether adding a cubic term in X (i.e., the term $\beta_3 X^3$) to the second-degree model significantly improves prediction of Y . In other words, we now consider a test of $H_0: \beta_3 = 0$ versus $H_A: \beta_3 \neq 0$. The appropriate test statistic here is computed to be

$$\begin{aligned} \text{LR} &= -2 \ln L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}_2^2) - (-2 \ln L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\sigma}_3^2)) \\ &= -6.8078 - (-17.0234) = 10.21 \end{aligned}$$

¹¹ Since we previously stated that $L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}_2^2) = (0.4270)^{-4}$, it follows that

$-2 \ln L(\mathbf{Y}; \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}_2^2) = -2(-4) \ln (0.4270) = -2(-4)(-0.8510) = -6.8078$

which corresponds to a P -value between .0005 and .005 based on a chi-square distribution with 1 d.f. Although this result argues for adding the term $\beta_3 X^3$ to the quadratic model, other information suggests the contrary. Specifically, the change in R^2 in going from a quadratic to a cubic model is negligible ($\Delta R^2 = 0.999 - 0.997 = 0.002$); a plot of the data clearly suggests no more than a second-degree model in X ; and the LR test under discussion is based on a data set with only $n = 8$ observations and so is not completely reliable. In light of these considerations, the best conclusion is that a second-degree model in X is appropriate for describing the relationship between X and Y with high precision.

21.3.4 Comparison of Computer Output Using Different SAS Procedures

Table 21.2 on the next page provides a summary of SAS computer output for fitting linear (i.e., straight-line) and quadratic multiple linear regression models using the data in Table 21.1 that relates dosage (X) to weight gain (Y). We have used three different SAS computer procedures that fit multiple linear regression models—SAS's REG, MIXED, and GENMOD procedures. The method of estimation used by the REG procedure is unweighted least squares. The MIXED procedure uses ML estimation and a variation of ML estimation called REML (i.e., Restricted ML estimation). The GENMOD procedure fits both linear and nonlinear models and allows for the analysis of repeated measures data in which responses (Y 's) on different subjects (or clusters) may be correlated (as discussed later in Chapters 25 and 26). When GENMOD is applied to a linear regression model in which all responses are assumed to be mutually independent, ML estimation is used to fit the model.

From Table 21.2, we highlight the following results:

1. For models 1 and 2 considered separately, all four SAS procedures (considering the REML and ML approaches using MIXED as different procedures) yield identical estimated regression coefficients; these results demonstrate that unweighted least-squares and ML estimates of corresponding regression coefficients are identical for multiple linear regression models (assuming normally distributed and mutually independent responses with homogeneous variance).
2. For model 1, the REG procedure and the REML option in the MIXED procedure yield identical standard errors, t statistics (e.g., 7.88 for dose), and F statistics (i.e., 62.05); these results demonstrate that the REML option in MIXED performs unweighted least-squares estimation for multiple linear regression models (assuming normally distributed and mutually independent responses with homogeneous variance).
3. For model 2, the REG procedure and the REML option in the MIXED procedure yield identical standard errors and t statistics (e.g., 11.09 for dosesq); these results also confirm that the REML option in MIXED performs unweighted least-squares estimation for multiple linear regression models.
4. For model 1, the ML option in the MIXED procedure and the GENMOD procedure yield identical standard errors, Wald chi-square statistics (e.g., 82.74 for dose), and $-2 \ln L_1$ values (i.e., 19.0); these results indicate that the ML

TABLE 21.2 Edited SAS computer output from SAS's REGRESSION, MIXED, and GENMOD procedures applied to the dose (X), weight gain (Y) data of Table 21.1

SAS Procedure	Model 1: $E(Y X) = \beta_0 + \beta_1 X$						Model 2: $E(Y X, X^2) = \beta_0 + \beta_1 X + \beta_2 X^2$					
REG (Least Squares)	Variable	DF	Estimate	SE	t	$\text{Pr} > t $	Variable	DF	Estimate	SE	t	$\text{Pr} > t $
	Intercept	1	-1.1964	0.7135	-1.68	0.1446	Intercept	1	1.3482	0.2767	4.87	0.0046
	dose	1	1.1131	0.1413	7.88	0.0002	dose	1	-0.4137	0.1411	-2.93	0.0326
	Analysis of Variance						dosesq	1	0.1696	0.0153	11.09	0.0001
	Source	DF	SS	MS	F	$\text{Pr} > F$	Source	DF	SS	MS	F	$\text{Pr} > F$
	Model	1	52.0372	52.0372	62.05	0.0002	Model	2	56.8720	28.4360	722.73	<.0001
	Error	6	5.0315	0.8386			Error	5	0.1967	0.0394		
	Total	7	57.06875				Total	7	57.06875			
	R-Square		0.9118				R-Square		0.9966			
	$-2 \ln L_1 = 21.8$						$-2 \ln L_1 = 9.0$					
MIXED (REML option)	Effect	Estimate	SE	DF	t	$\text{Pr} > t $	Effect	Estimate	SE	DF	t	$\text{Pr} > t $
	Intercept	-1.1964	0.7135	6	-1.68	0.1446	Intercept	1.3482	0.2767	5	4.87	0.0046
	dose	1.1131	0.1413	6	7.88	0.0002	dose	-0.4137	0.1411	5	-2.93	0.0326
	Type 3 Tests of Fixed Effects						dosesq	0.1696	0.0153	5	11.09	0.0001
	Effect	NDF	DDF	F		$\text{Pr} > F$	Effect	NDF	DDF	F	$\text{Pr} > F$	
	dose	1	6	62.05	.0002		dose	1	5	122.88	0.0001	
	$-2 \ln L_1 = 19.0$						$-2 \ln L_1 = -6.9$					
MIXED (ML option)	Effect	Estimate	SE	DF	t	$\text{Pr} > t $	Effect	Estimate	SE	DF	t	$\text{Pr} > t $
	Intercept	-1.1964	0.6179	6	-1.94	0.1010	Intercept	1.3482	0.2188	5	6.16	0.0016
	dose	1.1131	0.1224	6	9.10	<.0001	dose	-0.4137	0.1115	5	-3.71	0.0139
	Type 3 Tests of Fixed Effects						dosesq	0.1696	0.0121	5	14.02	<.0001
	Effect	NDF	DDF	F		$\text{Pr} > F$	Effect	NDF	DDF	F	$\text{Pr} > F$	
	dose	1	6	82.74	<.0001		dose	1	5	196.61	0.0139	
GENMOD (ML)	$\ln L_1 = -9.4967$ (i.e., $-2 \ln L_1 = 19.0$)						$\ln L_1 = 3.4700$ (i.e., $-2 \ln L_1 = -6.94$)					
	Parameter	DF	Estimate	SE			Parameter	DF	Estimate	SE	Wald CS	$\text{Pr} > \text{CS}$
	Intercept	1	-1.1964	0.6179	3.75	0.0529	Intercept	1	1.3482	0.2188	37.98	<.0001
	dose	1	1.1131	0.1224	82.74	<.0001	dose	1	-0.4137	0.1115	13.76	0.0002
	dosesq	1					dosesq	1	0.1696	0.0121	196.61	<.0001

option in MIXED and the GENMOD procedure both perform ML estimation for multiple linear regression models (assuming normally distributed and mutually independent responses with homogeneous variance).

5. For model 2, the ML option in the MIXED procedure and the GENMOD procedure yield identical standard errors, Wald chi-square statistics (e.g., 196.61 for dosesq), and $-2 \ln L_1$ values (i.e., -6.9); these results also indicate that the ML option in MIXED and the GENMOD procedure both perform ML estimation for multiple linear regression models (assuming normally distributed and mutually independent responses with homogeneous variance).

We have previously indicated that, when fitting multiple linear regression models, ML estimators of variance (and corresponding standard errors) are biased for small samples when compared to variance estimators based on unweighted least-squares estimation (which provides unbiased variance estimators when all assumptions are satisfied). Thus, particularly for small n , inference-making conclusions may differ depending on whether least-squares or ML estimation procedures are applied. Consequently, we recommend using unweighted least-squares estimation (e.g., as performed by SAS's REG or MIXED-REML procedures), rather than ML estimation (e.g., SAS's MIXED-ML or GENMOD procedures), when fitting multiple linear regression models assuming normally distributed and mutually independent responses with homogeneous variance. For large n , however, the bias in variance estimators from ML estimation will typically be small, so that inference-making conclusions using least-squares and ML estimation approaches typically will not differ very much.

Primarily for pedagogical reasons, we have illustrated the use of ML procedures in this chapter using multiple linear regression models, since we have only considered such models to this point in the text. Nevertheless, we emphasize that ML estimation is the method of choice for estimating the parameters in nonlinear models for nonnormal responses such as the logistic regression model (discussed in Chapters 22 and 23) and the Poisson regression model (discussed in Chapter 24).

21.4 Summary

In this chapter, we have discussed the maximum likelihood (ML) method for estimating parameters and for making statistical inferences. In ML estimation, the likelihood function to be maximized must be specified. For multiple linear regression involving mutually independent normally distributed response variables with homogeneous variance, the ML estimates of regression coefficients are identical to the unweighted least-squares estimates. Note, however, that the ML estimators of the variances of estimated regression coefficients are slightly biased when compared to the corresponding variance estimators obtained from unweighted least squares. Nevertheless, ML estimation is particularly useful for estimating regression coefficients in nonlinear models involving nonnormal responses (e.g., logistic regression or Poisson regression models discussed in subsequent chapters). The estimation procedure for such models typically requires a computer program that employs an iterative model-fitting algorithm.

Procedures for testing hypotheses and constructing confidence intervals use maximized likelihood values and related estimated covariance matrices for the various models under

study. Two alternative large-sample test procedures for testing hypotheses involve the Wald statistic and the likelihood ratio (LR) statistic. Both statistics frequently yield similar (though not identical) numerical results; when doubt exists, the LR statistic is preferred. Upper and lower limits, respectively, of large-sample confidence intervals for (linear functions of) regression coefficients are obtained by adding to and subtracting from an ML point estimate a percentile of the standard normal distribution multiplied by the estimated standard error of the estimated linear function under consideration.

Problems

1. A certain drug is suspected of lowering blood pressure as a side effect. A clinical trial is conducted to investigate this suspicion. Thirty-two patients are randomized into drug and placebo groups (16 per group). Their initial and posttreatment systolic blood pressures (SBPs, measured in mm Hg) and their body sizes as measured by the quetelet index (QUET) are recorded.
 The researchers conducted a multiple linear regression analysis to investigate whether the mean changes in SBP differ between the placebo and drug groups, controlling for initial SBP and QUET. The dependent variable is posttreatment SBP. One model under consideration involves three predictors: DRUG status, initial SBP, and QUET as main effects only (together with an intercept term). Suppose that a computer program that calculates unweighted least-squares estimates of the regression coefficients (e.g., the REG procedure in SAS) is used to fit this model.
 - a. Are the unweighted least-squares estimates of the regression coefficients identical to the ML estimates of these same coefficients for the above model? Explain briefly.
 - b. Assuming that the estimation procedure is ML, how would one carry out a Wald test for the effect of the DRUG status variable, controlling for initial SBP and QUET? (Specify the null hypothesis being tested, the form of the test statistic, and its large-sample distribution under the null hypothesis.)
 - c. Is the Wald test procedure described in part (b) equivalent to a partial F test obtained from unweighted least-squares estimation? Explain briefly.
 - d. How would one carry out an LR test for the effect of the DRUG status variable, controlling for initial SBP and QUET? (Specify the null hypothesis being tested, the form of the test statistic, and its large-sample distribution under the null hypothesis.)
 - e. Should one expect to obtain the same P -value for the LR test described in part (d) as for the Wald test described in part (c)? Explain briefly.
 - f. Assuming ML estimation, state the formula for a large-sample 95% confidence interval for the effect of the DRUG status variable, controlling for initial SBP and QUET.
2. The data for this question consist of a sample of 50 persons from the 1967–1980 Evans County Study (Schoenbach et al. 1986). Two basic independent variables are of interest: AGE and chronic disease status (CHR), where CHR is coded as 0 = none and 1 = presence of chronic disease. A product term of the form AGE \times CHR is also

considered. The dependent variable is time until death, a continuous variable. The primary question of interest is whether CHR, considered as the exposure variable, is related to survival time, controlling for AGE. The computer results, based on ML estimation,¹² are as follows:

Model 1

ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES				
Parameter	Estimate	Standard Error	Chi-Square	Pr > ChiSq
CHR	0.8595	0.3116	7.61	0.0058
$-2 \ln L = 285.74$				

Model 2

ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES				
Parameter	Estimate	Standard Error	Chi-Square	Pr > ChiSq
CHR	0.8051	0.3252	6.13	0.0133
AGE	0.0856	0.0193	19.63	< .0001
$-2 \ln L = 264.90$				

Model 3

ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES				
Parameter	Estimate	Standard Error	Chi-Square	Pr > ChiSq
CHR	1.0009	2.2556	0.20	0.6572
AGE	0.0874	0.0276	10.01	0.0016
CHRxAGE	-0.0030	0.0345	0.01	0.9301
$-2 \ln L = 264.89$				

- a. Assuming a regression model that contains the main effects of CHR and AGE, as well as the interaction effect between CHR and AGE, carry out a Wald test for significant interaction. (Specify the null hypothesis, the form of the test statistic, and its distribution under the null hypothesis.) What are the conclusions about interaction based on this test?
- b. For the same model considered in part (a), how would one carry out an LR test for significant interaction? (Specify the null hypothesis, the form of the test statistic, and its distribution under the null hypothesis.) What are the conclusions about interaction based on this test? How do these results compare with those in part (a)?

¹² The analysis described in Problem 2 is an example of a *survival analysis*, and the model considered is called the *Cox proportional hazards model*. See Kleinbaum and Klein (2012) for a detailed discussion of survival analysis.

- c. Assuming no interaction, carry out a Wald test for the significance of the CHR variable, controlling for AGE. (As in the preceding parts of this problem, state the null hypothesis, the form of the test statistic, and its distribution under the null hypothesis.) What are the conclusions based on this test?
- d. For the same no-interaction model considered in part (c), how would one carry out an LR test for significance of the CHR variable, controlling for AGE? (Specify the null hypothesis, the form of the test statistic, and its distribution under the null hypothesis.) Why can't one actually carry out this test given the output information provided earlier?
- e. For the no-interaction model considered in part (c), compute a 95% confidence interval for the coefficient of the CHR variable, controlling for AGE. What does the computed confidence interval say about the reliability of the point estimate of the effect of CHR in this model?
- f. What is the overall conclusion about the effect of CHR on survival time based on the computer results from this study?

References

- Kleinbaum, D. G., and Klein, M. 2012. *Survival Analysis: A Self-Learning Text*, Third Edition. New York: Springer.
- Schoenbach, V. J.; Kaplan, B. H.; Fredman, L.; and Kleinbaum, D. G. 1986. "Social Ties and Mortality in Evans County, Georgia." *American Journal of Epidemiology* 123(4): 577–91.

22

Logistic Regression Analysis

22.1 Preview

Logistic regression analysis is the most popular regression technique available for modeling dichotomous dependent variables. This chapter describes the logistic model form and several of its key features—particularly how an odds ratio can be estimated via its use. We also demonstrate how logistic regression may be applied using a real-life data set.

Maximum likelihood (ML) procedures are used to estimate the model parameters of a logistic model. Therefore, the general principles and inference-making procedures described in Chapter 21 on ML estimation directly carry over to the likelihood functions appropriate for logistic regression analysis. We will examine two alternative ML procedures for logistic regression—called unconditional and conditional—that involve different likelihood functions to be maximized. The latter (conditional) method is recommended when the amount of data available for analysis is not large relative to the number of parameters in the model, as is often the case when matching on potential confounders is employed when selecting subjects.

22.2 The Logistic Model

Logistic regression is a statistical modeling approach that can be used to describe the relationship of several predictor variables X_1, X_2, \dots, X_k to a *dichotomous* dependent variable Y , where Y is typically coded as 1 or 0 for its two possible categories. The logistic model describes the expected value of Y (i.e., $E(Y)$) in terms of the following “logistic” formula:

$$E(Y) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \sum_{j=1}^k \beta_j X_j\right)\right]}$$

For $(0, 1)$ random variables such as Y , it follows from basic statistical principles about expected values¹ that $E(Y)$ is equal to the probability $\text{pr}(Y = 1)$, so the formula for the logistic model can be written in a form that describes the probability of occurrence of one of the two possible outcomes of Y , as follows:

$$\text{pr}(Y = 1) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \sum_{j=1}^k \beta_j X_j\right)\right]} \quad (22.1)$$

The logistic model (22.1) is useful in many important practical situations where the response variable can take one of two possible values. For example, a study of the development of a particular disease in some human population could employ a logistic model to describe in probabilistic terms whether a given individual in the study group will ($Y = 1$) or will not ($Y = 0$) develop the disease in question during a fixed follow-up period of interest.

The first step in logistic regression analysis is to postulate (based on knowledge about, and experience with, the process under study) a mathematical model describing the mean of Y as a function of the X_j and the β_j . The model is then fit to the data by maximum likelihood, and eventually appropriate statistical inferences are made (after the model's adequacy of fit is verified, including consideration of relevant regression diagnostic indices).

The mathematical expression on the right side of the logistic model formula given by (22.1) is of the general mathematical form

$$f(z) = \frac{1}{1 + e^{-z}}$$

where $z = \beta_0 + \sum_{j=1}^k \beta_j X_j$. The function $f(z)$ is called the *logistic function*. This function is well suited to modeling a probability, since the values of $f(z)$ increase from 0 to 1 as z increases from $-\infty$ to $+\infty$. In epidemiologic studies, such a probability can be used to quantify an individual's risk of developing a disease. The logistic model, therefore, is set up to ensure that, whatever estimated value of risk we obtain, that value always lies between 0 and 1. This is not true for other possible models, which is why the logistic model is often used when a probability is to be estimated.

Another reason why the logistic model is popular relates to the general sigmoid shape of the logistic function (see Figure 22.1). A sigmoid shape is particularly appealing to

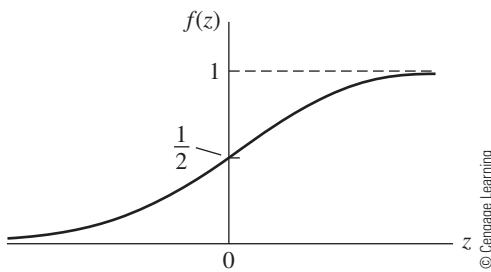


FIGURE 22.1 The logistic function $f(z) = \frac{1}{1 + e^{-z}}$

¹ For a $(0, 1)$ random variable Y , $E(Y) = [0 \times \text{pr}(Y = 0)] + [1 \times \text{pr}(Y = 1)] = \text{pr}(Y = 1)$.

epidemiologists if the variable z is viewed as representing an index that combines the contributions of several risk factors, so that $f(z)$ represents the risk for a given value of z . In this context, the risk is minimal for low z values, rises over a range of intermediate values of z , and remains close to 1 once z gets large enough. Epidemiologists believe that this sigmoid shape applies to a variety of disease processes.

22.3 Estimating the Odds Ratio Using Logistic Regression

As in any regression model, the regression coefficients β_j in the logistic model given by (22.1) provide important information about the relationships of the predictors in the model to the dichotomous dependent variable. For the logistic model, these coefficients are often used to estimate a parameter called the *odds ratio*.

The odds ratio (OR) is a widely used measure of effect in epidemiologic studies. By “measure of effect” we mean a measure that compares two or more groups with regard to the outcome (dependent) variable. To describe an odds ratio, we first define *odds* as the ratio of the probability that some event (e.g., developing lung cancer) will occur divided by the probability that the same event will not occur (e.g., not developing lung cancer). Thus, the odds for some event D are given by the formula

$$\text{Odds}(D) = \frac{\text{pr}(D)}{\text{pr}(\text{not } D)} = \frac{\text{pr}(D)}{1 - \text{pr}(D)}$$

For example, if $\text{pr}(D) = .25$, then

$$\text{Odds}(D) = \frac{.25}{1 - .25} = \frac{1}{3}$$

Odds of one-third can be interpreted to mean that the probability of event D occurring is 1/3 the probability of event D not occurring; equivalently, the odds are “3 to 1” that event D will not happen.

Any odds ratio, by definition, is a ratio of two odds; that is,

$$\text{OR}_{A \text{ vs. } B} = \frac{\text{Odds}(D_A)}{\text{Odds}(D_B)} = \frac{\text{pr}(D_A)}{1 - \text{pr}(D_A)} \Big/ \frac{\text{pr}(D_B)}{1 - \text{pr}(D_B)}$$

in which the subscripts A and B denote two groups of individuals being compared. For example, suppose that $A = S$ denotes a group of smokers and $B = NS$ denotes a group of nonsmokers; then D_S is the event that a smoker develops lung cancer, and D_{NS} is the event that a nonsmoker develops lung cancer. If $\text{pr}(D_S) = .0025$ and $\text{pr}(D_{NS}) = .0010$, the odds ratio that compares the odds of developing lung cancer for smokers with the odds of developing lung cancer for nonsmokers is given by

$$\text{OR}_{S \text{ vs. } NS} = \frac{.0025}{1 - .0025} \Big/ \frac{.0010}{1 - .0010} = 2.504$$

In other words, the odds of developing lung cancer for smokers are about 2.5 times the corresponding odds for nonsmokers, suggesting that smokers have roughly 2.5 times the risk of developing lung cancer than nonsmokers have.² An odds ratio of 1 would mean that the odds for the two groups are the same; that is, it would indicate that there is no effect of smoking on the development of lung cancer.

Where does logistic regression fit in here? To answer this question, we must consider an equivalent way to write the logistic regression model, called the *logit form* of the model. The “logit” is a transformation of the probability $\text{pr}(Y = 1)$, defined as the natural log of the odds of the event $D = \{Y = 1\}$. In other words,

$$\text{logit}[\text{pr}(Y = 1)] = \log_e[\text{odds}(Y = 1)] = \log_e\left[\frac{\text{pr}(Y = 1)}{1 - \text{pr}(Y = 1)}\right] \quad (22.2)$$

If we then substitute the logistic model formula (22.1) for $\text{pr}(Y = 1)$ into equation (22.2), it follows that

$$\text{logit}[\text{pr}(Y = 1)] = \beta_0 + \sum_{j=1}^k \beta_j X_j \quad (22.3)$$

Equation (22.3) is called the *logit form* of the model. The logit form is given by the linear function $(\beta_0 + \sum_{j=1}^k \beta_j X_j)$. For convenience, many authors describe the logistic model in its logit form given by (22.3) rather than in its original form defined by (22.1).

For example, if Y denotes lung cancer status (1 = yes, 0 = no) and there is only one (i.e., $k = 1$) predictor X_1 —say, smoking status (1 = smoker, 0 = nonsmoker)—then the logistic model (22.1) can be written equivalently in logit form (22.3) as

$$\text{logit}[\text{pr}(Y = 1)] = \beta_0 + \beta_1 X_1 = \beta_0 + \beta_1(\text{smoking status})$$

To obtain an expression for the odds ratio using a logistic model, we must compare the odds for two groups of individuals. For the preceding example involving one predictor, the two groups are smokers ($X_1 = 1$) and nonsmokers ($X_1 = 0$). Thus, for this example, the log odds for smokers and nonsmokers can be written as

$$\log_e \text{odds(smokers)} = \beta_0 + (\beta_1)(1) = \beta_0 + \beta_1$$

and

$$\log_e \text{odds(nonsmokers)} = \beta_0 + (\beta_1)(0) = \beta_0$$

respectively. It follows that the odds ratio comparing smokers to nonsmokers is given by

$$\text{OR}_{S \text{ vs. } NS} = \frac{\text{Odds(smokers)}}{\text{Odds(nonsmokers)}} = \frac{e^{(\beta_0 + \beta_1)}}{e^{\beta_0}} = e^{\beta_1}$$

² Note that the *risk ratio* of developing lung cancer (.0025/.0010) is exactly 2.5. This is consistent with the observation that, when an outcome is rare, the odds ratio approximates the risk ratio.

In other words, for the simple example involving one (0–1) predictor, the odds ratio comparing the two categories of the predictor is obtained by exponentiating the coefficient of the predictor in the logistic model.

Generally, when computing an odds ratio, we can define the two groups (or individuals) that are to be compared in terms of two different specifications of the set of predictors X_1, X_2, \dots, X_k . We do this by letting $\mathbf{X}_A = (X_{A1}, X_{A2}, \dots, X_{Ak})$ and $\mathbf{X}_B = (X_{B1}, X_{B2}, \dots, X_{Bk})$ denote the collection of X 's for groups (or individuals) A and B , respectively. For example, if $k = 3$, X_1 is smoking status (1 = yes, 0 = no), X_2 is age (continuous), and X_3 is race (1 = black, 0 = white), then \mathbf{X}_A and \mathbf{X}_B are two specifications of these three variables—say,

$$\mathbf{X}_A = (1, 45, 1) \text{ and } \mathbf{X}_B = (0, 45, 1)$$

Here \mathbf{X}_A denotes the group of 45-year-old black smokers, whereas \mathbf{X}_B denotes the group of 45-year-old black nonsmokers.

To obtain a general formula for the odds ratio, we must divide the odds for group (or individual) A by the odds for group (or individual) B and then appropriately employ the logit form of the logistic model given by (22.1) to obtain an expression involving the logistic model parameters. Using some algebra, the following result is obtained:

$$\text{OR}_{\mathbf{X}_A \text{ vs. } \mathbf{X}_B} = \frac{\text{Odds for } \mathbf{X}_A}{\text{Odds for } \mathbf{X}_B} = \frac{e^{(\beta_0 + \sum_{j=1}^k \beta_j X_{Aj})}}{e^{(\beta_0 + \sum_{j=1}^k \beta_j X_{Bj})}}$$

We can simplify the above expression further³ to obtain the following general formula for the odds ratio:

$$\text{OR}_{\mathbf{X}_A \text{ vs. } \mathbf{X}_B} = e^{\sum_{j=1}^k (X_{Aj} - X_{Bj})\beta_j} \quad (22.4)$$

The constant term β_0 in the logistic model (22.1) does not appear in the odds ratio expression (22.4), so the odds ratio depends only on the β_j coefficients and the corresponding differences in values of the X_j 's for groups (or individuals) A and B . Expression (22.4) describes a “population” odds ratio parameter because the β_j 's in this expression are themselves unknown population parameters. An estimate of this population odds ratio can be obtained by fitting the logistic model using ML estimation and substituting the ML estimates $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$, together with values of X_{Aj} and X_{Bj} , into the formula (22.4) to obtain a numerical value for the odds ratio.

For example, if \mathbf{X}_A and \mathbf{X}_B are two specifications of the three variables smoking status, age, and race, so that $\mathbf{X}_A = (1, 45, 1)$ and $\mathbf{X}_B = (0, 45, 1)$ as described earlier, then

$$\text{OR}_{\mathbf{X}_A \text{ vs. } \mathbf{X}_B} = e^{(1-0)\beta_1 + (45-45)\beta_2 + (1-1)\beta_3} = e^{\beta_1}$$

If the estimate of the β_1 coefficient from ML estimation turns out to be, say, $\hat{\beta}_1 = 1.32$, then the estimated odds ratio will be $e^{1.32} = 3.74$. Note that the logistic model in this example

³ For any two values a and b , it follows that $\frac{e^a}{e^b} = e^{(a-b)}$.

involves three variables, as follows (in logit form):

$$\text{logit}[\text{pr}(Y = 1)] = \beta_0 + \beta_1(\text{smoking}) + \beta_2(\text{age}) + \beta_3(\text{race})$$

So the value of the estimate $\hat{\beta}_1$ for this three-variable model will almost surely be numerically different from the value obtained for $\hat{\beta}_1$ in a model containing only smoking status.

In this latest example, the smoking status variable changes from 1 in group A to 0 in group B, whereas the other variables remain the same for each group—namely, age is 45 and race is 1. In general, whenever only one variable (e.g., smoking) changes, while the other variables are fixed, we say that the odds ratio comparing two categories of the changing variable (e.g., smokers versus nonsmokers) is an *adjusted odds ratio* that *controls* for the other variables (i.e., those that are fixed at specific values) in the model. The variable of interest—in this case, smoking status—is often referred to as the *exposure* (or *study*) variable; the other variables in the model are often called *control* (or *confounder*) variables. Thus, we have, as an important special case of the odds ratio expression (22.4), the following rule:

- An adjusted odds ratio can be obtained by exponentiating the coefficient of a (0–1) exposure variable in the logistic model (provided that there are no product [interaction] terms in the model involving the exposure variable).

The examples that we have considered so far have involved only *main effect variables* like smoking, age, and race; we have not considered product terms like “smoking \times age” or “smoking \times race,” nor have we considered exposure variables other than (0–1) variables. When the model contains product terms (like “smoking \times age”) or exposure variables that are not (0–1) variables, the preceding simple rule for obtaining an adjusted odds ratio does not work. In such instances, we must use the general formula given by (22.4).

For example, suppose, as before, that smoking is a (0–1) exposure variable and that age (continuous) and race (0–1) are control variables; but suppose now that the logistic model we want to fit is given (in logit form) as

$$\begin{aligned} \text{logit}[\text{pr}(Y = 1)] &= \beta_0 + \beta_1(\text{smoking}) + \beta_2(\text{age}) + \beta_3(\text{race}) \\ &\quad + \beta_4(\text{smoking} \times \text{age}) + \beta_5(\text{smoking} \times \text{race}) \end{aligned} \tag{22.5}$$

Then, to obtain an adjusted odds ratio for the effect of smoking adjusted for age and race, we need to specify two sets of values \mathbf{X}_A and \mathbf{X}_B for the collection of predictor variables defined by

$$\mathbf{X} = (\text{smoking}, \text{age}, \text{race}, \text{smoking} \times \text{age}, \text{smoking} \times \text{race})$$

If the two groups being compared are, as before, 45-year-old black smokers (i.e., group A) and 45-year-old black nonsmokers (i.e., group B), then \mathbf{X}_A and \mathbf{X}_B are given as

$$\mathbf{X}_A = (1, 45, 1, 1 \times 45, 1 \times 1) = (1, 45, 1, 45, 1)$$

and

$$\mathbf{X}_B = (0, 45, 1, 0 \times 45, 0 \times 1) = (0, 45, 1, 0, 0)$$

Applying the odds ratio formula (22.4) to model (22.5) with the given values for \mathbf{X}_A and \mathbf{X}_B , we obtain the following expression for the adjusted odds ratio:

$$\text{OR}_{\mathbf{X}_A \text{ vs. } \mathbf{X}_B} = e^{(1-0)\beta_1 + (45-45)\beta_2 + (1-1)\beta_3 + (45-0)\beta_4 + (1-0)\beta_5} = e^{\beta_1 + 45\beta_4 + \beta_5}$$

Since this odds ratio describes the effect of smoking adjusted for age = 45 and race = 1, we can alternatively denote the odds ratio as

$$\text{OR}_{(S \text{ vs. } NS | \text{age} = 45, \text{race} = 1)}$$

where the | sign is common mathematical notation for “given.” Thus, we have

$$\text{OR}_{(S \text{ vs. } NS | \text{age} = 45, \text{race} = 1)} = e^{\beta_1 + 45\beta_4 + \beta_5}$$

Rather than involving only β_1 , the preceding expression involves the three coefficients β_1 , β_4 , and β_5 , each of which is a coefficient of a variable in the model that contains the exposure variable (i.e., smoking) in some form. The coefficients β_4 and β_5 are included because they are coefficients of “interaction terms” in the model being fit. The odds ratio expression essentially says that the value of the odds ratio for the effect of smoking should vary, depending on the values of the variables age and race (which are components of the two interaction terms in model (22.5) and of the regression coefficients β_4 and β_5). Since the “fixed” value of age is 45 and the “fixed” value of race is 1, these two values multiply their corresponding regression coefficients in the odds ratio expression. In other words, the variables age and race in model (22.5) are *effect modifiers* (see Chapter 11) since the *effect* of the exposure variable smoking status, as quantified by the odds ratio $\text{OR}_{(S \text{ vs. } NS | \text{age, race})}$, changes (i.e., is modified), depending on the values of age and race.

To more generally represent how age and race are effect modifiers of smoking status in model (22.5), we let the values of age and race in \mathbf{X}_A and \mathbf{X}_B be fixed but unspecified, so we can write \mathbf{X}_A and \mathbf{X}_B as

$$\mathbf{X}_A = (1, \text{age}, \text{race}, 1 \times \text{age}, 1 \times \text{race}) = (1, \text{age}, \text{race}, \text{age}, \text{race})$$

and

$$\mathbf{X}_B = (0, \text{age}, \text{race}, 0 \times \text{age}, 0 \times \text{race}) = (0, \text{age}, \text{race}, 0, 0)$$

Then, substituting \mathbf{X}_A and \mathbf{X}_B into formula (22.4) for the odds ratio, we obtain

$$\begin{aligned}\text{OR}_{(S \text{ vs. } NS | \text{age, race})} &= e^{(1-0)\beta_1 + (\text{age} - \text{age})\beta_2 + (\text{race} - \text{race})\beta_3 + (\text{age} - 0)\beta_4 + (\text{race} - 0)\beta_5} \\ &= e^{\beta_1 + \beta_4(\text{age}) + \beta_5(\text{race})}\end{aligned}$$

This expression again shows that the value of the odds ratio for the effect of smoking varies, depending on the values of the variables age and race (which are components of the two interaction variables in the model) and of their coefficients β_4 and β_5 . So, for example, if we have age = 45 and race = 1, as in our earlier example, the adjusted odds ratio is given by

$$\text{OR}_{(S \text{ vs. } NS | \text{age} = 45, \text{race} = 1)} = e^{\beta_1 + 45\beta_4 + \beta_5}$$

Similarly, if we have age = 35 and race = 0, the adjusted odds ratio is given by

$$\text{OR}_{(S \text{ vs. } NS | \text{age} = 35, \text{race} = 0)} = e^{\beta_1 + 35\beta_4}$$

And if age = 20 and race = 1, then

$$\text{OR}_{(S \text{ vs. } NS | \text{age} = 20, \text{race} = 1)} = e^{\beta_1 + 20\beta_4 + \beta_5}$$

The preceding examples illustrate another general rule that describes an important special case of the general odds ratio formula (22.4):

- For a logistic model that contains a (0–1) exposure variable together with interaction terms that are products of the exposure variable with other control variables, the adjusted odds ratio is obtained by exponentiating a linear function of the regression coefficients involving the exposure alone and the product terms in the model involving exposure. Moreover, the numerical value of the adjusted odds ratio will vary, depending on the values of the control variables (i.e., effect modifiers) that are components of the product terms involving exposure.

We now consider one other important special case of the general odds ratio formula (22.4). Suppose that the exposure variable is a continuous variable, like systolic blood pressure (SBP), so that, for example, our model contains SBP and the control variables age and race; that is, the logit form of the no-interaction model is given by

$$\text{logit}[\text{pr}(Y = 1)] = \beta_0 + \beta_1(\text{SBP}) + \beta_2(\text{age}) + \beta_3(\text{race}) \quad (22.6)$$

To obtain an adjusted odds ratio for the effect of SBP, controlling for age and race, we must specify two values of the exposure variable (SBP) to be compared. Two specified values are needed, even when the exposure variable (like SBP) is continuous, because an odds ratio by definition compares the odds for two groups (or individuals). For example, if the two values of SBP are 160 and 120 and age and race are considered fixed, the general odds ratio expression (22.4) simplifies to

$$\text{OR}_{(\text{SBP} = 160 \text{ vs. } \text{SBP} = 120 | \text{age, race})} = e^{(160 - 120)\beta_1 + (\text{age} - \text{age})\beta_2 + (\text{race} - \text{race})\beta_3} = e^{40\beta_1}$$

More generally, if we specify the two values of SBP to be SBP_1 and SBP_0 , then the odds ratio is

$$\text{OR}_{(\text{SBP}_1 \text{ vs. } \text{SBP}_0 | \text{age, race})} = e^{(\text{SBP}_1 - \text{SBP}_0)\beta_1}$$

This odds ratio expression simplifies to e^{β_1} when the difference ($\text{SBP}_1 - \text{SBP}_0$) equals 1. Thus, exponentiating the coefficient of the exposure variable gives an odds ratio for comparing any two groups that differ by one unit of SBP. A one-unit difference in SBP is rarely of interest, however. A more typical choice of SBP values to be compared would be clinically meaningful different values of blood pressure, such as SBP values of 160 and 120. One possible strategy for choosing values of SBP to compare is to categorize the distribution of SBP in a data set into clinically different categories—say, quintiles. Then, using the mean or median SBP in each quintile, we could compute odds ratios for all possible pairs of mean or median SBP values. We would then obtain a table of odds ratios to consider in assessing the relationship of SBP to the outcome variable. (For a more thorough treatment of important special cases for computing the odds ratio, see Chapter 3 of Kleinbaum and Klein 2010).

ML estimates of the regression coefficients are typically obtained by using standard computer packages for logistic regression. These estimates can then be used in appropriate odds ratio formulas based on the general expression (22.4) to obtain numerical values for adjusted odds ratios. Since these are point estimates, researchers typically carry out statistical inferences about the odds ratios that are being estimated. For example, if the adjusted odds ratio is given by the simple expression e^{β_1} , involving a single coefficient, then the null hypothesis that this odds ratio equals 1 can be stated equivalently as $H_0: \beta_1 = 0$, since, under H_0 , $e^{\beta_1} = e^0 = 1$. The test for this null hypothesis can be carried out by using either the Wald test or the likelihood ratio (LR) test described in Chapter 21 on ML estimation methods. A confidence interval for the adjusted odds ratio can be obtained by first calculating a confidence interval for β_1 , as described in Section 21.3, and then exponentiating the lower and upper limits. For this simple situation, an appropriate large-sample $100(1 - \alpha)\%$ confidence interval for $\exp(\beta_1)$ is given by the formula

$$\exp(\hat{\beta}_1 \pm Z_{1-\alpha/2} S_{\hat{\beta}_1})$$

where $\hat{\beta}_1$ is the ML estimate of β_1 , $S_{\hat{\beta}_1}$ is the estimated standard error of $\hat{\beta}_1$, and $(1 - \alpha)$ is the confidence level.

More detailed discussion of the properties and applications of logistic regression may be found in several other textbooks, including Kleinbaum and Klein (2010), Hosmer and Lemeshow (1989), Collett (1991), and Kleinbaum, Kupper, and Morgenstern (1982). In particular, a general analysis strategy for selecting the variables to retain in a logistic model is described in Kleinbaum and Klein (2010), Chapters 6–8, and in Kleinbaum et al. (1982), Chapters 21–24. The goal of this strategy is similar to that of the procedure outlined in Section 16.9, which seeks to obtain a *valid* estimate (in the context of the analysis of epidemiologic research data) of the relationship between a specified exposure variable and a particular disease variable, while controlling for other covariates that, if not correctly taken into account, can lead to an incorrect assessment of the strength of the exposure–disease relationship of interest.

22.4 A Numerical Example of Logistic Regression

Dengue fever is an acute infectious disease caused by a virus transmitted by several species of the *Aedes* mosquito. A retrospective survey motivated by an epidemic of dengue fever was carried out in three Mexican cities in 1984 (Dantes et al. 1988). In this section, we review the analyses of a subset of the data from this survey obtained via a two-stage stratified random sample of 196 persons from the city of Puerto Vallarta, 57 of whom were determined to be suffering from dengue fever. The goal of the analyses was to identify risk factors associated with having the disease, especially the effect of the absence of mosquito netting around a subject's bed as a determinant of the disease. The following variables were recorded for each subject:

1: Subject ID

2: Dengue fever status (DENGUE): 1 = yes, 2 = no

3: AGE in years

4: Use of mosquito netting (MOSNET): 0 = yes, 1 = no

5: Geographical sector in which the subject lived (SECTOR): 1, 2, 3, 4, or 5

The variable SECTOR was treated as a categorical variable in the logistic regression analysis, so four dummy variables had to be created to distinguish the five geographical sectors. These variables were defined so that sector 5 was the referent group, as follows:

$$\text{SECTOR1} = \begin{cases} 1 & \text{if sector 1} \\ 0 & \text{if other} \end{cases}$$

$$\text{SECTOR2} = \begin{cases} 1 & \text{if sector 2} \\ 0 & \text{if other} \end{cases}$$

$$\text{SECTOR3} = \begin{cases} 1 & \text{if sector 3} \\ 0 & \text{if other} \end{cases}$$

$$\text{SECTOR4} = \begin{cases} 1 & \text{if sector 4} \\ 0 & \text{if other} \end{cases}$$

The following block of edited computer output comes from using SAS's LOGISTIC procedure to fit a logistic regression model that regresses the dichotomous outcome variable DENGUE on the predictors MOSNET, AGE, and SECTOR j for $j = 1, 2, 3, 4$. The logit form of the model being fit is given as

$$\begin{aligned} \text{logit}[\text{pr}(Y = 1)] = & \beta_0 + \beta_1(\text{AGE}) + \beta_2(\text{MOSNET}) + \beta_3(\text{SECTOR1}) \\ & + \beta_4(\text{SECTOR2}) + \beta_5(\text{SECTOR3}) + \beta_6(\text{SECTOR4}) \end{aligned} \quad (22.7)$$

where Y denotes the dependent variable DENGUE.

Edited SAS Output (PROC LOGISTIC) for Full Model: Dengue Data

The LOGISTIC Procedure

RESPONSE PROFILE		
Ordered Value	DENGUE	Total Frequency
1	1	57
2	2	139

MODEL FIT STATISTICS			
Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	238.329	217.706	.
SC	241.607	240.653	.
-2 Log L	236.329	203.706	32.623 with 6 DF ($p = 0.0001$)
Score	.	.	28.775 with 6 DF ($p = 0.0001$)

(continued)

ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES							
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Odds Ratio
Intercept	1	-1.9001	1.3254	2.0551	0.1517		0.150
AGE	1	0.0243	0.00906	7.1778	0.0074	0.252890	1.025
MOSNET	1	0.3335	1.2718	0.0688	0.7931	0.034212	1.396
SECTOR1	1	-2.2200	1.0723	4.2861	0.0384	-0.441811	0.109
SECTOR2	1	-0.6589	0.5536	1.4164	0.2340	-0.142513	0.517
SECTOR3	1	0.8121	0.4750	2.9235	0.0873	0.173824	2.253
SECTOR4	1	0.5310	0.4502	1.3911	0.2382	0.121456	1.701

Using the given computer output, we now focus on the information provided under the heading “Analysis of Maximum Likelihood Estimates.” From this information, we can see that the ML-based estimated regression coefficients obtained for the fitted model are

$$\hat{\beta}_0 = -1.9001, \hat{\beta}_1 = 0.0243, \hat{\beta}_2 = 0.3335, \hat{\beta}_3 = -2.2200,$$

$$\hat{\beta}_4 = -0.6589, \hat{\beta}_5 = 0.8121, \hat{\beta}_6 = 0.5310$$

so the fitted model is given (in logit form) by

$$\begin{aligned} \text{logit}[\hat{\text{pr}}(Y = 1)] &= -1.9001 + 0.0243(\text{AGE}) + 0.3335(\text{MOSNET}) \\ &\quad - 2.2200(\text{SECTOR1}) - 0.6589(\text{SECTOR2}) \\ &\quad + 0.8121(\text{SECTOR3}) + 0.5310(\text{SECTOR4}) \end{aligned}$$

Based on this fitted model and the information provided in the computer output, we can compute the estimated odds ratio for contracting dengue fever for persons who did not use mosquito netting relative to persons who did use mosquito netting, controlling for age and sector. We do this using the previously stated rule for adjusted odds ratios for (0–1) variables (when there is no interaction) by exponentiating the estimated coefficient ($\hat{\beta}_2$) of the MOSNET variable in the fitted model. We then obtain the following value for the adjusted odds ratio:

$$\widehat{\text{OR}}_{(\text{MOSNET} = 1 \text{ vs. } \text{MOSNET} = 0 | \text{age, sector})} = e^{0.3335} = 1.396$$

A 95% confidence interval for $e^{\hat{\beta}_2}$ can be obtained by computing

$$\exp[0.3335 \pm 1.96(1.2718)]$$

where 1.2718 is the estimated standard error of the estimator $\hat{\beta}_2$ of β_2 in model (22.7). The resulting lower and upper confidence limits are 0.115 and 16.89, respectively. The Wald (chi-square) statistic for testing the null hypothesis that the adjusted odds ratio $e^{\hat{\beta}_2}$ is equal to 1 (or, equivalently, that the coefficient β_2 of the MOSNET variable equals 0) is shown in the output to be 0.0688, with a P -value equal to .7931.

From these results, it can be concluded that the estimated odds of contracting dengue for a person who did not use mosquito netting are about 1.4 times (or about 40% higher

than) that of a person who did use mosquito netting. The Wald statistic is not statistically significant, however, and the 95% confidence interval is very wide (and includes the null value of 1). Thus, there is no statistical evidence in these data that the nonuse of mosquito netting significantly increases the probability (or risk) of contracting dengue fever.

The preceding computer output contains, under the heading “Model Fit Statistics,” the log likelihood statistic ($-2 \log \hat{L}$) of 203.706 for the fitted model (in the column labeled “Intercept and Covariates”). To compute the LR test for the null hypothesis $H_0: \beta_2 = 0$, we must compare this value of $-2 \log \hat{L}$ to the corresponding value of $-2 \log \hat{L}$ for the “reduced” model, which is obtained under the null hypothesis by dropping the exposure variable (MOSNET) from the “full” model given by (22.7). The reduced model is written in logit form as

$$\text{logit}[\text{pr}(Y = 1)] = \beta_0 + \beta_1(\text{AGE}) + \beta_3(\text{SECTOR1}) + \beta_4(\text{SECTOR2}) \\ + \beta_5(\text{SECTOR3}) + \beta_6(\text{SECTOR4})$$

The following block of edited computer output for the reduced model indicates that the log likelihood statistic for the fitted model has changed slightly to 203.778. The LR statistic for comparing these two models is given by the difference:

$$\text{LR} = -2 \log \hat{L}_R - (-2 \log \hat{L}_F) = 203.778 - 203.706 = 0.072$$

Because the null hypothesis $H_0: \beta_2 = 0$ involves only one parameter, the preceding LR statistic is distributed as approximately a chi-square variable with 1 degree of freedom under the null hypothesis. The test is nonsignificant, which agrees with the result from the Wald test described earlier, thereby supporting the previous conclusion that using mosquito netting did not significantly affect a person’s probability of contracting dengue fever.

Edited SAS Output (PROC LOGISTIC) for Reduced Model: Dengue Data

The LOGISTIC Procedure

RESPONSE PROFILE		
Ordered Value	DENGUE	Total Frequency
1	1	57
2	2	139

MODEL FIT STATISTICS			
Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	238.329	215.778	.
SC	241.607	235.447	.
-2 Log L	236.329	203.778	32.551 with 5 DF (p = 0.0001)
Score	.	.	28.766 with 5 DF (p = 0.0001)

(continued)

ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES							
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Odds Ratio
Intercept	1	-1.5717	0.4229	13.8144	0.0002		0.208
AGE	1	0.0240	0.00901	7.1129	0.0077	0.250447	1.024
SECTOR1	1	-2.2333	1.0715	4.3444	0.0371	-0.444446	0.107
SECTOR2	1	-0.6650	0.5531	1.4455	0.2293	-0.143843	0.514
SECTOR3	1	0.8224	0.4735	3.0169	0.0824	0.176013	2.276
SECTOR4	1	0.5439	0.4478	1.4755	0.2245	0.124412	1.723

Returning to model (22.7), involving the predictors AGE, MOSNET, and SECTOR1 through SECTOR4, we can also evaluate the effect of predictors other than MOSNET on the probability of getting dengue fever. For example, let us examine the effect of AGE, controlling for MOSNET and the four SECTOR dummy variables. From the first block of computer output, we find that the Wald chi-square statistic corresponding to the AGE variable is 7.1778. The null hypothesis being tested here is $H_0: \beta_1 = 0$, where β_1 is the coefficient of the AGE variable in model (22.7). The P -value for this test, as shown in the output, is .0074, so we can reject the null hypothesis at less than the .01 significance level and conclude that the AGE variable is a significant predictor of dengue fever status in model (22.7). In particular, the positive estimated coefficient for AGE suggests that the risk of dengue fever increases with increasing age.

The LR test of the same null hypothesis (i.e., $H_0: \beta_1 = 0$) would require subtracting the log likelihood statistic for model (22.7) from the log likelihood statistic for the “reduced” model, which is obtained by dropping the AGE variable from model (22.7). This reduced model has not been presented in any of the previously displayed blocks of computer output, so an additional computer run would be required to obtain the log likelihood statistic for this reduced model.⁴ The resulting LR test also gives a significant result, agreeing with (though not numerically identical to) the significant result obtained from the Wald test.

Since the Wald and LR tests just illustrated were assessing the statistical significance of the effect of a continuous variable (AGE), their specific purpose here is to assess whether the effect of the AGE variable on the response variable DENGUE can be described by a significant linear relationship between AGE and the log odds of dengue fever in model (22.7), controlling for the other variables in that model. That is, when no functions of AGE (e.g., AGE^2 or $AGE \times MOSNET$) other than AGE itself are in the model, then we are testing whether a linear effect in AGE is more plausible than no effect of AGE (i.e., we are testing $H_0: \beta_1 = 0$). Such a test is typically referred to as a (*linear*) *trend test*. Thus, using the results from the first block of computer output considered previously, we conclude that a linear trend test for AGE in model (22.7) is significant.

Since the AGE variable is continuous in model (22.7), the quantity $e^{\hat{\beta}_1}$, where $\hat{\beta}_1 = 0.0243$ is the estimated coefficient of the AGE variable, describes the odds ratio comparing two persons whose age differs by only one year. As mentioned earlier, if the exposure

⁴ Under SAS's LOGISTIC procedure, the log likelihood statistic for the reduced model is 211.143.

variable of interest is continuous, a one-unit difference in the variable is rarely of interest. Instead, we need to compare meaningfully (e.g., clinically) different categories of the continuous variable. For example, we might consider two persons whose ages differ by five years—say, a 35-year-old person and a 30-year-old person—as having a meaningful difference in age. Using the general odds ratio expression (22.4), we can compute the estimated odds ratio for two persons with a five-year age difference, controlling for mosquito netting and sector status, as follows:

$$\begin{aligned}\widehat{\text{OR}}_{(\text{AGE}_1 - \text{AGE}_0 = 5 | \text{mosnet, sector})} \\ &= e^{[5\hat{\beta}_1 + (\text{mosnet} - \text{mosnet})\hat{\beta}_2 + (\text{sector1} - \text{sector1})\hat{\beta}_3 + \dots + (\text{sector4} - \text{sector4})\hat{\beta}_6]} \\ &= e^{5\hat{\beta}_1} = e^{5(0.0243)} = 1.13\end{aligned}$$

Since we are controlling for mosquito netting and sector status, we assume that the MOS-NET variable and the four sector variables (SECTOR1 through SECTOR4) have the same values for the two persons being compared. The coefficients of these five variables then drop out of the odds ratio expression, and we have only the quantity $5\hat{\beta}_1$ to exponentiate. The estimated adjusted odds ratio is, therefore, given by $\exp(5\hat{\beta}_1)$, rather than by $\exp(\hat{\beta}_1)$, since we are considering the effect of a five-year, rather than a one-year, difference in age. Thus, for two persons whose age differs by five years, the adjusted odds ratio that controls for use of mosquito netting and sector status equals 1.13, which is quite close to the null value of 1. To obtain a 95% confidence interval for the adjusted odds ratio $e^{5\hat{\beta}_1}$, we make the following calculation:

$$95\% \text{ CI for } e^{5\hat{\beta}_1}: \exp[5\hat{\beta}_1 \pm 1.96(5S_{\hat{\beta}_1})] = \exp[5(0.0243) \pm 1.96(5)(0.0091)]$$

which yields 95% confidence limits of 1.03 and 1.23. Since the confidence limits do not include the null value of 1, the effect of a five-year difference in age is statistically significant, though small (i.e., the point estimate of the odds ratio is 1.13).

If, instead of a 5-year difference in age, we consider the effect of a 10-year difference in age, the formulas for the estimated adjusted odds ratio and 95% confidence interval become

$$\exp[10\hat{\beta}_1]$$

and

$$\exp[10\hat{\beta}_1 \pm 1.96(10S_{\hat{\beta}_1})]$$

respectively. The corresponding computed values are 1.28 for the estimated adjusted odds ratio and 1.07 and 1.52 for the confidence limits.

Table 22.1 presents point and interval estimates of the adjusted odds ratio for age differences from 5 to 40 years in 5-year increments. The estimated adjusted odds ratio for the effect of age increases from 1.13 for an age difference of 5 years to 2.64 for an age difference of 40 years. The width of the 95% confidence interval also increases as the age difference increases.

We now consider one more illustration of numerical calculations using the dengue fever data. This time we consider a model involving an interaction term. The following block of edited computer output applies to a model that contains the product term

TABLE 22.1 Point and interval estimates of the adjusted odds ratio for the effect of age based on the fit of model (22.7) from the dengue fever study

AGE ₁ – AGE ₀	$\widehat{OR}_{(AGE_1 - AGE_0 mosnet, sector)} = \exp[(AGE_1 - AGE_0)\beta_1]$	95% Confidence Limits for $OR_{(AGE_1 - AGE_0 mosnet, sector)}$
5	1.13	(1.03, 1.23)
10	1.28	(1.07, 1.52)
15	1.44	(1.10, 1.88)
20	1.63	(1.14, 2.32)
25	1.84	(1.18, 2.87)
30	2.07	(1.21, 3.54)
35	2.34	(1.25, 4.37)
40	2.64	(1.30, 5.39)

© Cengage Learning

MOSNET_AGE = MOSNET × AGE in addition to the predictors AGE, MOSNET, and SECTOR1 through SECTOR4. This “interaction” model is written in logit form as

$$\begin{aligned} \text{logit}[pr(Y = 1)] &= \beta_0 + \beta_1(\text{AGE}) + \beta_2(\text{MOSNET}) + \beta_3(\text{SECTOR1}) \\ &\quad + \beta_4(\text{SECTOR2}) + \beta_5(\text{SECTOR3}) + \beta_6(\text{SECTOR4}) \\ &\quad + \beta_7(\text{MOSNET_AGE}) \end{aligned} \quad (22.8)$$

Edited SAS Output (PROC LOGISTIC) for Interaction Model: Dengue Data

The LOGISTIC Procedure

RESPONSE PROFILE		
Ordered Value	DENGUE	Total Frequency
1	1	57
2	2	139

MODEL FIT STATISTICS			
Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	238.329	218.995	.
SC	241.607	245.220	.
-2 Log L	236.329	202.995	33.334 with 7 DF (p = 0.0001)
Score	.	.	29.492 with 7 DF (p = 0.0001)

(continued)

ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES							
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Odds Ratio
Intercept	1	-0.8080	1.6311	0.2454	0.6203		0.446
AGE	1	-0.00434	0.0362	0.0143	0.9048	-0.045186	0.996
MOSNET	1	-0.8043	1.6433	0.2396	0.6245	-0.082505	0.447
SECTOR1	1	-2.2929	1.0804	4.5042	0.0338	-0.456309	0.101
SECTOR2	1	-0.6813	0.5541	1.5118	0.2189	-0.147362	0.506
SECTOR3	1	0.8153	0.4756	2.9388	0.0865	0.174497	2.260
SECTOR4	1	0.5115	0.4515	1.2830	0.2573	0.116992	1.668
MOSNET_AGE	1	0.0306	0.0374	0.6689	0.4134	0.316912	1.031

Suppose that we wish to use the information in the computer output to compute the estimated odds ratio for contracting dengue fever for persons who did not use mosquito netting relative to persons who did use mosquito netting, controlling for age and sector status. We previously considered this question using the no-interaction model (22.7), but now we are using model (22.8) instead. We again use the general odds ratio formula (22.4) to compute the adjusted odds ratio, but this time we must take into account the coefficient of the product term MOSNET_AGE. Nevertheless, since the exposure variable of interest (MOSNET) is a (0–1) variable, the general odds ratio formula (22.4) simplifies to the previously stated (see Section 22.3) rule for calculating an adjusted odds ratio when the logistic model contains interaction terms: the adjusted odds ratio is obtained by exponentiating a linear function of those regression coefficients involving the exposure alone and those product terms in the model involving exposure. Thus, for interaction model (22.8), we compute the estimated adjusted odds ratio for the effect of mosquito net use, controlling for age and sector status, as follows:

$$\widehat{OR}_{(MOSNET=1 \text{ vs. } MOSNET=0 | \text{age, sector})} = \exp [\hat{\beta}_2 + \hat{\beta}_7(\text{AGE})] \\ = \exp [-0.8043 + 0.0306(\text{AGE})]$$

where $\hat{\beta}_2 = -0.8043$ is the estimated coefficient of the MOSNET variable and $\hat{\beta}_7 = 0.0306$ is the estimated coefficient of the MOSNET_AGE variable obtained from the last batch of computer output. The adjusted odds ratio formula says that the value of the odds ratio depends on the value we specify for AGE, which is exactly what we mean when we assume (using model 22.8) that AGE is an effect modifier of the relationship between mosquito net use status and risk of contracting dengue fever. Table 22.2 shows computed values for this estimated adjusted odds ratio for various specifications of the effect modifier AGE.

Table 22.2 indicates that, based on the interaction model (22.8), the adjusted odds ratio for the effect of mosquito net usage status varies from values below the null value of 1 when AGE is 25 or less to values increasingly above 1 (though below 3) when AGE is over 30. These results suggest, for example, that a person of age 20 who does not use a mosquito net (i.e., MOSNET = 1) is 0.83 times as likely to get dengue fever as a person of the same age who uses a mosquito net; in contrast, a person of age 40 who does not use a mosquito net is 1.52 times as likely to get dengue fever as a person of the same age who uses a mosquito net.

TABLE 22.2 Effect of age on the value of the estimated adjusted odds ratio

AGE	$\exp[-0.8043 + 0.0306(\text{AGE})]$
10	0.61
15	0.71
20	0.83
25	0.96
26.28	1.00
30	1.12
35	1.31
40	1.52
45	1.77
50	2.07
55	2.41
60	2.81

© Cengage Learning

We can obtain confidence intervals using any of the preceding estimated adjusted odds ratios by first computing a confidence interval for the general linear function $\beta_2 + \beta_7(\text{AGE})$, which is the log of the (population) adjusted odds ratio $\exp[\beta_2 + \beta_7(\text{AGE})]$, and then exponentiating the lower and upper limits of this confidence interval. For a 95% confidence interval, we need to use the following formula:

$$\exp(\hat{L} \pm 1.96S_{\hat{L}})$$

where $\hat{L} = \hat{\beta}_2 + \hat{\beta}_7(\text{AGE})$ and $S_{\hat{L}} = \sqrt{\widehat{\text{Var}}(\hat{L})}$ and where $\sqrt{\widehat{\text{Var}}(\hat{L})}$ is computed using the formula for the variance of a linear combination of random variables, which for this example turns out to be

$$\widehat{\text{Var}}(\hat{L}) = \widehat{\text{Var}}(\hat{\beta}_2) + (\text{AGE})^2 \widehat{\text{Var}}(\hat{\beta}_7) + 2(\text{AGE}) \widehat{\text{Cov}}(\hat{\beta}_2, \hat{\beta}_7)$$

The numerical values of the estimated variances and the covariances involving the $\hat{\beta}_j$'s are typically printed out as an option by the computer package used. For model (22.8), the numerical values needed to utilize the preceding variance formula are

$$\widehat{\text{Var}}(\hat{\beta}_2) = 2.7004, \quad \widehat{\text{Var}}(\hat{\beta}_7) = 0.001399, \quad \text{and} \quad \widehat{\text{Cov}}(\hat{\beta}_2, \hat{\beta}_7) = -0.0435$$

If, for example, we let AGE = 40, then

$$\hat{L} = -0.8043 + 0.0306(40) = 0.4197$$

$$\widehat{\text{Var}}(\hat{L}) = 2.7004 + (40)^2(0.001399) + 2(40)(-0.0435) = 1.4588$$

and

$$S_{\hat{L}} = \sqrt{\text{Var}(\hat{L})} = 1.2078$$

The 95% confidence interval for the adjusted odds ratio $e^{\beta_2 + (40)\beta_7}$ is then given by

$$\exp[0.4197 \pm (1.96)(1.2078)]$$

which yields lower and upper confidence limits of 0.14 and 16.23, respectively. This confidence interval is extremely wide, which implies that the point estimate $e^{0.4197} = 1.52$ when age is 40 is very unreliable. Furthermore, since the confidence interval includes the null value of 1, a corresponding (Wald or LR) test of significance for the adjusted odds ratio will not be significant at the 5% level. Computations of 95% intervals for values of age other than 40 also yield very wide intervals and nonsignificant findings.

Finally, we may wish to compare model (22.8), which contains the interaction term MOSNET_AGE, with the no-interaction model (22.7). We can do this by using an LR test that takes the following form:

$$\text{LR} = -2 \log \hat{L}_R - (-2 \log \hat{L}_F) = 203.706 - 202.995 = 0.711$$

The full model (F) in this case is model (22.8), and the reduced model (R) is model (22.7). The null hypothesis is $H_0: \beta_7 = 0$, which involves only one parameter, so that the LR statistic is approximately chi-square with 1 d.f. under H_0 . The test statistic value of 0.711 is non-significant, indicating that the no-interaction model (22.7) is preferable to the interaction model (22.8). In other words, the estimated odds ratio for the effect of MOSNET adjusted for age and sector status is best expressed by the single value $\exp(0.3335) = 1.396$ based on model (22.7), rather than by the expression $\exp(\hat{L})$, where $\hat{L} = \hat{\beta}_2 + \hat{\beta}_7(\text{AGE})$, based on model (22.8).

22.5 Theoretical Considerations

In this section, we consider the form of the likelihood function to be maximized for a logistic regression model. In particular, we will distinguish between two alternative ML procedures for estimation, called *unconditional* and *conditional*, that involve different likelihood functions. To describe these two procedures, we must first discuss the distributional properties of the dependent (i.e., outcome) variable underlying the logistic model.

For logistic regression, the basic dependent random variable of interest is a dichotomous variable Y taking the value 1 with probability θ and the value 0 with probability $(1 - \theta)$. Such a random variable is called *Bernoulli* (or point-binomial) and has the simple discrete probability distribution

$$\text{pr}(Y; \theta) = \theta^Y(1 - \theta)^{1-Y} \quad Y = 0, 1 \tag{22.9}$$

The name *point-binomial* arises because (22.9) is a special case of the binomial distribution ${}_nC_Y \theta^Y(1 - \theta)^{n-Y}$ when $n = 1$, where ${}_nC_Y$ denotes the number of combinations of n distinct objects selected Y -at-a-time.

In general, for a study sample of n subjects, suppose (for $i = 1, 2, \dots, n$) that Y_i denotes the Bernoulli random variable for the i th subject, having the Bernoulli distribution

$$\text{pr}(Y_i; \theta_i) = \theta_i^{Y_i} (1 - \theta_i)^{1 - Y_i} \quad Y_i = 0, 1 \quad (22.10)$$

For example, θ_i could represent the probability that individual i in a random sample of n individuals from some population will develop some particular disease during the follow-up period in question.

22.5.1 Unconditional ML Estimation

Given that Y_1, Y_2, \dots, Y_n are mutually independent, the likelihood function based on (22.10) is obtained as the product of the marginal distributions for the Y_i 's—namely,

$$L(\mathbf{Y}; \boldsymbol{\theta}) = \prod_{i=1}^n \text{pr}(Y_i; \theta_i) = \prod_{i=1}^n [\theta_i^{Y_i} (1 - \theta_i)^{1 - Y_i}] \quad (22.11)$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_n)$. Now suppose, without loss of generality, that the first n_1 out of the n individuals in our random sample actually develop the disease in question (so that $Y_1 = Y_2 = \dots = Y_{n_1} = 1$) and that the remaining $(n - n_1)$ individuals do not (so that $Y_{n_1+1} = Y_{n_1+2} = \dots = Y_n = 0$). Given this set of observed outcomes, the likelihood expression (22.11) takes the specific form

$$L(\mathbf{Y}; \boldsymbol{\theta}) = \left(\prod_{i=1}^{n_1} \theta_i \right) \left[\prod_{i=n_1+1}^n (1 - \theta_i) \right] \quad (22.12)$$

We can work with expression (22.12) to write the likelihood function in terms of the regression coefficients β_j in the logistic model. We let $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{ik})$ denote the set of values of the k predictors X_1, X_2, \dots, X_k specific to individual i . Then the logistic model assumes that the relationship between θ_i and the X_{ij} 's is of the specific form

$$\theta_i = \frac{1}{1 + \exp\left[-\left(\beta_0 + \sum_{j=1}^k \beta_j X_{ij}\right)\right]} \quad i = 1, 2, \dots, n \quad (22.13)$$

where $\beta_j, j = 0, 1, \dots, k$, are unknown regression coefficients that must be estimated. (The right side of (22.13) has the same form as the right side of the logistic model form (22.1), where the X_{ij} in (22.13) has been substituted for X_j in (22.1).)

Now, if we replace θ_i in the likelihood (22.12) with the logistic function expression (22.13), we obtain the so-called *unconditional likelihood function* characterizing standard

logistic regression analysis—namely,

$$L(\mathbf{Y}; \boldsymbol{\beta}) = \prod_{i=1}^{n_1} \left\{ 1 + \exp \left[-\left(\beta_0 + \sum_{j=1}^k \beta_j X_{ij} \right) \right] \right\}^{-1} \\ \times \prod_{i=n_1+1}^n \left(\exp \left[-\left(\beta_0 + \sum_{j=1}^k \beta_j X_{ij} \right) \right] \left\{ 1 + \exp \left[-\left(\beta_0 + \sum_{j=1}^k \beta_j X_{ij} \right) \right] \right\}^{-1} \right)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)$. The term *unconditional likelihood* refers to the unconditional probability of obtaining the particular set of data under consideration. More specifically, the unconditional likelihood function is the joint probability distribution for discrete data or the joint density function for continuous data. A *conditional likelihood*, which we discuss in the next subsection, gives the conditional probability of obtaining the data configuration *actually observed* given all possible configurations (i.e., permutations) of the observed data values.

With a little algebraic manipulation, we can verify that an equivalent expression for the preceding likelihood function is⁵

$$L(\mathbf{Y}; \boldsymbol{\beta}) = \frac{\prod_{i=1}^{n_1} \exp \left(\beta_0 + \sum_{j=1}^k \beta_j X_{ij} \right)}{\prod_{i=1}^n \left[1 + \exp \left(\beta_0 + \sum_{j=1}^k \beta_j X_{ij} \right) \right]} \quad (22.14)$$

Because (22.14) is a complex nonlinear function of the elements of $\boldsymbol{\beta}$, maximizing (22.14) to find the ML estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ must involve using appropriate computer algorithms. Such programs (e.g., SAS's LOGISTIC procedure) also produce the maximized likelihood value $L(\mathbf{Y}; \hat{\boldsymbol{\beta}})$ and the estimated (large-sample) covariance matrix $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})$ for a given model, which can then be used to make appropriate statistical inferences.

22.5.2 Conditional ML Estimation

An alternative to using the unconditional likelihood function (22.14) for estimating the elements of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_k)$ is to employ a conditional likelihood function. The primary reason for using conditional likelihood methods is that unconditional methods can lead to seriously biased estimates of the elements of $\boldsymbol{\beta}$ when the amount of data available for analysis is not large. Although “not large” is admittedly a vague term, it acknowledges

⁵ The likelihood (22.14) is based on the responses of *individual* subjects, with (22.10) pertaining to the i th subject. In contrast, a categorical data analysis involving the binomial distribution is based on the responses of *groups* of subjects, with, say, Y_i subjects out of n_i in the i th group contracting the disease in question. In this situation, the underlying distribution of Y_i is binomial; that is,

$$\text{pr}(Y_i; \theta_i) = {}_n C_{Y_i} \theta_i^{Y_i} (1 - \theta_i)^{n_i - Y_i}$$

and the validity of certain categorical data analyses requires that n_i be fairly large in each group. The effects of small samples on the validity of ML analyses based on (22.14) are briefly discussed in the next subsection.

the potential problem of using large-sample-based statistical procedures such as maximum likelihood when the number of parameters to be estimated constitutes a fair proportion of the available data. This is often the situation when data involving *matching* must be analyzed.

In some matched case-control studies, for example, each case (i.e., a person with the disease in question) is matched with one or more controls (i.e., persons not having the disease in question) who have the same values (or are in the same categories) as the cases for the covariates involved in the matching. Analyzing such (matched) data requires that the data be cast into strata corresponding to the matched sets (e.g., pairs) with stratum-specific sample sizes that are small (i.e., reflecting sparse data). In particular, for pair-matched case-control data, each stratum contains only two subjects. A logistic model for analyzing matched data requires that the matching be taken into account by including indicator (i.e., dummy) variables in the model to reflect the matching strata. Thus, the model will have as many parameters as there are matched sets (plus parameters for the exposure variable, any other unmatched variables, and possibly even product terms of unmatched variables with exposure), so the total number of parameters in the model is large relative to the number of subjects in the study. See Kleinbaum and Klein (2010), Chapter 8, and Kleinbaum et al. (1982), Chapter 24, for further discussion of the principles of matching and modeling.

To see why using a conditional likelihood function is appropriate for matched data, let us consider a *pair-matched case-control study* to assess the effect of a (0–1) exposure variable E on a (0–1) health outcome variable D . Suppose that the matching involves the variables age, race, and sex; then, for each case, a control subject is found who has the same age (or age category), race, and sex as the corresponding case. Suppose, too, that the study involves 100 matched pairs, so that the study sample size n is 200. If no predictor variables are considered in the study other than E and the matching variables age, race, and sex, then a (no-interaction) logistic regression model appropriate for analyzing these data is

$$\text{logit}[\text{pr}(D = 1)] = \beta_0 + \beta_1 E + \sum_{i=1}^{99} \gamma_i V_i \quad (22.15)$$

where V_i , $i = 1, \dots, 99$, denote a set of dummy variables distinguishing the collection of 100 matched pairs; for example, V_i may be defined as

$$V_i = \begin{cases} 1 & \text{for the } i\text{th matched set} \\ 0 & \text{otherwise} \end{cases}$$

Model (22.15) allows us to predict case-control status as a function of exposure status (E) and the matching variables age, race, and sex, where the matching variables are incorporated into the model as dummy variables.⁶ The number of parameters in model (22.15) is 101. This is an example of a model whose number of parameters (101) is “large” relative to the number of subjects (200), so use of a conditional likelihood function is appropriate.

⁶ The analysis of a matched-pair study can (equivalently) be carried out without using logistic regression if no variables are controlled other than those involved in the matching. Such an analysis is a “stratified” analysis and involves using a Mantel-Haenszel test, odds ratio, and confidence interval. Equivalently, the stratified analysis can be carried out using McNemar’s test and estimation procedure (see Kleinbaum and Klein 2010, Chapter 11, or Kleinbaum 2002, Lesson 15).

If an unconditional likelihood is used to fit model (22.15), the resulting (biased) estimated odds ratio for the exposure effect is the squared value of the estimated odds ratio obtained from using a conditional likelihood; that is, if $\hat{\beta}_{1U}$ and $\hat{\beta}_{1C}$ denote the estimates of β_1 using unconditional and conditional likelihoods, respectively, then

$$\widehat{OR}_U = e^{\hat{\beta}_{1U}} \equiv (\widehat{OR}_C)^2 = e^{2\hat{\beta}_{1C}}$$

Thus, for example, if a pair-matched analysis using a conditional likelihood to fit model (22.15) produces an estimated odds ratio of, say, 3, a corresponding analysis (involving the same model fit to the same data) using an unconditional likelihood would yield a biased estimate of 3^2 , or 9.

Consider the data in the following table, which comes from the “Agent Orange Study” (Donovan, MacLennan, and Adena 1984), a pair-matched case–control study involving 8,502 matched pairs (i.e., $n = 17,004$):

		$D = 0$	
		$E = 1$	$E = 0$
$D = 1$	$E = 1$	2	125
	$E = 0$	121	8,254

In this table, the data layout separates the 8,502 case–control pairs into four cells, depending on whether both the case and the control were exposed (2 *concordant* pairs), the case was exposed and the control was unexposed (125 *discordant* pairs), the case was unexposed and the control was exposed (121 *discordant* pairs), or both the case and the control were unexposed (8,254 *concordant* pairs). The D and E variables are defined as follows: D = case–control status (1 = baby born with genetic anomaly, 0 = baby born without genetic anomaly) and E = father’s status (1 = Vietnam vet, 0 = non-Vietnam vet). The matching variables are M_1 = time period of birth, M_2 = mother’s age, M_3 = health insurance status, and M_4 = hospital. Since only the matching variables are being controlled, the analysis can be carried out using Mantel–Haenszel (or McNemar) statistics for a stratified analysis (see Kleinbaum and Klein 2010, Chapter 8, or Kleinbaum 2002, Lesson 15) without the need to perform a logistic regression analysis.

For these data, the Mantel–Haenszel (i.e., McNemar) test of H_0 : “No (E, D) association” gives a 1-d.f. chi-square statistic of 0.0650 ($P = .80$) and an estimated Mantel–Haenszel odds ratio of 1.033.⁷ An equivalent analysis (yielding the same results) based on logistic regression with a conditional likelihood uses the following logistic model:

$$\text{logit}[\text{pr}(D = 1)] = \beta_0 + \beta_1 E + \sum_{i=1}^{8,501} \gamma_i V_i \quad (22.16)$$

⁷ McNemar’s test is computed from the data in the preceding table using only the information from the (125 + 121) discordant pairs. The test statistic is computed as $(125 - 121)^2 / (125 + 121) = 0.0650$. The (Mantel–Haenszel) odds ratio is computed as the ratio of discordant pairs: $125/121 = 1.033$.

TABLE 22.3 Edited computer output from SAS's PROC LOGISTIC for logistic regression of the "Agent Orange Study" data using conditional ML estimation for model (22.16)

Variable	d.f.	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Odds Ratio
E	1	0.0325	0.1275	0.0650	.7987	1.033

© Cengage Learning

where the V_i denote dummy variables that distinguish the 8,502 matched sets. Model (22.16) contains 8,503 parameters, which is a large number (just over 50%) relative to the number of subjects in the study ($n = 17,004$).

Summarized output from fitting model (22.16) using a conditional likelihood is presented in Table 22.3. The output indicates that the odds ratio estimate is $\exp(0.0325) = 1.033$ and that the Wald test statistic is 0.0650 with a P -value of .80. These are the same results we obtained from the stratified analysis. The output does not show estimated coefficients for β_0 and for the $\{\gamma_i\}$; this is because these parameters, which distinguish the different matching strata, drop out of the conditional likelihood function (given by equation (22.17)) for this model and, therefore, cannot be estimated.

The output in Table 22.3 was obtained by using SAS's PROC LOGISTIC. The conditional likelihood function used by the LOGISTIC procedure for the general logistic model (written in logit form)

$$\text{logit}[\text{pr}(Y = 1)] = \beta_0 + \sum_{j=1}^k \beta_j X_j$$

takes the form

$$L_C(\mathbf{Y}; \boldsymbol{\beta}) = \frac{\prod_{i=1}^{n_1} \exp\left(\beta_0 + \sum_{j=1}^k \beta_j X_{ij}\right)}{\sum_u \left(\prod_{l=1}^{n_1} \exp\left(\beta_0 + \sum_{j=1}^k \beta_j X_{ulj}\right) \right)} \quad (22.17)$$

where the sum in the denominator is over all partitions of the set $\{1, 2, \dots, n\}$ into two subsets, the first of which contains n_1 elements. Expression (22.17) assumes, without loss of generality, that the first n_1 of the n subjects actually develop the disease, so X_{ij} denotes the value of variable X_j for the i th of these first n_1 subjects. The X_{ulj} term in the denominator, on the other hand, denotes the value of X_j for the l th person in the u th partition of the data into n_1 cases and $(n - n_1)$ controls. There are $n_1 C_{n_1} = n!/(n_1)!(n - n_1)!$ such partitions and hence that many terms in the summation.⁸

⁸ In the likelihood (22.17), the constant term β_0 drops out, since $\exp(\beta_0)$ can be factored out of both the numerator and the denominator. For the pair-matched model given by (22.16), the γ_i parameters can be factored out similarly; thus, in addition to β_0 , the γ_i coefficients drop out of the conditional likelihood and so do not have to be estimated.

In words, the conditional likelihood function (22.17) can be considered analogous to a conditional probability. More specifically, if we let \mathbf{X}_l denote the set of predictor values observed on the l th subject, $L_C(\mathbf{Y}; \boldsymbol{\beta})$ is the conditional probability (based on the underlying logistic model assumption) that the first n_1 members of the observed set $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ actually go with the n_1 subjects who developed the disease in question *given* (or conditional on) the observed set $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ and *given* the fact that exactly n_1 of the n subjects under study actually developed the health outcome (e.g., disease under study). Expressed in yet another way, (22.17) compares the likelihood of what was actually observed (the numerator) relative to the likelihood of all possible arrangements of the given data set (the denominator). The reason that (22.17) is called a *conditional likelihood* is that it is completely analogous to a conditional probability and is conditional on (arrangements of) the data actually observed.

With regard to data analysis, the conditional likelihood (22.17) is employed just like any other (e.g., unconditional) likelihood used with ML procedures: Wald and/or LR tests are used to test hypotheses about regression coefficients in the model, and large-sample confidence intervals are constructed using a percentage point of the $N(0,1)$ distribution and using estimates of the variances and covariances of the estimated regression coefficients. The computational aspects for a conditional likelihood are, nevertheless, somewhat more involved than for an unconditional likelihood, mainly because of the permutations of the data required to evaluate the denominator in (22.17). For example, if $n = 20$ and $n_1 = 3$, then ${}_{20}C_3 = 1,140$, so the denominator in (22.17) involves a sum of more than a thousand terms.

22.6 An Example of Conditional ML Estimation Involving Pair-matched Data with Unmatched Covariates

In this section, we describe an application of a conditional likelihood in a logistic regression analysis of case-control data involving both matched and unmatched covariates. The data derive from a pair-matched case-control study of the effect of estrogen use on the development of endometrial cancer. This study was conducted on 63 matched pairs of women living in a Los Angeles retirement village from 1971 to 1975 (Breslow and Day 1980; McNeil 1996). The following variables were involved in the study (and are analyzed here):

Outcome variable: $D = \text{CASE}$ (endometrial cancer status: 1 = case, 0 = control)

Exposure variable: $E = \text{EST}$ (estrogen use: 1 = yes, 0 = no)

Matching variables: $M_1 = \text{age}$, $M_2 = \text{marital status}$, $M_3 = \text{date of entry into retirement village}$

Covariates not matched on: $GALL$ (gall bladder disease status: 1 = present, 0 = absent)

Table 22.4 contains an excerpt of the pair-matched case-control data described above.

The STRATUM variable distinguishes the different strata (i.e., matched pairs). The cases in a given stratum are listed on a separate line from the controls. Thus, the listing shows pairs of lines with the same stratum number. Note that strata 62 and 63 indicate pairs that

TABLE 22.4 Data excerpt from a pair-matched case-control study of the relationship of estrogen use to endometrial cancer status on 63 matched pairs of women living in a Los Angeles retirement village from 1971 to 1975

STRATUM	CASE	EST	GALL
1	1	0	0
1	0	0	0
2	1	0	0
2	0	1	0
3	1	0	0
3	0	1	1
...
61	1	1	1
61	0	0	1
62	1	1	1
62	0	1	1
63	1	1	1
63	0	1	1

© Cengage Learning

are identical with respect to the exposure EST and unmatched covariates GALL. This corresponds to two observed pairs of the following layout:

	EST	GALL
CASE = 1	1	1
CASE = 0	1	1

Indeed, since both EST and GALL are binary variables, only 16 patterns of EST and GALL values are possible for each case-control pair; thus, among the 63 observed pairs, there will necessarily be several replicates of the same patterns for these two covariates. Of course, the individuals that comprise each “replicate” stratum may differ from those in the other strata with respect to the three factors matched on.

In analyzing the data of Table 22.4, we first describe the results obtained when the variable GALL is ignored. The analysis then simplifies to a stratified analysis for pair-matched data with no unmatched covariates. The corresponding frequencies describing the numbers of concordant and discordant case-control pairs are given in Table 22.5.

For the data in Table 22.5, the Mantel-Haenszel test statistic is computed as

$$\chi^2_{MH} = \frac{(29 - 3)^2}{(29 + 3)} = 21.125 \quad (P < .01)$$

TABLE 22.5 Frequencies of case-control pairs by exposure status, ignoring the variable GALL, in the case-control study of the relationship of estrogen use to endometrial cancer status

		CASE = 0		63	
		EST = 1			
CASE = 1	EST = 1	27	29		
	EST = 0	3	4		

© Cengage Learning

and the Mantel-Haenszel odds ratio estimate is computed as

$$\widehat{\text{mOR}} = \frac{29}{3} = 9.67$$

These results indicate a strong and significant association between estrogen use and endometrial cancer, based on these study data. However, this analysis does not control for the variable GALL; we now control for GALL using a logistic model with a conditional likelihood.

Table 22.6 summarizes the SAS output for fitting the following logistic model:

$$\text{logit}[\text{pr}(\text{CASE} = 1)] = \beta_0 + \beta_1(\text{EST}) + \beta_2(\text{GALL}) + \sum_{i=1}^{62} \gamma_i V_i \quad (22.18)$$

where the V_i , $i = 1, \dots, 62$, denote 62 dummy variables that distinguish the 63 matched pairs. Model (22.18) is a no-interaction model that estimates the effect of the exposure variable EST on the outcome CASE, controlling for the confounding effects of the unmatched covariate GALL and the matching variables reflected by the V_i . If we wanted to consider

TABLE 22.6 Edited output from SAS's PROC LOGISTIC using model (22.18) from the pair-matched case-control study of the relationship of estrogen use to endometrial cancer status for 63 matched pairs of women living in a Los Angeles retirement village (1971–1975)

Variable	d.f.	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Odds Ratio
EST	1	2.209	0.610	13.127	.000	9.107
GALL	1	0.695	0.616	1.128	.259	2.003

© Cengage Learning

the possibility of interaction between EST and GALL, we could add the product term $\text{EST} \times \text{GALL}$ to the preceding model. We leave the assessment of interaction to the interested reader, however, and report only the results for the no-interaction model (22.18).

From Table 22.6, the odds ratio for the effect of EST, controlling for GALL and the matching variables (not shown in the table), is estimated as $\exp(2.209) = 9.107$. The Wald statistic for testing whether this odds ratio differs significantly from 1 is given by

$$\chi^2_{1 \text{ d.f.}} = \left(\frac{2.209}{0.610} \right)^2 = 13.127 \quad (P < .0001)$$

which is highly significant. The estimated odds ratio and corresponding test statistic previously obtained from a stratified analysis that ignored the variable GALL were 9.67 and 21.125, respectively, versus 9.107 and 13.127 when controlling for GALL. Nevertheless, both analyses lead to the same conclusion: these data strongly suggest that estrogen use has a strong effect on the development of endometrial cancer. We can also compute a confidence interval for the (adjusted) odds ratio, using the output of Table 22.6. To obtain a 95% confidence interval for the adjusted odds ratio, we calculate

$$\exp[2.209 \pm 1.96(0.610)]$$

which yields lower and upper limits of 2.755 and 30.102, respectively.

Table 22.6 does not provide estimates of the coefficients of the V_i variables that control for the matching. This is because these coefficients drop out of the conditional likelihood (22.17) and, therefore, cannot be estimated. Although the V_i variables are not explicitly specified in the data listing given by Table 22.4, the matched pairs are identified in the listing by the STRATUM variable, which is used by the conditional ML program (e.g., SAS's PROC LOGISTIC) to control for the matching.

22.7 Summary

This chapter described the key features of logistic regression analysis, the most popular regression technique available for modeling dichotomous dependent variables. The *logistic model* is defined as a probability of the occurrence of one of two possible outcomes (say, 0 and 1) using the following formula:

$$\text{pr}(Y = 1) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \sum_{j=1}^k \beta_j X_j\right)\right]}$$

The most important reason for the popularity of the logistic model is that the right-hand side of the preceding expression ensures that the predicted value of Y will always lie between 0 and 1. Using the logit form of the logistic model defined by the expression

$$\text{logit}[\text{pr}(Y = 1)] = \beta_0 + \sum_{j=1}^k \beta_j X_j$$

we can estimate an odds ratio; the general formula for an odds ratio comparing two specifications \mathbf{X}_A and \mathbf{X}_B of the set of predictors $\mathbf{X} = (X_1, X_2, \dots, X_k)$ is

$$\text{OR}_{\mathbf{X}_A \text{ vs. } \mathbf{X}_B} = e^{\sum_{j=1}^k (X_{Aj} - X_{Bj})\beta_j}$$

ML estimates of the regression coefficients and associated estimated standard errors of the regression coefficients in a logistic model are typically obtained by using computer packages for logistic regression. These statistics can then be used to obtain numerical values for estimated adjusted odds ratios, to test hypotheses, and to obtain confidence intervals for population odds ratios based on standard ML techniques.

When performing a logistic regression analysis, we must decide between two alternative ML procedures, called *unconditional* and *conditional*. The key distinction between the two involves whether the number of parameters in the model constitutes a “large” proportion of the total study size. If so, the situation is typical of matched data, requiring the use of the conditional approach to ensure validity of the odds ratio estimates.

Problems

1. A researcher was interested in determining risk factors for high blood pressure (hypertension) among women. Data from a sample group of 680 women were collected. The following table gives the observed relationship between hypertension and smoking:

		Hypertension		
		Yes	No	
Smokers	<i>a</i> = 28	<i>b</i> = 271		
	<i>c</i> = 13	<i>d</i> = 368		
		41	639	

Let π be the probability of having hypertension, and suppose that the researcher used logistic regression to model the relationship between smoking and hypertension.

One model the researcher considered is

$$\log_e \frac{\pi}{1 - \pi} = \beta_0 + \beta_1 X_1 \quad (\text{Model 0})$$

where X_1 = smoking status (1 = smoker, 0 = nonsmoker). For this model, the estimated odds ratio for the effect of smoking on hypertension status can be computed by using the simple formula for an odds ratio in a 2×2 table—namely, ad/bc , where a , b , c , and d are the cell frequencies in the preceding table.

- a. Based on this information, compute the point estimate $\hat{\beta}_1$ of β_1 .

The researcher ultimately decided to use the following logistic regression model:

$$\log_e \frac{\pi}{1 - \pi} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \quad (\text{Model 1})$$

where X_1 = smoking status (1 = smoker, 0 = nonsmoker) and X_2 = age. The following information was obtained:

Parameter	Estimate	SE
β_0	-2.8	1.2
β_1	0.706	0.311
β_2	0.0004	0.0001
β_3	0.0006	0.0003

$$-2 \ln \hat{L} = 303.84$$

The estimated covariance matrix for $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$ is as follows:

$$\begin{bmatrix} \hat{\beta}_0 & \hat{\beta}_1 & \hat{\beta}_2 & \hat{\beta}_3 \\ 1.44 & 0.0001 & 0.0001 & 0.0001 \\ & 0.0967 & 0.1 \times 10^{-8} & 2.0 \times 10^{-8} \\ & & 1.0 \times 10^{-8} & 3.0 \times 10^{-8} \\ & & & 9.0 \times 10^{-8} \end{bmatrix}$$

- b. What is the estimated logistic regression model for the relationship between age and hypertension for nonsmokers?
 - c. What is a 20-year-old smoker's predicted probability of having hypertension?
 - d. Estimate the odds ratio comparing a 20-year-old smoker to a 21-year-old smoker. Interpret this estimated odds ratio.
 - e. Find a 95% confidence interval for the population odds ratio being estimated in part (d).
 - f. The log likelihood ($-2 \ln \hat{L}$) for the model consisting of the intercept, age, and smoking status was 308.00. Use this information, plus other information provided earlier, to perform an LR test of the null hypothesis that $\beta_3 = 0$ in model 1.
2. A five-year follow-up study on 600 disease-free subjects was carried out to assess the effect of a (0–1) exposure variable E on the development or not of a certain disease. The variables AGE (continuous) and obesity status (OBS), the latter a (0–1) variable, were determined at the start of follow-up and were to be considered as control variables in analyzing the data. For this study, answer the following questions.
- a. State the logit form of a logistic model that assesses the effect of the (0, 1) exposure variable E , controlling for the confounding effects of AGE and OBS and for the interaction effects of AGE with E and OBS with E .
 - b. Given the model described in part (a), give a formula for the odds ratio for the exposure–disease relationship that controls for the confounding and interactive effects of AGE and OBS.
 - c. Use the formula described in part (b) to derive an expression for the estimated odds ratio for the exposure–disease relationship that considers both confounding and interaction when AGE = 40 and OBS = 1.

- d. Give a formula for a 95% confidence interval for the population adjusted odds ratio being estimated in part (c). In stating this confidence interval formula, write out the formula for the estimated variance of the log of the estimated adjusted odds ratio in terms of estimated variances and covariances of appropriate estimated regression coefficients.
 - e. State the null hypothesis (in terms of model parameters) for simultaneously testing for no interaction of either AGE with E or OBS with E .
 - f. State the formula for the LR statistic that tests the null hypothesis described in part (e). What is the large-sample distribution of this statistic, including its degrees of freedom, under the null hypothesis?
 - g. Give the formula for the Wald statistic for testing the null hypothesis described in part (e). What is the large-sample distribution of this statistic under the null hypothesis?
 - h. Assuming that both the LR statistic described in part (f) and the Wald statistic described in part (g) are nonsignificant, state the logit form of the no-interaction model that is appropriate to consider at this point in the analysis.
 - i. For the no-interaction model given in part (h), give the formula for the (adjusted) odds ratio for the effect of exposure, controlling for AGE and OBS.
 - j. For the no-interaction model given in part (h), give the formula for a 95% confidence interval for the adjusted odds ratio for the effect of exposure, controlling for AGE and OBS.
 - k. Describe the Wald test for the effect of exposure, controlling for AGE and OBS, in the no-interaction model given in part (h). Include a description of the null hypothesis being tested, the form of the test statistic, and its distribution under the null hypothesis.
3. A study was conducted on a sample of 53 patients presenting with prostate cancer who had also undergone a laparotomy to ascertain the extent of nodal involvement (Collett 1991). The result of the laparotomy is a binary response variable, where 0 signifies the absence of, and 1 the presence of, nodal involvement. The purpose of the study was to determine variables that could be used to forecast whether the cancer has spread to the lymph nodes. Five predictor variables were considered, each measurable without surgery. The following printout provides information for fitting two logistic models based on these data. The five predictor variables were age of patient at diagnosis, level of serum acid phosphatase, result of an X-ray examination (0 = negative, 1 = positive), size of the tumor as determined by a rectal exam (0 = small, 1 = large), and summary of the pathological grade of the tumor as determined from a biopsy (0 = less serious, 1 = more serious).
- a. Which method of estimation do you think was used to obtain estimates of parameters for both models—conditional or unconditional ML estimation? Explain briefly.
 - b. For model I, test the null hypothesis of no effect of X-ray status on response. State the null hypothesis in terms of an odds ratio parameter; give the formula for the test statistic; state the distribution of the test statistic under the null hypothesis; and, finally, carry out the test, using the computer output information for Model I. Is the test significant?

Model I

Variable				P-value	\widehat{OR}	95% CI
Name	Coeff	StErr				
0 constant	0.057	3.460		.987		
1 age	-0.069	0.058		.232	0.933	0.833
2 acid	2.434	1.316		.064	11.409	0.865
3 xray	2.045	0.807		.011	7.731	1.589
4 tsize	1.564	0.774		.043	4.778	1.048
5 tgrad	0.761	0.771		.323	2.141	0.473

d.f.: 47 Dev: 48.126*

*The deviance statistic is an LR statistic that compares a current model of interest to the baseline model containing as many parameters as there are data points. The difference in deviance statistics obtained for two (hierarchically ordered) models being compared is equivalent to the difference in log likelihood statistics for each model. Thus, an LR test can equivalently be carried out by using differences in deviance statistics. See Section 24.5 for further details about deviances.

Model II

Variable				P-value	\widehat{OR}	95% CI
Name	Coeff	StErr				
0 constant	2.928	4.044		.469		
1 age	-0.101	0.067		.132	0.904	0.792
2 acid	1.462	1.559		.349	4.314	0.203
3 xray	-19.171	20.027		.338	0.000	0.000
4 tsize	1.285	0.894		.151	3.613	0.626
5 tgrad	0.421	0.923		.648	1.523	0.250
7 age*xray	0.257	0.278		.356	1.292	0.750
8 acid*xray	7.987	9.197		.385	2,943.008	0.000
9 tsiz*xray	2.236	2.930		.445	9.356	0.030
10 tgrd*xray	-0.624	2.376		.793	0.536	0.005

d.f.: 43 Dev: 44.474

- c. Using the printout data for Model I, compute the point estimate and 95% confidence interval for the odds ratio for the effect of X-ray status on response for a person of age 50, with phosphatase acid level of 0.50, tsize equal to 0, and tgrad equal to 0.
- d. State the logit form of Model II given in the accompanying computer printout.
- e. Using the results for Model II, give an expression for the estimated odds ratio that describes the effect of X-ray status on the response, controlling for age, phosphatase acid level, tsize, and tgrad. Using this expression, compute the estimated odds ratio of the effect of X-ray status on the response for a person of age 50, with phosphatase acid level of 0.50, tsize equal to 0, and tgrad equal to 0.
- f. For Model II, give an expression for the estimated variance of the estimated adjusted odds ratio relating X-ray status to response for a person of age 50, with phosphatase acid level of 0.50, tsize equal to 0, and tgrad equal to 0. Write this

expression for the estimated variance in terms of estimated variances and covariances obtained from the variance–covariance matrix of regression parameter estimates.

- g. Using your answer to part (f), give an expression for a 95% confidence interval for the odds ratio relating X-ray status to response for a person of age 50, with phosphatase acid level of 0.50, tsize equal to 0, and tgrad equal to 0.
- h. For Model II, carry out a “chunk” test for the combined interaction of X-ray status with each of the variables age, phosphatase acid level, tsize, and tgrad. State the null hypothesis in terms of one or more model coefficients; give the formula for the test statistic and its distribution and degrees of freedom under the null hypothesis; and report the P -value. Is the test significant?
- i. If you had to choose between Model I and Model II, which would you pick as the “better” model? Explain.

Assume that the following model has been defined as the initial model to be considered in a backward-elimination strategy to obtain a “best” model:

$$\begin{aligned}\text{logit}[\text{pr}(Y = 1)] = & \alpha + \beta_1(\text{xray}) + \beta_2(\text{age}) + \beta_3(\text{acid}) + \beta_4(\text{tsize}) + \beta_5(\text{tgrad}) \\ & + \beta_6(\text{age} \times \text{acid}) + \beta_7(\text{xray} \times \text{age}) + \beta_8(\text{xray} \times \text{acid}) \\ & + \beta_9(\text{xray} \times \text{tsize}) + \beta_{10}(\text{xray} \times \text{tgrad}) \\ & + \beta_{11}(\text{xray} \times \text{age} \times \text{acid})\end{aligned}$$

- j. For the above model—and considering the variable xray to be the only “exposure” variable of interest, with the variables age, acid, tsize, and tgrad considered for control—which β 's are coefficients of (potential) effect modifiers? Explain briefly.
- k. Assume that the only interaction term found significant is the product term xray \times age. What variables are left in the model at the end of the interaction assessment stage?
- l. Based on the (reduced) model described in part (k) (where the only significant interaction term is xray \times age), what expression for the odds ratio describes the effect of xray on nodal involvement status?
- m. Suppose that, as a result of confounding and precision assessment, the variables age \times acid, tgrad, and tsize are dropped from the model described in part (k). What is your final model, and what expression for the odds ratio describes the effect of xray on nodal involvement status?

References

- Breslow, N., and Day, N. 1980. *Statistical Methods in Cancer Research. Vol. I: The Analysis of Case-Control Studies*. Lyon, France: IARC Scientific Publications, No. 32.
- Collett, D. 1991. *Modeling Binary Data*. London: Chapman & Hall.
- Cox, D. R. 1975. “Partial Likelihood.” *Biometrika* 62: 269–76.
- Dantes, H. G., et al. 1988. “Dengue Epidemics on the Pacific Coast of Mexico.” *International Journal of Epidemiology* 17(1): 178–86.
- Donovan, J. W., MacLennan, R., and Adena, M. 1984. “Vietnam Service and the Risk of Congenital Anomalies: A Case–Control Study.” *Medical Journal of Australia* 140: 394–97.

- Hosmer, D. W., and Lemeshow, S. 1989. *Applied Logistic Regression*. New York: John Wiley & Sons.
- Kleinbaum, D. G. 2002. *ActivEpi—A CD ROM Course on Fundamentals of Epidemiologic Research*. New York and Berlin: Springer Publishers.
- Kleinbaum, D. G., and Klein, M. 2010. *Logistic Regression: A Self-Learning Text*, Third Edition. New York and Berlin: Springer Publishers.
- Kleinbaum, D. G., and Klein, M. 2012. *Survival Analysis: A Self Learning Text*, Second Edition. New York and Berlin: Springer Publishers.
- Kleinbaum, D. G.; Kupper, L. L.; and Morgenstern, H. 1982. *Epidemiologic Research: Principles and Quantitative Methods*. New York: Van Nostrand Reinhold.
- McNeil, D. 1996. *Epidemiological Research Methods*. West Sussex, England: John Wiley and Sons.

23

Polytomous and Ordinal Logistic Regression

23.1 Preview

In this chapter, the standard logistic model is extended to consider outcome variables that have more than two categories. We first describe polytomous logistic regression, which is used when the categories of the outcome variable are nominal; that is, they do not have any natural order. We then describe ordinal logistic regression, which is appropriate when the categories of the outcome variable do have a natural order.

Examples of outcome variables with more than two levels might include (1) endometrial cancer subtypes: adenosquamous, adenocarcinoma, or other; (2) patients' preferred treatment regimen, selected from among three or more options; (3) tumor grade: well differentiated, moderately differentiated, or poorly differentiated; (4) disease symptoms that have been classified by subjects as being absent, mild, moderate, or severe; (5) invasiveness of a tumor classified as *in situ*, locally invasive, or metastatic. For the first two of these examples, the categories are essentially nominal (i.e., polytomous, without ordinality), whereas examples (3)–(5) consider ordinal categories.

When modeling an outcome variable, whether polytomous or ordinal, with more than two categories, the typical research question remains the same as when there are two outcome categories: What is the relationship of one or more predictor variables (X_1, X_2, \dots, X_k) to an outcome (Y) of interest? Also, since the model is an extension of binary logistic regression, the measure of effect of interest will be an odds ratio whose formula involves the exponential of linear functions of the regression coefficients in the model. Furthermore, since the method of estimation used to fit both polytomous and ordinal logistic regression models is maximum likelihood (ML) estimation, the procedures for carrying out hypothesis testing and confidence interval estimation are analogous to ML techniques used for binary logistic regression. The specific forms that the odds ratio takes, as well as examples of statistical inference procedures for polytomous and ordinal logistic regression, will be described and illustrated in the sections to follow.

23.2 Why Not Use Binary Regression?

A simple approach for the analysis of data with an outcome containing more than two categories is to choose an appropriate referent category out of the total collection of categories or combine categories into a single referent group, pool the remaining categories to form a second group, and then simply utilize the logistic modeling techniques for dichotomous outcomes. For example, if endometrial cancer subtypes is the outcome variable, one could pool “adenocarcinoma” and “other” into a referent category and then use “adenosquamous” as a second category. This can also be done for ordinal outcomes; for example, if the outcome symptom severity has four categories of severity, one might compare subjects with none or only mild symptoms to those with either moderate or severe symptoms.

One major disadvantage of dichotomizing a polytomous or ordinal outcome is loss of meaningful detail in describing the outcome of interest. For example, in the last scenario given above, we can no longer compare mild versus none or moderate versus mild. This loss of detail may, in turn, affect the conclusions made about the covariate–outcome relationships.

Another possible approach to the analysis of polytomous (or ordinal) outcomes would be to choose a referent category and then fit several separate dichotomous logistic models comparing the remaining categories individually with the chosen referent category. Such an approach may be criticized because the likelihood function for any separate dichotomous logistic model utilizes the data involving only the two categories of the outcome variable being considered. In contrast, the likelihood function for a polytomous or ordinal logistic regression model utilizes the data involving all categories of the outcome variable in a single model structure. In other words, different likelihood functions are used when fitting each dichotomous model separately than when fitting a polytomous model that considers all levels simultaneously. Consequently, both the estimation of the parameters and the estimation of the variances of the parameter estimates may differ when comparing the results from fitting separate dichotomous models to the results from the polytomous or ordinal model. In fact, in general, there is a power loss when doing separate dichotomous logistic models versus doing polytomous (or ordinal) logistic regression. (Note, however, that for the special case of a polytomous model with one dichotomous predictor, which is introduced in the next section, fitting separate logistic models yields the same parameter estimates and variance estimates as fitting the polytomous model.)

23.3 An Example of Polytomous Logistic Regression: One Predictor, Three Outcome Categories

In this section, we illustrate polytomous logistic regression with one dichotomous predictor variable and an outcome (Y) that has three categories. This is the simplest case of a polytomous logistic regression model. In a later section, we discuss extending the polytomous model to more than one predictor variable and then to outcomes with more than three categories.

End Stage Renal Disease (ESRD) is a condition in which there has been an irreversible loss of renal (i.e., kidney) function that makes the patient permanently dependent on renal replacement therapy (RRT), typically involving hemodialysis, in order to sustain life. ESRD patients suffer from increased morbidity and substantially reduced life expectancy. We consider data obtained from an ESRD surveillance system on 3,049 ESRD patients who initiated dialysis in (the U.S. state of) Georgia in 2002 (Volkova et al. 2006). For illustrative purposes, the outcome variable that we consider is “cause of ESRD,” which has been categorized into the following three nominal categories: Category 2 is hypertension, Category 1 is diabetes, and Category 0 is other disease.

The research question concerns whether there is an association between race and cause of ESRD. Thus, the predictor (i.e., exposure) variable that we consider here is “race,” which has been coded as 1 = black, 0 = white. A cross-tabulation of race and ESRD stage is presented in Table 23.1. Other variables that we consider as control variables include “age at dialysis initiation” and “gender” (1 = female, 0 = male). The data set is called *esrddata* and can be obtained from the publisher’s website.

With polytomous logistic regression, one of the categories of the outcome variable is designated as the reference category, and each of the other categories is compared with this referent. The choice of reference category can be arbitrary and is at the discretion of the researcher. Changing the reference category does not change the general structure of the polytomous logistic regression model but does change the interpretation of the parameter estimates in the model.

In our example with three levels of the outcome, the Other Disease group has been designated as the reference category. We are, therefore, interested in modeling two main comparisons. We want to compare subjects whose ESRD was caused by Hypertension (category 2) to those subjects whose ESRD was caused by Other Disease (category 0), and we also want to compare subjects whose ESRD was caused by Diabetes (category 1) to those subjects whose ESRD was caused by Other Disease (category 0).

If we consider these two comparisons separately, the crude odds ratios can be calculated using data from Table 23.1. The odds ratio comparing Hypertension (category 2) to Other Disease (category 0) is 2.105, and the odds ratio comparing Diabetes (category 1) to Other Disease (category 0) is 1.484. Thus, without considering statistical significance, blacks appear to be twice as likely as whites to have their ESRD being caused by Hypertension relative to the ESRD cause being Other Disease. Similarly, blacks appear to be 1.5 times as likely as whites to have their ESRD being caused by Diabetes relative to Other Disease.

TABLE 23.1 Crude data layout and odds ratios relating race to cause of End Stage Renal Disease (Georgia, 2002)

ESRD Cause	Black	White	Total	\widehat{OR}
Hypertension (2)	630	323	953	2.105
Diabetes (1)	773	562	1,335	1.484
Other Disease (0)	366	395	761	1.000
Total	1,769	1,280	3,049	

As previously described for a dichotomous outcome variable coded 0 and 1, the odds for developing an outcome of interest (e.g., Diabetes) equal

$$\text{Odds} = \frac{P(D = 1)}{1 - P(D = 1)} = \frac{P(D = 1)}{P(D = 0)}$$

namely, the probability that the outcome equals one divided by the probability that the outcome equals zero. Also, for a dichotomous outcome variable coded as 0 or 1, recall that the logit form of the logistic model, $\text{logit } P(\mathbf{X})$, is defined as the natural log of the odds for developing a disease for a person with a set of independent variables specified by \mathbf{X} ; that is,

$$\text{logit } P(\mathbf{X}) = \ln \left[\frac{P(D = 1|\mathbf{X})}{P(D = 0|\mathbf{X})} \right] = \alpha + \sum_{j=1}^k \beta_j X_j$$

However, when there are three categories of the outcome, there are three probabilities to consider:

$$P(D = 0|\mathbf{X}), \quad P(D = 1|\mathbf{X}), \quad \text{and} \quad P(D = 2|\mathbf{X})$$

and the sum of the probabilities for the three outcome categories must be equal to one. With three outcome categories, there are now two “odds-like” expressions to consider, one for each of the two pairwise ratio comparisons we are making:

$$\frac{P(D = 2|\mathbf{X})}{P(D = 0|\mathbf{X})}, \quad \frac{P(D = 1|\mathbf{X})}{P(D = 0|\mathbf{X})}$$

Because each comparison considers only two probabilities, the probabilities in each ratio do not sum to one. Thus, the two “odds-like” expressions are not true odds. However, if we restrict our interest to just the two categories being considered in a given ratio, we may still roughly interpret the ratio as an odds; for example,

$$\frac{P(D = 2|\mathbf{X})}{P(D = 0|\mathbf{X})}$$

is a measure of how likely it is for a person with covariate values \mathbf{X} to be in outcome category 2 rather than in outcome category 0. For ease of our subsequent discussion, we will use the term “odds,” rather than “odds-like,” for these expressions.

In polytomous logistic regression with three levels, we, therefore, define our model using two expressions for the natural log of these “odds-like” quantities. The first is the natural log of the probability that the outcome is in category 1 divided by the probability that the outcome is in category 0; the second is the natural log of the probability that the outcome is in category 2 divided by the probability that the outcome is in category 0; that is,

$$\ln \left[\frac{P(D = 1|\mathbf{X})}{P(D = 0|\mathbf{X})} \right], \quad \ln \left[\frac{P(D = 2|\mathbf{X})}{P(D = 0|\mathbf{X})} \right]$$

Because our example has three outcome categories, the polytomous model requires two regression expressions:

$$\text{Diabetes vs. Other Disease: } \ln \left[\frac{P(D = 1 | \text{race})}{P(D = 0 | \text{race})} \right] = \alpha_1 + \beta_{11}(\text{race}) \quad (23.1)$$

$$\text{Hypertension vs. Other Disease: } \ln \left[\frac{P(D = 2 | \text{race})}{P(D = 0 | \text{race})} \right] = \alpha_2 + \beta_{21}(\text{race})$$

In these two models, corresponding parameters for the intercept (α) and the slope (β) are allowed to be different; that is, when comparing outcome categories 1 and 0, the parameters are denoted α_1 and β_{11} , whereas the parameters are denoted as α_2 and β_{21} when comparing categories 2 and 0.

The results, based on fitting the polytomous model examining the association between “race” and “cause of ESRD,” are presented in the computer output that follows. There are two sets of parameter estimates. The output is listed in descending order, with α_2 labeled as intercept 2 and α_1 labeled as intercept 1.

Edited SAS Output (PROC LOGISTIC) for Polytomous Logistic Regression for ESRD Data Evaluating the Relationship between Race (X) and Cause of ESRD (Y)

MODEL FIT STATISTICS							
Criterion	Intercept Only	Intercept and Covariates					
-2 Log L	6534.134	6091.865					
TYPE 3 ANALYSIS OF EFFECTS							
Effect	DF	Wald Chi-Square	Pr > ChiSq				
race	1	55.6927	<.0001				
ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES							
Parameter	Cause	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Symbol
Intercept	2	1	-0.2013	0.0750	7.1976	0.0073	α_2
Intercept	1	1	0.3526	0.0657	28.8456	<.0001	α_1
race	2	1	0.7443	0.0997	55.6924	<.0001	β_{21}
race	1	1	0.3950	0.0913	18.7135	<.0001	β_{11}
ODDS RATIO ESTIMATES							
Effect	Cause	Point Estimate	95% Wald Confidence Limits				
race	2	2.105	1.731	2.559			
race	1	1.484	1.241	1.775			

Computer packages vary in the presentation of output; in particular, the coding of the polytomous outcome variable must be understood to correctly read and interpret the computer output for a given package. For example, when the reference category for the outcome variable (i.e., Cause) is 0 (i.e., Other Disease), the SAS output shown will list the (intercept and race) parameters in descending order. Thus, the output corresponding to α_2 , which corresponds to category 2 (i.e., Hypertension), precedes the output corresponding to α_1 , corresponding to category 1 (i.e., Diabetes); similarly, the output for β_{21} precedes the output for β_{11} .

From the SAS output, the two estimated logit-like expressions for the fitted polytomous model are given by

$$\text{Hypertension versus Other Disease: } \ln \left[\frac{\hat{P}(D = 2 | \text{race})}{\hat{P}(D = 0 | \text{race})} \right] = -0.2013 + (0.7443)(\text{race})$$

$$\text{and Diabetes versus Other Disease: } \ln \left[\frac{\hat{P}(D = 1 | \text{race})}{\hat{P}(D = 0 | \text{race})} \right] = 0.3526 + (0.3950)(\text{race})$$

Once a polytomous logistic regression model has been fit and the parameters (intercepts and regression coefficients) have been estimated, the estimates of the exposure–disease association can be calculated in a similar manner (using β -type formulas) as used for binary logistic regression.¹ For example, since the model considered in the SAS output contains a single (0,1) predictor variable (i.e., race), the two odds ratio estimates shown are calculated by exponentiating the corresponding estimated regression coefficient $\hat{\beta}_{21}$ and $\hat{\beta}_{11}$; that is,

$$\widehat{OR}_{\text{race}}(\text{Hypertension versus Other Disease}) = e^{\hat{\beta}_{21}} = e^{0.7443} = 2.105$$

$$\widehat{OR}_{\text{race}}(\text{Diabetes versus Other Disease}) = e^{\hat{\beta}_{11}} = e^{0.3950} = 1.484$$

Notice that these odds ratio estimates are identical to the estimates shown in Table 23.1 for the crude data in which logistic modeling was not performed. In general, when a polytomous logistic regression model contains only one dichotomous predictor variable, the estimated odds ratios comparing separate outcome categories to a referent category are identical to corresponding crude odds ratios calculated directly from appropriate two-way frequency tables.

¹ The special case of a dichotomous predictor can be generalized to include multicategory or continuous predictors. For a polytomous outcome with three categories (as in our example), the odds ratio formula that compares any two levels ($X = X_A$ vs. $X = X_B$) of a predictor (X) is given by

$$OR_g = \exp[\beta_{g1}(X_A - X_B)]$$

where $g = 1, 2$.

In our example, $X = \text{race}$ with $X_A = 1$ and $X_B = 0$, so the above formula simplifies to

$$OR_g = \exp[\beta_{g1}(1 - 0)] = \exp[\beta_{g1}] = e^{\beta_{g1}} \quad g = 1, 2$$

If race had been coded as $X_A = 1$ and $X_B = -1$, however, the odds ratio formula would change to

$$OR_g = \exp[\beta_{g1}(1 - (-1))] = \exp[2\beta_{g1}] = e^{2\beta_{g1}} \quad g = 1, 2$$

Procedures for testing hypotheses and computing confidence intervals are straightforward generalizations of techniques that apply to logistic regression modeling with a dichotomous outcome variable. As with a standard logistic regression, we can use a likelihood ratio (LR) test to assess the significance of the independent variable $X = \text{race}$ in our model. However, rather than testing one beta coefficient for each independent variable, we are now testing two at the same time in our example; that is, the null hypothesis for the effect of race is given by $H_0: \beta_{11} = \beta_{21} = 0$. The corresponding LR statistic is, therefore,

$$\text{LR}_{\text{race}} = -2 \ln L_R - (-2 \ln L_F) \quad (23.2)$$

which has approximately a chi-square distribution with 2 degrees of freedom (d.f.) under H_0 . (The test has 2 d.f. because H_0 involves setting two parameters, β_{11} and β_{21} , equal to 0.) The *full* model (F) identified in (23.2) refers to the polytomous logistic regression model defined by (23.1) and involving the single independent variable race. The *reduced* model (R) refers to the polytomous logistic regression model that (23.1) reduces to under the null hypothesis $H_0: \beta_{11} = \beta_{21} = 0$, namely, a model containing only intercept parameters:

$$\text{Reduced model: } \ln \left[\frac{P(D = 1 | \text{race})}{P(D = 0 | \text{race})} \right] = \alpha_1 \quad \ln \left[\frac{P(D = 2 | \text{race})}{P(D = 0 | \text{race})} \right] = \alpha_2 \quad (23.3)$$

The *Model Fit Statistics* in the SAS output provide the values of log likelihood statistics required to calculate the LR statistic (23.2). The LR statistic is, therefore, calculated as

$$\text{LR}_{\text{race}} = -2 \ln L_R - (-2 \ln L_F) = (6534.134) - (6091.865) = 442.269$$

which is highly significant (P -value < .0001) based on the chi-square distribution with 2 d.f., thus indicating a significant effect of race in model (23.1).

While the LR test allows for the assessment of the effect of an independent variable across all categories of the outcome simultaneously, it is possible that one might be interested in evaluating the effect of the independent variable on a single (nonreferent) outcome category (e.g., Diabetes or Hypertension compared individually to Other Disease in our example). A Wald test can be performed in this situation. Continuing with our example, the null hypothesis for comparing Diabetes and Other Disease (i.e., category 1 vs. 0) is that β_{11} equals zero. From the above SAS output, the Wald statistic for testing $H_0: \beta_{11} = 0$ is equal to 18.7135, with a P -value less than .0001. The null hypothesis for comparing Hypertension versus Other Disease (i.e., category 2 vs. 0) is that β_{21} equals zero. The Wald statistic for testing $H_0: \beta_{21} = 0$ is equal to 55.6924, with a P -value less than .0001. The reader should not be surprised to see such small P -values, since the sample size for this study ($n = 3,049$) is quite large.

From the above significance testing, we conclude that race is statistically significant for both the Diabetes versus Other Disease comparison (category 1 vs. 0) and the Hypertension versus Other Disease comparison (category 2 vs. 0). When using a polytomous logistic regression model for these data, we must either keep both betas (β_{11} and β_{21}) for the independent variable "race" or drop both betas. Even if only one estimated beta is significantly

different from zero, both betas must be retained if the independent variable is to remain in the model.

Large-sample confidence interval estimation using a polytomous logistic regression model is also analogous to the standard (dichotomous) logistic regression situation. However, since two odds ratios are involved for the variable *race* (i.e., β_{11} and β_{21}), we must calculate two confidence interval estimates.² Using the estimated standard errors 0.0997 and 0.0913 in the preceding SAS output for the estimated coefficients $\hat{\beta}_{21}$ and $\hat{\beta}_{11}$, respectively, the two large-sample 95% confidence intervals for the race variable are calculated as follows:

95% CI for $e^{\beta_{21}}$ (Hypertension vs. Other Disease):

$$\exp[\hat{\beta}_{21} \pm 1.96 S_{\hat{\beta}_{21}}] = \exp[0.7443 \pm 1.96(0.0997)] = (1.731, 2.559)$$

95% CI for $e^{\beta_{11}}$ (Diabetes vs. Other Disease):

$$\exp[\hat{\beta}_{11} \pm 1.96 S_{\hat{\beta}_{11}}] = \exp[0.3950 \pm 1.96(0.0913)] = (1.241, 1.775)$$

The above two 95% confidence intervals calculated above for $e^{\beta_{21}}$ and $e^{\beta_{11}}$ are shown in the SAS output given earlier for these data. In general, SAS's LOGISTIC procedure automatically provides output for confidence intervals for e^β corresponding to each β regression coefficient in a polytomous logistic regression model.³

23.4 An Example: Extending the Polytomous Logistic Model to Several Predictors

Expanding the polytomous logistic regression model to add more independent variables is straightforward. If the model contains k predictors, we simply include these k independent variables for each of the outcome comparisons. The procedures for calculation of odds ratios and confidence intervals and for hypothesis testing remain the same.

To illustrate, we return to our example that considers the relationship of race to cause of ESRD in 3,049 ESRD patients who initiated dialysis in Georgia in 2002. Suppose we now want to consider the effect of race, controlling for age at dialysis initiation and gender.

² As with point estimates of odds ratios, confidence interval estimation can be generalized from the special case of a single dichotomous predictor to include multicategory or continuous predictors. For a polytomous outcome with three categories (as in our example), the 95% confidence interval formula that compares any two levels ($X = X_A$ vs. $X = X_B$) of a predictor (X) is given by

$$\exp[\hat{\beta}_{g1}(X_A - X_B) \pm 1.96 (X_A - X_B)S_{\hat{\beta}_{g1}}] \quad g = 1, 2$$

³ As with dichotomous logistic regression models, the expression $e^{\hat{\beta}_{g1}}$ is a correct formula for an odds ratio estimate for a predictor that is coded as (0, 1). However, for multicategory or continuous predictors, $e^{\hat{\beta}_{g1}}$ gives the odds ratio for a one-unit change in the predictor, which might not reflect an estimated odds ratio of interest (e.g., if the predictor is systolic blood pressure). Similarly, the large-sample confidence interval formula $\exp[\hat{\beta}_{g1} \pm 1.96 S_{\hat{\beta}_{g1}}]$ will not generally be a confidence interval of interest for an odds ratio for a continuous predictor.

The age variable has been defined as a categorical variable called “ageg,” containing seven categories according to the following scheme:

ageg	age
1	≤ 22
2	23–39
3	40–49
4	50–59
5	60–69
6	70–79
7	≥ 80

The gender variable has been defined as gender = 1 if female and gender = 0 if male. The model now contains eight predictor variables:

$$X_1 = \text{race}$$

$X_j = \text{ageg}_j, j = 2, 3, \dots, 7$, denoting six (0, 1) dummy variables for seven age groups, with ageg = 1 as the referent group

$$X_8 = \text{gender}$$

Since there are three outcome categories, the polytomous logistic regression model is defined in terms of the following two logit-like functions:⁴

$$\begin{aligned} \left[\frac{P(D = 1 | \text{race, ageg, gender})}{P(D = 0 | \text{race, ageg, gender})} \right] &= \alpha_1 + \beta_{11}(\text{race}) + \sum_{j=2}^7 \beta_{1j}(\text{ageg}_j) + \beta_{18}(\text{gender}) \\ \left[\frac{P(D = 2 | \text{race, ageg, gender})}{P(D = 0 | \text{race, ageg, gender})} \right] &= \alpha_2 + \beta_{21}(\text{race}) + \sum_{j=2}^7 \beta_{2j}(\text{ageg}_j) + \beta_{28}(\text{gender}) \end{aligned}$$

(23.4)

The results from fitting the polytomous model defined by (23.4) are presented in the following SAS output. Again, there are two sets of parameter estimates corresponding to the intercept and each predictor in each of the two models in (23.4). There are a total of 18 parameters (including intercepts) that are estimated. The output is listed in descending

⁴ The general form of the polytomous logistic regression model that allows $G (\geq 2)$ categories of the outcome variable and k predictors is given as follows:

$$\ln \left[\frac{P(D = g | \mathbf{X})}{P(D = 0 | \mathbf{X})} \right] = \alpha_g + \sum_{j=1}^k \beta_{gj} X_j \quad g = 1, 2, \dots, (G - 1)$$

where $\mathbf{X} = (X_1, X_2, \dots, X_k)$ denotes the collection of predictors in the model. Note that the number of logit-like functions is $(G - 1)$; for example, if $G = 3$ (as in our example), then $(G - 1) = 2$; and if $G = 5$, then $(G - 1) = 4$.

The general formula for the collection of $(G - 1)$ odds ratios that compare two different specifications of \mathbf{X} is

$$\text{OR}_g = \exp \left[\sum_{j=1}^k \beta_{gj} (X_{A_j} - X_{B_j}) \right] \quad g = 1, 2, \dots, (G - 1)$$

where $\mathbf{X}_A = (X_{A_1}, X_{A_2}, \dots, X_{A_k})$ and $\mathbf{X}_B = (X_{B_1}, X_{B_2}, \dots, X_{B_k})$.

Edited SAS Output (PROC LOGISTIC) for Polytomous Logistic Regression Output for ESRD Data for Evaluating the Relationship Between Race (X_1) and Cause of ESRD (Y) Controlling for Age and Gender

MODEL FIT STATISTICS		
Criterion	Intercept Only	Intercept and Covariates
-2 Log L	6534.134	6136.928

TYPE 3 ANALYSIS OF EFFECTS			
Effect	DF	Wald Chi-Square	Pr > ChiSq
race	2	95.8805	<.0001
ageg	12	276.2004	<.0001
gender	2	23.2241	<.0001

ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES						
Parameter	Cause	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	2	1	-0.7752	0.1235	37.3888	<.0001
Intercept	1	1	-0.4820	0.1315	13.4437	0.0002
race	2	1	1.0652	0.1088	95.8639	<.0001
race	1	1	0.5745	0.0993	33.4818	<.0001
ageg1	2	1	-2.5520	0.5188	24.1958	<.0001
ageg1	1	1	-3.0331	0.6241	23.6128	<.0001
ageg2	2	1	-0.4855	0.1533	10.0333	0.0015
ageg2	1	1	-0.6062	0.1644	13.6034	0.0002
ageg3	2	1	-0.2032	0.1511	1.8092	0.1786
ageg3	1	1	0.1263	0.1531	0.6806	0.4094
ageg4	2	1	0.4442	0.1397	10.1153	0.0015
ageg4	1	1	0.9505	0.1439	43.6399	<.0001
ageg5	2	1	0.5459	0.1367	15.9360	<.0001
ageg5	1	1	1.0288	0.1413	53.0194	<.0001
ageg6	2	1	0.8655	0.1382	39.2336	<.0001
ageg6	1	1	0.9632	0.1453	43.9475	<0.0001
gender	2	1	-0.0642	0.1030	0.3891	0.5328
gender	1	1	0.3247	0.0958	11.4997	0.0007

ODDS RATIO ESTIMATES				
Effect	Cause	Point Estimate	95% Wald Confidence Limits	
race	2	2.901	2.344	3.591
race	1	1.776	1.462	2.158

(continued)

ODDS RATIO ESTIMATES				
Effect	Cause	Point Estimate	95% Wald Confidence Limits	
ageg 1 vs 7	2	0.020	0.006	0.066
ageg 1 vs 7	1	0.027	0.006	0.117
ageg 2 vs 7	2	0.154	0.102	0.232
ageg 2 vs 7	1	0.308	0.201	0.472
ageg 3 vs 7	2	0.204	0.136	0.306
ageg 3 vs 7	1	0.641	0.429	0.959
ageg 4 vs 7	2	0.390	0.267	0.570
ageg 4 vs 7	1	1.462	0.999	2.141
ageg 5 vs 7	2	0.432	0.299	0.625
ageg 5 vs 7	1	1.581	1.089	2.297
ageg 6 vs 7	2	0.595	0.412	0.859
ageg 6 vs 7	1	1.481	1.015	2.162
gender	2	0.938	0.766	1.148
gender	1	1.384	1.147	1.669

© Cengage Learning

order, with α_2 corresponding to the first intercept listed and α_1 corresponding to the second intercept; similarly, the estimates for each predictor are listed in descending order, with the estimate of β_{2j} preceding the estimate of β_{1j} for the j th predictor.

Since the primary predictor (i.e., exposure) variable of interest is race, the odds ratio values of interest in the following SAS output are 2.901 and 1.776, which are the two point estimates listed for race in the portion of the output labeled “Odds Ratio Estimates.” The value 2.901 gives the estimated odds ratio for the effect of race comparing subjects whose ESRD cause is Hypertension (cause = 2) versus Other Disease (cause = 0), controlling for age and gender. The value 1.776 gives the odds ratio for the effect of race comparing subjects whose ESRD cause is Diabetes (cause = 1) versus Other Disease (cause = 0), controlling for age and gender. Since race has a (0, 1) coding, these “adjusted” odds ratios can be computed alternatively by exponentiating the appropriate regression coefficients for race listed in the “Analysis of Maximum Likelihood Estimates” portion of the output; that is,

$$\widehat{OR}_{race}(\text{Hypertension versus Other Disease|age, gender}) = e^{\hat{\beta}_{21}} = e^{1.0652} = 2.901$$

$$\widehat{OR}_{race}(\text{Diabetes versus Other Disease|age, gender}) = e^{\hat{\beta}_{11}} = e^{0.5745} = 1.776$$

The 95% confidence intervals for $e^{\beta_{g1}}$, $g = 1, 2$, were calculated using the standard large-sample formula given by

$$\exp[\hat{\beta}_{g1} \pm 1.96 S_{\hat{\beta}_{g1}}] \quad g = 1, 2$$

The Wald chi-square statistics for testing $H_0: \beta_{g1} = 0, g = 1, 2$, were calculated using the standard format described in Chapter 21; that is, the Wald test statistic

$$(\hat{\beta}_{g1}/S_{\hat{\beta}_{g1}})^2 \quad g = 1, 2$$

has an approximate chi-square distribution with 1 d.f. for large samples under $H_0: \beta_{g1} = 0$.

The portion of the output that is labeled “Type 3 Analysis of Effects” gives generalized Wald tests for testing each of the three predictors in the model (race, ageg, and gender, with the six dummy variables for ageg being considered collectively as a single predictor). These test statistics are not equivalent to the LR statistics given by (23.2). In large samples, values obtained for the generalized Wald statistics will typically be close to corresponding LR values; nevertheless, LR tests are considered by statisticians to have better statistical properties than Wald tests, so that LR tests are typically preferred.

To carry out an LR test for the significance of the race variable in model (23.4), we would need to compare the log likelihood statistic ($-2 \ln L_F$) for model (23.4) with the corresponding log likelihood statistic ($-2 \ln L_R$) for the reduced model that would be obtained under the null hypothesis $H_0: \beta_{11} = \beta_{21} = 0$ in model (23.4). The reduced model will not contain the race variable and can be written as follows:

$$\begin{aligned} \ln \left[\frac{P(D = 1 | \text{ageg, gender})}{P(D = 0 | \text{ageg, gender})} \right] &= \alpha_1 + \sum_{j=2}^7 \beta_{1j} (\text{ageg}_j) + \beta_{18}(\text{gender}) \\ \ln \left[\frac{P(D = 2 | \text{ageg, gender})}{P(D = 0 | \text{ageg, gender})} \right] &= \alpha_2 + \sum_{j=2}^7 \beta_{2j} (\text{ageg}_j) + \beta_{28}(\text{gender}) \end{aligned} \quad (23.5)$$

From the SAS output, we find that $-2 \ln L_F = 6136.938$. The log likelihood statistic obtained from the reduced model (output not provided) is $-2 \ln L_R = 6236.806$. The corresponding LR statistic is then computed as

$$\text{LR}_{\text{race}} = -2 \ln L_R - (-2 \ln L_F) = (6236.806) - (6136.938) = 99.868$$

which is highly significant ($P\text{-value} < .0001$) based on the chi-square distribution with 2 d.f., thus indicating a significant effect of race in model (23.4) after adjusting for age and gender. Notice that the generalized Wald statistic for race is 95.8805, which is close but not identical to the LR statistic. Since race is the primary predictor of interest, the generalized Wald tests provided for ageg and gender, as well as corresponding LR tests for ageg and gender, are not discussed further. (Note, however, that the d.f. = 12 for the generalized Wald test for the significance of ageg because there are two sets of six dummy variables that define the ageg variable in model (23.4).)

Table 23.2 compares the estimated odds ratios and corresponding 95% confidence intervals for race in model (23.4), which controls for ageg and gender, with those for model (23.1), which does not control for ageg and gender. If we compare the model with three predictor variables (race, ageg, and gender) with the model with only race included, the effect of race in the reduced model (with race only) is weaker both for the comparison of Hypertension to Other Disease (odds ratio estimates: 2.901 vs. 2.105) and for the comparison of

TABLE 23.2 Odds ratios, 95% confidence intervals, and Wald P-values for race comparing models that control and do not control for ageg and gender

Comparison	Variables in Model					
	Race, ageg, gender			Race only		
	Odds Ratio	95% Conf. Int.	Wald P-value	Odds Ratio	95% Conf. Int.	Wald P-value
2 vs. 0 Hypertension vs. Other Disease	2.901	(2.344, 3.591)	< .0001	2.105	(1.731, 2.559)	< .0001
	1.776	(1.462, 2.158)	< .0001	1.484	(1.241, 1.775)	< .0001

© Cengage Learning

Diabetes to Other Disease (odds ratio estimates: 1.776 vs. 1.484). These results suggest that ageg and/or gender acts as a confounder of the relationship between race and cause of ESRD. The numerical results based on the model containing only race suggest a bias toward the null value of one when both ageg and gender are not included as covariates.

The 95% confidence intervals in the table are wider when ageg and gender are included in the models than when they are not (2 vs. 0 widths: 1.427 vs. 0.828; 1 vs. 0 widths: 0.696 vs. 0.534). This indicates a slight loss of precision when ageg and gender are included along with race in the model. Nevertheless, the model that contains all three predictors is preferred because the estimated odds ratios are meaningfully different when ageg and gender are included but not when ageg and gender are ignored (i.e., ageg and gender are apparently confounders; see Chapter 11).

23.5 Ordinal Logistic Regression: Overview

In this section, the standard logistic model is extended to consider ordinal outcome variables. We restrict attention to the most popular form of ordinal logistic regression model called the *proportional odds (or cumulative logit) model* (see Ananth and Kleinbaum 1997 for a description of other types of ordinal logistic regression models).

Ordinal variables have a natural ordering among the outcome levels. An example is cancer tumor grade, ranging from well-differentiated to moderately differentiated to poorly differentiated tumors. Another example is length of in-patient hospital stay (in days), ranging from short to intermediate to long. A third example is time since previous mammography, ranging from within one year to greater than one year to never having had a mammography.

An ordinal outcome variable can always be modeled with a polytomous logistic model, as discussed in the previous section, but it can also be modeled with more precision using ordinal logistic regression, provided that certain assumptions are met. Ordinal logistic regression, unlike polytomous regression, takes into account any inherent ordering of the levels in the outcome variable, thus making better use of the ordinal information.

23.6 A “Simple” Example: Three Ordinal Categories and One Dichotomous Exposure Variable

Table 23.3 provides a 3×2 table of hypothetical data on 275 hospital patients collected to assess the relationship between an ordinal outcome variable for length of hospital stay (D) and a $(0, 1)$ exposure variable (E) that distinguishes between two types of patients.

If we treat the outcome variable D as a polytomous outcome by ignoring the natural ordering of the categories of D , the analysis (without controlling for other variables) would consider two odds-ratio estimates obtained from two 2×2 sub-tables that compare separate outcome categories (say, $D = 1$ and $D = 2$) to a referent category (say, $D = 0$), as shown in Table 23.4.

TABLE 23.3 Crude data relating patient type (E) to hospital length of stay (D)

	$E = 1$	$E = 0$	Total
$D = 2$ (long)	20	30	50
$D = 1$ (medium)	30	70	100
$D = 0$ (short)	25	100	125
Total	75	200	275

© Cengage Learning

TABLE 23.4 Tables and estimated odds ratios treating D as a polytomous outcome

	$E = 1$	$E = 0$	Total
$D = 1$ (medium)	30	70	100
$D = 0$ (short)	25	100	125
Total	55	170	225
$\widehat{OR} = \frac{30 \times 100}{70 \times 25} = 1.71$			
	$E = 1$	$E = 0$	Total
$D = 2$ (long)	20	30	50
$D = 0$ (short)	25	100	125
Total	45	130	175
$\widehat{OR} = \frac{20 \times 100}{30 \times 25} = 2.67$			

© Cengage Learning

However, if the natural ordering of the outcome (D) variable is to be considered, the use of the ordinal logistic regression procedure that involves the proportional odds model requires a different partitioning of the overall data into two sub-tables. In particular, we must consider the possible ways to collapse the ordinal categories into 2×2 tables that preserve the natural ordering of the categories of D . Using the data in Table 23.3, there are two such tables that can be formed, which are shown in Table 23.5. These tables compare category 0 to categories 1 and 2 combined and also compare categories 0 and 1 combined to category 2. We have not combined categories 0 and 2 for comparison with category 1, since that would disrupt the natural ordering from 0 through 2.

The proportional odds model makes an important assumption. Under this model, the odds ratio for the effect of an exposure variable for any 2×2 table that is obtained by collapsing one or more of the rows of the 3×2 table given in Table 23.3 will be the same, regardless of where the cut-point is made, provided the natural ordering of the outcome variable is not disrupted. In other words, the odds ratio is *invariant* to how the outcome categories are dichotomized.

In Table 23.5, the odds ratios from the two collapsed tables are almost identical (2.00 versus 2.06) and thus provide evidence that the proportional odds assumption is not violated. It would be unusual for the collapsed-table odds ratios to match perfectly. The odds ratios do not have to be exactly equal; as long as they are “close,” the proportional odds assumption may be considered reasonable.

There is also a statistical test called the *score test* designed to evaluate whether a model constrained by the proportional odds assumption (i.e., an ordinal model) is significantly different from the corresponding model in which the odds ratio parameters are not constrained by the proportional odds assumption (i.e., a polytomous model). The test statistic is distributed approximately chi-square, with d.f. equal to $k(G - 2)$, where k is the number of predictor variables in the model and G is the number of ordinal categories of the outcome variable.

TABLE 23.5 Collapsed 2×2 tables obtained from Table 23.3 for a proportional odds model

	$E = 1$	$E = 0$	Total
$D = 1$ or 2	50	100	150
$D = 0$	25	100	125
Total	75	200	275
$\widehat{OR} = \frac{50 \times 100}{100 \times 25} = 2.00$			
	$E = 1$	$E = 0$	Total
$D = 2$	20	30	50
$D = 0$ or 1	55	170	225
Total	75	200	275
$\widehat{OR} = \frac{20 \times 170}{30 \times 35} = 2.06$			

For the ordinal 3×2 data given in Table 23.3, the score test is equivalent to testing the null hypothesis that the two odds ratios being estimated by collapsing one or more of the rows of the table are equal. In other words, for these data, the null hypothesis that the proportional odds assumption is satisfied is equivalent to the equality of the two population odds ratios estimated in Table 23.5. Since we have $G = 3$ ordinal outcome categories and $k = 1$ predictor (E), the d.f. for the score test is $k(G - 2) = 1$. The score test result for these data (chisq = 0.0082, P -value = .9281) is nonsignificant, so that we can conclude that the proportional odds assumption is satisfied.⁵

The proportional odds model for the analysis of the data in Table 23.3 is given by the following expression:

$$P(D \geq g|E) = \frac{1}{1 + \exp[-(\alpha_g + \beta E)]} \quad g = 1, 2 \quad (23.6)$$

where $\alpha_1 > \alpha_2$.

The above formula for the proportional odds model differs from the binary logistic model in that model (23.6) is formulated as the probability of an inequality (i.e., $P(D \geq g|E)$), whereas the formula for a binary outcome is the probability of an equality (i.e., $P(D = 1|E)$). Moreover, the proportional odds model differs from the polytomous logistic regression model (defined generally in footnote 4 of the previous section) in that the regression coefficient β in model (23.6) is not subscripted by g , as in the polytomous model.

An equivalent definition of the proportional odds model can be written in terms of the odds of inequalities as follows:

$$\text{Odds}_g(E) = \frac{P(D \geq g|E)}{1 - P(D \geq g|E)} = \frac{P(D \geq g|E)}{P(D < g|E)} \quad (23.7)$$

If we substitute the formula (23.6) for $P(D \geq g|E)$ into the expression (23.7), we find that the formula for the odds simplifies to

$$\text{Odds}_g(E) = \frac{P(D \geq g|E)}{P(D < g|E)} = \exp[\alpha_g + \beta E] \quad (23.8)$$

Using (23.8), we can then derive the formula for the odds ratio comparing an exposed person ($E = 1$) to an unexposed person ($E = 0$), as follows:

$$\text{OR}_{E=1 \text{ vs. } E=0} = \frac{\text{Odds}_g(E = 1)}{\text{Odds}_g(E = 0)} = \frac{\exp[\alpha_g + \beta(1)]}{\exp[\alpha_g + \beta(0)]} = \exp[(\alpha_g - \alpha_g) + \beta(1 - 0)] = e^\beta \quad (23.9)$$

⁵ If the proportional odds assumption is inappropriate, there are other ordinal logistic models available that make alternative assumptions about the ordinal nature of the outcome. Examples include a continuation ratio model, a partial proportional odds model, and a stereotype regression model. A discussion of these models is beyond the scope of the current presentation. (See the review by Ananth and Kleinbaum 1997.)

From (23.9), even though the intercept terms of the proportional odds model (α_g) are subscripted by g , a single odds ratio is obtained for the comparison of exposed and unexposed groups. This single odds ratio (given by e^β) is the common odds ratio that is assumed by the proportional odds model when the ordinal data (e.g., Table 23.3) are partitioned into the two 2×2 tables given in Table 23.5.

Using simple algebra, we can rewrite (23.9) to express the odds for exposed subjects in terms of the odds for unexposed subjects as follows:

$$\text{Odds}_g(E = 1) = e^\beta \text{Odds}_g(E = 0) \quad (23.10)$$

Equation (23.10) essentially says that the odds for exposed persons for any value of g can be obtained from the odds for unexposed persons for that same value of g by multiplying the latter by the proportionality constant e^β that does not depend on g . In other words, the odds for exposed persons are *proportional* to the odds for unexposed persons no matter which of the 2×2 collapsed tables is considered (i.e., no matter what value of g is used). This is the reason why model (23.6), and its more general form (given in footnote 11 later in this section), is called the *proportional odds* model.

The results from fitting model (23.6) to the data in Table 23.3 are presented in the following SAS output.

Edited SAS Output (PROC LOGISTIC) for Analysis of Data from Table 24.3 Using the Proportional Odds Model

SCORE TEST FOR THE PROPORTIONAL ODDS ASSUMPTION					
Chi-Square	DF	Pr > ChiSq			
0.0082	1	0.9281			
MODEL FIT STATISTICS					
-2 Log L		532.218			
ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.0228	0.2301	19.7665	<.0001
Intercept	1	0.7021	0.2241	9.8139	0.0017
E	1	0.7044	0.2543	7.6702	0.0056
ODDS RATIO ESTIMATES					
Effect		Point Estimate	95% Wald Confidence Limits		
E		2.023	1.229	3.333	

The output provides the estimates for two intercepts (α_2 and α_1 , respectively), but there is only one estimated β for the effect of E . The estimated odds ratio for E is $e^{0.7044} = 2.023$. Note that this estimated odds ratio lies between 2.00 and 2.06, the two estimated odds ratios obtained from partitioning the data into two 2×2 tables (Table 23.5) that preserve the ordering of the outcome variable.

The Wald chi-square test for the significance of E (i.e., the test of $H_0: \beta = 0$ versus $H_A: \beta \neq 0$) is highly significant (two-tailed P -value = .0056). The 95% confidence limits for a confidence interval for e^β are 1.229 and 3.333.

The results for this hypothetical data set indicate that exposed patients are about twice (i.e., 2.023 times) as likely as unexposed patients to have a “longer” rather than a “shorter” hospital stay, regardless of whether “longer” is categorized strictly or broadly (i.e., just long or medium and/or long).

23.7 Ordinal Logistic Regression Example Using Real Data with Four Ordinal Categories and Three Predictor Variables

Acinetobacter species are aerobic gram-negative bacilli that can cause health-care-associated infections and can survive for prolonged periods of time in the environment and on health-care workers’ hands. *Acinetobacter* infections have become increasingly difficult to treat due to the emergence of strains resistant to almost all commonly prescribed antimicrobial agents. Outbreaks caused by multidrug-resistant (MDR) *Acinetobacter* have been reported in hospitals all over the world; more recently, they have become a significant problem in military medical facilities.

We consider an example of the use of the proportional odds model for the Multidrug-Resistant (MDR) *Acinetobacter* Infection Study (Sunenshine et al. 2006), a retrospective matched-pair cohort study⁶ that investigated the impact of MDR *Acinetobacter* infection status on length of hospitalization stay using patient records from January 2003 through August 2004 from two tertiary care hospitals in Baltimore City. The data considered here compare the outcomes of 56 patients with MDR *Acinetobacter* infection (exposed patients) to 56 pair-matched patients with susceptible *Acinetobacter* infection (unexposed susceptible referents).

The predictor (exposure) variable of primary interest is *Acinetobacter* infection status (acstatus), coded as 0 for susceptible patients and 1 for infected patients. The outcome variable is the ordinal variable ordhospdays, defined from terciles of hospital length of stay (hospdays) using the following three categories:

Ordhospdays	Hospdays
1	≤ 9
2	10–23
3	≥ 24

⁶ In their paper, Sunenshine et al. (2006) did not use a proportional odds model to analyze their data but rather treated the outcome variable as a binary outcome and used conditional logistic regression to account for the matching.

The matching variable is exposure time prior to infection (*exptime*); in particular, a susceptible referent patient had to have a pre-infection hospital length of stay (LOS) within 5% of the matched MDR *Acinetobacter* patient's pre-infection hospital length of stay. The variable *exptime* was redefined⁷ as a categorical variable called *newexptime* with the following eight categories:

newexptime	exptime
0	0
1	1–4
2	5–9
3	10–14
4	15–19
5	20–24
6	25–30
7	≥ 30

There were two other control variables (not involved in the matching): a measure of severity of illness prior to *Acinetobacter* infection, called "apache" score, and a summary measure of comorbidities, called "charlson" score, calculated using data from the past medical history recorded in the medical chart of each patient.

Table 23.6 presents the 3×2 table of the crude data that describes the association between the exposure variable *Acinetobacter* infection status (*acstatus*) and the ordinal outcome variable hospital length of stay (*ordhospdays*). Here the coding of *ordhospdays* as 0, 1, or 2 reflects the ordinal nature of the outcome. For example, it is necessary that the interval " $9 < \text{hospdays} \leq 23$ " be coded (as *ordhospdays* = 1) between the interval " $0 \leq \text{hospdays} \leq 9$ " (coded as *ordhospdays* = 0) and the interval " $\text{hospdays} > 23$ " (coded as

TABLE 23.6 Crude data relating *Acinetobacter* infection status (*acstatus*) and hospital length of stay (*ordhospdays*)

E = acstatus			
D = ordhospdays	1 = infected	0 = susceptibles	Total
2 = (hospdays > 23)	24	13	37
1 = (9 < hospdays ≤ 23)	16	23	39
0 = (0 ≤ hospdays ≤ 9)	16	20	36
Total	56	56	112

© Cengage Learning

⁷ SAS's LOGISTIC procedure does not allow the use of conditional ML estimation (typically used for matched data when the number of matching strata is large relative to the study size) when fitting an ordinal logistic regression model. Consequently, the matching variable *exptime* was redefined as an eight-category variable (*newexptime*) in order to justify the use of unconditional ML estimation, which may be used when the number of matching strata (e.g., 8) is small relative to the total number of observations ($n = 112$). Such pooling of the original matched pairs strata is reasonable whenever a referent patient from one pair would have been eligible to be chosen as the referent in another pair because of similar exposure times (i.e., two such matched pairs are said to be *exchangeable*).

TABLE 23.7 Collapsed 2 × 2 tables obtained from Table 23.6 with the natural ordering of the ordhospdays variable preserved

<i>E</i> = acstatus		
<i>D</i> = ordhospdays	1 = infected	0 = susceptibles
1–2	40	36
0	16	20
$\widehat{OR} = \frac{40 \times 20}{36 \times 16} = 1.39$		
<i>E</i> = acstatus		
<i>D</i> = ordhospdays	1 = infected	0 = susceptibles
2	24	13
0–1	32	43
$\widehat{OR} = \frac{24 \times 43}{13 \times 32} = 2.48$		

© Cengage Learning

ordhospdays = 2). This contrasts with polytomous logistic regression, in which the coding is not necessarily reflective of an underlying order in the outcome variable.

Under the proportional odds assumption, the odds ratio obtained for any 2 × 2 table obtained by collapsing one of the rows of the 3 × 2 table given in Table 23.6 should be the same, regardless of where the cut-point is made, provided the natural ordering of the outcome variable is not disrupted. There are two such 2 × 2 tables, as shown in Table 23.7. These tables compare category 0 to categories 1 and 2 combined and categories 0 and 1 combined to category 2. Again, we cannot combine categories 0 and 2 for comparison with category 1, since that would disrupt the natural ordering from 0 through 2.

The two odds ratios shown in Table 23.7 are 1.39 and 2.48, which are different enough from one another to suggest that the proportional odds assumption may be violated. However, because three control variables (apache, charlson, and the categorical variable newextime) are being considered in the analysis in addition to the exposure variable (acstatus), a more appropriate comparison should compare two estimated *adjusted* odds ratios for the effect of acstatus on ordhospdays, controlling for apache, charlson, and newextime.⁸

Moreover, the score test (described earlier) for the proportional odds assumption that considers all four predictors in the same model (23.11) provides a comparison of such

⁸ Estimated adjusted odds ratios for the effect of acstatus, controlling for apache, charlson, and newextime, can be computed by fitting two separate binary logistic models using the data in the 2 × 2 tables shown in Table 23.7, where each model would contain all four predictor variables. The adjusted odds ratios are of the form e^{β_1} , where β_1 is the coefficient of acstatus in each model.

estimated adjusted odds ratios.⁹ The score test result for this model [chisq = 8.7459, d.f. = $k(G - 2) = 10(3 - 2) = 10$, P -value = .5564] is nonsignificant. Thus, even though the two unadjusted odds ratios shown in Table 23.7 are somewhat different, we can nevertheless argue from the score test result that the proportional odds assumption seems to be satisfied for these data.

The proportional odds model¹⁰ for the analysis of the *Acinetobacter* infection data described above is

$$\begin{aligned} P(D \geq g | \mathbf{X}) &= \frac{1}{1 + \exp\left[-\left(\alpha_g + \beta_1(\text{acstatus}) + \beta_2(\text{apache}) + \beta_3(\text{charlson}) + \sum_{j=0}^6 \gamma_j(\text{newexptime}_j)\right)\right]} \\ &\quad (23.11) \end{aligned}$$

where $g = 1, 2$; $\alpha_1 > \alpha_2$; $\text{newexptime}_0, \dots, \text{newexptime}_6$ are seven dummy variables for the eight categories of the categorical variable newexptime; and $\mathbf{X} = (\text{acstatus}, \text{apache}, \text{charlson}, \text{newexptime}_0, \text{newexptime}_1, \dots, \text{newexptime}_6)$.

As with the simpler version of the proportional odds model defined earlier by (23.6), the regression coefficients $\beta_1, \beta_2, \beta_3, \gamma_0, \gamma_1, \dots$, and γ_6 in model (23.11) are not subscripted by g .¹¹

⁹ The score test for the proportional odds model (23.11) containing the four predictor variables (acstatus, apache, charlson, and newexptime) simultaneously tests the proportional odds assumption for all four variables in the model rather than one variable at a time adjusted for the other three variables. The latter (single variable) test is not automatically provided in SAS's LOGISTIC procedure.

¹⁰ The general form of the proportional odds model (used by SAS) that allows $G (\geq 3)$ categories of the outcome variable and k predictors is given as follows:

$$P(D \geq g | \mathbf{X}) = \frac{1}{1 + \exp\left[-\left(\alpha_g + \sum_{j=1}^k \beta_j X_j\right)\right]} \quad g = 1, 2, \dots, G - 1$$

where $\mathbf{X} = (X_1, X_2, \dots, X_k)$ denotes the collection of predictors in the model and where $\alpha_1 > \alpha_2 > \dots > \alpha_{G-1}$. Note that the number of intercept terms is $(G - 1)$; for example, if $G = 4$ (as in the *Acinetobacter* example), then $(G - 1) = 3$.

¹¹ An alternate formulation of the proportional odds model (used by STATA) is given by

$$P(D^* \leq g | \mathbf{X}) = \frac{1}{1 + \exp\left[-\left(\alpha_g^* - \sum_{j=1}^k \beta_j^* X_j\right)\right]} \quad g = 1, 2, \dots, G - 1$$

where $D^* = 1, 2, \dots, G$. Two important features of this formula that differ from the general formula in footnote 10 are the direction of the inequality ($D^* \leq g$) and the negative sign before $\sum_{j=1}^k \beta_j^* X_j$. In terms of the beta coefficients, these two key differences "cancel out," so that $\beta_j = \beta_j^*$. Consequently, if the same data are fit for each formulation of the model, the same parameter estimates of the β_j 's would be obtained for each model. However, the intercepts for the two formulations differ as follows: $\alpha_g = -\alpha_g^*$.

Also, we can equivalently write model (23.11) in odds form as

$$\begin{aligned}\text{Odds}_g(\mathbf{X}) &= \frac{P(D \geq g | \mathbf{X})}{P(D < g | \mathbf{X})} \\ &= \exp \left[\alpha_g + \beta_1(\text{acstatus}) + \beta_2(\text{apache}) + \beta_3(\text{charlson}) \right. \\ &\quad \left. + \sum_{j=0}^6 \gamma_j(\text{newexptime}_j) \right] \quad g = 1, 2\end{aligned}\quad (23.12)$$

Using (23.12), the formula for the odds ratio¹² comparing an exposed person (*acstatus* = 1) to an unexposed person (*acstatus* = 0), controlling for *apache*, *charlson*, and *newexptime*, is given as follows:

$$\begin{aligned}\text{OR}_{E=1 \text{ vs. } E=0 | \text{apache, charlson, newexptime}} &= \frac{\text{Odds}_g(E=1 | \text{apache, charlson, newexptime})}{\text{Odds}_g(E=0 | \text{apache, charlson, newexptime})} \\ &= \frac{\exp \left[\alpha_g + \beta_1(1) + \beta_2(\text{apache}) + \beta_3(\text{charlson}) + \sum_{j=0}^6 \gamma_j(\text{newexptime}_j) \right]}{\exp \left[\alpha_g + \beta_1(0) + \beta_2(\text{apache}) + \beta_3(\text{charlson}) + \sum_{j=0}^6 \gamma_j(\text{newexptime}_j) \right]} \\ &= e^{\beta_1}\end{aligned}$$

(23.13)

From (23.13), even though the intercept terms of the proportional odds model (α_g) are subscripted by g , a single odds ratio, e^{β_1} , is obtained for the comparison of exposed and unexposed groups.

The results from fitting model (23.11) to the *Acinetobacter* data are presented below.

As mentioned earlier, the score test shown in the output is nonsignificant (*P*-value = .5564), which supports the use of the proportional odds model for this analysis. The degrees of freedom of 10 for this test are obtained from the formula $k(G - 2)$, where $k = 10$ is the number of predictors in the model and $G = 3$ is the number of ordinal categories.

The output shows estimates for two intercepts (α_2 and α_1 , respectively) but only one estimated coefficient for the effect of each predictor in the model. The estimated odds ratio for the *adjusted* effect of *acstatus* is $e^{0.6865} = 1.99$. This odds ratio value lies between 1.39 and

¹² The general formula for the odds ratio that compares two different specifications of \mathbf{X} is

$$\text{OR}_{\mathbf{X}_A, \mathbf{X}_B} = \exp \left[\sum_{j=1}^k \beta_j(X_{Aj} - X_{Bj}) \right]$$

where $\mathbf{X}_A = (X_{A1}, X_{A2}, \dots, X_{Ak})$ and $\mathbf{X}_B = (X_{B1}, X_{B2}, \dots, X_{Bk})$. Notice that the subscript g does not appear in the formula.

2.48, the two estimated odds ratios obtained by partitioning the data into two 2×2 tables (in Table 23.7) that preserve the ordering of the outcome variable.

Edited SAS Output (PROC LOGISTIC) for Analysis of *Acinetobacter* Data Using the Proportional Odds Model (23.11)

SCORE TEST FOR THE PROPORTIONAL ODDS ASSUMPTION								
	Chi-Square	DF	Pr > ChiSq					
	7.4396	10	0.5564					
MODEL FIT STATISTICS								
-2 Log L		206.617						
ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES								
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq			
Intercept 2	1	-2.0965	0.5248	15.9583	<.0001			
Intercept 1	1	-0.1671	0.4802	0.1211	0.7278			
acstatus	1	0.6865	0.3876	3.1380	0.0765			
apache	1	0.0392	0.0130	9.0900	0.0026			
charlson	1	-0.1755	0.0650	7.2931	0.0069			
newexptime 0	1	-1.9109	0.5169	13.6648	0.0002			
newexptime 1	1	-0.6888	0.4972	1.9192	0.1659			
newexptime 2	1	0.2996	0.4174	0.5153	0.4729			
newexptime 3	1	-0.7062	0.5713	1.5277	0.2165			
newexptime 4	1	-0.0715	0.5433	0.0173	0.8954			
newexptime 5	1	0.5357	0.5388	0.9885	0.3201			
newexptime 6	1	0.7573	0.5394	1.9715	0.1603			
ODDS RATIO ESTIMATES								
Effect		Point Estimate	95% Wald Confidence Limits					
acstatus		1.987	0.930	4.247				
apache		1.040	1.014	1.067				
charlson		0.839	0.739	0.953				
newexptime 0 vs 7		0.025	0.005	0.121				
newexptime 1 vs 7		0.084	0.018	0.388				
newexptime 2 vs 7		0.227	0.057	0.901				
newexptime 3 vs 7		0.083	0.016	0.441				
newexptime 4 vs 7		0.156	0.031	0.784				
newexptime 5 vs 7		0.287	0.060	1.367				
newexptime 6 vs 7		0.358	0.075	1.717				

The Wald chi-square test for the significance of $acstatus$ (i.e., $H_0: \beta_1 = 0$) is of borderline significance (the two-tailed P -value is .0765, whereas the one-tailed P -value is .0383). The 95% confidence limits for e^{β_1} are 0.930 and 4.247, which indicates a moderate lack of precision.

The results for this hypothetical data set indicate that patients with *Acinetobacter* infection were about twice (i.e., 1.99 times) as likely as susceptible patients to have a “long” rather than a “short” hospital stay, regardless of whether “long” is categorized strictly or broadly (e.g., just long or medium and/or long).

23.8 Summary

In this chapter, the standard “binary” logistic regression model has been extended to handle an outcome variable that has more than two categories. When the categories of the outcome have no natural order, polytomous logistic regression is appropriate, whereas ordinal logistic regression is preferred when the categories have a natural order.

We illustrated the use of polytomous logistic regression to assess the relationship of “race” to “cause” of End Stage Renal Disease (ESRD) in 3,049 patients who initiated dialysis in Georgia in 2002 (Volkova et al., 2006). The three categories of ESRD cause (2 = Hypertension, 1 = Diabetes, or 0 = Other Disease) had no natural ordering, so polytomous logistic regression was appropriate. The polytomous logistic regression model provides separate odds ratio estimates that compare different categories of the outcome with a referent category. The regression coefficients in this model are different for different outcome categories.

The proportional odds model is the most popular model used for carrying out ordinal logistic regression but requires a “proportional odds assumption” to be satisfied. A test procedure for assessing the proportional odds assumption is the *score test*, which is part of the output in most computer packages (e.g., SAS, STATA, SPSS) that fit the proportional odds model. We have referenced alternative models to consider if the proportional odds assumption is not satisfied (Ananth and Kleinbaum 1997).

In contrast to the polytomous logistic regression model, the regression coefficients for each predictor in the proportional odds model do not vary by outcome category, and a single odds ratio estimate is obtained for each (binary) predictor. Such a single odds ratio estimate represents a weighted average of estimated odds ratios obtained from different ways of combining the outcome categories into strata that preserve the natural ordering of the outcome variable.

We illustrated ordinal logistic regression using data from a Multidrug-Resistant (MDR) *Acinetobacter* Infection Study (Sunenshine et al. 2006); this is a retrospective matched-pair cohort study that investigated the impact of MDR *Acinetobacter* infection status on length of hospitalization stay using patient records from January 2003 through August 2004 from two tertiary care hospitals in Baltimore City. The outcome variable, days of hospital stay, was treated as an ordinal variable, so that ordinal logistic regression was an appropriate data analytic procedure.

Problems

1. A cross-sectional study was carried out to assess the relationship of alcohol and smoking to blood pressure in 2,500 men ages 20 years or older in four North American population groups, each group utilizing a different clinic. The outcome variable was blood pressure status (BP): normotensive = 0, moderate hypertensive = 1, and severe hypertensive = 2. The primary exposure variables were alcohol status (ALC: nondrinker = 0, light drinker = 1, and heavy drinker = 2, with 0 being the referent category) and smoking status (SMK: none = 0, less than 2 packs per day = 1, and 2+ packs per day = 2, with 0 being the referent category). Variables considered for control included AGE (under 40 = 0, at least 40 = 1), PA (physical activity: not regular = 0, regular = 1), and CLINIC (4 categories, with clinic 4 being the referent category).
 - a. State the logit form of a no-interaction ordinal (proportional odds) logistic model that would describe the relationship of the outcome BP (treated ordinally) with the predictors ALC, SMK, AGE, PA, and CLINIC. In stating this model, make sure to treat all predictors as categorical variables.

(Note: In questions to follow, we will refer to the exposure variables ALC and SMK as *E* variables; the control variables AGE, PA, and CLINIC will be referred to as *V* variables.)

- b. Consider the following (crude) odds ratio estimates obtained from the crude 3×3 table relating ALC to BP:

$$\widehat{OR}_1 = \widehat{OR}_{BP=2 \text{ vs. } 0-1}(\text{ALC} = 2 \text{ vs. } 0), \quad \widehat{OR}_2 = \widehat{OR}_{BP=2 \text{ vs. } 0-1}(\text{ALC} = 1 \text{ vs. } 0)$$

$$\widehat{OR}_3 = \widehat{OR}_{BP=1-2 \text{ vs. } 0}(\text{ALC} = 2 \text{ vs. } 0), \quad \widehat{OR}_4 = \widehat{OR}_{BP=1-2 \text{ vs. } 0}(\text{ALC} = 1 \text{ vs. } 0)$$

What relationships among the above odds ratios would one expect to see if the proportional odds model was appropriate for the variable ALC?

- c. What does the score test for the proportional odds model allow one to evaluate, and how does it work (i.e., state the null hypothesis and how to proceed, depending on whether you reject or do not reject the null hypothesis)?
- d. For the model described in part (a), give an expression for the odds of being a moderate or severe hypertensive (i.e., $BP \geq 1$) who is a heavy-drinking, 45-year-old, 2+ pack-a-day smoker who does regular physical activity and comes from clinic 4.
- e. For the model described in part (a), give an expression for the odds ratio for being a moderate or severe hypertensive (i.e., $BP \geq 1$) that compares a heavy-drinking, 2+ pack-a-day smoker to a light-drinking nonsmoker, controlling for AGE, PA, and CLINIC.
- f. For the model described in part (a), give an expression for the odds ratio for being a normotensive (i.e., $BP = 0$) that compares a heavy-drinking, 2+ pack-a-day smoker to a light-drinking nonsmoker, controlling for AGE, PA, and CLINIC.

- g.** Suppose that the model in part (a) was extended to allow for two-way interactions (i.e., products involving two variables) between alcohol and age, alcohol and physical activity, alcohol and clinic, smoking and age, smoking and physical activity, smoking and clinic, and alcohol and smoking. Assuming that alcohol status and smoking status are exposure (E) variables of interest and that AGE, PA, and CLINIC are control (V) variables, state which variables in the interaction model are E_iV_j variables and which are E_iE_j variables. (Make sure to define the variables explicitly, keeping in mind that all the variables in the model are being treated as categorical variables.)
- h.** Starting with the model described in part (g), suppose that it was decided to determine a “best” model by first testing for significant interaction among the E_iV_j variables. Describe how one would *simultaneously test* for the significance of at least one of the E_iV_j product terms in this model. Make sure to state the null hypothesis in terms of model parameters, describe the formula for the test statistic, and give the distribution and degrees of freedom of the test statistic under the null hypothesis.
- i.** Assume that the only significant interaction terms among the E_iV_j variables were product terms involving alcohol status with age and smoking status with physical activity. State the model that remains for further assessment after this interaction assessment. Make sure to write the model in terms of the variables alcohol status, smoking status, age, and physical activity and the appropriate product terms.
- j.** Considering the answer to part (i), what is the formula for the odds ratio that compares heavy-drinking, 2+ pack-a-day smokers to light-drinking nonsmokers, controlling for AGE, PA, and CLINIC?
- k.** Using the odds ratio formula described in part (j), give a formula for a 95% confidence interval for the odds ratio for a 30-year-old man who does not do regular physical activity.

Note: The general 95% CI formula for an odds ratio of the form $\exp[L]$, where $\hat{L} = c_1\hat{\beta}_1 + \dots + c_k\hat{\beta}_k$, is given by

$$\exp[\hat{L} \pm 1.96\sqrt{\text{Var}(\hat{L})}]$$

where

$$\text{Var}(\hat{L}) = \sum_{j=1}^k c_j^2 \text{Var}(\hat{\beta}_j) + 2 \sum_{j=1}^{k-1} \sum_{j'=j+1}^k c_j c_{j'} \text{Cov}(\hat{\beta}_j, \hat{\beta}_{j'})$$

- l.** Suppose that one decides that there is no interaction between alcohol and smoking, that CLINIC is not a confounder but needs to be considered for reasons of precision, and that the only two significant interaction terms are between ALC and AGE and between SMK and PA. What is your “final” odds ratio expression for a moderate or severe hypertensive (i.e., $\text{BP} \geq 1$) that compares a heavy-drinking, 2+ pack-a-day smoker to a light-drinking nonsmoker, controlling for AGE, PA, and CLINIC?

2. A 2005 study was conducted to evaluate the influence of fear avoidance beliefs (FAB), chronicity of low back pain (CHR), and severity of low back pain (SEV) on disability (DIS) in 209 patients from seven different regions in Spain who were treated for low back pain (LBP) in health-care centers belonging to the Spanish National Health Service. The sample was balanced for acute, subacute, and chronic LBP (i.e., the three categories of the CHR variable). Validated scales and questionnaires were used to assess DIS and FAB during the first visit (day 1) and also 14 days later (day 15). The analyses performed in the study were linear regression analyses using disability measurement (DIS) treated as a continuous outcome variable. However, in the questions to follow, we will treat the disability outcome (DIS) as either a polytomous or an ordinal outcome variable in a logistic regression model. The variables being considered are listed as follows:

DIS = disability level at day 15 (0 = none, 1 = mild, 2 = moderate, 3 = severe)

FAB = fear avoidance measure at day 1 (low = 0, high = 1)

CHR = chronicity of LBP at day 1 (1 = acute, 2 = subacute, 3 = chronic)

SEV = severity of LBP at day 1 (1 = low, 2 = medium, 3 = high)

SL = sick leave status at day 1 (0 = not on sick leave, 1 = on sick leave)

SEX = male (= 0) or female (= 1)

AGE (continuous variable)

- State the logit form of a no-interaction proportional odds logistic model for the relationship of disability outcome DIS to FAB, CHR, SEV, SL, SEX, and AGE. Make sure to treat the variables CHR and SEV as nominal (rather than ordinal) variables and to treat AGE as a continuous variable. In stating this model, let “acute” denote the referent category for CHR, and let “low” denote the referent category for SEV.
- Based on the answer to part (a) above,
 - What is the formula for the *odds* for moderate or severe disability (DIS = 2 or 3) to none or mild disability (DIS = 0 or 1) for a subject with high fear avoidance behavior (FAB = 1), chronic low back pain (CHR = 3), and severe low back pain (SEV = 3) who is on sick leave, is female, and is 40 years old?
 - What is the formula for the *odds ratio* that compares the odds for moderate or severe disability (DIS = 2 or 3) to none or mild disability (DIS = 0 or 1) for a subject with high fear avoidance behavior (FAB = 1), chronic low back pain (CHR = 3), and high severity of low back pain (SEV = 3) to the corresponding odds for a subject with low fear avoidance behavior (FAB = 0), acute low back pain (CHR = 1), and low severity of low back pain (SEV = 1), controlling for SL, SEX, and AGE?
- Consider the following 4×2 table that describes the crude relationship between fear avoidance behavior (FAB) and disability outcome (DIS):

		FAB	
		1	0
DIS	3	a	b
	2	c	d
	1	e	f
	0	g	h

- i. If the proportional odds assumption is satisfied for these data, describe the 2×2 sub-tables whose corresponding odds ratios are assumed to be equal (i.e., use the table above to assign letters to the appropriate cells and corresponding odds ratios in the sub-tables below):

		FAB	
		1	0
DIS	3		
	0-2		
		$\widehat{OR} =$	

		FAB	
		1	0
DIS	2 or 3		
	0 or 1		
		$\widehat{OR} =$	

		FAB	
		1	0
DIS	1-3		
	0		
		$\widehat{OR} =$	

- ii. Give two reasons why one might want to consider using a polytomous logistic regression model instead of a proportional odds logistic regression model for analyzing these data.

Suppose that one decided to use polytomous logistic regression, instead of ordinal logistic regression, to analyze these data. Also, suppose that the variables CHR and SEV are treated as ordinal variables. Suppose also that the variables FAB, CHR, and SEV are considered as *exposure* (i.e., E) *variables*, with the variables SL, SEX, and AGE treated as *control* (i.e., V) *variables*.

Suppose further that this revised model is expanded to allow for two-way interactions (i.e., products of two variables) between FAB and each of the control variables SL, SEX, and AGE; two-way interactions between CHR and each of the control variables SL, SEX, and AGE; two-way interactions between SEV and each of the control variables SL, SEX, and AGE; and two-way interactions among FAB, CHR, and SEV. This “initial” model can thus be written as follows:

$$\begin{aligned} \ln [P(\text{DIS} = g|X)/P(\text{DIS} = 0|X)] &= \alpha_g + \beta_{1g}(\text{FAB} + \beta_{2g}\text{CHR} + \beta_{3g}(\text{SEV} + \gamma_{1g}\text{SL}) \\ &\quad + \gamma_{2g}(\text{SEX} + \gamma_{3g}\text{AGE}) + \delta_{F1g}(\text{FAB} \times \text{SL}) \\ &\quad + \delta_{F2g}(\text{FAB} \times \text{SEX}) + \delta_{F3g}(\text{FAB} \times \text{AGE}) \\ &\quad + \delta_{C1g}(\text{CHR} \times \text{SL}) + \delta_{C2g}(\text{CHR} \times \text{SEX}) \\ &\quad + \delta_{C3g}(\text{CHR} \times \text{AGE}) + \delta_{S1g}(\text{SEV} \times \text{SL}) \\ &\quad + \delta_{S2g}(\text{SEV} \times \text{SEX}) + \delta_{S3g}(\text{SEV} \times \text{AGE}) \\ &\quad + \delta_{FC1g}(\text{FAB} \times \text{CHR}) + \delta_{FS2g}(\text{FAB} \times \text{SEV}) \\ &\quad + \delta_{CS3g}(\text{CHR} \times \text{SEV}) \\ g &= 1, 2, 3 \end{aligned}$$

- d. Using the above polytomous logistic regression model, describe how one would *simultaneously test* for the significance of all $E_i V_j$ product terms in this model. Make sure to state the null hypothesis in terms of model parameters, describe the formula for the test statistic, and give the distribution and degrees of freedom of the test statistic under the null hypothesis.
- e. At the end of the interaction assessment stage, suppose that it was determined that the variables $(\text{CHR} \times \text{SEV})$ and $(\text{CHR} \times \text{SL})$ are the only product terms remaining in the model as significant interaction effects. For the *reduced model* obtained

from the interaction results, give a formula for the odds ratio that compares the odds for severe disability ($\text{DIS} = 3$) to mild disability ($\text{DIS} = 1$) for a subject with high fear avoidance behavior ($\text{FAB} = 1$), chronic low back pain ($\text{CHR} = 3$), and high severity low back pain ($\text{SEV} = 3$) to the corresponding odds for a subject with low fear avoidance behavior ($\text{FAB} = 0$), acute low back pain ($\text{CHR} = 1$), and low severity low back pain ($\text{SEV} = 1$), controlling for SL, SEX, and AGE.

References

- Ananth, C. V., and Kleinbaum, D. G. 1997. "Regression Models for Ordinal Responses: A Review of Methods and Applications." *International Journal of Epidemiology*, 26:1323–33.
- Sunenshine, R. H.; Wright, M. O.; Maragakis, L. L.; Harris, A. D.; Song, X.; Hebden, J.; Cosgrove, S. E.; Anderson, A.; Carnell, J.; Jernigan, D. B.; Kleinbaum, D. G.; Perl, T. M.; Standiford, H. C.; and Srinivasan, A. 2006. "Impact of Multidrug-Resistant *Acinetobacter* Infection on Mortality and Length of Hospitalization." Unpublished.
- Volkova, N.; McClellan, W.; Kleinbaum, D. G.; Klein, M.; and Flanders, W. D. 2006. "Neighborhood Economic Deprivation and Racial Differences in ESRD Incidence." Unpublished.

24

Poisson Regression Analysis

24.1 Preview

Poisson regression analysis is a regression technique available for modeling dependent variables that describe *count* (i.e., *discrete*) *data*. The purpose of this chapter is to discuss the Poisson regression model and several key features of the model, particularly how a rate ratio (*RR*) can be estimated using Poisson regression. We also use real-life data to demonstrate how Poisson regression may be applied. Maximum likelihood (ML) procedures are used to estimate the parameters in a Poisson regression model. Therefore, the general principles and inference-making procedures described in Chapter 22 on ML estimation carry over directly to the likelihood functions appropriate for Poisson regression analysis.

24.2 The Poisson Distribution

The methodology of Poisson regression analysis assumes that the underlying distribution of the response variable Y under consideration is Poisson. The Poisson probability distribution with parameter μ is given by the formula

$$\text{pr}(Y; \mu) = \frac{\mu^Y e^{-\mu}}{Y!}, \text{ where } Y = 0, 1, \dots, \infty \quad (24.1)$$

Theoretically, a Poisson random variable can take any nonnegative integer value. From (24.1), for example, the probability that Y takes the value 10 is

$$\text{pr}(Y = 10; \mu) = \frac{\mu^{10} e^{-\mu}}{10!} = \frac{\mu^{10} e^{-\mu}}{3,628,800}$$

This value of this probability changes as a function of the value of μ .

The Poisson distribution is often used to model the occurrence of rare events, such as the number of new cases of lung cancer developing in some population over a certain period of time or the number of automobile accidents occurring at a certain location per year. An interesting statistical attribute of the Poisson distribution is that $E(Y) = \text{Var}(Y) = \mu$ when Y has the Poisson distribution (24.1). Also, from statistical theory, the Poisson distribution provides a good approximation to the binomial distribution for rare events.

To illustrate an application of the Poisson distribution, consider the results of a study to investigate whether exposure to magnetic fields is associated with the development of breast cancer in men (a notably rare disease). In a 1976–1980 study of 50,582 male telecom workers in New York (N.Y.), two cases of breast cancer were found out of 206,667 person-years of follow-up (Matanoski, Breysse, and Elliott 1991). The question being addressed, therefore, is whether observing two breast cancers in men working in the telecom industry is (statistically) excessive when compared with the breast cancer rate in the general adult male population.

To answer this question, we first must recognize that the parameter that we are estimating for the N.Y. male telecom workers is a *rate*, not a proportion. By *rate*, we mean the frequency of occurrence (i.e., *count*) of an event (i.e., *developing breast cancer*) over an accumulation of person-time follow-up information (i.e., *person-years*). Mathematically, we define a rate, which we denote by λ , as

$$\lambda = \frac{E(Y)}{PT}$$

where $E(Y)$ denotes the expected value (or true average value) of the frequency Y in the population under study and PT denotes the accumulated person-time of follow-up. Typically, PT is obtained by summing up the follow-up time contributions of all persons studied, recognizing that not all study subjects may have been followed for the same amount of time. The estimated rate obtained from a sample is defined as

$$\hat{\lambda} = \frac{Y}{PT}$$

Because the numerator (Y) of an estimated rate is not a subset of the denominator (PT), a rate is not a proportion. Thus, the range of possible values for a rate is $0 \leq \lambda < \infty$, whereas the range for a proportion π is $0 \leq \pi \leq 1$. (See Kleinbaum 2003 for further discussion about the terms *rate* and *risk*, where only the latter is a proportion.)

In order to answer the question at hand, let us assume that published literature about breast cancer in men indicated that, in the general adult male population in N.Y., the expected breast cancer rate for men was about 15 new cases for every 10,000,000 person-years of follow-up. We may, therefore, address our question by carrying out a statistical test of hypothesis, where our null hypothesis can be stated as follows:

$$H_0: \lambda = \frac{15}{10,000,000} (= \text{expected rate for N.Y. male telecom workers})$$

Using the above expected rate, we can alternatively state the null hypothesis in terms of the expected number of male breast cancer cases:

$$H_0: \mu = E(Y) = 206,667 \times \frac{15}{10,000,000} = 0.31 \text{ expected cases}$$

If we now assume that Y , the observed number of cases, has the Poisson distribution with expected value $\mu = 0.31$ under H_0 , we can compute the probability of observing 2 or more male breast cancer cases using the Poisson distribution formula (24.1). This probability will, therefore, give us the P -value for a one-sided alternative to H_0 —namely, $H_A: \mu > 0.31$.

Since $\mu = 0.31$ under H_0 , the probability of obtaining any value of Y under H_0 is computed using the formula

$$\text{pr}(Y; 0.31) = \frac{0.31^Y e^{-0.31}}{Y!}$$

We can then compute the desired P -value as follows:

$$\begin{aligned} P\text{-value} &= \text{pr}(Y \geq 2 | \mu = 0.31) \\ &= 1 - \text{pr}(Y = 0 | \mu = 0.31) - \text{pr}(Y = 1 | \mu = 0.31) \\ &= 1 - \frac{0.31^0 e^{-0.31}}{0!} - \frac{0.31^1 e^{-0.31}}{1!} \\ &= 1 - 0.7334 - 0.2274 \\ &= 0.0392 \end{aligned}$$

Thus, we would reject H_0 at the 5% level and conclude that the breast cancer rate for male telecom workers in N.Y. is higher than that for the general adult male population in N.Y.

24.3 An Example of Poisson Regression

To illustrate the utility of Poisson regression analysis, let us consider a data analysis situation where Poisson regression has been used quite successfully. Table 24.1 gives nonmelanoma skin cancer data for women stratified by age in two metropolitan areas: Dallas–Ft. Worth and Minneapolis–St. Paul (Scotto, Kopf, and Urbach 1974). In this example, the dependent variable Y is a count, the number of cases of skin cancer. Since eight age strata and two metropolitan areas are involved, we let Y_{ij} denote the count for the i th age stratum in the j th area, where i ranges from 1 to 8 for the eight age groups and $j = 0$ (Minneapolis–St. Paul) or $j = 1$ (Dallas–Ft. Worth). We also let ℓ_{ij} denote the population size for the i th age stratum in the j th area. For these data, one analysis goal is to determine whether the risk for skin cancer adjusted for age is higher in one metropolitan area than in the other. The term *risk* in this context essentially means the probability associated with an event of interest—for example, the probability of developing skin cancer. We will let λ_{ij} denote the true (i.e., population) risk in the (i, j) th group. The ratio

$$RR_i = \frac{\lambda_{i1}}{\lambda_{i0}}$$

is commonly referred to as the *relative risk* or *risk ratio*, which in this case is the population risk for Dallas–Ft. Worth in the i th age group divided by the population risk for Minneapolis–St. Paul in the i th age group. If $RR_i = 1$, then the population risks are the same in the i th age group; if $RR_i > 1$, however, then the risk for Dallas–Ft. Worth is higher than the risk for Minneapolis–St. Paul in this age group.

TABLE 24.1 Comparison of incidence of nonmelanoma skin cancer among women in Minneapolis–St. Paul and Dallas–Ft. Worth

Age Group (yr)	Minneapolis–St. Paul		Dallas–Ft. Worth		Estimated Risk Ratio*
	No. of Cases	Population Size	No. of Cases	Population Size	
15–24	1	172,675	4	181,343	3.81
25–34	16	123,065	38	146,207	2.00
35–44	30	96,216	119	121,374	3.14
45–54	71	92,051	221	111,353	2.57
55–64	102	72,159	259	83,004	2.21
65–74	130	54,722	310	55,932	2.33
75–84	133	32,185	226	29,007	1.89
85+	40	8,328	65	7,583	1.78

Source: Adapted from Scott, Kopf, and Urbach (1974).

*With Minneapolis–St. Paul as the reference group.

© Cengage Learning

As indicated in the last column of Table 24.1, the estimated risk ratios in all age groups are greater than 1, which clearly suggests that the Dallas–Ft. Worth area has a higher overall incidence of skin cancer than Minneapolis–St. Paul. Our objective here is to use Poisson regression analysis to determine whether such a data pattern is statistically significant and to obtain an estimate of the overall risk ratio that adjusts for the effect of age.

Where does the Poisson distribution enter into this problem? Notice, first, that the count Y_{ij} is, in theory, a binomial random variable with mean $\mu_{ij} = \ell_{ij}\lambda_{ij}$. We know from statistical theory that the binomial distribution can be approximated by a Poisson distribution with the same mean, provided that the population size is large and the binomial probability parameter is small, so that the expected binomial count (i.e., the mean μ) is small relative to the population size. In other words, the Poisson distribution provides a good approximation to the binomial distribution for rare events. The data in Table 24.1 satisfy this requirement reasonably well, since all stratum-specific counts are quite small relative to the corresponding population sizes.

To develop a Poisson regression model for the above situation, we need to define a model for the expected number of skin cancer cases, $E(Y_{ij})$, in terms of the predictor variables of interest. Here two underlying predictor variables are of interest: “age” and “area.” Since “age” has been categorized into eight groups, we will use seven dummy variables to index them.¹ The variable “area,” which contains two categories, requires only one dummy variable.

¹ Alternatively, the model can be defined by using eight dummy variables for “age” and one dummy variable for “area”; when eight dummy variables are used for “age,” using an intercept term is redundant. The eight-dummy-variable alternative was used in the published analysis of this data set.

Thus, one possible model for the expected number of skin cancer cases in the (i, j) th group can be written as

$$E(Y_{ij}) = \mu_{ij} = \ell_{ij}\lambda_{ij} \quad i = 1, 2, \dots, 8; j = 0, 1$$

where

$$\ln \lambda_{ij} = \alpha + \sum_{m=1}^7 \gamma_m U_m + \beta E$$

with

$$U_m = \begin{cases} 1 & \text{if } m = i \\ 0 & \text{otherwise} \end{cases} \quad m = 1, 2, \dots, 7$$

$$E = \begin{cases} 1 & \text{if } j = 1 \quad (\text{Dallas-Ft. Worth}) \\ 0 & \text{if } j = 0 \quad (\text{Minneapolis-St. Paul}) \end{cases}$$

Using this model, we can write the risks λ_{ij} in terms of the parameters α , γ_i , and β to obtain

$$\ln \lambda_{i0} = \alpha + \gamma_i \quad \text{and} \quad \ln \lambda_{i1} = \alpha + \gamma_i + \beta \quad i = 1, 2, \dots, 7$$

and, for $i = 8$,

$$\ln \lambda_{80} = \alpha \quad \text{and} \quad \ln \lambda_{81} = \alpha + \beta$$

since $U_m = 0$, $m = 1, 2, \dots, 7$ for $i = 8$. Hence,

$$\ln \lambda_{i1} - \ln \lambda_{i0} = (\alpha + \gamma_i + \beta - \alpha - \gamma_i) = \beta \quad i = 1, 2, \dots, 7$$

Also,

$$\ln \lambda_{81} - \ln \lambda_{80} = (\alpha + \beta - \alpha) = \beta$$

In other words,

$$RR_i = \frac{\lambda_{i1}}{\lambda_{i0}} = \exp\left[\ln\left(\frac{\lambda_{i1}}{\lambda_{i0}}\right)\right] = \exp[\ln \lambda_{i1} - \ln \lambda_{i0}] = \exp[\beta] = e^\beta \quad i = 1, 2, \dots, 8$$

Thus, using this model, we can estimate the risk ratio for any age group by fitting the model, estimating the coefficient (β) of the E -variable, and then exponentiating this estimate. Since the estimated risk ratio e^β is independent of the age group i , we can interpret e^β as being an estimate of an overall risk ratio adjusted for age.

The example just described illustrates the type of model used in performing a Poisson regression analysis. In general, instead of having two variables (like age and area) to consider, we may have several (say, k) predictor variables X_1, X_2, \dots, X_k to examine. Nevertheless, the general method of fitting a Poisson regression model is still to use the Poisson model formulation to derive a likelihood function that can then be maximized so that parameter estimates, estimated standard errors, maximized likelihood statistics, and other statistical

information can be produced. Since packaged programs can now carry out such analyses, a user need only specify the model to be fit; the program then determines the likelihood function, maximizes it, and computes relevant statistics. We shall return later to the same example to illustrate methods of Poisson regression analysis numerically.

The preceding example (strictly speaking) involves a model for estimating the *risk* of developing a disease. A more general and popular application of Poisson regression involves modeling *rates of disease development* for different subgroups of interest. An estimated rate is generally defined as

$$\hat{\lambda} = \frac{Y}{\ell}$$

where Y is the observed count of adverse health outcomes (e.g., the number of cases of skin cancer or the number of new cases of heart disease) for a subgroup of interest and ℓ denotes the accumulated length of (disease-free) follow-up time for all persons in the subgroup. Thus, $\hat{\lambda}$ measures the number of adverse health outcomes (i.e., health failures) relative to the total amount of follow-up time for all persons in a given subgroup. If, for example, the data in Table 24.1 were based on a one-year follow-up study of the Minneapolis-St. Paul and Dallas-Ft. Worth populations, then the numbers in the table giving population size might be considered as *person-years* of follow-up time for the age-area subgroups. The ratio of two rates (e.g., $\lambda_{i1}/\lambda_{i0}$) is commonly referred to as a *rate ratio*. Other terms used are *incidence density ratio (IDR)* and *hazard ratio*.

24.4 Poisson Regression

24.4.1 General Considerations

We are now ready to describe the general Poisson regression analysis framework. The dependent variable Y is, as already mentioned, typically a count (of health failures, in this example) obtained for each of a number of subgroups that are described by a set of predictor variables X_1, X_2, \dots, X_p . For subgroup i , $i = 1, 2, \dots, n$, let Y_i denote the observed number of failures, let $X_{i1}, X_{i2}, \dots, X_{ip}$ denote the set of covariate values specific to subgroup i , and let ℓ_i denote the total length of follow-up time for all persons in that subgroup (also called *PT* earlier). For subgroup i , the rate of failures is denoted by λ_i .² In the Poisson regression model, the natural log of λ_i is described by the linear function

$$\ln(\lambda_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} = \beta_0 + \sum_{j=1}^k \beta_j X_{ij} \quad (24.2)$$

² In this presentation, we presume a biomedical scenario, with counts and rates relating to health failures and follow-up time contributed by persons under study. The language may be easily adapted to accommodate examples of counts/rates from a broad range of fields. For example, the rate of traffic accidents at different intersections may be modeled using Poisson regression, with Y_i = the count of traffic accidents and ℓ_i = the number of cars passing through the i th intersection during a specified time period. Another example would be a study of the rate of manufacturing defects among factory assembly lines, where Y_i = the count of defective parts and ℓ_i = the total number of parts produced on the i th assembly line during a specified time period.

By exponentiation of (24.1), the equation for the rate λ_i is thus

$$\lambda_i = e^{\beta_0 + \sum_{j=1}^k \beta_j X_{ij}} \quad (24.3)$$

Note that this is identical (except for the subscript i that identifies a specific subject) to the expression for the odds of a binary outcome for logistic regression (Chapter 22). Accordingly, the general form of the RR comparing two different covariate specifications \mathbf{X}_A and \mathbf{X}_B is identical to that for the odds ratio (OR) seen in (22.4), namely,

$$RR_{\mathbf{X}_A \text{ vs. } \mathbf{X}_B} = e^{\sum_{j=1}^k (X_{Aj} - X_{Bj}) \beta_j} \quad (24.4)$$

From the relationship $\lambda = E(Y)/\ell$ described in Section 24.2, one can readily see that $E(Y_i) = \lambda_i \ell_i$.³ Applying the properties of logarithms, one can further show that $\ln E(Y_i) = \ln(\lambda_i) + \ln(\ell_i)$. Using this equality in expression (24.2), we arrive at a second equivalent formulation of the Poisson regression model:

$$\ln E(Y_i) = \beta_0 + \sum_{j=1}^k \beta_j X_{ij} + \ln(\ell_i) \quad (24.5)$$

The model form (24.5) is more recognizable as a classical regression equation involving the dependent and independent variables and is indeed the form utilized in most computer Poisson regression procedures (such as SAS PROC GENMOD). Note that the quantity $\ln(\ell_i)$ must also be specified as part of the modeled data and is called the model *offset* (with a regression coefficient set equal to 1).

24.4.2 Theoretical Considerations

Thus far, we have introduced the Poisson regression model without emphasizing the Poisson distribution, which the Y_i response variables are assumed to follow.⁴ In this section, we discuss the general likelihood function for Poisson regression and the process by which ML estimates are obtained for the Poisson model. To be concise, in this section we symbolize the function (24.3) for λ_i that contains the regression equation as $\lambda(\mathbf{X}_i, \boldsymbol{\beta}) (> 0)$.

Under the assumption that Y_i is Poisson with mean μ_i ,⁴ so that

$$\text{pr}(Y_i; \mu_i) = \frac{\mu_i^{Y_i} e^{-\mu_i}}{Y_i!} \quad i = 1, 2, \dots, n \quad (24.6)$$

³ Alternatively, this may be written as $\mu_i = \lambda_i \ell_i$.

⁴ That the Poisson distribution is useful for modeling certain types of health count data can be loosely argued on the basis of the well-known Poisson approximation to the binomial distribution (see, e.g., Remington and Schork (1985), chap. 5). If $Y \sim \text{Bin}(n, \pi)$, and if n is large and π is very small, then Y has an approximate Poisson($\mu = n\pi$) distribution. For many health outcomes (e.g., the development of a rare disease), the length ℓ_i of follow-up time (analogous to n) is large, and the rate $\lambda(\mathbf{X}_i, \boldsymbol{\beta})$ of occurrence of the health outcome in question (analogous to π) is small, thus suggesting the Poisson model.

it follows from (24.3) and (24.6) that

$$\text{pr}(Y_i; \boldsymbol{\beta}) = \frac{[\ell_i \lambda(\mathbf{X}_i, \boldsymbol{\beta})]^{Y_i} e^{-\ell_i \lambda(\mathbf{X}_i, \boldsymbol{\beta})}}{Y_i!} \quad (24.7)$$

where $Y_i = 0, 1, \dots, \infty$ and $i = 1, 2, \dots, n$.

Note that the only real conceptual difference between Poisson regression and standard multiple linear regression is that the former involves the Poisson distribution, whereas the latter involves the normal distribution. In each instance, the analysis goal is the same—namely, to fit to the data a regression equation that will accurately model $E(Y)$ as a function of a set of predictor variables X_1, X_2, \dots, X_k .⁵

In the most general sense, then, regression analysis pertains to modeling the mean of the dependent variable under consideration as a function of certain predictor variables. The form of likelihood function that is used to estimate the regression coefficient set $\boldsymbol{\beta}$ is determined by the assumptions made about the distribution of that dependent variable.

As we did earlier to obtain the likelihood function (21.12), let us assume that the set $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ constitutes a mutually independent set of Poisson random variables, with Y_i having the probability distribution (24.7). Then, the *likelihood function for Poisson regression analysis* is of the general form

$$\begin{aligned} L(\mathbf{Y}; \boldsymbol{\beta}) &= \prod_{i=1}^n \text{pr}(Y_i; \boldsymbol{\beta}) = \prod_{i=1}^n \left\{ \frac{[\ell_i \lambda(\mathbf{X}_i, \boldsymbol{\beta})]^{Y_i} e^{-\ell_i \lambda(\mathbf{X}_i, \boldsymbol{\beta})}}{Y_i!} \right\} \\ &= \frac{\left\{ \prod_{i=1}^n [\ell_i \lambda(\mathbf{X}_i, \boldsymbol{\beta})]^{Y_i} \right\} \exp \left[-\sum_{i=1}^n \ell_i \lambda(\mathbf{X}_i, \boldsymbol{\beta}) \right]}{\prod_{i=1}^n Y_i!} \end{aligned} \quad (24.8)$$

where $E(Y_i) = \mu_i = \ell_i \lambda(\mathbf{X}_i, \boldsymbol{\beta})$, $i = 1, 2, \dots, n$.

Recall that the ML estimators $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ of $\beta_0, \beta_1, \dots, \beta_k$ are obtained from (24.8) as the solutions of the $(k + 1)$ equations

$$\frac{\partial}{\partial \beta_j} [\ln L(\mathbf{Y}; \boldsymbol{\beta})] = 0 \quad j = 0, 1, \dots, k \quad (24.9)$$

The solution to the set of ML equations given by (24.9) must generally be obtained by a computer-based iteration procedure. Frome (1983) discusses the use of algorithms for

⁵ If, in the likelihood (21.12) of Chapter 21, we replace $E(Y_i) = \beta_0 + \beta_1 X_i$ with, say, $E(Y_i) = \beta_0 + e^{\beta_1 X_i}$, then we change from a *linear* (in the regression coefficients) model to a *nonlinear* model and hence from a *linear* regression analysis to a *nonlinear* one. The major effect of this change is that we have to solve a set of nonlinear (as opposed to linear) likelihood equations in the β 's. This solution generally requires some sort of computer-assisted iteration procedure.

solving the system of equations (24.9). In particular, he argues for the use of a computational algorithm referred to as *iteratively reweighted least squares* (IRLS).⁶ Several statistical packages, such as SAS (using PROC GENMOD), can be utilized to find the ML estimator $\hat{\beta}$ of β based on the likelihood (24.8). In addition, the estimated covariance matrix $\hat{V}(\hat{\beta})$ of $\hat{\beta}$, measures of goodness of fit of the model under consideration, and certain regression diagnostic statistics (i.e., indices useful for detecting influential observations and multicollinearity) can be obtained as part of the computer output.

24.4.3 Skin Cancer Example

For an application of the above procedures, we return to the data in Table 24.1 describing nonmelanoma skin cancer data for women stratified by age in Minneapolis–St. Paul and Dallas–Ft. Worth (adapted from Scott et al. [1974] and reanalyzed by Frome and Checkoway [1985]). We previously considered the following Poisson regression model for the expected number of skin cancer cases in subgroup (i, j) , $i = 1, 2, \dots, 8$ and $j = 0, 1$:

$$E(Y_{ij}) = \mu_{ij} = \ell_{ij}\lambda_{ij}$$

where

$$\ln \lambda_{ij} = \alpha + \sum_{m=1}^7 \gamma_m U_m + \beta E$$

Here the U_m 's were 0–1 dummy variables indexing the age strata, and E was a 0–1 variable delineating metropolitan area or city (1 = Dallas–Ft. Worth, 0 = Minneapolis–St. Paul). For this model, the risk (or, more generally, rate) ratio

$$RR_i = \frac{\lambda_{i1}}{\lambda_{i0}}$$

reduced to the expression

$$RR_i = e^\beta$$

where e^β is independent of i and represents an overall rate ratio parameter adjusted for age.

The likelihood function L for the preceding model, based on the assumption that the count Y_{ij} follows the Poisson distribution with mean $\mu_{ij} = \ell_{ij}\lambda_{ij}$ and that the $\{Y_{ij}\}$ are a set of mutually independent random variables, is given by the expression

$$L = \prod_{i=1}^8 \left\{ \left[\frac{(\ell_{i0}\lambda_{i0})^{Y_{i0}} e^{-\ell_{i0}\lambda_{i0}}}{Y_{i0}!} \right] \left[\frac{(\ell_{i1}\lambda_{i1})^{Y_{i1}} e^{-\ell_{i1}\lambda_{i1}}}{Y_{i1}!} \right] \right\}$$

⁶ The fact that $E(Y_i) = \text{Var}(Y_i) = \mu_i = \ell_i\lambda_i(\mathbf{X}_i, \beta)$ means that the variance of the response variable is *not* constant (i.e., it varies as a function of ℓ_i and \mathbf{X}_i , thus requiring a weighted-least-squares regression analysis). And because this variance is a mathematical function of β , the weights in such a weighted regression analysis necessarily change as a function of the change in the estimate $\hat{\beta}$ at each step of the iteration process (i.e., a reweighting is required at each step). This is the reason for the terminology *iteratively reweighted least squares*, or IRLS for short.

where $\lambda_{i0} = \exp(\alpha + \gamma_i)$, $\lambda_{il} = \exp(\alpha + \gamma_i + \beta)$ for $i = 1, \dots, 7$, $\lambda_{80} = \exp(\alpha)$, and $\lambda_{81} = \exp(\alpha + \beta)$.

The use of a Poisson regression computer package would then maximize this likelihood function to produce the nine parameter estimates

$$\{\hat{\alpha}, \hat{\gamma}_1, \hat{\gamma}_2, \dots, \hat{\gamma}_7, \hat{\beta}\}$$

along with a (9×9) estimated covariance matrix. The computer output for these data, using SAS's GENMOD procedure, is given next. From this output, we see that the estimates of β and its standard error are

$$\hat{\beta} = 0.804, S_{\hat{\beta}} = 0.0522$$

Thus, the point estimate of the adjusted rate ratio is given by

$$e^{\hat{\beta}} = e^{0.804} = 2.2342$$

SAS Output (PROC GENMOD) for No-Interaction Model (Model 1) for Non-Melanoma Skin Cancer Data

The GENMOD Procedure

CRITERIA FOR ASSESSING GOODNESS OF FIT			
Criterion	DF	Value	Value/DF
Deviance	7	8.2585	1.1798
Scaled Deviance	7	8.2585	1.1798
Pearson Chi-Square	7	8.1273	1.1610
Scaled Pearson X2	7	8.1273	1.1610
Log Likelihood	.	7201.8317	

ANALYSIS OF MAXIMUM LIKELIHOOD PARAMETER ESTIMATES

E →

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.4834	0.1037	-5.6866 -5.2802	2797.32	<.0001
CITY	1	0.8039	0.0522	0.7016 0.9062	237.13	<.0001
U1	1	-6.1742	0.4577	-7.0713 -5.2770	181.94	<.0001
U2	1	-3.5440	0.1675	-3.8723 -3.2157	447.75	<.0001
U3	1	-2.3268	0.1275	-2.5767 -2.0770	333.22	<.0001
U4	1	-1.5790	0.1138	-1.8021 -1.3559	192.41	<.0001
U5	1	-1.0869	0.1109	-1.3043 -0.8695	96.04	<.0001
U6	1	-0.5288	0.1086	-0.7417 -0.3159	23.70	<.0001
U7	1	-0.1157	0.1109	-0.3331 0.1018	1.09	0.2972
Scale	0	1.0000	0.0000	1.0000 1.0000		

An approximate large-sample 95% confidence interval for $e^{\hat{\beta}}$ is calculated as

$$\begin{aligned}\exp[\hat{\beta} \pm 1.96 S_{\hat{\beta}}] &= \exp[0.804 \pm 1.96(0.0522)] \\ &= \exp(0.804 \pm 0.1023)\end{aligned}$$

which gives the 95% confidence limits

$$(e^{0.7017}, e^{0.9063}) = (2.0172, 2.4751)$$

A large-sample test of $H_0: \beta = 0$ versus $H_A: \beta \neq 0$ can be based on the Wald statistic

$$Z = \frac{\hat{\beta} - 0}{S_{\hat{\beta}}}$$

which is approximately $N(0, 1)$ under $H_0: \beta = 0$.

For our example,

$$Z = \frac{0.804 - 0}{0.0522} = 15.40 \quad (P\text{-value} \approx 0)$$

Thus, this particular Poisson regression analysis indicates that there is a statistically significant effect due to area and that the overall (adjusted for age) rate of nonmelanoma skin cancer in women in Dallas–Ft. Worth is approximately 2.2 times the corresponding adjusted rate for women in Minneapolis–St. Paul; a 95% confidence interval for the (adjusted) rate ratio is (2.0172, 2.4751). We will return to this example to illustrate how to evaluate interaction, confounding, and goodness of fit.

24.5 Measures of Goodness of Fit

Measures of the goodness of fit of Poisson regression models are obtained via comparisons of maximized likelihood values. Suppose that Y_i has the Poisson distribution and that Y_1, Y_2, \dots, Y_n are mutually independent; then, expressed as a general function of $\mu_1, \mu_2, \dots, \mu_n$ (i.e., ignoring the predictors X_1, X_2, \dots, X_k completely), the likelihood function takes the form

$$L(\mathbf{Y}; \boldsymbol{\mu}) = \prod_{i=1}^n \frac{\mu_i^{Y_i} e^{-\mu_i}}{Y_i!} = \frac{\left(\prod_{i=1}^n \mu_i^{Y_i}\right) \exp\left(-\sum_{i=1}^n \mu_i\right)}{\prod_{i=1}^n Y_i!} \quad (24.10)$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)$. The system of ML equations

$$\frac{\partial}{\partial \mu_i} [\ln L(\mathbf{Y}; \boldsymbol{\mu})] = 0 \quad i = 1, 2, \dots, n$$

leads to the solution $\hat{\mu}_i = Y_i$, $i = 1, 2, \dots, n$. Thus, the maximized likelihood value for the likelihood function (24.10) is

$$L(\mathbf{Y}; \hat{\boldsymbol{\mu}}) = \frac{\left(\prod_{i=1}^n Y_i^{Y_i} \right) \exp\left(-\sum_{i=1}^n Y_i\right)}{\prod_{i=1}^n Y_i!} \quad (24.11)$$

where $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_n) = (Y_1, Y_2, \dots, Y_n)$.

The value of the maximized likelihood $L(\mathbf{Y}; \hat{\boldsymbol{\mu}})$ based on (24.10) will be larger (for any set of data) than that achieved by maximizing a likelihood such as (24.8) when $(k + 1) < n$. This is because (24.10) imposes no restrictions on the structure of μ_i , whereas (24.8) imposes the restriction $\mu_i = \ell_i \lambda(\mathbf{X}_i, \boldsymbol{\beta})$. In other words, (24.8) can be thought of as the likelihood function under H_0 : $\mu_i = \ell_i \lambda(\mathbf{X}_i, \boldsymbol{\beta})$, $i = 1, 2, \dots, n$, whereas (24.10) is the likelihood under H_A : “ μ_i is unrestricted in structure, $i = 1, 2, \dots, n$.”

Thus, if $L(\mathbf{Y}; \hat{\boldsymbol{\beta}})$ is the maximized likelihood value under (24.8), where $\hat{\boldsymbol{\beta}}$ is the ML estimator of $\boldsymbol{\beta}$, then

$$-2 \ln \left[\frac{L(\mathbf{Y}; \hat{\boldsymbol{\beta}})}{L(\mathbf{Y}; \hat{\boldsymbol{\mu}})} \right] \quad (24.12)$$

is a likelihood-ratio-type statistic reflecting the goodness of fit of the model $\mu_i = \ell_i \lambda(\mathbf{X}_i, \boldsymbol{\beta})$ relative to the model where no structure has been imposed on μ_i . Since the objective of any regression analysis is to obtain a parsimonious description of the data, the model $\mu_i = \ell_i \lambda(\mathbf{X}_i, \boldsymbol{\beta})$ involving $(k + 1)$ parameters will (we hope) provide a maximized likelihood value almost as large as can be obtained by the baseline (and uninformative) model that involves as many parameters (namely, n) as data points. By “almost as large,” we mean that $L(\mathbf{Y}; \hat{\boldsymbol{\beta}})$ will not be significantly smaller than $L(\mathbf{Y}; \hat{\boldsymbol{\mu}})$ based on a likelihood ratio (LR) test using (24.12).

The quantity

$$D(\hat{\boldsymbol{\beta}}) = -2 \ln \left[\frac{L(\mathbf{Y}; \hat{\boldsymbol{\beta}})}{L(\mathbf{Y}; \hat{\boldsymbol{\mu}})} \right] \quad (24.13)$$

is the goodness-of-fit statistic employed to assess whether $L(\mathbf{Y}; \hat{\boldsymbol{\beta}})$ is significantly less than $L(\mathbf{Y}; \hat{\boldsymbol{\mu}})$ and thus to suggest meaningful lack of fit to the data of the assumed regression model $\mu_i = \ell_i \lambda(\mathbf{X}_i, \boldsymbol{\beta})$. The quantity $D(\hat{\boldsymbol{\beta}})$ is also called the *deviance* for the Poisson regression model $\mu_i = \ell_i \lambda(\mathbf{X}_i, \boldsymbol{\beta})$, and it can be thought of as a measure of residual variation about (or deviation from) the fitted model. Under H_0 : $\mu_i = \ell_i \lambda(\mathbf{X}_i, \boldsymbol{\beta})$, the deviance $D(\hat{\boldsymbol{\beta}})$ is typically (although not strictly legitimately) assumed to have (for large samples) an approximate chi-square distribution with $(n - k - 1)$ degrees of freedom, where n is the number of parameters (i.e., the number of subgroups, cells, or categories) specified in the likelihood (24.10) and $(k + 1)$ is the number of parameters (i.e., β_j 's) in the likelihood (24.8).

Thus, a very approximate test for goodness of fit of the model $\mu_i = \ell_i \lambda(\mathbf{X}_i, \boldsymbol{\beta})$ to a given data set can be performed by comparing the calculated value of $D(\hat{\boldsymbol{\beta}})$ to an appropriate upper-tail value of the chi-square distribution with $(n - k - 1)$ degrees of freedom.

With $\hat{Y}_i = \ell_i \lambda(\mathbf{X}_i, \hat{\boldsymbol{\beta}})$ denoting the predicted response in cell i for the likelihood function (24.8), the quantity (24.13) can be written in the form

$$D(\hat{\boldsymbol{\beta}}) = 2 \sum_{i=1}^n \left[Y_i \ln\left(\frac{Y_i}{\hat{Y}_i}\right) - (Y_i - \hat{Y}_i) \right] \quad (24.14)$$

Hence, $D(\hat{\boldsymbol{\beta}})$ behaves like $\text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ in standard multiple linear regression analysis. When the fitted model exactly predicts the observed data (i.e., $Y_i = \hat{Y}_i, i = 1, 2, \dots, n$), then $D(\hat{\boldsymbol{\beta}}) = 0$; and the larger the discrepancy between observed and predicted responses, the larger the value of $D(\hat{\boldsymbol{\beta}})$.

When the predicted values are all of reasonable size (i.e., $\hat{Y}_i > 3, i = 1, 2, \dots, n$), then (24.14) can be reasonably approximated by the more familiar Pearson-type observed-versus-predicted chi-square statistic of the form

$$\chi^2 = \sum_{i=1}^n \frac{(Y_i - \hat{Y}_i)^2}{\hat{Y}_i} \quad (24.15)$$

As a word of caution, the statistic (24.15) can be misleadingly large when certain \hat{Y}_i -values are very small.

The deviances for various models in a hierarchical class can be used to produce LR tests. In particular, consider again the likelihood (24.8) involving the parameter set $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)$, with deviance $D(\hat{\boldsymbol{\beta}})$ given by (24.13). Now, for $0 < r < k$, suppose that we wish to test whether the last $(k - r)$ parameters in $\boldsymbol{\beta}$ are equal to 0; that is, our null hypothesis is $H_0: \beta_{r+1} = \beta_{r+2} = \dots = \beta_k = 0$. Under H_0 , the (null hypothesis) likelihood can be obtained by replacing $\boldsymbol{\beta}$ in (24.8) with $\boldsymbol{\beta}_r$, where

$$\boldsymbol{\beta}_r = (\beta_0, \beta_1, \dots, \beta_r, 0, 0, \dots, 0)$$

If we denote this likelihood function by $L(\mathbf{Y}; \boldsymbol{\beta}_r)$ and if $\hat{\boldsymbol{\beta}}_r$ is the ML estimator of $\boldsymbol{\beta}_r$ using $L(\mathbf{Y}; \boldsymbol{\beta}_r)$, then the LR test of H_0 is performed using the test statistic

$$-2 \ln \left[\frac{L(\mathbf{Y}; \hat{\boldsymbol{\beta}}_r)}{L(\mathbf{Y}; \hat{\boldsymbol{\beta}})} \right] \quad (24.16)$$

which has approximately a chi-square distribution with $(k - r)$ degrees of freedom for large samples when H_0 is true.

Furthermore, expression (24.16) is exactly equal to the deviance difference

$$D(\hat{\boldsymbol{\beta}}_r) - D(\hat{\boldsymbol{\beta}}) \quad (24.17)$$

To see this, recall the general definition of $D(\hat{\beta})$ given by expression (24.13). Using (24.13) and (24.17), we have

$$\begin{aligned} D(\hat{\beta}_r) - D(\hat{\beta}) &= -2 \ln \left[\frac{L(\mathbf{Y}; \hat{\beta}_r)}{L(\mathbf{Y}; \hat{\mu})} \right] + 2 \ln \left[\frac{L(\mathbf{Y}; \hat{\beta})}{L(\mathbf{Y}; \hat{\mu})} \right] \\ &= -2 \ln \left[\frac{L(\mathbf{Y}; \hat{\beta}_r)}{L(\mathbf{Y}; \hat{\beta})} \right] \end{aligned}$$

which is exactly the LR test statistic (24.16). Under $H_0: \beta_{r+1} = \beta_{r+2} = \cdots = \beta_k = 0$, the difference $D(\hat{\beta}_r) - D(\hat{\beta})$ has approximately a chi-square distribution with $(k - r)$ degrees of freedom in large samples.

Thus, when we use Poisson regression to analyze a set of data, members of a set of candidate models within a hierarchical class can be compared by considering differences between pairs for deviances for these models.

24.6 Continuation of Skin Cancer Data Example

We again consider the data in Table 24.1 giving skin cancer counts for women stratified by age in Minneapolis–St. Paul and Dallas–Ft. Worth. For these data, we used ML estimation to fit the following Poisson regression model for the expected number of skin cancer cases:

$$E(Y_{ij}) = \mu_{ij} = \ell_{ij}\lambda_{ij}$$

where

$$\ln \lambda_{ij} = \alpha + \sum_{m=1}^7 \gamma_m U_m + \beta E$$

The set $\{U_m\}$ contains 0–1 dummy variables indexing the age strata, and E contrasts metropolitan areas (1 = Dallas–Ft. Worth, 0 = Minneapolis–St. Paul). We will refer to this model as Model 1. For this model, the estimated rate ratio adjusted for age was $e^{\hat{\beta}} = 2.2342$, and a 95% confidence interval for the true adjusted rate ratio was (2.0172, 2.4751). Also, a large-sample test for $H_0: \beta = 0$ versus $H_A: \beta \neq 0$ yielded a Z -statistic of 15.40 (P -value ≈ 0), which is highly significant.

Two additional questions are of interest for these data:

1. Is “age” an effect modifier? That is, does the “area” or “city” effect (as measured by a rate ratio parameter) differ for different age strata?
2. If the answer to question 1 is no, is “age” a confounder? That is, does “age” need to be in the model in some form in order to produce a valid estimate of the “area” effect?

At this point, readers may wish to review the definitions and properties of effect modifiers and confounders discussed in Chapter 11.

To answer question 1 directly, we could modify model 1 to include interaction terms, as follows:

$$\text{Model 2: } \ln \lambda_{ij} = \alpha + \sum_{m=1}^7 \gamma_m U_m + \beta E + \sum_{m=1}^7 \delta_m (EU_m)$$

Using the preceding interaction model, we can test for effect modification by testing $H_0: \delta_1 = \delta_2 = \dots = \delta_7 = 0$, using an LR χ^2 statistic with 7 degrees of freedom. This test involves comparing model 1 (without interaction terms) to model 2 (with seven EU_m terms). Earlier, we looked at the computer output based on fitting model 1. The computer output resulting from fitting model 2 (again using SAS's GENMOD procedure) is presented next. From this latter block of output, we see that the deviance for model 2 is exactly zero because model 2 fits the data perfectly (i.e., we have fit a model with 16 parameters to a data set of size $n = 16$). Thus, the LR test statistic for testing $H_0: \delta_1 = \delta_2 = \dots = \delta_7 = 0$ is obtained by subtracting the deviance for model 2 (i.e., zero) from the deviance for model 1, which is just the deviance for model 1. Therefore, an equivalent way to carry out this particular interaction test is to use the deviance previously obtained for model 1, which has the value 8.26. When compared with upper-tail chi-square values for which

$$\begin{aligned} \text{d.f.} &= [\text{Number of } Y_{ij}'\text{'s}] - [\text{Number of parameters in model 1}] \\ &= 16 - 9 = 7 \end{aligned}$$

the value 8.26 provides no evidence for lack of fit of model 1 (i.e., there are no large deviations of observed Y_{ij} values from predicted \hat{Y}_{ij} values). This indicates that adding more terms (e.g., of the form EU_k) to model 1 will not significantly improve the fit of that model.

Edited SAS Output (PROC GENMOD) for Model 2 for Non-Melanoma Skin Cancer Data

The GENMOD Procedure

CRITERIA FOR ASSESSING GOODNESS OF FIT			
Criterion	DF	Value	Value/DF
Deviance	0	0.0000	.
Scaled Deviance	0	0.0000	.
Pearson Chi-Square	0	0.0000	.
Scaled Pearson X2	0	0.0000	.
Log Likelihood	.	7205.9610	

ANALYSIS OF MAXIMUM LIKELIHOOD PARAMETER ESTIMATES					
Parameter	DF	Estimate	Std Error	Chi-Square	Pr > ChiSq
Intercept	1	-5.3385	0.1581	1139.98	<.0001
CITY	1	0.5792	0.2010	8.31	0.0039
U1	1	-6.7207	1.0124	44.07	<.0001
U2	1	-3.6094	0.2958	148.89	<.0001

(continued)

ANALYSIS OF MAXIMUM LIKELIHOOD PARAMETER ESTIMATES					
Parameter	DF	Estimate	Std Error	Chi-Square	Pr > ChiSq
U3	1	-2.7347	0.2415	128.20	<.0001
U4	1	-1.8289	0.1977	85.58	<.0001
U5	1	-1.2232	0.1866	42.99	<.0001
U6	1	-0.7040	0.1808	15.16	<.0001
U7	1	-0.1504	0.1803	0.70	0.4042
CU1	1	0.7581	1.1360	0.45	0.5045
CU2	1	0.1135	0.3594	0.10	0.7523
CU3	1	0.5664	0.2866	3.91	0.0481
CU4	1	0.3659	0.2429	2.27	0.1320
CU5	1	0.2126	0.2325	0.84	0.3604
CU6	1	0.2679	0.2265	1.40	0.2368
CU7	1	0.0549	0.2288	0.06	0.8102

EU_m
 $m = 1, 2, \dots, 7$

To answer question 2 about confounding, we need to see whether $\hat{\beta}$, or $e^{\hat{\beta}}$, changes meaningfully if we ignore (i.e., don't control for) "age." In particular, we need to drop the age terms (i.e., the $\sum_{m=1}^7 \gamma_m U_m$ component) from model 1 to see whether the estimated coefficient of E changes meaningfully from its value of 0.804 (or whether the estimate of the rate ratio changes meaningfully from its value of $e^{\hat{\beta}} = 2.2342$). If we fit

Model 3: $\ln \lambda_{ij} = \alpha + \beta_0 E$

to these data, we obtain $\hat{\beta}_0 = 0.743$ and a *crude* rate ratio estimate of

$$\widehat{RR}_c = e^{\hat{\beta}_0} = e^{0.743} = 2.1024$$

There is enough change here to suggest that "age" is a confounder and so should be controlled at the analysis stage.

To this point, we have treated age as a *categorical* variable, using seven terms of the form $\gamma_m U_m$ ($m = 1, 2, \dots, 7$) plus a constant α in model 1 to reflect the eight age strata. An alternative analysis that treats age as an interval variable is suggested by plotting $\ln \lambda_{ij}$ versus $\ln T_i$, where

$$T_i = \frac{[\text{Midpoint of } i\text{th age interval}] - 15}{35} \quad i = 1, 2, \dots, 8$$

These plots are presented in Figure 24.1, and the values used for the plots are given in Table 24.2. Figure 24.1 shows that the plotted points for each city graph essentially as a straight line, with the two lines being essentially parallel. These results suggest using a parsimonious model involving only a single linear effect of "age," with no interaction terms involving "age" and E .

In particular, consider the following model:

Model 4: $\ln \lambda_{ij} = \alpha + \theta \ln T_i + \beta E$

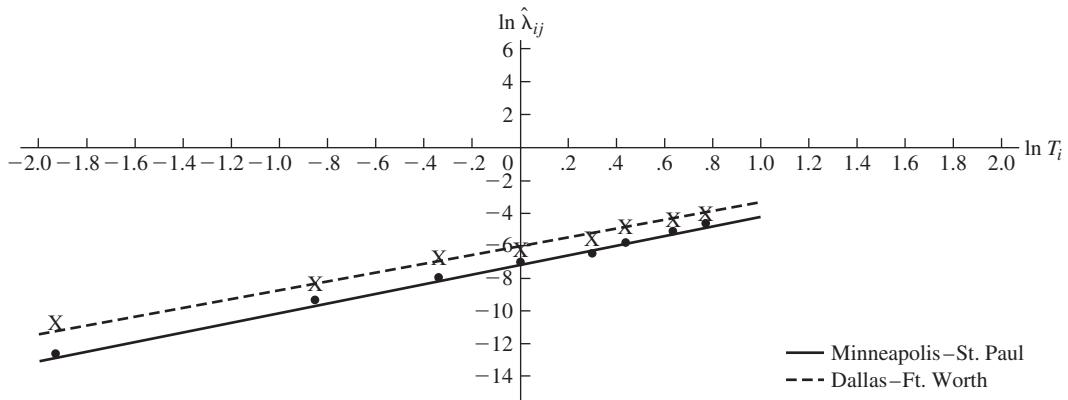


FIGURE 24.1 Plots of $\ln \hat{\lambda}_{ij}$ versus $\ln T_i$, where $T_i = [(\text{Midpoint of } i\text{th age interval}) - 15]/35$, using values in Table 24.2 from the study of nonmelanoma skin cancer in two cities (Scotto et al. 1974)

TABLE 24.2 Values of $\ln \hat{\lambda}_{ij}$ and $\ln T_i$, where $T_i = [(\text{Midpoint of } i\text{th age interval}) - 15]/35$, using the data in Table 24.1 from the study of nonmelanoma skin cancer in two cities (Scotto et al. 1974)

Age Group	Midpoint	T_i	$\ln T_i$	$\ln \hat{\lambda}_{i0}$	$\ln \hat{\lambda}_{i1}$
15–24	20	0.1429	-1.94591	-11.5266	-10.7239
25–34	30	0.4286	-0.84730	-9.0137	-8.2110
35–44	40	0.7143	-0.33647	-7.8453	-7.0426
45–54	50	1.0000	0.00000	-7.0757	-6.2730
55–64	60	1.2857	0.25131	-6.5009	-5.6982
65–74	70	1.5714	0.45199	-6.0419	-5.2392
75–84	80	1.8571	0.61904	-5.6598	-4.8571
85+	85–95	2.1429	0.76214	-5.3325	-4.5298

$\hat{\lambda}_{i0} = Y_{i0}/\ell_{i0}$ = Crude rate for Minneapolis-St. Paul

$\hat{\lambda}_{i1} = Y_{i1}/\ell_{i1}$ = Crude rate for Dallas-Ft. Worth

© Cengage Learning

Model 4 says that

$$\lambda_{i0} = e^\alpha T_i^\theta \quad \text{and} \quad \lambda_{i1} = e^\alpha T_i^\theta e^\beta$$

so that

$$RR_i = \frac{\lambda_{i1}}{\lambda_{i0}} = e^\beta$$

When $i = 4$ (so that $T_4 = 1$) and $j = 0$, we have $\lambda_{40} = e^\alpha T_4^\theta = e^\alpha(1)^\theta = e^\alpha$, so $\ln \lambda_{40} = \alpha$. Hence, α (the intercept term in model 4) is the natural logarithm of the rate for the 45- to 54-year-old age group ($i = 4$) in Minneapolis-St. Paul ($j = 0$). The following block of computer output (using SAS's GENMOD procedure) is based on fitting model 4.

Edited SAS Output (PROC GENMOD) for Model 4 for Non-Melanoma Skin Cancer Data

The GENMOD Procedure

CRITERIA FOR ASSESSING GOODNESS OF FIT			
Criterion	DF	Value	Value/DF
Deviance	13	14.5693	1.1207
Scaled Deviance	13	13.0000	1.0000
Pearson Chi-Square	13	14.3803	1.1062
Scaled Pearson X2	13	12.8314	0.9870
Log Likelihood		6423.2802	

ANALYSIS OF MAXIMUM LIKELIHOOD PARAMETER ESTIMATES						
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept	1	-7.0757	0.0504	-7.1745 -6.9769	19696.9	<.0001
city	1	0.8027	0.0552	0.6944 0.9109	211.18	<.0001
InT	1	2.2873	0.0664	2.1573 2.4174	1188.11	<.0001
Scale	0	1.0586	0.0000	1.0586 1.0586		

From the above computer output, the following point estimates and standard errors can be obtained:

$$\hat{\alpha} = -7.076 \quad S_{\hat{\alpha}} = 0.050$$

$$\hat{\beta} = 0.803 \quad S_{\hat{\beta}} = 0.055$$

$$\hat{\theta} = 2.287 \quad S_{\hat{\theta}} = 0.066$$

The deviance value for model 4 is 14.57, with $(16 - 3) = 13$ degrees of freedom, indicating a good fit to the data. The estimated adjusted rate ratio for model 4 is

$$e^{\hat{\beta}} = e^{0.803} = 2.2315$$

The 95% confidence interval for e^β is given by

$$\begin{aligned} \exp[0.803 \pm 1.96(0.055)] &= \exp(0.803 \pm 0.1078) \\ &= (e^{0.6952}, e^{0.9108}) \\ &= (2.0041, 2.4863) \end{aligned}$$

TABLE 24.3 Poisson ANOVA table for the skin cancer data of Table 24.1

	Model for $\ln \lambda_{ij}$	Number of Parameters	D(β)	d.f.
Model 3:	α	1	2789.7	15
	$\alpha + \beta E$	2	2569.1	14
	$\alpha + \theta \ln T_i$	2	272.6	14
Model 4:	$\alpha + \theta \ln T_i + \beta E$	3	14.6	13
	$\alpha + \sum_{m=1}^7 \gamma_m U_m$	8	266.7	8
Model 1:	$\alpha + \sum_{m=1}^7 \gamma_m U_m + \beta E$	9	8.3	7
	α_{ij}	16	0.0	0

© Cengage Learning

Thus, using model 4 leads to essentially the same conclusions obtained using model 1, with a very small gain in precision (since the confidence interval for e^β using model 4 is slightly narrower than the one using model 1). Since model 4 contains fewer parameters than model 1, without compromising on validity or precision and since model 4 describes the linear effect of “age” in a clear-cut manner, it is the model of choice!

A summary of the results of fitting various models to the data set under consideration can be presented in a Poisson ANOVA table, as illustrated in Table 24.3. As this table indicates, models 1 and 4 clearly have the best deviance values, but model 4 is our choice for the “best” model. Model 1 has a deviance of 8.3, whereas model 4 has a deviance of 14.6—but model 1 *should* have a smaller deviance (i.e., should fit the data better) than model 4, since it contains nine parameters, whereas model 4 contains only three. The issue is really whether model 1 fits the data *significantly* better than model 4. To address this issue, we can look at the difference in the two deviance values (namely, $14.6 - 8.3 = 6.3$), just as we would look at the difference in SSE values in standard multiple regression analysis. Let us assume that

$$(\text{Deviance for model 4}) - (\text{Deviance for model 1})$$

is very approximately a chi-square random variable for large n with $(13 - 7) = 6$ degrees of freedom. Since $\chi^2_{6,0.60} = 6.211$, the P -value for testing H_0 : “Model 1 and model 4 fit the data equally well” is about .40. Hence, there is absolutely no evidence that model 1 provides a better fit to the data than model 4.

Finally, a pseudo- R^2 for model 4 can be computed as

$$\text{Pseudo-}R^2 = \frac{2789.7 - 14.6}{2789.7} = 0.9948$$

which indicates a superb fit to the data.

24.7 A Second Illustration of Poisson Regression Analysis

We now consider an example given by Frome (1983). The data to be used appear in Table 24.4. The basic response variable Y is “number of lung cancer deaths,” which is assumed to have a Poisson distribution. More specifically, Y_{ij} denotes the observed number of lung cancer deaths in row i and column j , $i = 1, 2, \dots, 9$ and $j = 1, 2, \dots, 7$; thus, there are $n = 9 \times 7 = 63$ subgroups. The rows represent “years of smoking” (defined as age minus 20 years) in 5-year categories from 15 through 19 to 55 through 59; the columns represent “number of cigarettes smoked per day,” starting at 0 for nonsmokers and going up to 35 or more for the heaviest smokers. The variable ℓ_{ij} denotes the number of man-years at risk for cell (i, j) . The variable T_i , defined as the midpoint of the i th “years of smoking” category divided by 42.5, will be employed when we fit some dose-response models to the data in Table 24.4; the variable D_j will denote the dosage level variable for the j th “number of cigarettes smoked per day” category.

One particular form of failure rate model $\lambda(\mathbf{X}_{ij}, \boldsymbol{\beta})$ to be fit to these data is the standard two-way cross-classification model *in exponentiated form*. (Failure rates are always

TABLE 24.4 Man-years at risk (ℓ_{ij}) and observed number (Y_{ij}) of lung cancer deaths (in parentheses)

Years of Smoking*	$42.5T_i$	Number of Cigarettes Smoked per Day						
		0 ($D_1 = 0$)	1–9 ($D_2 = 5.2$)	10–14 ($D_3 = 11.2$)	15–19 ($D_4 = 15.9$)	20–24 ($D_5 = 20.4$)	25–34 ($D_6 = 27.4$)	35+ ($D_7 = 40.8$)
15–19	17.5	10,366 (1)	3,121 (0)	3,577 (0)	4,317 (0)	5,683 (0)	3,042 (0)	670 (0)
20–24	22.5	8,162 (0)	2,937 (0)	3,286 (1)	4,214 (0)	6,385 (1)	4,050 (1)	1,166 (0)
25–29	27.5	5,969 (0)	2,288 (0)	2,546 (1)	3,185 (0)	5,483 (1)	4,290 (4)	1,482 (0)
30–34	32.5	4,496 (0)	2,015 (0)	2,219 (2)	2,560 (4)	4,687 (6)	4,268 (9)	1,580 (4)
35–39	37.5	3,512 (0)	1,648 (1)	1,826 (0)	1,893 (0)	3,646 (5)	3,529 (9)	1,336 (6)
40–44	42.5	2,201 (0)	1,310 (2)	1,386 (1)	1,334 (2)	2,411 (12)	2,424 (11)	924 (10)
45–49	47.5	1,421 (0)	927 (0)	988 (2)	849 (2)	1,567 (9)	1,409 (10)	556 (7)
50–54	52.5	1,121 (0)	710 (3)	684 (4)	470 (2)	857 (7)	663 (5)	255 (4)
55–59	57.5	826 (2)	606 (0)	449 (3)	280 (5)	416 (7)	284 (3)	104 (1)

Note: If ℓ_{ij} really equalled the number of people in cell (i, j) from which the observed number Y_{ij} of lung cancer cases developed, then Y_{ij} could be treated as a binomial random variable with sample size ℓ_{ij} and unknown probability (or risk) of lung cancer death π_{ij} , in which case categorical data analysis methods might be utilized (although having several cells with very few deaths is problematic).

The quantity T_i is the midpoint of the i th “years of smoking” category divided by 42.5; D_j is the dosage level variable for the j th “cigarettes per day” category.

*Age minus 20 years

nonnegative, and using an exponential function ensures that this will be the case for all estimated rates.) In particular, consider modeling the rate $\lambda(\mathbf{X}_{ij}, \boldsymbol{\beta}) \equiv \lambda_{ij}$ for cell (i, j) as

$$\lambda_{ij} = e^{\mu + \alpha_i + \delta_j} \quad (24.18)$$

where μ is the overall mean, α_i is the fixed effect of the i th row, and δ_j is the fixed effect of the j th column. Here, as in standard two-way ANOVA, we impose the constraints $\sum_{i=1}^9 \alpha_i = \sum_{j=1}^7 \delta_j = 0$, so a total of $1 + (9 - 1) + (7 - 1) = 15$ parameters must be estimated using model (24.18).

As with standard regression model representations of ANOVA-type data (see elsewhere in this book), the “ X ” variables underlying model (24.18) are the usual dummy variables used to index the various rows and columns; their appearance has been suppressed for notational convenience.

The Poisson model-based likelihood function for the data in Table 24.4 and under model (24.18) is

$$\prod_{i=1}^9 \prod_{j=1}^7 \left[\frac{(\lambda_{ij})^{Y_{ij}} e^{-\lambda_{ij}}}{(Y_{ij})!} \right]$$

where $\lambda_{ij} = \exp(\mu + \alpha_i + \delta_j)$, $\alpha_9 = -\sum_{i=1}^8 \alpha_i$, and $\delta_7 = -\sum_{j=1}^6 \delta_j$. For the data in Table 24.4, using IRLS methods leads to the following estimates (in exponentiated form):

$$\begin{aligned} e^{\hat{\mu}} &= 7.69 \times 10^{-5} \\ e^{\hat{\alpha}_1} &= 0.039, & e^{\hat{\alpha}_2} &= 0.117, & e^{\hat{\alpha}_3} &= 0.247, & e^{\hat{\alpha}_4} &= 1.105, & e^{\hat{\alpha}_5} &= 1.144, \\ e^{\hat{\alpha}_6} &= 3.017, & e^{\hat{\alpha}_7} &= 3.823, & e^{\hat{\alpha}_8} &= 6.047, & e^{\hat{\alpha}_9} &= 10.052, \\ e^{\hat{\delta}_1} &= 1.00, & e^{\hat{\delta}_2} &= 3.39, & e^{\hat{\delta}_3} &= 8.16, & e^{\hat{\delta}_4} &= 10.10, & e^{\hat{\delta}_5} &= 18.20, \\ e^{\hat{\delta}_6} &= 22.60, & e^{\hat{\delta}_7} &= 36.80 \end{aligned}$$

Given these estimates, the predicted number \hat{Y}_{ij} of cancer deaths in cell (i, j) is $\hat{Y}_{ij} = \ell_{ij} e^{\hat{\mu} + \hat{\alpha}_i + \hat{\delta}_j}$. For example, when $i = 4$ and $j = 5$, then

$$\begin{aligned} \hat{Y}_{45} &= \ell_{45} e^{\hat{\mu}} e^{\hat{\alpha}_4} e^{\hat{\delta}_5} \\ &= (4,687)(7.69 \times 10^{-5})(1.105)(18.20) \\ &= 7.25 \end{aligned}$$

which should be compared with the actual observed value $Y_{45} = 6$. It can be shown that the deviance for this fitted model, as calculated using (24.14), has the numerical value of 51.47. (Formula (24.15) is not appropriate for these data, since several \hat{Y}_{ij} 's are close to 0 in value.) When compared with critical values of the chi-square distribution with $(63 - 15) = 48$ degrees of freedom, the value 51.47 does not suggest any significant lack of fit for the cross-classification model (24.18).

However, using model (24.18) involves estimating 15 parameters (whereas n is only 63); hence, we are not achieving a parsimonious description of the data. In addition, it is of considerable interest with these data to fit a model whose parameters can realistically be interpreted in terms of the mathematical theory of carcinogenesis.

In what follows, we will consider the *four-parameter nonlinear* model described by Frome (1983)—namely,

$$\lambda_{ij} = \lambda(T_i, D_j; \gamma, \alpha, \theta, \delta) = (\gamma + \alpha D_j^\theta) T_i^\delta \quad (24.19)$$

where T_i and D_j are as defined earlier.⁷

By using the mathematical identity $e^{\ln a} = a$ for $a > 0$, we can write (24.19) in an equivalent form considered by Frome (1983):

$$\lambda_{ij} = [e^{\ln \gamma} + e^{(\ln \alpha + \theta \ln D_j)}] e^{\delta \ln T_i} = [e^{\beta_3} + e^{(\beta_1 + \beta_2 X_{2j})}] e^{\beta_0 X_{1i}} \equiv \lambda(\mathbf{X}_{ij}, \boldsymbol{\beta}) \quad (24.20)$$

where $\mathbf{X}_{ij} = (X_{1i}, X_{2j}) = (\ln T_i, \ln D_j)$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) = (\delta, \ln \alpha, \theta, \ln \gamma)$. Expressing (24.19) in the exponential form (24.20) ensures that predicted rates are always positive.

The fitting of model (24.20) by IRLS gives the estimates $\hat{\beta}_0 = 4.46$, $\hat{\beta}_1 = 1.82$, $\hat{\beta}_2 = 1.29$, and $\hat{\beta}_3 = 2.94$. The estimated standard errors for these four estimators (obtained as the square roots of the appropriate diagonal elements of the estimated covariance matrix) are, respectively, 0.33, 0.66, 0.20, and 0.58. For example, an approximate large-sample 95% confidence interval for $\alpha (= e^{\beta_1})$ would be

$$\exp(\hat{\beta}_1 \pm 1.96 \sqrt{\text{Var}(\hat{\beta}_1)})$$

giving $\exp[1.82 \pm 1.96(0.66)] = \exp(1.82 \pm 1.29)$ and thus the interval $(e^{0.53}, e^{3.11}) = (1.70, 22.42)$.

Finally, a summary of the results of fitting various subset models of (24.19) is provided in Table 24.5.

TABLE 24.5 Summary of analyses of data in Table 24.4

Model for λ_{ij}	Number of Parameters	$D(\beta)$	d.f.*
γ	1	445.10	62
γT_i^δ	2	180.82	61
$(\gamma + \alpha D_j) T_i^\delta$	3	61.84	60
$(\gamma + \alpha D_j^\theta) T_i^\delta$	4	59.58	59

*d.f. = 63 – (Number of parameters in fitted model).

⁷ In model (24.19), γ represents the background (i.e., $D_j = 0$ for a nonsmoker) rate at age 62.5 (i.e., $T_i = [\text{Age} - 20]/42.5 = 1$ at age 62.5), αD_j^θ describes the effect of dosage (i.e., amount smoked) on lung cancer death rates, and T_i^δ is the multiplicative effect (on $\gamma + \alpha D_j^\theta$) of the time elapsed since the smoking habit was started. Frome (1983) provides further discussion of scientific evidence in favor of using a model like (24.19) to describe the incidence of lung cancer.

As discussed earlier, calculating the various differences between deviances in the ANOVA-type table presented in Table 24.5 will provide LR tests for the importance of the parameters γ , δ , α , and θ . First of all, the difference $(445.10 - 180.82) = 264.28$, when compared with upper-tail values of the χ^2_1 distribution, argues strongly for rejecting $H_0: \delta = 0$ in favor of $H_A: \delta \neq 0$. This means that the (multiplicative) effect of time since first exposure is important. Second, the difference $(180.82 - 61.84) = 118.98$ demands rejection of $H_0: \alpha = 0$ in favor of $H_A: \alpha \neq 0$, suggesting that the amount smoked is an important variable. Finally, the difference $(61.84 - 59.58) = 2.26$ does not lead to rejection of $H_0: \theta = 1$, so the first power of dosage seems most appropriate. Since the deviance value of 61.84 with 60 degrees of freedom for the model $\lambda_{ij} = (\gamma + \alpha D_j) T_i^\delta$ does not suggest any lack of fit, this model seems preferable.⁸

24.8 Summary

This chapter described the general form and several key features of the Poisson regression model. The “typical” Poisson regression model expresses a log *rate* as a linear function of a set of predictors. The Poisson regression method, nevertheless, allows for more-complicated nonlinear models as well. The dependent variable is a count of the number of occurrences of an event of interest, such as the number of cases of a disease that occur over a given follow-up time period. For the “typical” Poisson regression model, the natural measures of effect that are estimated are *rate ratios*.

ML estimation is used to estimate the regression coefficients of a Poisson regression model. The likelihood function assumes that the underlying response (a count) has the Poisson distribution. A measure of goodness of fit of a Poisson regression model is obtained by using the deviance statistic, which is an LR-type statistic reflecting the fit of a current model of interest relative to the baseline (uninformative or “saturated”) model (which involves as many parameters as there are data points). The difference in deviance statistics obtained for two (hierarchically ordered) models being compared is equivalent to the difference in log likelihood statistics for each model, so an LR test for Poisson regression can be carried out equivalently by using differences in deviance statistics. Finally, a tabular summary of the results of fitting various models to the same data set can be presented in a *Poisson ANOVA table*, which contains deviance and corresponding d.f. information for each model being fit.

It is important to emphasize again that the mean and variance of the Poisson distribution are equal. However, it is common to observe count data for which sample variances are either larger than sample means (called “over-dispersion”) or smaller than sample means (called “under-dispersion”), with the former phenomenon being more prevalent. For most popular categorical data analysis procedures (e.g., logistic and Poisson regression), there are statistical methods available to account appropriately for over-dispersed or under-dispersed data (e.g., see the excellent book by McCullagh and Nelder [1989]).

⁸ The reduction in the deviance due to adding a parameter (or a group of parameters) is *order-dependent*; consequently, a different order of parameter additions can lead to a different final model.

Problems

1. Suppose that, for each of three age groups (25–34, 35–44, and 45–54), we have recorded yearly sex-specific lung cancer mortality rates for the five-year period 1990 through 1994. These data are to be analyzed by Poisson regression methods to see whether the change (if any) in log rate over time varies by age–sex group and, if so, to quantify that variation. In what follows, we will index the six age–sex groups, as follows:

group 1($i = 1$):	25–34-year-old females
group 2($i = 2$):	35–44-year-old females
group 3($i = 3$):	45–54-year-old females
group 4($i = 4$):	25–34-year-old males
group 5($i = 5$):	35–44-year-old males
group 6($i = 6$):	45–54-year-old males

Consider the following model for the expected cell count $E(Y_{ik})$ for age–sex group i in year $(1990 + k)$, where $i = 1, 2, 3, 4, 5, 6$ and $k = 0, 1, 2, 3, 4$:

$$E(Y_{ik}) = \ell_{ik}\lambda_{ik}$$

where

$$\ln \lambda_{ik} = \sum_{i=1}^6 \alpha_i A_i + \beta k \quad (1)$$

Here ℓ_{ik} and λ_{ik} are, respectively, the person-years at risk and the (unknown) population lung cancer mortality rate in cell (i, k) . The independent variables in model (1) are defined as follows:

$$A_i = \begin{cases} 1 & \text{if age-sex group } i \\ 0 & \text{otherwise} \end{cases}$$

$$k = (\text{year} - 1990)$$

- a. What is the total number n of data points, or pairs (ℓ_{ik}, Y_{ik}) , for this data set?
- b. Based on model (1), what is the expected cell count for a 40-year-old male in 1992, written as a function of α_5 and β ?
- c. For the i th age–sex group, how does model (1) describe the *change* in log rate over time?
- d. What does model (1) assume about the effect of age–sex group on the *change* in log rate over time?
- e. Find a general expression for

$$RR_{ik} = \frac{\lambda_{ik}}{\lambda_{10}}$$

the rate ratio comparing the mortality rate for age group i and year $(1990 + k)$ to the mortality rate for the reference category “25–34-year-old females in 1990.”

- f. Suppose that it is of interest to assess whether the *change* in log rate over time actually varies by age–sex group (i.e., whether there is a group-by-time interaction). By adding appropriate cross-product terms to model (1), construct a model that will permit such an assessment and then discuss how one would interpret this model.

Consider the following Poisson regression ANOVA table, based on fitting various models to the data under study:

Model for $\ln\lambda_{ik}$	Number of Parameters	$D(\hat{\beta})$	d.f.
(1) α	1	300	29
(2) $\alpha + \beta k$	2	200	28
(3) $\sum_{i=1}^6 \alpha_i A_i$	6	175	24
(4) $\sum_{i=1}^6 \alpha_i A_i + \beta k$	7	60	23
(5) $\sum_{i=1}^6 \alpha_i A_i + \beta_1 k + \beta_2 k^2$	8	59	22
(6) $\sum_{i=1}^6 \alpha_i A_i + \beta k + \gamma_1(A_i k)$	8	25	22
(7) $\sum_{i=1}^6 \alpha_i A_i + \beta k + \sum_{i=1}^5 \gamma_i(A_i k)$	12	20	18
(8) α_{ik}	30	0	0

Using this table, answer the following questions:

- g. Ignoring (for now) the variable “time,” is there evidence that average mortality rates differ among the six age–sex groups?
- h. Is adding the linear time term (βk) to model (3) worthwhile?
- i. Assuming that the change in log rate over time is the same for all age–sex groups, do the data argue for adding a quadratic time term to a model that already contains a linear component of time?
- j. Is there evidence in the data that the change in log rate over time differs for different age–sex groups?
- k. Carry out a test of H_0 : “All six age–sex groups have the same slope” versus H_A : “All age–sex groups except group 1 have the same slope.”
- l. Of the models that fit the data well (use $\alpha = .1$ for any test of lack of fit that you perform), which one would you choose as your final model? Why?
- m. For your final model chosen in part (l), calculate a pseudo- R^2 -value.
- n. Now, assume that model (6) has been fit to the data, resulting in the following regression coefficient estimates and standard errors:

Parameter	Point Estimate	Standard Error
$\hat{\alpha}_1$	0.50	0.25
$\hat{\alpha}_2$	1.00	0.40
$\hat{\alpha}_3$	1.50	0.30
$\hat{\alpha}_4$	1.25	0.30
$\hat{\alpha}_5$	1.50	0.50
$\hat{\alpha}_6$	1.75	0.40
$\hat{\beta}$	0.50	0.20
$\hat{\gamma}_1$	-3.00	0.50

- o.** Find and interpret approximate 95% confidence intervals for the following time-related parameters: (1) the common slope for age–sex groups 2 through 6 and (2) the slope for age–sex group 1. [Note: The estimated covariance between $\hat{\beta}$ and $\hat{\gamma}_1$ is $\widehat{\text{Cov}}(\hat{\beta}, \hat{\gamma}_1) = -0.10$.]
- 2.** A five-year follow-up study was carried out to assess the relationship of diet and weight to the incidence of stomach cancer in 40- to 50-year-old males in a certain metropolitan area. Let Y_{ij} denote the number of cases of stomach cancer found in the i th weight category of diet type j , as indicated in the following table:

	Low-cholesterol Diet ($j = 1$)	High-cholesterol Diet ($j = 2$)
Low weight ($i = 1$)	Y_{11}	Y_{12}
Medlow weight ($i = 2$)	Y_{21}	Y_{22}
Medhigh weight ($i = 3$)	Y_{31}	Y_{32}
High weight ($i = 4$)	Y_{41}	Y_{42}

Consider the following Poisson regression model for these data:

$$\text{Model 1: } E(Y_{ij}) = \ell_{ij}\lambda_{ij} \quad i = 1, 2, 3, 4 \quad \text{and} \quad j = 1, 2$$

where ℓ_{ij} = man-years at risk in the (i, j) th cell and λ_{ij} = rate of development of stomach cancer in the (i, j) th cell, where λ_{ij} is modeled as follows:

$$\ln \lambda_{ij} = \sum_{i=1}^4 \alpha_i V_i + \beta C \quad \text{where} \quad V_i = \begin{cases} 1 & \text{if weight group } i \\ 0 & \text{otherwise} \end{cases}$$

$$C = \begin{cases} 1 & \text{if low-cholesterol diet } (j = 1) \\ 0 & \text{if high-cholesterol diet } (j = 2) \end{cases}$$

- a.** Based on this model, give an expression for the rate ratio (RR) comparing low-cholesterol diet subjects to high-cholesterol diet subjects, adjusted for weight group.
- b.** Give a large-sample formula for the 95% confidence interval for the RR described in part (a).
- c.** How would one test whether there is a significant interaction between diet type and weight group?
- d.** Assuming that there is no interaction between weight group and diet type, consider the following model as an alternative to model 1:

$$\text{Model 2: } \ln \lambda_{ij} = \alpha + \gamma W_i + \beta C$$

where W_i denotes the midpoint (in pounds) of weight group i .

Describe how to use a “difference between deviances” approach to evaluate whether model 2 fits the data significantly better than model 1. (State the null hypothesis, describe the test statistic in terms of deviances, and state the distribution of the test statistic, including degrees of freedom, under the null hypothesis.) How might one criticize the use of this approach?

- 3.** The following set of questions relates to using Poisson regression methods to analyze data from an in vitro study of human chromosome damage. In this study, using

Poisson regression is appropriate because we have “count” response data, where each count is the number of broken chromosomes in a sample of 100 cells taken from each of $n = 40$ individuals. (A Poisson distribution is appropriate for describing counts per unit of *space*—e.g., 100 cells—as well as for describing counts per unit of *time*.)

The design of the study described here involved randomly assigning each of the cell samples for the 40 individuals to one of four treatment groups consisting of 10 individuals each. The 10 sets of 100 cells in each treatment group were exposed to a particular combination of two drugs, A and B, as follows:

Treatment Group j	Number of Individuals n_j	Drug Combination
1	10	neither A nor B
2	10	A only
3	10	B only
4	10	both A and B
	40	

The exposed cells were then examined for chromosome breakage, and the number of broken chromosomes in the 100 cells for each individual was counted. *The investigator wishes to assess the effects of drug use, and of drug interaction, on the rate of chromosome damage.*

Let Y_{ij} denote the random variable representing the number of chromosome breaks counted for each individual i ($i = 1, 2, \dots, 10$) within treatment group j ($j = 1, 2, 3, 4$). Also, let $\ell_{ij} = 100$, where ℓ_{ij} denotes the amount of “person-space” of observation. Assume that the drugs act in a multiplicative fashion on the rate of chromosome breakage (λ_{ij}). The Poisson regression model under consideration can be written as

$$\ln \lambda_{ij} = \beta_0 + \sum_{k=1}^p \beta_k X_{ijk}$$

where X_{ijk} denotes the value of the k th predictor for the i th individual in the j th treatment group.⁹

⁹ Note that Chapter 24 provides only an introduction to Poisson regression and primarily covers the common application of Poisson regression to the modeling of group rates. However, there is a broader set of scenarios in which Poisson regression may be appropriate. Problem 3 presents an example where the level of analysis is the individual level rather than the group level. In this problem, the outcome of broken chromosome count was enumerated on a per-person basis for individuals receiving different combinations of two drugs. Instead of representing a unit of time, the offset ℓ_{ij} represents 100 cells sampled from each person. Individual-level analyses involving time are also possible, such as studies that count recurrent events in individuals over a period of followup. An example of this would be a 10-year study of the number of cardiac episodes a person with cardiovascular disease experiences, in which individuals may be observed for differing amounts of time due to study drop-out or other reasons (called *censoring*).

An additional aspect of Problem 3 is that exactly 100 cells were observed for every person, and thus a constant value of $\ln(100)$ is added to every observation in the $\ln \mu_{ij}$ form of the regression model. Because there is no change in ℓ_{ij} from person to person, we are effectively adding the same value to the intercept for every person, so the estimates of the other model regression coefficients (and the *RR* estimates calculated from them) are not affected. It then follows that having a constant value for the offset does not affect estimated relationships between groups, so we would obtain the same estimated rate differences between treatment groups whether we sampled 50, 100, or 100,000 cells per subject or whether we left out the offset entirely from the model.

(continued)

- Based on the preceding scenario, write out the form of the Poisson regression model for $\ln \lambda_{ij}$ as a function of the predictors of interest in this study. In doing so, *make sure to explicitly define a set of categorical predictors to include in this model; these predictors should reflect the primary objective of the study.*
- Based on the model specified in part (a), describe *two* alternative ways one can test the null hypothesis of no interaction between the drugs. Include in the answer the null hypothesis (in terms of regression coefficients), the formula for the test statistic(s), the degrees of freedom for each test, and the distribution of the test statistic under the null hypothesis.
- Fill in the blanks in the following Poisson regression ANOVA table, which is based on fitting various models to describe the rate of chromosome breakage:

Model for $\ln \lambda_{ij}$	Number of Parameters	Deviance	Deviance d.f.
(I) Baseline (i.e., β_0)	1	90	39
(II) Main effect of drug A only	—	85	—
(III) Main effect of drug B only	—	84	—
(IV) Main effects of both drug A and drug B	—	40	—
(V) All main effects and interaction	—	30	—

- According to the ANOVA table, which models provide good fit to the data? Explain briefly.
 - According to the ANOVA table, does drug A have a significant main effect? Explain briefly.
 - According to the ANOVA table, does drug B have a significant effect over and above the effect of drug A? Explain briefly.
 - According to the ANOVA table, do drugs A and B have a significant interaction effect? Explain briefly.
 - According to the ANOVA table (and in light of your answers to parts (d) through (g)), provide an appropriate expression for the rate ratio for the effect of drug A given drug B status relative to baseline and an expression for the rate ratio for the effect of drug B given drug A status relative to baseline.
4. For the nonmelanoma skin cancer data considered in this chapter, consider the following additional model that was fit to these data:

$$\text{Model 5: } \ln \lambda_{ij} = \alpha + \theta \ln T_i + \beta E + \delta E(\ln T_i)$$

where T_i is defined as

$$T_i = \frac{[\text{Midpoint of the } i\text{th age interval}] - 15}{35} \quad i = 1, 2, \dots, 8$$

and E denotes city (1 = Dallas–Ft. Worth, 0 = Minneapolis–St. Paul).

This invariance property extends to models where $\ln(\ell)$ is assumed to be 0 for all individuals—that is, where time is considered irrelevant (or constant, as discussed above) and the Poisson regression analysis considers a per-individual count as the response. Examples of this situation would be the number of moles found on a patient during a dermatologic examination and the number of angina episodes that a study subject experiences over a fixed time period.

Edited SAS Output (PROC GENMOD) for Skin Cancer Data, Model 5

CRITERIA FOR ASSESSING GOODNESS OF FIT			
Criterion	DF	Value	Value/DF
Deviance	12	10.6302	0.8859
Scaled Deviance	12	10.6302	0.8859
Pearson Chi-Square	12	10.6319	0.8860
Scaled Pearson X2	12	10.6319	0.8860
Log Likelihood		7200.6459	

ANALYSIS OF MAXIMUM LIKELIHOOD PARAMETER ESTIMATES					
Parameter	DF	Estimate	Std Error	Chi-Square	Pr > ChiSq
Intercept	1	-7.1400	0.0597	14324.9448	<.0001
CITY	1	0.8869	0.0686	167.2316	<.0001
LNT	1	2.4916	0.1237	405.9034	<.0001
CITYLNT	1	-0.2811	0.1435	3.8386	0.0501
Scale	0	1.0000	0.0000		

Note: CITY = E , LNT = $\ln T_i$, and CITYLNT = $E(\ln T_i)$.

- Using the available edited computer output, calculate the estimated rate ratio comparing the two cities for the following two age-groups: (i) 15–24 and (ii) 45–54.
(Note: For the 15–24 age-group, $T_i = 1/7$; for the 45–54 age-group, $T_i = 1$.)
- What do your results in part (a) suggest about interaction?
- Carry out both a Wald test and an LR test for the significance of the coefficient of the product term $E(\ln T_i)$. [Hint: To carry out the LR test, use the deviance statistic $D(\beta)$ given in Table 24.3 for model 4].
- Do the results from part (c) support the conclusion that the product term $E(\ln T_i)$ needs to remain in the model?
- Are the tests conducted in part (c) equivalent to testing whether either model 4 or model 5 fits the data?
- Which model appears to be the best model among models 3, 4, and 5? Briefly justify any answer.
- The following questions concern Poisson regression models fit to fictitious follow-up study data in which rates of disease are modeled as a function of age and smoking status. The SAS program codes and output for six models are provided on pages 775–780 following the last problem in this chapter. The dependent variable is called COUNT and the independent variables are AGE and SMOKE. There are three age groups (defined by age midpoint) and two levels of smoking status. The data are given in the following table:

	SMOKE = 0			SMOKE = 1		
	AGE = 25	AGE = 45	AGE = 75	AGE = 25	AGE = 45	AGE = 75
Person-time	1,000	3,500	2,000	100	1,500	1,000
Count	15	36	35	4	28	32

Note: AGE = 25 if $15 \leq \text{age} < 35$

AGE = 45 if $35 \leq \text{age} < 55$

AGE = 75 if $55 \leq \text{age} < 95$

- Using the problem code for the available computer output, note that the difference between model 1 and model 2 is that model 2 does not contain an offset term. Why does the lack of an offset term for model 2 suggest that smoking is preventive of the disease, while model 1 suggests that smoking is a risk factor?
- Suppose that both model 1 and model 2 were fit to a different data set, and it was found that there was no difference in the estimated effects of smoking and age. What would that say about the structure of the data?
- Use the Wald test to evaluate the statistical significance of the interaction term in model 3. In answering this question, state the model, state the null hypothesis, perform the statistical test, show the calculations, and draw conclusions.
- What assumption is made for model 1 that is not made for model 5 concerning the effect of age on the rate of disease? Based on the output, should there be a preference for either of these two models? Explain.
- Use deviance information to perform an LR test to assess the significance of the age variables in model 5 (controlling for smoking). Make sure to discuss appropriate conclusions.
- Use deviance information to perform an LR test to assess the significance of the interaction terms in model 6. Make sure to discuss appropriate conclusions.
- Why is the deviance for model 6 equal to zero?
- Determine the estimated rate ratio and 95% confidence interval for comparing the oldest age group ($55 \leq \text{age} < 95$) to the youngest age group ($15 \leq \text{age} < 35$) controlling for smoking and using Model 1.
- Determine the estimated rate ratio and 95% confidence interval for comparing the oldest age group ($55 \leq \text{age} < 95$) to the youngest age group ($15 \leq \text{age} < 35$) controlling for smoking and using model 5.
- Are the results for part (h) different from the results for part (i)? Explain briefly.
- Using the computer output, fill in the blanks in the following Poisson regression "ANOVA" table based on fitting models 1–6.

Model number	Number of Parameters	Deviance	Deviance d.f.	Deviance/d.f.
1	3	6.3636	—	—
2	—	—	—	—
3	—	—	—	—
4	—	—	—	—
5	—	—	—	—
6	—	—	—	—

- Based on answers to the previous questions and on any other considerations related to the computer output from fitting models 1–6, which of the six models seems to be the "best" model? Explain.

6. A five-year follow-up study was carried out in a certain metropolitan area to assess the relationship of diet and weight to the incidence of stomach cancer. Data were obtained on $n = 2,000$ subjects. The variables of interest for these data were

T = time (in months) until stomach cancer (SCA) was detected or time (in months) until either the subject was lost to follow-up or the study ended (often called the censoring time);

ST = event indicator status (1 if SCA detected, 0 if SCA not detected);

$WTGP$ = weight group (1 = low, 2 = medlow, 3 = medhigh, 4 = high), with “low” the referent group;

DT = diet type (1 = high fiber diet, 2 = medium fiber diet, 3 = low fiber diet), with “high fiber diet” the referent group;

GEN = gender (0 = male, 1 = female);

$AGEGP$ = agegroup (1 = 40–54 years, 2 = 55–69 years, 3 = 70+ years), with 40–54 years being the referent group.

Suppose that one considers doing a Poisson regression analysis to assess the effects of diet type and weight on the development of stomach cancer (SCA), controlling for age and gender.¹⁰ To carry out such an analysis, organize the data as follows:

Step 1 Form combinations of categories over all four predictors ($WTGP$, DT , GEN , $AGEGP$) being considered; these category combinations will define the subgroups to be analyzed using Poisson regression. Since there are four categories of $WTGP$, three categories of DT , two categories of GEN , and three categories of $AGEGROUP$, the total number of subgroups will be $(4 \times 3 \times 2 \times 3) = 72$.

Step 2 For the 72 subgroups, count the number of persons who develop SCA in each subgroup, and denote this count variable as Y . Also, sum up the person-time information over all the persons in each subgroup, and call this variable PT .

Step 3 Use the 72 Y values as the counts and the 72 PT values as the person-time information to fit Poisson regression models to these data.

- a. Based on the data organization just described, what is the “sample size” to be used for fitting a Poisson regression model to these data?
- b. State a Poisson regression model (called model 1) that would model the natural log of the rate of development of stomach cancer as a linear function of the risk factors DT and $WTGP$, controlling for potential confounding and effect modification by the variables GEN and $AGEGP$. Consider only two-factor product terms involving exposure variables and control variables.
- c. How would one modify the model in part (b) so that both the $WTGP$ variable and the DT variable are treated as ordinal variables on a natural logarithmic scale? (In stating this modified model, called model 2, make sure to explicitly define the “transformed” $WTGP$ and DT variables that would need to be used.)

¹⁰ An alternative approach to a Poisson regression analysis is to conduct a survival analysis using a Cox Proportional Hazards (PH) Model (see Kleinbaum and Klein, 2005, for further details). The PH modeling approach may be more appropriate than using Poisson regression if the “baseline hazards” component of the PH model is not constant over follow-up time.

- d. Provide the model statement, including required options, that one would use with SAS's PROC GENMOD (or a program from a different computer package) to fit model 2, described in part (c) above.
- e. Based on model 2 defined in part (c), give a formula for the rate ratio that compares a subject who has a low fiber diet and is in the high weight group to a subject who has a high fiber diet and is in the low weight group, controlling for GEN and AGE GP. (Assume nonzero interaction effects.)
- f. Provide an expression for a 95% confidence interval for the rate ratio that compares a subject who has a low fiber diet and is in the high weight group to a subject who has a high fiber diet and is in the low weight group, controlling for GEN and AGE GP. (Assume nonzero interaction effects.)
- g. Based on model 2, describe how one would carry out an overall test for significant interaction involving deviance statistics. (Make sure to state the null hypothesis, the test statistic, and the d.f. for the test statistic under the null hypothesis.)
- h. Is the test described in part (g) for model 2 equivalent to carrying out a goodness of fit test for a no-interaction version of model 2 that does not contain any product terms? Explain briefly.
- i. If model 1 is considered instead of model 2, is an overall test for significant interaction equivalent to carrying out a goodness-of-fit test for a no-interaction version of model 1 that does not contain any product terms? Explain briefly.

Suppose that the following Poisson ANOVA table resulted from fitting several different Poisson regression models to these data.

Model Number	Variables in Model*	Number of Parameters	Deviance
1	None (i.e., constant term only)	1	a
2	(ordinal) DT only	2	b
3	(ordinal) WTGP only	2	c
4	GEN only	2	d
5	(nominal) DT only	3	e
6	(ordinal) DT and (ordinal) WTGP	3	f
7	AGE GP only	3	g
8	AGE GP and GEN	4	h
9	(nominal) WTGP only	4	i
10	(nominal) DT and (nominal) WTGP	6	j
11	AGE GP, GEN, (ordinal) DT, and (ordinal) WTGP	6	k
12	AGE GP, GEN, (nominal) DT, and (nominal) WTGP	9	l
13	Model 2 (from part (c))	—	m
14	Model 1 (from part (b))	—	n
15	Saturated	—	p

*(ordinal) DT is represented by a single ordinal variable.

(ordinal) WTGP is represented by a single ordinal variable.

(nominal) DT is represented by 2 dummy variables.

(nominal) WTGP is represented by 3 dummy variables.

- j. Assuming no interaction of any kind between the risk factors (DT and WTGP) and either AGE GP and/or GEN, use the deviance values (e.g., a, b, c) in the above table to give an expression for the LR statistic that tests whether there is a significant

- difference between the (joint) effects of the nominal exposure variables—that is, (nominal) DT and (nominal) WTGP—controlling for AGE GP and GEN.
- k.** What are the degrees of freedom for the LR statistic described in part (j)?
 - l.** Use the deviance scores in the above table to give an expression for the LR statistic for testing whether there is at least one significant interaction effect in model 1 (as defined in part (b)); that is, describe a chunk test for the interaction terms in model 1.
 - m.** What are the degrees of freedom for the LR statistic described in part (l)?
 - n.** Describe how one might carry out a (single) test of hypothesis to determine whether model 1 (as defined in part (b)) or model 2 (as defined in part (c)) fits the data better. In answering this question, state the null hypothesis, the test statistic, and its d.f. under the null hypothesis.
 - o.** Assuming that a Poisson model is appropriate for these data, how could one *criticize* the significance testing method described in part (n) for comparing model 1 with model 2?
 - p.** In what other way, using deviance statistic information (other than the test of hypothesis described in part (n)), can one evaluate whether model 1 or model 2 is better?

SAS Program Code for Problem 5

```
Below is SAS program code followed by SAS output for six Poisson models.
Models 1-4 are performed on "data A", whereas Models 5-6 are performed
on "data B".

data A;
    input age smoke count p_time;
    log_time=log(p_time);
    age_smk=age*smoke;
cards;
  25 0 15 1000
  25 1 4   100
  45 0 36 3500
  45 1 28 1500
  75 0 35 2000
  75 1 32 1000
  ;
run;

proc genmod data=A;
    model count=age smoke/dist=poisson link=log offset=log_time;
    title 'MODEL 1';
run;
proc genmod data=A;
    model count=age smoke/dist=poisson link=log;
    title 'MODEL 2';
run;
proc genmod data=A;
    model count=age smoke age_smk/dist=poisson link=log offset=log_time;
    title 'MODEL 3';
run;
proc genmod data=A;
    model count=smoke/dist=poisson link=log offset=log_time;
    title 'MODEL 4';
run;
```

Six variables are created in "data A":

- age:** three age groups: 25, 45, and 75
- smoke:** smoker = 1, non-smoker = 0
- count:** the number of individuals in the cohort with disease
- p_time:** total person time for each age-by-smoke level
- log_time:** the natural log of p_time
- age_smk:** age × smk interaction term

(continued)

```

data B;
  set A;
  if age=25 then do; age1=0; age2=0; end;
  if age=45 then do; age1=1; age2=0; end;
  if age=75 then do; age1=0; age2=1; end;
  age1_smk=age1*smoke;
  age2_smk=age2*smoke;
run;

proc genmod data=B;
  model count=age1 age2 smoke/covb dist=poisson link=log
    offset=log_time;
  title 'MODEL 5';
run;

proc genmod data=B;
  model count=age1 age2 smoke age1_smk age2_smk/dist=poisson link=log
    offset=log_time;
  title 'MODEL 6';
run;

```

Four additional variables are created in "data B":
age1 and **age2**: two dummy variables for the
 three age groups; age=25 is referent group
age1_smk and **age2_smk**: interaction terms
 for the two age dummy variables with smoke

Edited SAS Output (PROC GENMOD) for Problem 5

MODEL 1
 The SAS System
 The GENMOD Procedure

MODEL INFORMATION	
Data Set	WORK.A
Distribution	Poisson
Link Function	Log
Dependent Variable	count
Offset Variable	log_time
Observations Used	6

CRITERIA FOR ASSESSING GOODNESS OF FIT			
Criterion	DF	Value	Value/DF
Deviance	3	6.3636	2.1212
Scaled Deviance	3	6.3636	2.1212
Pearson Chi-Square	3	7.2783	2.4261
Scaled Pearson X2	3	7.2783	2.4261
Log Likelihood	.	350.6333	

(continued)

ANALYSIS OF MAXIMUM LIKELIHOOD PARAMETER ESTIMATES					
Parameter	DF	Estimate	Std Error	Chi-Square	Pr > ChiSq
Intercept	1	-4.8391	0.2839	290.5207	<.0001
Age	1	0.0098	0.0049	4.0497	0.0442
Smoke	1	0.5784	0.1663	12.1051	0.0005

MODEL 2
The SAS System
The GENMOD Procedure

MODEL INFORMATION	
Data Set	WORK.A
Distribution	Poisson
Link Function	Log
Dependent Variable	count
Observations Used	6

CRITERIA FOR ASSESSING GOODNESS OF FIT			
Criterion	DF	Value	Value/DF
Deviance	3	18.7971	6.2657
Scaled Deviance	3	18.7971	6.2657
Pearson Chi-Square	3	17.3596	5.7865
Scaled Pearson X2	3	17.3596	5.7865
Log Likelihood	.	344.4165	

ANALYSIS OF MAXIMUM LIKELIHOOD PARAMETER ESTIMATES					
Parameter	DF	Estimate	Std Error	Chi-Square	Pr > ChiSq
Intercept	1	2.4336	0.2487	95.7434	<.0001
Age	1	0.0177	0.0040	19.4544	<.0001
Smoke	1	-0.2955	0.1651	3.2033	0.0735

MODEL 3
The SAS System
The GENMOD Procedure

MODEL INFORMATION	
Data Set	WORK.A
Distribution	Poisson
Link Function	Log
Dependent Variable	count
Offset Variable	log_time
Observations Used	6

(continued)

CRITERIA FOR ASSESSING GOODNESS OF FIT			
Criterion	DF	Value	Value/DF
Deviance	2	6.2371	3.1186
Scaled Deviance	2	6.2371	3.1186
Pearson Chi-Square	2	7.2916	3.6458
Scaled Pearson X2	2	7.2916	3.6458
Log Likelihood	.	350.6965	

ANALYSIS OF MAXIMUM LIKELIHOOD PARAMETER ESTIMATES					
Parameter	DF	Estimate	Std Error	Chi-Square	Pr > ChiSq
Intercept	1	-4.7668	0.3474	188.2664	0.0001
Age	1	0.0084	0.0061	1.8770	0.1707
Smoke	1	0.3763	0.5936	0.4017	0.5262
Age_smk	1	0.0036	0.0100	0.1264	0.7222

MODEL 4

The SAS System

The GENMOD Procedure

MODEL INFORMATION	
Data Set	WORK.A
Distribution	Poisson
Link Function	Log
Dependent Variable	count
Offset Variable	log_time
Observations Used	6

CRITERIA FOR ASSESSING GOODNESS OF FIT			
Criterion	DF	Value	Value/DF
Deviance	4	10.3962	2.5991
Scaled Deviance	4	10.3962	2.5991
Pearson Chi-Square	4	10.6195	2.6549
Scaled Pearson X2	4	10.6195	2.6549
Log Likelihood	.	348.6170	

ANALYSIS OF MAXIMUM LIKELIHOOD PARAMETER ESTIMATES					
Parameter	DF	Estimate	Std Error	Chi-Square	Pr > ChiSq
Intercept	1	-4.3252	0.1078	1608.8402	0.0001
Smoke	1	0.6208	0.1651	14.1426	0.0002

(continued)

MODEL 5
The SAS System
The GENMOD Procedure

MODEL INFORMATION					
Data Set					WORK.B
Distribution					Poisson
Link Function					Log
Dependent Variable					count
Offset Variable					log_time
Observations Used					6

CRITERIA FOR ASSESSING GOODNESS OF FIT					
Criterion	DF	Value	Value/DF		
Deviance	2	0.3925	0.1962		
Scaled Deviance	2	0.3925	0.1962		
Pearson Chi-Square	2	0.4225	0.2112		
Scaled Pearson X2	2	0.4225	0.2112		
Log Likelihood	.	353.6188			

ANALYSIS OF MAXIMUM LIKELIHOOD PARAMETER ESTIMATES					
Parameter	DF	Estimate	Std Error	Chi-Square	Pr > ChiSq
Intercept	1	-4.1355	0.2310	320.5959	0.0001
Age1	1	-0.4568	0.2657	2.9545	0.0856
Age2	1	0.0769	0.2657	0.0839	0.7721
Smoke	1	0.6305	0.1688	13.9433	0.0002

ESTIMATED COVARIANCE MATRIX					
	Prm1	Prm2	Prm3	Prm4	
Prm1	0.05334	-0.05133	-0.05116	-0.004509	
Prm2	-0.05133	0.07062	0.05531	-0.008207	
Prm3	-0.05116	0.05531	0.07059	-0.009300	
Prm4	-0.004509	-0.008207	-0.009300	0.02851	

MODEL 6
The SAS System
The GENMOD Procedure

MODEL INFORMATION	
Data Set	WORK.B
Distribution	Poisson
Link Function	Log
Dependent Variable	count
Offset Variable	log_time
Observations Used	6

(continued)

CRITERIA FOR ASSESSING GOODNESS OF FIT			
Criterion	DF	Value	Value/DF
Deviance	0	0.0000	.
Scaled Deviance	0	0.0000	.
Pearson Chi-Square	0	0.0000	.
Scaled Pearson X2	0	0.0000	.
Log Likelihood	.	353.8151	

ANALYSIS OF MAXIMUM LIKELIHOOD PARAMETER ESTIMATES					
Parameter	DF	Estimate	Std Error	Chi-Square	Pr > ChiSq
Intercept	1	-4.1997	0.2582	264.5628	0.0001
Age1	1	-0.3773	0.3073	1.5072	0.2196
Age2	1	0.1542	0.3086	0.2495	0.6174
Smoke	1	0.9808	0.5627	3.0380	0.0813
Age1_smk	1	-0.3848	0.6166	0.3896	0.5325
Age2_smk	1	-0.3773	0.6136	0.3781	0.5386

References

- Baker, R. J., and Nelder, J. A. 1978. *Generalized Linear Interactive Modeling (GLIM), Release 3*. Oxford: Numerical Algorithms Group.
- Frome, E. L. 1983. "The Analysis of Rates Using Poisson Regression Models." *Biometrics* 39: 665–74.
- Frome, E. L., and Checkoway, H. 1985. "Use of Poisson Regression Models in Estimating Incidence Rates and Ratios." *American Journal of Epidemiology* 121(2): 309–23.
- Kleinbaum, D. G. 2003. *ActivEpi*. New York and Berlin: Springer Publisher.
- Kleinbaum, D. G., and Klein, M. 2005. *Survival Analysis: A Self-Learning Text*, Second Edition. New York and Berlin: Springer Publishers.
- Matanossi, G. M.; Breysse, P. N.; and Elliott, E. A. 1991. "Electromagnetic Field Exposure and Male Breast Cancer." *Lancet* 337(8743): 737.
- McCullagh, P., and Nelder, J. A. 1989. *Generalized Linear Models*, Second Edition. London and New York: Chapman and Hall.
- Remington, R. D., and Schork, M. A. 1985. *Statistics with Applications to the Biological and Health Sciences*. Englewood Cliffs, N.J.: Prentice-Hall.
- Scotto, J.; Kopf, A. W.; and Urbach, F. 1974. "Non-Melanoma Skin Cancer among Caucasians in Four Areas of the United States." *Cancer* 34: 1333–38.
- Stokes, M. E., and Koch, G. G. 1983. "A Macro for Maximum Likelihood Fitting of Log-Linear Models to Poisson and Multinomial Counts with Contrast Matrix Capability for Hypothesis Testing." *Proceedings of the Eighth Annual SAS Users Group International Conference*, pp. 795–800.

25

Analysis of Correlated Data Part 1: The General Linear Mixed Model

25.1 Preview

Up to this point in the text, we have described regression-modeling techniques for which a *single response* is measured on each observational unit (e.g., subject), where the response may be continuous (e.g., multiple linear regression), categorical (e.g., logistic regression), or a count variable (e.g., Poisson regression). For many studies, however, *two or more responses* are observed on each unit. The responses on each unit will generally be correlated, thereby requiring an analysis that accounts for such correlation.

As a simple example of *correlated data*, several blood pressure measurements may be taken over time on each of several subjects. Since the same response variable (i.e., blood pressure, treated as a continuous variable) is being measured at different times, we typically refer to such data as *repeated measures data*, and the collection of responses on the same subject is often called a *cluster* of responses. When such repeated measures are taken over time, the study is called a *longitudinal study*. In this chapter, we will only consider continuous responses, although more general methods are available (e.g., generalized linear mixed models; see Diggle et al. 2002; Zeger and Liang 1986; or Kleinbaum and Klein 2010).

Figure 25.1 illustrates repeated measures data on blood pressure (BP) for two different persons. Notice that person 1 has blood pressure measurements taken at five different times, whereas person 2 has blood pressure measurements taken at four times. Moreover, the times at which person 1 is observed are not all the same as the times at which person 2 is observed. This graph illustrates that, *in a repeated measures study, it is possible that different subjects can have different numbers of observations and that the observations on different subjects can occur at different times*.

The above example also leads us to introduce the following general mathematical notation for a continuous outcome/response variable for the i th subject/cluster in any type of

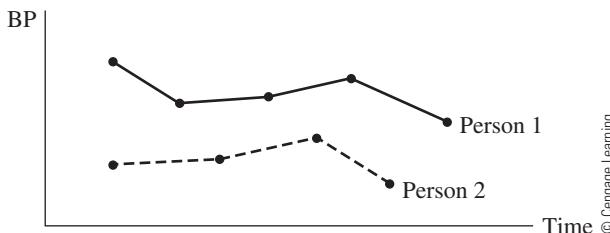


FIGURE 25.1 Longitudinal data on two subjects

repeated measures/correlated data study:

$$\mathbf{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{bmatrix} \quad i = 1, 2, \dots, K \quad (25.1)$$

where K denotes the number of subjects and n_i denotes the number of observations on the i th subject. \mathbf{Y}_i is often referred to as the *response vector* for the i th subject.¹

For the $K = 2$ subjects in Figure 25.1, the two response vectors can then be denoted as

$$\mathbf{Y}_1 = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{15} \end{bmatrix} \quad \text{for } i = 1 \text{ and } n_1 = 5 \text{ (person 1)} \quad \text{and } \mathbf{Y}_2 = \begin{bmatrix} Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \end{bmatrix} \quad \text{for } i = 2 \text{ and } n_2 = 4 \text{ (person 2)}$$

The above notation, nevertheless, does not distinguish the different times at which these two subjects are measured. To make such a distinction, we would have to lay out the data as shown in Table 25.1.

In Table 25.1, note that there is a separate line of data for each response on each subject. Also, the columns of this table describe the different variables that may be used in the analysis. The last column shown gives the times at which each subject is observed and indicates that different subjects can be observed at different times.

Table 25.1 provides an example of the general data layout for repeated measures data that we present in Table 25.2 below. Table 25.2 is organized so that observations are grouped by subjects/clusters. The first column denotes the subject ID. The second column indicates which repeated measure is observed. The third column indicates the time at which the

¹ The term *response vector* used to describe \mathbf{Y}_i derives from "matrix" terminology (see Appendix B on Matrices and Their Relationship to Regression Analysis). The analysis of repeated measures data can be conveniently described using matrices; in fact, most published literature on this subject uses matrix notation almost exclusively. Although this chapter will not emphasize a matrix-based discussion of repeated measures analysis, we will provide alternative matrix formulas for the interested reader.

TABLE 25.1 Data layout for longitudinal data on two subjects

Subject ID	Y	Repeat Number	Time
1	Y_{11}	1	T_{11}
1	Y_{12}		T_{12}
1	Y_{13}		T_{13}
1	Y_{14}		T_{14}
1	Y_{15}		T_{15}
2	Y_{21}	1	T_{21}
2	Y_{22}		T_{22}
2	Y_{23}		T_{23}
2	Y_{24}		T_{24}

© Cengage Learning

repeated measure is observed; this column would be omitted if the study is not longitudinal (i.e., if the data are not gathered over time). The fourth column describes the response variable. The remaining columns (X 's) describe any variables to be considered as predictors (i.e., independent variables) of interest; that is, X_{ijg} denotes the j th value of the g th predictor for subject i , where $i = 1, 2, \dots, K; j = 1, 2, \dots, n_i$; and $g = 1, 2, \dots, s$.

TABLE 25.2 General data layout for repeated measures data

Subject (i)	Repeat (j)	Time (T_{ij})	Y_{ij}	X_{ij1}	X_{ij2}	...	X_{ijp}
1	1	T_{11}	Y_{11}	X_{111}	X_{112}	...	X_{11s}
1	2	T_{12}	Y_{12}	X_{121}	X_{122}	...	X_{12s}
:	:	:	:	:	:	⋮	⋮
1	n_1	T_{1n_1}	Y_{1n_1}	X_{1n_11}	X_{1n_12}	...	X_{1n_1s}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	1	T_{i1}	Y_{i1}	X_{i11}	X_{i12}	...	X_{i1s}
i	2	T_{i2}	Y_{i2}	X_{i21}	X_{i22}	...	X_{i2s}
:	:	:	:	:	:	⋮	⋮
i	n_i	T_{in_i}	Y_{in_i}	X_{in_i1}	X_{in_i2}	...	X_{in_is}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	1	T_{K1}	Y_{K1}	X_{K11}	X_{K12}	...	X_{K1s}
K	2	T_{K2}	Y_{K2}	X_{K21}	X_{K22}	...	X_{K2s}
:	:	:	:	:	:	⋮	⋮
K	n_K	T_{Kn_K}	Y_{Kn_K}	X_{Kn_K1}	X_{Kn_K2}	...	X_{Kn_Ks}

© Cengage Learning

Clustered data need not be longitudinal. For example, suppose the responses in each cluster are measurements of blood pressure taken within a few minutes of each other on members of the same family. Then, for all practical purposes, the data are gathered at one point in time (i.e., cross-sectionally) and so are not longitudinal in nature. Nevertheless, responses on persons in the same family are unlikely to be independent, since family members typically share both lifestyle and genetic factors. For this scenario, the data (Table 25.2) will not contain a column for Time; moreover, the Repeat column in the table will not necessarily contain observations listed by the time order in which they were observed.

Prior to the development of high-speed computer programs that allow iterative solution of a complex system of equations required for maximum likelihood (ML) estimation, the classical approach to the analysis of repeated measures (i.e., clustered) data required the partitioning of sums of squares according to various sources of variation analogous to the classical ANOVA techniques described in Chapters 17 through 20. This ANOVA approach is equivalent to using a regression (ANOVA) model that contains both fixed and random effects (defined in Chapter 17), where the factor Subjects is considered a random factor. Inferences about main effects and interactions are carried out using F tests based on appropriate ratios of mean square terms.

The ANOVA approach to repeated measures analysis has a number of limitations. First, ANOVA considers only continuous outcomes assumed to be normally distributed and requires that all independent variables in the model must be categorized, including interval variables. Such categorization generally forfeits information on interval predictor variables that are validly measured. Furthermore, the ANOVA approach works most efficiently when the study design is *balanced* in the sense that each subject is observed the same number of times and at the same times; however, repeated measures studies often involve unbalanced data.

In light of the above limitations, other approaches to the analysis of repeated measures data have been developed, including the use of a more general form of the multiple linear regression model with both fixed and random effects, called the *general linear mixed model*. The general linear mixed model allows for independent variables to be of any type (not just categorical) and uses the method of maximum likelihood (described in Chapter 21) to estimate parameters and to make statistical inferences. In this chapter, we focus on the general linear mixed model (demonstrated with SAS's MIXED procedure rather than SAS's REG or GLM procedures), and we illustrate this approach to model correlated data when only fixed effects are considered. In the next chapter (Chapter 26), which is Part 2 of this topic, we focus on linear mixed models that contain random effects. We also briefly describe the ANOVA "partitioning" approach and, using an example, compare the results between the ANOVA and mixed model approaches. Also, in Chapter 26, we give a brief overview of how such analyses can be carried out for discrete-type (e.g., binary or count) outcomes using *generalized linear mixed models*.

25.2 Examples

In this section, we introduce three examples of repeated measures data. The analysis of each data set will be considered in later sections of this chapter.

25.2.1 A Study of the Effect of an Air Pollution Episode on Pulmonary Function

Let's first consider a hypothetical (but realistic) study of 40 school children who were examined under normal conditions, then during the week of an air pollution alert, and then on three successive weeks following the alert. The study objective was to determine whether FEV1 (often referred to as "lung capacity"), the volume of air exhaled in the first second of a forced exhalation, was depressed during the alert. A secondary objective was the identification of sensitive subgroups or individuals most severely affected by the pollution episode. Table 25.3 lists the data for the first 15 subjects in this study for the two primary variables of interest: FEV1 (the outcome variable) and Week, the primary predictor variable indicating

TABLE 25.3 Pulmonary function scores (FEV1) on the first 15 of 40 school children over five weeks with a pollution episode during week two

Subj.	Week	FEV1	Subj.	Week	FEV1	Subj.	Week	FEV1
1	1	9.43	6	1	9.12	11	1	12.44
1	2	5.71	6	2	7.71	11	2	8.68
1	3	5.86	6	3	6.75	11	3	7.48
1	4	7.70	6	4	9.55	11	4	6.40
1	5	5.89	6	5	8.20	11	5	7.61
2	1	11.15	7	1	10.30	12	1	7.88
2	2	9.48	7	2	7.24	12	2	4.88
2	3	10.11	7	3	6.99	12	3	6.43
2	4	8.89	7	4	7.66	12	4	3.47
2	5	9.19	7	5	5.67	12	5	5.77
3	1	8.40	8	1	8.67	13	1	11.81
3	2	4.42	8	2	4.10	13	2	6.03
3	3	4.89	8	3	6.51	13	3	4.45
3	4	2.80	8	4	5.80	13	4	5.85
3	5	4.34	8	5	5.46	13	5	4.43
4	1	9.20	9	1	9.75	14	1	9.17
4	2	4.51	9	2	9.41	14	2	5.59
4	3	6.20	9	3	9.15	14	3	5.45
4	4	4.01	9	4	9.42	14	4	5.93
4	5	5.16	9	5	7.99	14	5	7.84
5	1	10.40	10	1	9.05	15	1	8.30
5	2	9.17	10	2	6.45	15	2	8.87
5	3	9.22	10	3	5.33	15	3	8.75
5	4	7.48	10	4	5.94	15	4	8.68
5	5	8.66	10	5	6.46	15	5	6.46

the five weeks during which measurements were taken.² The first week's measurements were taken prior to the pollution alert, the second week's measurements were taken during the pollution alert, and the remaining measurements were taken during the next three weeks.

This study is an example of a *repeated measures (longitudinal) study*, in which each child (i.e., subject) has an FEV1 measurement taken at each of five consecutive weeks. The study design is *balanced* in that each subject has the *same number of repeated measurements* (i.e., five) that are *measured at the same times*.

As mentioned above, the primary study objective was to determine whether FEV1 levels were significantly lowered during and possibly after the week of the pollution alert—that is, whether the FEV1 levels measured during the first week (prior to the pollution alert) were lowered during the week of the pollution episode and possibly in the three weeks following the episode. A direct way to address this analysis objective would be to compare the mean FEV1 scores for all 40 children for each of the five weeks (see Table 25.4).

From Table 25.4, we can see that the mean FEV1 value of 9.81 for week 1, prior to the pollution alert, is higher than each of the mean FEV1 values for weeks 2 through 5. These summary data suggest that FEV1 levels were depressed (i.e., lowered) during and after the pollution alert and that, moreover, average FEV1 levels didn't seem to meaningfully change during the three weeks subsequent to the week of the pollution alert. Statistically, we might hope to show that

- a. There is a statistically significant difference among the average FEV1 scores over all five weeks.
- b. There is a statistically significant difference between the average FEV1 at week 1 and the average of the average FEV1 values over weeks 2 through 5; that is, the *linear contrast*

$$\overline{\text{FEV1}_1} - \frac{\overline{\text{FEV1}_2} + \overline{\text{FEV1}_3} + \overline{\text{FEV1}_4} + \overline{\text{FEV1}_5}}{4}$$

is statistically significantly different from zero.

TABLE 25.4 Mean FEV1 scores on 40 school children over five weeks with a pollution episode during week 2

Week <i>j</i>	<i>K</i>	Mean FEV1
1	40	9.81
2	40	6.85
3	40	7.00
4	40	6.95
5	40	7.00
Average	40	7.52

² A file containing the complete data set can be obtained from the publisher's website. The data in Table 25.3, and in the complete data set, are hypothetical, although real study data from an analogous study were described and analyzed by Laird and Ware (1982).

Using the data from Table 25.4, we obtain the following calculation:

Contrast (Week 1 vs. Weeks 2–5)

$$= 9.81 - \frac{6.85 + 7.00 + 6.95 + 7.00}{4} = 2.86$$

As we now discuss, evaluating statistical significance is a little more complicated than we might think.

The comparison being considered in item (a) above essentially is a one-way ANOVA-type comparison: the single factor (i.e., nominal independent variable) of interest is Week, the outcome Y is FEV1, and the null hypothesis of interest can be stated as

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

where μ_j denotes the population mean FEV1 score for week j that is being estimated by FEV1_j . A one-way ANOVA model, using a multiple linear regression form with reference cell coding, for this situation (see Chapter 17) can be written as follows:

$$Y = \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_3 + \beta_4 W_4 + E \quad (25.2)$$

where Y denotes FEV1, W_1 through W_4 represent four dummy variables for distinguishing among the five weeks (with the referent group chosen as desired), and E denotes the error term in the model.

We can alternatively write the model in classical one-way ANOVA format as follows:

$$Y_{ij} = \mu + \tau_j + E_{ij} \quad i = 1, \dots, 40 \text{ and } j = 1, \dots, 5 \quad (25.3)$$

where Y_{ij} denotes the FEV1 measurement at the j th week for subject i , μ denotes the overall mean FEV1 over all five weeks, τ_j denotes the effect of the j th week, E_{ij} denotes the error at the j th week for subject i , and $\sum_{j=1}^5 \tau_j = 0$. This ANOVA model assumes that the variable Week is a fixed factor.

Nevertheless, there is a problem with using a classical multiple linear regression analysis program (e.g., PROC REG in SAS), or an ANOVA program, to analyze these data: the classical regression methodology assumes that the responses for subjects in the study are mutually independent. Such an assumption is clearly questionable in this particular study because the five FEV1 measurements on each subject are likely to be correlated with one another. In other words, for the ANOVA model (25.3), we may reasonably assume that $\text{corr}(E_{ij}, E_{i'j'}) = 0$ whenever $i \neq i'$, but we must allow $\text{corr}(E_{ij}, E_{i'j'}) \neq 0$ whenever $j \neq j'$. Moreover, the five sample means (see Table 25.4) are likely correlated with each other because each of these means involves responses from the same 40 subjects in the study; consequently, any test about a contrast of such sample means must also account for such correlations. Thus, we need to use a data analysis approach that takes into account the correlations among the observations on the same subject.

The approach that we describe in this chapter generalizes the classical linear model approach by allowing for such correlations among observations on the same subject. Note, however, that the goal of both the uncorrelated response approach and the correlated

response approach is essentially the same in this setting: to describe the effect of one or more predictors (i.e., independent variables) on an outcome (i.e., dependent variable) of interest.

So far, regarding our FEV1 example involving a pollution alert, we have only focused on the primary question of interest. Recall that a *secondary objective* concerned identifying sensitive subgroups or individuals most severely affected by the pollution episode. To address this secondary question, we must carry out an analysis that provides us with information about the effect of the pollution alert on each of the 40 school children individually rather than about the effect of the pollution alert on the entire sample of school children. As we will see in Chapter 26, when we return to this question, such an individual-specific analysis requires us to consider a regression model involving one or more random effects for subjects (i.e., school children) in addition to fixed effects for the variable Week. Without specifying such a model at this point, we can nevertheless provide some insight into the structure of such a subject-specific analysis. For example, let's compare the data on the first 2 of the 40 school children in Table 25.3:

Subj.	Week	FEV1
1	1	9.43
1	2	5.71
1	3	5.86
1	4	7.70
1	5	5.89
2	1	11.15
2	2	9.48
2	3	10.11
2	4	8.89
2	5	9.19

The mean FEV1 value over the five weeks for subject 1 is 6.92, whereas the corresponding mean FEV1 value for subject 2 is 9.76. Thus, over all five weeks, subject 1 had an average that was 2.84 units lower than the average for subject 2. However, simply comparing five-week averages does not directly address how the pollution alert comparably affected each subject because the pollution alert did not begin until week 2. A more appropriate way to compare these two subjects, therefore, would be to consider the estimated linear contrast

$$L_i = \text{FEV1}_{i1} - \frac{(\text{FEV1}_{i2} + \text{FEV1}_{i3} + \text{FEV1}_{i4} + \text{FEV1}_{i5})}{4}$$

for each subject, where FEV1_{ij} is the FEV1 measurement for the i th subject ($i = 1, 2$) at week j ($j = 1, 2, 3, 4, 5$). It is easy to see that

$$L_1 = 9.43 - 6.29 = 3.14 \text{ for subject 1 and } L_2 = 11.15 - 9.42 = 1.73 \text{ for subject 2}$$

Thus, when comparing week 1 with the average of weeks 2 through 5 separately for each subject, we see that the FEV1 value for subject 1 was depressed 1.41 units more than the FEV1 value for subject 2. In other words, subject 1 was somewhat more adversely affected by the pollution alert than was subject 2.

We have just illustrated a type of subject-specific comparison that could be done to answer the secondary question. In Section 25.4, we will illustrate the analysis of the primary

question involving these data by fitting a linear mixed model containing only fixed effects using SAS's MIXED procedure. In Chapter 26, we will return to these data to consider the analysis of the secondary objective involving a mixed model containing random effects.

25.2.2 A Study of the Posture of Computer Operators

Over the past half-century, the use of computers for all kinds of professional and personal activities has grown exponentially. Accompanying this growth, public health specialists have become increasingly concerned about the possible health consequences of spending long hours operating a computer. In particular, persons whose daily work assignments require intensive computer usage may be at increased risk for developing chronic musculoskeletal disorders associated with degraded posture. Proper study of such health consequences requires a reliable measure of posture.

Ergonomists have proposed various approaches to measuring posture. In a study by Ortiz et al. (1997), the subjects were computer operators recruited from a major utility company and a large hospital in Atlanta. A primary goal of this study was to evaluate the effects of selected factors on the measurement of posture. Two such factors were *day of the work week* (early, middle, or late), denoted as Day, and time of day (AM or PM), denoted as Time. The investigators were interested in determining whether or not posture measurements taken on a given subject exhibited little within-subject variability over the days and time periods of measurement.

The Ortiz study considered several different measurements (responses) of posture on 19 subjects. Without going into full detail here, one of the measurements was labeled as Shoulder Flexion (SF), measured in degrees, with higher SF scores reflecting worsening posture. The study was designed so that each of the 19 subjects was measured sequentially in the AM and PM on Monday, Wednesday, and Friday of the same week. Thus, a total of six repeated measurements were made on each subject for this response (SF), involving all $6 (= 3 \times 2)$ combinations of the two factors (Day and Time). In ANOVA terminology, there were three levels of Day (Monday, Wednesday, and Friday), and two levels of Time (AM and PM).

The study just described is an example of *a repeated measures design involving two factors*, in which each subject is observed at each combination of levels of each of the two factors.³ This study design is *balanced* in that each subject has the same number of repeated measurements—namely, six—measured at the same times.

Table 25.5 lists the data for the posture measurement SF. In this table, the six repeated measures on subject 1 are given by the first row of observations in the table; thus, the cluster of possibly correlated measurements for subject 1 is given by

$$\{17, 5, 10, 1, 5, 1\}$$

The sample means given at the bottom of the table suggest that the SF scores are lower on Friday than on the other two days; that is, there appears to be a possible main effect of the factor Day. Also, there is some suggestion of interaction between the factors Day and Time,

³ The Ortiz study used two different ergonomists to take measurements on each subject, thus introducing a third factor (Raters) that can also be considered in the analysis of these data. However, to simplify the discussion of the repeated measures ANOVA approach, we have ignored the factor Raters in our discussion of these data. However, the analysis involving the three factors Days, Time, and Raters is considered in a problem at the end of Chapter 26.

TABLE 25.5 Shoulder flexion (SF) data from repeated measures study of posture among computer operators (Ortiz et al. 1997)

Subject Number	Sample Size	Monday		Wednesday		Friday	
		AM	PM	AM	PM	AM	PM
1	6	17	5	10	1	5	1
2	6	1	7	7	4	19	12
3	6	36	22	31	30	27	25
4	6	21	24	22	30	20	24
5	6	10	14	7	10	8	5
6	6	15	18	22	26	2	17
7	6	19	10	12	8	12	14
8	6	46	48	46	52	41	52
9	6	9	11	7	15	5	8
10	6	31	31	39	28	38	40
11	6	17	26	27	24	19	16
12	6	25	24	29	5	24	31
13	6	8	14	7	9	6	9
14	6	23	18	21	24	27	25
15	6	7	7	7	12	0	0
16	6	22	19	7	13	6	14
17	6	24	20	32	27	22	17
18	6	28	28	42	34	9	7
19	6	1	5	0	0	6	2
Sample Means:		18.95	18.47	19.74	18.53	15.58	16.79

© Cengage Learning

since the Wednesday AM mean score is higher than the Wednesday PM score, but the Friday AM mean score is lower than the Friday PM mean score.

The analysis of these data using a linear mixed model is described in the next chapter (Chapter 26), where we evaluate whether or not there are significant main effects or interaction effects involving the two factors Day and Time. We also show how this analysis can be alternatively carried out using the repeated measures ANOVA approach that partitions sums of squares due to appropriate sources of variation.

25.2.3 A Study of Treatments for Heartburn

In this example, we consider a (fictitious) study to compare two treatments for the relief of heartburn. Suppose that $K = 30$ subjects are each given two symptom-provoking meals that are spaced three days apart; upon completing each meal, each subject is given either

an active treatment (A) for heartburn or a placebo (P). The subjects have been randomly allocated to one of two groups (each containing 15 subjects), and each subject gets the same treatment for both meals. The response of interest is an index of physical discomfort measured by a questionnaire administered to each subject two hours after receiving a given treatment. This index provides a numerical value between 0 and 100, where 0 denotes no sign of any discomfort and 100 denotes extreme discomfort. The data are presented in Table 25.6.

The study considered here is a repeated measures study because two measurements are made on each subject. This study differs from the previous (posture measurement) study in that each subject in this design receives only one of the two treatments; in the posture measurement study, each subject was observed on all three days and at both time periods.

The overall averages provided at the bottom of Table 25.6 indicate that the active treatment (\bar{Y}_A) produces, on the average, lower discomfort scores than the placebo (\bar{Y}_P). The (repeated measures) ANOVA method used to assess whether the observed treatment mean difference is statistically significant is considered in a problem at the end of this chapter.

TABLE 25.6 Data layout for repeated measures study to compare treatments for heartburn

Subject	Treatment	Meal		Subject	Treatment	Meal	
		1	2			1	2
1	Active	65	55	16	Placebo	85	75
2	Active	60	70	17	Placebo	60	60
3	Active	70	70	18	Placebo	80	75
4	Active	35	30	19	Placebo	55	50
5	Active	50	50	20	Placebo	50	45
6	Active	40	40	21	Placebo	70	65
7	Active	50	65	22	Placebo	70	80
8	Active	55	35	23	Placebo	65	50
9	Active	20	50	24	Placebo	60	40
10	Active	30	70	25	Placebo	90	35
11	Active	65	80	26	Placebo	85	50
12	Active	45	70	27	Placebo	65	55
13	Active	30	45	28	Placebo	60	50
14	Active	55	65	29	Placebo	75	70
15	Active	25	45	30	Placebo	55	40
Average		46.33	56.00			68.33	56.00

Overall averages: $\bar{Y}_A = 51.17$, $\bar{Y}_P = 62.17$

25.3 General Linear Mixed Model Approach

In this section, we define and illustrate the general linear mixed model for analyzing correlated (i.e., repeated measures/cluster) data. We restrict attention to linear models that assume normally distributed outcome variables and, correspondingly, normally distributed error terms. We allow for both balanced and unbalanced data: in a *balanced* repeated measures design, each subject is observed the same number of times and at the same times; in an *unbalanced* design, the number of responses per subject may differ, and/or the different subjects may be observed at different times.

25.3.1 The Model

The general linear mixed model for analyzing correlated data can be stated in regression model format as follows:

$$Y = (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_s X_s) + (b_{i0} + b_{i1} Z_{i1} + b_{i2} Z_{i2} + \cdots + b_{iq} Z_{iq}) + E \quad (25.4)$$

We often refer to this model as the *scalar form* of the regression model to distinguish it from alternative ways to specify the model (e.g., in *matrix form*). The outcome/response variable Y is assumed to be measured several (i.e., n_i) times on the i th subject (or cluster), and there are K subjects (or clusters) in the entire study sample. The X 's in this model denote covariates that contribute *fixed effects* (the β 's) to quantify their relationships to Y , whereas the Z 's in the model denote covariates that contribute *random effects* (the b 's) to quantify their relationships to Y .⁴ Note that some X 's and Z 's can be the same variables. The E in the model denotes the error term for each observation on each subject; and we assume that the n_i responses on the same (i th) subject are correlated, so that the classical linear model assumption of mutually independent responses does not hold.

The above model can alternatively be written in *subject-specific scalar form* as follows:

$$\begin{aligned} Y_{ij} = & (\beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_s X_{ijg}) + (b_{i0} + b_{i1} Z_{ij1} + b_{i2} Z_{ij2} \\ & + \cdots + b_{iq} Z_{ijq}) + E_{ij} \end{aligned} \quad (25.5)$$

where $i = 1, 2, \dots, K$ and $j = 1, 2, \dots, n_i$. Here Y_{ij} denotes the j th response on the i th subject, X_{ijg} denotes the value of the (fixed factor) predictor X_g ($g = 1, 2, \dots, s$) for the j th response on the i th subject, Z_{ijb} denotes the value of the (random factor) predictor Z_b ($b = 1, 2, \dots, q$) for the j th response on the i th subject, and E_{ij} denotes the error term for the j th response on the i th subject. Note that the β fixed effects are not subscripted by i or j , whereas the b random effects are subscripted by i only; the β 's represent population parameters, whereas the b 's represent subject-specific random effects that may vary by subject. The use of

⁴We have introduced fixed and random effects previously in Chapter 17. In the next chapter, we will refine this explanation regarding correlated data analysis.

the word *mixed* is meant to convey the fact that the regression model for Y_{ij} contains a mixture of fixed effects (the β 's) and random effects (the b_i 's), in addition to the error term (E_{ij}).

A third form of the general linear mixed model is the *subject-specific matrix* form:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{E}_i \quad (25.6)$$

Using a matrix formulation is the most mathematically convenient and efficient way to write the general linear mixed model; this formulation treats the set of observations for each subject (the Y 's, the X 's, and the Z 's) using simplified matrix notation in **bold** lettering, as well as collectively summarizing the β 's for different X 's using $\boldsymbol{\beta}$, the b_i 's for different Z 's using \mathbf{b}_i , and the E 's for different j 's using \mathbf{E}_i . We will use the matrix format (25.6) for notational convenience at various times in this chapter; we also think that the matrix version makes this topic more understandable, even for the reader who is not completely comfortable with matrix mathematics.⁵

We now illustrate the general linear mixed model using the example described in Section 25.2.1 concerning the effect of an air pollution episode on pulmonary function measurements (FEV1) taken on $K = 40$ school children over five weeks; here $n_i = 5$, $i = 1, 2, \dots, 40$. For this example, recall that we considered the following model:

$$Y = \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_3 + \beta_4 W_4 + E \quad (25.2 \text{ repeated})$$

where $Y = \text{FEV1}$ and W_1 through W_4 represent four dummy variables for distinguishing the five weeks. This model contains only fixed factors (and correspondingly only fixed effects), so it is a special case of the scalar form of the general linear mixed model (25.4) in which there are four X 's (namely, $X_1 = W_1, \dots, X_4 = W_4$) and no Z variables (and correspondingly no random effects).

The subject-specific scalar form of this model is given by

$$Y_{ij} = \beta_0 + \beta_1 W_{ij1} + \beta_2 W_{ij2} + \beta_3 W_{ij3} + \beta_4 W_{ij4} + E_{ij} \quad i = 1, \dots, 40, j = 1, \dots, 5 \quad (25.7)$$

where W_{ijg} denotes the dummy variable value (either 0 or 1) corresponding to the j th week for the g th dummy variable, $g = 1, 2, 3, 4$. If the (scalar) dummy variables W_1, W_2, W_3 , and W_4 are defined so that the referent group is week 5, then for measurements taken during the third week,

$$W_{i33} = 1 \text{ and } W_{i31} = W_{i32} = W_{i34} = 0$$

and for measurements taken during the fifth (referent) week,

$$W_{i51} = W_{i52} = W_{i53} = W_{i54} = 0$$

⁵ See Appendix B on Matrices and Their Relationship to Regression Analysis for a description of matrix mathematics (e.g., matrix multiplication).

The subject-specific matrix form of this model is

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{E}_i \quad i = 1, \dots, 40 \quad (25.8)$$

where \mathbf{Y}_i denotes the collection of five FEV1 measurements on the i th child, \mathbf{X}_i denotes the collection of dummy variable values for each of the five weeks for subject i , $\boldsymbol{\beta}$ denotes the parameter vector for the set of β 's in this model, and \mathbf{E}_i denotes the set of error terms for the i th subject. Assuming again that the referent group is week 5, then \mathbf{X}_i , $\boldsymbol{\beta}$, and \mathbf{E}_i can be written as follows:

$$\mathbf{X}_i = \begin{bmatrix} 1 & W_{i11} & W_{i12} & W_{i13} & W_{i14} \\ 1 & W_{i21} & W_{i22} & W_{i23} & W_{i24} \\ 1 & W_{i31} & W_{i32} & W_{i33} & W_{i34} \\ 1 & W_{i41} & W_{i42} & W_{i43} & W_{i44} \\ 1 & W_{i51} & W_{i52} & W_{i53} & W_{i54} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}, \quad \mathbf{E}_i = \begin{bmatrix} E_{i1} \\ E_{i2} \\ E_{i3} \\ E_{i4} \\ E_{i5} \end{bmatrix} \quad (25.9)$$

Note that the \mathbf{X}_i matrix contains five rows (for the five weeks) and five columns (for the intercept variable, always taking the value 1, and the four dummy variables).

If the referent group is week 5, it also follows (as for any one-way ANOVA model) that the estimated coefficients (i.e., the $\hat{\beta}$'s) in all three models (25.2, 25.7, 25.8) are determined as follows:

$$\begin{aligned} \hat{\beta}_0 &= \overline{\text{FEV1}}_5 \\ \hat{\beta}_1 &= \overline{\text{FEV1}}_1 - \overline{\text{FEV1}}_5, \quad \hat{\beta}_2 = \overline{\text{FEV1}}_2 - \overline{\text{FEV1}}_5, \quad \hat{\beta}_3 = \overline{\text{FEV1}}_3 - \overline{\text{FEV1}}_5, \\ \hat{\beta}_4 &= \overline{\text{FEV1}}_4 - \overline{\text{FEV1}}_5 \end{aligned} \quad (25.10)$$

In other words, if the referent group is week 5, then the estimated intercept $\hat{\beta}_0$ will be the mean FEV1 value at week 5, and the estimated regression coefficient $\hat{\beta}_g$, $g = 1, 2, 3, 4$, will be the difference between the mean FEV1 value at week g and the mean FEV1 value at week 5. These estimated coefficients are identical to the estimates that would result if we assumed that the observations on the same subject were mutually independent (i.e., the classical one-way ANOVA assumption). Nevertheless, the variances of the estimates in (25.10) will differ, depending on the manner and extent to which the observations on the same subject are correlated; consequently, statistical inference procedures about the regression coefficients will vary with the “correlation structure” of the observations on the same subject.

For the FEV1 data, there was a secondary study objective: to identify the extent to which the pollution episode severely affected some study subjects more than other subjects. This secondary objective requires a regression model different from the fixed effect model given by (25.2) or, equivalently (25.7) and (25.8). This new (subject-specific) model now requires one or more random effects for subjects in addition to the fixed effects. The subject-specific scalar form of an appropriate linear mixed model for this situation is

$$Y_{ij} = \beta_0 + \beta_1 W_{ij1} + \beta_2 W_{ij2} + \beta_3 W_{ij3} + \beta_4 W_{ij4} + b_{i0} + b_{i1} Z_{ij1} + E_{ij} \quad (25.11)$$

which contains two random effects b_{i0} and b_{i1} and the random coefficient factor Z_{ij1} in addition to the fixed factor dummy variables W_{ij1} through W_{ij4} and their corresponding fixed effects (i.e., β_1 through β_4). The Z_{ij1} variable is defined as follows:

$$Z_{ij1} = 0 \text{ if week 1 } (j = 1) \text{ and } Z_{ij1} = 1 \text{ if weeks 2, 3, 4 or 5 } (j \neq 1)$$

(We can equivalently restate b_{i0} in model (25.11) as $b_{i0}Z_{ij0}$, where Z_{ij0} is identically equal to 1 for each subject at each week; i.e., model (25.11) effectively considers two random coefficient factors Z_{ij0} and Z_{ij1}).

In Chapter 26, where we consider random effects for these data, we will describe and illustrate with computer results why this model is appropriate. Here we show the subject-specific matrix form of this model:

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \mathbf{E}_i \quad i = 1, 2, \dots, 40 \quad (25.12)$$

where \mathbf{Y}_i , \mathbf{X}_i , $\boldsymbol{\beta}$, and \mathbf{E}_i are the same as in model (25.8) and matrix expression (25.9) and where \mathbf{b}_i and \mathbf{Z}_i are defined as follows:

$$\mathbf{b}_i = \begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix}, \quad \mathbf{Z}_i = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \quad (25.13)$$

The first column (i.e., the column of 1's) of the \mathbf{Z}_i matrix denotes the values of the variable Z_{i0} corresponding to the random effect b_{i0} for each of the five observations on subject i , and the second column denotes the values of the variable Z_{i1} corresponding to the random effect b_{i1} for the five observations on subject i .

In Chapter 26, we show that use of the above random effects model (25.11) will lead to identical estimated regression coefficients given by (25.10) as when using the fixed effects model (25.7). Nevertheless, the estimated variances of these estimated coefficients based on (25.11) will differ from corresponding estimated variances for the fixed effects model (25.7) even when correlation structures are appropriately taken into account. As we shall discuss in the next section, corresponding regression coefficient variances in the two models are different because the two models have different correlation structures.

25.3.2 Correlation Structures

The correlation structure involving a set of observations

$$\mathbf{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{bmatrix}$$

on the same subject (i) is defined by the set of all pairwise correlations of the form

$$\rho_{ijj'} = \text{Corr}(Y_{ij}, Y_{ij'})$$

where j and j' denote two possibly different observations of Y for the i th subject. Recall that, in general, $|\rho_{ijj'}| \leq 1$ for any i, j , and j' and that $\rho_{jj'} \equiv 1$ whenever $j = j'$.

For example, in the FEV1 study, the correlation structure for the i th child can be generally described by the following (5 rows \times 5 columns) correlation matrix:

$$\mathbf{C}_i = \begin{bmatrix} 1 & \rho_{i12} & \rho_{i13} & \rho_{i14} & \rho_{i15} \\ \rho_{i12} & 1 & \rho_{i23} & \rho_{i24} & \rho_{i25} \\ \rho_{i13} & \rho_{i23} & 1 & \rho_{i34} & \rho_{i35} \\ \rho_{i14} & \rho_{i24} & \rho_{i34} & 1 & \rho_{i45} \\ \rho_{i15} & \rho_{i25} & \rho_{i35} & \rho_{i45} & 1 \end{bmatrix} \quad (25.14)$$

Any such correlation matrix is always “symmetric,” since $\rho_{ijj'} \equiv \rho_{j'j}$; in other words, the set of correlations above the diagonal of 1’s is the identical mirror image of the set of correlations below this diagonal.

The simplest form that a subject-specific correlation structure (\mathbf{C}_i) can take is the *independent correlation structure* (or matrix); that is,

$$\mathbf{C}_i^{(\text{IND})} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (25.15)$$

This is the structure that is assumed in classical ANOVA when it is assumed that all observations on all subjects are mutually independent. The above matrix, with ones on the diagonal and zeros elsewhere, is typically referred to as an *identity matrix* and is denoted as \mathbf{I} . In particular, the above (5×5) identity matrix would be written as \mathbf{I}_5 ; and, in general, an $(n \times n)$ identity matrix would be written as \mathbf{I}_n .

When one wants to consider nonzero correlations among observations on the same subject, there is a large variety of choices for the correlation structure. For example,

an exchangeable correlation structure assumes that all the correlations between pairs of observations on a given subject are identical; that is,

$$\mathbf{C}_i^{(\text{EXCH})} = \begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{bmatrix} \quad (25.16)$$

Notice that the above exchangeable correlation structure (or matrix) also assumes that the correlation between any pair of observations on any subject *does not vary from subject to subject*. This assumption, stated mathematically, says that

$$\rho_{ij'} = \rho_{jj'} \equiv \rho \quad \text{for all } i \quad (25.17)$$

An alternative version of an exchangeable correlation structure would allow the pairwise correlations to be identical for a given subject but different for different subjects; that is,

$$\rho_{ij'} \equiv \rho_i \quad (25.18)$$

Another correlation structure called *autoregressive 1 (AR1)* assumes that observations “closer to each other” (e.g., in time or space) are more highly correlated than observations that are “farther apart.” The pairwise correlations decrease according to the following formula:

$$\rho_{ij'} = \rho^{|T_{ij} - T_{j'}|} \quad \text{when } j \neq j' \quad (25.19)$$

where T_{ij} and $T_{j'}$ denote the times at which the j th and j' th observations, respectively, are measured on the i th subject. If subjects are measured at equally spaced time intervals (i.e., $|T_{ij} - T_{j'}| = c|j - j'|$ for some positive constant c), the corresponding (5×5) AR1 correlation matrix is given by

$$\mathbf{C}_i^{(\text{AR1})} = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix} \quad (25.20)$$

For this matrix, the correlation ρ between the 1st and 2nd observations is the same as the correlations between the 2nd and 3rd, 3rd and 4th, and 4th and 5th observations; this is because the two responses in each of these pairs are exactly one equally spaced time occasion apart. In contrast, the correlation ρ^2 between the 1st and 3rd observations, which are two equally spaced time occasions apart, is the same as the correlations between the 2nd and 4th and the 3rd and 5th observations. Also, notice that, as with our example of an exchangeable

correlation structure, we have assumed that the above AR1 correlation structure is the same for all subjects.

There are many other correlation structures that one might consider. Another structure often considered, called the *unstructured correlation structure*, takes the following form, again assuming that such a structure does not vary by subject:

$$\mathbf{C}_i^{(UN)} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{15} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} & \rho_{25} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} & \rho_{35} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 & \rho_{45} \\ \rho_{15} & \rho_{25} & \rho_{35} & \rho_{45} & 1 \end{bmatrix} \quad (25.21)$$

An unstructured correlation matrix has the property that all possible correlations are allowed to be different. Such a correlation matrix is often considered as a starting point for assessing whether a simpler correlation structure (e.g., exchangeable) might suffice or when it is unclear what simpler structure might be specified.

The *assumption* in the above examples that the *underlying correlation structure is the same for all subjects* (i.e., $\rho_{ijj'} = \rho_{jj'}$) is typically made to avoid considering a model that has more parameters than can be estimated validly and precisely with the available data (i.e., the model is *over-parameterized*, and there is not enough information to yield reliable parameter estimates). In contrast, assuming that each subject has the same set of correlation parameters reduces the number of correlation parameters substantially for a study containing K subjects. In our FEV1 pollution study, for example, with $K = 40$, if we assume the same AR1 correlation structure for each subject, then the number of possible correlation parameters to be estimated is reduced from 40 to 1.

We have so far considered a correlation matrix \mathbf{C}_i at the subject-specific level. Alternatively, we might consider the larger correlation matrix for *all* study subjects. If, for example, there are K subjects, each of whom has n observations on Y , then the correlation matrix—say, \mathbf{C} —for all K subjects will have Kn rows and Kn columns (e.g., for $K = 40$ and $n = 5$, the dimensions of \mathbf{C} will be 200×200). When considering the overall correlation matrix \mathbf{C} , an important characteristic of repeated measures/clustered data is the assumption that *different responses within the same cluster* (e.g., Y_{ij} and $Y_{ij'}$, $j \neq j'$) are correlated, whereas responses from *different clusters* (e.g., Y_{ij} and $Y_{i'j'}$, $i \neq i'$) are uncorrelated. Thus, in the FEV1 study involving $n = 5$ responses on each of $K = 40$ subjects, there are $Kn = 200$ response observations in total, and so the corresponding correlation matrix for all study subjects will be of the following form:

$$\mathbf{C}(200 \times 200) = \begin{bmatrix} \mathbf{C}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{C}_3 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C}_{40} \end{bmatrix} \quad (25.22)$$

Each of the 40 components (written in bold) within \mathbf{C} is a (5×5) matrix. The 40 \mathbf{C}_i matrices within \mathbf{C} are the correlation matrices for all $K = 40$ subjects. The $\mathbf{0}$ matrices, each of which is (5×5) and has all entries equal to zero, reflect the assumption that the observations for any one subject are independent of the observations for any other subject. This large \mathbf{C} matrix is called a *block diagonal matrix*, since the \mathbf{C}_i 's on the diagonal of the matrix are the only nonzero matrices.

If we assume that all subjects have the same set of correlation parameters and that $n_i \equiv n$ for all i , the above block diagonal matrix will simplify to the following block diagonal matrix:

$$\mathbf{C}(200 \times 200) = \begin{bmatrix} \mathbf{C} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{C} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C} \end{bmatrix} \quad (25.23)$$

A crucial question about the above discussion of correlation structures is “How do we choose the correlation structure appropriate for the data set we are analyzing?”

The answer to this question is not straightforward but rather requires consideration of several approaches from which the investigator(s) must make a reasoned decision. Typically, but not always, the investigator makes an “educated guess” about the appropriate correlation structure (e.g., exchangeable), fits a model using the guessed or postulated structure, tries other “guesses” that might also be reasonable, and compares numerical results based on the use of different structures. A postulated correlation structure is typically referred to as a *working correlation structure*. Choices among different correlation structures might be based on which correlation structures are most biologically/clinically reasonable, on which models have better “goodness of fit” to the data, or on the extent to which the modeling results vary for different structures. In particular, if the conclusions about the study questions turn out to be essentially the same, regardless of which correlation structure (among reasonable guesses) is chosen, then any of these correlation structures could suffice. (Note that conclusions can still be invalid if the wrong regression model is used.) In general, it is more important to make reasonably valid statistical conclusions about the study questions than to determine exactly the perfectly correct correlation structure.

One other frequently used method related to the choice of the correlation structure involves accounting for the possibility that a chosen working correlation structure is actually not the correct correlation structure. The method used is generally referred to as *empirical (or robust) variance estimation*. A so-called robust/empirical standard error estimator is a modified standard error estimator for an estimated regression coefficient that combines an estimator of the chosen working correlation structure with an estimator of the correct (but unknown) correlation structure to produce an asymptotically unbiased estimator of the standard error.

When testing for the significance of a specific regression coefficient—say, $\hat{\beta}$ —the test statistic used is of the form

$$T = \frac{\hat{\beta}}{S_{\hat{\beta}, \text{emp}}}$$

where $S_{\hat{\beta}, \text{emp}}$ denotes the empirical standard error estimate for the estimated regression coefficient $\hat{\beta}$. If the assumed linear mixed model is correctly specified, then T will have an approximate t distribution⁶ (and, for large samples, an approximate Z distribution) under the null hypothesis $H_0: \beta = 0$. Because this empirical standard error estimate involves in its computation the choice of the working correlation structure, the numerical value of this estimate can be somewhat different for different choices of the working correlation structure, especially for small and intermediate size samples. In other words, it is possible that the use of, say, an exchangeable working correlation structure may yield a significant test result, whereas the use of, say, an AR1 working correlation structure may yield a nonsignificant test result, or vice versa. So, even though the use of an empirical standard error estimate helps to correct for possible misspecification of the correlation structure, the choice of the working correlation structure can still be an important concern when sample sizes are not large.

25.3.3 Covariance Structures and Their Relationship to Correlation Structures

As we have previously described (in Chapter 6), the correlation parameter is a measure of the linear association between two random variables, X and Y , that generally ranges between -1 and 1 . It is a dimensionless, scale-invariant measure. In a repeated measures study, these two random variables can be two different responses, Y_{ij} and $Y_{ij'}$, for the same outcome variable Y measured on the same subject i . In this chapter, we have used the notation $\rho_{ijj'} = \text{corr}(Y_{ij}, Y_{ij'})$ for such a correlation.

An alternative measure of association between two variables that has a specific mathematical relationship with the correlation is the *covariance*. The covariance between X and Y is defined as the expected value, or population average, of the product of X minus its mean (μ_X) and Y minus its mean (μ_Y); that is,

$$\nu_{X,Y} = \text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Note that $\nu_{X,Y}$ has dimensions and is not scale-invariant. Also, the covariance of X with itself is called the variance of X ; that is,

$$\nu_{X,X} = \text{Cov}(X, X) \equiv \text{Var}(X)$$

Analogously, if we are considering repeated measures data

$$\mathbf{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{bmatrix} \quad \text{for subject } i$$

⁶ A discussion of the options for the appropriate denominator degrees of freedom (DDFM) associated with t and F statistics for the predictors in one's model is provided in Section 26.4.7.

we can describe the covariance between any two observations for subject i , and the corresponding variance of any one of these observations, as

$$\nu_{ijj'} = \text{Cov}(Y_{ij}, Y_{ij'}) = E[(Y_{ij} - \mu_{Y_{ij}})(Y_{ij'} - \mu_{Y_{ij'}})]$$

and

$$\sigma_{ij}^2 = \nu_{ijj} = \text{Cov}(Y_{ij}, Y_{ij}) = \text{Var}(Y_{ij})$$

The mathematical relationship between the correlation and the covariance can be stated as follows:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}, \quad \text{that is, } \rho_{XY} = \frac{\nu_{XY}}{\sigma_X\sigma_Y} \quad (25.24)$$

Analogously, for repeated measures data, we can write this relationship as

$$\text{Corr}(Y_{ij}, Y_{ij'}) = \frac{\text{Cov}(Y_{ij}, Y_{ij'})}{\sqrt{\text{Var}(Y_{ij})\text{Var}(Y_{ij'})}}, \quad \text{that is, } \rho_{ijj'} = \frac{\nu_{ijj'}}{\sigma_{ij}\sigma_{ij'}} \quad (25.25)$$

We can alternatively, using simple algebra, write the covariance in terms of the correlation as follows:

$$\text{Cov}(Y_{ij}, Y_{ij'}) = \text{Corr}(Y_{ij}, Y_{ij'})\sqrt{\text{Var}(Y_{ij})\text{Var}(Y_{ij'})} \quad (25.26)$$

The above scalar relationships between covariance and correlation can also be described in matrix terms. In particular, corresponding to a given covariance matrix (also called a variance–covariance matrix) \mathbf{V} , there is a correlation matrix \mathbf{C} .

The simplest form of *covariance* matrix corresponds to an independent *correlation* matrix $\mathbf{C}_i^{(\text{IND})}$, as illustrated by (25.15) for the identity matrix \mathbf{I}_5 . The corresponding independent covariance matrix $\mathbf{V}_i^{(\text{IND})}$ will have all response variable covariances (located off the diagonal) equal to zero, with the (response) variances on the diagonal. If all the variances are assumed to be equal to σ^2 , say, for all ($n_i = 5$) responses, then we can write $\mathbf{V}_i^{(\text{IND})} = \sigma^2\mathbf{I}_5$. If, however, the variances for different responses are not assumed to be equal (i.e., $\text{Var}(Y_{ij}) = \sigma_j^2$ for $j = 1, \dots, n_i$), then the covariance matrix corresponding to a (5×5) independent correlation structure ($n_i = 5$) will take the following form:

$$\mathbf{V}_i^{(\text{IND})} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_4^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_5^2 \end{bmatrix} \quad (25.27)$$

As another example, for the FEV1 study, an exchangeable correlation matrix for the i th subject (assuming $\rho_{ijj'} = \rho_{jj'} \equiv \rho$ for all i) would be given by the following (5×5) correlation matrix:

$$\mathbf{C}_i^{(\text{EXCH})} = \begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{bmatrix} \quad (25.28)$$

A covariance matrix that corresponds to the above correlation matrix is given by

$$\mathbf{V}_i^{(\text{Het EXCH})} = \begin{bmatrix} \sigma_j^2 & v_{12} & v_{13} & v_{14} & v_{15} \\ v_{12} & \sigma_j^2 & v_{23} & v_{24} & v_{25} \\ v_{13} & v_{23} & \sigma_j^2 & v_{34} & v_{35} \\ v_{14} & v_{24} & v_{34} & \sigma_j^2 & v_{45} \\ v_{15} & v_{25} & v_{35} & v_{45} & \sigma_j^2 \end{bmatrix} \quad (25.29)$$

where $v_{ijj'} = v_{jj'} = \rho\sigma_j\sigma_{j'}$ for $j \neq j'$. The covariance matrix in (25.29) allows for different covariances for all $j \neq j'$, even though the corresponding exchangeable correlation matrix $\mathbf{C}_i^{(\text{EXCH})}$ in (25.28) assumes that all pairwise correlations are identical. Unequal covariances are obtained here because there are different variances σ_j^2 for different observations on subject i . We call such a covariance matrix a *heterogeneous exchangeable covariance matrix* (hence the notation **Het EXCH**).

If all the variances are assumed to be the same (i.e., $\sigma_j^2 \equiv \sigma^2$), the covariance structure would simplify as follows:

$$\mathbf{V}_i^{(\text{Hom EXCH})} = \begin{bmatrix} \sigma^2 & v & v & v & v \\ v & \sigma^2 & v & v & v \\ v & v & \sigma^2 & v & v \\ v & v & v & \sigma^2 & v \\ v & v & v & v & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{bmatrix} = \sigma^2 \mathbf{C}_i^{(\text{EXCH})} \quad (25.30)$$

where $v_{jj'} \equiv \rho\sigma^2 \equiv v$; that is, the covariance is the same between any two observations j and j' for subject i . The above matrix is a *homogeneous exchangeable covariance matrix*. (A popular special case of this matrix is a *compound symmetric (CS) matrix*, which requires ρ to be nonnegative.) Notice that the two covariance structures $\mathbf{V}_i^{(\text{Het EXCH})}$ and $\mathbf{V}_i^{(\text{Hom EXCH})}$ yield the same exchangeable correlation structure $\mathbf{C}_i^{(\text{EXCH})}$.

The scalar relationship between covariance and correlation given in equation (25.26) can be expressed generally in matrix form as follows:

$$\mathbf{V}_i = \mathbf{D}_i^{1/2} \mathbf{C}_i \mathbf{D}_i^{1/2} \quad (25.31)$$

where \mathbf{V}_i and \mathbf{C}_i denote the covariance matrix and its corresponding correlation matrix, respectively, and $\mathbf{D}_i^{1/2}$ denotes a diagonal matrix whose diagonal elements are the square roots of corresponding variances obtained from the \mathbf{V}_i matrix. For example, for the (5×5) heterogeneous covariance matrix (25.29), the corresponding (5×5) diagonal matrix $\mathbf{D}_i^{1/2}$ is given by

$$\mathbf{D}_i^{1/2} = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 & 0 \\ 0 & 0 & 0 & \sigma_4 & 0 \\ 0 & 0 & 0 & 0 & \sigma_5 \end{bmatrix} = \text{Diag}(\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5) \quad (25.32)$$

For the (5×5) exchangeable covariance matrix given by (25.30), in which all the variances are assumed equal (to σ^2), the corresponding (5×5) diagonal matrix $\mathbf{D}_i^{1/2}$ is given by

$$\mathbf{D}_i^{1/2} = \begin{bmatrix} \sigma & 0 & 0 & 0 & 0 \\ 0 & \sigma & 0 & 0 & 0 \\ 0 & 0 & \sigma & 0 & 0 \\ 0 & 0 & 0 & \sigma & 0 \\ 0 & 0 & 0 & 0 & \sigma \end{bmatrix} = \sigma \mathbf{I}_5$$

Thus, we have seen that the pairwise associations among repeated continuous responses on the same subject can alternatively be described by correlation or covariance structures, which can be described easily in matrix form. Further, the mathematical relationship between correlation and covariance matrices can be expressed succinctly using (25.31).

25.3.4 Covariance Structure for the General Linear Mixed Model

We have previously defined the general linear mixed model in both scalar and matrix forms, as follows:

Regression Scalar Form:

$$Y = (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_s X_s) + (b_0 + b_1 Z_1 + b_2 Z_2 + \cdots + b_q Z_q) + E$$

or, equivalently, using summation notation:

$$Y = \beta_0 + \sum_{g=1}^s \beta_g X_g + b_0 + \sum_{b=1}^q b_b Z_b + E \quad (25.33)$$

Subject-specific Scalar Form:

$$Y_{ij} = (\beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_s X_{ijq}) + (b_{i0} + b_{i1} Z_{ij1} + b_{i2} Z_{ij2} + \cdots + b_{iq} Z_{ijq}) + E_{ij}$$

where $i = 1, 2, \dots, K$ and $j = 1, 2, \dots, n_i$, or, equivalently, using summation notation:

$$Y_{ij} = \beta_0 + \sum_{g=1}^s \beta_g X_{ijg} + b_{i0} + \sum_{b=1}^q b_{ib} Z_{ijb} + E_{ij} \quad (25.34)$$

Subject-specific Matrix Form:

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \mathbf{E}_i \quad i = 1, 2, \dots, K \quad (25.35)$$

In each of the above (equivalent) models, there are two random components: the collection of random effects (denoted by \mathbf{b}_i in the matrix version) and the collection of random errors (denoted by \mathbf{E}_i in the matrix version).

Since \mathbf{b}_i contains $(q + 1)$ scalar random components—that is,

$$\mathbf{b}_i = \begin{bmatrix} b_{i0} \\ b_{i1} \\ b_{i2} \\ \vdots \\ b_{iq} \end{bmatrix}$$

its associated covariance and correlation matrices are $(q + 1) \times (q + 1)$ matrices that we denote as \mathbf{G} and \mathbf{C}^G , respectively.

Similarly, \mathbf{E}_i contains n_i scalar random components—that is,

$$\mathbf{E}_i = \begin{bmatrix} E_{i1} \\ E_{i2} \\ E_{i3} \\ \vdots \\ E_{in_i} \end{bmatrix}$$

and its associated covariance and correlation matrices are $(n_i \times n_i)$ matrices that we denote as \mathbf{R}_i and \mathbf{C}^{R_i} , respectively. We typically assume that each subject has the same number $(q + 1)$ of random components and that the matrices \mathbf{G} and \mathbf{C}^G do not vary with i . However, since the number of observations per subject (n_i) may vary with i , we need to retain the subscript i when denoting the $(n_i \times n_i)$ \mathbf{R}_i matrix.

Since the subject-specific response vector \mathbf{Y}_i can be written as the linear sum of three vectors, one of which ($\mathbf{X}_i \boldsymbol{\beta}$) involves fixed (i.e., nonrandom) parameters and the other two of which ($\mathbf{Z}_i \mathbf{b}_i$ and \mathbf{E}_i) involve random variables, it is possible to express the covariance structure \mathbf{V}_i of \mathbf{Y}_i (as well as the corresponding correlation structure \mathbf{C}_i) in terms of the covariance structures \mathbf{G} and \mathbf{R}_i .

The matrix version for this relationship is

$$\mathbf{V}_i = \mathbf{Z}_i \mathbf{G} \mathbf{Z}'_i + \mathbf{R}_i \quad (25.36)$$

Equation (25.36) tells us that we can specify the covariance structure \mathbf{V}_i of \mathbf{Y}_i by choosing the \mathbf{G} and \mathbf{R}_i matrices that correspond to the random effects and the error terms in our model. When using computer procedures (such as SAS PROC MIXED) to fit a general linear mixed model, the user is required to do exactly this—that is, to specify the \mathbf{G} and \mathbf{R}_i matrices that will identify the working covariance structure (or matrix) \mathbf{V}_i to be used for model fitting.

From equation (25.36), if the user does not want to consider any random effects in the model, then the user specifies that $\mathbf{G} = \mathbf{0}$; the covariance structure is then determined entirely by the \mathbf{R}_i matrix, which involves variances and covariances for the n_i error components in \mathbf{E}_i . Such a model (without random effects) is typically referred to as a *marginal model*.⁷ As mentioned previously, the choice of covariance/correlation structure, in this case (when $\mathbf{G} = \mathbf{0}$) only for \mathbf{R}_i , represents a “guess” from consideration of several reasonable choices. We have already mentioned several possible choices, including exchangeable (EXCH), autoregressive (AR1), and unstructured (UN). Most computer packages offer a large selection of choices in addition to these for a correlation/covariance structure for \mathbf{R}_i and/or \mathbf{G} . We have also mentioned that an *empirical* (or *robust*) estimator of the standard error of any estimated regression coefficient can be computed to account for a possibly incorrect choice of one’s “guessed” covariance/correlation structure.

Whether or not the model contains random effects (i.e., whether or not $\mathbf{G} = \mathbf{0}$), the simplest choice for the \mathbf{R}_i matrix is $\sigma_e^2 \mathbf{I}_{n_i}$, where σ_e^2 denotes the “common variance” for each observation (Y_{ij}) on the i th subject. If, in fact, the researcher specifies a nonzero \mathbf{G} matrix when using $\sigma_e^2 \mathbf{I}_{n_i}$ for the \mathbf{R}_i matrix, we then say that the covariance matrix \mathbf{V}_i for \mathbf{Y}_i has a “conditionally independent” structure; by this, we mean that once we “condition on” (i.e., fix) the random effects, the observations $Y_{i1}, Y_{i2}, \dots, Y_{in_i}$ are (conditionally) mutually independent.

In general, since \mathbf{G} may be chosen not equal to $\mathbf{0}$ (and \mathbf{R}_i is always nonzero), the form that the correlation/covariance structure for \mathbf{Y}_i takes will depend on the combined choices for both the \mathbf{G} and the \mathbf{R}_i matrices. The fact that both matrices may be specified for a given model, therefore, allows the investigator great flexibility in choosing the working covariance matrix \mathbf{V}_i .

There is at least one situation⁸ in which two different choices for \mathbf{G} and \mathbf{R}_i will result in the same overall covariance structure \mathbf{V}_i . This situation occurs when either

- (i) $\mathbf{G} = \mathbf{0}$ and $\mathbf{R}_i = \sigma_e^2 \mathbf{C}_i^{(\text{EXCH})} = \mathbf{CS}$, $\rho \geq 0$ (a marginal model with a compound symmetric correlation structure), or
- (ii) $\mathbf{G} = \sigma_0^2$ a scalar, and $\mathbf{R}_i = \sigma_e^2 \mathbf{I}_{n_i}$ (a random effects model with a single random intercept)

⁷ A marginal model can be defined using the general linear mixed model formula $\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \mathbf{E}_i$ by re-expressing this formula as $\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{E}_i^*$, where $\mathbf{E}_i^* = \mathbf{Z}_i \mathbf{b}_i + \mathbf{E}_i$. If one is not interested in differentiating $\mathbf{Z}_i \mathbf{b}_i$ from \mathbf{E}_i , then the variances and covariances among the elements of \mathbf{E}_i^* are summarized in a \mathbf{V}_i matrix (which is still a function of \mathbf{G} and \mathbf{R}_i but where now the specific forms for \mathbf{G} and \mathbf{R}_i are not of interest).

⁸ Using (i) together with the variance–covariance formula (25.36), it follows that

$$\mathbf{V}_i = \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i' + \mathbf{R}_i = \mathbf{0} + \mathbf{CS} = \mathbf{CS}$$

Using (ii) and matrix algebra (details omitted, but see Appendix B), where \mathbf{Z}_i is a $(n_i \times 1)$ column matrix of ones, it follows that

$$\mathbf{V}_i = \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i' + \mathbf{R}_i = \sigma_0^2 \mathbf{Z}_i \mathbf{Z}_i' + \sigma_e^2 \mathbf{I}_{n_i} = \sigma^2 \mathbf{C}_i^{(\text{EXCH})} = \mathbf{CS}$$

since $\sigma^2 = (\sigma_0^2 + \sigma_e^2)$ and $\rho = \sigma_0^2 / (\sigma_0^2 + \sigma_e^2) \geq 0$.

For a given data set, the investigator then must use a computer program (e.g., SAS's MIXED procedure) to carry out the analysis, specifying the following for each i , $i = 1, 2, \dots, K$:

1. The outcome vector \mathbf{Y}_i ;
2. The predictors \mathbf{X}_i (covariates for fixed effects $\boldsymbol{\beta}$) and \mathbf{Z}_i (covariates for random effects \mathbf{b}_i);
3. The correlation/covariance structure of the model in terms of the matrices \mathbf{G} and \mathbf{R}_i ; and
4. Whether or not robust/empirical standard error estimates are to be used to correct for the possibility of incorrectly specifying the \mathbf{G} and \mathbf{R}_i matrices.

Also, the data layout for the computer should agree with the general format described in Table 25.2. In the next section, we illustrate such an analysis for the Air Pollution Study previously described.

25.3.5 ML Estimation in the General Linear Mixed Model

By exploiting the assumption that the random components in \mathbf{b}_i and \mathbf{E}_i for the general linear mixed model (25.35) are jointly normally distributed, estimation of effects (the β 's and b 's) in the general linear mixed model is typically carried out using likelihood-based methods. Two such likelihood-based methods implemented in SAS's MIXED procedure (and the analogous procedures found in other statistical programs) are *maximum likelihood* (ML) and *restricted maximum likelihood* (REML), both of which involve maximizing (log) likelihood functions (Littell et al. 1996). REML estimation differs from ML estimation; the former method adjusts certain parameter estimates for the number ($p + 1$) of fixed effect parameters being estimated, so that REML estimators generally are equivalent to unbiased estimators obtained from ANOVA methods (via partitioning sums of squares) for balanced data sets. In contrast, ML estimators are always slightly biased and, when the data are unbalanced, are typically more biased than REML estimators. This is why REML is the default choice for SAS's MIXED procedure, although both REML and ML estimates will differ negligibly for large samples.

Regardless of whether REML or ML is used, tests of hypotheses about fixed effects parameters can be carried out using Wald and/or likelihood ratio (LR) tests, as described previously in Chapter 21. If, as is typically assumed, the response variable (Y_{ij}) is normally distributed, then (approximate) t tests and F tests can be used instead of Wald tests and LR tests, respectively. Also, confidence interval estimation for fixed effects parameters is carried out using previously described large-sample methods. Statistical inference methods for random effects, however, are somewhat more complicated and are described in Section 26.4.6.

25.4 Example: Study of Effects of an Air Pollution Episode on FEV1 Levels

In this section, we illustrate the use of a linear mixed model using the previously described data involving FEV1 measurements on 40 school children before, during, and after an air pollution episode. We describe and compare the results from fitting marginal models to the

data, considering both model-based (non-empirical) standard errors and empirical standard errors. We also illustrate results for unbalanced data obtained by arbitrarily deleting several observations from the original FEV1 data set. The computational details for fitting these models using SAS's MIXED procedure are discussed in Appendix C.

25.4.1 Model-based Results for Marginal Models: FEV1 Data

To illustrate the results from fitting a linear mixed model, we once again consider the analysis of the FEV1 data. Table 25.7 provides edited computer output for several choices of correlation/covariance structures for marginal models (i.e., $\mathbf{G} = \mathbf{0}$) considered for these data. The “marginal” linear mixed model used here takes the form

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{E}_i \quad i = 1, \dots, 40 \quad (25.8 \text{ repeated})$$

where \mathbf{Y}_i denotes the collection of five FEV1 measurements on the i th child and where \mathbf{X}_i , $\boldsymbol{\beta}$, and \mathbf{E}_i are defined in (25.9). Note that the dummy variables (X_1, X_2, X_3, X_4) are defined so that the referent group is week 5. Also, *model-based* (non-empirical) standard errors are shown here; empirical standard errors will be shown separately later. The (marginal model) covariance structure used for each computer run is of the form $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$, where $\mathbf{G} = \mathbf{0}$ always and where \mathbf{R}_i is chosen to be independent (IND), compound symmetric (CS), autoregressive (AR1), or unstructured (UN).

Column 1 in Table 25.7 presents the estimated regression coefficients and associated standard errors for the intercept and for each of the four dummy variables. The estimated regression coefficients are identical for all four choices of the correlation structure (i.e., \mathbf{R} matrix), yet the standard errors are different for each different choice of \mathbf{R} . In general, the estimated regression coefficients also differ with the choice of \mathbf{R} , but for these balanced data, all choices for \mathbf{R} yield the same estimated coefficients. We will later show that, when the data set is unbalanced, the estimated regression coefficients vary as a function of \mathbf{R} . The fact that the standard errors differ with the choice of \mathbf{R} should not be surprising, since the primary reason for carrying out a correlated data analysis is to account for the possible effects that different correlation structures can have on statistical inferences.

Since the estimated regression coefficients are identical for each choice of correlation structure, let's focus on the interpretation of the estimated model for a given correlation structure—say, compound symmetric (CS). Since the referent group for the Week variable is week 5, the intercept β_0 in the model represents the true FEV1 mean for week 5, whereas β_j is the difference in true FEV1 mean levels between week j and week 5 ($j = 1, 2, 3, 4$).

Thus, from the output (Table 25.7), the estimated intercept is $\hat{\beta}_0 = \overline{\text{FEV1}}_5 = 6.9985$, and the estimated differences between mean FEV1 levels for week j and week 5, $j = 1, \dots, 4$, are, respectively,

$$\hat{\beta}_1 = \overline{\text{FEV1}}_1 - \overline{\text{FEV1}}_5 = 2.8153, \quad \hat{\beta}_2 = \overline{\text{FEV1}}_2 - \overline{\text{FEV1}}_5 = -0.1502$$

$$\hat{\beta}_3 = \overline{\text{FEV1}}_3 - \overline{\text{FEV1}}_5 = 0.00325, \quad \hat{\beta}_4 = \overline{\text{FEV1}}_4 - \overline{\text{FEV1}}_5 = -0.0470$$

TABLE 25.7 Edited output based on FEV data for different marginal models ($G = 0$) without empirical standard error option, using REML in the SAS MIXED procedure

Column 1			Column 2*			Column 3		
Model-based IND								
Effect	Week	Estimate	Standard Error	Type 3 Tests of Fixed Effects				
Intercept		6.9985	0.2590	Num DF	Den DF	F Value	Pr > F	
week	1	2.8153	0.3663	Effect week	4	195	24.50	<.0001
week	2	-0.1502	0.3663					
week	3	0.003250	0.3663					
week	4	-0.04700	0.3663	contrast	1	195	97.79	<.0001
Model-based CS								
Effect	Week	Estimate	Standard Error	Type 3 Tests of Fixed Effects				
Intercept		6.9985	0.2590	Num DF	Den DF	F Value	Pr > F	
week	1	2.8153	0.2439	Effect week	4	195	55.26	<.0001
week	2	-0.1502	0.2439					
week	3	0.003250	0.2439					
week	4	-0.04700	0.2439	contrast	1	195	220.51	<.0001
Model-based AR1								
Effect	Week	Standard Estimate	Error	Type 3 Tests of Fixed Effects				
Intercept		6.9985	0.2526	Num DF	Den DF	F Value	Pr > F	
week	1	2.8153	0.3359	Effect week	4	195	44.51	<.0001
week	2	-0.1502	0.3198					
week	3	0.003250	0.2902					
week	4	-0.04700	0.2306	contrast	1	195	133.91	<.0001

*The F statistic shown for Type 3 Tests of Fixed Effects tests $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$. The F statistic shown for the "contrast" tests $H_0: \mu_1 - \mu_2 + \mu_3 + \mu_4 + \mu_5 = 0$. Although the estimation procedure used to fit the models is REML (a likelihood procedure), the resulting test statistics are (either exact or approximate) F statistics given the normality assumptions. Similarly, such t statistics would be obtained if MLE is used instead of REML, although the numerical values of the F statistics may differ slightly from REML values. Correspondingly, test statistics for individual regression coefficients can be described as (exact or approximate) t statistics. Even if random effects are added to the model, such F and t statistics will be obtained if both the random components and the error terms are assumed to be normally distributed. For large samples, these F and t statistics essentially become χ^2 and Z statistics, respectively.

TABLE 25.7 Edited output based on FEV data for different marginal models ($G = 0$) without empirical standard error option, using REML in the SAS MIXED procedure (continued)

Column 1			Column 2			Column 3		
Model-based UN								
Effect	Week	Standard Estimate	Error	Type 3 Tests of Fixed Effects				
			Effect	Num DF	Den DF	F Value	Pr > F	
Intercept		6.9985	0.2449	week	4	195	37.64	<.0001
week	1	2.8153	0.2946					
week	2	-0.1502	0.2078					
week	3	0.003250	0.2102					
week	4	-0.04700	0.2419	contrast	1	195	143.78	<.0001

*The F statistic shown for Type 3 Tests of Fixed Effects tests $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$. The F statistic shown for the "contrast" tests $H_0: \mu_1 - \mu_2 + \mu_3 + \mu_4 + \mu_5 = 0$. Although the estimation procedure used to fit the models is REML (a likelihood procedure), the resulting test statistics are (either exact or approximate) F statistics given the normality assumptions. Similarly, such t statistics would be obtained if ML is used instead of REML, although the numerical values of the F statistics may differ slightly from REML values. Correspondingly, test statistics for individual regression coefficients can be described as (exact or approximate) t statistics. Even if random effects are added to the model, such F and t statistics will be obtained if both the random components and the error terms are assumed to be normally distributed. For large samples, these F and t statistics essentially become χ^2 and Z statistics, respectively.

Also, regardless of the correlation structure (i.e., \mathbf{R} matrix) shown in Table 25.7, the predicted FEV1 mean values for each of the five weeks are simply the average FEV1 levels for these five weeks (as previously shown in Table 25.4).⁹

$$\overline{\text{FEV1}}_1 = 9.81375, \overline{\text{FEV1}}_2 = 6.84825, \overline{\text{FEV1}}_3 = 7.00175,$$

$$\overline{\text{FEV1}}_4 = 6.96150, \overline{\text{FEV1}}_5 = 6.99850$$

These results indicate that the average FEV1 level prior to the pollution alert (namely, $\overline{\text{FEV1}}_1$) was almost 3 units higher than during the week of the pollution alert (namely, $\overline{\text{FEV1}}_2$); further, for weeks 3, 4, and 5, mean FEV1 levels remained depressed to about the same level as for week 2.

Column 2 in Table 25.7 gives hypothesis-testing results about the overall effect of “week” as well as about the contrast that compares the FEV1 mean for week 1 with the average of the FEV means for the other four weeks. The null hypothesis for the first of these tests can be stated as

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \text{ or, equivalently, } H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

where μ_j denotes the true (i.e., population) mean FEV1 value at week j and β_j denotes the coefficient of the j th dummy variable in the (marginal) model (25.8). The corresponding F statistics for different choices of \mathbf{R} are all different¹⁰ but are nevertheless all highly significant. Thus, we can conclude that there are significant differences among the estimated mean FEV1 levels over the five weeks.

The estimated average FEV1 scores for each of the five weeks suggest that the pollution episode may have caused a significant overall *reduction* in lung capacity (i.e., FEV1 level) *that remained essentially the same* over the three weeks following the episode. A significance test about this assertion can be carried out by comparing the observed FEV1 mean value for the first week with the average of the observed FEV1 mean values for the other four weeks. We previously showed in Section 25.2 that this “contrast” is estimated from the data to be

$$\begin{aligned} \overline{\text{FEV1}}_1 - \frac{\overline{\text{FEV1}}_2 + \overline{\text{FEV1}}_3 + \overline{\text{FEV1}}_4 + \overline{\text{FEV1}}_5}{4} \\ = 9.81375 - \frac{6.84825 + 7.00175 + 6.95150 + 6.99850}{4} = 2.8637 \end{aligned}$$

⁹ The estimated FEV1 mean level for each week is identical to the corresponding predicted mean FEV1 level for that week because the only predictors in the marginal model are the four dummy variables for the five weeks, with no other covariates being included in the model.

¹⁰ The denominator degrees of freedom (Den DF) for each test in column 2 is 195, which is determined from the typical formula $(n - p^*)$, where $n (= 200)$ is the total number of observations in the study and p^* is the number of fixed effect parameters in the model (in our example, $p = 4$ for the four dummy variables X_1, X_2, X_3 , and X_4 , so that $p^* = p + 1 = 5$). This choice for Den DF (also called DDFM) is called the *residual option*. MIXED also allows the user to make other choices for DDFM. For example, if the model contains random effects, it might be argued that the DDFM should involve subtracting from $(n - p^*)$ the number (q^*) of random effects in the model times one less than the number of subjects in the study; that is, $\text{DDFM} = (n - p^*) - q^*(K - 1)$. Thus, for the FEV1 data (where $K = 40$), the DDFM for a model containing a single random effect ($q^* = 1$) for the intercept would be $\text{DDFM} = (200 - 5) - (1 \times 39) = 156$; for two random effects ($q^* = 2$), $\text{DDFM} = (200 - 5) - (2 \times 39) = 117$. See Section 26.4.7 for further discussion on choices for DDFM.

which is meaningfully different from zero. The null hypothesis for testing for the statistical significance of this contrast is given by either of the following equivalent expressions,¹¹ with results shown below for a model using a compound symmetric (**CS**) correlation structure:

$$H_0: \mu_1 - \frac{\mu_2 + \mu_3 + \mu_4 + \mu_5}{4} = 0 \text{ or, equivalently, } H_0: \beta_1 - \frac{\beta_2 + \beta_3 + \beta_4}{4} = 0$$

CONTRAST OUTPUT (USING CS STRUCTURE)				
Label	Num DF	Den DF	F Value	Pr > F
Week 1 vs. 2-5	1	195	220.51	<.0001

From the above output, we conclude that the estimated contrast is highly significantly different from zero, so that the estimated mean FEV1 level for week 1 is statistically different from the average of the other four estimated means. Column 2 of Table 25.7 provides corresponding *F* statistics for this contrast for four different choices of the **R** matrix. All four *F* statistics are numerically different from one another because different **R** matrices are used. Nevertheless, for each of these choices for **R**, the contrast comparing the estimated mean FEV1 value at week 1 and the average of the estimated mean FEV1 values over the other four weeks is highly significant.

25.4.2 Results Using Empirical Standard Errors: FEV1 Data

From the model-based output given in Table 25.7, we concluded that, for any of four stated choices for the **R** matrix, there was a statistically significant difference among the estimated mean FEV1 values for the five weeks; moreover, an estimated contrast of interest was also statistically significantly different from zero. These analyses do not clearly identify which of the four correlation structures considered is most appropriate for these data, nor do they consider the wide variety of other correlation structures that could be used.

As mentioned previously, the choice of appropriate correlation/covariance structure requires consideration of several approaches from which the investigator(s) must make a “reasoned” decision. Such approaches include possibly identifying a particular correlation structure that seems the most biologically or clinically reasonable, deciding which model and associated correlation structure have the best “goodness of fit” to the data, and assessing the extent to which numerical results (and attendant statistical inferences and study conclusions) differ among different correlation structures. The fact that we reached the same general conclusions about the overall effects of the pollution alert on FEV1 mean levels for the four **R** matrices that we considered is possibly reason enough not to investigate other

¹¹ The two null hypotheses are equivalent, since it follows that

$$\begin{aligned} \beta_1 - \frac{\beta_2 + \beta_3 + \beta_4}{4} &= (\mu_1 - \mu_5) - \frac{(\mu_2 - \mu_5) + (\mu_3 - \mu_5) + (\mu_4 - \mu_5)}{4} \\ &= \mu_1 - \frac{\mu_2 + \mu_3 + \mu_4 + \mu_5}{4} \end{aligned}$$

correlation structures. We note, however, that a logical choice of correlation structure for these data is autoregressive (**ARI**), the logic being that the farther apart in time observations on the same subject are, the smaller the pairwise correlations should be.

Nevertheless, the availability of empirical standard errors suggests that it would be useful to assess the effects on numerical results of possibly making an incorrect choice of correlation/covariance structure. For this purpose, Table 25.8 has been produced, which provides corresponding output for the same four marginal models ($\mathbf{G} = \mathbf{0}$) considered in Table 25.7, but now using empirical standard error estimates. From column 1 of Table 25.8, we can see that corresponding estimated regression coefficients are identical for the four choices of \mathbf{R} and also that corresponding empirical standard errors are identical (although not shown, corresponding empirical estimated regression coefficient covariance estimates are also identical for all four choices of \mathbf{R}). Thus, for these data, once we account for a possibly incorrect choice for the correlation/covariance structure, any statistical inferences about model parameters are identical, regardless of the choice of working correlation structure. Such a finding is not generally true when using empirical standard errors, as we will illustrate in the next section for unbalanced data.

From column 2 of Table 25.8, we see that all F statistics for testing the null hypothesis of equal mean FEV1 scores over the five weeks—that is,

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \text{ (or, equivalently, } H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0\text{)}$$

as well as the F statistics for testing

$$H_0: \mu_1 - \frac{\mu_2 + \mu_3 + \mu_4 + \mu_5}{4} = 0 \text{ (or, equivalently, } H_0: \beta_1 - \frac{\beta_2 + \beta_3 + \beta_4}{4} = 0\text{)}$$

are identical for all the different choices of \mathbf{R} . These results are obtained because the empirical standard errors (as well as empirical covariance estimates) for the estimated regression coefficients do not vary with the choice for \mathbf{R} .

The column 3 numerical results in Table 25.8 are exactly the same as those in Table 25.7; this is because these computations of estimated working correlation matrices are done independently of empirical standard error calculations.

25.4.3 Results for Unbalanced Data: FEV1 Study

So far, we have provided numerical illustrations based on fitting a particular linear mixed model to a balanced data set; that is, each subject has the same number of repeated observations that are measured at the same times. In this section, we show that the same linear mixed model can also be applied to an unbalanced data set. To illustrate this, we arbitrarily removed 24 observations from 15 of the 40 subjects in the FEV1 study. Table 25.9 gives the data for the 15 subjects that now each have less than 5 observations; the remaining 25 subjects still have FEV1 data for all five weeks. Notice, for example, that subjects 1, 2, and 24 are each missing one FEV1 measurement at week 5, subject 3 is missing FEV1 data at weeks 3 and 4, and subject 4 is missing FEV1 measurements at weeks 4 and 5. The use of a “.” denotes a week for a given subject that is missing a value of FEV1.

TABLE 25.8 Edited output based on FEV data for different marginal models ($G = 0$) with empirical standard error option, using REML in the SAS MIXED procedure

Column 1			Column 2			Column 3		
<i>Empirical CS</i>			Type 3 Tests of Fixed Effects			$\mathbf{V} \equiv \mathbf{R} = \sigma_e^2 \mathbf{I}$, where $\hat{\mathbf{C}}^{(EXCH)}$ is		
Effect	Week	Estimate	Standard Error	Effect	Num DF	F Value	Pr > F	
Intercept		6.9985	0.2418	Effect week	4	195	38.61	<.0001
week	1	2.8152	0.2514					
week	2	-0.1502	0.2052					
week	3	0.003250	0.2076					
week	4	-0.04700	0.2389	contrast	1	195	147.46	<.0001
<i>Empirical AR1</i>			Type 3 Tests of Fixed Effects			$\mathbf{V} \equiv \mathbf{R} = \sigma_e^2 \mathbf{C}^{(EXCH)}$ where $\hat{\mathbf{C}}^{(AR1)}$ is		
Effect	Week	Standard Estimate	Error	Effect week	Num DF	F Value	Pr > F	
Intercept		6.9985	0.2418	Effect week	4	195	38.61	<.0001
week	1	2.8152	0.2514					
week	2	-0.1503	0.2052					
week	3	0.003250	0.2076					
week	4	-0.04700	0.2389	contrast	1	195	147.46	<.0001
<i>Empirical UN</i>			Type 3 Tests of Fixed Effects			$\mathbf{V} \equiv \mathbf{R} = \mathbf{UN}$ where $\hat{\mathbf{C}}^{(UN)}$ is		
Effect	Week	Standard Estimate	Error	Effect week	Num DF	F Value	Pr > F	
Intercept		6.9985	0.2418	Effect week	4	195	38.61	<.0001
week	1	2.8153	0.2514					
week	2	-0.1502	0.2052					
week	3	0.003250	0.2076					
week	4	-0.04700	0.2389	contrast	1	195	147.46	<.0001

TABLE 25.9 Unbalanced data for 15 subjects from FEV1 data set

Subj.	Week	FEV1	Subj.	Week	FEV1	Subj.	Week	FEV1
1	1	9.43	7	1	10.30	24	1	10.63
1	2	5.71	7	2	.	24	2	5.15
1	3	5.86	7	3	6.99	24	3	6.07
1	4	7.70	7	4	.	24	4	7.44
1	5	.	7	5	5.67	24	5	.
2	1	11.15	8	1	8.67	26	1	.
2	2	9.48	8	2	4.10	26	2	7.51
2	3	10.11	8	3	.	26	3	7.56
2	4	8.89	8	4	.	26	4	4.88
2	5	.	8	5	5.46	26	5	7.54
3	1	8.40	10	1	9.05	28	1	.
3	2	4.42	10	2	.	28	2	6.87
3	3	.	10	3	5.33	28	3	5.79
3	4	.	10	4	5.94	28	4	8.30
3	5	4.34	10	5	.	28	5	8.19
4	1	9.20	11	1	12.44	30	1	10.24
4	2	4.51	11	2	8.68	30	2	8.48
4	3	6.20	11	3	.	30	3	.
4	4	.	11	4	6.40	30	4	.
4	5	.	11	5	7.61	30	5	.
6	1	9.12	12	1	7.88	32	1	11.50
6	2	7.71	12	2	4.88	32	2	.
6	3	6.75	12	3	6.43	32	3	6.69
6	4	.	12	4	.	32	4	.
6	5	8.20	12	5	.	32	5	.

© Cengage Learning

Table 25.10 presents results obtained from fitting marginal linear mixed models to the unbalanced FEV1 data set using the same four \mathbf{R} matrices previously considered for the balanced data set.

From column 1 of Table 25.10, it can be seen that the sets of estimated regression coefficients differ for each choice of \mathbf{R} ; in contrast, for the balanced data set, the sets of estimated

TABLE 25.10 Edited output based on unbalanced FEV data for different marginal models ($\mathbf{G} = \mathbf{0}$), using REML in the SAS MIXED procedure

Column 1				Column 2				Column 3*			
IND				Type 3 Tests of Fixed Effects				Type 3 Tests of Fixed Effects			
Effect	Week	Estimate	Model-based SE	Num DF	Den DF	Model-based F Value	Empirical F Value	Num DF	Den DF	Model-based F Value	Empirical F Value
Intercept		7.0347	0.2856	0.2660	4	170	22.08				
week	1	2.7748	0.3876	0.2838							
week	2	-0.1633	0.3899	0.2630							
week	3	0.00559	0.3925	0.2656							
week	4	0.2075	0.4038	0.2874	contrast	1	195	86.91			
CS				Type 3 Tests of Fixed Effects				Type 3 Tests of Fixed Effects			
Effect	Week	Estimate	Model-based SE	Num DF	Den DF	Model-based F Value	Empirical F Value	Num DF	Den DF	Model-based F Value	Empirical F Value
Intercept		7.0104	0.2727	0.2533	4	170	47.42				
week	1	2.8027	0.2704	0.2729							
week	2	-0.1572	0.2706	0.2453							
week	3	-0.01109	0.2739	0.2421							
week	4	0.1070	0.2809	0.2710	contrast	1	195	187.33			
ARI				Type 3 Tests of Fixed Effects				Type 3 Tests of Fixed Effects			
Effect	Week	Estimate	Model-based SE	Num DF	Den DF	Model-based F Value	Empirical F Value	Num DF	Den DF	Model-based F Value	Empirical F Value
Intercept		7.0552	0.2715	0.2521	4	170	41.57				
week	1	2.7642	0.3514	0.2718							
week	2	-0.1900	0.3370	0.2580							
week	3	-0.08110	0.3119	0.2424							
week	4	0.05857	0.2609	0.2749	contrast	1	195	124.38			
UN				Type 3 Tests of Fixed Effects				Type 3 Tests of Fixed Effects			
Effect	Week	Estimate	Model-based SE	Num DF	Den DF	Model-based F Value	Empirical F Value	Num DF	Den DF	Model-based F Value	Empirical F Value
Intercept		7.0145	0.2542	0.2506	4	171	36.89				
week	1	2.8020	0.2716	0.2677							
week	2	-0.1762	0.2488	0.2451							
week	3	-0.04812	0.2427	0.2390							
week	4	0.09776	0.2737	0.2693	contrast	1	195	136.94			

*The estimated working covariance matrices given in column 3 are all (4×4) matrices because the first subject in the unbalanced data set has only four repeated observations.

coefficients were identical (see Tables 25.7 and 25.8). Also, as with the balanced data set, the sets of model-based standard errors vary, as expected, with varying choices for \mathbf{R} . Note, however, that the sets of empirical standard errors also vary as \mathbf{R} varies; in contrast, empirical standard errors did not change as a function of \mathbf{R} when the data set was balanced (see Table 25.8).

From column 2 of Table 25.10, we can also see that the model-based F statistics for the overall comparison among the five weeks, and for the contrast of interest, differ, as expected, as \mathbf{R} differs. However, in contrast to the balanced data set results, the empirical-based F statistics also differ as \mathbf{R} differs. Although the overall conclusion for the unbalanced data set is the same (i.e., a significant effect of “week” and a significant “contrast”), the fact that the empirical standard errors and corresponding F statistics differ with \mathbf{R} indicates that, in general, numerical results using the empirical standard error option can vary when analyzing data sets of small to intermediate size that are unbalanced (e.g., there are missing data).

Finally, since Table 25.10 is based on the use of the unbalanced data set, the estimated working correlation matrices in column 3 of Table 25.10 differ (as expected) from corresponding numerical results in Tables 25.7 and 25.8.

25.4.4 Summary of Analyses for the FEV1 Data

From the analyses of the FEV1 data set that we have carried out in this section, we summarize the results as follows:

1. With regard to the primary question of whether or not lung capacity as measured by FEV1 was depressed during the pollution episode:
 - a. The observed FEV1 mean level prior to the pollution alert (week 1) was almost 3 units higher than during week 2 of the pollution alert; further, the observed mean FEV1 levels for weeks 3, 4, and 5 remained depressed at about the same level as for week 2.
 - b. There is a significant difference among the estimated FEV1 mean levels over the five weeks.
 - c. There is also a significant difference between the FEV1 sample mean for week 1 and the average of the FEV1 sample means over the other four weeks.
 - d. The statistical conclusions in (b) and (c) above are essentially the same, regardless of the choice of working correlation structure.
2. The following results were obtained based on using marginal models (varying \mathbf{R} , $\mathbf{G} = \mathbf{0}$) to analyze the complete (*balanced*) data set:
 - a. The estimated regression coefficients for the effect of “week” were identical, regardless of the choice of \mathbf{R} and, as expected, regardless of whether model-based or empirical standard errors were used.
 - b. The set of empirical standard errors was identical, regardless of the choice of \mathbf{R} . For a given \mathbf{R} , however, the estimated empirical standard errors were

numerically different from corresponding model-based estimated standard errors. (The finding of identical estimated empirical standard errors for different choices of \mathbf{R} for these balanced data, however, is not true in general for unbalanced data.)

- c. Hypothesis testing (using F statistics) resulted in finding a significant difference among estimated mean FEV1 levels over the five weeks, regardless of the choice of \mathbf{R} and regardless of whether model-based or empirical standard error options were used. Nevertheless, using model-based standard errors, F statistic values differed numerically as \mathbf{R} differed. In contrast, using empirical standard errors, F statistic values were identical, regardless of the choice for \mathbf{R} .
 - d. The contrast that compared the estimated mean FEV1 level for week 1 with the estimated average mean FEV1 level over the other four weeks was found to be statistically (as well as meaningfully) different from zero when either model-based or empirical standard errors options were used.
 - e. The estimated \mathbf{R} matrix differed, as expected, as \mathbf{R} differed. Nevertheless, for a given choice of \mathbf{R} , the estimated \mathbf{R} matrix was identical when either model-based or empirical standard error options were used; this is because the \mathbf{R} matrix, being a “working-covariance structure,” is estimated independently of empirical standard error calculations.
 - f. A reasonable choice of correlation structure for these data is autoregressive (AR1); because of the longitudinal nature of the study, the farther apart in time subject-specific observations are, the smaller the pairwise correlations should be. Nevertheless, the global statistical results essentially do not differ for the four \mathbf{R} matrices considered, which suggests that the choice of correlation structure does not affect the general conclusions to be drawn from the study data.
3. The following results were obtained based on using marginal models (varying $\mathbf{R}, \mathbf{G} = \mathbf{0}$) to analyze an unbalanced subset of the original data:
- a. The set of estimated regression coefficients for the effect of “week” differed (slightly) as the choice of \mathbf{R} differed when either model-based or empirical standard error options were used. This illustrates the general principle that the estimated regression coefficients may differ in value for different choices of the working covariance structure, especially for unbalanced data sets.
 - b. The set of estimated empirical standard errors differed (slightly) with the choice of \mathbf{R} . This illustrates the general principle that the empirical standard errors may differ in value for different choices of the working covariance structure, especially for small to intermediate size studies with unbalanced data.
 - c. The general linear mixed model defined by (25.4) can be used to analyze unbalanced data in which different subjects can have different numbers of responses measured at different times.

25.5 Summary—Analysis of Correlated Data: Part 1

In this chapter, we have extended the multiple linear regression model to consider outcomes (i.e., responses) that are observed two or more times on each unit of analysis (e.g., a subject or, more generally, a *cluster*). The responses on each unit will generally be correlated, thereby requiring an analysis that accounts for such correlation. We illustrated this type of data using three examples:

- a. A Study of the Effect of an Air Pollution Episode on Pulmonary Function
- b. A Study of the Posture of Computer Operators
- c. A Study of Treatments for Heartburn

We provided general notation for the response vector \mathbf{Y}_i that characterizes the collection of n_i responses within the i th cluster, and we described the general data layout for such clustered data.

We then described in detail the general linear mixed model approach assuming normally distributed outcome variables and, correspondingly, normally distributed random effects and error terms. We described this model in both general scalar form and general matrix form as follows:

$$\text{Subject-specific scalar form: } Y_{ij} = (\beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_s X_{ij s}) \\ + (b_{i0} + b_{i1} Z_{ij1} + b_{i2} Z_{ij2} + \cdots + b_{iq} Z_{ijq}) + E_{ij}$$

$$\text{Subject-specific matrix form: } \mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \mathbf{E}_i$$

The X_{ij} 's (or \mathbf{X}_i) in the above models denote covariates for fixed effects, and the Z_{ij} 's (or \mathbf{Z}_i) denote covariates for random effects. Correspondingly, the β 's (or $\boldsymbol{\beta}$) denote fixed effects, and the b 's (or \mathbf{b}_i) denote random effects.

We distinguished among different *working correlation structures* that could be considered for a given model (including independent, exchangeable, autoregressive, and unstructured matrices), and we described the relationship between a covariance structure and its corresponding correlation structure. We also showed that the covariance structure (or matrix) \mathbf{V}_i for \mathbf{Y}_i corresponding to the general linear mixed model had the general matrix form $\mathbf{V}_i = \mathbf{Z}_i \mathbf{G} \mathbf{Z}'_i + \mathbf{R}_i$, where \mathbf{G} denotes the covariance structure (or matrix) for the random component (\mathbf{b}_i) and \mathbf{R}_i denotes the covariance structure (or matrix) for the error component (\mathbf{E}_i). And we described the use of a robust or empirical standard error option to correct for the possibility of misspecifying the correlation structure in a given analysis.

Finally, we described the analysis of the Air Pollution Study using SAS's MIXED procedure for marginal models ($\mathbf{G} = \mathbf{0}$) with different choices for the working correlation structure; also, we considered the entire (balanced) data set as well as an unbalanced data set in which not all subjects had the same number of FEV1 observations.

In the next chapter, which is the second part of our discussion on the analysis of correlated data, we focus on the use of random effects when fitting a linear mixed model.

Problems

- 1.** Consider the following four correlation matrices that apply to clusters having five responses per cluster:

$$\mathbf{A} = \begin{bmatrix} 1 & 0.61 & 0.45 & 0.39 & 0.25 \\ 0.61 & 1 & 0.43 & 0.51 & 0.35 \\ 0.45 & 0.43 & 1 & 0.29 & 0.22 \\ 0.39 & 0.51 & 0.29 & 1 & 0.53 \\ 0.25 & 0.35 & 0.22 & 0.53 & 1 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 1 & 0.50 & 0.25 & 0.125 & 0.0625 \\ 0.50 & 1 & 0.50 & 0.25 & 0.125 \\ 0.25 & 0.50 & 1 & 0.50 & 0.25 \\ 0.125 & 0.25 & 0.50 & 1 & 0.50 \\ 0.0625 & 0.125 & 0.25 & 0.50 & 1 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 1 & 0.46 & 0.46 & 0.46 & 0.46 \\ 0.46 & 1 & 0.46 & 0.46 & 0.46 \\ 0.46 & 0.46 & 1 & 0.46 & 0.46 \\ 0.46 & 0.46 & 0.46 & 1 & 0.46 \\ 0.46 & 0.46 & 0.46 & 0.46 & 1 \end{bmatrix}$$

- a. Matrix A is an example of what kind of correlation structure?
 - b. Matrix B is an example of what kind of correlation structure?
 - c. Matrix C is an example of what kind of correlation structure?
 - d. Matrix D is an example of what kind of correlation structure?
- 2.** Consider the following four covariance matrices that apply to clusters having four responses per cluster:

$$\mathbf{P} = \begin{bmatrix} 2.50 & 1.25 & 0.625 & 0.3125 \\ 1.25 & 2.50 & 1.25 & 0.625 \\ 0.625 & 1.25 & 2.50 & 1.25 \\ 0.3125 & 0.625 & 1.25 & 2.50 \end{bmatrix} \quad \mathbf{Q} = \begin{bmatrix} 4.00 & 1.952 & 1.721 & 1.593 \\ 1.952 & 4.50 & 1.826 & 1.690 \\ 1.721 & 1.826 & 3.50 & 1.491 \\ 1.593 & 1.690 & 1.491 & 3.00 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 4.00 & 1.84 & 1.84 & 1.84 \\ 1.84 & 4.00 & 1.84 & 1.84 \\ 1.84 & 1.84 & 4.00 & 1.84 \\ 1.84 & 1.84 & 1.84 & 4.00 \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} 4.00 & 1.732 & 1.061 & 0.234 \\ 1.732 & 3.00 & 1.837 & 0.810 \\ 1.061 & 1.837 & 4.50 & 1.984 \\ 0.234 & 0.810 & 1.984 & 3.50 \end{bmatrix}$$

- a. Which of the above matrices are heterogeneous (as compared to homogeneous) covariance matrices?
 - b. What is the correlation structure defined by each of the above matrices?
- 3.** Consider the data layout for a correlated data analysis provided in Table 25.2. Using this data layout as a guiding framework, describe a study involving longitudinal data (and/or other types of correlated data) that relates either to a study in which you are already involved or to a study that you might wish to undertake in an area of your own interest. Assume that your outcome (i.e., response) variable is continuous.

In your description, make sure to

- a. state whether the study design is longitudinal or of some other type involving correlated data;
 - b. state the primary research objective of the study;
 - c. give a brief outline of the study design;
 - d. describe the variables being measured or observed and summarize such information using the same format as the data layout in Table 25.2;
 - e. specify at least one “marginal” regression model that you would like to fit to the data under consideration; also, describe any random effects model that you might wish to consider; and
 - f. describe what you think is the appropriate correlation structure for each cluster in your analysis (e.g., exchangeable, autoregressive, other).
4. Consider again the study described in the main body of this chapter concerning the effect of an air pollution episode on pulmonary function measurements (FEV1) taken on each of $K = 40$ school children over five weeks. The subject-specific scalar and matrix versions of the model defined by expressions (25.7) and (25.8) use week 5 as the reference level. In this exercise, you are to consider how the model and analysis would be modified if week 1 is used as the reference level.
- a. State the subject-specific scalar form of the model (i.e., analogous to expression (25.7)) in which week 1 is used as the reference level. (*Note:* Denote the coefficients in this model by β'_g , $g = 0, 1, 2, 3, 4$.)
 - b. State the subject-specific matrix form of the model (i.e., analogous to expression (25.8)) in which week 1 is used as the reference level, making sure to specify the components and corresponding dimensions of the matrices involved.
 - c. Using the subject-specific scalar model defined in part (a), express each of the estimated regression coefficients ($\hat{\beta}'_g$, $g = 0, 1, 2, 3, 4$) in terms of sample mean FEV1 levels

$$\overline{\text{FEV1}}_j \quad j = 1, 2, 3, 4, 5$$

- d. Express the null hypothesis

$$H_0: \mu_1 - \frac{\mu_2 + \mu_3 + \mu_4 + \mu_5}{4} = 0$$

in terms of the regression coefficients β'_g ($g = 0, 1, 2, 3, 4$) for the model specified in part (a).

- e. Will the computed test statistic resulting from testing H_0 using the model in part (a) be identical to the computed test statistic resulting from testing H_0 using model (25.2)—or, equivalently, model (25.7)—in which the reference level is week 5? Explain briefly.
- f. What are the predicted FEV1 mean values for weeks 1 through 5, respectively?
- g. What is the estimate of the contrast that compares week 1 with the average of the other four weeks? Is this estimated contrast statistically significantly different from zero?

Edited SAS Output (PROCMIXED) for Problem 4

ESTIMATED R CORRELATION MATRIX FOR SUBJ 1					
Row	Col1	Col2	Col3	Col4	Col5
1	1.0000	0.5565	0.5565	0.5565	0.5565
2	0.5565	1.0000	0.5565	0.5565	0.5565
3	0.5565	0.5565	1.0000	0.5565	0.5565
4	0.5565	0.5565	0.5565	1.0000	0.5565
5	0.5565	0.5565	0.5565	0.5565	1.0000

COVARIANCE PARAMETER ESTIMATES		
Cov Parm	Subject	Estimate
CS	subj	1.4936
Residual		1.1901

SOLUTION FOR FIXED EFFECTS					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	9.8137	0.1873	195	52.39	<.0001
d1	-2.9655	0.2610	195	-11.36	<.0001
d2	-2.8120	0.2782	195	-10.11	<.0001
d3	-2.8623	0.2920	195	-9.80	<.0001
d4	-2.8153	0.2514	195	-11.20	<.0001

TYPE 3 TESTS OF FIXED EFFECTS				
Effect	Num DF	Den DF	F Value	Pr > F
d1	1	195	129.14	<.0001
d2	1	195	102.14	<.0001
d3	1	195	96.08	<.0001
d4	1	195	125.36	<.0001

CONTRASTS				
Label	Num DF	Den DF	F Value	Pr > F
week	1	195	147.46	<.0001

- h.** The output shown above does not provide the results for testing whether or not there is a significant difference among the estimated mean FEV1 levels for all five weeks.
- i.** Specify a “contrast” statement (SAS code) that would perform such a test.

- ii. Will the results obtained from the contrast statement specified in part (h.i) be the same as the results previously obtained in Table 25.7, for which the referent group is week 5? Explain briefly.
 - i. i. What correlation structure has been assumed to generate the above output?
 - ii. What is the estimated correlation between any two responses on the same subject?
 - iii. How can the correlation in part (i.ii) be computed using the output information provided for “Covariance Parameter Estimates”?
 - iv. Does the output shown above provide test results that allow for the possibility that the assumed correlation structure may be incorrectly specified? Explain briefly.
 - v. Based on the output summarized in Table 25.7, would you expect the numerical test results obtained for different working correlation structures to be identical when the referent group is week 1? Explain briefly.
5. This problem considers the (fictitious) study designed to compare two treatments for the relief of heartburn, described by the data set of Table 25.6. One of two treatments ($A = \text{active} = 1$, $P = \text{placebo} = 0$) has been randomly allocated to two groups of 15 subjects each, and each subject receives the same treatment for both meals. The outcome variable is a continuous measure of physical discomfort measured by a questionnaire administered to each subject two hours after receiving the treatment. Edited computer output from SAS's MIXED procedure is provided below.
- Why is it appropriate to do a correlated data analysis here?
 - Why doesn't it matter what correlation structure is chosen (other than “independence”) for carrying out a correlated data analysis of these data?
 - Specify the subject-specific scalar model form used in this analysis.
 - Specify the subject-specific matrix model form used in this analysis.
 - What do you conclude about whether or not there is a significant difference between the effects of the two treatments on the amount of discomfort experienced by study subjects? Explain briefly.

Edited SAS Output (PROC MIXED) for Problem 5

SOLUTION FOR FIXED EFFECTS EMPIRICAL REML					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	62.17		28		
trt	-11.00	4.4168	28	2.4905	0.0190

TYPE 3 TESTS OF FIXED EFFECTS				
Effect	Num DF	Den DF	F Value	Pr > F
trt	1	28	6.2026	0.0190

6. Consider a different (fictitious) study to compare two treatments for the relief of heartburn. This new study involves a two-period “cross-over” design in which each subject is given two symptom-provoking meals and receives (with random assignment) an active treatment (A) after one meal and a placebo (P) after the other meal. In other words, each subject receives both treatments in this study; in contrast, for the previously described study (see Problem 5), each subject received either the active treatment or the placebo after both meals. The data set is given below.

A:P			P:A		
Subject Number	A	P	Subject Number	P	A
1	65	75	16	85	55
2	60	60	17	60	70
3	70	75	18	80	70
4	35	50	19	55	30
5	50	45	20	50	50
6	40	65	21	70	40
7	50	80	22	70	65
8	55	50	23	65	35
9	20	40	24	60	50
10	30	35	25	90	70
11	65	50	26	85	80
12	45	55	27	65	70
13	30	50	28	60	45
14	55	70	29	75	65
15	25	40	30	55	45

Mean: 46.33 56.00 Mean: 68.33 56.00

Overall Mean A: 51.17 Overall Mean P: 62.17

Edited computer output from SAS's MIXED procedure is provided below.

- a. Specify the subject-specific scalar model form used for this analysis.
- b. What do you conclude about whether or not there is a significant difference between the effects of the two treatments on the amount of discomfort experienced by study subjects? Explain briefly.

Edited SAS Output (PROC MIXED) for Problem 6

SOLUTION FOR FIXED EFFECTS REML					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	62.1667	2.7822	29	22.34	<.0001
trt	-11.0000	2.2540	29	-4.88	<.0001

TYPE 3 TESTS OF FIXED EFFECTS				
Effect	Num DF	Den DF	F Value	Pr > F
trt	1	29	23.82	<.0001

- c. Why are the Den DF equal to 29 in this analysis, whereas the Den DF were equal to 28 in the analysis used for the data of Problem 5? Suppose that the investigator wished to consider the possibility that the sequence used—that is, (A:P) versus (P:A)—had an effect on the response in addition to a possible treatment effect. To carry out such an analysis, he/she modified the model by adding a dichotomous variable, denoted as “seq,” to the model. Edited output for this latter model is provided below:

Edited SAS Output (PROC MIXED) for Problem 6 (continued)

SOLUTION FOR FIXED EFFECTS EMPIRICAL REML					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	67.1667	3.5373	28	19.13	<.0001
trt	-11.0000	2.2540	29	-4.88	<.0001
seq	-11.0000	4.7417	28	-2.32	0.0279

TYPE 3 TESTS OF FIXED EFFECTS				
Effect	Num DF	Den DF	F Value	Pr > F
trt	1	29	23.82	<.0001
seq	1	28	5.38	0.0279

- d. Specify the subject-specific scalar model form used that includes both “trt” and “seq.”
e. What do you conclude about whether or not there is a significant difference between the effects of the two treatments on the amount of discomfort experienced by study subjects, controlling for the variable “seq”? Explain briefly.
f. How would you evaluate whether or not it was necessary to control for the “seq” variable in order to determine whether or not there was a significant difference between the two treatments?

References

- Diggle, P. J.; Heagerty, P.; Liang, K. Y.; and Zeger, S. L. 2002. *Analysis of Longitudinal Data*, Second Edition. Oxford: Oxford University Press.
- Kleinbaum, D. G., and Klein, M. 2010. *Logistic Regression: A Self-Learning Text*, Third Edition. New York and Berlin: Springer Publishers.
- Laird, N. M., and Ware, J. H. 1982. “Random Effects Models for Longitudinal Data, *Biometrics*, 38(4): 963–74.
- Littell, R. C.; Milliken, G. A.; Stroup, W. W.; and Wolfinger, R. D. 1996. *SAS System for Mixed Models*. Cary, N.C: SAS Institute.
- Ortiz, D. J.; Marcus, M.; Gerr, F.; Jones, W.; and Cohen, S. 1997. “Measurement Variability in Upper Extremity Posture among VDT Users.” *Applied Ergonomics* 28(2): 139–43.
- Zeger, S. L., and Liang, K. Y. 1986. “Longitudinal Data Analysis for Discrete and Continuous Outcomes.” *Biometrics* 42: 121–30.

26

Analysis of Correlated Data Part 2: Random Effects and Other Issues

26.1 Preview

In this chapter, we continue our discussion of the general linear mixed model for analyzing correlated data by focusing on the use of random effects in such models. Recall that in Chapter 17 on analysis of variance methods, we introduced and distinguished between *fixed effects* and *random effects*. Here we revisit this distinction in the context of correlated data. We then return to the analysis of the data from the Air Pollution Study described in the previous chapter, and we consider a model containing random effects to address the secondary objective of the analysis, the identification of sensitive subgroups or individuals most severely affected by the pollution episode. We then consider the use of random effects to analyze the data from the Study of the Posture of Computer Operators previously introduced in Chapter 25. We also use these data to describe the ANOVA “partitioning” approach for analyzing repeated measures data. Finally, we give a brief overview of how such analyses can be carried out for discrete-type (e.g., binary or count) outcomes using *generalized linear mixed models*. The MIXED procedure in SAS continues to be the default computer package referenced, due to both its robust capabilities and its historical prominence for fitting random-effects models. Analogous procedures exist in SPSS (MIXED procedure), R (LME and LMER procedures), and STATA (XTMIXED procedure) and are described in the online computer appendices that are available on the publisher’s website for the text.

26.2 Random Effects Revisited

We begin by reviewing the “classical” definitions of a *fixed factor* and a *random factor*, as described previously in Chapter 17:

Fixed Factor: A variable in a regression model whose possible values (i.e., levels) are the only ones of interest.

Random Factor: A variable in a regression model whose levels are regarded as a random sample from some large population of levels.

The above definitions can then be used to distinguish a fixed effect from a random effect:

Fixed Effect: A coefficient in a regression model corresponding to a fixed factor. Such an effect is considered a population parameter and is typically denoted by a Greek symbol (e.g., α , β , γ).

Random Effect (Definition 1): A term in a regression model corresponding to a random factor. Such an effect is considered a random variable and is typically denoted by a Latin letter (e.g., a , b , g).

When applying the above definitions to epidemiologic studies, we typically postulate that

- a. Subjects, litters, observers, families, and households are *random factors*;
- b. Gender, age, marital status, day of the week, and education are *fixed factors*; and
- c. Locations, treatments, clinics, exposures, and time may be considered as *either random or fixed factors*, depending on the context of the study.

When in doubt, one approach for deciding how to classify a particular study variable is to consider the following question: “If I was able to replicate the study, would I want a given factor to have the exact same categories as observed in the current study?” Equivalently, “Would I want a replicate study to use the same treatments, days of the week, or subjects as used in the current study?” If your answer is

yes: treat the factor as fixed
no: treat the factor as random

An alternative way to define a random effect, which gives another perspective on how to interpret a random effect, is given by the following second definition.

Random Effect (Definition 2): A random effect is a random variable that is included in a mixed model *to account for the effect of the natural heterogeneity among subjects* on the prediction of a response variable (Y) of interest—for example,

- i. One random effect for the intercept: $Y_{ij} = (\beta_0 + \beta_1 X_{ij}) + b_{j0} + E_{ij}$
- ii. Two random effects (intercept and slope): $Y_{ij} = (\beta_0 + \beta_1 X_{ij}) + (b_{j0} + b_{j1} X_{ij}) + E_{ij}$

The idea of heterogeneity among subjects when there is a single random effect for the intercept is illustrated in Figures 26.1 and 26.2. Figure 26.1 considers three subjects (Harry, Barry, and Gary), each of whom has his own straight-line model relating $X = \text{AGE}$ to $Y = \text{SBP}$. These subject-specific straight-line models differ in their subject-specific intercepts because of the random effects b_{H0} , b_{B0} , and b_{G0} , for which $E(b_{H0}) = E(b_{B0}) = E(b_{G0}) = 0$. Furthermore, each subject-specific model differs from the population model identified by the equation $E(Y|X) = \beta_0 + \beta_1 X$.

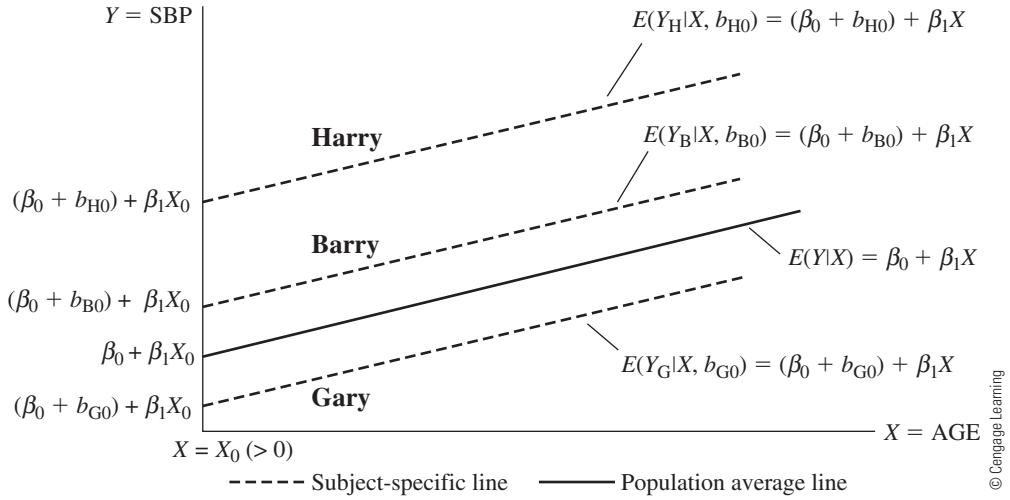


FIGURE 26.1 Heterogeneity among subjects with random intercepts

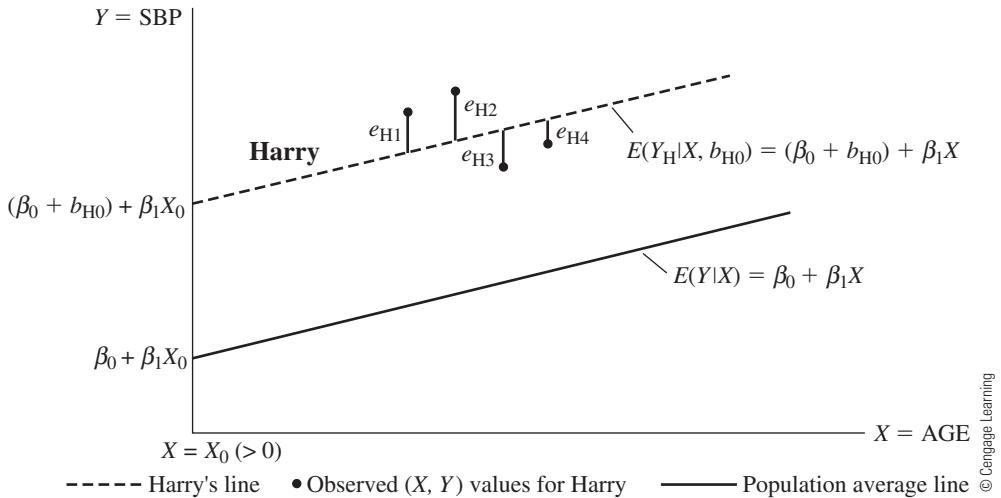
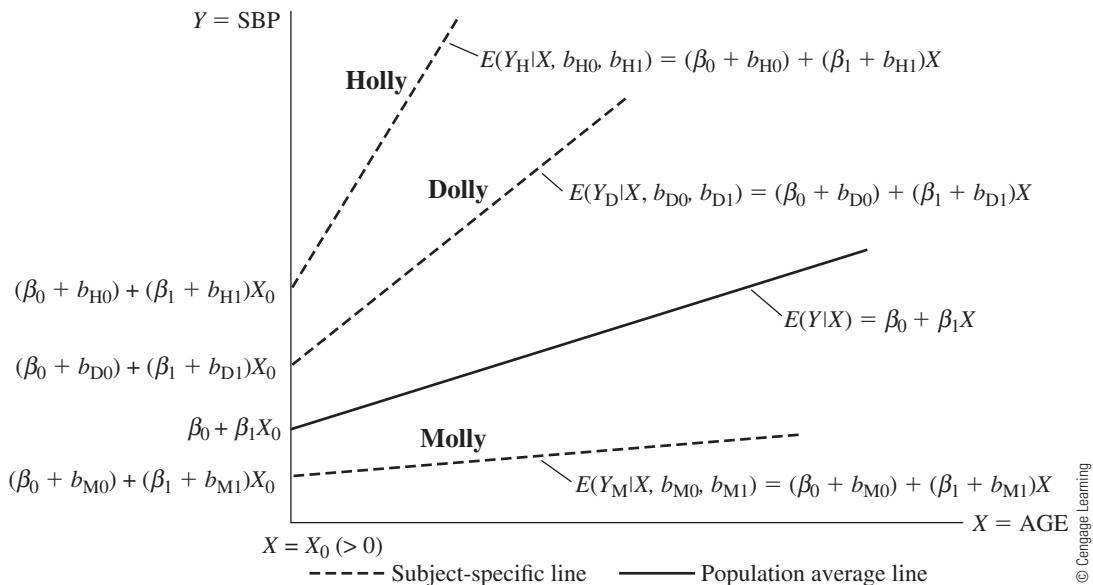


FIGURE 26.2 Heterogeneity with a random intercept

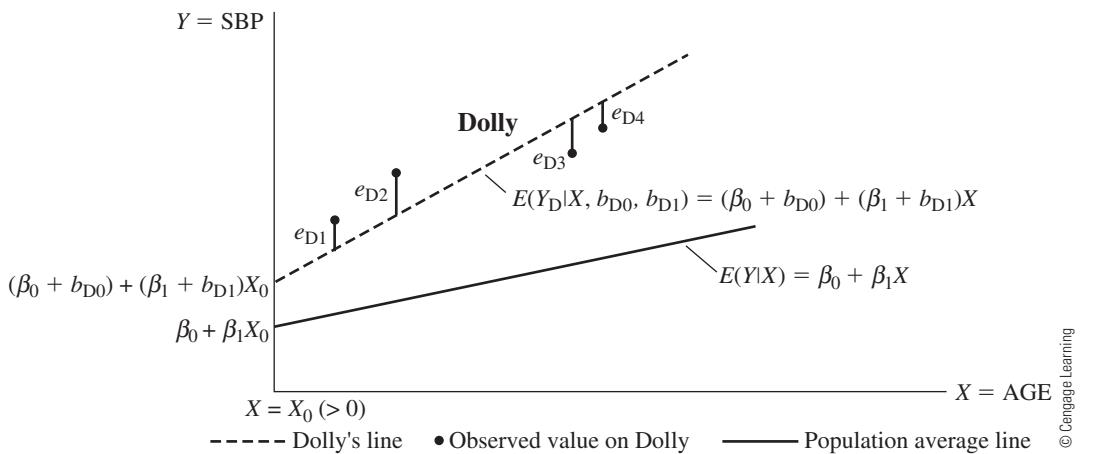
Figure 26.2 plots four repeated (X, Y) measurements on Harry, where again Harry's line differs from the population line because of Harry's subject-specific intercept. Each of these four measurements is a sample observation of SBP taken at a different age. Because of random variation, each repeated (X, Y) measurement does not lie exactly on Harry's line, and the differences between the measurements and Harry's line are denoted by the error terms $e_{Hj} = Y_{Hj} - E(Y_H|X_{Hj}, b_{H0}), j = 1, 2, 3, 4$.

Figure 26.3 considers three different subjects, Holly, Dolly, and Molly, each of whom has her own straight-line model relating $X = \text{AGE}$ to $Y = \text{SBP}$. These subject-specific

**FIGURE 26.3** Heterogeneity among subjects with random intercepts and slopes

straight-line models differ in both their intercepts (because of the random effects b_{H0} , b_{D0} , and b_{M0}) and their slopes (because of the random effects b_{H1} , b_{D1} , and b_{M1}); all these random effects are assumed to have expected (or true average) values of zero. Furthermore, each subject-specific model differs from the population model identified by the equation $E(Y|X) = \beta_0 + \beta_1 X$.

Figure 26.4 plots four repeated (X , Y) measurements on Dolly, where again Dolly's line differs from the population line because of Dolly's subject-specific intercept and

**FIGURE 26.4** Heterogeneity with a random intercept and a random slope

subject-specific slope. As with Harry in Figure 26.2, each of the four measurements on Dolly is a sample observation of SBP taken at a different age. Because of random variation, each repeated measurement does not lie exactly on Dolly's line, and the differences between the measurements and Dolly's line are denoted by the error terms $e_{Dj} = Y_{Dj} - E(Y_D | X_{Dj}, b_{D0}, b_{D1})$, $j = 1, 2, 3, 4$.

From the above discussion and Figures 26.1 through 26.4, we see that "heterogeneity of effects among subjects" indicates that each subject in the population under study has his or her own regression model—that is,

$$E(Y|X, b_{i0}, b_{i1}) = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})X \quad \text{for subject } i$$

whose general form will depend on the number and type of random effects. Each subject-specific model will differ from a population average model—that is,

$$E(Y|X) = \beta_0 + \beta_1 X$$

whose form contains population (average) parameters only, since $E(b_{i0}) = E(b_{i1}) = 0$.

Thus, it may be argued that *another reason for considering random effects in a mixed model, other than to identify individual subject effects as previously mentioned in Section 25.2.1 for the FEV1 data, is to determine whether or not the appropriate model needs to account for such "heterogeneity of effects among subjects."* That is, we may want to answer this question: "Is a model that incorporates subject-specific effects more appropriate than a model that ignores subject-specific effects?" A decision to include specific types of random effects (e.g., random slopes) in a mixed model should always involve subject-matter considerations. A related question is: "Assuming that random effects are needed, how many random effects need to be used, and how should they be defined in the model?" These questions suggest that tests about the overall significance of all random effects considered together, as well as tests of significance about individual random effects, are of interest. We describe such tests in the context of a second example in the next section.

26.3 Results for Models with Random Effects Applied to Air Pollution Study Data

The mixed linear model analyses previously described in Chapter 25, Section 25.4 for the FEV1 data set have addressed the following primary objective: determine whether or not true average FEV1 levels were depressed during and after the pollution alert when considering the 40 school children as a sample from a larger population of school children. In this section, we address the secondary objective: identify sensitive subgroups or individuals most severely affected by the pollution episode. This second objective focuses on drawing conclusions about individual subjects in the study rather than on making statistical inferences about the population mean FEV1 levels over the five weeks of the study. When individuals are the focus of the study, the appropriate approach to the analysis is to consider a model with one or more random effects, since such random effects are defined to represent subject-specific effects.

26.3.1 A Subject-specific Model with a Random Intercept

The simplest form of subject-specific model for the FEV1 data involves adding to model (25.7) of Chapter 25 a single subject-specific random effect b_{i0} to the population-average intercept β_0 , an addition that may be stated in scalar terms as follows:

$$Y_{ij} = \beta_0 + \beta_1 W_{ij1} + \beta_2 W_{ij2} + \beta_3 W_{ij3} + \beta_4 W_{ij4} + b_{i0} + E_{ij} \quad (26.1)$$

which contains one random effect b_{i0} for the i th subject in addition to the fixed factors W_1 through W_4 and their corresponding fixed effects (i.e., $\beta_1, \beta_2, \beta_3, \beta_4$). The subject-specific matrix form of this model is given as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{E}_i \quad i = 1, 2, \dots, 40 \quad (26.2)$$

where \mathbf{Y}_i denotes the set of five FEV1 measurements on the i th child, \mathbf{X}_i denotes the set of intercept and dummy variable values for each of the five weeks for subject i , $\boldsymbol{\beta}$ denotes the set of five β 's in this model, and \mathbf{E}_i denotes the set of five error terms for the i th subject. (The \mathbf{X}_i matrix and the $\boldsymbol{\beta}$ and \mathbf{E}_i vectors were defined earlier in (25.9).) The \mathbf{Z}_i matrix in matrix model (26.2) is the (5×1) column vector containing all ones—that is,

$$\mathbf{Z}_i = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

and the \mathbf{b}_i vector is simply the (1×1) scalar b_{i0} . (We will see shortly that the above simple subject-specific model is not appropriate for answering the secondary objective.)

Since model (26.2) contains both fixed and random effects, we need to specify both an \mathbf{R}_i matrix and a \mathbf{G} matrix in order to define and fit this model. Here, the \mathbf{G} matrix is a scalar [i.e., a (1×1) matrix] that specifies the variance (say, σ_0^2) of the random intercept effect b_{i0} . The simplest form of \mathbf{R}_i matrix that can be specified is an independence structure—that is, $\sigma_e^2 \mathbf{I}_5$, where σ_e^2 denotes the variance of any error component (E_{ij}). The assumption of “conditional” independence (i.e., $\mathbf{R}_i = \sigma_e^2 \mathbf{I}_5$) is a typical choice when the model contains random effects; this assumption basically states that, once the random effects are accounted for (i.e., conditional on b_{i0} being fixed), the remaining error terms (E_{ij} 's) are (conditionally) independent of one another.

Table 26.1 shows the estimated fixed effects, the overall F test concerning the equality of the five FEV1 population means, the F test for the significance of the contrast of interest, and the estimated correlation structures for model (26.2). This table also shows the previous output (Table 25.7) for a marginal model in which \mathbf{R} is \mathbf{CS} and $\mathbf{G} = \mathbf{0}$.

From Column 1, we can see that the estimated regression coefficients, as well as both model-based and empirical standard errors, are identical for both sets of \mathbf{R} and \mathbf{G} matrices. From Column 2, we see that the F -statistic values of 55.26 (model-based) and 38.61 (empirical)

TABLE 26.1 Edited output based on balanced FEV1 data—random intercept, conditional independence model versus marginal compound symmetric model, using REML in the SAS MIXED procedure

Column 1			Column 2			Column 3*		
G = σ_0^2, R = $\sigma_e^2 \mathbf{I}_5$								
Effect	Week	Estimate	Model-based SE	Empirical SE		Type 3 Tests of Fixed Effects		
Intercept		6.9985	0.2590	0.2418		Num DF	Den DF	Empirical F Value
week	1	2.8152	0.2439	0.2514	Effect week	4	156	55.26
week	2	-0.1503	0.2439	0.2052				38.61
week	3	-0.00325	0.2439	0.2076				
week	4	0.0047	0.2439	0.2710	contrast	1	156	147.46
G = 0, R = CS						Type 3 Tests of Fixed Effects		
Effect	Week	Estimate	Model-based SE	Empirical SE		Num DF	Den DF	Empirical F Value
Intercept		6.9985	0.2590	0.2418	Effect week	4	195	55.26
week	1	2.8152	0.2439	0.2514				38.61
week	2	-0.1503	0.2439	0.2052				
week	3	-0.00325	0.2439	0.2076				
week	4	0.0047	0.2439	0.2710	contrast	1	195	220.51

*C^V denotes the correlation matrix that corresponds to the covariance matrix V in the model with one random intercept effect (i.e., G = σ_0^2 , R = $\sigma_e^2 \mathbf{I}_5$); C^(EXCH) denotes the correlation matrix that corresponds to the covariance matrix V in the marginal CS model (i.e., G = 0, R = CS).

are also identical for the two choices of covariance structures. These results are as expected, since we have previously stated that the general covariance matrix formula

$$\mathbf{V}_i = \mathbf{Z}_i \mathbf{G} \mathbf{Z}'_i + \mathbf{R}_i$$

yields the same overall covariance structure for \mathbf{V}_i when the \mathbf{G} and \mathbf{R} matrices are either

$$\{\mathbf{G} = \sigma_0^2, \mathbf{R} = \sigma_e^2 \mathbf{I}_5\} \quad \text{or} \quad \{\mathbf{G} = \mathbf{0}, \mathbf{R} = \mathbf{C}\mathbf{S}\}$$

The only difference in the output for the two fitted models is the 156 denominator degrees of freedom (Den DF) for the random intercept model versus the 195 Den DF for the marginal model in Column 2. We have previously mentioned that SAS's MIXED procedure allows the user to choose among several options for the Den DF, so we could have used 195 for both models. However, when a random effects model is used (see footnote 10 in Chapter 25), the popular choice for the Den DF involves subtracting from the residual DF of 195 the value $q^*(K - 1)$ to account for the number (q^*) of random effects being estimated. Since $q^* = 1$ and there are $K = 40$ values of b_{i0} to be estimated (one for each subject), $q^*(K - 1) = 39$, so that the resulting Den DF value is $(195 - 39) = 156$. Since the marginal model does not require estimation of random effects, one does not need to subtract additional DF from the residual DF of 195 for this model.

Table 26.2 provides the values for the estimated random effects for the first 10 of the 40 subjects in the FEV1 data set. The estimated random effects are shown in the "Estimate"

TABLE 26.2 Edited output based on balanced FEV data—estimates of random effects from random intercept, conditional independence model, using REML in the SAS MIXED procedure

Estimates for Random Effects							Comparison of Mean FEV1 values		
Effect	Subject	Estimate	Std Error	DF	t Value	Pr > t	FEV1 _i	FEV1	Difference
b_{10}	1	-0.5216	0.4874	156	-1.07	0.2861	6.92	7.52	-0.60
b_{20}	2	1.9332	0.4874	156	3.97	0.0001*	9.76	7.52	2.24
b_{30}	3	-2.2019	0.4874	156	-4.52	<.0001*	4.97	7.52	-2.55
b_{40}	4	-1.4722	0.4874	156	-3.02	0.0029*	5.82	7.52	-1.70
b_{50}	5	1.2621	0.4874	156	2.59	0.0105 [†]	8.99	7.52	1.47
b_{60}	6	0.6411	0.4874	156	1.32	0.1903	8.27	7.52	0.75
b_{70}	7	0.0425	0.4874	156	0.09	0.9307	7.57	7.52	0.05
b_{80}	8	-1.2203	0.4874	156	-2.50	0.0133 [†]	6.17	7.52	-1.35
b_{90}	9	1.3984	0.4874	156	2.87	0.0047*	9.14	7.52	1.62
$b_{10,0}$	10	-0.7562	0.4874	156	-1.55	0.1228	6.65	7.52	-0.87

*Significant at .01 level

[†]Significant at .05 level

column. From this table, among the first 10 subjects, subjects 2, 3, 4, 5, 8, and 9 each have estimated random effects that are significantly different from zero. The last three columns of Table 26.2 give the observed mean FEV1 value for each subject ($\bar{FEV1}_i$) over all five weeks, the mean FEV1 value ($\bar{FEV1}$) of 7.52 for all subjects over the five weeks, and the difference between these two observed means. We can see that the 6 subjects with significant estimated random effects also have the largest observed mean differences. This information supports the fact that the estimated random effect for the i th subject (denoted as \hat{b}_{i0}) reflects the observed difference between $FEV1_i$ for subject i and $FEV1$.

It can be shown mathematically¹ that the formula for the estimated random effect for the i th subject is

$$\hat{b}_{i0} = \frac{5\hat{\sigma}_0^2}{5\hat{\sigma}_0^2 + \hat{\sigma}_1^2} (\bar{FEV1}_i - \bar{FEV1}) \quad i = 1, \dots, 40 \quad (26.3)$$

26.3.2 A Model with Two Random Effects

Unfortunately, the random intercepts model (26.1) or (26.2) does not adequately address the secondary objective of the FEV1 study: to identify those subjects most severely affected by the pollution episode. In Chapter 25, Section 25.2.1, we indicated that a more appropriate way to determine how much a subject was affected by the pollution episode was to compute for each subject the linear contrast

$$L_i = FEV1_{i1} - \frac{FEV1_{i2} + FEV1_{i3} + FEV1_{i4} + FEV1_{i5}}{4}$$

which compares that subject's FEV1 value at week 1 (prior to the pollution alert) with the average of the FEV1 values for that subject over the other four weeks (during and after the pollution alert). For example, for subjects 1 and 2, we found that

$$L_1 = 9.43 - 6.29 = 3.14 \text{ for subject 1} \quad \text{and} \quad L_2 = 11.15 - 9.42 = 1.73 \text{ for subject 2}$$

Thus, subject 1 seemed to be more adversely affected by the pollution alert than was subject 2.

In order to compare all 40 subjects in the study, we suggest that the estimated linear contrast L_i for each subject should be compared to the estimated average contrast \bar{L} over all 40 subjects in the study, which we previously calculated to be

$$\bar{L} = \frac{1}{40} \sum_{i=1}^{40} L_i = 9.81375 - \frac{(6.84825 + 7.00175 + 6.95150 + 6.99850)}{4} = 2.8637$$

¹ The general formula for the estimate of the set (i.e., vector) of random effects \mathbf{b}_i for the general linear mixed model (25.35) is called the "best linear unbiased predictor" (BLUP) and is given by the following matrix expression (see Littell et al. 1996):

$$\hat{\mathbf{b}}_i = \hat{\mathbf{G}}\mathbf{Z}_i \hat{\mathbf{V}}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}})$$

Thus, we can evaluate the extent to which each subject was affected by the pollution episode relative to the average effect over all 40 subjects by calculating the difference score

$$\text{DIFF}_{L_i} = 2.8637 - L_i \quad (26.4)$$

for each subject. Those subjects with “large”² *negative* difference scores can then be identified as subjects most adversely affected by the pollution episode.³

For example, for subjects 1 and 2, respectively, we obtain

$$\text{DIFF}_{L_1} = 2.8637 - 3.14 = -0.2763 \text{ and } \text{DIFF}_{L_2} = 2.8637 - 1.73 = 1.1337$$

Thus, subject 1, with a “small” negative difference score of -0.2763 , was (slightly) adversely affected by the pollution episode when compared to all 40 subjects. In contrast, subject 2, with a relatively “large” positive score of 1.1337 , was not adversely affected by the pollution episode when compared to all 40 subjects.

To carry out the above analysis using a linear mixed model, which will provide significance tests for the subject-specific differences just illustrated, we consider the following model:

$$Y_{ij} = \beta_0 + \beta_1 W_{ij1} + \beta_2 W_{ij2} + \beta_3 W_{ij3} + \beta_4 W_{ij4} + b_{i0} + b_{i1} Z_{ij1} + E_{ij} \quad (26.5)$$

which contains two random effects b_{i0} and b_{i1} and the covariate Z_{ij1} in addition to the fixed factors W_1 through W_4 and their corresponding fixed effects (i.e., β 's); here W_1 through W_4 represent the same four dummy variables for distinguishing the five weeks. The Z_{ij1} variable is defined as follows:

$$Z_{ij1} = \begin{cases} 0 & \text{if the } j\text{th measurement on subject } i \text{ occurs during week 1} \\ 1 & \text{otherwise (i.e., if the } j\text{th measurement on subject } i \text{ occurs} \\ & \text{during weeks 2, 3, 4, or 5)} \end{cases} \quad (26.6)$$

The subject-specific matrix form of this model is given by

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \mathbf{E}_i, \quad i = 1, 2, \dots, 40 \quad (26.7)$$

² We have used “large” in quotes to indicate that a statistical test of whether a particular difference score is significantly different from zero should be considered. Note that *positive* differences will identify subjects whose FEV1 scores increased on average over weeks 2 through 5 relative to week 1 when compared to all 40 subjects; a statistical test is not required in such cases.

³ The difference formula (26.4) subtracts the estimated subject-specific contrast L_i from the average of all estimated contrasts (2.8637) rather than the other way around. Mathematically, the sign of this difference score conforms to the sign of the estimated random effect b_{i1} defined by mixed model (26.5); in particular, negative estimates of b_{i1} will correspond to *negative* values of (26.4). For further details, see footnote 5.

where \mathbf{Y}_i , \mathbf{X}_i , $\boldsymbol{\beta}$, and \mathbf{E}_i are the same as in matrix model (26.2) and where \mathbf{b}_i and \mathbf{Z}_i are given by

$$\mathbf{b}_i = \begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix}, \quad \mathbf{Z}_i = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \quad (25.13 \text{ repeated})$$

To run this more customized model, we need to specify that \mathbf{R} is $\sigma_e^2 \mathbf{I}_5$ (i.e., we assume conditional independence) and that \mathbf{G} is a (2×2) matrix (since the model now contains two random effects) of the form

$$\mathbf{G} = \begin{bmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}$$

The \mathbf{G} matrix has 0's for the two off-diagonal elements, indicating that the random effects b_{i0} and b_{i1} in model (26.5) are assumed to be uncorrelated.⁴

Because the \mathbf{G} matrix is no longer a scalar, as it was with the simple random intercept model previously considered, the covariance structure for the subject-specific model (26.5) or, equivalently, (26.7) is no longer compound symmetric; but rather, it is determined by using the \mathbf{Z}_i and \mathbf{G} matrices above and the $\mathbf{R}_i = \sigma_e^2 \mathbf{I}_5$ matrix in the general covariance formula $\mathbf{V}_i = \mathbf{Z}_i \mathbf{G} \mathbf{Z}'_i + \mathbf{R}_i$.

Table 26.3 gives the estimated fixed effects, overall F statistic for assessing equality of the five mean FEV1 values, and estimated correlation/covariance structures for model (26.5), in which there are two random effects. This table also shows, for comparison purposes, the previous output (from Table 26.1) for the random intercept subject-specific mixed model in which \mathbf{R} is $\sigma_e^2 \mathbf{I}_5$ and \mathbf{G} is the scalar σ_0^2 .

From Column 1 of Table 26.3, we can see that the set of model-based estimated standard errors for model (26.7) containing random intercept and slope effects is different from the corresponding set of model-based estimated standard errors for the random intercept only model (26.2). This result is expected, since the addition of the second random (slope) effect in model (26.7) changes the \mathbf{G} matrix and corresponding overall covariance structure \mathbf{V} from those used in model (26.2).⁵

⁴ If the user wanted to allow for a possible correlation between b_{i0} and b_{i1} , the "random" statement could be modified as follows: *random intercept Z1/ type=cs subject=subj s g*. The latter statement will replace the 0's in the above matrix with a covariance parameter—say, σ_{01} —which would have to be estimated in addition to σ_0^2 and σ_1^2 .

⁵ There are a number of ways to prove that the second random effect b_{i1} in random-effects model (26.7) actually provides an estimate of the contrast difference (26.4) that describes the extent to which a subject is affected by the pollution alert. One approach, which is difficult theoretically, is to solve for b_{i0} and b_{i1} in model (26.7) using the general formula

$$\hat{\mathbf{b}}_i = \hat{\mathbf{G}} \mathbf{Z}_i \hat{\mathbf{V}}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}})$$

An alternative approach is to use empirical Bayes' formulas (Littell et al. 1996) for the conditional expected values $E(\text{FEV1}_{ij} | b_{i1})$ to solve for b_{i1} in terms of a difference between a contrast involving population mean μ_j values and an analogous contrast involving these empirical Bayes' conditional expected values:
 $b_{i1} = \mu_i - EB_{L_i}$

(continued on page 837)

TABLE 26.3 Edited output based on balanced FEV data—mixed model (26.7) with random intercept and slope versus random intercept only model, using REML in the SAS MIXED procedure

Column 1				Column 2				Column 3				
$\mathbf{G} = \begin{bmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}$, $\mathbf{R} = \sigma_e^2 \mathbf{I}_5$				Type 3 Tests of Fixed Effects				$\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R} \equiv \sigma^2 \mathbf{C}^{\mathbf{v}}$ where $\hat{\mathbf{C}}^{\mathbf{v}}$ is				
Effect	Week	Estimate	Model-based SE	Effect	Den DF	Model-based FValue	Empirical FValue	Effect	Den DF	Model-based FValue	Empirical FValue	
Intercept		6.9985	0.2686	0.2418	4	117	32.99	0.2931	1.0000	0.2931	0.2931	
week	1	2.8152	0.2834	0.2514				0.2931	0.2931	0.6705	0.6705	
week	2	-0.1503	0.2181	0.2052				0.2931	0.6705	0.6705	0.6705	
week	3	0.00325	0.2181	0.2076				0.2931	0.6705	0.6705	1.0000	
week	4	-0.0047	0.2181	0.2710	contrast	1	195	131.30				
							147.46					
								$\hat{\sigma}^2 = \hat{\sigma}_0^2 + \hat{\sigma}_1^2 + \hat{\sigma}_e^2 = 2.8856$, where $\hat{\sigma}_0^2 = 0.6250$, $\hat{\sigma}_1^2 = 1.3097$, $\hat{\sigma}_e^2 = 0.9509$				
								$\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R} \equiv \sigma^2 \mathbf{C}^{\mathbf{v}}$ where $\hat{\mathbf{C}}^{\mathbf{v}}$ is				
$\mathbf{G} = \sigma_0^2, \mathbf{R} = \sigma_e^2 \mathbf{I}_5$				Type 3 Tests of Fixed Effects				$\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R} \equiv \sigma^2 \mathbf{C}^{\mathbf{v}}$ where $\hat{\mathbf{C}}^{\mathbf{v}}$ is				
Effect	Week	Estimate	Model-based SE	Effect	Den DF	Model-based FValue	Empirical FValue	Effect	Den DF	Model-based FValue	Empirical FValue	
Intercept		6.9985	0.2590	0.2418	4	156	55.26	0.5565	1.0000	0.5565	0.5565	
week	1	2.8152	0.2439	0.2514				0.5565	0.5565	0.5565	0.5565	
week	2	-0.1503	0.2439	0.2052				0.5565	0.5565	0.5565	0.5565	
week	3	0.00325	0.2439	0.2076				0.5565	0.5565	0.5565	1.0000	
week	4	-0.0047	0.2439	0.2710	contrast	1	195	220.51				
							147.46					
								$\hat{\sigma}^2 = \hat{\sigma}_0^2 + \hat{\sigma}_e^2 = 2.6837$, where $\hat{\sigma}_0^2 = 1.4936$, $\hat{\sigma}_e^2 = 1.1901$				

* $\mathbf{C}^{\mathbf{v}}$ denotes the correlation matrix that corresponds to the covariance matrix \mathbf{V} in the mixed model (26.7) with two uncorrelated random effects b_{0j} and b_{1j} ; that is, $\mathbf{G}(2 \times 2) = \mathbf{diag}(\sigma_0^2, \sigma_1^2)$, $\mathbf{R} = \sigma_e^2 \mathbf{I}_5$.

$\mathbf{C}^{\mathbf{v}}$ denotes the correlation matrix that corresponds to the covariance matrix \mathbf{V} in the mixed model (26.2) with a random intercept; that is, $\mathbf{G} = \sigma_0^2 \mathbf{I}_2$, $\mathbf{R} = \sigma_e^2 \mathbf{I}_5$. When a random statement is used in SAS's MIXED procedure, the overall covariance structure \mathbf{V} , or its corresponding overall correlation structure \mathbf{C} , is provided in the output using a "v" or "vcov" statement, respectively, as an option to the random statement.

The difference in correlation structures between model (26.2) and model (26.7) is also reflected in Column 2 of Table 26.3 by the different model-based F -statistic values. Also, the Den DF value shown for model (26.7) is

$$195 - q^*(K - 1) = 195 - 2(39) = 117$$

a value adjusted for the $q^* = 2$ random effects being estimated for each of the 40 subjects.

Column 3 of Table 26.3 compares the estimated correlation structures for models (26.7) and (26.2). Notice that the correlation structure for model (26.7) is not “exchangeable,” as reflected by the estimated correlations not being identical.

Table 26.4 provides the values for the estimated random effects for model (26.7) for the first 10 subjects based on model (26.7). The estimated random effects are shown in the “Estimate” column. The last column of the table shows the contrast difference scores (26.4) for these 10 subjects. Those subjects with “large” negative difference scores, which correspond to negative \hat{b}_{i1} estimated random effects, can then be identified as subjects most adversely affected by the pollution episode.

From Table 26.4, subjects 1, 3, 4, 7, 8, and 10 have contrast difference scores that are negative, but only subject 3 has an estimated random effect \hat{b}_{i3} that is significantly different from zero. These results indicate that only subject 3 of the first 10 subjects was significantly more adversely affected by the pollution episode than the average subject. A similar analysis is required to assess whether some of the remaining 30 subjects were also significantly adversely affected by the pollution episode.

26.3.3 Summary of Subject-specific Analyses for the FEV1 Data

We now summarize the results obtained in the previous section for the subject-specific analyses of the FEV1 data:

1. With regard to the secondary question of how to identify sensitive subgroups or individuals most severely affected by the pollution episode:
 - a. One possible answer to this question involves deciding whether the estimated contrast involving the FEV mean level at week 1 and the average FEV level over weeks 2 through 5 for a given subject differs statistically from an estimated average of such contrasts over all 40 subjects.
 - b. Since the above question requires making statistical conclusions about individual subjects, a mixed model involving random effects should be used rather than a marginal model without random effects.
 - c. A simple random intercept model is not appropriate for addressing the secondary question, since this particular subject-specific model does not allow for

(continued from page 835)

where

$$\mu_L = \left[\mu_1 - \frac{\mu_2 + \mu_3 + \mu_4 + \mu_5}{4} \right]$$

and

$$EB_{L_i} = \left[E(\text{FEV1}_{i1} | \mathbf{b}_i) - \frac{E(\text{FEV1}_{i2} | \mathbf{b}_i) + E(\text{FEV1}_{i3} | \mathbf{b}_i) + E(\text{FEV1}_{i4} | \mathbf{b}_i) + E(\text{FEV1}_{i5} | \mathbf{b}_i)}{4} \right]$$

TABLE 26.4 Edited output based on balanced FEV data—estimates of random effects from mixed model (26.7) with random intercept and slope, using REML in the SAS MIXED procedure

Effect	Subject	Estimates for Random Effects					DIFF_{L_i} = 2.8637 - L_i
		Estimate	Std Error	DF	t Value	Pr > t 	
b ₁₀	1	-0.2517	0.5579	117	-0.45	0.6527	
b ₁₁	1	-0.3456	0.6639	117	-0.52	0.6037	-0.2763[†]
b ₂₀	2	0.9096	0.5579	117	1.63	0.1057	
b ₂₁	2	1.3185	0.6639	117	1.99	0.0494*	1.1337
b ₃₀	3	-1.0069	0.5579	117	-1.80	0.0737	
b ₃₁	3	-1.5494	0.6639	117	-2.33	0.0213*	-1.9238[†]
b ₄₀	4	-0.5837	0.5579	117	-1.05	0.2976	
b ₄₁	4	-1.1818	0.6639	117	-1.78	0.0776	-1.3663[†]
b ₅₀	5	0.5166	0.5579	117	0.93	0.3563	
b ₅₁	5	0.9867	0.6639	117	1.49	0.1399	1.0963
b ₆₀	6	-0.00517	0.5579	117	-0.01	0.9926	
b ₆₁	6	0.9375	0.6639	117	1.41	0.1606	1.7963
b ₇₀	7	0.1433	0.5579	117	0.26	0.7978	
b ₇₁	7	-0.1721	0.6639	117	-0.26	0.7959	-0.5463[†]
b ₈₀	8	-0.6552	0.5579	117	-1.17	0.2426	
b ₈₁	8	-0.7002	0.6639	117	-1.05	0.2937	-0.3388[†]
b ₉₀	9	0.3799	0.5579	117	0.68	0.4972	
b ₉₁	9	1.4071	0.6639	117	2.12	0.0362*	2.1063
b _{10,0}	10	-0.4209	0.5579	117	-0.75	0.4521	
b _{10,1}	10	-0.4097	0.6639	117	-0.62	0.5383	-0.1413[†]

*Significant at .01 level

[†]Difference score indicating subject with estimated larger adverse effect of pollution episode than the average subject

© Cengage Learning

consideration of the contrast between the week 1 FEV1 mean and the means for the other four weeks for each subject.

- d. A subject-specific linear mixed model containing two random effects, as defined by (26.5) or, equivalently, (26.7), can be used to address the secondary question, determining those individuals significantly adversely affected by the pollution alert.
- 2. The following comments are relevant based on the use of these subject-specific models (varying $\mathbf{R}_i, \mathbf{G} \neq \mathbf{0}$):
 - a. The general matrix formula for the covariance structure is given by $\mathbf{V}_i = \mathbf{Z}_i \mathbf{G} \mathbf{Z}'_i + \mathbf{R}_i$, where \mathbf{Z}_i specifies the random factors in the matrix version of the model.

- b. The same covariance structure is obtained when either $\{\mathbf{R} = \mathbf{C}\mathbf{S}$ and $\mathbf{G} = \mathbf{0}\}$ or $\{\mathbf{R} = \sigma_e^2\mathbf{I}$ and \mathbf{G} is the scalar $\sigma_0^2\}$.
- c. When fitting a subject-specific model, the output will include estimated random effects for all K subjects. If there are $K = 40$ subjects and one random effect b_{i0} (i.e., a random intercept) for each subject, the output will list 40 random-effect estimates; for a model with two random effects, b_{i0} and b_{i1} , for each subject, there will be 80 estimated random effects in total, 40 for each random effect in the model (i.e., two for each of the $K = 40$ subjects).
- d. When there are two or more random effects in a mixed linear model, the overall covariance structure defined by the formula $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}'_i + \mathbf{R}_i$ is not compound symmetric (see the estimated covariance structures provided in Table 26.3).

26.4 Second Example—Analysis of Posture Measurement Data

So far, we have considered two reasons for considering random effects in the analysis of correlated data. There are two other reasons to mention, which we also summarize in Table 26.5.

Reasons 3 and 4 in Table 26.5 are illustrated in this section, in which we discuss the analysis of the posture measurement data introduced earlier in Section 25.2.

In Section 25.2.2 and Table 25.5, we described data from a study by Ortiz et al. (1997) involving a random sample of 19 computer operators recruited from a major utility company in Atlanta. A primary goal of this study was to evaluate the effects of two factors on a measurement of posture called Shoulder Flexion (SF). The two factors were *day of the work week* (early, middle, or late), denoted as “Day,” and time of day (AM or PM), denoted as “Time.” Each of the 19 subjects was measured sequentially both in the AM and in the PM on Monday, Wednesday, and Friday of the same week. Thus, a total of six repeated measurements were made on each subject for this response (SF), involving all combinations of the two factors (Day and Time). In ANOVA terminology, there were three levels of Day (Monday, Wednesday, and Friday), and two levels of Time (AM and PM), and each subject was observed at each level for each of the two factors. This study design is *balanced* in that each subject has the same number of repeated measurements (namely, six), with all subjects

TABLE 26.5 Some reasons for using random effects in a mixed model for correlated data

1. To estimate cluster-specific (i.e., subject-specific) effects.
2. To account for effects of natural heterogeneity among subjects on the prediction of the response variable.
3. To use a correlation structure not available as a choice in available computer packages (e.g., as for a “repeated measures ANOVA” approach).
4. To represent one or more exposure/study/design variables (other than cluster-related variables) as a random factor.

measured at the same times. The study design here is often called a *two-factor crossover design*, where the crossover factors are Day and Time. By *crossover factor*, we mean any variable for which each study subject is observed at two or more levels of that variable. In a *completely balanced* crossover design, each subject is observed at *every* combination of levels of the crossover factors, as is the case for the crossover factors Day and Time in the Ortiz study. (More generally, a balanced design, which may or may not involve crossover factors, involves the same number of responses [i.e., Y values] measured at the same times for each subject.)

26.4.1 Arranging the Data for a Repeated Measures Analysis

We previously (Chapter 25, Section 25.2.2) noted that the sample mean SF scores were lower on Friday (15.58, 16.79) than on either Monday (18.95, 18.47) or Wednesday (19.74, 18.53); that is, there appears to be a possible main effect of the factor Day. Also, there was some suggestion of interaction between the factors Day and Time: the Wednesday AM mean score (19.74) was higher than the Wednesday PM score (18.53), whereas the Friday AM mean score (15.58) was lower than the Friday PM mean score (16.79). Nevertheless, to determine whether or not there are statistically significant main effects or interaction effects involving the two factors Day and Time, a repeated measures analysis is required, which we now describe. Since the outcome variable, SF, is continuous, we will use a general linear mixed model approach to carry out the analysis.

First, we rearrange the data in Table 25.5, which contains one line of data for each individual, to conform to the data layout format given by Table 25.2, which involves a different line of data for each different observed response on the same individual. Table 26.6 gives the rearranged table. Both the Day and Time variables are time-dependent variables, since their values change over the different times of SF measurement on each subject. Nevertheless, both Day and Time should be treated as *fixed factors*, since these were the only days of the week and times of day of interest to the investigators. However, if the variable (i.e., factor) "Subjects" is also contained in the model, then this variable should be treated as a *random factor*, since the investigators considered the study subjects to be a random sample from a larger population of subjects.

26.4.2 Results from Fitting Marginal Models: SF Data

A *marginal model* that contains the main effects of the factors Day and Time, plus a Day-by-Time interaction, can be given by the following subject-specific scalar equation:

$$SF_{ij} = \beta_0 + \beta_1 D_{ij1} + \beta_2 D_{ij2} + \beta_3 T_{ij} + \beta_4(D_{ij1}T_{ij}) + \beta_5(D_{ij2}T_{ij}) + E_{ij} \quad (26.8)$$

where D_{ij1} and D_{ij2} denote the values of two dummy variables D_1 and D_2 for Day and T_{ij} denotes the values of a binary variable T for Time, for the j th observation on the i th subject, $i = 1, 2, \dots, 19 (= K)$ and $j = 1, 2, \dots, 6 (= n_i)$. Here the dummy variables for Day are coded as

$$\begin{aligned} D_1 &= 1 \text{ if Monday, 0 otherwise} \\ D_2 &= 1 \text{ if Wednesday, 0 otherwise} \end{aligned}$$

TABLE 26.6 Repeated measures data layout for posture study data

Subject(<i>i</i>)	Repeat(<i>j</i>)	<i>Y</i> = SF	Day	Time
1	1	$Y_{11} = 17$	M = 1	AM = 1
1	2	$Y_{12} = 5$	M = 1	PM = 2
1	3	$Y_{13} = 10$	W = 2	AM = 1
1	4	$Y_{14} = 1$	W = 2	PM = 2
1	5	$Y_{15} = 5$	F = 3	AM = 1
1	6	$Y_{16} = 1$	F = 3	PM = 2
2	1	$Y_{21} = 1$	M = 1	AM = 1
2	2	$Y_{22} = 7$	M = 1	PM = 2
2	3	$Y_{23} = 7$	W = 2	AM = 1
2	4	$Y_{24} = 4$	W = 2	PM = 2
2	5	$Y_{25} = 19$	F = 3	AM = 1
2	6	$Y_{26} = 12$	F = 3	PM = 2
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
19	1	$Y_{19,1} = 1$	M = 1	AM = 1
19	2	$Y_{19,2} = 5$	M = 1	PM = 2
19	3	$Y_{19,3} = 0$	W = 2	AM = 1
19	4	$Y_{19,4} = 0$	W = 2	PM = 2
19	5	$Y_{19,5} = 6$	F = 3	AM = 1
19	6	$Y_{19,6} = 2$	F = 3	PM = 2

© Cengage Learning

The dummy variable for Time is coded as

$$T = 1 \text{ if AM, } 0 \text{ if PM}$$

Since *SF* is a continuous variable and there are six SF measurements on each of the 19 subjects, we will assume for simplicity that the error term E_{ij} has the normal distribution $N(0, \sigma_e^2)$ for all *i* and *j* and that $\text{corr}(E_{ij}, E_{ij'}) \neq 0$ for $j \neq j'$.

The subject-specific matrix version of (26.8) can be written as follows:

$$\mathbf{SF}_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}_i \quad (26.9)$$

where \mathbf{SF}_i is a (6×1) vector of SF observations $\{SF_{i1}, SF_{i2}, SF_{i3}, SF_{i4}, SF_{i5}, SF_{i6}\}$ and $\mathbf{X}_i = \mathbf{X}$ is a (6×6) matrix of predictor values given by

$$\mathbf{X}(6 \times 6) = \begin{bmatrix} 1 & D_{i11} & D_{i12} & T_{i1} & D_{i11}T_{i1} & D_{i12}T_{i1} \\ 1 & D_{i21} & D_{i22} & T_{i2} & D_{i21}T_{i2} & D_{i22}T_{i2} \\ 1 & D_{i31} & D_{i32} & T_{i3} & D_{i31}T_{i3} & D_{i32}T_{i3} \\ 1 & D_{i41} & D_{i42} & T_{i4} & D_{i41}T_{i4} & D_{i42}T_{i4} \\ 1 & D_{i51} & D_{i52} & T_{i5} & D_{i51}T_{i5} & D_{i52}T_{i5} \\ 1 & D_{i61} & D_{i62} & T_{i6} & D_{i61}T_{i6} & D_{i62}T_{i6} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$\boldsymbol{\beta}$ is a (6×1) vector of fixed-effect parameters $\{\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$, and \mathbf{E}_i is a (6×1) vector of random error terms $\{E_{i1}, E_{i2}, E_{i3}, E_{i4}, E_{i5}, E_{i6}\}$.

In the above coding for \mathbf{X} , we have assumed that the referent group for Day is Friday and the referent group for Time is PM, although other choices for the referent groups are possible.

For the marginal model defined by (26.8) or (26.9), the covariance/correlation structure is determined entirely by the \mathbf{R} matrix (since $\mathbf{G} = \mathbf{0}$), which gives the variances and covariances for the error terms E_{ij} . Table 26.7 summarizes the output for three choices of \mathbf{R} : compound symmetric ($\mathbf{R} = \mathbf{CS}$), autoregressive ($\mathbf{R} = \mathbf{AR1}$), and unstructured ($\mathbf{R} = \mathbf{UN}$).

From the output (see Table 26.7), the same conclusions about the significance of the effects are made, regardless of the choice of \mathbf{R} ; that is, the main effects of Day and Time, as well as the interaction effects of Day with Time, are all nonsignificant ($P > .05$). As expected, model-based estimated standard errors differ as \mathbf{R} differs. As with the FEV1 data set previously analyzed, empirical estimated standard errors are identical, regardless of the choices for \mathbf{R} . Similarly, model-based F -statistic values vary as \mathbf{R} varies, whereas empirical F -statistic values are identical, regardless of the choice for \mathbf{R} . Furthermore, as expected, the estimated correlations in the unstructured correlation matrix are all different. Nevertheless, since all such correlations are relatively high and roughly similar in value, a CS/exchangeable correlation structure seems reasonable to use.

The above analyses, involving three choices for \mathbf{R} , support the conclusion that there are no statistically significant main effects or interaction effects involving the predictors Day and Time. If one decided to postulate a correlation structure appropriate for these data before considering the numerical results in Table 26.7, one might initially consider an autoregressive (**AR1**) structure, since the study is longitudinal (so that correlations decrease as the time between repeated observations increases). Nevertheless, we might expect repeated observations on the same day to be more correlated than observations on different days, which suggests a more complex correlation structure than either **AR1** or compound symmetric (**CS**). There are many other choices for the correlation structure still available, including structures that result from the use of random effects. In particular, we might want to account for the natural heterogeneity among “Subjects,” a third predictor variable, when modeling the response variable *SF*.

TABLE 26.7 Edited output based on modeling SF data using predictors Day and Time for different marginal models ($G = 0$), using REML in SAS's MIXED procedure

Column 1			Column 2^{*†}			Column 3[‡]		
CS	Model-based Estimate	Empirical SE		Type 3 Tests of Fixed Effects				
Effect	Estimate	SE		Num DF	Den DF	Model-based F Value	Empirical F Value	$\mathbf{R} = \sigma^2 \mathbf{C}^{(\text{EXCH})}$ where $\hat{\mathbf{C}}^{(\text{EXCH})}$ is
Intercept	16.7895	2.8884	3.0368	Effect				1.0000 0.7792 0.7792 0.7792 0.7792 0.7792
day 1	1.6842	1.9257	1.6274	day	2	36	2.74	0.7792 1.0000 0.7792 0.7792 0.7792 0.7792
day 2	1.7368	1.9257	2.4243	time	1	18	0.02	0.7792 0.7792 1.0000 0.7792 0.7792 0.7792
time	-1.2105	1.9257	1.2990	day × time	2	36	0.42	0.7792 0.7792 0.7792 1.0000 0.7792 0.7792
day × time 1	1.6842	2.7234	1.7771	day × time 2	2	36	0.84	0.7792 0.7792 0.7792 0.7792 1.0000 0.7792
day × time 2	2.4211	2.7234	1.9879					$\hat{\sigma}^2 = \hat{\sigma}_0^2 + \hat{\sigma}_1^2$, where $\hat{\sigma}_0^2 = 124.38$, $\hat{\sigma}_1^2 = 35.2304$
ARI	Model-based Estimate	Empirical SE		Type 3 Tests of Fixed Effects				
Effect	Estimate	SE		Num DF	Den DF	Model-based F Value	Empirical F Value	$\mathbf{R} = \mathbf{A}\mathbf{R}1$ where $\hat{\mathbf{C}}^{(\text{ARI})}$ is
Intercept	16.7895	2.9017	3.0368	Effect				1.0000 0.8042 0.6467 0.5201 0.4183 0.3364
day 1	1.6842	3.1298	1.6274	day	2	36	1.05	0.8042 1.0000 0.8042 0.6467 0.5201 0.4183
day 2	1.7368	2.4390	2.4243	time	1	18	0.03	0.6467 0.8042 0.6467 0.5201 0.4183 0.3364
time	-1.2105	1.8158	1.2990	day × time	2	36	0.02	0.4183 0.5201 0.6467 0.8042 1.0000 0.8042
day × time 1	1.6842	2.6325	1.7771	day × time 2	2	36	0.44	0.3364 0.4183 0.5201 0.6467 0.8042 1.0000
day × time 2	2.6671	1.9879						
UN	Model-based Estimate	Empirical SE		Type 3 Tests of Fixed Effects				
Effect	Estimate	SE		Num DF	Den DF	Model-based F Value	Empirical F Value	$\mathbf{R} = \mathbf{U}\mathbf{N}$ where $\hat{\mathbf{C}}^{(\text{UN})}$ is
Intercept	16.7895	3.1200	3.0368	Effect				1.0000 0.8525 0.8655 0.7820 0.7301 0.7953
day 1	1.6842	1.6720	1.6274	day	2	36	1.26	0.8525 1.0000 0.8821 0.8660 0.7401 0.8466
day 2	1.7368	2.4907	2.4243	time	1	18	0.02	0.8655 0.8821 1.0000 0.8333 0.7307 0.7442
time	-1.2105	1.3346	1.2990	day × time	2	36	0.80	0.7820 0.8660 0.8333 1.0000 0.6006 0.6802
day × time 1	1.6842	1.8218	1.7771	day × time 2	2	36	0.84	0.7301 0.7401 0.7307 0.6006 1.0000 0.9041
day × time 2	2.4211	2.0424	1.9879					0.7953 0.8466 0.7442 0.6802 0.9041 1.0000

*All F statistics are nonsignificant with $P > .05$.

†Denominator DFs (i.e., DDFM) were determined using MIXED's "between-within" option, which is the default method used when the program code uses a repeated statement. See Section 26.4.7 for further discussion of DDFM.

‡All three correlation matrices are (6×6) because each subject is observed six times.

26.4.3 A Random Intercept Model for the SF Data

The simplest random-effects model involving Subjects uses a single random effect associated with the intercept. The subject-specific scalar form for this model can be written as follows:

$$SF_{ij} = \beta_0 + b_{i0} + \beta_1 D_{ij1} + \beta_2 D_{ij2} + \beta_3 T_{ij} + \beta_4 (D_{ij1} T_{ij}) + \beta_5 (D_{ij2} T_{ij}) + E_{ij} \quad (26.10)$$

where D_{ij1} , D_{ij2} and T_{ij} are the values of variables D_1 , D_2 , and T , respectively, for the j th observation on the i th subject as defined previously in the marginal model (26.8). Model (26.10) contains two random components, E_{ij} and b_{i0} . As for the marginal model, E_{ij} denotes a random error term. The new term, b_{i0} , denotes a random effect for the (random) Subjects factor, which gives a random intercept ($\beta_0 + b_{i0}$) for subject i . The typical distributional assumptions made about E_{ij} and b_{i0} are that they are each normally distributed as $N(0, \sigma_e^2)$ and $N(0, \sigma_0^2)$, respectively, and that b_{i0} and E_{ij} are mutually independent for all i, j .

The subject-specific matrix version of (26.10) can be written as follows:

$$\mathbf{SF}_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}_i + \mathbf{E}_i \quad (26.11)$$

where $\mathbf{SF}_i(6 \times 1)$, $\mathbf{X}(6 \times 6)$, $\boldsymbol{\beta}(6 \times 1)$, and $\mathbf{E}_i(6 \times 1)$ are defined as for the marginal model (26.9) and where

$$\mathbf{Z}(6 \times 1) = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{b}_i(1 \times 1) = b_{i0}$$

Table 26.8 provides edited computer output based on fitting the random intercept model (26.10) or, equivalently, (26.11) together with the previous output shown in Table 26.7 for a compound symmetric ($\mathbf{R} = \mathbf{CS}$) marginal model fit to these same data. The SAS PROC MIXED code for fitting this model is described in Appendix C.

From Column 1 of Table 26.8, we can see that the estimated regression coefficients, as well as both model-based and empirical estimated standard errors, are identical for both choices of covariance structures.

From Column 2, we also see that corresponding model-based and empirical F statistics differ *within* each separate model, yet both models produce identical sets of F statistics. These results are as expected, since we have previously shown that the general covariance formula

$$\mathbf{V}_i = \mathbf{Z}_i \mathbf{G} \mathbf{Z}'_i + \mathbf{R}_i$$

yields the same overall covariance structure for \mathbf{V}_i when the \mathbf{G} and \mathbf{R} matrices are either

$$\{\mathbf{G} = \sigma_0^2, \mathbf{R} = \sigma_e^2 \mathbf{I}\} \quad \text{or} \quad \{\mathbf{G} = \mathbf{0}, \mathbf{R} = \mathbf{CS}\}$$

TABLE 26.8 Edited output based on modeling SF data using predictors Day and Time comparing random intercept model (26.10) with marginal CS model using REML in SAS's MIXED procedure

		Column 1						Column 2*†						Column 3									
$\mathbf{G} = \sigma^2_{\text{Z}}, \mathbf{R} = \sigma^2_{\text{e6}}$		Model-based Estimate		Empirical SE		Type 3 Tests of Fixed Effects						Type 3 Tests of Fixed Effects						Type 3 Tests of Fixed Effects					
Effect	Estimate	SE	Effect	DF	Den	Model-based	Empirical	F Value	Effect	DF	Den	Model-based	Empirical	F Value	Effect	DF	Den	Model-based	Empirical	F Value			
Intercept	16.7895	2.8984	3.0368						day	2	36	2.74	1.33		0.7792	1.0000	0.7792	0.7792	0.7792	0.7792			
day 1	1.6842	1.9257	1.6274						time	1	18	0.02	0.02		0.7792	0.7792	0.7792	0.7792	0.7792	0.7792			
day 2	1.7368	1.9257	2.4243						day × time	2	36	0.42	0.84		0.7792	0.7792	0.7792	0.7792	0.7792	0.7792			
time	-1.2105	1.9257	1.2990						day × time 1	1.6842	2.7934	1.7771									1.0000		
day × time 1	1.6842	2.7934	1.7771						day × time 2	2.4211	2.7234	1.9879										0.7792	
day × time 2	2.4211	2.7234	1.9879																				

$\mathbf{G} = \mathbf{0}, \mathbf{R} = \mathbf{CS}$	Model-based Estimate						Type 3 Tests of Fixed Effects						Type 3 Tests of Fixed Effects										
Effect	Estimate	SE	Effect	DF	Den	Model-based	Empirical	F Value	Effect	DF	Den	Model-based	Empirical	F Value	Effect	DF	Den	Model-based	Empirical	F Value			
Intercept	16.7895	2.8984	3.0368						day	2	36	2.74	1.33		0.7792	1.0000	0.7792	0.7792	0.7792	0.7792			
day 1	1.6842	1.9257	1.6274						time	1	18	0.02	0.02		0.7792	0.7792	0.7792	0.7792	0.7792	0.7792			
day 2	1.7368	1.9257	2.4243						day × time	2	36	0.42	0.84		0.7792	0.7792	0.7792	0.7792	0.7792	0.7792			
time	-1.2105	1.9257	1.2990						day × time 1	1.6842	2.7934	1.7771											
day × time 1	1.6842	2.7934	1.7771						day × time 2	2.4211	2.7234	1.9879											
day × time 2	2.4211	2.7234	1.9879																				

*All F statistics are nonsignificant with $P > .05$.

†Denominator DF's (i.e., DDFM) for random intercept model (26.11) were determined using MIXED's "contain" option, which is the default method used when the program code uses a random statement. In contrast, the DDFMs for the marginal CS model were determined using MIXED's "between-within" option, which is the default method used when the program code uses a repeated statement. See Section 26.4.7 for further discussion about DDFM.

$$\hat{\sigma}^2 = \hat{\sigma}_0^2 + \hat{\sigma}_e^2, \text{ where } \hat{\sigma}_0^2 = 124.38, \hat{\sigma}_e^2 = 35.2304$$

$$\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R} \text{ where } \hat{\mathbf{C}}_V \text{ is}$$

$$\mathbf{V} = \sigma^2 \hat{\mathbf{C}}^{(\text{EXCH})} \text{ where } \hat{\mathbf{C}}^{(\text{EXCH})} \text{ is}$$

The estimated working correlation structures for the two models are seen to be identical in Column 3 of Table 26.8.

The denominator degrees of freedom (i.e., den DF or DDFM) shown in Column 2 are also identical for both models even though two different (default) methods were used: “contain” for the random intercept model versus “between-within” for the marginal model. We have previously mentioned that SAS’s MIXED procedure allows the user to choose among several options for DDFM. We provide further discussion on these different DDFM options in Section 26.4.7.

For the SF posture measurement data, the investigator was not specifically interested in identifying individual subject heterogeneity but rather was focused on assessing the effects of the fixed factors Day and Time. Consequently, output pertaining to the subject-specific random-effect estimate (\hat{b}_{ij}) for each of the 19 study subjects was not obtained.

The results from using the random intercept model indicate that neither the main effects of Day or Time nor the interaction effects of Day with Time were statistically significant. This finding supported the investigators’ interest in using posture measurements that were not influenced by either the day of the week or the time of day when attempting to evaluate the extent to which the postures of computer operators are associated with muscular-skeletal disorders in a subsequent study.

26.4.4 Random-Effects 3-Way ANOVA Model for the SF Data

A more complex fixed-effects model involving Subjects allows for the random-effect interactions with the two fixed-effect predictor variables Day and Time. This model can be written in subject-specific scalar form as follows:

$$SF_{ij} = \beta_0 + b_{i0} + \beta_1 D_{ij1} + \beta_2 D_{ij2} + \beta_3 T_{ij} + \beta_4 (D_{ij1} T_{ij}) + \beta_5 (D_{ij2} T_{ij}) + b_{i1} D_{ij1} + b_{i2} D_{ij2} + b_{i3} T_{ij} + E_{ij} \quad (26.12)$$

where D_{ij1} , D_{ij2} , and T_{ij} are the values of variables D_1 , D_2 , and T , respectively, for the j th observation on the i th subject as defined in previous models (26.8) and (26.10). Model (26.12), however, now contains random effects b_{i0} , b_{i1} , b_{i2} , and b_{i3} in addition to the random error term E_{ij} . As in model (26.10), b_{i0} denotes a random intercept effect for the (random) Subjects factor. The random components b_{i1} and b_{i2} denote two random effects that reflect the interaction of Subjects with Day, and the random component b_{i3} denotes a fourth random effect that reflects the interaction of Subjects with Time. The typical distributional assumptions made about the E_{ij} , b_{i0} , b_{i1} , b_{i2} , and b_{i3} are that they are each normally distributed as $N(0, \sigma_e^2)$, $N(0, \sigma_0^2)$, $N(0, \sigma_{D1}^2)$, $N(0, \sigma_{D2}^2)$, and $N(0, \sigma_T^2)$, respectively, and that they are mutually independent for all i, j . Moreover, since the random effects b_{i1} and b_{i2} concern the two dummy variables (D_{ij1} and D_{ij2}) for the three-level nominal variable Day, it is typically assumed (as in SAS’s MIXED procedure) that the corresponding variance components σ_{D1}^2 and σ_{D2}^2 are equal; that is, $\sigma_{D1}^2 = \sigma_{D2}^2 \equiv \sigma_D^2$, say.

The subject-specific matrix version of (26.12) can be written as follows:

$$\mathbf{SF}_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}_i + \mathbf{E}_i \quad (26.13)$$

where $\mathbf{SF}_i(6 \times 1)$, $\mathbf{X}(6 \times 6)$, $\boldsymbol{\beta}(6 \times 1)$, and $\mathbf{E}_i(6 \times 1)$ are defined as in models (26.9) and (26.11) and where

$$\mathbf{Z}(6 \times 4) = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{b}_i(4 \times 1) = \begin{bmatrix} b_{i0} \\ b_{i1} \\ b_{i2} \\ b_{i3} \end{bmatrix}$$

Table 26.9 provides edited computer output based on fitting model (26.12) or, equivalently, (26.13) together with the previous output shown in Table 26.8 for a random intercept model (26.10) involving these same data.

In Column 1 of Table 26.9, the set of model-based estimated standard errors for model (26.12) containing four random effects is seen to be different from the corresponding set of model-based estimated standard errors for the random intercept model (26.10). This result is as expected, since the addition of the three random effects in model (26.12) changes the \mathbf{G} matrix and corresponding overall covariance structure \mathbf{V} .

The difference in correlation structures between model (26.12) and model (26.10) is also reflected in Column 2 of Table 26.9 by the different sets of model-based F -statistic values—that is, {1.78, 0.02, 0.79} versus {2.74, 0.02, 0.42}.

From Column 3 of Table 26.9, the estimated correlation structure for model (26.12) is not “exchangeable,” as is the correlation structure for the random intercept model (26.10).

We may wonder why regression model (26.12), which contains four random effects, is being considered at all for this analysis. The answer is that such a model makes sense if we consider the analysis problem as involving the three predictors Day, Time, and Subjects, each of which explains to some extent the total variability involved in all 114 observations on the 19 subjects in the study. In other words, we may wish to consider model (26.12) as equivalent to a *three-way ANOVA model*, with the factor Subjects treated as random and the factors Day and Time treated as fixed.

By using such an ANOVA approach, we are effectively allowing for a correlation structure that reflects the partitioning of the total variability (as described in Chapters 17–20 for classical one- and two-way ANOVA data) into variance contributions for the three factors being considered. The ANOVA approach for the analysis of repeated measures data is described in the next section. The reader should realize, nevertheless, that the maximum likelihood (ML) approach for fitting a linear mixed model is preferred, since it includes the ANOVA approach as a special case and also applies to a much larger variety of correlated data analysis problems.

26.4.5 Repeated Measures ANOVA

The ANOVA approach for the analysis of correlated/repeated measures data typically involves the following steps:

- 1. Specify the ANOVA model;** make sure to distinguish fixed factors from random factors and to state the distributional (i.e., normality and independence) assumptions for the random factors (including defining the relevant variance components).

TABLE 26.9 Edited output based on modeling SF data using predictors Day and Time and comparing mixed model (26.12) with four random effects with random intercept model (26.10), using REML in SAS's MIXED procedure

				Column 1						Column 2*†						Column 3					
$\mathbf{G}(\mathbf{4} \times \mathbf{4}), \mathbf{R} = \sigma_{\mathbf{e}}^2 \mathbf{I}$				Type 3 Tests of Fixed Effects						$\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$ where $\hat{\mathbf{C}}^{\mathbf{V}}$ is						$\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$ where $\hat{\mathbf{C}}^{\mathbf{V}}$ is					
Effect	Estimate	Model-based SE	Empirical SE	Effect	Num DF	Den DF	Model-based F Value	Empirical F Value	Effect	Num DF	Den DF	Model-based F Value	Empirical F Value	Effect	Num DF	Den DF	Model-based F Value	Empirical F Value			
Intercept	16.7895	2.8984	3.0368	Effect					Effect					Effect							
day	1	1.6842	1.9577	1.6274	3	2	36	1.78	day	2	36	1.78	1.33	day	2	36	0.02	0.7719			
day	2	1.7368	1.9577	2.4243	time	1	18	0.02	time	1	18	0.02	0.02	time	1	18	0.7467	0.7467			
time	-1.2105	1.5374	1.2990	day × time	1	1.6967	1.7771	0.79	day × time	2	36	0.79	0.84	day × time	2	36	0.7719	0.7719			
day × time	1	1.6842	1.9697	1.7771	day × time	2	1.9697	1.9879	day × time	2	36	0.79	0.84	day × time	2	36	0.8593	0.8593			
day × time	2	2.4211	1.9879																		1.0000
$\mathbf{G} = \sigma_0^2 \mathbf{I}, \mathbf{R} = \sigma_{\mathbf{e}}^2 \mathbf{I}$																					
Effect	Estimate	Model-based SE	Empirical SE	Effect	Num DF	Den DF	Model-based F Value	Empirical F Value	Effect	Num DF	Den DF	Model-based F Value	Empirical F Value	Effect	Num DF	Den DF	Model-based F Value	Empirical F Value			
Intercept	16.7895	2.8984	3.0368	Effect					day	2	36	2.74	1.33	day	2	36	0.02	0.7792			
day	1	1.6842	1.9257	1.6274	time	1	18	0.02	time	1	18	0.02	0.02	time	1	18	0.42	0.7792			
day	2	1.7368	1.9257	2.4243	day × time	2	36	0.84	day × time	2	36	0.84	0.84	day × time	2	36	0.7792	0.7792			
time	-1.2105	1.9257	1.2990	day × time	1	1.7771	1.9879		day × time	2	36	0.84	0.84	day × time	2	36	0.7792	0.7792			
day × time	1	1.6842	2.7234	2.7234	day × time	2	2.7234	2.7234	day × time	2	36	0.84	0.84	day × time	2	36	0.7792	0.7792			
day × time	2	2.4211	1.9879																		

*All F statistics are nonsignificant with $P > .05$.

†Denominator DF's (i.e., DDFM) for random intercept model (26.10) were determined using MIXED's "contain" option, which is the default method used when the program code uses a random statement. See Section 26.4.7 for further discussion about DDFM.

2. **Partition the total sums of squares** according to the various sources of variation, including subject-to-subject variation.
3. **Form the ANOVA table** that corresponds to the ANOVA model specified in Step 1 and that incorporates the partitioning of the total sums of squares delineated in Step 2.
4. **Determine the formulas for expected mean squares** in terms of fixed effects and variance component parameters.
5. **Determine the appropriate F statistic** for each test of hypothesis of interest as the ratio of the mean square for the factor being evaluated to the mean square for the source of variation that yields the same expected mean square as the numerator expected mean square under the null hypothesis.
6. **Carry out F tests of interest using a convenient computer program for ANOVA**; then **summarize your results**, using, as appropriate, the hypothesis-testing results, sample means, estimated standard errors for estimated effects of interest, and confidence intervals for contrasts of interest involving population means.

Classical linear regression procedures such as SAS's GLM procedure are well suited for carrying out balanced (and some unbalanced) repeated measures ANOVA procedures for continuous outcome data. Nevertheless, SAS's MIXED procedure, and its analogs in the other major statistical packages, uses an ML approach applied to a more general model than the (ANOVA) model used by GLM.

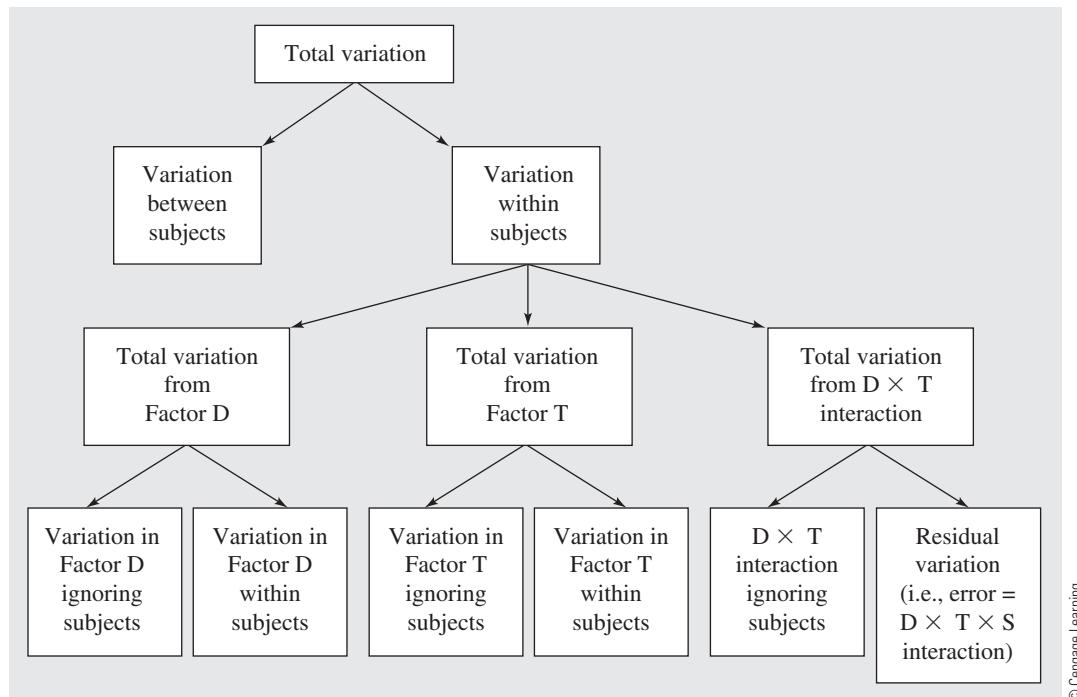
Returning to the posture measurement study involving (SF) data (introduced in Section 25.2), we can write the classical ANOVA effects version of this model as follows:

$$Y_{ijk} = \mu + b_i + \delta_j + \tau_k + (\delta\tau)_{jk} + b_{ij} + b_{ik} + E_{ijk} \quad (26.14)$$

where μ denotes the overall population mean, $\{\delta_j\}$ are the fixed effects of the Day factor, $\{\tau_k\}$ are the fixed effects of the Time factor, $\{(\delta\tau)_{jk}\}$ are the fixed interaction effects of Day with Time, and $\{b_i\}$, $\{b_{ij}\}$, and $\{b_{ik}\}$ are the appropriate random effects due to Subjects and their interactions with Day and Time, respectively. The error term E_{ijk} is a three-way interaction term that is assumed to represent only random error.

For the above model, it is assumed that b_i , b_{ij} , b_{ik} , and E_{ijk} are mutually independent, that b_i is distributed as $N(0, \sigma_0^2)$, that b_{ij} is distributed as $N(0, \sigma_D^2)$, that b_{ik} is distributed as $N(0, \sigma_T^2)$, and that E_{ijk} is distributed as $N(0, \sigma_e^2)$. The terms σ_0^2 , σ_D^2 , and σ_T^2 are called the *variance components* of the model, and σ_e^2 is called the *error component*.

The total variability, in terms of sums of squares, associated with model (26.14) can be partitioned as shown in Figure 26.5. We first partition the total variation into between-subjects variation and within-subjects variation. We conduct a further partitioning by dividing the within-subjects variation into variations contributed separately by Factor D (Day), by Factor T (Time), and by their $(D \times T)$ interaction. This partitioning comes from the within-subjects box because the variation explained by the two factors and their interaction is derived from responses "within" subjects. A final partitioning is then carried out by



© Cengage Learning

FIGURE 26.5 Partitioning the total sums of squares for three-way ANOVA of the SF posture measurement data involving Day (Factor D), Time (Factor T), and Subjects (Factor S)

separately dividing the total variation of each factor and associated interactions into two sources: the variation in the factor (or interaction) “ignoring” subjects and the variation in the factor (or interaction) “within” subjects.

The boxes at the bottom of Figure 26.5 indicate that the “Variation in Factor D ignoring subjects” describes the main effect of Factor D (i.e., Day), whereas the “Variation in Factor D within subjects” describes the Subjects-by-Factor D interaction; a similar interpretation holds for Factor T. The residual variation (σ_e^2) is actually the variation due to the interaction of Factor D with Factor T within Subjects (i.e., $D \times T \times S$ interaction); but because there is only one observation per subject at each combination of the levels of Factor D and Factor T, we must assume that these three-factor interaction effects represent random error E .

From the above partitioning, Table 26.10 can be produced to summarize the repeated measures ANOVA based on model (26.14). A pure estimate of the error for each subject is not possible because each subject is observed only once at each (day, time) combination. Therefore, we must assume that the mean square for this three-way interaction estimates σ_e^2 .

From Table 26.10, we see that all three F tests are nonsignificant. The F statistics in this table are correct because the expected mean squares of the numerator and denominator for each F statistic both estimate the same quantity under the null hypothesis being tested; that is, the mean square terms in each F statistic represent two independent estimates of the same variance under the null hypothesis of interest.

TABLE 26.10 ANOVA table for repeated measures ANOVA of SF data in Table 25.5

Source	d.f.	MS	F
Between Subjects	18	781.51	
Within Subjects	95		
Day	2	96.56	1.78 ($P = .18$)
Subjects \times Day	36	54.39	
Time	1	0.71	0.02 ($P = .88$)
Subjects 3 Time	18	30.51	
Day \times Time	2	14.63	0.79 ($P = .50$)
Subjects \times Day \times Time (i.e., Error)	36	18.43	
Total (corrected)	113		

© Cengage Learning

For example, under the null hypothesis of no Day-by-Time interaction, $MS(Day \times Time)$ estimates σ_e^2 as does $MS(\text{Error})$. Similarly, under the null hypothesis of no main effect of Day, $MS(\text{Day})$ estimates the same quantity that is estimated by $MS(\text{Subjects} \times \text{Day})$. We can, therefore, conclude that neither Day nor Time, nor the Day \times Time interaction, is a statistically significant factor in predicting SF response for these data.

The repeated measures ANOVA results in Table 26.10 are identical to the model-based results in Table 26.9 previously obtained from SAS's MIXED procedure for the linear mixed model (26.12) containing random effects for Subjects, Subjects-by-Day, and Subjects-by-Time. Thus, the correlation structure chosen for model (26.12) is identical to the (unstated) correlation structure being assumed for the three-way repeated measures ANOVA analysis. Note, however, that the ANOVA analysis derived from partitioning sums of squares does not directly provide empirical standard errors, nor does it allow for the large variety of other choices for the correlation structure available with SAS's MIXED procedure.

26.4.6 Testing Hypotheses about Random Effects

In this section, we describe three methods for testing hypotheses about random effects—namely, assessing whether or not one or more variance components corresponding to random effects in a mixed linear model are equal to zero.

Method 1: Approximate LR test (Snijders and Bosker 1999)

Method 1 involves a likelihood ratio (LR) test that compares two models (one with and one without the random effects of interest) using a (large sample) *chi-square statistic with r d.f.*; r is the number of variances and covariances in the G matrix that are restricted to be equal to 0 under the null hypothesis that the variance components of the random effects in question are zero. The two-tailed approximate P-value for this test statistic must be halved to get the approximate one-tailed P-value, since variance components cannot be negative.

■ **Example 26.1** We consider a simple random intercept linear mixed model for the SF data set involving the single predictor variable Time (T):

$$SF_{ij} = \beta_0 + b_{i0} + \beta_1 T_{ij} + E_{ij}$$

where b_{i0} is $N(0, \sigma_0^2)$ and E_{ij} is $N(0, \sigma_e^2)$ and where b_{i0} and E_{ij} are mutually independent. The null and alternative hypotheses of interest are

$$H_0: \sigma_0^2 = 0 \quad \text{and} \quad H_A: \sigma_0^2 > 0$$

To carry out the approximate LR test, the above full (F) model would be compared to the following reduced (R) model:

$$Y_{ij} = \beta_0 + \beta_1 T_{ij} + E_{ij}$$

The LR statistic is then computed in the usual way as follows:

$$\text{LR} = -2 \ln L_R - (-2 \ln L_F)$$

which is approximately χ^2_1 (i.e., chi-square with 1 d.f.) under H_0 .

To justify that d.f. = $r = 1$, we note that the \mathbf{G} matrix for the full model equals the scalar σ_0^2 , so that $r = \{\text{number of variances and covariances equal to zero under } H_0\} = 1$. Also, the two-tailed approximate P -value must be halved, since only the one-sided alternative hypothesis $H_A: \sigma_0^2 > 0$ is relevant. ■

■ **Example 26.2** We now consider a linear mixed model containing a random intercept and a random slope. Here we will test for the significance of the variance component for the random slope; that is, the null hypothesis of interest is $H_0: \sigma_T^2 = 0$, and the alternative hypothesis is $H_A: \sigma_T^2 > 0$. The full (F) model is

$$Y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})T_{ij} + E_{ij}$$

where b_{i0} is $N(0, \sigma_0^2)$, b_{i1} is $N(0, \sigma_T^2)$, and E_{ij} is $N(0, \sigma_e^2)$ and where b_{i0} and b_{i1} are independent of E_{ij} but not necessarily of each other. ■

To carry out the approximate LR test for this model, we need to compare the above full (F) model with the following reduced (R) model:

$$Y_{ij} = (\beta_0 + b_{i0}) + \beta_1 T_{ij} + E_{ij}$$

The LR statistic has the following structure:

$$\text{LR} = -2 \ln L_R - (-2 \ln L_F) \sim \chi^2_2$$

under $H_0: \sigma_T^2 = 0$, where $\text{Var}(b_{i1}) = \sigma_T^2$.

To justify that d.f. = $r = 2$, the \mathbf{G} matrix is a (2×2) matrix

$$\mathbf{G} = \begin{bmatrix} \sigma_0^2 & \sigma_{0T} \\ \sigma_{0T} & \sigma_T^2 \end{bmatrix}$$

involving the three parameters σ_0^2 , σ_T^2 , and $\text{Cov}(b_{i0}, b_{i1}) = \sigma_{0T}$. Under the null hypothesis that $\sigma_T^2 = \text{Var}(b_{i1}) = 0$, it must also follow that $\sigma_{0T} = \text{Cov}(b_{i0}, b_{i1}) = 0$, so that there are actually two parameters, σ_T^2 and σ_{0T} , that must be set equal to zero under H_0 . Thus, $r = \{\text{number of variances and covariances equal to zero under } H_0\} = 2$. Also the two-tailed approximate P -value must be halved.

Method 2: Approximate Wald test

SAS's MIXED procedure provides point estimates, estimated standard errors, Wald test Z statistics, two-tailed P -values for each variance component, and the error variance for the mixed model under consideration. As with the approximate LR test described above, the two-tailed P -value for the Wald test statistic must be halved to get the approximate P -value for a one-tailed test.

In particular, for a variance component— σ_R^2 , say—the null and alternative hypotheses being tested are $H_0: \sigma_R^2 = 0$ and the one-sided alternative $H_A: \sigma_R^2 > 0$, respectively, and the corresponding test statistic is

$$Z = \frac{\hat{\sigma}_R^2}{S_{\hat{\sigma}_R^2}}$$

which is approximately $N(0, 1)$ for large samples under H_0 .

■ Example 26.3 We now return to the random-effects model (26.12) involving mutually independent random effects for Subjects, Subjects \times Day, and Subjects \times Time:

$$\begin{aligned} SF_{ij} = & \beta_0 + b_{i0} + \beta_1 D_{ij1} + \beta_2 D_{ij2} + \beta_3 T_{ij} + \beta_4 (D_{ij1} T_{ij}) + \beta_5 (D_{ij2} T_{ij}) + b_{i1} D_{ij1} \\ & + b_{i2} D_{ij2} + b_{i3} T_{ij} + E_{ij} \end{aligned}$$

where

- b_{i0} is distributed $N(0, \sigma_0^2)$,
- b_{i1} and b_{i2} are independently distributed $N(0, \sigma_D^2)$,
- b_{i3} is distributed $N(0, \sigma_T^2)$, and
- E_{ij} is distributed $N(0, \sigma_e^2)$

and where σ_0^2 , σ_D^2 , and σ_T^2 are the *variance components* associated with these random effects.

Table 26.11 provides computer output for estimated variance component parameters not previously shown in the output of Table 26.9 for this model. (This extra output can be obtained with SAS's MIXED procedure using the *covtest* option with the PROC statement.) Using the approximate Wald test procedure of method 2, this table indicates that the σ_T^2 (Subjects \times Time) variance component (i.e., Cov Parm = TIME in the output) is not significant (one-tailed $P = .2747/2 = .1374$) and, therefore, b_{i3} may be dropped from the model. For a reduced model without this random component b_{i3} , we could then determine whether or not the variance component for σ_D^2 (Subjects \times Day) is significant and, if not, reduce the model further to evaluate whether the random intercept for Subjects is important. ■

TABLE 26.11 Variance component information from SAS's MIXED procedure for model (26.12) fit to SF data

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > z	Wald Tests* for
Intercept (σ_0^2)	6.4667	119.1725	43.5088	2.74	.0062	σ_0^2 S
Day (σ_D^2)	0.9758	17.9834	6.7684	2.66	.0079	σ_D^2 S
Time (σ_T^2)	0.2185	4.0263	3.6859	1.09	.2747	σ_T^2 NS
Residual (σ_e^2)	1.0000	18.4279	4.3435	4.24	.0000	σ_e^2

*S = significant; NS = nonsignificant

Estimated Variance Components

© Cengage Learning

Method 3: Approximate mixture test (Verbeke and Molenberghs 2001)

The test statistic here is a “mixture” of two chi-square statistics. The two-tailed P -value is not halved when using this approximate mixture test. Generally, SAS’s MIXED procedure does not directly compute this test statistic, but it does provide the information necessary to carry out the mixture test. However, when comparing a random intercept only model to a model without a random intercept, as illustrated in Example 26.1 above for the approximate LR method (i.e., method 1 above), the mixture test is identical to the approximate LR test. (Correspondingly, the P -value for the mixture test equals the halved P -value for the approximate LR test.) We, therefore, illustrate the mixture test procedure using the second of the two examples previously considered for method 1 (the approximate LR test).

■ **Example 26.2 revisited** We again consider a linear mixed model for the SF data involving the single predictor Time and containing a random intercept and a random slope. We will test for the significance of the variance component for the random slope; that is, we will test $H_0: \sigma_T^2 = 0$ versus $H_A: \sigma_T^2 > 0$. The full (F) model is

$$Y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})T_{ij} + E_{ij}$$

where b_{i0} is $N(0, \sigma_0^2)$, b_{i1} is $N(0, \sigma_T^2)$, and E_{ij} is $N(0, \sigma_e^2)$ and where it is assumed that the \mathbf{R}_i matrix is $\sigma_e^2 \mathbf{I}_6$ and that the \mathbf{G} matrix is

$$\mathbf{G} = \begin{bmatrix} \sigma_0^2 & \sigma_{0T} \\ \sigma_{0T} & \sigma_T^2 \end{bmatrix}$$

As with method 1, there are two parameters, σ_1^2 and σ_{0T} , that must be set equal to zero under H_0 . To test H_0 using the mixture test, we again, as with method 1, compute the LR statistic that compares the above full (F) model with the following reduced (R) model:

$$Y_{ij} = (\beta_0 + b_{i0}) + \beta_1 T_{ij} + E_{ij}$$

The LR statistic thus has the following formula:

$$\text{LR} = -2 \ln L_R - (-2 \ln L_F)$$

where L_R and L_F denote the maximized likelihoods for the reduced and full models, respectively. The mixture test then assumes this LR statistic has approximately the following “mixture” distribution:

$$\text{LR} \sim 0.5 \chi_2^2 + 0.5 \chi_1^2$$

where χ_2^2 and χ_1^2 denote chi-square random variables with 2 and 1 d.f., respectively. Essentially, this “mixture” distribution reflects the viewpoint that the appropriate d.f. for LR should be some average of 2 d.f. (corresponding to two parameters, σ_T^2 and σ_{0T} , that need to be set equal to zero under H_0) and 1 d.f. (corresponding to the one parameter, σ_T^2 , that is the primary parameter of interest).

The P -value for this “mixture” test is given by the expression

$$P = .5 \Pr(\chi_2^2 > \text{LR}) + .5 \Pr(\chi_1^2 > \text{LR})$$

For example, if LR is computed to be, say, 4.6, then

$$P = .5 \Pr(\chi_2^2 > 4.6) + .5 \Pr(\chi_1^2 > 4.6) \approx .5(0.10) + .5(0.032) = .066$$

which is nonsignificant at the 5% level of significance. ■

26.4.7 Denominator Degrees of Freedom in SAS's MIXED Procedure

We previously indicated in the footnotes to Tables 26.8 and 26.9 that SAS's MIXED procedure provides several options for determining the denominator degrees of freedom (DDFM) associated with F statistics for testing hypotheses about the predictors in the linear mixed model. Four of these options are

1. Residual: default for independence correlation structure
2. BetWithin: default for marginal model; that is, $\mathbf{G} = \mathbf{0}$
3. Contain: default for variance components (VC) model; that is, $\mathbf{G} \neq \mathbf{0}$
4. Satterthwaite: needed when denominator MS in F statistic is a linear function of MS terms

Note that SAS (MIXED procedure) offers greater flexibility for DDFM choices relative to other statistical packages; SPSS (MIXED) offers only the Satterthwaite method, R (LME) offers the containment method, and STATA (XTMIXED) provides no DDFM choices.

Table 26.12 presents, for the SF posture measurement data, the DDFM values and corresponding F statistics that result from the different DDFM options when both model-based and empirical standard errors are used with different correlation structures. The following items attempt to summarize the results in Table 26.12:

1. When an independent (**IND**) correlation structure is chosen, SAS's MIXED procedure only allows the residual option to be used with model-based standard errors. The corresponding F statistics, {0.60, 0.00, 0.09}, are computed using

TABLE 26.12 Denominator degrees of freedom (DDFM) and *F* test results for different DDFM options for the SF posture data for using the predictors Day (D), Time (T), and Day × Time interaction (DT)

		Correlation Structure-Specific DDFM Values					<i>F</i> statistics				
		IND	CS	AR1	VC1	VC3	IND	CS	AR1	VC1	VC3
DDFM Option with Model-based SE	Residual	D	108	108	108	108	0.60	2.74	1.05	2.74	1.78
		T	108	108	108	108	0.00	0.02	0.03	0.02	0.02
		DT	108	108	108	108	0.09	0.42	0.44	0.42	0.79
	BetWithin	D	36	36	36	36		2.74	1.05	2.74	1.78
		T	NA-RES	18	18	18	NA-RES	0.02	0.03	0.02	0.02
		DT	36	36	36	36		0.42	0.44	0.42	0.79
	Contain	D	108	108	90	36		2.74	1.05	2.74	1.78
		T	NA-RES	108	108	90	NA-RES	0.02	0.03	0.02	0.02
		DT	108	108	90	36		0.42	0.44	0.42	0.79
DDFM Option with Empirical SE	Satterthwaite	D	90	97.5	90	36		2.74	1.05	2.74	1.78
		T	NA-RES	90	99.8	90	NA-RES	0.02	0.03	0.03	0.02
		DT	90	82.7	90	36		0.42	0.44	0.42	0.79
	Residual	D	108	108	108	108		1.33	1.33	1.33	1.33
		T	NA	108	108	108	NA	0.02	0.02	0.02	0.02
		DT	108	108	108	108		0.84	0.84	0.84	0.84
	BetWithin	D	36	36	36	36	NA	1.33	1.33	1.33	1.33
		T	NA	18	18	18	NA	0.02	0.02	0.02	0.02
		DT	36	36	36	36		0.84	0.84	0.84	0.84
	Contain	D	108	108	90	36		1.33	1.33	1.33	1.33
		T	NA	108	108	90	NA	0.02	0.02	0.02	0.02
		DT	108	108	90	36		0.84	0.84	0.84	0.02
	Satterthwaite	D						1.33	1.33	1.33	1.33
		T	NA	NA-BW	NA-BW	NA-CO	NA-CO	NA	0.02	0.02	0.02
		DT						0.84	0.84	0.84	0.84

IND: independent, **CS:** compound symmetric, **AR1:** autoregressive

VC1: random intercept model, **VC3:** 3-way repeated measures ANOVA model

NA: not applicable

NA-RES: uses Residual DDFM

NA-BW: uses BetWithin DDFM

NA-CO: uses Contain DDFM

© Cengage Learning

a two-way fixed-effects ANOVA (as described in Chapter 19), where the two factors are Day (D) and Time (T) and where there are 19 replications of the SF response for each of the six ($= 3 \times 2$) combinations of Day with Time (thus giving $18 \times 6 = 108$ DDFM).

2. Using the residual option, the DDFM values equal 108 for D, T, and DT, regardless of the choice of correlation structure; the value 108 can be determined using the standard formula for residual DDFM—namely, $(n - p^*)$, where $n = 114$ (total number of observations) and $p^* = 6$ (number of fixed-effect model parameters) for model (26.12).
3. For a given choice of (non-IND) correlation structure (CS, AR1, VC1, VC3) using model-based standard errors, the same set of F statistics for D, T, and DT is obtained, regardless of which DDFM option is used. The set of F statistics obtained always corresponds to the three-way repeated measures ANOVA (**VC3**) model (26.12). However, as expected, the set of F statistics differs for different correlation structures—for example, {2.74, 0.02, 0.42} for **CS**, {1.05, 0.03, 0.44} for **AR1**, and {1.78, 0.02, 0.79} for **VC3**.
4. When empirical standard errors are used, the same set of F statistics—namely, {1.33, 0.02, 0.84}—is obtained, regardless of the DDFM option and regardless of the (non-IND) correlation structure chosen.
5. The BetWithin DDFM option always yields the same set of DDFM values for D, T, and DT (namely, 36, 18, and 36), regardless of the (non-IND) correlation structure chosen. This set of DDFM values corresponds to the DDFM values obtained for the three-way repeated measures ANOVA model (26.12).
6. For a random intercept (**VC1**) model, the Contain (default) option, using either model-based or empirical standard errors, yields DDFM values of 90, 90, and 90 for D, T, and DT, respectively. The value 90 is obtained using the formula $n - p^* - q^*(K - 1)$, where $n = 114$ observations, $p^* = 6$ fixed-effect parameters, $q^* = 1$ random effect, and $K = 19$ clusters.
7. For the three-way repeated measures ANOVA (**VC3**) model (26.12), the Contain (default) option, using either model-based or empirical standard errors, yields DDFM values of 36, 18, 36 for D, T, and DT, respectively.
8. When model-based standard errors are used, the Satterthwaite option yields DDFM values identical to those for the Contain option for the random-effects models **VC1** ({90, 90, 90}) and **VC3** ({36, 18, 36}). When empirical standard errors are used, specifying the Satterthwaite option in the code allows only the default option to be used (BW for marginal models and CO for random-effects models).

The above summary describes a complicated pattern of results corresponding to different DDFM options and different choices for the working correlation structure. One problematic feature of these results (from summary items 3 and 4 above) indicates that, for a given correlation structure, the set of F statistics does not change even when the DDFM values change, regardless of whether or not model-based or empirical standard errors are used. This is somewhat problematic because conclusions about statistical significance can vary, depending on the choice of the DDFM value used.

The above discussion indicates that it is difficult to provide firm recommendations on how to choose the appropriate DDFM option when analyzing continuous correlated data.

Perhaps the simplest recommendation is to let the user/investigator decide based on the study under consideration. Nevertheless, we offer the following suggestions, recognizing that the reader may wish to pursue this topic further before accepting these suggestions:

- I. If a marginal model is used (i.e., $\mathbf{G} = \mathbf{0}$), choose the Residual DDFM option, regardless of the \mathbf{R} matrix chosen.
- II. If a linear mixed model with random effects ($\mathbf{G} \neq \mathbf{0}$) is used and all the effects in the model other than the cluster effects are treated either as all fixed or as all random factors, choose the Contain DDFM option, regardless of the number of random effects in the model and regardless of whether or not an \mathbf{R} matrix other than $\sigma_e^2 \mathbf{I}$ is chosen.
- III. Use the Satterthwaite option only if the variables in the model other than cluster effects are a mixture of fixed and random effects (e.g., if Day is a fixed factor and Time is a random factor, or vice versa, for the SF example).

We recommend the Residual DDFM option for marginal models ($\mathbf{G} = \mathbf{0}$) because the DDFM reflects only the difference between the total number of observations and the number of fixed-effect model parameters when there are no random effects to consider.

The Contain DDFM option is recommended for linear mixed models with random effects to reflect the partitioning of sums of squares corresponding to the random effects chosen. Thus, for example, if only a random effect for the intercept is used to model the SF data, the DDFM values will each be 90 for testing the D, T, and DT effects; in contrast, if the three-way repeated measures ANOVA model (26.12) is used, the DDFMs will be 36, 18, and 36 for assessing the D, T, and DT effects.

We recommend the Satterthwaite option only for situations in which the denominator mean square is a linear function of mean square terms for two or more predictor variables, at least one of which is a fixed factor and at least one of which is a random factor. In such cases, the DDFM is determined by a complex formula that is readily computed by SAS's MIXED procedure. The use of this formula often, in fact, leads to DDFM values that are not integer values. For the balanced SF data set considered in this chapter, however, the two factors Day and Time are both fixed factors, so that the use of the Satterthwaite option is not necessary.

26.4.8 Summary of Analyses for the SF Data

From the analyses of the SF posture measurement data set that we have carried out, we summarize the results with regard to the primary question of whether or not there are statistically significant effects of Day, Time, and/or (Day \times Time) interaction:

1. Based on the sample means provided in Chapter 25, Table 25.5, there appears to be a possible main effect of the factor Day and some suggestion of a (Day \times Time) interaction effect but very little indication of a main effect of Time.
2. Regardless of the correlation structure chosen (for marginal or random-effects models) and regardless of whether model-based or empirical standard errors were used, there were no significant effects of Day, Time, and (Day \times Time) interaction in this sample of computer operators. These results support the investigators' contention that, in their subsequent study of the possible relationship of posture

measurements to muscular-skeletal disorders, they will not need to control for the effects of day of the week or time of day.

3. Regardless of the DDFM option used and the correlation structure chosen, the set of F -statistic values obtained for the effects of Day, Time and (Day \times Time) interaction was the same as the set obtained for a three-way repeated measures ANOVA, where Day and Time are fixed crossover factors and Subjects is a random factor.
4. Although the same conclusions were found, regardless of correlation structure chosen, we suggest that the most appropriate correlation structure for these data derives from the three-way repeated measures ANOVA model given by (26.14) or, equivalently, by (26.12) and (26.13). This model allows for a correlation structure that cannot be specified for any choice of \mathbf{R} matrix using a marginal model ($\mathbf{G} = \mathbf{0}$). Moreover, the correlation structure for model (26.14) makes sense, since it allows for random effects for Subjects, Subjects \times Day, and Subjects \times Time, in addition to fixed effects for Day, Time, and Day \times Time, thus reflecting all possible main effects and two-way interaction effects for the three factors Day, Time, and Subjects.

26.5 Recommendations about Choice of Correlation Structure

The choice of a working correlation structure is a very important part of the process when analyzing continuous response correlated data, especially since an investigator has to make a hopefully educated guess about the correlation structure and thus needs to be concerned that an inappropriate choice might be made. As we discussed in Chapter 25, Section 25.3, the choice process can involve several approaches. Here is a list of several options that may be considered when attempting to specify a working correlation structure:

1. Use what makes sense clinically/biologically.

This option should certainly be considered when analyzing any correlated data set, although the choice of correlation structure may not be obvious, especially when the study design has both longitudinal and cross-sectional features (e.g., when obtaining repeated measures on family members at several different times). Possible choices are *exchangeable* (e.g., when clusters are households studied cross-sectionally) and *autoregressive* (e.g., when clusters are individual subjects who are measured over time).

2. Try several different correlation structures.

Even if one has a logical choice for the correlation structure based on clinical/biological features of the study, if the same overall study conclusions are obtained for a variety of different plausible correlation structures, one can be more confident that intracluster correlations have been taken into account appropriately in the data analysis.

3. Use an unstructured correlation matrix.

This choice makes sense if it is believed that the correlation structure is quite complicated and there is little support for a simpler (e.g., more patterned) structure. Nevertheless, choosing

the *unstructured* option is actually “structured” to provide separate estimates for all possible pairwise correlations. As a result, the number of parameters to be estimated is often too large for the data to support a valid and stable numerical solution (e.g., the model-fitting algorithm sometimes does not converge to a solution when the *unstructured* option is chosen).

4. Start with an unstructured correlation matrix and simplify.

The estimated correlations based on using an *unstructured* correlation matrix may suggest a more simplified (or patterned) structure (e.g., if all estimated correlations are numerically close in value, this might suggest an exchangeable correlation structure).

5. Start with a Toeplitz (i.e., stationary) structure and simplify.

A *Toeplitz* (i.e., *stationary*) m -dependent correlation structure, which we have not previously described, has the property that correlations k units apart are the same for $k = 1, 2, \dots, m$, while correlations more than m units apart are zero. For example, a correlation structure that is *stationary 1-dependent* (i.e., $m = 1$) has the following property:

$$\text{corr}(Y_{ij}, Y_{i(j+1)}) = \rho_1 \quad \text{whereas} \quad \text{corr}(Y_{ij}, Y_{i(j+m')}) = 0 \quad \text{for } m' > 1$$

A subject with $n_i = 4$ repeated observations would, therefore, have a *stationary 1-dependent* structure of the following form:

$$\mathbf{C}(\rho_1) = \begin{bmatrix} 1 & \rho_1 & 0 & 0 \\ \rho_1 & 1 & \rho_1 & 0 \\ 0 & \rho_1 & 1 & \rho_1 \\ 0 & 0 & \rho_1 & 1 \end{bmatrix}$$

As another example, a correlation structure that is *stationary 2-dependent* (i.e., $m = 2$) has the following property:

$$\text{corr}(Y_{ij}, Y_{i(j+1)}) = \rho_1, \text{corr}(Y_{ij}, Y_{i(j+2)}) = \rho_2 \text{ whereas } \text{corr}(Y_{ij}, Y_{i(j+m')}) = 0 \text{ for } m' > 2$$

A subject with $n_i = 5$ repeated observations would, therefore, have a *stationary 2-dependent* structure of the following form:

$$\mathbf{C}(\rho_1, \rho_2) = \begin{bmatrix} 1 & \rho_1 & \rho_2 & 0 & 0 \\ \rho_1 & 1 & \rho_1 & \rho_2 & 0 \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ 1 & \rho_2 & \rho_1 & 1 & \rho_1 \\ 1 & 1 & \rho_2 & \rho_1 & 1 \end{bmatrix}$$

Toeplitz (i.e., *stationary*) is an alternative to **AR1** that allows for correlations to be different as occasions are farther apart in time, yet does not follow the **AR1** structure of ρ, ρ^2, ρ^3 , and so on. If a mixed model analysis does not yield stable numerical results when an unstructured correlation matrix is used, then a **Toeplitz** structure using the maximum value of m possible for the number of repeated measures observed on a given subject is a reasonable substitute choice. For example, if a balanced design involves six repeated measures per subject, then a *Toeplitz 5-dependent* correlation structure can serve as a starting structure for possible simplification.

6. Use goodness of fit (GOF) procedures AIC and/or BIC.

Akaike's Information Criterion (AIC) and Bayes' Information Criterion (BIC) are two goodness of fit (GOF) measures that are functions of the likelihood function adjusted for the number of fixed-effect parameters in the fitted linear mixed model. The smaller the value of AIC and/or BIC is, the better the model is said to fit. Consequently, either or both of these criteria can be used to decide on the appropriate correlation structure using the following process. For each correlation structure chosen, rank-order the AIC and/or BIC values from smallest to largest. Choose the correlation structure that yields the smallest value of AIC and/or BIC, or choose a structure that makes etiologic sense and also yields AIC or BIC values close to the smallest values observed.

7. Use the estimated correlation matrix of the Y 's as a fixed (numerical) structure.

The main idea here is to let the data suggest a fixed correlation structure to use in a marginal model. However, using a correlation structure derived from the Y 's does not account for the fixed effects in the model (i.e., a more appropriate approach would estimate the correlation structure using residual error values after accounting for fixed effects).

8. Start with an independence correlation structure, fit the model, obtain residuals, and then use the correlation matrix of the residuals as a fixed (numerical) structure.

This approach estimates correlations using residual error values as described in item 7. It effectively gives the starting correlation structure used for estimating an unstructured correlation matrix.

9. Use a random-effects model.

This approach is appropriate for accounting for heterogeneity of subject-specific effects, estimating subject-specific effects, defining a correlation structure not clearly specified by an **R** matrix in a marginal model, and modeling the effects of predictor variables that are considered to be random factors.

10. For any of the above options, decide between using model-based and empirical standard errors.

A safe (i.e., conservative) choice is to always use empirical standard errors, assuming that one can never be certain as to what the appropriate working correlation structure is. However, if one can make a strong case for a specific correlation structure, then standard errors based on that specific correlation structure may provide more efficient statistical inferences (i.e., more powerful hypothesis tests and narrower confidence intervals).

26.6 Analysis of Data for Discrete Outcomes

To address situations where the (repeated measures) response is binary, a count variable, or continuous but not normally distributed, a method involving the use of *generalized estimating equations* (GEE) has been developed (Zeger and Liang 1986). The GEE approach uses a procedure called *quasi-likelihood estimation*, which is a generalization of ML estimation. In recent years, the GEE approach has become very popular among researchers who

need to consider nonlinear regression models for correlated binary or count response data (e.g., the logistic model discussed in Chapters 22 and 23 and the Poisson regression model discussed in Chapter 24). Some currently available software can apply the GEE method, such as SAS's GENMOD and GLIMMIX procedures. The GENMOD procedure does not allow random effects, whereas the GLIMMIX procedure allows both fixed and random effects. SAS's NL MIXED procedure is an approximate likelihood-based procedure (i.e., it does not use GEE) that is also available. A thorough discussion of GEE methods is beyond the scope of this text. For further information, see Kleinbaum and Klein (2010) and/or Diggle et al. (2002).

Problems

1. The data set for this problem derives from the posture measurement study described in the main body of this chapter. Here we consider the data on shoulder flexion (SF) for 19 subjects that were each observed by 2 different raters on each of the 3 days (Monday, Wednesday, and Friday) and at 2 time periods (AM and PM) during each day. Note that our previous version of this example considered only one rater, but now there are two raters who have independently taken the posture measurements. Thus, we now consider 12 observations to have been made on each subject, with each observation corresponding to one of the 12 combinations of 3 days, 2 times, and 2 raters. In this problem, *we assume that the investigator is interested only in assessing whether the SF measurements vary significantly by day of measurement*. For this purpose, we have provided in the table below the average of the 4 SF scores taken on each subject (at 2 times and by 2 raters) for each of the 3 days.

Average SF score by day of week

Subject	Monday	Wednesday	Friday
1	10.50	7.78	6.00
2	4.00	2.50	16.75
3	25.75	28.00	22.75
4	20.00	22.50	20.00
5	4.75	8.25	6.50
6	17.25	22.00	18.00
7	24.00	9.25	12.50
8	45.50	50.00	41.75
9	10.00	10.00	5.50
10	28.25	31.25	41.25
11	21.75	22.25	15.50
12	23.00	29.00	26.00
13	10.50	7.00	7.00
14	22.00	19.50	22.50
15	7.25	5.00	0.00
16	19.50	9.00	9.00
17	23.25	26.50	17.50
18	28.50	35.25	8.75
19	3.00	1.75	2.50
Mean of Averages:	18.36	18.25	15.78

Program Statements for SAS's MIXED Procedure for Problem 1

```

proc mixed data= mnsf covtest;
  class subj day;
  model mnsf = day/s;
  random intercept/ subject=subj;
run;
proc mixed data= mnsf covtest empirical;
  class subj day;
  model mnsf = day/s;
  random intercept/ subject=subj;
run;
proc mixed data= mnsf covtest empirical;
  class subj day;
  model mnsf = day/s;
  repeated day/ subject=subj type=ar(1) rcorr;
run;

```

Edited SAS Output (PROC MIXED) for Problem 1

COVARIANCE PARAMETER ESTIMATES					
Cov Parm	Subject	Estimate	Std Error	Z Value	Pr > Z
Intercept	subj	113.90	40.8944	2.79	0.0027
Residual		25.8839	6.1009	4.24	<.0001
TYPE 3 TESTS OF FIXED EFFECTS (MODEL-BASED)					
Effect		Num DF	Den DF	F Value	Pr > F
day		2	36	1.5646	0.2231
TYPE 3 TESTS OF FIXED EFFECTS (EMPIRICAL – RANDOM INTERCEPT)					
Effect		Num DF	Den DF	F Value	Pr > F
day		2	36	1.19	0.3172
TYPE 3 TESTS OF FIXED EFFECTS (EMPIRICAL – AR(1))					
Effect		Num DF	Den DF	F Value	Pr > F
day		2	36	1.19	0.3172

- a. Assuming that the only important effect is that of Day (i.e., the factors Time and Rater are assumed *not* to have important effects), state the subject-specific scalar form of a random intercept model for analyzing the above data. In stating this model, make sure to describe the the assumptions made on the random effects (including the error term) in the model.

- b. State the null hypothesis that there is no significant effect of the factor Day in terms of a statement about parameters in your model given in part (a).
 - c. Based on a comparison of averages at the bottom of the table, does there appear to be a meaningful effect of the factor Day? Explain.
 - d. Describe how the data in the above table need to be reorganized in order for the MIXED procedure to carry out the analysis. (*Hint:* Use the format given in Table 25.2 of Chapter 25.)
 - e. Based on the computer output provided below, is there a significant effect of the factor Day? Explain by specifying the F statistic, its degrees of freedom, and its P -value appropriate for these data.
 - f. Based on the computer output, compute the estimate of the (exchangeable) correlation assumed by the model. (*Hint:* The output provides estimates of the variance of the subject-specific random intercept effect and the variance of the error term; you need to combine this information to compute the correlation coefficient estimate.)
 - g. Use the output to test whether there is a significant random effect for Subjects. If such a test was nonsignificant, why might you be concerned regarding the model that has been fit, and how might you redo the analysis?
2. The analysis described in Problem 1 for the posture measurement data on Shoulder Flexion (SF) may be criticized because information is lost when the 4 observations for a given subject on a given day are combined into an average score rather than being treated individually in the analysis. The data set shown below allows for an analysis that considers all 12 observations per subject. In analyzing this data set, we assume that the 4 observations for a given subject on a given day are true *replicates* (i.e., we assume that neither the time of day observed nor the rater used, factors that combine to provide the 4 observations for a given subject, is an important factor for predicting SF response).

Shoulder flexion scores by day of week

Subject	Sample Size	Monday				Wednesday				Friday			
1	12	15	5	17	5	11	9	10	1	8	10	5	1
2	12	0	8	1	7	-2	1	7	4	18	18	19	12
3	12	29	16	36	22	25	26	31	30	13	26	27	25
4	12	17	18	21	24	14	24	22	30	16	20	20	24
5	12	6	17	10	14	8	8	7	10	9	4	8	5
6	12	18	18	15	18	16	24	22	26	20	1	2	17
7	12	12	7	19	10	10	7	12	8	7	17	12	14
8	12	41	49	46	48	50	52	46	52	30	44	41	52
9	12	8	12	9	11	7	11	7	15	1	8	5	8
10	12	27	34	31	31	36	22	39	28	42	45	38	40
11	12	17	27	17	26	20	18	27	24	17	10	19	16
12	12	23	20	25	24	20	32	29	5	19	30	24	31
13	12	5	15	8	14	6	6	7	9	9	4	6	9
14	12	25	18	23	18	15	18	21	24	16	22	27	25
15	12	5	10	7	7	0	1	7	12	2	-2	0	0
16	12	20	17	22	19	5	11	7	13	3	13	6	14
17	12	28	21	24	20	27	20	32	27	16	15	22	17
18	12	33	25	28	28	38	27	42	34	5	14	9	7
19	12	2	4	1	5	2	5	0	0	2	0	6	2

Averages:

18.36

18.25

15.78

- a. How should the subject-specific scalar model for Problem 1 be modified to consider the data layout below? (*Hint:* You will need to add a second random effect to the model that reflects the interaction of Day with Subjects.)
- b. Use the computer information provided below to test whether there is a significant main effect of the factor Day. (*Hint:* See the ANOVA table provided below.) What do you conclude?
- c. Use the computer output provided below to test whether there is a significant interaction effect between Subjects and Day. If such a test is nonsignificant, why might you be concerned regarding the model that has been fit, and how might you redo the analysis?

Program Statements for SAS's GLM Procedure for Problem 2

```

proc mixed data=sf2 covtest;
  class subj day;
  model sf= day/s;
  random intercept day/subject=subj g;
run;
proc mixed data=sf2 covtest empirical;
  class subj day;
  model sf= day/s;
  random intercept day/subject=subj g;
run;
proc mixed data=sf2 covtest empirical;
  class subj day;
  model sf= day/s;
  repeated day /subject=subj g;
run;

```

Edited SAS Output (PROC MIXED) for Problem 2

COVARIANCE PARAMETER ESTIMATES					
Cov Parm	Subject	Estimate	Std Error	Z Value	Pr > Z
Intercept	Subj	113.75	40.5659	2.80	0.0025
day	subj	17.9537	5.5630	3.23	0.0006
Residual		22.0482	2.3845	9.25	<.0001

FIT STATISTICS					
-2 Res Log Likelihood			1475.1		

(continued)

TYPE 3 TESTS OF FIXED EFFECTS (MODEL-BASED)				
Effect	Num DF	Den DF	F Value	Pr > F
day	2	36	1.95	0.1571

TYPE 3 TESTS OF FIXED EFFECTS (EMPIRICAL)				
Effect	Num DF	Den DF	F Value	Pr > F
day	2	36	1.84	0.1730

3. A study by Holder, Plikaytis, and Carbone (1996) compared two laboratory protocols designed to measure antibody levels in an enzyme-linked immunosorbent assay (ELISA) for *Streptococcus pneumoniae*. One protocol incorporates a “blocking step” that is thought to increase the specificity of the assay, maximizing the yield of the specific antibody; the other protocol does not use the blocking step. The data shown below provide the ELISA results on six specimens (i.e., samples), each with *Streptococcus pneumoniae* Serogroup 4. This is a balanced repeated measures ANOVA design with six measurements on each sample, three of which use the protocol with the blocking step and three of which do not use the blocking step.

Antibody yields for two protocols measuring

Streptococcus pneumoniae serogroup 4

Specimen (or Sample) Number

		1	2	3	4	5	6	Sample Mean
Blocking Step	Yes	7.3	1.0	2.8	2.3	28.4	0.4	7.66
	9.3	1.0	3.1	3.2	29.2	0.5		
	6.4	1.0	2.8	2.8	35.9	0.4		
	No	9.3	1.5	3.6	3.7	27.2	0.6	7.67
		9.1	1.4	3.4	3.6	25.5	0.6	
		9.3	2.0	4.5	3.6	28.3	0.8	

- Should the factor Blocking Step be considered a fixed or random factor? Explain.
- State the subject-specific scalar regression model for this analysis. (*Hint:* For these data, the subjects are the samples.)
- State the subject-specific matrix model for this analysis.
- State the null hypothesis that there is no significant effect of Blocking Step in terms of a statement about the parameters in the regression model given in part (b).
- Based on a comparison of sample means, does there appear to be a meaningful effect of the Blocking Step factor? Explain.
- Below are given computer results for a repeated measures analysis of these data. Based on these results, is there a significant effect of the Blocking Step factor? Explain by specifying the *F* statistic, its degrees of freedom, and its *P*-value appropriate for these data.

Program Statements for SAS's MIXED Procedure for Problem 3

```

proc mixed data=serogroup4 covtest;
  class block sample;
  model result= block/s;
  random intercept block/subject=sample g;
run;
proc mixed data=serogroup4 covtest empirical;
  class block sample;
  model result= block/s;
  random intercept block/subject=sample g;
run;
proc mixed data=serogroup4 covtest;
  class block sample;
  model result= block/s;
  random intercept/subject=sample g;
run;
proc mixed data=serogroup4 covtest empirical;
  class block sample;
  model result= block/s;
  random intercept/subject=sample g;
run;
proc mixed data=serogroup4 covtest empirical;
  class block sample;
  model result= block/s;
  repeated block/subject=sample type=ar(1);
run;

```

Edited SAS Output (PROC MIXED) for Problem 3

COVARIANCE PARAMETER ESTIMATES (RANDOM INTERCEPT BLOCK)					
Cov Parm	Subject	Estimate	Std Error	Z Value	Pr > Z
Intercept	sample	116.64	74.4682	1.57	0.0586
block	sample	1.5818	1.3957	1.13	0.1285
Residual		1.8225	0.5261	3.64	0.0003

TYPE 3 TESTS OF FIXED EFFECTS (MODEL-BASED)				
Effect	Num DF	Den DF	F Value	Pr > F
block	1	5	0.00	0.9901

TYPE 3 TESTS OF FIXED EFFECTS (EMPIRICAL)				
Effect	Num DF	Den DF	F Value	Pr > F
block	1	29	0.00	0.9838

(continued)

COVARIANCE PARAMETER ESTIMATES (RANDOM INTERCEPT ONLY)					
Cov Parm	Subject	Estimate	Std Error	Z Value	Pr > Z
Intercept	sample	117.30	74.4651	1.58	0.0576
Residual		2.6407	0.6935	3.81	<.0001

TYPE 3 TESTS OF FIXED EFFECTS (MODEL-BASED – RANDOM INTERCEPT ONLY)				
Effect	Num DF	Den DF	F Value	Pr > F
Block	1	29	0.00	0.9838

TYPE 3 TESTS OF FIXED EFFECTS (EMPIRICAL – RANDOM INTERCEPT ONLY)				
Effect	Num DF	Den DF	F Value	Pr > F
block	1	29	0.00	0.9887

TYPE 3 TESTS OF FIXED EFFECTS (EMPIRICAL – REPEATED AR1)				
Effect	Num DF	Den DF	F Value	Pr > F
block	1	5	0.13	0.7321

- g. Use the computer output to test whether there is a significant interaction effect between Specimen (or Samples) and Blocking Step. If such a test is nonsignificant, why might you be concerned about the model that has been fit, and how might you redo the analysis?
4. A study by Heffner, Drawbaugh, and Zigmond (1974) investigated the effects of an amphetamine on the behavior of rats. Before the study began, 24 “thirsty” rats were trained to press a lever to obtain water. The rats were categorized into three groups (slow, medium, and fast) of equal size according to their initial press rates. Each rat received three doses of the drug (i.e., the amphetamine under study), as well as a placebo, on separate occasions and in random order. One hour after the drug injection, an experimental session began in which the rat received water after pressing the lever a prespecified number of times. Half of the rats received water after two presses of the lever, whereas the other half received water after five presses.

The response measured was the lever press rate (i.e., LPR, the total number of lever presses divided by elapsed time in seconds) used by a thirsty rat to press the lever and receive water. The primary research question was whether the drug affected the LPR. Also of interest was the question of whether there was an effect on the response (LPR) of the number of presses (PRS) required to obtain water (two versus five) and/or the initial press rate (IPR) (slow, medium, or fast) and whether there were any interaction effects. The data are given below.

- a. Which of the factors Drug, IPR, and PRS should be treated as fixed factors and which as random factors? Explain.
- b. Consider an analysis of these data that focuses only on the effect of the factor Drug (i.e., ignore the factors IPR and PRS in the analysis). Thus, for each of the 24 rats, we have four responses corresponding to the administration of each of the

four levels of the factor Drug. State a subject-specific scalar model for analyzing these data, making sure to state the assumptions typically made about the random effects (including the error term) in the model. (*Note:* The subjects are the rats in this example.)

- c. State the subject-specific matrix model that corresponds to the subject-specific scalar model stated in part (b).
- d. Based on a comparison of sample means, does there appear to be a meaningful effect of the factor Drug? Explain.
- e. Use the computer output provided below to test for the significance of the factor Drug. Make sure to describe the null hypothesis being tested, the *F* statistic used, its degrees of freedom, and the resulting *P*-value for this test. What do you conclude?
- f. Use the computer output below to test for the significance of the subject (i.e., Rat) factor. Why might you expect the results for this test to be significant?
- g. Explain why it is not possible to test whether there is a significant interaction between Drug and Rat.

Lever press rate (LPR) for 24 thirsty rats

IPR	PRS	Rat Number	Drug				Mean
			1 (Placebo)	2	3	4	
Slow	2	1	0.81	0.80	0.82	0.50	0.76
		2	0.77	0.78	0.79	0.51	
		3	0.80	0.82	0.83	0.52	
		4	0.95	0.95	0.91	0.60	
	5	5	2.18	2.44	1.92	0.92	1.80
		6	2.02	2.20	1.75	0.82	
		7	2.06	2.28	1.86	0.80	
		8	2.28	2.46	1.90	0.90	
Medium	2	9	1.03	1.13	1.04	0.82	0.98
		10	0.96	0.93	1.02	0.63	
		11	0.98	1.00	0.98	0.74	
		12	1.17	1.20	1.18	0.91	
	5	13	2.62	2.58	2.21	1.03	2.09
		14	2.60	2.60	2.34	1.14	
		15	2.39	2.41	2.09	0.90	
		16	2.70	2.64	2.23	1.02	
Fast	2	17	1.20	1.24	1.27	0.96	1.20
		18	1.25	1.23	1.30	1.01	
		19	1.23	1.20	1.18	0.95	
		20	1.31	1.42	1.41	1.08	
	5	21	2.98	2.64	2.34	1.28	2.34
		22	3.10	2.85	2.40	1.35	
		23	2.80	2.48	2.16	1.01	
		24	3.21	2.92	2.56	1.40	
			Mean	1.81	1.80	1.60	0.91

IPR = initial press rate; PRS = number of presses

Program Statements for SAS's MIXED Procedure for Problem 4

```

proc mixed data=rats covtest;
  class rat drug;
  model lpr = drug /s;
  random intercept /subject=rat g;
run;
proc mixed data=rats covtest empirical;
  class rat drug;
  model lpr = drug /s;
  random intercept /subject=rat g;
run;
proc mixed data=rats covtest empirical;
  class rat drug;
  model lpr = drug /s;
  repeated drug /subject=rat type=ar(1);
run;

```

Edited SAS Output (PROC MIXED) for Problem 4

COVARIANCE PARAMETER ESTIMATES					
Cov Parm	Subject	Estimate	Std Error	Z Value	Pr > Z
Intercept	rat	0.3419	0.1079	3.17	0.0008
Residual		0.09489	0.01616	5.87	<.0001
TYPE 3 TESTS OF FIXED EFFECTS (MODEL-BASED)					
Effect	Num DF	Den DF	F Value	Pr > F	
drug	3	69	45.72	<.0001	
TYPE 3 TESTS OF FIXED EFFECTS (EMPIRICAL – RANDOM INTERCEPT)					
Effect	Num DF	Den DF	F Value	Pr > F	
drug	3	69	47.63	<.0001	
TYPE 3 TESTS OF FIXED EFFECTS (EMPIRICAL – REPEATED AR1)					
Effect	Num DF	Den DF	F Value	Pr > F	
drug	3	69	47.63	<.0001	

5. Consider the same study by Heffner et al. (1974) and the data set described in Problem 4, in which the response measured was the lever press rate (LPR) used by a thirsty rat to press a lever and receive water. The three factors identified as possible predictors were Drug (three doses plus a placebo), IPR (initial press rate, categorized as slow, medium, or fast), and PRS (number of presses required to obtain water):

two or five). Another factor of importance is the random factor Rat (with 24 levels). Furthermore, the factors IPR and PRS were combined into a single Factor A with six levels according to the following categorization:

IPR	PRS	Level of Factor A
slow	2	1
slow	5	2
medium	2	3
medium	5	4
fast	2	5
fast	5	6

In this problem, we consider both Factor A and Drug together in a repeated measures analysis to evaluate whether there is an effect of either (or both) of the factors and their interaction on the response.

- a. Describe the data layout for this analysis using the format given by Table 25.2 of Chapter 25.
- b. Using the data provided in Problem 4 together with the above definition of Factor A, we obtained following table of sample means:

Factor A

Drug	1	2	3	4	5	6	Average
1 (Placebo)	0.8325	2.1350	1.0350	2.5775	1.2475	3.0225	1.8083
2	0.8375	2.3450	1.0650	2.5575	1.2725	2.7225	1.8000
3	0.8375	1.8575	1.0550	2.2175	1.2900	2.3650	1.6038
4	0.5325	0.8600	0.7750	1.0225	1.0000	1.2600	0.9083
Average	0.7600	1.7994	0.9825	2.0938	1.2025	2.3425	1.5301

Based on this table, describe whether there appears to be a meaningful main effect of Factor A, a main effect of Drug, and/or an interaction effect between Factor A and Drug. Explain.

- c. State the formula for a subject-specific scalar model for the analysis of these data, and specify the assumptions typically made about the random effects (and error term) for this model.
- d. For the model stated in part (c), use a flow diagram to describe how the variation (sums of squares) is partitioned into contributions from various sources of predictors. (*Hint:* Consider Figure 26.5 in the text.)
- e. Using the computer output provided below, carry out tests of whether the main effect of Factor A, the main effect of Drug, and the Factor A-by-Drug interaction are significant. For each test, state the null hypothesis being tested, the form of the F statistic, its degrees of freedom under the null hypothesis, and the resulting P -value. What do you conclude?

Program Statements for SAS's MIXED Procedure for Problem 5

```

proc mixed data=rats covtest;
  class rat factor a drug;
  model lpr= factor a drug factor*a*drug/s;
  random intercept/subject=rat(factor a) g;
run;
proc mixed data=rats covtest;
  class rat factor a drug;
  model lpr= factor a drug factor*a*drug/s;
  random intercept/subject=rat g;
run;
proc mixed data=rats covtest empirical;
  class rat factor a drug;
  model lpr= factor a drug factor*a*drug/s;
  random intercept/subject=rat(factor a) g;
run;
proc mixed data=rats covtest empirical;
  class rat factor a drug;
  model lpr= factor a drug factor*a*drug/s;
  repeated drug /subject=rat(factor a) type=ar(1);
run;

```

Edited SAS Output (PROC MIXED) for Problem 5 (*same output obtained for first two program codes*)

COVARIANCE PARAMETER ESTIMATES					
Cov Parm	Subject	Estimate	Std Error	Z Value	Pr > Z
Intercept	rat (factor a)	0.01108	0.003823	2.90	0.0019
Residual		0.001539	0.000296	5.20	<.0001

TYPE 3 TESTS OF FIXED EFFECTS (MODEL-BASED – RANDOM INTERCEPT)					
Effect	Num DF	Den DF	F Value	Pr > F	
factor a	5	18	143.06		<.0001
drug	3	54	2818.00		<.0001
factor a*drug	15	54	10261.5		<.0001

TYPE 3 TESTS OF FIXED EFFECTS (EMPIRICAL – RANDOM INTERCEPT)					
Effect	Num DF	Den DF	F Value	Pr > F	
factor a	5	18	227.13		<.0001
drug	3	54	3486.01		<.0001
factor a*drug	15	54	10261.5		<.0001

(continued)

TYPE 3 TESTS OF FIXED EFFECTS (EMPIRICAL – REPEATED AR1)				
Effect	Num DF	Den DF	F Value	Pr > F
factora	5	18	227.13	<.0001
drug	3	54	3486.01	<.0001
factora*drug	15	54	10261.5	<.0001

6. Consider the same study by Heffner et al. (1974) and the data set described in Problems 4 and 5, where the response measured was the lever press rate (LPR) taken by a thirsty rat to press a lever and receive water. The three factors identified as possible predictors were Drug (three doses plus a placebo), IPR (initial press rate, categorized as slow, medium, or fast), and PRS (number of presses required to obtain water: two or five). Another factor of importance is the random factor Rat (with 24 levels). In this problem, we consider the effects of IPR and PRS individually, as well as the effects of Drug and Rat and various interactions among these factors, in a repeated measures analysis.

The ANOVA model for this situation is a modification of the model used in Problem 5, except that the effect of Factor A is split into individual components. The structure for the subject-specific scalar version of this model involves the following fixed-effect parameters and random effects:

μ = overall mean

α_j = j th fixed effect of IPR, $j = 1, 2, 3$

β_k = k th fixed effect of PRS, $k = 1, 2$

γ_l = l th fixed effect of Drug, $l = 1, 2, 3, 4$

$(\alpha\beta)_{jk}$ = fixed interaction effect of the j th level of IPR with the k th level of PRS

$(\alpha\gamma)_{jl}$ = fixed interaction effect of the j th level of IPR with the l th level of Drug

$(\beta\gamma)_{kl}$ = fixed interaction effect of the k th level of PRS and the l th level of Drug

Program Statements for SAS's MIXED Procedure for Problem 6

```

proc mixed data=rats covtest;
  class ipr prs drug rat;
  model lpr = ipr prs drug ipr*prs ipr*drug
             prs*drug ipr*prs*drug/s;
  random intercept/subject=rat(ipr*prs) g;
run;
proc mixed data=rats covtest;
  class ipr prs drug rat;
  model lpr = ipr prs drug ipr*prs ipr*drug
             prs*drug ipr*prs*drug;
  random intercept/subject=rat g;
run;

```

Edited SAS Output (PROC MIXED) for Problem 6 (same output obtained for first two program codes)

COVARIANCE PARAMETER ESTIMATES					
Cov Parm	Subject	Estimate	Std Error	Z Value	Pr > Z
Intercept	rat (ipr*prs)	0.01108	0.003823	2.90	0.0019
Residual		0.001539	0.000296	5.20	<.0001
TYPE 3 TESTS OF FIXED EFFECTS (MODEL-BASED – RANDOM INTERCEPT)					
Effect	Num DF	Den DF	F Value	Pr > F	
ipr	2	18	42.40	<.0001	
prs	1	18	629.59	<.0001	
drug	3	54	2818.00	<.0001	
ipr*prs	2	18	0.47	0.6333	
ipr*drug	6	54	17.02	<.0001	
prs*drug	3	54	1319.73	<.0001	
ipr*prs*drug	6	54	22.94	<.0001	
TYPE 3 TESTS OF FIXED EFFECTS (EMPIRICAL – RANDOM INTERCEPT)					
Effect	Num DF	Den DF	F Value	Pr > F	
ipr	2	18	60.14	<.0001	
prs	1	18	839.45	<.0001	
drug	3	54	3486.01	<.0001	
ipr*prs	2	18	0.76	0.6333	
ipr*drug	6	54	47.89	<.0001	
prs*drug	3	54	1474.66	<.0001	
ipr*prs*drug	6	54	63.23	<.0001	

$(\alpha\beta\gamma)_{jkl}$ = three-way fixed interaction effect for the (j, k, l) combination of IPR, PRS, and Drug

$S_{i(jk)}$ = random effect of rat i within levels j and k of IPR and PRS, respectively

$E_{l(ljk)}$ = random error of the l th level of Drug within levels j and k of IPR and PRS, respectively, for rat i

- a. Using the computer output based on fitting the above model, carry out tests for main effects and interactions of each predictor in the model. (Assume that all such tests are orthogonal.) What do you conclude about whether or not the Drug factor has a significant effect on the response?
- b. Use the output to test whether there is a significant random effect of the factor Rat.

7. A study by Rikkers et al. (1978) involved a prospective randomized surgical trial that compared cirrhotic patients who had bled from use of either a nonselective shunt (a standard operation) or a selective shunt (a new operation). The response is a measure of the maximal rate of urea synthesis (MRUS), a measure of liver function; low values of MRUS indicate poor liver function. The study sample consisted of 8 selective shunt patients and 13 nonselective shunt patients. MRUS was measured both preoperatively and early postoperatively on each of the 21 patients. The purpose of the study was to compare the change in liver function in each of the two groups. The data are described below:

Pre- and postoperative maximal rate of urea synthesis level (MRUS) by group

Group	Subject Number	PreOp	PostOp
Selective Shunt	1	51	48
	2	35	55
	3	66	60
	4	40	35
	5	39	36
	6	46	43
	7	52	46
	8	42	54
Means		46.375	47.125
Subject Number	PreOp	PostOp	
Nonselective Shunt	9	34	16
	10	40	36
	11	34	16
	12	36	18
	13	38	32
	14	32	14
	15	44	20
	16	50	43
	17	60	45
	18	63	67
	19	50	36
	20	42	34
	21	43	32
Means		43.538	31.462

- Decide whether each of the factors Group (Selective Shunt versus Nonselective Shunt) and Time (PreOp versus PostOp) should be considered as a fixed or as a random factor. Explain.
- State a subject-specific scalar model for analyzing the above data, and specify the assumptions typically made about the random effects (including the error term) in the model.
- State the subject-specific matrix model that corresponds to the model you provided in your answer to part (b).
- State the null hypothesis that there is no significant effect of the factor Group in terms of a statement about parameters in your model given in part (b). What

other null hypothesis should be tested prior to testing whether there is a significant effect of Group?

- e. Based on a comparison of sample means, does there appear to be a meaningful difference in the change in liver function between the two groups? Explain.
- f. Below are given computer results for a repeated measures analysis of these data. Based on these results, is there a significant effect of the factor Group? Explain by describing the F statistic, its degrees of freedom, and its P -value appropriate for these data.
- g. Use the output to test whether there is a significant random effect for Subjects.

Program Statements for SAS's MIXED Procedure for Problem 7

```
proc mixed data=mrus covtest;
  class subj group time;
  model mrus = group time group*time;
  random intercept/subject=subj(group) g;
run;
proc mixed data=mrus covtest empirical;
  class subj group time;
  model mrus= group time group*time;
  random intercept/subject=subj(group) g;
run;
```

Edited SAS Output (PROC MIXED) for Problem 7

COVARIANCE PARAMETER ESTIMATES					
Cov Parm	Subject	Estimate	Std Error	Z Value	Pr > Z
Intercept	subj (group)	98.9155	38.3522	2.58	0.0050
Residual		35.8532	11.6323	3.08	0.0010
TYPE 3 TESTS OF FIXED EFFECTS (MODEL-BASED)					
Effect	Num DF	Den DF	F Value	Pr > F	
group	1	19	3.63	0.0721	
time	1	19	8.86	0.0078	
group*time	1	19	11.36	0.0032	
TYPE 3 TESTS OF FIXED EFFECTS (EMPIRICAL)					
Effect	Num DF	Den DF	F Value	Pr > F	
group	1	19	4.87	0.0399	
time	1	19	8.85	0.0078	
group*time	1	19	11.34	0.0032	

8. In this problem, we consider the analysis of the combined information from both raters on the shoulder flexion (SF) scores in the posture measurement study. Thus, the questions below concern the data on 19 subjects, where there are 12 observations on each subject, with each observation corresponding to one of the 12 combinations of 3 days, 2 times, and 2 raters. The design is a balanced repeated measures design.

The data are given below:

- a. Using the information provided in the data layout, supply the values of the sample means in the following tables:

	Rater1			Rater2			
	Monday	Wednesday	Friday	Monday	Wednesday	Friday	
AM	_____	_____	_____	_____	_____	_____	_____
PM	_____	_____	_____	_____	_____	_____	_____
	_____	_____	_____	_____	_____	_____	_____

© Cengage Learning

- b. Based on the results obtained for the tables in part (a), does there appear to be a meaningful main effect of either Day, Time, and/or Rater? Explain.
 c. Based on the results obtained for the tables in part (a), does there appear to be a meaningful interaction effect between Day and Time, between Day and Rater, and/or between Time and Rater? Explain.

Shoulder flexion scores by day, time, and rater

Subject	Sample Size	Monday				Wednesday				Friday			
		Rater 1		Rater 2		Rater 1		Rater 2		Rater 1		Rater 2	
		AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM
1	12	15	5	17	5	11	9	10	1	8	10	5	1
2	12	0	8	1	7	-2	1	7	4	18	18	19	12
3	12	29	16	36	22	25	26	31	30	13	26	27	25
4	12	17	18	21	24	14	24	22	30	16	20	20	24
5	12	6	17	10	14	8	8	7	10	9	4	8	5
6	12	18	18	15	18	16	24	22	26	20	14	21	17
7	12	12	7	19	10	10	7	12	8	7	17	12	14
8	12	41	49	46	48	50	52	46	52	30	44	41	52
9	12	8	12	9	11	7	11	7	15	1	8	5	8
10	12	27	34	31	31	36	22	39	28	42	45	38	40
11	12	17	27	17	26	20	18	27	24	17	10	19	16
12	12	23	20	25	24	20	32	29	35	19	30	24	31
13	12	5	15	8	14	6	6	7	9	9	4	6	9
14	12	25	18	23	18	15	18	21	24	16	22	27	25
15	12	5	10	7	7	0	1	7	12	2	-2	0	0
16	12	20	17	22	19	5	11	7	13	3	13	6	14
17	12	28	21	24	20	27	20	32	27	16	15	22	17
18	12	33	25	28	28	38	27	42	34	5	14	9	7
19	12	2	4	1	5	2	5	0	0	2	0	6	2
Averages:		17.4	18.0	19.0	18.5	16.2	17.0	19.7	18.5	13.3	15.7	15.6	16.8

- d. Based on the results obtained for the tables in part (a), does there appear to be a meaningful three-way interaction effect among Day, Time, and Rater? Explain.
- e. We will now consider the analysis of the SF data in which we have treated the Day, Time, and Rater factors all as fixed.
 - i. How can the investigators justify a decision to treat the Rater factor as fixed?

One set of analyses that was carried out for these data is described by the following computer code and edited output using SAS's MIXED procedure.

Program Statements for SAS's MIXED Procedure for Problem 8

```

proc mixed data = sf covtest;
  class subj day time rater;
  model sf = day time rater day*time day*rater time*rater
day*time*rater/s;
①    random intercept day time rater day*time day*rater time*rater/
subject=subj;
run;

proc mixed data = sf covtest empirical;
  class subj day time rater;
  model sf = day time rater day*time day*rater time*rater
day*time*rater/s;
②    random intercept day time rater day*time day*rater time*rater/
subject=subj;
run;

```

Edited SAS Output (PROC MIXED) for Problem 8

NOTE: Estimated G matrix is not positive definite

COVARIANCE PARAMETER ESTIMATES					
Cov Parm	Subject	Estimate	Std Error	Z Value	Pr > Z
Intercept	subj	115.70	41.3109	2.80	0.0025
day	subj	12.6223	5.6911	2.22	0.0133
time	subj	0.6983	2.5024	0.28	0.3901
rater	subj	0.4851	0.8433	0.58	0.2826
day*time	subj	14.4403	4.0274	3.59	0.0002
day*rater	subj	2.6646	1.3158	2.03	0.0214
time*rater	subj	0	.	.	.
Residual		5.0489	0.9717	5.20	<.0001

(continued)

(1)

TYPE 3 TESTS OF FIXED EFFECTS (MODEL-BASED)				
Effect	Num DF	Den DF	F Value	Pr > F
day	1.69	0.1986	1.69	0.1986
time	0.83	0.3739	0.83	0.3739
rater	18.23	0.0005	18.23	0.0005
day*time	0.39	0.6810	0.39	0.6810
day*rater	2.54	0.0931	2.54	0.0931
time*rater	5.70	0.0281	5.70	0.0281
day*time*rater	1.63	0.2108	1.63	0.2108

(2)

TYPE 3 TESTS OF FIXED EFFECTS (EMPIRICAL)				
Effect	Num DF	Den DF	F Value	Pr > F
day	2	36	1.17	0.3222
time	1	18	0.88	0.3613
rater	1	18	19.24	0.0004
day*time	2	36	0.35	0.7080
day*rater	2	36	2.60	0.0883
time*rater	1	18	8.00	0.0111
day*time*rater	2	36	1.84	0.1742

- ii. State the subject-specific scalar model that is being fit by each of the two program statements (① and ②). In what way do the program statements differ?

Program Statements for SAS's MIXED Procedure for Problem 8

```

(3) proc mixed data = sf covtest;
   class subj day time rater;
   model sf= day time rater day*time day*rater time*rater
   day*time*rater/s;
   random intercept /subject = subj;
run;

(4) proc mixed data = sf covtest empirical;
   class subj day time rater;
   model sf= day time rater day*time day*rater time*rater
   day*time*rater/s;
   random intercept/subject=subj;
run;

```

Edited SAS Output (PROC MIXED) for Problem 8 (continued)

COVARIANCE PARAMETER ESTIMATES					
Cov Parm	Subject	Estimate	Std Error	Z Value	Pr > Z
Intercept	subj	121.26	41.2644	2.94	0.0016
Residual		30.4270	3.0580	9.95	<.0001

(3)

TYPE 3 TESTS OF FIXED EFFECTS (MODEL-BASED)				
Effect	Num DF	Den DF	F Value	Pr > F
day	2	198	4.99	0.0077
time	1	198	1.04	0.3087
rater	1	198	7.96	0.0053
day*time	2	198	0.43	0.6492
day*rater	2	198	0.87	0.4224
time*rater	1	198	0.95	0.3320
day*time*rater	2	198	0.27	0.7638

(4)

TYPE 3 TESTS OF FIXED EFFECTS (EMPIRICAL)				
Effect	Num DF	Den DF	F Value	Pr > F
day	2	198	1.17	0.3128
time	1	198	0.88	0.3500
rater	1	198	19.24	<.0001
day*time	2	198	0.35	0.7061
day*rater	2	198	2.60	0.0770
time*rater	1	198	8.00	0.0051
day*time*rater	2	198	1.84	0.1623

- iii. Is the correlation structure assumed by the model stated in part (e.ii) an exchangeable correlation structure? Explain.
- iv. Assuming (perhaps incorrectly) that the F test for each effect is “orthogonal,” what would you conclude about which effects are significant and which effects are not significant?
- v. Based on the information provided by the “NOTE” and by the “Covariate Parameter Estimates” in the output, should you be concerned about the appropriateness of these analyses? Explain.

Another set of analyses that was carried out for these data is described by the following computer code and edited output using SAS’s MIXED procedure.

- vi. State the subject-specific scalar model that is being fit by each of the two program statements (③) and (④). In what way do these program statements differ?

- vii. Is the correlation structure assumed by the model stated in part (e,vi) an exchangeable correlation structure? Explain.
- viii. Assuming (perhaps incorrectly) that the F test for each effect is “orthogonal,” what would you conclude about which effects are significant and which effects are not significant?
- ix. Based on the information provided by the “Covariate Parameter Estimate,” in the output, should you be concerned about the appropriateness of these analyses? (Note: There is no “NOTE” statement output for these analyses.) Explain.
- f. We will now consider the analysis of the SF data in which we have treated the Day and Time factors as fixed and the Rater factor as random.
 - i. How can the investigators justify a decision to treat the Rater factor as random? Consider the following computer program code:

Edited SAS Output (PROC MIXED) for Problem 8 (continued)

```
(5) proc mixed data = sf covtest;
  class subj day time rater;
  model sf = day time day*time/s;
  random rater day*rater time*rater day*time*rater;
  random intercept day time rater day*time day*rater time*rater/
subject=subj;
run;
```

- ii. State the subject-specific scalar model that is being fit by the above program statements (5).
- iii. Is the correlation structure assumed by the model stated in part (f.ii) an exchangeable correlation structure? Explain.

The code described in (5) can be shown from the output to provide an inappropriate analysis. Another code and its corresponding output are given as follows:

Edited SAS Output (PROC MIXED) for Problem 8 (continued)

```
(6) proc mixed data=sf covtest;
  class subj day time rater;
  model sf= day time day*time;
  random rater/g;
run;
```

COVARIANCE PARAMETER ESTIMATES				
Cov Parm	Estimate	Std Error	Z Value	Pr > Z
Rater	0.8204	3.0073	0.27	0.3925
Residual	148.69	14.1452	10.51	<.0001

(continued)

(6)

TYPE 3 TESTS OF FIXED EFFECTS (MODEL-BASED)				
Effect	Num DF	Den DF	F Value	Pr > F
day	2	221	1.02	0.3620
time	1	221	0.21	0.6448
day*time	2	221	0.09	0.9152

- iv. State the subject-specific scalar model that is being fit by the above program code (6).
- v. Is the correlation structure assumed by the model stated in part (f.iv) an exchangeable correlation structure? Explain.
- vi. Assuming that the above analysis is appropriate, what can you conclude about whether there are significant effects of Day, Time, Day \times Time, and/or Rater? Explain.
- vii. What other approaches might you take to carry out the analysis when considering Day and Time as fixed effects and Rater as a random effect?

References

- Diggle, P. J.; Heagerty, P.; Liang, K. Y.; and Zeger, S. L. 2002. *Analysis of Longitudinal Data*, Second Edition. Oxford: Oxford University Press.
- Heffner, T. G.; Drawbaugh, R. B.; and Zigmond, M. J. 1974. "Amphetamine and Operant Behavior in Rats: Relationship between Drug Effect and Control Response Rate." *Journal of Comparative and Physiological Psychology*. 86(6): 1031–43.
- Holder, P.; Plikaytis, B. D.; and Carbone, G. 1996. "Need for Blocking to Reduce Non-specific Binding." Paper presented at WHO, Workshop on Pneumococcal ELISA Standardization, Atlanta, May 15–16.
- Kleinbaum, D. G., and Klein, M. 2010. *Logistic Regression: A Self-Learning Text*, Third Edition. New York and Berlin: Springer Publishers.
- Littell, R. C.; Milliken, G. A.; Stroup, W. W.; and Wolfinger, R. D. 1996. *SAS System for Mixed Models*. Cary, N.C.: SAS Institute.
- Ortiz, D. J.; Marcus, M.; Gerr, F.; Jones, W.; and Cohen, S. 1997. "Estimation of Within Subject Variability in Upper Extremity Posture Measured with Manual Goniometry among Video Display Terminal Users." *Applied Ergonomics* 28(2): 139–43.
- Rikkens, L. F.; Rudman, D.; Galambos, J. T.; Fulenwidr, J. T.; Millikan, W. J.; Kutner, M. H.; Smith, R. B.; Salam, A. A.; Sones, P. J.; and Warren, W. D. 1978. "A Randomized Controlled Trial of Distal Spenoral Shunt." *Annals of Surgery* 188: 271–82.
- Snijders, T., and Bosker, R. 1999. *Multilevel Analysis*. London: Sage Publishers.
- Verbeke, G., and Molenberghs, G. 2001. *Linear Mixed Models for Longitudinal Data*. New York and Berlin: Springer Publishers.
- Zeger, S. L., and Liang, K. Y. 1986. "Longitudinal Data Analysis for Discrete and Continuous Outcomes." *Biometrics* 42: 121–30.

27

Sample Size Planning for Linear and Logistic Regression and Analysis of Variance

27.1 Preview

Sample size planning is an essential component of study design. In this chapter, we present practical approaches for sample size determination for linear and logistic regression and for ANOVA models. Unfortunately, the mathematical theory for sample size calculations can be fairly complicated for some of these multivariable methods. As a result, in practice, researchers sometimes choose to depend on simple but error-prone approaches, such as approximate power and sample size charts (Pearson and Hartley 1951; Kutner, Neter, Nachtsheim, and Li 2004) and even rules of thumb (Tabachnick and Fidell 2001). Such shortcuts can be costly. In an undersized study, effects that are large enough to be scientifically important may turn out not to be statistically significant. In both undersized and oversized studies, resources are wasted.

With recently developed methods and continual improvements in computer software, rigorous sample size planning for regression methods is now much more feasible in everyday practice. In this chapter, we begin with a review of basic concepts and formulas relevant to sample size planning. We then present two approaches for sample size calculation for linear and logistic regression. The first is an approximate, but simple, approach that has been shown to yield fairly accurate sample sizes and for which even manual computation is feasible. The second is based on more traditional theory for sample size determination and is best implemented using computer software. For the latter approach, we present examples for linear regression and ANOVA using two popular programs, SAS, version 9.3 (SAS Institute 2002–2010), and PASS® 11 (Hintz 2011).

27.2 Review: Sample Size Calculations for Comparisons of Means and Proportions

Sample size formulas for hypothesis tests comparing the means of two normally distributed populations and comparing proportions in two populations are often covered in introductory statistical methods courses. Their presentation here serves two purposes. First, the review provides a convenient way to illustrate the main statistical concepts relevant to sample size planning. Second, these same approaches may be used for simple linear and logistic regression models and, with a slight adjustment, for approximate but reasonably accurate results in the case of multiple linear and logistic regression.

27.2.1 Sample Size Determination for Tests Comparing Two Means

The classical sample size formula for a test of $H_0: \mu_1 - \mu_2 = 0$ versus $H_A: \mu_1 - \mu_2 \neq 0$ when a random sample of size n is to be collected from each of two normally distributed populations with means μ_1 and μ_2 and known common variance σ^2 is as follows:

$$n \geq 2 \left[\frac{\sigma(Z_{1-\alpha/2} + Z_{1-\beta})}{\Delta} \right]^2 \quad (27.1)$$

For chosen values of the significance level (α), power ($1 - \beta$), and σ^2 , formula (27.1) provides the minimum sample size required to detect an absolute difference $|\mu_1 - \mu_2|$ of size at least equal to Δ between the two population means (i.e., to reject H_0 in favor of H_A when $|\mu_1 - \mu_2| \geq \Delta$, $\Delta > 0$).¹ The researcher, in addition to specifying the desired levels of α and β , must provide values for σ^2 and Δ . An educated guess about the value of σ^2 can sometimes be made by using information obtained from prior research. To specify Δ intelligently, the researcher has to decide on the smallest absolute population mean difference, $|\mu_1 - \mu_2|$, that is practically (as opposed to statistically) significant for the study.

- **Example 27.1** Researchers want to compare the average quantity of the active ingredient in doses of a new formulation of a popular pain reliever with the amount in the standard formulation. Independent random samples of each type of drug will be analyzed. They wish to test $H_0: \mu_1 - \mu_2 = 0$ versus $H_A: \mu_1 - \mu_2 \neq 0$, where μ_1 = true mean quantity of the ingredient (in mg), in a dose of the new formulation and μ_2 = true mean quantity in a dose of the standard formulation. They decide to use a significance level of 5% and want to have a power of 0.8 to detect an absolute difference in average quantity

¹ Formula 27.1 assumes equal sample sizes for both groups. However, it is easily adapted to the unequal sample size situation. If n_1 and n_2 are the two sample sizes to be determined and n_2 is a multiple k (a positive integer) of n_1 (i.e., $n_2 = kn_1$), the sample size formula for n_2 is

$$n_2 \geq (k+1) \left[\frac{\sigma(Z_{1-\alpha/2} + Z_{1-\beta})}{\Delta} \right]^2.$$

of ingredient of at least 1.5 mg. The standard deviation of the quantity of ingredient is thought to be no more than 2 mg.

The sample size, n , to be collected for each group is then given by

$$n \geq 2 \left[\frac{\sigma(Z_{1-\alpha/2} + Z_{1-\beta})}{\Delta} \right]^2 = 2 \left[\frac{(2)(Z_{0.975} + Z_{0.8})}{1.5} \right]^2 = 2 \left[\frac{(2)(1.96 + 0.842)}{1.5} \right]^2 = 27.9$$

At least 28 doses of each formulation of the drug need to be sampled. ■

27.2.2 Sample Size Determination for Tests Comparing Two Proportions

The sample size formula for a test of $H_0: \pi_1 - \pi_2 = 0$ versus $H_A: \pi_1 - \pi_2 \neq 0$ when large random samples of size n are to be collected from each of two Bernoulli (or point binomial) populations with proportions π_1 and π_2 is

$$n \geq \left[\frac{Z_{1-\alpha/2} \sqrt{2\bar{\pi}(1 - \bar{\pi})} + Z_{1-\beta} \sqrt{\pi_1(1 - \pi_1) + \pi_2(1 - \pi_2)}}{\Delta} \right]^2 \quad (27.2)$$

where $\bar{\pi} = \frac{\pi_1 + \pi_2}{2}$. For chosen values of α and β , formula (27.2) provides the minimum sample size required to detect an absolute difference between proportions π_1 and π_2 of a size at least equal to Δ (i.e., to reject H_0 in favor of H_A when $|\pi_1 - \pi_2| \geq \Delta$, $\Delta > 0$).² The researcher, in addition to specifying the desired values of α and β , must make reasoned guesses about the size of either π_1 or π_2 and specify the smallest absolute difference in population proportions, Δ , that would be of practical importance to detect. The formula uses a normal approximation to the binomial distribution and, therefore, is appropriate for large sample sizes.

■ **Example 27.2** Another way to investigate whether the drug formulations in Example 27.1 were similar would be to compare the proportions of doses for which the quantity of the active ingredient was within an acceptable range. One could then perform a test of $H_0: \pi_1 - \pi_2 = 0$ versus $H_A: \pi_1 - \pi_2 \neq 0$, where π_1 and π_2 are the true proportions of acceptable doses of the new and standard formulations, respectively. Suppose that previous studies of the standard formulation suggest that $\pi_1 = 0.90$ and that, if the proportion of acceptable doses for the new drug was 0.95, this would

² Formula (27.2) can also be adapted to the unequal sample size case, where n_2 is a multiple k (a positive integer) of n_1 (i.e., $n_2 = kn_1$):

$$n_1 \geq \left[\frac{Z_{1-\alpha/2} \sqrt{\bar{\pi}(1 - \bar{\pi}) \left(1 + \frac{1}{k}\right)} + Z_{1-\beta} \sqrt{\pi_1(1 - \pi_1) + \frac{\pi_2(1 - \pi_2)}{k}}}{\Delta} \right]^{-2} \quad \text{where } \bar{\pi} = \frac{(\pi_1 + k\pi_2)}{(1 + k)}$$

represent a difference that was important to detect. Then, for $\alpha = 0.05$ and $\beta = 0.20$, the sample size, n , for each type of drug is given by

$$\begin{aligned} n &\geq \left[\frac{(1.96)\sqrt{(2)(0.925)(1 - 0.925)} + (0.842)\sqrt{(0.90)(1 - 0.90) + (0.95)(1 - 0.95)}}{0.05} \right]^2 \\ &= 434.6 \end{aligned}$$

At least 435 doses of each formulation are needed. ■

27.2.3 Sample Size Planning: The Relationship among α , $(1 - \beta)$, Δ , σ^2 , and n

In real-world applications, sample size planning does not involve a single calculation. Instead, it is usually a process involving examination of the required sample sizes for various different combinations of the following quantities: α , the Type I error rate or significance level; $(1 - \beta)$ or power (β is the Type II error rate); Δ , the minimum effect size that would be of scientific significance and that, therefore, scientists would like to be able to detect; and σ^2 , the population variance.

For a fixed sample size, α and β are inversely related in the following sense. If one tries to guard against making a Type I error by choosing a small rejection region, the nonrejection region (and hence β) will be large. Conversely, protecting against a Type II error necessitates using a large rejection region, leading to a large value for α .

Increasing the sample size generally decreases β for any given value of α . For fixed α and β , the larger Δ is, the smaller the sample size required to detect that effect.

Typically, researchers will study tables and graphs of sample sizes for various combinations of β and Δ (α is usually fixed) and will select the sample sizes that result in high power for desirable combinations of these parameters and that are also practically feasible given resource constraints.

Researchers can also observe the variation in sample sizes for a range of values of σ^2 , the population variance, in order to understand the impact this parameter has on sample size; in practice, of course, the researcher generally does not have control over the population variance.

27.3 Sample Size Planning for Linear Regression

In this section, we present simple approaches to sample size calculations for linear regression where manual calculation is feasible and specialized software is not necessary. We begin with the simple linear regression model where the main independent variable of interest is binary and where the dependent variable is assumed to be normally distributed. The inference-making tool in this case is an equal population variances t test comparing two population means; for the large sample case, the sample size calculation will be approximately the same as discussed in Section 27.2.1. Next, we consider the simple linear regression model with a continuous covariate; in this setting, a test for zero slope of the straight line is mathematically

equivalent to a test for zero correlation between the dependent and independent variables; from this relationship, another simple sample size formula arises. For multiple regression with a single predictor of primary interest, and p other predictors mainly for control of confounding, it has been shown that the simple linear regression sample size formulas can be easily adjusted to give approximate but fairly accurate results.

27.3.1 Sample Size Determination for Simple Linear Regression with a Binary Predictor

Consider the simple linear regression model (5.3) in Chapter 5:

$$Y = \beta_0 + \beta_1 X + E$$

Let X be a binary independent variable indexing two groups or treatments, coded as 1 for one group and 0 for the other. The main inference of interest in this regression, performed to investigate whether or not independent variable X is a statistically significant predictor of Y , is the test for zero slope ($H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$), which was described in Section 5.7.

As described in Chapter 5, the simple linear regression model relates the mean value of Y to values of X . Therefore, the t test for the slope compares the average value of Y for two values of X , which are one unit apart. This is analogous to a two-sample t test (since Y is assumed to be normally distributed and to have constant variance σ^2 for all values of X). As such, the sample size formula (27.1) can be used for simple linear regression with a binary predictor. Although the formula is based on the standard normal distribution rather than the t distribution, for large sample sizes the approaches converge, and the sample size calculations are accurate (Kupper and Hafner 1989). Also, Δ , the effect size of interest, is replaced by $|\beta_1^*|$ ($\beta_1^* \neq 0$), the minimum absolute value of the slope for which it would be important to reject the null hypothesis of zero slope.

■ Example 27.3 Example 27.1 can be modeled with the following simple linear regression:

$$Y = \beta_0 + \beta_1 X + E$$

where

Y = quantity of active ingredient, in mg

X = 1 if new formulation of drug, 0 if standard formulation

To test whether the true average quantities of the active ingredient are different, a test of $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$ will be conducted. For $\alpha = 0.05$, $\sigma = 2.0$ mg, and a power of 0.8 to detect a slope value $|\beta_1|$ of at least 1.5 (corresponding to an absolute difference in average quantity of 1.5 mg), the sample size calculation is

$$n \geq 2 \left[\frac{(2.0)(Z_{0.975} + Z_{0.8})}{1.5} \right]^2 = 2 \left[\frac{(2.0)(1.96 + 0.842)}{1.5} \right]^2 = 27.9$$

At least 28 doses of each formulation of the drug need to be sampled. ■

27.3.2 Sample Size Determination for Simple Linear Regression with a Continuous Predictor

Consider again the simple linear regression model (5.3):

$$Y = \beta_0 + \beta_1 X + E$$

This time, let X be a normally distributed independent variable representing the continuous predictor of interest. The main inference of interest in this regression again involves a t test for zero slope. In Section 6.6.1, it was shown that this test is mathematically equivalent to the test of $H_0: \rho_{YX} = 0$ versus $H_A: \rho_{YX} \neq 0$, the test for the correlation between Y and X . For the latter test, it can be shown that the minimum sample size, n_s , necessary to reject the null hypothesis H_0 in favor of the alternative H_A when $\rho_{XY} = \rho$ ($\rho \neq 0$) with significance level α and power $1 - \beta$ is

$$n_s \geq \left[\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{C(\rho)} \right]^2 + 3 \quad (27.3)$$

where $C(\rho)$ is the Fisher's Z transformation discussed in Section 6.6.2—namely,

$$C(\rho) = \frac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right)$$

Example 27.4 Researchers are planning to investigate the association between APGAR scores taken five minutes after birth and infant blood pressure five hours after birth. They will use $\alpha = 0.05$ and $\beta = 0.20$ and want to detect a correlation of at least $\rho = 0.30$. Using formula (27.3), we can determine that (X, Y) data should be collected from at least 85 newborns:

$$n_s \geq \left[\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{C(\rho)} \right]^2 + 3 = \left[\frac{1.96 + 0.842}{\frac{1}{2} \ln\left(\frac{1+0.3}{1-0.3}\right)} \right]^2 + 3 = 85.0 \quad \blacksquare$$

27.3.3 Sample Size Determination for Multiple Linear Regression: A Simple Approach

Hsieh, Bloch, and Larsen (1998) have shown that sample sizes for multiple regression models where a single independent variable is primarily of interest and other predictors are included mainly for control of confounding can be easily approximated from the simple linear regression sample size formulas above. In particular, if n_s is the sample size for a simple linear regression of Y on X_1 , the sample size, n_m , for a multiple regression of Y regressed on X_1, \dots, X_k is approximately

$$n_m = \frac{n_s}{1 - \rho_{X_1(X_2, \dots, X_k)}^2} \quad (27.4)$$

where $\rho_{X_1(X_2, \dots, X_k)}^2$ is the chosen value of the population squared multiple correlation between the main independent variable of interest, X_1 , and the control variables, X_2, \dots, X_k . The basic idea is that the simple linear regression sample size is adjusted by multiplying by the variance inflation factor for the primary independent variable of interest. Hsieh et al. (1998) show the approximate calculation in formula (27.4) to be reasonably accurate for normally and nonnormally distributed independent variables.

■ **Example 27.5** Researchers want to conduct a study evaluating two types of surgery for nearsightedness. The outcome will be the patient's refractive error, measured in diopters, one year after surgery. Researchers want to control for the patient's age and baseline (i.e., presurgical) refractive error. They want to use $\alpha = 0.05$ and $\beta = 0.20$ and want to detect an absolute difference in average error of at least 1 diopter (i.e., $|\beta_1^*| \geq 1$). The variance, σ^2 , of the refractive error is believed to be at most 1, and the population squared multiple correlation between the dummy variable indexing surgery type and the control variables age and baseline refractive error is chosen to be no more than 0.25.

Assuming equal sample sizes for the two surgeries, the sample size per surgery-type group for the simple linear regression using the dummy variable for surgery type is

$$n \geq 2 \left[\frac{\sigma(Z_{1-\alpha/2} + Z_{1-\beta})}{|\beta_1^*|} \right]^2 = 2 \left[\frac{(1)(Z_{0.975} + Z_{0.8})}{1} \right]^2 = 2(1.96 + 0.842)^2 = 15.70$$

or 16 patients per type of surgery. Using the adjustment suggested by Hsieh et al. (1998), the total sample size for the multiple regression would be $32/(1 - 0.25)$ or 43 (i.e., 21 or 22 patients for each of the two surgery groups). ■

■ **Example 27.6** Consider again Example 27.4. Suppose that researchers want to control for mother's age and whether or not the birth was cesarean. They believe that the squared multiple correlation between the main independent variable, APGAR score, and the two control variables is 0.5. Using the simple linear regression sample size determined in Example 27.4 and formula (27.4), the minimum required sample size is $85/(1 - 0.5) = 170$. ■

27.4 Sample Size Planning for Logistic Regression

The theory behind power and sample size calculations is more difficult for logistic regression than for linear regression. We present sample size calculations for the simple logistic regression model and then for the multiple logistic model using the approach of Hsieh et al. (1998), which has already been discussed in previous sections. Both sets of calculations are simple enough to do manually.

27.4.1 Sample Size Determination for Simple Logistic Regression with a Binary Predictor

Consider the simple logistic regression model

$$\text{logit}[\text{pr}(Y = 1)] = \log_e\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 X$$

where Y is the binary outcome (coded $Y = 1$ for the event of interest, $Y = 0$ otherwise), π is the probability of the event of interest, and X is a binary covariate indexing two groups or treatments (coded $X = 1$ for one group and $X = 0$ for the other). The main inference of interest in this regression, performed to investigate whether or not covariate X is a statistically important predictor of Y , involves a hypothesis test for zero slope: $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$. The Wald chi-square test for the slope was described in Section 21.3.1. Also, it can easily be shown that this large sample test is mathematically equivalent to the test comparing two proportions described in Section 27.2.2. Therefore, sample size calculations for a logistic regression with a single binary covariate can be performed using formula (27.2).

■ Example 27.7 Consider again Example 27.2. The researchers want to conduct a logistic regression of Y (coded 1 if a dose is unacceptable and 0 if it is acceptable) on X (coded 1 if the new formulation and 0 if the standard formulation). They plan to perform a test of $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$, where β_1 = slope of the logit model. Suppose that previous studies of the standard formulation suggest that π_1 , the true proportion of acceptable doses for the standard formulation, is 0.90. If an absolute difference, Δ , in proportions of at least 0.05 would be important to detect, then, for $\alpha = 0.05$ and $\beta = 0.2$, the sample size for each group is given by equation (27.2)—at least 435 doses of each formulation would be required. ■

27.4.2 Sample Size Determination for Simple Logistic Regression with a Continuous Predictor

Consider again the simple logistic regression model

$$\text{logit}[\text{pr}(Y = 1)] = \log_e\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 X$$

where Y is the response variable (coded 1 if the event of interest occurs and 0 if it does not), π is the probability of the event of interest, and X is a standardized, normally distributed covariate. The main inference of interest in this regression, performed to investigate whether or not covariate X is a statistically significant predictor, involves testing $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$. This test is equivalent to testing whether or not the true average value of X is the same at the two values of Y and, as such whether it can be considered to be a two-sample test comparing means. Hsieh et al. (1998) present a modified version of equation (27.1) for

sample size calculations. The minimum sample size necessary, n_s , is given by

$$n_s \geq \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\pi_1(1 - \pi_1)\beta_1^{*2}} \quad (27.5)$$

where π_1 is the probability of the outcome at the mean value of covariate X and $|\beta_1^*|$ is the minimum absolute value of the slope that would be important to detect if the alternative hypothesis were true. Note that, since the variable X is entered in standardized form, β_1^* measures the effect of a one-standard-deviation change in X .

■ Example 27.8 Consider again Examples 27.1 and 27.2. Suppose that the researchers want to conduct a logistic regression of Y (coded 1 if a dose is unacceptable and 0 if it is acceptable) on X , the weight (in mg) of doses of the new formulation of the drug. X is assumed to be normally distributed and will be entered into the regression as a standardized variable. The researchers plan to perform a test of $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$, where β_1 = slope of the logit model. Suppose that the manufacturing process for each formulation is calibrated to provide an average dose weight of 15 mg with a standard deviation of 2 mg; and, at this average weight, experience has shown that 3 out of every 100 doses of the standard drug would be deemed unacceptable. If a minimum difference, β_1^* , in the slope of one unit (corresponding to a large odds ratio of 2.72) would be important to detect, then, for $\alpha = 0.05$ and $\beta = 0.2$, the approximate sample size for each group is given by

$$n_s \geq \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\pi_1(1 - \pi_1)\beta_1^{*2}} = \frac{(Z_{0.975} + Z_{0.8})^2}{(0.03)(1 - 0.03)(1)} = \frac{(1.96 + 0.842)^2}{0.0291} = 269.8$$

At least 270 doses of the new formulation need to be sampled. ■

27.4.3 Sample Size Determination for Multiple Logistic Regression

For multiple logistic regression, when one covariate is primarily of interest and one or more other covariates are included mainly for control, the sample size for the simple linear regression, n_s , may be calculated using whichever of (27.2) or (27.5) above is appropriate, and then the adjustment of Hsieh et al. (1998) shown in equation (27.4) may be applied to obtain the approximate sample size for the multiple logistic regression model.

■ Example 27.9 Tearing of the anterior cruciate ligament (ACL) is a common knee injury suffered by athletes in many sports. Surgery to replace the torn ACL with a muscle graft or cadaver ligament is often performed to repair the injury. Suppose that an orthopedic surgeon wants to evaluate two forms of the surgery: the standard surgery, in which a single replacement ligament is used, versus the “double-bundle” technique, in which two replacement ligaments are used. The following logistic regression is planned:

$$\text{logit}[\text{pr}(Y = 1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

where $Y = 1$ if the replacement ligament tears within two years after surgery, $Y = 0$ if it does not tear; $X_1 = 1$ if standard surgery is performed, $X_1 = 0$ if the double-bundle technique is used; $X_2 = \text{patient's age at the time of surgery}$; and $X_3 = \text{patient's weight at the time of surgery}$. X_1 is the primary predictor of interest; X_2 and X_3 are control variables.

The researchers want to use $\alpha = 0.05$ and $\beta = 0.2$. They believe that the re-tear rate is 2% for the standard surgery and would like to have a power of 0.8 of detecting an absolute difference in re-tear rates of at least 1%. Since the patients will be randomized to the two types of surgery, there will be little correlation between X_1 and the two control variables; a population multiple squared correlation, $\rho_{X_1(X_2, X_3)}^2$, of 0.0001 will be assumed.

Using formula (27.2), the required minimum sample size, n_s , for a simple logistic regression of Y on X_1 would be

$$n_s \geq 2 \left[\frac{(1.96)\sqrt{(2)(0.015)(1 - 0.015)} + (0.842)\sqrt{(0.02)(1 - 0.02) + (0.01)(1 - 0.01)}}{0.01} \right]^2 \\ = 4,637.7$$

At least 4,638 patients will be needed (i.e., at least 2,319 patients will have to be evaluated for each type of surgery). For the multiple logistic regression, the approximate minimum sample size, n_m , is found using formula (27.3):

$$n_m = \frac{n_s}{1 - \rho_{X_1(X_2, X_3)}^2} = \frac{4,638}{1 - 0.0001} = 4,638.5$$

At least 4,639 patients will be needed (i.e., at least 2,320 patients will have to be evaluated for each type of surgery). ■

Note that the required sample size in this example would, practically speaking, be impossible to achieve. In such situations, as described in Section 27.3.2, researchers would need to study tables and graphs of sample sizes for various combinations of power and effect size, the purpose being to determine whether acceptable combinations of these parameters exist for which the required sample size is practically feasible.

■ **Example 27.10** Planning for emergency medical care needs in environments involving mass gatherings (e.g., large sports events and concerts) requires an understanding of the type of care that will most likely be needed: trauma (i.e., injury) care or care for non-injury medical conditions. The type of care needed will probably be associated with a patient's age (assumed to be normally distributed). Suppose that planning for the relevant study to investigate this association is being conducted. Other covariates that will need to be controlled are type of event (concert versus sports event), patient's gender, and type of weather during the event.³ The logistic regression model of interest is

$$\text{logit}[\text{pr}(Y = 1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

³ A similar study has been conducted by Milsten et al. (2003).

where $Y = 1$ if the patient requires trauma care, $Y = 0$ if care for a non-injury medical condition is required; X_1 = age of the patient, in years (the primary covariate of interest); $X_2 = 1$ if the patient is male, $X_2 = 0$ if female; $X_3 = 1$ if the event is a concert, $X_3 = 0$ if it is a sporting event; and X_4 = amount of precipitation, in inches, on the day of the event.

How many patients will need to be studied? The researchers will use $\alpha = 0.05$ and $\beta = 0.2$. They believe that 20% of the care necessary will be for traumas, and, once again, believe that a minimum difference in the slope of one unit would be important to detect. The multiple correlation between X_1 and the control variables is assumed to be low; a multiple squared correlation, $\rho_{X_1(X_2, X_3, X_4)}^2$, of 0.01 will be assumed.

The minimum necessary sample size is given by formula (27.5) and by Hsieh et al.'s (1998) formula (27.4):

$$n_s \geq \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\pi_1(1 - \pi_1)\beta_1^2} = \frac{(Z_{0.975} + Z_{0.8})^2}{(0.2)(1 - 0.2)(1)} = \frac{(1.96 + 0.842)^2}{0.16} = 49.1$$

$$n_m = \frac{n_s}{1 - 0.01} = \frac{49.1}{0.99} = 49.6$$

At least 50 patients should be included in the study. ■

27.5 Power and Sample Size Determination for Linear Models: A General Approach

The sample size formulas presented in the preceding sections are conceptually simple and are also easy to implement, even without the use of specialized software. Indeed, these approaches are probably the simplest approaches available that yield reasonably accurate sample size calculations. However, these simple approaches are not based on the standard theory for power and sample size calculations for regression models. The standard theory provides a more general approach that allows for power (and sample size) determination for a variety of experimental designs and models and a variety of contrasts related to the models. For example, in a multiple linear regression, a *multiple* partial F test for the joint importance of several independent variables, controlling for several other variables, may be of interest rather than a partial F test about just one particular independent variable. As another example, in ANOVA, it may be of more interest to determine the sample size necessary for powerful contrasts in a multiple comparisons analysis rather than for the overall test for the significance of a single factor. The theory also generalizes to more complicated models such as ANOVA models involving two or more factors, models with repeated measures, and MANOVA models (Koele 1982; Muller et al. 1992).

This theory is well documented in numerous books and papers (e.g., Scheffé 1959; Cohen 1977; Muller et al. 1992). We present just the technical details that are necessary for the implementation of the theory. Analysts typically will rely on the use of specialized software to implement these methods. PASS 11 and SAS examples will be presented.

27.5.1 Power and Sample Size Determination for Multiple Linear Regression

Consider the multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_s X_s + \beta_1^* X_1^* + \cdots + \beta_q^* X_q^* + E \quad (27.6)$$

with the usual assumptions (see Section 8.4.1). Suppose that the hypothesis test of interest is a test for the importance of X_1, \dots, X_s given X_1^*, \dots, X_q^* (i.e., a test of $H_0: \beta_1 = \cdots = \beta_s = 0$ versus H_A : “At least one of these $\beta_j \neq 0$, given X_1^*, \dots, X_q^* are in the model”). It was shown in Section 10.7.1 that, for sample size n , the test statistic can be written as

$$F = \frac{R_{Y(X_1, \dots, X_s) | X_1^*, \dots, X_q^*}^2 / s}{(1 - R_{Y | X_1, \dots, X_s, X_1^*, \dots, X_q^*}^2) / (n - q - s - 1)}$$

where $F \sim F_{s, n-q-s-1}$ under the null hypothesis. The steps necessary to determine the power of this size α test are as follows:

1. Determine the critical value $F_{s, n-q-s-1, 1-\alpha}$.
2. Estimate (or provide values for), using a reasoned approach, $\rho_{Y | X_1, \dots, X_s, X_1^*, \dots, X_q^*}^2$ (the population squared multiple correlation) and $\rho_{Y(X_1, \dots, X_s) | X_1^*, \dots, X_q^*}^2$ (the population squared multiple partial correlation for variables X_1, \dots, X_s given X_1^*, \dots, X_q^*).
3. Calculate the power as $\Pr(F_\lambda > F_{1-\alpha})$, where F_λ follows a non-central F distribution with non-centrality parameter

$$\lambda = \frac{n \rho_{Y(X_1, \dots, X_s) | X_1^*, \dots, X_q^*}^2}{1 - \rho_{Y | X_1, \dots, X_s, X_1^*, \dots, X_q^*}^2}.$$

This parameter λ captures the effect size that is to be detected. Typically, users rely on software to perform this probability calculation.

There is no straightforward way, in this general setting, to calculate sample sizes for multiple linear regression models. Instead, the power calculation above has to be repeated with different values of n until a sample size that yields an acceptable power is achieved. Computer software makes this iterative approach feasible.

- **Example 27.11** This example illustrates the use of PASS 11 software to perform a sample size calculation for multiple linear regression. The presentation here will not provide insight into all the capabilities of the software. Instead, it will focus simply on the solution to the current problem. It is left to the reader to explore the additional capabilities of PASS 11.

Consider again Example 27.5. Researchers want to conduct a study evaluating two surgeries for nearsightedness. The outcome will be the patient’s refractive error, measured in diopters, one year after surgery. Researchers want to control for the patient’s age and baseline refractive error. They want to use $\alpha = 0.05$ and $\beta = 0.20$ and to detect a minimum difference in average error of 1 diopter.

Suppose that the population squared multiple correlation between the dependent and all independent variables is projected to be between 0.25 and 0.70; the population squared multiple partial correlation between the dependent variable and the control variables, age and baseline refractive error, given that the main independent variable of interest (surgery type) is in the model, is projected to be 0.05. (Note that, therefore, the population squared partial correlation between the dependent variable and main independent variable of interest, surgery type, given that age and baseline refractive error are in the model, will, therefore, be between 0.20 and 0.65.)

In PASS 11, the user should select “Multiple Regression” from the list of procedures available under the “Regression” category and then complete the subsequent dialog box as shown in Figure 27.1. To perform the calculation, the “RUN” button (near the top left of the dialog box) must be clicked.

Figure 27.2 shows a screenshot of the output from PASS 11. Both a table and a graph are produced, by default. They show that a sample size of between 8 and 32 is required.

PASS 11 could be re-run to explore how the required sample size would vary for different values of power, significance level, and squared partial correlations.

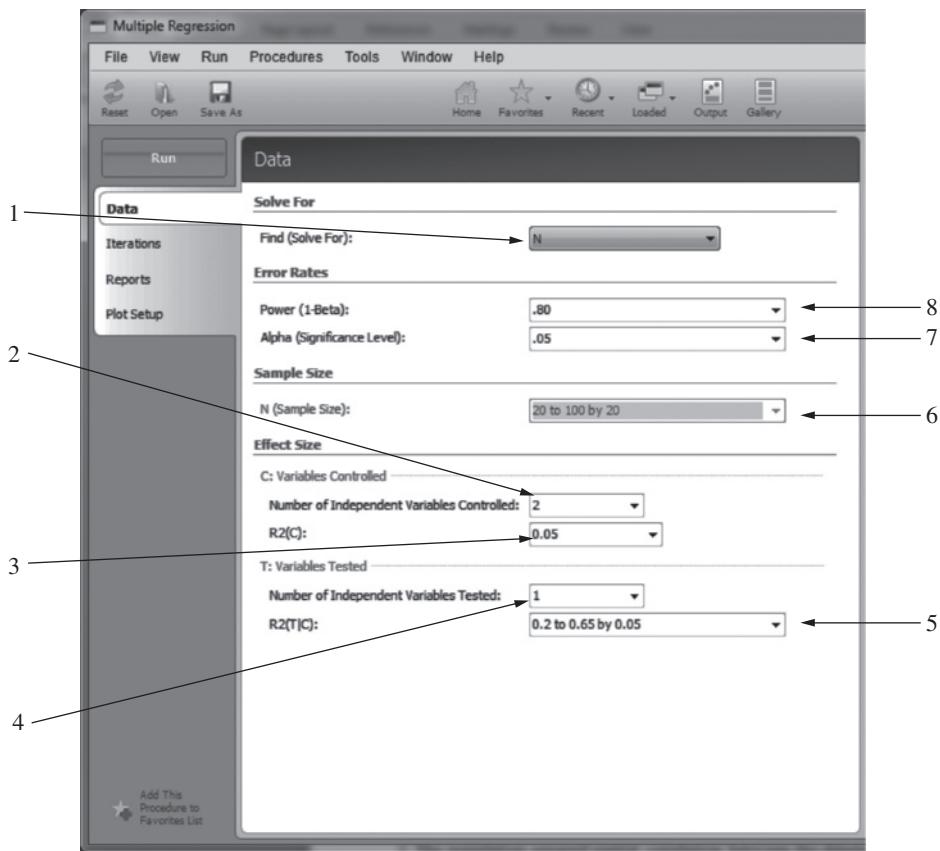


FIGURE 27.1 PASS 11 dialog box for the multiple linear regression sample size calculation in Example 27.11

Multiple Regression Power Analysis

Numeric Results

Power	N	Alpha	Beta	Ind. Variables Tested		Ind. Variables Controlled	
				Cnt	R2	Cnt	R2
0.80504	32	0.05000	0.19496	1	0.20	2	0.05
0.81303	25	0.05000	0.18697	1	0.25	2	0.05
0.81376	20	0.05000	0.18624	1	0.30	2	0.05
0.80061	16	0.05000	0.19939	1	0.35	2	0.05
0.81959	14	0.05000	0.18041	1	0.40	2	0.05
0.81996	12	0.05000	0.18004	1	0.45	2	0.05
0.84882	11	0.05000	0.15118	1	0.50	2	0.05
0.80084	9	0.05000	0.19916	1	0.55	2	0.05
0.87644	9	0.05000	0.12356	1	0.60	2	0.05
0.86891	8	0.05000	0.13109	1	0.65	2	0.05

References

Cohen, Jacob. 1988. Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates. Hillsdale, New Jersey.

Report Definitions

Power is the probability of rejecting a false null hypothesis.

N is the number of observations on which the multiple regression is computed.

Alpha is the probability of rejecting a true null hypothesis. It should be small.

Beta is the probability of accepting a false null hypothesis. It should be small.

Cnt refers to the number of independent variables in that category.

R2 is the amount that is added to the overall R-Squared value by these variables.

Ind. Variables Tested are those variables whose regression coefficients are tested against zero.

Ind. Variables Controlled are those variables whose influence is removed from experimental error.

Summary Statements

A sample size of 32 achieves 81% power to detect an R-Squared of 0.20 attributed to 1 independent variable(s) using an F-Test with a significance level (alpha) of 0.05000. The variables tested are adjusted for an additional 2 independent variable(s) with an R-Squared of 0.05.

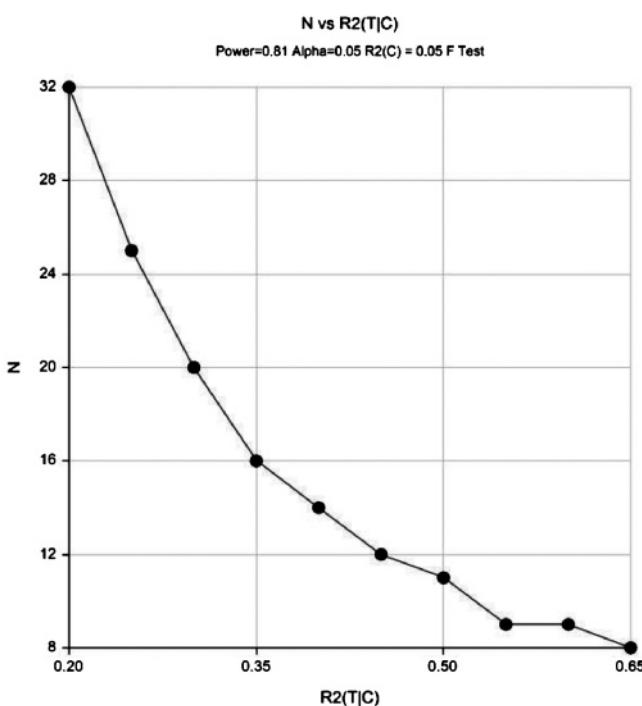


FIGURE 27.2 PASS 11 output for Example 27.11

Explanation of values in the dialog box:

1. PASS 11 allows you to calculate the power of a test or the sample size necessary to achieve that power. We have selected the sample size option.
2. The number of control variables, q , is specified here. PASS 11 labels these C.
3. The population squared multiple partial correlation between the dependent and p control variables, labeled R2(C) in PASS 11, is specified here. This must be a single value.
4. The number of independent variables being tested, s , is specified here. PASS 11 labels these variables T.
5. The population squared partial correlation between the dependent variable and the s independent variable(s) that are being tested, controlling for the remaining q independent variables, is specified here. PASS labels this value R2(T|C). The sum of this value and the value in box 3 above must be less than 1. A range of values may be specified.
6. This box is inactive in the current calculation. If, in box 1, we had chosen to calculate power, a range of sample sizes could have been specified here.
7. The desired significance level or the desired range of values is specified here.
8. The desired power is specified here. A single value or a range of values may be specified. ■

■ **Example 27.12** We re-do the previous example using the PROC POWER procedure in SAS software. The SAS program (Figure 27.3) and output are shown below. The SAS code is annotated with explanatory notes. The output shows the same results as seen in the PASS output.

```
proc power;
  multreg ← Specifies that the model is a regression model
    model = fixed ← Specifies fixed effects
    alpha = 0.05 ← Specifies significance level
    power = 0.80 ← Specifies power
    nfullpredictors = 3 ← Specifies the number of independent
                          variables in the full model ( $p+k$ )
    nreducedpredictors = 2 ← Specifies the number of control variables ( $p$ )
    rsquarefull = 0.25 to 0.70 by 0.05 ← Specifies the population squared multiple
                                         correlation for the model with all independent
                                         variables
    rsquarereduced = 0.05 ← Specifies the population squared multiple
                           correlation for the control variables
    ntotal = .; ← Specifies that the required sample size
                  is to be computed
run;
```

FIGURE 27.3 SAS program for the multiple linear regression sample size calculation in Example 27.12

Edited SAS Output (PROC POWER) for Example 27.12

The SAS System
 The POWER Procedure
 Type III F Test in Multiple Regression

FIXED SCENARIO ELEMENTS	
Method	Exact
Model	Fixed X
Number of Predictors in Full Model	3
Number of Predictors in Reduced Model	2
Alpha	0.05
R-square of Reduced Model	0.05
Nominal Power	0.8

COMPUTED N TOTAL			
Index	R-square Full	Actual Power	N Total
1	0.25	0.805	32
2	0.30	0.813	25
3	0.35	0.814	20
4	0.40	0.801	16
5	0.45	0.820	14
6	0.50	0.820	12
7	0.55	0.849	11
8	0.60	0.801	9
9	0.65	0.876	9
10	0.70	0.869	8

27.5.2 Power and Sample Size for One-way Fixed Effects ANOVA

In Chapter 17, it was pointed out that ANOVA can be viewed as a special case of multiple linear regression. As such, the sample size determination methods already presented for multiple linear regression can be applied to ANOVA models, including the simple approaches in Sections 27.3.1 and 27.3.3. Indeed, in its simplest form, one-way ANOVA with a two-level, fixed effects factor is the equivalent of a two-sample *t* test with equal population variances (see Sections 17.1 and 17.3.2); therefore, formula (27.1) could be used in sample size planning. For more general ANOVA models, the procedure for determining power (and sample size) described in Section 27.5.1 can be used. However, practical implementation of the procedure will be easier if we restate the procedure specifically for ANOVA models.

Consider the fixed effects one-way ANOVA model presented in Chapter 17:

$$Y_{ij} = \mu + \alpha_i + E_{ij} \quad \text{for } i = 1, 2, \dots, k \quad \text{and} \quad j = 1, 2, \dots, n_i \quad (27.7)$$

where

Y_{ij} = j th observation from the i th population

μ_i = mean for population i

$$\mu = \frac{\mu_1 + \mu_2 + \cdots + \mu_k}{k} = \text{true overall mean response}$$

$\alpha_i = \mu_i - \mu$ = differential effect of population i

E_{ij} = error term for the j th observation from the i th population

and where the usual assumptions are made (see Section 17.2). A hypothesis test that is always of interest is the overall F test with null hypothesis $H_0: \mu_1 = \mu_2 \dots = \mu_k$ versus H_A : “The k population means are not all equal”. It was shown in Section 17.3 that the test statistic can be written as

$$F = \frac{\text{MST}}{\text{MSE}} = \frac{\sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y})^2 / (k - 1)}{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 / (n - k)}$$

where $F \sim F_{k-1, n-k}$ under the null hypothesis. The steps necessary to determine the power of this size α test are as follows:

1. Determine the critical value $F_{k-1, n-k, 1-\alpha}$.
2. Using a reasoned approach, provide a value for the non-centrality parameter $\lambda = \frac{n\sigma_m^2}{\sigma^2}$, where n is the total sample size ($n = \sum_{i=1}^k n_i$), σ^2 is the variance of the outcome that is common to each population (i.e., the within-population variance), and σ_m^2 is the between-population variance

$$\sigma_m^2 = \sum_{i=1}^k \frac{n_i(\mu_i - \mu^*)^2}{n}, \text{ where } \mu^* \text{ is the weighted average population mean, } \mu^* = \sum_{i=1}^k \frac{n_i \mu_i}{n}.$$

Once again, λ captures the effect size that is to be detected. In practice, the value of σ^2 that is used is based on prior research or pilot studies or is a sound guess based on expert knowledge. The value of σ_m^2 is obtained by specifying both the values of $\mu_1, \mu_2, \dots, \mu_k$ under the alternative hypothesis that would be of interest to detect and the values of the sample sizes n_1, n_2, \dots, n_k .

3. Calculate the power as $\Pr(F_\lambda > F_{k-1, n-k, 1-\alpha})$, where F_λ follows a *non-central F distribution* with non-centrality parameter λ . Typically, users rely on software to perform this probability calculation.

As in the case of linear regression, iterative implementation of this procedure with different values of n allows the user to determine the overall sample size that yields an acceptable power. Note that this procedure is easily adapted to the case where the main hypothesis test of interest involves some other contrast of interest, as illustrated in Example 27.13.

■ Example 27.13 Researchers in Bangladesh want to compare the average temperature of surface waters (ponds) in three rural regions of Bangladesh. It is believed that, with higher temperatures, elevated cholera bacteria counts can be expected, which can then put rural inhabitants at increased risk for cholera infection. How many ponds should be studied in each region? Previous studies show that the standard deviation of pond water temperature is approximately 6 degrees Celsius. If the average pond water temperature of one well-studied southern region is 29 degrees Celsius and, in actuality, in the two other northern regions it is 24 degrees Celsius, the researchers would like to be able to reject the null hypothesis with $\alpha = 0.05$ and power 0.8.

Again, we use PASS 11 to perform the sample size determination. The user selects “One Way Analysis of Variance” from the list of available procedures in the “Means-ANOVA” category and then completes the subsequent dialog box as shown in Figure 27.4. The output from PASS 11, shown in Figure 27.5, indicates that 22 ponds need to be sampled in each of the three regions of the study.

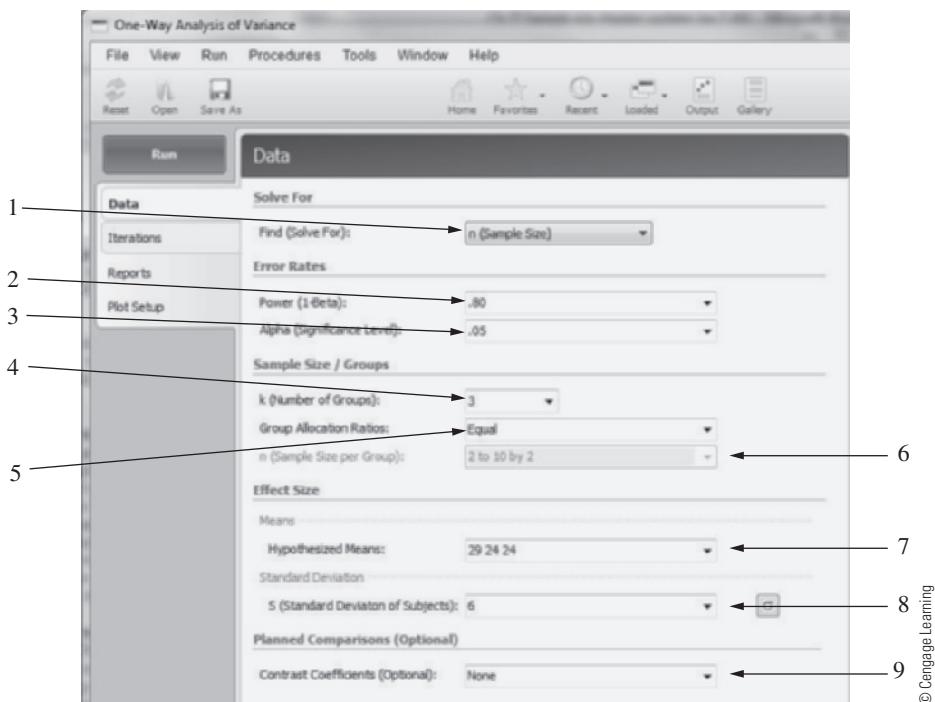


FIGURE 27.4 PASS 11 dialog box for the one-way ANOVA sample size calculation in Example 27.13

Explanation of values in the dialog box:

- PASS 11 allows you to determine the power of a test or the sample size necessary to achieve that power. We have selected the sample size option.
- The desired power is specified here. A single value or a range of values may be specified.

One Way ANOVA Power Analysis									
Numeric Results									
	Average	Total				Std Dev	Standard		
Power	n	k	N	Alpha	Beta	of Means (Sm)	Deviation (S)	Effect Size	
0.80319	22.00	3	66	0.05000	0.19681	2.36	6.00	0.3928	

References

Desu, M. M. and Raghavarao, D. 1990. *Sample Size Methodology*. Academic Press. New York.
 Fleiss, Joseph L. 1986. *The Design and Analysis of Clinical Experiments*. John Wiley & Sons. New York.
 Kirk, Roger E. 1982. *Experimental Design: Procedures for the Behavioral Sciences*. Brooks/Cole. Pacific Grove, California.

Report Definitions

Power is the probability of rejecting a false null hypothesis. It should be close to one.
 n is the average group sample size.
 k is the number of groups.
 Total N is the total sample size of all groups combined.
 Alpha is the probability of rejecting a true null hypothesis. It should be small.
 Beta is the probability of accepting a false null hypothesis. It should be small.
 Sm is the standard deviation of the group means under the alternative hypothesis.
 Standard deviation is the within group standard deviation.
 The Effect Size is the ratio of Sm to standard deviation.

Summary Statements

In a one-way ANOVA study, sample sizes of 22, 22, and 22 are obtained from the 3 groups whose means are to be compared. The total sample of 66 subjects achieves 80% power to detect differences among the means versus the alternative of equal means using an F test with a 0.05000 significance level. The size of the variation in the means is represented by their standard deviation which is 2.36. The common standard deviation within a group is assumed to be 6.00.

FIGURE 27.5 Edited PASS 11 output for Example 27.13

3. The desired significance level is specified here. A single value or a range of values may be specified.
4. The number of populations or groups, k , for which means are to be compared is specified here.
5. The sample sizes for the groups can be specified as being equal, or a pattern for the sample sizes can be specified. If a pattern is specified, the number of values provided should equal the number of groups. If it does not, then the last provided value is repeated for remaining groups. Also, if a pattern is specified and power is being calculated, the pattern in box 5 is multiplied by the sample size specified in box 6 in order to determine the per-group sample sizes.
6. When power is chosen for the calculation specified in box 1, the per-group sample size (or a range of sample sizes) for which power is to be calculated is specified here.
7. The values of the population means under the alternative hypothesis are specified here.
8. The value of α , the within-group standard deviation, is specified here.
9. If the calculation is for a user-defined contrast, rather than the standard inference that all means are equal, the contrast coefficients are specified here. ■

■ **Example 27.14** We repeat the previous example using PROC POWER in SAS. The SAS program (Figure 27.6) and output follow. The SAS code is again annotated with explanatory notes. The output shows once again that 22 observations per region are required.

```
proc power;
  onewayanova ← Specifies one-way ANOVA model
    test = overall ← Specifies that the overall test, rather
                      than a particular contrast, is of interest
    alpha = 0.05 ← Specifies the significance level or a
                    range of levels of interest
    power = 0.80 ← Specifies the desired power or a range of
                    levels of power
    stddev = 6 ← Specifies value of  $\sigma$ 
    groupmeans = 29 | 24 | 24 ← Group means of interest under  $H_A$ 
    npergroup = :, ← Indicates that the per-group sample
                     size is to be determined
  run;
```

© Cengage Learning

FIGURE 27.6 SAS program for the multiple linear regression sample size calculation in Example 27.14

Edited SAS Output (PROC POWER) for Example 27.14

The SAS System
 The POWER Procedure
 Overall F Test for One-Way ANOVA

FIXED SCENARIO ELEMENTS	
Method	Exact
Alpha	0.05
Group Means	29 24 24
Standard Deviation	6
Nominal Power	0.8

COMPUTED N PER GROUP	
Actual Power	N Per Group
0.803	22

Example 27.15 In Example 27.13, suppose the researchers are particularly interested in whether the true average water temperature in the northern regions (regions 2 and 3) is different from the true average water temperature in the southern region (region 1). The null and alternative hypotheses of interest can be written as

$$H_0: \mu_1 - \frac{\mu_2 + \mu_3}{2} = 0 \text{ versus } H_A: \mu_1 - \frac{\mu_2 + \mu_3}{2} \neq 0.$$

The coefficients, c_i , $i = 1, 2, 3$, for this contrast are 1, -0.5 , and -0.5 . In this case, the non-centrality parameter for the relevant power calculation is

$$\lambda = \frac{n\sigma_{mc}^2}{\sigma^2}, \text{ where}$$

$$\sigma_{mc} = \frac{\left| \sum_{i=1}^3 c_i \mu_i \right|}{\sqrt{n \sum_{i=1}^3 \frac{c_i^2}{n_i}}}.$$

Figure 27.7 shows how the PASS 11 dialog box is completed for this example. The PASS 11 output in Figure 27.8 shows that 18 ponds need to be sampled in each region.

The contrast coefficients are specified here. The rest of the dialog box is completed as in Figure 27.4.

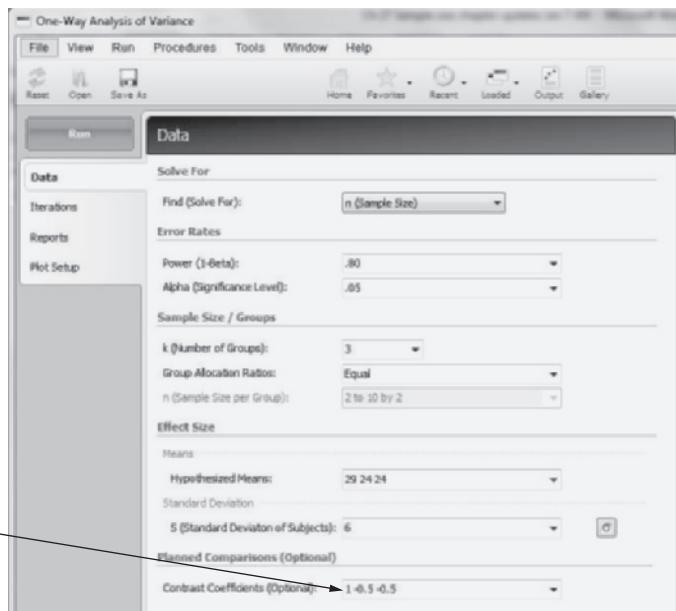


FIGURE 27.7 PASS dialog box for the one-way ANOVA sample size calculation in Example 27.15

One Way ANOVA Power Analysis								
Numeric Results						Std Dev of Means	Standard Deviation	Effect Size
Power	Average n	Total k	N	Alpha	Beta	(Sm)	(S)	
0.80841	18.00	3	54	0.05000	0.19159	2.36	6.00	0.3928

References

Desu, M. M. and Raghavarao, D. 1990. Sample Size Methodology. Academic Press. New York.
 Fleiss, Joseph L. 1986. The Design and Analysis of Clinical Experiments. John Wiley & Sons. New York.
 Kirk, Roger E. 1982. Experimental Design: Procedures for the Behavioral Sciences. Brooks/Cole. Pacific Grove, California.

Report Definitions

Power is the probability of rejecting a false null hypothesis. It should be close to one.
 n is the average group sample size.
 k is the number of groups.
 Total N is the total sample size of all groups combined.
 Alpha is the probability of rejecting a true null hypothesis. It should be small.
 Beta is the probability of accepting a false null hypothesis. It should be small.
 Sm is the standard deviation of the group means under the alternative hypothesis.
 Standard deviation is the within group standard deviation.
 The Effect Size is the ratio of Sm to standard deviation.

Summary Statements

In a one-way ANOVA study, sample sizes of 18, 18, and 18 are obtained from the 3 groups whose means are to be compared using a planned comparison (contrast). The total sample of 54 subjects achieves 81% power to detect a non-zero contrast of the means versus the alternative that the contrast is zero using an F test with a 0.05000 significance level. The value of the contrast of the means is 5.00. The common standard deviation within a group is assumed to be 6.00.

© Cengage Learning

FIGURE 27.8 Edited PASS 11 output for Example 27.15

- **Example 27.16** We re-do Example 27.15 using SAS to show the code necessary for user-defined contrasts. The SAS code is shown in Figure 27.9; the output (not shown) yields the same answer as PASS 11: 18 ponds need to be sampled per region. ■

27.5.3 Sample Size for Randomized Blocks ANOVA with Two Groups

In Section 18.2, it was pointed out that a randomized blocks ANOVA with two groups (say, a treatment [T] group and a control [C] group) is equivalent to the analysis of a matched-pairs experiment—the *F* test for differences in population means in the ANOVA would result in the exact same conclusion as would a paired-difference *t* test. As a result, sample sizes for a test of $H_0: \mu_T = \mu_C$ versus $H_A: \mu_T \neq \mu_C$ in a randomized blocks ANOVA with two groups can be estimated using the sample size formula for a one-sample test for the true mean difference.

For the situation where large sample sizes are assumed, the latter formula is based on the normal distribution, is very similar to formula (27.1), and is easy to implement by hand:

$$n \geq \left[\frac{\sigma(Z_{1-\alpha/2} + Z_{1-\beta})}{\Delta} \right]^2 \quad (27.8)$$

```

proc power;
onewayanova

test = contrast           ←
contrast = (1 -0.5 -0.5) ←
alpha = 0.05
power = 0.80
stddev = 6
groupmeans = 29 | 24 | 24
npergroup = .;
run;

```

User-defined contrasts are specified using these options. SAS allows for multiple contrasts to be specified by including sets of contrast coefficients in parentheses.

© Cengage Learning

FIGURE 27.9 SAS program for the one-way ANOVA sample size calculation in Example 27.16

where n is the number of pairs of observations, α is the desired significance level, $(1 - \beta)$ is the desired minimum power, σ is the standard deviation of the population of paired differences, and Δ is the minimum absolute difference between the two population means that would be important to detect. Once again, the researcher, in addition to specifying the desired levels of α , β , and Δ , must provide a reasoned guess for σ .

■ **Example 27.17** Suppose we are designing a study to investigate whether there is a difference in the true mean self-perception score for a treatment group versus a control group. A paired experimental design will be used, with pairs of subjects, one from the treatment group and one from the control group, being matched on some covariates. As a result, the analysis will involve a randomized blocks ANOVA. The number of pairs of people, n , necessary to achieve a power of at least 0.8, at a significance level of 0.05, with a population variance of $\sigma^2 = 13$ and a minimum absolute mean difference to be detected of $\Delta = 1$, is

$$n \geq \left[\frac{\sigma(Z_{1-\alpha/2} + Z_{1-\beta})}{\Delta} \right]^2 = \left[\frac{(3.6056)(1.96 + 0.842)}{1} \right]^2 = 102.1$$

At least 103 pairs of subjects would be needed in order to detect an absolute population mean difference of at least one unit in self-perception scores, for a significance level of 0.05 and a power of at least 0.8. ■

The sample size formula given by (27.8) for a matched-pairs design with a continuous outcome assumes that the only variables being controlled in the analysis are the variables

involved in the matching. If variables other than those involved in the matching are to be measured and controlled for during the analysis, then a more complicated sample size calculation is required. One possible method would be to use PASS 11 or SAS for power calculations for a multiple linear regression model involving dummy variables that incorporate the matching as well as other variables not involved in the matching. The use of this latter approach would require an educated guess for the population squared partial correlation between the dependent variable and the main independent variable (i.e., treatment group) of interest, controlling for the remaining independent variables, including both the dummy variables reflecting the matching and the other variables being controlled. As another alternative, provided there is only one other categorical predictor to control, one might use the PASS 11 procedure for randomized blocks ANOVA; this latter procedure allows for up to two categorical predictors and is described in the next subsection.

27.5.4 Sample Size for Randomized Blocks ANOVA with More than Two Groups

In general, sample size estimates for randomized blocks ANOVA are obtained using an approach similar to that described in Section 27.5.2, and computer software will typically be needed to perform the computations. We present examples using PASS 11.

Example 27.18 Suppose we are designing a study to investigate whether there are any differences in true mean self-perception scores for two treatment groups and a control group, where a block consists of three subjects (one subject from each of the two treatment groups and one subject from the control group), each matched on certain covariates. The self-perception score means are taken to be 10 and 12 for the two treatment groups and 9 for the control group. We want to be able to reject the null hypothesis of no difference in group means with power at least 0.8, at a significance level of 0.05. The variance of the population means

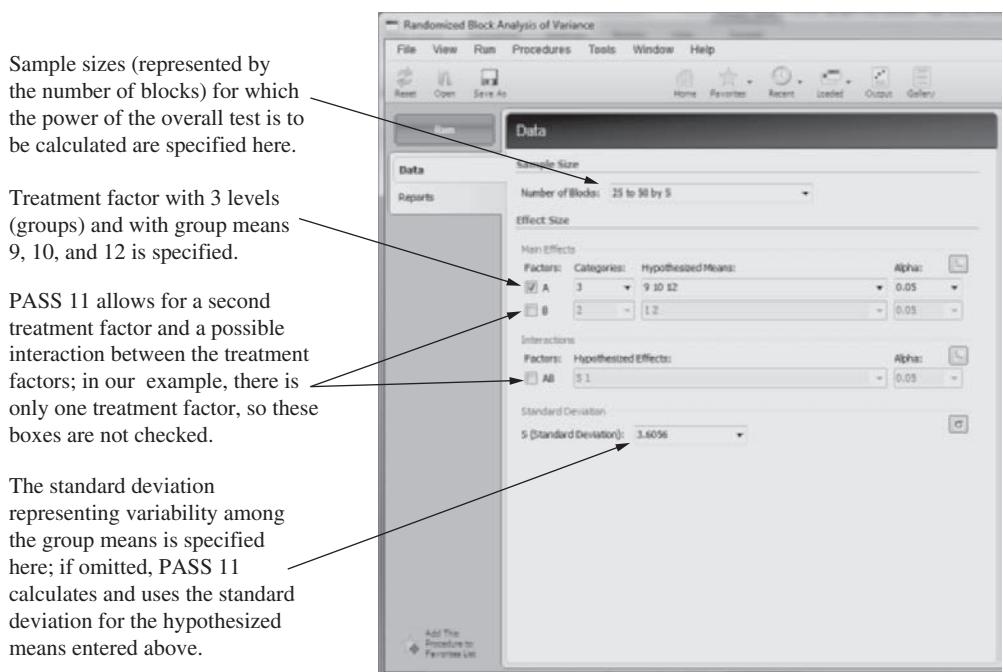
$$\left[\frac{1}{k} \sum_{i=1}^k (\mu_j - \bar{\mu})^2 \right]$$

is to be conservatively set at 13 based on previous experience.

Figure 27.10 shows how the PASS 11 dialog box is completed for this example. PASS determines the power of the overall F test for different sample sizes; we have chosen to investigate the power for sample sizes ranging from 25 blocks (i.e., 75 total observations, since there are three groups) to 50 blocks (150 observations) in increments of 5.

The PASS 11 output in Figure 27.11 shows that 30 blocks (90 total observations) are needed to achieve a power of 0.82. Therefore, approximately 30 subjects are needed per group.

As mentioned in the previous section, PASS 11 software for randomized blocks can be used to determine power calculations for matched data in which a single variable not involved in the matching is also controlled. Such an approach is possible by designating a second treatment factor (B) in the appropriate dialog box to represent the variable not involved in the matching. We have omitted an example of this approach, however.



© Cengage Learning

FIGURE 27.10 PASS 11 dialog box for the randomized block ANOVA sample size calculation in Example 27.18

Randomized Block ANOVA Power Analysis										
Numeric Results										
	Term	Power	Blocks	Units	df1	df2	Std Dev of Means	Effect Size	Alpha	Beta
30 blocks (90 total observations)	A	0.74128	25	75	2	48	1.247	0.3459	0.05000	0.25872
are needed to achieve a power of 0.82.	A	0.82429	30	90	2	58	1.247	0.3459	0.05000	0.17571
	A	0.88363	35	105	2	68	1.247	0.3459	0.05000	0.11637
	A	0.92460	40	120	2	78	1.247	0.3459	0.05000	0.07540
	A	0.95207	45	135	2	88	1.247	0.3459	0.05000	0.04793
	A	0.97004	50	150	2	98	1.247	0.3459	0.05000	0.02996
Standard Deviation Within Blocks (block-treatment interaction): 3.606										
References										
Odeh, R.E. and Fox, M. 1991. <i>Sample Size Choice</i> . Marcel Dekker, Inc. New York, NY.										
Winer, B.J. 1991. <i>Statistical Principles in Experimental Design</i> . Third Edition. McGraw-Hill. New York, NY.										
Report Definitions										
Power is the probability of rejecting a false null hypothesis.										
Blocks are the number of blocks in the design.										
Units are the number of experimental units in the design.										
df1 is the numerator degrees of freedom.										
df2 is the denominator degrees of freedom.										
Sm is the standard deviation of the group means at which the power is calculated.										
Effect Size is the ratio of Sm to standard deviation.										
Alpha is the probability of rejecting a true null hypothesis.										
Beta is the probability of accepting a false null hypothesis.										
Standard Deviation Within Blocks is the pooled block-treatment interaction.										
Summary Statements										
A randomized-block design with one treatment factor at 3 levels has 25.0 blocks each with 3.0 treatment combinations. The square root of the block-treatment interaction is 3.606. This design achieves 74% power when an F test is used to test factor A at a 5% significance level and the actual standard deviation among the appropriate means is 1.247 (an effect size of 0.3459).										

© Cengage Learning

FIGURE 27.11 PASS 11 output for Example 27.18

27.6 Sample Size Determination for Matched Case-control Studies with a Dichotomous Outcome

In Section 22.5.2, the logistic regression model appropriate for analyzing pair-matched case-control studies (without unmatched covariates) was presented. The model was used to assess the effect of a binary exposure variable on a binary outcome variable. The matched design provided a method of controlling for the covariates upon which the matching was based, with the idea being that well-performed matching would improve the power of the inferences of interest in the logistic regression analysis.

We present here an example of sample size calculations for such pair-matched case-control studies, without the consideration of unmatched covariates. The main inference of interest in the study is whether the odds ratio, which compares the odds of the outcome of interest for exposed (or treatment group) subjects and unexposed (control) subjects, is 1 (null hypothesis) or different from 1 (alternative hypothesis). The mathematical details of power and sample size calculations for this type of inference are beyond the scope of this book; we refer readers to Dupont (1988) for these details. However, PASS 11 software performs the necessary calculations both for pair-matching and for situations where more than one control subject is matched to each case. PASS 11 does not currently perform calculations for models involving unmatched covariates (discussed in Section 22.6).

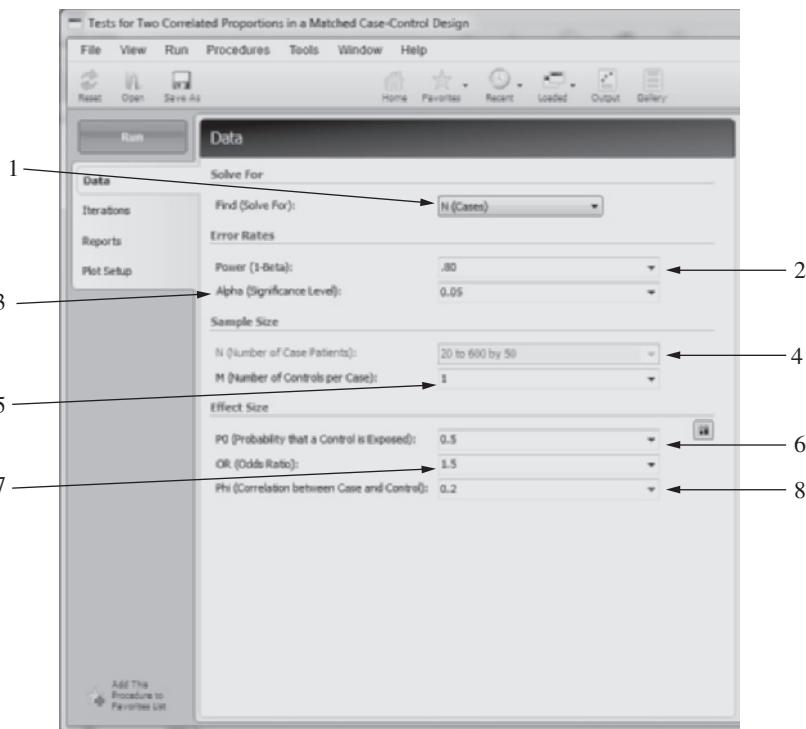
■ Example 27.19 Suppose that we are designing a pair-matched case-control study to investigate the association between exposure to a toxin (two levels: exposed or not exposed) and occurrence of a disease (two levels: disease or no disease). Researchers would like to reject the null hypothesis of no association between exposure and disease occurrence if the population odds ratio for disease (comparing exposed subjects to unexposed subjects) is at least 1.5. A significance level of 0.05 and power of 0.8 are desired. Cases and controls are matched on age, race, and sex.

The PASS 11 dialog box for this matched case-control analysis is shown in Figure 27.12. The results, shown in Figure 27.13, indicate that a fairly large sample size of at least 484 cases (and, therefore, 484 controls) would be required.

Explanation of values in the dialog box:

1. PASS 11 can solve either for power or for the number of cases to be sampled, N. Here the latter option is chosen.
2. The desired value of power.
3. The desired significance level.

4. This box is left blank; the number of cases sampled would be supplied if the power was to be calculated.
5. Number of controls per case.
6. A reasonable value for the probability that a control will be exposed. A range of values can be supplied (e.g., 0.4 to 0.6 in increments of 0.05).
7. The smallest odds ratio (comparing the odds of disease for exposed and unexposed people) that is to be detected with the desired power. A value greater than one is usually specified.
8. A reasonable value for the magnitude of the correlation of exposure status for cases and controls. The value must be between 0 and 1. Dupont (1988) suggests a default of at least 0.2 when the true value is not known. ■



© Cengage Learning

FIGURE 27.12 PASS 11 dialog box for the matched case-control analysis sample size calculation in Example 27.19

Matched Case-Control Power Analysis							
Numeric Results							
Power	Cases (N)	Controls Per Case (M)	Odds Ratio (OR)	Probability Exposed (P0)	Correlation (Phi)	Alpha	Beta
0.80078	484	1	1.50	0.50000	0.20000	0.05000	0.19922

References
 'Power Calculations for Matched Case-Control Studies', Biometrics, Volume 44, pages 1157-1168.

Report Definitions
 Power is the probability of rejecting a false null hypothesis.
 N is the size of the sample drawn from the treatment (case) group.
 M is the number of matching control patients drawn for each case patient.
 OR is the odds ratio of for subjects exposed to the risk factor.
 P0 is the probability of exposure among sampled control patients.
 Phi is the correlation of exposure between matched individuals.
 Alpha is the probability of rejecting a true null hypothesis.
 Beta is the probability of accepting a false null hypothesis.

Summary Statements
 In a matched case-control study, the probability of exposure among sampled control patients is 0.50000 and the correlation coefficient for exposure between matched case and control patients is 0.20000. A sample of 484 case patients is obtained. For each case patient, a matching sample of 1 control patient(s) is also obtained. This sample of 968 patients achieves 80% power to detect an odds ratio of 1.50 versus the alternative of equal odds using a Chi-Square test with a 0.05000 significance level.

© Cengage Learning

FIGURE 27.13 PASS 11 output for Example 27.19

27.7 Practical Considerations and Cautions

This chapter has focused on simple sample size determination techniques which can be easily and quickly implemented by both statisticians and non-statisticians. It is important to emphasize, once again, that sample size planning will rarely involve a single, quick calculation in real-world applications. Indeed, there are several important and potentially difficult practical issues that must be carefully considered before the methods in this chapter are applied. As a result, sample size planning should be viewed as a process rather than a single calculation.

The most difficult issue in sample size planning is the requirement that underlying population variances and correlations (if needed) and the minimum significant difference or effect size be specified. Statistical analysts and subject-matter experts should work together to determine these values. Prior research or, better still, a pilot study will be useful in estimating unknown population variances and correlations. Where such research is not available, conservative but scientifically reasoned guesses may be used. For example, it may be possible to specify the theoretical range of the outcome variable based on subject-matter knowledge; this range divided by 4 is sometimes used as a very rough and conservative estimate of the population standard deviation. The effect size of interest must be elicited from the subject-matter experts. Lenth (2001) has an excellent discussion on this subject. By varying both the variance and the effect size, sensitivity analyses can be performed to provide insight into the variation in required sample sizes across different plausible and practical combinations of effect size and variance.

Sensitivity analyses involving the significance level and power can also be useful. In Example 27.9, researchers would find that the cost required for a sample size of over 4,600 patients

was far too high to be feasible. They could compute required samples sizes for smaller values of power, larger effect sizes, or even higher significance levels in order to decide whether an acceptable combination of these indices exists for which the required sample size is practically feasible.

Another issue of practical importance concerns the increasing availability of software for power and sample size calculations, including SAS and PASS 11 but also including many other shareware or freeware packages and web applets. Many of these programs have easy-to-use “point-and-click” interfaces and thus put rigorous sample size planning tools within the reach of more analysts. At the same time, there is a danger that the applications will be employed inappropriately by users who are not adequately aware of details about the underlying models and the methods of calculation associated with the software. As a further complication, the levels of documentation differ greatly for different software packages and applets. Finally, not all software applications are equally accurate, as some will rely on approximations more than others (Hsieh et al. 1998; Thomas and Krebs 1997). Indeed, the simulation studies of Hsieh et al. showed that results from the simple formulas discussed in Section 27.2 were, in some cases, more accurate than results produced by certain software packages. We think it is essential that users choose well-documented software, in which all the mathematical details of the power and sample size calculations are given; moreover, we encourage users to read and understand the software documentation so that they are well aware of the methodology being implemented, the correct technique for using the software, and any potential limitations of the software.

Finally, a word of caution is needed about *a posteriori* (or retrospective) power calculations. Analysts sometimes calculate the power of a hypothesis-testing method *after* the study has been conducted and the null hypothesis has not been rejected, using the effect size that was estimated based on the data set under consideration. Often, these calculations are performed in order to explain or “excuse” the nonsignificant finding. This practice is controversial. It can be argued that there is little to be learned from such calculations, since it can be assumed that, if the null hypothesis of no effect was not rejected, then presumably the power of the method was not high enough to conclude that the observed effect size was significantly different from zero (Lenth, 2000, 2001).

A situation in which retrospective power calculations may be useful is when pilot study results are used to conduct power calculations for a future, larger study. Of course, it could be argued that these calculations are actually prospective with regard to the future study. Regardless of how they are viewed, software packages often do allow for such power calculations.

Problems

In the problems below, assume that a significance level, α , of 0.05 and a power, $(1 - \beta)$, of 0.8 are desired unless otherwise stated.

1. A clinical trial is to be conducted to investigate and compare the effects of a new cholesterol-lowering drug with the effects of a standard drug. Determine the sample sizes necessary for the analysis scenarios in parts (a) and (b) below.
 - a. A two-tailed, two-sample t test will be performed to compare the average cholesterol levels of hypercholesterolemic patients treated with the new drug and hypercholesterolemic patients treated with the standard drug. Under the homogeneous

variance assumption, the population-specific standard deviation of cholesterol levels is assumed to be 25 mg/dl. It will be important to detect an absolute difference in average true cholesterol levels of 25 mg/dl or larger.

- b. A test comparing the proportions of new-drug-treated patients and standard-drug-treated patients who complain of side effects is to be performed. Previous studies suggest that 25% of patients treated with the standard drug complain of side effects. It is important to detect an absolute difference in population proportions of 0.1 or more.
2. For the scenario in Problem 1(a),

 - a. Determine the sample size for power levels ranging from 0.6 to 0.9, in increments of 0.05. Plot the sample size versus power. Comment on the observed relationship.
 - b. Determine the sample size for absolute differences in average cholesterol levels ranging from 20 mg/dl to 70 mg/dl, in increments of 10 mg/dl. Plot the sample size versus the absolute difference. Comment on the observed relationship.
3. For the scenario in Problem 1(b),

 - a. Determine the required sample size for power levels ranging from 0.6 to 0.9, in increments of 0.10. Plot the sample size versus power. Comment on the observed relationship.
 - b. Determine the required sample size for absolute differences in proportions of patients who report side effects, ranging between 0.01 and 0.10, in increments of 0.01. Comment on the observed relationship.
4. In Example 27.3, the manufacturer of the drugs would like to determine how much more powerful the hypothesis-testing method would be if larger samples were collected. Plot the sample size versus the power for an appropriate range of levels of power. What, approximately, would the power of the testing method be if

 - a. 50 doses of each drug were sampled?
 - b. 100 doses of each drug were sampled?
5. A study is being planned to investigate the association between systolic blood pressure (SBP) and independent variables age (AGE), smoking history (SMK = 0 if non-smoker, SMK = 1 if a current or previous smoker), and body size as measured by the quetelet index (QUET) described in Problem 2 of Chapter 5. SBP, AGE and QUET are continuous variables. The statistical analysis will involve multiple linear regression with SBP as the dependent variable.

 - a. Suppose that SMK is the only independent variable to be included in the regression model. Determine the approximate sample size necessary to detect a true slope with an absolute value of at least 1.5. From previous research, the population standard deviation of SBP is approximately 3 mm Hg.
 - b. Suppose that SMK, AGE, and QUET are all included as predictors in the regression analysis, with SMK being the main predictor of interest and AGE and QUET being included to control for confounding. Using the method described in Section 27.3.3, determine the approximate sample size necessary to detect a true slope for SMK that has an absolute value of at least 1.5. From previous research,

- the population standard deviation of SBP is about 3 mm Hg, and the population squared multiple correlation between SMK and the other two independent variables is approximately 0.10.
- c. In part (b), recalculate the required sample size for population squared multiple correlations ranging from 0.10 to 0.90, in increments of 0.1. Comment on the relationship between the required sample size and the population squared multiple correlation.
 - d. Suppose that AGE is the only independent variable to be included in the regression model. Determine the sample size necessary to detect a true correlation between AGE and SBP of at least 0.40 in absolute value.
 - e. Suppose that AGE, SMK, and QUET are all included as predictors in the regression analysis, with AGE being the main predictor of interest. Using the method described in Section 27.3.3, determine the sample size necessary to detect a correlation between SBP and AGE of at least 0.40 in absolute value, assuming that the population squared multiple correlation between AGE and the other two independent variables is 0.10.
6. Researchers in Example 27.9 feel that at most 200 patients can be enrolled in the study for each type of surgery. Redo the sample size calculations for that example and identify at least one combination of power and effect size (absolute difference in tear rates) for which the required sample size is 200 or less per surgery type.

For the remaining problems, it is recommended that a software package such as PASS 11 or SAS be employed.

- 7–9.** Redo Examples 27.11, 27.13, and 27.15 to determine the approximate sample size if
- a. The desired power is 0.9 (all other parameters are as given in the examples).
 - b. The desired significance level is 0.1 (all other parameters are as given in the examples).

References

- Cohen, J. 1977. *Statistical Power Analysis for the Behavioral Sciences*, Revised Edition. New York: Academic Press.
- Dupont, W. 1988. "Power Calculations for Matched Case–Control Studies." *Biometrics* 44: 1157–68.
- Hintze, J. 2011. PASS 11. Kaysville, UT: NCSS, LLC.
- Hsieh, F. Y.; Bloch, D. A.; and Larsen, M. D. 1998. "A Simple Method of Sample Size Calculation for Linear and Logistic Regression." *Statistics in Medicine* 17: 1623–34.
- Koele, P. 1982. "Calculating Power in Analysis of Variance." *Psychological Bulletin* 92(2): 513–16.
- Kupper, L. K., and Hafner, K. B. 1989. "How Appropriate Are Popular Sample Size Formulas?" *The American Statistician* 43: 101–5.
- Kutner, M. H.; Neter, J.; Nachtsheim, C. J.; and Li, W. 2004. *Applied Linear Statistical Models*, Fifth Edition. New York: McGraw-Hill/Irwin.
- Lenth, R. V. 2001. "Some Practical Guidelines for Effective Sample Size Determination." *The American Statistician* 55(3): 187–93.

- Lenth, R. V. 2003. "Two Sample-size Practices That I Don't Recommend." Available: <http://www.stat.uiowa.edu/~rlenth/Power/2badHabits.pdf>. Accessed June 5, 2013.
- Milstén, A. M.; Seaman, K. G.; Liu, P.; Bissell, R. A.; and Maguire, B. J. 2003. "Variables Influencing Medical Usage Rates, Injury Patterns, and Levels of Care for Mass Gatherings." *Prehospital and Disaster Medicine* 18(4): 334–45.
- Muller, K. E.; LaVange, L. M.; Ramey, S. L.; and Ramey, C. T. 1992. "Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications." *Journal of the American Statistical Association* 87: 1209–26.
- Pearson, E. S., and Hartley, H. O. 1951. "Charts of the Power Function for Analysis of Variance Tests, Derived from the Non-central F Distribution." *Biometrika* 38: 112–30.
- SAS Institute Inc. 2002–2012. SAS 9.3. Cary, N.C.
- Scheffé, H. 1959. *The Analysis of Variance*. New York: Wiley.
- Tabachnick, B. G., and Fidell, L. S. 2001. *Using Multivariate Statistics*, Fourth Edition. Boston: Allyn and Bacon.
- Thomas, L., and Krebs, C. J. 1997. "A Review of Statistical Power Analysis Software." *Bulletin of the Ecological Society of America* 78(2): 126–39.

A

Appendix—Tables

TABLE A.1 Standard Normal Cumulative Probabilities

<i>z</i>	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0014	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0076	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1057	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

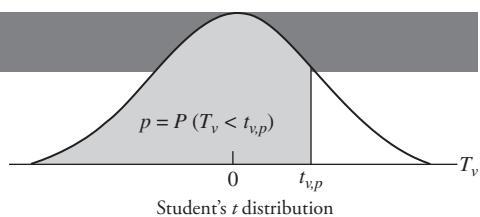
Note: Table entry is the area under the standard normal curve to the left of the indicated *z*-value, thus giving $P(Z < z)$.

TABLE A.1 Standard Normal Cumulative Probabilities (*continued*)

<i>z</i>	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8943	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9986	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000									

TABLE A.1 Standard Normal Cumulative Probabilities (*continued*)

<i>z</i>	$P(Z < z)$	<i>z</i>	$P(Z < z)$
-4.265	0.00001	0	0.50
-3.891	0.00005	0.126	0.55
-3.719	0.0001	0.253	0.60
-3.291	0.0005		
-3.090	0.001	0.385	0.65
-2.576	0.005	0.524	0.70
-2.326	0.01	0.674	0.75
		0.842	0.80
-2.054	0.02	1.036	0.85
-1.960	0.025		
-1.881	0.03	1.282	0.90
-1.751	0.04	1.341	0.91
-1.645	0.05	1.405	0.92
		1.476	0.93
-1.555	0.06	1.555	0.94
-1.476	0.07		
-1.405	0.08	1.645	0.95
-1.341	0.09	1.751	0.96
-1.282	0.10	1.881	0.97
		1.960	0.975
-1.036	0.15	2.054	0.98
-0.842	0.20		
-0.674	0.25	2.326	0.99
-0.524	0.30	2.576	0.995
-0.385	0.35	3.090	0.999
		3.291	0.9995
-0.253	0.40	3.719	0.9999
-0.126	0.45	3.891	0.99995
0	0.50	4.265	0.99999

TABLE A.2 Percentiles of the t Distribution

$\frac{100p}{df}$	55	65	75	85	90	95	97.5	99	99.5	99.95
1	0.158	0.510	1.000	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.142	0.445	0.816	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.137	0.424	0.765	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.134	0.414	0.741	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.408	0.727	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.404	0.718	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.402	0.711	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.130	0.399	0.706	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.398	0.703	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.397	0.700	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.396	0.697	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.395	0.695	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.394	0.694	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.393	0.692	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.393	0.691	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.392	0.690	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.392	0.689	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.392	0.688	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.391	0.688	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.391	0.687	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.391	0.686	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.390	0.686	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.390	0.685	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.127	0.390	0.685	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.390	0.684	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.390	0.684	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0.389	0.684	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.389	0.683	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.389	0.683	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.127	0.389	0.683	1.055	1.310	1.697	2.042	2.457	2.750	3.646
35	0.127	0.388	0.682	1.052	1.306	1.690	2.030	2.438	2.724	3.591
40	0.126	0.388	0.681	1.050	1.303	1.684	2.021	2.423	2.704	3.551
45	0.126	0.388	0.680	1.049	1.301	1.679	2.014	2.412	2.690	3.520
50	0.126	0.388	0.679	1.047	1.299	1.676	2.009	2.403	2.678	3.496
60	0.126	0.387	0.679	1.045	1.296	1.671	2.000	2.390	2.660	3.460
70	0.126	0.387	0.678	1.044	1.294	1.667	1.994	2.381	2.648	3.435
80	0.126	0.387	0.678	1.043	1.292	1.664	1.990	2.374	2.639	3.416
90	0.126	0.387	0.677	1.042	1.291	1.662	1.987	2.368	2.632	3.402
100	0.126	0.386	0.677	1.042	1.290	1.660	1.984	2.364	2.626	3.390
120	0.126	0.386	0.677	1.041	1.289	1.658	1.980	2.358	2.617	3.373
140	0.126	0.386	0.676	1.040	1.288	1.656	1.977	2.353	2.611	3.361
160	0.126	0.386	0.676	1.040	1.287	1.654	1.975	2.350	2.607	3.352
180	0.126	0.386	0.676	1.039	1.286	1.653	1.973	2.547	2.603	3.345
200	0.126	0.386	0.676	1.039	1.286	1.653	1.972	2.345	2.601	3.340
∞	0.126	0.385	0.674	1.036	1.282	1.645	1.960	2.326	2.576	3.291

TABLE A.3 Percentiles of the Chi-square Distribution



χ^2 distribution										
df	100p	1	2.5	5	10	20	30	40	50	60
1	0.00001	0.0002	0.001	0.004	0.016	0.064	0.148	0.2275	0.455	0.708
2	0.010	0.020	0.051	0.103	0.214	0.446	0.713	1.022	1.386	1.833
3	0.072	0.115	0.216	0.352	0.584	1.005	1.424	1.869	2.366	2.946
4	0.207	0.297	0.484	0.711	1.084	1.649	2.195	2.753	3.357	4.045
5	0.412	0.554	0.831	1.145	1.610	2.343	3.000	3.655	4.351	5.132
6	0.676	0.872	1.237	1.635	2.204	3.070	3.828	4.570	5.348	6.211
7	0.989	1.239	1.690	2.167	2.833	3.822	4.671	5.493	6.346	7.231
8	1.344	1.646	2.180	2.733	3.490	4.594	5.527	6.423	7.344	8.351
9	1.735	2.098	2.705	3.225	4.168	5.390	6.393	7.357	8.414	9.524
10	2.156	2.558	3.247	3.940	4.865	6.179	7.267	8.295	9.342	10.473
11	2.603	3.053	3.816	4.575	5.578	6.989	8.148	9.237	10.341	11.530
12	3.074	3.571	4.404	5.226	6.304	7.807	9.034	10.182	11.340	12.584
13	3.565	4.107	5.009	5.892	7.042	8.634	9.926	11.129	12.340	13.636
14	4.075	4.660	5.629	6.571	7.594	9.467	10.821	12.078	13.339	14.685
15	4.601	5.229	6.262	7.261	8.259	10.307	11.721	13.030	14.339	15.733
16	5.142	5.812	6.908	7.962	9.312	11.152	12.624	13.983	15.338	16.780
17	5.697	6.408	7.564	8.672	10.085	12.002	13.531	14.937	16.338	17.824
18	6.265	7.015	8.231	9.390	10.865	12.857	14.440	15.883	17.338	18.868
19	6.844	7.633	8.907	10.117	11.651	13.716	15.352	16.850	18.338	19.910
20	7.434	8.260	9.591	10.851	12.443	14.578	16.266	17.809	19.337	20.951
21	8.034	8.897	10.283	11.591	13.240	15.445	17.182	18.788	20.337	21.991
22	8.643	9.542	10.982	12.338	14.041	16.314	18.101	19.797	21.337	23.031
23	9.260	10.196	11.689	13.091	14.848	17.187	19.021	20.690	22.337	24.069
24	9.886	10.856	12.401	13.848	15.659	18.062	19.943	21.752	23.337	25.106
25	10.520	11.524	13.120	14.611	16.473	18.940	20.867	22.616	24.337	26.143
26	11.160	12.198	13.844	15.379	17.292	19.820	21.792	23.579	25.336	27.179
27	11.808	12.879	14.573	16.151	18.114	20.703	22.647	24.544	26.336	28.214
28	12.461	13.565	15.308	16.928	18.939	21.588	23.647	25.509	27.336	29.249
29	13.121	14.256	16.047	17.708	19.768	22.475	24.577	26.435	28.336	30.283
30	13.787	14.953	16.791	18.493	20.599	23.364	25.508	27.442	29.336	31.316
35	17.192	18.509	20.569	22.465	24.797	27.836	30.178	32.282	34.336	36.475
40	20.707	22.164	24.433	26.599	29.051	32.345	34.817	37.134	39.334	41.622
45	24.311	25.901	28.366	30.612	33.350	36.884	39.585	41.995	44.335	46.761
50	27.991	29.707	32.357	34.764	37.689	41.449	44.313	46.864	49.335	51.892
60	35.534	37.485	40.482	43.188	46.459	50.641	53.809	56.620	59.335	62.135
80	51.172	53.540	57.153	60.391	64.278	73.291	78.558	82.511	85.993	89.334
90	59.196	61.754	65.647	69.126	77.929	82.358	87.945	91.429	95.808	99.334
100	67.328	70.065	74.222	76.705	80.624	86.086	91.149	95.486	99.334	102.946
120	83.852	91.573	95.205	100.624	106.806	111.456	115.485	119.334	123.289	127.616
140	100.655	104.034	109.137	113.659	119.029	125.758	130.766	135.149	143.604	153.654
160	117.679	121.346	126.810	131.56	137.946	144.783	150.158	154.856	159.334	163.898
180	134.884	138.820	144.741	149.969	156.153	163.868	169.588	179.354	184.173	189.446
200	152.241	156.432	162.728	168.279	174.335	183.003	189.049	194.319	199.334	204.434

TABLE A.4 Percentiles of the *F* Distribution

		DEGREES OF FREEDOM FOR NUMERATOR																			DEGREES OF FREEDOM FOR DENOMINATOR		DEGREES OF FREEDOM FOR DENOMINATOR		DEGREES OF FREEDOM FOR DENOMINATOR				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	40	50	100	150	200	
1	5.83	750	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32	9.37	9.41	9.44	9.47	9.49	9.52	9.53	9.55	9.57	9.58	9.63	9.67	9.71	9.74	9.80	9.81	9.87		
2	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.39	3.39	3.41	3.41	3.42	3.43	3.42	3.43	3.44	3.45	3.46	3.47	3.47	3.47	3.47	3.47	3.47	3.47		
3	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.45	2.45	2.46	2.46	2.46	2.46	2.46	2.46	2.46	2.46	2.46	2.46	2.47	2.47	2.47	2.47	2.47	2.47		
4	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08		
5	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.88	1.88	1.88	1.88	1.87	1.87	1.87		
6	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.77	1.77	1.77	1.76	1.76	1.76	1.76	1.76	1.76	1.75	1.75	1.74	1.74	1.74	1.74	1.74	1.74		
7	1.57	1.70	1.72	1.72	1.71	1.70	1.69	1.69	1.68	1.68	1.68	1.68	1.68	1.67	1.67	1.67	1.67	1.67	1.67	1.66	1.66	1.65	1.65	1.65	1.65	1.65	1.65		
8	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.63	1.62	1.62	1.62	1.61	1.61	1.61	1.61	1.61	1.61	1.60	1.59	1.58	1.58	1.58	1.58	1.58	1.58		
9	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.59	1.59	1.58	1.58	1.58	1.57	1.57	1.56	1.56	1.56	1.56	1.56	1.56	1.55	1.54	1.54	1.54	1.53	1.53	1.53		
10	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.54	1.54	1.53	1.53	1.53	1.53	1.53	1.53	1.52	1.52	1.51	1.51	1.50	1.49	1.49			
11	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52	1.52	1.51	1.51	1.50	1.50	1.50	1.50	1.50	1.50	1.49	1.49	1.49	1.49	1.48	1.47	1.46	1.46		
12	1.48	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50	1.49	1.49	1.49	1.48	1.47	1.47	1.46	1.46	1.46	1.45	1.45	1.45	1.44	1.43	1.43	1.43	1.43		
13	1.45	1.55	1.55	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.47	1.46	1.46	1.45	1.45	1.45	1.45	1.44	1.44	1.43	1.42	1.42	1.41	1.41	1.40	1.40		
14	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.45	1.44	1.44	1.43	1.43	1.43	1.43	1.42	1.42	1.41	1.40	1.39	1.38	1.38	1.38	1.38		
15	1.43	1.52	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.46	1.45	1.45	1.44	1.44	1.43	1.43	1.43	1.42	1.42	1.42	1.41	1.40	1.39	1.38	1.37	1.37	1.37		
16	1.42	1.51	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.43	1.42	1.42	1.41	1.41	1.41	1.40	1.40	1.40	1.39	1.38	1.37	1.36	1.35	1.35	1.35		
17	1.42	1.51	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.43	1.42	1.42	1.41	1.41	1.40	1.40	1.39	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.34		
18	1.41	1.50	1.50	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.42	1.41	1.40	1.40	1.39	1.39	1.38	1.38	1.37	1.36	1.35	1.34	1.33	1.33	1.32	1.32		
19	1.41	1.49	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.37	1.37	1.37	1.37	1.37	1.36	1.35	1.34	1.33	1.32	1.31	1.30		
20	1.40	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.39	1.38	1.37	1.37	1.37	1.37	1.37	1.37	1.37	1.36	1.35	1.34	1.33	1.32	1.31	1.30		
21	1.40	1.48	1.46	1.44	1.43	1.42	1.41	1.40	1.39	1.39	1.38	1.37	1.37	1.36	1.36	1.36	1.36	1.36	1.36	1.35	1.35	1.34	1.33	1.32	1.31	1.30	1.29	1.29	
22	1.40	1.48	1.46	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.34	1.34	1.34	1.34	1.33	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.27	
23	1.39	1.47	1.46	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.35	1.35	1.35	1.35	1.35	1.34	1.33	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.27
24	1.39	1.47	1.46	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.35	1.34	1.34	1.34	1.34	1.33	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.27	
25	1.39	1.47	1.46	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.35	1.34	1.34	1.34	1.33	1.33	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	
26	1.38	1.46	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.35	1.34	1.34	1.34	1.34	1.33	1.33	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	
27	1.38	1.46	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26		
28	1.38	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.26		
29	1.38	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.26		
30	1.38	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.26		
31	1.37	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.26		
32	1.37	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.26		
33	1.37	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.26	1.26		
34	1.37	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.26	1.26		
35	1.37	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.26	1.26		
36	1.36	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.26	1.26		
37	1.36	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.26	1.26		
38	1.36	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.26	1.26		
39	1.36	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.26	1.26		
40	1.36	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.26	1.26		
41	1.36	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.26	1.26			
42	1.36	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.26	1.26			
43	1.36	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.26	1.26			
44	1.36	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.33	1.32	1.32	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.26	1.26			
45	1.36	1.43	1.42	1.41	1.40	1.39</td																							

TABLE A.4 Percentiles of the *F* Distribution (continued)Upper 10% point of the *F* distribution

		DEGREES OF FREEDOM FOR NUMERATOR																											
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	40	50	100	150	200	
DENOMINATOR		1.399	4.95	5.68	5.72	58.2	58.9	59.4	59.9	60.2	60.5	60.7	60.9	61.1	61.2	61.3	61.5	61.6	61.7	61.7	62.1	62.3	62.5	62.7	63.0	63.1	63.2		
	2.853	9.00	9.16	9.24	9.29	9.33	9.39	9.40	9.42	9.43	9.43	9.44	9.44	9.44	9.44	9.45	9.45	9.45	9.45	9.46	9.46	9.47	9.47	9.48	9.48	9.49			
	3.554	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.22	5.21	5.20	5.19	5.19	5.18	5.17	5.17	5.16	5.15	5.15	5.14	5.14	5.14	5.14			
	4.454	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.89	3.88	3.87	3.86	3.85	3.84	3.83	3.82	3.81	3.80	3.78	3.77	3.77	3.77	3.77			
	5.406	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.28	3.27	3.26	3.25	3.24	3.23	3.22	3.21	3.21	3.21	3.21	3.15	3.13	3.12	3.12	3.12			
	6.378	3.46	3.29	3.18	3.11	3.06	3.01	2.98	2.96	2.94	2.92	2.90	2.89	2.88	2.87	2.86	2.85	2.84	2.84	2.81	2.80	2.78	2.77	2.75	2.74	2.73			
	7.359	3.26	3.07	2.96	2.86	2.83	2.78	2.75	2.72	2.67	2.62	2.59	2.56	2.54	2.52	2.50	2.49	2.46	2.45	2.44	2.43	2.40	2.38	2.36	2.35	2.31			
	8.346	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.52	2.50	2.48	2.46	2.44	2.42	2.40	2.38	2.36	2.34	2.32	2.31	2.29	2.28	2.27	2.21			
	9.336	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.40	2.38	2.36	2.35	2.34	2.33	2.32	2.31	2.30	2.29	2.27	2.25	2.23	2.22	2.19	2.18			
	10.329	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.30	2.28	2.27	2.26	2.24	2.23	2.22	2.21	2.20	2.16	2.13	2.12	2.09	2.08	2.07	2.07	2.07		
	11.323	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.23	2.21	2.19	2.18	2.17	2.16	2.15	2.14	2.13	2.12	2.10	2.08	2.07	2.06	2.03	2.01	1.99	1.99	
	12.318	2.81	2.61	2.48	2.39	2.33	2.28	2.23	2.20	2.17	2.17	2.15	2.13	2.12	2.10	2.09	2.08	2.07	2.06	2.05	2.04	2.03	2.02	2.01	1.99	1.97	1.97		
	13.314	2.76	2.56	2.43	2.35	2.28	2.20	2.16	2.12	2.10	2.12	2.10	2.08	2.07	2.06	2.05	2.04	2.02	2.01	2.00	1.99	1.97	1.96	1.95	1.94	1.93	1.93		
	14.310	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.10	2.07	2.06	2.05	2.04	2.03	2.02	2.00	1.99	1.97	1.96	1.95	1.94	1.93	1.92	1.91	1.91		
	15.307	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.04	2.02	2.00	1.99	1.97	1.96	1.95	1.94	1.93	1.92	1.89	1.87	1.85	1.83	1.79	1.78	1.77		
D.F. FOR DENOMINATOR		16.305	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	2.01	1.99	1.97	1.95	1.94	1.93	1.92	1.91	1.90	1.89	1.86	1.84	1.81	1.79	1.76	1.74		
	17.303	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.98	1.96	1.94	1.93	1.91	1.90	1.89	1.88	1.87	1.86	1.85	1.84	1.83	1.81	1.78	1.76	1.74		
	18.301	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.95	1.93	1.92	1.90	1.89	1.87	1.86	1.85	1.84	1.83	1.82	1.81	1.78	1.75	1.73	1.71			
	19.299	2.61	2.40	2.27	2.18	2.18	2.10	2.06	2.02	1.98	1.96	1.94	1.93	1.91	1.89	1.87	1.86	1.85	1.84	1.83	1.82	1.81	1.80	1.79	1.76	1.74	1.71		
	20.297	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.91	1.89	1.87	1.86	1.84	1.83	1.82	1.81	1.80	1.79	1.76	1.74	1.71	1.69	1.66	1.64	1.63		
	21.296	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.90	1.87	1.85	1.83	1.81	1.80	1.79	1.78	1.78	1.78	1.77	1.76	1.75	1.74	1.72	1.69	1.67		
	22.295	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.88	1.86	1.84	1.83	1.81	1.80	1.79	1.78	1.77	1.76	1.75	1.74	1.73	1.72	1.70	1.67	1.65		
	23.294	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.87	1.84	1.83	1.81	1.80	1.78	1.77	1.76	1.75	1.74	1.73	1.72	1.71	1.69	1.66	1.64	1.62		
	24.293	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.85	1.83	1.81	1.80	1.78	1.77	1.76	1.75	1.74	1.73	1.72	1.71	1.69	1.67	1.65	1.64	1.62		
	25.292	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.84	1.82	1.80	1.78	1.77	1.75	1.74	1.73	1.72	1.71	1.70	1.68	1.66	1.64	1.63	1.61	1.59		
	26.291	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.83	1.81	1.79	1.77	1.75	1.73	1.72	1.71	1.71	1.71	1.70	1.68	1.66	1.65	1.63	1.61	1.59		
	27.290	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.82	1.80	1.78	1.76	1.75	1.74	1.73	1.72	1.71	1.70	1.68	1.66	1.64	1.62	1.60	1.58	1.57		
	28.289	2.50	2.29	2.16	2.06	1.99	1.93	1.89	1.86	1.83	1.81	1.78	1.76	1.75	1.73	1.72	1.71	1.70	1.69	1.68	1.67	1.66	1.65	1.63	1.62	1.60	1.59		
	29.288	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.80	1.78	1.75	1.73	1.71	1.70	1.69	1.68	1.67	1.66	1.65	1.64	1.63	1.62	1.60	1.59	1.58		
	30.288	2.48	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.80	1.78	1.75	1.73	1.71	1.70	1.69	1.68	1.67	1.66	1.65	1.64	1.63	1.62	1.61	1.60	1.59		
D.F. OF FREEDOM		31.287	2.47	2.26	2.13	2.04	1.97	1.91	1.87	1.83	1.81	1.78	1.76	1.74	1.72	1.71	1.69	1.68	1.67	1.66	1.65	1.64	1.63	1.62	1.61	1.60	1.59		
	32.287	2.46	2.26	2.13	2.04	1.97	1.91	1.87	1.83	1.81	1.78	1.76	1.74	1.72	1.71	1.69	1.68	1.67	1.66	1.65	1.64	1.63	1.62	1.61	1.60	1.59	1.58		
	34.286	2.47	2.25	2.12	2.02	1.96	1.90	1.86	1.82	1.78	1.75	1.73	1.71	1.69	1.67	1.65	1.64	1.63	1.62	1.61	1.60	1.59	1.58	1.57	1.56	1.55	1.54		
	36.285	2.46	2.24	2.11	2.01	1.94	1.89	1.85	1.81	1.78	1.75	1.73	1.71	1.69	1.67	1.65	1.64	1.63	1.62	1.61	1.60	1.59	1.58	1.57	1.56	1.55	1.54		
	38.284	2.45	2.23	2.10	2.01	1.94	1.88	1.84	1.80	1.77	1.75	1.72	1.70	1.68	1.66	1.64	1.63	1.62	1.61	1.60	1.59	1.58	1.57	1.56	1.55	1.54	1.53		
	40.284	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.75	1.72	1.70	1.68	1.66	1.64	1.63	1.62	1.61	1.60	1.59	1.58	1.57	1.56	1.55	1.54	1.53	1.52		
	42.283	2.43	2.22	2.08	1.99	1.92	1.86	1.82	1.78	1.75	1.73	1.71	1.69	1.67	1.65	1.64	1.63	1.62	1.61	1.60	1.59	1.58	1.57	1.56	1.55	1.54	1.53		
	44.282	2.42	2.21	2.08	1.98	1.91	1.86	1.81	1.78	1.75	1.73	1.71	1.69	1.67	1.65	1.64	1.63	1.62	1.61	1.60	1.59	1.58	1.57	1.56	1.55	1.54	1.53		
	48.281	2.42	2.20	2.07	1.97	1.90	1.85	1.80	1.77	1.73	1.71	1.69	1.67	1.65	1.63	1.62	1.61	1.60	1.59	1.58	1.57	1.56	1.55	1.54	1.53	1.52	1.51		
	50.281	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.70	1.68	1.66	1.64	1.63	1.62	1.61	1.60	1.59	1.58	1.57	1.56	1.55	1.54	1.53	1.52	1.51		
	52.280	2.40	2.19	2.05	1.95	1.87	1.82	1.77	1.74	1.71	1.68	1.66	1.64	1.62	1.60	1.59	1.58	1.57	1.56	1.55	1.54	1.53	1.52	1.51	1.50	1.49	1.48	1.47	
	60.279	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.68	1.66	1.64	1.62	1.60	1.59	1.58	1.57	1.56	1.55	1.54	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46
	70.278	2.38	2.16	2.03	1.93	1.86	1.80	1.76	1.72	1.69	1.66	1.64	1.62	1.60	1.59	1.57	1.56	1.55	1.54	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.45	
	80.277	2.37	2.15	2.02	1.92	1.85	1.79	1.75	1.71	1.68	1.65	1.63	1.61	1.59	1.57	1.55	1.54	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42
	90.276	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67	1.64	1.62	1.60	1.58	1.56	1.55	1.54	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1			

TABLE A.4 Percentiles of the *F* Distribution (continued)Upper 5% point of the *F* distribution

		DEGREES OF FREEDOM FOR NUMERATOR																				DEGREES OF FREEDOM FOR DENOMINATOR		DEGREES OF FREEDOM FOR DENOMINATOR							
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	40	50	100	150	200			
1	161	200	216	225	230	234	239	241	242	244	245	246	247	248	249	248	248	247	247	246	252	253	253	254	254	254	254	254			
2	185	190	192	192	193	193	194	194	194	194	194	194	194	194	194	194	194	194	194	194	195	195	195	195	195	195	195	195			
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.73	8.71	8.70	8.69	8.68	8.67	8.66	8.63	8.62	8.59	8.58	8.56	8.54	8.54	8.54	8.54	8.54			
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.94	5.89	5.87	5.85	5.83	5.82	5.75	5.75	5.75	5.75	5.75	5.66	5.66	5.65	5.65	5.65	5.65	5.65	5.65			
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62	4.60	4.59	4.58	4.57	4.56	4.52	4.50	4.46	4.44	4.41	4.39	4.39	4.39			
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.92	3.91	3.90	3.88	3.87	3.83	3.81	3.77	3.75	3.71	3.70	3.69	3.69			
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.49	3.48	3.47	3.46	3.45	3.40	3.38	3.34	3.32	3.32	3.32	3.32	3.32			
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.20	3.19	3.17	3.16	3.15	3.11	3.08	3.04	3.02	2.97	2.96	2.95	2.95	2.95			
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.98	2.95	2.94	2.92	2.86	2.85	2.83	2.81	2.80	2.79	2.77	2.75	2.75			
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.88	2.86	2.85	2.83	2.81	2.79	2.77	2.73	2.70	2.66	2.64	2.64	2.64	2.64	2.64				
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.70	2.69	2.67	2.66	2.65	2.60	2.57	2.53	2.51	2.46	2.44	2.43				
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.60	2.58	2.56	2.54	2.50	2.47	2.47	2.43	2.40	2.35	2.33	2.32	2.32			
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.51	2.50	2.48	2.47	2.46	2.41	2.38	2.34	2.31	2.26	2.24	2.23	2.23			
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.44	2.43	2.41	2.40	2.39	2.34	2.31	2.27	2.24	2.19	2.17	2.16	2.16			
15	4.54	3.68	3.29	3.05	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.38	2.37	2.35	2.34	2.33	2.28	2.25	2.20	2.18	2.16	2.14	2.10	2.10			
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35	2.33	2.31	2.29	2.27	2.26	2.24	2.23	2.18	2.15	2.12	2.07	2.06	2.04			
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31	2.29	2.27	2.25	2.23	2.21	2.19	2.14	2.11	2.08	2.06	2.04	1.98	1.96			
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.28	2.26	2.23	2.21	2.20	2.18	2.17	2.11	2.07	2.03	2.00	1.94	1.92	1.91				
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.21	2.20	2.18	2.17	2.11	2.07	2.03	2.00	1.94	1.92	1.91	1.91				
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.18	2.17	2.15	2.14	2.12	2.07	2.04	1.99	1.97	1.91	1.89	1.88				
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.23	2.20	2.17	2.12	2.10	2.08	2.05	2.01	1.96	1.94	1.88	1.86	1.84	1.83	1.82	1.82				
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.20	2.17	2.13	2.10	2.06	2.04	2.01	1.99	1.97	1.92	1.88	1.84	1.83	1.82	1.82				
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.18	2.15	2.13	2.11	2.09	2.08	2.06	2.05	2.00	1.98	1.96	1.91	1.88	1.86	1.85				
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.15	2.13	2.11	2.09	2.07	2.05	2.04	2.03	1.97	1.94	1.89	1.85	1.83	1.82	1.81				
25	4.24	3.39	2.99	2.75	2.60	2.50	2.41	2.34	2.28	2.24	2.20	2.16	2.14	2.10	2.08	2.04	2.01	2.00	2.05	2.04	2.02	2.01	1.96	1.92	1.87	1.84	1.82	1.81			
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.12	2.09	2.07	2.05	2.03	2.02	2.00	1.99	1.94	1.85	1.82	1.76	1.74	1.73	1.73				
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.26	2.21	2.15	2.10	2.09	2.06	2.04	2.02	2.00	1.99	1.97	1.96	1.91	1.88	1.84	1.81	1.79	1.77	1.76				
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.10	2.09	2.06	2.04	2.02	2.00	1.99	1.96	1.94	1.90	1.88	1.85	1.80	1.76	1.73	1.72				
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.26	2.21	2.16	2.11	2.06	2.02	2.00	1.97	1.95	1.92	1.90	1.89	1.87	1.85	1.84	1.81	1.78	1.75	1.73	1.72				
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.25	2.19	2.12	2.08	2.04	2.00	1.96	1.92	1.89	1.86	1.84	1.81	1.79	1.77	1.75	1.73	1.72	1.70	1.68	1.66				
31	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14	2.10	2.06	2.01	1.98	1.94	1.91	1.89	1.86	1.84	1.81	1.78	1.75	1.72	1.69	1.66	1.63	1.61	1.60			
32	4.14	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12	2.08	2.04	2.00	1.97	1.94	1.91	1.89	1.87	1.85	1.83	1.80	1.78	1.75	1.72	1.69	1.66	1.63	1.62	1.61		
33	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.14	2.10	2.05	2.01	1.96	1.93	1.90	1.88	1.86	1.84	1.82	1.80	1.78	1.75	1.72	1.69	1.66	1.63	1.61	1.60			
34	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.99	1.96	1.93	1.90	1.88	1.86	1.84	1.82	1.80	1.78	1.75	1.72	1.69	1.66	1.63	1.62	1.61		
35	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.97	1.93	1.90	1.87	1.85	1.83	1.81	1.79	1.77	1.75	1.72	1.69	1.66	1.63	1.62	1.61	1.60		
36	4.06	3.22	2.83	2.59	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.96	1.91	1.88	1.84	1.81	1.78	1.75	1.72	1.70	1.67	1.65	1.63	1.61	1.59	1.57	1.55	1.53	1.52		
37	4.05	3.20	2.81	2.57	2.40	2.30	2.22	2.15	2.09	2.04	2.00	1.97	1.94	1.91	1.88	1.85	1.82	1.79	1.77	1.75	1.72	1.69	1.66	1.63	1.61	1.59	1.57	1.55	1.53	1.52	
38	4.04	3.19	2.80	2.57	2.41	2.32	2.21	2.14	2.08	2.03	2.00	1.97	1.94	1.91	1.88	1.85	1.82	1.79	1.77	1.75	1.72	1.69	1.66	1.63	1.61	1.59	1.57	1.55	1.53	1.52	
39	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.99	1.95	1.92	1.89	1.87	1.85	1.83	1.81	1.78	1.75	1.72	1.69	1.66	1.63	1.61	1.59	1.57	1.55	1.53	1.52	
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.97	1.93	1.90	1.87	1.84	1.82	1.80	1.78	1.75	1.72	1.69	1.66	1.63	1.61	1.59	1.57	1.55	1.53	1.52
41	4.07	3.22	2.83	2.59	2.44	2.31	2.21	2.13	2.06	2.01	1.98	1.94	1.91	1.88	1.85	1.82	1.79	1.77	1.75	1.72	1.69	1.66	1.63	1.61	1.59	1.57	1.55	1.53	1.52	1.51	1.50
42	4.07	3.22	2.83	2.59	2.44	2.31	2.21	2.13	2.06	2.01	1.98	1.94	1.91	1.88	1.85	1.82	1.79	1.77	1.75	1.72	1.69	1.66	1.63	1.61	1.59	1.57	1.55	1.53	1.52	1.51	1.50
43	4.06	3.21	2.82	2.58	2.43	2.30	2.21	2.13	2.05	2.01	1.98	1.94	1.91	1.88	1.85	1.82	1.79	1.77	1.75	1.72	1.69	1.66	1.63	1.61	1.59	1.57	1.55	1.53	1.52	1.51	1.50

TABLE A.4 Percentiles of the *F* Distribution (continued)Upper 2.5% point of the *F* distribution

		DEGREES OF FREEDOM FOR NUMERATOR																				30		40		50		100		150		200	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	40	50	100	150	200					
	1	548	800	864	900	922	937	948	957	963	969	973	977	980	983	985	987	989	990	992	993	998	1001	1006	1008	1013	1015	1016					
	2	385	39.0	39.2	39.3	39.3	39.4	39.4	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5					
	3	174	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.4	14.3	14.3	14.3	14.3	14.3	14.2	14.2	14.2	14.2	14.2	14.1	14.1	14.0	14.0	13.9	13.9					
	4	122	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.79	8.75	8.71	8.68	8.66	8.63	8.61	8.59	8.58	8.56	8.50	8.46	8.41	8.38	8.32	8.30	8.29					
	5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.62	6.57	6.49	6.46	6.43	6.40	6.36	6.34	6.33	6.27	6.23	6.18	6.08	6.06	6.05									
	6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.41	5.37	5.33	5.30	5.27	5.24	5.22	5.20	5.18	5.17	5.11	5.07	5.01	4.98	4.92	4.89	4.88					
	7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.71	4.67	4.63	4.60	4.57	4.54	4.52	4.50	4.48	4.47	4.40	4.36	4.31	4.28	4.21	4.19	4.18					
	8	7.57	6.06	5.42	5.05	4.82	4.64	4.45	4.36	4.30	4.10	4.03	3.96	3.91	3.87	3.83	3.80	3.77	3.74	3.70	3.68	3.67	3.60	3.56	3.51	3.47	3.40	3.38					
	9	7.21	5.71	5.08	4.72	4.42	4.07	3.95	3.85	3.78	3.72	3.66	3.62	3.58	3.55	3.52	3.50	3.47	3.45	3.44	3.42	3.35	3.31	3.26	3.22	3.15	3.13	3.12					
	10	6.94	5.46	4.83	4.47	4.24	3.98	3.76	3.66	3.55	3.53	3.51	3.49	3.47	3.43	3.39	3.36	3.33	3.30	3.28	3.26	3.24	3.23	3.16	3.12	3.06	3.03	2.96	2.92				
	11	6.72	5.26	4.63	4.28	4.04	3.88	3.73	3.66	3.55	3.53	3.51	3.49	3.47	3.43	3.39	3.36	3.33	3.30	3.28	3.26	3.24	3.23	3.16	3.12	3.06	3.03	2.96	2.92				
	12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.49	3.37	3.34	3.32	3.30	3.28	3.25	3.21	3.18	3.15	3.13	3.11	3.09	3.07	3.01	2.96	2.91	2.87	2.80	2.78	2.76			
	13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.95	2.88	2.84	2.81	2.74	2.67	2.65	2.63	2.63				
	14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.16	3.09	3.05	3.02	3.01	2.98	2.95	2.92	2.88	2.86	2.84	2.81	2.77	2.76	2.69	2.64	2.64	2.64	2.63				
	15	6.20	4.77	4.15	3.80	3.58	3.34	3.20	3.12	3.06	3.01	2.96	2.92	2.89	2.86	2.84	2.81	2.79	2.77	2.74	2.71	2.69	2.65	2.62	2.59	2.55	2.47	2.45					
	16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.93	2.89	2.85	2.82	2.79	2.76	2.74	2.72	2.70	2.68	2.61	2.57	2.51	2.47	2.40	2.37	2.36					
	17	6.04	4.62	4.01	3.66	3.44	3.28	3.18	3.06	3.01	2.93	2.87	2.82	2.77	2.72	2.67	2.64	2.62	2.60	2.58	2.56	2.49	2.44	2.38	2.35	2.27	2.24	2.23					
	18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.72	2.67	2.62	2.58	2.54	2.50	2.47	2.44	2.41	2.39	2.36	2.32	2.26	2.21	2.19	2.18					
	19	5.92	4.52	4.11	3.80	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.76	2.72	2.68	2.62	2.59	2.57	2.55	2.53	2.51	2.44	2.39	2.33	2.30	2.22	2.19	2.18					
	20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.72	2.68	2.64	2.60	2.57	2.50	2.48	2.46	2.40	2.35	2.29	2.25	2.17	2.14	2.13	2.13						
	21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.68	2.64	2.60	2.56	2.53	2.51	2.48	2.46	2.42	2.36	2.31	2.25	2.21	2.13	2.10	2.09	2.08					
	22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.65	2.60	2.56	2.53	2.50	2.47	2.45	2.43	2.41	2.39	2.32	2.27	2.21	2.17	2.09	2.06	2.05					
	23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.62	2.57	2.53	2.50	2.47	2.44	2.42	2.39	2.37	2.35	2.32	2.26	2.21	2.18	2.14	2.06	2.03					
	24	5.72	4.32	3.72	3.38	3.15	3.09	2.98	2.87	2.73	2.70	2.64	2.59	2.54	2.50	2.47	2.44	2.41	2.39	2.36	2.34	2.32	2.26	2.21	2.15	2.11	2.02	1.98					
	25	5.69	4.29	3.69	3.35	3.13	3.07	2.97	2.85	2.75	2.68	2.61	2.56	2.51	2.48	2.44	2.41	2.38	2.36	2.34	2.32	2.30	2.23	2.18	2.12	2.08	2.00	1.95					
	26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.64	2.59	2.54	2.49	2.45	2.42	2.39	2.36	2.34	2.31	2.29	2.28	2.21	2.16	2.10	2.05	1.97	1.94	1.92					
	27	5.63	4.24	3.63	3.32	3.06	2.90	2.78	2.69	2.61	2.55	2.49	2.45	2.41	2.37	2.34	2.32	2.29	2.27	2.25	2.23	2.16	2.11	2.05	2.01	1.92	1.88						
	28	5.61	4.22	3.61	3.32	3.04	2.88	2.76	2.67	2.59	2.53	2.48	2.43	2.39	2.36	2.32	2.30	2.27	2.25	2.23	2.21	2.14	2.09	2.03	1.99	1.90	1.87						
	29	5.59	4.19	3.59	3.30	3.03	2.87	2.75	2.65	2.57	2.51	2.46	2.41	2.37	2.34	2.31	2.28	2.26	2.23	2.21	2.19	2.12	2.07	2.01	1.97	1.88	1.85						
	30	5.57	4.18	3.59	3.30	3.03	2.87	2.75	2.65	2.57	2.51	2.46	2.41	2.37	2.34	2.31	2.28	2.26	2.23	2.21	2.19	2.12	2.08	2.00	1.97	1.88	1.85						
	31	5.53	4.15	3.56	3.22	3.00	2.84	2.71	2.62	2.54	2.48	2.43	2.38	2.34	2.31	2.28	2.25	2.22	2.20	2.17	2.15	2.13	2.08	2.04	1.98	1.93	1.85	1.82					
	32	5.50	4.12	3.53	3.19	2.97	2.81	2.69	2.59	2.52	2.45	2.40	2.35	2.31	2.28	2.25	2.22	2.20	2.17	2.15	2.13	2.11	2.04	1.99	1.92	1.88	1.85	1.82					
	33	5.47	4.09	3.50	3.17	2.94	2.78	2.66	2.57	2.49	2.43	2.37	2.33	2.29	2.25	2.22	2.20	2.17	2.15	2.13	2.11	2.09	2.04	1.99	1.92	1.88	1.85	1.82					
	34	5.45	4.06	3.48	3.16	3.04	2.90	2.74	2.62	2.53	2.47	2.41	2.33	2.27	2.23	2.20	2.16	2.14	2.11	2.09	2.06	2.03	2.01	1.98	1.94	1.90	1.86	1.82					
	35	5.42	4.03	3.45	3.11	2.99	2.73	2.61	2.51	2.43	2.37	2.32	2.27	2.23	2.20	2.16	2.12	2.10	2.06	2.03	2.00	1.97	1.93	1.89	1.85	1.81	1.77	1.73					
	36	5.45	4.07	3.48	3.19	2.95	2.76	2.64	2.55	2.46	2.38	2.32	2.26	2.22	2.18	2.14	2.11	2.08	2.06	2.03	2.00	1.97	1.94	1.89	1.84	1.80	1.76	1.74					
	37	5.42	4.06	3.46	3.13	2.94	2.70	2.62	2.53	2.45	2.39	2.31	2.26	2.21	2.18	2.15	2.12	2.09	2.06	2.03	2.00	1.97	1.94	1.89	1.85	1.81	1.77	1.73					
	38	5.40	4.03	3.43	3.09	2.87	2.71	2.59	2.47	2.38	2.30	2.24	2.18	2.14	2.10	2.06	2.03	2.00	1.97	1.95	1.92	1.88	1.84	1.80	1.76	1.72	1.68	1.64					
	39	5.37	4.00	3.42	3.08	2.86	2.70	2.58	2.47	2.38	2.31	2.24	2.18	2.12	2.08	2.03	2.00	1.97	1.95	1.92	1.89	1.85	1.81	1.77	1.73	1.69	1.65	1.62					
	40	5.34	3.97	3.39	3.05	2.83	2.67	2.51	2.43	2.37	2.32	2.27	2.23	2.19	2.15	2.11	2.08	2.04	2.00	1.97	1.94	1.91	1.88	1.85	1.81	1.77	1.73	1.69	1.66				
	41	5.39	3.92	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.22	2.17	2.13	2.09	2.05	2.01	1.98	1.96	1.94	1.91	1.87	1.82	1.78	1.74	1.70	1.67	1.65					
	42	5.40	4.03	3.45	3.11	2.89	2.73	2.61	2.51	2.43	2.37	2.32	2.27	2.23	2.19	2.15	2.11	2.08	2.04	2.00	1.97	1.93	1.89	1.85	1.81	1.77	1.73	1.69	1.67				
	43	5.37	3.98	3.31	2.97	2.75	2.61	2.51	2.41	2.33	2.28	2.22	2.18	2.14	2.10	2.06	2.03	2.00	1.97	1.95	1.92	1.88	1.84	1.80	1.76	1.72	1.68	1.65					
	44	5.39	3.93	3.34	3.09	2.87	2.71	2.59	2.50	2.42	2.34	2.29	2.24	2.19</td																			

TABLE A.4 Percentiles of the *F* Distribution (continued)Upper 1% point of the *F* distribution

		DEGREES OF FREEDOM FOR NUMERATOR																											
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	40	50	100	150	200	
DEGREES OF FREEDOM FOR DENOMINATOR																													
1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6083	6126	6146	6157	6170	6181	6192	6201	6209	6240	6261	6287	6303	6334	6345	6350			
2	98.5	99.0	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5	99.5			
3	34.1	30.8	29.5	28.7	28.2	27.9	27.5	27.3	27.1	27.0	26.9	26.8	26.7	26.7	26.7	26.7	26.7	26.7	26.7	26.5	26.5	26.4	26.4	26.2	26.2	26.2			
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.3	14.2	14.2	14.1	14.1	14.0	14.0	14.0	13.9	13.7	13.7	13.7	13.6	13.5	13.5			
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.1	10.0	9.9	9.8	9.7	9.7	9.6	9.6	9.6	9.6	9.6	9.5	9.5	9.5	9.5	9.4	9.4	9.4			
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.66	7.60	7.56	7.52	7.48	7.45	7.42	7.40	7.30	7.23	7.14	7.09	6.99	6.95	6.93		
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.80	6.62	6.54	6.47	6.41	6.36	6.31	6.28	6.24	6.21	6.18	6.16	6.06	5.99	5.91	5.86	5.75	5.72	5.70			
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.61	5.56	5.52	5.48	5.44	5.41	5.38	5.36	5.26	5.20	5.12	5.07	4.96	4.93	4.91		
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.25	5.26	5.18	5.11	5.05	5.01	4.96	4.92	4.89	4.86	4.83	4.81	4.71	4.65	4.57	4.52	4.41	4.38	4.36		
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.65	4.60	4.56	4.52	4.49	4.46	4.41	4.31	4.25	4.17	4.1	4.01	3.98	3.96	3.96		
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.34	4.29	4.25	4.21	4.18	4.15	4.12	4.10	4.01	3.94	3.86	3.81	3.71	3.67	3.66		
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.10	4.05	4.01	3.97	3.94	3.91	3.88	3.86	3.76	3.70	3.62	3.57	3.47	3.43	3.41		
13	9.07	6.70	5.74	5.21	4.86	4.62	4.46	4.30	4.19	4.10	4.02	3.96	3.91	3.86	3.82	3.78	3.75	3.72	3.69	3.66	3.57	3.51	3.43	3.38	3.32	3.24			
14	8.88	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.75	3.70	3.66	3.62	3.59	3.56	3.53	3.51	3.41	3.35	3.27	3.22	3.11	3.08	3.06		
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.61	3.56	3.52	3.49	3.45	3.42	3.40	3.37	3.28	3.21	3.13	3.08	2.98	2.94	2.92		
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.56	3.50	3.46	3.40	3.35	3.31	3.24	3.21	3.19	3.16	3.13	3.08	2.98	2.92	2.87	2.71		
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46	3.40	3.35	3.31	3.27	3.23	3.19	3.16	3.13	3.10	3.06	3.00	2.92	2.87	2.76	2.73		
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.32	3.27	3.23	3.19	3.16	3.13	3.10	3.08	3.03	3.00	2.95	2.89	2.84	2.78	2.68		
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.24	3.19	3.15	3.10	3.08	3.05	3.03	3.00	2.98	2.94	2.89	2.84	2.76	2.71	2.66	2.62		
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.18	3.13	3.09	3.05	3.02	2.99	2.96	2.94	2.84	2.78	2.72	2.69	2.64	2.54	2.50		
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.24	3.17	3.12	3.07	3.03	2.99	2.96	2.93	2.90	2.88	2.79	2.72	2.67	2.64	2.58	2.48	2.44		
22	7.92	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.07	3.02	2.98	2.94	2.91	2.88	2.85	2.83	2.73	2.67	2.58	2.53	2.42	2.38	2.36		
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	3.02	2.97	2.93	2.89	2.86	2.83	2.80	2.78	2.69	2.62	2.54	2.48	2.42	2.37	2.34		
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03	3.02	2.98	2.93	2.89	2.85	2.82	2.79	2.76	2.74	2.64	2.58	2.49	2.44	2.33	2.29		
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	3.07	3.02	2.93	2.86	2.80	2.74	2.70	2.65	2.62	2.58	2.55	2.50	2.45	2.40	2.39	2.35	2.29		
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	3.02	2.96	2.90	2.86	2.81	2.78	2.75	2.72	2.69	2.66	2.57	2.50	2.42	2.36	2.25	2.21	2.19		
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.99	2.93	2.87	2.82	2.78	2.75	2.72	2.68	2.66	2.63	2.54	2.47	2.38	2.33	2.22	2.18	2.16		
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96	2.90	2.84	2.79	2.75	2.72	2.68	2.65	2.63	2.60	2.51	2.44	2.35	2.30	2.19	2.15	2.13		
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.93	2.87	2.81	2.77	2.73	2.69	2.66	2.63	2.60	2.57	2.48	2.41	2.33	2.27	2.16	2.12	2.10		
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84	2.79	2.74	2.70	2.66	2.63	2.60	2.57	2.55	2.49	2.42	2.39	2.37	2.27	2.20	2.13		
31	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02	2.93	2.86	2.80	2.74	2.70	2.65	2.62	2.58	2.55	2.50	2.45	2.40	2.34	2.25	2.18	2.09	2.03	1.91	1.85	
32	7.45	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98	2.89	2.82	2.76	2.70	2.66	2.61	2.58	2.54	2.51	2.49	2.46	2.37	2.30	2.21	2.16	2.04	2.00	1.98		
33	7.44	5.25	4.38	3.89	3.57	3.35	3.18	3.05	2.95	2.87	2.81	2.75	2.70	2.64	2.60	2.56	2.51	2.48	2.45	2.43	2.33	2.28	2.18	2.12	2.07	1.94	1.84	1.80	1.78
34	7.40	5.25	4.38	3.86	3.54	3.32	3.15	3.02	2.92	2.83	2.77	2.72	2.67	2.62	2.58	2.54	2.50	2.48	2.45	2.42	2.34	2.28	2.18	2.12	2.07	1.94	1.84	1.80	1.76
35	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.92	2.83	2.77	2.72	2.67	2.62	2.58	2.54	2.50	2.48	2.45	2.42	2.34	2.28	2.18	2.12	2.07	1.94	1.84	1.80	1.76
36	7.30	5.17	4.26	3.72	3.41	3.19	3.02	2.91	2.80	2.71	2.64	2.59	2.54	2.50	2.46	2.42	2.38	2.35	2.32	2.28	2.21	2.17	2.10	2.01	1.95	1.82	1.78	1.76	
37	7.25	5.13	4.21	3.68	3.38	3.15	3.02	2.91	2.80	2.71	2.64	2.59	2.54	2.50	2.46	2.42	2.38	2.35	2.32	2.28	2.21	2.17	2.10	2.03	1.98	1.93	1.89	1.85	
38	7.20	5.08	4.13	3.65	3.34	3.12	3.02	2.91	2.80	2.71	2.64	2.59	2.54	2.50	2.46	2.42	2.38	2.35	2.32	2.28	2.22	2.17	2.10	2.03	1.98	1.93	1.88	1.83	
39	7.15	5.06	4.20	3.72	3.41	3.19	3.02	2.91	2.80	2.71	2.64	2.59	2.54	2.50	2.46	2.42	2.38	2.35	2.32	2.28	2.22	2.17	2.10	2.03	1.98	1.93	1.88	1.83	
40	7.11	5.01	4.17	3.66	3.34	3.12	3.02	2.91	2.80	2.71	2.64	2.59	2.54	2.50	2.46	2.42	2.38	2.35	2.32	2.28	2.22	2.17	2.10	2.03	1.98	1.93	1.88	1.83	
41	7.06	4.97	4.07	3.60	3.29	3.04	2.87	2.74	2.64	2.55	2.48	2.42	2.36	2.31	2.27	2.22	2.17	2.14	2.11	2.09	2.04	1.99	1.94	1.89	1.83	1.79	1.75	1.70	
42	7.28	5.15	4.29	3.80	3.47	3.24	3.08	2.95	2.84	2.75	2.68	2.62	2.56	2.51	2.45	2.40	2.35	2.31	2.27	2.23	2.17	2.14	2.11	2.09	2.02	1.97	1.92	1.87	1.83
43	7.25	5.12	4.26	3.78	3.44	3.22	3.06	2.93	2.82	2.79	2.72	2.66	2.61	2.55	2.50	2.45	2.40	2.37	2.31	2.27	2.22	2.17	2.14	2.11	2.07	2.02	1.97	1.92	1.87
44	7.22	5.10	4.24	3.76	3.43	3.21	3.04	2.93	2.82	2.79	2.72	2.66	2.60	2.54	2.50	2.45	2.40	2.37	2.31	2.27	2.22	2.19	2.15	2.12	2.07	2.02	1.97	1.92	1.87
45																													

TABLE A.4 Percentiles of the *F* Distribution (*continued*)Upper 0.5% point of the *F* distribution

	DEGREES OF FREEDOM FOR NUMERATOR																										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	40	50	100	150	200
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000		
2	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	
3	5.65	49.8	47.5	46.2	45.4	44.8	44.1	43.9	43.7	43.5	43.4	43.3	43.2	43.1	43.0	42.9	42.8	42.6	42.5	42.3	42.2	42.0	42.0	41.9	41.9		
4	31.3	24.3	23.2	22.5	22.0	21.6	21.4	21.1	21.0	20.8	20.7	20.6	20.5	20.4	20.4	20.3	20.2	20.2	20.2	20.1	19.9	19.7	19.7	19.4	19.4		
5	22.8	18.3	16.5	15.6	14.9	14.5	14.0	13.8	13.6	13.5	13.4	13.3	13.2	13.1	13.0	12.9	12.9	12.8	12.7	12.5	12.5	12.3	12.3	12.2	12.2		
6	18.6	14.5	12.9	12.0	11.5	11.1	10.8	10.6	10.4	10.3	10.1	10.0	9.95	9.88	9.81	9.76	9.71	9.66	9.62	9.59	9.45	9.36	9.24	9.17	9.03	8.98	8.95
7	16.2	12.4	10.9	10.1	9.52	9.16	8.89	8.68	8.51	8.38	8.18	8.03	7.97	7.91	7.87	7.83	7.79	7.75	7.62	7.53	7.42	7.35	7.22	7.17	7.15	7.15	
8	14.7	11.0	9.60	8.81	8.30	7.95	7.65	7.34	7.21	7.10	6.94	6.87	6.81	6.76	6.72	6.68	6.64	6.61	6.58	6.50	6.40	6.29	6.09	6.04	6.02	6.02	
9	13.6	10.1	8.72	7.96	7.47	7.13	6.88	6.59	6.34	6.21	6.15	6.09	6.03	5.98	5.94	5.90	5.86	5.83	5.71	5.62	5.52	5.45	5.32	5.28	5.26	5.26	
10	12.8	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.75	5.66	5.59	5.53	5.47	5.42	5.38	5.34	5.31	5.27	5.15	5.07	4.97	4.90	4.77	4.71	
11	12.2	8.91	7.60	6.88	6.42	6.10	5.86	5.56	5.34	5.22	5.24	5.16	5.10	5.05	5.00	4.96	4.92	4.89	4.86	4.83	4.74	4.65	4.55	4.49	4.36	4.21	
12	11.8	8.51	7.23	6.52	6.07	5.76	5.52	5.25	5.08	4.94	4.82	4.72	4.64	4.57	4.51	4.46	4.41	4.37	4.33	4.27	4.15	4.07	3.97	3.91	3.78	3.74	
13	11.4	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.72	4.64	4.57	4.51	4.46	4.41	4.36	4.30	4.25	4.20	4.16	4.12	4.09	4.06	3.94	3.86	
14	11.1	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.51	4.43	4.36	4.30	4.24	4.19	4.12	4.07	4.02	3.98	3.95	3.91	3.88	3.77	3.69		
15	10.8	7.10	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.33	4.25	4.16	4.12	4.07	4.02	3.98	3.95	3.91	3.88	3.77	3.69	3.58	3.52	3.39		
16	10.6	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	4.18	4.10	4.03	3.97	3.92	3.87	3.83	3.80	3.76	3.73	3.62	3.54	3.44	3.37	3.25		
17	10.4	7.35	6.16	5.50	5.07	4.78	4.46	4.26	4.05	3.97	3.90	3.84	3.79	3.75	3.70	3.67	3.64	3.61	3.59	3.51	3.43	3.31	3.25	3.12	3.07		
18	10.2	7.21	6.03	5.37	5.06	4.76	4.44	4.24	4.03	3.94	3.86	3.79	3.73	3.68	3.64	3.60	3.56	3.53	3.50	3.48	3.40	3.32	3.20	3.14	3.01		
19	10.1	7.09	5.92	5.27	4.95	4.65	4.34	4.18	4.04	3.93	3.84	3.76	3.70	3.64	3.59	3.54	3.50	3.46	3.43	3.39	3.32	3.20	3.12	3.02	2.96		
20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.76	3.68	3.61	3.55	3.50	3.46	3.42	3.38	3.35	3.32	3.20	3.12	3.02	2.96	2.78		
21	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	3.68	3.60	3.54	3.48	3.43	3.38	3.34	3.31	3.27	3.24	3.13	3.05	2.95	2.88	2.75	2.68	
22	9.74	6.73	6.57	6.11	5.61	5.22	4.84	4.47	4.11	3.84	3.70	3.61	3.52	3.41	3.32	3.24	3.18	3.12	3.07	3.02	2.97	2.92	2.88	2.82	2.76	2.62	
23	9.63	6.73	6.58	6.15	5.67	5.26	4.86	4.48	4.19	3.90	3.75	3.64	3.55	3.47	3.37	3.25	3.20	3.18	3.12	3.06	3.02	2.95	2.91	2.87	2.81	2.75	
24	9.55	6.66	6.52	6.09	5.62	5.20	4.80	4.42	4.12	3.90	3.73	3.63	3.55	3.47	3.37	3.25	3.20	3.16	3.12	3.09	3.06	3.01	2.97	2.92	2.87	2.81	
25	9.48	6.60	6.46	6.04	5.61	5.19	5.34	4.94	4.63	4.35	4.15	3.94	3.78	3.64	3.54	3.45	3.37	3.26	3.20	3.15	3.11	3.08	3.04	3.01	2.97	2.92	
26	9.41	6.54	6.41	6.11	5.79	5.38	5.02	4.73	4.41	4.10	3.89	3.73	3.60	3.49	3.40	3.33	3.26	3.20	3.15	3.11	3.07	3.03	2.97	2.92	2.87	2.81	
27	9.34	6.49	6.36	6.14	5.74	5.36	5.05	4.76	4.44	4.13	3.89	3.74	3.60	3.49	3.40	3.33	3.26	3.20	3.15	3.11	3.07	3.03	2.99	2.96	2.93	2.88	
28	9.28	6.44	6.32	6.07	5.70	5.31	5.01	4.71	4.39	4.08	3.81	3.65	3.52	3.41	3.32	3.26	3.18	3.12	3.07	3.03	2.99	2.95	2.92	2.87	2.82	2.77	
29	9.23	6.40	6.28	5.96	5.56	5.16	4.86	4.56	4.26	3.96	3.78	3.68	3.58	3.48	3.39	3.29	3.21	3.15	3.09	3.04	2.99	2.95	2.92	2.88	2.83	2.78	
30	9.18	6.35	6.24	5.92	5.52	5.12	4.82	4.52	4.22	3.92	3.73	3.63	3.53	3.43	3.34	3.25	3.18	3.11	3.06	3.01	2.96	2.92	2.88	2.83	2.78	2.73	
31	9.09	6.28	6.17	5.86	5.46	5.17	4.89	4.60	4.30	4.01	3.74	3.54	3.43	3.32	3.22	3.12	3.03	2.96	2.90	2.85	2.81	2.77	2.73	2.69	2.65	2.62	
32	9.09	6.28	6.17	5.86	5.46	5.17	4.89	4.60	4.30	4.01	3.74	3.54	3.43	3.32	3.22	3.12	3.03	2.96	2.90	2.85	2.81	2.77	2.73	2.69	2.65	2.62	
33	8.94	6.16	5.96	5.61	5.20	4.81	4.51	4.21	3.91	3.62	3.34	3.15	3.04	2.93	2.82	2.72	2.62	2.57	2.53	2.47	2.43	2.37	2.33	2.27	2.23	2.19	
34	8.91	6.22	6.11	5.81	5.41	5.01	4.72	4.42	4.12	3.83	3.54	3.34	3.15	3.04	2.93	2.82	2.72	2.68	2.64	2.60	2.57	2.52	2.47	2.43	2.38	2.34	
35	8.88	6.11	5.92	5.62	5.22	4.82	4.52	4.22	3.92	3.63	3.34	3.15	3.04	2.93	2.82	2.72	2.62	2.57	2.53	2.49	2.44	2.39	2.35	2.30	2.26	2.22	
36	8.84	6.07	5.98	5.67	5.27	4.87	4.57	4.27	3.97	3.68	3.39	3.19	3.08	2.97	2.86	2.75	2.71	2.67	2.63	2.59	2.55	2.51	2.47	2.43	2.38	2.34	
37	8.83	6.07	5.98	5.67	5.27	4.87	4.57	4.27	3.97	3.68	3.39	3.19	3.08	2.97	2.86	2.75	2.71	2.67	2.63	2.59	2.55	2.51	2.47	2.43	2.38	2.34	
38	8.88	6.07	5.98	5.67	5.27	4.87	4.57	4.27	3.97	3.68	3.39	3.19	3.08	2.97	2.86	2.75	2.71	2.67	2.63	2.59	2.55	2.51	2.47	2.43	2.38	2.34	
39	8.88	6.07	5.98	5.67	5.27	4.87	4.57	4.27	3.97	3.68	3.39	3.19	3.08	2.97	2.86	2.75	2.71	2.67	2.63	2.59	2.55	2.51	2.47	2.43	2.38	2.34	
40	8.83	6.07	5.98	5.67	5.27	4.87	4.57	4.27	3.97	3.68	3.39	3.19	3.08	2.97	2.86	2.75	2.71	2.67	2.63	2.59	2.55	2.51	2.47	2.43	2.38	2.34	
41	8.78	6.03	5.94	5.64	5.24	4.83	4.53	4.23	3.93	3.64	3.35	3.16	3.06	2.95	2.84	2.73	2.70	2.66	2.63	2.60	2.57	2.54	2.50	2.46	2.42	2.38	
42	8.78	6.03	5.94	5.64	5.24	4.83	4.53	4.23	3.93	3.64	3.35	3.16	3.06	2.95	2.84	2.73	2.70	2.66	2.63	2.60	2.57	2.54	2.50	2.46	2.42	2.38	
43	8.74	5.98	4.91	4.31	3.92	3.60	3.42	3.26	3.14	3.03	2.94	2.87	2.80	2.75	2.70	2.65	2.61	2.58	2.54	2.51	2.47	2.43	2.39	2.35	2.31	2.27	
44	8.74	5.98	4.91	4.31	3.92	3.60	3.42	3.26	3.14	3.03	2.94	2.87	2.80	2.75	2.70	2.65	2.61	2.58	2.54	2.51	2.47	2.43	2.39	2.35	2.31	2.27	
45	8.67	5.92	4.88	4.28	3.90	3.62	3.40	3.24	3.11	3.01	2.92	2.85	2.78	2.72	2.67	2.63	2.59	2.55	2.52	2.49	2.45	2.41	2.37	2.33	2.29	2.25	
46	8.66	5.93	4.85	4.25	3.87	3.60	3.40	3.24	3.11	3.01	2.92	2.85	2.78	2.72	2.67	2.63	2.59	2.55	2.52	2.49	2.45	2.41	2.37	2.33	2.29	2.25	
47	8.63	5.90	4.83	4.23	3.85	3.58	3.38	3.22	3.09	2.99	2.92	2.85	2.78	2.72	2.67	2.63	2.59	2.55	2.52								

TABLE A.4 Percentiles of the *F* Distribution (continued)Upper 0.1% point of the *F* distribution

		DEGREES OF FREEDOM FOR NUMERATOR																										
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	40	50	100	150	200
1	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••		
2	969	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	
3	167	148	141	137	135	133	132	131	130	129	128	128	127	127	127	127	127	127	127	126	126	125	125	124	124	124	124	
4	741	612	562	53.1	51.7	49.7	46.1	44.1	41.7	40.7	39.7	38.7	37.7	36.7	35.7	34.7	33.7	32.7	31.7	30.7	30.7	30.7	30.7	30.7	30.7	30.7		
5	47.2	37.1	33.2	31.1	29.8	28.8	27.5	26.2	25.0	24.7	24.4	24.1	23.7	23.3	22.9	22.4	21.9	21.4	20.9	20.4	20.4	20.4	20.4	20.4	20.4	20.4		
6	35.5	27.0	23.7	21.9	20.8	19.5	19.0	18.7	18.4	18.2	18.0	17.8	17.7	17.6	17.4	17.3	17.2	17.1	16.9	16.7	16.3	16.0	15.9	15.9	15.9	15.9		
7	29.2	21.7	18.8	17.2	16.2	15.0	14.5	14.3	14.1	13.9	13.6	13.4	13.2	13.1	13.0	12.9	12.7	12.5	12.3	12.2	12.0	11.9	11.8	11.8	11.8			
8	25.4	18.5	15.8	14.4	13.5	12.9	12.0	11.8	11.5	11.4	11.2	11.1	10.9	10.8	10.7	10.6	10.5	10.5	10.3	10.1	9.92	9.80	9.57	9.49	9.45	9.45		
9	22.9	16.4	13.9	12.6	11.7	11.1	10.7	10.4	9.9	9.72	9.44	9.33	9.15	9.08	9.01	8.95	8.89	8.80	8.75	8.69	8.55	8.37	8.26	8.04	7.96	7.93		
10	21.0	14.9	12.6	11.3	10.5	9.93	9.52	9.20	8.96	8.75	8.59	8.45	8.32	8.22	8.13	8.05	7.98	7.91	7.86	7.80	7.74	7.30	7.19	6.98	6.91	6.87		
11	19.7	13.8	11.6	10.3	9.58	9.05	8.66	8.35	8.12	7.92	7.76	7.63	7.51	7.41	7.32	7.24	7.17	7.11	7.06	7.01	6.88	6.52	6.42	6.21	6.14	6.10		
12	18.6	13.0	10.8	9.68	8.89	8.38	8.00	7.71	7.48	7.29	7.14	7.00	6.89	6.79	6.71	6.63	6.57	6.51	6.45	6.40	6.22	6.09	5.93	5.83	5.56	5.52		
13	17.8	12.3	10.2	9.07	8.35	7.86	7.49	7.21	6.98	6.80	6.65	6.52	6.41	6.31	6.23	6.16	6.09	6.03	5.98	5.93	5.75	5.63	5.47	5.37	5.17	5.10		
14	17.1	11.8	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40	6.26	6.13	6.02	5.93	5.85	5.78	5.71	5.66	5.60	5.56	5.38	5.25	5.10	4.80	4.74	4.71		
15	16.6	11.3	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08	5.94	5.81	5.67	5.54	5.46	5.35	5.29	5.25	5.07	4.95	4.80	4.70	4.51	4.44	4.41	4.41		
16	16.1	11.0	9.01	7.94	7.27	6.80	6.46	6.19	5.98	5.81	5.67	5.55	5.44	5.35	5.27	5.20	5.14	5.09	5.00	4.99	4.87	4.82	4.76	4.60	4.48	4.34		
17	15.7	10.7	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58	5.44	5.32	5.22	5.13	5.05	4.99	4.92	4.87	4.82	4.74	4.65	4.54	4.46	4.36	4.19	4.16		
18	15.1	10.2	8.49	7.46	6.81	6.35	6.02	5.76	5.59	5.39	5.25	5.13	5.04	4.94	4.87	4.76	4.70	4.64	4.58	4.52	4.47	4.33	4.19	4.06	3.87	3.80		
19	15.1	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08	4.94	4.82	4.72	4.64	4.56	4.49	4.44	4.38	4.33	4.29	4.12	4.00	3.86	3.77	3.58	3.51		
20	14.8	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08	4.94	4.82	4.72	4.64	4.56	4.49	4.44	4.38	4.33	4.29	4.12	4.00	3.86	3.77	3.58	3.48		
21	14.6	9.77	6.94	6.32	5.88	5.56	5.31	5.11	4.95	4.81	4.70	4.60	4.51	4.43	4.34	4.26	4.17	4.00	3.88	3.74	3.64	3.46	3.39	3.36	3.36	3.36		
22	14.4	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83	4.70	4.58	4.49	4.40	4.33	4.24	4.17	4.10	4.05	3.96	3.89	3.78	3.63	3.54	3.35	3.28	3.25	
23	14.2	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89	4.73	4.60	4.48	4.39	4.30	4.23	4.16	4.10	4.05	3.96	3.88	3.79	3.68	3.53	3.44	3.35	3.28		
24	14.0	9.34	7.55	6.59	5.98	5.55	5.23	4.98	4.80	4.64	4.51	4.39	4.30	4.21	4.14	4.07	3.96	3.92	3.87	3.71	3.62	3.53	3.45	3.36	3.27	3.19		
25	13.9	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56	4.42	4.31	4.20	4.13	4.06	3.99	3.94	3.88	3.74	3.68	3.58	3.53	3.49	3.38	3.27	3.19		
26	13.7	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64	4.48	4.35	4.24	4.14	4.06	3.99	3.92	3.86	3.81	3.77	3.72	3.66	3.54	3.44	3.30	3.21	3.20		
27	13.6	9.02	7.27	6.33	5.73	5.31	4.90	4.66	4.45	4.28	4.11	4.01	3.93	3.86	3.78	3.74	3.69	3.64	3.60	3.54	3.49	3.38	3.27	3.18	3.14	3.12		
28	13.5	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50	4.35	4.22	4.11	4.01	3.93	3.86	3.79	3.74	3.68	3.63	3.58	3.53	3.49	3.38	3.28	3.20	3.18		
29	13.4	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45	4.29	4.16	4.05	3.96	3.88	3.80	3.74	3.68	3.63	3.58	3.53	3.49	3.38	3.28	3.23	3.19	3.15		
30	13.3	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24	4.11	4.00	3.91	3.82	3.75	3.69	3.63	3.58	3.53	3.49	3.38	3.28	3.23	3.19	3.15	3.12		
31	13.1	8.64	6.94	5.64	5.07	4.68	4.38	4.16	3.93	3.78	3.60	3.59	3.54	3.50	3.45	3.39	3.33	3.28	3.23	3.18	3.14	3.10	2.98	2.89	2.79	2.69		
32	13.0	8.52	6.83	5.92	5.34	4.93	4.63	4.40	4.17	3.94	3.74	3.65	3.52	3.47	3.42	3.37	3.31	3.25	3.20	3.15	3.10	3.06	3.02	2.96	2.87	2.74		
33	12.9	8.42	6.74	5.84	5.26	4.86	4.56	4.33	4.14	3.99	3.87	3.76	3.67	3.56	3.51	3.46	3.40	3.34	3.29	3.24	3.20	3.15	3.10	3.06	3.02	2.91		
34	12.8	8.33	6.66	5.76	5.19	4.79	4.49	4.26	4.08	3.93	3.80	3.70	3.60	3.52	3.45	3.39	3.34	3.28	3.24	3.20	3.14	3.09	3.04	2.98	2.92	2.85		
35	12.7	8.23	6.56	5.66	5.07	4.69	4.37	4.16	3.96	3.86	3.75	3.64	3.55	3.47	3.40	3.34	3.28	3.23	3.19	3.14	3.08	3.03	2.97	2.91	2.85	2.78		
36	12.6	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87	3.75	3.64	3.55	3.47	3.40	3.34	3.28	3.24	3.20	3.14	3.08	3.03	2.97	2.91	2.85	2.78		
37	12.5	8.18	6.53	5.64	5.07	4.68	4.38	4.16	3.93	3.78	3.70	3.59	3.55	3.46	3.38	3.31	3.25	3.19	3.14	3.08	3.03	2.97	2.91	2.85	2.79	2.73		
38	12.4	8.06	6.42	5.54	4.98	4.59	4.30	4.07	3.84	3.74	3.62	3.51	3.43	3.34	3.27	3.21	3.14	3.08	3.02	2.96	2.91	2.85	2.79	2.73	2.67	2.61		
39	12.3	7.90	6.38	5.50	4.94	4.55	4.26	4.03	3.85	3.70	3.59	3.48	3.38	3.29	3.21	3.14	3.07	3.01	2.96	2.91	2.85	2.79	2.73	2.67	2.61	2.55		
40	12.2	7.84	6.34	5.46	4.90	4.51	4.22	4.00	3.82	3.67	3.55	3.44	3.35	3.27	3.20	3.14	3.08	3.02	2.96	2.91	2.85	2.79	2.73	2.67	2.61	2.55		
41	12.1	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.66	3.54	3.43	3.32	3.23	3.14	3.06	2.99	2.93	2.88	2.83	2.77	2.72	2.66	2.61	2.55	2.50	2.45		
42	12.0	7.64	6.06	5.20	4.66	4.28	3.98	3.77	3.60	3.45	3.33	3.23	3.14	3.06	2.99	2.93	2.88	2.83	2.78	2.74	2.67	2.62	2.57	2.52	2.47	2.42	2.39	
43	11.9	7.54	5.97	5.12	4.58	4.20	3.92	3.70	3.53	3.39	3.27	3.16	3.07	3.00	2.93	2.87	2.81	2.76	2.72	2.68	2.63	2.57	2.52	2.47	2.42	2.39	2.35	
44	11.8	7.41	6.42	5.54	4.93	4.50	4.20	3.91	3.74	3.57	3.46	3.34	3.24	3.17	3.10	3.03	2.97	2.91	2.84	2.78	2.73	2.68	2.63	2.59	2.54	2.50	2.45	
45	11.7	7.24	5.71	4.88	4.35	3.98	3.67	3.40	3.21	3.02	2.87	2.73	2.61	2.51	2.43	2.36	2.29	2.24	2.19	2.14	2.10	2.06	2.01	1.97	1.93	1.87	1.81	
46	11.6	7.30	5.77	4.93	4.40	4.03	3.75	3.54	3.37	3.23	3.11	3.00	2.92	2.84	2.77	2.71	2.66	2.61	2.56	2.52	2.46	2.41	2.36	2.32	2.27	2.22	2.17	2.12
47	11.5	7.20	5.63	4.81	4.35	3.98	3.67</td																					

TABLE A.5 Values of $\frac{1}{2} \ln \frac{1+r}{1-r}$

<i>r</i>	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.000	0.0000	0.0010	0.0020	0.0030	0.0040	0.0050	0.0060	0.0070	0.0080	0.0090
0.010	0.0100	0.0110	0.0120	0.0130	0.0140	0.0150	0.0160	0.0170	0.0180	0.0190
0.020	0.0200	0.0210	0.0220	0.0230	0.0240	0.0250	0.0260	0.0270	0.0280	0.0290
0.030	0.0300	0.0310	0.0320	0.0330	0.0340	0.0350	0.0360	0.0370	0.0380	0.0390
0.040	0.0400	0.0410	0.0420	0.0430	0.0440	0.0450	0.0460	0.0470	0.0480	0.0490
0.050	0.0501	0.0511	0.0521	0.0531	0.0541	0.0551	0.0561	0.0571	0.0581	0.0591
0.060	0.0601	0.0611	0.0621	0.0631	0.0641	0.0651	0.0661	0.0671	0.0681	0.0691
0.070	0.0701	0.0711	0.0721	0.0731	0.0741	0.0751	0.0761	0.0771	0.0782	0.0792
0.080	0.0802	0.0812	0.0822	0.0832	0.0842	0.0852	0.0862	0.0872	0.0882	0.0892
0.090	0.0902	0.0912	0.0922	0.0933	0.0943	0.0953	0.0963	0.0973	0.0983	0.0993
0.100	0.1003	0.1013	0.1024	0.1034	0.1044	0.1054	0.1064	0.1074	0.1084	0.1094
0.110	0.1105	0.1115	0.1125	0.1135	0.1145	0.1155	0.1165	0.1175	0.1185	0.1195
0.120	0.1206	0.1216	0.1226	0.1236	0.1246	0.1257	0.1267	0.1277	0.1287	0.1297
0.130	0.1308	0.1318	0.1328	0.1338	0.1348	0.1358	0.1368	0.1379	0.1389	0.1399
0.140	0.1409	0.1419	0.1430	0.1440	0.1450	0.1460	0.1470	0.1481	0.1491	0.1501
0.150	0.1511	0.1522	0.1532	0.1542	0.1552	0.1563	0.1573	0.1583	0.1593	0.1604
0.160	0.1614	0.1624	0.1634	0.1644	0.1655	0.1665	0.1676	0.1686	0.1696	0.1706
0.170	0.1717	0.1727	0.1737	0.1748	0.1758	0.1768	0.1779	0.1789	0.1799	0.1810
0.180	0.1820	0.1830	0.1841	0.1851	0.1861	0.1872	0.1882	0.1892	0.1903	0.1913
0.190	0.1923	0.1934	0.1944	0.1954	0.1965	0.1975	0.1986	0.1996	0.2007	0.2017
0.200	0.2027	0.2038	0.2048	0.2059	0.2069	0.2079	0.2090	0.2100	0.2111	0.2121
0.210	0.2132	0.2142	0.2153	0.2163	0.2174	0.2184	0.2194	0.2205	0.2215	0.2226
0.220	0.2237	0.2247	0.2258	0.2268	0.2279	0.2289	0.2300	0.2310	0.2321	0.2331
0.230	0.2342	0.2353	0.2363	0.2374	0.2384	0.2395	0.2405	0.2416	0.2427	0.2437
0.240	0.2448	0.2458	0.2469	0.2480	0.2490	0.2501	0.2511	0.2522	0.2533	0.2543
0.250	0.2554	0.2565	0.2575	0.2586	0.2597	0.2608	0.2618	0.2629	0.2640	0.2650
0.260	0.2661	0.2672	0.2682	0.2693	0.2704	0.2715	0.2726	0.2736	0.2747	0.2758
0.270	0.2769	0.2779	0.2790	0.2801	0.2812	0.2823	0.2833	0.2844	0.2855	0.2866
0.280	0.2877	0.2888	0.2898	0.2909	0.2920	0.2931	0.2942	0.2953	0.2964	0.2975
0.290	0.2986	0.2997	0.3008	0.3019	0.3029	0.3040	0.3051	0.3062	0.3073	0.3084
0.300	0.3095	0.3106	0.3117	0.3128	0.3139	0.3150	0.3161	0.3172	0.3183	0.3195
0.310	0.3206	0.3217	0.3228	0.3239	0.3250	0.3261	0.3272	0.3283	0.3294	0.3305
0.320	0.3317	0.3328	0.3339	0.3350	0.3361	0.3372	0.3384	0.3395	0.3406	0.3417
0.330	0.3428	0.3439	0.3451	0.3462	0.3473	0.3484	0.3496	0.3507	0.3518	0.3530
0.340	0.3541	0.3552	0.3564	0.3575	0.3586	0.3597	0.3609	0.3620	0.3632	0.3643
0.350	0.3654	0.3666	0.3677	0.3689	0.3700	0.3712	0.3723	0.3734	0.3746	0.3757
0.360	0.3769	0.3780	0.3792	0.3803	0.3815	0.3826	0.3838	0.3850	0.3861	0.3873
0.370	0.3884	0.3896	0.3907	0.3919	0.3931	0.3942	0.3954	0.3966	0.3977	0.3989
0.380	0.4001	0.4012	0.4024	0.4036	0.4047	0.4059	0.4071	0.4083	0.4094	0.4106
0.390	0.4118	0.4130	0.4142	0.4153	0.4165	0.4177	0.4189	0.4201	0.4213	0.4225
0.400	0.4236	0.4248	0.4260	0.4272	0.4284	0.4296	0.4308	0.4320	0.4332	0.4344
0.410	0.4356	0.4368	0.4380	0.4392	0.4404	0.4416	0.4429	0.4441	0.4453	0.4465
0.420	0.4477	0.4489	0.4501	0.4513	0.4526	0.4538	0.4550	0.4562	0.4574	0.4587
0.430	0.4599	0.4611	0.4623	0.4636	0.4648	0.4660	0.4673	0.4685	0.4697	0.4710
0.440	0.4722	0.4735	0.4747	0.4760	0.4772	0.4784	0.4797	0.4809	0.4822	0.4835
0.450	0.4847	0.4860	0.4872	0.4885	0.4897	0.4910	0.4923	0.4935	0.4948	0.4961
0.460	0.4973	0.4986	0.4999	0.5011	0.5024	0.5037	0.5049	0.5062	0.5075	0.5088
0.470	0.5101	0.5114	0.5126	0.5139	0.5152	0.5165	0.5178	0.5191	0.5204	0.5217
0.480	0.5230	0.5243	0.5256	0.5279	0.5282	0.5295	0.5308	0.5321	0.5334	0.5347
0.490	0.5361	0.5374	0.5387	0.5400	0.5413	0.5427	0.5440	0.5453	0.5466	0.5480

TABLE A.5 Values of $\frac{1}{2} \ln \frac{1+r}{1-r}$ (continued)

<i>r</i>	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.500	0.5493	0.5506	0.5520	0.5533	0.5547	0.5560	0.5573	0.5587	0.5600	0.5614
0.510	0.5627	0.5641	0.5654	0.5668	0.5681	0.5695	0.5709	0.5722	0.5736	0.5750
0.520	0.5763	0.5777	0.5791	0.5805	0.5818	0.5832	0.5846	0.5860	0.5874	0.5888
0.530	0.5901	0.5915	0.5929	0.5943	0.5957	0.5971	0.5985	0.5999	0.6013	0.6027
0.540	0.6042	0.6056	0.6070	0.6084	0.6098	0.6112	0.6127	0.6141	0.6155	0.6170
0.550	0.6184	0.6198	0.6213	0.6227	0.6241	0.6256	0.6270	0.6285	0.6299	0.6314
0.560	0.6328	0.6343	0.6358	0.6372	0.6387	0.6401	0.6416	0.6431	0.6446	0.6460
0.570	0.6475	0.6490	0.6505	0.6520	0.6535	0.6550	0.6565	0.6579	0.6594	0.6610
0.580	0.6625	0.6640	0.6655	0.6670	0.6685	0.6700	0.6715	0.6731	0.6746	0.6761
0.590	0.6777	0.6792	0.6807	0.6823	0.6838	0.6854	0.6869	0.6885	0.6900	0.6916
0.600	0.6931	0.6947	0.6963	0.6978	0.6994	0.7010	0.7026	0.7042	0.7057	0.7073
0.610	0.7089	0.7105	0.7121	0.7137	0.7153	0.7169	0.7185	0.7201	0.7218	0.7234
0.620	0.7250	0.7266	0.7283	0.7299	0.7315	0.7332	0.7348	0.7364	0.7381	0.7398
0.630	0.7414	0.7431	0.7447	0.7464	0.7481	0.7497	0.7514	0.7531	0.7548	0.7565
0.640	0.7582	0.7599	0.7616	0.7633	0.7650	0.7667	0.7684	0.7701	0.7718	0.7736
0.650	0.7753	0.7770	0.7788	0.7805	0.7823	0.7840	0.7858	0.7875	0.7893	0.7910
0.660	0.7928	0.7946	0.7964	0.7981	0.7999	0.8017	0.8035	0.8053	0.8071	0.8089
0.670	0.8107	0.8126	0.8144	0.8162	0.8180	0.8199	0.8217	0.8236	0.8254	0.8273
0.680	0.8291	0.8310	0.8328	0.8347	0.8366	0.8385	0.8404	0.8423	0.8442	0.8461
0.690	0.8480	0.8499	0.8518	0.8537	0.8556	0.8576	0.8595	0.8614	0.8634	0.8653
0.700	0.8673	0.8693	0.8712	0.8732	0.8752	0.8772	0.8792	0.8812	0.8832	0.8852
0.710	0.8872	0.8892	0.8912	0.8933	0.8953	0.8973	0.8994	0.9014	0.9035	0.9056
0.720	0.9076	0.9097	0.9118	0.9139	0.9160	0.9181	0.9202	0.9223	0.9245	0.9266
0.730	0.9287	0.9309	0.9330	0.9352	0.9373	0.9395	0.9417	0.9439	0.9461	0.9483
0.740	0.9505	0.9527	0.9549	0.9571	0.9594	0.9616	0.9639	0.9661	0.9684	0.9707
0.750	0.9730	0.9752	0.9775	0.9799	0.9822	0.9845	0.9868	0.9892	0.9915	0.9939
0.760	0.9962	0.9986	1.0010	1.0034	1.0058	1.0082	1.0106	1.0130	1.0154	1.0179
0.770	1.0203	1.0228	1.0253	1.0277	1.0302	1.0327	1.0352	1.0378	1.0403	1.0428
0.780	1.0454	1.0479	1.0505	1.0531	1.0557	1.0583	1.0609	1.0635	1.0661	1.0688
0.790	1.0714	1.0741	1.0768	1.0795	1.0822	1.0849	1.0876	1.0903	1.0931	1.0958
0.800	1.0986	1.1014	1.1041	1.1070	1.1098	1.1127	1.1155	1.1184	1.1212	1.1241
0.810	1.1270	1.1299	1.1329	1.1358	1.1388	1.1417	1.1447	1.1477	1.1507	1.1538
0.820	1.1568	1.1599	1.1630	1.1660	1.1692	1.1723	1.1754	1.1786	1.1817	1.1849
0.830	1.1870	1.1913	1.1946	1.1979	1.2011	1.2044	1.2077	1.2111	1.2144	1.2178
0.840	1.2212	1.2246	1.2280	1.2315	1.2349	1.2384	1.2419	1.2454	1.2490	1.2526
0.850	1.2561	1.2598	1.2634	1.2670	1.2708	1.2744	1.2782	1.2819	1.2857	1.2895
0.860	1.2934	1.2972	1.3011	1.3050	1.3089	1.3129	1.3168	1.3209	1.3249	1.3290
0.870	1.3331	1.3372	1.3414	1.3456	1.3498	1.3540	1.3583	1.3626	1.3670	1.3714
0.880	1.3758	1.3802	1.3847	1.3892	1.3938	1.3984	1.4030	1.4077	1.4124	1.4171
0.890	1.4219	1.4268	1.4316	1.4366	1.4415	1.4465	1.4516	1.4566	1.4618	1.4670
0.900	1.4722	1.4775	1.4828	1.4883	1.4937	1.4992	1.5047	1.5103	1.5160	1.5217
0.910	1.5275	1.5334	1.5393	1.5453	1.5513	1.5574	1.5636	1.5698	1.5762	1.5825
0.920	1.5890	1.5956	1.6022	1.6089	1.6157	1.6226	1.6296	1.6366	1.6438	1.6510
0.930	1.6584	1.6659	1.6734	1.6811	1.6888	1.6967	1.7047	1.7129	1.7211	1.7295
0.940	1.7380	1.7467	1.7555	1.7645	1.7736	1.7828	1.7923	1.8019	1.8117	1.8216
0.950	1.8318	1.8421	1.8527	1.8635	1.8745	1.8857	1.8972	1.9090	1.9210	1.9333
0.960	1.9459	1.9588	1.9721	1.9857	1.9996	2.0140	2.0287	2.0439	2.0595	2.0756
0.970	2.0923	2.1095	2.1273	2.1457	2.1649	2.1847	2.2054	2.2269	2.2494	2.2729
0.980	2.2976	2.3223	2.3507	2.3796	2.4101	2.4426	2.4774	2.5147	2.5550	2.5988
0.990	2.6467	2.6996	2.7587	2.8257	2.9031	2.9945	3.1063	3.2504	3.4534	3.8002

$$\begin{array}{cc} r & z \\ 0.9999 & 4.95172 \\ 0.99999 & 6.10303 \end{array}$$

Source: Albert E. Waugh, *Statistical Tables and Problems*, McGraw-Hill Book Company, New York, 1952, Table A11, pp. 40–41, with the kind permission of the author and publisher.

Note: To obtain $\frac{1}{2} \ln(1+r)/(1-r)$ when r is negative, use the negative of the tabulated value corresponding to the absolute value of r ; e.g., if $r = -0.241$, $\frac{1}{2} \ln(1-0.242)/(1+0.242) = -0.2469$

Table A.6 Upper α Point of Studentized Range, $q_{k,v} = R/S$, k = sample size for range R , v = number of degrees of freedom for S (entry = $q_{k,v,1-\alpha}$, where $P(q_{k,v} > q_{k,v,1-\alpha}) = \alpha$)

$\alpha = 0.05$

v	k	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	17.97	26.98	32.82	37.09	40.41	43.12	45.40	47.36	49.07	50.59	51.96	53.20	54.33	55.37	56.32	57.21	58.04	58.82	59.56	
2	6.08	8.33	9.80	10.88	11.73	12.43	13.03	13.54	13.99	14.39	14.75	15.08	15.37	15.65	15.91	16.14	16.36	16.57	16.77	
3	4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46	9.72	9.95	10.15	10.35	10.52	10.69	10.84	10.98	11.11	11.24	
4	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83	8.03	8.21	8.37	8.52	8.66	8.79	8.91	9.03	9.13	9.23	
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17	7.32	7.47	7.60	7.72	7.83	7.93	8.03	8.12	8.21	
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.65	6.79	6.92	7.03	7.14	7.24	7.34	7.43	7.51	7.59	
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30	6.43	6.55	6.66	6.76	6.85	6.94	7.02	7.10	7.17	
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05	6.18	6.29	6.39	6.48	6.57	6.65	6.73	6.80	6.87	
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.87	5.98	6.09	6.19	6.28	6.36	6.44	6.51	6.58	6.64	
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72	5.83	5.93	6.03	6.11	6.19	6.27	6.34	6.40	6.47	
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61	5.71	5.81	5.90	5.98	6.06	6.13	6.20	6.27	6.33	
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	5.51	5.61	5.71	5.80	5.88	5.95	6.02	6.09	6.15	6.21	
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43	5.53	5.63	5.71	5.79	5.86	5.93	5.99	6.05	6.11	
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36	5.46	5.55	5.64	5.71	5.79	5.85	5.91	5.97	6.03	
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31	5.40	5.49	5.57	5.65	5.72	5.78	5.85	5.90	5.96	
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26	5.44	5.52	5.59	5.66	5.73	5.79	5.84	5.89	5.90	
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.21	5.31	5.39	5.47	5.54	5.61	5.67	5.73	5.79	5.84	
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17	5.27	5.35	5.43	5.50	5.57	5.63	5.69	5.74	5.79	
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14	5.23	5.31	5.39	5.46	5.53	5.59	5.65	5.70	5.75	
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11	5.20	5.28	5.36	5.43	5.49	5.55	5.61	5.66	5.71	
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	4.92	5.00	5.08	5.15	5.21	5.27	5.33	5.38	5.43	5.47	
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.82	4.90	4.98	5.04	5.11	5.16	5.22	5.27	5.31	5.36	
50	2.84	3.42	3.76	4.00	4.19	4.34	4.47	4.58	4.68	4.77	4.85	4.92	4.98	5.04	5.10	5.15	5.20	5.24	5.29	
100	2.81	3.36	3.70	3.93	4.11	4.26	4.38	4.48	4.58	4.66	4.73	4.80	4.86	4.92	4.97	5.02	5.07	5.11	5.15	
150	2.79	3.35	3.67	3.90	4.08	4.23	4.35	4.45	4.54	4.62	4.70	4.76	4.82	4.88	4.93	4.98	5.02	5.06	5.10	
1000	2.78	3.32	3.64	3.86	4.04	4.18	4.3	4.4	4.48	4.56	4.63	4.7	4.75	4.81	4.86	4.9	4.95	4.99	5.03	

Table A.6 Upper α Point of Studentized Range (*continued*) $\alpha = .01$

v	k	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	90.03	135.04	164.26	185.58	202.21	215.78	227.18	236.97	245.55	253.15	259.98	266.17	271.82	277.01	281.81	286.27	290.44	294.34	298.01	
2	14.04	19.02	22.29	24.72	26.63	28.20	29.53	30.68	31.69	32.59	33.40	34.13	34.81	35.43	36.00	36.53	37.03	37.50	37.94	
3	8.26	10.62	12.17	13.32	14.24	15.00	15.64	16.20	16.69	17.13	17.52	17.88	18.22	18.52	18.80	19.07	19.31	19.54	19.76	
4	6.51	8.12	9.17	9.96	10.58	11.10	11.54	11.93	12.26	12.57	12.84	13.09	13.32	13.53	13.72	13.91	14.08	14.24	14.39	
5	5.70	6.98	7.80	8.42	8.91	9.32	9.67	9.97	10.24	10.48	10.70	10.89	11.08	11.24	11.40	11.55	11.68	11.81	11.93	
6	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30	9.48	9.65	9.81	9.95	10.08	10.21	10.32	10.43	10.54	
7	4.95	5.92	6.54	7.00	7.37	7.68	7.94	8.17	8.37	8.55	8.71	8.86	9.00	9.12	9.24	9.35	9.46	9.55	9.65	
8	4.75	5.64	6.20	6.62	6.96	7.24	7.47	7.68	7.86	8.03	8.18	8.31	8.44	8.55	8.66	8.76	8.85	8.94	9.03	
9	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.32	7.49	7.65	7.78	7.91	8.03	8.13	8.23	8.32	8.41	8.49	8.57	
10	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36	7.48	7.60	7.71	7.81	7.91	7.99	8.08	8.15	8.23	
11	4.39	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13	7.25	7.36	7.46	7.56	7.65	7.73	7.81	7.88	7.95	
12	4.32	5.05	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94	7.06	7.17	7.26	7.36	7.44	7.52	7.59	7.66	7.73	
13	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79	6.90	7.00	7.10	7.19	7.27	7.34	7.42	7.48	7.55	
14	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66	6.77	6.87	6.96	7.05	7.12	7.20	7.27	7.33	7.39	
15	4.17	4.84	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55	6.66	6.76	6.84	6.93	7.00	7.07	7.14	7.20	7.26	
16	4.13	4.79	5.19	5.49	5.72	5.91	6.08	6.22	6.35	6.46	6.56	6.66	6.74	6.82	6.90	6.97	7.03	7.09	7.15	
17	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38	6.48	6.57	6.66	6.73	6.81	6.87	6.94	7.00	7.05	
18	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31	6.41	6.50	6.58	6.65	6.73	6.79	6.85	6.91	6.97	
19	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25	6.34	6.43	6.51	6.58	6.65	6.72	6.78	6.84	6.89	
20	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19	6.28	6.37	6.45	6.52	6.59	6.65	6.71	6.77	6.82	
30	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85	5.93	6.01	6.08	6.14	6.20	6.26	6.31	6.36	6.41	
40	3.82	4.37	4.69	4.93	5.11	5.26	5.39	5.50	5.60	5.69	5.76	5.83	5.90	5.96	6.02	6.07	6.12	6.16	6.21	
50	3.79	4.32	4.63	4.86	5.04	5.19	5.31	5.41	5.51	5.59	5.67	5.73	5.80	5.85	5.91	5.96	6.01	6.05	6.09	
100	3.71	4.22	4.52	4.73	4.90	5.03	5.14	5.24	5.33	5.40	5.47	5.54	5.59	5.65	5.70	5.74	5.79	5.83	5.86	
150	3.69	4.18	4.48	4.69	4.85	4.98	5.09	5.19	5.27	5.34	5.41	5.47	5.53	5.58	5.63	5.67	5.71	5.75	5.79	
1000	3.65	4.13	4.41	4.62	4.77	4.90	5.00	5.09	5.17	5.24	5.31	5.37	5.42	5.47	5.51	5.56	5.59	5.63	5.67	

TABLE A.7 Orthogonal Polynomial Coefficients

k	POLYNOMIAL	X										$(\sum p_i^2)$
		1	2	3	4	5	6	7	8	9	10	
3	Linear	-1	0	1								2
	Quadratic	1	-2	1								6
4	Linear	-3	-1	1	3							20
	Quadratic	1	-1	-1	1							4
	Cubic	-1	3	-3	1							20
5	Linear	-2	-1	0	1	2						10
	Quadratic	2	-1	-2	-1	2						14
	Cubic	-1	2	0	-2	1						10
	Quartic	1	-4	6	-4	1						70
6	Linear	-5	-3	-1	1	3	5					70
	Quadratic	5	-1	-4	-4	-1	5					84
	Cubic	-5	7	4	-4	-7	5					180
	Quartic	1	-3	2	2	-3	1					28
	Quintic	-1	5	-10	10	-5	1					252
7	Linear	-3	-2	-1	0	1	2	3				28
	Quadratic	5	0	-3	-4	-3	0	5				84
	Cubic	-1	1	1	0	-1	-1	1				6
	Quartic	3	-7	1	6	1	-7	3				154
	Quintic	-1	4	-5	0	5	-4	1				84
	Sextic	1	-6	15	-20	15	-6	1				924
8	Linear	-7	-5	-3	-1	1	3	5	7			168
	Quadratic	7	1	-3	-5	-5	-3	1	7			168
	Cubic	-7	5	7	3	-3	-7	-5	7			264
	Quartic	7	-13	-3	9	9	-3	-13	7			616
	Quintic	-7	23	-17	-15	15	17	-23	7			2,184
	Sextic	1	-5	9	-5	-5	9	-5	1			264
	Septic	-1	7	-21	35	-35	21	-7	1			3,432
9	Linear	-4	-3	-2	-1	0	1	2	3	4		60
	Quadratic	28	7	-8	-17	-20	-17	-8	7	28		2,772
	Cubic	-14	7	13	9	0	-9	-13	-7	14		990
	Quartic	14	-21	-11	9	18	9	-11	-21	14		2,002
	Quintic	-4	11	-4	-9	0	9	4	-11	4		468
	Sextic	4	-17	22	1	-20	1	22	-17	4		1,980
	Septic	-1	6	-14	14	0	-14	14	-6	1		858
	Octic	1	-8	28	-56	70	-56	28	-8	1		12,870
10	Linear	-9	-7	-5	-3	-1	1	3	5	7	9	330
	Quadratic	6	2	-1	-3	-4	-4	-3	-1	2	6	132
	Cubic	-42	14	35	31	12	-12	-31	-35	-14	42	8,580
	Quartic	18	-22	-17	3	18	18	3	-17	-22	18	2,860
	Quintic	-6	14	-1	-11	-6	6	11	1	-14	6	780
	Sextic	3	-11	10	6	-8	-8	6	10	11	3	660
	Septic	-9	47	-86	92	56	-56	-42	86	-47	9	29,172
	Octic	1	-7	20	-28	14	14	-28	20	-7	1	2,860
	Novic	-1	9	-36	84	-126	126	-84	36	-9	1	48,620

**TABLE A.8A Bonferroni Corrected Jackknife
Residual Critical Values**

$\alpha = .05$		n = 5										n = 10										n = 15										n = 20										n = 25										n = 30										n = 35										n = 40										n = 45										n = 50										n = 55										n = 60										n = 65										n = 70										n = 75										n = 80									
1	22.33	5.41	4.55	4.29	4.17	4.03	4.06	4.15	4.27	4.40	1	1.73	2.54	2.87	3.06	3.19	3.51	3.77	3.99	4.18	4.35	2	1.41	2.45	2.83	3.03	3.17	3.51	3.77	3.99	4.18	4.35	3	2.33	2.78	3.01	3.15	3.51	3.77	3.99	4.18	4.35	4	2.18	2.72	2.98	3.14	3.50	3.77	3.99	4.18	4.35	5	5	1.98	2.65	2.94	3.11	3.50	3.77	3.99	4.18	4.35	6	6	1.73	2.57	2.91	3.09	3.49	3.77	3.99	4.18	4.35	7	7	1.41	2.47	2.86	3.07	3.49	3.76	3.99	4.18	4.35	8	8	2.35	2.81	3.04	3.48	3.76	3.98	4.18	4.35	9	9	2.19	2.75	3.01	3.47	3.76	3.98	4.18	4.35	10	10	1.99	2.68	3.07	3.38	3.61	3.81	3.99	4.17	4.35	15	15	1.99	2.70	3.43	3.75	3.98	4.17	4.35	4.35	20	20	1.99	2.74	3.38	3.74	3.98	4.17	4.35	4.35	40	40	2.75	3.69	3.97	4.17	4.35	4.35	4.35	4.35	80	80	4.97	4.21	4.28	4.40	4.31	3.94	4.17	4.34					

TABLE A.8B Bonferroni Corrected Studentized Residual Critical Values

$\alpha = .01$		k = 1										k = 2										k = 3										k = 4										k = 5										k = 6										k = 7										k = 8										k = 9										k = 10										k = 11										k = 12										k = 13										k = 14										k = 15										k = 16										k = 17										k = 18										k = 19										k = 20										k = 21										k = 22										k = 23										k = 24										k = 25										k = 26										k = 27										k = 28										k = 29										k = 30										k = 31										k = 32										k = 33										k = 34										k = 35										k = 36										k = 37										k = 38										k = 39										k = 40										k = 41										k = 42										k = 43										k = 44										k = 45										k = 46										k = 47										k = 48										k = 49										k = 50										k = 51										k = 52										k = 53										k = 54										k = 55										k = 56										k = 57										k = 58										k = 59										k = 60										k = 61										k = 62										k = 63										k = 64										k = 65										k = 66										k = 67										k = 68										k = 69										k = 70										k = 71										k = 72										k = 73										k = 74										k = 75										k = 76										k = 77										k = 78										k = 79										k = 80									
1	22.33	5.41	4.55	4.29																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													

TABLE A.9 Critical Values for Leverages, n = sample size, k = number of predictors $\alpha = .10$

n	k	1	2	3	4	5	6	7	8	9	10	15	20	40	80
10	0.626	0.759	0.847	0.911	0.956	0.984	0.997	1.000	0.959	0.980	0.988				
15	0.481	0.595	0.679	0.748	0.806	0.855	0.897	0.932	0.851	0.883	0.918				
20	0.394	0.491	0.565	0.627	0.682	0.731	0.775	0.815	0.751	0.784	0.837				
25	0.335	0.419	0.484	0.540	0.589	0.635	0.676	0.715	0.669	0.701	0.837				
30	0.293	0.366	0.424	0.474	0.519	0.560	0.599	0.635	0.669	0.701	0.837				
40	0.236	0.295	0.342	0.383	0.420	0.455	0.487	0.518	0.547	0.576	0.806				
60	0.172	0.214	0.248	0.279	0.306	0.332	0.356	0.380	0.402	0.424	0.524				
80	0.137	0.170	0.197	0.221	0.242	0.263	0.283	0.301	0.319	0.337	0.418				
100	0.114	0.141	0.164	0.183	0.201	0.219	0.235	0.250	0.266	0.280	0.348				
200	0.064	0.079	0.091	0.102	0.111	0.121	0.130	0.138	0.146	0.155	0.227				
400	0.036	0.043	0.050	0.055	0.060	0.065	0.070	0.075	0.079	0.083	0.104				
800	0.020	0.024	0.027	0.030	0.032	0.035	0.037	0.040	0.042	0.044	0.055				

 $\alpha = .05$

n	k	1	2	3	4	5	6	7	8	9	10	15	20	40	80
10	0.683	0.802	0.879	0.933	0.969	0.990	0.999	1.000	0.946	0.969	0.986				
15	0.531	0.639	0.719	0.782	0.835	0.880	0.916	0.939	0.872	0.901	0.911				
20	0.436	0.531	0.602	0.662	0.714	0.761	0.802	0.839	0.807	0.831	0.934				
25	0.372	0.454	0.518	0.573	0.621	0.665	0.705	0.742	0.776	0.807	0.947				
30	0.325	0.398	0.455	0.505	0.549	0.589	0.627	0.662	0.695	0.726	0.855				
40	0.261	0.321	0.368	0.409	0.446	0.480	0.512	0.543	0.572	0.600	0.722				
60	0.190	0.233	0.268	0.298	0.326	0.352	0.376	0.400	0.422	0.444	0.543				
80	0.151	0.185	0.212	0.236	0.258	0.279	0.299	0.318	0.336	0.353	0.435				
100	0.126	0.154	0.176	0.196	0.215	0.232	0.248	0.264	0.279	0.294	0.363				
200	0.070	0.085	0.098	0.108	0.119	0.128	0.137	0.146	0.154	0.162	0.236				
400	0.039	0.047	0.053	0.059	0.064	0.069	0.074	0.079	0.083	0.088	0.162				
800	0.021	0.025	0.029	0.032	0.034	0.037	0.039	0.042	0.044	0.046	0.057				

 $\alpha = .01$

n	k	1	2	3	4	5	6	7	8	9	10	15	20	40	80
10	0.785	0.875	0.930	0.965	0.986	0.997	1.000	1.000	0.969	0.984	0.994				
15	0.629	0.724	0.792	0.844	0.887	0.921	0.948	0.969	0.883	0.910	0.933				
20	0.524	0.612	0.677	0.731	0.777	0.817	0.852	0.883	0.824	0.851	0.953				
25	0.450	0.529	0.589	0.640	0.685	0.724	0.761	0.794	0.824	0.851	0.989				
30	0.394	0.466	0.521	0.568	0.610	0.648	0.683	0.716	0.746	0.774	0.889				
40	0.318	0.377	0.424	0.464	0.501	0.534	0.565	0.595	0.622	0.649	0.763				
60	0.231	0.275	0.310	0.341	0.369	0.395	0.420	0.443	0.465	0.487	0.584				
80	0.183	0.218	0.246	0.271	0.283	0.314	0.334	0.353	0.372	0.389	0.471				
100	0.152	0.181	0.205	0.225	0.244	0.262	0.279	0.295	0.310	0.325	0.394				
200	0.085	0.100	0.113	0.124	0.135	0.145	0.154	0.163	0.172	0.180	0.219				
400	0.046	0.054	0.061	0.067	0.073	0.078	0.083	0.088	0.092	0.097	0.138				
800	0.025	0.029	0.033	0.036	0.039	0.041	0.044	0.046	0.049	0.051	0.062				

TABLE A.10 Critical Values for the Maximum of n Values of Cook's ($n - k - 1$) d_i (Bonferroni correction used), n observations and k predictors

$\alpha = 0.1$

k	$n = 5$	10	15	20	25	50	100	200	400	800
1	14.96	11.13	11.84	12.68	13.46	16.39	19.97	23.94	28.70	33.80
2	40.53	12.21	12.09	12.63	13.22	15.65	18.64	22.09	25.96	30.12
3		13.30	12.09	12.35	12.79	14.84	17.48	20.52	23.86	27.50
4		15.21	12.18	12.14	12.45	14.23	16.62	19.36	22.30	25.97
5		19.33	12.44	12.03	12.21	13.76	15.95	18.49	21.39	24.51
6		31.06	12.94	12.01	12.04	13.39	15.43	17.81	20.36	23.51
7		96.01	13.79	12.08	11.94	13.10	15.02	17.27	19.75	22.42
8			15.26	12.26	11.90	12.85	14.70	16.83	19.20	21.73
9			18.00	12.55	11.91	12.66	14.40	16.52	18.62	21.45
10			23.93	13.02	11.97	12.50	14.16	16.16	18.43	20.55
15				27.66	13.60	12.01	13.39	15.16	17.00	19.34
20					30.94	11.83	12.92	14.53	16.31	18.35
40						15.95	12.26	13.56	15.10	16.83
80							13.49	13.05	14.39	15.85

$\alpha = 0.05$

k	$n = 5$	10	15	20	25	50	100	200	400	800
1	24.97	15.24	15.55	16.37	17.18	20.41	24.31	28.83	33.88	40.15
2	82.06	16.56	15.63	16.01	16.56	19.08	22.33	26.05	30.20	33.96
3		18.16	15.50	15.49	15.85	17.93	20.72	24.14	27.57	32.06
4		21.28	15.59	15.14	15.33	17.06	19.63	22.49	25.83	29.31
5		28.40	15.94	14.95	14.96	16.41	18.70	21.39	24.42	28.24
6		50.22	16.70	14.91	14.70	15.91	17.97	20.54	23.48	26.68
7		192.90	17.99	15.00	14.55	15.50	17.49	20.00	22.35	25.67
8			20.32	15.25	14.48	15.19	17.05	19.31	22.06	24.44
9			24.78	15.69	14.49	14.92	16.69	18.85	21.34	24.29
10			34.72	16.38	14.58	14.70	16.38	18.42	20.49	23.33
15				39.98	16.94	14.03	15.36	17.16	19.39	21.75
20					44.63	13.79	14.81	16.52	18.46	20.32
40						19.50	13.92	15.22	16.83	18.76
80							15.55	14.58	15.99	17.52

$\alpha = 0.01$

p	$n = 5$	10	15	20	25	50	100	200	400	800
1	77.29	28.72	26.88	27.24	27.92	31.46	36.10	41.22	49.42	68.39
2	415.27	30.97	26.13	25.65	25.81	28.12	32.61	37.34	44.99	57.70
3		35.12	25.66	24.22	24.33	26.17	29.15	34.23	37.55	52.58
4		44.09	25.82	23.58	23.20	24.56	27.31	31.26	35.28	40.60
5		66.83	26.66	23.20	22.49	23.39	25.84	29.44	34.14	36.91
6		150.47	28.48	23.12	22.00	22.55	24.35	28.42	31.04	36.91
7		964.09	31.80	23.34	21.71	21.79	24.19	26.87	31.04	33.55
8			37.84	23.93	21.59	21.26	23.28	25.83	29.31	33.55
9			50.10	24.93	21.64	20.76	22.23	25.62	28.21	30.50
10			80.67	26.54	21.83	20.37	22.11	24.53	28.21	30.50
15				92.09	27.02	19.16	20.22	22.40	25.64	27.73
20					102.32	18.82	19.18	21.32	23.31	25.21
40						29.95	18.04	19.32	21.17	22.91
80							20.67	18.57	20.12	22.90

References

- Cook, R. D., and Weisberg, S. 1982. *Residuals and Influence in Regression*. New York: Chapman & Hall.
- Jensen, D. R., and Ramirez, D. E. 1996. "Computing the CDF of Cook's D(I) Statistic." In A. Prat and E. Ripoll, eds., *Proceedings of the 12th Symposium in Computational Statistics*, pp. 65–66. Barcelona, Spain: Institut d'Estadistica de Catalunya.
- Muller, K. E., and Chen Mok, M. 1997. "The Distribution of Cook's Statistic." *Communications in Statistics: Theory and Methods* 26(3): 525–46.
- Obenchain, R. L. 1977. Letter to the Editor. *Technometrics* 19: 348–49.

B

Appendix—Matrices and Their Relationship to Regression Analysis

B.1 Preview

Statisticians have found matrix mathematics to be a very useful vehicle for compactly presenting the concepts, methods, and formulas of regression analysis and other multivariable methods. Moreover, matrix formulation of such topics has had the important practical implication of permitting extremely efficient and accurate use of the computer for carrying out multivariable analyses on large data sets.

This appendix will summarize some of the more elementary but important notions and manipulations of matrix algebra and will use this tool to describe the general least-squares procedures of multiple regression. Admittedly, the material in this appendix is somewhat more mathematical than that in the main body of the text, but it is not absolutely necessary knowledge for the applied user of the multivariable methods we describe. Nevertheless, the reader who is comfortable with the mathematical level used here should find the matrix formulation of regression analysis a powerful and unifying supplement that may facilitate the learning of more advanced multivariable methods.

B.2 Definitions

A *matrix* may be simply defined as a rectangular array of numbers. For example,

$$\mathbf{A} = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 1 & 2 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 2 & 1 \\ 3 & 2 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}$$

are all matrices.

The *dimensions* of a matrix are the number of rows and the number of columns that it has. For example, the matrix **A** above has two rows and three columns:

$$\begin{array}{ccc} \text{column} & \text{column} & \text{column} \\ 1 & 2 & 3 \\ \downarrow & \downarrow & \downarrow \\ \text{row 1} \rightarrow & \left[\begin{matrix} 2 & 3 & 1 \end{matrix} \right] \\ \text{row 2} \rightarrow & \left[\begin{matrix} 1 & 1 & 2 \end{matrix} \right] \end{array}$$

It is customary to say that **A** is a 2×3 matrix or to write $\mathbf{A}_{2 \times 3}$. The dimensions of the matrices **B** and **C** above are 2×2 and 3×1 , respectively. An example of a 1×4 matrix is any matrix with one row and four columns, such as $\mathbf{D}_{1 \times 4} = [-2 \ 3 \ 3 \ 0]$. Matrices that contain only one row or only one column are often referred to as *vectors*; thus, the matrices

$$\mathbf{C} = \begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix} \quad \text{and} \quad \mathbf{D} = [-2 \ 3 \ 3 \ 0]$$

are examples of a column vector and a row vector, respectively. Also, the matrix **B** is called a *square matrix* because it has the same number of rows and columns.

The numbers forming the rectangular array of a matrix are called the *elements* of the matrix. If we let a_{ij} denote the element in the i th row and j th column of the matrix **A** above,

$$\begin{aligned} a_{11} &= 2, & a_{12} &= 3, & a_{13} &= 1 \\ a_{21} &= 1, & a_{22} &= 1, & a_{23} &= 2 \end{aligned}$$

It is often informative to write

$$\mathbf{A}_{2 \times 3} = ((a_{ij}))$$

indicating that the matrix **A** (represented by a capital letter) with two rows and three columns has typical element a_{ij} (represented by the corresponding lowercase letters). Thus, if you were given that

$$\mathbf{B}_{3 \times 2} = ((b_{ij}))$$

where $b_{21} = 3$, $b_{31} = 2$, $b_{11} = -2$, $b_{12} = 6$, $b_{22} = 0$, $b_{32} = 1$, you should construct **B** to be

$$\mathbf{B} = \begin{bmatrix} -2 & 6 \\ 3 & 0 \\ 2 & 1 \end{bmatrix}$$

B.3 Matrices in Regression Analysis

Given any set of multivariable data suitable for a regression analysis, a number of key matrices can be defined that directly correspond to the basic components of the regression model being postulated. For example, consider the following $n = 12$ pairs of observations on $Y = \text{WGT}$ and $X = \text{HGT}$:

Variable	Child											
	1	2	3	4	5	6	7	8	9	10	11	12
$Y(\text{WGT})$	64	71	53	67	55	58	77	57	56	51	76	68
$X(\text{HGT})$	57	59	49	62	51	50	55	48	42	42	61	57

We can, in correspondence with the straight-line regression model $Y = \beta_0 + \beta_1 X + E$, define four matrices: \mathbf{Y} , the vector of observations on Y ; \mathbf{X} , the matrix of independent variables; $\boldsymbol{\beta}$, the vector of parameters to be estimated; and \mathbf{E} , the vector of errors.

For the data given, these matrices are defined as follows:

$$\mathbf{Y}_{12 \times 1} = \begin{bmatrix} 64 \\ 71 \\ 53 \\ 67 \\ 55 \\ 58 \\ 77 \\ 57 \\ 56 \\ 51 \\ 76 \\ 68 \end{bmatrix}, \quad \mathbf{X}_{12 \times 2} = \begin{bmatrix} 1 & 57 \\ 1 & 59 \\ 1 & 49 \\ 1 & 62 \\ 1 & 51 \\ 1 & 50 \\ 1 & 55 \\ 1 & 48 \\ 1 & 42 \\ 1 & 42 \\ 1 & 61 \\ 1 & 57 \end{bmatrix}, \quad \boldsymbol{\beta}_{2 \times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \mathbf{E}_{12 \times 1} = \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \\ E_6 \\ E_7 \\ E_8 \\ E_9 \\ E_{10} \\ E_{11} \\ E_{12} \end{bmatrix}$$

Notice that the first column of the \mathbf{X} matrix of independent variables contains only 1's. This is the general convention used for any regression model containing a constant term (or intercept) β_0 ; motivation for adopting this convention follows by imagining the β_0 term to be of the form $\beta_0 X_0$, where X_0 is a dummy variable always taking the value 1. The vector of errors \mathbf{E} contains random (and unobservable) error values, one for each pair of observations; these errors represent the differences between the observed Y -values and their (unknown) expected values under the given model.

B.4 Transpose of a Matrix

The transpose \mathbf{A}' of a matrix \mathbf{A} is defined to be that matrix whose (i, j) th element a'_{ij} is equal to the (j, i) th element of \mathbf{A} . For example, if

$$\mathbf{A} = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

then

$$\mathbf{A}' = \begin{bmatrix} 2 & 1 \\ 3 & 1 \\ 1 & 2 \end{bmatrix}$$

since $a'_{11} = a_{11} = 2$, $a'_{12} = a_{21} = 1$, $a'_{21} = a_{12} = 3$, $a'_{22} = a_{22} = 1$, $a'_{31} = a_{13} = 1$, $a'_{32} = a_{23} = 2$. Another way of looking at this is that the first column of \mathbf{A} becomes the first row of \mathbf{A}' , the second column of \mathbf{A} becomes the second row of \mathbf{A}' , and so on. Thus, if \mathbf{A} is $r \times c$, then \mathbf{A}' is $c \times r$.

As examples, the transposes of

$$\mathbf{A}_{3 \times 2} = \begin{bmatrix} 1 & 2 \\ 3 & 1 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{B}_{3 \times 3} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 4 & -5 \\ 2 & -5 & 3 \end{bmatrix}, \quad \mathbf{C}_{4 \times 1} = \begin{bmatrix} 0 \\ 1 \\ 2 \\ -2 \end{bmatrix}$$

are

$$\mathbf{A}'_{2 \times 3} = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 1 & 1 \end{bmatrix}, \quad \mathbf{B}'_{3 \times 3} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 4 & -5 \\ 2 & -5 & 3 \end{bmatrix}, \quad \mathbf{C}'_{1 \times 4} = [0 \ 1 \ 2 \ -2]$$

Also, the transpose of the matrix $\mathbf{X}_{12 \times 2}$ of the previous section is given by

$$\mathbf{X}'_{2 \times 12} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 57 & 59 & 49 & 62 & 51 & 50 & 55 & 48 & 42 & 42 & 61 & 57 \end{bmatrix}$$

In the examples above, note that the matrix \mathbf{B} is such that $\mathbf{B} = \mathbf{B}'$ (equality here means that corresponding elements are equal). A matrix satisfying this condition is said to be a *symmetric matrix*. Note that a symmetric matrix \mathbf{A} must always be square, since otherwise \mathbf{A} and \mathbf{A}' would have different dimensions and so could not possibly be equal. A necessary and sufficient condition for the square matrix $\mathbf{A} = ((a_{ij}))$ to be symmetric is that $a_{ij} = a_{ji}$ for every $i \neq j$.

Correlation matrices such as

$$\mathbf{R}_1 = \begin{bmatrix} 1 & r_{XY} \\ r_{XY} & 1 \end{bmatrix} \quad \text{or} \quad \mathbf{R}_2 = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{bmatrix}$$

are always symmetric.

An important special case where the above condition for symmetry is satisfied is when $a_{ij} = a_{ji} = 0$ for every $i \neq j$. A square matrix having this property is said to be a *diagonal matrix*, the general form of which is given (for the 3×3 case) by

$$\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}$$

Diagonal

The most often used diagonal matrix is the *identity matrix* \mathbf{I} , which has 1's on the diagonal; for example, the 3×3 identity matrix is

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

We will see shortly that an identity matrix serves the same algebraic function for matrix multiplication that the number 1 serves for ordinary scalar multiplication.

B.5 Matrix Addition

The sum of two matrices—say, \mathbf{A} and \mathbf{B} —is obtained by adding together the corresponding elements of each matrix. Clearly, such addition can be performed only when the two matrices have the same dimensions. Thus, for example, we can add the matrices

$$\mathbf{A}_{2 \times 3} = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 1 & 2 \end{bmatrix} \quad \text{and} \quad \mathbf{B}_{2 \times 3} = \begin{bmatrix} 4 & 1 & 5 \\ 1 & 3 & 1 \end{bmatrix}$$

(since they have the same dimensions) to get

$$\mathbf{A}_{2 \times 3} + \mathbf{B}_{2 \times 3} = \begin{bmatrix} 2 + 4 & 3 + 1 & 1 + 5 \\ 1 + 1 & 1 + 3 & 2 + 1 \end{bmatrix} = \begin{bmatrix} 6 & 4 & 6 \\ 2 & 4 & 3 \end{bmatrix}$$

We could not, however, add the matrices \mathbf{A} and \mathbf{C} , where

$$\mathbf{C}_{3 \times 2} = \begin{bmatrix} 5 & 4 \\ 1 & 4 \\ 2 & 6 \end{bmatrix}$$

since the dimensions of these matrices are not the same.

An example of a more abstract use of matrix addition would be to sum the two 12×1 vectors

$$\mathbf{D}_{12 \times 1} = \begin{bmatrix} \beta_0 + 57\beta_1 \\ \beta_0 + 59\beta_1 \\ \vdots \\ \beta_0 + 57\beta_1 \end{bmatrix} \quad \text{and} \quad \mathbf{E}_{12 \times 1} = \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_{12} \end{bmatrix}$$

to obtain

$$\mathbf{D}_{12 \times 1} + \mathbf{E} = \begin{bmatrix} \beta_0 + 57\beta_1 + E_1 \\ \beta_0 + 59\beta_1 + E_2 \\ \vdots \\ \beta_0 + 57\beta_1 + E_{12} \end{bmatrix}$$

Actually, you may recognize that each element of the matrix $\mathbf{D} + \mathbf{E}$ is obtained by substituting for X in the right side of the straight-line regression equation

$$Y = \beta_0 + \beta_1 X + E$$

each of the 12 X (HGT) values given in the data set of Section B.3.

B.6 Matrix Multiplication

Multiplication of two matrices is somewhat more complicated than addition. The first rule to remember is that the product \mathbf{AB} of two matrices \mathbf{A} and \mathbf{B} can exist if and only if the *number of columns of A is equal to the number of rows of B*. Thus, if \mathbf{A} is 2×3 and \mathbf{B} is 3×4 , the product \mathbf{AB} exists since \mathbf{A} has 3 columns and \mathbf{B} has 3 rows. However, the product \mathbf{BA} does not exist, since the number of columns of \mathbf{B} (i.e., 4) is not equal to the number of rows of \mathbf{A} (i.e., 2). Notationally, therefore, a matrix product can exist only if the dimensions of the matrices can be represented as follows:

$$\begin{array}{ccc} \text{Equal numbers} & & \\ \downarrow & \downarrow & \\ \mathbf{A}_{m \times n} \times \mathbf{B}_{n \times p} & = & \mathbf{AB}_{m \times p} \\ \swarrow & \nearrow & \\ \text{Product dimensions} & & \end{array}$$

Note also from the expression above that the dimensions $m \times p$ of the product matrix \mathbf{AB} are given by the number of rows of the *premultiplier* (i.e., \mathbf{A}) and the number of columns of the *postmultiplier* (i.e., \mathbf{B}). For example, if \mathbf{A} is 2×3 and \mathbf{B} is 3×4 , the dimensions of the product \mathbf{AB} are 2×4 .

Now, to carry out matrix multiplication, consider the two matrices

$$\mathbf{A}_{2 \times 3} = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 3 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{B}_{3 \times 2} = \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 3 & 2 \end{bmatrix}$$

Since \mathbf{A} is 2×3 and \mathbf{B} is 3×2 , the product \mathbf{AB} will be a 2×2 matrix. If we let the elements of \mathbf{AB} be denoted by

$$\mathbf{AB}_{2 \times 2} = ((c_{ij})) = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

we can obtain the upper-left-hand corner element c_{11} by working with the first row of \mathbf{A} and the first column of \mathbf{B} , as follows:

$$\begin{array}{ccc} \mathbf{A} & \mathbf{B} & \mathbf{AB} \\ \left[\begin{array}{ccc} 2 & 1 & 0 \end{array} \right] & \left[\begin{array}{cc} 1 & -2 \\ 1 & 0 \\ 3 & 2 \end{array} \right] & \left[\begin{array}{cc} (2 \times 1) + (1 \times 1) & c_{12} \\ + (0 \times 3) = 3 & c_{21} \\ c_{22} \end{array} \right] \end{array}$$

What we have done here is to calculate the product of each element in row 1 of **A** with the corresponding element in column 1 of **B**, and then add up these three products to obtain $c_{11} = 3$:

$$\begin{array}{c} \text{Column 1 of } \mathbf{B} \\ c_{11} = (2 \times 1) + (1 \times 1) + (0 \times 3) = 3 \\ \text{Row 1 of } \mathbf{A} \end{array}$$

To find the element in the second row and first column of **AB** (i.e., c_{21}), we work with the second row of **A** and the first column of **B** as follows:

$$\begin{array}{ccc} \mathbf{A} & \mathbf{B} & \mathbf{AB} \\ \left[\begin{array}{ccc} 2 & 1 & 0 \\ 0 & 3 & 1 \end{array} \right] \left[\begin{array}{cc} 1 & -2 \\ 1 & 0 \\ 3 & 2 \end{array} \right] & = & \left[\begin{array}{c} 3 \\ (0 \times 1) + (3 \times 1) \\ + (1 \times 3) = 6 \end{array} \right] \\ & & \begin{array}{l} c_{12} \\ c_{22} \end{array} \end{array}$$

Thus, for the element c_{21} , we find

$$\begin{array}{c} \text{Column 1 of } \mathbf{B} \\ c_{21} = (0 \times 1) + (3 \times 1) + (1 \times 3) = 6 \\ \text{Row 2 of } \mathbf{A} \end{array}$$

Continuing this process, we find

$$c_{12} = (2 \times -2) + (1 \times 0) + (0 \times 2) = -4$$

$$c_{22} = (0 \times -2) + (3 \times 0) + (1 \times 2) = 2$$

Thus,

$$\mathbf{AB}_{2 \times 2} = \left[\begin{array}{ccc} 2 & 1 & 0 \\ 0 & 3 & 1 \end{array} \right] \left[\begin{array}{cc} 1 & -2 \\ 1 & 0 \\ 3 & 2 \end{array} \right] = \left[\begin{array}{cc} 3 & -4 \\ 6 & 2 \end{array} \right]$$

In general, if $\mathbf{A}_{m \times n} = ((a_{ij}))$ and $\mathbf{B}_{n \times p} = ((b_{ij}))$, the (i, j) th element c_{ij} of the product

$$\mathbf{AB}_{m \times p} = ((c_{ij}))$$

is defined to be

$$c_{ij} = \sum_{l=1}^n a_{il} b_{lj} \quad i = 1, 2, \dots, m; \quad j = 1, 2, \dots, p$$

Thus, as another example, if

$$\mathbf{A}_{2 \times 2} = \begin{bmatrix} -1 & 3 \\ 2 & 2 \end{bmatrix} \quad \text{and} \quad \mathbf{B}_{2 \times 1} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

then $m = 2$, $p = 1$, $n = 2$, and

$$c_{11} = \sum_{l=1}^2 a_{1l}b_{l1} = (-1 \times 0) + (3 \times 1) = 3$$

$$c_{21} = \sum_{l=1}^2 a_{2l}b_{l1} = (2 \times 0) + (2 \times 1) = 2$$

so that

$$\mathbf{AB}_{2 \times 2} = \begin{bmatrix} -1 & 3 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

Here are a few other examples for additional practice:

1. Find \mathbf{AI} and \mathbf{IA} , where

$$\mathbf{A}_{2 \times 2} = \begin{bmatrix} -1 & 3 \\ 2 & 2 \end{bmatrix} \quad \text{and} \quad \mathbf{I}_{2 \times 2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

2. Find $\mathbf{X}'\mathbf{X}$, where

$$\mathbf{X}_{3 \times 2} = \begin{bmatrix} 1 & 10 \\ 1 & 15 \\ 1 & 20 \end{bmatrix}$$

The answer to problem 1 is

$$\mathbf{AI} = \mathbf{IA} = \mathbf{A}$$

which indicates why \mathbf{I} is generally referred to as the identity matrix, since, like the scalar identity 1, the product of any square matrix (e.g., \mathbf{A}) with an appropriate identity matrix (of the right dimensions) will always yield the original matrix \mathbf{A} , whether \mathbf{I} premultiplies or postmultiplies \mathbf{A} .

The answer to problem 2 is

$$\mathbf{X}'\mathbf{X}_{2 \times 2} = \begin{bmatrix} 1 & 1 & 1 \\ 10 & 15 & 20 \end{bmatrix} \begin{bmatrix} 1 & 10 \\ 1 & 15 \\ 1 & 20 \end{bmatrix} = \begin{bmatrix} 3 & 45 \\ 45 & 725 \end{bmatrix}$$

which is a symmetric matrix, as will be the case whenever any matrix is pre- or postmultiplied by its own transpose.

B.7 Inverse of a Matrix

The definition of an *inverse matrix* parallels the basic property of the *reciprocal* of an ordinary (scalar) number. That is, for any nonzero number a , its reciprocal $1/a$ satisfies the equation

$$a \times \frac{1}{a} = \frac{1}{a} \times a = 1$$

In words, when a is either pre- or postmultiplied by its reciprocal, the result is the scalar 1. Analogously, we say that a square matrix \mathbf{A} has an inverse \mathbf{A}^{-1} if and only if

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

That is, *the product of \mathbf{A} by its inverse must be equal to an identity matrix*. We hasten to add at this point that *only the inverses of square matrices are being considered*. Thus, if \mathbf{A} is $n \times n$, \mathbf{A}^{-1} must also be $n \times n$.

We shall not describe here any of the many algorithms available for actually computing the inverse of a matrix.¹ For most practical applications, one can use standard computer packages for such computations. We wish to emphasize instead that one important attribute of inverses in statistical analyses is that their use permits the solution of matrix equations for a matrix of unknowns (e.g., the vector $\boldsymbol{\beta}$ of unknown regression parameters) in a manner similar to how division is used in ordinary algebra. Most specifically, in regression analysis, the use of an inverse is crucial because it provides a means for efficiently solving the least-squares equations, as well as providing compact formulas for additional components of the analysis (e.g., the variances of and covariances among the estimated regression coefficients).

Some examples of matrix inverses are given as follows:

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 1 \\ 1 & -1 & 2 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{A}^{-1} = \begin{bmatrix} 0 & 0 & 1 \\ 2 & -1 & -3 \\ 1 & 0 & -2 \end{bmatrix}$$

(The reader can check this out by multiplying \mathbf{A} and \mathbf{A}^{-1} together to obtain $\mathbf{I}_{3 \times 3}$.) The inverse of

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix}$$

is

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \frac{\sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2} & \frac{-\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ \frac{-\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} & \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \end{bmatrix}$$

¹For further details, see N. R. Draper and H. Smith, *Applied Regression Analysis* (New York: Wiley, 1966); W. Mendenhall, *Introduction to Linear Models and the Design and Analysis of Experiments* (Belmont, Calif.: Wadsworth, 1968).

B.8 Matrix Formulation of Regression Analysis

We have previously seen in Section B.3 that when fitting a straight-line model

$$Y = \beta_0 + \beta_1 X + E$$

to a set of data consisting of n pairs of observations on the variables X and Y , we can define several matrices to characterize the regression problem under consideration: \mathbf{Y} , the $n \times 1$ vector of observations on Y ; \mathbf{X} , the $n \times 2$ matrix of independent variables; $\boldsymbol{\beta}$, the 2×1 vector of parameters; and \mathbf{E} , the $n \times 1$ vector of random errors.

In general, whenever we are considering a regression problem involving p independent variables using a model such as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + E$$

we can analogously construct appropriate matrices based on the multivariable data set being considered. Thus, if the data on the i th individual consist of the $p + 1$ values

$$Y_i, X_{i1}, X_{i2}, \dots, X_{ip} \quad i = 1, 2, \dots, n$$

the following matrices can be constructed:

$$\mathbf{Y}_{n \times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{array}{l} \text{vector of} \\ \text{observations on } Y \end{array}$$

$$\mathbf{X}_{n \times (p+1)} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} = \begin{array}{l} \text{matrix of} \\ \text{independent} \\ \text{variables} \end{array}$$

$$\boldsymbol{\beta}_{(p+1) \times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \text{vector of parameters}$$

$$\mathbf{E}_{n \times 1} = \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{bmatrix} = \text{vector of random errors}$$

Using the matrices above in conjunction with the notions of matrix addition and multiplication, we can formulate the general regression model in matrix terms as follows:

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times (p+1)} \boldsymbol{\beta}_{(p+1) \times 1} + \mathbf{E}_{n \times 1} \quad (B.1)$$

This compact equation summarizes in a single statement the n equations

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + E_i \quad i = 1, 2, \dots, n$$

Note that this equivalence follows from the following matrix calculations:

$$\begin{aligned} \mathbf{X}\boldsymbol{\beta} + \mathbf{E} &= \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{bmatrix} \\ &= \begin{bmatrix} \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_p X_{1p} \\ \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_p X_{2p} \\ \vdots \\ \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_p X_{np} \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{bmatrix} \\ &= \begin{bmatrix} \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_p X_{1p} + E_1 \\ \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_p X_{2p} + E_2 \\ \vdots \\ \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_p X_{np} + E_n \end{bmatrix} \end{aligned}$$

Based on the general matrix equation given by (B.1), a description of all the essential features of regression analysis can be expressed in matrix notation. In particular, the least-squares solution for the estimates of the regression coefficients in the parameter vector $\boldsymbol{\beta}$ can now be compactly written. This least-squares solution, in matrix terms, is that vector $\hat{\boldsymbol{\beta}}$ which minimizes the error sum of squares (given in matrix terms)

$$(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

The solution to this minimization problem, which is obtained via the use of matrix calculus, yields the following easy-to-remember matrix formula:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

where $\hat{\boldsymbol{\beta}}'_{1 \times (p+1)} = [\hat{\beta}_0 \quad \hat{\beta}_1 \quad \cdots \quad \hat{\beta}_p]$ denotes the vector of estimated regression coefficients.

Thus, although we have pointed out in the text the futility of *explicitly* giving the solutions to the least-squares equations for models of more complexity than a straight line, we can at least *implicitly* express these solutions in matrix notation and, because of modern computer technology, conveniently carry through with the computation of the least-squares solutions using this matrix representation.

At this point it is not our intention to carry through completely with the matrix formulation of every other aspect of a regression analysis. The reader is referred to Draper and Smith (1966; see footnote 1) for a fuller treatment of this matrix approach. However, some additional matrix results will be summarized below to give more of an indication of the utility of the matrix approach:

1. The *vector of predicted responses* is given by $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.
2. The *error sum of squares* SSE is given by $SSE = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$.
3. The *regression sum of squares* is given by $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}' - n\bar{Y}^2$.
4. The *variances of the regression coefficients*—that is, $\sigma_{\hat{\beta}_j}^2, j = 0, 1, 2, \dots, p$ —are given by the diagonal elements of the matrix $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$.

Finally, although we do not present the details here, any test of a (linear) statistical hypothesis concerning some subset of the regression coefficients can be formulated and carried out entirely by means of matrix operations.

C

SAS Computer Appendix

C.1 Introduction

SAS is a powerful and widely used software system for conducting statistical analyses. The current version as of this edition is SAS 9.3, the PC version of which was used to conduct the majority of the analyses in this text. In this appendix we seek to provide the user who has at least a basic familiarity with SAS program syntax and structure the tools needed to conduct the major analyses described in this text. We do not attempt to provide an exhaustive guide to SAS programming or to multivariable modeling in SAS. However, the SAS 9.3 online documentation is an extremely thorough resource that we recommend consulting for any further details or information. It may be found at <http://support.sas.com/documentation/93/index.html>.

C.1.1 SAS Syntax and Structure

Almost every statistically oriented SAS program consists of a series of two elements, **DATA** steps and **PROC** (procedure) steps. Within both DATA and PROC steps, instructions (called *statements*) are specified that direct SAS to perform the desired tasks. All statements must end with a semicolon.

DATA steps contain statements used to manipulate data sets through tasks such as creating new observations (records), creating new data sets through the combination of other data sets, and creating new variables using formulas and logical evaluations that may involve variables already contained in the data. The first line of a DATA step names the data set that is to be produced, while subsequent lines may contain statements in which the contents of the data set are created.

PROC steps contain statements in which statistical or data management tasks may be performed on data sets. The procedures that we are concerned with in this appendix all

conduct statistical analyses on data sets. The first statement in a PROC step indicates which procedure is to be executed and which data set is to be utilized. Additional general display and analysis options for the PROC may appear on this line. Subsequent statements in the step are usually specific to the PROC and provide SAS with additional details necessary to conduct the relevant analysis. These statements often may be customized with additional options that are specified at the end of the line following a forward slash (/).

Both DATA steps and PROC steps usually end with the RUN statement, which instructs SAS to end the block of syntax and run the code.

C.1.2 Conventions for SAS Coding Used in This Appendix

1. Line breaks: While it is good programming style to do so, one need not place only one statement per line. Sometimes it can save considerable space to place several statements on a single line, and this does not alter SAS's processing at all. For example, consider this entire sorting routine:

```
proc sort data = my_data;      by study_id;      run;
```

On the other hand, if a line of code is becoming too long, it is permissible to hit “enter” and break the line into two or more lines, with a semicolon appearing only once on the several lines you've created. For example:

```
proc glm data = low_drinkers;
    model bmi = drink_days poor_sleep_days age
        drink_days * poor_sleep_days drink_days * age
    /solution;
run;
```

2. Case sensitivity: SAS statements are not case-sensitive, except where values of character variables are included in the statement. That is, “PROC PRINT” and “proc print” are equivalent. When discussing SAS syntax in the text body, we use uppercase characters so that the code stands out from the prose. When demonstrating segments of SAS code, we typically use lowercase characters throughout.
3. Indentation: It is a good (and often overlooked) programming habit to indent the lines of one's SAS programs in a way that demarcates the code into blocks. We use the first level for the DATA or PROC and RUN lines, and indent further for each nested level of statements. For example:

```
proc means data = my_data;
    class gender;
    var bmi;
run;
```

is preferable to

```
proc means data = my_data;
class gender;
var bmi;
run;
```

C.1.3 Output Style

SAS is capable of producing output in a variety of formats. Using the Output Delivery System (ODS), SAS can send output in a standard monospace text format to the Listing or Output window or can write the output to pdf, rtf, html, and/or a host of other file types. A full description of the ODS is beyond the scope of this appendix; an example of how to save your output in a pdf file is given here:

```
ods pdf file = "c:\xyz.pdf";
proc means data = my_data;
  class gender;
  var bmi;
run;
ods pdf close;
```

In the current release of SAS, the SAS Listing window is now turned off by default, and an on-screen html feed (with the “HTMBLue” style) is the new default output mode. The following SAS statements can be used to produce output in the format shown in earlier versions of our textbook:

```
ods listing;
ods html close;
```

C.2 Straight-line Regression

Many SAS statistical procedures may be used to fit a straight-line regression model. In the chapters and in this appendix, we illustrate the two procedures most commonly used for this analysis (as well as for multiple linear regression): **PROC GLM** and **PROC REG**. Each has its strengths. The **GLM** procedure is analytically more robust, with features such as the automatic creation of dummy variables for categorical predictors (**CLASS** statement) and the specification of linear sums of predictors (**ESTIMATE** and **CONTRAST** statements). On the other hand, the **REG** procedure readily produces more extensive model output and diagnostic plots. For standard regression tasks, the syntax and output for both procedures is quite similar, and we will primarily demonstrate these via **PROC GLM**.

Using the BRFSS example of Section 5.12, the most basic syntax for fitting a straight-line is composed of three lines:¹

```
proc glm data = low_drinkers;
  model bmi = drink_days / solution;
run;
```

¹ Compared to other SAS statistical procedures, both PROC REG and PROC GLM are unique in that they will keep “running” in the background after their syntax has been executed. This is to allow them to be run “interactively,” with the user supplying further commands to generate more model output. In practice this can be a nuisance, so we recommend following the RUN line with “QUIT.”

The first line requests that SAS run PROC GLM on the data set “low_drinkers.” The **MODEL** statement is used to specify the dependent and independent variables of the model, in the format *<y-variable> = <x-variable>*.

By default, SAS will provide for this model:

- A table of observations in the data set (called “read” data by SAS) and those with nonmissing values of the model variables (called “used” data by SAS),
- An ANOVA table (see Chapter 7),
- A table of summary statistics such as r^2 (see Chapter 6), and
- Tables of the independent variable-specific Type I (*variables-added-in-order*) and Type III (*variables-added-last*) sums of squares with *F* tests (see Chapter 9).

One may specify numerous model computation and output options after the slash on the MODEL line. The **SOLUTION** option (or just “S”) requests a table of estimates for the β ’s, their estimated standard errors, and individual independent variable hypothesis tests of the general form $H_0: \beta = 0$. In PROC REG, this last table is given by default, and it is the Type I and Type III sums of squares that need to be requested using the **SS1 SS2** options (the Type II sum of squares is a less frequently utilized quantity, and stemming from PROC REG’s inability to directly specify interaction terms, the Type II sum of squares produced from the SS2 option is exactly equal to the Type III sum of squares). The GLM output is shown below. Note that the test findings for Types I and III and the overall model are identical, since the model has only one independent variable.

The GLM Procedure

Number of Observations Read		1099			
Number of Observations Used		1056			
Dependent Variable: bmi					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1175.78658	1175.78658	34.07	<.0001
Error	1054	36377.54464	34.51380		
Corrected Total	1055	37553.33122			
R-Square		Coeff Var	Root MSE	bmi Mean	
0.031310		21.88936	5.874845	26.83882	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
drink_days	1	1175.786580	1175.786580	34.07	<.0001
Source	DF	Type III SS	Mean Square	F Value	Pr > F
drink_days	1	1175.786580	1175.786580	34.07	<.0001
Parameter	Estimate	Standard Error	t Value	Pr > t	
Intercept	27.77496308	0.24167816	114.93	<.0001	
drink_days	-0.14968629	0.02564568	-5.84	<.0001	

C.3 Correlation Coefficients

PROC CORR is the primary SAS procedure for estimating, and making statistical inferences about, the population correlation coefficient ρ between two variables (see Chapters 6 and 10). The syntax for this procedure is quite simple, requiring that the first variable be specified using the **VAR** statement and the second variable using the **WITH** statement:

```
proc corr data = low_drinkers;
    var bmi;
    with drink_days;
run;
```

Multiple variables may be specified on the VAR and WITH lines, causing SAS to compute all pairwise correlation estimates between these two sets of variables; for example,

```
proc corr data = low_drinkers;
    var bmi age;
    with drink_days poor_sleep_days;
run;
```

The output for the second CORR example is shown below. By default, SAS provides univariate statistics for all four variables. This is followed by a table that shows all possible comparisons of the VAR and WITH variables with one another. Since we specified two of each, four total pairwise comparisons are presented. In each cell of the table, CORR vertically displays, from the top, the sample correlation coefficient r between the two variables, the associated P -value for the hypothesis test $H_0: \rho = 0$, and the sample size used for these calculations.

For *BMI* and *drink_days*, note that the square of r , as well as its associated P -value, exactly matches the value for r^2 shown in the straight-line regression output above and further confirms the significant linear relationship. Thus, GLM can also be used to make certain inferences about ρ .

SIMPLE STATISTICS							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	
drink_days	1099	6.18263	7.02559	6795	1.00000	30.00000	
poor_sleep_days	1095	8.79909	9.93690	9635	0	30.00000	
bmi	1056	26.83882	5.96620	28342	15.40000	53.27000	
AGE	1099	50.57962	15.31740	55587	7.00000	94.00000	
PEARSON CORRELATION COEFFICIENTS PROB > R UNDER H0: RHO=0 NUMBER OF OBSERVATIONS							
		bmi	AGE				
drink_days		-0.17695 <.0001 1056	0.22226 <.0001 1099				
poor_sleep_days		0.07690 0.0126 1052	-0.22461 <.0001 1095				

Custom hypothesis tests of $H_0: \rho = \rho_0$ and confidence intervals using *Fisher's z transformation* may be requested by adding the FISHER option to the PROC CORR line. The exact values of ρ_0 and α can be specified in parentheses using the RHO0 and ALPHA options. For example:

```
proc corr data = low_drinkers fisher(alpha = 0.01 rho0= -0.5);
```

yields an additional table that provides a 99% confidence interval for ρ and provides two-sided P -values for each of four hypothesis tests $H_0: \rho = -0.5$:

PEARSON CORRELATION STATISTICS (FISHER'S Z TRANSFORMATION)									
Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate			H0:Rho=Rho0
							99% Confidence Limits		Rho0 p Value
bmi	drink_days	1056	-0.17695	-0.17883	-0.0000839	-0.17686	-0.252539	-0.099040	-0.50000 <.0001
AGE	drink_days	1099	0.22226	0.22604	0.0001012	0.22217	0.147056	0.294733	-0.50000 <.0001
bmi	poor_sleep_days	1052	0.07690	0.07705	0.0000366	0.07686	-0.002517	0.155277	-0.50000 <.0001
AGE	poor_sleep_days	1095	-0.22461	-0.22851	-0.0001027	-0.22452	-0.297118	-0.149334	-0.50000 <.0001

Rather than computing a simple correlation coefficient between two variables, one may also wish to compute partial correlation coefficients, as discussed in Section 10.5. In that section, PROC REG and its **PCORR1** and **PCORR2** options were used to compute Type I and III squared partial correlations. To understand the correlations between BMI and drinking days, age, and total drinks, after adjustment for the preceding or all other predictors, the syntax would be

```
proc reg data = low_drinkers ;
model bmi = drink_days age total_drinks /pcorr1 pcorr2;
run;
```

Alternatively, with PROC CORR one may directly compute the correlation between two variables, adjusting for *any* set of other variables. This is done with the addition of the **PARTIAL** statement below the **WITH** statement. First-order through p th-order partials may be generated by simply adding the control variables Z_1, Z_2, \dots, Z_p after PARTIAL. The above statistics may all be computed for these partial correlation coefficients. For example:

```
proc corr data = low_drinkers;
var bmi;
with drink_days;
partial age total_drinks;
run;
```

This yields an estimated partial correlation between *BMI* and *drink_days* of -0.12, after adjustment for *age* and *total_drinks*:

PEARSON PARTIAL CORRELATION COEFFICIENTS, N = 1906
PROB > |R| UNDER H0: PARTIAL RHO=0

bmi	
drink_days	-0.12014 <.0001

C.4 Multiple Linear Regression

In SAS, the syntax for fitting a multiple linear regression model in PROC GLM and PROC REG is essentially the same as for a simple linear regression model, except that extra independent variables are added to the right of the equal sign on the MODEL statement. For example, the BMI model with predictors of drinking days, age, and days of poor sleep from Chapter 8 would be fit using the code

```
proc glm data = low_drinkers;
    model bmi = drink_days age poor_sleep_days /solution;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1289.79375	429.93125	12.46	<.0001
Error	1045	36065.26233	34.51221		
Corrected Total	1048	37355.05608			

R-Square	Coeff Var	Root MSE	bmi Mean
0.034528	21.87405	5.874710	26.85699

Source	DF	Type I SS	Mean Square	F Value	Pr > F
drink_days	1	1073.403915	1073.403915	31.10	<.0001
AGE	1	11.595517	11.595517	0.34	0.5623
poor_sleep_days	1	204.794319	204.794319	5.93	0.0150

Source	DF	Type III SS	Mean Square	F Value	Pr > F
drink_days	1	1068.207272	1068.207272	30.95	<.0001
AGE	1	42.497569	42.497569	1.23	0.2674
poor_sleep_days	1	204.794319	204.794319	5.93	0.0150

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	26.66350677	0.70642133	37.74	<.0001
drink_days	-0.14945246	0.02686344	-5.56	<.0001
AGE	0.01404807	0.01265963	1.11	0.2674
poor_sleep_days	0.04593183	0.01885564	2.44	0.0150

Note that the output looks nearly identical to that obtained for straight-line regression. The order in which the independent variables are specified does not affect the overall model-fitting results. Yet this order is used to determine the sequence of the variables-in-order Type I sum of squares. Indeed, changing the order of the variables specified is useful for calculating custom partial (conditional) sums of squares and partial F tests, seen in regression models in Chapter 9 and in unbalanced two-way ANOVA regression models in Chapter 20. Manipulating the variable order changes their display order in the Type III sum of squares and estimated β 's tables, although this does not affect the results shown.

C.5 Dummy Variables (see Chapter 12)

There are two options available in SAS for the creation of dummy variables that represent the levels of a categorical independent variable. The first method is to manually create the new variables in one's data set. As discussed in Chapter 12, we recommend *reference cell* coding, which uses 0/1 coding and $(k - 1)$ dummy variables to represent the categories of interest. The simplest way to program reference cell coding is by using a series of if/else statements that evaluate the original categorical variable and assign values to a new one. For example, suppose we have a variable that represents three animals—dogs, cats, and mice—and wish to include it in a regression model for predicting abdominal circumference. Then we add to our data set two dummy variables: *animal_1* representing dogs and *animal_2* representing cats, with the referent group being mice:

```
data menagerie;
  set menagerie;
  if animal = "dog" then do;      animal_1 = 1; animal_2 = 0; end;
  else if animal = "cat" then do;  animal_1 = 0; animal_2 = 1; end;
  else if animal = "mouse" then do; animal_1 = 0; animal_2 = 0; end;
run;
```

To save space and enhance readability, multiple statements have been placed on one line, separated by semicolons (SAS has no problem with this). A variety of ways exist to accomplish this recoding, but the above method ensures that observations with missing values of *animal* are not inadvertently assigned values of the new dummy variables.

Once the dummy variables *animal_1* and *animal_2* have been created, simply use them on the MODEL line in PROC GLM or PROC REG as you would any other independent variables—for example,

```
proc glm data = menagerie;
  model abdomen_circ = weight animal_1 animal_2 / solution;
run;
```

To save the time needed to manually create new variables, most SAS regression procedures (except for REG) also provide the **CLASS** statement, which instructs the procedure to create dummy variables from any character or numeric variable (i.e., if one has coded/formatted numeric variables) for one or more categorical variables. These dummy variables are created only temporarily for use in the procedure. Using this syntax below, PROC GLM will detect the number of unique levels for *animal* in the data set, fit a model using $(k - 1)$

dummy variables² in place of *animal*, and display the relevant model output for these dummy variables.

```
proc glm data = menagerie;
  class animal;
  model abdomen_circ = weight animal /solution;
run;
```

An additional advantage of using the CLASS statement method is that SAS will automatically treat all dummy variables for *animal* as a single entity and provide Type I and Type III chunk tests (numerator d.f. = 2 for *F* tests with three categories of *animal*) in the appropriate tables.

Individual β estimates for each dummy variable will still be provided in a table obtained from the **SOLUTION (S)** option. SAS will decide which dummy variables represent which levels of *animal* and which is the referent level. In the table below, the first *animal* dummy variable represents “cat” and the second “dog,” while the referent is “mouse” (denoted by $\hat{\beta}$ [“Estimate”] = 0). The order of the variables and the determination of the referent group may be manipulated using the **ORDER=** option on the PROC GLM line (consult the SAS Help feature for the list and explanations of options).

Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	6.43076923	B	0.70004790	9.19	0.0003
Weight	-0.09230769		0.08675614	-1.06	0.3360
animal cat	3.80000000	B	1.10346773	3.44	0.0184
animal dog	18.33846154	B	4.15675681	4.41	0.0069
animal mouse	0.00000000	B	.	.	.

Note that the default scheme for dummy variable coding using the CLASS statement and the ways in which referent groups are selected vary between SAS procedures. For example, the logistic regression procedure (LOGISTIC) by default uses effect coding (based on values of -1/1) rather than 0/1 coding. We discuss this in the description of PROC LOGISTIC later.

C.6 Interaction (see Chapter 11)

As with the creation of dummy variables in the preceding section, we can include interaction (i.e., product) terms in the model by creating the variables either in the source data set or at the time the statistical procedure is run (except for PROC REG).

² Technically, PROC GLM creates k dummy variables coded as 0/1, but only $(k - 1)$ of these variables are used in model fitting. This distinction is important when using the **ESTIMATE** and **CONTRAST** statements in conjunction with **CLASS**—typically for an ANOVA or ANCOVA model. In these cases, values for all k dummy variables must be specified in order to estimate linear functions of model parameters. We demonstrate this in the ANOVA section.

Returning to the *low_drinkers* example, we will create a new variable that can be used to represent the interaction term in a regression model in the **DATA** step. This is done by simply assigning the product of the two variables, using an asterisk (*) to define the new product variable, and then including the new variable on the MODEL line of PROC GLM:

```
data low_drinkers;
    set low_drinkers;

    drink_sleep = drink_days * poor_sleep_days;
run;

proc glm data = low_drinkers;
    model bmi = drink_days poor_sleep_days drink_sleep /solution;
run;
```

A more convenient alternative is to have PROC GLM produce this new variable by specifying the product desired directly on the MODEL line, as shown below. As with the CLASS statement, SAS will create the new variable and include it in the analysis.

```
proc glm data = low_drinkers;
    model bmi = drink_days poor_sleep_days drink_days * poor_sleep_days
/solution;
run;
```

In the case where the interaction involves one or more CLASS variables, SAS will create new product terms involving the appropriate dummy variables it created from the use of CLASS.

C.7 Statistical Inference in Multiple Linear Regression

In Chapters 5 and 9, a variety of methods for statistical inference was presented for the linear regression model. These included the overall F and partial F (and t) tests, which are given in standard output as shown earlier. Here we discuss the conduct of other techniques using model options and the **ESTIMATE**, **CONTRAST**, and **TEST** statements found in PROC GLM and PROC REG. These statements provide overlapping functionality and also find application in the ANACOVA and ANOVA contexts, as discussed in the sections for those models. Therefore, we present here only some applications of these statements.

C.7.1 Multiple Partial F Test (see Section 9.4)

To conduct a multiple partial F test, assessing whether the addition of a set of independent variables collectively aids in predicting an outcome, PROC REG provides the **TEST** statement. All variables whose coefficients are hypothesized to equal 0 under H_0 follow the TEST statement, separated by commas.

```
proc reg data = low_drinkers;
    model bmi = drink_days age poor_sleep_days;
    test age, poor_sleep_days;
run;
```

Accordingly, the following results for the F test with two numerator degrees of freedom are appended to the output:

TEST 1 RESULTS FOR DEPENDENT VARIABLE BMI				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	96.08254	3.36	0.0350
Denominator	1889	28.61682		

PROC GLM provides no method for directly conducting the multiple partial F test. However, as illustrated by equation (9.6) in Chapter 9, one may indirectly do so by running two models, one with and one without the variables being tested. The resulting sums of squares for the “full” and “reduced” models are then used to compute a multiple partial F statistic.

C.7.2 Hypothesis Tests About Regression Coefficients (see Sections 9.3.3 and 9.6.1)

Tests of the general form $H_0: \beta^* = 0$ are provided in standard regression output as both partial F and t tests. To conduct the hypothesis test of $H_0: \beta^* = q$, where q is a particular value of interest, presented in Section 9.3.3, the TEST statement in PROC REG may be adapted to generate an F distribution–based hypothesis test. To test the null hypothesis that the slope for AGE is equal to 0.04 in the above BRFSS model, the TEST syntax would be:

```
test age = 0.04;
```

This prompts PROC REG to report a table similar to that shown above. Hypothesized values for several β 's can be simultaneously tested by using commas. For example:

```
test age = 0.04, poor_sleep_days = 0.1;
```

Although PROC GLM has no TEST statement, the parameter estimates provided by the SOLUTION option allow one to compute single β hypothesis tests, as shown in Section 9.3.3.

C.7.3 Confidence Intervals About Regression Coefficients (see Section 9.6.2)

Confidence intervals may be added to the parameter estimate tables by adding the MODEL line option CLB in PROC REG and both CLPARM and SOLUTION in PROC GLM. Confidence intervals are displayed in additional columns to the right. Additional interval estimation may be done with the ESTIMATE statement, as described further below. The REG and GLM syntaxes are, respectively,

```
model bmi = drink_days age poor_sleep_days /clb;
```

and

```
model bmi = drink_days age poor_sleep_days /solution clparm;
```

Parameter	Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	27.34824315	0.48431579	56.47	<.0001	26.39839303	28.29809326
drink_days	-0.09169169	0.01542156	-5.95	<.0001	-0.12193678	-0.06144661
AGE	0.00721213	0.00854608	0.84	0.3988	-0.00954862	0.02397287
poor_sleep_days	0.03334739	0.01295782	2.57	0.0101	0.00793425	0.05876053

C.7.4 Inference for Linear Functions of Regression Coefficients (see Section 9.6.5)

PROC GLM uses the **ESTIMATE** statement to compute linear functions (L) of the β coefficients, as well as hypothesis tests and confidence intervals concerning these linear functions. The **CONTRAST** statement in PROC GLM and the **TEST** statement in PROC REG (described earlier) may be used to produce the hypothesis tests about linear functions. **ESTIMATE** is, therefore, most versatile for multiple regression, and we describe its syntax.

Following the ESTIMATE statement, a string in quotes is required to label the linear function in the output. Anything may be put in these quotes. Then one specifies the coefficients of the variables in the linear function by typing each variable name and its coefficient in an alternating fashion. A variable whose associated β is 0 in the linear function of the β coefficients need not be entered, as all variables are assumed to have coefficients of 0 by default. For the indicated linear function, SAS will provide t distribution-based hypothesis tests of the general form $H_0: L = 0$ and confidence intervals, if the CLPARM option is specified.

For example, to compute the confidence interval for the effect of one-unit increases in both height and age discussed in Section 9.6.5, the following syntax may be used, resulting in the output shown below it.

```
proc glm data= nutr_def;
  model WGT = HGT AGE /clparm;
  estimate "1-unit increase HGT and AGE" HGT 1 AGE 1 ;
run;
```

Parameter	Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits	
1-unit increase HGT and AGE	2.73516786	7.00889522	0.39	0.7065	-13.42737349	18.89770922

Since a linear function is set equal to 0 under H_0 , some null hypotheses may require algebraic manipulation in order to be run in SAS. Consider situation #1 in Section 9.6.5: for two drugs A and B, the null hypothesis is $H_0: \beta_A = 2\beta_B$. The corresponding ESTIMATE syntax would thus look like this:

```
estimate "Test of H0: A = 2B" drug_A 1 drug_B -2 ;
```

C.7.5 The Mean Value of Y at $X_{1,0}, X_{2,0}, \dots, X_{k,0}$ (see Section 9.6.3)

A convenient method for estimating and making statistical inferences about the mean value of Y at a given combination of independent variables is to use the ESTIMATE statement, introduced in the preceding section. In this application of ESTIMATE, we are effectively interested in a specific linear sum of the β coefficients in which we input values of X for every term in the model. In addition to specifying each value of X , one must also type “intercept 1” on the ESTIMATE line. As illustrated with the nutritional deficiency example below, this will generate an estimate of \hat{Y} at $X_{1,0}, X_{2,0}, \dots, X_{k,0}$, and its standard error, as well as its 95% confidence interval.

```
proc glm data= nutr_def;
    model WGT = HGT AGE AGESQ /clparm;
    estimate "L=b0+55*b1+9*b2+81*b3" intercept 1 HGT 55 AGE 9 AGESQ 81 ;
run;
```

Parameter	Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits
L=b0+55*b1+9*b2+81*b3	64.8550177	2.07261500	31.29	<.0001	60.0755589 69.6344765

Results for the hypothesis test of $H_0: \mu_{Y|X_{1,0}, X_{2,0}, \dots, X_{k,0}} = 0$ (i.e., $H_0: L = 0$) are provided as well, but this specific test of the mean is often of little interest. Rather, one is typically interested in $H_0: \mu_{Y|X_{1,0}, X_{2,0}, \dots, X_{k,0}} = \mu_{Y|X_{1,0}, X_{2,0}, \dots, X_{k,0}}^{(0)}$, a specific nonzero value of interest. To conduct this hypothesis test, the point estimate and standard error provided by ESTIMATE may be plugged into the formulas provided in Section 9.3.3 along with the hypothesized value. Alternatively, one may directly compute this hypothesis test (but not the associated confidence interval) using the TEST statement in PROC REG, discussed earlier.

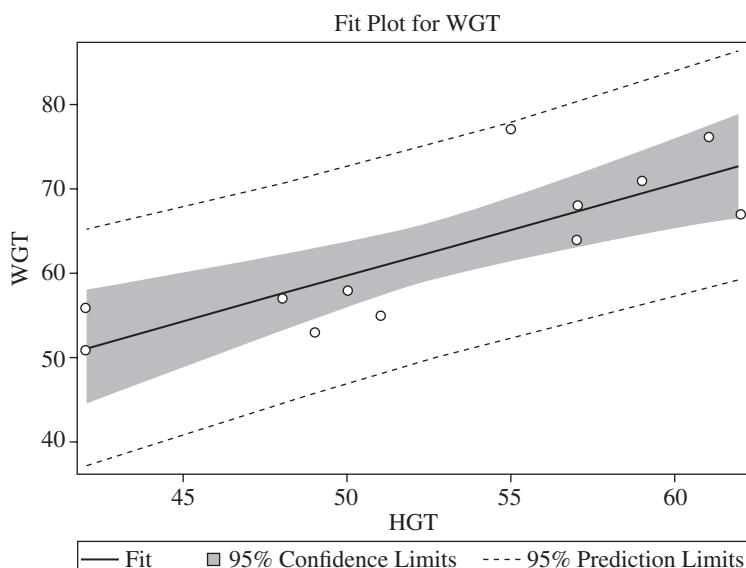
To illustrate, using the hypothesized value of 60 selected in the main text, the code would be

```
proc reg data= nutr_def;
    model WGT = HGT AGE AGESQ ;
    test intercept + 55*HGT + 9*AGE + 81*AGESQ = 60;
run;
```

In the next section, we describe a method for computing prediction intervals at values of the independent variables. This method may also be used to generate confidence intervals and is particularly useful if intervals are desired for combinations of independent variables observed in the data.

C.7.6 Confidence/Prediction Bands and Prediction of a New Value of Y at $X_{1,0}, X_{2,0}, \dots, X_{k,0}$ (see Section 9.6.4)

For straight-line regression procedures, both the GLM and the REG regression procedures now automatically output an html scatterplot for the modeled data, with the estimated regression line and both 95% confidence and prediction bands. Below is the plot generated for the regression model of $WGT = HGT$ (model 1 in Section 8.7).



To create further statistically and visually customized confidence/prediction plots, the SGPlot procedure and its REG statement are recommended.

To obtain the predicted \hat{Y} from the regression line or the values of the confidence and prediction bands at the independent variable values observed in the data, specify the CLM or CLI option, respectively, on the MODEL line. SAS will then output a table of observed and predicted values of Y , along with upper/lower interval limits for each observation (this can be a lot of output!). If you desire these model predictions and intervals for any value of an independent variable X , including a value *not* observed in the data (but within the range of values of X), follow these steps, which we illustrate from Section 5.12 of Chapter 5 for the BMI prediction interval (i.e., using the CLI option) for an individual with four drinking days/month. (*Note:* Here we are predicting Y at $X_{1,0} = 4$ when $k = 1$ in $X_{1,0}, X_{2,0}, \dots, X_{k,0}$.)

1. Create a new data set with the desired value of the independent variable(s) but missing the outcome variable so that it will not be included as a true observation by GLM and thus not affect the analysis.

```
data new_obs;
    drink_days = 4;
    BMI = .;
    output;
run;
```

2. Create a new analysis data set called *low_drinkers_new* that appends this new observation to the real analytical data set using the SET statement. It is helpful to place the data set with the new observation below *low_drinkers*:

```
data low_drinkers_new;
  set low_drinkers
    new_obs;
run;
```

3. Run PROC GLM or PROC REG on *low_drinkers_new*, specifying the CLI options.

```
proc glm data = low_drinkers_new;
  model bmi = drink_days / cli;
run;
```

4. The predicted value and the associated 95% prediction interval are printed below the table of predictions for the rest of the data set (observation 1100 signifies that this new observation is the 1100th in the data set).

Observation		Observed	Predicted	Residual	95% Confidence Limits for Individual Predicted Value
1100	*	.	27.17621793	.	15.64248170 38.70995415

C.7.7 Changing the α Level for Confidence and Prediction Intervals

By default, an α level of .05 is assumed for all interval estimation and plots generated by PROC REG and PROC GLM. To change the alpha/confidence levels produced by these procedures, the option **ALPHA** = <alpha-level> should be specified on the MODEL line. For example to produce 90% intervals for a model fit to the BRFSS data, we would specify:

```
model bmi = drink_days / alpha = .10;
```

C.8 ANACOVA

Analysis of covariance (ANACOVA) is closely related to multiple regression, and thus it may be conducted with either PROC GLM or PROC REG. However, PROC GLM's offering of both the **CLASS** and the **LSMEANS** statements makes it more ideally suited to the category-focused ANACOVA applications. Tests of coincidence, of the parallelism assumption, and of differences in least-squares means (LS-means) can be accomplished through *F* tests obtained in standard PROC GLM output and through statements described earlier.

C.8.1 Estimated Means (Least-squares Means)

We describe two methods for calculating the \bar{Y} for each group of interest, using the example of systolic blood pressure among males and females, controlling for age, discussed in Example 13.1 of Chapter 13.

The first method views this as a special case of finding the mean value of Y at specific X values, as discussed in the multiple regression context of Section 9.6.3. In this instance, we wish to obtain two such mean estimates by specifying X values of the sample-wide mean age of 46.14 years³ in conjunction with the two dummy variable values of sex (0 for males and 1 for females). This is readily done with the **ESTIMATE** statement:

```
proc glm data = table_11_1;
model sbp = sex age /solution;
estimate "Male mean SBP:" intercept 1 sex 0 age 46.14;
estimate "Female mean SBP:" intercept 1 sex 1 age 46.14;
run;
```

A more convenient second method is to use the **LSMEANS** statement, which will automatically output these adjusted means at each level of a modeled categorical variable. The categorical variable must be indicated as such using a **CLASS** statement rather than including it as a manually created dummy variable. For example,

```
proc glm data = table_11_1;
class sex;
model sbp = sex age /solution;
lsmeans sex;
run;
```

produces the additional LS-means output shown below. Note that these values are more precise than those shown in Chapter 13, which are subject to rounding error:

Sex	sbp LSMEAN
0	154.404206
1	140.890751

Confidence intervals for each \bar{Y} may be requested for the **ESTIMATE** method using the **CLPARM** option on the **MODEL** line and for the **LSMEANS** method by specifying the **CL** option on the **LSMEANS** line. That is:

```
model sbp = sex age /solution clparm;
estimate "Male mean SBP:" intercept 1 sex 0 age 46.14;
```

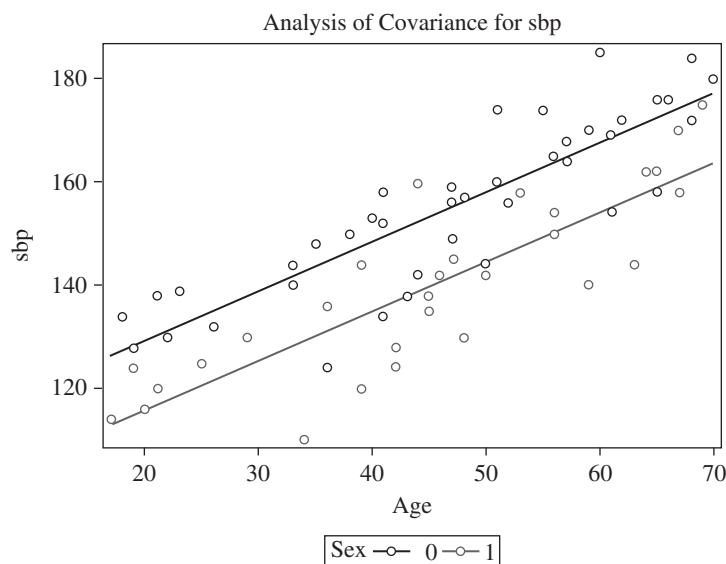
and

```
lsmeans sex /cl;
```

³ In general, for all continuous covariates being controlled, one inputs each sample-wide mean, which may be found using PROC MEANS or UNIVARIATE.

C.8.2 ANACOVA Plot

If, as illustrated in the previously stated code, we specify a model in PROC GLM with a single continuous independent variable and another in a CLASS statement, SAS 9.3 now automatically produces an “Analysis of Covariance” plot, which is a scatterplot of the continuous Y and X variables, with differently colored symbols that represent the levels of the categorical independent variable. (Such a plot is not automatically produced by PROC REG and cannot be produced in PROC GLM for two or more continuous independent variables and/or CLASS statements.) Each level of the variable has a parallel line of fit, based on the model estimates, as indicated in Figure 13.1 of Chapter 13. For the systolic blood pressure example, the following ANACOVA plot is generated:



C.9 ANOVA

Although SAS offers a PROC ANOVA, the PROC GLM procedure can fit a greater variety of models and offers far more analytic options. Therefore, PROC GLM is the primary procedure for ANOVA that we endorse and demonstrate in this section.

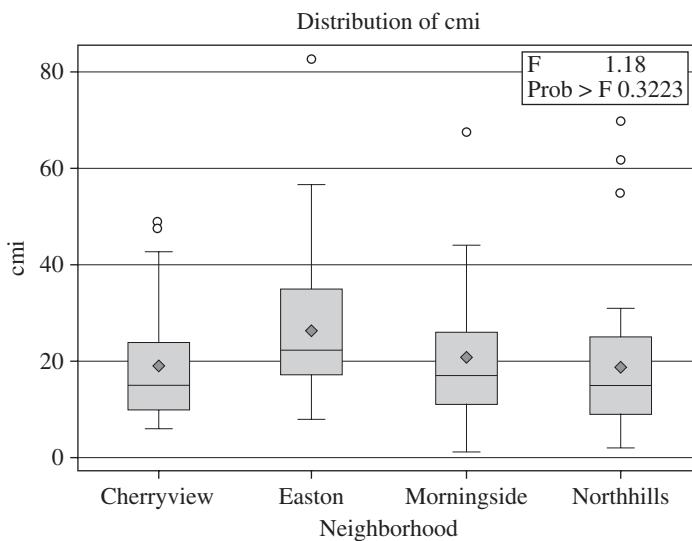
C.9.1 One-way Fixed-effects Model

As with ANACOVA, the fixed-effects ANOVA model for both equal and unequal cell sizes is fit in PROC GLM using virtually the same syntax as the syntax used for a regression model. Since PROC GLM is fundamentally a multiple linear regression procedure, SAS will fit the ANOVA model using a regression model with reference cell coding (Section 17.4.2 of Chapter 17) rather than the classical fixed-effects ANOVA model (Section 17.5).

Accordingly, in any ANOVA model, all independent variables are now strictly categorical ones, and one should ensure that these variables are represented using 0/1 dummy variables created in a DATA step or, more conveniently, are created automatically using the CLASS statement. The PROC GLM syntax to run a one-way (four-level) ANOVA model on Daly's neighborhood CMI data of Example 17.1 would be

```
proc glm data = daly;
  class neighborhood;
  model CMI = neighborhood;
run;
```

In addition to the typical regression/ANOVA-style model output, SAS will detect that a one-way ANOVA is being fit and will display simultaneous boxplots of the response variable at each level of the categorical predictor, annotated with the test statistic and P -value for the overall model F test:



C.9.2 Estimated Means, Mean Differences, and Multiple Comparisons of Means

We have earlier described both the LSMEANS and the ESTIMATE statements in PROC GLM as methods for estimating and conducting statistical inference on the mean response at values of the independent variables. In the ANOVA context, where all independent variables frequently appear on a CLASS statement, LSMEANS and the similar MEANS statements are typically more convenient and offer extra features, such as multiple-comparison procedures. By adding the LSMEANS syntax shown below, a variety of information may be requested for the $k = 4$ neighborhood-specific means (i.e., LSMEANS).

```
lsmeans neighborhood /cl pdiff;
```

The CL option requests that 95% confidence intervals be added to the standard table of LS-means:

neighborhood	cmi LSMEAN	95% Confidence Limits	
Cherryview	18.920000	12.603611	25.236389
Easton	26.840000	20.523611	33.156389
Morningside	20.720000	14.403611	27.036389
Northhills	20.840000	14.523611	27.156389

The PDIFF option requests two-sided t tests of the hypothesis $H_0: \mu_i = \mu_j$, for all possible pairs of the k levels of neighborhood, yielding two tables of output. The first table identifies the \bar{Y}_i (i.e., LSMEANS) for each level of neighborhood and assigns each a numeric alias 1–4. Following this, a $(k \times k)$ grid is constructed with each cell containing the P -value for a specific test of $H_0: \mu_i = \mu_j$, at the i and j values shown in the table margins. Note that each pair is repeated, and thus one only needs to examine this grid above or below the diagonal. Using the output below, we see that the P -value for the t test comparing the means of Easton ($i = 2$) and Morningside ($j = 3$) is 0.6901. Thus, we fail to find evidence of a significant difference in the neighborhood mean CMI, although the estimated means differ by 6.12 ($26.84 - 20.72$).

neighborhood	cmi LSMEAN	LSMEAN Number
Cherryview	18.920000	1
Easton	26.840000	2
Morningside	20.720000	3
Northhills	20.840000	4

LEAST SQUARES MEANS FOR EFFECT NEIGHBORHOOD PR > T FOR H0: LSMEAN(I)=LSMEAN(J)				
Dependent Variable: cmi				
i/j	1	2	3	4
1		0.0816	0.6901	0.6706
2	0.0816		0.1770	0.1856
3	0.6901	0.1770		0.9788
4	0.6706	0.1856	0.9788	

The joint specification of both **CL** and **PDIFF** causes SAS to add a table that complements the above tests, providing estimated pair-wise differences in neighborhood means ($\bar{Y}_i - \bar{Y}_j$) and their 95% confidence intervals.

LEAST SQUARES MEANS FOR EFFECT NEIGHBORHOOD				
i	j	Difference Between Means	95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-7.920000	-16.852723	1.012723
1	3	-1.800000	-10.732723	7.132723
1	4	-1.920000	-10.852723	7.012723
2	3	6.120000	-2.812723	15.052723
2	4	6.000000	-2.932723	14.932723
3	4	-0.120000	-9.052723	8.812723

As discussed in Section 17.7, as we make more such comparisons, the chance of falsely rejecting at least one H_0 increases (i.e., of making a Type I error). The multiple-comparisons adjustment techniques described in Chapter 17—Bonferroni, Tukey–Kramer, and Scheffé—may all be requested as LSMEANS options. For example, a table of Tukey–Kramer simultaneous 95% confidence intervals for $(\mu_i - \mu_j)$ is requested with

```
lsmeans neighborhood /pdiff cl adjust=tukey ;
```

Adjustment for Multiple Comparisons: Tukey

LEAST SQUARES MEANS FOR EFFECT NEIGHBORHOOD				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-7.920000	-19.686119	3.846119
1	3	-1.800000	-13.566119	9.966119
1	4	-1.920000	-13.686119	9.846119
2	3	6.120000	-5.646119	17.886119
2	4	6.000000	-5.766119	17.766119
3	4	-0.120000	-11.886119	11.646119

The **ADJUST** = keywords for Bonferroni and Scheffé adjustment are **BON** and **SCHEFFE**, respectively.

An alternative to LSMEANS is the **MEANS** statement, which can also produce many of the above estimates and their comparisons, albeit in a different output style. MEANS output is displayed throughout Section 17.7 and features convenient linear groupings of letters to denote nonsignificant mean differences. We provide below the MEANS syntax for the Daly neighborhood CMI data that is analogous to the LSMEANS syntax explained earlier. The same BON and SCHEFFE keywords may be specified for those multiple correction estimates.

```
means neighborhood / cldiff lines tukey;
```

MEANS is less suited than LSMEANS for other model contexts (such as ANOVA models with interaction and ANACOVA), and we recommend instead mastering the more capable LSMEANS for general usage.

C.9.3 Two-way Fixed-effects Model

Conducting a two-way fixed-effects ANOVA model is a straightforward extension of the one-way case. Consider the syntax below for fitting the two-way fixed-effects ANOVA model, with interaction, using the Daly CMI neighborhood data, shown in Sections 19.1–19.2.

```
proc glm data = daly_2way;
  class nht psn ;
  model cmi = nht psn nht*psn ;
  lsmeans nht psn nht*psn / cl pdiff;
run;
```

By having two independent variables in the **CLASS** and **MODEL** statements, SAS will fit a model involving both NHT and PSN. As discussed in Section 19.6 on interaction, including a third term that is the “product” of the two factors, denoted by an asterisk, instructs SAS to create and fit dummy variables representing such interaction. With more variables being modeled, one may specify after the **LSMEANS** statement any categorical predictors and their interactions. The LSMEANS code illustrated here yields the main-effects means for the two levels of NHT (in Table 19.2, the row means $\hat{Y}_{1..}$ and $\hat{Y}_{2..}$) and the two levels of PSN (in Table 19.2, the column means $\hat{Y}_{.1..}$ and $\hat{Y}_{.2..}$). Including the interaction term also yields the four means at the combinations of NHT and PSN ($\hat{Y}_{11..}$, $\hat{Y}_{12..}$, $\hat{Y}_{21..}$, $\hat{Y}_{22..}$). As before, **CL** and **PDIFF** will provide confidence intervals and tests of hypothesis that compare the means for each factor.

C.9.4 Contrasts for Fixed-effects Models

For regression and ANACOVA, we have illustrated how the ESTIMATE statement may be used to compute both means and linear contrasts. For ANOVA models, LSMEANS offers the better alternative for mean estimation, but ESTIMATE is still desirable for contrasts. For the one-way ANOVA Daly CMI example in Chapter 17, suppose we want a test of hypothesis that compares the average CMI of both Cherryview and Easton to that of both Morningside and Northhills, corresponding to the contrast

$$L = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2} = \frac{1}{2}(\mu_1) + \frac{1}{2}(\mu_2) - \frac{1}{2}(\mu_3) - \frac{1}{2}(\mu_4)$$

When a categorical CLASS variable is used in the ESTIMATE statement, one simply follows that variable with the coefficients to be used for each level of that factor, in the order that these levels are included in the model.⁴ The above contrast and its hypothesis test are obtained with the following code, where the E option requests that SAS print a table confirming the coefficients used in the contrast:

```
proc glm data = daly;
  class neighborhood;
  model CMI = neighborhood /solution;

  estimate "C & E vs. M & N" neighborhood .5 .5 -.5 -.5 /e;
run;
```

COEFFICIENTS FOR ESTIMATE C & E VS. M & N	
	Row 1
Intercept	0
neighborhood Cherryview	0.5
neighborhood Easton	0.5
neighborhood Morningside	-0.5
neighborhood Northhills	-0.5

Parameter	Estimate	Standard Error	t Value	Pr > t
C & E vs. M & N	2.1000000	3.18208527	0.66	0.5109

Note that ESTIMATE does not directly allow coefficients to be specified as fractions and 0.5 was used in place of $\frac{1}{2}$. But what if we were interested in the comparison of Cherryview with the average of the other three neighborhoods? This corresponds to

$$L = \mu_1 - \frac{\mu_2 + \mu_3 + \mu_4}{3} = \mu_1 - \frac{1}{3}(\mu_2) - \frac{1}{3}(\mu_3) - \frac{1}{3}(\mu_4)$$

In this case, a repeating decimal (0.333 ...) would be necessary to represent the above fractions. Fortunately, SAS offers the DIVISOR option on the ESTIMATE statement, which allows all listed coefficients to be divided by a common number. If we input our desired coefficients *multiplied by 3* and then instruct SAS to *divide by 3* using DIVISOR, we will yield the correct contrast coefficients. This is performed below:

```
estimate "C vs. E & M & N" neighborhood 3 -1 -1 -1 /e divisor = 3;
```

⁴ You might observe that this syntax technically conflicts with the form of the regression model that SAS is using to conduct the ANOVA. The computer is fitting a model with referent cell coding [(k - 1) levels] for each categorical variable rather than the k-level classical ANOVA model implied by this syntax. However, it is easier to specify contrasts in this classical way; that is, coefficients for all k levels are used in the ESTIMATE statement.

COEFFICIENTS FOR ESTIMATE C VS. E & M & N	
	Row 1
Intercept	0
neighborhood Cherryview	1
neighborhood Easton	-0.333333333
neighborhood Morningside	-0.333333333
neighborhood Northhills	-0.333333333

Parameter	Estimate	Standard Error	t Value	Pr > t
C vs. E & M & N	-3.88000000	3.67435557	-1.06	0.2936

C.9.5 Random-effects Models

In PROC GLM, factors in the model that are random effects are specified using the **RANDOM** statement. Consider the two-way random effects DalyCMI neighborhood model, shown in Section 19.2. All factors, including the interaction, are treated as random. This is indicated in SAS by putting all terms on the RANDOM line:

```
proc glm data = daly_2way;
   class nht psn ;
   model cmi = nht psn nht*psn ;
   random nht psn nht*psn /test;
run;
```

By default, SAS provides Type I and III tests of all factors as if they were fixed effects, using the mean square error [MS(error)] as the denominator variance estimate. While these tests of hypothesis are accurate in the one-way random-effects ANOVA case, they often are not appropriate for more complicated random-effects models. This is because the denominator expected mean squares (EMS) may not equal the MS(error), as explored in Section 19.7.4. When the RANDOM statement is used, the computer will automatically output the form of the EMS for each random effect, as determined by the MODEL and RANDOM statements. These may be used to determine the correct denominator mean square to use to test the null hypothesis that a particular random effect of interest is equal to 0. Based on the table below, MS(NHT) and MS(PSN) each reduce to MS(NHT × PSN) when σ_{NHT}^2 and σ_{PSN}^2 , respectively, equal 0. Thus, as shown in Section 19.2, for NHT and PSN, the correct denominator mean square is MS(NHT × PSN) for Type III hypothesis tests of $H_0: \sigma_{NHT}^2 = 0$ and $H_0: \sigma_{PSN}^2 = 0$.

Source	Type III Expected Mean Square
nht	Var(Error) + 25 Var(nht*psn) + 50 Var(nht)
psn	Var(Error) + 25 Var(nht*psn) + 50 Var(psn)
nht*psn	Var(Error) + 25 Var(nht*psn)

The **TEST** option on the **RANDOM** statement requests that SAS automatically compute Type III *F* tests of the general form $H_0: \sigma^2 = 0$ for each random effect, using the appropriate denominator mean square deduced from this table. The results are reproduced below:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
nht	1	110.250000	110.250000	0.27	0.6935
psn	1	380.250000	380.250000	0.94	0.5096
Error: MS(nht*psn)	1	404.010000	404.010000		

Source	DF	Type III SS	Mean Square	F Value	Pr > F
nht*psn	1	404.010000	404.010000	1.60	0.2095
Error: MS(Error)	96	24302	253.141667		

Some complex random-effects designs necessitate hypothesis tests using mean square ratios that are different than those generated by the TEST option on the RANDOM statement. In conjunction with the table of expected means squares above, one may use the **TEST statement (not option)** to conduct customized tests of significance for random effects. Consider this syntax:

```
test h=nht e=nht*psn      / htype=3 etype=3;
```

This requests a test of hypothesis calculated as $MS(NHR) / MS(NHT*PSN)$, specified by the **H=** and **E=** options for numerator and denominator, respectively. The **HTYPE=** and **ETYPE=** options control which sums of squares are used in the calculation. Here we have specified Type III SS. This syntax yields identical results to the test of NHT obtained using the RANDOM statement's TEST option displayed above:

TESTS OF HYPOTHESES USING THE TYPE III MS FOR NHT*PSN AS AN ERROR TERM

Source	DF	Type III SS	Mean Square	F Value	Pr > F
nht	1	110.250000	110.250000	0.27	0.6935

C.9.6 Random Effects with Unequal Cells (Unbalanced Data)

Although PROC GLM is suitable for fixed-effects ANOVA with unequal cell numbers, the procedure may give misleading results for a two-way random-effects ANOVA using unbalanced data. Therefore, we recommend using **PROC MIXED** to conduct these analyses, which is described in the section on correlated data.

C.10 Regression Diagnostics

In Chapter 14, a number of techniques were presented for evaluating one's data and their appropriateness for regression analyses according to the assumptions discussed in Chapter 8. Here we describe how to conduct these methods in SAS, as used in the water pollutants example of Section 14.7 (Example 14.6).

C.10.1 Descriptive Statistics for Variables

The standard procedure for examining the distributional characteristics of variables is PROC UNIVARIATE. For each variable indicated on the VAR line, SAS will provide measures of center and spread, as well as extreme observations. Providing **PLOT** on the PROC UNIVARIATE line requests (visually crude) boxplots, stem-and-leaf plots, and normal probability plots.

```
proc univariate data = pollutant      plot;
   var y x1 x2;
run;
```

C.10.2 Linearity

As detailed in Chapter 14, there are multiple ways to explore the linear relationships between one's dependent and independent variables.

1. Correlation and partial correlation coefficients: These may be computed using **PROC CORR** as described earlier in the appendix.
2. Scatterplots: **PROC GPLOT** and the new **PROC SGPlot** are most commonly used for making high-resolution scatterplots and a variety of other statistical graphics. We provide the basic scatterplot syntax to create Figure 14.11 for each procedure below, although the full range of figure customization options is beyond the scope of this text.

```
proc gplot data = pollutant;
   plot y * x1;
run;

proc sgplot data = pollutant;
   scatter Y = y      X = x1;
run;
```

3. Partial regression plots: These plots show a particular Y - X relationship after controlling for all other independent variables in a specific multiple regression model. Accordingly, these plots need to be generated in a modeling procedure, and **PROC REG** is best suited for this. The addition of the **PARTIAL** option on the MODEL line requests partial regression plots for each independent variable.

```
proc reg data = pollutant;
   model y = x1 x2      / partial;
run;
```

4. Residual plots: These plots may be used to assess multiple assumptions and are described directly below.

C.10.3 Other Model Assumptions

Homoscedasticity, Independence, and Linearity

As illustrated in Chapter 14, these three model assumptions and the existence of outliers may be assessed through the creation of plots of the model residuals (ordinary, studentized, or jackknife) versus predicted values of the dependent variables or each independent variable. By default, PROC REG provides a grid of plots of the ordinary residuals versus predicted values and each independent variable, as well as a plot of the jackknife (“rstudent”) residuals versus predicted values. These plots allow for quick inspection of patterns that suggest deviations from model assumptions but are rather small and may not allow the full exploration that one desires. Two alternatives exist in REG.

Using the **PLOT** statement, one may explicitly request larger, visually customizable residual plots in the “*y* * *x*” form used by **G PLOT** and **S G PLOT**. This is illustrated below. Data set variables, such as the independent variables, may be specified. To plot against residual values, one uses the keywords **RESIDUAL.**, **STUDENT.**, and **RSTUDENT.** (with periods) to represent the ordinary, studentized, and jackknife residuals. The keyword **PREDICTED.** represents predicted values of *Y*. Note that the use of parentheses is a shortcut that represents repetitive plot requests. (*a b*) * (*x1 x2*) asks for the plots *a* * *x1*, *a* * *x2*, *b* * *x1*, and *b* * *x2*. Thus, in the example below, nine total residual plots are requested from REG.

```
proc reg data = pollutant;
  model y = x1 x2      ;
  plot (residual. student. rstudent.) * (predicted. x1 x2);
run;
```

A second method, also available in PROC GLM, is to request that the procedure output the residual values into a data set that may be fed into G PLOT or S G PLOT for plotting. This additionally allows the individual values to be inspected and/or summarized, a feature discussed more in the next section. This is done using the **OUTPUT** statement. The option **OUT=** specifies the data set name and is followed by at least one of the above keywords (now without the periods) and its desired name in the data set. Below, a data set *diag* is created, with variables *jack* and *yhat* that contain the jackknife residuals and predicted values, respectively. This data set also contains all of the original information in *pollutant*, allowing for any number of plots to be created in G PLOT or S G PLOT.

```
proc reg data = pollutant;
  model y = x1 x2      ;
  output out = diag rstudent = jack predicted = yhat ;
run;
```

Normality

The usual normality assumption for multiple linear regression (that is, the residuals are normally distributed) may be assessed using **PROC UNIVARIATE** and the residuals data set

obtained from REG or GLM. Using the *diag* data set below, the **NORMAL** option on the PROC UNIVARIATE requests tests of normality (e.g., Shapiro–Wilks, Kolmogorov–Smirnov) for the jackknife residuals, which are specified using **VAR**. The **PROBPLOT** syntax requests a normal probability plot for the jackknife residuals.

```
proc univariate data = diag normal;
    var jack;
    probplot jack / normal (mu=est sigma=est);
run;
```

C.10.4 Outliers

In addition to utilizing the univariate and bivariate statistics and plots described above to identify outliers, values of *leverage* (h_i) and *Cook's distance* (d_i) can be calculated by either REG or GLM and included in a data set using the **OUTPUT** statement, discussed earlier. These metrics have the keywords **H** and **COOKD**, respectively. One may then look for extreme values of leverage and Cook's distance using PROC UNIVARIATE. Alternatively, the observations may be directly inspected for high values using PROC PRINT:

This sample code assumes that a variable *cook_dist* was created using OUTPUT in REG or GLM and then uses the WHERE statement to only print values above a cutoff of 1:

```
proc print data = diag;
    where (cook_dist > 1);
    var cook_dist y x1 x2;
run;
```

To avoid choosing a cutoff point, one may sort the data set by *cook_dist* and then print the sorted values:

```
proc sort data = diag;
    by cook_dist;
run;

proc print data = diag;
    var cook_dist y x1 x2;
run;
```

C.10.5 Collinearity

Model collinearity, which results in unreliable parameter estimates and standard errors, is caused by the presence of one or more strong linear relationships among the independent variables in the multiple linear regression model. Collinearity may be assessed using the methods described in Section 14.5. The first thing to do is to examine pair-wise correlations, which may be done using the **PROC CORR** syntax described above. Next, the inspection of values of variance-inflation factors (VIFs) and condition indices (CIs) with corresponding variance proportions is done with the assistance of PROC REG. The VIF option in the MODEL statement adds a column of VIF values to the table of parameter estimates, and the **COLLIN** option will generate a table with columns for eigenvalues, condition indices, and variance proportions for each independent variable. The **COLLINOINT** option will

generate a version without the intercept, and all independent variables are centered accordingly. These are depicted in Section 14.5.3, and sample syntax is shown below.

```
proc reg data = pollutant;
  model y = x1 x2 /vif collin;
run;
```

For modeling techniques other than linear regression that use maximum likelihood (ML) methods, such as logistic regression and Poisson regression, the appropriateness of using PROC REG for collinearity diagnostics is questionable. SAS does not provide a built-in method for assessing collinearity for these regression procedures, but SAS macros have been developed for this purpose. An example is the `%collingenmod` macro developed at CDC and Emory University (Zack et al. 2011).

C.11 Model Selection Techniques

Most of the prediction-oriented modeling selection procedures described in Chapter 16 may be automated in PROC REG, through the specification of options on the MODEL line. The **SELECTION** = option, followed by BACKWARD, FORWARD, or STEPWISE, instructs SAS to use one of these *F*-test-based selection procedures. The additional option **SLENTRY** = *<alpha-level>* allows for control over the significance level to use for allowing predictors into the model when using the forward selection and stepwise regression methods. Similarly, **SLSTAY** = *<alpha-level>* allows one to specify the significance level for keeping predictors in backward elimination and stepwise regression.

The following code illustrates the forward selection procedure conducted on the nutritional deficiency data in Section 16.5.3.

```
proc reg data = nutr_def;
  model WGT = HGT AGE AGE2 / selection = forward slentry = 0.10;
run;
```

Setting the **SELECTION** = option to RSQUARE or CP requests that the *all possible regressions* procedure be conducted, using R_p^2 or C_p , respectively, as the selection criterion. The option **STOP** = *<number>* further lets the user set the maximum number of predictors to be allowed into the model in order to control model complexity. For example, consider the case where one is designing a predictive model, beginning with 20 candidate independent variables but would only realistically use a model with a maximum of 10 “final” ones. Setting STOP = 10 would then cause SAS to only show the best models containing 1 to 10 predictors rather than 1 to 20.

Other options allow for finer control over the candidate predictors in the model. Setting the option **INCLUDE** = to a given number instructs SAS to always keep that first number of predictors in the MODEL statement in the model throughout the selection process. This is helpful when there are certain predictors that must be always be considered in the final model. For example, specifying INCLUDE = 1 in the above code would require SAS to keep HGT in the model during all stages of model selection.

Similarly, *chunkwise* testing may be requested in the following way. The variables involved in each chunk should be first listed consecutively and enclosed in braces: { and }. Then the **GROUPNAMES** = option can be used to name the chunks of predictors in the order that they are specified. These names are then used in the display of the model selection

results. For example, the following code groups both age variables into a single chunk for testing in forward selection.

```
proc reg data = nutr_def;
    model WGT = HGT {AGE AGE2} / selection = forward slentry = 0.10
        groupnames = 'Height' 'Age';
run;
```

C.12 Logistic Regression

C.12.1 Binary Logistic Regression

The primary SAS procedure for fitting a binary logistic regression model is PROC LOGISTIC, the syntax structure of which is similar to that of GLM and REG. We illustrate below the code for the dengue fever example, shown in Chapter 22. In this example, the binary outcome of being stricken with dengue fever is modeled with the independent variables mosquito net use (MOSNET), age (AGE), and sector of residence (SECTOR).

```
proc logistic data = dengue ;
    model dengue = age mosnet sector1 sector2 sector3 sector4 ;
run;
```

The five levels of SECTOR may be represented with four binary dummy variables, as done in Chapter 22, or by a single five-level variable in combination with the CLASS statement. As earlier, the use of CLASS instructs SAS to create and use four dummy variables to represent the five-level categorical predictor. Yet, unlike PROC GLM, which uses 0/1 reference cell coding, LOGISTIC uses a -1/1 *effect coding* scheme. Thus, for a given dummy variable that represents a level of a nominal predictor, the odds ratio (OR) comparing that level to the referent group would be $e^{(1-(-1))\beta} = e^{2\beta}$, rather than e^β . To undo this nuisance, we recommend using the PARAM = REF option on the CLASS statement, which instructs SAS to use GLM-style reference cell coding and to provide output that confirms the dummy variables created. Note that the reference group for this example is SECTOR 5.

```
proc logistic data = dengue;
    class sector /param = ref;
    model dengue = age mosnet sector ;
run;
```

CLASS LEVEL INFORMATION					
Class	Value	Design Variables			
SECTOR	1	1	0	0	0
	2	0	1	0	0
	3	0	0	1	0
	4	0	0	0	1
	5	0	0	0	0

The *MOSNET* outcome for the dengue example is coded with values 1 (Yes)/2 (No), with 1 being the response level desired as the modeled outcome. Commonly, binary variables are coded as 0 (No)/1 (Yes), and SAS will, by default, choose the lower value of 0 as the outcome. From the definition of the logistic model in Section 22.2, inverting the outcome has the effect of negating all β estimates—and thus inverting all odds-ratio estimates (i.e., obtaining OR^{-1}). This can be changed by recoding the variable in a DATA step or by adding the DESCENDING option to the PROC LOGISTIC line. This statement instructs SAS to reverse the outcome ordering:

```
proc logistic data = dengue descending;
```

Note that, even with the above modification, the reference group for the SECTOR variable is still SECTOR 5.

Odds Ratios in Binary Logistic Regression

A key analytical task in logistic regression is the estimation of model-adjusted odds ratios between independent variables and the outcome. By default, SAS will generate a table labeled “Odds Ratio Estimates,” containing all exponentiated estimated coefficients ($e^{\hat{\beta}}$) and their 95% confidence intervals. If using a CLASS statement, the rows labels include the appropriate referent group:

ODDS RATIO ESTIMATES			
Effect	Point Estimate	95% Wald Confidence Limits	
AGE	1.025	1.007	1.043
MOSNET	1.396	0.115	16.882
SECTOR 1 vs 5	0.109	0.013	0.888
SECTOR 2 vs 5	0.517	0.175	1.531
SECTOR 3 vs 5	2.253	0.888	5.715
SECTOR 4 vs 5	1.701	0.704	4.110

This table is often helpful, as it saves the effort of manual calculation using each $\hat{\beta}$, found elsewhere in the output. Yet this table may not contain the appropriate or the most useful estimated OR for the analysis at hand. Consider the estimated OR for AGE, for which the standard table provides the estimated OR for only a one-year increase in age. As was the case in Section 22.4, the estimated ORs for 5 and 10-year age increments were of interest. These estimated ORs and their confidence intervals may be requested by adding two additional lines with ODDSRATIO and UNITS statements:

```
oddsratio age ;
units age = 5 10;
```

ODDS RATIO ESTIMATES AND WALD CONFIDENCE INTERVALS			
Label	Estimate	95% Confidence Limits	
AGE units=5	1.129	1.033	1.234
AGE units=10	1.275	1.067	1.522

An alternative method for customizing OR estimates is to use the ESTIMATE statement, which can be used to compute any linear function of the model parameter estimates (\hat{L}). As earlier, this statement begins with a descriptive label for the linear function, followed by each independent variable whose coefficient is to be included and the constant by which it is to be multiplied. By default, LOGISTIC will only compute point estimates on the linear scale. Thus, we supply EXP and CL options, which, respectively, instruct SAS to exponentiate the linear function ($e^{\hat{L}}$) and to produce confidence intervals.

```
estimate "Age, 5 yr" age 5 / exp cl;
```

ESTIMATE											
Label	Estimate	Standard Error	z Value	Pr > z	Alpha	Lower	Upper	Exponentiated	Exponentiated Lower	Exponentiated Upper	
Age, 5 yr	0.1213	0.04528	2.68	0.0074	0.05	0.03257	0.2101	1.1290	1.0331	1.2338	

The standard OR table produced is also inadequate when a factor of interest is a component of an interaction term with a second factor. In these cases, the estimated OR for one factor varies according to the value of the second factor. In the dengue fever example of Chapter 24, a model involving the interaction between MOSNET and AGE was considered. For this model, the estimated OR for the effect of mosquito net use was $e^{\beta_2 + \beta_7(AGE)}$. If we were interested in the relationship between mosquito nets and dengue fever for a 20-year-old, we would specify coefficients of 1 and 20 for MOSNET and AGE in the following ESTIMATE statement:

```
estimate "OR for MOSNET, Age = 20" mosnet 1 mosnet * age 20 / exp cl;
```

ESTIMATE											
Label	Estimate	Standard Error	z Value	Pr > z	Alpha	Lower	Upper	Exponentiated	Exponentiated Lower	Exponentiated Upper	
OR for Mosnet, Age = 20	-0.1931	1.2325	-0.16	0.8755	0.05	-2.6087	2.2225	0.8244	0.07363	9.2302	

As with other discussed procedures, the α level used by LOGISTIC may be adjusted with the ALPHA= option on the MODEL line.

C.12.2 Conditional Logistic Regression

Sections 22.5 and 22.6 consider the situation in which the number of factors or groups included as model predictors is large relative to the sample size, such as in matched designs. In these cases, conditional ML logistic regression is typically required and is easily implemented in PROC LOGISTIC via the use of the STRATA statement. This statement is used to indicate the independent variable(s) used to partition the data into sets (i.e., to identify stratum membership). The code used for the endometrial cancer example of Section 22.6 is shown below and is used to generate the results of Table 22.6.

```
proc logistic data = endomet;
    model case = est gall ;
    strata stratum;
run;
```

For conditional logistic regression, the individual effects of the stratification variables cannot be estimated. Parameter estimates are, therefore, only produced for the factors on the MODEL line.

C.12.3 Polytomous Logistic Regression

Polytomous logistic regression is one of two methods discussed in Chapter 23 for dealing with a categorical outcome that has more than two categories. In the polytomous model, the outcome is considered to be unordered. The end-stage renal disease (ESRD) example in Chapter 23 featured a polytomous model with a three-level response. This corresponded to two unique regression equations that separately compared one of two outcome levels (hypertension and diabetes ESRD causes) to a referent level (other cause). A multivariable model with independent variables race, age, and gender was fit in PROC LOGISTIC, with output provided in Table 23.3 and code shown below.

The option LINK = GLOGIT is used to instruct the computer to fit a *generalized logit* model, another term used for the polytomous logistic regression model. By placing the outcome of CAUSE in the CLASS statement and following it on the MODEL line with a (REF =) option, one can best control which outcome level is treated as the referent group.

```
proc logistic data = esrd;
    class cause race ageg gender /param = ref;
    model cause (ref = "0") = race ageg gender /link = glogit;
run;
```

Since two equations are being fit, with unique sets of coefficients being estimated, two parallel forms of OR expressions are required for the set of covariates included in the model. Both the standard odds ratio output and ESTIMATE statement output from PROC LOGISTIC will automatically produce extra lines of OR estimates to reflect each comparison to the reference level of the outcome.

C.12.4 Ordinal Logistic Regression

In Chapter 23, the other regression technique presented for analyzing a categorical outcome is the proportional-odds ordinal logistic regression model. Unlike polytomous regression,

the outcome has a natural ordering that is explicitly modeled. The proportional odds assumption specifies that, while the odds of the outcome might increase or decrease, the ORs for each of the model covariates' effects are assumed to be identical across cut-points for dichotomizing the outcome. This yields a single OR estimate for each independent variable in the model.

To fit a proportional odds model, we simply fit a model using the model syntax for binary logistic regression but with an ordinal response variable. SAS will detect that more than two levels of the outcome exist and will automatically fit a proportional odds model. To explicitly specify this model, one may add the option LINK = CLOGIT to the MODEL statement. Since the levels of the outcome variable are strictly ordered, we recommend that the outcome variable be numerically coded to ensure that the correct ordering is used.⁵ The direction of the ordinal variable may be inverted using the DESCENDING option on the PROC LOGISTIC line, just as for binary logistic regression. The model code for the *Acinetobacter* example of Section 23.7 is given below. This code yields the output provided in Table 23.11, including the results of the *score test*.

```
proc logistic data = acineto;
  class newexptime;
  model ordhospdays = acstatus apache charlson newexptime;
run;
```

C.13 Poisson Regression

Covered in Chapter 24, Poisson regression is a method for analyzing data on counts of events (Y) and rates (λ), which are the average number of events per unit of time (or per other units such as persons, person-time (PT), or surface area, all represented by the symbol ℓ).

In this section, we describe how to fit the skin cancer model of Chapter 24, the data for which are presented in Table 24.1, using PROC GENMOD in SAS. In PROC GENMOD, the MODEL statement's general structure is the same as for the other procedures discussed. Because GENMOD may be used to fit most of the linear-format models demonstrated in this text, three options may be specified to request the Poisson regression model. The LINK = LOG option specifies that a natural logarithm is to be applied to $E(Y)$ in our model. The DIST = POISSON option indicates that the Poisson distribution is to be used for the outcome (rather than, say, the normal distribution). Finally, the OFFSET = option allows us to select which variable contains the offset information. GENMOD will not automatically compute the logarithm for the offset, and this must be done in a previous DATA step, as illustrated below with the creation of the new variable LN_N, which represents the $\log_e(\text{population})$ in each city-age grouping.

⁵ It is possible to fit a proportional odds model to a nonnumerically coded outcome variable (e.g., "low," "medium," "high"); however, the computer will order these responses alphabetically in its configuration of the proportional odds model, which is often not the desired order. Thus, we recommend coding these responses numerically; one may add these text descriptors to the outcome levels using formats defined through PROC FORMAT and assigned using the FORMAT statement. See the SAS Help feature for more information.

```

data skin_cancer;
  set skin_cancer;

  ln_n = log(n);
run;

proc genmod data = skin_cancer ;
  model cases = city u1 u2 u3 u4 u5 u6 u7 /link=log dist=poisson
  offset = ln_n ;
run;

```

This code generates the output shown in Section 24.4, including a table containing values of $\hat{\alpha}$, $\hat{\beta}$, and their estimated standard errors. The deviance statistic is displayed in a separate table.

In order to compute estimated adjusted rate ratios (RR), a key analytical objective, we use the ESTIMATE statement in a similar fashion as for PROC LOGISTIC.⁶ The EXP option is required in order to compute exponentiated linear sums $e^{\hat{L}}$. The following code requests the RR comparing the cancer rate in Dallas–Ft. Worth to Minneapolis–St. Paul, $e^{\hat{\beta}}$.

```
estimate "DFW vs. MSP" city 1 /exp;
```

The second line of the table below gives the exponentiated results. The column “L-Beta Estimate” contains the estimated rate ratio of 2.2342, and the “L-Beta Confidence limits” contains the CI for the RR. The “Chi-Square” columns contain the results of hypothesis tests of the general form $H_0: L = 0$, which, in this instance, denotes $H_0: \beta = 0$, a result provided in the standard output. Since the result of a test of $H_0: L = 0$ is always the same as one for $H_0: e^L = 1$, the hypothesis test results are identical for both rows and are shown only once in the table

CONTRAST ESTIMATE RESULTS										
Label	Mean Estimate	Mean		L'Beta Estimate	Standard Error	Alpha	L'Beta		Chi-Square	Pr > ChiSq
		Confidence Limits					Confidence Limits			
DFW vs. MSP	2.2342	2.0169	2.4749	0.8039	0.0522	0.05	0.7016	0.9062	237.13	<.0001
Exp(DFW vs. MSP)				2.2342	0.1166	0.05	2.0169	2.4749		

C.14 Analysis of Correlated Data

In this section, we demonstrate the conduct of general linear mixed model analyses for correlated data in SAS, introduced in Chapters 25 and 26. This model considers a continuous outcome in an ANOVA or multiple regression setting, where observations are clustered

⁶ Although PROC LOGISTIC automatically provides tables of $e^{\hat{\beta}}$ as estimated odds ratios, no such table is provided in PROC GENMOD.

within subjects. We exclusively discuss PROC MIXED, which can be used to fit both repeated measures and random-effects models.

As mentioned in Section 26.6, other SAS procedures can be used to analyze clustered binary data (e.g., logistic regression) and count outcome data (e.g., Poisson regression), such as PROC GENMOD for repeated measures and PROC GLIMMIX for random effects models. The details of these procedures and the methods underlying them are beyond the scope of this text.

C.14.1 Marginal Models

The code below shows the syntax for fitting the repeated-measures marginal FEV1 model in Section 25.4. In this example, FEV1 lung capacity measurements were taken on 40 subjects (SUBJ) over the course of five consecutive weeks (WEEK). The analysis seeks to understand the between-week differences in FEV1, while controlling for the repeated observations on the subjects. To do this in PROC MIXED, we use nearly identical CLASS and MODEL syntax to that used in GLM. The SOLUTION option requests a table of individual β estimates, standard errors, and t tests. The DDFM= option sets the denominator degrees of freedom for hypothesis tests involving fixed effects. In the repeated measures models, all model effects are fixed, and RESIDUAL is the appropriate denominator DF option value. See Sections 25.4.2, 26.3.1, and 26.4.7 for more details.

The **REPEATED** statement causes MIXED to fit a marginal model and to estimate an **R** matrix that represents the covariances among the FEV1 responses. In specifying the WEEK variable after REPEATED, the computer will know to fit a 5×5 R matrix and how to link observations from a given week on a specific person to cells in the **R** matrix.⁷ The required SUBJECT= and TYPE= options, respectively, specify the variables that uniquely identify the cluster variable and the covariance structure. The compound symmetric covariance structure (all covariances are equal) is indicated using CS. The independent (IND; all covariances equal 0), autoregressive-1 (AR1), and unstructured (UN) structures may also be specified.

```
proc mixed data = fev1;
  class week subj;
  model fev = week /solution ddfm=residual;
  repeated week /subject = subj type = cs r ;
run;
```

The primary output obtained from running the above code is shown below. It contains familiar-looking tables of Type III F tests of the overall model effects and of individual β estimates.

⁷ Technically, WEEK is optional, and by excluding it, SAS will attempt to determine the dimensions of **R** by detecting the maximum number of repeated observations within a cluster. The order of the repeated observations will then be dependent on the sorted order of the data. A potential pitfall to this approach is the case where observations within a cluster are missing. Failing to include placeholder observations (with a value of “.” for the dependent variable) at the appropriate week may cause later values of week to be mistaken for earlier ones. While it is the style of some statistical programmers to leave out such an indexing variable on the REPEATED line, we recommend including them whenever possible.

TYPE 3 TESTS OF FIXED EFFECTS

Effect	Num DF	Den DF	F Value	Pr > F
week	4	195	55.26	<.0001

SOLUTION FOR FIXED EFFECTS

Effect	week	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		6.9985	0.2590	195	27.02	<.0001
week	1	2.8153	0.2439	195	11.54	<.0001
week	2	-0.1502	0.2439	195	-0.62	0.5387
week	3	0.003250	0.2439	195	0.01	0.9894
week	4	-0.04700	0.2439	195	-0.19	0.8474
week	5	0

PROC MIXED also provides estimates of the unique covariance parameters that appear in the *R* matrix. Since a compound symmetric covariance structure was used, only one covariance was estimated.

COVARIANCE PARAMETER ESTIMATES

Cov Parm	Subject	Estimate
CS	subj	1.4936
Residual		1.1901

In the case of more complicated covariance structures, MIXED will print all additional covariance estimates vertically in the above table. This display can be difficult to read. The option *R* for the REPEATED statement causes SAS to print out the estimated **R** matrix, as a square matrix, for a given subject. Inspection of the **R** matrix is often helpful for exploring the relationships between repeated observations and assessing the goodness of fit of a given covariance structure.

ESTIMATED R MATRIX FOR SUBJ 1

Row	Col1	Col2	Col3	Col4	Col5
1	2.6837	1.4936	1.4936	1.4936	1.4936
2	1.4936	2.6837	1.4936	1.4936	1.4936
3	1.4936	1.4936	2.6837	1.4936	1.4936
4	1.4936	1.4936	1.4936	2.6837	1.4936
5	1.4936	1.4936	1.4936	1.4936	2.6837

Note that, when using the REPEATED statement, SAS does not automatically produce the often-easier-to-interpret correlation matrix **C**, although it may be computed by dividing all cells of the **R** matrix by estimated variances found on the diagonal of the **R** matrix. For the reader's convenience, we have written a macro called PRINT_C_MATRIX, which will automatically compute and print the **C** matrix based on the **R** matrix generated by MIXED. One needs only to add the following ODS OUTPUT statement to PROC MIXED and to execute the call to the macro, inputting the name of the stored **R** matrix, as written below:

```
proc mixed data = fev1;
  class week subj;
  model fev = week /solution ddfm=residual;
  repeated week /subject = subj type = cs r;

  ods output Mixed.R = r_matrix; * store R matrix in dataset;
run;

%print_c_matrix (r_mat = r_matrix);
```

This code for PRINT_C_MATRIX is to be run beforehand:

```
%macro print_c_matrix (r_mat = );
  * Read and store the dimensions of R matrix in macro variable (= 5 in FEV1 ex);
  data _NULL_;
    set &r_mat;
    by index;
    if last.index then call symput('dimension', compress(put(_N_, 2.)));
  run;

  * compute C matrix, using R matrix;
  data c_matrix;
    set &r_mat;
    retain s2; * remember est variance on all rows;

    if _N_ = 1 then s2 = col1; * store variance from upper-left cell;

    * divide each entry in matrix by variance;
    array cols {&dimension} col1 - col&dimension;

    do i = 1 to &dimension;
      cols{i} = cols{i} / s2;
    end;

  run;

  * Print C matrix ;
  proc print data = c_matrix;
    title "C matrix";
    var row col1 - col&dimension;
  run;
%mend print_c_matrix;
```

The following **C** matrix is produced for the above FEV1 example and matches results found in Table 25.7:

C matrix

Obs	Row	Col1	Col2	Col3	Col4	Col5
1	1	1.0000	0.5565	0.5565	0.5565	0.5565
2	2	0.5565	1.0000	0.5565	0.5565	0.5565
3	3	0.5565	0.5565	1.0000	0.5565	0.5565
4	4	0.5565	0.5565	0.5565	1.0000	0.5565
5	5	0.5565	0.5565	0.5565	0.5565	1.0000

Contrasts

In PROC MIXED, we compute estimates and tests of linear contrasts in the same manner as done for fixed-effects ANOVA. In Section 25.4, a contrast was computed to compare the mean FEV1 at week 1 to the average mean for the other four weeks. This corresponded to a test of the hypothesis:

$$H_0: \mu_1 - \frac{\mu_2 + \mu_3 + \mu_4 + \mu_5}{4} = 0$$

We can use the ESTIMATE statement to estimate this contrast and to compute the associated *t* test and 95% CI.

```
estimate "Week 1 vs. 2-5" week 1 -.25 -.25 -.25 -.25 / cl;
```

ESTIMATES								
Label	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper
Week 1 vs. 2-5	2.8638	0.1928	195	14.85	<.0001	0.05	2.4834	3.2441

This output is similar to that provided in Chapter 25; however, results provided in the chapter are of corresponding *F* tests and are generated by the identically coded CONTRAST statements. In MIXED, the CONTRAST statement provides the additional ability to test multiple hypotheses simultaneously (through the use of commas), but it provides only hypothesis test results and not estimated contrasts and associated 95% CIs. It additionally does not have the DIVISOR option that ESTIMATE has available, which allows for the input of more complicated fraction coefficients, which is found in ESTIMATE.

As explained in the section on contrasts for fixed-effects ANOVA, using a variable in a CLASS statement requires that the coefficients be entered in an ESTIMATE statement in terms of a cell means (μ) representation rather than a regression coefficient (β) one. In Section 25.4, the regression representation of this contrast was found to be:

$$H_0: \beta_1 - \frac{\beta_2 + \beta_3 + \beta_4}{4} = 0$$

When a variable is not included in a CLASS statement, and is instead specified with manually created dummy variables (e.g., week1, ..., week4), the syntax for ESTIMATE should be changed to reflect the regression-coefficients version of the contrast:

```
estimate "Week 1 vs. 2-5" week1 1 week2 -.25 week3 -.25 week4 -.25;
```

Empirical Standard Errors

Described in Section 25.4, the use of empirical standard errors provides for the possibility that the specified covariance structure is incorrect. These are requested with EMPIRICAL option on the PROC MIXED line.

```
proc mixed data = fev1 empirical;
```

C.14.2 Random-effects models

Random Intercept Only

To fit models that include one or more random effects, the RANDOM statement is used. The code below produces a single random intercept model that allows for a single effect that varies from subject to subject (b_{j0}), with a variance of σ^2_0 , contained in a 1×1 **G** matrix. The use of the keyword INTERCEPT in conjunction with the SUBJECT = option below instructs SAS to fit a random effect for the variable *SUBJ*.⁸

```
proc mixed data = fev1;
  class week subj;
  model fev = week /solution ;
  random intercept / subject = subj g s;
run;
```

TYPE 3 TESTS OF FIXED EFFECTS				
Effect	Num DF	Den DF	F Value	Pr > F
week	4	156	55.26	<.0001

(continued)

⁸ An alternative way to code this random effect is “random subj / g s”; this method is discussed more in the next section.

SOLUTION FOR FIXED EFFECTS						
Effect	week	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		6.9985	0.2590	39	27.02	<.0001
week	1	2.8152	0.2439	156	11.54	<.0001
week	2	-0.1503	0.2439	156	-0.62	0.5388
week	3	0.003250	0.2439	156	0.01	0.9894
week	4	-0.04700	0.2439	156	-0.19	0.8475
week	5	0

The **G** option causes **G** matrix to be printed. Note that the single variance estimate of 1.4936 for this model matches the covariance estimate in the **R** matrix of the compound symmetric marginal model and that many of the other estimates between the two models are the same (with an important distinction in the d.f. used). This equivalence is discussed in detail in Section 26.3.1.

ESTIMATED G MATRIX			
Row	Effect	subj	Col1
1	Intercept	1	1.4936

The **S** option requests the full random-effects solution to be produced, which includes each subject's estimated value of the random effect (complete results appear in Table 26.2).

SOLUTION FOR RANDOM EFFECTS						
Effect	subj	Estimate	Std Err Pred	DF	t Value	Pr > t
Intercept	1	-0.5216	0.4874	156	-1.07	0.2861
Intercept	2	1.9332	0.4874	156	3.97	0.0001
Intercept	3	-2.2019	0.4874	156	-4.52	<.0001
...

Models with Multiple Random Effects

In the posture and shoulder flexion example of Section 26.4, two random-effects models were presented. The first was a random intercept model, with subjects being the random factor and with day, time, and day*time as fixed effects (Section 26.4.3). Similar to the preceding example, the corresponding SAS code for this model would be

```
proc mixed data = sf;
   class subj day time;
   model sf = day time day*time /solution;
   random intercept / subject = subj g;
run;
```

The above syntax for the RANDOM statement directly reads as creating a random intercept and maintains similar style to the REPEATED statement. When not using the EMPIRICAL option, an alternative way to write this line exists that more explicitly makes SUBJ the random effect:

```
random subj /g;
```

Written this way, the MIXED syntax is readily interpreted as placing all fixed effects on the MODEL line and all random effects after RANDOM.⁹

The differences between these two syntactical styles may be clearly seen in more complicated random-effects models, such as the three-way ANOVA model of Section 26.4.4. In this model, the random factor of subjects is involved in interaction terms with the fixed day and time effects. These subject*day and subject*time factors are themselves random effects.

Using the syntax first introduced for the random intercept model, the code for the three-way ANOVA model is shown below. This represents the three random effects of interest because SUBJECT = SUBJ essentially causes MIXED to multiply each of the terms INTERCEPT, DAY, and TIME by SUBJ. This yields SUBJ, SUBJ*DAY, and SUBJ*TIME random effects.¹⁰

```
proc mixed data = sf;
   class subj day time;
   model sf = day time day*time /solution;
   random intercept day time / subject = subj g;
run;
```

Yet this formulation of the model seems difficult to read in the context of several random effects. We recommend the alternative formulation, which directly states the random effects on the RANDOM statement.

⁹ Note that both of these styles for specifying random effects are different from the way that random effects are specified for random-effects ANOVA models in PROC GLM. For those models, the random effects are specified on *both* the MODEL and the REPEATED lines.

¹⁰ Recall that the first column of the **Z** matrix has all values equal 1 and so the product INTERCEPT*SUBJ = SUBJ.

```
proc mixed data = sf;
  class subj day time;
  model sf = day time day*time /solution;
  random subj subj*day subj*time / g;
run;
```

Reference

Zack, M.; Singleton, J.; Satterwhite, C.; and Delaney, K. P. 2011. "Collinearity Macro (SAS)." Unpublished, Department of Epidemiology, Emory University Rollins School of Public Health (contact dkleinb@emory.edu).

D

Appendix—Answers to Selected Problems

Chapter 3

2. nominal, ordinal, interval, ratio
3. a. 0.8413 b. -0.842
4. a. 18.475 b. 0.699
5. a. -1.350 b. 0.05
6. a. 2.51 b. 0.025
7. a. 0 b. 0 c. 0
8. standard normal
9. a. 3.0 b. 3 c. 2.8 or 3.11
10. e
11. a. 5.0 b. (187.44, 192.56)
12. $t_{0.975, 27} = 2.052$
13. (24.66, 35.33)
14. 3.6858
15. a. significant difference b. significant difference
16. nonsignificant difference
17. $t_{0.995, 10} = 3.619$
18. b
19. a. Type I error b. correct decision c. correct decision d. Type II error
20. a, b
21. c
22. b

23. a. $1 - \alpha$ b. α c. β d. $1 - \beta$

24. Accept H_0 .

25. b

Chapter 5

- 1.** a. Dry Weight (Y) does increase with increasing Age (X), but the relationship may not be linear. An exponential relationship between X and Y may better fit the data. Log Dry Weight (Z) increases linearly with increasing Age (X).
 b. $Y = \beta_0 + \beta_1 X + E$ $Z = \beta'_0 + \beta'_1 X + E$
 c. $\hat{Y} = -1.885 + 0.235X$ $\hat{Z} = -2.689 + 0.196X$
 d. The regression line for \log_{10} Dry Weight regressed on Age has a better fit. It is more appropriate to run a linear regression of Z on X .
 e. 95% confidence intervals: for β'_1 : (0.190, 0.202), for β'_0 : (-2.759, -2.620)
 f. (-1.149, -1.096)
- 3.** a. The relationship between Time (Y) and Inc (X) does not appear to be linear.
 b. $\hat{\beta}_0 = 19.626$ $\hat{\beta}_1 = 0.0007$
 c. $\hat{Y} = 19.626 + 0.0007X$. The regression line fits the data poorly.
 d. The linearity assumption is not met.
 e. $T = 2.023$, $P = .0582$ (from SAS output). We do not reject H_0 , since P -value > .05.
 f. The scatter plot suggests that a parabola would better fit the data.
- 5.** a. $\hat{Y} = 2.174 + 1.177X$. The line fits the data well.
 b. No.
 c. $T = 13.5$ $P < .0001$ (from SAS output).
 Since P -value < .05, we reject H_0 .
 d. $T = 0.954$ $.15 < P < .25$.
 Since P -value > .05, we do not reject H_0 .
 e. (44.146, 47.276)
- 7.** a. $\hat{Y}_1 = -122.345 + 6.227X$ $\hat{Y}_2 = -1.697 + 0.299X$
 b. Y_2 regressed on X
 c. $T = -57.934$. Critical value: $t_{17} \sim 2.898$ under H_0 at $\alpha = .01$. We see that $|T| = 57.934 > 2.898$, so we reject H_0 at $\alpha = .01$.
 d. (0.264, 0.334)
 e. (11.18, 12.32)
- 9.** a. $\hat{\beta}_0 = 2.936$ $\hat{\beta}_1 = -1.785$
 b. $\hat{Y} = 2.936 - 1.785X$. The line fits the data well.
 d. $\hat{Y} = 862.979\hat{X}^{-1.785}$
 e. (6.266, 9.311), (321.366, 528.445)
 f. One could plot both logarithm-transformed data (X , Y) and their estimated regression line on the same plot, and then do the same for the original data values (X' , Y') and their corresponding new estimated regression line. Then one could visually compare the lines to the patterns of the data. In doing this comparison, it is quickly apparent that the plot of (X' , Y') has a very curvilinear pattern and a straight-line is a poor fit, relative to (X , Y), which fits a line well.

(More formal methods for assessing straight-line fit are discussed in Chapters 6 and 14.)

- 11.** **a.** $\hat{Y} = 3.707 - 0.012X$.
- b.** $T = -8.684 \quad P = .001$ (from SAS output).
Since P -value < .05, we reject H_0 .
- c.** Including the data from the three experiments, rather than just using the average values, would provide more information and might improve the sensitivity of the analysis.
- d.** (1.243, 3.737)
- e.** It is inappropriate, since the estimated model relies on data that do not include any information for average growth rate when exposed to a gas with a molecular weight of 200.
- f.** The chosen X values are not uniformly distributed in the experiment. There are large gaps between the X values of 39.9, 83.8, and 131.3. This may result in a fitted line subject to inaccuracies for predicting Y based on X .
- 13.** **b.** From the SAS output: $\hat{\beta}_0 = 0.116 \quad \hat{\beta}_1 = 0.005$
- c.** $T = 0.811 \quad P = .433$ (from SAS output).
Since P -value > .10, we do not reject H_0 .
- d.** $T = 0.046 \quad P = .964$ (from SAS output).
Since P -value > .10, we do not reject H_0 .
- e.** $\hat{Y} = 0.116 + 0.005X$
- f.** The line does not differ from the line plotted in part (e). The evidence suggests that there is no significant linear relationship. Determining a well-fitting line is difficult, given the dispersion of the data.
- 15.** **a.** As X increases, the dispersion of Y increases.
- b.** $\hat{\beta}_0 = -2.546 \quad \hat{\beta}_1 = 0.032 \quad \hat{Y} = -2.546 + 0.032X$
- c.** This chart implies a better linear relationship between X and Y .
- d.** $\hat{\beta}_0 = -6.532 \quad \hat{\beta}_1 = 1.430 \quad \hat{Y} = -6.532 + 1.430X$
- e.** The natural log transformation provides the best representation. The natural log plot better illustrates the linear relationship, and the dispersion of the data is more similar at each level of toluene exposure. The first plot indicates that there may be a violation of homoscedasticity for the untransformed data.
- 17.** **a.** Yes.
- b.** $\hat{Y} = 1.643 + 1.057X$. The line appears to fit the data well.
- c.** 95% CI for β_1 : (0.580, 1.534). The 95% CI does not include the null value of zero, indicating that there is a significant linear relationship at $\alpha = .05$.
- d.** No, it is not appropriate, since the data used to estimate the regression line do not include a \$10 million advertising expenditure in their range.
- 19.** **a.** There may be a slight negative linear relationship.
- b.** $Y = \beta_0 + \beta_1 X + E \quad \hat{\beta}_0 = 76.008 \quad \hat{\beta}_1 = -0.015$.
The baseline OWNEROCC = 76%, and as OWNCOST increases by \$1,000, the percentage of OWNEROCC decreases by ~2%.
- c.** $\hat{Y} = 76.008 - 0.015X$. The line fits the data well.
- d.** $T = -2.607 \quad P = .0155$ (from SAS output).
Since P -value < .05, we reject H_0 .

- e. $(-0.027, -0.003)$. We are 95% confident that the true slope is between -0.027 and -0.003 . Since the interval does not contain zero, we conclude that the slope is not equal to zero (at $\alpha = .05$).
21. a. Substituting the parameter estimates from the computer output into the regression equation (5.6), we find
- $$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 (\text{drink days}) = 27.530 - 0.133 (\text{drink days})$$
- b. The value for $S_{\hat{\beta}_1}$ is provided in the output, so we calculate
- $$T = \frac{\hat{\beta}_1 - 0}{S_{\hat{\beta}_1}} = \frac{-0.133}{0.0367} = -3.62$$
- To perform a two-tailed test at the $\alpha = .05$ level, we examine if $|T| \geq t_{420993, 0.975}$. Indeed, 3.62 exceeds 1.96 , and we reject the null hypothesis that there is no linear relationship between alcohol consumption frequency and BMI. The P -value is $.0003$.
- c. At 4 drinking days per month, the estimated BMI is 27.00 . The 95% CI is thus $27.00 \pm 1.96(0.2658) = (26.48, 27.52)$.
At 15 days, the estimated BMI is 25.54 , so the 95% CI is $25.54 \pm 1.96(0.3950) = (24.77, 26.31)$
- d. The global statistical conclusions about the linear relationship between alcohol consumption frequency and BMI are the same for both the unweighted and the weighted analyses, although the latter analysis is the correct one. As expected, the standard error of the estimated slope increases when the sampling design is taken into account; this increase inflates the P -value and the width of the confidence interval for the slope parameter.

Chapter 6

1. a. (1) $r = 0.86$ (2) $r = 0.999$
 b. (1) $(0.546, 0.964)$ (2) $(0.996, 1.000)$
 c. (1) $r^2 = 0.744$, so 74% of the variation in Y is explained with the help of X .
 (2) $r^2 = 0.998$, so 99% of the variation in Z is explained with the help of X .
 d. The regression using $\text{Log}_{10} \text{Dry Weight}$ appears to fit better. This agrees with Chapter 5, Problem 1(d).
3. a. $r = 0.980$
 b. Substitute $r(S_Y/S_X)$ and $\frac{(n-1)}{(n-2)}(S_Y^2 - \hat{\beta}_1^2 S_X^2)$ for $\hat{\beta}_1$ and $S_{Y|X}^2$ in formula (5.9).
 c. $T' = 13.929$; $T \sim t_8$ under $H_0: \rho = 0$. The P -value for this test is less than $.001$; therefore, we reject H_0 .
 d. The graph of Y versus X does not illustrate a linear relationship.
5. a. $r^2 = 0.601$, $r = 0.775$. 60% of the variation in SBP (Y) is explained by AGE (X).
 b. $(0.504, 0.907)$. Since $\rho = 0$ is not included in the interval, we reject $H_0: \rho = 0$ at $\alpha = .01$.

7. a. $r^2 = 0.1853$, $r = 0.430$. 19% of the variation in SBP (Y) is explained by AGE (X).
 b. $T = 2.02$; critical value: $t_{18, 0.975} = 2.101$. At $\alpha = .05$, since $|T| <$ critical value, we would not reject H_0 .
 c. $(-0.015, 0.733)$. Since $\rho = 0$ is included in the interval, we do not reject $H_0: \rho = 0$ at $\alpha = .05$.
9. a. $r^2 = 0.9101$, $r = 0.954$. 91% of the variation in Y is explained by X .
 b. $T = 13.49$; critical value: $t_{18, 0.975} = 2.101$. At $\alpha = .05$, since $|T| >$ critical value, we would reject H_0 .
 c. $(0.885, 0.982)$. Since $\rho = 0$ is not included in the interval, we reject $H_0: \rho = 0$ at $\alpha = .05$.
11. a. $r^2 = 0.9728$, $r = 0.986$, so 98% of the variation in Y_2 is explained by X .
 b. $T = 24.640$; critical value: $t_{17, 0.975} = 2.110$. At $\alpha = .05$, since $|T| >$ critical value, we would reject H_0 .
 c. $(0.963, 0.995)$. Since $\rho = 0$ is not included in the interval, we reject $H_0: \rho = 0$ at $\alpha = .05$.
13. a. $r = 0.971$
 b. $(0.813, 0.996)$. Since $\rho = 0$ is not included in the interval, we reject $H_0: \rho = 0$ at $\alpha = .05$.
15. $Z = 3.62$; critical value = 1.96. Since $|Z|$ is $>$ critical value, we reject the null hypothesis.
17. a. $r^2 = 0.9558$, $r = 0.978$. 96% of the variation in Y is explained by X . This includes the newly added 16th observation.
 b. $T = 17.424$; critical value: $t_{14, 0.975} = 2.145$. At $\alpha = .05$, we would reject H_0 .
 c. $(0.936, 0.993)$. Since $\rho = 0$ is not included in the interval, we reject $H_0: \rho = 0$ at $\alpha = .05$.
19. a. $r^2 = 0.0310$, $r = -0.176$. 3% of the variation in Y is explained by X .
 b. $T = 0.759$; critical value: $t_{18, 0.975} = 2.101$. At $\alpha = .05$, we would not reject H_0 .
 c. $(-0.574, 0.289)$. Since $\rho = 0$ is included in the interval, we do not reject $H_0: \rho = 0$ at $\alpha = .05$.
21. a. $r^2 = 0.9813$, $r = 0.991$. 98% of the variation in Y is explained by X .
 b. $T = 55.088$; critical value: $t_{38, 0.975} = 2.0$. At $\alpha = .05$, we would reject H_0 .
 c. $(0.985, 0.995)$. Since $\rho = 0$ is not included in the interval, we reject $H_0: \rho = 0$ at $\alpha = .05$.
23. a. $r^2 = 0.9044$, $r = 0.951$. 90% of the variation in Y is explained by X .
 b. $T = 6.155$; critical value: $t_{4, 0.975} = 2.776$. At $\alpha = .05$, we would reject H_0 .
 c. $(0.611, 0.995)$. Since $\rho = 0$ is not included in the interval, we reject $H_0: \rho = 0$ at $\alpha = .05$.
25. a. $r^2 = 0.2207$, $r = 0.470$. 22% of the variation in Y is explained by X .
 b. $T = 2.608$; critical value: $t_{24, 0.975} = 2.064$. At $\alpha = .05$, we would reject H_0 .
 c. $(0.101, 0.725)$. Since $\rho = 0$ is not included in the interval, we reject $H_0: \rho = 0$ at $\alpha = .05$.

Chapter 7**1. a. (1)**

	d.f.	Sum of Squares	Mean Sum of Squares	F
Regression	1	6.0785	6.0785	
Residual	9	2.0896	0.2322	
	10	8.1681		

(2)

	d.f.	Sum of Squares	Mean Sum of Squares	F
Regression	1	4.2211	4.2211	
Residual	9	0.0071	0.0008	
	10	4.2282		

- b.** (1) $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$; critical value: $F_{1, 9, 0.95} = 5.12$. Since $26.18 > 5.12$, we reject H_0 at $\alpha = .05$.
 (2) $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$; critical value: $F_{1, 9, 0.95} = 5.12$. Since $5355.59 > 5.12$, we reject H_0 at $\alpha = .05$.

5. a.

	d.f.	Sum of Squares	Mean Sum of Squares	F
Regression	1	450.8673	450.8673	
Residual	18	1,982.9141	110.1619	
	19	2,433.7814		

- b.** $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$; critical value: $F_{1, 18, 0.95} = 4.41$. $4.09 < 4.41$, so we do not reject H_0 at $\alpha = .05$.
c. $T^2 = (2.023)^2 = 4.09$. The values are the same.
d. The hypotheses for each test are equivalent. As mentioned in the text, the F statistic and T statistic are equivalent after squaring T . Using the information that the tests of hypotheses are equivalent (as are the test statistics), one may infer that the resulting P -values for each test should also be equivalent.

9. a.

	d.f.	Sum of Squares	Mean Sum of Squares	F
Regression	1	76,858,486.13	76,858,486.13	
Residual	28	35,419,546.54	1,264,983.00	
	29	112,278,032.67		

- b.** $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$; critical value: $F_{1, 28, 0.95} = 4.20$. Since $60.76 > 4.20$, we reject H_0 at $\alpha = .05$.

13. a.

	d.f.	Sum of Squares	Mean Sum of Squares	F
Regression	1	12,846,354	12,846,354	
Residual	14	593,585	42,399	
	15	13,439,939		

- b.** $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$; critical value: $F_{1, 14, 0.95} = 4.60$. Since $302.99 > 4.60$, we would reject H_0 and conclude that there is a significant linear relationship of Y on X at $\alpha = .05$.

	d.f.	Sum of Squares	Mean Sum of Squares	F
Regression	1	177.5297	177.5297	
Residual	58	3.3809	0.0583	
	59	180.9106		

- b. $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$; critical value: $F_{1, 58, 0.95} = 4.00$. Since $3,045.56 > 4.00$, we reject H_0 at $\alpha = .05$.

	d.f.	Sum of Squares	Mean Sum of Squares	F
Regression	1	132.6203	132.6203	
Residual	24	468.3412	19.5142	
	25	600.9615		

- b. $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$; critical value: $F_{1, 24, 0.95} = 4.26$. Since $6.80 > 4.26$, we reject H_0 at $\alpha = .05$.

Chapter 8

1. a. i. $\hat{Y} = 145.771$ ii. $\hat{Y} = 135.825$ iii. $\hat{Y} = 141.475$

As QUET increases from 3.0 to 3.5, average SBP increases by an estimated 4.296 points, from 141.475 to 145.771.

- b. SBP on AGE: $R^2 = 0.601$; SBP on AGE and SMK: $R^2 = 0.730$; SBP on AGE, SMK, and QUET: $R^2 = 0.761$.

The model using AGE and SMK to predict SBP appears to be the best choice.

	d.f.	Sum of Squares	Mean Sum of Squares	F
Regression	3	25,974.00	8,658.00	
Residual	21	2,248.23	107.06	
Total	24	28,222.23		

- b. $R^2 = 0.920$. There is a strong positive linear relationship between education resources and student performance.

9. a. As temperature increases from 20 to 25 degrees, the average oxygen consumption increases by an estimated 0.197 units.

- b. As weight increases from 0.25 to 0.5, the average oxygen consumption increases by an estimated 0.148 units.

- c. i. $R^2 = 0.019$ ii. $R^2 = 0.814$ iii. $R^2 = 0.943$

13. a. $\hat{Y} = 6.874 - 0.004X_2 - 0.234X_3$ b. $\hat{Y} = 3.734$

- c. $R^2 = 0.553$. The model explains about half of the variation in Yield (Y). The model has a limited ability to predict the yield for a company using 1989 ranking and P-E ratio as predictors.

15. a. $\hat{Y} = 130.468 + 0.089 \text{ zooplankton} - 0.025 \text{ phytoplankton}$

- b. $R^2 = 0.1332$. The model fit is poor.

Chapter 9

- 1. a.** i. $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ (Full model: $Y = \beta_0 + \beta_1 X_1 + E$)
 $F(X_1)_{1,30} = 45.18, P = .0001$. At $\alpha = .05$, we reject H_0 .
- ii. $H_0: \beta_1 = \beta_2 = 0$ $H_A:$ At least one $\beta_i \neq 0$ ($i = 1, 2$)
(Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$)
 $F(X_1, X_2)_{2,29} = 39.16, P < .0001$. At $\alpha = .05$, we reject H_0 .
- iii. $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ $H_A:$ At least one $\beta_i \neq 0$ ($i = 1, 2, 3$)
(Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$)
 $F(X_1, X_2, X_3)_{3,28} = 29.71, P < .0001$. At $\alpha = .05$, we reject H_0 .
- b. One would choose the parsimonious model tested in (a(i)).
5. a. i. $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ (Full model: $Y = \beta_0 + \beta_1 X_1 + E$)
 $F(X_1)_{1,40} = .40, P > .25$. At $\alpha = .05$, we do not reject H_0 .
- ii. $H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$ (Full model: $Y = \beta_0 + \beta_2 X_2 + E$)
 $F(X_2)_{1,40} = .19, P > .25$. At $\alpha = .05$, we do not reject H_0 .
- iii. $H_0: \beta_3 = 0$ $H_A: \beta_3 \neq 0$ (Full model: $Y = \beta_0 + \beta_3 X_3 + E$)
 $F(X_3)_{1,40} = 7.58, P = .009$. At $\alpha = .05$, we reject H_0 .
- b. $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ $H_A:$ At least one $\beta_i \neq 0$ ($i = 1, 2, 3$)
(Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$)
 $F(X_1, X_2, X_3)_{3,38} = 2.74, .05 < P < .10$. At $\alpha = .05$, we do not reject H_0 .
- c. $H_0: \beta_4 = \beta_5 = 0$ $H_A:$ At least one $\beta_i \neq 0$ ($i = 4, 5$)
(Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 X_3 + \beta_5 X_2 X_3 + E$)
 $F(X_1 X_3, X_2 X_3 | X_1, X_2, X_3)_{2,36} = 0.362, P > .25$. At $\alpha = .05$, we do not reject H_0 .
- d. $H_0: \beta_3 = 0$ $H_A: \beta_3 \neq 0$ (Full model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$)
 $F(X_3 | X_1, X_2)_{1,38} = 6.85, .01 < P < .025$. At $\alpha = .05$, we reject H_0 .
- e. X_3 is associated with Y , but the other two independent variables are not.
9. a. $H_0: \beta_1 = \beta_2 = 0$ $H_A:$ At least one $\beta_i \neq 0$ ($i = 1, 2$)
(Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$)
 $F(X_1, X_2)_{2,4} = 20.03, P = .0082$. At $\alpha = .05$, we reject H_0 .
- b. i. $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ in the model $Y = \beta_0 + \beta_1 X_1 + E$.
 $F(X_1)_{1,5} = 40.06, P = .0032$. At $\alpha = .05$, we reject H_0 . Note: We use the residual MS from the largest model. See Section 9.5.2.
- ii. $H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$ in the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$.
 $F(X_2 | X_1)_{1,4} = .01, P = .9344$. At $\alpha = .05$, we do not reject H_0 .
- c. i. $H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$ in the model $Y = \beta_0 + \beta_2 X_2 + E$.
 $F(X_2)_{1,5} = 37.83, .001 < P < .005$. At $\alpha = .05$, we reject H_0 .
- ii. $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ (Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$)
 $F(X_1 | X_2)_{1,4} = 9.80, P = .0352$. At $\alpha = .05$, we reject H_0 .
- d.
- | Source | d.f. | SS | MS | F | R ² |
|-------------|------|-----------|-----------|------|----------------|
| $X_1 X_2$ | 1 | 1,402.315 | 1,402.315 | 9.8 | 0.91 |
| $X_2 X_1$ | 1 | 1.098 | 1.098 | 0.01 | |
| Residual | 4 | 572.393 | 143.098 | | |
- e. X_1 is the only necessary predictor.

- 13.** **a.** $H_0: \beta_1 = \beta_2 = 0$ $H_A:$ At least one $\beta_i \neq 0$ ($i = 1, 2$)
 $(X_1 = \text{OWNCOST}, X_2 = \text{URBAN})$
(Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$)
 $F(X_1, X_2)_{2, 23} = 8.52, P = .017$. At $\alpha = .05$, we reject H_0 .
- b. i.** $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ (Full model: $Y = \beta_0 + \beta_1 X_1 + E$)
 $F(X_1)_{1, 24} = 8.84, P = .0068$. At $\alpha = .05$, we reject H_0 .
- ii.** $H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$ (Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$)
 $F(X_2 | X_1)_{1, 23} = 8.21, P = .0088$. At $\alpha = .05$, we reject H_0 .
- c. i.** $H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$ (Full model: $Y = \beta_0 + \beta_2 X_2 + E$)
 $F(X_1)_{1, 24} = 13.17, .001 < P < .005$. At $\alpha = .05$, we reject H_0 .
- ii.** $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ (Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$)
 $F(X_1 | X_2)_{1, 23} = 4.42, P = .0467$. At $\alpha = .05$, we reject H_0 .

Source	d.f.	SS	MS	F	R ²
$X_1 X_2$	1	66.334	66.334	4.42	0.43
$X_2 X_1$	1	123.158	123.158	8.21	
Residual	23	345.183	15.008		

e. Both predictors are necessary.

- 15.** **a.** $\hat{L} = 3\hat{\beta}_1 + 2\hat{\beta}_2 = 3(0.7220) + 2(2.0501) = 6.2662$
- b.** $S_L^2 = 3^2(0.0680) + 2^2(0.8784) + 2(3)(2)(-0.1500) = 2.3256$ and
 $S_L = \sqrt{2.3256} = 1.5250$.
- c.** The 95% CI is $6.2552 \pm 2.2622\sqrt{2.3256} = (2.7496, 9.7828)$.
- 17.** Additional models that include the predictors in different orders would need to be fit to assess the significance of
 - Drinking frequency given age
 - Drinking frequency given poor sleep
 - Age given sleep quality
 - Age alone
 - Sleep quality given age
 - Sleep quality given drinking frequency
 - Sleep quality alone

Chapter 10

- 1. a.** Age with an r value of 0.7752
- b. i.** $r_{\text{SBP, SMK}|AGE} = 0.568$ **ii.** $r_{\text{SBP, QUET}|AGE} = 0.318$
- c.** $H_0: \rho_{\text{SBP, SMK}|AGE} = 0$ $H_A: \rho_{\text{SBP, SMK}|AGE} \neq 0$
 $F(\text{SMK} | \text{AGE})_{1, 29} = 13.83, P < .001$. At $\alpha = .05$, we reject H_0 .
- d.** $H_0: \rho_{\text{SBP, QUET}|AGE, SMK} = 0$ $H_A: \rho_{\text{SBP, QUET}|AGE, SMK} \neq 0$
 $T_{28} = 1.91, P = .066$. At $\alpha = .05$, we do not reject H_0 .
- e.** Based on the results for a–d, we find that the following variables (ranked in order of their significance) helped explain the variation in SBP: (1) AGE, (2) SMK, (3) QUET.
- f.** $r^2_{\text{SBP}(QUET, SMK)|AGE} = 0.401$
 $H_0: \rho_{\text{SBP}(QUET, SMK)|AGE} = 0$ $H_A: \rho_{\text{SBP}(QUET, SMK)|AGE} \neq 0$

$$F(\text{QUET, SMK} \mid \text{AGE})_{2,28} = 9.371, P < .001.$$

The highly significant P -value suggests that both SMK and QUET are important variables, and the 0.160 increase in r^2 going from model 1, with only AGE (0.601), to model 3, with all 3 variables (0.761), indicates a small-moderate gain in prediction. There is good evidence for retaining these variables in the model.

- 4. a.** i. $r_{YX_1|X_3}^2 = 0.076$. ii. $r_{YX_2|X_3}^2 = 0.214$.
 iii. $r_{YX_3|X_3}^2 = 0.215$. The computations are nearly equivalent with the difference being due to round-off error.
- b. X_2 should be considered next for entry into the model because X_2 has a higher partial correlation than does X_1 .
- c. $H_0: \rho_{YX_2|X_3} = 0$ $H_A: \rho_{YX_2|X_3} \neq 0$
 $T_{17} = 2.154, .02 < P < .05$. At $\alpha = .05$, we reject H_0 .
- d. $r_{YX_1|X_2,X_3}^2 = 0.082$
 $H_0: \rho_{YX_1|X_2,X_3} = 0$ $H_A: \rho_{YX_1|X_2,X_3} \neq 0$
 $T_{16} = 1.199, P = .248$. At $\alpha = .05$, we do not reject H_0 .
- e. $r_{Y(X_1,X_2)|X_3}^2 = 0.279$
 $H_0: \rho_{Y(X_1,X_2)|X_3} = 0$ $H_A: r_{Y(X_1,X_2)|X_3} \neq 0$
 $F(X_1, X_2|X_3)_{2,16} = 3.098, .05 < P < .10$. At $\alpha = .05$, we do not reject H_0 .
- f. Based on the above results, only X_3 should be included in the model at $\alpha = .05$.
- 8. a.** i. $H_0: \rho_{YX_1} = 0$ $H_A: \rho_{YX_1} \neq 0$
 $F(X_1)_{1,45} = 0.89, P = .35$. At $\alpha = .05$, we do not reject H_0 .
 ii. $H_0: \rho_{YX_2} = 0$ $H_A: \rho_{YX_2} \neq 0$
 $F(X_2)_{1,45} = 197.58, P < .0001$. At $\alpha = .05$, we reject H_0 .
- b. i. $H_0: \rho_{YX_1|X_2} = 0$ $H_A: \rho_{YX_1|X_2} \neq 0$
 $F(X_1, X_2)_{1,44} = 98.605, P < .001$. At $\alpha = .05$, we reject H_0 .
 ii. $H_0: \rho_{YX_2|X_1} = 0$ $H_A: \rho_{YX_2|X_1} \neq 0$
 $F(X_2|X_1)_{1,44} = 709.641, P < .001$. At $\alpha = .05$, we reject H_0 .
- c. Both X_1 and X_2 should be included in the model with X_2 being more important than X_1 .
- 12. a.** $r_{Y|X_1, X_2}^2 = 0.909$ **b.** $r_{YX_2|X_1} = -\sqrt{0.002} = -0.045$
 c. $r_{YX_1|X_2} = \sqrt{0.710} = 0.843$
 d. $H_0: \rho_{YX_2|X_1} = 0$ $H_A: \rho_{YX_2|X_1} \neq 0$
 $T_4 = -0.09, P > .90$. At $\alpha = .05$, we do not reject H_0 .
 e. $H_0: \rho_{YX_1|X_2} = 0$ $H_A: \rho_{YX_1|X_2} \neq 0$
 $T_4 = 3.131, .02 < P < .05$. At $\alpha = .05$, we reject H_0 .
 f. X_1 should be included in the model, while X_2 should not be included.
- 16. a.** $r_{Y|X_1, X_2}^2 = 0.426$ **b.** $r_{YX_2|X_1} = -\sqrt{0.263} = -0.513$
 c. $r_{YX_1|X_2} = -\sqrt{0.161} = -0.401$
 d. $H_0: \rho_{YX_2|X_1} = 0$ $H_A: \rho_{YX_2|X_1} \neq 0$
 $T_{23} = -2.865, .001 < P < .01$. At $\alpha = .05$, we reject H_0 .
 e. $H_0: \rho_{YX_1|X_2} = 0$ $H_A: \rho_{YX_1|X_2} \neq 0$
 $T_{23} = -2.1, .02 < P < .05$. At $\alpha = .05$, we reject H_0 .
 f. Both variables should be included in the model with X_2 being the more important predictor of Y .

Chapter 11

- 1. a.** $\text{WGT} = \beta_0 + \beta_1 \text{HGT} + \beta_2 \text{AGE} + \beta_3 \text{AGE}^2 + E$
- b.** $\hat{\beta}_1$ does not change when either AGE or AGE² is removed from the model. However, $\hat{\beta}_1$ changes meaningfully when both AGE and AGE² are removed from the model. Thus, there is evidence of confounding due to AGE and AGE².
- c.** AGE² can be dropped from the model because $\hat{\beta}_1$ does not change meaningfully.
- d.** AGE² should not be retained in the model because the 95% CI for β_1 is narrower when AGE² is absent from the model.
- e.** Considering the change in $\hat{\beta}_1$ and the width of the 95% CI, the final model should be $\text{WGT} = \beta_0 + \beta_1 \text{HGT} + \beta_2 \text{AGE} + E$.
- f.** Revise the initial model as
- $$\text{WGT} = \beta_0 + \beta_1 \text{HGT} + \beta_2 \text{AGE} + \beta_3 \text{AGE}^2 + \beta_4 \text{HGT} * \text{AGE} + \beta_5 \text{HGT} * \text{AGE}^2 + E.$$
- g.** We would test for interaction by performing a F test for $H_0: \beta_4 = \beta_5 = 0$. If this test proved significant, we would perform separate partial F tests to assess $H_0: \beta_4 = 0$ and $H_0: \beta_5 = 0$.
- 3. a.** There is no confounding due to X_2 because $\hat{\beta}_1$ does not change when X_2 is removed from the model.
- b.** $r_{YX_1} = 0.265 \quad r_{YX_1|X_2} = \sqrt{0.5} = 0.707$
Since the two correlation coefficients are qualitatively and statistically significantly different, we conclude that confounding exists.
- c.** The conclusions for confounding depend on the definition of confounding.
- d.** Since $H_0: \beta_2 = 0$ is rejected ($P = .0005$), we might conclude that confounding exists, which is contradictory to part (a). This illustrates the potential problem of using significance testing to make judgments about confounding.
- 5. a. i.** $H_0: \beta_1 = 0 \quad H_A: \beta_1 \neq 0 \quad F(X_1)_{1,40} = 0.4, P = .528$. At $\alpha = .05$, we do not reject H_0 .
- ii.** $H_0: \beta_2 = 0 \quad H_A: \beta_2 \neq 0 \quad F(X_2)_{1,40} = 0.19, P = .669$. At $\alpha = .05$, we do not reject H_0 .
- iii.** $H_0: \beta_3 = 0 \quad H_A: \beta_3 \neq 0 \quad F(X_3)_{1,40} = 7.58, P = .009$. At $\alpha = .05$, we reject H_0 .
- b.** $H_0: \beta_1 = \beta_2 = \beta_3 = 0 \quad H_A: \text{At least one } \beta_i \neq 0 \ (i = 1, 2, 3)$
 $F(X_1, X_2, X_3)_{3,38} = 2.737, .05 < P < .1$. At $\alpha = .05$, we do not reject H_0 .
- c.** $H_0: \rho_{Y(X_1X_3, X_2X_3)|X_1, X_2, X_3} = 0 \quad H_A: \rho_{Y(X_1X_3, X_2X_3)|X_1, X_2, X_3} \neq 0$
 $F(X_1X_3, X_2X_3|X_1, X_2, X_3)_{2,36} = 0.362, P > .25$. At $\alpha = .05$, we do not reject H_0 . The two regression lines are parallel.
- d.** $H_0: \rho_{YX_3|X_1, X_2} = 0 \quad H_A: \rho_{YX_3|X_1, X_2} \neq 0$
 $F(X_3|X_1, X_2)_{1,38} = 6.855, P = .014$. At $\alpha = .05$, we do not reject H_0 .
- e.** First fit the full model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E \quad (1)$$

Next, fit the reduced models

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + E \quad (2)$$

$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + E \quad (3)$$

$$Y = \beta_0 + \beta_3 X_3 + E \quad (4)$$

We assess confounding by noting how $\hat{\beta}_3$ changes for the different models. In particular, if $\hat{\beta}_3$ from model (2), (3), or (4) differs from $\hat{\beta}_3$ from model (1), then X_1 , X_2 , or X_1 and X_2 , respectively, are confounders. To assess precision, we note how the $100(1 - \alpha)\%$ CI's for $\hat{\beta}_3$ change. We only eliminate potential confounders from the model if the width of the CI for $\hat{\beta}_3$ does not widen significantly.

- f. From the information provided, we can assess the confounding effects of X_1 or X_2 alone with respect to X_3 but not for X_1 and X_2 taken together.
7. a. No, there is no meaningful change in the estimate for $\hat{\beta}_1$ when X_2 is added to the model.
- b. No, the confidence interval for $\hat{\beta}_1$ is narrower when only X_1 is in the model.
- c. No, there is not evidence suggesting that including X_2 the model improves the precision and/or the validity of the estimated relationship between X_1 and Y .
9. a. Let $\text{OWNCOST} = X_1$ and $\text{INCOME} = X_2$
 $H_0: \beta_1 = \beta_2 = 0 \quad H_A: \text{At least one } \beta_i \text{ does not equal 0}$
(Full model: $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + E$)
 $F(X_1, X_2)_{2, 23} = 6.38, P = .006$. At $\alpha = .05$, we reject H_0 .
- b. $H_0: \beta_1 = 0 \quad H_A: \beta_1 \neq 0$ (Full model: $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + E$)
 $F(X_1|X_2)_{1, 23} = 11.47, P = .003$. At $\alpha = .05$, we reject H_0 .
- c. $H_0: \beta_2 = 0 \quad H_A: \beta_2 \neq 0$ (Full model $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + E$)
 $F(X_2|X_1)_{1, 23} = 4.87, P = .038$. At $\alpha = .05$, we reject H_0 .
- d. Including X_2 does meaningfully change $\hat{\beta}_1$, and it should therefore be included in the model as a confounder, assuming there is no interaction between X_1 and X_2 .
11. a. The test statistic value is $[(4.92 + 8.58 + 106.03)/3]/34.50 = 1.155$. Based on the $F_{3, 1042}$ distribution, the P -value is about 0.20, indicating no statistical evidence of interaction.
- b. Because there is virtually no difference in the estimated regression coefficient for drinking frequency between the above output and that found in Section 5.12 (-0.015 in both), one may conclude that there is no confounding of the drinking frequency–BMI relationship due to age and sleep quality. These factors need not be controlled for.
- c. Although the P -value associated with sleep quality of .015 indicates that this factor is statistically significantly related to BMI, it is not needed to attain an unbiased estimate of the drinking frequency–BMI association. If that is the objective, then control for sleep quality is not needed.

Chapter 12

1. a. For smokers: $\hat{Y} = 79.225 + 20.118X$. For nonsmokers: $\hat{Y} = 49.312 + 26.303X$.
- b. $H_0: \beta_{1(\text{SMK}=1)} = \beta_{1(\text{SMK}=0)} \quad H_A: \beta_{1(\text{SMK}=1)} < \beta_{1(\text{SMK}=0)}$
 $T_{28} = 0.892, .15 < P < .25$. At $\alpha = .05$, we do not reject H_0 that the slopes for smokers and nonsmokers are the same.
- c. $H_0: \beta_{0(\text{SMK}=1)} = \beta_{0(\text{SMK}=0)} \quad H_A: \beta_{0(\text{SMK}=1)} \neq \beta_{0(\text{SMK}=0)}$
 $T_{28} = -1.24, .20 < P < .30$. At $\alpha = .05$, we do not reject H_0 , and we conclude that the two intercepts are equal.
- d. The straight lines for smokers and nonsmokers are coincident since both tests failed to reject H_0 .

5. a. For NY: $\hat{Y} = 2.174 + 1.177X$. For CA: $\hat{Y} = 8.030 + 1.036X$.
- b. $H_0: \beta_{1NY} = \beta_{1CA}$ $H_A: \beta_{1NY} > \beta_{1CA}$
 $T_{33} = 1.115, .10 < P < .15$. At $\alpha = .05$, we do not reject H_0 , and we conclude that the slopes are the same for NY and CA.
- c. $H_0: \beta_{0NY} = \beta_{0CA}$ $H_A: \beta_{0NY} > \beta_{0CA}$
 $T_{33} = 1.219, .85 < P < .90$. At $\alpha = .05$, we do not reject H_0 that the two intercepts are equal for NY and CA.
- d. Since the tests for equal slopes and equal intercepts did not lead to rejection, we can conclude that the lines are coincident.
- e. $H_0: \rho_{NY} = \rho_{CA}$ $H_A: \rho_{NY} \neq \rho_{CA}$
 $Z = 0.252, P > .80$. At $\alpha = .05$, we do not reject H_0 , and we conclude that the correlation coefficients for each straight line regression are not significantly different.
9. a. $SBP = \beta_0 + \beta_1AGE + \beta_2QUET + \beta_3SMK + \beta_4AGE*SMK + \beta_5QUET*SMK + E$
Smokers: $SBP = (\beta_0 + \beta_3) + (\beta_1 + \beta_4)AGE + (\beta_2 + \beta_5)QUET + E$
Nonsmokers: $SBP = \beta_0 + \beta_1AGE + \beta_2QUET + E$
- b. Smokers: $\widehat{SBP} = 48.076 + 1.466(AGE) + 6.744(QUET)$
Nonsmokers: $\widehat{SBP} = 48.613 + 1.029(AGE) + 10.451(QUET)$
- c. $H_0: \beta_4 = \beta_5 = 0$ $H_A: \text{At least one } \beta_i \neq 0$ (Full model given in part (a).)
 $F(QUET*SMK, AGE*SMK | AGE, QUET, SMK)_{2, 26} = 0.222, P > .25$
At $\alpha = .05$, we do not reject H_0 and conclude the two lines are coincident.
- d. $H_0: \beta_3 = \beta_4 = \beta_5 = 0$ $H_A: \text{At least one } \beta_i \neq 0$ (Full model given in part (a).)
 $F(SMK, QUET*SMK, AGE*SMK | AGE, QUET)_{3, 26} = 4.562, .01 < P < .025$
At $\alpha = .05$, we reject H_0 and conclude the two lines are not coincident.
13. a. $R = 1$ and $TD = 1$:
 $\widehat{SBPL} = (\hat{\beta}_0 + \hat{\beta}_2 + \hat{\beta}_4 + \hat{\beta}_9) + \hat{\beta}_1(SBP1) + (\hat{\beta}_6 + \hat{\beta}_{11})RW$
 $R = 0$ and $TD = 1$:
 $\widehat{SBPL} = (\hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_4 + \hat{\beta}_7) + \hat{\beta}_1(SBP1) + (\hat{\beta}_6 + \hat{\beta}_{13})RW$
- b. $H_0: \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = 0$ $H_A: \text{At least one } \beta_i \neq 0$.
 $F(X_{11}, X_{12}, X_{13}, X_{14} | X_1, \dots, X_{10})_{4, 89} = 1.410, .1 < P < .25$. At $\alpha = .05$, we do not reject H_0 , and we conclude that the lines are parallel.
- c. H_0 : The three regression lines corresponding to rural, town, and urban background are parallel (i.e., $H_0: \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = 0$)
 $H_A: \text{At least one } \beta_i \neq 0$
 $F(X_7, \dots, X_{14} | X_1, \dots, X_6)_{8, 89} = 1.103$ with $P > .25$. At $\alpha = .05$, we do not reject H_0 , and we conclude that the three regression lines are parallel.
- d. $H_0: \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = 0$
 $H_A: \text{At least one } \beta_i \neq 0$
 $F(X_2, X_3, X_4, X_5, X_7, \dots, X_{14} | X_1, X_6)_{12, 89}$

$$\frac{\text{SS Regression (full model)} - \text{SS Regression } (X_1, X_6)}{12} = \frac{}{\text{MS Residual (full model)}}$$

- 17.** **a.** $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + E$, where $Z = 1$ if cool, 0 if warm.
- b.** For cool: $\hat{Y} = 104.003 + 2.465X$. For warm: $\hat{Y} = 96.830 + 3.485X$.
- c.** $H_0: \beta_2 = \beta_3 = 0$ $H_A: \text{At least one } \beta_i \neq 0$ (Full model given in part (a).) $F(Z, XZ | X)_{2, 14} = 41.875, P = .0076$. At $\alpha = .05$, we reject H_0 and conclude that the lines do not coincide.
- d.** $H_0: \beta_3 = 0$ $H_A: \beta_3 \neq 0$ (Full model given in part (a).) $F(XZ | X, Z)_{1, 14} = 9.699, P < .01$. At $\alpha = .05$, we reject H_0 and conclude that the lines are not parallel.
- e.** Baseline sales are higher during the warm season than in the cool season. Advertising expenditures are higher in the cool season than in the warm season. By spending more money in advertising during the cool season, retailers are able to surpass the sales revenue of the warm season.
- 21.** **a.** $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + E$
- b.** $H_0: \beta_2 = \beta_3 = 0$ $H_A: \text{At least one } \beta_i \neq 0$ (Full model given in part (a).) $F(Z, X_1 Z | X_1)_{2, 50} = 2.704, .05 < P < .10$. At $\alpha = .05$, we do not reject H_0 , and we conclude that the lines coincide.
- d.** $H_0: \beta_3 = 0$ $H_A: \beta_3 \neq 0$ (Full model given in part (a).) $F(X_1 Z | X_1, Z)_{1, 50} = 3.587, P = .064$. At $\alpha = .05$, we do not reject H_0 , and we conclude that the lines are parallel.
- e.** The change in refraction–baseline refractive relationship is the same for males and females.

Chapter 13

- 1.** **a.** $\text{SBP} = \beta_0 + \beta_1(\text{QUET}) + \beta_2\text{SMK} + E$
- b.** For smokers: $\text{SBP} = (\beta_0 + \beta_2) + \beta_1(\text{QUET}) + E$; $\overline{\text{SBP}}_{(\text{adj})} = 148.548$
For nonsmokers: $\text{SBP} = \beta_0 + \beta_1(\text{QUET}) + E$; $\overline{\text{SBP}}_{(\text{adj})} = 139.977$
- c.** $H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$ in the model $\text{SBP} = \beta_0 + \beta_1(\text{QUET}) + \beta_2\text{SMK} + E$ $T_{29} = 2.707, P = .011$. At $\alpha = .05$, we reject H_0 and conclude that mean SBP differs for smokers and for nonsmokers, after adjusting for QUET.
- d.** Finding the 95% confidence interval for the true difference in adjusted mean SBP is equivalent to finding the 95% confidence interval for $\hat{\beta}_2$. The 95% confidence interval for $\hat{\beta}_2$ is (2.094, 15.048).
- 5.** **a.** $\text{VIAD} = \beta_0 + \beta_1\text{IQM} + \beta_2\text{IQF} + \beta_3 Z + E$, where $Z_1 = 1$ if female, 0 if male.
- b.** For males: $\text{VIAD} (\text{adj}) = -3.307$ versus -3.00 unadjusted
For females: $\text{VIAD} (\text{adj}) = 1.889$ versus 1.60 unadjusted
- c.** $H_0: \beta_3 = 0$ $H_A: \beta_3 \neq 0$ in the model $\text{VIAD} = \beta_0 + \beta_1\text{IQM} + \beta_2\text{IQF} + \beta_3 Z + E$ $T_{16} = 1.659, P = .117$. At $\alpha = .05$, we do not reject H_0 , and we conclude that the mean scores do not significantly differ by gender, after adjusting for IQM and IQF.
- d.** 95% CI: $(-1.446, 11.838)$
- 9.** **a.** $\text{LN_BRNTL} = \beta_0 + \beta_1\text{WGT} + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 Z_3 + E$, where $Z_1 = 1$ if 100 ppm, 0 otherwise; $Z_2 = 1$ if 500 ppm, 0 otherwise; $Z_3 = 1$ if 1000 ppm, 0 otherwise.
- b.** $\hat{\beta}_0 = -0.764$ $\hat{\beta}_1 = 0.0006$ $\hat{\beta}_2 = 0.828$ $\hat{\beta}_3 = 3.571$ $\hat{\beta}_4 = 4.214$

c. PPM_TOLU	Adjusted Means	Unadjusted Means
50	-0.537	-0.548
100	0.291	0.282
500	3.034	3.019
1000	3.677	3.668

- d. $H_0: \beta_2 = \beta_3 = \beta_4 = 0$ $H_A: \text{At least one } \beta_i \neq 0 (i = 2, 3, 4)$
(Full model: $\text{LN_BRNTL} = \beta_0 + \beta_1 \text{WGT} + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 Z_3 + E$)
 $F(Z_1, Z_2, Z_3 | \text{WGT})_{3, 55} = 1662.526, P < .001$. At $\alpha = .05$, we reject H_0 and conclude that the adjusted means significantly differ.

11. a. Using PPM_TOLU as the reference group, we define the cross-product terms as

$$\begin{aligned} XZ_1 &= \text{WGT} && \text{if PPM_TOLU} = 100, \quad 0 \text{ otherwise} \\ XZ_2 &= \text{WGT} && \text{if PPM_TOLU} = 500, \quad 0 \text{ otherwise} \\ XZ_3 &= \text{WGT} && \text{if PPM_TOLU} = 1000, \quad 0 \text{ otherwise} \end{aligned}$$

in which X stands for WGT.

- b. The appropriate regression model is

$$\begin{aligned} \text{LN_BRNTL} &= \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 Z_3 + \beta_5 XZ_1 + \beta_6 XZ_2 \\ &\quad + \beta_7 XZ_3 + E \end{aligned}$$

in which X stands for WGT.

- c. The null hypothesis for this test is represented by:

$$H_0: \beta_5 = \beta_6 = \beta_7 = 0.$$

12. a. $Y = \beta_0 + \beta_1 \text{AGE} + \beta_2 Z + E$

- b. $Y = \beta_0 + \beta_1 \text{AGE} + \beta_2 Z + \beta_3 \text{AGE}^*Z + E$

$$\begin{aligned} H_0: \beta_3 &= 0 && H_A: \beta_3 \neq 0 \text{ (Full model: } Y = \beta_0 + \beta_1 \text{AGE} + \beta_2 Z \\ &&& + \beta_3 \text{AGE}^*Z + E) \end{aligned}$$

$T_{26} = -1.677, P = .106$. At $\alpha = .05$, we do not reject H_0 , and we conclude that the ANACOVA model in part (a) is appropriate.

c. Location	Adjusted Means	Unadjusted Means
Intown/inner	81.526	82.187
Outer	85.46	84.807

$$H_0: \beta_2 = 0 \quad H_A: \beta_2 \neq 0$$

$T_{27} = 1.212, P = .236$. At $\alpha = .05$, we do not reject H_0 , and we conclude that the adjusted means do not significantly differ.

14. a. Let X denote OWNCOST and Y denote OWNEROCC.

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + E$$

- b. $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + E$

$$H_0: \beta_3 = 0 \quad H_A: \beta_3 \neq 0$$

$$F(XZ | X, Z)_{1, 22} = \frac{0.696}{17.580} = 0.04, \text{ with } P \gg .25.$$

At $\alpha = .05$, we do not reject H_0 and conclude the ANACOVA model in part (a) is appropriate.

c. Location	Adjusted Means	Unadjusted Means
Urban $\geq 75\%$	56.15%	64.5%
Urban $\leq 75\%$	60.06%	69.25%

$$H_0: \beta_2 = 0 \quad H_A: \beta_2 \neq 0$$

$$T_{23} = -2.191, P = .039.$$

At $\alpha = .05$, we reject H_0 and conclude the adjusted means significantly differ.

16. a. The problem describes a hypothesis test of $H_0: \beta_3 = 0$ versus the alternative hypothesis $H_A: \beta_3 \neq 0$. The given output only allows for a *t* test, with the test statistic value equal to $-2.470/0.449 = -5.50$. The associated *P*-value of $< .0001$ is also given, and we would reject the null hypothesis.
- b. Both β_3 and β_4 would need to be equal to zero for these adjusted means to be equal. Chapter 9 presented methods for testing such a null hypothesis. Using the provided output we compute the linear contrast $\hat{l} = \hat{\beta}_3 + \hat{\beta}_4$, which is equal to $-2.470 - 1.108 = -3.578$. Using the covariance matrix, we compute the estimated standard error of this contrast as $S_{\hat{l}} = \sqrt{(1)(0.2019) + (1)(0.2421) + 2(1)(1)(0.0233)} = 0.4906$. The test statistic is then equal to $3.578/0.4906 = -7.29$, which provides strong statistical evidence for rejecting the null hypothesis that both regression coefficients are equal to zero.
- c. The estimated linear function of interest is \hat{Y}_{11} , given earlier; its estimated variance $\widehat{\text{var}}(\hat{Y}_{11})$ can be obtained as described in Section 9.6.5. The 99% confidence interval is $\hat{Y}_{11} \pm t_{1048, 0.995} \sqrt{\widehat{\text{var}}(\hat{Y}_{11})}$.
- d. If we consider the estimate $\hat{\beta}_4$, which is adjusted for the two continuous covariates, to be a more valid adjusted estimate of the effect of tobacco_now on BMI, then dropping the continuous variables from the model changes this estimate by 23%, which is a rather sizable change. Based on this numerical finding, there seems to be sufficient evidence that confounding is an issue, and so controlling for drinking frequency and sleep quality is a reasonable approach.

Chapter 14

1. a. The plot of jackknife residuals versus the predicted values shows a distinct pattern, indicating that an assumption has been violated. In this case, the obvious curvilinear pattern indicates a violation of the linearity assumption (much as the simple plot of Y versus X did in Chapter 5, Problem 1). A linearizing transformation should be applied; for example, $\log_{10}(\text{dry weight})$ can be used as the dependent variable.
- b. The skewness statistic ($=1.66$) and kurtosis statistic ($=3.21$) suggest that at least a moderate violation of the normality assumption has occurred. The normal probability plot also looks fairly nonlinear. With such few data values, it is difficult to conclude that a gross violation of the linearity assumption has occurred. Since a violation of the linearity assumption has occurred (see part (a)), no attempt should be made to correct for the possible violation of normality until the linearity issue is addressed.
- c. Observation 11 has a Cook's distance value greater than 1. No observations have leverage values greater than $2(k + 1)/n = 0.36$. The data for observation 11

should be double-checked and corrected, if necessary; if the recorded value is correct, it should be judged as to plausibility. If judged to be implausible, the observation can be removed from the analysis. If plausible, no corrective action is taken; instead, the analysis can be run with and without the observation included and the regression results compared to judge the impact of the outlier.

3. a. The plot of jackknife residuals versus the predicted values shows a distinct pattern, indicating that an assumption has been violated. In this case, the obvious curvilinear pattern indicates a violation of the linearity assumption (much as the simple plot of Y versus X did in Chapter 5, Problem 3). In this problem, it may help to add an X^2 term to the model.
 - b. The skewness and kurtosis statistics are both less than 1 in magnitude; the normal probability plot is not grossly nonlinear. This suggests that there is no gross violation of the normality assumption.
 - c. No Cook's distance values are greater than 1. Four observations have leverage values greater than $2(k + 1)/n = 0.20$. The data for these observations should be double-checked and corrected, if necessary; if the recorded values are correct, they should be judged as to plausibility. If judged to be implausible, an observation can be removed from the analysis. If plausible, no corrective action is taken; instead, the analysis can be run with and without the observation included and the regression results compared to judge the impact of the outlier.
19. a. The plot of jackknife residuals versus predicted values looks like a random scatter of points; since no pattern is evident, no assumptions appear to be violated based on this plot.
 - b. The skewness and kurtosis statistics are both less than 1 in magnitude; the normal probability plot is fairly linear. This suggests that there is no gross violation of the normality assumption.
 - c. No Cook's distance values are greater than 1. The leverage value for observation 18 is greater than $2(k + 1)/n = 0.42$. The data for this observation should be double-checked and corrected, if necessary.
 - d. None of the variance inflation factors is larger than 10, and none of the condition indexes is greater than 30. Therefore, no collinearity problem exists.
25. a. $Y = \beta_0 + \beta_1 (\text{ADVERTISING}) + E$, where Y denotes sales.
 - b. The largest studentized residual (absolute value) = 1.527, which is no cause for alarm.
 - c. The plot of the jackknife residuals versus the predictor does not suggest any problems.

Chapter 15

1. b. From the computer output, we find
 - (1) Degree 1: $\hat{Y} = -1.932 + 0.246X$
 - (2) Degree 2: $\hat{Y} = 3.172 - 0.781X + 0.047X^2$
 - (3) $\ln Y$ on X : $\ln \hat{Y} = -6.21 + 0.451X$
 - (4) The above fitted equations are plotted on the graphs presented for part 1(a).

c.	Source	d.f.	SS	MS	F
	Regression	1	12.7054	12.7054	43.69
	Residual Lack of fit Pure error	4 12	4.4196 0.2325	1.1049 0.0194	56.95
	Total	17	17.3575		

d.	Source	d.f.	SS	MS	F
	Regression Degree 1(X) Degree 2 ($X^2 X$)	1 1	12.7054 3.9051	12.7054 3.9051	255.13 78.42
	Residual Lack of fit Pure error	3 12	0.5145 0.2325	0.1715 0.0194	8.84
	Total	17	17.3575		

e. $r_{XY}^2 = 0.732$; $r^2(\text{quadratic}) = 0.957$

f. **Test for significance of straight-line regression of Y on X**

H_0 : The straight-line regression is not significant.

$F_{1, 16} = 43.69$, $P < .001$. At $\alpha = .05$, we reject H_0 and conclude that the straight-line regression is significant.

Test for adequacy of straight-line model

H_0 : The straight-line model is adequate.

$F_{4, 12} = 56.95$, $P < .001$. At $\alpha = .05$, we reject H_0 and conclude that the straight-line model is not adequate.

g. **Test for significance of quadratic regression**

H_0 : The quadratic regression is not significant.

$F_{2, 15} = 166.77$, $P = .0001$. At $\alpha = .05$, we reject H_0 and conclude that the quadratic regression is significant.

Test for addition of X^2 term

H_0 : The addition of X^2 to a model already containing X is not significant.

Partial $F(X^2|X)_{1, 15} = 78.42$, $P = .0001$. At $\alpha = .05$, we reject H_0 and conclude that the addition of X^2 is significant.

Test for adequacy of quadratic model

H_0 : The quadratic model is adequate.

$F_{3, 12} = 8.84$, $P = .002$. At $\alpha = .05$, we reject H_0 and conclude that the quadratic model is not adequate.

h. **Test for significance of straight-line regression of $\ln Y$ on X**

H_0 : The straight-line regression is not significant.

$F_{1, 16} = 4,277.167$, $P = .0001$. At $\alpha = .05$, we reject H_0 and conclude that the straight-line regression is significant.

Test for adequacy of straight-line model of $\ln Y$ on X

H_0 : The straight-line model is adequate.

$F_{4, 12} = 0.896$, $P = .471$. At $\alpha = .05$, we do not reject H_0 and conclude that the straight-line model is adequate.

- i. R^2 (straight-line regression of $\ln Y$ on X) = 0.9965
 R^2 (quadratic regression of Y on X) = 0.957
A comparison of the above two results for R^2 shows that the straight-line fit of $\ln Y$ on X provides a better fit.
- j. (1) Homoscedasticity assumption appears to be much more reasonable when using $\ln Y$ on X than when using Y on X .
(2) The straight-line regression of $\ln Y$ on X is preferred.
- k. The independence assumption is violated.
5. b. **Test for significance of straight-line regression**
 H_0 : The straight-line regression is not significant.
 $F_{1, 24} = 978.04, P < .001$. At $\alpha = .05$, we reject H_0 and conclude that the straight-line regression is significant.
- Test for adequacy of straight-line model**
 H_0 : The straight-line model is adequate.
 $F_{16, 8} = 1.57, P > .25$. At $\alpha = .05$, we do not reject H_0 and conclude that the straight-line model is adequate.
- c. **Test for addition of X^2 to the model**
 H_0 : The addition of X^2 is not significant.
Partial $F(X^2|X)_{1, 23} = 0.55, P > .25$. At $\alpha = .05$, we do not reject H_0 .
- d. The straight-line model is most appropriate.
9. a. Fit the model $VOC_SIZE = \beta_0 + \beta_1(AGE^*) + \beta_2(AGE^*)^2 + \beta_3(AGE^*)^3 + E$, where $AGE^* = AGE - 2.867$.
 $\hat{\beta}_0 = 741.84$ $\hat{\beta}_1 = 645.60$ $\hat{\beta}_2 = 70.43$ $\hat{\beta}_3 = -31.18$
- b. Using variables-added-in-order tests, the best model includes AGE^* , $(AGE^*)^2$, and $(AGE^*)^3$.
- c. Using variables-added-last tests, the best model includes AGE^* , $(AGE^*)^2$, and $(AGE^*)^3$.
- d. The only large predictor correlation is between AGE^* and $(AGE^*)^3$. The largest condition index ($CI_3 = 6.05$) suggests that the centered data do not have any serious collinearity problems.
- f. The estimated regression coefficients differ from those in Problem 8, but the best model includes the linear, quadratic, and cubic terms as in Problem 8. Also, the sums of squares for the variables-added-in-order test are the same as in Problem 8. The centering of AGE greatly reduced the previous collinearity problems. Centering does not affect the residual diagnostics.
13. a. The estimated equation is
 $\hat{Y} = 10.64 + 94.42(LIN_TOL) + 14.59(QUAD_TOL) - 0.73(CUB_TOL)$, where LIN_TOL , $QUAD_TOL$, and COL_TOL denote the centered PPM_TOL orthogonal polynomials.
- b. Using the variable-added-in-order tests, the best model includes LIN_TOL and $QUAD_TOL$.
- c. Using the variable-added-last tests, the best model includes LIN_TOL and $QUAD_TOL$, which is the same model as in part (b).
- d. The orthogonal polynomials are uncorrelated with each other, which implies that any collinearities are eliminated as shown by the condition indices.
- e. The residual plots suggest that the variances increase as the predicted values increase.

Chapter 16

1. a. The final model by forward selection is $\widehat{WGT} = 37.600 + 0.053(AGE*HGT)$
 b. The final model by backward elimination is $\widehat{WGT} = 37.600 + 0.053(AGE*HGT)$
 c. The best model resulting from this approach is

$$\widehat{WGT} = 6.553 + 0.722(HGT) + 2.050(AGE)$$

 d. The first two approaches result in the model including only the AGE*HGT interaction. It is difficult to interpret results for such a model, since the variables that make up the interaction term are not included. The third approach results in a model that is easier to interpret. It is possible to conduct modified forward and backward stepwise strategies in which interaction terms are only considered for inclusion if the terms that make up the interaction term are already in the model.
3. For smokers: the final estimated model is $\widehat{SBP} = 102.200 + 0.252(AGE*QUET)$
 For nonsmokers: the final estimated model is $\widehat{SBP} = 93.073 + 0.250(AGE*QUET)$
 These two models are different from those obtained from Problem 2(d) by putting SMK = 1 (for smokers) and SMK = 0 (for nonsmokers).
5. The best regression models using the sequential procedure of adding AGE first for Females and Males are
 Females: $\widehat{DEP} = 190.012 - 1.099AGE - 1.217MC$
 Males: $\widehat{DEP} = 270.056 - 2.514AGE - 1.065MC$
 For females, the above model was selected as best because of its high r^2 (0.401), satisfactory $C(P)$ (2.238), and low MSE (3,274.364). The next best model was the full model with $r^2 = 0.413$, $C(P) = 4.0$, and MSE = 3,478.318. Emphasizing parsimony, we find that the above model is more favorable than the full model.
 The reasoning is similar for selecting the model shown above for males.
7. a. Using the $C(P)$ criterion exclusively, we find that the three models with the most favorable $C(P)$ values contain AGE alone (1.23), AGE and WGT (2.63), or all three variables (4). None of these models is particularly impressive, and since it would be hard to argue that either of the multiple-variable models is better than the model with AGE alone, we could take AGE alone as the best model. Upon further investigation, the difficulty is seen to be partly due to none of these models having a significant overall F test.
 b. Since there seems to be no rationale for grouping the variables (age, weight, and height) in any way, chunks are taken to be the three variable-specific pairs of linear/quadratic terms (i.e., AGE_C and AGE_CSQ; HGT_C and HGT_CSQ; WGT_C and WGT_CSQ, where the _C terms are the centered variables and the _CSQ terms are the squared centered terms). A plausible forward chunkwise strategy is to treat each chunk/pair as a distinct entity that cannot be split and then proceed in the usual forward manner (use $\alpha = .10$):
 Step 1: WGT_C, WGT_CSQ added to the model. F test = 2.78
 Step 2: AGE_C, AGE_CSQ added to the model. F test = 3.49
 Step 3: Stop HGT_C, HGT_CSQ not significant. F test = 0.16

- c. The all-possible-regressions method yields the following “best” models for each of the model sizes:

Number in Model	Variables	$C(P)$	R^2	MSE
1	WGT_CSQ	8.026	0.078	0.771
2	WGT_C, WGT_CSQ	5.213	0.300	0.631
3	WGT_C, WGT_CSQ, AGE_CSQ	4.337	0.432	0.554
4	WGT_C, WGT_CSQ, AGE_C, AGE_CSQ	3.312	0.571	0.456
5	WGT_C, WGT_CSQ, AGE_C, AGE_CSQ, HGT_CSQ	5.225	0.576	0.497
6	Full model	7.00	0.586	0.539

The best model is most likely the four-variable model including WGT_C, WGT_CSQ, AGE_C, and AGE_CSQ. The R^2 , $C(P)$, and MSE for this model are better than for any of the smaller models; and they are similar to, if not better than, the statistics for the larger models, which, of course, are less parsimonious.

- d. Any model containing only first-order terms (part (a)) is seriously deficient. The model in parts (b) and (c) is the best one.
9. a. A model containing X_1 , X_3 , and X_4 is the best model by this method. The model has a relatively high R^2 , a satisfactory $C(P)$, and one of the lowest MSEs. The model also has the benefit of parsimony compared to the larger models.
- b. The selected model contains X_1 , X_3 , and X_4 .
- c. The selected model contains X_1 , X_3 , and X_4 .
- d. All three methods selected the same model, and it appears to be the best model for the reasons cited in part (a).
11. a. The model containing X_1 and X_3 would be the recommended model. Its R^2 , $C(P)$, and MSE are clearly superior to the best one-variable model statistics, and they are similar to the statistics for the less parsimonious full model.
- b. The results of the stepwise regression show that the model containing X_1 and X_3 is the best model.
- c. The model: $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + E$ appears to be best. This model is chosen given the results in parts (a) and (b).

Chapter 17

1. a. Treatment	Mean	S
1	7.5	1.643
2	5	1.265
3	4.333	1.033
4	5.167	1.472
5	6.167	2.041

- b. ANOVA table:

Source	d.f.	SS	MS	F
Treatment	4	36.467	9.117	3.90
Error	25	58.50	2.34	
Total	29	94.967		

- c. $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = 0$

H_A : At least two treatments have different population means

$$F_{4,25} = \frac{9.117}{2.34} = 3.896, \text{ with } P = .0136$$

At $\alpha = .05$, we reject H_0 and conclude that at least two treatments have different population means.

- d. Estimates of true effects $(\mu_i - \mu)$, where μ is the overall mean:

Treatment	
<i>i</i>	$(\bar{Y}_i - \bar{Y})$
1	1.8667
2	-0.6333
3	-1.3000
4	-0.4667
5	0.5333
Total $\sum_{i=1}^5 (\bar{Y}_i - \bar{Y})$	0.000

- e. We define X_i such that

$X_i = 1$, for treatment i ; 0 otherwise

where $i = 1, 2, 3, 4$. Then the appropriate regression model is

$$Y = \beta_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + E$$

where the regression coefficients are as follows:

$$\beta_0 = \mu_5$$

$$\alpha_1 = \mu_2 - \mu_5, \alpha_2 = \mu_2 - \mu_5, \alpha_3 = \mu_3 - \mu_5, \alpha_4 = \mu_4 - \mu_5$$

$X_i = 1$ for treatment i ($i = 1, 2, 3, 6$); $X_5 = -1$ for treatment 5; $X_i = 0$ otherwise.

The regression coefficients are

$$\beta_0 = \mu, \alpha_1 = \mu_1 - \mu, \alpha_2 = \mu_2 - \mu, \alpha_3 = \mu_3 - \mu, \alpha_4 = \mu_4 - \mu, \text{ and also}$$

$$-(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) = \mu_5 - \mu$$

- f. Hand-calculated 95% confidence intervals for $(\mu_1 - \mu_3)$ and $(\mu_1 - \mu_2)$ are as follows:

Comparison	Scheffé	Tukey–Kramer	Bonferroni
$(\mu_1 - \mu_3)$	(0.238, 6.096)	(0.57, 5.764)	(0.34, 5.994)
$(\mu_1 - \mu_2)$	(-0.429, 5.429)	(-0.097, 5.097)	(-0.327, 5.327)

Conclusions: Treatments 1 and 3 significantly differ; treatments 1 and 2 do not significantly differ. We conclude that all other remaining comparisons, which involve smaller (in absolute value) pairwise sample mean differences, are not significant using $\alpha = .05$. Of the three methods, Scheffé's method gives the widest interval, the Bonferroni method gives the next widest, and the Tukey–Kramer method gives the narrowest interval.

Source	d.f.	SS	MS	F
Parties	3	5,625	1,875.00	37.31
Error	16	804	50.25	
Total	19	6,429		

b. $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_A : At least two parties have different mean authoritarianism scores

$F_{(3, 16)} = 37.31$, with $P < .001$. At $\alpha = .05$, we reject H_0 and conclude that the authoritarianism scores of the members of different political parties significantly differ.

c. Regression model:

$$Y = \beta_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + E \quad \text{where } X_i = 1 \text{ if party } \#i, 0 \text{ otherwise}; \\ i = 1, 2, 3$$

d. The Tukey–Kramer method confidence intervals for different comparisons are then given as follows:

Comparison	Confidence Interval	Remark
P_3 vs. P_4 :	$(\bar{Y}_3 - \bar{Y}_4) \pm 12.93 = (32.07, 57.93)$	significant
P_3 vs. P_2 :	$(\bar{Y}_3 - \bar{Y}_2) \pm 12.93 = (2.07, 27.93)$	significant
P_3 vs. P_1 :	$(\bar{Y}_3 - \bar{Y}_1) \pm 12.93 = (-2.93, 22.93)$	not significant, since 0 in interval
P_1 vs. P_4 :	$(\bar{Y}_1 - \bar{Y}_4) \pm 12.93 = (22.07, 47.93)$	significant
P_1 vs. P_2 :	$(\bar{Y}_1 - \bar{Y}_2) \pm 12.93 = (-7.93, 17.93)$	not significant
P_2 vs. P_4 :	$(\bar{Y}_2 - \bar{Y}_4) \pm 12.93 = (17.07, 42.93)$	significant

At $\alpha = .05$, all differences except for $(\bar{Y}_1 - \bar{Y}_2)$ and $(\bar{Y}_3 - \bar{Y}_1)$ are significant.

9. a. $H_0: \mu_1 = \mu_2 = \mu_3 \quad H_A$: At least two mean generation times differ
 $F_{(3, 20)} = 46.99$, with $P < .001$. At $\alpha = .05$, we reject H_0 and conclude that true mean generation times between strains significantly differ.

Source	d.f.	SS	MS	F
Parties	3	134,713.00	44,904.33	46.99
Error	20	19,113.244	955.662	
Total	23	153,826.244		

- c. Using the Tukey–Kramer method confidence intervals for pairwise comparisons, we conclude that

$$\mu_C > (\mu_D = \mu_A) > \mu_B$$

where μ_A , μ_B , μ_C , and μ_D are population means from Strains A, B, C, and D respectively.

Source	d.f.	SS	MS	F
Dosage	3	566.628	188.876	14.71
Error	44	564.96	12.84	
Total	47	1,131.588		

17. a. $Y_{ij} = \mu + \alpha_i + E_{ij}$, $i = 1, 2, 3$; $j = 1, \dots, 6$.
b. $MST = 4.667 \quad F = 5.53 \quad MSE = 0.844$

- c. $H_0: \mu_1 = \mu_2 = \mu_3$
 $H_A:$ There is a significant difference in attitude toward advertising by practice type
 $F_{2, 15} = 5.53$, with $P = .0159$. At $\alpha = .05$, we reject H_0 and conclude that there are significant differences in attitude toward advertising by practice type.

- d. GP vs. IM: $3.6667 - 2.00 \pm 1.4398$ (0.2269, 3.1065) significant
GP vs. FP: $3.6667 - 2.333 \pm 1.4398$ (-0.1061, 2.7735) not significant
FP vs. IM: $2.333 - 2.00 \pm 1.4398$ (-1.1068, 1.7728) not significant

- e. Same as Problem 17(a).

Source	d.f.	SS	MS	F
Parties	2	26.333	13.167	20.43
Error	15	9.667	0.6444	
Total	17	35.999		

- g. $H_0: \mu_1 = \mu_2 = \mu_3$ $H_A:$ At least two means differ by practice type
 $F_{2, 15} = 20.43$, $P = .0001$. At $\alpha = .05$, we reject H_0 and conclude that there is a significant difference in influence on prescription writing habits by practice type.
- h. GP vs. IM: $4.3333 - 1.5000 \pm 1.2578$ (1.5755, 4.0911) significant
GP vs. FP: $4.3333 - 2.1667 \pm 1.2578$ (0.9088, 3.4244) significant
FP vs. IM: $2.1667 - 1.500 \pm 1.2578$ (-0.5911, 1.9245) not significant

21. a. $Y_{ij} = \mu + \alpha_i + E_{ij}$, $i = 1, \dots, 3; j = 1, \dots, n_i$. Clear zone sizes are fixed-effect factors.

Source	d.f.	SS	MS	F	P-value
Model	2	14.704	7.352	5.462	
Error	48	64.592	1.346		
Total	50	79.296			

- c. $H_0: \mu_1 = \mu_2 = \mu_3$ $H_A:$ At least two mean five-year changes differ by clear zone
 $F_{2, 48} = 5.462$, with $P = .0073$. At $\alpha = .05$, we reject H_0 and conclude that mean five-year changes significantly differ by clear zone size.

- d. μ_1 vs. μ_3 : (0.3368, 2.2188) significant
 μ_1 vs. μ_2 : (-0.2562, 1.6603) not significant
 μ_2 vs. μ_3 : (-1.6603, 0.2562) not significant

The mean five-year refractive changes significantly differ between 3.0 mm and 4.0 mm.

Chapter 18

1. a. The rats are the blocks, and the three chemicals are the treatments.
- b. $SSE = 25.0$ $MSE = 1.786$ Type I SS (chem) = 25.0 MS (chem) = 12.5
- c. $H_0: \mu_1 = \mu_2 = \mu_3$ $H_A:$ The mean irritation scores differ by chemical type
 $F_{2, 14} = 7.00$ with $P = .0078$. At $\alpha = .05$, we reject H_0 and conclude the toxic effects of the three chemicals significantly differ.
- d. 98% CI on $\mu_1 - \mu_{1j}$: (-3.0, 0.5) e. $R^2 = 0.635$
- f. Fixed-effects ANOVA model: $Y_{ij} = \mu + \tau_i + \beta_j + E_{ij}$, $i = 1, 2, 3; j = 1, 2, \dots, 8$ where

Y_{ij} = observation on the j th rat for the i th chemical effect

μ = overall mean

τ_i = i th chemical effect

β_j = j th rat effect

E_{ij} = error due to observation on the j th rat for the i th chemical effect (E_{ij} 's are independent and assumed to be normally distributed).

$$\text{Regression model: } Y = \beta_0 + \alpha_1 X_1 + \alpha_2 X_2 + \sum_{j=1}^7 \beta_j Z_j + E$$

$$\text{where } X_1 = \begin{cases} 1 & \text{if chemical I} \\ 0 & \text{if chemical II}, \\ -1 & \text{if chemical III} \end{cases} \quad X_2 = \begin{cases} 1 & \text{if chemical I} \\ 0 & \text{if chemical II}, \\ -1 & \text{if chemical III} \end{cases} \quad Z_j = \begin{cases} -1 & \text{if rat 8} \\ 1 & \text{if rat } (j = 1, 2, \dots, 7) \\ 0 & \text{otherwise} \end{cases}$$

- g.** Assumptions underlying the model: (i) additivity of the model (no interaction), (ii) homogeneity of variance, (iii) normality of the errors, (iv) independence of the errors.
5. $H_0: \mu_1 = \mu_2 = \mu_3 \quad H_A:$ At least two mean ESP scores differ by person.
 $F_{2,8} = 15.12$ with $P = .0019$. At $\alpha = .05$, we reject H_0 and conclude that there are significant differences in ESP ability by person. Note that, since the F test for blocking on days is not significant, one may conclude that blocking on days is not necessary.
9. **a.–f.**
- | Source | d.f. | SS | MS | F |
|------------|------|-----|----|------|
| Treatments | 4 | 160 | 40 | 5.00 |
| Blocks | 5 | 240 | 48 | 6.00 |
| Error | 20 | 160 | 8 | |
- g.** Test of treatments: $F_{4,20} = 5.00$ with $.005 < P < .01$. At $\alpha = .05$, we reject H_0 and conclude that there is a significant main effect of treatments.
Test of blocks: $F_{5,20} = 6.00$ with $.001 < P < .005$. At $\alpha = .05$, we reject H_0 and conclude that there is a significant main effect of blocks.

Chapter 19

1. **a.** The two factors are levorphanol and epinephrine.
b. Both factors should be considered fixed.
c. Rearrangement of the data into a two-way ANOVA layout:

Levels of Epinephrine

		Absence	Presence
		—	+
Levels of Levorphanol	Absence	(Control) 1.90, 1.80, 1.54, 4.10, 1.89	(Epinephrine only) 5.33, 4.85, 5.26, 4.92, 6.07
	Presence	(Levorphanol only) 0.82, 3.36, 1.64, 1.74, 1.21	(Levorphanol and Epinephrine) 3.08, 1.42, 4.54, 1.25, 2.57

d. Table of sample means:

		Epinephrine		
		-	+	Total (Row Mean)
Levorphanol	-	2.25	5.28	3.77
	+	1.75	2.57	2.16
Total (Col. Mean)		2.00	3.93	2.96

The presence of levorphanol appears to reduce stress, whereas the presence of epinephrine appears to increase stress. In the presence of epinephrine, levorphanol reduces stress by an average of 2.71 units, whereas in the absence of epinephrine, levorphanol reduces stress by only 0.50 units, suggesting the possibility of an interaction.

e. Source	d.f.	SS	MS	F
Levorphanol	1	12.832	12.932	12.60
Epinephrine	1	18.586	18.586	18.25
Interaction	1	6.161	6.161	6.05
Error	16	16.298	1.019	
Total	19	53.877		

f. Main effect for levorphanol

H_0 : Levorphanol has no significant main effect on stress

H_A : Levorphanol has a significant main effect on stress

$$F_{1, 16} = 12.60, P = .0027$$

At $\alpha = .05$, we reject H_0 and conclude that levorphanol has a significant main effect on stress.

Main effect for epinephrine

H_0 : Epinephrine has no significant main effect on stress

H_A : Epinephrine has a significant main effect on stress

$$F_{1, 16} = 18.25, P = .0066$$

At $\alpha = .05$, we reject H_0 and conclude that epinephrine has a significant main effect on stress.

Interaction

H_0 : There is no significant interaction between levorphanol and epinephrine on stress

H_A : There is a significant interaction between levorphanol and epinephrine on stress

$$F_{1, 16} = 6.05, P = .0267$$

At $\alpha = .05$, we reject H_0 and conclude that there is significant interaction between levorphanol and epinephrine on stress.

- 5. a.** Yes. There is an apparent same-direction interaction, which is reflected in the fact that the difference in mean waiting times between suburban and rural court locations is larger for State 1 than for State 2.

b. **Main effect of State:** $F_{1,594} = 217.45, P < .001$

Main effect of Court Location: $F_{2,594} = 184.86, P < .001$

Interaction effect: $F_{2,594} = 10.96, P < .001$

All effects are highly statistically significant.

c. Regression model:

$$Y = \beta_0 + \beta_1 S + \beta_2 C_1 + \beta_3 C_2 + \beta_4 SC_1 + \beta_5 SC_2 + E$$

where

$S = 1$ if State 1, -1 if State 2 $C_1 = -1$ if Urban, 1 if Rural, 0 if Suburban

$C_2 = -1$ if Urban, 1 if Suburban, 0 if Rural

d. We might consider a model of the form

$$Y = \beta'_0 + \beta'_1 S + \beta'_2 C + \beta'_3 SC + E$$

where S is as defined in part (c) and where C is a variable taking on three values (i.e., 0, 1, and 2) that increase directly with the degree of urbanization. One difficulty here is how to determine the appropriate values for C .

9. a. “Species”: fixed factor; “Locations”: random factor, unless only the four locations chosen are of interest.

b. Source	F (Mixed)	P-value	F (Both fixed)	P-value
Species	9.109 (d.f. = 2, 6)	.01 < $P < .025$	9.455 (d.f. = 2, 48)	$P < .001$
Locations	0.811 (d.f. = 3, 6)	$P > .25$	0.841 (d.f. = 3, 48)	$P > .25$
Interaction	1.038 (d.f. = 6, 48)	$P > .25$	1.038 (d.f. = 6, 48)	$P > .25$

13. a. Source	d.f.	SS	F	P
AIRTEMP	3	0.03	0.32	0.813
DOSE	2	0.06	1.04	0.370
A*D	6	0.05	0.29	0.936
Error	24	0.70		
TOTAL	35			

- b. From the ANOVA table, we see that the main effects are not significant, and neither is the interaction term.

- c. An appropriate multiple regression model is

$$Y = \mu + \sum_{i=2}^4 \alpha_i X_i + \sum_{j=2}^3 \beta_j Z_j + \sum_{i=2}^4 \sum_{j=2}^3 X_i Z_j \gamma_{ij} + E$$

in which

$$X_i \begin{cases} -1 & \text{for AIRTEMP} = 21 \\ 1 & \text{for level } i \text{ of AIRTEMP, } i = 2, 3, 4 \\ 0 & \text{otherwise} \end{cases}$$

$$Z_j \begin{cases} -1 & \text{for DOSE} = 21 \\ 1 & \text{for level } j \text{ of DOSE, } j = 2, 3 \\ 0 & \text{otherwise} \end{cases}$$

$$\alpha_i = \mu_{i\cdot} - \mu_{..}, i = 2, 3, 4$$

$$\beta_j = \mu_{j\cdot} - \mu_{..}, j = 2, 3$$

$$\gamma_{ij} = \mu_{ij} - \mu_{i\cdot} - \mu_{j\cdot} + \mu_{..}, i = 2, 3, 4; j = 2, 3$$

AIRTEMP = 21 and DOSE = 0 correspond to control levels and hence are of least interest. The larger AIRTEMPS and DOSES are of primary interest, so they are directly parameterized in the model.

- d. A natural polynomial model is

$$Y = \beta_0 + \beta_1(\text{AIRTEMP}) + \beta_2(\text{DOSE}) + \beta_3(\text{AIRTEMP} * \text{DOSE}) + E.$$

Note that higher-order terms can be added to the model if deemed reasonable.

17. a. The ANOVA model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk}, i = 1, 2; j = 1, 2; k = 1, \dots, 6$$

α_i = effect for the i th school type

β_j = effect for the j th reputation category

γ_{ij} = interaction effect for the i th school type and j th reputation category

E_{ijk} = error term for the observation ijk

The factors are fixed.

- b. d.f. = 3 $F_{3,20} = 2.572$, with $.05 < P < .10$

- c. **Main effect of school type:**

H_0 : There is no significant main effect of school type on starting salary

H_A : There is a significant main effect of school type on starting salary

$F_{1,20} = 0.52$, $P = .4786$. At $\alpha = .05$, we do not reject H_0 , and we conclude that school type does not have a significant main effect on starting salary.

Main effect of reputation rank:

H_0 : There is no significant main effect of reputation rank on starting salary

H_A : There is a significant main effect of reputation rank on starting salary

$F_{1,20} = 6.21$, $P = .0216$. At $\alpha = .05$, we reject H_0 , and we conclude that reputation rank has a significant main effect on starting salary.

Test of interaction:

H_0 : There is no significant interaction between school type and reputation rank with respect to starting salary

H_A : There is significant interaction between school type and reputation rank with respect to starting salary

$F_{1,20} = 0.99$, $P = .3323$. At $\alpha = .05$, we do not reject H_0 , and we conclude that there is no significant interaction between school type and reputation rank.

Chapter 20

1. a. Table of sample means:

Traditional Rank

		HI	MED	LO	Total
		Modern	Med	Mod	
Rank	HI	135	147.5	165	152.5
	MED	145	155	163	155.91
	LO	161.67	143.33	121.67	142.22
Total		147.22	148.75	154.23	

The preceding table indicates that the sample mean blood pressure for males with low modern rank is lower than for other modern rank categories. It also illustrates that the sample mean blood pressure for males with low traditional rank is higher than the other traditional rank categories. Finally, the table hints at an interaction effect: for persons with high modern rank, mean blood pressure increases with decreasing traditional rank, whereas mean blood pressure for persons with low modern rank decreases with decreasing traditional rank. In other words, this interaction suggests that persons with incongruous cultural roles (i.e., HI-LO, LO-HI) tend to have higher blood pressures than persons with more congruent cultural roles.

b. Regression model:

$$Y = \beta_0 + \alpha_1 X_1 + \alpha_2 X_2 + \beta_1 Z_1 + \beta_2 Z_2 + \gamma_{11} X_1 Z_1 + \gamma_{12} X_1 Z_2 + \gamma_{21} X_2 Z_1 + \gamma_{22} X_2 Z_2 + E$$

where

$$X_i = \begin{cases} -1 & \text{if LO modern rank} \\ 1 & \text{if modern rank } i \text{ (} i = 1 \text{ for HI, } i = 2 \text{ for MED)} \\ 0 & \text{otherwise} \end{cases}$$

and

$$Z_i = \begin{cases} -1 & \text{if LO traditional rank} \\ 1 & \text{if traditional rank } i \text{ (} i = 1 \text{ for HI, } i = 2 \text{ for MED)} \\ 0 & \text{otherwise} \end{cases}$$

c. Modern main effect:

- (i) $F(X_1, X_2)_{2,27} = 2.076, .10 < P < .25$, which is not significant
- (ii) $F(X_1, X_2 | Z_1, Z_2)_{2,25} = 1.768, .10 < P < .25$, which is not significant

Traditional main effect:

- (i) $F(Z_1, Z_2)_{2,27} = 0.578, P > .25$, which is not significant
- (ii) $F(Z_1, Z_2 | X_1, X_2)_{2,25} = 0.397, P > .25$, which is not significant

Interaction:

$F(X_1 Z_1, X_1 Z_2, X_2 Z_1, X_2 Z_2 | X_1, X_2, Z_1, Z_2)_{4,21} = 15.069, P < .001$, which is highly significant

d. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + E$, where

$$X_1 = \begin{cases} 0 & \text{if LO modern rank} \\ 1 & \text{if MED modern rank} \\ 2 & \text{if HI modern rank} \end{cases} \quad X_1 = \begin{cases} 0 & \text{if LO traditional rank} \\ 1 & \text{if MED traditional rank} \\ 2 & \text{if HI traditional rank} \end{cases}$$

The difficulty arises with regard to assigning numerical values to the categories of each factor. The coding scheme for X_1 and X_2 given here assumes that the categories are “equally spaced,” which may not really be the case.

5. a. Table of means:

		Social Class			
		Lo	Med	Hi	Row means
Number of times victimized	0	14.571	12.00	13.50	13.5
	1	9.00	12.25	7.40	9.4
	2+	8.00	2.33	5.75	5.8
Col means	Total	11.3	9.7	8.8	

The preceding table suggests a downward trend in confidence with increasing number of victimizations and a slight downward trend with increasing social class status score.

- b. From the ANOVA table given in the question, we obtain

$$F(\text{Victim})_{2,31} = 8.81, P < .001, \text{ which is significant}$$

$$F(\text{SCLS})_{2,31} = 0.501, P > .25, \text{ which is not significant}$$

$$F(\text{Interaction})_{4,31} = 1.21, P > .25, \text{ which is not significant}$$

- c. We can use the following regression model:

$$Y = \beta_0 + \alpha_1 X_1 + \alpha_2 X_2 + \beta_1 Z_1 + \beta_2 Z_2 + \gamma_{11} X_1 Z_1 + \gamma_{12} X_1 Z_2 + \gamma_{21} X_2 Z_1 + \gamma_{22} X_2 Z_2 + E$$

where

$$X_1 = \begin{cases} -1 & \text{if no. of times victimized} = 0 \\ 1 & \text{if no. of times victimized} = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} -1 & \text{if no. of times victimized} = 0 \\ 1 & \text{if no. of times victimized} = 2+ \\ 0 & \text{otherwise} \end{cases}$$

$$Z_j = \begin{cases} -1 & \text{if social class status} = \text{Lo} \\ 1 & \text{if social class status} = j (j = 1 \text{ for Med}, j = 2 \text{ for Hi}) \\ 0 & \text{otherwise} \end{cases}$$

We then obtain the following F values from the regression results given in the question:

Main effect of VICTIM:

$$F(X_1, X_2)_{2,37} = 9.03, P < .001, \text{ which is significant}$$

$$F(X_1, X_2 | Z_1, Z_2)_{2,35} = 8.61, P < .001, \text{ which is significant}$$

Main effect of SCLS:

$$F(Z_1, Z_2)_{2,37} = 0.695, P > .25, \text{ which is not significant}$$

$$F(Z_1, Z_2 | X_1, X_2)_{2,35} = 0.707, P > .25, \text{ which is not significant}$$

Interaction:

$$F(X_1 Z_1, X_1 Z_2, X_2 Z_1, X_2 Z_2 | X_1, X_2, Z_1, Z_2)_{4,31} = 1.10,$$

$$P > .25, \text{ which is not significant}$$

- d. The error term may have a nonconstant variance.

9. a. A tabulation of DOSAGE * SEX sample sizes reveals that all combinations of these factors have 10 subjects except for 9 in the DOSAGE = 0, SEX = Male combination. We could call this a nearly orthogonal design, since we would only need one more observation in the cell with 9 to make a perfectly orthogonal design (equal cell sizes).

b.

Source	TYPE I		TYPE III	
	F	P-value	F	P-value
DOSAGE	2.65	.05 < P < .10	2.64	.05 < P < .10
SEX	0.22	P > .25	0.20	P > .25
DOSAGE*SEX	0.14	P > .25	0.14	P > .25

- c. At $\alpha = .05$, no effects are significant.
d. Since no effects are significant, no multiple-comparison tests are justified.

13. a. $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk}$, $i = 1, 2, 3; j = 1, 2, 3; k = 1, \dots, n_{ij}$
 $n_{11} = 3, n_{12} = 6, n_{13} = 11, n_{21} = 4, n_{22} = 5, n_{23} = 6, n_{31} = 4, n_{32} = 4, n_{33} = 8$
 α_i = effect for the i th clear zone ($i = 1$: 3.0 mm, $i = 2$: 3.5 mm, $i = 3$: 4.0 mm)
 β_j = effect for the j th baseline curvature category
 γ_{ij} = interaction effect for the i th clear zone and j th curvature
 E_{ijk} = error term for the observation ijk
b. d.f. = 8 $F_{8, 42} = 3.595$ $.001 < P < .005$
c. Using the SAS output, at $\alpha = .10$, we would conclude that there is a significant interaction between CLRZONE and BASECURV. We would also conclude that CLRZONE and BASECURV have significant main effects on refractive error.

Chapter 21

1. a. Yes, they are identical if we assume that the linear regression model is fit to normally distributed data.
b. $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$
(Full model: $Y = \beta_0 + \beta_1 \text{DRUG} + \beta_2 \text{SBP} + \beta_3 \text{QUET} + E$ where DRUG = 1 if drug, 0 if placebo)

Test statistic: $Z = \frac{\hat{\beta}_1}{\sqrt{\text{Var}_{\hat{\beta}_1}}}$, which is approximately standard normal under H_0 .

Critical value: If $Z > 1.96$, then reject H_0 at $\alpha = .05$.

- c. No; however, as N increases, they approach one another.
d. $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$
Test statistic: $-2 \log L(\text{reduced}) - (-2 \log L(\text{full})) = \chi^2$ with 1 d.f. where
Full model equals: $Y = \beta_0 + \beta_1 \text{DRUG} + \beta_2 \text{SBP} + \beta_3 \text{QUET} + E$
Reduced model equals: $Y = \beta_0 + \beta_2 \text{SBP} + \beta_3 \text{QUET} + E$
Under H_0 , the test statistic is approximately chi-square with 1 d.f.
Critical value: if $\chi^2 > 3.841$, then reject H_0 at $\alpha = .05$.

- e. Yes, since $\chi^2 = Z^2$ when Z is normally distributed.
f. 95% confidence interval for β_1 : $\hat{\beta}_1 \pm 1.96 * \sqrt{\text{Var}_{\hat{\beta}_1}}$

Chapter 22

1. a. $\hat{OR} = 2.925 \Rightarrow \hat{\beta}_1 = \ln(2.925) = 1.073$
- b. $\text{logit}[\Pr(Y=1)] = -2.8 + 0.706X_1 + 0.0004X_2 + 0.0006X_3$
- c. $\Pr(Y=1) = 0.112$
- d. $\hat{OR}_{20\text{vs.}21} = 0.999$
The odds for hypertension for a 20-year-old smoker are essentially equal to those for a 21-year-old smoker.
- e. 95% CI for $\hat{OR}_{20\text{yr old smoker vs. 21 yr old smoker}}$: (0.9982, 0.9998)
- f. $H_0: \beta_3 = 0 \quad H_A: \beta_3 \neq 0$
Test statistic: $-2 \log L(\text{reduced}) - (-2 \log L(\text{full})) = \chi^2$ with 1 d.f.
 $\chi^2 = 308.00 - 303.84 = 4.16$, with $.025 < P < .05$.
At $\alpha = .05$, we reject H_0 and conclude that there is significant interaction between age and smoking.

Chapter 23

1. a. $\ln \frac{P(D \geq g|X)}{P(D < g|X)} = \alpha_g + \beta_1 \text{ALC1} + \beta_2 \text{ALC2} + \beta_3 \text{SMK1} + \beta_4 \text{SMK2}$
 $+ \beta_5 \text{AGE} + \beta_6 \text{PA} + \beta_7 \text{CL1} + \beta_8 \text{CL2} + \beta_9 \text{CL3}$
where $g = 1$ (moderate hypertensive), $g = 2$ (severe hypertensive)
 $\text{ALC1} = 1$ if light drinker, 0 otherwise
 $\text{ALC2} = 1$ if heavy drinker, 0 otherwise
 $\text{SMK1} = 1$ if <2 packs per day, 0 otherwise
 $\text{SMK2} = 1$ if 2+ packs per day, 0 otherwise
 $\text{CL1} = 1$ if Clinic 1, 0 otherwise
 $\text{CL2} = 1$ if Clinic 2, 0 otherwise
 $\text{CL3} = 1$ if Clinic 3, 0 otherwise
- b. $\hat{OR}_1 = \hat{OR}_3$ and $\hat{OR}_2 = \hat{OR}_4$
- c. The Score test allows you to evaluate the proportional odds (p.o.) assumption collectively for all predictor variables in a p.o. model. H_0 : The p.o. assumption is satisfied. If the Score test does not reject H_0 , then you can conclude that the p.o. model is appropriate. If the Score test rejects H_0 , then you have to use some other approach for carrying out ordinal logistic regression, or you can use polytomous logistic regression instead.
- d. $\exp[\alpha_1 + \beta_2 + \beta_4 + \beta_5 + \beta_6]$
- e. $\text{OR}_{\text{BP} \geq 1} = \exp[-\beta_1 + \beta_2 + \beta_4]$
- f. $\text{OR}_{\text{BP}=0} = \text{OR}_{\text{BP}<1} = 1/\text{OR}_{\text{BP} \geq 1} = \exp[\beta_1 - \beta_2 - \beta_4]$
- g. $E_i E_j$ variables: $\text{ALC1} \times \text{SMK1}$, $\text{ALC1} \times \text{SMK2}$, $\text{ALC2} \times \text{SMK1}$,
 $\text{ALC1} \times \text{SMK2}$
 $E_i V_j$ variables: $\text{ALC1} \times \text{AGE}$, $\text{ALC2} \times \text{AGE}$, $\text{ALC1} \times \text{PA}$, $\text{ALC2} \times \text{PA}$,
 $\text{ALC1} \times \text{CL1}$, $\text{ALC2} \times \text{CL1}$, $\text{ALC1} \times \text{CL2}$, $\text{ALC2} \times \text{CL2}$, $\text{ALC1} \times \text{CL3}$,
 $\text{ALC2} \times \text{CL3}$,
 $\text{SMK1} \times \text{AGE}$, $\text{SMK2} \times \text{AGE}$, $\text{SMK1} \times \text{PA}$, $\text{SMK2} \times \text{PA}$,
 $\text{SMK1} \times \text{CL1}$, $\text{SMK2} \times \text{CL1}$, $\text{SMK1} \times \text{CL2}$, $\text{SMK2} \times \text{CL2}$, $\text{SMK1} \times \text{CL3}$,
 $\text{SMK2} \times \text{CL3}$

- h.** H_0 : The coefficients of the 20 $E_i V_j$ variables listed above are all equal to 0
 LR statistic $= (-2 \ln L_R) - (-2 \ln L_F) \sim \chi^2$ with 20 d.f. under H_0 ,
 where R denotes the reduced model that contains the following predictor variables:

ALC1, ALC2, SMK1, SMK2, AGE, PA, CL1, CL2, CL3, and
 $ALC1 \times SMK1, ALC1 \times SMK2, ALC2 \times SMK1, ALC1 \times SMK2$

i. For $g = 1, 2, \ln \frac{P(D \geq g|X)}{P(D < g|X)}$

$$\begin{aligned} &= \alpha_g + \beta_1 ALC1 + \beta_2 ALC2 + \beta_3 SMK1 + \beta_4 SMK2 + \beta_5 AGE + \beta_6 PA \\ &\quad + \beta_7 CL1 + \beta_8 CL2 + \beta_9 CL3 + \beta_{10}(ALC1 \times AGE) + \beta_{11}(ALC2 \times AGE) \\ &\quad + \beta_{12}(SMK1 \times PA) + \beta_{13}(SMK2 \times PA) + \beta_{14}(ALC1 \times SMK1) \\ &\quad + \beta_{15}(ALC1 \times SMK2) + \beta_{16}(ALC2 \times SMK1) + \beta_{17}(ALC2 \times SMK2) \end{aligned}$$

j. $OR_{BP \geq 1} = \exp[-\beta_1 + \beta_2 + \beta_4 + \beta_{17} - \beta_{10}AGE + \beta_{11}AGE + \beta_{13}PA]$

k. 30-year-old: AGE = 0; no regular physical activity: PA = 0

95% CI: $\hat{L} \pm 1.96 \sqrt{\text{Var}(\hat{L})}$, where

$$\begin{aligned} \text{Var } \hat{L} &= \text{Var}\hat{\beta}_1 + \text{Var}\hat{\beta}_2 + \text{Var}\hat{\beta}_4 + \text{Var}\hat{\beta}_{17} - 2 \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) - 2 \text{Cov}(\hat{\beta}_1, \hat{\beta}_4) \\ &\quad - 2 \text{Cov}(\hat{\beta}_1, \hat{\beta}_{17}) + 2 \text{Cov}(\hat{\beta}_2, \hat{\beta}_4) + 2 \text{Cov}(\hat{\beta}_2, \hat{\beta}_{17}) + 2 \text{Cov}(\hat{\beta}_4, \hat{\beta}_{17}) \end{aligned}$$

l. $OR_{BP \geq 1} = \exp[-\beta_1 + \beta_2 + \beta_4 - \beta_{10}AGE + \beta_{11}AGE + \beta_{13}PA]$

2. a. $\ln \Pr(\text{DIS} \geq g|\mathbf{X})/\Pr(\text{DIS} < g|\mathbf{X}) = \alpha_g + \beta_1 FAB + \beta_2 CHR1 + \beta_3 CHR2$
 $+ \beta_4 SEV1 + \beta_5 SEV2 + \gamma_1 SL + \gamma_2 SEX + \gamma_3 AGE$

where $g = 1, 2, 3$ and $\alpha_1 > \alpha_2 > \alpha_3$

and $CHR1 = 1$ if $CHR = 2$ and 0 otherwise

$CHR2 = 1$ if $CHR = 3$ and 0 otherwise

$SEV1 = 1$ if $SEV = 2$ and 0 otherwise

$SEV2 = 1$ if $SEV = 3$ and 0 otherwise

b. i. $ODDS(\mathbf{X}) = \exp[\alpha_2 + \beta_1 + \beta_3 + \beta_5 + \gamma_1 + \gamma_2 + 40\gamma_3]$,

where

ii. $OR(\mathbf{X}^*, \mathbf{X}) = \exp[\beta_1 + \beta_3 + \beta_5]$

d. $H_0: \delta_{F1g} = \delta_{F2g} = \delta_{F3g} = \delta_{C1g} = \delta_{C2g} = \delta_{C3g} = \delta_{S1g} = \delta_{S2g} = \delta_{S3g} = 0$,
 for $g = 1, 2, 3$

LR statistic $= (-2 \ln L_R) - (-2 \ln L_F) \sim \chi^2$ with $9 \times 3 = 27$ d.f. under H_0

e. $OR_{\mathbf{X}^*, \mathbf{x}}(\text{DIS} = 3 \text{ vs } \text{DIS} = 1)$

$$= OR_{\mathbf{X}^*, \mathbf{x}}(\text{DIS} = 3 \text{ vs } \text{DIS} = 0) OR_{\mathbf{X}^*, \mathbf{x}}(\text{DIS} = 1 \text{ vs } \text{DIS} = 0)$$

$$= \exp [\beta_{13} + 2\beta_{23} + 2\beta_{33} + 2\delta_{C13}SL + 8\delta_{CS33}] \neq \exp [\beta_{11} + 2\beta_{21} + 2\beta_{31} + 2\delta_{C11}SL + 8\delta_{CS31}]$$

$$= \exp [(\beta_{13} - \beta_{11}) + 2(\beta_{23} - \beta_{21}) + 2(\beta_{33} - \beta_{31}) + 2(\delta_{C13} - \delta_{C11})SL + 8(\delta_{CS33} - \delta_{CS31})]$$

Chapter 24

1. a. $n = 30$ (6 age-sex groups \times 5 years)

b. Here, $i = 5$ and $k = 1992 - 1990 = 2$, so $E(Y_{52}) = l_{52}\lambda_{52} = l_{52}e^{(\alpha_5 + 2\beta)}$.

c. Log rate changes linearly with time. In particular, for the i th group,

$$\ln \lambda_{ik} = \alpha_i + \beta k$$

so α_i is the intercept and β is the slope of the straight line relating the response $\ln \lambda_{ik}$ to the time variable $k = [\text{year}] - 1960$.

- d.** Model (1) assumes no interaction between age-sex group and time in the sense that the change in log rate over time (as measured by β) does not depend on i . Since $\ln \lambda_{ik} = \alpha_i + \beta k$, a graph of $\ln \lambda_{ik}$ versus k for each i would plot as a series of parallel straight lines—that is, lines all with the same slope (β) but possibly different intercepts (the α_i 's). A lack of parallelism would reflect interaction between age-sex groups and time because the change in log rate over time would differ for different age-sex groups.
- e.** $\ln \text{IDR}_{ik} = \ln \lambda_{ik} - \ln \lambda_{10} = (\alpha_i + \beta k) - (\alpha_1 + \beta * 0) = (\alpha_i - \alpha_1) + \beta k$, so that

$$\text{IDR}_{ik} = e^{\alpha_i - \alpha_1} e^{\beta k}$$

Note that this is a function of both age-sex group (i) and time (k).

- f.** An appropriate model is

$$E(Y_{ik}) = i_{ik} \lambda_{ik}$$

where

$$\ln \lambda_{ik} = \sum_{i=1}^6 \alpha_i A_i + \beta k + \sum_{i=1}^5 \gamma_i (A_i k)$$

For age-sex group i , then,

$$\begin{aligned} \ln \lambda_{ik} &= \alpha_i + \beta k + \gamma_i k \\ &= \alpha_i + (\beta + \gamma_i) k \\ &= \alpha_i + \delta_i k \end{aligned}$$

where $\delta_i = \beta + \gamma_i$. Hence, the slope for group i (namely, δ_i) is now a function of i . Only when all six δ_i 's (or, equivalently, all six γ_i 's) are equal will the straight lines be parallel.

- g.** Yes, since $D(\hat{\beta})_{(1)} - D(\hat{\beta})_{(3)} = 300 - 175 = 125$, which is highly significant when compared to appropriate upper-tail χ^2 -values with $29 - 24 = 5$ d.f.
- h.** Yes, since $D(\hat{\beta})_{(3)} - D(\hat{\beta})_{(4)} = 175 - 60 = 115$, which is highly significant when compared to appropriate upper-tail χ^2 -values with $24 - 23 = 1$ d.f.
- i.** No, since $D(\hat{\beta})_{(4)} - D(\hat{\beta})_{(5)} = 60 - 59 = 1$, which is clearly not significant when compared to appropriate upper-tail χ^2 -values with $23 - 22 = 1$ d.f.
- j.** Yes, since $D(\hat{\beta})_{(4)} - D(\hat{\beta})_{(7)} = 60 - 20 = 40$, which is highly significant when compared to appropriate upper-tail χ^2 -values with $23 - 18 = 5$ d.f.
- k.** H_0 is rejected, since $D(\hat{\beta})_{(4)} - D(\hat{\beta})_{(6)} = 60 - 25 = 35$, which is highly significant when compared to appropriate upper-tail χ^2 -values with $23 - 22 = 1$ d.f.
- l.** Only models (6) and (7) are candidates to be the final model. Model (6) has a deviance of 25 based on 22 d.f., indicating a good fit to the data. Model (7) has a deviance of 20 based on 18 d.f., also indicating a good fit. All other candidate models have significant lack of fit. Note that $D(\hat{\beta})_{(6)} - D(\hat{\beta})_{(7)} = 25 - 20 = 5$, which is clearly not significant when compared to appropriate upper-tail χ^2 -values with $22 - 18 = 4$ d.f. Hence, model (6) certainly fits the data as well as

model (7), and it also characterizes very specifically the type of interaction present in the data (namely, that the group 1 slope differs from the slope common to the other five groups). Model (6) is our choice as the final model.

- m.** For model (6),

$$\text{pseudo } R^2 = \frac{300 - 25}{300} = 0.917$$

which is indicative of a good model.

- n.** $\hat{\beta} \pm 1.96 * \text{SE}_{\hat{\beta}} = 0.50 \pm 1.96(0.20) = (0.108, 0.892)$
- o.** The point estimate of δ_1 is $\hat{\delta}_1 = \hat{\beta} + \hat{\gamma}_1 = 0.50 - 3.00 = -2.50$. The variance of the estimator $\hat{\delta}_1$ is $\text{Var}(\hat{\delta}_1) = \text{Var}(\hat{\beta}) + \text{Var}(\hat{\gamma}_1) + 2 \text{Cov}(\hat{\beta}_1, \hat{\gamma}_1)$, which equals $\text{Var}(\hat{\delta}_1) = (0.20)^2 + (0.50)^2 + 2(-0.10) = 0.09$

Finally, an approximate 95% CI for δ_1 is

$$\hat{\delta}_1 \pm 1.96 \sqrt{\text{Var}(\hat{\delta}_1)} = -2.50 \pm 1.96 \sqrt{0.09} = (-3.088, -1.912)$$

These two confidence intervals suggest that log rate increases linearly ($\beta > 0$) for groups 2 through 6 but decreases ($\delta_1 < 0$) for group 1.

- 4. a.** $\text{RR} = \exp[0.8869 + (-0.2811) \ln T]$
 - i.** 15–24: $T = 1/7$, so $\widehat{\text{RR}}_{15-24} = \exp[0.8869 + (-0.2811) \ln(1/7)] = 4.20$
 - ii.** 45–54: $T = 1$, so $\widehat{\text{RR}}_{45-54} = \exp[0.8869 + (-0.2811)(\ln 1)] = 2.43$
- b.** The results in part (a) suggest that there is interaction of City with the variable $\ln T$, which is the natural log of Age Group centered around its midpoint.
- c.** $H_0: \beta = 0$, where β is the coefficient of the product term $E \times \ln T_i$ in model 5. Wald test: Chi-square statistic (1 d.f. under H_0) = 3.8386; two-tailed P -value = .0501; LR test: Chi-square statistic (1 d.f. under H_0) $C = \text{Dev}(\text{model 4}) - \text{Dev}(\text{model 5}) = 14.8 - 10.63 = 4.17$ (.025 < P < .05)
- d.** Yes, the results suggest that the product term is significant at the .05 level.
- e.** No. A test for fit uses the Dev statistic for one model only.
- f.** Model 4 is likely the best model choice, since the graph shown by Figure 24.1 shows excellent fit using the linear model involving $\ln T$ —that is, model 4.
- 5. b.** The person-time information for each subgroup would have to be the same, which is what is being assumed by model 2. Otherwise, different effects would result from model 1 and model 2.
- c.** The model is $\ln[\lambda] = \beta_0 + \beta_1 \text{AGE} + \beta_2 \text{SMOKE} + \beta_3 \text{AGE} \times \text{SMOKE}$, where λ is the rate for a given subgroup. The null hypothesis is $H_0: \beta_3 = 0$. The Wald statistic is given by $Z^2 = [0.0036/0.0100]^2 = .1264$. The two-tailed P -value for this test is .7222, which is > .05. We conclude from the Wald test that the interaction term in model 3 is not significant.
- d.** The deviance for model 1 is of borderline significance (.05 < P < .10), whereas the deviance for model 5 is not significant. This means that model 1 questionably fits the data, whereas model 5 fits the data very well. Thus, there should be a strong preference for model 5.
- e.** $\text{LR} = \text{Deviance model 4} - \text{Deviance model 5} = 10.3962 - 0.3925 = 10.0037$

The LR statistic has a chi-square distribution with 2 degrees of freedom under H_0 . The P -value is less than .01. The two age variables should not be dropped from model 5.

- g.** Model 6 is a saturated model.
 - h.** $\mathbf{X}^* = (75, \text{SMOKE}), X = (25, \text{SMOKE})$, so the rate ratio is $\exp[0.0098(75 - 25) + .5784(\text{SMOKE} - \text{SMOKE})] = 1.632$; 95% CI for $\exp[(50)\beta_1]$ is $\exp[0.49 \pm 1.96(0.0049)(50)] = (1.01, 2.64)$
 - i.** The best model is model 5. It fits the data well. Its only competitor of the models provided is model 1, which does not fit nearly as well. There is no significant interaction. Model 2 is out because it does not use an offset and, as a result, does not fit the data well.
- 6. a.** The sample size is 72.
- c.** Model 2: $\ln(E(C)) = \alpha + \beta'_1 \ln(DT) + \beta'_2 \ln(WTGP) + \gamma'_1(\text{GEN}) + \sum_{k=1}^2 \gamma'_{k+1} (\text{AGEGP}_k) + \delta'_1 \ln(DT) \times (\text{GEN}) + \delta'_2 \ln(WTGP) \times (\text{GEN}) + \sum_{k=1}^2 \delta'_k \ln(DT) \times (\text{AGEGP}_k) + \sum_{k=1}^2 \delta''_k \ln(WTGP) \times (\text{AGEGP}_k) + \ln(PT)$

- d.** SAS code:

```
model C = LDT LWTGP GEN AGEGP1 AGEGP2 LDTGEN
           LWTGPGEN SDTAGEGP1 LDTAGEGP2 LWTGPAGEGP1
           LWTGPAGEGP2/dist=poisson link=log offset=LPT;
```

where

$LDT = \ln(DT)$, $LWTGP = \ln(WTGP)$, $AGEGP_k = AGEGP_k$ for $k = 1, 2$
 $LDTGEN = \ln(DT) \times (\text{GEN})$, $LWTGPGEN = \ln(WTGP) \times (\text{GEN})$
 $LDTAGEGP_k = \ln(DT) \times (AGEGP_k)$, $LWTGPAGEGP_k = \ln(WTGP) \times (AGEGP_k)$ for $k = 1, 2$

- e.** $\mathbf{X}^* = (\ln 3, \ln 4, \text{GEN}, AGEGP1, AGEGP2)$, $\mathbf{X} = (\ln 1, \ln 1, \text{GEN}, AGEGP1, AGEGP2)$

Using the model defined in part (c):

$$RR = \exp [L] \quad \text{where}$$

$$L = \beta'_1(\ln 3) + \beta'_2(\ln 4) + \delta'_1(\ln 3) \times (\text{GEN}) + \delta'_2(\ln 4) \times (\text{GEN}) + \sum_{k=1}^2 \delta'_k(\ln 3) \times (AGEGP_k) + \sum_{k=1}^2 \delta''_k(\ln 4) \times (AGEGP_k)$$

- f.** 95% CI: $\exp[\hat{L} \pm 1.96\sqrt{\text{Var}(\hat{L})}]$, where L is given in the answer to part (e).
- g.** $LR = \text{Dev}(\text{Reduced Model}) - \text{Dev}(\text{Full Model}) \sim \chi^2_{(66-60)} = \chi^2_6$ under H_0 , where Reduced Model = Model 2 without six interaction terms and Full Model = Model 2
 H_0 : Coefficients of the six interaction terms in model 2 are all 0

- h.** No. model 2 is not the saturated model since model 2 contains 12 parameters rather than the 72 parameters required for a saturated model in this example.
- i.** No. model 1 is not the saturated model since model 1 contains 24 parameters, rather than the 72 parameters required for a saturated model.

Chapter 25

- 1. a.** Matrix **A** is unstructured.
- b.** Matrix **B** is independence.
- c.** Matrix **C** is autoregressive.
- d.** Matrix **D** is exchangeable.
- 2. a.** **Q** and **S** are heterogeneous whereas **P** and **R** are homogeneous covariance matrices.
- b.** **P:** autoregressive ($\rho = .5$); **Q:** exchangeable ($\rho = .46$); **R:** exchangeable ($\rho = .46$); **S:** autoregressive ($\rho = .5$)
- 3.** The answer will vary with the reader.
- 4. a.** Either of two ways to write the subject-specific scalar model:
 - i.** $Y = \beta_0 + \beta_1 R_1 + \beta_2 R_2 + \beta_3 R_3 + \beta_4 R_4 + E$,
where $R_g = 1$ if week $g = 1$, otherwise = 0, $g = 1, \dots, 4$
 - ii.** $Y_{ij} = \beta_0 + \beta_1 R_{ij1} + \beta_2 R_{ij2} + \beta_3 R_{ij3} + \beta_4 R_{ij4} + E_{ij}$, $i = 1, \dots, 40, j = 1, \dots, 5$
If the (scalar) dummy variables R_1, R_2, R_3 , and R_4 are defined so that the referent group is week 1, then $R_{ijg} = 1$ if $g = j - 1$ whereas $R_{ijg} = 0$ if $g \neq j - 1$.
- b.** The subject-specific matrix form of the model in which week 1 is used as the reference level is given by

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{E}_i, i = 1, \dots, 40$$

where \mathbf{Y}_i denotes the collection of 5 FEV1 measurements on the i th child, \mathbf{X}_i denotes the intercept variable (1.0, 1) together with the collection of dummy variable values for each of the five weeks for subject i , $\boldsymbol{\beta}$ denotes the parameter vector for the collection of β 's in this model, and \mathbf{E}_i denotes the collection of error terms on the i th subject. Assuming again that the referent group is week 5, then \mathbf{X}_i , $\boldsymbol{\beta}$, and \mathbf{E}_i can be written as follows:

$$\mathbf{X}_i = \begin{bmatrix} 1 & R_{i11} & R_{i12} & R_{i13} & R_{i14} \\ 1 & R_{i21} & R_{i22} & R_{i23} & R_{i24} \\ 1 & R_{i31} & R_{i32} & R_{i33} & R_{i34} \\ 1 & R_{i41} & R_{i42} & R_{i43} & R_{i44} \\ 1 & R_{i51} & R_{i52} & R_{i53} & R_{i54} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}, \quad \mathbf{E}_i = \begin{bmatrix} E_{i1} \\ E_{i2} \\ E_{i3} \\ E_{i4} \\ E_{i5} \end{bmatrix}$$

where \mathbf{Y}_i is (5×1) , \mathbf{X}_i is (5×5) , $\boldsymbol{\beta}$ is (5×1) , and \mathbf{E}_i is (5×1) .

- c.** $\hat{\beta}_0 = \overline{\text{FEV1}}_1, \hat{\beta}_1 = \overline{\text{FEV1}}_2 - \overline{\text{FEV1}}_1, \hat{\beta}_2 = \overline{\text{FEV1}}_3 - \overline{\text{FEV1}}_1,$
 $\hat{\beta}_3 = \overline{\text{FEV1}}_4 - \overline{\text{FEV1}}_1, \hat{\beta}_4 = \overline{\text{FEV1}}_5 - \overline{\text{FEV1}}_1$
- d.** $\beta'_0 = \mu_1, \beta'_1 = \mu_2 - \mu_1, \beta'_2 = \mu_3 - \mu_1, \beta'_3 = \mu_4 - \mu_1, \beta'_4 = \mu_5 - \mu_1$, so

$$\begin{aligned}\mu_1 - \frac{\mu_2 + \mu_3 + \mu_4 + \mu_5}{4} &= \beta'_0 - \frac{(\beta'_1 + \beta'_0) + (\beta'_2 + \beta'_0) + (\beta'_3 + \beta'_0) + (\beta'_4 + \beta'_0)}{4} \\ &= -\frac{\beta'_1 + \beta'_2 + \beta'_3 + \beta'_4}{4}\end{aligned}$$

Thus, the null hypothesis can alternatively be stated as H_0 :

$$-\frac{\beta'_1 + \beta'_2 + \beta'_3 + \beta'_4}{4} = 0.$$

- e. Yes. The coding of the dummy variables will not affect the overall test statistic, even though the estimates of individual regression coefficients will change with the coding.
- f. $\widehat{\text{FEV}}_{11} = \hat{\beta}'_0 = 9.8137$, $\widehat{\text{FEV}}_{12} = \hat{\beta}'_0 + \hat{\beta}'_1 = 9.8137 + (-2.9655) = 6.8482$, $\widehat{\text{FEV}}_{13} = \hat{\beta}'_0 + \hat{\beta}'_2 = 9.8137 + (-2.8120) = 7.0017$, $\widehat{\text{FEV}}_{14} = \hat{\beta}'_0 + \hat{\beta}'_3 = 9.8137 + (-2.8623) = 6.9514$, $\widehat{\text{FEV}}_{15} = \hat{\beta}'_0 + \hat{\beta}'_4 = 9.8137 + (-2.8153) = 6.9984$
- g. $\hat{\mu}_1 - \frac{\hat{\mu}_2 + \hat{\mu}_3 + \hat{\mu}_4 + \hat{\mu}_5}{4}$
 $= 9.8137 - \frac{6.8482 + 7.0017 + 6.9514 + 6.9984}{4} = 2.8637$
- h. The output does not provide the results for testing whether or not there is a significant difference among the mean FEV1 scores for all five weeks.
 - i. contrast "week" d1 1, d2 1, d3 1, d4 1.
 - ii. Yes, the results obtained from the contrast statement specified in the preceding question will be the same as the results previously obtained in Table 25.8.
 - i. i. exchangeable
 - ii. $\hat{\rho} = 0.5565$
 - iii. $\hat{\rho} = \frac{1.4936}{1.1901 + 1.4936} = 0.5565$
 - iv. No, the output uses a model-based estimate of the standard error, rather than an empirical or robust standard error; the latter corrects for possible misspecification of the correlation structure.
 - v. Yes, the estimated correlation structure should not depend on how the predictor variables are coded for the computer.
- 5. c. $Y_{ij} = \beta_0 + \beta_1(\text{trt})_{ij} + E_{ij}$, $i = 1, \dots, 15$, $j = 1, 2$, where $(\text{trt})_{ij} = 1$ if the j th treatment on the i th subject is the active treatment, $= 0$ if the j th treatment on the i th subject is the placebo.
- d. $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{E}_i$, $i = 1, \dots, 15$, where
 $\mathbf{X}_i(2 \times 2) = \begin{bmatrix} 1 & \text{trt}_{i1} \\ 1 & \text{trt}_{i2} \end{bmatrix}$, $\boldsymbol{\beta}(2 \times 1) = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$, $\mathbf{E}_i(2 \times 1) = \begin{bmatrix} E_{i0} \\ E_{i1} \end{bmatrix}$
- 6. a. $Y_{ij} = \beta_0 + \beta_1(\text{trt})_{ij} + E_{ij}$, $i = 1, \dots, 15$, $j = 1, 2$, where $(\text{trt})_{ij} = 1$ if the j th treatment on the i th subject is the active treatment, $= 0$ if the j th treatment on the i th subject is the placebo.
- c. In this analysis, each subject gets both treatment and placebo, so there are 30 independent difference scores contributing to the variance due to the effect of

treatment, which results in 29 d.f. for error. In the analysis for the data of Problem 5, each subject gets either the treatment or the placebo for both meals, so the 15 subjects receiving the active treatment contribute $15 - 1 = 14$ d.f. to error and the 15 subjects receiving the placebo contribute another 14 d.f. to error, so the error d.f. = $14 + 14 = 28$ for all subjects together. (Note: The error d.f. of 4.4168 for treatment from the data of Problem 5 is considerably higher than the error d.f. of 2.2540 for treatment from the data of Problem 6; the latter error is smaller because each subject is serving as its own control, which is not the case for the data of Problem 5.)

d. $Y_{ij} = \beta_0 + \beta_1(\text{trt})_{ij} + \beta_1(\text{seq})_{ij} + E_{ij}, i = 1, \dots, 15, j = 1, 2,$

$$\text{where } (\text{seq})_{ij} = \begin{cases} 1 & \text{for subjects receiving the active treatment first and the placebo second} \\ 0 & \text{for subjects receiving the placebo first and the active treatment second} \end{cases}$$

Chapter 26

1. a. $\text{SF}_{ij} = \beta_0 + b_{i0} + \beta_1 D_{ij1} + \beta_2 D_{ij2} + E_{ij}, i = 1, \dots, 19; j = 1, 2$
where E_{ij} and b_{i0} are each assumed to be normally distributed as $N(0, \sigma^2)$ and $N(0, \sigma_0^2)$, respectively, and b_{i0} and E_{ij} are mutually independent for all i, j . Also, D_{ij1} and D_{ij2} are dummy variables that distinguish the three days, and SF_{ij} denotes the shoulder flexion measurement for the i th subject on the j th day.
- b. $H_0: \beta_1 = \beta_2 = 0$
- c. Yes, there appears to be a larger difference between Friday (15.78) and the other days (18.36 and 18.25).
- d. To use SAS's MIXED procedure, the table needs to be reorganized so that there are $19 \times 3 = 57$ lines of data, with three lines per subject. There should be three columns of data per subject, containing the data for the following variables, subject id, day, and sf, where day denotes the day on which measurements are taken (coded, say, as 1 = Monday, 2 = Wednesday, 3 = Friday) and sf denotes the shoulder flexion measurement on a subject on a given day.
- e. The factor Day is not significant, regardless of whether model-based standard errors with a random intercept model, empirical standard errors with a random intercept model, or empirical standard errors with a marginal model and an AR1 correlation structure are used. The F statistic obtained when an empirical standard error is used is $F = 1.19 (P = .317)$ for both the random intercept model and the AR1 marginal model. The F statistic obtained for the model-based random intercept model is $F = 1.5646 (P = .2231)$. The use of an empirical standard error makes most sense here, since such use considers the possibility of misspecification of the correlation structure.
- f. $\hat{\rho} = \frac{113.90}{113.90 + 25.8839} = 0.8148$
- g. A test of significance for the random effect of Subjects can be performed using an approximate Wald test based on the output provided for "Covariance Parameter Estimates." The null hypothesis is $H_0: \sigma_0^2 = 0$, where σ_0^2 is the variance

component associated with the random intercept b_{i0} in the random intercept only model stated in the answer to part (a). For this test, the Z value is $Z = 2.79$, which has a two-tailed P -value of .0027. However, because the alternative hypothesis for this test is $H_A: \sigma_0^2 > 0$, the two-tailed P -value must be halved, so that the correct P -value is .0014, which is clearly significant at the .01 level. Consequently, using the approximate Wald test, we would conclude that the random-effect of Subjects is significant, which justifies using a model with at least a random intercept. If this test had been nonsignificant, there would be several sources of concern:

- i. Perhaps a different conclusion would have been obtained from the more appropriate mixture test (in this case, also equivalent to the approximate LR test).
 - ii. Perhaps the use of either a marginal model with an AR1 correlation structure or simply an independence correlation structure would be more appropriate. (Note, however, that the computer output suggests that the results of the test for the effect of the Day variable would not change from nonsignificance no matter what other model or correlation structure is used.)
- 2. a.** $SF_{ij} = \beta_0 + b_{i0} + (\beta_1 + b_{i1})D_{ij1} + (\beta_2 + b_{i2})D_{ij2} + E_{ij}$, $i = 1, \dots, 19; j = 1, 2, \dots, 12$, where E_{ij} , b_{i0} , b_{i1} , and b_{i2} are each assumed to be normally distributed as $N(0, \sigma^2)$, $N(0, \sigma_0^2)$, $N(0, \sigma_1^2)$, and $N(0, \sigma_2^2)$, respectively, noting that the variance components corresponding to b_{i1} and b_{i2} are typically assumed to be equal. It is also assumed that b_{i0} , b_{i1} , b_{i2} , and E_{ij} are mutually independent for all i, j . Also, the variables D_{ij1} and D_{ij2} are dummy variables that distinguish the three days, and SF_{ij} denotes the j th shoulder flexion measurement for the i th subject.
- b.** The factor Day is not significant. The F statistics are 1.69 ($P = .1986$) and 1.17 ($P = .3222$) for model-based and empirical standard errors, respectively.
- c.** A test of significance for the interaction random effect of Subjects \times Day can be performed using an approximate Wald test based on the output provided for “Covariance Parameter Estimates.” The null hypothesis is $H_0: \sigma_1^2 = \sigma_2^2 = 0$, where σ_1^2 and σ_2^2 are the variance components associated with the random slopes b_{i1} and b_{i2} , respectively, in the random mixed model stated in the answer to part (a). For this test, the Z value is $Z = 3.35$, which has a two-tailed P -value of .0004. However, because the alternative hypothesis for this test is $H_A: \sigma_1^2 = \sigma_2^2 > 0$, the two-tailed P -value must be halved, so that the correct P -value is .0004, which is clearly significant at the .01 level. Consequently, using the approximate Wald test, we would conclude that the interaction random effect of Subjects \times Day is significant, which justifies using random slopes in the model. If this test had been nonsignificant, then
 - i. Perhaps a different conclusion would have been obtained from the more appropriate mixture test (in this case, also equivalent to the approximate LR test).
 - ii. The use of a random intercept only model would be considered next.
- 3. a.** The factor Blocking Step should be considered a fixed factor because there are only two categories (i.e., levels) of this factor of interest.
- b.** $Y_{ij} = \beta_0 + b_{i0} + (\beta_1 + b_{i1})B_{ij} + E_{ij}$, $i = 1, \dots, 6$ (sample); $j = 1, \dots, 6$ (replicate), where E_{ij} , b_{i0} , and b_{i1} are each assumed to be normally distributed as

$N(0, \sigma^2)$, $N(0, \sigma_0^2)$, and $N(0, \sigma_1^2)$, respectively. It is also assumed that b_{i0} , b_{i1} , and E_{ij} are mutually independent for all i, j . (Note: This assumes that the random effects b_{i0} and b_{i1} are uncorrelated whereas a more general assumption [i.e., an unstructured \mathbf{G} matrix] would allow the random effect of b_{i0} to be correlated with b_{i1} .)

Also, the variable B_{ij} is a (0,1) dummy variable that distinguishes the two Blocking Step categories, and Y_{ij} denotes the j th antibody measurement for the i th sample.

- c. The subject-specific matrix form of the model is given by

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{E}_i, i = 1, \dots, 40$$

where \mathbf{Y}_i (6×1) denotes the collection of six antibody measurements from the i th sample, \mathbf{X}_i (6×2) denotes the values corresponding to the constant (always 1) together with values (0 or 1) for the Blocking Step variable, \mathbf{Z}_i (6×2) denotes the values corresponding to the two random effects (for intercept and slope), $\boldsymbol{\beta}$ (2×1) denotes the parameter vector for the collection for β 's in this model, \mathbf{b}_i (2×1) is the vector of random effects (b_{i0} and b_{i1}), and \mathbf{E}_i (6×1) denotes the collection of error terms on the i th subject.

- d. $H_0: \beta_1 = 0$

- e. No, the sample means for each Blocking Step category (7.66, 7.67) are almost identical.

5. a. To use SAS's MIXED procedure, the data need to be reorganized so that there are $24 \times 4 = 96$ lines of data, with four lines per rat. There should be five columns of data per rat, containing the data for the following variables: rat id, IPR, PRS, Drug, and LPR. The variable Factor A can be defined as shown in the programming statements provided.

c.
$$\begin{aligned} \text{LPR}_{ij} = & \beta_0 + b_{i0} + \beta_1 D_{ij1} + \beta_2 D_{ij2} + \beta_3 D_{ij3} + \beta_4 A_{ij1} + \beta_5 A_{ij2} + \beta_6 A_{ij3} + \beta_7 A_{ij4} \\ & + \beta_8 A_{ij5} + \delta_{11} D_{ij1} \times A_{ij1} + \delta_{21} D_{ij2} \times A_{ij1} + \delta_{31} D_{ij3} \times A_{ij1} + \delta_{12} D_{ij1} \times A_{ij2} \\ & + \delta_{22} D_{ij2} \times A_{ij2} + \delta_{32} D_{ij3} \times A_{ij2} + \delta_{13} D_{ij1} \times A_{ij3} + \delta_{23} D_{ij2} \times A_{ij3} + \delta_{33} D_{ij3} \times A_{ij3} \\ & + \delta_{14} D_{ij1} \times A_{ij4} + \delta_{24} D_{ij2} \times A_{ij4} + \delta_{34} D_{ij3} \times A_{ij4} + \delta_{15} D_{ij1} \times A_{ij5} \\ & + \delta_{25} D_{ij2} \times A_{ij5} + \delta_{35} D_{ij3} \times A_{ij5} + E_{ij}, i = 1, \dots, 24; j = 1, 2, 3, 4, \end{aligned}$$

where E_{ij} and b_{i0} are each assumed to be normally distributed as $N(0, \sigma^2)$ and $N(0, \sigma_0^2)$, respectively, and b_{i0} and E_{ij} are mutually independent for all i, j . Also, the variables D_{ij1} , D_{ij2} , and D_{ij3} are dummy variables that distinguish the four treatments (i.e., three drugs and a placebo) that are administered to each rat; the variables A_{ij1} , A_{ij2} , A_{ij3} , A_{ij4} , and A_{ij5} are dummy variables that distinguish the six categories of Factor A; and LPR_{ij} denotes the lever press rate for the i th rat on the j th treatment.

- d. The null hypotheses, F statistics, and P -values can be stated as follows:

Main effect of Drug: $H_0: \beta_1 = \beta_2 = \beta_3 = 0$, $F_{3, 54}(\text{empirical}) = 3486.01$ ($P < .0001$)

Main effect of Factor A: $H_0: \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$, $F_{5, 18}(\text{empirical}) = 227.13$ ($P < .0001$)

Interaction of Drug with Factor A: $H_0: \delta_{gb} = 0$ for $g = 1, 2, 3; b = 1, 2, 3, 4, 5$, $F_{15, 54}(\text{empirical}) = 279.93$ ($P < .0001$)

Conclusion: The interaction of Factor A with Drug and both main effects of Factor A and Drug are all highly significant.

7. a. Both Group and Time should be considered as fixed factors because the two levels of each of these variables are the only levels of interest.
- b. $\text{MRUS}_{ij} = \beta_0 + b_{j0} + \beta_1 G_{ij} + \beta_2 T_{ij} + \beta_3 G_{ij} \times T_{ij} + E_{ij}$, where $i = 1, \dots, 21$ (subject); $j = 1, 2$ (time), G represents a fixed effect for group, and T represents a fixed effect for time. E_{ij} and b_{j0} are assumed mutually independent and respectively normally distributed as $N(0, \sigma^2)$ and $N(0, \sigma_0^2)$.
- c. The subject-specific matrix form of this model would be:

$$\text{MRUS}_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}_i + \mathbf{E}_i$$
 where MRUS_i (2×1) denotes a participant's PreOp and PostOp MRUS measurements, \mathbf{X} (2×4) the (fixed-effect) predictor values, $\boldsymbol{\beta}$ (4×1) the fixed-effect parameters, \mathbf{Z} (2×1) = $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ the coefficients for the subject-specific random intercept, \mathbf{b}_i (1×1) = $[b_{i0}]$ the random intercept, \mathbf{E}_i (6×1) = the random error for each observation.
- d. Main effect of Group $H_0: \beta_1 = 0$. However, should first test for Group-by-Time interaction—that is, $H_0: \beta_3 = 0$ —since this test considers whether the change over time differs between the two groups.
- g. The approximate Wald test for the random intercept is highly significant ($P = .0050/2 = .0025$) at the 5% level, which suggests that a random intercept only model is appropriate.

8. a. Sample means:

	Rater 1			Rater 2				
	Monday	Wednesday	Friday		Monday	Wednesday	Friday	
AM	17.4	16.2	13.3	15.6	18.9	19.7	16.6	18.4
PM	17.9	16.9	16.4	17.1	18.5	20.1	16.8	18.5
	17.7	16.6	14.9	16.4	18.7	19.9	16.7	18.4

- c. Day-by-Time Interaction: Small amount [AM: ($M = 18.2$, $W = 18.0$, $F = 14.9$), PM: ($M = 18.2$, $W = 18.5$, $F = 16.6$)], since bigger drop from W to F for AM than for PM.
 Day-by-Rater Interaction: Small amount [Rater1: ($M = 17.7$, $W = 16.6$, $F = 14.9$), Rater 2: ($M = 18.7$, $W = 19.9$, $F = 16.7$)], since decrease from M to W for Rater 1 but increase from M to W for Rater 2.
 Time-by-Rater Interaction: Yes [Rater 1: (AM = 15.6, PM = 17.1), Rater 2: (AM = 18.4, PM = 18.5)], since increase from AM to PM for Rater 1 but slight decrease from AM to PM for Rater 1.
- d. Day-by-Time-by-Rater Interaction: Small amount since Day-by-Time interaction for Rater 1 is slightly different than Day-by-Time interaction for Rater 2.
- e. i. The Rater factor should be considered fixed if the two raters being considered are the only two raters about which the investigators want to draw conclusions.

$$\text{ii. } \begin{aligned} \text{SF}_{ij} = & (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})D_{ij1} + (\beta_2 + b_{i2})D_{ij2} + (\beta_3 + b_{i3})T_{ij} + (\beta_4 + b_{i4})R_{ij} \\ & + (\beta_5 + b_{i5})(D_{ij1} \times T_{ij}) + (\beta_6 + b_{i6})(D_{ij2} \times T_{ij}) + (\beta_7 + b_{i7})(T_{ij} \times R_{ij}) \\ & + (\beta_8 + b_{i8})(D_{ij1} \times R_{ij}) + (\beta_9 + b_{i9})(D_{ij2} \times R_{ij}) + \beta_{10}(D_{ij1} \times T_{ij} \times R_{ij}) \\ & + \beta_{11}(D_{ij2} \times T_{ij} \times R_{ij}) + E_{ij}, \end{aligned} \quad i = 1, \dots, 19 \text{ (subject); } j = 1, 2, \dots, 12,$$

where D_{ij1} and D_{ij2} denote the values of two dummy variables D_1 and D_2 for the three days (Monday, Wednesday, Friday), T_{ij} denotes the values of a binary variable T for time of day (typically coded as 0/1), and R_{ij} denotes the values of a binary variable T for rater (typically coded as 0/1) for the j th observation on the i th subject. It is also assumed that E_{ij} , b_{i0} , b_{i1}, \dots, b_{i9} are each normally distributed as $N(0, \sigma^2)$, $N(0, \sigma_0^2)$, $N(0, \sigma_1^2), \dots, N(0, \sigma_9^2)$, respectively, noting that the variance components for Subjects-by-Day—that is, σ_1^2 and σ_2^2 —corresponding to b_{i1} and b_{i2} are typically assumed to be equal, that the variance components for Subjects-by-Day-by-Time—that is, σ_5^2 and σ_6^2 —corresponding to b_{i5} and b_{i6} are typically assumed to be equal, and that the variance components for Subjects-by-Day-by-Rater—that is, σ_8^2 and σ_9^2 —corresponding to b_{i8} and b_{i9} are typically assumed to be equal.

- iii.** No, the correlation structure is much more complicated than exchangeable. The correlation structure would be exchangeable only if the model contained a single random effect for Subjects (i.e., random intercept only).
- iv.** The test for the main effect of Rater is the only test statistic that is significant—that is, $F_{1,18}$ (Empirical) = 19.24($P = .0004$).
- v.** Yes, the “NOTE” indicates that the iteration process used to obtain ML estimates was faulty, particularly because one or more of the matrices involved in the calculations did not have an inverse (see Appendix B on matrices). Thus, the results for this model are questionable, and one or more alternative models should be considered instead.

$$\text{vi. } \begin{aligned} \text{SF}_{ij} = & (\beta_0 + b_{i0}) + \beta_1 D_{ij1} + \beta_2 D_{ij2} + \beta_3 T_{ij} + \beta_4 R_{ij} + \beta_5(D_{ij1} \times T_{ij}) \\ & + \beta_6(D_{ij2} \times T_{ij}) + \beta_7(T_{ij} \times R_{ij}) + \beta_8(D_{ij1} \times R_{ij}) + \beta_9(D_{ij2} \times R_{ij}) \\ & + \beta_{10}(D_{ij1} \times T_{ij} \times R_{ij}) + \beta_{11}(D_{ij2} \times T_{ij} \times R_{ij}) + E_{ij} \end{aligned}$$

where $i = 1, \dots, 19$ (subject); $j = 1, 2, \dots, 12$. The two programs differ in that the first program computes model-based standard errors whereas the second program computes empirical standard errors.

- f. i.** The Rater factor could be considered a random factor if the investigators were interested in drawing conclusions to a larger population of raters instead of just the two raters that were used in the study.

$$\text{ii. } \begin{aligned} \text{SF}_{ij} = & (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})D_{ij1} + (\beta_2 + b_{i2})D_{ij2} + (\beta_3 + b_{i3})T_{ij} + r_{i1} \\ & + (\beta_4 + b_{i4})(D_{ij1} \times T_{ij}) + (\beta_5 + b_{i5})(D_{ij2} \times T_{ij}) + r_{i2}T_{ij} + r_{i3}D_{ij1} \\ & + r_{i4}D_{ij2} + r_{i5}(D_{ij1} \times T_{ij}) + r_{i6}(D_{ij2} \times T_{ij}) + r_{i01} + r_{i02}T_{ij} + r_{i03}D_{ij1} \\ & + r_{i04}D_{ij2} + E_{ij} \end{aligned}$$

where E_{ij} , b_{i0} , b_{i1} , b_{i2} , b_{i3} , b_{i4} , and b_{i5} are each assumed to be normally distributed as $N(0, \sigma^2)$, $N(0, \sigma_0^2)$, $N(0, \sigma_1^2), \dots, N(0, \sigma_5^2)$, respectively, and r_{i1} , r_{i2} , r_{i3} , r_{i4} , r_{i5} , r_{i6} , r_{i01} , r_{i02} , r_{i03} , and r_{i04} are each assumed to be normally distributed as $N(0, \sigma_{1r}^2), \dots, N(0, \sigma_{10,r}^2)$, respectively.

$$\text{iv. } \begin{aligned} \text{SF}_{ij} = & \beta_0 + \beta_1 D_{ij1} + \beta_2 D_{ij2} + \beta_3 T_{ij} + r_{i1} + \beta_4(D_{ij1} \times T_{ij}) \\ & + \beta_5(D_{ij2} \times T_{ij}) + E_{ij} \end{aligned}$$

Chapter 27

- 1. a.** $n \geq 2 \left[\frac{\sigma(Z_{1-\alpha/2} + Z_{1-\beta})}{\Delta} \right]^2 = 2 \left[\frac{(25)(1.96 + 0.842)}{25} \right]^2 = 15.7$. At least 16 patients will be needed in each treatment group.
- b.** $n \geq \left[\frac{(1.96)\sqrt{(2)(0.3)(1 - 0.3)} + (0.842)\sqrt{(0.25)(1 - 0.25) + (0.35)(1 - 0.35)}}{0.10} \right]^2 = 328.6$. At least 329 patients will be needed in each treatment group.

Power	Sample Size Required in Each Group
0.6	206
0.7	259
0.8	329
0.9	440

The required sample size increases as the power increases.

Differences in Proportions	Sample Size Required in Each Group
0.01	29,281
0.02	7,550
0.03	3,397
0.04	1,933
0.05	1,251
0.06	878
0.07	652
0.08	504
0.09	402
0.10	329

The required sample size increases dramatically as the difference in proportions decreases.

- 5. a.** $n \geq 2 \left[\frac{(3.0)(Z_{0.975} + Z_{0.8})}{1.5} \right]^2 = 2 \left[\frac{(3.0)(1.96 + 0.842)}{1.5} \right]^2 = 62.8$. At least 63 smokers and 63 nonsmokers are required.
- b.** Using the adjustment suggested by Hsieh et al., the total sample size for the multiple regression would be $126/(1 - 0.10)$ or 140 (i.e., 70 smokers and 70 nonsmokers).

Squared Multiple Correlation	Total Sample Size Required
0.1	140
0.2	158
0.3	180
0.4	210
0.5	252
0.6	315
0.7	420
0.8	630
0.9	1260

The required sample size increases as the squared multiple correlation between the main predictor and the other predictors increases.

$$\text{d. } n_s \geq \left[\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{C(\rho)} \right]^2 + 3 = \left[\frac{1.96 + 0.842}{\frac{1}{2} \ln \left(\frac{1+0.4}{1-0.4} \right)} \right]^2 + 3 = 46.7.$$

At least 47 subjects should be sampled.

- e. Using the adjustment suggested by Hsieh et al., the total sample size for the multiple regression would be $47/(1 - 0.10)$ or 53.

Reference

Hsieh, F. Y.; Bloch, D. A.; and Larsen, M. D. 1998. "A Simple Method of Sample Size Calculation for Linear and Logistic Regression." *Statistics in Medicine* 17: 1623–34.

INDEX

Page numbers followed by “f” or “t” indicate figures or tables, respectively.

- Additivity
adding, Tukey’s F (usage), 572
Tukey’s test, 569–570
- Adjusted means
BRFSS data, 318
computation, 311
differences, 329
equivalence, 330
examples, 317–318
increases, determination, 332
obtaining, 312
tolerance scores, determination, 324
- Adjusted odds ratio, 686
confidence interval, obtaining, 694, 697
estimated ratio, effects, 697t
interval estimates, 695t
obtaining, 688–689, 696
point estimates, 695t
- Adjustment, covariance method
(application), 313
- Akaike’s Information Criterion
(AIC), 861
- Alternative hypothesis, 30
- Analysis
choice, 12–14
prediction-oriented process, 439
weighted-least-squares method, 357–358
- Analysis of covariance (ANACOVA), 4–5, 227, 308, 310–313
adjustment
problem, 309–310
procedure, 318
rationale, 318–319
alternatives, 320–321
- BRFSS example, 316, 336
comments/cautions, 318–321
control design, 322
covariates, 314–316
dependent variable, usage, 325–326
examples, 312–313, 315–316, 316t
- groups, 314–316
methods, usage, 334, 335
model, analysis scenario, 317
nominal independent variables, 316–318
parallelism assumption, 319
plot, 965
regression model, usage, 323, 330, 334
SAS, usage, 963–965
usage, 315, 327, 404, 568
validity/precision, 319–320
variable control, example, 311f
- Analysis of variance (ANOVA)
ANOVA-type data, regression
model representations, 763
computer programs, usage, 641
effects (estimation), computer
package (usage caution), 564
factors/levels, 482–483
fixed-effects model, usage, 524
model, specification, 847
multiple-comparison technique,
selection procedure, 515f
name, selection (reasons), 482
one-way ANOVA, 481
partitioning approach, 825
presentation, 312–313
regression, contrast, 482
sample size planning, 883
SAS, usage, 965–972
sum of squares, orthogonal squares/
partitioning, 516–521
- Analysis of variance (ANOVA) table, 129, 166
alternative, 132t
basis, 194–195
column effects, impact, 640t
completion, example, 148
determination/completion, 624, 626
edited SAS output, 130
multiple regression usage, 145–146
- numerical examples, 148–151
problems, 133–135
regression sum of squares,
components, 168t
row effects, impact, 640t
variables-added-in-order tests, 270t
variables, regression, 175t
- Apache score, 721
- Arcsin transformation, 356
- Association
causality, contrast, 42–45
linear association, index, 109
measure, correlation coefficient (r)
usage, 109–112, 110f
specificity, 44
strength, 44
- Assumptions
statement, 51–55
two-way ANOVA, 592–594
- Autoregressive clusters, 849
- Autoregressive structure, 842
- Backward approach, usage, 466, 468
- Backward elimination
analysis, 458
procedure, 447–449
SAS output, 448–449
strategy, 172
usage, 459t
- Backward method, 48
- Backward testing strategy, 273f
- Balanced crossover design, 840
- Balanced patterns, 630
- Bayes’ Information Criterion
(BIC), 861
- Behavioral Risk Factor Surveillance
System (BRFSS), 5
analysis, examples, 71–74, 122–123
analysis, multiple regression analysis
example, 146–148
data, unadjusted/adjusted means, 318
linear regression model, 106

- Behavioral Risk Factor Surveillance System (*continued*)
 multiple regression, 186–188
 output, example, 72
- Bends, consideration, 410
- Bernoulli populations (point-binomial populations), 885
- Bernoulli random variable (point-binomial random variable), 698–699
- Best-fit problem, solution, 57–59
- Best fitting, meaning, 51
- Best-fitting regression model, 438
- Best-fitting regression surface, 348
- Best-fitting straight line
 data, SAS output, 59t
 determination, 49, 55–59
 example, 58f
- Best regression equation, 438, 470–471
 backward elimination, 459t, 462t
 computer programs, usage, 452–453
- cross-validation analysis, 462
- data analysis, example, 458–463
- data set
 analysis, examples, 457–463
 correlations, 458t
- descriptive statistics, 458t, 462t
- first-order interactions, 476–477
- five-variable model, preference, 461
- fixed factors, random factors
 (contrast), 483–484
- interaction-ordered fitting, 460t
- maximum model, specification
 (prediction goal), 439–442
- model selection, criterion
 specification, 442–444
- multiple-comparison technique,
 selection, 515–516
- power-ordered fitting, 460t
- problems, 466–480
- selection
 criteria, 442–443
 steps, 439
- Between-population variance, 899
- Binary logistic regression, 977–978
 extension, 714
 odds ratios, 978–979
- Binary outcome variable, binary
 exposure variable (effect), 908
- Binary predictor, usage, 887
- Binary regression, usage (reason), 715
- Binomial distribution, 698–699
- Binomial probability parameter, 746
- Biracial dyads, outcomes, 3–4
- Bivariate normal distribution, 112–113
 example, 112f
 parameters, 200
- Block diagonal matrix, 799
- Blocking, 579
- incorrectness, example, 554f
 principle, 553–554
 step, 866
- Blocks, formation, 554f
- Block sum of squares (SSB),
 definition, 559
- Bonferroni approach, 504–507
 ANOVA table, example, 505t
 data, examples, 505t
 example, 504
 multiple-comparison analysis,
 505–506
 sample means, crude comparison,
 505f, 506f
- Bonferroni corrected jackknife
 residual critical values, 933t
- Bonferroni corrected studentized
 residual critical values, 933t
- Bonferroni multiple-comparison
 procedures, 522
- Box-and-whisker plot (boxplot), 19
- Candidate sets, example, 241t
- Case-control pairs, frequencies
 (exposure status), 706t
- Categorical variable, 758
- Categories, ordering, 9
- Causal conjectures, criteria, 44–45
- Causality, association (contrast), 42–45
- Cell
 frequency, obtaining, 547
 means
 data, 582t
 matrix, 563t
 table, usage, 581–586
 sample sizes, 635
 single observation
 experimental situations, 548f
 impact, 547–549
- Centering, 365
 usage, 367–368
- Central tendency, measure, 17
- Chi-square (χ^2) distribution, 24, 26,
 755–756
 example, 25f
 percentiles, 920t
- Chi-square (χ^2) random variable,
 761, 855
- Chi-square (χ^2) statistic, 755, 851
- Chunk tests, usage, 637–638, 712
- Chunkwise methods, 453–454
 single-variable selection methods,
 contrast, 454
- Classification schemes, overlapping,
 11–12
- Class intervals, grouping, 8
- Cluster, 818
- Coefficient, *F* test (association),
 212–214
- Coincidence, testing, 267–268, 272
- Coincident lines, tests (differences),
 272
- Collinearity, 231–232, 358–372
 assessment, 362, 366
 concepts, 362–368
 decrease, centering (usage),
 367–368
 diagnostics, 368–372
 output, 371
 examples, 358–359, 369–371
 intercept, role, 367
 mathematical concepts, 359
 near collinearity, 365
 perfect collinearity, 362–363
 predictors, usage, 360–362
 problem, 339, 435
 checking, 382
 interaction terms, impact,
 371–372
 reduction, 372
 treatment, 371–372
- SAS scatterplot, 370f
- term, usage, 362
- Collinear, term (usage), 361
- Column factor, 545
 effects, 596–597, 598
- Column main effects, 601–602
- Column marginal means, 597, 598
- Common variance, 805
- Completely balanced crossover
 design, 840
- Complete pattern, 630
- Complete second-order model, 213
- Compound symmetric (CS), 807
 correlation structure, 811
 marginal model, 844
 matrix, 802–803
- Computer-generated standard error
 estimates, 183
- Coincidence (test), single-model
 approach (usage), 271
- Concordant pairs, 702
- Conditional distribution, 112
 univariate normal distribution,
 203–204
- Conditional independence,
 assumption, 830
- Conditional likelihood, 700
 function, conditional probability
 (comparison), 704
- Conditional logistic regression, 980
- Conditional margins, 337
- Conditional ML estimation/
 procedure, 698, 700–704
- matched case-control studies, 701
- pair-matched data, unmatched
 covariates (combination),
 704–707
- Conditional (multiple partial *F*) test,
 usage, 641

- Condition index (CI), 366
 determination, 385, 390
 statistics, examination, 368–369
 usage, 371
- Condition numbers, determination, 385
- Confidence bands, 67, 962–963
 determination, 85, 91
 example, 68f
 simplification, 67
 sketching, 85, 91
- Confidence intervals, 27, 67, 183
 advantage, 182
 calculations, sample, 64t
 construction, 86
 determination, 94
 equations, 63t
 estimated values, usage, 185
 expression, 774
 formula, square root part, 70
 interpretation, 83
 α level, change, 963
 lower limit, denotation, 119
 models, 726t
 obtaining, 697
 ρ (rho), 119–120
 SAS output, 69
 usage, 313
- Confidence level, 27
- Confounders, 11, 756
 control, 171–172
 difference, 239
- Confounder (control) variable, 686
- Confounding, 226, 758
 assessment, 226–227, 237
 no-interaction model, 242
- BRFSS data, example, 255
- control, 888–889
 candidate sets, example, 241t
 variables, usage, 240
- defining, 239
- existence, 227
- problems, 243–256
- statistical test, usage, 238
- usage, 236–241, 308
- Continuous data, adjustment
 methods, 308
- Continuous independent variable, impact, 481
- Continuous predictor
 linear regression, 341
 usage, 888
- Continuous variables
 discrete variables
 contrast, 7f
 categorization, decision, 9
 frequency distributions, 8f
- Contrast, 986–987
 definition, 510
 evaluation, 511
 statement, 821–822
- sum of squares, association, 516
 target set, 515
 usage, 509–515
- Control (confounder) variable, 686
- Control variable (Z), 227
 second expression, numerator, 208–209
- Cook's distance, 349–351
 examination, 383
 example, 350
 value, 351t
- Cook's values, maximum (critical values), 935t
- Cornell Medical Index scores, example, 487t
- Correction factor, 132
- Correlated data, analysis, 781, 825
 ANOVA approach, 784
 data layout, 783t
 problems, 819–824, 862–882
 SAS, usage, 982–990
 summary, 818, 858–859
- Correlated data (mixed model), random effects (usage), 839t
- Correlation
 equality, testing, 120–122
 matrix, 801–802
 methods, focus, 2
 problems, 215–225
 range, examples, 111f
 SAS output, 206
 spurious correlation, 216–217
- Correlation coefficient (r), 199
 computation, 123
 confidence intervals, tests, 117–120
 definition, 108–109
 estimate, 122–123
 hypotheses, tests, 117–120
 population correlation coefficient, 109–110
 problems, 125–127
 sample
 defining, 109
 square, 114
 SAS, usage, 953–955
 terms, 165
 usage, 109–112
- Correlation matrix, 200–201, 365
 determination, 390
 examination, 385
 form, 200
 usage, 221, 819
- Correlation structure, 794–800, 812
 assumption, 881
 covariance structures, relationship, 800–803
 determination, 842
 differences, 847
 exchangeable correlation structure, 797–798
- independent correlation structure, 796–797
 recommendations, 859–861
 underlying correlation structure, 798
 unstructured correlation structure, 798
 usage, 859
 working correlation structure, 799, 818
- Count data (discrete data), 743
- Covariance, 800
 adjustment, 312
 analysis, 4, 5, 309
 problems, 321–338
 estimates, examples, 336
- matrix
 estimated covariance matrix, 709
 form, 801–802
- structures, 812
 correlation structures, relationship, 800–803
 determination, 842
- Covariates, 11, 227, 308
 coefficients, examination, 314–315
 control, 239–240
 impact, 314–316, 320–321, 792
 specifications, comparison, 749
 values, 717
- Critical point, 31
 example, 32f
- Critical region (rejection region), 31
 example, 34f
- Cross-classification model, 763
- Crossover factor, 840
- Cross-validation analysis, 462
- Crude rate ratio estimate, 758
- Cubic model, regression ANOVA table, 409t
- Cumulative logit model (proportional odds model), 726
- Data problems, diagnosis
 (approaches), 340–347
 example, 342
- SAS output (PROC UNIVARIATE), 343
 scatterplots, 344f–347f
- Data set
 confounding/interaction, usage, 228
 variables, inclusion, 302–303, 333–334
- Data transformations, usage
 (reasons), 356
- Data value, examination, 340–341
- Delinquency index (DI), 83
- Denominator degrees of freedom (denDF) (DDFM), 846, 855–858
 results, 856t

- Dependent variable, 1, 10–11
 dichotomous dependent variable, 681–682
 example, 317
 plots, 341–342
 usage, 325–326
- Descriptive analysis, 18
- Descriptive orientation, 10–11
 example, 11f
- Descriptive statistics, 16–19
 example, 458
- Deterministic models, statistical models (contrast), 45
- Deviance, 754–755
 hierarchical class, usage, 755
 information, 772
 value, 760–761
- Diagnostics
 analysis, 347–348
 regression diagnostics, 339
- Diagonal matrix, 940
- Dichotomous dependent variable, 681–682
- Dichotomous outcome variable, coding, 717
- Dichotomous variable, addition, 824
- Differences of means, interaction method, 599–600
- Directed acrylic graph (DAG), 43–44
- Discordant pairs, 702
- Discrete data (count data), 743
- Discrete random variable, probability distribution, 662
- Discrete variables
 continuous variables
 contrast, 7f
 treatment, 8f
 frequency distributions, 8f
 treatment, 8
- Dose-response effect, 44
- Double-bundle technique, 891–892
- Double-subscript notation, usage, 486
- Dummy variables, 257
 change, 563–564
 coded example, 644
 coefficients, 494
 defining, 275–276, 300, 621
 rule, 258–259
 definitions, 257–258
 allowance, 499
 examples, 257–258
 indexing type, 889
 $k - 1$ dummy variables, defining, 258–259
 methods, comparison, 271–272
 model, 268–269
 alternatives, 276–277
 usage, 273–275
 problems, 283–307
 SAS, usage, 956–957
- usage, 309, 568
 problem, 371
 value, denotation, 793–794
- variables-added-in-order test, 270
- variables-added-last test, 270
- Effect coding
 model, 496
 scheme, 494
- Effect measure modification, 235
- Effect modification, 235
 interaction, contrast, 235–236
- Effect modifiers, 235, 236, 687
 identification, 756
- Effects, heterogeneity, 829
- Eigenvalues, 365
 determination, 385, 390
- Empirical estimator (robust estimator), 805
- Empirical *F* statistics, difference, 844, 846
- Empirical standard errors, 857
 example, 813t
 usage, 811–812
- Empirical variance estimation (robust variance estimation), 799–800
- Equal cell numbers, 579
 example, 580
 index scores, 581t
 regression model, usage, 637
- Equal-cell-number two-way ANOVA, methodology, 650
- Equal cell sample sizes, three-way ANOVA table, 642t
- Equal intercepts, test (single-model approach), 269–270
- Equality of block means, *F* test (usage), 561
- Equality of treatment means, *F* test (usage), 560
- Error component, 849–850
 example, 54
- Error rates, 35–37
- Error sum of squares, 144, 948
- Estimated correlation matrix, usage, 861
- Estimated covariance matrix, 665, 667, 764
- Estimated Mantel-Haenszel odds ratio, 702–703
- Estimated means (least-squares means), 964, 966–969
- Estimated odds ratio, 731
 adjusted effect, 735–736
 comparison, 725–726
 impact, 727t
- Estimated standard error, 25, 987
- Estimates, covariance, 186
- Estimation, 2, 16
 statistical inference, 27–30
- Estimators
 estimated standard error, 691–692
 linear functions, 61
- Evidence, coherence, 44
- Exchangeable clusters, 859
- Exchangeable correlation structure, 797–798
- Exchangeable covariance matrix, 803
- Existence, assumption statement, 51, 141
- Expected mean squares, 501, 608–610
 ratio, 608
- Experimental design, 580
- Experimental error, estimate, 551
- Experimental evidence, 45
- Experimental research, 1
- Experiment, usage, 548
- Exposure (study) variable, 686, 738
- Extraneous variables, 227
 control, 171–172, 237–241
- Extra-sum-of-squares principle, 634
- Extra-sum-of-squares test, 210
- Factor analysis, 4
- Factor classification schemes, 611
- Factors
 classification schemes, 611, 613, 617
 determination, 622
 fixed/random, 593
 numerical values, assignation, 643
 quantitative analysis, 247
- F* distribution, percentiles, 921t–927t
- Fisher's *F* distributions, 24, 26–27
 example, 25f
- Fisher's *Z* transformation, 118, 888
 usage, 119, 120, 123
- Fitted quadratic model, plotting, 429
- Fitted regression equations, 199, 314
 determination, 289
- Fitted regression lines
 observed points, deviations, 56f
 slopes, variances (estimation), 264
- Fitted regression model, 167–168
 basis, 291
- Fitting marginal models, results, 840–843
- Fixed-effect predictor variables, 846
- Fixed effects, 825
 covariates, impact, 792
 parameters, 842
 confidence interval estimation, 806
 tests, 585
- Fixed-effects ANOVA
 data, example, 498
 model
 statement, 566
 usage, 565–566
- Fixed-effects factor, impact, 537

- Fixed-effects model, 497–499, 566
 contrasts, 969–971
 interaction method, 600
 writing, 619
- Fixed-effects one-way ANOVA, 484, 497–499
 application, 485
 assumptions, 485
 Bonferroni approach, 504–507
 format, 499
 means (comparisons), contrasts (usage), 509–515
 multiple-comparison procedures, 503–515
 regression model, 494–497
 example, 494
 Scheffé method, 509–515
 Tukey–Kramer method, 507–509
- Fixed-effects two-way ANOVA, regression model, 594–599
- Fixed factors, 826, 840
 examples, 483t
 random factors, contrast, 483–484
 treatment, 840
- Fixed model, 583–586
- Fixed order, usage, 472
- Fixed structure (numerical structure), 861
- Fixed value, normal distribution, 54
- Forward approach, usage, 466, 468
- Forward method, 48
 flow diagram, 49f
- Forward selection procedure, 449–451
 SAS output, 450–451
- Forward-selection testing
 approach, 422
- Four-parameter nonlinear model, 764
- Frequency histogram, 18f
- F* statistic, 130
 calculation, computer output (usage), 615
 comparison, 133
 degrees of freedom, 869
 quantity, 268
 determination, 849
 $n - k - 1$ degrees of freedom, 181
 numerator degrees of freedom, 276
 usage, 560
 variance estimates, ratio, 491–492
- F* test
 computation, ANOVA information (requirement), 406
 conducting, 188
 expression, 165
 findings, 188
 rationale, 490–493
 results, 856
 σ^2 estimate, impact, 446
- F*-to-enter, 453
- F*-to-leave, 453
- Full model (*F*), 672, 698
- Full partial correlation, 209
- Fully specified model, 583
- Fundamental equation of regression analysis, 132
- Gappiness, 7–9
 classification, 11–12
- Gaussian errors, yield, 651
- Generalized estimating equation (GEE), 52, 861–862
 techniques, 141
- Generalized linear mixed models, usage, 784, 825
- Generalized squared correlation, 364
- General linear mixed model, 781
 ANOVA approach, 784
 approach, 792–806
 correlated data, analysis, 818
 covariance structure, usage, 803–806
 data layout, 783t
 empirical standard errors, usage, 811–812
 ML estimation, 806
 problems, 819–824
 random effects, 792
 regression scalar form, 803
 study
 data, 785t, 786t, 790t
 data layout, 791t
 examples, 785–791, 806–817
- subject-specific matrix form, 804–806
- subject-specific scalar form, 804
 unbalanced data, 812–816
 analyses, summary, 816–817
 examples, 814t, 815t
 unbalanced design, 792
- General (balanced) two-way ANOVA table, 590t
- Goodness of fit (GOF)
 assessment, 674–675
 measures, 753–756
 procedures, 861
- Groups
 variability, *F* statistic comparison, 492–493
- Groups, impact, 314–316
- Hazard ratio, 748
- Health status, change (example), 29
- Heterogeneous exchangeable covariance matrix, 802–803
- Heterogeneous experimental units, 554f
- Heteroscedasticity, 54
- Hierarchical class, impact, 671–672
- Hierarchically well-formulated (HWF) model, 231, 464
- Hierarchical sequence, interaction test, 232
- Higher-order indices, 200
- Higher-order partial correlations, computation, 209, 211
- Higher-order polynomial models
 collinearity problem, 420
 fitting/testing, 410
- Higher-order terms
 addition, 137
 collection/chunk, 213–214
- Higher-way ANOVA, 641–642
- Homogeneous exchangeable covariance matrix, 802–803
- Homoscedasticity
 assessment, residuals (graphical analyses), 352–353
 assumption
 statement, 54, 142
 violation, 352
- Hybrid limits, 70
- Hypersurface ($k + 1$)-dimensional space, 139
- Hypothesis (hypotheses)
 biological/theoretical plausibility, 44
 testing
 likelihood ratio tests, usage, 671–675
 Wald statistics, usage, 668
 tests, 183, 185, 276
 calculations, sample, 64t
 equations, 63t
- Hypothesis testing, 16
 example, 33
 outcomes, 35t
 probabilities, 35t
 results, 810
 statistical inference, 27, 30–34
- Hypothesis-testing method, 912
- Identity matrix, 796, 941
 equivalence, 945
- Incidence density ratio (IDR), 748
- Incomplete pattern, 630
- Independence assumption
 assessment, residuals (graphical analyses), 352–353
 statement, 52, 141
- Independent correlation structure, usage, 861
- Independent correlation structure, 796–797
- Independent variables, 1, 10–11
 coding schemes, 562
 examination, 290–291
 importance, determination, 42
 interactive effects, assessment, 42
 linear association, 346–347
 matrix, 946

- Independent variables (*continued*)
n × 2 matrix, 946
 pairwise correlations, calculation, 342
 relationship, 190
 assessment, 242
 separation, 137
 significance, 720
- Indicator variable, 257
- Inflation factor, 360
- Instrument error data, stem-and-leaf diagram, 18f
- Interacting independent variables, graph, 229f
- Interaction, 226, 341
 absence, 251–252, 680
 response curves, 230f
 assessment, 226–227, 234–235, 308
 BRFSS data, example, 255
 concept, 599–601
 condition, 228–229
 disability, 3–4
 effects, 598, 606t
 absence, 601–602
 evaluation, statistical testing (usage), 232
 examples, 228–230, 232–235
 model, 231–232, 757
 modeling, 231–232
 no-interaction model, 241
 a priori knowledge, 231
 problems, 243–256
 SAS, usage, 957–958
 statistical significance, 772
- Interaction effect
 description, 582
 detection, 580
 modification, contrast, 235–236
 representation, 231
- Interaction-ordered fitting, example, 460t
- Intercept
 comparison, 265–267
 example, 266–267
 equal intercepts, test (single-model approach), 269–270
 estimates, 266
 inferences, 61–64
 least-squares estimates, determination, 79, 82, 86
 role, 367
 tests, 180–181
 interpretations, 64–66
 treatment, 363–364
 VIF, impact, 364
- Intercept-added-in-order test, 180–181
- Intercept-added-last test, 180–181
- Intercept-adjusted collinearity diagnostic statistics, examination, 368–369
- Interquartile range (IQR), 19f
- Interval
 level, measurement, 9
 testing, 669–670
 variable, 10
- Iteratively reweighted least squares (IRLS), 751
 methods, usage, 763
- Jackknife residuals [$r_{(-i)}$], 348, 349
 cutoff value, 379–380
 examination, 383
 plots
 examination, 383
 examples, 352f, 394–395, 398
 plotting, 432
 probability plots, 354
 SAS normal quantile–quantile plot, 382f
- SAS probability plots, example, 354f
- SAS scatterplot, 380f–381f
 usage, example, 353f
 value, 351t
- Joint density function, 112
- Kolmogorov–Smirnov test, 354
- k* predictors, 362
- Kurtosis statistic, 354
- Lack of fit, testing, 49
- Lack-of-fit (LOF) tests, 411–412
 conducting, 420
 pure-error estimates, 411t
 regression ANOVA table, 412t
 replicates, 411t
- Large-sample chi-square distributions, 669
- Large-sample confidence interval estimation, polytomous logistic regression model (usage), 721
- Large-sample distribution, 710
- Large-sample test, 753
- Least significant chunk test, usage, 638
- Least-squares approach, 143–144
- Least-squares equation, independent variables (involvement), 153
- Least-squares estimates, determination, 75, 79, 86, 495
 formulas, 403
 multiple regression analysis example, 160
- Least-squares fitting, 312
- Least-squares formula, development, 113
- Least-squares line, determination, 283
- Least-squares means (estimated means), 964
- Least-squares method, 55–56, 212
- Least-squares parabola, observed points (deviations), 403f
- Least-squares regression equation, 144
- Least-squares solutions, comments, 144–145
- Level of confidence, 27
- Level-of-measurement classification, 11–12
- Leverage, 348–349
 critical values, 934t
 statistics, examination, 383
 values, 351t
 determination, 385, 390, 393, 396
- Likelihood
 conditional likelihood, 700
 function, 662–663
 function, derivation, 747–748, 750
 unconditional likelihood, 700
- Likelihood ratio (LR), 720
 hierarchical class, impact, 671
 statistic, 852
 computation, 854–855
 format, 672
 Wald statistic, discrepancy, 673–674
- test, 725, 806
 computation, 692
 execution, 678
 usage, 693–694
- Likelihood ratio (LR) tests, usage, 671–675
- Likelihood-ratio-type statistic, 754
- Likert Scale measurements, usage, 532
- Linear association, index, 109
- Linear combination, polynomial form, 413
- Linear contrast, 185, 786–787
- Linear function, expression, 185
- Linearity assessment, residuals (graphical analyses usage), 352–353
- Linearity assumption
 statement, 52–54, 141–142
 violation, 341
- Linear model
 power, approach, 893–907
 regression ANOVA table, 408t
 sample size determination, 893–907
- Linear orthogonal polynomial score, 416–417
- Linear prediction, variable (usage), 205
- Linear regression
 analysis, 740
 assumptions, violations, 393, 396
 context, 227
 model
 statement, 75
 usage, 666

- problems, 911–913
 procedures, 849
 sample size planning, 883,
 886–889
- Linear relationship
 statistical evidence, 73–74
 strength (description), correlations
 (usage), 201
 strength (measurement), partial
 correlation coefficient (usage),
 204–212
- Linear test (trend test), 693–694
- Logistic function, 682, 682f
- Logistic model, 681–683
 exposure variable, presence, 688
- Logistic regression
 computer output (PROC LOGISTIC), 703t
 model, 892–893
 logit form, 684 709
 numerical example, 689–698
 problems, 811–813
 sample size planning, 883,
 889–893
 SAS, usage, 977–981
 statistical modeling, 681–682
 usage, 683–689, 891
- Logistic regression analysis, 3, 681
 categorical variable, treatment,
 690–691
 mathematical model,
 postulation, 682
 problems, 708–712
 theoretical considerations, 698–704
- Logit form, 684, 709
 statement, 711
- Log-like functions, usage, 722–724
- Log likelihood statistic, 692
- Log-likelihood statistics, 673
- Log transformation, 356
- Longitudinal data, 782f
 data layout, 783t
- Longitudinal study, 781
- Lower confidence limits, yield, 698
- Lower-order terms, elimination/
 retention, 465
- Main-effect chunk tests, usage,
 637–638
- Main effects, 229
 layouts
 alternatives, 601t
 no-interaction example, 602t
- Mantel-Haenszel odds ratio
 estimate, computation, 706
 estimated Mantel-Haenszel odds
 ratio, 702–703
- Marginal methods, 337–338
- Marginal models, 805, 840, 983–986
 fitting marginal models, results,
 840–843
- model-based results, 807–811
 output, example, 808t, 809t
 usage, 858
- Marginal regression model, 820
- Matched case-control analysis, sample
 size calculation (PASS 11
 dialog box), 909f
- Matched case-control studies,
 dichotomous outcome (sample
 size determination), 908–910
- Matched data (analysis), logistic
 model (usage), 701
- Matched-pairs data, ANOVA table,
 551t
- Matched-pairs design, example, 550t
- Matched-pairs experiment, equivalent
 analysis, 549–553
- Matching variable, usage, 732
- Mathematical model
 determination, 42
 transition, 402
- Matrix (matrices)
 addition, 941–942
 calculations, 947
 definitions, 937–938
 diagonal matrix, 940
 dimensions, 938
 formulation, 946–948
 identity matrix, 796, 941
 inverse, 944–945
 multiplication, 942–944
 regression analysis, relationship,
 937
 symmetric matrix, 940
 transposition, 939–941
 usage, 938–939
- Maximized likelihood
 function, numerical value,
 665–666
 value, 665, 754
 involvement (LR test), 674
- Maximum likelihood (ML), 661
 computer output, SAS procedures
 (comparison), 675–677
 conditional ML estimation, 698,
 700–704
 equations, computer-based iteration
 procedure, 750–751
 estimates, usage, 689, 749–750
 estimation, 685–686, 714,
 784, 806
- estimators
 large-sample properties, 668
 likelihood function,
 maximization, 667
 general definition, 664–665
 hypothesis testing
 likelihood ratio tests, usage,
 671–675
 Wald statistics, usage, 668–669
- interval estimation, 669–670
- ML-based estimated regression
 coefficients, 691
- ML-based large-sample confidence
 interval, 670
- ML-based test statistics,
 usage, 669
- principle, 661–665
- problems, 678–680
- procedures, usage, 681, 698
- restricted maximum likelihood
 (REML), 806
- SAS computer output, 676t
- unconditional ML estimation,
 698–700
 usage, 665–677
- Maximum model
 analysis, conducting, 454
 backward elimination procedure,
 447–449
 chunkwise methods, 453–454
 computer programs, usage,
 452–453
 consideration, 440
 constraints, 441, 442
 data analysis, example, 457–463
 definition, 439
 degrees of freedom, requirement,
 441–442
 forward-selection procedure,
 449–451
 gold standard, 465
 polynomial terms, impact, 442
 reduced model, 443
 regressions procedure, 445–447
 reliability evaluation, split samples
 (usage), 454–457
- selection criterion
 candidates, 443–444
 specification, 442–444
 specification (prediction goal),
 439–442
- stepwise regression procedure, 452
 validity, selection, 463–465
 variables, selection strategy
 (specification), 444–454
- Mean differences, 966–969
- Mean response
 confidence interval, 76
 linear trend, test, 520–521
- Means
 comparison
 contrasts, usage, 509–515
 sample size calculations,
 884–886
 tests, sample size determination,
 884–885
 differences, interaction method,
 599–600
 multiple comparisons, 966–969
 pairs (differences), one-way
 ANOVA (usage), 528

- Means ANOVA (MANOVA)
 category, procedures, 900
 models, 893
- Mean scores, adjustment, 324
- Mean squared error (MSE), square root, 348
- Mean-square quantities, provision, 166
- Mean-square residual/regression, statistical independence, 131
- Mean squares, obtaining, 552, 591
- Mean-square test, 166
- Mean value
 estimate, 74
 specified values, 182–183
 straight-line function, 52–53
- Measurement, levels, 9–11
- Measures data, data layout, 783t
- Measures design, repetition, 789
- Minimum sum of squares, 56
- Minimum-variance approach, 144
- Minimum-variance method, 56
- Mixed effects, tests, 586
- Mixed-effects two-way ANOVA model, 607–610
 fixed row factor/random column factor, inclusion, 607–608
 mixed-effects model, random row factor/fixed column factor (inclusion), 608
- Mixed model, 583–586
 random effects, 829
 techniques, 141
- Mixture distribution, 855
- Mixture test, usage, 854–855
- Model-based estimated standard errors, 835, 847
 differences, 842
- Model-based *F* statistics, difference, 844, 846
- Model-based standard errors (non-empirical standard errors), 807, 857
- Model Fit Statistics, 720
- Model selection techniques, 976–977
- Multiple-comparison analysis, 505–506
- Multiple comparisons, Tukey–Kramer method (application), 527
- Multiple-comparison technique, selection, 515–516
 procedure, 515f
- Multiple correlation coefficient, 144, 201–203
 analog, 204
 generalization, 202
- Multiple correlations, 199, 200
- Multiple linear regression
 analysis, 755
 power, 894–898
- sample size calculation, 888–889, 894–898
- PASS 11 dialog box, 895f
- PASS 11 output, 896f
- SAS program, 897f, 902f
- SAS, usage, 955–956
- statistical inference, 958–963
- Multiple partial correlations, 199, 200, 212–214
 coefficient, 201
 usage, 212–213
- Multiple partial *F* test, 172–175
 comments, 174–175
 null hypothesis, 173
 procedure, 173–174
 provision, 329, 573
 usage, 641
- Multiple random effects, models (usage), 989–990
- Multiple regression
 ANOVA table, 145–146
 assumptions, 141–143
 BRFSS analysis, 186–188
 confidence intervals, 183
 equation, best estimate (determination), 143–145
 hypothesis tests, 183
 inference methods, 180–186
 mean value, 182–183
 new values, prediction, 184
 relationships, example, 151–153
 SAS output, 189, 191, 192–194
 statement of assumptions, 141–142
 statistical inference, 188–198
- Multiple regression analysis, 5, 136
 BRFSS analysis, example, 146–148
 problems, 151–164
 graphical examination, 138–140
- Multiple regression models, 137–138, 308
 defining, 288–289
 dependent variable, plots, 341–342
 examination, 345
 output, 186
- Multivariable methods, 13t
 application, 14t
- Multivariable techniques, 1
- Multivariate normal, 203
- Multivariate normal distribution, relationship, 203–204
- Mutually orthogonal, term (usage), 517
- Mutual orthogonality, establishment, 521
- Naive predictor, deviations, 114
- Natural heterogeneity, 826, 842
- Natural polynomials, 371
 model, 413
 transformation, 414–417
- Near collinearity, 365
- Negative quadrants, 110
- New value, prediction, 184
- No-interaction assumption, violation, 559
- No-interaction effects, 601–602
- No-interaction model, 241, 602
- No-interaction proportional odds logistic model, logit form, 740
- Nominal independent variables, 316–318
 impact, 481
- Nominal level (measurement), 9
- Nominal variables, involvement, 277–283
- Noncentered predictor data, 419
- Non-central *F* distribution, 899
- Non-empirical standard errors (model-based standard errors), 807
- Noninteracting independent variables, graph, 229f
- Nonorthogonality, 632–636
 occurrence, 635–636
- Nonrejection region, 31, 36
- Normal distribution, assumption statement, 54–55
- Normality assumption, 142
 assessment, 354–355
 example, 354
- Nuisance variable, 211t
- Null hypothesis, 30, 168–169
 chi-square distribution, 672
 consideration, 123
 expression, 181
 false rejection, 508
 LR test, 693
 neighborhood effects, absence, 501
 partial *F* test, 405–406
 prediction, involvement, 214
 regression coefficients, terms, 330
 rejection, 85, 89, 207, 484
F test, impact, 498, 505
 statement, 165, 321, 744–745
 test, 79, 82, 84, 85, 569, 710
 two-sided alternative, 336
 testing, 529
 equivalent statistic, usage, 171
- two-way ANOVA, 592
- Numerical descriptive statistics, examination, 383
- Numerical structure (fixed structure), 861
- Observational research, 1
- Observational study, 2
 usage, 548
- Observed margins weighting, 337
- Odds
 defining, 683
 estimation, 691–692
- Odds-like expressions, 717

- Odds ratio (OR)
 adjusted odds ratio, 686
 comparison, 726t
 computation, 685
 correspondence, 741, 891
 crude data layout, 716t
 descriptive ability, 687
 estimated Mantel-Haenszel odds ratio, 702–703
 estimates, comparison, 719–720
 estimation, 709
 logistic regression, usage, 683–689
 expression
 obtaining, 684–685
 usage, 694
 formula, 696, 735, 739
 derivation, 729–730
 usage, 683
 value, variation, 687–688
 $(1/2) \ln [(1+r)/(1-r)]$, values, 928t–929t
 One-tailed alternative, 118
 One-tailed test, usage, 127
 One-way analysis of variance (one-way ANOVA), 480
 ANOVA table, 493, 493t
 assumptions, 485–486
 data configuration, example, 486t
 data presentation, 486–487
 fixed-effects model, 497–499
 random-effects model, combination, 502t
 fixed-effects one-way ANOVA, 484
 format, 787
 F test, 580–581
 rationale, 490–493
 problem, 484
 problems, samples, 522–543
 random-effects model, 500–502
 format, 502
 sample size calculation
 PASS dialog box, 903f
 SAS program, 905f
- One-way fixed-effects analysis of variance (one-way ANOVA)
 F test, 490–491
 methodology, 488–493
 numerical illustration, 488–490
 power/sample size, 898–904
 SAS, usage, 965–966
- Ordinal level (measurement), 9
 Ordinal logistic regression, 714, 980–981
 ordinal categories/one dichotomous exposure variable, 727–731
 overview, 726
 problems, 738–742
 real data usage (four ordinal categories/three predictor variables), 731–737
- Ordinal outcome, dichotomization (disadvantage), 715
 Orthogonal contrasts, 516–521
 usage, example, 519–521
 Orthogonality, meaning, 632–633
 Orthogonal polynomials, 412–422
 coefficients, 932t
 model of order k , partial F test, 414
 transformation, 414–417
 usage, 417–422
 values, example, 416t
 variables, information, 414
- Outcome variables
 treatment, 727
 usage, 532, 716
- Outliers, 18, 59
 Cook's distance values, inclusion, 351t
 definition, 348
 detection, 347, 348–351
 statistics (computation), SAS (usage), 350
 example, 380t
 jackknife residual values, inclusion, 351t
 leverage values, inclusion, 351t
 removal, example, 403–404
- Output Delivery System (ODS), 951
- Overall confidence level, 504, 508
- Overall mean (coefficient μ), 562
- Overall regression, test, 166–167
- Overparameterized model, 363
- Paired-difference t test, 550–551
 usage, 552, 567–568
- Pair-matched analysis, conditional likelihood (usage), 702
- Pair-matched case-control data
 example, 705t
 usage, 704–705
 model, output (PROC LOGISTIC), 706t
 studies, 701–702, 908
 design, 908–909
- Pair-matched data, unmatched covariates (combination), 704–707
- Pair matching, usage, 549
- Pairwise comparisons
 computation, 531
 consideration, 514
- Pairwise correlations
 calculation, 342
 decrease, 797
- Pairwise Scheffé method confidence intervals, 512
- Parabola, 401
 fit, ANOVA table (example), 404t
 fitting, least-squares procedure, 402–404
- Parabolic curve, appropriateness, 407
 Parabolic model, assumption, 295–296
- Parabolic relationship, regression/strength (test), 405
- Parallelism
 assumption, 319
 drawback, 313
 test, 315
 testing, 263–265, 282
 single-model approach, 268–269
- Parallel lines, tests, 271
- Parameters, 16, 55
 vector, 946
- Partial correlations, 199, 200
 coefficient, 201, 204–211
 features, summary, 210–211
 data, 205t
 description, 208–209
 determination, 215, 221
 test of significance, 207
- Partial F statistic
 determination, 447–448
 form, 177–178
 significance, test, 450
 usage, 211, 216
- Partial F test, 167–173
 acceptance, 316
 application, 171–172
 comments, 171–172
 partial correlation, test (relationship), 207–208
 principles, 176–178
 procedure, 169–170
 P -value, 170
- SAS output data, 176
 source tables, models, 179–180
 strategies, commentary, 178–179
 t test, alternative, 170–171
 usage, strategies, 175–180
- Partial regression, plots, 341–342
 examples, 379f
 scatterplot, example, 345f, 346f
- PASS 11 software, 906–908
- Path analysis, 44
- Path diagram, 43–44
- Pearson correlation, 341
 SAS output (PROC CORR), 345
- Perfect collinearity, 362–363
 overparameterized model, 363
- Per-group sample sizes, 901
- Persuasion scores, 578
- Plane, 139
- Point-and-click interfaces, usage, 911
- Point-binomial populations (Bernoulli populations), 885
- Point-binomial random variable (Bernoulli random variable), 698–699
- Point estimates, 55

Poisson ANOVA table, 761t
 Poisson distribution, 743–745
 application, 744
 appropriateness, 769
 formula, usage, 745
 Poisson model, assumption, 775
 Poisson probability distribution, 743–744
 Poisson regression, 748–753, 781
 ANOVA table, example, 767, 770
 application, 748
 computer package, usage, 752
 considerations, 748–749
 development, 746–747
 example, 745–748, 751–753,
 756–761
 model, 768, 773
 SAS, usage, 981–982
 theoretical considerations, 749–751
 Poisson regression analysis, 743
 data, analyses (summary), 764t
 example, 762–765
 framework, 748–749
 incidence, comparison, 746t
 problems, 766–780
 utility, 745
 values, example, 759t
 Political reasons, impact, 465
 Polynomial models, 401, 402
 higher-order polynomial models,
 fitting/testing, 410
 predictor correlation matrices,
 eigenvalues, 420t
 selection strategies, 422–423
 Polynomial regression, 401
 problems, 423–437
 second-order polynomial
 regression, 404–405
 Polynomial, term (usage), 402
 Polynomial values, example, 415t
 Polytomous logistic regression,
 714, 980
 example (one predictor/three
 outcome categories), 715–721
 extension, example, 721–726
 model
 fitting, 719
 usage, 721
 problems, 738–742
 PROC LOGISTIC, 718, 723–724
 usage, example, 741
 Polytomous outcome
 data, 727t
 dichotomization, disadvantage, 715
 Pooled sample variance, 26
 Population, 16
 correlation coefficient, 109–110
 covariance (σ_{XY}), 110
 Positive quadrants, 110
 Power, 35–37

Power of the test, 35
 Power-ordered fitting, example, 460t
 Predicted response
 variance, 70
 vector, 948
 Prediction
 bands, 70, 962–963
 example, 68f
 equation, usage, 456
 intervals
 α level, change, 963
 sample, 64t
 SAS output, 69
 question, 438–439
 Predictors
 assessment, 187
 correlation matrix eigenvalues, 419
 examples, 420t
 correlations, example, 419t
 effect, evaluation, 693
 factors, identification, 873
 intercorrelations, 459
 polytomous logistic model,
 extension, 721–726
 principal component analysis, 365
 testing, fixed order, 472
 usage, 442–444, 843t, 845t
 values, variability amount
 (presence), 442
 Predictor variables, 1, 10–11
 computer output, example,
 473–476
 identification, 384
 linear relationships, 339
 perfect linear combination, 361
 usage, 174
 variables-added-last tests, 393, 396
 Principal component analysis, 365
 Principal components, regression, 372
 Probability distribution, 16
 random variable, 51
 Probability pattern, 17
 Product terms, coefficients, 251
 Proportional cell frequencies, 630
 allocation, 636
 Proportional odds model (cumulative
 logit model), 726–728
 collapsed 2x2 tables, 728t, 733t
 usage, 731, 734
 Proportionate reduction, 115–116
 Proportions comparison
 sample size calculations, usage,
 884–886
 tests, sample size determination,
 885–886
 Prospective Evaluation of Radial
 Keratotomy (PERK) study,
 examples, 390, 479–480, 539
 Pseudo- R^2 -value, calculation, 767
 Pure-error estimates, 411t
 Pure, term (usage), 580
 P -value, 32, 406
 estimation, 34
 examples, 32f, 34f
 significance level, comparison, 448
 size, approximation, 33
 P -values, basis, 371
 P -value, usage, 167, 745
 Quadratic model, regression ANOVA
 table, 408t
 Quadratic regression
 estimated equation, 423
 determination, 433, 436
 F tests, 426
 significance, 437
 Quantitative formula, usage, 41
 Quantitative Graduate Record
 Examination (QGRE)
 score, 32
 example, 28, 30
 variable score, 31
 Quartiles, 18–19
 Quasi-experimental research, 1–2
 Quasi-likelihood estimation, 861–862
 R (software), 57
 r^2 (square of the sample correlation
 coefficient)
 nonmeasurement, 115–116
 size, determination, 123–125
 straight-line relationship, strength,
 113–115
 value
 characteristics, problem, 443
 determination, example,
 155–156, 163
 increase, 409
 Random effects, 825–829
 assumptions, 863–864
 discrete outcomes, data analysis,
 861–862
 hypotheses, testing, 851–855
 impact, 833–837
 measurement data, analysis,
 839–859
 problems, 862–882
 results, 829–839
 unequal cells (unbalanced data),
 972
 usage, 848t
 reasons, 839t
 Random-effects factor, impact, 537
 Random-effects model, 971–972,
 987–990
 usage, 861
 writing, 619
 Random-effects one-way ANOVA
 model, 500
 Random effects, tests, 586

- Random-effects three-way ANOVA model, 846–847
- Random-effects two-way ANOVA models, 607
- Random errors, vector, 946
- Random factors, 826, 840, 871 examples, 483t
fixed factors, contrast, 483–484
- Random intercept, 987–988 effect, 846 heterogeneity, 827f, 828f model, 832, 844–846, 857 random effects, impact, 833–837 subject-specific model, 830–833
- Randomized blocks, 545 analysis, 570
- ANOVA conducting, 573–574 partitioning, 551 PASS 11, usage, 906 sample size, examples, 904–906, 906–907
- design assumption, 572 usage, 553 formation, steps, 554f
- F* test, 551–553 usage, 552 layout, 562 cell means, matrix, 563t
- PASS 11 procedure/software, 906, 907f
- Randomized blocks, study analysis, 555–557 ANOVA table, 557–561 *k* treatments/*b* blocks, impact, 557t
- fixed-effects ANOVA model, usage, 565–566 regression models, usage, 561–564
- Random model, 583–586
- Random variables, 17, 55 distribution, 16 relationship, 19–24
- Rank-ordering, 172
- Rate discrepancy, data, 571–572
- Rate ratio (RR), 743
- Ratio variable (ratio-scale variable), 10
- Reduced model (*R*), 443, 672
- Reference cell coding, 564, 956 example, 496 model, 496–497 usage, 597
- Regression, 2 ANOVA, contrast, 482 ANOVA table, 408t usage, 643 assumptions, 264 violations, strategies, 355–358
- confounding, 226 problems, 243–256 usage, 236–241
- correlation, connection (extension), 200 data, example, 286–287 dummy variables, 257 fits, usage, 263–268
- F* tests, conducting, 192–193, 195 interaction, 226, 228–236 examples, 228–230, 232–235 problems, 243–256
- jackknife residual plot, example, 353f
- least-squares equation, sketching, 102 observed/predicted values, 202t procedures, 445–447, 465 preference, 445 results, summary, 446t relationships comparison, 42 examination, 189–190
- residuals correlation, partial correlation, 209 examination, 347–348 results, usage (example), 155, 254
- SAS output, examples, 148–151 scalar form, 803
- single independent variable problem, 47–48 questions, 48 strategy, 48–50 usage, 47–50
- standard errors (estimation), regression parameter estimates (impact), 368
- terminology, usage, 228 test, 166–167, 405
- Regression analysis, 3, 41, 227 assumption, variance homoscedasticity, 424 definitions, 937–938 description, 242–244
- F* tests, involvement, 165–166 fundamental equation, 132 matrices formulation, 946–948 relationship, 937 usage, 938–939
- orthogonal polynomials, usage, 417–422
- variables, consideration, 245
- Regression coefficients confidence intervals, 182, 959–960 defined dummy variables, 633–634 estimation/estimates, 137, 358, 360 *k* estimated variances, 185 function, 596
- hypothesis tests, 959 linear functions, inference, 960–961 methods, 184–186
- ML estimates, 689 obtaining, 42 terms, 330 variances, 948 estimates, 363
- Regression diagnostics, 339 analysis, 347–348 data example, 373t problems, diagnosis (approaches), 340–347 example, 372–382 intercept, treatment, 364 model assumptions, violations, 347–355 problems, 382–398
- SAS (software) output, 374–377 usage, 973–976 scatterplots, examples, 344f, 378f–379f statistics, 379–380 steps, 340–342
- Regression equations coincidence, 278, 279 comparison, 275–277 defining, 138 examples, 52f, 293–295 formation, 51 nominal variables, involvement, 277–283 data, 281t examples, 280–283 parallel comparison, 276 parallelism, 279 selection, 438 problems, 466–480
- Regression line differences, 282–283 estimate, sketching, 96 least-squares estimates (recomputation), 95 parameters, least-squares estimates (calculation), 101 slope (magnitude measure), *r*² (relationship), 115–116
- Regression model appropriateness, 648 creation, dummy variables (usage), 522 dummy variables, coefficients, 494 effects, presence, 679–680 expansion, 573 fixed-effects two-way ANOVA, 594–599 identification, 298

- Regression model (*continued*)
 linearization, 356
 parallelism/coincidence, test, 289
 parameter estimates, 359t
 representation, alternative, 212
 scalar form, 792
 term, interaction method, 600
 usage, 561–564
- Regression sum of squares
 (SSY – SSE), 130, 145, 948
 components, 168t
 computations, data, 180t
 equation, 146, 559
- Regressor variable, 10–11
- Rejection region (critical region), 31
- Relative extrema, considerations, 410
- Relative risk, 745–746
- Reliability, assessment, 455
- Repeatability/reproducibility, terms (usage), 526
- Repeated measures analysis, data arrangement, 840
- Repeated measures ANOVA, 847–851
 ANOVA table, 851t
- Repeated measures data, 781, 801
 data layout, 791t, 841t
 example, 790t
- Replicates, 411t
- Research
 classification, 1
 examples, 2–5
- Residual option, usage, 857
- Residuals
 analysis, 347–355
 graphical analyses, usage, 352–353
 jackknife residuals, 348
 mean-square errors, combination, 264
 plot, indication, 353
 probability plots, 354
 SAS output (PROC UNIVARIATE), 355
 standardized residuals, 348
 studentized residuals, 348
 time, contrast, 352–353
 unstandardized residuals, 463
 values, determination, 385, 390, 393, 396
 variance, 352
 variation, proportion (determination), 217
- Residual sum of squares (residual SSE), 56, 144, 145
 data, 179t, 180t
 denotation, 181
 equation, 146, 559
 partitioning, 411–412
- Response, predictors (relationship), 440
- Response variable, 1, 10–11
- Response vector, 782
- Restricted maximum likelihood (REML), 806
- Reverse interaction, 604–605
 layouts, 604t
- ρ (rho), confidence interval, 119–120
- Ridge regression, 372
- Risk factors, 238
- Risk ratio, 745–746
- Robust estimator (empirical estimator), 805
- Robustness, 485
- Robust variance estimation (empirical variance estimation), 799–800
- Row factor, 545
 effects, 598
- Row main effects, 601–602
- Row marginal means, 597
- Rows and columns
 fixed, 593
 random, 593
- R-square value, square root, 109
- Same-direction interaction, 602–604
- Sample, 16
 observations, 58t
- Sample proportion, 663
- Sample size, 35–37
 calculation, 893–907, 913
 formula, 36
 minimum, 893
 planning, 883
 relationships, 886
 range, 930t
- Sample standard deviation, 17
- Sample variance, 17
- Sampling distributions, 25–27
- SAS (software), 57
 coding, conventions, 950
 computer output, 232–233
 confidence/prediction interval output, 69
 data output, 59
 GLM procedure, 489
 MIXED procedure, 141
 output, 109
 style, 951
 PROC GLM, 71, 644–647
 PROC MEANS, 71–72
 PROC REG, usage, 57, 68, 71
 PROC UNIVARIATE, 342
 syntax/structure, 949–950
- Satterthwaite option, 858
 recommendation, 858
- Scalar form, 792
- Scaling, 364–365
 centering, 365
 problems, avoidance, 416–417
- Scatter diagram, 47, 78
 evaluation, 110
- examples, 48f, 111f, 140f
 observation, example, 75
- Scatterplots
 examples, 72f, 138f, 344f, 347f
 partial regression plot, example, 345f, 346f
 usefulness, 341
- Scheffé confidence intervals, comparison, 513t
- Scheffé method, 509–515
 advantage, 512
 pairwise Scheffé method confidence intervals, 512
 usage, 534
- Scheffé multiple-comparison procedures, 522, 529
- Score test, 733–734
 design, 728–729
- Secondary objective, 788
- Second-degree model, goodness of fit (assessment), 674–675
- Second-order partial, determination, 216
- Second-order polynomial fit, LOF test (regression ANOVA table), 412t
- Second-order polynomial regression adequacy, testing, 406
 ANOVA table, usage, 404
 data, scatter diagram, 407f
 inferences, 405–406
 model, X^2 term addition (term), 405–406
 quadratic model, regression ANOVA table, 408t
 requirement, example, 406–410
- Second-order variables, multiple partial correlation, 213
- Semipartial correlations, 209–210
 zero-order correlations, 210
- Sensitivity analyses, 910–911
- Seventh-order orthogonal polynomial model, 421t
- Shrinkage on cross-validation, 456
- σ^2 , estimate, 60–61
- Significance level
 alpha (Type I error), 31
 P-value, comparison, 448
 repair, 504
- Simple linear regression (sample size determination)
 binary predictor, usage, 887
 continuous predictor, usage, 888
- Single independent variable
 scatterplot, 138f
 usage, 47–50
- Single regression
 equation, usage, 268–271
 model, identification, 301–302, 304

- Single variable, addition (test), 165
 Single-variable selection methods, chunkwise methods (contrast), 454
 Skewness, statistics, 354
 Slope, 50
 inferences, 61–64
 least-squares estimates, 264
 determination, 79, 82, 86
 tests. *See* Zero slope.
 interpretations, 64–66
 Social influence activity, measures, 3
 Split-sample analysis, 455
 Split-sample approach, 472
 SPSS (software), 57
 Spurious correlation, 216–217
 Squared correlation coefficient (r^2), 199
 Squared multiple correlation, 895
 Squared multiple partial correlation, determination, 218, 219, 221–225
 Square of partial correlation, measurement, 210
 Square root transformation, 356
 Square transformation, 356
 Standard errors
 empirical standard errors, usage, 811–812
 estimate, 67
 Standard errors of regression
 parameter estimates, predictor variables linear associations (impact), 368
 Standardization, 318
 methods, 337
 Standardized residuals, 348
 Standardized scores (z scores), computation, 365
 Standard normal cumulative probabilities, 916t–918t
 Standard stepwise regression program (SPSS)
 output, 246
 usage, example, 296–298
 usage, 245
 STATA (software), 57
 Statement of assumptions, 51–55
 summary/comments, 55
 Stationary m -dependent correlation structure (Toeplitz m -dependent correlation structure), 860
 Stationary 1-dependent structure, 860
 Stationary structure (Toeplitz structure), 860
 Stationary 2-dependent correlation structure, 860
 Statistical independence, 592
 Statistical inference, 27–34, 958–963
 maximum likelihood, usage, 665–677
 mean value, 961
 Statistically significant association, finding, 43
 Statistical methodology, description, 50
 Statistical models, 402
 deterministic models (contrast), 45
 Statistical null hypothesis, testing procedure, 30
 Statistical tests
 performing, 63
 usage, 238
 Statistics, 16
 problems, 37–40
 Stem-and-leaf diagram, 18f
 Stepwise regression approach, usage, 468
 Stepwise regression procedure, 452
 Stepwise variable selection decision-making algorithms, 452–453
 Straight-line fits, 310–311
 basis, 287–288
 coincidence, testing, 267–268
 example, 60
 quality, measurement, 60–61
 Straight-line model
 adequacy, 428
 adequacy of fit, 424
 appropriateness, assessment, 70–71
 appropriateness, r^2 (nonmeasurement), 116f
 assumption, examination, 49
 defining, 299
 extension, 401
 fit, 386
 implications, 66
 plotting, 429
 significance, 176–177
 statistical assumptions, 51–55
 Straight-line regression
 ANOVA table, 129–133
 computer output, 426
 comparison, 235f, 260f
 conclusions, 261f
 equations, comparison, 233t
 example, 259–260
 estimate, 79
 estimated equation, 423
 determination, 433, 436
 fits, comparison, 321–322
 F tests, 426
 model
 fitting, 125, 168
 lack of fit, test, 520
 relationship, basis, 126
 residuals, correlation, 209
 response, estimated value (computation), 182–183
 results
 display, 284–285
 results, comparison, 260t
 SAS, usage, 951
 separation, 140
 significance, 429, 437
 slope/intercept, least-squares estimates (determination), 89, 100
 weighted least-squares solution, 357
 Straight-line regression analysis, 5
 assumptions, impact, 121–122
 prediction intervals, equations, 63t
 problems, 74–106, 125–127
 Straight lines
 assumption, example, 53f, 311
 coincidence, testing, 291
 comparison
 backward testing strategy, 273f
 methods, 262–263
 questions, 261–262
 regression fits, usage, 263–268
 strategies/interpretation, testing, 272–273
 equations, usage, 305
 intercepts, comparison, 265–267
 example, 266–267
 mathematical model, 53
 mathematical properties, 50–51
 model (appropriateness), r^2
 nonmeasurement, 116
 parallelism, testing, 263–265, 291
 example, 265
 plots, 50f
 relationship, 309
 single regression equation, usage, 268–271
 Stratified methods, 337–338
 Stratified random sample, usage, 647
 Stratifying, 579
 Strength, test, 405
 Structural equation modeling, 44
 Studentized range, upper α point, 930t–931t
 Studentized residuals, 348
 plot, examination, 383
 plotting, 432
 probability plots, 354
 Student's t distribution, 24–27
 example, 25f
 $n - 1$ degrees of freedom, 25
 $n - k - 1$ degrees of freedom, 348
 Study (exposure) variable, 686
 Subgroups, identification, 788
 Subject-specific analyses, summary, 837–839
 Subject-specific comparison, 788–789
 Subject-specific matrix form, 793–794, 804–806, 820
 equation, 834–835

- Subject-specific models, 831–833
 covariance structure, 835
 fitting, 839
 statement, 875
- Subject-specific random
 intercept, 864
- Subject-specific response vector, 804
- Subject-specific scalar form, 792, 795, 804, 820, 846
 equation, 840
- Subject-specific scalar model, modification, 865
- Sufficient component cause (SCC) model, 43–44
- Sum of squares
 collection, 633
 degrees of freedom,
 correspondence, 552, 591
 equations, 408–409
 results, 154
- Sum of squares about regression, 144
- Sum of squares due to error (SSE), 56, 145, 443
- Sum of squares due to/explained by regression ($SSY - SSE$), 130
- Symmetric matrix, 940
- Table of cell means, usage, 581–586
- t* distribution, 61–62
 percentiles, 919t
 usage, 524
- t* distribution-based hypothesis tests, provision, 960
- Temporal ambiguity, absence, 44
- Testing, 2
- Test of coincidence, single-model approach, 271
- Test of significance, 187
 partial correlations, 207
- Test power, 35
- Tests, data-set-specific series, 36–37
- Tests of hypotheses, 276
- Test statistic
 computation, 187, 269
 distribution, 36f
 usage, 62
 value, obtaining, 32
- Third-order natural polynomial model, examples, 417t, 418t
- Third-order orthogonal polynomial model, fit, 414
- Third-order polynomial model, 414
- Three-dimensional data, best-fitting plane, 139f
- Three-dimensional space, best-fitting surface, 139
- Three-factor interaction effects, 850
- Three-level nominal variable, 846–847
- Three-way ANOVA
 table, 642t
 total sums of squares, partitioning, 850f
- Three-way interaction effect, 878
- Toepelitz m -dependent correlation structure (stationary m -dependent correlation structure), 860
- Toeplitz structure (stationary structure), 860
- Total sum of squares (SST), 147
 equation, 146
 measurement, 558
- Total sum of squares about/corrected for the mean (SSY), 130, 145
 decomposition, 634–635
 measurement, 558
- Total uncorrected sum of squares, 132
- Total unexplained variation (SSY), 130
- Transformations, 356
 arcsin transformation, 356
 log transformation, 356
 square root transformation, 356
 square transformation, 356
- Treatment combinations, 549 formation, 579
- Trend, assessment, 519–521
- Trend test (linear test), 693–694
- Trivariate normal distribution, 203
- True slope, null hypothesis (test), 84, 102, 104
- t* statistic, square, 170–171
- t* test
 alternative, 170–171
 technique, usage, 221
- Tukey–Kramer confidence intervals comparison, 513t
 narrowness, 509
- Tukey–Kramer method, 507–509
 preference, 515
 results, ambiguity, 508
 usage, 537, 539
- Tukey–Kramer multiple-comparison procedures, 522
- Tukey–Kramer pairwise confidence interval, 508
- Tukey's test for additivity, 569–570
- Two-dimensional representation, 138
- Two-dimensional space, curve (representation), 139
- Two-factor crossover design, 840
- Two-factor pattern, example, 545–546
- Two-factor summary tables, formation, 649
- Two-sample *t* test, 887, 911
 statistic, generalization, 490
- Two-stage stratified random sample, usage, 689–690
- Two-tailed alternative, 118, 122
- Two-tailed *P*-values, 853
- Two-way analysis of variance (two-way ANOVA), 4–5
 ANOVA table, 589–591
 assumptions, 592–594
 blocking principle, 553–554
 cell frequencies, demonstration, 650
 cell means, sample, 587t, 605t
 consideration, 483
 dependent variable, usage, 658
 equal cell numbers, 579
 cell means, 589t
 data, presentation, 586–589
 example, 588t
 methodology, 586–591
 problems, 610–628
- equal observations per cell, data layout, 587t
- expected mean squares, 609t
- flow diagram, 636f
- F* tests, 592–594
- interactions, 599–606
 concept, 599–601
 effects, 605–606
 hypothetical examples, 601–605
- matched-pairs experiment, equivalent analysis, 549–553
- models, null hypotheses/expected mean squares, 608–610
- null hypotheses, 592, 609t
- P*-values, 620
- randomized blocks, 545
 problems, 566–578
- regression model, defining, 616
- results, interaction (inclusion), 583–585
- reverse interaction, 602–604
- same-direction interaction, 602–604
 layouts, 603t
- significance patterns, equal cell numbers, 594t
- significance tests, determination, 583
- table, general (balanced) two-way ANOVA table, 590t
- treatment combinations, 549
- two-way layout, 595t
- unequal cell numbers, 630
 data, presentation, 630–632
 nonorthogonality, 632–636
 problems, 642–660
 usage, 655, 656
- Two-way data patterns, 545–547
 examples, 546f
- Two-way fixed-effects ANOVA model, 598–599, 969
- Two-way layouts, equal cell numbers (usage), 579
- Two-way tables, result, 548f
- Type I error, 35, 36
 rate, 886

- Type II error, 35, 36
rate, 886
- Type III sums of squares, usage, 188
- Unadjusted means, BRFSS data, 318
- Unbalanced patterns, 630
- Unbalanced two-way ANOVA data analysis, regression approach (example), 640–641
regression analysis, flow diagram, 639f
- Unconditional likelihood function, 699–700
- Unconditional ML estimation/ procedure, 698–700
- Unconditional sums of squares, 633
- Underfitting bias, avoidance, 422–423
- Underlying correlation structure, 798
- Unequal cell numbers, 630
data layout, 631t
experimental studies usage, 632
nonorthogonality, 632–636
problems, 642–660
regression model, usage, 637
- Unequal cell sample sizes, regression approach, 637–641
- Univariate correlation, computation, 456
- Univariate normal distribution, 203–204
generalization, 112
- Unmatched covariates, usage, 704–707
- Unstandardized residuals, 463
- Unstructured correlation matrix, usage, 859–860
- Unstructured correlation structure, 798
- Unweighted least-squares, 357
- Upper confidence limits, yield, 698
- User-defined contrast, 901
- Validity-based strategy, 463–464
- Validity, question, 438–439
- Variability (dispersion), measure, 17
- Variables
addition, significance, 216
classification, 7
overlap, 12f
control, consideration, 238
descriptive statistical analysis, 340
descriptive statistics, examination, 341
effects, control, 41–42
fixed value, 51
- group, addition (test), 165
ordering, 10
pair, example, 87
prediction, 114
random sample, 137t
relationship, 211t
assessment, 2
removal, 371
selection, 206
strategy, specification, 444–454
validity-based strategy, 463–464
- sets, identification, 240
- types, 1
- usage, example, 293–294
- value pairs, perfectly collinear set, 361f
- Variables-added-in-order regression models, 180t
- Variables-added-in-order squared partial correlations, 206
- Variables-added-in-order sum of squares, 418, 952
characteristic, 178
- Variables-added-in-order testing approach, 178
- Variables-added-in-order tests, 175, 270t
conducting, 433, 435
method, 176–177
usage, 197–198, 432
- Variables-added-last-regression models, 179t
- Variables-added-last statistic, 272
- Variables-added-last sum of squares, 418, 952
- Variables-added-last tests
dummy variable consideration, 270
performing, 393, 395
usage, 197–198
- Variables-added-last-tests, 175
ANOVA table, 178, 179t
method, 177
- Variance components (VC) model, default, 855
- Variance inflation factor (VIF)
calculations, 368
determination, 390, 393, 396, 435
usage, 363
values, 371
- Variances
estimation, 795
homogeneity, 515–516
assumption, 592
homoscedasticity, 424
proportion, 366
- statistics, examination, 368–369
 σ^2 value, estimate, 551
three-way analysis, 4
two-way analysis, 4–5
- Variances components, 849–850
- Variation, straight-line regression (explanation/nonexplanation), 131f
- Wald chi-square statistics, calculation, 725
- Wald chi-square test, usage, 737, 890
- Wald P-values, 726t
- Wald statistic
formula, 710
LR statistic, discrepancy, 673–674
testing, 720
usage, 668–669, 707
- Wald tests
procedures, 853
usage/description, 693–694, 710, 772
- Weighted least-squares analysis, 352, 357–358
- Within-group standard deviation, 901
- Within-subject variability, 789, 849–850
- Working correlation structure, 799, 818
- X (specified value), 66–68
- X^2 term, addition (test), 405–406
- X -values, distinctiveness, 410
- Y (straight line), determination/drawing, 92
- Yield, regression (results), 254
- Y -intercept, 50
estimate, impact, 73
- Y , usage (predictions), 115f
- Y -values, 52
- Y (mean value), X (specified value), 66–68
- Y (new value prediction), X_0 value, 68–70
- Zero intercept, test, 66
- Zero-order correlations, 201
matrix, 205
- Zero slope
null hypothesis, test, 79, 83
test, 65–66
interpretation, 65f
- Z -statistic, yield, 756

