

SPRINGER TEXTS IN STATISTICS

# Statistical Analysis and Data Display

## An Intermediate Course with Examples in S-PLUS, R, and SAS



Richard M. Heiberger  
Burt Holland



Springer

## *Springer Texts in Statistics*

*Advisors:*

George Casella Stephen Fienberg Ingram Olkin

# Springer Texts in Statistics

---

|                                   |  |
|-----------------------------------|--|
| <i>Alfred Berger</i>              | Elements of Statistics for the Life and Social Sciences  |
| <i>Bilodeau and Brenner</i>       | Introduction to Probability and Stochastic Processes, Second Edition   |
| <i>Blom</i>                       | Theory of Multivariate Statistics  |
| <i>Brockwell and Davis</i>        | Probability and Statistics: Theory and Applications  |
| <i>Carmona</i>                    | Introduction to Time Series and Forecasting, Second Edition  |
| <i>Chow and Teicher</i>           | Statistical Analysis of Financial Data in S-Plus   |
|                                   | Probability Theory: Independence, Interchangeability, Martingales, Third Edition   |
| <i>Christensen</i>                | Advanced Linear Modeling: Multivariate, Times Series, and Spatial Data; Nonparametric Regression and Response Surface Maximization, Second Edition |
| <i>Christensen</i>                | Log-Linear Models and Logistic Regression, Second Edition  |
| <i>Christensen</i>                | Plane Answers to Complex Questions: The Theory of Linear Models, Second Edition  |
| <i>Creighton</i>                  | A First Course in Probability Models and Statistical Inference   |
| <i>Davis</i>                      | Statistical Methods for the Analysis of Repeated Measurements  |
| <i>Dean and Voss</i>              | Design and Analysis of Experiments   |
| <i>du Toit, Steyn, and Stumpf</i> | Graphical Exploratory Data Analysis  |
| <i>Durrett</i>                    | Essential of Stochastic Processes  |
| <i>Edwards</i>                    | Introduction to Graphical Modeling, Second Edition   |
| <i>Finkelstein and Levin</i>      | Statistics for Lawyers   |
| <i>Flury</i>                      | A First Course in Multivariate Statistics  |
| <i>Heiberger and Holland</i>      | Statistical Analysis and Data Display: An Intermediate Course with Examples in S-PLUS, R, and SAS  |
| <i>Jobson</i>                     | Applied Multivariate Data Analysis, Volume I:<br>Regression and Experimental Design  |
| <i>Jobson</i>                     | Applied Multivariate Data Analysis, Volume II:<br>Categorical and Multivariate Methods   |
| <i>Kalbfleisch</i>                | Probability and Statistical Inference, Volume I:<br>Probability, Second Edition  |
| <i>Kalbfleisch</i>                | Probability and Statistical Inference, Volume II:<br>Statistical Interference, Second Edition  |
| <i>Karr</i>                       | Probability  |
| <i>Keyfitz</i>                    | Applied Mathematical Demography, Second Edition  |
| <i>Kiefer</i>                     | Introduction to Statistical Inference  |
| <i>Kokoska and Nevison</i>        | Statistical Tables and Formulae  |
| <i>Kulkarni</i>                   | Modeling, Analysis, Design, and Control of Stochastic Systems  |
| <i>Lange</i>                      | Applied Probability  |

*Continued after index*

Richard M. Heiberger      Burt Holland

# Statistical Analysis and Data Display

An Intermediate Course with Examples  
in S-PLUS, R, and SAS

With 200 Figures

Richard M. Heiberger  
Department of Statistics  
Temple University  
Philadelphia, PA 19122  
USA  
rmh@temple.edu

Burt Holland  
Department of Statistics  
Temple University  
Philadelphia, PA 19122  
USA  
bholland@temple.edu

*Editorial Board*

George Casella  
Department of Statistics  
University of Florida  
Gainesville, FL 32611-8545  
USA

Stephen Fienberg  
Department of Statistics  
Carnegie Mellon University  
Pittsburgh, PA 15213-3890  
USA

Ingram Olkin  
Department of Statistics  
Stanford University  
Stanford, CA 94305  
USA

*Cover illustration:* Cover art is a variation of Figure 14.14d. The data source is (Williams, 2001).

Cygwin: Copyright © 1996, 1998, 2001, 2003 Free Software Foundation, Inc.

EMACS: Copyright © 1989, 1991 Free Software Foundation, Inc.

Excel: Copyright © 1985–1999, Microsoft Corp.

Ghostscript: Copyright © 1994, 1995, 1997, 1998, 1999, 2000 Aladdin Enterprises, Menlo Park, California, U.S.A. All rights reserved.

GSview: Copyright © 1993–2001 Ghostgum Software Pty Ltd.

Internet Explorer: Copyright © 1995–2001 Microsoft Corp.

Linux: Copyright © 2004, Eklektix, Inc.

LogXact: Copyright © Cytel Software Corporation

MathType: Copyright © 1990–1999 Design Science, Inc.

Microsoft Windows: Copyright © 1981–2001 Microsoft Corp.

MiKTeX: Copyright © 1999 Christian Schenk

MS\_DOS: Copyright © 1985–2001 Microsoft Corp.

MS\_Word: Copyright © 1983–1999, Microsoft Corp.

PostScript: Copyright © Adobe Systems Incorporated

R: Copyright © 2002, The R Development Core Team

SAS: Copyright © 2002 by SAS Institute Inc., Cary, NC, USA.

sas.library/code/ischeffe.sas: copyright holder unknown.

S-Plus: Copyright © 1988, 2002 Insightful Corp.

Stata: Copyright © 1984–2002 Stata Corp.

TeX is a trademark of the American Mathematical Society.

Unix: Copyright © 1998 The Open Group

Windows XP: Copyright © 2001 Microsoft Corporation. All rights reserved.

XLISP-STAT 2.1 Copyright © 1990, by Luke Tierney

ISBN 978-1-4419-2320-2 ISBN 978-1-4757-4284-8 (eBook)

DOI 10.1007/978-1-4757-4284-8

© 2004 Springer Science+Business Media New York

Originally published by Springer Science+Business Media Inc. in 2004.

Softcover reprint of the hardcover 1st edition 2004

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher Springer Science+Business Media, LLC , except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

*In loving memory of Mary Morris Heiberger*

*To my family: Margaret, Irene, Andrew, and Ben Holland*

---

# Preface

## 1 Audience

Students seeking master's degrees in applied statistics in the late 1960s and 1970s typically took a year-long sequence in statistical methods. Popular choices of the course textbook in that period prior to the availability of high-speed computing and graphics capability were those authored by Snedecor and Cochran, and Steel and Torrie.

By 1980, the topical coverage in these classics failed to include a great many new and important elementary techniques in the data analyst's toolkit. In order to teach the statistical methods sequence with adequate coverage of topics, it became necessary to draw material from each of four or five text sources. Obviously, such a situation makes life difficult for both students and instructors. In addition, statistics students need to become proficient with at least one high-quality statistical software package.

This book can serve as a standalone text for a contemporary year-long course in statistical methods at a level appropriate for statistics majors at the master's level or other quantitatively oriented disciplines at the doctoral level. The topics include both concepts and techniques developed many years ago and a variety of newer tools not commonly found in textbooks.

This text requires some previous studies of mathematics and statistics. We suggest some basic understanding of calculus including maximization or minimization of functions of one or two variables, and the ability to undertake definite integrations of elementary functions. We recommend acquired knowledge from an earlier statistics course, including a basic understanding

of statistical measures, probability distributions, interval estimation, and hypothesis testing.

## 2 Structure

The book is organized around statistical topics. Each chapter introduces concepts and terminology, develops the rationale for its methods, presents the mathematics and calculations for its methods, and gives examples supported by graphics and computer output, culminating in a writeup of conclusions. Some chapters have greater detail of presentation than others, based on our personal interests and expertise.

Our emphasis on graphical display of data is a distinguishing characteristic of this book. Many of our graphical displays appear here for the first time. Appendix G summarizes those new graphs that are based on Cartesian products. We show graphs, how to construct and interpret them, and how they relate to the tabular outputs that appear automatically when a statistical program “analyzes” a data set. The graphs are not automatic and so must be requested. Gaining an understanding of a data set is always more easily accomplished by looking at appropriately drawn graphs than by examining tabular summaries. In our opinion, graphs are the heart of most statistical analyses; the corresponding tabular results are formal confirmations of our visual impressions. We advanced this point of view in seminars and presentations ((Heiberger, 1998), (Heiberger and Holland, 2002), (Heiberger and Holland, 2003b), and (Heiberger and Holland, 2003a)) and so have others, for example (Gelman et al., 2002). A vivid demonstration of it appears in Section 4.2.

We have chosen to work with both of what we believe are the two leading statistical languages available today: S (available as both S-PLUS and R), and SAS. S is an exceptionally well-developed tool for statistical research and analysis, that is for exploring and designing new techniques of analysis, as well as for analysis. S is especially strong for statistical graphics, the output of data analysis through which both the raw data and the results are displayed for the analyst and the client. SAS is the most widely used package for serious and extensive statistical analysis and data management. Because of our heavy use of graphics as an essential part of most analyses, we make somewhat heavier use of S than SAS. We frequently mention the package name S-PLUS, rather than the language name S, in situations where S-PLUS and R could equally well be used.

Although we do not explicitly teach S-PLUS or SAS, we make the reader aware of their powerful capabilities by using them to perform the data anal-

yses we present. Sections B.5 and C.4 contain our currently recommended references for learning S-PLUS and SAS. All S-PLUS and SAS code used in the book appears in the companion online files that readers are expected to download from the Springer website (see Preface Section 3). We anticipate that readers will wish to adapt our code to their own data analyses. The code files used to produce the book's numerous graphs are identified alongside each graph. Readers are encouraged to examine these code files in the online files in order to gain full understanding of what has been plotted.

We believe that a firm control of the language gives the analyst the tools to think about the ideal way to detect and display the information in the data. We focus our presentation on the written command languages, the most flexible descriptors of the statistical techniques. The written languages provide the opportunity for growth and understanding of the underlying techniques. The point-and-click technology is convenient for routine tasks. However, many interesting data analyses are not routine and therefore cannot be accomplished by pointing and clicking the icons provided by the program developers.

### 3 Data and Programs

The data for all examples and exercises in this book, and the sample code in both languages [S (meaning S-PLUS and R) and SAS] for all examples and figures, are provided on the accompanying online files (Heiberger and Holland, 2004b). Occasionally we produce listing (output) files that are too big to include in this text. In such situations we place the complete file in the online files and only excerpts in the text. (The collection of directories and files in the online files is distributed from the Springer web page <http://springeronline.com> as a downloadable zipped file. Search for "Heiberger Holland". We recommend that readers burn a CD of the unzipped directories for reference and copy the entire directory structure to their hard disk for use. See the file **README.HH** on the website for details.)

The filename in the online files is given in the text for every code fragment, function, and macro presented. The code and the PostScript file for every figure in the text is in the online files. Transcripts (\*.st files for S-PLUS, and \*.1st and occasionally \*.log files for SAS) are included for code fragments that produce printed output.

---

The directories are structured by chapter, with three subdirectories for each.

```
chapter/code/  
chapter/transcript/  
chapter/figure/
```

The filename is indicated at the time the example is presented.

In addition, there are several directories not associated with specific chapters.

```
datasets/  
splus.library/  
sas.library/  
software/
```

All datasets are in the `datasets` directory. The `splus.library` and `sas.library` directories contain general utilities and new analysis and display functions. All our code and examples assume that these libraries are attached.

In S-PLUS and R, the libraries are attached by running the `.First` function described in Appendix B. The `.First` function must be customized for the individual computer.

In SAS, the macros are made available by running the file `hh.sas` described in Appendix C. The `hh.sas` file must be customized for the individual computer.

Both customizations are simple and these are the only customizations required. All our functions and input statements are defined relative to the paths defined in these customizations. Once these customizations have been made, all examples in the book work as written, with no changes.

## 4 Software

We include in the Software Appendix A and the (`sftw/code/url.htm`) file the `urls` to the software we recommend:

- S-PLUS, Insightful's implementation of the S language
- R, the GNU-licensed implementation of the S language
- SAS
- Ghostscript/Ghostview for displaying PostScript graphs

- Emacs, the extensible text editor from the Free Software Foundation
- ESS (Emacs Speaks Statistics), an intelligent environment for statistical analysis
- Springer, the online files for this book are distributed from the Springer website
- L<sup>A</sup>T<sub>E</sub>X. We wrote this book in L<sup>A</sup>T<sub>E</sub>X (Lamport, 1986), the best mathematical typesetting software (and the one required by several statistics journals), so we provide the url for that as well.

## 4.1 Microsoft Windows

We include urls for

- Cygwin, an implementation of the Unix shell and other user tools for Microsoft Windows
- Standalone utilities (`gunzip`, `gzip`, `tar`) that work in the MS-DOS prompt window
- `gnuserv` and `ispell`, utilities that work with Emacs
- MathType fonts, for improved appearance of mathematics written in Microsoft Word.

## 4.2 Unix

Most of the software listed above is distributed as part of Unix systems and is probably already available on the Unix system you are using. The statistical programs S-PLUS, R, and SAS, and the ESS interface between Emacs and the statistical software will be needed.

## 5 Exercises

Learning requires that the student work a fair selection of the exercises provided, using, where appropriate, one of the statistical software packages we discuss. Beginning with the exercises in Chapter 5, even when not specifically asked to do so, the student should routinely plot the data in a way that illuminates its structure, and state all assumptions made and discuss their reasonableness.

## Acknowledgments

We are indebted to many people for providing us advice, comments and assistance with this project. Among them are our editor John Kimmel and the production staff at Springer, our colleagues Francis Hsuan and Byron Jones, our current and former students (particularly Paolo Teles who coauthored the paper on which Chapter 18 is based, Kenneth Swartz, and Yuo Guo), and Sara R. Heiberger. Each of gratefully acknowledges the support of a study leave from Temple University. We are also grateful to Insightful Corp. for providing us with current copies of S-PLUS software for ourselves and our student, and to the many professionals who reviewed portions of early drafts of this manuscript.

---

# Contents

|  |            |
|--|------------|
| <b>Preface</b>   | <b>vii</b> |
| 1 Audience . . . . .                                       | vii        |
| 2 Structure . . . . .                                      | viii       |
| 3 Data and Programs . . . . .                              | ix         |
| 4 Software . . . . .                                       | x          |
| 4.1 Microsoft Windows . . . . .                            | xi         |
| 4.2 Unix . . . . .   | xi         |
| 5 Exercises . . . . .                                      | xi         |
| <br>   |            |
| <b>1 Introduction and Motivation</b>                       | <b>1</b>   |
| 1.1 Statistics in Context . . . . .                        | 3          |
| 1.2 Examples of Uses of Statistics . . . . .               | 4          |
| 1.2.1 Investigation of Salary Discrimination . . . . .     | 4          |
| 1.2.2 Measuring Body Fat . . . . .                         | 5          |
| 1.2.3 Minimizing Film Thickness . . . . .                  | 5          |
| 1.2.4 Surveys . . . . .                                    | 6          |
| 1.2.5 Bringing Pharmaceutical Products to Market . . . . . | 6          |
| 1.3 The Rest of the Book . . . . .                         | 7          |
| 1.3.1 Fundamentals . . . . .                               | 7          |
| 1.3.2 Linear Models . . . . .                              | 7          |
| 1.3.3 Other Techniques . . . . .                           | 9          |
| 1.3.4 New Graphical Display Techniques . . . . .           | 9          |
| <br>   |            |
| <b>2 Data and Statistics</b>                               | <b>11</b>  |
| 2.1 Types of Data . . . . .                                | 11         |
| 2.2 Data Display and Calculation . . . . .                 | 12         |
| 2.2.1 Presentation . . . . .                               | 13         |

|          |   |           |
|----------|---|-----------|
| 2.2.2    | Rounding . . . . .  | 13        |
| 2.3      | Importing Data . . . . .  | 14        |
| 2.3.1    | S-PLUS . . . . .  | 14        |
| 2.3.2    | SAS . . . . .   | 15        |
| 2.3.3    | Data Rearrangement . . . . .  | 15        |
| 2.4      | Analysis with Missing Data . . . . .  | 16        |
| 2.4.1    | Missing Data in S-PLUS . . . . .  | 16        |
| 2.4.2    | Missing Data in SAS . . . . .   | 17        |
| 2.5      | Tables and Graphs . . . . .   | 17        |
| 2.6      | Files for <i>Statistical Analysis and Data Display</i> (HH) . . . . .                                   | 18        |
| 2.6.1    | Datasets . . . . .  | 18        |
| 2.6.2    | Code, Transcripts, and Figures . . . . .  | 18        |
| 2.6.3    | Functions and Macros . . . . .  | 19        |
| 2.6.4    | Software . . . . .  | 19        |
| <b>3</b> | <b>Statistics Concepts</b> . . . . .  | <b>21</b> |
| 3.1      | A Brief Introduction to Probability . . . . .   | 21        |
| 3.2      | Random Variables and Probability Distributions . . . . .  | 22        |
| 3.2.1    | Discrete Versus Continuous Probability Distributions . . . . .  | 23        |
| 3.2.2    | Displaying Probability Distributions . . . . .  | 24        |
| 3.3      | Concepts That Are Used When Discussing Distributions . . . . .  | 27        |
| 3.3.1    | Expectation and Variance of Random Variables . . . . .  | 27        |
| 3.3.2    | Median of Random Variables . . . . .  | 28        |
| 3.3.3    | Symmetric and Skewed Distributions . . . . .  | 28        |
| 3.3.4    | Displays of Univariate Data . . . . .   | 30        |
| 3.3.5    | Multivariate Distributions—Covariance and Correlation . . . . .   | 34        |
| 3.4      | Three Probability Distributions . . . . .   | 37        |
| 3.4.1    | The Binomial Distribution . . . . .   | 37        |
| 3.4.2    | The Normal Distribution . . . . .   | 38        |
| 3.4.3    | The (Student's) <i>t</i> Distribution . . . . .   | 39        |
| 3.5      | Sampling Distributions . . . . .  | 40        |
| 3.6      | Estimation . . . . .  | 41        |
| 3.6.1    | Statistical Models . . . . .  | 41        |
| 3.6.2    | Point and Interval Estimators . . . . .   | 42        |
| 3.6.3    | Criteria for Point Estimators . . . . .   | 42        |
| 3.6.4    | Confidence Interval Estimation . . . . .  | 43        |
| 3.6.5    | Example—Confidence Interval on the Mean $\mu$ of a Population Having Known Standard Deviation . . . . . | 44        |
| 3.6.6    | Example—One-Sided Confidence Intervals . . . . .  | 44        |
| 3.7      | Hypothesis Testing . . . . .  | 45        |
| 3.8      | Examples of Statistical Tests . . . . .   | 47        |
| 3.9      | Power and Operating Characteristic (O.C.) Curves . . . . .  | 49        |
| 3.10     | Sampling . . . . .  | 52        |
| 3.10.1   | Simple Random Sampling . . . . .  | 53        |

---

|          |  |           |
|----------|--|-----------|
| 3.10.2   | Stratified Random Sampling . . . . .   | 53        |
| 3.10.3   | Cluster Random Sampling . . . . .  | 54        |
| 3.10.4   | Systematic Random Sampling . . . . .   | 55        |
| 3.10.5   | Standard Errors of Sample Means . . . . .  | 56        |
| 3.10.6   | Sources of Bias in Samples . . . . .   | 56        |
| 3.11     | Exercises . . . . .  | 57        |
| <b>4</b> | <b>Graphs</b>  | <b>63</b> |
| 4.1      | Definition . . . . .   | 64        |
| 4.2      | Example—Ecological Correlation . . . . .   | 64        |
| 4.3      | Scatterplots . . . . .   | 65        |
| 4.4      | Scatterplot Matrix . . . . .   | 67        |
| 4.5      | Example—Life Expectancy . . . . .  | 71        |
| 4.6      | Scatterplot Matrices—Continued . . . . .   | 74        |
| 4.7      | Data Transformations . . . . .   | 78        |
| 4.8      | Life Expectancy Example—Continued . . . . .  | 82        |
| 4.9      | SAS Graphics . . . . .   | 85        |
| 4.10     | Exercises . . . . .  | 87        |
| <b>5</b> | <b>Introductory Inference</b>  | <b>91</b> |
| 5.1      | Normal ( $z$ ) Intervals and Tests . . . . .   | 91        |
| 5.1.1    | Test of a Hypothesis Concerning the Mean of a Population Having Known Standard Deviation . . . . . | 92        |
| 5.1.2    | Confidence Intervals for Unknown Population Proportion $p$ . . . . .                               | 93        |
| 5.1.3    | Tests on an Unknown Population Proportion $p$ . . . . .  | 94        |
| 5.1.4    | Example—One-Sided Hypothesis Test Concerning a Population Proportion . . . . .                     | 94        |
| 5.2      | $t$ -intervals and Tests for the Mean of a Population Having Unknown Standard Deviation . . . . .  | 95        |
| 5.3      | Confidence Interval on the Variance or Standard Deviation of a Normal Population . . . . .         | 96        |
| 5.4      | Comparisons of Two Populations Based on Independent Samples . . . . .                              | 97        |
| 5.4.1    | Confidence Intervals on the Difference Between Two Population Proportions . . . . .                | 98        |
| 5.4.2    | Confidence Interval on the Difference of Between Two Means . . . . .                               | 98        |
| 5.4.3    | Tests Comparing Two Population Means When the Samples Are Independent . . . . .                    | 99        |
| 5.4.4    | Comparing the Variances of Two Normal Populations . . . . .  | 100       |
| 5.5      | Paired Data . . . . .  | 101       |
| 5.6      | Sample Size Determination . . . . .  | 105       |
| 5.6.1    | Sample Size for Estimation . . . . .   | 105       |
| 5.6.2    | Sample Size for Hypothesis Testing . . . . .   | 106       |
| 5.7      | Goodness of Fit . . . . .  | 106       |
| 5.7.1    | Chi-square Goodness-of-Fit Test . . . . .  | 107       |

|          |  |            |
|----------|--|------------|
| 5.7.2    | Example—Test of Goodness-of-Fit to a Discrete Uniform Distribution . . . . .         | 108        |
| 5.7.3    | Example—Test of Goodness-of-Fit to a Binomial Distribution . . . . .                 | 108        |
| 5.8      | Normal Probability Plots and Quantile Plots . . . . .                                | 110        |
| 5.8.1    | Normal Probability Plots . . . . .   | 112        |
| 5.8.2    | Example—Comparing <i>t</i> -Distributions . . . . .                                  | 113        |
| 5.9      | Kolmogorov–Smirnov Goodness-of-Fit Tests . . . . .                                   | 114        |
| 5.10     | Maximum Likelihood . . . . .   | 117        |
| 5.10.1   | Maximum Likelihood Estimation . . . . .  | 118        |
| 5.10.2   | Likelihood Ratio Tests . . . . .   | 119        |
| 5.11     | Exercises . . . . .  | 119        |
| <b>6</b> | <b>One-Way Analysis of Variance</b>  | <b>123</b> |
| 6.1      | Example—Catalyst Data . . . . .  | 123        |
| 6.2      | Fixed Effects . . . . .  | 127        |
| 6.3      | Multiple Comparisons—Tukey Procedure for Comparing All Pairs of Means .              | 130        |
| 6.4      | Random Effects . . . . .   | 135        |
| 6.5      | Expected Mean Squares (EMS) . . . . .  | 135        |
| 6.6      | Example—Catalyst Data—Continued . . . . .  | 136        |
| 6.7      | Example—Batch Data . . . . .   | 137        |
| 6.8      | Example—Turkey Data . . . . .  | 137        |
| 6.8.1    | Analysis . . . . .   | 139        |
| 6.8.2    | Interpretation . . . . .   | 143        |
| 6.8.3    | Specification of Analysis . . . . .  | 143        |
| 6.9      | Contrasts . . . . .  | 144        |
| 6.9.1    | Mathematics of Contrasts . . . . .   | 144        |
| 6.9.2    | Scaling . . . . .  | 146        |
| 6.10     | Tests of Homogeneity of Variance . . . . .   | 147        |
| 6.11     | Exercises . . . . .  | 148        |
| 6.A      | Appendix: Computation for the Analysis of Variance . . . . .                         | 153        |
| 6.A.1    | Computing Notes . . . . .  | 153        |
| 6.A.2    | Computation . . . . .  | 153        |
| <b>7</b> | <b>Multiple Comparisons</b>  | <b>155</b> |
| 7.1      | Multiple Comparison Procedures . . . . .   | 156        |
| 7.1.1    | Bonferroni Method . . . . .  | 156        |
| 7.1.2    | Tukey Procedure for All Pairwise Comparisons . . . . .                               | 157        |
| 7.1.3    | The Dunnett Procedure for Comparing One Mean with All Others .                       | 157        |
| 7.1.4    | Simultaneously Comparing All Possible Contrasts—Scheffé and Extended Tukey . . . . . | 162        |
| 7.2      | The Mean–Mean Multiple Comparisons Display (MMC Plot) . . . . .                      | 168        |
| 7.2.1    | Difficulties with Standard Displays . . . . .  | 168        |
| 7.2.2    | Hsu and Peruggia’s Mean–Mean Scatterplot . . . . .                                   | 173        |
| 7.2.3    | Extensions of the Mean–Mean Display to Arbitrary Contrasts . . . . .                 | 178        |

---

|          |   |            |
|----------|---|------------|
| 7.2.4    | Display of an Orthogonal Basis Set of Contrasts . . . . .       | 180        |
| 7.2.5    | Hsu and Peruggia's Pulmonary Example . . . . .                  | 182        |
| 7.3      | Exercises . . . . .   | 184        |
| <b>8</b> | <b>Linear Regression by Least Squares</b>                       | <b>187</b> |
| 8.1      | Introduction . . . . .  | 187        |
| 8.2      | Example—Body Fat Data . . . . .                                 | 188        |
| 8.3      | Simple Linear Regression . . . . .                              | 190        |
| 8.3.1    | Algebra . . . . .   | 190        |
| 8.3.2    | Normal Distribution Theory . . . . .                            | 192        |
| 8.3.3    | Calculations . . . . .  | 193        |
| 8.3.4    | Residual Mean Square in Regression Printout . . . . .           | 199        |
| 8.3.5    | New Observations . . . . .                                      | 199        |
| 8.4      | Diagnostics . . . . .   | 205        |
| 8.5      | Graphics . . . . .  | 209        |
| 8.6      | Exercises . . . . .   | 210        |
| 8.A      | Appendix: Computation for Regression Analysis . . . . .         | 213        |
| 8.A.1    | S-PLUS Functions . . . . .                                      | 213        |
| 8.A.2    | SAS Macros and Procs . . . . .                                  | 213        |
| <b>9</b> | <b>Multiple Regression—More Than One Predictor</b>              | <b>215</b> |
| 9.1      | Regression with Two Predictors—Least-Squares Geometry . . . . . | 215        |
| 9.2      | Multiple Regression—Algebra . . . . .                           | 217        |
| 9.2.1    | The Hat Matrix and Leverage . . . . .                           | 220        |
| 9.3      | Multiple Regression—Two- $X$ Analysis . . . . .                 | 221        |
| 9.4      | Geometry of Multiple Regression . . . . .                       | 223        |
| 9.5      | Programming . . . . .   | 223        |
| 9.5.1    | Model Specification . . . . .                                   | 223        |
| 9.5.2    | Printout Idiosyncrasies . . . . .                               | 224        |
| 9.6      | Example—Albuquerque Home Price Data . . . . .                   | 225        |
| 9.7      | Partial $F$ -Tests . . . . .                                    | 228        |
| 9.8      | Polynomial Models . . . . .                                     | 230        |
| 9.9      | Models Without a Constant Term . . . . .                        | 233        |
| 9.10     | Prediction . . . . .  | 235        |
| 9.11     | Example—Longley Data . . . . .                                  | 236        |
| 9.12     | Collinearity . . . . .  | 241        |
| 9.13     | Variable Selection . . . . .                                    | 243        |
| 9.13.1   | Manual Use of the Stepwise Philosophy . . . . .                 | 244        |
| 9.13.2   | Automated Stepwise Regression . . . . .                         | 247        |
| 9.13.3   | Automated Stepwise Modeling of the Longley Data . . . . .       | 250        |
| 9.14     | Residual Plots . . . . .  | 254        |
| 9.14.1   | Partial Residuals . . . . .                                     | 254        |
| 9.14.2   | Partial Residual Plots . . . . .                                | 256        |
| 9.14.3   | Partial Correlation . . . . .                                   | 256        |

|           |   |            |
|-----------|---|------------|
| 9.14.4    | Added Variable Plots . . . . .  | 256        |
| 9.14.5    | Interpretation of Residual Plots . . . . .                                | 257        |
| 9.15      | Example—U.S. Air Pollution Data . . . . .                                 | 259        |
| 9.16      | Exercises . . . . .   | 264        |
| <b>10</b> | <b>Multiple Regression—Dummy Variables and Contrasts</b>                  | <b>267</b> |
| 10.1      | Dummy (Indicator) Variables . . . . .                                     | 267        |
| 10.2      | Example—Height and Weight . . . . .                                       | 268        |
| 10.3      | Equivalence of Linear Independent $X$ -Variables for Regression . . . . . | 275        |
| 10.4      | Polynomial Contrasts and Orthogonal Polynomials . . . . .                 | 277        |
| 10.4.1    | Specification and Interpretation of Interaction Terms . . . . .           | 282        |
| 10.5      | Analysis Using a Concomitant Variable (Analysis of Covariance) . . . . .  | 283        |
| 10.6      | Example—Hot Dog Data . . . . .  | 284        |
| 10.6.1    | One-Way ANOVA . . . . .   | 284        |
| 10.6.2    | Concomitant Explanatory Variable . . . . .                                | 286        |
| 10.6.3    | Tests of Equality of Regression Lines . . . . .                           | 292        |
| 10.7      | <code>ancova</code> Function . . . . .                                    | 294        |
| 10.8      | Exercises . . . . .   | 294        |
| <b>11</b> | <b>Multiple Regression—Regression Diagnostics</b>                         | <b>297</b> |
| 11.1      | Example—Rent Data . . . . .   | 297        |
| 11.1.1    | Rent Levels . . . . .   | 298        |
| 11.1.2    | Alfalfa Rent Relative to Other Rent . . . . .                             | 303        |
| 11.2      | Checks on Model Assumptions . . . . .                                     | 309        |
| 11.2.1    | Scatterplot Matrix . . . . .  | 309        |
| 11.2.2    | Residual Plots . . . . .  | 309        |
| 11.3      | Case Statistics . . . . .   | 312        |
| 11.3.1    | Leverage . . . . .  | 315        |
| 11.3.2    | Deleted Standard Deviation . . . . .                                      | 316        |
| 11.3.3    | Standardized and Studentized Deleted Residuals . . . . .                  | 317        |
| 11.3.4    | Cook's Distance . . . . .   | 319        |
| 11.3.5    | DFFITS . . . . .  | 321        |
| 11.3.6    | DFBETAS . . . . .   | 322        |
| 11.3.7    | Calculation of Regression Diagnostics . . . . .                           | 323        |
| 11.4      | Exercises . . . . .   | 324        |
| <b>12</b> | <b>Two-Way Analysis of Variance</b>                                       | <b>329</b> |
| 12.1      | Example—Display Panel Data . . . . .                                      | 329        |
| 12.2      | Statistical Model . . . . .   | 336        |
| 12.3      | Main Effects and Interactions . . . . .                                   | 336        |
| 12.4      | Two-Way Interaction Plot . . . . .  | 338        |
| 12.5      | Sums of Squares in the Two-Way ANOVA Table . . . . .                      | 339        |
| 12.6      | Treatment and Blocking Factors . . . . .                                  | 339        |
| 12.7      | Fixed and Random Effects . . . . .  | 341        |

---

|           |  |            |
|-----------|--|------------|
| 12.8      | Randomized Complete Block Designs . . . . .                  | 342        |
| 12.9      | Example—The Blood Plasma Data . . . . .                      | 344        |
| 12.10     | Random Effects Models and Mixed Models . . . . .             | 346        |
| 12.11     | Introduction to Nesting . . . . .                            | 347        |
| 12.11.1   | Example—Workstation Data . . . . .                           | 347        |
| 12.12     | Example—Display Panel Data—Continued . . . . .               | 349        |
| 12.13     | Example—The <i>Rhizobium</i> Data . . . . .                  | 353        |
| 12.13.1   | First <i>Rhizobium</i> Experiment: Alfalfa Plants . . . . .  | 354        |
| 12.13.2   | Second <i>Rhizobium</i> Experiment: Clover Plants . . . . .  | 354        |
| 12.13.3   | Initial Plots . . . . .                                      | 355        |
| 12.13.4   | Alfalfa Analysis . . . . .                                   | 357        |
| 12.13.5   | Clover Analysis . . . . .                                    | 362        |
| 12.14     | Models Without Interaction . . . . .                         | 371        |
| 12.15     | Example—Animal Feed Data . . . . .                           | 372        |
| 12.16     | Exercises . . . . .  | 374        |
| 12.A      | Appendix: Computation for the Analysis of Variance . . . . . | 379        |
| <b>13</b> | <b>Design of Experiments—Factorial Designs</b>               | <b>381</b> |
| 13.1      | A Three-Way ANOVA—Muscle Data . . . . .                      | 381        |
| 13.2      | Latin Square Designs . . . . .                               | 389        |
| 13.2.1    | Example—Latin Square . . . . .                               | 390        |
| 13.3      | Simple Effects for Interaction Analyses . . . . .            | 396        |
| 13.3.1    | Example—The <i>filmcoat</i> Data . . . . .                   | 397        |
| 13.4      | Nested Factorial Experiment . . . . .                        | 401        |
| 13.4.1    | Example—Gunload Data . . . . .                               | 401        |
| 13.4.2    | Example—Turkey Data (continued) . . . . .                    | 410        |
| 13.5      | Specification of Model Formulas . . . . .                    | 413        |
| 13.6      | Squential and Conditional Tests . . . . .                    | 417        |
| 13.6.1    | SAS Types of Sums of Squares . . . . .                       | 418        |
| 13.6.2    | Example—Application to Body Fat Data . . . . .               | 419        |
| 13.7      | Exercises . . . . .  | 421        |
| 13.A      | Appendix: Orientation for Boxplots . . . . .                 | 427        |
| <b>14</b> | <b>Design of Experiments—Complex Designs</b>                 | <b>429</b> |
| 14.1      | Confounding . . . . .  | 429        |
| 14.2      | Split Plot Designs . . . . .                                 | 431        |
| 14.3      | Example—Yates Oat Data . . . . .                             | 432        |
| 14.3.1    | Alternate Specification . . . . .                            | 439        |
| 14.3.2    | Polynomial Effects for Nitrogen . . . . .                    | 440        |
| 14.4      | Introduction to Fractional Factorial Designs . . . . .       | 442        |
| 14.4.1    | Example— $2^{8-2}$ Design . . . . .                          | 442        |
| 14.4.2    | Example— $2^{5-1}$ Design . . . . .                          | 444        |
| 14.5      | Introduction to Crossover Designs . . . . .                  | 448        |
| 14.6      | Example—Apple Tree Data . . . . .                            | 452        |

|           |   |            |
|-----------|---|------------|
| 14.6.1    | Models in Table 14.17 . . . . .   | 453        |
| 14.6.2    | Multiple Comparisons . . . . .  | 458        |
| 14.6.3    | Models in Figure 14.5 . . . . .   | 460        |
| 14.7      | Example— <i>testscore.dat</i> . . . . .   | 466        |
| 14.8      | The Tukey One Degree of Freedom for Nonadditivity . . . . .                             | 472        |
| 14.8.1    | Example—Crash Data . . . . .  | 472        |
| 14.8.2    | Theory . . . . .  | 481        |
| 14.9      | Exercises . . . . .   | 483        |
| <b>15</b> | <b>Bivariate Statistics—Discrete Data</b>   | <b>487</b> |
| 15.1      | Two-Dimensional Contingency Tables—Chi-Square Analysis . . . . .                        | 487        |
| 15.1.1    | Example—Drunkenness Data . . . . .  | 487        |
| 15.1.2    | Chi-Square Analysis . . . . .   | 490        |
| 15.2      | Two-Dimensional Contingency Tables—Fisher’s Exact Test . . . . .                        | 492        |
| 15.2.1    | Example—Do Juvenile Delinquents Eschew Wearing Eyeglasses? . .                          | 493        |
| 15.3      | Simpson’s Paradox . . . . .   | 495        |
| 15.4      | Relative Risk and Odds Ratios . . . . .   | 498        |
| 15.4.1    | Glasses (Again) . . . . .   | 499        |
| 15.4.2    | Large Sample Approximations . . . . .   | 500        |
| 15.4.3    | Example—Treating Cardiac Arrest with Therapeutic Hypothermia .                          | 500        |
| 15.5      | Retrospective and Prospective Studies . . . . .   | 503        |
| 15.6      | Mantel–Haenszel Test . . . . .  | 504        |
| 15.7      | Example—Salk Polio Vaccine . . . . .  | 506        |
| 15.8      | Exercises . . . . .   | 508        |
| <b>16</b> | <b>Nonparametrics</b>   | <b>511</b> |
| 16.1      | Introduction . . . . .  | 511        |
| 16.2      | Sign Test for the Location of a Single Population . . . . .                             | 512        |
| 16.3      | Comparing the Locations of Paired Populations . . . . .                                 | 514        |
| 16.3.1    | Sign Test . . . . .   | 514        |
| 16.3.2    | Wilcoxon Signed-Ranks Test . . . . .  | 516        |
| 16.4      | Mann–Whitney Test for Two Independent Samples . . . . .                                 | 520        |
| 16.5      | Kruskal–Wallis Test for Comparing the Locations of at Least Three Populations . . . . . | 523        |
| 16.6      | Exercises . . . . .   | 526        |
| <b>17</b> | <b>Logistic Regression</b>  | <b>527</b> |
| 17.1      | Example—The Space Shuttle Challenger Disaster . . . . .                                 | 529        |
| 17.1.1    | Graphical Display . . . . .   | 530        |
| 17.1.2    | Numerical Display . . . . .   | 533        |
| 17.2      | Estimation . . . . .  | 537        |
| 17.3      | Example—Budworm Data . . . . .  | 540        |
| 17.4      | Example—Lymph Nodes . . . . .   | 542        |
| 17.4.1    | Data . . . . .  | 542        |

---

|           |   |            |
|-----------|---|------------|
| 17.4.2    | Data Analysis . . . . .   | 543        |
| 17.4.3    | Additional Techniques . . . . .   | 546        |
| 17.4.4    | Diagnostics . . . . .   | 553        |
| 17.5      | Numerical Printout . . . . .  | 553        |
| 17.6      | Graphics . . . . .  | 553        |
| 17.6.1    | Conditioned Scatterplots . . . . .  | 553        |
| 17.6.2    | Scatterplot Matrix . . . . .  | 554        |
| 17.6.3    | Common Scaling in Comparable Plots . . . . .  | 554        |
| 17.6.4    | Functions of Predicted Values . . . . .   | 555        |
| 17.7      | Model Specification . . . . .   | 556        |
| 17.7.1    | S-PLUS . . . . .  | 556        |
| 17.7.2    | SAS . . . . .   | 557        |
| 17.8      | Fitting Models When the Response Is a Sample Proportion . . . . .                   | 557        |
| 17.9      | LogXact . . . . .   | 558        |
| 17.10     | Exercises . . . . .   | 558        |
| <b>18</b> | <b>Time Series Analysis</b>   | <b>565</b> |
| 18.1      | Introduction . . . . .  | 565        |
| 18.2      | The ARIMA Approach to Time Series Modeling . . . . .                                | 567        |
| 18.3      | Autocorrelation . . . . .   | 570        |
| 18.3.1    | Autocorrelation Function (ACF) . . . . .  | 570        |
| 18.3.2    | Partial Autocorrelation Function (PACF) . . . . .                                   | 570        |
| 18.4      | Analysis Steps . . . . .  | 571        |
| 18.5      | Some Algebraic Development, Including Forecasting . . . . .                         | 573        |
| 18.5.1    | The General ARIMA Model . . . . .   | 573        |
| 18.5.2    | Special case—The AR(1) model . . . . .  | 574        |
| 18.5.3    | Special Case—The MA(1) Model . . . . .  | 575        |
| 18.6      | Graphical Displays for Time Series Analysis . . . . .                               | 575        |
| 18.7      | Models with Seasonal Components . . . . .   | 580        |
| 18.7.1    | Multiplicative Seasonal ARIMA Models . . . . .                                      | 580        |
| 18.7.2    | Example— <i>co2</i> ARIMA(0, 1, 1) $\times$ (0, 1, 1) <sub>12</sub> Model . . . . . | 581        |
| 18.7.3    | Determining the Seasonal AR and MA Parameters . . . . .                             | 581        |
| 18.8      | Example of a Seasonal Model—The Monthly <i>co2</i> Data . . . . .                   | 582        |
| 18.8.1    | Identification of the Model . . . . .   | 582        |
| 18.8.2    | Parameter Estimation and Diagnostic Checking . . . . .                              | 584        |
| 18.8.3    | Forecasting . . . . .   | 589        |
| 18.9      | Exercises . . . . .   | 589        |
| 18.A      | Appendix: Graphical Displays for Time Series Analysis . . . . .                     | 618        |
| 18.A.1    | Characteristics of This Presentation of the Time Series Plot . . . . .              | 619        |
| 18.A.2    | Characteristics of This Presentation of the Sample ACF and PACF Plots . . . . .     | 619        |
| 18.A.3    | Construction of Graphical Displays . . . . .  | 620        |
| 18.A.4    | User Functions Written for S-PLUS . . . . .   | 620        |

|   |            |
|---|------------|
| <b>A Software</b>   | <b>623</b> |
| A.1 Statistical Software . . . . .  | 623        |
| A.2 Text Editing Software . . . . .   | 624        |
| A.2.1 Emacs . . . . .   | 624        |
| A.2.2 Microsoft Word . . . . .  | 625        |
| A.3 Word Processing Software . . . . .  | 625        |
| A.3.1 L <sup>A</sup> T <sub>E</sub> X . . . . .                               | 626        |
| A.3.2 Microsoft Word . . . . .  | 626        |
| A.4 Graphics Display Software . . . . .                                       | 626        |
| A.5 Operating Systems . . . . .   | 627        |
| A.6 Mathematical Fonts . . . . .  | 627        |
| A.7 Directory Structure . . . . .   | 627        |
| A.7.1 HOME Directory . . . . .  | 627        |
| A.7.2 HH Book Online Files . . . . .  | 629        |
| <b>B S-PLUS and R</b>   | <b>631</b> |
| B.1 Create Your Working Directory and Make the HH Library Available . . . . . | 632        |
| B.1.1 Windows—Both S-PLUS and R . . . . .                                     | 632        |
| B.1.2 Windows and S-PLUS . . . . .  | 633        |
| B.1.3 Windows and R . . . . .   | 634        |
| B.1.4 Unix—Both S-PLUS and R . . . . .  | 635        |
| B.1.5 Unix and S-PLUS . . . . .   | 636        |
| B.1.6 Unix and R . . . . .  | 636        |
| B.2 Using S-PLUS and R with HH . . . . .                                      | 637        |
| B.3 S-PLUS for Windows—Recommended Options . . . . .                          | 638        |
| B.4 HH Library Functions . . . . .  | 640        |
| B.5 Learning the S Language . . . . .   | 640        |
| B.6 S Language Style . . . . .  | 643        |
| B.7 S-PLUS Inexplicable Error Messages . . . . .                              | 645        |
| B.8 Using S-PLUS with Emacs and ESS . . . . .                                 | 647        |
| B.9 Constructing the HH Library with S-PLUS and R . . . . .                   | 647        |
| <b>C SAS</b>  | <b>649</b> |
| C.1 Make the HH Library Available . . . . .                                   | 649        |
| C.1.1 Windows . . . . .   | 649        |
| C.1.2 Unix . . . . .  | 650        |
| C.2 Using SAS with HH . . . . .   | 652        |
| C.2.1 Reading HH Datasets . . . . .   | 652        |
| C.2.2 Any Other Data Files . . . . .  | 653        |
| C.2.3 ASCII Data Files with TAB Characters . . . . .                          | 653        |
| C.2.4 Windows and Unix EOL (End-of-Line) Conventions . . . . .                | 654        |
| C.3 Macros . . . . .  | 655        |
| C.4 Learning the SAS Language . . . . .                                       | 655        |
| C.5 SAS Coding Conventions . . . . .  | 656        |

---

|   |            |
|---|------------|
| <b>D Probability Distributions</b>  | <b>657</b> |
| D.1 Common Probability Distributions with S-PLUS and SAS Commands . . . . . | 657        |
| D.1.1 An Example Involving Calculations with the Binomial Distribution .    | 661        |
| D.2 Noncentral Probability Distributions . . . . .                          | 661        |
| <b>E Editors</b>  | <b>663</b> |
| E.1 Working Style . . . . .   | 664        |
| E.2 Typography . . . . .  | 665        |
| E.3 Emacs and ESS . . . . .   | 667        |
| E.3.1 ESS . . . . .   | 670        |
| E.3.2 Mouse and Keyboard . . . . .  | 671        |
| E.3.3 Learning Emacs . . . . .  | 672        |
| E.3.4 Requirements . . . . .  | 672        |
| E.4 Microsoft Word . . . . .  | 673        |
| E.4.1 Learning Word . . . . .   | 673        |
| E.4.2 Requirements . . . . .  | 673        |
| E.5 Microsoft Excel . . . . .   | 674        |
| E.5.1 Database Management . . . . .   | 674        |
| E.5.2 Organizing Calculations . . . . .                                     | 674        |
| E.5.3 Excel as a Statistical Calculator . . . . .                           | 674        |
| E.6 Exhortations, Some of Which Are Writing Style . . . . .                 | 677        |
| E.6.1 Writing Style . . . . .   | 677        |
| E.6.2 Programming Style and Common Errors . . . . .                         | 678        |
| E.6.3 Presentation of Results . . . . .                                     | 679        |
| <b>F Mathematics Preliminaries</b>  | <b>683</b> |
| F.1 Algebra Review . . . . .  | 683        |
| F.2 Elementary Differential Calculus . . . . .                              | 685        |
| F.3 An Application of Differential Calculus . . . . .                       | 686        |
| F.4 Topics in Matrix Algebra . . . . .                                      | 687        |
| F.4.1 Elementary Operations . . . . .                                       | 688        |
| F.4.2 Linear Independence . . . . .   | 690        |
| F.4.3 Rank . . . . .  | 691        |
| F.4.4 Quadratic Forms . . . . .   | 692        |
| F.4.5 Orthogonal Transformations . . . . .                                  | 692        |
| F.4.6 Orthogonal Basis . . . . .  | 693        |
| F.4.7 Matrix Factorization— $QR$ . . . . .                                  | 693        |
| F.4.8 Matrix Factorization—Cholesky . . . . .                               | 695        |
| F.4.9 Orthogonal Polynomials . . . . .                                      | 695        |
| F.4.10 Projection Matrices . . . . .  | 695        |
| F.4.11 Geometry of Matrices . . . . .                                       | 695        |
| F.4.12 Eigenvalues and Eigenvectors . . . . .                               | 696        |
| F.4.13 Singular Value Decomposition . . . . .                               | 698        |
| F.4.14 Generalized Inverse . . . . .  | 698        |

|                         |   |            |
|-------------------------|---|------------|
| F.4.15                  | Solving Linear Equations . . . . .  | 699        |
| F.5                     | Combinations and Permutations . . . . .                                     | 700        |
| F.5.1                   | Factorial . . . . .   | 700        |
| F.5.2                   | Permutations . . . . .  | 700        |
| F.5.3                   | Combinations . . . . .  | 700        |
| F.6                     | Exercises . . . . .   | 701        |
| <b>G</b>                | <b>Graphs Based on Cartesian Products</b>                                   | <b>703</b> |
| G.1                     | Structured Sets of Graphs . . . . .   | 704        |
| G.1.1                   | Cartesian Products . . . . .  | 704        |
| G.1.2                   | Trellis Paradigm . . . . .  | 704        |
| G.2                     | Scatterplot Matrices: <code>splom</code> and <code>xysplom</code> . . . . . | 705        |
| G.3                     | Cartesian Products of Sets of Functions . . . . .                           | 706        |
| G.4                     | Graphs Requiring Multiple Calls to <code>xysplom</code> . . . . .           | 706        |
| G.5                     | Asymmetric Roles for the Row and Column Sets . . . . .                      | 707        |
| G.6                     | Rotated Plots . . . . .   | 707        |
| G.7                     | Squared Residual Plots . . . . .  | 708        |
| G.8                     | Alternate Presentations . . . . .   | 708        |
| <b>References</b>       |   | <b>709</b> |
| <b>List of Datasets</b> |   | <b>721</b> |
| <b>Index</b>            |   | <b>723</b> |

# Introduction and Motivation

Statistics is the science and art of making decisions based on quantitative evidence. This introductory chapter motivates the study of statistics by describing where and how it used in all endeavors. It gives examples of applications, a little history of the subject, and a brief overview of the structure and content of the remaining chapters.

Almost all fields of study (including but not limited to physical science, social science, business, and economics) collect and interpret numerical data. Statistical techniques are the standard ways of summarizing and presenting the data, of turning data from an accumulation of numbers into usable information. Not all numbers are the same. No group of people are all the same height, no group has an identical income, not all cars get the same gas mileage, not all manufactured parts are absolutely identical. How much do they differ? Variability is the key concept that statistics offers. It is possible to measure how much things are not alike. We use standard deviation, variance, range, interquartile range, and MAD (median absolute deviation from the median) as measures of not-the-sameness. When we compare groups we compare their variability as well as their range.

Statistics uses many mathematical tools. The primary tools—algebra, calculus, matrix algebra, analytic geometry—are reviewed in Appendix F. Statistics is not purely mathematics. Mathematics problems are usually well-specified and have a single correct answer on which all can agree. Data interpretation problems calling for statistics are not yet well-specified. Part of the data analyst's task is to specify the problem clearly enough that a mathematical tool may be used. Different answers to the same initial decision problem may be valid because a statistical analysis requires as-

sumptions about the data and its manner of collection, and analysts can reasonably disagree about the plausibility of such assumptions.

Statistics uses many computational tools. We have a general discussion of software and give [urls](#) for the software we use in Appendix A. We emphasize two software systems for statistical analysis: the S language (Appendix B) as implemented in S-PLUS and in R, and the SAS system (Appendix C). SAS is the most widely used package for serious and extensive statistical analysis and data management. S-PLUS is an exceptionally well-developed tool for statistical research and analysis, that is for exploring and designing new techniques of analysis, as well as for analysis.

We make liberal use of graphs in our presentations. Data analysts are responsible for the *display* of data with graphs and tables that summarize and represent the data and the analysis. Graphs are often the output of data analysis that provide the best means of communication between the data analyst and the client. We study a variety of display techniques. The captions to our graphs contain the names of both the code file that produced the graph and the PostScript file for the graph itself. All code, transcript, and PostScript files are included in the online files. We consider the online files to be an integral part of the book.

While producing this book, we designed many innovative graphical displays of data and analyses. We introduce our displays in Section 1.3.4. We discuss the displays throughout the book in the context of their associated statistical techniques. These discussions are indexed under the term *graphical design*. In Appendix G, we summarize the large class of newly created graphs that are based on Cartesian products.

Statistics is an art. Skilled use of the mathematical tools is necessary but not sufficient. The data analyst must also know the subject area under study (or must work closely with a specialist in the subject area) to ensure an appropriate choice of statistical techniques for solving a problem. Experience, good judgment, and considerable creativity on the part of the statistical analyst are frequently needed.

Statistics is “the science of doing science” and is perhaps the only discipline that interfaces with all other sciences. Most statisticians have training or considerable knowledge in one or more areas other than statistics. The statistical analyst needs to communicate successfully both orally and in writing with the client for the analysis.

Statistics uses many communications skills, both written and oral. Results must be presented to the client and to the client’s management. We discuss some of the mechanics of writing programs and technical reports in Appendix E on Editors.

A common statistical problem is to discover the characteristics of an unobservable population by examining the corresponding characteristics of a sample *randomly* selected from the population and then (inductively) inferring the population characteristics (parameters) from the corresponding sample characteristics (statistics). The task of selecting a random sample is not trivial. The discipline of statistics has developed a vast array of techniques for inferring from samples to populations, and for using probabilities to quantify the quality of such inferences.

Most statistical problems involve simultaneous consideration of several related measurements. Part of the statistician's task is to determine the interdependence among such measures, and then to account for it in the analysis.

The word "statistics" derives from the political science collections of numerical data describing demographics, business, politics that are useful for management of the "state". The development of statistics as a scientific discipline dates from the end of the 19<sup>th</sup> century with the design and analysis of agricultural experiments aimed at finding the best combination of fertilization, irrigation, and variety to maximize crop yield. Early in the 20<sup>th</sup> century, these ideas began to take hold in industry, with experiments designed to maximize output or minimize cost. Techniques for statistical analysis are developed in response to the needs of specific subject areas. Most of the techniques developed in one subject field can be applied unchanged to other subjects.

## 1.1 Statistics in Context

We write as if the statistician and the client are two separate people. In reality they are two separate roles and the same person often plays both roles. The client has a problem associated with the collection and interpretation of numerical data. The statistician is the expert in designing the data collection procedures and in calculating and displaying the results of statistical analyses.

The statistician's contribution to a research project typically includes the following steps:

1. Help the client phrase the question(s) to be answered in a manner that leads to sensible data collection and that is amenable to statistical analysis.
2. Design the experiment, survey, or other plan to approach the problem.
3. Gather the data.
4. Analyze the data.

## 5. Communicate the results.

In most statistics courses, including the one for which this book is designed, much of the time is spent learning how to perform step 4, the science of statistics. However, step 2, the art of statistics, is very important. If step 2 is poorly executed, the end results in step 5 will be misleading, disappointing, or useless. On the other hand, if step 4 is performed poorly following an excellent plan from step 2 and a correct execution of step 3, a reanalysis of the data (a new step 4) can “save the day”.

Today there are more than 15,000 statisticians practicing in the United States. Most fields in the biological, physical, and social sciences require training in statistics as educational background. Over 100 U.S. universities offer graduate degrees in statistics. Most firms of any size and most government agencies employ statisticians to assist in decision making. For example, the Merck & Co. locations in Greater Philadelphia and New Jersey employ over 150 statisticians with advanced degrees. The profession of *statistician* is highly placed in the *Jobs Rated Almanac* (Krantz, 1999). A shortage of qualified statisticians to fill open positions is expected to persist for some time (American Statistical Association, 2002).

## 1.2 Examples of Uses of Statistics

Below are a few examples of the countless situations and problems for which statistics plays an important part in the solution.

### 1.2.1 Investigation of Salary Discrimination

When a group of workers believes that their employer is illegally discriminating against the group, legal remedies are often available. Usually such groups are minorities consisting of a racial, ethnic, gender, or age group. The discrimination may deal with salary, benefits, an aspect of working conditions, mandatory retirement, etc. The statistical evidence is often crucial to the development of the legal case.

To illustrate the statistician’s approach, we consider the case of claimed salary discrimination against female employees. The legal team and statistician begin by developing a defensible list of criteria that the defendant may legally use to determine a worker’s salary. Suppose such a list includes years of experience (`yrsexp`), years of education (`yrsed`), a measure of current job responsibility or complexity (`respon`), and a measure of the worker’s current productivity (`product`). The statistician then obtains from a sample of employees, possibly following a subpoena by the legal team,

data on these four criteria and a fifth criterion that distinguishes between male and female employees (**gender**). Using regression analysis techniques we introduce in Chapter 9, the statistician considers two statistical models, one that explains **salary** as a function of the four stipulated permissible criteria, and another that explains **salary** as a function of these four criteria plus **gender**. If the model containing the predictor **gender** predicts salary appreciably better than does the model excluding **gender** and if, according to the model with **gender** included, females receive significantly less salary than males, then this may be regarded as statistical evidence of discrimination against females. Tables and graphs based on techniques discussed in Chapters 15, 17, and 4 (and other chapters) are often used in legal proceedings.

In the previous section it is pointed out that two statisticians can provide different analyses because of different assumptions made at the outset. In this discrimination context, the two legal teams may disagree over the completeness or relevance of the list of permissible salary determinants. For example, the defense team may claim that females are “less ambitious” than males, or that women who take maternity or child care leaves have less continuous or current experience than men. If the court accepts such arguments, this will undermine the plaintiff statistician’s finding of the superiority of the model with the extra predictor.

### 1.2.2 Measuring Body Fat

In Chapters 8, 9, and 13 we discuss an experiment designed to develop a way to estimate the percentage of fat in a human body based only on body measurements that can be made with a simple tape measure. The motivation for this investigation is that measurement of body fat is difficult and expensive (it requires an underwater weighing technique), but tape measurements are easy and inexpensive to obtain. At the outset of this investigation, the client offered data consisting of 15 inexpensive measurements and the expensive body fat measurement on each of 252 males of various shapes and sizes. Our analysis in Chapter 9 demonstrates that essentially all of the body fat information in the 15 other measurements can be captured by just three of these other measurements. We develop a regression model of body fat as a function of these three measurements, and then we examine how closely these three inexpensive measurements alone can estimate body fat.

### 1.2.3 Minimizing Film Thickness

In Section 13.3.1 we discuss an experiment that seeks to find combinations of **temperature** and **pressure** that minimize the thickness of a film

deposited on a substrate. Each of these factors can affect thickness, and the complication here is the possibility that the optimum amount of one of these factors may well depend on the chosen amount of another factor. Modeling such *interaction* between factors is key to a proper analysis. The statistician is also expected to advise on the extent of sensitivity of thickness to small changes in the optimum mix of factors.

#### 1.2.4 Surveys

Political candidates and news organizations routinely sample potential voters for their opinions on candidates and issues. Results gleaned from samples selected by contemporary methods are often regarded as sufficiently reliable to influence candidate behavior or public policy.

The marketing departments of retail firms often sample potential customers to decide issues such as product composition or packaging, and the best matches between specialized products and locales for their sale.

Manufacturers sample production to determine if the proportion of output that is defective is excessive. If so, this may lead to the decision the output should be scrapped, or at least that the production process be inspected and corrected for problems.

All three of these examples share statistical features. The data are collected using techniques discussed in Section 3.10. The initial analysis is usually based on techniques of Chapter 5.

#### 1.2.5 Bringing Pharmaceutical Products to Market

The successful launching of a new pharmaceutical drug is a huge undertaking in which statisticians are key members of the investigative team. After candidate drugs are found to be effective for alleviation of a condition, experiments must be run to check them for toxicity, safety, side effects, and interactions with other drugs. Once these tests are passed, statisticians help to determine the optimum quantity and spacing of dosages. Much of the testing is done on lab animals; only at the later stages are human subjects involved. The entire process is performed in a manner mandated by government regulatory agencies (such as the FDA in the United States). Techniques are based on material developed in all chapters of this book.

## 1.3 The Rest of the Book

### 1.3.1 Fundamentals

Chapters 2 through 5 discuss data, types of data analysis, and graphical display of data and of analyses.

Chapter 2 describes data acquisition and how to get the data ready for its analysis. We emphasize that an important early step in any data analysis is graphical display of the data.

Chapter 3 provides an overview of basic concepts—probability, distributions, estimation, testing, principles of inference, and sampling—that are background material for the remainder of the book. Several common distributions are discussed and illustrated here. Others appear in Appendix D. Two important fitting criteria—least squares and maximum likelihood—are introduced. Random sampling is a well-defined technique for collecting data on a subset of the population of interest. Random sampling provides a basis for making inferences that a haphazard collection of data cannot provide.

A variety of graphical displays are discussed and illustrated in Chapter 4. The graphs themselves are critically important analysis tools, and we show examples where different display techniques help in the interpretation of the data. On occasion we display graphs that are intermediate steps leading to other graphs. For example, Figure 14.14 belongs in a final report, but Figure 14.12, which suggests the improved and expanded Figure 14.14, should not be shown to the client.

Chapter 5 introduces some of the elementary inference techniques that are used throughout the rest of the book. We focus on tests on data from one or two normal distributions. We show the algebra and graphics for finding the center and spread of the distributions. These algebraic and graphical techniques are used in all remaining chapters.

### 1.3.2 Linear Models

Chapters 6 through 13 build on the techniques developed in Chapter 5. The word “linear” means that the equations are all linear functions of the model parameters and that graphs of the analyses are all straight lines or planes.

In Chapter 6 we extend the *t*-tests of Chapter 5 to the comparison of the means of several (more than two) populations.

With  $k > 2$  populations, there are only  $k - 1$  independent comparisons possible, yet we often wish to make  $\binom{k}{2}$  comparisons. In Chapter 7 we discuss

the concept of *multiple comparisons*, the way to make valid inferences when there are more comparisons of interest than there are degrees of freedom. We introduce the fundamental concept of “contrasts”, direct comparisons of linear combinations of the means of these populations, and show several potentially sensible ways to choose  $k - 1$  independent contrasts. We introduce the *MMC* plot, the mean–mean plot for displaying arbitrary multiple comparisons.

Chapters 8 through 11 cover regression analysis, the process of modeling a continuous response variable as a linear function of one or more predictor variables.

In Chapter 8 we plot a continuous response variable against a single continuous predictor variable and develop the least-squares procedure for fitting a straight line to the points in the plot. We cast the algebra of least squares in matrix notation (relevant matrix material is in Appendix F) and apply it to more than one predictor variable. We introduce the statistical assumptions of a normally distributed error term and show how that leads to estimation and testing procedures similar to those introduced in Chapter 5.

Chapter 9 builds on Chapter 8 by allowing for more than one predictor for a response variable and introducing additional structure, such as interactions, among the predictor variables. We show techniques for studying the relationships of the predictors to each other as well as to the response.

Chapter 10 shows how dummy variables are used to incorporate categorical predictors into multiple regression models. We begin to use dummy variables to encode the contrasts introduced in Chapter 6, and we continue using dummy variables and contrasts in Chapters 12, 13, and 14. We show how the use of continuous (concomitant) variables (also known as covariates) can enhance the modeling of designed experiments.

Chapter 11 evaluates the models, introduces diagnostic techniques for checking assumptions and detecting outliers, and uses tools such as transformation of the variables to respond to the problems detected.

In Chapter 12 we extend the analysis of one-way classifications of continuous data to several types of two-way classifications. We cast the analysis of variance into the regression framework with dummy variables that code the classification factors with sets of contrasts.

In Chapters 13 and 14 we consider the principles of experimental design and their application to more complex classifications of continuous data. We discuss the analysis of data resulting from designed experiments.

### 1.3.3 Other Techniques

The analysis of tabular categorical data is considered in Chapter 15. We discuss contingency tables, tables in which frequencies are classified by two or more factors. For  $2 \times 2$  tables or sets of  $2 \times 2$  tables we use odds ratios or the Mantel-Haenszel test. For larger tables we use  $\chi^2$  analysis. We discuss several situations in which contingency tables arise, including sample surveys and case-control studies.

In Chapter 16 we briefly survey nonparametric testing methods that don't require the assumption of an underlying normal distribution.

Chapter 17 is concerned with logistic regression, the statistical modeling of a response variable which is either dichotomous or which represents a probability. We place logistic regression in the setting of generalized linear models (although we do not fully discuss generalized linear models in this volume). We extend the graphical and algebraic analysis of linear regression to this case.

We conclude in Chapter 18 with an introduction to ARIMA modeling of time series. Time series analysis makes the explicit assumption that the observations are *not* independent and models the structure of the dependence.

### 1.3.4 New Graphical Display Techniques

This book presents many new graphical display techniques for statistical analysis. Most of our new displays are based on defining the panels of a multipanel graphical display by a Cartesian product of sets of variables, of transformations of a variable, of functions of a fit, of models for a fit, of numbers of parameters, or of levels of a factor. Appendix G summarizes how we use the Cartesian products to design these new displays and gives a reference to an example in the book for each. The displays, introduced throughout this book's 18 chapters, cover a wide variety of statistical methods. The construction and interpretation of each display is provided in the chapter where it appears.

We produced these displays with the S-PLUS functions listed in Table B.3 that we wrote expressly for this book and that are included in the on-line files. We use S-PLUS because it is especially strong for designing and programming statistical graphics. We encourage readers and software developers to write and publish functions and macros for these displays in other software systems that have a similarly rich graphics environment.

# Data and Statistics

Statistics is the field of study whose objective is the transformation of data (usually sets of numbers along with identifying characteristics) into information (usually in the form of tables, graphs, and written and verbal summaries) that can inform sound policy decisions. We give examples of applications of statistics to many fields in Chapter 1. Here we focus on the general concepts describing the collection and arrangement of the numbers themselves.

## 2.1 Types of Data

Traditionally, we refer to five different *types* of data: count, categorical, ordered, interval, and ratio.

**count data:** The observational unit either has, or does not have, a particular property. For example, tossed coins can come up heads or tails.

We count the number of  $n$  of heads when a total of  $N$  coins are tossed.

**categorical data:** The data values are distinct from each other. Categorical variables are also referred to as *nominal* variables, *class* variables, or *factors*. The various categories or classes of a categorical variable are called its *levels*. An example of a factor, from the introductory paragraph of Chapter 6, is `factory` having six levels. That is, the investigation takes place at six factories. If we code `factory` as  $\{1, 2, 3, 4, 5, 6\}$ , meaning that we arbitrarily assign these six numbers to the six factories, we must be careful not to interpret these codes as ratio data. Coding the

factory levels as integers doesn't give us the right to do arithmetic on the code numbers.

**ordered data:** The data values can be placed in a rank ordering. For any two observations, the analyst knows which of the two is larger, but not necessarily the magnitude of the difference between them. There is a distinct concept of *first*, *second*, ..., *last*. There is no way to measure the distance between values. An example is military ranks: A general is higher-ranked than a colonel, which in turn is higher than a major. There is no easy way to say something like, "A general is twice as far above a colonel as a colonel is above a major."

**interval data:** The data values have well-defined distances between them. School grades are an example. Students in 10<sup>th</sup> grade have studied one year longer than students in 9<sup>th</sup> grade; similarly, students in 9<sup>th</sup> grade have studied one year longer than students in 8<sup>th</sup> grade. It is not meaningful to say a 10<sup>th</sup>-grade student is twice as knowledgeable as a 5<sup>th</sup>-grade student.

**ratio data:** The data values are measured by real numbers: There are a well-defined origin and a well-defined unit. Height of people is an example. There is a well-defined 0 height. We can speak of one person being 1 inch taller than another or of being 10% taller than another.

We also have another categorization of data as *discrete* or *continuous*. Discrete data have a finite or countably infinite number of possible values the data can take. Continuous data take any real number value in an interval; the interval may be either closed or open.

Many of the datasets we will study, both in this book and in the data analysis situations that this book prepares you for, have several variables. Frequently, there are one or more ratio-scaled numeric variables associated with each value of a categorical variable. When only one numeric variable is measured for each observational unit, the dataset is said to be *univariate*. When there are  $k$  ( $k > 1$ ) variables measured on each observational unit, the dataset is said to be *multivariate*. Multivariate datasets require additional techniques to identify and respond to correlations among the observed variables.

## 2.2 Data Display and Calculation

Data are often collected and presented as tables of numbers. Analysis reports are also frequently presented as numbers. Tables of numbers can be presented on a page in ways that make them easier to read or harder to read. We illustrate some of each here and will identify some of the formatting decisions that affect the legibility of numerical tables.

TABLE 2.1. Legible and illegible tabular displays of the same numerical data: In panel a the numbers are aligned on the decimal point and are displayed to the same precision (the same number of decimal digits). In panel b the numbers are centered or left justified—with the effect of hiding the comparability—and displayed with different precisions—which further hides comparability.

| a. Correct |           |           | b. Incorrect |           |           |
|------------|-----------|-----------|--------------|-----------|-----------|
| 109.20931  | 133.50234 | 112.21950 | 109.209      | 133.50234 | 112.21    |
| 153.91753  | 78.97100  | 109.31152 | 153.9        | 78        | 109.31152 |
| 80.26995   | 83.76253  | 77.03695  | 80.26        | 83.76253  | 77.036    |
| 74.81323   | 112.72001 | 119.71915 | 74.81323     | 112.72001 | 119.7     |
| 84.22883   | 103.84942 | 85.58610  | 84.2         | 103.      | 85.58     |
| 80.55801   | 100.94474 | 115.13436 | 80.55801     | 100.94474 | 115.13436 |
| 85.51940   | 89.28095  | 109.24788 | 85.51940     | 89.28095  | 109.24788 |

### 2.2.1 Presentation

There are two general principles:

**alignment of decimal points:** Units digits of each number are in the same vertical column. Larger numbers extend farther to the left.

**display comparable numbers with common precision:** Numbers to be compared are displayed so the positions to be compared are in the same column.

Table 2.1 shows two tables with identical numerical information. The first is legible because it follows both principles; the second is not because it doesn't.

### 2.2.2 Rounding

The number of decimal digits in a number indicates the precision with which the number was observed. Any analysis normally retains the same precision. Any changes in the number of decimal digits that are not justified by the method of analysis implicitly suggests that the data are either more or less precise than they actually are. This can lead to misleading inferences and wrong policy decisions.

There are simple rules:

1. DO NOT ROUND intermediate values! Keep all 16 digits of double precision arithmetic in a computer program and all 12 digits on pocket calculators. For example, if a correct calculation  $7.1449/3.6451 = 1.9601$  is rounded to  $7.14/3.65$ , the quotient is less than 1.96 and a decision based on whether or not the result exceeds 1.96 will reach an incorrect conclusion.

2. Final answers may be rounded to the SAME number of significant digits as the original data. You may never have final answers with more digits than any intermediate value or with more digits than the original data.
3. Standard deviations can give a hint as to the number of believable digits. For example, if  $\bar{x} = 1.23456$  and  $s_{\bar{x}} = .0789$ , then we can justifiably round to  $\bar{x} \approx 1.234$  (using the logic that  $t = 1.23456/.0789 = 15.64715 \approx 15.647$  is good to three decimal positions).

## 2.3 Importing Data

The datasets discussed in this book and provided in our online files are in ASCII format. In consulting environments data is often collected and stored in a database management system. It is currently uncommon for the data analyst to be offered ASCII data. Fortunately, S-PLUS and SAS have facilities for importing data from a variety of formats.

### 2.3.1 S-PLUS

In S-PLUS use the `importData` command to see the list of data types that are supported and to get details on the syntax of the command. For relatively simple data structures, it is also possible to use the GUI: Go to “File/Import Data/From File” and fill in the requested menu items. S-PLUS can also write data back in various formats, either with the `exportData` command or with the GUI “File/Export Data/To File” menu. By alternately using both the `importData` and `exportData` commands, S-PLUS can be used as a versatile utility for changing data from one file type to another.

We often receive data as Microsoft Excel files (with extension `.xls`). We sometimes have to take some precautions to assure that S-PLUS is able to work with the data as the producer of the Excel file intended.

The easiest situation is where the data appear as a rectangle on a single sheet, that is, there are the same number of observations on each variable. In this situation, when calling the data with S-PLUS code, the user must be careful to exclude rows with extraneous information—such as multiple rows of header or rows containing column summary statistics.

If the Excel file contains multiple sheets, each sheet must be individually imported to S-PLUS by saving a new Excel file with the sheet in the first position and then importing the new file.

When the data do not appear as a rectangle, that is, the variables have differing numbers of observations, S-PLUS will record as missing (`NA`) the entries in any column between the last datum in that column and the

row number of the last row in the longest column in the spreadsheet. For example, if the longest column, say A, has 100 entries while column B has 83 entries, cells 84 through 100 in column B will contain NA's.

We work with NA in the usual ways, either by telling the functions about them, for example,

```
import.data(Book1, "Book1.xls", "EXCEL")
lm(C1 ~ C2, data=Book1, na.action=na.omit)
```

or by defining new variables that include only the nonmissing values,

```
C1 <- Book1$C1; C2 <- Book1$C2[1:83]
```

### 2.3.2 SAS

SAS offers a GUI process for data import and export similar to that provided by S-PLUS. SAS Version 9.0 can handle a far shorter list of formats than S-PLUS Version 6.1. Excel is a supported format. We begin with the GUI option “File/Import Data/” which invites an opportunity to browse for the file.

### 2.3.3 Data Rearrangement

Datasets are not necessarily arranged in the most convenient way for the analysis you have in mind.

We usually work with one of two data arrangements. One arrangement is a set of multiple columns, one per variable, with factor levels explicitly indicated by data values in the appropriate column. Each observation has all its values listed in the same row of all columns.

The other is a structured arrangement of values for the response variable, with levels of factors implied by the position of each observation in the data file. We have several examples with implicit factor levels. See, for example, Exercise 6.5 with data file (`(datasets/patient.dat)`). The data in (`(datasets/patient.dat)`) are in five columns, one column per treatment level, with variable numbers of rows per treatment. The columns are separated by space characters. We provide annotated code files (`(oway/code/patient.s)` and `(oway/code/patient.sas)`) to read this data.

The file (`(datasets/draft70mn.dat)`) contains daily data arranged in 12 columns, 1 per month, with columns having 29, 30, or 31 rows. The

data file must be read with a fixed format to ensure that, for example, the blanks for February 30 and 31 are correctly detected. See files (`grap/code/draft70mn-read.s`) and (`grap/code/draft70mn-read.sas`).

The file (`datasets/weightloss.dat`) also has the levels of the factor implied by the column in which the data appear. We show in code files (`mcomp/code/weightloss.s`) and (`mcomp/code/weightloss.sas`) how to convert that to the two-column format used by `aov()` and PROC ANOVA.

There are many formats for ASCII data. In addition to the space-separated format used in this book, there are also tab-separated and comma-separated formats. There are also many binary formats, usually associated with the software package that originated them. S-PLUS has objects inside a `.Data` directory, SAS has datasets inside a working directory, and Excel has a `.xls` format.

## 2.4 Analysis with Missing Data

Statisticians frequently encounter situations where an analysis cannot be completed in straightforward fashion because some portion of the data is missing. For some analyses, missing or unbalanced data cause no difficulty beyond the need to calculate results with a revised formula. Examples include the two-sample  $t$ -test of Section 5.4.3 and the one-way analysis of variance of Chapter 6. In other circumstances, such as multiple regression analyses discussed in Chapters 9 to 11, the analyst must either discard the observations carrying incomplete information or use sophisticated techniques beyond the scope of this book to impute, or estimate, the missing portions of the existing data. If the reasons for “missingness” are related to the problem being addressed, abandoning observations is likely to lead to incorrect inferences. If the data are missing at random, discarding a few observations may be a satisfactory solution, but the smaller the ultimate sample size, the less likely the analysis will produce useful and correct results. Imputing the values of missing data is usually preferable to discarding cases with incomplete information. We recommend (Little and Rubin, 2002) as a comprehensive reference on how to handle missing data, particularly techniques of imputation.

### 2.4.1 Missing Data in S-PLUS

The S-PLUS convention for missing data is `NA` (a standard abbreviation for “Not Available” or “No Answer”).

When S-PLUS knows that a value is missing it prints “NA” (without the quotes). When S-PLUS is reading an ASCII data file, it will recognize by default the character sequence “NA” as a missing observation.

If the ASCII data file uses some other convention (such as the “.” that SAS uses by default), then we must tell S-PLUS to use a different convention for reading missing data either with an argument to the `read.table` function or, after the reading, by some logical investigation of the data values.

See our annotated code file (`oway/code/patient.s`) for an example of how to use the S language to read a complicated data structure.

### 2.4.2 Missing Data in SAS

The SAS convention for missing data is `.` and, more generally, `.A-Z`. Logical statements in the DATA step can also be used to identify and revalue missing observations.

SAS does not have an easy way to tell the program that a different missing-value convention is in use. Instead, we must read the field into a character-valued variable and then explicitly convert it to a numeric with a DATA step statement.

See our annotated code file (`oway/code/patient.sas`) for an example of how to use the SAS language to read a complicated data structure.

## 2.5 Tables and Graphs

Graphs constructed from data arranged in a table are generally more useful and informative than the table. The human eye and brain can quickly discern patterns from a well-constructed picture that would be far from obvious from the underlying tabular data. Excellent examples are contained in (Tufte, 2001) and (Wainer, 1997).

Characteristics that we wish to reveal with our graphs are location, variability, scale, shape, correlation, interaction, clustering, and outliers. In Chapter 4 we illustrate many of these characteristics, primarily through our discussion of scatterplots and scatterplot matrices. Additional types of displays are presented in many subsequent chapters. We discuss both the information about the data that we obtain from the graphs and the structure of the graphs. We introduce many new types of graphs throughout the book. In Appendix G we provide a summary on those new graphs that are based on Cartesian products.

## 2.6 Files for Statistical Analysis and Data Display (HH)

We have many types of files in the companion online files (Heiberger and Holland, 2004b) that are an integral part of this book.

### 2.6.1 Datasets

This book is on the topic of data analysis. We have many datasets that we analyze in examples or make available for analysis in exercises. Most datasets are real, taken from journal articles; data repositories of governments, corporations and organizations; data libraries; or our own consulting experience. Citations to these datasets are included in the text. As befits a text, most data we present are structured for the techniques of the chapter in which we present it. Our datasets are frequently used in more than one chapter. We have an index of datasets, with which you can locate all references to a specific dataset across chapters.

All datasets used in either examples or exercises are included in the accompanying online files in the directory `hh/datasets`. We have usually structured the datasets on disk in a way that is relatively easy to read.

### 2.6.2 Code, Transcripts, and Figures

We include in the online files the computer programs for all *examples*, and for occasional *exercises*. Thus you can duplicate all tables and figures in the book and you can use these as templates for analyzing other data. The files for the examples are in a subdirectory named after the chapter. For example, the files associated with Chapter 6, the one-way analysis of variance chapter, are in subdirectories of the chapter's directory `hh/oway`. The code, both S-PLUS and SAS, is in directory `hh/oway/code`. The transcripts of the analyses (the printed output from running the examples) are in the directory `hh/oway/transcript`. All figures in the chapter are in the subdirectory `hh/oway/figure`. The code and transcript files are in ASCII. The figure files are gzipped PostScript files with filenames of the form `name.eps.gz` or `name.ps.gz`.

The easiest way to see gzipped PostScript files on Windows is with Ghostscript (Aladdin Enterprises Software Pty Ltd., 2001) and GSView (Ghostgum Software Pty Ltd., 2001). Either drop the `name.eps.gz` file from Windows Explorer onto the GSview icon or open the file from the GSview File menu. On Windows you may need to download GhostScript and GSview. On Unix, these programs are normally included. There is no need to gunzip the files as the viewer reads the gzipped versions directly.

### 2.6.3 Functions and Macros

We include many S-PLUS and R functions and SAS macros in the online files. All of these are collected into a single directory in the file structure. When you attach the directory to your S-PLUS or SAS session, you get immediate access to the functions.

All data and code files are mentioned in the text with pathnames relative to the `hh` directory. We describe how to reference the data and code files from S-PLUS and R in Appendix B and from SAS in Appendix C. When you first set up your computer for the text, you define the S-PLUS function `hh()` or the SAS macro `&hh`. Forever after, you use the `hh` referent to get all datasets and code. We recommend that you copy the `hh` directory tree to your hard disk. You may choose instead to work directly from the CD of the unzipped downloaded online files (see Preface Section 3). In either case, proper definition of `hh` makes the referencing easy.

### 2.6.4 Software

The software we use is discussed in Appendix A. URLs for the software are in the `hh` directory tree in the file (`sftw/code/url.htm`).

# Statistics Concepts

## 3.1 A Brief Introduction to Probability

The quality of inferences are commonly conveyed by probabilities. Therefore, before discussing inferential techniques later in this chapter, we briefly digress to discuss *probability* in this section and *random variables* in Section 3.2.

If  $A$  is any *event*,  $P(A)$  represents the probability of occurrence of  $A$ . Always,  $0 \leq P(A) \leq 1$ . The odds in favor of the occurrence of event  $A$  are

$$\frac{P(A)}{1 - P(A)} \quad (3.1)$$

and the odds against the occurrence of event  $A$  are

$$\frac{1 - P(A)}{P(A)} \quad (3.2)$$

Thus, if  $P(A) = \frac{3}{4}$ , then the odds in favor of  $A$  are 3, also referred to as 3 to 1, and the odds against  $A$  are  $\frac{1}{3}$ .

If  $B$  is a second event,  $A \cup B$  represents the event that “either  $A$  or  $B$  occurs”, that is, the *union* of  $A$  and  $B$ , then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (3.3)$$

where  $A \cap B$  is the event that “both  $A$  and  $B$  occur”, that is the *intersection* of  $A$  and  $B$ . Events  $A$  and  $B$  are said to be *mutually exclusive* events if they cannot both occur; in this case,  $A \cap B = \emptyset$  (the impossible event) and

so  $P(A \cap B) = 0$ . Events  $A$  and  $B$  are said to be *independent* events if the occurrence or nonoccurrence of one of them does not affect the probability of occurrence of the other one; for independent events,

$$P(A \cap B) = P(A) P(B)$$

The *conditional probability* of  $B$  given  $A$ , written  $P(B | A)$ , is the probability of occurrence of  $B$  given that  $A$  occurs. If  $P(A) \neq 0$ ,

$$P(B | A) = \frac{P(A \cap B)}{P(A)}$$

Note that  $P(B | A) = P(B)$  if  $A$  and  $B$  are independent events, but not otherwise.

To illustrate these ideas, imagine a box containing six white and four red billiard balls, identical to the touch. Suppose we select two balls from the box and let  $A$  = “the first ball is white” and  $B$  = “the second ball is white”.  $A$  and  $B$  are independent events if the first ball is replaced in the box prior to drawing the second ball, but not otherwise. Let us assume that the first ball is not replaced so that the two events are dependent. Various sets of events are listed with their probabilities in Table 3.1.

In this table we demonstrate two ways to calculate the probability  $\frac{78}{90}$  that we get a white ball in either the first selection or second selection or both selections. One way is with the formula for  $P(A \cup B)$  in Equation (3.3). Another method is to recognize that the event “at least one white” can be partitioned into three mutually exclusive events: First draw white and second draw red; first draw red and second draw white; and both draws white. The probability of “at least one white” is seen to be the sum of the probabilities of the events comprising this partitioning.

## 3.2 Random Variables and Probability Distributions

A *random variable*, abbreviated as r.v., is a function that associates events with real numbers. For example, if we toss a coin 10 times, we can define an r.v.  $X$  to be the number of heads observed in these 10 tosses. This r.v. has *possible values*  $x = 0, 1, 2, \dots, 10$ . Observing 7 heads among the 10 tosses is an event, and “7” is the number that this r.v.  $X$  associates with it.

A closely related concept is the r.v.’s *probability distribution*, which indicates how the total probability, 1, is distributed or allocated to the possible values of the r.v. It is usual to denote an r.v. with a capital letter and a possible value of this r.v. with the corresponding lowercase letter.

TABLE 3.1. Probability of Events

| Event      | Position   |     | Probability                             |   | Probability of event  |
|------------|--|-----|---|---|---|
|            | 1  | 2   | 1                                       | 2   |   |
| $A$        | $W$  | $?$ | $\frac{6}{10}$                          | 1   | $\frac{6}{10}$  |
| $B$        | $?$  | $W$ | 1                                       | $\frac{6}{10}$  | $\frac{6}{10}$  |
| $B \cap A$ | $W$  | $W$ | $\frac{6}{10}$                          | $\frac{5}{9}$   | $\frac{30}{90}$   |
| $B   A$    | $[W]$  | $W$ | $\frac{(\frac{6}{10})}{(\frac{6}{10})}$ | $\frac{5}{9}$   | $\frac{5}{9}$   |
| $B \cup A$ | $\left\{ \begin{array}{ll} W & R \\ R & W \\ W & W \end{array} \right.$                                  |     | $\frac{6}{10}$                          | $\frac{4}{9}$   | $P(WR) + P(RW) + P(WW) = P(A) + P(B) - P(B \cap A) = P(B \cup A)$ |
|            | $\left. \begin{array}{ll} \frac{4}{10} & \frac{6}{9} \\ \frac{6}{10} & \frac{5}{9} \end{array} \right\}$ |     | $\frac{24}{90}$                         | $+$   | $\frac{24}{90}$   |
|            | $\left. \begin{array}{ll} & \frac{30}{90} \\ & \frac{6}{10} \end{array} \right\}$                        |     | $=$                                     | $\frac{6}{10} + \frac{6}{10} - \frac{30}{90} = \frac{78}{90}$ | $= \frac{78}{90}$   |

### 3.2.1 Discrete Versus Continuous Probability Distributions

There are essentially two distinct types of probability distribution of a quantitative variable: discrete and continuous. (Random variables are also classified as discrete or continuous according to the classification of their probability distributions.) It is important to distinguish between the two types because they differ in their methods of display and calculation.

The key distinction between these two types relates to the spacings between adjacent possible values of the data. For discrete data, the distance separating consecutive possible values of the variable does not depend on a measurement device; indeed it may be completely arbitrary. For continuous data, the distances may (theoretically) assume all possible values in some interval.

For example, the number of times an archer hits a target in 10 attempts is a discrete variable because the answer is a count of the number of occurrences. It is impossible for there to be 3.5 hits. A discrete variable need not be integer-valued. The proportion of hits in 10 attempts is also discrete. It is impossible for this proportion to be .35. It is possible for a discrete variable to have a *countably infinite* number of possible values. An example would be the number of attempts needed for the archer to achieve her ninth hit. This

variable can assume any positive integer value; it is possible but unlikely that the archer will need 100 attempts.

On the other hand, the archer's height in inches is a continuous variable because it can be anything between perhaps 3 feet and 8 feet. While as a practical matter it would be difficult to measure height to within  $\frac{1}{4}$ -inch accuracy, it is not theoretically impossible for someone to be  $68\frac{3}{4}$  inches tall.

In summary, it is possible to make a list of the possible values of a discrete random variable, but this is not true for a continuous random variable.

For completeness, we also point out that it is possible for data to be a mixture of discrete and continuous types. Let  $Y$  = the total measurable daily precipitation measured at Philadelphia International Airport. On some fraction of all days, roughly 70% of them, there is no precipitation. So  $P(Y = 0) \approx .7$ . But considering only those dates with measurable precipitation,  $Y$  is continuous, i.e., the distribution of  $(Y \mid Y > 0)$  is continuous.

### 3.2.2 Displaying Probability Distributions

The display of a probability distribution varies according to whether the r.v. is discrete or continuous. We can make an ordered list of the possible values of a discrete r.v. For example, if  $X$  denotes the number of heads in two tosses of a fair coin, then  $X$  has three possible values  $\{0,1,2\}$ . We will see later that for this coin, the probabilities are as given in the following table:

| $x$ | $P(X = x)$ |
|-----|------------|
| 0   | .25        |
| 1   | .50        |
| 2   | .25        |

Notice that the total probability, 1, has been *distributed* to the three possible values.

However, is not possible to list the possible values of a continuous r.v. We cannot list all conceivable archer heights.

Sometimes we choose to study several interdependent random variables at the same time. In such instances, we require their bivariate or multivariate probability distribution.

We now consider an example of a discrete bivariate and conditional distribution. Here p.m.f. stands for *probability mass function*.

TABLE 3.2. Example of Discrete Bivariate and Conditional Distributions

| Joint p.m.f. $f(x, y)$   |     |     |     |                          |
|--------------------------|-----|-----|-----|--------------------------|
| $x$                      | $y$ |     |     | $y(x) = x\text{-margin}$ |
|                          | 0   | 1   | 2   |                          |
| 1                        | .10 | .20 | .30 | .60                      |
| 2                        | .05 | .10 | .25 | .40                      |
| $g(y) = y\text{-margin}$ | .15 | .30 | .55 | 1.00                     |

| Conditional p.m.f. $f(x   y)$ |               |               | Conditional p.m.f. $g(y   x)$ |     |               |               |               |   |
|-------------------------------|---------------|---------------|-------------------------------|-----|---------------|---------------|---------------|---|
| $x$                           | $y$           |               | $x$                           | $y$ |               | all           |               |   |
|                               | 0             | 1             |                               | 0   | 1             |               |               |   |
| 1                             | $\frac{2}{3}$ | $\frac{2}{3}$ | $\frac{6}{11}$                | 1   | $\frac{1}{6}$ | $\frac{2}{6}$ | $\frac{3}{6}$ | 1 |
| 2                             | $\frac{1}{3}$ | $\frac{1}{3}$ | $\frac{5}{11}$                | 2   | $\frac{1}{8}$ | $\frac{2}{8}$ | $\frac{5}{8}$ | 1 |
| all                           | 1             | 1             | 1                             |     |               |               |               |   |

Here  $X$  and  $Y$  are dependent r.v.'s because, e.g.,  $f(1, 0) = .10$ , which differs from  $f(1) \times g(0) = .60 \times .15 = .09$ . Alternatively,  $f(1 | 0) = \frac{2}{3}$ , which differs from  $f(1) = .6$ . In general, if  $U$  and  $V$  are discrete random variables, then  $U$  and  $V$  are independent r.v.'s if

$$P((U = u) \cap (V = v)) = P(U = u) \times P(V = v)$$

for all possible values  $u$  of  $U$  and  $v$  of  $V$ , i.e., the distribution of  $U$  doesn't depend on the value of  $V$ .

The probability distribution of a continuous random variable cannot be described in the above manner (listing its possible values alongside their associated probabilities) because a continuous r.v. has an *uncountably infinite* number of possible values. Instead the probability distribution of a continuous r.v.  $X$  is described by its probability density function (p.d.f.), say  $f(x)$ . This function has the properties that

1.  $f(x) \geq 0$
2. the probability that  $X$  lies in any interval is given by the area under  $f(x)$  above this interval.

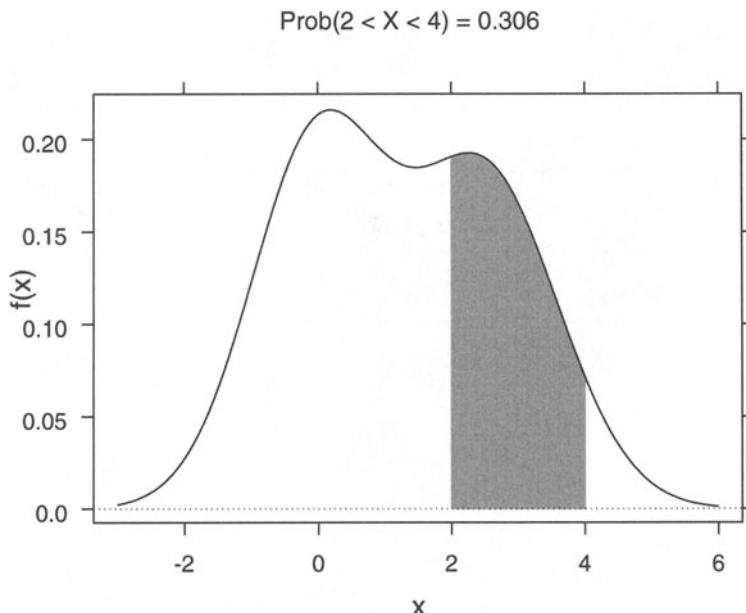


FIGURE 3.1.  $P(2 < X < 4)$  equals the area under the density between 2 and 4.  
 (conc/code/simple-pdf.s), (conc/figure/bimodal.shade.eps.gz)

In the p.d.f. in Figure 3.1, the shaded area under the density and above the horizontal axis represents the probability that the random variable lies between 2 and 4.

The cumulative distribution  $\mathcal{F}$  of a continuous random variable is calculated as

$$\mathcal{F}(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$$

The cumulative distribution  $\mathcal{F}$  of a discrete random variable is calculated as

$$\mathcal{F}(x) = P(X \leq x) = \sum_{t \leq x} f(t)$$

where the sum is taken over all possible values  $t$  of  $X$  that are less than or equal to  $x$ .

Continuous r.v.'s  $U$  and  $V$  are also independent if the distribution of  $U$  doesn't depend on the value of  $V$  or, equivalently, if the distribution of  $V$  doesn't depend on the value of  $U$ . In this case, we can express the

independence condition as

$$P((U \leq u) \cap (V \leq v)) = P(U \leq u) \times P(V \leq v) \quad (3.4)$$

for all  $u$  and  $v$ .

Appendix D catalogs frequently encountered probability distributions, including details about how to perform calculations with them using SAS and S-PLUS.

## 3.3 Concepts That Are Used When Discussing Distributions

### 3.3.1 Expectation and Variance of Random Variables

The expectation of an r.v.  $X$ , denoted  $E(X)$ , is its expected or long-run average value; alternatively it is the mean of the probability distribution of  $X$  and so we write  $E(X) = \mu$ . If  $X$  is discrete with p.m.f.  $p(x)$ , then  $E(X) = \sum x p(x)$ . If  $X$  is continuous, then  $E(X) = \int x f(x) dx$ , where the range of integration extends over the set of real numbers that  $X$  may assume. The variance of  $X$  is defined by  $\sigma^2 = \text{var}(X) = E(X - \mu)^2 = E(X^2) - \mu^2$ . The square root  $\sigma$  of the variance is called the *standard deviation*, abbreviated s.d. It is a more useful measure of variability than the variance because it is measured in the same units as  $X$ , rather than in artificial squared units.

If  $x_1, x_2, \dots, x_n$  is a random sample of  $n$  items selected from some population, the sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3.5)$$

estimates the population mean  $\mu$ , and the sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3.6)$$

estimates the population variance  $\sigma^2$ . In addition, the sample standard deviation  $s = \sqrt{s^2}$  estimates the population standard deviation  $\sigma$ .

It can be shown that if  $a_1$  and  $a_2$  are constants and  $x_1$  and  $x_2$  are any two random variables, then

$$E(a_1 x_1 \pm a_2 x_2) = a_1 E(x_1) \pm a_2 E(x_2) \quad (3.7)$$

If, in addition,  $x_1$  and  $x_2$  are uncorrelated random variables, then

$$\text{var}(a_1 x_1 \pm a_2 x_2) = a_1^2 \text{var}(x_1) + a_2^2 \text{var}(x_2) \quad (3.8)$$

### 3.3.2 Median of Random Variables

The median of an r.v.  $X$ , denoted  $\text{median}(X) = \eta$ , is the middle value of the distribution. The population median is defined as the value  $\eta$  such that

$$\int_{-\infty}^{\eta} f(x) dx = .5 \quad \text{for continuous distributions} \quad (3.9)$$

or

$$\sum_{x \leq \eta} p(x) \leq .5 \quad \text{for discrete distributions.} \quad (3.10)$$

The order statistics  $X_{(i)}$  are the values of the observed  $X_i$  ordered from smallest to largest. The middle order statistic  $\dot{\bar{X}}$  is called the sample median and is defined as

$$\dot{\bar{X}} = \begin{cases} X_{(\frac{n+1}{2})} & \text{odd } n \\ (X_{(\frac{n}{2})} + X_{(\frac{n+1}{2})})/2 & \text{even } n \end{cases} \quad (3.11)$$

The notation  $\dot{\bar{X}}$  for the sample median is intended to be self-descriptive, with an overbar split in the middle into two equal halves. We believe the notation is due to Tukey.

### 3.3.3 Symmetric and Skewed Distributions

Symmetry and skewness are classifications applicable to both continuous and discrete distributions. The mean of a symmetric distribution coincides with its median. A continuous distribution example is the normal distribution having a density function such as that plotted in Figure 3.8. A symmetric distribution has equivalent behavior on either side of its mean. In particular, its *tails*, the values of the density function away from the center, are mirror images.

A skewed distribution is one that is not symmetric. Unimodal distributions (ones having a single point where the probability mass is higher than at adjacent points) that are skewed are further classified as being positively or negatively skewed. A positively skewed distribution has a long, thin tail on its right side and a short, fat tail on its left side. Its mean exceeds its median. A negatively skewed distribution has a long, thin tail on its left side and a short, fat tail on its right side. Its median exceeds its mean. Note that the left/right naming convention for skewed distributions is based on the side containing the long, thin tail. We illustrate a negatively skewed, symmetric, and positively skewed distribution in Figure 3.2.

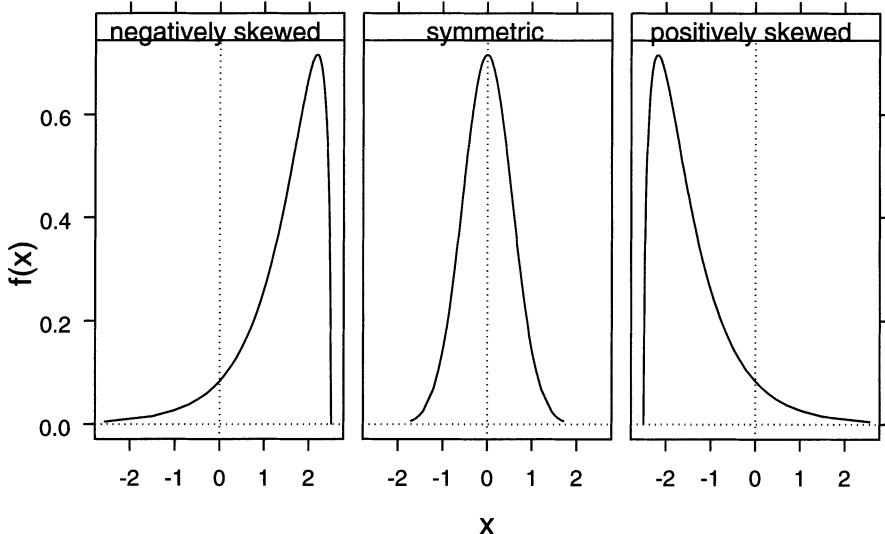


FIGURE 3.2. Negatively skewed, symmetric, and positively skewed distributions.  
 (conc/code/simple-pdf.s), (conc/figure/skewdens2.eps.gz)

The  $\chi^2$  distribution described in Section D.1 is an example of a continuous positively skewed distribution. The (discrete) binomial distribution to be described in Section 3.4.1 is negatively skewed, symmetric, or positively skewed according to whether its parameter  $p$  is less than, equal to, or greater than 0.5.

The skewness terminology often comes into play because many statistics procedures work best when underlying distributions are symmetric, and tactics that move the distribution toward symmetry (for example, with data transformations such as the power transformations described in Section 4.7) are frequently used in the analysis of skewed distributions.

Each of the densities in Figure 3.2 has a single mode. Some densities have more than one mode. Figure 3.1 is an example of a bimodal density, with one mode between 0 and 1 and another mode between 2 and 3. Multimodal distributions, ones having more than two modes, are occasionally encountered. Sometimes bimodality and multimodality arise as a result of interpreting samples coming from two or more populations with different locations as having arisen from a single population. Therefore, bimodality or multimodality may suggest a need for disaggregation of samples.

### 3.3.4 Displays of Univariate Data

It is difficult to gain an understanding of data presented as a list. Summary statistics such as those presented in the preceding sections are helpful for this purpose but may fail to capture some important features. In this section we present three displays for univariate data that are basic tools for studying both the distributional shape and unusual data values. We illustrate these displays with the variable `male.life.exp`, 1990 male life expectancy in each of 40 countries, part of the datafile (`datasets/tv.dat`) examined in more detail in Section 4.5.

#### 3.3.4.1 Histogram

The construction of a histogram begins with a frequency distribution, a partitioning of the data into  $k$  mutually exclusive and nonoverlapping categories, and a tally of the number or proportion of items in each category. Usually the number of categories is between 6 and 12—the use of fewer than 6 categories tends to undersummarize the data while the use of more than 12 categories tends to oversummarize the data. For `male.life.exp` we choose 6 categories that encompass all 40 countries and display the frequencies in Table 3.3.

The corresponding histogram in Figure 3.3 is a graph consisting of rectangles with width covering the breadth of the classes and heights equal to the class frequencies.

Alternatively we could have displayed a *relative frequency* histogram, which would have the same appearance as Figure 3.3 but with vertical axis giving the *proportion* of countries in each category, for example  $\frac{6}{40} = 0.15$  in the first category for ages 50–54.

Figure 3.3 is an example of a bimodal distribution, one having two peaks. In this example, the lower peak may correspond to economically poorer countries and the upper peak to wealthier countries, with relatively few countries falling between these extremes. In general, bimodal distributions

TABLE 3.3. Frequency Distribution of Male Life Expectancy

| <code>male.life.exp</code> | Frequency |
|----------------------------|-----------|
| 50–54                      | 6         |
| 55–59                      | 4         |
| 60–64                      | 9         |
| 65–69                      | 11        |
| 70–74                      | 7         |
| 75–79                      | 3         |
| Total                      | 40        |

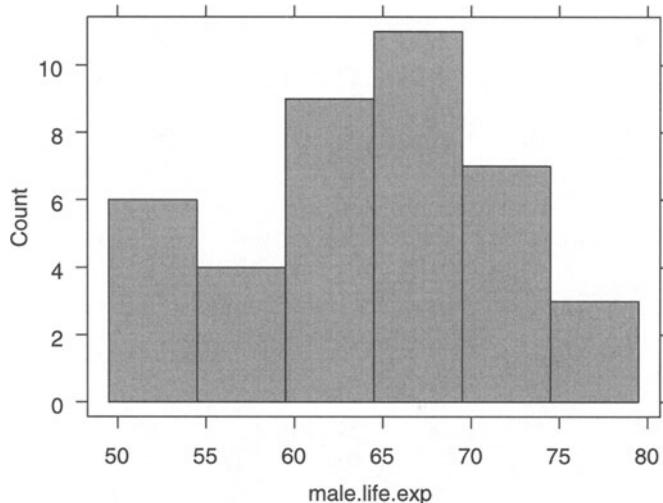


FIGURE 3.3. Life Expectancy for Males  
`(conc/code/tv-graphs.s)`, `(conc/figure/tv-hist.eps.gz)`

sometimes suggest an amalgamation of samples from two separate populations that perhaps should be investigated separately. An advantage of histograms is that they can be constructed from huge datasets with no more effort than small data sets. A disadvantage is that the data used to construct a histogram cannot be recovered from the histogram itself.

#### 3.3.4.2 Stem-and-Leaf Display

Stem-and-leaf displays resemble histograms in that they portray the shape of a distribution. The stem-and-leaf display is usually preferable because it is possible to recover the data used to construct a stem-and-leaf display (at least to some degree of precision). Unlike histograms, stem-and-leaf displays are limited to data sets of not more than a few hundred observations in order that the display fits entirely on one page or one computer monitor.

A stem-and-leaf display for male life expectancy is in Table 3.4.

The numbers in the third column of this display, immediately to the left of the colons, represents the tens digit of each of the life expectancies. This column is the *stem*. The numbers to the right of the colons, one for each countries, are the leaves, the unit digits of the life expectancies for the 40 countries. A  $90^\circ$  counterclockwise rotation of the stem and leaves gives a picture that closely resembles Figure 3.3. The second column in this display

TABLE 3.4. Stem-and-Leaf Display of Male Life Expectancy  
(conc/code/tv-graphs.s)

---

```
> stem(tv$male.life.exp, nl=5, scale=-1, twodig=F, depth=T)

N = 40 Median = 66
Quartiles = 59.5, 70

Decimal point is 1 place to the right of the colon

    6     6   5 : 002234
    10    4   5 : 6799
    19    9   6 : 012223344
          11   6 : 66777888899
    10    7   7 : 1223334
      3    3   7 : 556
```

---

gives the number of leaves emanating from each portion of the stem. The first column shows cumulative totals of the second column both below and above the class containing the median life expectancy. The legend locating the decimal point tells the reader that 5:0 in the display stands for 50, rather than .05 or 500.

Stem-and-leaf displays can accommodate measurements containing more than two significant digits. This is accomplished either by suppressing the values of trailing digits or by allowing more than a single digit for each leaf. For example, suppose in a different problem the measurement is 568. This can be represented as 5 : 6, ignoring the units digit with a legend locating the decimal point 2 places to the right of the colon. Alternatively, it can be represented as 5 : 68, again locating the decimal point two places to the right of the colon.

#### 3.3.4.3 Boxplots

Boxplots, also known as box-and-whisker plots, are among the many inventions of John Tukey. Their main use is as a compact, simultaneous display to compare several related data sets. Many examples of side-by-side boxplots appear in Chapter 4 and elsewhere in this book. Boxplots may be arranged along either a vertical or horizontal scale. This book contains examples illustrating both options.

Boxplots make use of the sample median  $\bar{x}$ , first quartile  $Q_1$ , and third quartile  $Q_3$ . The statistics  $Q_1, \bar{x}, Q_3$  divide the sample into four equal parts.

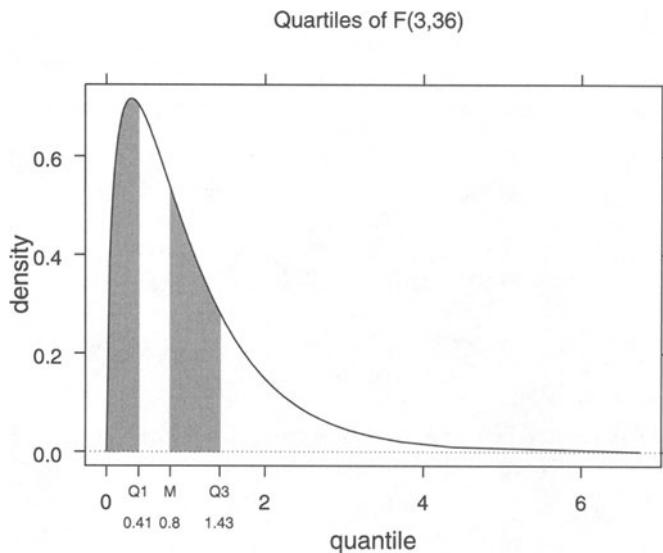


FIGURE 3.4. Illustration of median and quartiles for a continuous distribution.  
 (conc/code/simple-pdf.s), (conc/figure/quartiles.eps.gz)

$Q_1$  is the median of the sample values that are less than or equal to  $\bar{x}$  and  $Q_3$  is the median of the sample values that are greater than or equal to  $\bar{x}$ . At least 25% of the sample lies within each of the four intervals

$$(-\infty, Q_1], \quad [Q_1, \bar{x}], \quad [\bar{x}, Q_3], \quad [Q_3, \infty)$$

See the illustration in Figure 3.4 for a continuous distribution. The interquartile range

$$\text{IQR} = Q_3 - Q_1$$

is a measure of dispersion of the central portion of a distribution. When  $X$  is normally distributed  $X \sim N(\mu, \sigma^2)$ , we have  $\text{IQR} = 1.34898\sigma$ .

A rectangle (box) is drawn so that when placed against a numerical scale its edges occur at  $Q_1$  and  $Q_3$ . A line is drawn, parallel to the edges, through the inside of the box at the median  $\bar{x}$ . Lines perpendicular to the edges of the box extend outward from the midpoints of the edges. These lines are sometimes called “whiskers”. The lower whisker extends to the lowest sample item not more than  $1.5 \times \text{IQR}$  below  $Q_1$ . The upper whisker extends to the largest sample item not more than  $1.5 \times \text{IQR}$  above  $Q_3$ . Points outside the range of the whiskers are plotted as filled-in circles. Such points are deemed extreme or outlying values (“outliers”). In general, outliers should be carefully scrutinized. Sometimes they are due to transcription errors and

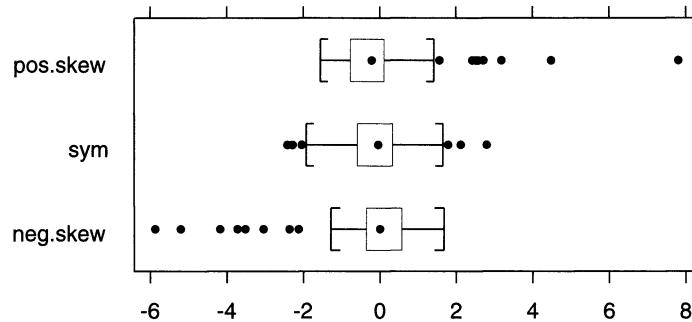


FIGURE 3.5. Boxplots illustrating negatively skewed, symmetric, and positively skewed distributions. (conc/code/skew.s), (conc/figure/skew.eps.gz)

are not legitimately part of the data under consideration (in which case you should attempt to correct the data). Other times, they are the critical data points that provide the key to an explanation of the study.

Figure 3.5 contains parallel boxplots depicting three samples on a common scale, illustrating the distinctions between boxplots for negatively skewed, symmetric, and positively skewed distributions. This parallels the density presentations in Figure 3.2.

Several more elaborate versions of the boxplot exist. For example, adding a *notch* to the sides of a box provides information on the variability of the sample median. For details, see (Hoaglin et al., 1983).

Boxplots are generally unsuccessful in conveying the existence of multiple modes. For such data, histograms and stem-and-leaf displays are often preferred choices.

### 3.3.5 Multivariate Distributions—Covariance and Correlation

In the previous section we give an example of a discrete multivariate (actually bivariate) probability distribution. We now touch on the notion of the continuous multivariate distribution of a continuous random vector  $X = (X_1, X_2, \dots, X_p)'$ . The mean or expectation of  $X$  is  $\mu = (\mu_1, \mu_2, \dots, \mu_p)'$ , the vector of means of the univariate distribution of the  $X_i$ 's. The variance-covariance matrix of  $X$ , say  $V$ , also called the covariance matrix or dispersion matrix, is the symmetric  $p \times p$  matrix having the variances of the  $X_i$ 's on its main diagonal, and the covariances of different

$X'_i$ 's elsewhere. The covariance of  $X_i$  and  $X_j$  is

$$V_{ij} = \sigma_{ij} = \text{cov}(X_i, X_j) = E((X_i - \mu_i)(X_j - \mu_j))$$

is the element in the row  $i$  column  $j$  position of  $V$ . If we denote the standard deviations of  $X_i$  and  $X_j$  by  $\sigma_i$  and  $\sigma_j$ , respectively, then the *correlation* between  $X_i$  and  $X_j$  is

$$\rho_{ij} = \frac{\text{cov}(X_i, X_j)}{\sigma_i \sigma_j} = \frac{V_{ij}}{\sqrt{V_{ii} V_{jj}}}$$

This is a rescaling of the covariance, interpreted as a measure of the strength of the (straight line) linear relationship between  $X_i$  and  $X_j$ . It can be shown that  $-1 \leq \rho_{ij} \leq 1$ . If this correlation is close to  $\pm 1$ ,  $X_i$  and  $X_j$  are closely linearly associated; the association is direct if  $\rho_{ij} > 0$  and inverse if  $\rho_{ij} < 0$ . If  $\rho_{ij} = 0$ , then  $X_i$  and  $X_j$  are said to be *uncorrelated*, i.e., the  $X$ 's are not linearly related. Figure 3.6 illustrates the interpretation of the correlation coefficient. A dynamic illustration of the effect of the correlation coefficient can be constructed by plotting a sequence of panels similar to those in Figure 3.6 and cycling through them. We give the code for this construction in S-PLUS in file (`conc/code/corrscat.s`).

Matrix algebra plays an important role in the study of multivariate distributions. For example, in matrix notation, the covariance matrix is

$$V = E((X - \mu)(X - \mu)')$$

and the correlation matrix  $P$  (uppercase  $\rho$ ) is given by

$$P = (\text{diag}(V))^{-\frac{1}{2}} V (\text{diag}(V))^{-\frac{1}{2}}$$

When the individual  $x_i$  are normally distributed, their joint distribution is called the multivariate normal and is notated  $x \sim N(\mu, V)$ . The bivariate

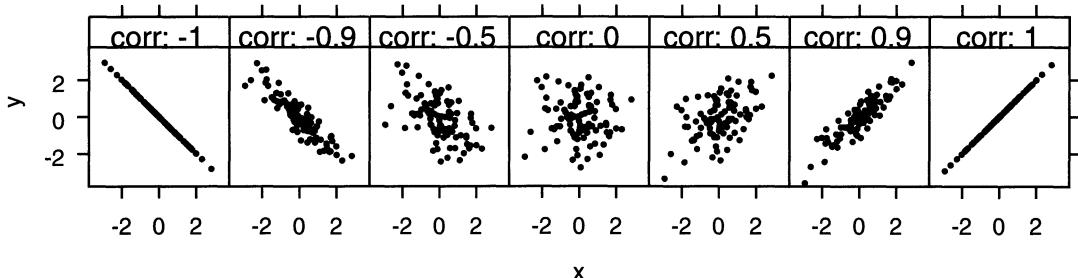


FIGURE 3.6. Bivariate Normal distribution—scatterplot at various correlations. The distributions in the panels are related. The  $x$ -variable in all panels is the same. The  $y$  are generated from a common  $e$ -variable by the formula  $y = \rho x + (1 - \rho^2)^{\frac{1}{2}} e$  for a sequence of values for  $\rho$ . The  $x$ - and  $e$ -variables were independently generated from the  $N(0,1)$  distribution.

(`conc/code/corrscat.s`), (`conc/figure/corr.eps.gz`)

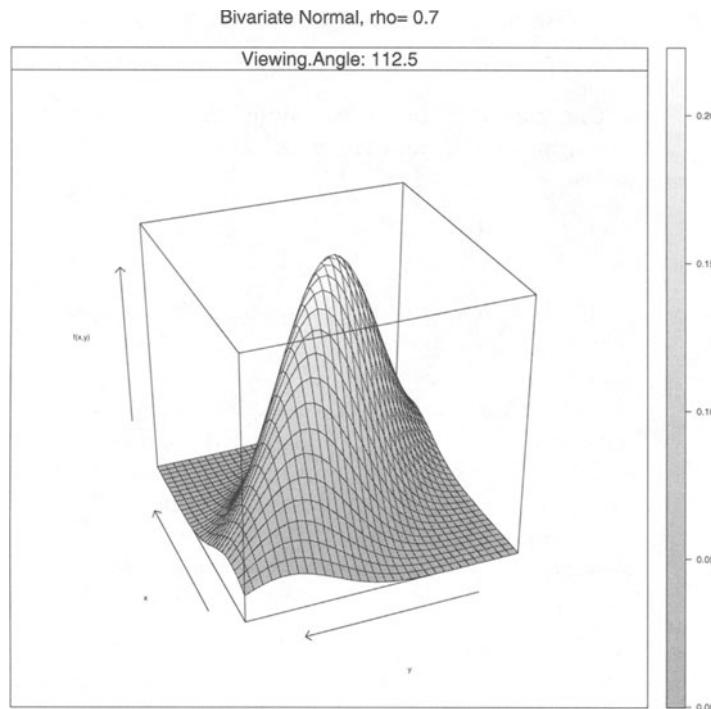


FIGURE 3.7. Bivariate Normal density with  $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \rho) = (0, 1, 0, 1, .7)$  in 3D space with viewing angle  $= 112.5^\circ$ . A complete set of eight viewing angles is available in the files (`conc/figure/bivnorm.eps.gz`) and (`conc/figure/bivnorm-color.eps.gz`).  
`(conc/code/bivnorm.s)`, (`conc/figure/bivnorm1125.eps.gz`),  
`(conc/figure/bivnorm1125-color.eps.gz)`

$(p = 2)$  normal distribution with means  $\mu_i = 0$ , variances  $\sigma_i^2 = 1$ , and correlation  $\rho = .7$  [hence  $V = \begin{pmatrix} 1.0 & 0.7 \\ 0.7 & 1.0 \end{pmatrix}$ ] is plotted as a three-dimensional object in Figure 3.7. This is actually one panel of a set of rotated views of the density. The complete set can be constructed with the function `example.bivariate.normal` in file (`conc/code/bivnorm.s`). See the file for details.

A rotating version of the bivariate normal runs in S-PLUS with the function in file (`conc/code/bivnorm-rotate.s`).

If  $X$  and  $Y$  are random vectors with

$$Y = B + CX$$

for some vector  $B$  and some matrix  $C$ , then

$$E(Y) = B + C E(X)$$

and

$$\text{var}(Y) = C \text{ var}(X) C'$$

If, moreover,  $X$  has a multivariate normal distribution, then so does  $Y$ . In other words, linear functions of normal r.v.'s are normal.

If  $Y$  has a multivariate normal distribution with mean  $\mu$  and covariance matrix  $V$ , then

$$Q = (Y - \mu)' V^{-1} (Y - \mu)$$

has a  $\chi^2$  distribution with  $k$  degrees of freedom.

## 3.4 Three Probability Distributions

In this section we introduce three probability distributions, the (discrete) binomial distribution and the (continuous) Normal and  $t$  distributions, that frequently arise in practice. Details of how to perform probability-related calculations for these and other frequently encountered distributions are discussed in Appendix D.

### 3.4.1 The Binomial Distribution

The binomial distribution is perhaps the most commonly encountered discrete distribution in statistics. Consider a sequence of  $n$  independent trials, or mini-experiments, each of which can result in one of just two possible outcomes. For convenience these outcomes are labeled *success* and *failure* although in context the success outcome may not connote a favorable event. Further assume that the probability of success,  $p$ , is the same for each trial. Let  $X$  denote the number of successes observed in the  $n$  trials. Then  $X$  has a binomial distribution with parameters  $n$  and  $p$ . This distribution has mean  $\mu = np$  and standard deviation  $\sigma = \sqrt{np(1-p)}$ .

The above scenario is widely applicable. If one randomly samples with replacement from a population with a proportion  $p$  of successes, then the number of successes in the sample is binomially distributed. Even if the sampling is *without* replacement, the number of successes is approximately binomial if the population size is much greater than the sample size; in this case the first two assumptions above are only mildly violated. Applications include the number of voters favoring a candidate in a political poll, the number of patients in a population that suffer from a particular illness, and the number of defective items in one day's output from an assembly line.

However, it is not unusual for one or more of the binomial assumptions to be violated. For example, suppose we sample *without* replacement from a

population of successes and failures and the population size is not much greater than the sample size, say less than 20 times as large as the sample. Then the trials are not independent and the *success* probability is not constant from trial to trial. (In this situation the correct distribution to use for  $X$  is the *hypergeometric distribution*.)

Similarly, the binomial model is unlikely to apply to the number of hits by the archer in Section 3.2.1 because her shots (trials) may not be independent and may not have the same probability of a hit.

Usually in practice, we need to calculate not just  $P(X = x)$ , the probability of achieving *exactly*  $x$  successes, but probabilities of an interval of successes such as  $P(X \leq x)$ , the probability of *at most*  $x$  successes, or  $P(a \leq X \leq b)$ , the probability of observing between  $a$  and  $b$  successes inclusive.

If we have available a table of binomial probabilities, this can be used if  $n$  and  $p$  appear in the table. Otherwise, as explained in Appendix D, SAS and S functions can easily be used to produce accurate results almost instantaneously even for large values of  $n$ .

### 3.4.2 The Normal Distribution

Many natural phenomena follow this distribution, whose probability density function is the familiar “bell-shaped” curve, symmetric about the mean  $\mu$ . In addition, a celebrated theoretical result called the Central Limit Theorem says that the sampling distributions of sample means, sample proportions, and sample totals each are approximately normally distributed if the sample size is “sufficiently large.” Since this theorem applies to almost all possible probability distributions from which a sample might be selected, including discrete distributions, the theorem brings the normal distribution into play in a wide variety of circumstances.

If  $X$  has a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ , and we define the standardization of  $X$  as  $Z = \frac{X-\mu}{\sigma}$ , then  $Z$  is normally distributed with mean 0 and standard deviation 1, *i.e.*, the *standard normal distribution*. We write  $X \sim N(\mu, \sigma^2)$  to indicate that  $X$  has a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ . In this notation, the standard normal distribution is  $N(0, 1)$ .

The normal distribution is “bell-shaped” and symmetrically distributed about  $\mu$ , which is also this distribution’s median and mode. Almost all of the probability is concentrated in the interval  $\mu \pm 3\sigma$ . We use  $z_\alpha$  to be the solution to the equation  $P(Z > z_\alpha) = \alpha$ . This is the value on the horizontal axis that has area  $\alpha$  under the curve and to its right. For example,  $z_{.05} = 1.645$ . Figure 3.8 shows the normal density function for a  $N(100, 25)$  distribution. If  $X$  has this distribution, the shaded area in

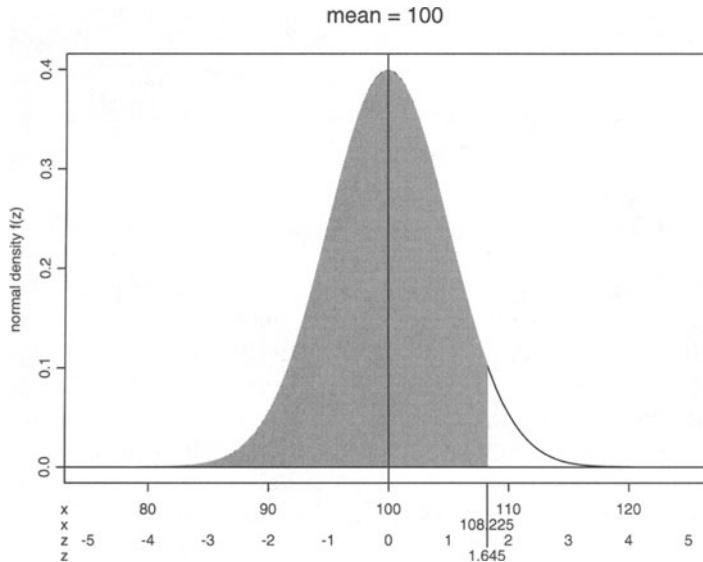


FIGURE 3.8. Plots a normal curve centered on the assumed true mean  $\mu = 100$ . We assume  $\sigma = 5$  and  $\alpha = .05$ . Shaded area is  $.95 = P((X - \mu)/\sigma \leq \Phi^{-1}(1 - \alpha) = 1.645)$ .  
`(conc/code/normpdf.s), (conc/figure/normpdf.eps.gz)`

Figure 3.8 represents 95% of the area under the density function. That is,

$$P(Z < 1.645) = P\left(\frac{X - \mu}{\sigma} < 1.645\right) = P(X < 108.225) = .95$$

after substituting  $\mu = 100$  and  $\sigma = 5$ .

### 3.4.3 The (Student's) $t$ Distribution

This distribution is similar to the standard normal distribution in that its density is a bell-shaped curve symmetric about 0. However, as compared to the normal distribution, its probability density function is lower in the center and “heavier” in the tails. If the mean of a sample of size  $n$  is standardized with a sample standard deviation  $s$  rather than with a population standard deviation  $\sigma$ , then the resulting standardization,  $\frac{\bar{X} - \mu}{s/\sqrt{n}}$ , has a Student’s  $t$  distribution with *degrees of freedom* parameter  $n - 1$ . The  $t$  distribution is used for inference on population means and regression coefficients.

That  $\frac{\bar{X} - \mu}{s/\sqrt{n}}$  has a  $t$  distribution rests on the fact that  $\bar{X}$  and  $s$  are independent random variables when sampling from a normal population.

As the sample size  $n$  and hence the degrees of freedom get large, the sample standard deviation  $s$  increasingly approximates  $\sigma$  so that  $\frac{\bar{X}-\mu}{s/\sqrt{n}}$  increasingly approximates  $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ . In other words, as the degrees of freedom increases, a  $t$  distribution increasingly resembles a standard normal distribution.

### 3.5 Sampling Distributions

In Chapter 1 we learn that knowledge about characteristics of populations can be gleaned from analogous characteristics of random samples from these populations. Also recall that population characteristics are called parameters and sample characteristics are called statistics. In the next two sections we discuss the two main techniques for using statistics to infer about parameters, estimation, and hypothesis testing. Implementation of these techniques requires that we use knowledge about the likely values of statistics. Such information about statistics is contained in their *sampling distribution*. The sampling distribution of a statistic depends on our assumed knowledge of the distribution of values in the population to which we are inferring. The term *standard error* is used to refer to the standard deviation of a sampling distribution.

Consider first the mean  $\bar{X}$  of a sample of  $n$  items randomly selected from a normal population,  $N(\mu, \sigma^2)$ . It can be shown that the sampling distribution of  $\bar{X}$  is also normally distributed with this same mean but with a much smaller variance:

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

This result says that both an individual sampled item and the mean of  $n$  sampled items are on average equal to the population mean, but that the sample mean is typically much closer to the population mean than is an individual sampled item. When  $\sigma^2$  is a known quantity, we use this Normal distribution to make probability statements about the closeness of  $\bar{X}$  to  $\mu$ . In the more likely situation where  $\sigma^2$  is unknown, analogous probability statements are made with reference the Student's  $t$  distribution.

Next suppose that the population is not necessarily normal. Then under fairly general conditions, a statistical theory result called the Central Limit Theorem states that  $\bar{X}$  has “approximately” a  $N(\mu, \sigma^2/n)$  distribution if the sample size  $n$  is “sufficiently large”. Thus, the inferential statements concerning  $\mu$  made in the normal distribution case are also approximately valid if the population is not normal.

What is meant here by “approximately” and “sufficiently large”? We mean that the closer the population is to a normal population the smaller the

sample size needs to be for the approximation to be acceptably accurate. Unless the population is multimodal or severely skewed, a sample size of 30 to 50 is usually sufficient for the approximation to hold.

Another application of the Central Limit Theorem implies that the sampling distribution of the proportion  $\hat{p} = X/n$  of successes in  $n$  binomial trials is approximately normally distributed with mean  $\mu = np$  and variance  $\sigma^2 = npq$ , where  $q = 1 - p$ . This result is used for inferences concerning the proportion of successes in a dichotomous population where the binomial assumptions apply.

If  $S^2$  is the variance of a random sample of size  $n$  from a normal population having variance  $\sigma^2$ , then the sampling distribution of  $(n - 1)S^2/\sigma^2$  is  $\chi^2$  with  $n - 1$  degrees of freedom. We use this result for inferences concerning the population standard deviation  $\sigma$ .

## 3.6 Estimation

A fundamental task of statistical analysis is inference of the characteristics of a large population from a sample of  $n$  items or individuals selected at random from the population. Sampling is commonly undertaken because it is

- a. cheaper and
- b. less prone to error

than examining the entire population. Estimation is one of the two broad categories of statistical techniques used for this purpose. The other is hypothesis testing, discussed in Section 3.7.

An *estimator* is a formula that can be evaluated with numbers from the sample. When the sample values are plugged into the formula, the result becomes an *estimate*. An estimator is a particular example of a statistic.

### 3.6.1 Statistical Models

A key component of statistical analysis involves proposing a statistical model. A statistical model is a relatively simple approximation to account for complex phenomena that generate data. A statistical model consists of one or more equations involving both random variables and parameters. The random variables have stated or assumed distributions. The parameters are unknown fixed quantities. The random components of statistical models account for the inherent variability in most observed phenomena.

Subsequent chapters of this book contain numerous examples of statistical models.

The term *estimation* is used to describe the process of determining specific values for the parameters by fitting the model to the data. This is followed by determinations of the quality of the fit, often via hypothesis testing or evaluation of an index of goodness-of-fit.

Model equations are often of the form

$$\text{data} = \text{model} + \text{residual}$$

where **model** is an equation that explains most of the variation in the data, and **residual**, or lack-of-fit, represents the portion of the data that is not accounted for by the model. A good-quality model is one where **model** accounts for most of the variability in the data, that is, the data are well-fitted by the model.

A proposed model provides a framework for the statistical analysis. Experienced analysts know how to match models to data and the method of data collection. They are also prepared to work with a wide variety of models, some of which are discussed in subsequent chapters of this book. Statistical analysis then proceeds by estimating the model and then providing figures and tables to support a discussion of the model fit.

### 3.6.2 Point and Interval Estimators

There are essentially two types of estimation: point estimation and interval estimation. Point estimates are single numbers calculated from the sample. Interval estimates are intervals within which the parameter is expected to fall, with a certain degree of confidence. Interval estimates are generally more useful than point estimates because they indicate the precision of the estimate. Often interval estimates are of the form:

$$\text{point estimate} \pm \text{constant} \times \text{standard error}$$

where “standard error” is the observed standard deviation of the statistic used as the point estimate. The constant is a percentile of the standardized sampling distribution of the point estimator.

### 3.6.3 Criteria for Point Estimators

There are a number of criteria for what constitutes “good” point estimators. Here is a heuristic description of some of these.

**unbiasedness:** the mean of the sampling distribution of the estimator is the parameter being estimated. Not too crucial if the bias, defined as:

$$\text{bias} = \text{mean of sampling distribution} - \text{parameter}$$

is small and if the bias decreases with increasing  $n$ . The sample mean  $\bar{x}$  is an unbiased estimator of the population mean  $\mu$  and the sample variance  $s^2$  is an unbiased estimate of the population variance  $\sigma^2$ . The sample standard deviation  $s$  is a biased estimator of the population standard deviation  $\sigma$ . However, the bias of  $s$  decreases toward zero as the sample size increases; we say that  $s$  is an *asymptotically unbiased* estimator of  $\sigma$ .

**small variance:** higher precision. For example, for estimating the mean of a normal population, the variance of the sample mean is less than the variance of the sample median.

**consistency:** the quality of the estimator improves as  $n$  increases.

**sufficiency:** the estimator fully uses all the sample information. Example: If  $X$  is distributed as continuous uniform on  $[0, a]$ , how would you estimate  $a$ ? Since the population mean is  $a/2$ , you might think that  $2\bar{x}$  is a “good” estimator for  $a$ . The largest item in the sample of size  $n$ , denoted  $x_{(n)}$ , is a better and *sufficient* estimator of  $a$ . This estimator cannot overestimate  $a$  while  $2\bar{x}$  can either underestimate or overestimate  $a$ . If  $x_{(n)}$  exceeds  $2\bar{x}$ , then it must be closer to  $a$  than is  $2\bar{x}$ .

### 3.6.4 Confidence Interval Estimation

A confidence interval estimate of a parameter is an interval that has a certain probability, called its *confidence coefficient*, of containing the parameter. The confidence coefficient is usually denoted  $1 - \alpha$  or as a percentage,  $100(1 - \alpha)\%$ . Common values for the confidence coefficient are 95% and 99%, corresponding to  $\alpha = .05$  or  $.01$ , respectively.

If we construct a 95% confidence interval (CI), what is the meaning of 95%? It is easy to incorrectly believe that 95% is the probability that the CI contains the parameter. This is false because the statement “*CI contains the parameter*” is not an event, but rather a situation that is certainly either true or false. The correct interpretation refers to the *process used to construct the CI*: If, hypothetically, many people were to use this same formula to construct this CI, plugging in the results of their individual random samples, about 95% of the CI’s of these many people would contain the parameter and about 5% of the CI’s would exclude the parameter.

It is important to appreciate the tradeoff between three quantities:

- confidence coefficient (the closer to 1 the better)

- interval width (the narrower the better)
- sample size (the smaller the better)

In practice it is impossible to optimize all three quantities simultaneously. There is an interrelationship among the three so that specification of two of them uniquely determines the third. A common practical problem is to seek the sample size required to attain a given interval width and confidence. Examples of such formulas appear in Section 5.6.

### 3.6.5 Example—Confidence Interval on the Mean $\mu$ of a Population Having Known Standard Deviation

The interpretation of the confidence coefficient may be further clarified by the following illustration of the construction of a  $100(1 - \alpha)\%$  confidence interval on an unknown mean  $\mu$  of a normal population having known standard deviation  $\sigma$ , using a random sample of size  $n$  from this population. If  $\bar{X}$  denotes the sample mean, then  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  has a standard normal distribution. Let  $z_{\frac{\alpha}{2}}$  denote the  $100(1 - \frac{\alpha}{2})^{\text{th}}$  percentile of this distribution. Then

$$P\left(-z_{\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

After a bit of algebraic rearrangement, this becomes

$$P\left(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

The endpoints of the interval  $\left(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$  are random variables, so the probability statement refers to the probability that the interval contains the parameter, not the probability that the parameter is contained in the interval.

In practice, we replace the random variable  $\bar{X}$  with  $\bar{x}$ , the realized value from the sample, and wind up with the  $100(1 - \alpha)\%$  confidence interval for  $\mu$ :

$$\left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) \quad (3.12)$$

### 3.6.6 Example—One-Sided Confidence Intervals

One-sided confidence intervals correspond to one-sided tests of hypotheses. Such intervals have infinite width and therefore are much less commonly used in practice than two-sided confidence intervals, which have finite

width. The rationale for using one-sided intervals matches that for one-sided tests—sometimes the analyst believes the value of a parameter is at least or at most some value rather than on either side. One-sided confidence intervals on the mean of a population having known standard deviation are shown in Table 5.1. Other examples of one-sided confidence intervals appear in Tables 5.2 and 5.3.

## 3.7 Hypothesis Testing

The statistician sets up two competing hypotheses, the null hypothesis  $H_0$  and the alternative hypothesis  $H_1$ , for example in a binomial testing situation,  $H_0: p \leq .5$  vs  $H_1: p > .5$ . The task is to decide whether the sample evidence better supports  $H_0$  (decision to “retain  $H_0$ ”) or  $H_1$  (decision to “reject  $H_0$ ”).

There are two types of errors: the Type I error of rejecting  $H_0$  when  $H_0$  is true, and the Type II error of retaining  $H_0$  when  $H_1$  is true. In the classical hypothesis setup, the statistician prespecifies  $\alpha =$  the maximum probability of committing a Type I error. Subject to this constraint, we select a testing procedure that gives good control over  $\beta =$  the probability of committing a Type II error. This probability is a function of the unknown parameter being tested. A plot of the probability against the parameter is called an *operating characteristic curve* (O.C. curve) of the test.

The *power* of a hypothesis test is the probability of correctly rejecting a false null hypothesis, equivalently, the probability  $1 - \beta$ . A *power curve* is a plot of the probability of rejecting  $H_0$  against the parameter. It contains information identical to that conveyed by an O.C. curve.

Statisticians can determine the sample size needed to conduct a test that has a high probability of detecting a departure from  $H_0$  by studying O.C. or power curves for a variety of proposed sample sizes. Examination of these curves displays the tradeoffs between Type I error, Type II error, and sample size. The calculation of O.C. and power curves is discussed in Section 3.9.

Commonly selected values of  $\alpha$  are .05 or .01. The choice is sometimes governed by what is traditional in a research area.

With the prespecification of  $\alpha$ , the statistician maintains better control over Type I error than Type II error. When we have a choice,  $H_0$  should be taken to be the hypothesis so that the more serious error is the Type I error and the less serious error is a Type II error. Often in many applications,  $H_0$  is essentially the statement that the status quo is better, while  $H_1$  is that an innovation is better. The Type I error of incorrectly deciding

TABLE 3.5. Comparison of Hypothesis Testing with the Decision Options in a Court of Law

| Decision     | True situation |               | Decision | True situation |              |
|--------------|----------------|---------------|----------|----------------|--------------|
|              | $H_0$ true     | $H_0$ false   |          | Innocent       | Guilty       |
| Reject $H_0$ | Type I error   | correct       | Convict  | greater error  | correct      |
| Retain $H_0$ | correct        | Type II error | Acquit   | correct        | lesser error |

in favor of an innovation is typically more serious than the error of incorrectly maintaining the status quo because innovation is usually costly. As a result, classical testing puts the burden of proof on the innovation  $H_1$ ;  $H_0$  is retained unless there is compelling evidence not to do so.

The preceding rules for deciding which hypothesis is  $H_0$  are based on the fact that classical hypothesis testing places more control over Type I error at the cost of reduced control over Type II error. The logic for this approach is seen by comparing in Table 3.5 the definitions of these two errors in the hypothesis testing context with the potential errors in a U.S. courtroom.

In the United States, the error of convicting an innocent defendant is viewed as far more serious than the error of acquitting a guilty defendant. Accordingly, the U.S. legal system places the burden of proof on the prosecution to establish guilt beyond a reasonable doubt. If sufficient evidence is not presented to the court, the defendant is acquitted. Similarly, in hypothesis testing, the burden is placed on the analyst to provide convincing evidence that  $H_0$  is false; in the absence of such evidence,  $H_0$  is accepted. Continuing the analogy, in the hypothesis testing framework, the way to reduce the probability of committing a Type II error without compromising control of Type I error is to seek an increased sample size. In the legal framework, courts can best reduce the probability of acquitting guilty defendants by obtaining as much relevant evidence as possible.

Do not confuse the decision to retain  $H_0$  with the statement that  $H_0$  is true. We might be committing a Type II error. Similarly, the decision to reject  $H_0$  is not the same as saying that  $H_0$  is false because we might be committing a Type I error.

Table 3.5 also demonstrates that if we modify a hypothesis testing procedure to less readily reject a null hypothesis, this results in both greater control of Type I error and reduced control of Type II error.

Tests of hypotheses are conducted by determining what sample results would be likely if  $H_0$  is true. If then a sufficiently unlikely sample statistic is observed, doubt is cast on the truth of  $H_0$ ; i.e.,  $H_0$  is rejected.

Most tests are constructed by calculating a test statistic from a random sample. This is compared to a critical value, or values. If the test statistic

is on one side of the critical value(s),  $H_0$  is retained; if on the other side,  $H_0$  is rejected. If the value of the test statistic leads to rejection of  $H_0$ , the test statistic is said to be (statistically) *significant*.

A criticism of classical hypothesis testing is the requirement that  $\alpha$  be pre-specified. One way around this is to calculate the  $p$ -value of the test. For most testing procedures, this calculation requires the use of the computer. The  $p$ -value is the probability of observing, in hypothetical repeated samples, a value of the test statistic at least as extreme in the direction of  $H_1$  as the test statistic calculated from the present sample. Then we reject  $H_0$  if  $\alpha > p$ -value and retain  $H_0$  otherwise. Then the analyst needs only to know how  $\alpha$  compares with the  $p$ -value, and does not have to commit to a particular value of  $\alpha$ . Most software provides  $p$ -values as part of the output rather than requesting  $\alpha$  as part of the input.

Another criticism of classical hypothesis testing is that if  $H_0$  is barely false, it is always possible to reject  $H_0$  simply by taking a large enough sample size. For example, if we test  $H_0: \mu = 32$ , where  $\mu$  is the mean amount of soda a bottling plant puts into 32-ounce bottles, and if in reality,  $\mu = 32.001$  ounces,  $H_0$  can be rejected even though as a practical matter it makes no sense to act as though anything is wrong with the filling mechanism. This would be an instance of a statistically significant result that is not of practical significance. Because of this criticism, many statisticians are much more comfortable using CI's than tests.

In practice, a very small  $p$ -value may be regarded as sufficiently strong evidence against  $H_0$  to convince us to act as though  $H_0$  is false. However, even in this situation and especially if the sample size is large, we should be mindful of the possibility that one is making a Type I error. Also, we should always be alert to the possibility that an underlying assumption about the population is incorrect; if so, the  $p$ -value calculation may be distorted.

## 3.8 Examples of Statistical Tests

Suppose in the example of the previous section, the standard deviation of fill volume is known to be 0.3 ounces, and that a sample of 100 bottles yields a mean of 31.94 ounces. If the alternative hypothesis is  $H_1: \mu \neq 32$ , then we should reject  $H_0$  if  $\bar{x}$  is sufficiently above or below 32. In this example, in order to maintain Type I error probability at  $\alpha = .01$ , we should reject  $H_0$  if

$$\begin{aligned}\bar{x} &< 32 - z_{.005} \sigma / \sqrt{n} \\ &= 32 - 2.576 (0.3) / 10 = 31.923\end{aligned}$$

or

$$\begin{aligned}\bar{x} &> 32 + z_{.005} \sigma / \sqrt{n} \\ &= 32 + 2.576 (0.3) / 10 = 32.077\end{aligned}$$

Since  $\bar{x}$  meets neither condition, we should retain  $H_0$  when testing at  $\alpha = .01$ . This is an example of a “two-tailed” (or “two-sided”) test because we reject  $H_0$  if  $\bar{x}$  lies sufficiently far on either tail of the  $Z$  distribution with the null hypothesized mean.

At this point we might ask whether a larger choice of  $\alpha$  would have led to the “retain  $H_0$ ” decision. This is answered by finding the  $p$ -value, here equal to  $2P(Z > |z_{\text{calc}}|)$  for  $z_{\text{calc}} = (\bar{x} - \mu_0)/(\sigma/\sqrt{n}) = -2$ . Thus  $p\text{-value} = 2P(Z > 2.00) = 0.046$ . Then any choice of  $\alpha \leq 0.046$  requires retention of  $H_0$ ; i.e., decide that filling machine is in control.

A two-tailed test can be conducted as follows. Reject the null hypothesis at level  $\alpha$  if the null hypothesized value of the parameter lies outside the  $100(1 - \alpha)\%$  confidence interval for the parameter.

Sometimes analysts prefer to conduct a “one-tailed” (or “one-sided”) test where the alternative hypothesis statement is a one-sided inequality. Suppose in the soda bottling example it was felt that the error of incorrectly claiming bottles are being underfilled is much more serious than an error of incorrectly claiming bottles are being overfilled. Then we might test  $H_0: \mu \geq 32$  vs  $H_1: \mu < 32$ , because this way the more serious error is the better controlled Type I error. Now  $H_0$  will be rejected only when  $\bar{x}$  is sufficiently below 32. If once again we take  $\alpha = .01$ , we reject  $H_0$  if

$$\begin{aligned}\bar{x} &< 32 - z_{.01} \sigma / \sqrt{n} \\ &= 32 - 2.326 (0.3) / 10 = 31.302\end{aligned}$$

As with the two-tailed test,  $H_0$  is retained.

Note that, if instead we had had  $\bar{x} = 31.2$  ounces, we would have rejected  $H_0$  with the one-tailed alternative but retained it with the two-tailed alternative. The explanation for this distinction is that the portion of the parameter space where  $H_1$  is true is larger under the two-tailed setup than under the analogous one-tailed setup. Hence the two-tailed setup has more “territory” to protect against a Type I error than does the one-tailed test.

Not all tests of hypotheses fit into the framework of being one- or two-tailed. Examples are goodness-of-fit tests, discussed in Chapter 5, of the form

$H_0$ : the population is of a particular form,

vs

$H_1$ : the population is not of this particular form.

### 3.9 Power and Operating Characteristic (O.C.) Curves

These two types of curves are used to assess the degree of Type II error control of a proposed test. The O.C. curve is a plot of the probability of retaining  $H_0$  vs the parameter being tested, and the power curve is a plot of the probability of rejecting  $H_0$  vs the parameter being tested. These two plots give equivalent information, and the choice of which to use is a matter of taste or tradition in one's discipline.

Power and O.C. curves are used to display the menu of competing choices of sample size,  $\alpha$ , and Type II error probability. One desires that all of these three quantities be as small as possible, but fixing any two of them uniquely determines the third. Analysts commonly use one of these curves to assess the needed sample size to achieve desired control over the two errors. If the required sample size is infeasibly large, the analyst can see what combinations of diminished control over the two errors are possible with the maximum attainable sample size. Note that  $P(\text{Type II error})$  is a function of the true value of the unknown parameter being tested and that  $\alpha$  is the *maximum* probability of committing a Type I error.

We illustrate the formulation of an O.C. curve and its construction using SAS and S-PLUS. SAS uses PROBNORM() and S-PLUS uses pnorm() for the normal c.d.f.  $\Phi$ . SAS uses PROBIT() and S-PLUS uses qnorm() for the inverse normal c.d.f.  $\Phi^{-1}$ .

Consider a situation where we have a normal population with unknown mean  $\mu$  and known s.d.  $\sigma = 2.0$ . Suppose we wish to test  $H_0: \mu \leq 10$  vs  $H_1: \mu > 10$ , using  $\alpha = .05$  and a sample of  $n = 64$  items. Here we retain  $H_0$  if

$$\begin{aligned}\bar{X} &\leq \mu_0 + \Phi^{-1}(.95) \sigma / \sqrt{n} \\ &= 10 + 1.645 \cdot 2/8 \\ &= 10.411\end{aligned}$$

i.e.,  $H_0$  is retained if  $\bar{X} \leq 10.411$ . Since the true  $\mu$  is unknown, the probability that  $H_0$  is retained is a function of this  $\mu$ :

$$\begin{aligned}P(\bar{X} \leq 10.411 | \mu) &= P\left[\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq \frac{10.411 - \mu}{(2/8)}\right] \\ &= P[Z \leq 4(10.411 - \mu)] \\ &= \Phi(41.644 - 4\mu)\end{aligned}$$

where  $Z$  is  $N(0, 1)$ . Then the O.C. curve for this problem is simply a plot of  $\Phi(41.644 - 4\mu)$  vs  $\mu$ . For any value of  $\mu$  when  $H_1$  is true, i.e.,  $\mu \geq 10$ ,  $\Phi(41.644 - 4\mu)$  is the probability of committing a Type II error.

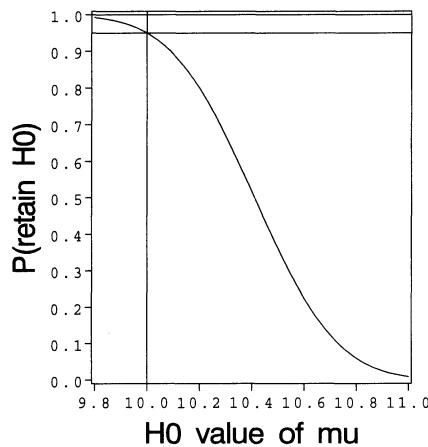
Note the close relationship between the O.C. curve and the power curve: The power curve is a plot of  $y = P(Z > 41.644 - 4\mu) = 1 - \Phi(41.644 - 4\mu)$  vs  $\mu$ .

To achieve a pretty graph, we do this plot for values of  $\mu$  from 9.8 to 11.0 in steps of 0.01. SAS and S-PLUS programs for generating plots of both the O.C. and power curves are in (`conc/code/conc.oc.sas`) and (`conc/code/conc.oc.s`), respectively. As usual, the interactive graphics device is used first to display the plots on a monitor. Then if we are satisfied we print it to paper, in this instance in Figure 3.9.

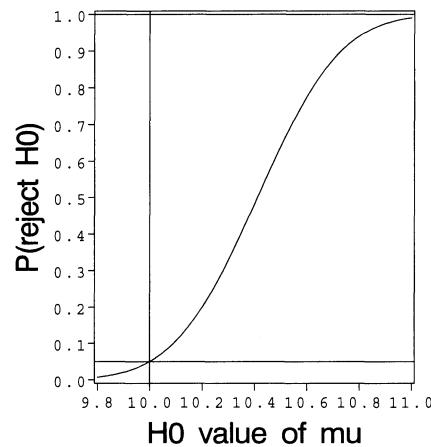
For most tests of hypotheses, calculation of Type II error probabilities and construction of O.C. and power curves involves the use of a *noncentral* probability distribution. This class of distributions is discussed in Appendix D. This is not an issue for tests using the normal distribution, which does not have a noncentral form.

FIGURE 3.9. O.C. and power curves drawn with SAS and S-PLUS for  
 $H_0: \mu \leq 10, \alpha = .05, n = 64, \mu_c = 10.411$   
 Operating Characteristic Curve:  $\beta(\mu) =$  Probability of retaining  $H_0$  for specified value of  $\mu$ .  
 Power Curve:  $1 - \beta(\mu) =$  Probability of rejecting  $H_0$  for specified value of  $\mu$ .  
 (`conc/code/conc.oc.sas`), (`conc/code/conc.oc.s`),  
 (`conc/figure/conc.f1.ps.gz`), (`conc/figure/conc.f2.ps.gz`),  
 (`conc/figure/conc.f3.ps.gz`), (`conc/figure/conc.f4.ps.gz`)

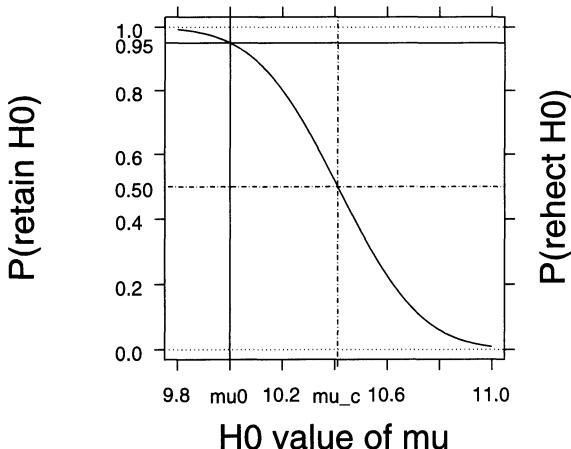
a. Operating Characteristic Curve—SAS.



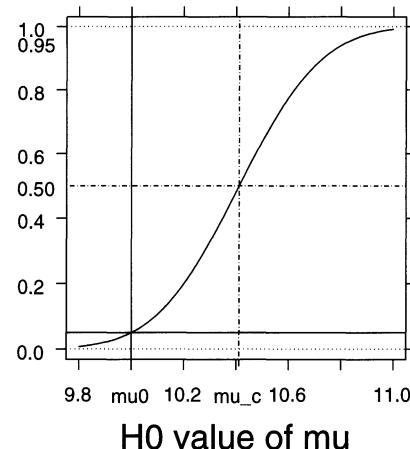
b. Power Curve—SAS.



c. Operating Characteristic Curve—S-PLUS.



d. Power Curve—S-PLUS.



## 3.10 Sampling

Whenever we wish to learn the characteristics of a large *population* or *universe* that is unwieldy or expensive to completely examine, we may instead select a *sample* from the population. If the sample has been selected by a *random* mechanism, it is usually possible to infer population characteristics from the analogous characteristics in the sample. Much of the remainder of this volume deals with methods for conducting such inferences. In this section we discuss methods for selecting random samples. Only rarely is it practical to sample the entire population; such a sample is called a *census* of the population.

Here are some examples of situations where we would learn about a population by choosing a random sample from it.

- A factory wishes to know if the proportion of today's output that is defective is sufficiently small that the output may be shipped for sale rather than scrapped. Examining the entire output stream is likely to be impractical and expensive, and clearly impossible if examining an item results in its destruction. Instead, a quality-control worker may suggest a random sample of the output, with a size of sample that is sufficient to accurately estimate the proportion of defectives without being excessively costly. [Formula (5.14) may be used for determining the sample size in this situation.]
- A candidate for statewide political office wants to assess whether more than half of the electorate will vote for her. An accurate estimate of the proportion favoring her would greatly influence her future campaign strategy. She obviously must contract for a sample because her campaign cannot afford to contact all potential voters. A complication in this situation is that the population of voters and their opinions are apt to be somewhat different on election day from what they are at the time the sample is selected.
- A timber company wishes to estimate the average height of the trees in a forest under its control. Such measurements are expensive to obtain because they involve sighting a tree's top at a fixed ground distance from the tree. Therefore, a census of the forest would be prohibitively expensive and some type of random sample of trees is preferred.

If an arbitrary sample is used, there is no guarantee that it will truly represent the population. To ensure that the sample adequately reflects the population, a randomization mechanism must be used. The techniques for inferring from sample to population discussed in the following chapters rest on the assumption that samples are randomly selected. If this assumption is unjustified, the probability-based statements that accompany the inferences will be incorrect.

For a given sample size  $n$  the analyst seeks to maximize the likely precision of the inference from sample to population while minimizing the cost of selecting and using the sample information. The most straightforward random sampling plan is termed *simple random sampling*. Sometimes, however, a different sampling plan can afford greater precision, or lower cost, or be easier to administer. We discuss simple random sampling and several commonly used alternatives.

### 3.10.1 Simple Random Sampling

A simple random sample of size  $n$  from a population of size  $N$  is one selected according to a mechanism guaranteeing that each of the  $\binom{N}{n}$  potential samples have the same probability,  $1/\binom{N}{n}$ , of being the sample actually selected.

If, as is usually the case, the population is already identified with a numbering from 1 to  $N$ , or if it is easy to set up such a numbering, then statistical software can be used to select  $n$  distinct integers in the range 1 to  $N$  so that all potential selections are equally likely to occur.

Such a sample is easily produced in S with the statement `sample(N,n)`. If the population is not numbered but exists as a character vector `x` [where  $n \leq \text{length}(x)$ ], then `sample(x,n)` produces the required sample from `x`.

In contrast, selecting a random sample using SAS requires at least 10 lines of code and so is not discussed here.

### 3.10.2 Stratified Random Sampling

Sometimes the population of interest is meaningfully partitioned into groups or *strata* and in addition to inferring about the entire population it is desired to learn about each *stratum* (the singular of *strata*). When this is the case, we may wish to select a random sample within each stratum. Then sample estimates are available for each stratum, and these can be combined into estimates for the entire population.

Suppose there are  $k$  strata and the number of population items in stratum  $i$  is  $N_i$ ,  $i = 1, \dots, k$ , where  $\sum_{i=1}^k N_i = N$ . The analyst then needs to decide how many of the  $n$  total sample items should be selected from stratum  $i$ . One popular possibility, called *proportional allocation*, stipulates sampling  $n_i = (\frac{N_i}{N}) n$  items from the  $i^{\text{th}}$  stratum. Since  $n_i$  need not be an integer, it is customary to round this calculation to the nearest integer. The mean estimated from the stratified random sample is  $\bar{x}_{\text{ST}} = \frac{1}{N} \sum_i N_i \bar{x}_i$ , i.e., a weighted average of the stratum sample means using the relative strata sizes as weights.

As an example, suppose it is desired to estimate the average annual malpractice premium paid by physicians licensed to practice in Pennsylvania. Since the risk of malpractice differs across medical specialties, it is likely also to be of interest to determine such estimates for each medical specialty. A physician considering relocation to Pennsylvania from elsewhere will be more interested in the estimated premium for her own medical specialty than the average premium of all Pennsylvania physicians. Accordingly, an investigator first decides the size  $n$  of a statewide sample she can afford. Then she obtains a directory of Pennsylvania physicians classified according to specialty and notes the number  $N_i$  of Pennsylvania physicians in each specialty  $i$ ,  $i = 1, \dots, k$ , where  $k$  is the number of distinct medical specialties. (Such a directory may be available for purchase from the American Medical Association.) Then a sample of approximately  $n_i = (\frac{N_i}{N}) n$  physicians is selected from among the Pennsylvania practitioners of specialty  $i$ .

Stratified sampling has the virtue of avoiding an undersampling of any stratum and so guarantees some minimum degree of precision for estimates from each stratum. When the population exhibits minimal variability within strata but considerable variability between units in different strata, estimates based on stratified random sampling are likely to be more precise than ones based on simple random samples of comparable total size. This fact will be demonstrated in Section 3.10.5.

### 3.10.3 Cluster Random Sampling

This technique is designed to control the cost of sampling in exchange for some decrease in precision of estimation. It is most frequently used when it is necessary to make personal contact with the *sampling units* (entity that is to be sampled), and the sampling units are physically dispersed to the extent that traveling from one unit to another is an appreciable cost.

As with stratified sampling, cluster sampling involves two stages. Assume that the population is partitioned into  $c$  clusters. A cluster is typically formed from geographically contiguous units so that sampling units within the same cluster are much closer to one another than two units in different clusters. In stage 1 the analyst selects  $c_0$  of these clusters, where  $c_0$  is considerably less than  $c$ . Then in stage 2 the analyst randomly samples  $n_i$  items from each selected cluster  $i$ , where  $\sum_{i=1}^{c_0} n_i = n$ . The samples within each cluster can be simple random samples, stratified random samples, etc. As in the case of stratified random sampling, we must decide on a rule for allocating the total sample size  $n$  to the clusters.

If  $T_i$  is the total for all observations in cluster  $i$ , then the mean estimated from the cluster random sample is  $\bar{y}_{\text{CRS}} = (\sum_i T_i)/(\sum_i N_i)$ , where both sums extend from 1 to  $c_0$ .

Cluster random sampling saves costs because it involves much less travel from one cluster to another than other sampling methods. But precision is sacrificed because this method prevents a large part of the population from appearing in the sample. In contrast to stratified sampling of strata, cluster sampling of clusters is most efficient when the variation within clusters is large compared to the variation between clusters.

When it is required to personally interview persons sampled from a city's population of eligible voters, a good strategy would be to identify voting districts as clusters and use cluster sampling. If, instead, we wanted to interview city residents as to their product preferences, an analyst might prefer to use zip codes as clusters because geography-based marketing strategies are more likely to be segmented by zip code than by voting district.

### 3.10.4 Systematic Random Sampling

This method may be considered when simplicity of the sampling design and administration is of prime importance.

Order the population from 1 to  $N$  and initially assume that  $N$  is an integral multiple of  $n$ , say  $N = mn$ . Then randomly select an integer  $i$ ,  $1 \leq i \leq m$ . Then sample population item  $i$  and every  $m^{\text{th}}$  item thereafter. For example, if  $N = 120$ ,  $n = 20$ ,  $m = 6$ , we might randomly sample items 4, 10, 16, ..., 118.

Suppose instead that  $N = mn + l$ ,  $1 \leq l < n$ . The analyst may then seek to move toward the  $N$  proportional to  $n$  situation. Suppose we modify the preceding illustration to  $N = 132$ . A possibility is to accept a larger  $n = 22$ . Another option that maintains  $n = 20$  is to randomly remove  $l = 12$  observations from sampling consideration and then proceed as before with the  $mn$  remaining observations.

This method should not be used if the population displays a periodic characteristic with the same period as  $m$ . For example, if we wish to randomly sample 20 houses in a subdivision consisting of 120 houses where each block has exactly 6 houses, then the preceding plan would either contain, or avoid, sampling houses on the end of blocks. Such houses tend to be on larger lots than ones in the middle of blocks and the plan would either include them exclusively or miss them entirely.

### 3.10.5 Standard Errors of Sample Means

In this section we provide standard errors for the means of random samples selected by various methods. Then according to the Central Limit Theorem, an approximate large-sample  $100(1 - \alpha)\%$  confidence interval for the population mean is of the form

$$\text{sample mean} \pm \text{standard error} \cdot z_{(1-\frac{\alpha}{2})}$$

For a simple random sample, the standard error is

$$s_{\text{SRS}} = \sqrt{\frac{s^2}{n} \left( \frac{N-n}{N-1} \right)}$$

For a stratified random sample with sample variance  $s_i^2$  from stratum  $i$ , the standard error is

$$s_{\text{ST}} = \frac{1}{N} \sqrt{\sum_i N_i^2 \left( \frac{N_i - n_i}{N_i - 1} \right) \frac{s_i^2}{n_i}}$$

If the  $\{s_i^2\}$  tend to be smaller than  $s$ , then  $s_{\text{ST}}$  will tend to be smaller than  $s_{\text{SRS}}$  with the conclusion that stratification was worthwhile.

To present the standard error for the mean of a cluster random sample, define  $\bar{N} = N/c$  to be the average cluster size. The standard error is

$$s_{\text{CRS}} = \sqrt{\left( \frac{c - c_0}{c_0 c \bar{N}^2} \right) \frac{\sum_i (T_i - \bar{y}_{\text{CRS}} N_i)^2}{c_0 - 1}}$$

The summation extends from 1 to  $c_0$ .

### 3.10.6 Sources of Bias in Samples

Sampling error is the discrepancy between the estimate and parameter being estimated. This error decreases as the sample size increases. Non-sampling errors are more serious than sampling errors because they can't be minimized by increasing the sample size. Continuing the example discussed in Section 3.10.2, we discuss two such sources of bias in the context of randomly sampling physicians who practice in Pennsylvania. *Selection bias* occurs when it is impossible to sample some members of the population. *Nonresponse bias* occurs if responses are not obtained from some members of the sample.

In order to randomly sample from the population consisting of all physicians licensed to practice medicine in Pennsylvania, we must obtain a list or computer file of such physicians. Even if we could obtain a list of physicians licensed to practice, there is no way to know which physicians on such a list

are in fact practicing medicine (as opposed to performing medical research or administrative tasks). Therefore, use of such a list would introduce selection bias. A better approach might be to obtain a list of the Pennsylvania membership of the American Medical Association (AMA). This list does indicate the nature of the physician's practice, if any, so nonpractitioners on the list can be ignored. However, not all physicians practicing in Pennsylvania are AMA members; such membership is not legally required in order to practice medicine. Thus some selection bias would still be present with this approach. Selection bias would be eliminated if the client can be persuaded to amend the target population to AMA members practicing in Pennsylvania.

Next suppose that this amendment is accepted and that a random sample of  $n$  practicing physicians is selected from the list. How should the physicians be contacted? Since physicians are busy individuals; visiting them in person or contacting them by telephone is unlikely to yield a response. Ignoring nonrespondents is likely to result in nonresponse bias because busier physicians are less likely to respond, and busyness may be associated with the survey questions.

Mail contact of the sampled physicians is preferred for several reasons. Since a written questionnaire can be answered at the physician's convenience, the physician is more likely to respond. Second, the questionnaire can be placed under a cover letter that encourages participation, written by a person respected by the respondents. Third, it is possible to keep track of who does not initially respond so that such individuals can be contacted again. This is accomplished by asking respondents to mail in a signed postcard indicating that they have participated, and to return the anonymous questionnaire in an envelope mailed separately.

Even this elaborate mail questionnaire approach does not eliminate the possibility of nonresponse bias. The extent of any remaining bias can be judged by comparing characteristics of the sampled physicians with those of the physician population reported in the AMA membership directory.

## 3.11 Exercises

**3.1.** Refer to the discrete bivariate distribution considered in Table 3.2.

- a. Let  $Z = X + 1$ . Find the distribution of  $Z$ .
- b. Find  $E(2X + 1)$  and  $2E(X) + 1$ . Then find  $E(X^2)$  and  $[E(X)]^2$ .
- c. Find  $P(X < Y)$ .

- d. Let  $X_1$  and  $X_2$  be independent and identically distributed as  $X$ . Make a table of the joint distribution of  $X_1$  and  $X_2$ , and use this to find  $P(X_1 < X_2 + 1)$ .
- 3.2.** How large a random sample is required for there to be a 92% probability of sampling at least one defective from a lot of 100,000 items which contains 100 defectives? (Hints: What is the random variable here? Consider the event that is the complement of “at least one defective”.)
- 3.3.** Suppose  $X$  is binomial(50, .10), and  $Y$  is binomial(20, .25). Draw the distribution functions of  $X$  and  $Y$ . Which one has a bigger mean? Which one has a bigger standard deviation?
- 3.4.** If  $X$ ,  $Y$  are each standard normal random variables, and they are independent of one another, what is the distribution of  $Z = 3X + 2Y$ ?
- 3.5.** Suppose that  $Y$  is a  $2 \times 1$  random vector such that
- $$W = \begin{pmatrix} 80 \\ 40 \end{pmatrix} + \begin{pmatrix} 10 & 7 \\ 7 & 5 \end{pmatrix} Y$$
- has a bivariate normal distribution with mean  $\begin{pmatrix} 60 \\ 70 \end{pmatrix}$  and covariance matrix  $\begin{pmatrix} 100 & 40 \\ 40 & 50 \end{pmatrix}$ . Find the probability distribution of  $Y$ , including its mean vector and covariance matrix.
- 3.6.** In class #1, 32 out of 40 students earned fewer than 70 points on the final exam. In class #2, 40 out of 50 students earned fewer than 75 points on the same exam. Restate the given class information in terms of percentiles. Is it possible to tell which class had a higher average score?
- 3.7.** Somebody tells you that a 95% confidence interval for the mean number of customers per day is (74.2, 78.5), and that this indicates that 95% is the probability that the mean is between 74.2 and 78.5. Criticize this statement and replace it with one correct sentence.
- 3.8.** Acme, Inc. thinks it has a new way of manufacturing a key product. It is trying to choose between  $A$  = “new way is better than old way” or  $B$  = “old way is better than new way”. Acme plans to reach its tentative conclusion by sampling some of the product produced the new way and conducting a statistical test. The new way is much more expensive than the old way. Which statement,  $A$  or  $B$ , should be the null hypothesis? Justify your answer.
- 3.9.** The probability that a project succeeds in New York is .4, the probability that it succeeds in Chicago is .5, and the probability that it succeeds in

at least one of these cities is .6. Find the probability that this project succeeds in Chicago given that it succeeds in New York.

- 3.10.** You are considering two projects, A and B. With A you estimate a payoff of \$60,000 with probability .6 and \$30,000 with probability .4. With B you estimate a payoff of \$80,000 with probability .5 or \$30,000 with probability .5. Answer the following questions after performing appropriate calculations.
- Which project is better in terms of expected payoff?
  - Which project is better in terms of variability of payoff?
- 3.11.** If  $X$  has a mean of 15 and a standard deviation of 4, and if  $Y = 5 - 3X$ , what are the mean and standard deviation of  $Y$ ?
- 3.12.** State the two ways in which a data analyst can modify a statistical test in order to decrease its Type II error probability.
- 3.13.** An analyst makes three independent inferences. For each of these inferences, the probability is .05 that it is *incorrect*. Find the probability that *all three* inferences are *correct*.
- 3.14.** Let  $A$  = “a McDonald’s franchise in Kansas is profitable” and let  $B$  = “the Philadelphia Eagles will have a winning season next year”. If  $P(A) = .8$  and  $P(B) = .6$ , find the probability that *either A or B* occurs.
- 3.15.** Use statistical software commands to do this problem. A new medicine has probability .70 of curing gout. If a random sample of 10 people with gout are to be given this medicine, what is the probability that among the 10 people in the sample, between 5 and 8 people will be cured?
- 3.16.** Use statistical software commands to do this problem. The daily output of a production line is normally distributed with a mean of 163 units and a standard deviation of 4 units.
- Find the probability that a particular day’s output will be 160 units or less.
  - The production manager wants to tell her supervisor, “80% of the time our production is at least  $x$  units”. What number should she use for  $x$ ?

- 3.17.** Find the expected value and standard deviation of a random variable  $U$  if its probability distribution is as follows:

| $u$ | $P(U = u)$ |
|-----|------------|
| 1   | .6         |
| 2   | .3         |
| 3   | .1         |

- 3.18.** A random variable  $W$  has probability density function  $f(w) = 2 - 2w$ ,  $0 < w < 1$ , and  $f(w) = 0$  for all other values of  $w$ .
- Verify that  $f(w)$  is indeed a probability density function.
  - Find the corresponding cumulative distribution function,  $\mathcal{F}(w)$ .
  - Find the expectation of  $W$ .
  - Find the standard deviation of  $W$ .
  - Find the median of this distribution, i.e., the number  $w_m$  such that  $P(W < w_m) = .5$ .
- 3.19.** Use a statistical software command to approximate the value of  $z_{.08}$ .
- 3.20.** State the two things that a data analyst can do in order to make a confidence interval *narrower*.
- 3.21.** A data analyst tentatively decides on values for  $\alpha$  and  $n$  for a statistical test. Before performing the test she investigates its Type II error control and finds this to be unsatisfactory. What two options does she have to improve Type II error control?
- 3.22.** In the discussion of *sufficiency* of a point estimator in Section 3.6.3, we indicated that  $2\bar{x}$  is not a good estimator of  $a$  from a sample of  $n$  items from a continuous uniform distribution on  $[0, a]$ . Can you suggest a better estimator of  $a$  and explain why it is better than  $2\bar{x}$ ?
- 3.23.** The dataset ([datasets/salary.dat](#)), from (Forbes Magazine, 1993), contains the ages and salaries of the chief executives of the 60 most highly ranked firms among *Forbes Magazine's* “Best small firms in 1993.” Consider the variable `age`.
- Produce a boxplot and a stem-and-leaf plot for `age`.
  - Construct a 95% confidence interval for the mean age. What assumptions were made in your construction?
  - Test  $H_0: \mu \leq 50$  against  $H_1: \mu > 50$ , reporting and interpreting the *p*-value for this test.

- d. Find the power of this test for the alternative  $\mu_1 = 53$ .
- 3.24.** The dataset (`datasets/cereals.dat`) contains various nutritional measurements for 77 breakfast cereals. We are concerned here with the variable `carbo` (carbohydrates) measured in grams per serving. Begin by eliminating the cereal `Quaker Oatmeal`, which was erroneously reported as having a negative value for carbohydrates.
- Produce boxplots and stem-and-leaf plot for `carbo`. Do these plots suggest that this variable comes from a normal population?
  - Construct at 99% confidence interval for the mean carbohydrate content.
  - Test  $H_0: \mu \geq 16$  against  $H_1: \mu < 16$ , reporting and interpreting the *p*-value for this test.
  - Find the probability of committing a Type II error for the alternative  $\mu_1 = 15$ .
- 3.25.** The sampling bias in the December 1969 U.S. Draft Lottery, with data in file (`datasets/draft70mn.dat`), is described in Exercise 4.1. Suppose you had been the administrator of that lottery. Explain how you would have performed the sampling without incurring such bias.
- 3.26.** Royalties paid to authors of novels have sometimes been based on the number of words contained in the novel. Recommend to an old-fashioned author how to estimate the number of words in a handwritten manuscript she is planning to give to her publisher.
- 3.27.** Samples are taken from two strata. Suppose the variance of the two samples combined is  $s^2 = 7.6$  and the following within-stratum information is known:

| Stratum | $N_i$ | $n_i$ | $s_i^2$ |
|---------|-------|-------|---------|
| 1       | 100   | 30    | 1.2     |
| 2       | 120   | 40    | 1.4     |

Observe that there is far less variability within the two strata than between the two strata. Calculate  $s_{SRS}$  and  $s_{ST}$  to verify that for estimating the common population mean in this situation,  $\bar{x}_{SRS}$  is much preferred to  $\bar{x}_{ST}$ .

- 3.28.** The organization of a candidate for a city political office wishes to poll the electorate. For this purpose, discuss the relative advantages and disadvantages of personal interview polling vs telephone polling.

- 3.29.** Explain how it is possible for a census to yield less accurate results than a random sample from the same population.
- 3.30.** A student claims that a random sample of  $n$  items from a population of  $N$  items is one selected so that each item in the population has the same probability  $\frac{n}{N}$  of appearing in the sample. Demonstrate that this definition is inadequate.
- 3.31.** A four-drawer file cabinet contains several thousand sheets of paper, each containing a statement of the dollar amount due to be paid to your company. The sheets are arranged in the order that the debt was incurred. You are asked to spend not more than one hour to estimate the average dollar amount on all sheets in the file cabinet. Propose a plan for accomplishing this.

# Graphs

Graphs are used to inspect and display patterns in data. Appropriately drawn graphs are, in our opinion, the best way to gain an understanding of what data have to say. In this chapter we present several of the types of graphs and plots we will be using throughout. We discuss the visual impact of the graphs and relate them to the tabular presentation of the same material.

Statistical techniques have underlying assumptions. An important use of graphs is to aid in the checking of a list of assumptions a technique requires in order for an analysis using the technique to be correct. For example, regression analysis, discussed in Chapters 8 to 11, requires that model residuals are randomly distributed. Residual plots, discussed in these chapters, must show random scatter rather than a systematic pattern.

We discuss the construction of graphs and pay attention to each of the components of graphs that can aid (or hinder) the interpretation of the data. We show good (and bad) examples of plots and discuss why we make those value judgments. Appendix G summarizes many graphs new to this book that are based on Cartesian products.

We see graphs as the heart of most statistical analyses; the corresponding tabular results are formal confirmations of our visual impressions. The graphs are not automatically produced by most software; instead it is up to the analyst to request them.

## 4.1 Definition

A graph is a geometrical representation of the information in a table of numbers. Our prototype graph is the traditional scatterplot—a two-dimensional plot of two variables ( $x$  on the abscissa or horizontal axis and  $y$  on the ordinate or vertical axis). For each observation  $(x, y)$  in the data we locate a point on the graphing surface with coordinates  $(x, y)$ . For a dataset with  $n$  observations we mark  $n$  points on the graphing surface. Figure 4.2 is an example of such a plot.

The goal of statistical graphics is to make evident the characteristics of the data such as location, variability, scale, shape, correlation, interaction, and clustering. Once we have a visual understanding of the data, we usually attempt to model it with formal algebraic procedures. We will normally translate our algebraic understanding back to a graphical presentation and to a verbal discussion of our findings.

## 4.2 Example—Ecological Correlation

Examination of plots of the data at early stages of the analysis, before requesting and examining tabular output, is an essential part of data analysis. This point is demonstrated in Figure 4.1, which illustrates what is known as the Ecological Fallacy. If without examining a plot of these (simulated) data we perform a simple regression of  $y$  on  $x$ , we find that  $y$  and  $x$  are directly related. The plot strongly suggests that what we have is the amalgamation of three disparate groups. Within each of the groups it is clear that  $y$  and  $x$  are inversely related, the opposite conclusion from the amalgamated result. In practice it is likely that the existence of the groups is meaningful information that must be accounted for by the analyst. In this case the individual within-group results are what should be reported.

(Robinson, 1950) apparently coined the terminologies Ecological Fallacy and Ecological Correlation. He noted that the correlation between percentage illiterate and percentage black racial group for the United States as a whole, based on the 1930 U.S. Census, is different from this correlation within various subgroups of the U.S. population. Human ecology is a branch of sociology dealing with the relationship between human groups and their environments. The fallacy is that we cannot necessarily use a finding from an entire population to reach conclusions from various subsets of the population.

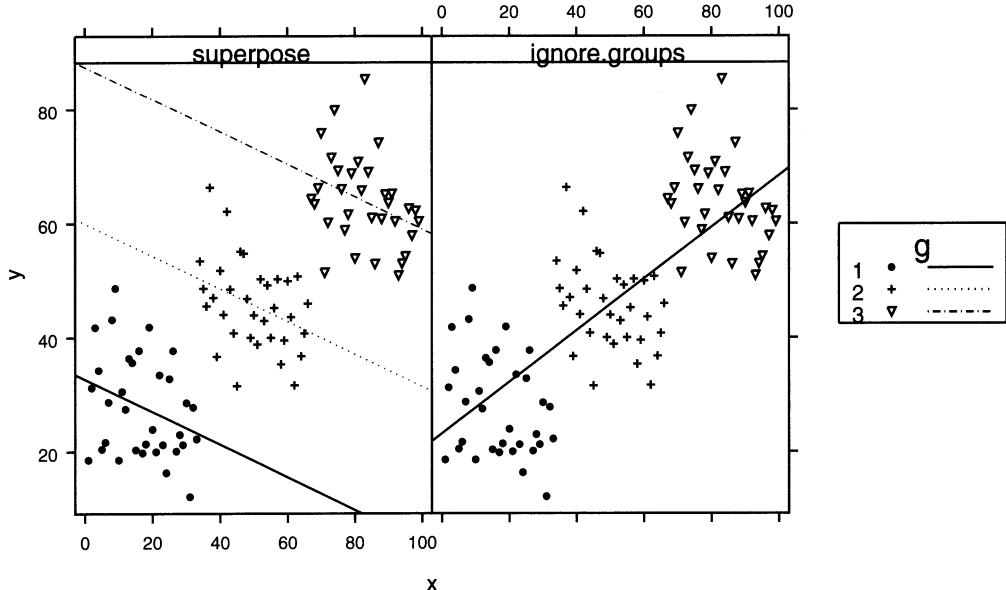


FIGURE 4.1. Ecological Correlation. The overall slope ignoring groups is strongly positive. The slope within each group is strongly negative. These are simulated data.  
 (grap/code/ecolo.s), (grap/figure/grap.ecop.eps.gz)

### 4.3 Scatterplots

Figure 4.2 shows the selling price by lot size for 105 single-family homes in Mount Laurel, New Jersey, from March 1992 through September 1994. The data in file (`datasets/njgolf.dat`), from (Asabere and Huffman, 1996), are read into S-PLUS with file (`grap/code/njgolf-read.s`).

There is much information in Figure 4.2. We start by listing the most obvious items, and then we will look at the less obvious and more puzzling items. The range of lot sizes is 0–30,000 square feet, with most of the lots in the 8,000–15,000-square-foot range. But what is that large cluster of lot sizes at 0 square feet? The range of sale prices is \$50,000–\$250,000, with most of the 0-size lots selling for under \$130,000 and most of the nonzero lots selling above \$130,000. Within the 5,000–25,000-square-foot range price seems independent of size of lot, that is, for any lot size in that range the best estimate of sales price is the same, about \$165,000.

The scatterplot is an ordinary 2-dimensional plot with one variable `lotsize` on the  $x$ -axis (horizontal axis or abscissa) and the other variable `sprice` on the  $y$ -axis (vertical axis or ordinate). The plotting routine automatically

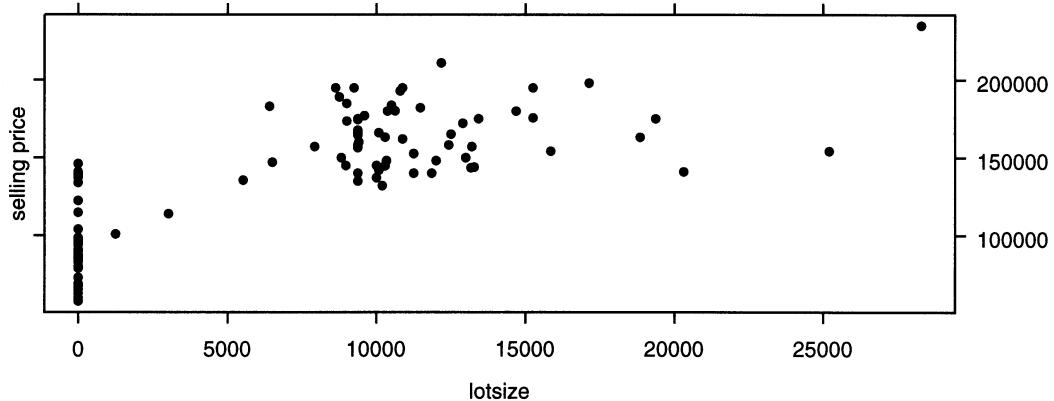


FIGURE 4.2. Selling price by lot size for 105 single-family homes in Mount Laurel, New Jersey, from March 1992 through September 1994.

([grap/code/njgolf-read.s](#)), ([grap/code/njgolf.graphs.s](#)),  
 ([grap/figure/grap.pric.lot.eps.gz](#)), ([grap/figure/grap.pric.lot.color.eps.gz](#))

determines the appropriate scale and tick locations for both axes and prints the variable names as the default labels for the axes.

We raised many questions in our perusal of Figure 4.2. Answering them requires us to look carefully at the definitions of the variables we displayed in the figure. We find that the variable labeled `lotsize` is actually a conflation of two distinct concepts. If the property is a condominium (a form of ownership of an apartment that combines single ownership of the residence unit with joint ownership of the building and associated grounds), the variable `lotsize` was arbitrarily coded to 0. If the property is a single-family house, then the variable `lotsize` contains the actual lot size in square feet. This explains the numerous observations having `lotsize=0` in Figure 4.2.

We must also look at additional variables. We will start with three measures of the size of the dwelling unit, rather than of the lot on which it is built. In Figure 4.3 we look at selling price against the number of bedrooms, the dining room area, and the kitchen area. All three plots show a rise in selling price as the  $x$ -variable increases. We can also see a hint in Figure 4.3 that selling price increases with  $x$  for both the lower-priced properties (the condominiums) and the higher-priced ones (the single-family houses).

We investigate that possibility in Figure 4.4 where we show all three plots conditioned on whether the property is a condominium or house. Now we

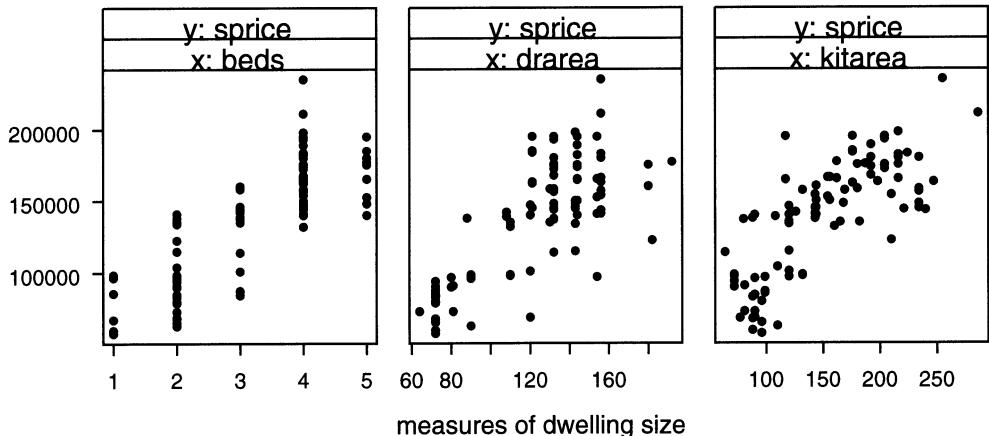


FIGURE 4.3. Selling price by number of bedrooms, by dining room area, and by kitchen area for 105 single-family homes in Mount Laurel, New Jersey, from March 1992 through September 1994.  
`(grap/code/njgolf-read.s)`, `(grap/code/njgolf.graphs.s)`,  
`(grap/figure/grap.pric.bdk.eps.gz)`

see very clear uphill trends of price on the measures of size within each of the panels of the figure.

## 4.4 Scatterplot Matrix

We looked at five variables in Figure 4.4 and nominally two, but actually three, variables in Figure 4.2. In both figures we used selling price as the  $y$ -variable and the others as either  $x$ -variables or as conditioning variables. In Figure 4.5 we look at all six variables together. This display shows all the individual panels that we looked at in the previous graphs in the `sprice` row and also shows the relationships among the other variables.

The display type is a *scatterplot matrix* or *splom* (Scatter PLOT Matrix), a matrix of scatterplots with each of the six variables taking the role of  $x$ -variable and  $y$ -variable against all the others. Thus there are  ${}_6P_5 = 30$  distinct plots in Figure 4.5. Each of these 30 plots is a plot of a pair of variables comparable to Figure 4.2. Since each of the six variables appears in both the  $x$ - and  $y$ -position, there are are only  ${}_6P_5/2 = 30/2 = \binom{6}{2} = 15$  distinct pairs of variables in the plots. We see that the  $(i, j)$  panel of the splom (counting from the lower-left corner) is the reflection of the  $(j, i)$  panel.

A defining property of the scatterplot matrix is that all panels in the same row have identical  $y$ -scaling and all panels in the same column have identical  $x$ -scaling. It is therefore easy to trace an interesting point in one panel

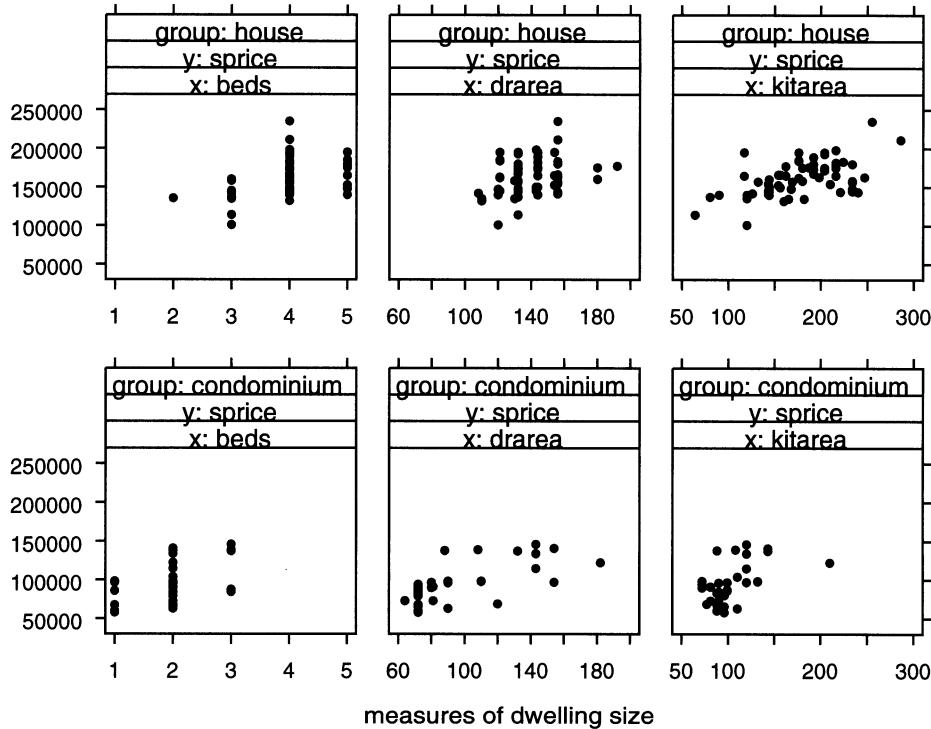


FIGURE 4.4. Selling price by number of bedrooms, by dining room area, and by kitchen area for 105 single-family homes, conditioned on whether the property is a condominium or house, in Mount Laurel, New Jersey, from March 1992 through September 1994.

(`grap/code/njgolf-read.s`), (`grap/code/njgolf.graphs.s`),  
`(grap/figure/grap.pric.bdkc.eps.gz)`

across to the other panels. For example, the single point visible in the condominium position of the `lotsize ~ cond.house` panel is recognized as an overplotting of many condominium points when we trace it in the other panels to the left and see that the dining area of condominiums runs the full range of dining areas for the entire dataset.

Unfortunately, Figure 4.5 has also lost the distinction between the condominiums and houses that we worked so hard to find. We recover that distinction in Figure 4.6 where we now show the five numeric variables separately for condominiums and houses. We can look across the subpanels in each main panel of Figure 4.6 and see relationships among multiple

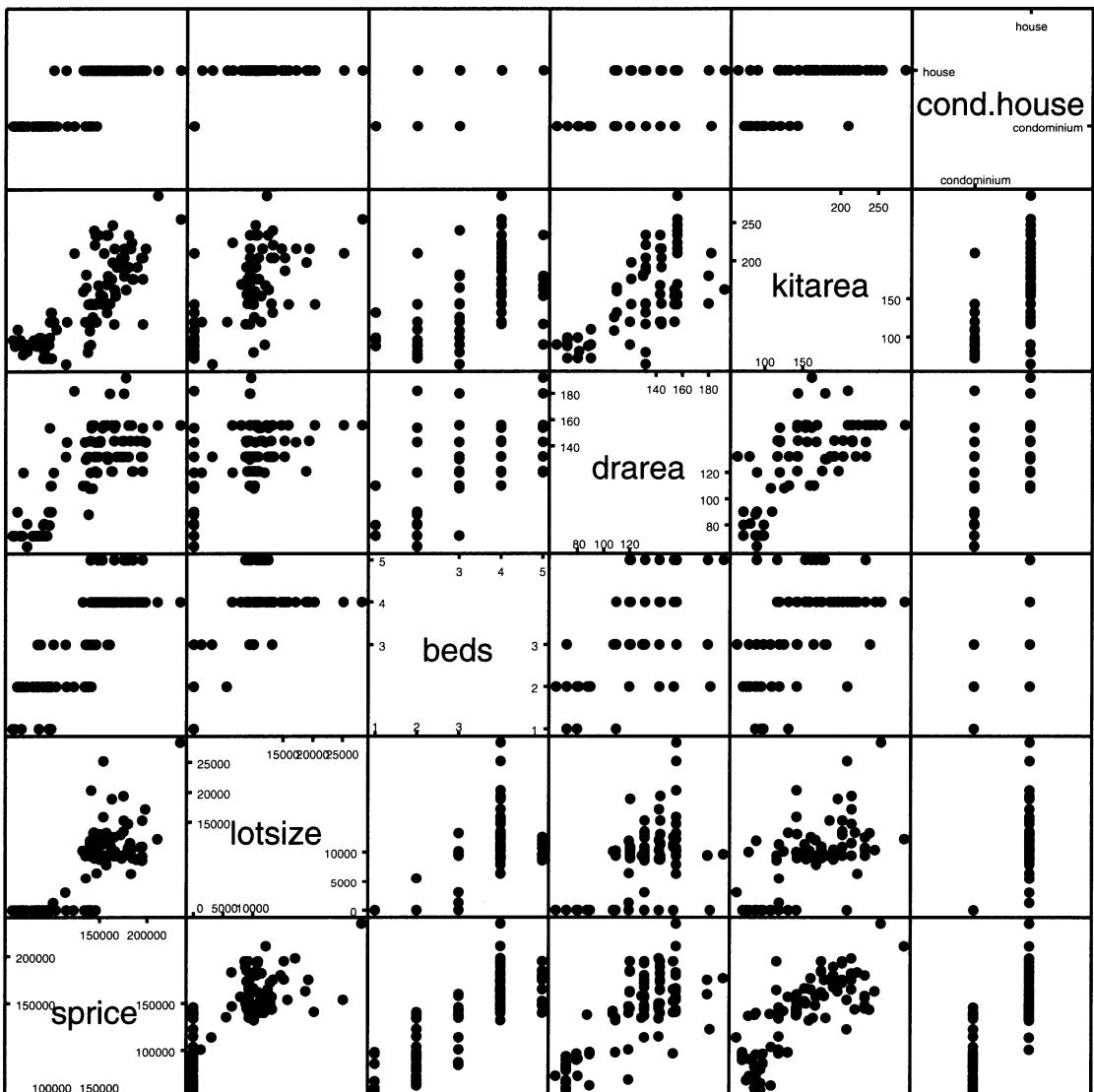


FIGURE 4.5. Scatterplot matrix of six variables for the 105 single-family homes in Mount Laurel, New Jersey, from March 1992 through September 1994.  
 (grap/code/njgolf-read.s), (grap/code/njgolf.graphs.s),  
 (grap/figure/grap.pbdkcl.eps.gz), (grap/figure/grap.pbdkcl.color.eps.gz)

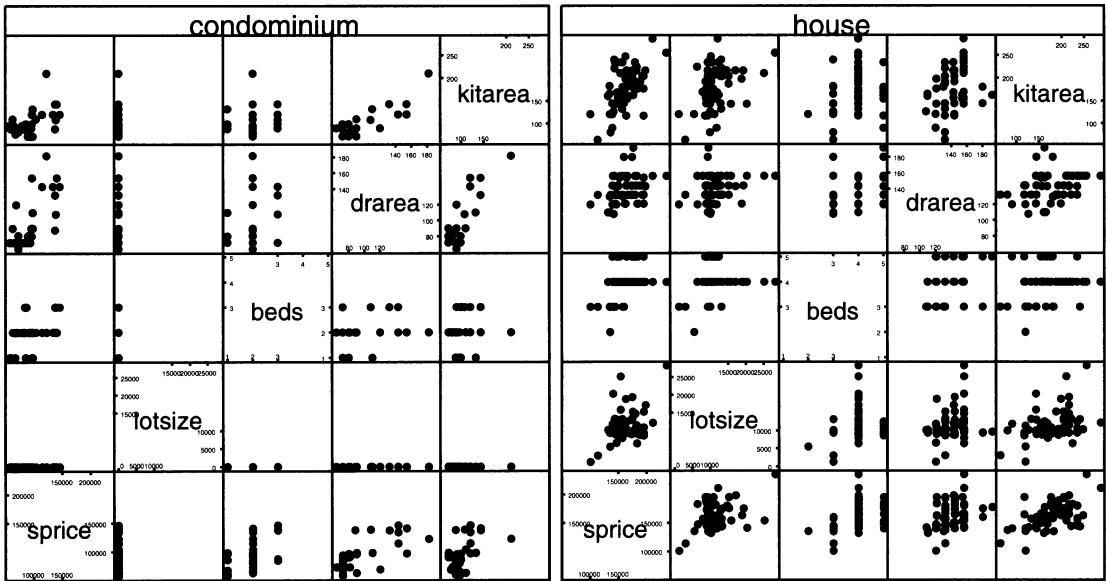


FIGURE 4.6. Scatterplot matrix of five variables for the 105 single-family homes, conditioned on whether the property is a condominium or house, in Mount Laurel, New Jersey, from March 1992 through September 1994.

(`grap/code/njgolf-read.s`), (`grap/code/njgolf.graphs.s`),  
 (`grap/figure/grap.pbdkc-1.eps.gz`)

variables. On the condominium panel of Figure 4.6 we see that the condominium with largest kitchen and dining room is one of the higher-priced properties (but not the highest) and it has only two bedrooms. On the house panel of Figure 4.6 we see that the highest-priced house has the largest lot size, but not the largest dining area and only four bedrooms.

Additional discussion of scatterplot matrices appears in Section 4.6.

## 4.5 Example—Life Expectancy

### Study Objectives

For each of the 40 largest countries in the world (according to 1990 population figures), data file (`datasets/tv.dat`) gives the country's life expectancy at birth partitioned by gender, number of people per television set, and number of people per physician (Rossman, 1994).

### Data Description

`life.exp`: Life expectancy at birth

`ppl.per.tv`: Number of people per television set

`ppl.per.phys`: Number of people per physician

`fem.life.exp`: Female life expectancy at birth

`male.life.exp`: Male life expectancy at birth

We read the data file with either (`grap/code/grap.read.le.s`) or (`grap/code/grap.read.le.sas`). We initially focus on the male and female life expectancies in Table 4.1 and Figure 4.7.

TABLE 4.1. Life expectancy.

| Abbrev | Country      | Female | Male | Abbrev | Country         | Female | Male |
|--------|--------------|--------|------|--------|-----------------|--------|------|
| Argn   | Argentina    | 74     | 67   | M(B)   | Myanmar (Burma) | 56     | 53   |
| Bngl   | Bangladesh   | 53     | 54   | Pkst   | Pakistan        | 57     | 56   |
| Brzl   | Brazil       | 68     | 62   | Peru   | Peru            | 67     | 62   |
| Cand   | Canada       | 80     | 73   | Phlp   | Philippines     | 67     | 62   |
| Chin   | China        | 72     | 68   | Plnd   | Poland          | 77     | 69   |
| Clmb   | Colombia     | 74     | 68   | Romn   | Romania         | 75     | 69   |
| Egyp   | Egypt        | 61     | 60   | Russ   | Russia          | 74     | 64   |
| Ethp   | Ethiopia     | 53     | 50   | StAf   | South Africa    | 67     | 61   |
| Frnc   | France       | 82     | 74   | Span   | Spain           | 82     | 75   |
| Grmn   | Germany      | 79     | 73   | Sudn   | Sudan           | 54     | 52   |
| Indi   | India        | 58     | 57   | Tawn   | Taiwan          | 78     | 72   |
| Indn   | Indonesia    | 63     | 59   | Tnzn   | Tanzania        | 55     | 50   |
| Iran   | Iran         | 65     | 64   | Thln   | Thailand        | 71     | 66   |
| Itly   | Italy        | 82     | 75   | Trky   | Turkey          | 72     | 68   |
| Japn   | Japan        | 82     | 76   | Ukrn   | Ukraine         | 75     | 66   |
| Keny   | Kenya        | 63     | 59   | UnKn   | United Kingdom  | 79     | 73   |
| K,Nr   | Korea, North | 73     | 67   | UnSt   | United States   | 79     | 72   |
| K,St   | Korea, South | 73     | 67   | Vnzl   | Venezuela       | 78     | 71   |
| Mexc   | Mexico       | 76     | 68   | Vtnm   | Vietnam         | 67     | 63   |
| Mrc    | Morocco      | 66     | 63   | Zair   | Zaire           | 56     | 52   |

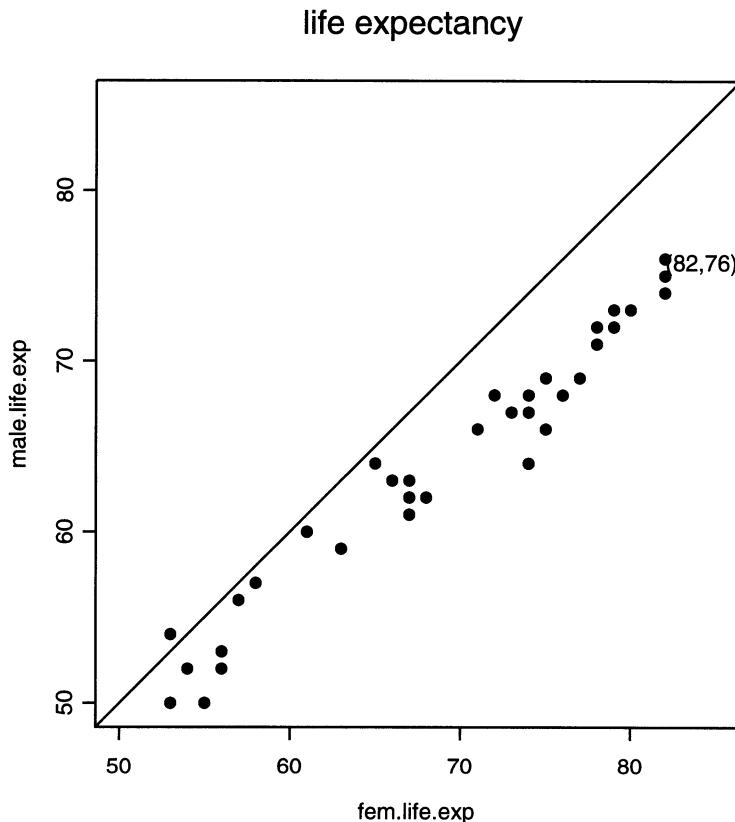


FIGURE 4.7. Life Expectancy  
 (grap/code/grap.f1.s), (grap/code/grap.f1.sas), (grap/figure/grap.f1.eps.gz)

Figure 4.7 shows each data row of Table 4.1 as a distinct point. For example, the point for Japan is located at  $(x, y) = (82, 76)$ . In a good graphical system we have control of the plotting symbols. We plotted the points with a solid dot  $\bullet$  and labeled one point with text to show its coordinates.

The first impression we get from reading Figure 4.7 is that most of the points are below the  $45^\circ$  line. This is such an important part of the interpretation of this graph that we drew the  $45^\circ$  line. Once the line is there for reference we immediately note that one country's point is above the  $45^\circ$  line. Which one? The easiest way to find out is to plot the abbreviated country names instead of dots (Figure 4.8a). Bangladesh is the country that has a longer life expectancy for males than females. On an interactive graphics system we merely click on the point and the system will label it

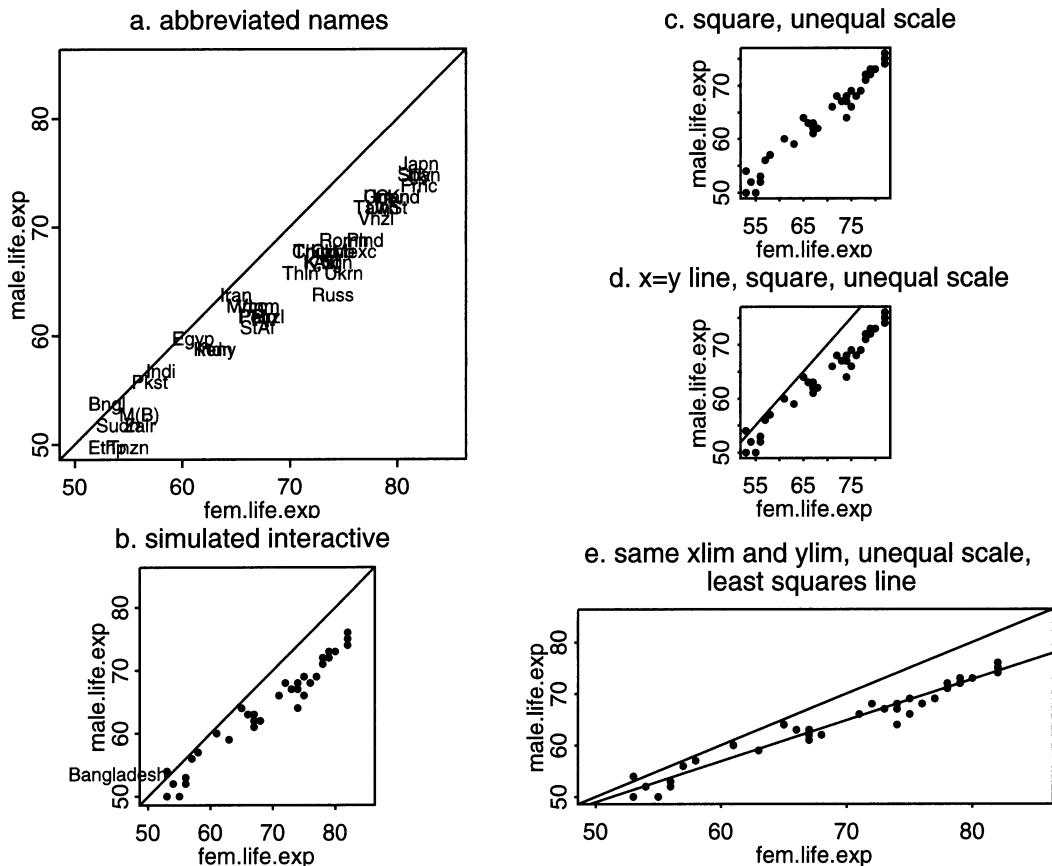


FIGURE 4.8. Life expectancy—variations on the plot.  
 (graph/code/graph.f2.s), (graph/figure/graph.f2.eps.gz)

[see file (graph/code/graph.identify.s)]. We have simulated the interactive appearance in Figure 4.8b.

We see from the figures that life expectancy for males and females is related; as one goes up the other tends to go up as well. We have done several other fine tunings on Figure 4.7. Life expectancy is measured on the same numerical scale for both male and female; therefore, we forced both scales to have the same range and we forced the graph to be square. By default, most plotting systems independently determine the  $x$ - and  $y$ -scales and use the maximum available area for the graph. Figure 4.8c releases the constraint on the ranges and we see that the male and female ranges are different (the female range is offset from the male range by 5 years). Since the graph goes from the lower-left corner to the upper-right corner, it falsely gives the

visual impression that the two ranges are the same. When we plot the  $45^\circ$  line in Figure 4.8d we get much of the correct impression back. In Figure 4.8e, where we no longer constrain the graph to be square, we lose the visual effect of forcing the ranges to be the same on both axes. In Figure 4.8e we have plotted the least-squares line through the points in addition to the  $45^\circ$  line. Least squares will be discussed in detail in Chapter 8. For now we note that this line attempts to get close to most of the points. It is used as an indicator of the linear relationship between the two variables male and female life expectancy.

## 4.6 Scatterplot Matrices—Continued

There are five variables in the `tv` dataset. Figure 4.9 plots them all in a scatterplot matrix.

Continuing with the discussion begun in Section 4.4, the scatterplot matrix is a coordinated set of scatterplots, one for each pair of variables in the dataset. We refer to the individual scatterplots comprising the matrix as *panels*. The panels are labeled by their  $Y \sim X$ , that is Row  $\sim$  Column, variable names. Thus, in Figure 4.9, the panel in the upper-left-hand corner (also called the NW or Northwest corner) is called the `ppl.per.phys ~ fem.life.exp` panel. Variable names are unambiguous and are constant across multiple views of the data: The `male.life.exp ~ fem.life.exp` panel refers to the same data values all of Figures 4.7, 4.8, and 4.9. We would NOT say “row 1 by column 4” because the sequencing of variables and the direction of ordering the rows and columns (is row 1 at the top or bottom?) are unclear.

There are several possible orientations of the panels; we display the best in Figure 4.9 and will discuss other orientations in Figures 4.10 and 4.11. There are five variables; hence the matrix consists of a  $5 \times 5$  array of scatterplots. Indexing for the set of plots is sorted in the same way as the axes in each individual panel. Indexing begins at the lower left and proceeds from left to right and from bottom to top. The main diagonal runs from southwest to northeast. Each panel containing one scatterplot is square. Each pair of variables appears twice, once below the main diagonal and again as a mirror image above the main diagonal. There is a single axis of symmetry for the entire *splom*.

The variables in Figure 4.9 are all continuous measurements. When using a *splom* to display data with categorical variables, we recommend avoiding inclusion of categorical variables among the variables comprising the *splom* itself, particularly for categorical variables having few categories, as they will usually appear as a noninformative regular lattice (see, for example, the

## Televisions, Physicians, and Life Expectancy

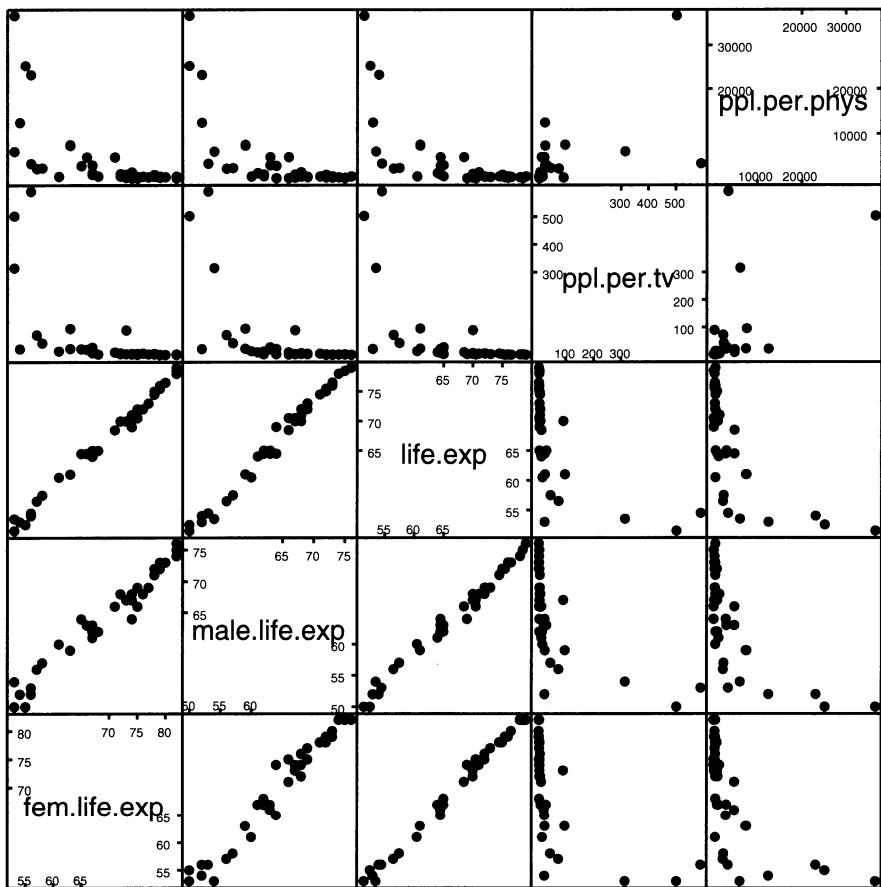


FIGURE 4.9. Televisions, physicians, and life expectancy.  
 (grap/code/grap.f3.s), (grap/code/grap.f3.sas), (grap/figure/grap.f3.eps.gz)

`customf` × `cornerf` panel of Figure 9.3). It is usually more informative to produce two or more adjacent *sploms*, by conditioning on the categorical variables, or to use different plotting symbols for the different levels of one of the factors. We use both strategies in Figure 9.4, conditioning on the levels of `corner` and using different plotting symbols for the levels

of `custom`. Another example is Figure 11.1, which contains two adjacent *sploms* conditioned on the two levels of the categorical variable `lime`.

We have presented what we consider to be the best orientation of the splom in Figure 4.9. Two other orientations are commonly used. When scatterplot matrices were first invented, the importance of a single axis of symmetry was not yet realized. Older scatterplot matrix programs (and some current ones) default, or are limited, to the more difficult main diagonal from northwest to southeast. The S-PLUS function `splom` can't use the alternate diagonal. The older S-PLUS function `pairs` defaults to the NW-SE diagonal but provides the option to change it with the `invert=F` argument. `pairs` also defaults to rectangular panels (the goal is maximal use of the plotting surface) but fortunately provides an option to force square panels. We show the `pairs` plot with both suboptimal choices in Figure 4.10. The SAS PROC INSIGHT `scatter` statement also uses the NW-SE diagonal and rectangular panels.

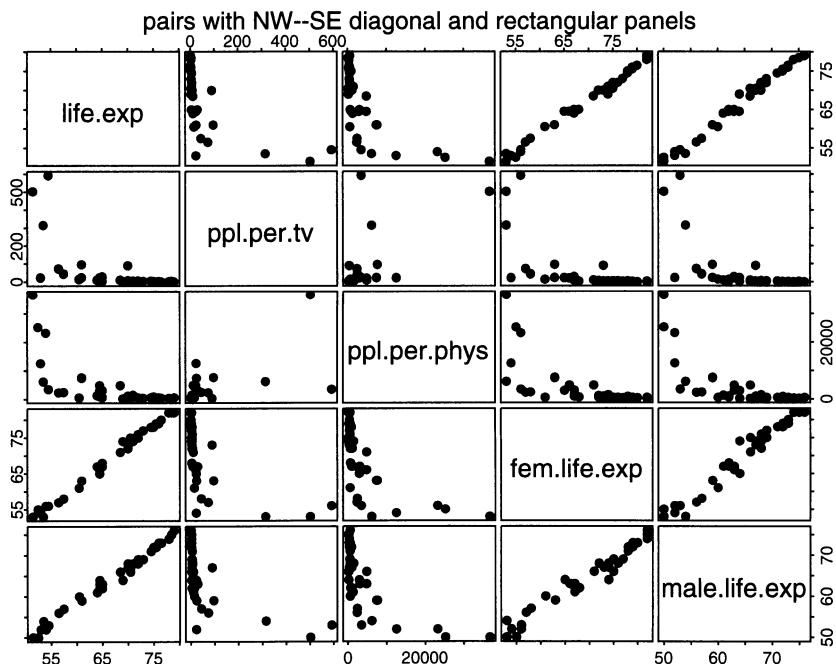


FIGURE 4.10. Alternate orientation with rectangular panels for `splom`. The downhill diagonal is harder to read (see Figure 4.11). The rectangular panels make it hard to compare each panel with its transpose.

(`grap/code/grap.f11.s`), (`grap/figure/grap.f11a.eps.gz`)

a. multiple axes of symmetry. b. single axis of symmetry.

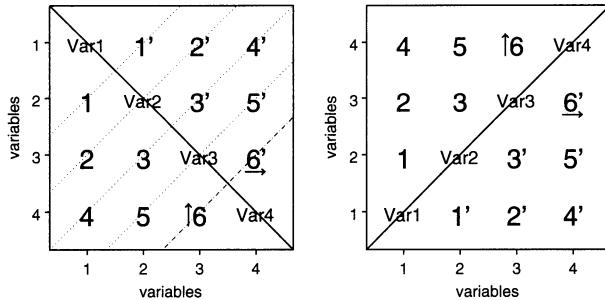


FIGURE 4.11. Axes of symmetry for splom. Figure 4.11a has six axes of symmetry. We focus on panel 6, which appears in positions reflected about the main NW–SE axis. The individual points within panels 6 and 6' are reflected about the dashed SW–NE line, as indicated by the position of the arrow. The other four axes, which reflect respectively panel 5, panels 3 and 4, panel 2, and panel 1 are indicated with dotted lines. Figure 4.11b has only one axis of symmetry. The arrow for panel 6 is reflected by the same SW–NE axis that reflects panels 6 and 6'.

([grap/code/grap.f12.s](#)), ([grap/figure/grap.f12.eps.gz](#))

The major difficulty with Figure 4.10 is that the multiple axes of symmetry are hard to find. The axes of symmetry are illustrated in Figure 4.11. The confusion in Figure 4.11a occurs because pairs of plots with the same variable names appear to the lower left and upper right of the NW–SE main axis of the matrix of plots. Within each pair, the upper plot needs to be reflected about its own SW–NE axis to match the lower plot. By comparison, Figure 4.11b has a single axis of symmetry for the entire plot. All pairs of plots and reflections within each pair occur around a single SW–NE axis of symmetry. Note also that the individual panels of the display are square to help ease the eye's task of seeing the symmetry.

Earlier versions of S (Becker et al., 1988) defaulted to printing just one triangle of the two mirror image triangles in `pairs` and had an option `full=T` to print the full matrix. From the manual, “By default, only the lower triangle is produced, saving space and plot time, but making interpretation harder.” That option made sense with typewriter terminals at 10 characters per second. It no longer makes sense with desktop workstations, windowing terminals, and laser printers. The single triangle of a scatterplot matrix is no longer available in S-PLUS. It is still occasionally seen in other programs and in texts.

Older programs sometimes display a very confusing subset of the lower triangle in which the rows and columns of the display show different sets of variables. The intent is to save space by suppressing a presumably non-informative main diagonal. The effect on the reader is to add confusion by

|           |   |           |           |           |
|-----------|---|-----------|-----------|-----------|
|           | 2 | <b>21</b> |           |           |
| variables | 3 | <b>31</b> | <b>32</b> |           |
|           | 4 | <b>41</b> | <b>42</b> | <b>43</b> |
|           |   | 1         | 2         | 3         |
|           |   | variables |           |           |

FIGURE 4.12. Symbolic form of very confusing subset of panels for the scatterplot matrix. This form has different variables along the rows and columns and has very little symmetry that might aid the reader. Note, for example, that panels 31 and 42 are positioned such that the eye wants to treat them as symmetric. This form is mostly obsolete and is strongly not recommended.

breaking symmetry. A symbolic version of this form of the plot is in Figure 4.12.

## 4.7 Data Transformations

Since the three life expectancy variables are summarized by just one, let us look at the simplified splom in Figure 4.13. The bottom row of the splom, with `life.exp` as the  $y$ -coordinate, shows an L-shaped pattern against both `ppl.per.tv` and `ppl.per.phys` as the  $x$ -variables. We have learned (or will learn in this chapter and again in Chapter 8) that straight lines are often helpful in understanding a plot. There is no sensible way to draw a straight line here. The plot of the two potential  $x$ -variables against each other is bunched up in the lower-left corner. The bunching suggests that a log transformation of the `ppl.*` variables will straighten out the plot. We see in Figure 4.14 that it has done so.

We also see that the log transformation has *stabilized the variance*. By this we mean that the `ppl.per.phys ~ life.exp` panel of Figure 4.13 has a range that fills the vertical dimension of the panel for values of `life.exp` near 50 and that is almost constant for values of `life.exp` larger than 65. After the log transformation of `ppl.per.phys` shown in Figure 4.14, for any given value of `life.exp` we observe that the vertical range of the response is about  $\frac{1}{3}$  of the vertical dimension of the panel.

There are several issues associated with data transformations. In the life expectancy example the natural logarithm  $\ln$  was helpful in straightening out the plots. In other examples other transformations may be helpful. We will take a first look at a family of power transformations. We recommend (Emerson and Stoto, 1983) for a more complete discussion. We identify some of the issues here and then focus on the use of graphics to help determine which transformation in the family of power transformation would be most helpful in any given situation.

- Stabilize variance. This chapter and also Chapters 6 and 14.

### Televisions, Physicians, and Life Expectancy

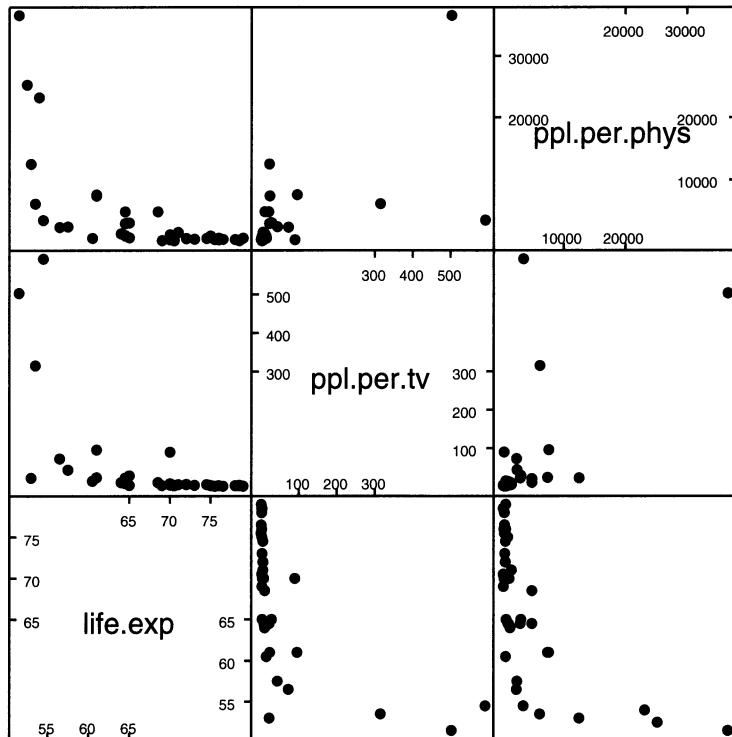


FIGURE 4.13. Televisions, physicians, and life expectancy.  
 (grap/code/grap.f5.s), (grap/code/grap.f5.sas), (grap/figure/grap.f5.eps.gz)

- Remove curvature. This chapter.
- Remove asymmetry. This chapter.
- Respond to systematic residuals. Chapters 8 and 11.

The family of power transformations  $T_p(x)$ , often called the *Box-Cox transformations* (Box and Cox, 1964), are given by

$$T_p(x) = \begin{cases} x^p & (p > 0) \\ \ln(x) & (p = 0) \\ -x^p & (p < 0) \end{cases} \quad (4.1)$$

Notice that the family includes both positive and negative powers, with the logarithm taking the place of the 0 power. The negative powers have a

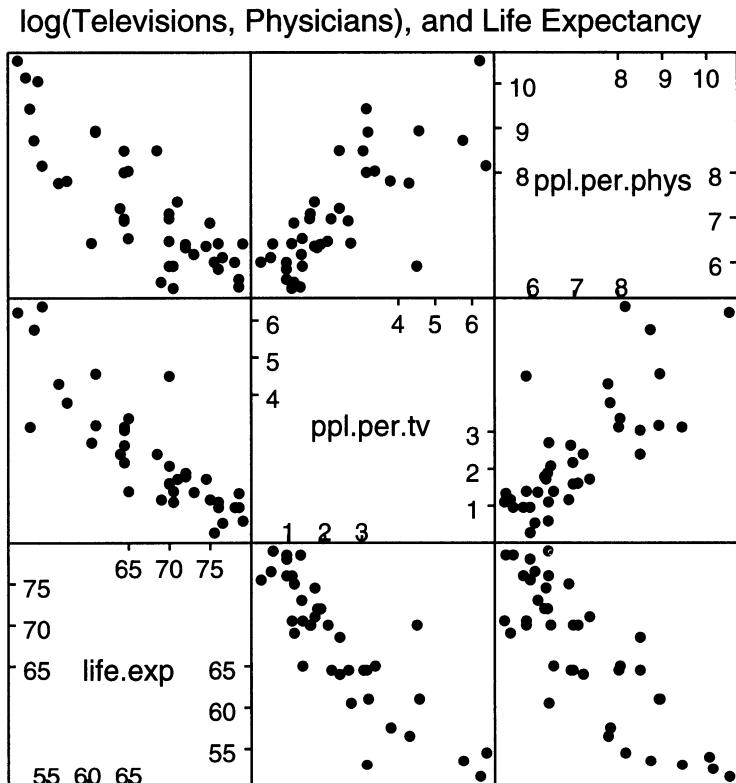
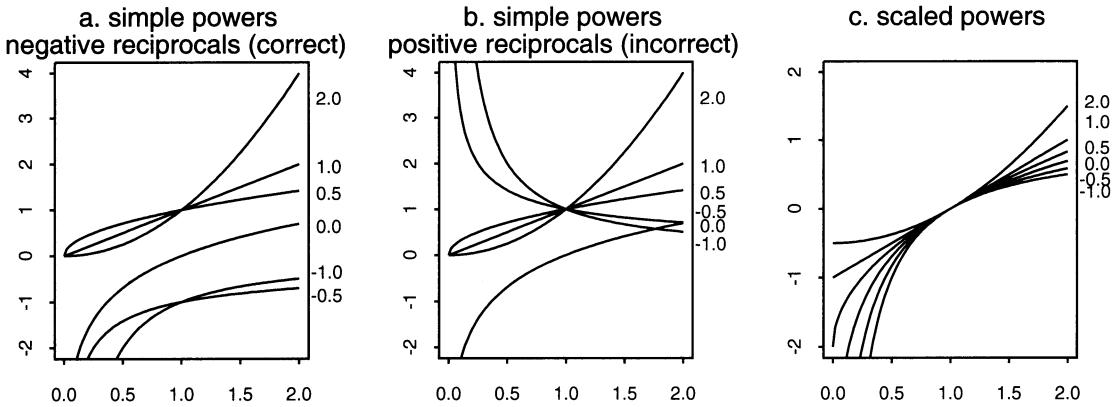


FIGURE 4.14.  $\log(\text{televisions})$ ,  $\log(\text{physicians})$ , and life expectancy.  
 (graph/code/grap.f6.s), (graph/code/grap.f6.sas), (graph/figure/grap.f6.eps.gz)

negative sign to maintain the same direction of the monotonicity; if  $x_1 < x_2$ , then  $T_p(x_1) < T_p(x_2)$  for all  $p$ . When the data are nonnegative but contain zero values, logarithms and negative powers are not defined. In this case we often add a “start” value, frequently  $\frac{1}{2}$ , to the data values before taking the log or power transformation.

When we wish to study the mathematical properties of these transformations, we use the related family of scaled power transformations  $T_p^*(x)$  given by

$$T_p^*(x) = \begin{cases} \frac{x^p - 1}{p} & (p \neq 0) \\ \ln(x) & (p = 0) \end{cases} \quad (4.2)$$



$$T_p(x) = \frac{x^p}{\text{sign}(p)}$$

$$W_p(x) = x^p$$

$$T_p^*(x) = \frac{x^p - 1}{p}$$

FIGURE 4.15. Power Transformations. The smooth transitions between the scaled curves in Figure 4.15c is the justification for using the family of power transformations  $T_p^*(x)$  in Equation (4.2). This is the only one of three panels in which both (a) the monotonicity of the individual powers is visible and (b) the simple relation between the curves and the sequence of powers in the ladder of powers  $p = -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2$  is retained. Figure 4.15a keeps the monotonicity but loses the sequencing. Figure 4.15b, which doesn't negate the reciprocals, is very hard to read because two of the curves are monotone decreasing and four are monotone increasing. Figure 4.15 is based on Figures 4-2 and 4-3 of (Emerson and Stoto, 1983).

(grap/code/grap.f8.s), (grap/figure/grap.f8.eps.gz)

The scaling in  $T_p^*(x)$  gives the same value  $T_p^*(1) = 0$  and derivative  $\frac{d}{dx} T_p^*(1) = 1$  for all  $p$ .

There is also a third family of power transformations  $W_p(x)$  given by

$$W_p(x) = \begin{cases} x^p & (p \neq 0) \\ \ln(x) & (p = 0) \end{cases} \quad \begin{array}{l} \text{Do not use this form,} \\ \text{the reciprocal is not negated.} \end{array} \quad (4.3)$$

that is occasionally (and incorrectly) used. This family does not negate the reciprocals; hence, as we see in Figure 4.15b, it is very difficult to read.

Figure 4.15 shows the plots of all three families: the two parameterizations of the Box–Cox power transformations  $T_p(x)$  and  $T_p^*(x)$ , and the third, poorly parameterized power family  $W_p(x)$ . There are several things to note in these graphs.

1. Figure 4.15a, the plots of  $T_p(x)$ , correctly negates the reciprocals, thereby maintaining a positive slope for all curves and permitting the perception that these are all monotone transformations.
2. In Figure 4.15b, the plots of  $W_p(x)$ , we see that the plots of the two reciprocal transformations have negative slope and that all the others have positive slope. This reversal interferes with the perception of the monotonicity of the transformations.
3. Figure 4.15c, the plots of  $T_p^*(x)$ , is used to study the mathematical and geometric properties of the family of transformations. The individual formulas in Equations (4.1) and (4.2) are linear functions of each other; hence the properties and appearance of the individual lines in the graphs based on them are equivalent. Equation (4.1) is simpler for hand arithmetic. Equation (4.2) makes evident that the powers (including 0 and negative) are simply and systematically related. Taking the negative of the reciprocal explains how the negative powers fits in. Showing how the 0 power or logarithm fits in is trickier; we use l'Hôpital's rule:

$$\lim_{p \rightarrow 0} \frac{x^p - 1}{p} = \lim_{p \rightarrow 0} \frac{\frac{d}{dp}(x^p - 1)}{\frac{d}{dp}p} = \lim_{p \rightarrow 0} x^p \ln x = \ln x$$

The *ladder of powers* is the sequential set of power transformations with  $p = -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2$ .

## 4.8 Life Expectancy Example—Continued

We look again at the plot of `life.exp` vs `ppl.per.phys` from Figures 4.13 and 4.14 where we see that taking the logarithm of `ppl.per.phys` straightened out the graph. In Figure 4.16 we use the functions in (`splus.library/ladder.cartesian.s`) to take the full set of powers (in the ladder of powers) of each of these two variables and plot them against each other. It is apparent from these plots that any power of `life.exp` plots as a straight line against the log of the number of physicians (power = 0). This is unusual behavior. More typically the shape of the plot shifts as the power of either variable shifts. This calls for further investigation.

We plot in Figure 4.17 the various powers against `life.exp`. Equation (4.2) is plotted twice. In panel a this equation is plotted for each of the values  $p = -2, -1, 0, .5, 1, 2$ , where  $p = 0$  represents the log transformation. Panel b shows more detail than panel a; each plot in panel b is rescaled by multiplying the transformed `life.exp` by a constant chosen to raise or lower the plot to have a vertical range in the vicinity of 10. In both panels we see that within the observed range of values (51, 79) of `life exp`, all

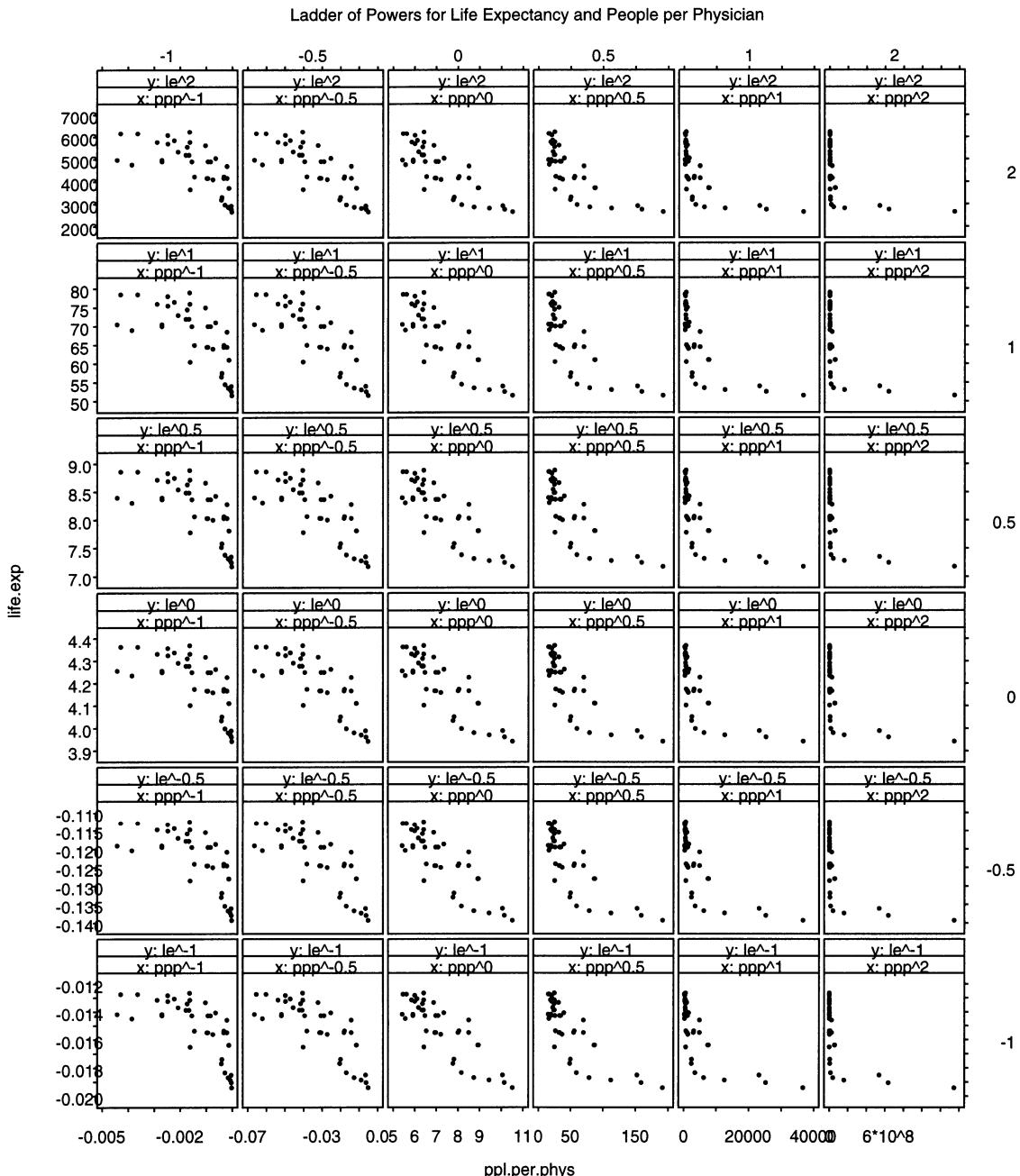


FIGURE 4.16. Ladder of powers for life expectancy and people per physician.  
(`grap/code/grap.f7.s`), (`grap/figure/grap.f7.eps.gz`)

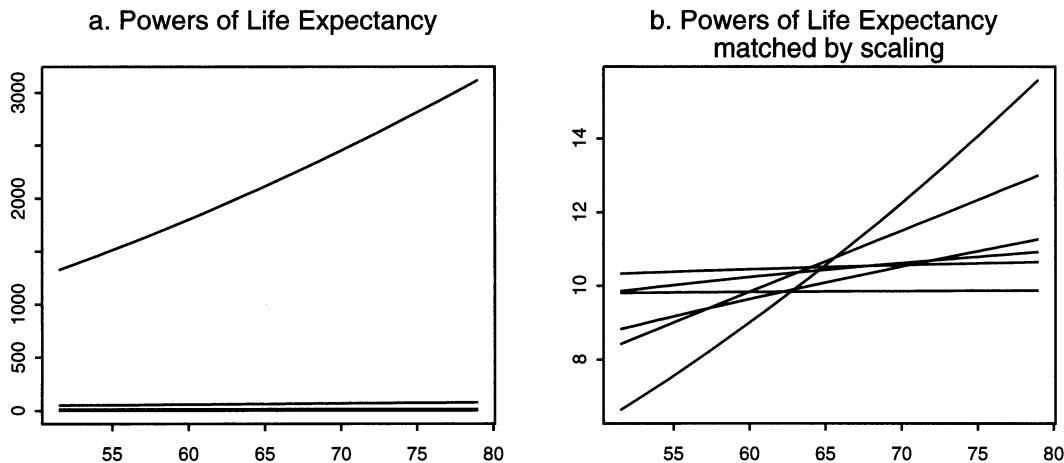


FIGURE 4.17. Powers of life expectancy.  
 (grap/code/grap.f9.s), (grap/figure/grap.f9.eps.gz)

the simple power transformations are essentially linear. This explains why all panels in the  $ppp^0$  column of Figure 4.16 are almost identical.

The columns of Figure 4.16 look different from each other. We look (for convenience) at row `life.exp^1`, with the original scaling of life expectancy, and note that the shape of the graphs shifts from concave-SW through diagonal to concave-NE as the power of `ppp` (people per physician) increases. We need to look at just this single variable as it moves through the series of powers. We do so in Figure 4.18. Panel 4.18a shows the boxplots, panel 4.18b shows the dotplots, and panel 4.18c shows the stem-and-leaf plots. All three panels show the same information. At the positive powers, the data for `ppp` are extremely asymmetric; they are bunched up at the low end of the scale. As the power moves from positive to negative, the center moves toward the higher end and the distribution becomes more symmetric. At the negative powers the data become asymmetric again; this time they are bunched up at the high end. If symmetry for just one variable were the only objective, we might try the  $-3$  power  $-(x^{-3})$ .

The boxplots show the shift in center as the dot for the median moves from one side to the other. They show the shift in symmetry as the center box increases from a small portion to a large fraction of the total width and as the whisker and outliers shift from one side to the other. The dotplots show the same information by density and spread of dots. Both dotplots and boxplots have the same scale. The stem-and-leaf is essentially a density plot. It shows the points bunched up at the low values for positive powers,

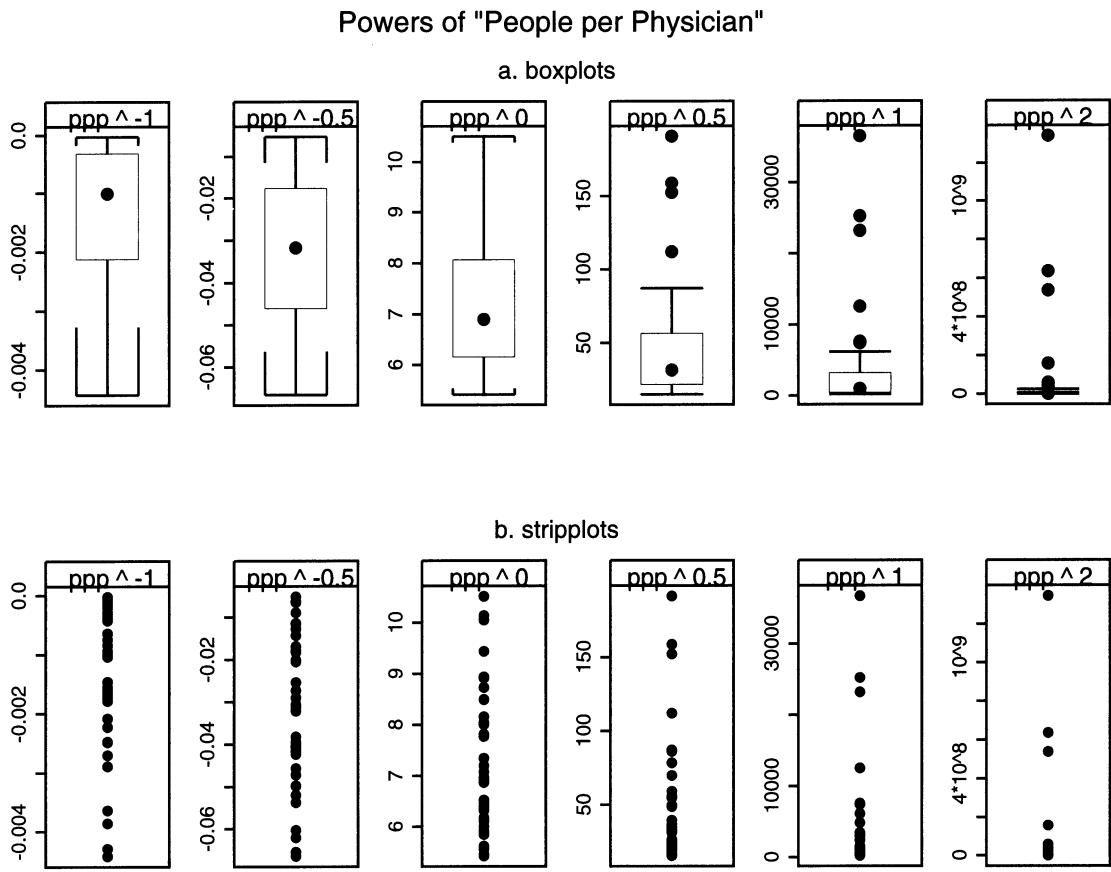


FIGURE 4.18. Powers of life expectancy: boxplots, strip plots, and stem-and-leaf. Stem-and-leaf appears in continuation of figure.

(`grap/code/grap.f10.s`), (`grap/figure/grap.f10.eps.gz`)

centered and symmetric for 0 power, and bunched at the high values for the negative powers.

## 4.9 SAS Graphics

SAS also has a substantial graphics capability. Interested readers should consult the excellent collection of SAS graphics macros described by (Friendly, 1991) and available in his companion website (Friendly, 2004).

c. stem-and-leaf

---

S-PLUS (grap/transcript/grap.f10.ste):

| $(\text{People per Physician})^{-1}$<br>(3 places to the left) | $(\text{People per Physician})^{-0.5}$<br>(2 places to the left) | $(\text{People per Physician})^0$<br>(at the colon) |
|--|--|---|
| -4 : 43  | -6 : 7620  | 5 : 4566899   |
| -3 : 96  | -5 : 42200   | 6 : 00123444445599                                  |
| -3 :   | -4 : 76221100  | 7 : 0012388   |
| -2 : 97755   | -3 : 982111  | 8 : 00255799  |
| -2 : 21  | -2 : 97510   | 9 : 4   |
| -1 : 87766665  | -1 : 88744321  | 10 : 115  |
| -1 : 00  | -0 : 9765  |   |
| -0 : 99876   |  |   |
| -0 : 44333222111000  |  |   |

| $(\text{People per Physician})^{0.5}$<br>(1 place to the right) | $(\text{People per Physician})^1$<br>(3 places to the right) | $(\text{People per Physician})^2$<br>(6 places to the right) |
|---|--|--|
| 1 : 5567999   | 0 : 2233344444566666667                                      | 0 : 11111112222334444459                                     |
| 2 : 001244455556  | 1 : 0011236  | 1 : 01148  |
| 3 : 1233479   | 2 : 45   | 2 : 4  |
| 4 : 9   | 3 : 015  | 3 :  |
| 5 : 0569  | 4 : 99   | 4 :  |
| 6 :   | 5 :  | 5 : 6  |
| 7 : 009   | 6 : 2  | 6 : 1  |
| 8 : 67  | 7 : 46   | 7 :  |
| 9 :   |  | 8 :  |
| 10 :  | High: 12550 23193 25229                                      | 9 : 06   |
| 11 : 2  | 36660  |  |
|   |  | High: 1.2e7 2.3e7 2.3e7 3.8e7                                |
| High: 152.2925 158.8364   |  | 5.5e7 5.7e7 1.5e8 5.3e8                                      |
| 191.4680  |  | 6.3e8 1.3e9  |

---

FIGURE 4.18 continued. Powers of life expectancy: stem-and-leaf.  
(grap/code/grap.f10.s)

## 4.10 Exercises

We recommend that you begin all exercises by examining a scatterplot matrix of the variables. Based on the scatterplot matrix, you might wish to consider transforming some of the variables.

- 4.1.** The U.S. Draft Lottery held in December 1969 was meant to prioritize the order in which young men would be drafted during 1970 for service in the Vietnam War. Each of the 366 dates was written on a small piece of paper and placed in a capsule. In chronological order the capsules were placed in a vessel and the vessel was stirred. The capsules were then drawn one at a time, thereby assigning ranks 1 to 366 to the dates. But because of inadequate stirring, men with birthdays toward the end of the year tended to have higher rank and thus greater vulnerability to the draft than men born early in the year. The data file (`datasets/draft70mn.dat`), available from (Data Archive, 1997), contains 12 columns for the months January through December. For each month, the  $m^{\text{th}}$  entry represents the rank between 1 and 366 for the  $m^{\text{th}}$  day of that month. You can read the data with (`grap/code/draft70mn-read.s`) or (`grap/code/draft70mn-read.sas`). Produce parallel boxplots for the months arranged chronologically, and draw the line segments connecting the medians of adjacent months. This illustrates the claim that the drawing was not random.
- 4.2.** (Sokal and Rohlf, 1981), later in (Hand et al., 1994), examined factors contributing to air pollution in 41 U.S. cities, as assessed by sulfur dioxide content. The dataset appears in (`datasets/usair.dat`). The variables are

`SO2`: SO<sub>2</sub> content of air in mcg per cubic meter

`temp`: average annual temperature in degrees Fahrenheit

`mfgfirms`: number of manufacturing firms employing at least 20 workers

`popn`: 1970 census population size, in thousands

`wind`: average annual wind speed in mph

`precip`: average annual precipitation in inches

`raindays`: average number of days per year having precipitation

Produce a scatterplot matrix for these data both before and after log-transforming all 7 variables. Compare the sploms and explain why the log transformation is appropriate for these data. Which of the 6 pre-

dictor variables are most highly correlated with the logged response `S02`? Which of the 15 pairs of logged predictors appear to be highly correlated?

- 4.3.** (Vandaele, 1978), also in (Hand et al., 1994), contains data on the reported 1960 crime rate per million population and 13 potential explanatory variables for each of 47 states. The data appear in the file (`datasets/uscrime.dat`). The variables are

`R`: reported crime rate per million population

`Age`: the number of males aged 14 to 24

`S`: 1 if Southern state, 0 otherwise

`Ed`: 10 times mean years of schooling of population age 25 or older

`Ex0`: 1960 per capital expenditures by state and local government on police protection

`Ex1`: same as `Ex0` but for 1959

`LF`: number of employed urban males aged 14–24 per 1000 such individuals

`M`: number of males per 1000 females

`N`: state population size in hundred thousands

`NW`: number of nonwhites per 1000 population

`U1`: unemployment rate per 1000 among urban males aged 14–24

`U2`: same as `U1` but for age group 25–39

`W`: a measure of wealth, units = 10 dollars

`X`: number of families per 1000 earning below one half of the median income (a measure of income inequality)

Construct a scatterplot matrix for these data. The variables other than `R` will be referred to as predictors in Exercise 9.6. Based on this plot, which pairs of predictors are highly correlated? Which predictors are most closely linearly associated with `R`?

- 4.4.** (Hand et al., 1994) contains data on the average `mortality` rate for males per 100,000 and the `calcium` concentration (ppm) in the public drinking water in 61 large towns in England and Wales, averaged over the years 1958 to 1964. Each town was also identified as being at least as far north as the town Derby (`derbynor=1`) or south of Derby (`derbynor=0`). The data are in the file (`datasets/water.dat`). Exer-

cise 10.4 will request investigation of the relationship between water hardness (`calcium`) and `mortality`. The sampling units are towns in two regions. Produce two separate but adjacent plots of `mortality` vs `calcium` for the two regions specified by `derbynor`. Discuss the differences you see in the two plots.

- 4.5. (Williams, 1959), also in (Hand et al., 1994), presents data on the `density` and `hardness` of 36 Australian eucalyptus trees. The data file is (`datasets/hardness.dat`). Determine a transformation from the Box–Cox family that will make `hardness` as close as possible to normally distributed. The result will be useful for Exercise 11.2, which requests a model of `hardness` as a function of `density`.
- 4.6. Following a severe water shortage in Concord, New Hampshire, during the late 1970s, conservation measures were instituted there in 1980. The shortage became especially acute during the summer of 1981. (Hamilton, 1983) and (Hamilton, 1992) discuss models of the 1981 household water consumption in Concord, New Hampshire, in terms of several other variables. The data file, (`datasets/concord.dat`), contains information on the following variables from each of 496 households:

`water81`: cubic feet of household water use in 1981

`water80`: cubic feet of household water use in 1980

`income`: 1981 household income in \$1000s

`educat`: education of head of household, in years

`peop81`: number of people living in household in summer 1981

`retired`: 1 if head of household is retired, otherwise 0

Exercise 11.3 requests the modeling of household water use in 1981 in Concord as a function of 5 predictors. To assist with this task, investigate which transformation from the ladder of powers family will bring the response variable, `water81`, as close as possible to normality.

# Introductory Inference

In this chapter we discuss selected topics and issues dealing with statistical inferences from samples to populations, building upon the brief introduction to these ideas in Chapter 3. The discussion here is at an intermediate technical level and at a speed appropriate for review of material learned in the prerequisite course.

We provide procedures for constructing confidence intervals and conducting hypothesis tests for several frequently encountered situations.

## 5.1 Normal ( $z$ ) Intervals and Tests

A confidence interval and test concerning a population mean were briefly described in Chapter 3. This is a more extensive presentation.

The confidence interval on the mean  $\mu$  of a normal population when the standard deviation is known was given in Equation (3.12). The development there assumed that the population was normal. However, since the Central Limit Theorem discussed in Section 3.4.2 guarantees that  $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$  is approximately normally distributed if  $n$  is “sufficiently large”, the interval

$$\left( \bar{y} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) \quad (5.1)$$

is an approximate two-sided  $100(1-\alpha)\%$  confidence interval when the population is not normal. The closer the population is to a normal population, the closer will be this interval's coverage probability to  $1-\alpha$ . Thus, in the nonnormal case, this interval is an approximate CI for  $\mu$ .

TABLE 5.1. Confidence intervals and tests with known standard deviation  $\sigma$ , where  $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$  and  $z_{\text{calc}} = \frac{\bar{y} - \mu_0}{\sigma_{\bar{y}}}$ .

| $H_0$            | $H_1$            | Tests                                      |   | $p$ -value                  | Confidence interval  |       |  |
|------------------|------------------|--|---|-----------------------------|--|-------|--|
|                  |                  | Rejection region                           |   |                             | Lower  | Upper |  |
|                  |                  | $z$ -scale                                 | $y$ -scale  |                             |  |       |  |
| $\mu \leq \mu_0$ | $\mu > \mu_0$    | $z_{\text{calc}} > z_\alpha$               | $\bar{y} > \mu_0 + z_\alpha \sigma_{\bar{y}}$               | $P(Z > z_{\text{calc}})$    | $(\bar{y} - z_\alpha \sigma_{\bar{y}}, \infty)$  |       |  |
| $\mu \geq \mu_0$ | $\mu < \mu_0$    | $z_{\text{calc}} < -z_\alpha$              | $\bar{y} < \mu_0 - z_\alpha \sigma_{\bar{y}}$               | $P(Z < z_{\text{calc}})$    | $(-\infty, \bar{y} + z_\alpha \sigma_{\bar{y}})$   |       |  |
| $\mu = \mu_0$    | $\mu \neq \mu_0$ | $ z_{\text{calc}}  > z_{\frac{\alpha}{2}}$ | $ \bar{y} - \mu_0  > z_{\frac{\alpha}{2}} \sigma_{\bar{y}}$ | $2P(Z >  z_{\text{calc}} )$ | $(\bar{y} - z_{\frac{\alpha}{2}} \sigma_{\bar{y}}, \bar{y} + z_{\frac{\alpha}{2}} \sigma_{\bar{y}})$ |       |  |

Also shown in the rightmost column of Table 5.1 are one-sided confidence intervals for  $\mu$ . These are less commonly used than two-sided intervals because they have infinite width. But they are sometimes encountered in contexts where an upper or lower bound for  $\mu$  is required.

### 5.1.1 Test of a Hypothesis Concerning the Mean of a Population Having Known Standard Deviation

We consider three pairs of null and alternative hypotheses in Table 5.1.

The first two pairs are called *one-tailed* or *one-sided* tests because their rejection regions lie on one side of the normal distribution. The third pair has a two-sided rejection region and hence is termed a two-tailed or two-sided test. In any given problem, only one of these three is applicable. For expository purposes, it is convenient to discuss them together.

Some authors formulate the one-sided tests with sharp null hypotheses

| $H_0$         | $H_1$         |
|---------------|---------------|
| $\mu = \mu_0$ | $\mu > \mu_0$ |
| $\mu = \mu_0$ | $\mu < \mu_0$ |

However, with the sharp formulation it can happen that neither the null nor alternative hypothesis is true, in which case the action of rejecting the null hypothesis has an uncertain interpretation.

For the first pair of hypotheses, we reject  $H_0$  if the sample mean is sufficiently greater than  $\mu_0$ , specifically, if  $\bar{y} > (\mu_0 + z_\alpha \sigma / \sqrt{n})$ . Otherwise,  $H_0$  is retained. Equivalently, if we define the calculated  $Z$  statistic under the

null hypothesis,

$$z_{\text{calc}} = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \quad (5.2)$$

then we reject  $H_0$  if  $z_{\text{calc}} > z_\alpha$ ; otherwise  $H_0$  is retained. The  $p$ -value of this test is  $P(Z > z_{\text{calc}})$ .

The testing procedure for the second pair of hypotheses is the mirror image of the first pair.  $H_0$  is rejected if  $\bar{y} < (\mu_0 - z_\alpha \sigma / \sqrt{n})$  and retained otherwise. Equivalently, we reject  $H_0$  if  $z_{\text{calc}} < -z_\alpha$ . The  $p$ -value of this test is  $P(Z < z_{\text{calc}})$ .

For the third pair, the two-sided test, we reject  $H_0$  if either

$$\bar{y} < (\mu_0 - z_{\frac{\alpha}{2}} \sigma / \sqrt{n}) \quad \text{or} \quad \bar{y} > (\mu_0 + z_{\frac{\alpha}{2}} \sigma / \sqrt{n});$$

equivalently, if  $|z_{\text{calc}}| > z_{\frac{\alpha}{2}}$ . The  $p$ -value of this two-sided test is  $2P(Z > |z_{\text{calc}}|)$ . Hence  $H_0$  is rejected if  $\bar{y}$  is sufficiently above or sufficiently below  $\mu_0$ . Another equivalent rule is to reject  $H_0$  if and only if  $\mu_0$  falls outside the  $100(1 - \alpha)\%$  confidence interval for  $\mu$ .

The rejection region for all three pairs is included in Table 5.1.

### 5.1.2 Confidence Intervals for Unknown Population Proportion $p$

We consider a confidence interval on the unknown proportion  $p$  of successes in a population consisting of items or people labeled as successes and failures. Such populations are very frequently encountered in practice. For example, we might wish to estimate the proportion  $p$  of voters who will ultimately vote for a particular candidate, based on a random sample from a population of likely voters. Inspectors of industrial output may wish to estimate the proportion  $p$  of a day's output that is defective based on a random sampling of this output.

Suppose the sample size is  $n$ , of which  $Y$  items are successes and that  $\hat{p} = \frac{Y}{n}$ , a point estimator of  $p$ , is the proportion of sampled items that fall into the success category. Until recently, the usual  $100(1 - \alpha)\%$  confidence interval for  $p$  suggested in the statistics literature was

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

This interval is satisfactory when  $n \geq 100$  unless  $p$  is close to either 0 or 1. The large sample is needed for the Central Limit Theorem to assure us that the discrete probability distribution of  $\hat{p}$  is adequately approximated by the continuous normal distribution.

(Agresti and Caffo, 2000) suggest the following alternative confidence interval for  $p$ , where  $\tilde{p} = \frac{Y+2}{n+4}$  and  $\tilde{n} = n + 4$ :

$$\tilde{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} \quad (5.3)$$

Agresti and Caffo show that their interval has coverage probability that typically is much closer to the nominal  $1 - \alpha$  than the usual confidence interval. It differs from the usual interval in that we artificially add two successes and two failures to the original sample. For  $p$  near 0 or 1, the usual interval, which is symmetric about  $\hat{p}$ , may extend beyond one of these extremes and hence not make sense, while the alternative interval is likely to remain entirely between 0 and 1.

Conventional one-sided confidence intervals for  $p$  are shown in Table 5.2. Comparable to Agresti and Caffo's proposed two-sided interval, (Cai, 2003) proposes improved one-sided confidence intervals for  $p$  having coverage probabilities closer to  $1 - \alpha$  than the conventional intervals. These lower and upper intervals, respectively, are

$$[0, \mathcal{F}_{\text{Be}}^{-1}(1 - \alpha | Y + .5, n - Y + .5)] \quad (5.4)$$

and

$$[\mathcal{F}_{\text{Be}}^{-1}(\alpha | Y + .5, n - Y + .5), 1] \quad (5.5)$$

where  $\mathcal{F}_{\text{Be}}^{-1}(\alpha | a, b)$  denotes the value  $x$  of a random variable corresponding to the  $100\alpha$  percentile of the beta distribution with parameters  $a$  and  $b$ . See Section D.1 for a brief discussion of the beta distribution.

### 5.1.3 Tests on an Unknown Population Proportion $p$

Assume we have a sample of  $n \geq 100$  items from a population of successes and failures, and we wish to test a hypothesis about the proportion  $p$  of successes. Paralleling the previous discussion of tests on a population mean, there are two one-tailed tests and one two-tailed test as detailed in Table 5.2. As in the discussion of the confidence interval on  $p$ , the normal approximation to the distribution of  $\hat{p}$  requires that  $n$  not be too small.

### 5.1.4 Example—One-Sided Hypothesis Test Concerning a Population Proportion

As an illustration, suppose a pollster wishes to test the hypothesis that at least 50% of a city's voting population favors a certain bond issue. Let us conduct this test at  $\alpha = 0.01$  if only 222 of a random sample of 500 persons in the population favors this bond issue.

TABLE 5.2. Conventional confidence intervals and tests with unknown population proportion  $p$ , where  $\sigma_{p_0} = \sqrt{\frac{p_0(1-p_0)}{n}}$  and  $z_{\text{calc}} = \frac{\hat{p} - p_0}{\sigma_{p_0}}$  for tests, and  $s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$  for confidence intervals.

| $H_0$        | $H_1$        | Tests                                      |   | $p$ -value                  | Confidence interval  |  |  |
|--------------|--------------|--|---|-----------------------------|--|--|--|
|              |              | Rejection region                           |   |                             |  |  |  |
|              |              | z-scale                                    | p-scale   |                             |  |  |  |
| $p \leq p_0$ | $p > p_0$    | $z_{\text{calc}} > z_\alpha$               | $\hat{p} > p_0 + z_\alpha \sigma_{p_0}$               | $P(Z > z_{\text{calc}})$    | $(\hat{p} - z_\alpha s_{\hat{p}}, 1)$  |  |  |
| $p \geq p_0$ | $p < p_0$    | $z_{\text{calc}} < -z_\alpha$              | $\hat{p} < p_0 - z_\alpha \sigma_{p_0}$               | $P(Z < z_{\text{calc}})$    | $(0, \hat{p} + z_\alpha s_{\hat{p}})$  |  |  |
| $p = p_0$    | $p \neq p_0$ | $ z_{\text{calc}}  > z_{\frac{\alpha}{2}}$ | $ \hat{p} - p_0  > z_{\frac{\alpha}{2}} \sigma_{p_0}$ | $2P(Z >  z_{\text{calc}} )$ | $(\hat{p} - z_{\frac{\alpha}{2}} s_{\hat{p}}, \hat{p} + z_{\frac{\alpha}{2}} s_{\hat{p}})$ |  |  |

Here  $H_1$  is of the form  $H_1: p < .50$ . We reject  $H_0$  if

$$\hat{p} < p_0 - z_{.01} \sqrt{\frac{p_0(1-p_0)}{n}} \quad (5.6)$$

With  $p_0 = .50$ ,  $\hat{p} = 222/500 = 0.444$ ,  $z_{.01} = 2.326$ , and

$$\sqrt{p_0(1-p_0)/n} = .0224 \quad (5.7)$$

we find that the right side of (5.6) is 0.448 so that  $H_0$  is (barely) rejected. In this example,  $z_{\text{calc}} = -2.500$  so that the  $p$ -value =  $P(Z < -2.500) = .0062$ . Hence we reject  $H_0$  when  $\alpha > .0062$ .

## 5.2 *t*-intervals and Tests for the Mean of a Population Having Unknown Standard Deviation

When we wish to construct a confidence interval or test a hypothesis about an unknown population mean  $\mu$ , more often than not the population standard deviation  $\sigma$  is also unknown. Then we must use the sample standard deviation  $s$  from Equation 3.8 in place of  $\sigma$  when standardizing  $\bar{y}$ . But while  $\frac{\bar{y}-\mu}{\sigma/\sqrt{n}}$  has an approximate normal distribution if  $n$  is sufficiently large,  $\frac{\bar{y}-\mu}{s/\sqrt{n}}$  has an approximate *t* distribution with  $n - 1$  degrees-of-freedom. The latter standardization with  $s$  in the denominator has more variability than the former standardization with  $\sigma$  in the denominator. The *t* distribution reflects this increased variability because it has less probability concentrated near zero than does the standard normal distribution.

The confidence interval and tests for  $\mu$  using the *t* distribution are identical to those using the normal (*Z*) distribution, with  $t_{\text{calc}}$  replacing  $z_{\text{calc}}$  and  $t_\alpha$  replacing  $z_\alpha$ . For this problem, the degrees-of-freedom parameter for the *t* distribution is always  $n - 1$ .

For example, to test  $H_0: \mu \geq \mu_0$  vs  $H_1: \mu < \mu_0$ , we reject  $H_0$  if

$$t_{\text{calc}} = \frac{\bar{y} - \mu}{s/\sqrt{n}} < -t_\alpha \quad (5.8)$$

Here the  $p$ -value =  $P(t < t_{\text{calc}})$  is calculated from the  $t$  distribution with  $n - 1$  degrees of freedom.

The approximate confidence interval on  $\mu$  is  $\bar{y} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ .

### Example—Inference on a Population Mean $\mu$

(Hand et al., 1994) presents a data set, reproduced in ([datasets/vocab.dat](#)), containing the scores on a vocabulary test of a sample of 54 students from a study population. Assume that the test was constructed to have a mean score of 10 in the general population. We desire to assess whether the mean score of the study population is also  $\mu = 10$ . Assuming that standard deviation for the study population is not known, we wish to calculate a 95% confidence interval for  $\mu$  and to test  $H_0: \mu = 10$  vs  $H_1: \mu \neq 10$ .

We begin by looking at a stem-and-leaf display of the sample data to see if the underlying assumption of normality is tenable. We observe in Figure 5.1 that the sample is slightly positively skewed with one high value that may be considered an outlier. Based on the Central Limit Theorem, the  $t$ -based procedures are justified here. The small  $p$ -value ( $p \approx 3 \times 10^{-14}$ ) is strong evidence that  $\mu$  is not 10. The 95% confidence interval (12.3, 13.4) suggests that the mean score is close to 12.9 in the study population.

We examine a nonparametric approach to this problem in Section 16.2.

## 5.3 Confidence Interval on the Variance or Standard Deviation of a Normal Population

Let the (unbiased) estimator of  $\sigma^2$  based on a sample of size  $n$  be denoted  $s^2$ . Then  $(n - 1)s^2/\sigma^2$  has a  $\chi^2$  distribution with  $\text{df} = n - 1$ . Thus

$$P\left(\chi^2_{\frac{\alpha}{2}, n-1} < (n-1)s^2/\sigma^2 < \chi^2_{1-\frac{\alpha}{2}, n-1}\right) = 1 - \alpha$$

Inverting this statement leads to the  $100(1 - \alpha)\%$  confidence interval for  $\sigma^2$ :

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}, n-1}}, \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}, n-1}}\right)$$

---



---

```
S-PLUS (iinf/transcript/vocab-stem.st):
> stem(vocab$score)

N = 54 Median = 13
Quartiles = 11, 14

Decimal point is at the colon

 9 : 0
10 : 0000
11 : 0000000000000000
12 : 0000000
13 : 000000000
14 : 000000000
15 : 0000
16 : 00000
17 : 0
18 :
19 : 0
```

---

FIGURE 5.1. Stem-and-leaf display of vocabulary scores.  
(iinf/code/vocab.s), (iinf/transcript/vocab.st)

If instead a CI on  $\sigma$  is desired, take the square roots of both the lower and upper limits in the above.

The distribution of  $(n - 1)s^2/\sigma^2$  can also be used to conduct a test about  $\sigma^2$  (or  $\sigma$ ). For example, to test  $H_0: \sigma^2 \leq \sigma_0^2$  vs  $H_1: \sigma^2 > \sigma_0^2$ , the  $p$ -value is  $1 - F_{\chi_{n-1}^2}[(n - 1)s^2/\sigma_0^2]$ . Tests of the equality of two or more variances are addressed in Section 6.10.

## 5.4 Comparisons of Two Populations Based on Independent Samples

Two populations are often compared by constructing confidence intervals on the difference of the population means or proportions. In this discussion it is assumed that random samples are independently selected from each population.

### 5.4.1 Confidence Intervals on the Difference Between Two Population Proportions

The need for confidence intervals on the difference of two proportions is frequently encountered. We might wish to estimate the difference in the proportions of voters in two populations who favor a particular candidate, or the difference in the proportions of defectives produced by two company locations.

Labeling the populations as 1 and 2, the traditional confidence interval, assuming that both populations are large and that neither proportion is close to either 0 or 1, is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \quad (5.9)$$

Agresti and Caffo also provided an improved confidence interval for this situation, which again provides confidence closer to  $100(1 - \alpha)\%$  than the preceding interval. For  $i = 1, 2$ , let  $\tilde{p}_i = \frac{Y_i + 1}{n_i + 2}$ , i.e., revise the estimate of  $p_i$  by adding one success and one failure to both samples. Then the improved interval is

$$(\tilde{p}_1 - \tilde{p}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{n_2 + 2}}$$

### 5.4.2 Confidence Interval on the Difference of Between Two Means

For a CI on a difference of two means under the assumption that the population variances are unknown, there are two cases. If the variances can be assumed to be equal, their common value is estimated as a weighted average of the two individual sample variances. In general, the process of calculating such meaningfully weighted averages is referred to as *pooling*, and the result in this context is called a pooled variance:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (5.10)$$

The pooled estimator  $s_p^2$  has more degrees of freedom (uses more information) than either  $s_1^2$  or  $s_2^2$  for the estimation of the common population variance. When the pooled variance is used as the denominator of  $F$ -tests it provides a more powerful test than either of the components, and therefore it is preferred for this purpose. Then the CI is

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\frac{\alpha}{2}, n_1 + n_2 - 2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

In the case where the variances cannot be assumed equal, there are two procedures. The Satterthwaite option is

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\frac{\alpha}{2}, df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where  $df$  is the integer part of

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$$

The Satterthwaite option is sometimes referred to as the Welch option.

The Cochran option is

$$(\bar{y}_1 - \bar{y}_2) \pm t \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where  $t = \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2}$ ,  $w_i = s_i^2/n_i$ , and  $t_i$  is  $t_{\frac{\alpha}{2}, (n_i-1)}$ .

The Satterthwaite option is more commonly used than the Cochran option. In practice, they lead to similar results.

### 5.4.3 Tests Comparing Two Population Means When the Samples Are Independent

There are two situations to consider with independent samples. When the populations may be assumed to have a common unknown variance  $\sigma$ , the calculated  $t$  statistic is

$$t_{\text{calc}} = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (5.11)$$

where  $s_p$  was defined in Equation (5.10) and  $t_{\text{calc}}$  has  $n_1 + n_2 - 2$  degrees of freedom.

When the two samples might have different unknown variances, then the test is based on

$$s_{(\bar{y}_1 - \bar{y}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \text{and} \quad t_{\text{calc}} = \frac{\bar{y}_1 - \bar{y}_2}{s_{(\bar{y}_1 - \bar{y}_2)}} \quad (5.12)$$

In either case, we consider one of the three tests in Table 5.3.

TABLE 5.3. Confidence intervals and tests for two population means. When the samples are independent and we can assume a common unknown variance, use  $s_{\Delta\bar{y}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$  and  $t_{\text{calc}}$  as given by Equation (5.11). When the samples are independent and we assume different unknown variances, use  $s_{\Delta\bar{y}} = s_{(\bar{y}_1 - \bar{y}_2)}$  and  $t_{\text{calc}}$  as given by Equation (5.12). When the samples are paired, use  $s_{\Delta\bar{y}} = s_d$  and  $t_{\text{calc}}$  as given by Equation (5.13).

| Tests              |                    |  |                             | Confidence interval  |       |
|--------------------|--------------------|--|-----------------------------|--|-------|
|                    |                    | Rejection                                  |                             |  |       |
| $H_0$              | $H_1$              | region                                     | $p$ -value                  | Lower  | Upper |
| $\mu_1 \leq \mu_2$ | $\mu_1 > \mu_2$    | $t_{\text{calc}} > t_\alpha$               | $P(t > t_{\text{calc}})$    | $((\bar{y}_1 - \bar{y}_2) - t_\alpha s_{\Delta\bar{y}}, \infty)$   | )     |
| $\mu_1 \geq \mu_2$ | $\mu_1 < \mu_2$    | $t_{\text{calc}} < -t_\alpha$              | $P(t < t_{\text{calc}})$    | ( $-\infty, (\bar{y}_1 - \bar{y}_2) + t_\alpha s_{\Delta\bar{y}}$ )  |       |
| $\mu_1 = \mu_2$    | $\mu_1 \neq \mu_2$ | $ t_{\text{calc}}  > t_{\frac{\alpha}{2}}$ | $2P(t >  t_{\text{calc}} )$ | $((\bar{y}_1 - \bar{y}_2) - t_{\frac{\alpha}{2}} s_{\Delta\bar{y}}, (\bar{y}_1 - \bar{y}_2) + t_{\frac{\alpha}{2}} s_{\Delta\bar{y}})$ |       |

The test may be carried out in SAS with PROC TTEST using the Satterthwaite option. The  $p$ -value reported by TTEST assumes the two-tailed test. For a one-tailed test, the correct  $p$ -value is half that given by TTEST. The use by PROC TTEST of a prior  $F$ -test of equality of the unknown variances is controversial because using it to decide on the equal variances  $t$ -test distorts the  $p$ -value of the  $t$ -test.

S-PLUS uses the `t.test()` function which performs a one-sample, two-sample, or paired  $t$ -test, or a Welch modified two-sample  $t$ -test. The Welch modification is synonymous with the Satterthwaite option used by SAS.

#### 5.4.4 Comparing the Variances of Two Normal Populations

We assume here that independent random samples are available from both populations. The  $F$  distribution is used to compare the variances  $\sigma_1^2$  and  $\sigma_2^2$  of two normal populations. Let  $s_1^2$  and  $s_2^2$  be the variances of independent random samples of size  $n_i$ ,  $i = 1, 2$  from these populations.

To test

$$H_0: \sigma_1^2 \leq \sigma_2^2$$

vs

$$H_1: \sigma_1^2 > \sigma_2^2$$

define  $F = s_1^2/s_2^2$  and reject  $H_0$  if  $F$  is sufficiently large. The  $p$ -value of the test is  $1 - \mathcal{F}_{F(n_1-1, n_2-1)}(F)$ . The power of this and other  $F$ -tests is sensitive to the second (denominator) df parameter and is usually not adequate unless this  $\text{df} \geq 20$ .

A  $100(1 - \alpha)\%$  confidence interval for a ratio of variances of two normal populations,  $\sigma_1^2/\sigma_2^2$ , is

$$\left( \frac{s_1^2}{s_2^2} \frac{1}{F_{\text{low}}}, \frac{s_1^2}{s_2^2} F_{\text{high}} \right)$$

where

$F_{\text{low}}$  is  $F_{1-\frac{\alpha}{2}, n_1-1, n_2-1}$ , the upper  $100(1 - \frac{\alpha}{2})$  percentage point of an  $F$  distribution with  $n_1 - 1$  and  $n_2 - 1$  degrees of freedom, and

$F_{\text{high}}$  is  $F_{1-\frac{\alpha}{2}, n_2-1, n_1-1}$ , the upper  $100(1 - \frac{\alpha}{2})$  percentage point of an  $F$  distribution with  $n_2 - 1$  and  $n_1 - 1$  degrees of freedom.

An extension to testing the homogeneity of more than two population variances will be presented in Section 6.10.

## 5.5 Paired Data

Sometimes we wish to compare the mean change in a measurement observed on an experimental unit under two different conditions. For example:

1. Compare the subject knowledge of students before and after they receive instruction on the subject.
2. Compare the yield per acre of a population of farms for a crop grown with two different fertilizers.
3. Compare the responses of patients to both an active drug and a placebo, when they are administered each of them in sequential random order with a suitable “washout” period between the administrations.

This “matched pairs” design is superior to a design of the same total size using independent samples because (in illustrations 1 and 3 above) the person to person variation is removed from the comparison of the two administrations, thereby improving the precision of this comparison. The principles of designing experiments to account for and remove extraneous sources of variation are discussed in more detail in Chapter 13.

It is assumed that the populations have a common variance and are approximately normal. Let  $y_{11}, y_{12}, \dots, y_{1n}$  be the sample of  $n$  items from the population under the first condition, having mean  $\mu_1$ , and similarly let  $y_{21}, y_{22}, \dots, y_{2n}$  be the sample from the population under the second condition, having mean  $\mu_2$ .

Define the  $n$  differences  $d_1 = y_{11} - y_{21}$ ,  $d_2 = y_{12} - y_{22}, \dots, d_n = y_{1n} - y_{2n}$ . Let  $\bar{d}$  and  $s_d$  be the mean and standard deviation, respectively, of the sample of  $n$   $d$ 's. Then an approximate  $100(1 - \alpha)\%$  confidence interval on

the mean difference  $\mu_1 - \mu_2$  is  $\bar{d} \pm t_{\frac{\alpha}{2}, n-1} s_{\bar{d}}$  where  $s_{\bar{d}} = s_d / \sqrt{n}$ . Tests of hypotheses proceed similarly to  $t$ -tests for two independent samples. Table 5.3 can still be used, but with

$$s_{\bar{d}} = s_d / \sqrt{n}, \quad \text{and} \quad t_{\text{calc}} = \frac{\bar{d}}{s_{\bar{d}}} \quad (5.13)$$

with degrees of freedom  $n - 1$ .

### Example— $t$ -test on Matched Pairs of Means

(Woods et al., 1986), later in (Hand et al., 1994), investigate whether native English speakers find it easier to learn Greek than native Greek speakers learning English. Thirty-two sentences are written in both languages. Each sentence is scored according to the quantity of errors made by an English speaker learning Greek and by a Greek speaker learning English. It is desired to compare the mean scores of the two groups. The data appear in the file (`datasets/teachers.dat`); the first column is the error score on the English version of the sentence and the second column is the error score on the Greek version of the sentence.

These are 32 pairs of observations because the same sentence is evaluated in both languages. It would be incorrect to regard these as independent samples. The dotplot in Figure 5.2 reveals that for most sentences the English version shows fewer errors. The stem-and-leaf of the differences in Figure 5.3a shows the difference variable is positively skewed so that a transformation is required. Care must be used with a power transformation because many of the differences are negative. The smallest difference is  $-16$ . Therefore, we investigate a square root transformation following the addition of 17 to each value. The second stem-and-leaf in Figure 5.3b illustrates that this transformation succeeds in bringing the data close to normality. Since a difference of zero in the original scale corresponds to a transformed difference of  $\sqrt{17} \approx 4.123$ , the null hypothesis of equal difficulty corresponds to a comparison of the sample means in the transformed scale to 4.123, not to 0. The observed  $p$ -value is .0073, showing a very clear difference in difficulty of learning the two languages. For comparison, the  $t$ -test on the untransformed differences show a  $p$ -value of only .0346.

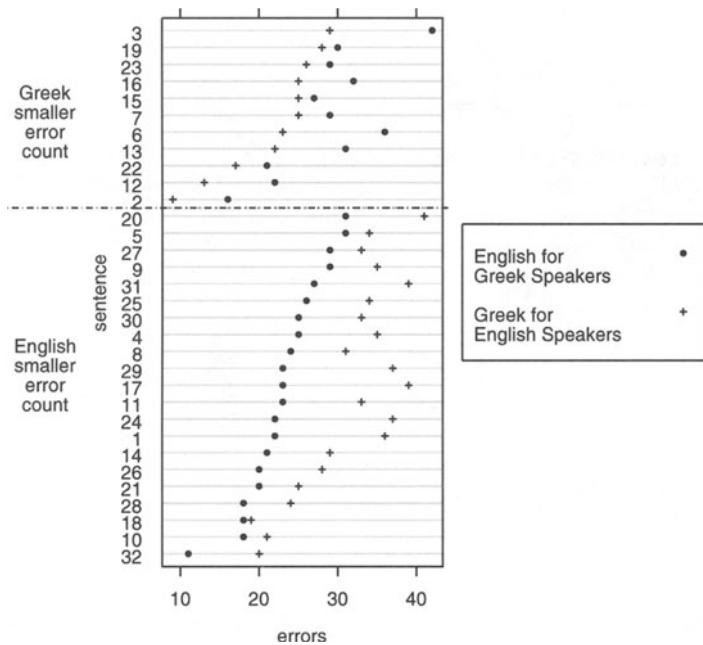


FIGURE 5.2. Dotplot of language difficulty scores. The difficulty in learning each of 32 sentences written in English for Greek speakers (marked English) and written in Greek for English speakers (marked Greek) is noted. The dashed line separates sentences in which the English version showed fewer errors (bottom) from those in which the Greek version showed fewer errors.

(*iinf/code/teachers.s*), (*iinf/code/teachers2.s*),  
 (*iinf/figure/teachers-dot.eps.gz*)

| S-PLUS   | S-PLUS   |
|--|--|
| (iinf/transcript/teachers-stem-a.st):  | (iinf/transcript/teachers-stem-b.st):                                |
| > stem(teachers.diff, nl=5, scale=-1)  | > stem(sqrt(teachers.diff + 17))                                     |
| N = 32 Median = -5.5   | N = 32 Median = 3.390363   |
| Quartiles = -9.5, 3.5  | Quartiles = 2.73709, 4.527356  |
| -1 : 65<br>-1 : 442000<br>-0 : 988887665<br>-0 : 4331<br>0 : 22344<br>0 : 7799<br>1 : 33 | 1 : 0477<br>2 : 26668<br>3 : 00002335677<br>4 : 04456699<br>5 : 1155 |
| a. Original Scale  | b. Transformed Scale   |

FIGURE 5.3. Stem-and-leaf display of vocabulary scores in original and transformed scales.  
(iinf/code/teachers.s), (iinf/transcript/teachers.st)

## 5.6 Sample Size Determination

Deciding on an appropriate sample size is a fundamental aspect of experimental design. In this section we provide discussions of the minimum required sample size for some situations of inference about population means:

- A confidence interval on  $\mu$  with specified width  $W$  and confidence coefficient  $100(1 - \alpha)\%$ .
- A test about  $\mu$  having specified Type I error  $\alpha$ , and power  $1 - \beta$  at a specified distance,  $\delta$ , from the null hypothesized parameter.

These are key design objectives for many experiments with modest inferential goals. Specialized software exists for the purpose of determining sample sizes in a vast array of inferential situations. But our discussion here is limited to a few commonly encountered situations for which the formulas are sometimes mentioned in elementary statistics texts.

We assume throughout this discussion that the sample size will be large enough to guarantee that the standardized test statistic is approximately normally distributed. If as is usual, a sample size calculation does not yield an integer, it is conservative to take  $n$  as the next-higher integer. The sample size formulas here are all the result of explicitly solving a certain equation for  $n$ . In situations not discussed here, an explicit solution for  $n$  may not exist, and the software may produce an iterative solution for  $n$ .

### 5.6.1 Sample Size for Estimation

Since the width of a confidence interval can be expressed as a function of the sample size, the solution of the problem of sample size for a confidence interval is straightforward in the case of a single sample.

For a CI on a single mean, assuming a known population variance  $\sigma^2$ ,

$$n = \frac{4\sigma^2(\Phi^{-1}(1 - \frac{\alpha}{2}))^2}{W^2}$$

where  $\Phi^{-1}$  is the inverse cumulative distribution of a standard normal distribution defined in Section D.1. If  $\sigma^2$  is unknown, a reasonable guess may be made in its place. (Note that the sample variance is not known prior to selecting the sample.) If we are unable to make a reasonable guess, an ad hoc strategy would be to take a small pilot sample of  $n_0$  items and replace  $\sigma$  in the formula with the standard deviation of the pilot sample. Then if the calculation results in a recommended  $n$  greater than  $n_0$ , one samples  $n - n_0$  additional items.

The required sample size for the Agresti and Caffo CI on a single proportion, Equation (5.10), is

$$n = \frac{\left(\Phi^{-1}(1 - \frac{\alpha}{2})\right)^2}{W^2} - 4 \quad (5.14)$$

This formula is based on the normal approximation to the binomial distribution. Many statistics texts contain charts for finding the required sample size based on the exact binomial distribution.

### 5.6.2 Sample Size for Hypothesis Testing

For hypothesis testing we are interested in controlling the specified Type II error probability  $\beta$  when the unknown parameter being tested is a distance  $\delta$  from the null hypothesized value. For a one-tailed test on the mean of a population with known variance  $\sigma^2$ , use

$$n = \sigma^2 \left( \Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta) \right)^2 / \delta^2$$

For a two-tailed test, use

$$n = \sigma^2 \left( \Phi^{-1}(1 - \frac{\alpha}{2}) + \Phi^{-1}(1 - \beta) \right)^2 / \delta^2$$

For testing the equality of the means of two populations with a common variance, with  $\delta$  now equal to the mean difference under the alternative hypothesis, use

$$n = 2 \sigma^2 \left( \Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta) \right)^2 / \delta^2$$

for the one-tailed test, and

$$n = 2 \sigma^2 \left( \Phi^{-1}(1 - \frac{\alpha}{2}) + \Phi^{-1}(1 - \beta) \right)^2 / \delta^2$$

for the two-tailed test.

Lastly, consider attempting to detect a difference between a proportion  $p_1$  and a proportion  $p_2$ . The required common sample size for the one-tailed test is

$$n = \frac{(p_1(1 - p_1) + p_2(1 - p_2)) \left( \Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta) \right)^2}{(p_1 - p_2)^2} \quad (5.15)$$

From the preceding pattern, you should be able to deduce the modification for the two-tailed test (see Exercise 5.16).

## 5.7 Goodness of Fit

Goodness-of-fit tests are used to assess whether a dataset is consistent with having been sampled from a designated hypothesized distribution. In

this section we discuss two general goodness-of-fit tests, the Chi-Square Goodness-of-Fit Test and the Kolmogorov–Smirnov Goodness-of-Fit Test. For testing goodness of fit to specific distributions, there may be better (more powerful) specialized tests than these. For example, the Shapiro–Wilk test of normality (in PROC UNIVARIATE) is more powerful than either general test.

Since many statistics procedures assume an underlying normal distribution, a test of goodness of fit to normal, either before or after transformation, is frequently performed. Occasionally, analysts need to check for fit to other distributions. For example, it often is the case that the distribution of a test statistic is known asymptotically (i.e., if the sample is “large”), but not if the sample is of modest size. It is therefore of interest to investigate how large a sample is needed for the asymptotic distribution to be an adequate approximation. This requires a series of goodness-of-fit tests to the asymptotic distribution. In Chapter 15, we will learn in our discussion of the analysis of contingency table data that the distribution of  $\chi^2 = \sum \frac{(O-E)^2}{E}$  is approximately chi-square provided that no cell sizes are too small. A determination of the ground rule for “too small” required tests of goodness of fit to chi-square distributions with appropriate degrees of freedom.

This class of tests assesses whether a sample may be assumed to be taken from a null hypothesized distribution.

### 5.7.1 Chi-square Goodness-of-Fit Test

The chi-square distribution may be used to conduct goodness-of-fit tests, i.e., ones of the form

$H_0$ : the data are from a [specified population]

vs

$H_1$ : the data are from some other population

For certain specific populations, including normal ones, other specialized tests are more powerful.

The test begins by partitioning the population into  $k$  classes or categories. For a discrete population the categories are the possible values; for a continuous population the choice of a decomposition is rather arbitrary, and the ultimate conclusion may well depend on the selected size of  $k$  and the selected partition.

The test statistic is the same as that used for contingency tables. For each category, calculate from the probability distribution the theoretical or expected frequency  $E$ . If over all  $k$  categories, there is a substantial discrepancy between the  $k$  observed frequencies  $O$  and the  $k$   $E$ 's, then  $H_0$  is

rejected. The measure of discrepancy is the test statistic  $\chi^2 = \sum \frac{(O-E)^2}{E}$ . A “large” value of  $\chi^2$  is evidence against  $H_0$ . If the total sample size,  $n = \sum O = \sum E$ , is sufficiently “large”,  $\chi^2$  is approximately chi-square distributed and the  $p$ -value is approximately the chi-square tail probability associated with  $\chi^2$  with  $k - 1$  degrees of freedom.

For adequacy of the chi-square approximation it is suggested that all expected frequencies be at least 5. If this is not the case, the analyst may consider combining adjacent categories after which this condition is met. Then  $k$  represents the number of categories following such combining.

Sometimes, the statement of the null hypothesis is so vague that calculation of expected frequencies requires that some parameters be estimated from the data. In such instances, the df is further reduced by the number of such parameters estimated. This possibility is illustrated in Example 5.7.3.

### 5.7.2 Example—Test of Goodness-of-Fit to a Discrete Uniform Distribution

A six-sided die (singular of the word *dice*) is rolled 30 times with the following outcomes: 1, 3 times; 2, 7 times; 3, 5 times; 4, 8 times; 5, 1 time; and 6, 6 times. Test whether the die is fair.

A fair die is one that has a discrete uniform distribution on 1, 2, 3, 4, 5, 6. Each of these six possibilities has  $\frac{1}{6}$  chance of occurring, and all six  $E$ 's are  $30(\frac{1}{6}) = 5$ . Then

$$\chi^2 = \frac{(3-5)^2}{5} + \dots + \frac{(6-5)^2}{5} = 6.8$$

and the  $p$ -value from  $\chi^2_5$  is 0.236. Hence these 30 observations do not provide evidence to refute the fairness of the die.

### 5.7.3 Example—Test of Goodness-of-Fit to a Binomial Distribution

In a certain community, there were 80 families containing exactly five children. It was noticed that there was an excess of boys among these. It was desired to test whether  $Y = \text{“number of girls in family”}$  is a binomial r.v. with  $n = 5$  and  $p = .4$ . The expected frequencies calculated from this binomial distribution are shown in Table 5.4 along with the observed frequencies and the calculated  $\chi^2_5$  statistic. Then the  $p$ -value is,  $8.510^{-6}$ , calculated as the tail probability at 31.215 for a chi-square distribution with 5 df. We conclude that the sample data contain more dispersion than does binomial(5, .4).

TABLE 5.4. Observed and expected frequencies for the goodness-of-fit example in Section 5.7.3.

| <i>Y</i> | <i>O</i> | <i>E</i> | $\frac{(O - E)^2}{E}$ |
|----------|----------|----------|-----------------------|
| 0        | 13       | 6.221    | 7.388                 |
| 1        | 18       | 20.736   | 0.361                 |
| 2        | 20       | 27.648   | 2.116                 |
| 3        | 18       | 18.432   | 0.010                 |
| 4        | 6        | 6.144    | 0.003                 |
| 5        | 5        | 0.819    | 21.337                |
|          |          |          | 31.215                |

TABLE 5.5. Calculation of *p*-value for chi-square test with known *p* and with estimated  $\hat{p}$ .  
(iinf/transcript/ex.chisq.st)

```
S-PLUS (iinf/transcript/ex.chisq.st):
> o <- c(13, 18, 20, 18, 6, 5)
> n <- 5
> y <- 0:n

> # if we know p=.4
> p1 <- .4
> e1 <- sum(o)*dbinom(y, n, p1)
> chisq1 <- sum((o-e1)^2/e1)
> chisq1
[1] 31.21459
> 1-pchisq(chisq1, n)
[1] 8.496273e-06

> # if we estimate p from the data
> p2 <- sum(o*y)/(n*sum(o))
> e2 <- sum(o)*dbinom(y, n, p2)
> chisq2 <- sum((o-e2)^2/e2)
> chisq2
[1] 30.71593
> 1-pchisq(chisq2, n-1)
[1] 3.498248e-06
```

In this example, the value of the binomial proportion parameter, *p*, was specified. If instead it had to be estimated, the df would decrease from 5 to 4. We illustrate the calculation of both tests in S-PLUS in Table 5.5.

## 5.8 Normal Probability Plots and Quantile Plots

Quantile plots (Q-Q plots) are visual diagnostics used to assess whether (a) a dataset may reasonably be treated as if it were a sample from a designated probability distribution, or (b) whether two datasets show evidence of coming from a common unspecified distribution.

The normal probability plot, an important special case of the more general quantile plot, is used to assess whether data are consistent with a normal distribution. The normal probability plot is a standard diagnostic plot in regression analysis (Chapters 8–11) used to check the assumption of normally distributed residuals. This condition is required for the validity of many of the usual inferences in a regression analysis. If the normality assumption appears to be violated, it is often possible to retain a simple analysis by transforming the data scale, for example by a power transformation, and then reanalyzing and replotting to see if the residuals from the transformed data are close to normal. The choice of transformation may be guided by the interpretation of the normal probability plot.

In S, a normal probability plot is produced with the `qqnorm()` function. Normal probability plots are included in the default plots for the results of linear model analyses.

A quantile plot to assess consistency of observed data  $y_i$  with a designated distribution is easily constructed. We sort the observed data to get  $y_{[i]}$ , find the quantiles of the distribution by looking up the fractions  $(i - \frac{1}{2})/n$  in the inverse cumulative distribution function to get  $q_i = F^{-1}((i - \frac{1}{2})/n)$ , and then plotting the sorted data  $y_{[i]}$  against the quantiles  $q_i$ . Consistency is suggested if the points tend to fall along a straight line. A pattern of a departure from a straight-line quantile plot usually suggests the nature of the departure from the assumed distribution. Both S-PLUS and SAS one-sample quantile plots default to the usual convention of plotting the data against the theoretical values. Other software and a number of references reverse the axes. Readers of presentations containing quantile plots should be alert to which convention is used, and writers must be sure to label the axes to indicate the convention, because the choice matters considerably for interpretation of departures from compatibility.

A general Q-Q (or quantile-quantile) plot is invoked in S-PLUS with the command `qqplot(x, y, plot=T)`, whereby the quantiles of two samples,  $x$  and  $y$ , are compared. As with a normal probability case, the straightness of the Q-Q plot indicates the degree of agreement of the distributions of  $x$  and  $y$ , and departure from a well-fitting straight line on an end of the plot indicates the presence of outlier(s). Quoting from S-PLUS online help for `qqplot`:

A Q-Q plot with a “U” shape means that one distribution is skewed relative to the other. An “S” shape implies that one distribution has longer tails than the other. In the default configuration (data on the  $y$ -axis) a plot from `qqnorm` that is bent down on the left and bent up on the right means that the data have longer tails than the Gaussian [normal].

For a normal probability plot with default configuration, a plot that is bent up on the left and bent down on the right indicates that the data have shorter tails than the normal. A curved plot that opens upward suggests positive skewness and curvature opening downward suggests negative skewness.

It is possible to construct a Q-Q plot comparing a sample with any designated distribution, not just the normal distribution. In S-PLUS this is accomplished with the function `ppoints(y)`, which returns a vector of  $n=\text{length}(y)$  fractions uniformly spaced between 0 and 1 which will be used as input to the quantile (inverse cumulative distribution) function. For example, the statement `plot(sort(y) ~ qlnorm(ppoints(y)))` produces a lognormal Q-Q plot of the data in  $y$ .

If it is unclear from a normal probability plot whether the data are in fact normal, the issue may be further addressed by a specialized goodness-of-fit test to the normal distribution, the Shapiro–Wilk test. This test works by comparing

$S(y) =$  the empirical distribution function of the data, the fraction of the data that is less than or equal to  $y$

with

$\Phi((y - \bar{y})/s) =$  the probability that a normal r.v.  $Y$  (with mean  $\bar{y}$  and s.d.  $s$ ) is less than or equal to  $y$

Over the observed sample,  $S(y)$  and  $\Phi((y - \bar{y})/s)$  should be highly correlated if the data are normal, but not otherwise. The Shapiro–Wilk statistic  $W$  is closely related to the square of this correlation. If the normal probability plot is nearly a straight line,  $W$  will be close to 1. A small value of  $W$  is evidence of nonnormality. The Shapiro–Wilk test is available in SAS `PROC UNIVARIATE` if we request a test for goodness of fit to the normal distribution with the `NORMAL` option. It is available in S-PLUS with the `shapiro.test` function. For this specific purpose the Shapiro–Wilk test is more powerful than a general goodness-of-fit test such as the Kolmogorov–Smirnov procedure discussed in Section 5.9.

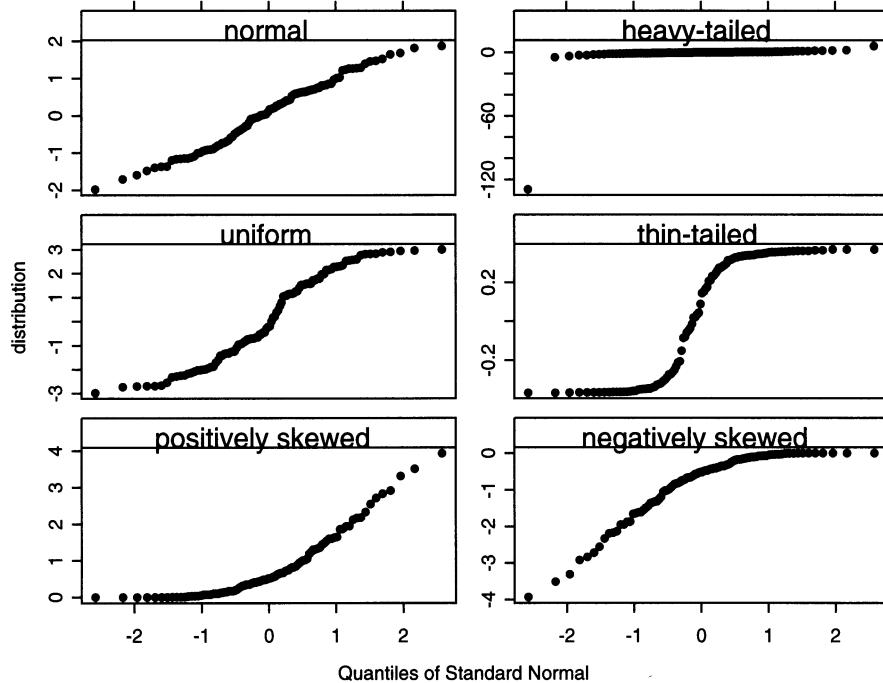


FIGURE 5.4. Normal probability plots.  
`(iinf/code/ex.qqnorm.s)`, `(iinf/figure/iinf.f.qqnorm.ps.gz)`

### 5.8.1 Normal Probability Plots

Figure 5.4 contrasts the appearance of normal probability plots for the normal distribution and various departures from normality. Typically, the plot has these appearances:

- An “S” shape for distributions with thinner tails than the normal.
- An inverted “S” shape for distribution with heavier tails than the normal.
- A “J” shape for positively skewed distributions.
- An inverted “J” shape for negatively skewed distributions.
- Isolated points at the extremes of a plot for distributions having outliers.

# One-Way Analysis of Variance

In Chapter 5 we consider ways to compare the means of two populations. Now we extend these procedures to comparisons of means from several populations. For example, we may wish to compare the average hourly production of a company's six factories. We say that the investigation has a *factor* `factory` that has six *levels*, namely the six identifiers distinguishing the factories from one another. Or we may wish to compare the yields per acre of five different varieties of wheat. Here, the factor is `wheat`, and the levels of `wheat` are `variety1` through `variety5`. This chapter discusses investigations having a single factor. Experiments having two factors are discussed in Chapter 12, while situations with two or more factors are discussed in Chapters 13 and 14.

One-way analysis of variance (ANOVA) is the natural generalization of the two-sample *t*-test to more than two groups. Suppose that we have a factor *A* with *a* levels. We select independent samples from each of these *a* populations, where  $n_i$  is the size of the sample from population *i*. We distinguish between two possible assumptions about these populations comprising the single factor. We discuss *fixed effects* beginning in Section 6.1 and *random effects* beginning in Section 6.4.

## 6.1 Example—Catalyst Data

With the catalyst data from (Montgomery, 1997) we are interested in comparing the concentrations of one component of a liquid mixture in the presence of each of four catalysts. We investigate whether the cata-

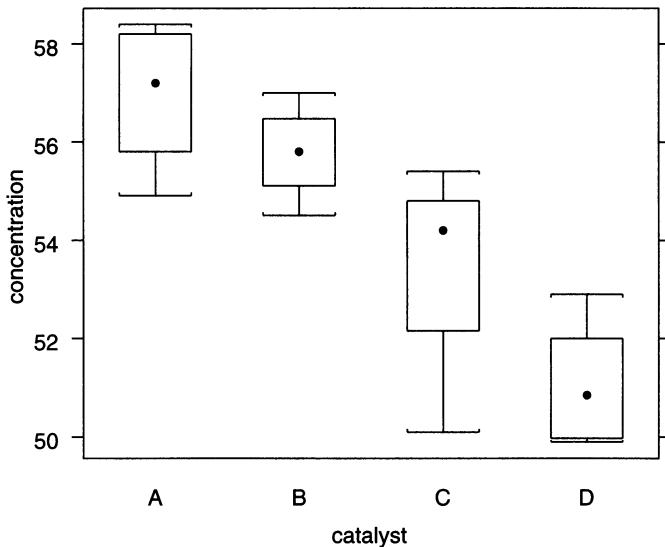


FIGURE 6.1. Boxplots Comparing the Concentrations for each Catalyst  
 (oway/code/catalystm.s), (oway/figure/catalystm1.eps.gz)

lists provide for equal mean concentrations, and then since this does not appear to be true, we study the extent of differences among the mean concentrations. We read the data from file (datasets/catalystm.dat) with file (oway/code/catalystm.s) or with file (oway/code/catalystm1.sas) [which calls(oway/code/catalystm.read.sas)] and plot it in Figure 6.1. We see that group D does not overlap groups A and B and that group C has a wider spread than the others.

The ANOVA (analysis of variance) table and the table of means are in Tables 6.1 (calculated with S-PLUS) and 6.2 (calculated with SAS). The  $F$ -test in the ANOVA table addresses the null hypothesis that the four catalysts have equal mean concentrations. We see immediately, from the small  $p$ -value ( $p = .0014$ ), that these four catalysts do not provide the same average concentrations.

TABLE 6.1. ANOVA Table for Catalyst Data (S-PLUS)  
(oway/code/catalystm.s)

---

```
S-PLUS (oway/transcript/catalystm.aov1.st):
> catalystm1.aov <- aov(concent ~ catalyst, data=catalystm)
> anova(catalystm1.aov)

Analysis of Variance Table

Response: concent

Terms added sequentially (first to last)
  Df Sum of Sq  Mean Sq   F Value    Pr(F)
catalyst    3  85.67583 28.55861 9.915706 0.001435628
Residuals 12  34.56167  2.88014

> model.tables(catalystm1.aov, "means")

Tables of means
Grand mean

54.487

catalyst
      A       B       C       D
  56.900 55.775 53.233 51.125
rep  5.000  4.000  3.000  4.000
```

---

TABLE 6.2. ANOVA Table for Catalyst Data (SAS)

SAS (oway/code/catalystm.aov1.sas):

```
proc anova data=catalystm;
  class cat;
  model concent = cat;
  means cat;
run;
```

SAS (oway/transcript/catalystm.aov1.lst):

The ANOVA Procedure

Dependent Variable: concent

| Source          | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|-----------------|----|----------------|-------------|---------|--------|
| Model           | 3  | 85.6758333     | 28.5586111  | 9.92    | 0.0014 |
| Error           | 12 | 34.5616667     | 2.8801389   |         |        |
| Corrected Total | 15 | 120.2375000    |             |         |        |

| R-Square | Coeff Var | Root MSE | concent Mean |
|----------|-----------|----------|--------------|
| 0.712555 | 3.114654  | 1.697097 | 54.48750     |

| Source | DF | Anova SS   | Mean Square | F Value | Pr > F |
|--------|----|------------|-------------|---------|--------|
| cat    | 3  | 85.6758333 | 28.5586111  | 9.92    | 0.0014 |

| Level of<br>cat | -----concent----- |            |            |
|-----------------|-------------------|------------|------------|
|                 | N                 | Mean       | Std Dev    |
| A               | 5                 | 56.9000000 | 1.51986842 |
| B               | 4                 | 55.7750000 | 1.09962115 |
| C               | 3                 | 53.2333333 | 2.77908858 |
| D               | 4                 | 51.1250000 | 1.44308697 |

## 6.2 Fixed Effects

Initially we assume that the  $a$  stated levels of  $A$  are the totality of all levels of interest to us. We call  $A$  a *fixed factor*. We model the  $j^{\text{th}}$  observation from population  $i$  as

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad \text{for } i = 1, \dots, a \quad \text{and } j = 1, \dots, n_i \quad (6.1)$$

where  $\mu$  and the  $\alpha_i$  are fixed quantities with the constraint

$$\sum_i \alpha_i = 0 \quad (6.2)$$

and the  $\epsilon_{ij}$  are assumed to be normally and independently distributed (NID) with common mean 0 and common variance  $\sigma^2$ , which we denote by

$$\epsilon_{ij} \sim \text{NID}(0, \sigma^2) \quad (6.3)$$

We interpret  $\mu$  as the grand mean of all  $a$  populations,  $\alpha_i$  as the deviation of the mean of population  $i$  from  $\mu$ , and assume that the responses from all  $a$  populations have a normal distribution with a common variance. If the normality assumption is more than mildly violated, we must either transform the response variable to one for which this assumption is satisfied, perhaps with a power transformation such as those discussed in Section 4.7, or use a nonparametric procedure as described in Chapter 16. The common variance assumption may be examined with the hypothesis test described in Section 6.10. If the variances are not homogeneous, a transformation such as those discussed in Section 4.7 sometimes can fix the inhomogeneity of variance problem as well as the nonnormality problem by changing to a scale in which the transformed observations show homogeneity of variance.

We discuss in Appendix 6.A.1 the correspondence between the notation of Equation (6.1) and the software notations in Tables 6.1 and 6.2.

The initial question of interest is the equality of the  $a$  population means, which we investigate with the test of

$$\begin{aligned} H_0: \alpha_1 &= \alpha_2 = \dots = \alpha_a \\ \text{vs} \\ H_a: \text{the } \alpha_i &\text{ are not all equal.} \end{aligned} \quad (6.4)$$

When  $a = 2$ , the test is the familiar

$$t_{n_1+n_2-2} = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$$s_p^2 = ((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2) / (n_1 + n_2 - 2)$$

from Equations (5.11) and (5.10). By squaring both sides, we can show

$$F_{1,n_1+n_2-2} = t_{n_1+n_2-2}^2 = \frac{n_1(\bar{y}_1 - \bar{\bar{y}})^2 + n_2(\bar{y}_2 - \bar{\bar{y}})^2}{s_p^2} \quad (6.5)$$

where

$$\bar{\bar{y}} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}$$

In the special case where  $n_1 = n_2$ , Equation (6.5) is easily proved by using these hints:

1.  $\bar{\bar{y}} = \frac{\bar{y}_1 + \bar{y}_2}{2}$
2.  $(\bar{y}_1 - \bar{\bar{y}}) = -(\bar{y}_2 - \bar{\bar{y}})$
3.  $\frac{1}{n_1} + \frac{1}{n_2} = \frac{2}{n_1}$

The equality (6.5) is also true for unequal  $n_i$ , but the proof is messier.

When  $a \geq 2$ , we generalize formula (6.5) to

$$F_{a-1, (\Sigma n_i)-a} = \frac{(\sum n_i(\bar{y}_i - \bar{\bar{y}})^2)/(a-1)}{s_p^2} \quad (6.6)$$

where  $\bar{\bar{y}}$  and  $s_p^2$  are the weighted mean

$$\bar{\bar{y}} = \frac{\sum n_i \bar{y}_i}{\sum n_i} \quad (6.7)$$

and pooled variance

$$s^2 = s_p^2 = \frac{\sum (n_i - 1)s_i^2}{\sum (n_i - 1)} = \text{MS}_{\text{residual}} \quad (6.8)$$

over all  $a$  samples.

The usual display of this formula is in the analysis of variance table and the notation is

$$F_{(a-1), (\Sigma n_i)-a} = \frac{\text{SS}_{\text{treatment}}/\text{df}_{\text{treatment}}}{\text{SS}_{\text{residual}}/\text{df}_{\text{residual}}} = \frac{\text{MS}_{\text{treatment}}}{\text{MS}_{\text{residual}}} = \frac{\text{MS}_{\text{Tr}}}{\text{MS}_{\text{Res}}} \quad (6.9)$$

The sample ANOVA table in Table 6.3 illustrates the structure.

As in Section 5.4.4, this  $F$ -test of the pair of hypotheses in Equation (6.4) compares two estimates of the population variance  $\sigma^2$ .  $\text{MS}_{\text{Res}}$  is an unbiased estimator of  $\sigma^2$  whether or not  $H_0$  is true.  $\text{MS}_{\text{Tr}}$  is unbiased for  $\sigma^2$  when  $H_0$  is true but an overestimate of  $\sigma^2$  when  $H_a$  is true. Hence, the larger the variance ratio  $F = \text{MS}_{\text{Tr}}/\text{MS}_{\text{Res}}$ , the stronger the evidence in support

TABLE 6.3. Sample Table to Illustrate Structure of the ANOVA Table

| Analysis of Variance of Dependent Variable $y$ |                    |                |             |          |          |
|--|--------------------|----------------|-------------|----------|----------|
| Source   | Degrees of Freedom | Sum of Squares | Mean Square | F        | p-value  |
| Treatment                                      | $df_{Tr}$          | $SS_{Tr}$      | $MS_{Tr}$   | $F_{Tr}$ | $p_{Tr}$ |
| Residual                                       | $df_{Res}$         | $SS_{Res}$     | $MS_{Res}$  |          |          |
| Total  | $df_{Total}$       | $SS_{Total}$   |             |          |          |

The terms of the table are defined by

| Treatment |  |
|-----------|--|
| $df_{Tr}$ | $a - 1$  |
| $SS_{Tr}$ | $\sum_{i=1}^a n_i (\bar{y}_i - \bar{\bar{y}})^2$ |
| $MS_{Tr}$ | $SS_{Tr}/df_{Tr}$                                |
| $F_{Tr}$  | $MS_{Tr}/MS_{Res}$                               |
| $p_{Tr}$  | $1 - \mathcal{F}_F(F_{Tr}   df_{Tr}, df_{Res})$  |

| Residual   |  |
|------------|--|
| $df_{Res}$ | $\left( \sum_{i=1}^a n_i \right) - a$                  |
| $SS_{Res}$ | $\sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ |
| $MS_{Res}$ | $SS_{Res}/df_{Res}$                                    |

| Total        |   |
|--------------|---|
| $df_{Total}$ | $\left( \sum_{i=1}^a n_i \right) - 1 = df_{Tr} + df_{Res}$                |
| $SS_{Total}$ | $\sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = SS_{Tr} + SS_{Res}$ |

of  $H_a$ . Comparing two variances facilitates the comparison of  $a$  means. For this reason, the foregoing procedure is called *analysis of variance*. It involves decomposing the total sum of squares  $SS_{\text{Total}}$  into the variances used to conduct this  $F$ -test. The  $p$ -value in this table is calculated as the probability that a central  $F$  random variable with  $df_{\text{Tr}}$  and  $df_{\text{Res}}$  degrees of freedom exceeds the calculated  $F_{\text{Tr}}$ .

### 6.3 Multiple Comparisons—Tukey Procedure for Comparing All Pairs of Means

Multiple comparisons refers to procedures for simultaneously conducting all inferences in a family of related inferences, while keeping control of a Type I error concept that relates to the entire family. This class of inferential procedures is discussed in detail in Chapter 7. In the present chapter, we introduce the Tukey procedure, used for the family of all  $\binom{a}{2}$  pairwise comparisons involving  $a$  population means.

We illustrate the Tukey procedure with a continuation of the analysis of the catalyst data. We seek to determine which catalyst mean differences are responsible for the overall conclusion that the catalyst means are not identical.

Under the assumption that **catalyst** is a fixed factor, we investigate the nature of the differences among the four catalysts. There are  $\binom{4}{2} = 6$  pairs of catalysts, and for each of these pairs we wish to determine whether there is a significant difference between the concentrations associated with the two catalysts comprising the pair. (If the levels of **catalyst** had instead been quantitative or bore a structural relationship to one another, a different follow-up to the analysis of variance table would have been more appropriate. An example of such a situation is the analysis of the turkey data presented in Section 6.8.)

We seek to control at a designated level  $\alpha$  the familywise error rate, FWE, defined as the probability of incorrectly rejecting at least one true null hypothesis under any configuration of true and false null hypotheses. For the family consisting of all pairs of means, the Tukey procedure maximizes, in various senses, the probability of detecting truly differing pairs of means while controlling the FWE at  $\alpha$ .

The Tukey procedure uses a critical value  $q_\alpha$  from the Studentized range distribution (see Section D.1), i.e., the distribution of standardized difference between the maximum sample mean and the minimum sample mean, rather than an ordinary  $t$  distribution for comparing two means discussed in Section 5.4.3. The Tukey output may be presented in the form of simul-

taneous confidence intervals on each of the mean differences rather than, or in addition to, tests on each difference. The interpretation is that the confidence coefficient  $1 - \alpha$  is the probability that all of the  $\binom{a}{2}$  pairwise confidence intervals among  $a$  sample means contain their respective true values of the difference between the two population means:

$$\begin{aligned} 1 - \alpha &\leq P(\text{CI}_{12} \cap \text{CI}_{13} \cap \dots \cap \text{CI}_{1a} \\ &\quad \cap \text{CI}_{23} \cap \dots \cap \text{CI}_{2a} \\ &\quad \cap \dots \cap \text{CI}_{(a-1)a}) \end{aligned} \quad (6.10)$$

where

$$\text{CI}_{ii'}: (\bar{y}_i - \bar{y}_{i'}) - \frac{q_\alpha}{\sqrt{2}} s_{(\bar{y}_i - \bar{y}_{i'})} \leq (\mu_i - \mu_{i'}) \leq (\bar{y}_i - \bar{y}_{i'}) + \frac{q_\alpha}{\sqrt{2}} s_{(\bar{y}_i - \bar{y}_{i'})} \quad (6.11)$$

and

$$s_{(\bar{y}_i - \bar{y}_{i'})} = s \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}$$

$$s = \sqrt{\text{MS}_{\text{residual}}}$$

If the sample sizes are unequal, the confidence intervals (6.11) are conservative in the sense that the coverage probability in Equation (6.10) exceeds  $1 - \alpha$ . If the sample sizes are equal, the inequality in Equation (6.10) is instead an equality and the simultaneous  $1 - \alpha$  confidence for the set of intervals in (6.11) is exact.

We show both S-PLUS and SAS listings for the Tukey test of the catalyst data in Tables 6.4 and 6.5, and the standard S-PLUS multiple comparisons plot in Figure 6.2. A new display, the *MMC* plot (the *mean-mean multiple comparisons* display discussed in Section 7.2), is in Figure 6.3. Denoting the mean concentration associated with catalyst  $i$  as  $\mu_i$ , since the confidence intervals on  $\mu_A - \mu_D$  and  $\mu_B - \mu_D$  lie entirely above 0 while all other confidence intervals include 0, we conclude that both catalysts A and B provide, on average, a significantly greater concentration than catalyst D; no other significant differences between catalysts were uncovered.

In view of this finding one might be tempted to focus on the differences demonstrated to be significant in Tables 6.4 and 6.5, and construct hypothesis tests or confidence intervals using a method from Section 5.4.2. A more general framing of this temptation is to ask, “Is it permissible to use preliminary examinations of the data to develop subsequent hypotheses about the data” (a practice referred to as *data snooping*)? With few exceptions, the answer is *no* because the two-stage nature of the procedure distorts the claimed significance levels or confidence coefficients of the analyses in the second stage. Inferential strategies should be developed before

TABLE 6.4. Tukey Multiple Comparisons for Catalyst Data  
 (oway/code/catalystm.s)

---

```

S-PLUS (oway/transcript/catalystm.aov2.st):
> catalystm.mca <- multicomp(catalystm1.aov, focus="catalyst")
> plot(catalystm.mca)
> catalystm.mca

95 % simultaneous confidence intervals for specified
linear combinations, by the Tukey method

critical point: 2.9691
response variable: catalyst

intervals excluding 0 are flagged by '****'

      Estimate Std.Error Lower Bound Upper Bound
A-B      1.12     1.14    -2.2600     4.51
A-C      3.67     1.24    -0.0132     7.35
A-D      5.78     1.14     2.3900     9.16 ****
B-C      2.54     1.30    -1.3100     6.39
B-D      4.65     1.20     1.0900     8.21 ****
C-D      2.11     1.30    -1.7400     5.96
  
```

---

the data are collected—based entirely on the structure of the data and the sampling method used. Strategies should not depend on the observed data. Here one should be content with the analyses in Tables 6.4 and 6.5 and supporting graphical displays such as Figure 6.2 or Figure 6.3, assuming the correctness of the assumptions underlying their construction.

Although in this example there were equal sample sizes from the levels of catalyst, neither the basic analysis of variance nor the Tukey multiple comparison procedure requires that the factor levels have the same sample size. Analyses of one-way data having unequal sample sizes are requested in the Exercises.

TABLE 6.5. Tukey Multiple Comparisons for Catalyst Data

SAS (oway/code/catalystm.aov2.sas):

```
proc anova data=catalystm;
  class cat;
  model concent = cat;
  means cat / tukey;
run;
```

SAS (oway/transcript/catalystm.aov2.lst):

The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for concent

NOTE: This test controls the Type I experimentwise error rate.

|                                     |          |
|-------------------------------------|----------|
| Alpha                               | 0.05     |
| Error Degrees of Freedom            | 12       |
| Error Mean Square                   | 2.880139 |
| Critical Value of Studentized Range | 4.19852  |

Comparisons significant at the 0.05 level are indicated by \*\*\*.

| cat<br>Comparison |  | Difference       | Simultaneous             |        |     |
|-------------------|--|------------------|--------------------------|--------|-----|
|                   |  | Between<br>Means | 95% Confidence<br>Limits |        |     |
| A - B             |  | 1.125            | -2.255                   | 4.505  |     |
| A - C             |  | 3.667            | -0.013                   | 7.346  |     |
| A - D             |  | 5.775            | 2.395                    | 9.155  | *** |
| B - A             |  | -1.125           | -4.505                   | 2.255  |     |
| B - C             |  | 2.542            | -1.306                   | 6.390  |     |
| B - D             |  | 4.650            | 1.087                    | 8.213  | *** |
| C - A             |  | -3.667           | -7.346                   | 0.013  |     |
| C - B             |  | -2.542           | -6.390                   | 1.306  |     |
| C - D             |  | 2.108            | -1.740                   | 5.956  |     |
| D - A             |  | -5.775           | -9.155                   | -2.395 | *** |
| D - B             |  | -4.650           | -8.213                   | -1.087 | *** |
| D - C             |  | -2.108           | -5.956                   | 1.740  |     |

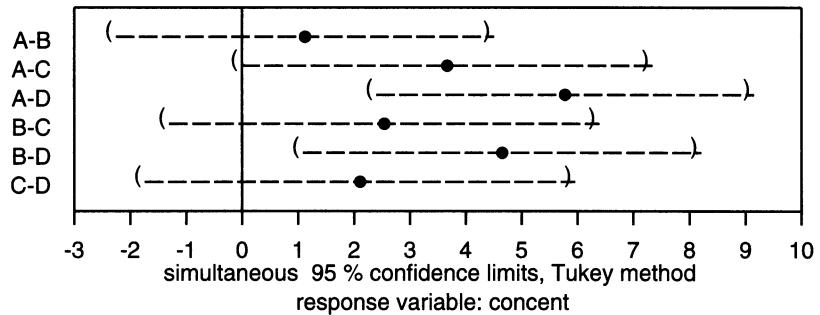


FIGURE 6.2. Tukey Multiple Comparisons of Catalyst Means  
 (oway/code/catalystm.s), (oway/code/catalystm.aov2.sas),  
 (oway/figure/catalystm2.eps.gz)

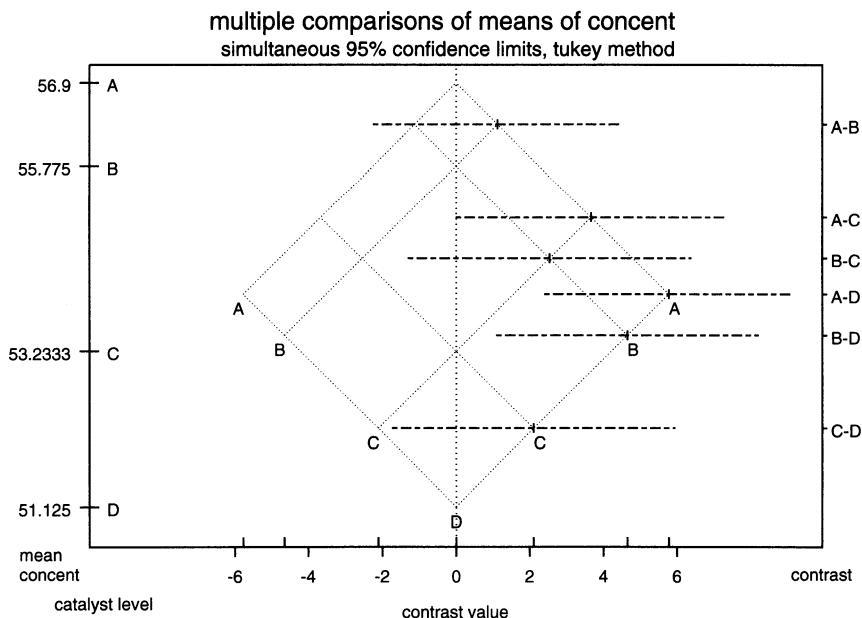


FIGURE 6.3. Tukey Multiple Comparisons of All Pairwise Comparisons of Catalyst Means with the MMC Display  
 (mcomp/code/catalystm-mmc3.s), (mcomp/figure/catalystm-mmc-mca.eps.gz)

## 6.4 Random Effects

We could assume that the  $a$  observed levels of  $A$  are a random sample from a large or conceptually infinite population of levels. We call  $A$  a *random factor*. For example, in a study to compare the daily productivity of assembly line workers in a large firm, the workers in the study may be a random sample of  $a$  employees from among thousands of employees performing identical tasks.

We still work with Equation (6.1), and still maintain the same assumptions about  $\mu$  and the  $\epsilon_{ij}$ 's. We have a different interpretation of the  $\alpha_i$ . Now the term  $\alpha_i$  in Equation (6.1) is assumed to be a  $N(0, \sigma_A^2)$  random variable and the restriction  $\sum_i \alpha_i = 0$  no longer applies. Instead we work with the hypotheses

$$\begin{aligned} H_0: \sigma_A^2 &= 0 \\ \text{vs} \\ H_a: \sigma_A^2 &> 0 \end{aligned} \tag{6.12}$$

The sample ANOVA table in Table 6.3 still applies. The  $F$  statistic now compares the hypotheses in Equation 6.12. In the context of the worker productivity example, the factor `worker` is referred to as a *random factor*. We are using the  $a$  sampled workers to assess whether the entire population of workers has identical or nonidentical productivity.

## 6.5 Expected Mean Squares (EMS)

To better understand the distinction between the  $F$ -test in the fixed and random factor cases, it is useful to compare the expected mean squares (EMS) for the ANOVA table under the two assumptions. The EMS are algebraically displayed in Table 6.6.

TABLE 6.6. Expected Mean Squares in One-Way Analysis of Variance

| Source      | df                 | $E(\text{MS})$              | EMS, factor A fixed  | EMS, factor A random  |
|-------------|--------------------|-----------------------------|--|---|
| Treatment A | $a - 1$            | $E(\text{MS}_{\text{Tr}})$  | $\sigma^2 + \left(\frac{1}{a-1}\right) \sum_i n_i (\alpha_i - \bar{\alpha})^2$ | $\sigma^2 + \frac{1}{a-1} \left( \sum_i n_i - \frac{\sum_i n_i^2}{\sum_i n_i} \right) \sigma_A^2$ |
| Residual    | $\sum_i (n_i - 1)$ | $E(\text{MS}_{\text{Res}})$ | $\sigma^2$   | $\sigma^2$  |
| Total       | $(\sum_i n_i) - 1$ |                             |  |   |

In the case of factor A fixed, the  $F$  statistic is testing whether  $\sum_i n_i(\alpha_i - \bar{\alpha})^2 = 0$ , where  $\bar{\alpha} = (\sum_i n_i \alpha_i) / (\sum_i n_i)$ . This statement is true if and only if the  $\alpha_i$  are identical. In the case of factor A random, the  $F$  statistic tests whether  $\sigma_A^2 = 0$  because the coefficient of  $\sigma_A^2$  is positive whether or not  $H_0$  is true.

The power of the  $F$ -test is an increasing function of the noncentrality parameter of the  $F$  statistic, which in turn is an increasing function of  $\text{EMS}_{\text{Treatment}}/\text{EMS}_{\text{Residual}}$ . When factor A is random, it follows that the power is an increasing function of  $\sum_i n_i - \sum_i n_i^2 / \sum_i n_i$ . For fixed total sample size  $\sum_i n_i$ , this quantity and hence power is maximized when  $n_i = \sum_i n_i/a$ , that is, when the sample is equally allocated to the levels, or nearly equal allocation if  $\sum_i n_i/a$  is not an integer.

In general in Analysis of Variance tables, examination of expected mean squares suggests the appropriate numerator and denominator mean squares for conducting tests of interest. We look for  $\text{EMS}_{\text{Treatment}}/\text{EMS}_{\text{Residual}}$  that exceeds 1 if and only if the null hypothesis of interest is false. This idea is especially useful in analyzing mixed models (i.e. ones containing both fixed and random factors) as is discussed in Chapter 14.

## 6.6 Example—Catalyst Data—Continued

In Section 6.1 the four levels of the factor `catalyst` were assumed to be qualitative rather than quantitative. It was also assumed that these are the only catalysts of interest. In this situation `catalyst` is a *fixed* factor since the four catalyst levels we study are the only levels of interest.

If instead these four catalysts had been regarded as a random sample from a large population of catalysts, then `catalyst` would have been considered a *random* factor. Figure 6.1 provides a tentative answer to the question of whether the four distributions are homogeneous. This figure also addresses the reasonableness of the assumption that the data come from normal homoskedastic populations, that is, populations having equal variances. The boxplots hint at the possibility that catalyst 3 has a more variable concentration than the others, but the evidence is not substantial in view of the small sample sizes (5,4,3,4). We look more formally at the homogeneity of the variances of these four catalysts in Section 6.10.

The  $F$ -test in Tables 6.2 and 6.1 addresses the null hypothesis that the four catalysts have equal mean concentrations. The small  $p$ -value suggests that these four catalysts provide different average concentrations.

If instead, the factor `catalyst` in this experiment had been a random factor rather than a fixed factor, the  $F$ -test would be addressing the hypothesis

that there is no variability in concentration over the population of catalysts from which these four catalysts are a random sample.

## 6.7 Example—Batch Data

In the batch data (`datasets/batch.dat`) taken from (Montgomery, 1997), the 5 sampled batches constitute a random sample from a large population of batches. Thus `batch` is a random factor, not a fixed factor. The response variable is `calcium` content. The ANOVA is in Table 6.7. The small  $p$ -value, .0036, leads us to conclude that the population of batches, from which these 5 batches were a random sample, had nonhomogeneous calcium content.

Explicitly telling SAS that `batch` is a random factor results in an algebraic expression

$$\text{Var}(\text{Error}) + 5 \text{Var}(\text{Batch})$$

for the expected value of the expected mean square for `Batch`. Equating the expected values of both `Error` and `Batch` to their corresponding calculated mean squares leads to an unbiased estimate of  $\sigma_A^2$ :

$$\begin{aligned}\hat{\sigma}^2 + 5\hat{\sigma}_A^2 &= .024244 \\ \hat{\sigma}^2 &= .004380\end{aligned}$$

which implies

$$\hat{\sigma}_A^2 = .003973$$

## 6.8 Example—Turkey Data

### Study Objectives

The goal in many agricultural experiments is to increase yield. In the Turkey experiment (data from (Ott, 1993)) (`datasets/turkey.dat`) the response is weight gain (in pounds) of turkeys and the treatments are diet supplements.

### Data Description

Six turkeys were randomly assigned to each of 5 diet groups and fed for the same length of time. The diets have a structure such that it is possible and desirable to undertake an *orthogonal contrast analysis*, a systematic set of comparisons among their mean responses. A contrast is a comparison of two or more means such that the expected value of the comparison is zero

TABLE 6.7. Batch ANOVA

See also (oway/code/batch.s), (oway/transcript/batch.st),  
 (oway/figure/batch-data.eps.gz), (oway/figure/batch-hov.eps.gz).

SAS (oway/code/batch.sas):

```
data batch;
  infile "&hh/datasets/batch.dat" firstobs=2;
  input Batch $ Calcium;
run;

proc glm;
  class Batch;
  model Calcium = Batch;
  random Batch;
run;
```

SAS (oway/transcript/batch.lst):  
 The GLM Procedure

Dependent Variable: Calcium

| Source          | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|-----------------|----|----------------|-------------|---------|--------|
| Model           | 4  | 0.09697600     | 0.02424400  | 5.54    | 0.0036 |
| Error           | 20 | 0.08760000     | 0.00438000  |         |        |
| Corrected Total | 24 | 0.18457600     |             |         |        |

| R-Square | Coeff Var | Root MSE | Calcium Mean |
|----------|-----------|----------|--------------|
| 0.525399 | 0.282301  | 0.066182 | 23.44360     |

| Source | DF | Type I SS  | Mean Square | F Value | Pr > F |
|--------|----|------------|-------------|---------|--------|
| Batch  | 4  | 0.09697600 | 0.02424400  | 5.54    | 0.0036 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|--------|----|-------------|-------------|---------|--------|
| Batch  | 4  | 0.09697600  | 0.02424400  | 5.54    | 0.0036 |

| Source | Type III Expected Mean Square<br>Var(Error) + 5 Var(Batch) |
|--------|--|
| Batch  |  |

when the null hypothesis is true. (Contrasts and orthogonal contrasts are discussed in Section 6.9.) The diets are

```
control: control
A1: control + amount 1 of additive A
A2: control + amount 2 of additive A
B1: control + amount 1 of additive B
B2: control + amount 2 of additive B
```

The data are read with (`oway/code/turkey-oway.s`) in S-PLUS or with (`oway/code/turkey.read.sas`) in SAS and plotted in Figure 6.4.

### 6.8.1 Analysis

The ANOVA table and table of means are in Table 6.8. The first thing we notice is that the diets differ significantly in their promotion of weight gain ( $F_{4,25} = 81.7$ ,  $p\text{-value} \approx 0$ ). Then we observe that the diets are structured so that particular comparisons among them are of special interest. We make these comparisons by partitioning the sum of squares to reflect several well-defined contrasts. The contrasts and the ANOVA table using them are displayed in Table 6.9.

The interaction line `diet: A.vs.B.by.amount` in Table 6.9 asks the question, “Does the increase from amount 1 to amount 2 of additive A have the same effect as the increase from amount 1 to amount 2 of additive B?” This question may equivalently be stated as: “Does the change from amount 1 of additive A to amount 1 of additive B have the same effect as the change from amount 2 of additive A to amount 2 of additive B?” (We interpret the description of the experiment to mean that the amounts 1 and 2 of the additives are measured in the same units). The concept of interaction is discussed in detail in Chapter 12.

These contrasts decompose the 4-df sum of squares for diet into four single-df sums of squares, one for each of the four contrasts. This set of contrast sums of squares is additive because we have defined the contrasts in such a way that they are *mutually orthogonal*. In essence this means that the information contained in one of the contrasts is independent of the information contained in any of the other contrasts. The independence of information makes each of the contrasts more readily interpretable than they would be if the contrasts had been defined without the property of orthogonality.

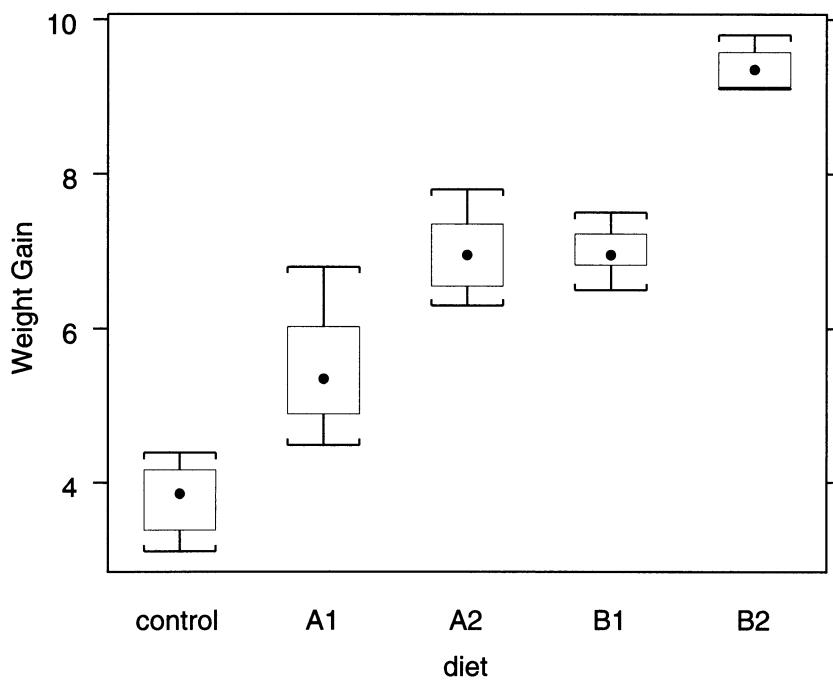


FIGURE 6.4. Turkey Data: Boxplots of Weight Gain for each Diet  
(`oway/code/turkey-oway.s`), (`oway/figure/turkey.f1.eps.gz`)

TABLE 6.8. ANOVA Table for Turkey Data  
(*oway/code/turkey-oway.s*), (*oway/code/turkey.aov1.sas*)

---

```
S-PLUS (oway/transcript/turkey.aov1.st):
> turkey.aov <- aov(wt.gain ~ diet, data=turkey)
> summary(turkey.aov)
   Df Sum of Sq Mean Sq F Value      Pr(F)
diet    4   103.038 25.7595 81.67248 5.595524e-014
Residuals 25     7.885  0.3154
> model.tables(turkey.aov, type="means", se=T)

Tables of means
Grand mean

6.53

diet
control A1     A2 B1     B2
3.7833 5.5 6.9833 7 9.3833

Standard errors for differences of means
diet
0.32424
replic. 6.00000
```

---

TABLE 6.9. ANOVA Table for Turkey Data with Contrasts. The sum of the individual sums of squares from each “diet:” line is the sum of squares for diet. The point estimates of the contrasts are in Table 7.3. Development of this table’s  $F$  statistic for diet: A.vs.B is explained in the discussion surrounding Equations (6.13)–(6.15).  
 (oway/code/turkey-oway.s)

```

S-PLUS (oway/transcript/turkey.contrasts.st):
> contrasts(turkey$diet)
      control.vs.treatment A.vs.B amount A.vs.B.by.amount
control           1.00    0.0    0.0          0.0
      A1             -0.25   0.5    0.5          0.5
      A2             -0.25   0.5   -0.5         -0.5
      B1             -0.25  -0.5    0.5         -0.5
      B2             -0.25  -0.5   -0.5          0.5
>
> tapply(turkey$wt.gain, turkey$diet, mean) %*% contrasts(turkey$diet)
      control.vs.treatment A.vs.B     amount A.vs.B.by.amount
[1,]           -3.433333  -1.95 -1.933333            0.45
>
> turkey2.aov <- aov(wt.gain ~ diet, data=turkey)
> summary(turkey2.aov)
      Df Sum of Sq Mean Sq F Value    Pr(F)
diet    4 103.038 25.7595 81.67248 5.595524e-014
Residuals 25    7.885  0.3154
> summary(turkey2.aov,
+   split=list(diet=list(
+     control.vs.treatment=1,
+     A.vs.B=2,
+     amount=3,
+     A.vs.B.by.amount=4)))
      Df Sum of Sq Mean Sq F Value    Pr(F)
diet    4 103.0380 25.75950 81.6725 0.000000000
diet: control.vs.treatment 1  56.5813 56.58133 179.3955 0.000000000
      diet: A.vs.B 1  22.8150 22.81500 72.3367 0.000000001
      diet: amount 1  22.4267 22.42667 71.1055 0.000000001
      diet: A.vs.B.by.amount 1   1.2150  1.21500  3.8523 0.06089888
Residuals 25    7.8850  0.31540

```

### 6.8.2 Interpretation

We tentatively interpret the contrast analysis as follows:

1. trt.vs.control: averaged over the 4 treatments, turkeys receiving a dietary additive gain significantly more weight than ones not receiving an additive.
2. additive: turkeys receiving additive B gain significantly more weight than turkeys receiving additive A.
3. amount: turkeys receiving amount 2 gain significantly more weight than turkeys receiving amount 1.
4. interaction between additive and amount: the extent of increased weight gain as a result of receiving amount 2 rather than amount 1 is not significantly different for the two additives.

Our conclusions derive from the definitions of the contrasts, the signs of their estimates in Table 6.8, and the results of the tests that each contrast is 0, shown in Table 6.9. We give further discussion of appropriate techniques for simultaneously testing the point estimates of the contrasts in Section 7.1.4.1. We illustrate the conclusions in Figure 7.4. In general, conclusions such as these are tentative because we are making several simultaneous inferences. Therefore, it may be appropriate to use a form of Type I error control that accounts for the simultaneity. See the discussion of multiple comparisons in Chapter 7. In this example, with very small *p*-values, the use of simultaneous error control will not lead to different conclusions.

### 6.8.3 Specification of Analysis

#### S-PLUS

The partitioned ANOVA table in Table 6.9 is constructed and displayed in two separate steps in file (`oway/code/turkey-oway.s`).

We specify the contrasts in several steps: We display the default contrasts, we define new contrasts, we display the new contrasts. Sometimes several iterations are needed until we get it right. Table 6.9 shows the display of the new contrasts.

Once the contrasts are defined, we use them in the `aov()` command. The `aov()` command uses the contrasts that are in the `data.frame` when it is called. Redefining the contrasts after the `aov()` command has been used has no effect on the `aov` object that has already been created.

The `split` argument to the `summary` command indexes the columns of the contrasts that are in the `aov` object. The index numbers in the `list`

argument are necessary. The names of the items in the list are optional. They are used to provide pretty labels in the ANOVA table.

SAS

We show in file (`mcomp/code/turkey.contrasts2.sas`) the specification of the same contrasts for SAS.

## 6.9 Contrasts

Once we have determined that there are differences among the means of the groups, that is, that the null hypothesis is rejected, we must follow through by determining the pattern of differences. Is one specific group responsible for the differences? Are there subsets of groups that behave differently than other subsets? We make this determination by partitioning the treatment sum of squares  $SS_{\text{treatment}}$  into single degree-of-freedom components, each associated with a *contrast* among the group means.

The concept of a contrast among group means was first encountered in Section 6.8. Contrasts are chosen primarily from the structure of the levels, for example, the average effect of Treatment A at several levels compared to the average effect of Treatment B at several levels (the `A.vs.B` contrast in Tables 6.9, 7.3, and 7.4 and in Figure 7.4). Or, for another example, a linear effect of the response to a linear increase in speed (the `.L` contrast in Section 10.4).

### 6.9.1 Mathematics of Contrasts

The mathematics of contrasts follows directly from the mathematics of the independent two-sample *t*-test:

$$t_{\text{calc}} = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (5.11)$$

The residual mean square  $s_{\text{Resid}}^2$  from the ANOVA table takes the place of  $s_p^2$ .

We look closely at the `A.vs.B` contrast in Table 6.9 comparing the average of the A treatments  $\bar{Y}_1 = (\bar{Y}_{A1} + \bar{Y}_{A2})/2$  to the average of the B treatments  $\bar{Y}_2 = (\bar{Y}_{B1} + \bar{Y}_{B2})/2$  with  $n_1 = n_{A1} + n_{A2} = n_2 = n_{B1} + n_{B2}$ .

Direct substitution of these values into Equation (5.11) with  $n \stackrel{\text{def}}{=} n_{\text{control}} = n_{A1} = n_{A2} = n_{B1} = n_{B2}$ , followed by simplification (see Exercise 6.12)

leads to

$$t_{\text{calc}} = \frac{(\bar{Y}_{A1} + \bar{Y}_{A2})/2 - (\bar{Y}_{B1} + \bar{Y}_{B2})/2}{\frac{1}{2} s_{\text{Resid}} \sqrt{\frac{1}{n} + \frac{1}{n} + \frac{1}{n} + \frac{1}{n}}} \stackrel{\text{def}}{=} t_{\mathbf{A}.\mathbf{vs}.\mathbf{B}} \quad (6.13)$$

We can write the numerator of Equation (6.13) as the dot product

$$\begin{aligned} C_{\mathbf{A}.\mathbf{vs}.\mathbf{B}} &= (\bar{Y}_{\text{control}} \quad \bar{Y}_{A1} \quad \bar{Y}_{A2} \quad \bar{Y}_{B1} \quad \bar{Y}_{B2}) \cdot (0 \quad \frac{1}{2} \quad \frac{1}{2} \quad -\frac{1}{2} \quad -\frac{1}{2}) \\ &= (\bar{Y}_j) \cdot (c_j) \end{aligned} \quad (6.14)$$

and then recognize the denominator of Equation (6.13) as the square root of the estimator of the variance of the numerator when the null hypothesis is true

$$\widehat{\text{var}}(C_{\mathbf{A}.\mathbf{vs}.\mathbf{B}}) = \frac{1}{4} s_{\text{Resid}}^2 \left( \frac{1}{n} + \frac{1}{n} + \frac{1}{n} + \frac{1}{n} \right) \quad (6.15)$$

When we do the arithmetic, the value

$$t_{\mathbf{A}.\mathbf{vs}.\mathbf{B}} = \frac{C_{\mathbf{A}.\mathbf{vs}.\mathbf{B}}}{\sqrt{\widehat{\text{var}}(C_{\mathbf{A}.\mathbf{vs}.\mathbf{B}})}} = 8.5051 = \sqrt{72.3367} = \sqrt{F_{\mathbf{A}.\mathbf{vs}.\mathbf{B}}}$$

is recognized as the square root of the  $F$ -statistic for the `diet: A.vs.B` line of the ANOVA table in Table 6.9.

The vector  $c = (c_j) = (0 \quad \frac{1}{2} \quad \frac{1}{2} \quad -\frac{1}{2} \quad -\frac{1}{2})$  is called a contrast vector and the product  $C_{\mathbf{A}.\mathbf{vs}.\mathbf{B}}$  is called a contrast. The numbers  $c_j$  in a contrast vector satisfy the constraint that

$$\sum_j c_j = 0 \quad (6.16)$$

Under the null hypothesis that  $\mu_1 = \mu_2 = \dots = \mu_5$ , we have  $E(C_{\mathbf{A}.\mathbf{vs}.\mathbf{B}}) = E(\sum_j c_j \mu_j) = 0$ . Under both the null and alternative hypotheses, assuming that all  $\sigma_j^2$  are identical and equal to  $\sigma^2$ , we see that

$$\text{var}(C_{\mathbf{A}.\mathbf{vs}.\mathbf{B}}) = \frac{\sigma^2}{n} \sum c_j^2$$

A similar argument shows that each of the columns listed under `contrasts(turkey$diet)` in Table 6.9 can be used to construct the correspondingly named row of the ANOVA table (Exercise 6.13).

This set of contrasts has an additional property. They are orthogonal. This means that the dot product of each column with any of the others is 0, for example,

$$C_{\mathbf{A}.\mathbf{vs}.\mathbf{B}} \cdot c_{\text{amount}} = (0 \quad \frac{1}{2} \quad \frac{1}{2} \quad -\frac{1}{2} \quad -\frac{1}{2}) \cdot (0 \quad \frac{1}{2} \quad -\frac{1}{2} \quad \frac{1}{2} \quad -\frac{1}{2}) = 0 \quad (6.17)$$

The implication is that the covariance of the contrasts is zero (for example,  $\text{cov}(C_{\mathbf{A}.\mathbf{vs}.\mathbf{B}}, C_{\text{amount}}) = 0$ ), that is, the contrasts are uncorrelated. As a

consequence, the sum of sums of squares for each of the four contrasts in Table 6.9 is the same as the sum of squares for **diet** given by the  $SStreatment$  term in Equations (6.6) and (6.9).

The  $SStreatment$ , and the sum of squares for each of the single degree-of-freedom contrasts comprising it, is independent of the  $MS_{residual} = s^2_{\text{Resid}}$ . The  $F$ -tests for each of the orthogonal contrasts are not independent of each other because all use the same denominator term.

In general, the  $n_j$  are not required to be identical. The general statement for a contrast vector

$$(c_j) = (c_1, \dots, c_J) \quad (6.18)$$

is that the contrast  $C = \sum c_j \bar{Y}_j$  has variance

$$\text{var}(C) = \sigma^2 \sum \frac{c_j^2}{n_j} \quad (6.19)$$

### 6.9.2 Scaling

The contrasts displayed here were scaled to make the sum of the positive values and the sum of the negative values each equal to 1. This scaling is consistent with the phrasing that a contrast is a comparison of the average response over several levels of a factor to the average response over several different levels of the factor. Any alternate scaling is equally valid and will give the same sum of squares.

#### 6.9.2.1 Absolute-Sum-2 Scaling

We recommend the *absolute-sum-2* scaling where the sum of the absolute values of the coefficients equals 2,

$$\sum_j |c_j| = 2 \quad (6.20)$$

Equivalently, the sum of the positive coefficients equals 1 and the sum of the negative coefficients also equals 1. The *absolute-sum-2* scaling makes it possible to extend the mean–mean multiple comparisons plots to arbitrary sets of contrasts. See Section 7.2.3 for details on the mean–mean multiple comparisons plots.

#### 6.9.2.2 Normalized Scaling

The normalized scaling, with  $c_j^* = c_j / \sqrt{\sum c_j^2}$ , is frequently used because the corresponding dot product

$$C_{A.\text{vs.}B}^* = (\bar{Y}_{\text{control}} \quad \bar{Y}_{A1} \quad \bar{Y}_{A2} \quad \bar{Y}_{B1} \quad \bar{Y}_{B2}) \cdot (0 \quad \frac{1}{2} \quad \frac{1}{2} \quad -\frac{1}{2} \quad -\frac{1}{2})$$

$$= (\bar{Y}_j) \cdot (c_j^*) \quad (6.21)$$

is simply related to the A.vs.B sum of squares by

$$\text{SS}_{\text{A.vs.B}} = n(C_{\text{A.vs.B}}^*)^2 = 22.815. \quad (6.22)$$

Under the null hypothesis

$$\text{var}(C_{\text{A.vs.B}}^*) = \sigma_{\text{Resid}}^2 \quad (6.23)$$

and under the alternate hypothesis

$$\text{var}(C_{\text{A.vs.B}}^*) \gg \sigma_{\text{Resid}}^2 \quad (6.24)$$

This provides the justification for the  $F$ -test.

In this example, the normalized scaling in Equation (6.21) is identical to the scaling in Equation (6.14) that makes the positive and negative sums each equal to 1. That is not always the case. The `control.vs.treatment` contrast with positive and negative values each summing to 1 as displayed in Table 6.9 is

$$(1 - .25 - .25 - .25 - .25)$$

The same `control.vs.treatment` contrast with normalized scaling is

$$\sqrt{.8} (1 - .25 - .25 - .25 - .25)$$

### 6.9.2.3 Integer Scaling

Another frequently used scaling makes each individual value  $c_j$  an integer. For the examples shown here, this gives

|                                   |                                     |
|-----------------------------------|-------------------------------------|
| <code>A.vs.B</code>               | (0      1      1      -1      -1)   |
| <code>control.vs.treatment</code> | (4      -1      -1      -1      -1) |

This scaling eases hand arithmetic and was very important prior to digital computers.

## 6.10 Tests of Homogeneity of Variance

In Sections 5.3, 5.4.4, 6.2, and 6.6 we mention that the assumption that several populations have a common variance can be checked via a statistical test. Assuming there are  $a$  populations having variances  $\sigma_i^2$  for  $i = 1, 2, \dots, a$ , the test is of the form

$$\begin{aligned} H_0: \sigma_1^2 &= \sigma_2^2 = \dots = \sigma_a^2 \\ \text{vs} \\ H_1: \text{not all the } \sigma_i^2 &\text{ are identical to each other.} \end{aligned} \quad (6.25)$$

For this purpose, (Brown and Forsyth, 1974) present the recommended test. Intensive simulation investigations, including (Conover et al., 1981), have found that this test performs favorably compared with all competitors in terms of Type I error control and power for a wide variety of departures from Normality.

The Brown and Forsyth test statistic is the  $F$  statistic resulting from an ordinary one-way analysis of variance on the absolute deviations from the median

$$Z_{ij} = |Y_{ij} - \bar{Y}_i^{\perp}| \quad (6.26)$$

where  $\bar{Y}_i^{\perp}$  is the median of  $\{Y_{i1}, \dots, Y_{in_i}\}$ .

This test may be requested in SAS with a `means` statement following a `model` statement in either PROC ANOVA or PROC GLM. The statement is of the form

```
means A / hovtest=bf;
```

where `A` is a class variable declared by the PROC.

The test is available as the `hov` function in files (`splus.library/hov-bf.s`) and (`splus.library/hov.plot.s`) with the form

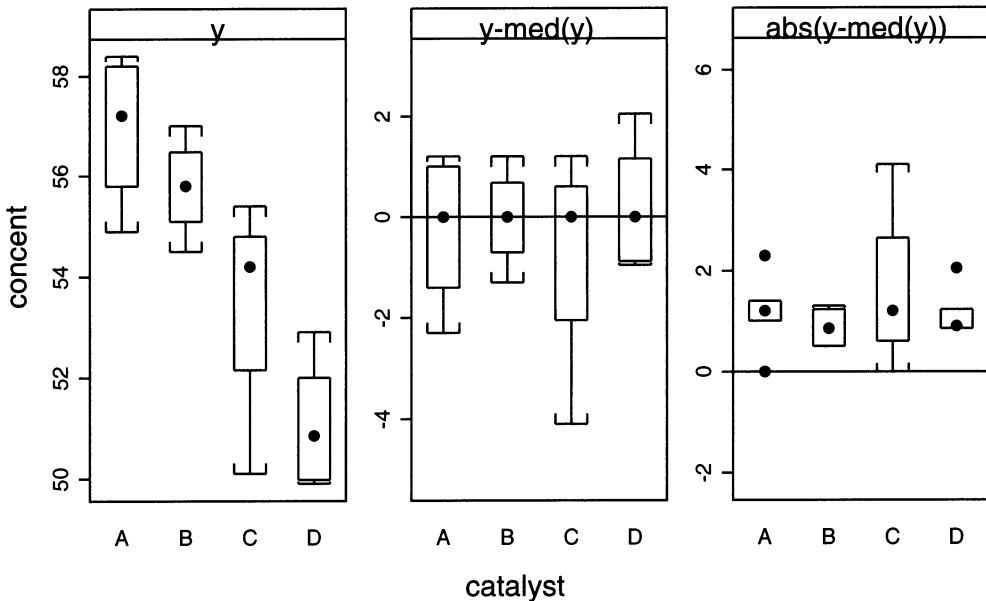
```
hov( y ~ A )
```

where `A` is a factor.

We continue the (`datasets/catalystm.dat`) example of Sections 6.1 and 6.6. Our impression from Figure 6.1 is that that catalyst 3 has a larger variance than the other three catalysts. We formally test this possibility with the Brown–Forsyth test, illustrated in Figure 6.5. Because of the large  $p$ -value, .74, we are unable to conclude that the variances of the concentrations are not identical.

## 6.11 Exercises

- 6.1. (Till, 1974), also cited in (Hand et al., 1994), compared the salinity (in parts per 1000) for three distinct bodies of water in the Bimini Lagoon, Bahamas. The data are in the file (`datasets/salinity.dat`). Analyze the data under the assumption that the 3 bodies of water constitute a random sample from a large number of such bodies of water.
- 6.2. (Milliken and Johnson, 1984) report on an experiment to compare the rates of workers' pulses during 20-second intervals while performing one of 6 particular physical tasks. Here 68 workers were randomly assigned to one of these tasks. The data are contained in the file



```
S-PLUS (oway/transcript/catalystm.aov3.st):
> hov(concent ~ catalyst, data=catalystm)
```

```
hov: Brown-Forsyth

data: concent
F = 0.4202, df:catalyst = 3, df:Residuals = 12, p-value = 0.7418
alternative hypothesis: variances are not identical

> plot.hov(concent ~ catalyst, data=catalystm)
```

FIGURE 6.5. Catalyst data: Brown-Forsyth test of the hypothesis of equal variances in Equation (6.25). The left panel shows the original data. The middle panel shows the deviations from the median for each group, hence is a recentering of the left panel. The right panel shows absolute deviations from the median. The central dots for each catalyst in the right panel, the MAD (median absolute deviation from the median), are approximately equal, reflecting the null distribution of the Brown-Forsyth test statistic. Hence we do not reject this test's null hypothesis. We conclude that the variances of the concentrations in the four catalyst groups are approximately equal.

(oway/code/catalystm.s), (oway/figure/catalystm-hov.eps.gz),  
 (oway/code/catalystm.hov.sas), (oway/transcript/catalystm.hov.lste)

(`datasets/pulse.dat`). Investigate differences between the mean pulse rates associated with the various tasks.

- 6.3.** (Johnson and Leone, 1967) provide (`datasets/operator.dat`). Five operators randomly selected from all company operators are each given four equal time slots in which their production is measured. Perform an appropriate analysis. Does it appear as if the population of operators has homogeneous productivity?
- 6.4.** (Anionwu et al., 1981), also reprinted in (Hand et al., 1994), examine whether hemoglobin levels for patients with sickle cell anemia differ across three particular types of sickle cell disease. Here `type` is a fixed factor and its three qualitative levels are “*HB SS*”, “*HB S/thalassaemia*”, and “*HB SC*”. The data appear in the file (`datasets/sickle.dat`). Perform an analysis of variance and multiple comparison with the Tukey procedure to compare the patients’ hemoglobin for the three types.
- 6.5.** (Cameron and Pauling, 1978), also reprinted in (Hand et al., 1994), compare the survival times in days of persons treated with supplemental ascorbate following a diagnosis of cancer at five organ sites: *Stomach*, *Bronchus*, *Colon*, *Ovary*, and *Breast*. The dataset in file (`datasets/patient.dat`) is easily read by people because the columns are aligned. The file is less easily read by a program because the lines are of unequal length and have missing values for some of the variables. See files (`oway/code/patient.s`) and (`oway/code/patient.sas`) to read the data.
  - a.** Perform a log transformation of the response `days` for each of the five levels of the factor `site` in order to improve conformity with the required assumption that the data be approximately normally distributed with equal within-site variance. Produce and compare boxplots to compare the response before and after the transformation.
  - b.** Perform an analysis to assess differences in mean survival between the different cancer sites.
- 6.6.** (NIST, 2002) reports the result of an experiment comparing the absorbed energy produced by each of four machines. The machines are labeled *Tinius1*, *Tinius2*, *Satec*, and *Tokyo*. The data are contained in (`datasets/notch.dat`). Assuming that these were the only machines of interest, compare the responses on the four machines and use the Tukey procedure to assess significant differences among them.
- 6.7.** An experiment was designed to examine the effect of storage temperature on the potency of an antibiotic. Fifteen antibiotic samples were

obtained and three samples were stored at each of the five indicated temperatures (degrees F). The potencies of each sample were checked after 30 days. The dataset, taken from (Peterson, 1985), is contained in the file (`datasets/potency.dat`).

- a. Perform an analysis of variance to confirm that potency changes with storage temperature.
  - b. Set up two orthogonal contrasts to assess the nature of the dependence of potency on temperature. You may use the contrast  $(-2, -1, 0, 1, 2)$  to assess linearity and the contrast  $(2, -1, -2, -1, 2)$  to assess whether there is a quadratic response. (Further discussion of polynomial contrasts is in Section 10.4.)
  - c. Test whether each of the contrasts you proposed in part b) is significantly different from 0.
  - d. Report your recommendations for the temperature at which this antibiotic should be stored.
- 6.8.** (Anderson and McLean, 1974) report the results of an experiment to compare the disintegration times in seconds of four types of pharmaceutical tablets labeled A, B, C, D. These were the only tablet types of interest. The data appear in (`datasets/tablet1.dat`). Perform an analysis of variance to see if the tablets have equivalent disintegration times. The time to disintegration determines when the medication begins to work. Shorter times mean the tablet will begin disintegrating in the stomach. Longer times mean the tablet will disintegrate in the small intestines where it more easily absorbed and less susceptible to degradation from the digestive enzymes. Assuming that longer times to disintegration are desirable, use the Tukey procedure to prepare a recommendation to the tablet manufacturer.
- 6.9.** The data (`datasets/blood.dat`) contain the results of an experiment reported by (Box et al., 1978) to compare the coagulation times in seconds of blood drawn from animals fed four different diets labeled A, B, C, D. Assuming that these were the only diets of interest, set up an analysis of variance to compare the effects of the diets on coagulation. Use the Tukey procedure to investigate whether any pairs of the diets can be considered to provide essentially equivalent coagulation times.
- 6.10.** Reconsider (`datasets/draft70mn.dat`) from (Data Archive, 1997), previously visited in Exercises 4.1 and 3.25. Assuming that the ranks were randomly assigned to the dates of the year, construct a one-way analysis of variance with the ranks as response and the months as groups. Isolate the linear effect of `month`.

- 6.11. Westfall and Rom (Westfall and Rom, 1990) considered the nonbirth litter weights of mice whose mothers were previously subjected to one of three treatments or a control, with the objectives of relating weight differences to treatment dosages. (It is conjectured that “nonbirth weight” refers to weight at some definite time following birth.) The data appear in the file (`datasets/mice.dat`). Perform a Brown-Forsyth homogeneity of variance test on these data and carefully state your conclusions.
- 6.12. Derive Equation (6.13) from Equation (5.11) by substitution and simplification as outlined in Section 6.9.
- 6.13. Verify that the four single degree-of-freedom lines in the ANOVA table in Table 6.9 can be obtained from the four contrasts in the `contrasts(turkey$dist)` section of Table 6.9.

## 6.A Appendix: Computation for the Analysis of Variance

### 6.A.1 Computing Notes

In both S-PLUS and SAS the design specification languages use simplifications of the more extended traditional notation

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad \text{for } i = 1, \dots, a \quad \text{and } j = 1, \dots, n_i \quad (6.1)$$

The intercept term  $\mu$  and the error term  $\epsilon_{ij}$  are assumed in both statistical languages. The existence of the subscripts is implied and the actual values are specified by the data values.

With S-PLUS we will be using `aov` for the calculations and `anova` and related commands for the display of the results. `aov` can be used with equal or unequal cell sizes  $n_i$ . Model (6.1) is denoted in S-PLUS by the formula

$$Y \sim A$$

The operator  $\sim$  is read as “is modeled by”.

With SAS we use `PROC GLM` and `PROC ANOVA`. `PROC GLM` is always appropriate. `PROC ANOVA` is appropriate for all one-factor designs (this chapter) with equal or unequal cell sizes  $n_i$ , but is limited to equal sample size cases when there is more than one factor as in Chapter 12. Model (6.1) is denoted in SAS by the expression

$$Y = A$$

The operator  $=$  is read as “is modeled by”.

### 6.A.2 Computation

Two different algorithms are used to calculate the analysis of variance for data with one factor: sums of squared differences of cell means and regression on dummy variables. Both give identical results.

The intuition of the analysis is most easily developed with the sums of squared differences algorithm. We began there in Equation 6.6 and the definitions in the notes to Table 6.3. The regression formulation is easier to work with and generalizes better. Once we have developed our intuition we will usually work with the regression formulation. The discussion of contrasts in Section 6.9 leads in to the regression formulation in Chapter 10.

# Multiple Comparisons

In Exercise 3.13 we discover that the probability of simultaneously making three correct inferences, when each of the three individually has  $P(\text{correct inference}) = 1 - \alpha = 0.95$ , is only  $(1 - \alpha)^3 = .95^3 = 0.857$ . Alternatively, the probability of making at least one incorrect inference is  $1 - 0.857 = 0.143 \approx 3\alpha$ . In general, the more simultaneous inferences we make at one time, the smaller the probability that all are correct. In this chapter we learn how to control the probability that all inferences are simultaneously correct. We usually phrase the goal as controlling the probability of making at least one incorrect inference.

We consider all inferences in a related *family* of inferences. Such a family is typically a natural and coherent collection; for example, all inferences resulting from a single experiment. The inferences can be individual tests of hypotheses or confidence intervals. In the context of a family of hypothesis tests, if we control the Type I error probability for each test at level  $\alpha$ , the probability of committing at least one Type I error in the family will be much larger than  $\alpha$ . For example, if the tests are independent and  $\alpha = .05$ , then the probability of at least one Type I error is  $1 - (1 - .05)^6 \approx .26$ , which seems an unacceptably large error threshold.

A way to avoid such errors when conducting many related inferences simultaneously is to employ a *multiple comparison procedure*. Such a procedure for *simultaneous hypothesis testing* may seek to (strongly) control the familywise error rate (FWE), defined as  $P(\text{reject at least one true hypothesis under any configuration of true and false hypotheses})$ . A procedure for *simultaneous confidence intervals* should control the probability that at least one member of the family of confidence intervals does not contain the

parameter being estimated by the interval. When a multiple comparison procedure is used, it is said that the analyst is *controlling for multiplicity*.

In order to exert FWE control over a family of related hypothesis tests, it is necessary to have a reduced probability of rejecting any particular null hypothesis in the family. As explained in Section 3.7, reducing the probability of rejecting particular hypotheses results in an increased probability of retaining them, and therefore reduced power for tests of these hypotheses. This implies that, as compared with testing hypotheses in isolation from one another, a multiple comparison procedure has a diminished ability to reject false null hypotheses. In other words, a test of a particular hypothesis using a multiple comparison procedure will be less powerful than the test of the same hypothesis in isolation. In deciding whether to use a multiple comparison procedure, the protection against the possibility of an excessive number of incorrect hypothesis rejections must be weighted against this loss of power. An analogous statement holds for simultaneous versus isolated confidence intervals.

In general, the choice of multiple comparison procedure to be used depends on the structure of the *family* of related inferences and the nature of the collection of statistics from which the confidence intervals or tests will be calculated.

Section 7.1 summarizes the most frequently used multiple comparisons procedures. Section 7.2 presents a graphical procedure for looking at the results of the multiple comparisons procedures.

## 7.1 Multiple Comparison Procedures

### 7.1.1 Bonferroni Method

A very general way to control the FWE is based on the Bonferroni inequality,  $P(\bigcup E_i) \leq \sum_i P(E_i)$ , where the  $E_i$  are arbitrary events. If the family consists of  $m$  related tests, conducting each test at level  $\frac{\alpha}{m}$  ensures that  $FWE \leq \alpha$ . If the family consists of  $m$  related confidence intervals, maintaining confidence  $100(1 - \frac{\alpha}{m})\%$  for each interval will ensure that the overall confidence of all  $m$  intervals will be at least  $100(1 - \alpha)\%$ . The Bonferroni method should be considered for use when the family of related inferences is unstructured (e.g., not like the structured families required for the procedures discussed in Sections 7.1.2–7.1.4), or when the statistics used for inference about each family member have nonidentical probability distributions.

The Bonferroni inequality is very blunt in the sense that its right side is typically much larger than its left. One reason for this is that it does not

seek to take into account information about the intersections of the events  $E_i$ . As a result, the Bonferroni approach is very conservative in the sense of typically guaranteeing an FWE substantially less than its nominal value of  $\alpha$ , and the extent of this conservativeness increases with  $m$ . The value of this approach is that it is very generally applicable, for example, when the pivotal statistics associated with the  $m$  inferences have nonidentical probability distributions. (Hochberg, 1988) provides an easy-to-understand improvement to the Bonferroni approach for hypothesis testing that tends to reject more false null hypotheses than Bonferroni. Hochberg's procedure has been proven to be applicable to a wide variety of testing situations; see (Sarkar, 1998).

### 7.1.2 Tukey Procedure for All Pairwise Comparisons

Often a family of inferences has a special structure that allows us to use available information about the joint distributions of the pivotal statistics, thus enabling the use of a less conservative approach than Bonferroni. An example of this, discussed in Section 6.3, is the family consisting of all  $m = \binom{k}{2}$  comparisons among all pairs of means of  $k$  populations. For this family, Tukey's Studentized range test is usually recommended.

### 7.1.3 The Dunnett Procedure for Comparing One Mean with All Others

The Dunnett procedure is used when the family of inferences of interest is the comparisons of the mean of one designated population with each of the means of the remaining populations, all populations being at least approximately normal with approximately the same variance. Often in practice the designated population is a control and the others are active treatments. The Dunnett procedure uses the percentiles of a multivariate  $t$  distribution rather than a univariate  $t$  distribution discussed in Section 5.4.3.

For purposes of illustration of the Dunnett procedure, we use weight-loss data. A random sample of 50 men who were matched for pounds overweight was randomly separated into 5 equal groups. Each group was given exactly one of the weight loss agents A, B, C, D, or E. After a fixed period of time, each man's weight loss was recorded. The data, taken from (Ott, 1993), appear in the file (`datasets/weightloss.dat`) and in Figure 7.1.

The  $F$ -statistic tests the null hypothesis that the five groups have identical mean weight loss vs the alternative that the groups do not have identical mean weight loss. The small  $p$ -value from the  $F$  test in the basic ANOVA in Table 7.1 suggests that the agents have differing impacts on weight loss.

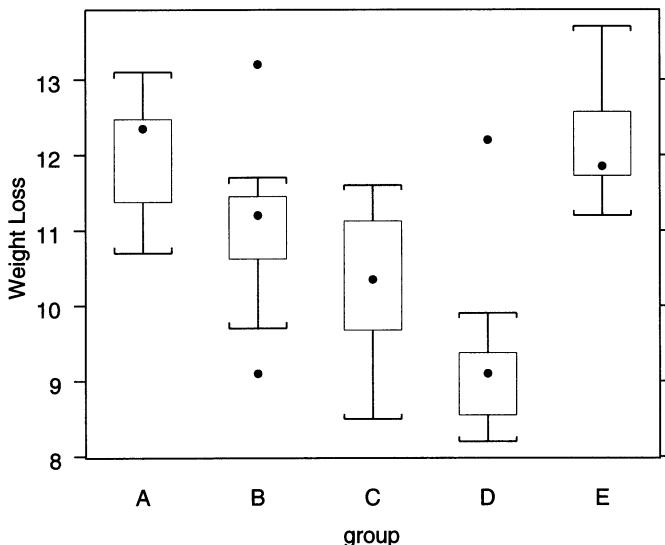


FIGURE 7.1. Weigh-loss data: Boxplots of weight loss for each group.  
`(mcomp/code/weightloss.s)`, `(mcomp/figure/weightloss-data.eps.gz)`

When we regard agent D as the control, we seek to investigate whether any of the other four agents appear to promote significantly greater weight loss than agent D. From Figure 7.1 we see that the five populations are approximately normal with approximately the same variance. Therefore, we may proceed with the Dunnett procedure. Since we are investigating whether the other agents *improve* on D, we display infinite upper one-sided confidence intervals against D in Table 7.2 and Figures 7.2 and 7.3.

The (default) 95% confidence level in Table 7.2 applies simultaneously to all four confidence statements. The fact that all four confidence intervals lie entirely above zero suggests that D is significantly inferior to the other four weight-loss agents.

Figure 7.3 is a mean–mean display of the result of applying Dunnett’s multiple comparison procedure to the weight-loss data, analogous to Figure 6.3 in Section 6.3. The mean–mean display technique is discussed in detail in Section 7.2. In Figure 7.3, reflecting the results for upper one-sided Dunnett confidence intervals, all horizontal lines except that for comparing groups D and C fall to the right of zero. Consistent with the boxplots in Figure 7.1, we conclude that all weight-loss agents (except possibly C) provide superior mean weight loss to that provided by agent D.

The Dunnett procedure is used in Exercises 7.7 and 12.4.

TABLE 7.1. Weight-loss ANOVA  
 (mcomp/code/weightloss.s), (mcomp/transcript/weightloss.st)

---

SAS (mcomp/code/weightloss.sas):

```

data weightloss;
  infile "&hh/datasets/weightloss.dat" firstobs=2;
  input Group $ Loss;
run;

proc anova;
  class Group;
  model Loss = Group;
  means Group / dunnett('D');
run;
```

---

SAS (mcomp/transcript/weightloss-a.lst):  
 Dependent Variable: Loss

| Source          | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|-----------------|----|----------------|-------------|---------|--------|
| Model           | 4  | 59.8792000     | 14.9698000  | 15.07   | <.0001 |
| Error           | 45 | 44.7040000     | 0.9934222   |         |        |
| Corrected Total | 49 | 104.5832000    |             |         |        |

---

### Computing Note

Note the difference in notation and conventions between the two programs.

SAS PROC ANOVA needs the option

dunnett('D')

SAS uses the suffix "u" to indicate an infinite upper interval. The corresponding SAS function for two-sided intervals is `dunnett()` and for infinite lower one-sided intervals `dunnett1()`.

The S-PLUS function `multicomp` needs the arguments

method="dunnett", comparisons="mcc",  
 bounds="lower", control=4

S-PLUS uses the argument "lower" to indicate a finite lower bound. The S-PLUS argument for finite upper bounds is "upper" and for two-sided intervals is "both".

TABLE 7.2. Weight loss using the Dunnett procedure.  
 (mcomp/code/weightloss.sas)

---

SAS (mcomp/transcript/weightloss-b.lst):  
 Dunnett's One-tailed t Tests for Loss

NOTE: This test controls the Type I experimentwise error for comparisons  
 of all treatments against a control.

|                                |          |
|--------------------------------|----------|
| Alpha                          | 0.05     |
| Error Degrees of Freedom       | 45       |
| Error Mean Square              | 0.993422 |
| Critical Value of Dunnett's t  | 2.22241  |
| Minimum Significant Difference | 0.9906   |

Comparisons significant at the 0.05 level are indicated by \*\*\*.

| Group Comparison | Difference    |              |                       |
|------------------|---------------|--------------|-----------------------|
|                  | Between Means | Simultaneous | 95% Confidence Limits |
| E - D            | 2.9000        | 1.9094       | Infinity ***          |
| A - D            | 2.7800        | 1.7894       | Infinity ***          |
| B - D            | 1.7500        | 0.7594       | Infinity ***          |
| C - D            | 1.0000        | 0.0094       | Infinity ***          |

---

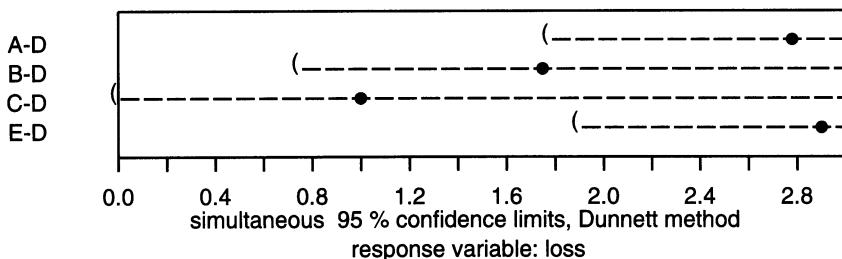


FIGURE 7.2. Weight-loss data: Standard display of one-sided multiple comparisons using the Dunnett method against the control treatment D.

(mcomp/code/weightloss.s), (mcomp/figure/weightloss.dunnet.eps.gz)

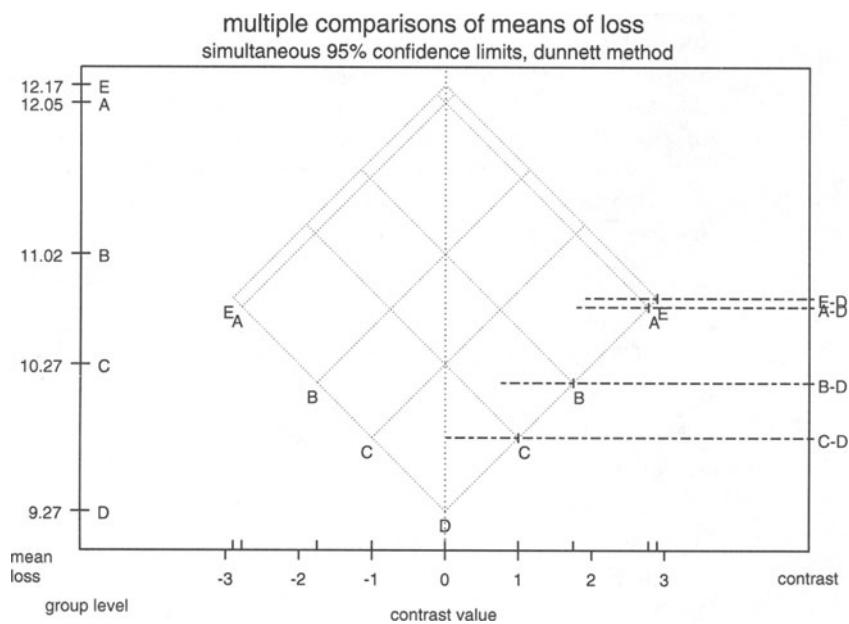


FIGURE 7.3. Weight-loss data: Mean–mean display of one-sided multiple comparisons using the Dunnett method against the control treatment D. Please see the discussion of the mean–mean display in Section 7.2.

([mcomp/code/weightloss.s](#)), ([mcomp/figure/weightloss.dunnet.mmc.eps.gz](#))

### 7.1.4 Simultaneously Comparing All Possible Contrasts—Scheffé and Extended Tukey

#### 7.1.4.1 The Scheffé Procedure

In the context of comparing the means of  $a$  populations, the Scheffé multiple comparison procedure controls the familywise error rate over the infinite-sized family consisting of all possible contrasts  $\sum_{j=1}^a c_j \mu_j$  involving the population means. The Scheffé procedure is therefore appropriate for exerting simultaneous error control over the set of four contrasts in our analysis of the turkey data (`datasets/turkey.dat`) from Section 6.8. In exchange for maintaining familywise error control over so large a family, the Scheffé method gives rise to wide confidence limits and relatively unpowerful tests. Therefore, we recommend its use only in the narrowly defined situation of simultaneously inferring about mean contrasts more complex than a comparison of two means. The Scheffé procedure uses a percentile of an  $F$  distribution, derived as the distribution of the most significant standardized contrast among the sample means.

The confidence interval formula by the Scheffé procedure is

$$\text{CI} \left( \sum_{j=1}^a c_j \mu_j \right) = \sum_{j=1}^a c_j \bar{y}_j \pm \sqrt{(a-1)F_{0.05,a-1,N-a}} s \sqrt{\sum_{j=1}^a \frac{c_j^2}{n_j}} \quad (7.1)$$

This provides the set of  $100(1 - \alpha)\%$  simultaneous confidence intervals for all possible contrasts among the population means. In this equation  $N = \sum_{j=1}^a n_j$ .

The Scheffé test is one of the methods built-in to the S-PLUS `multicomp` command. In the file (`sas.library/code/ischeffe.sas`) we provide a SAS macro.

#### Scheffé Intervals with the Turkey Data

Table 6.9 provides  $F$ -tests of the hypotheses that the members of a basis set of four contrasts are zero. These four tests do not control for multiplicity. The finding in Table 6.9 is that three of these contrasts differ significantly from zero. We do not declare the fourth contrast significantly different from zero because its  $p$ -value exceeds 0.05.

The Scheffé procedure allows us to make inferences about these same contrasts while controlling for multiplicity. The confidence interval and testing results are shown in Tables 7.3 and 7.4 and in Figure 7.4. An additional advantage of the Scheffé analysis is that the results specify the direction of contrasts' significant difference from zero. For example, in Table 7.3, the fact that the confidence interval on A vs. B lies entirely below zero implies that, on average, the mean weight gain from diet B exceeds that from diet

TABLE 7.3. Scheffé Test for Turkey Data Contrasts by S-PLUS. See also Figure 7.4.  
(oway/code/turkey-oway.s)

---

```
S-PLUS (mcomp/transcript/turkey.contrasts2.st):
> multicompare(turkey2.aov,
+                 focus="diet",
+                 lmat=rbind(0,contrasts(turkey$diet)),
+                 method="scheffe",
+                 bounds="both",
+                 plot=T,
+                 comparisons="none")

95 % simultaneous confidence intervals for specified
linear combinations, by the Scheffe method

critical point: 3.3219
response variable: wt.gain
rank used for Scheffe method: 4

intervals excluding 0 are flagged by '****'

      Estimate Std.Error Lower Bound Upper Bound
control.vs.treatment   -3.43     0.256    -4.280    -2.58 ****
          A.vs.B       -1.95     0.229    -2.710    -1.19 ****
          amount       -1.93     0.229    -2.690    -1.17 ****
A.vs.B.by.amount        0.45     0.229    -0.312     1.21
```

---

A. The  $F$ -statistics in Table 6.9 are essentially squared  $t$ -statistics, and this obscures information on directionality unless the definitions of the contrasts being tested are carefully examined alongside the test results.

We may use either SAS or S-PLUS to assess the extent to which, if any, of the Scheffé simultaneous confidence intervals cause us to modify our previous conclusions about the contrasts. When doing so it is important to observe the contrast codings, that is, the numerical values defining the contrast. Observing that the first three of the four Scheffé intervals exclude 0 while the last one includes 0, the Scheffé results reinforce our original impressions from the nonsimultaneous  $F$ -tests of these contrasts in Table 6.9.

In this example, examination of the Scheffé results did not cause us to revise our earlier results ignoring multiplicity. In general, use of a multiple comparison procedure is an appropriately conservative approach that

TABLE 7.4. Scheffé test for turkey data contrasts by SAS.  
 (mcomp/transcript/turkey.contrasts2.lst), (sas.library/code/ischeffe.sas)

---

SAS (mcomp/code/turkey.contrasts2.sas):

```
/*
   The %ischeffe macro requires the output dataset
   that MIXED produces.
   When there is no RANDOM statement,
   the MIXED and GLM give the same estimates.
*/
proc mixed data=turkey;
  class diet;
  model wtgain = diet;
  make 'estimates' out=estout;
  estimate 'control vs rest'    diet 1 -.25 -.25 -.25 -.25 ;
  estimate 'A vs B'            diet 0 .5 .5 -.5 -.5 ;
  estimate 'level 1 vs level 2' diet 0 .5 -.5 .5 -.5 ;
  estimate 'A x B'             diet 0 .5 -.5 -.5 .5 ;
run;

%ischeffe(d_indata=turkey,
           v_class=diet,
           d_estout=estout,
           v_fmt=8.5);
```

---

SAS (mcomp/transcript/turkey.contrasts2a.lst):

| Label              | Scheffe Interval |             |             |                |    |        | t<br>Value | Pr> t |
|--------------------|------------------|-------------|-------------|----------------|----|--------|------------|-------|
|                    | Estimate         | Lower Limit | Upper Limit | Standard Error | DF |        |            |       |
| control vs rest    | -3.4333          | -4.2848     | -2.5818     | 0.2563         | 25 | -13.39 | <.0001     |       |
| A vs B             | -1.9500          | -2.7116     | -1.1884     | 0.2293         | 25 | -8.51  | <.0001     |       |
| level 1 vs level 2 | -1.9333          | -2.6949     | -1.1717     | 0.2293         | 25 | -8.43  | <.0001     |       |
| A x B              | 0.4500           | -0.3116     | 1.2116      | 0.2293         | 25 | 1.96   | 0.0609     |       |

---

may not declare a difference found by nonsimultaneous tests or confidence intervals.

Figures 7.5 and 7.6 are graphic presentations of the Scheffé procedure applied to comparisons of all pairs of means. We use Scheffé intervals here because these pairwise comparisons are part of a larger family of contrasts

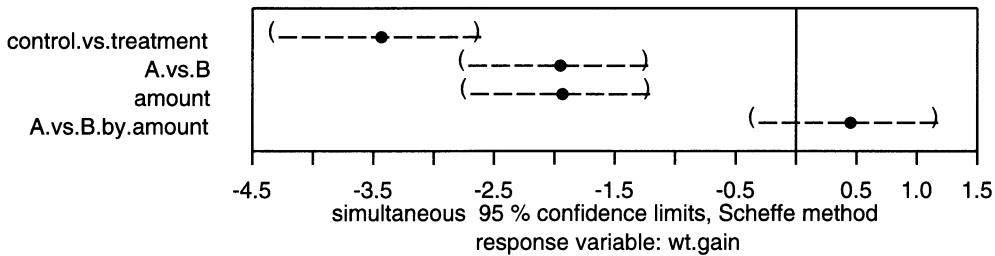


FIGURE 7.4. Scheffé plot for turkey data. See also Table 7.3.  
`(oway/code/turkey-oway.s)`, `(mcomp/code/turkey.contrasts2.sas)`,  
`(mcomp/figure/turkey.scheffe.eps.gz)`

that includes those displayed in Figures 7.7 and 7.8. There are  $10 = \binom{5}{2}$  pairwise differences among the means of the 5 diet combinations studied. Figure 7.5 is a mean–mean display of Scheffé simultaneous confidence intervals on these mean differences.

Figure 7.5 contains overprinting of labels for mean comparisons. In situations with such overprinting, we augment the mean–mean display with a traditional S-PLUS display of these same confidence intervals. This “tiebreaker” plot lists the contrasts in the same vertical order as in the mean–mean plot. The conclusions here, based on the fact that 9 of the 10 intervals lie entirely above zero, are

- For both amount 1 and 2, the mean weight gain from additive B is significantly greater than the mean weight gain from additive A.
- For both additive A or B, the mean weight gain from amount 2 significantly exceeds the mean weight gain from amount 1.
- The weight gain from the control diet is significantly below that from any of the other 4 diets.

We graphically summarize these conclusions with the orthogonal contrasts in Figures 7.7 and 7.8. The 3 contrasts that differ significantly from zero do not cross the vertical  $d = 0$  axis. The nonsignificant contrast does cross the  $d = 0$  axis.

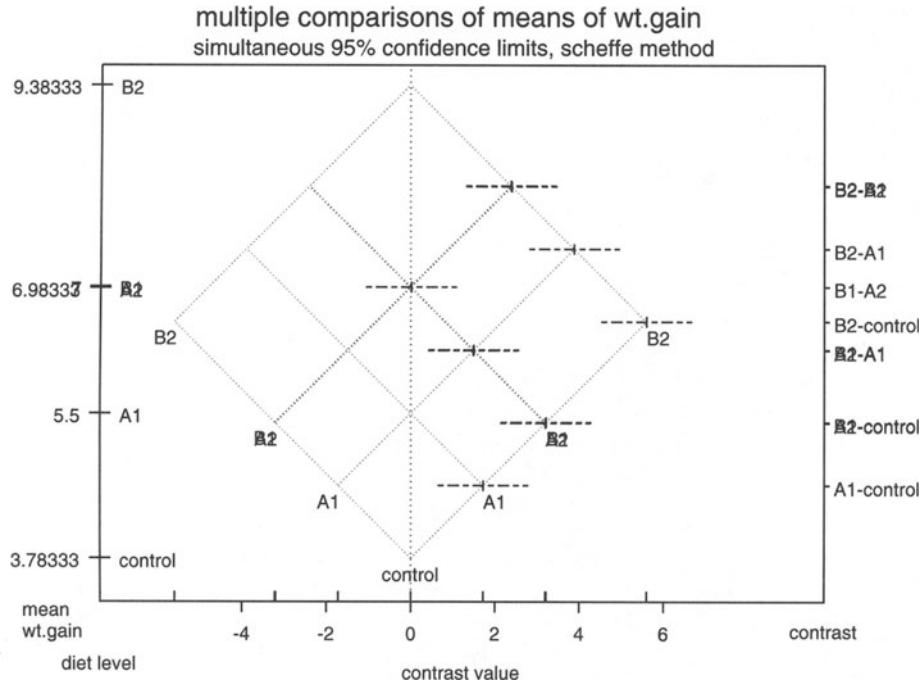


FIGURE 7.5. MMC: mca plot for Turkey data. Overprinting of contrasts at the same height are separated in Figure 7.6 by a standard multiple comparisons plot.  
(mcomp/code/turkey-mmcc.s), (mcomp/figure/turkey-mca.eps.gz)

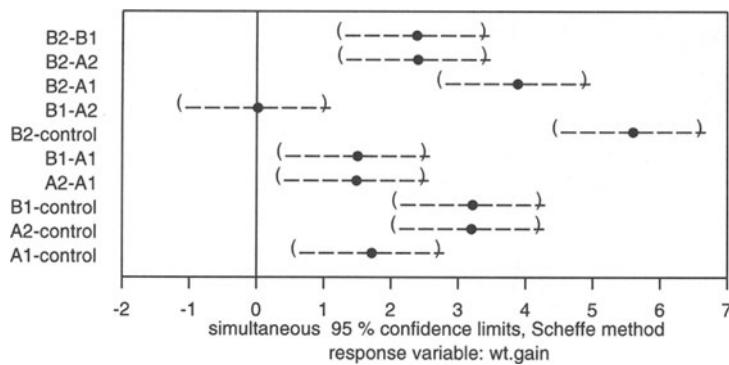


FIGURE 7.6. MMC: Tiebreaker multicomp plot for Turkey data. The overprinting in Figure 7.5 is resolved here with a standard multiple comparisons plot ordered to match the order of the MMC plot.  
(mcomp/code/turkey-mmcc.s), (mcomp/figure/turkey-mca2.eps.gz)

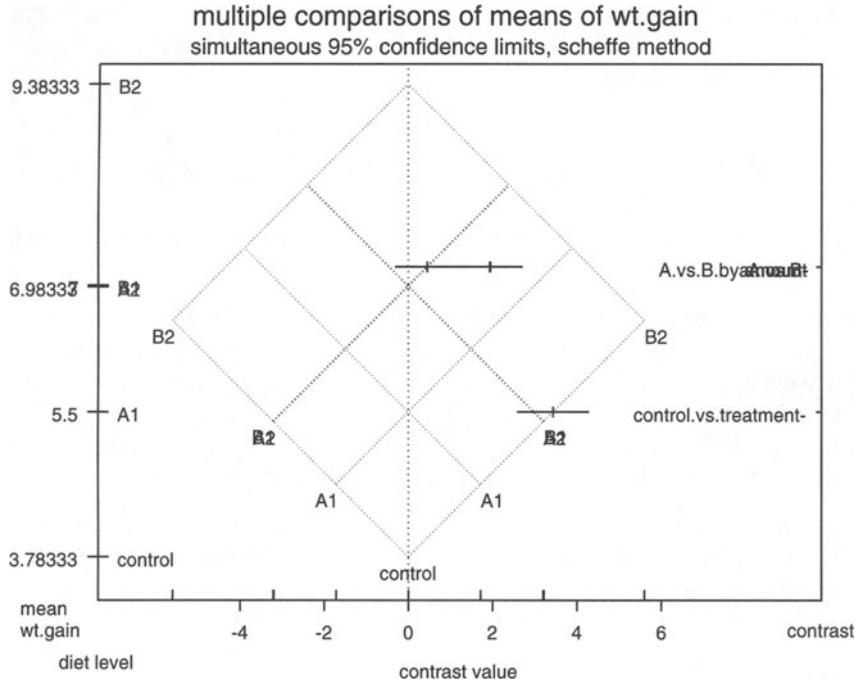
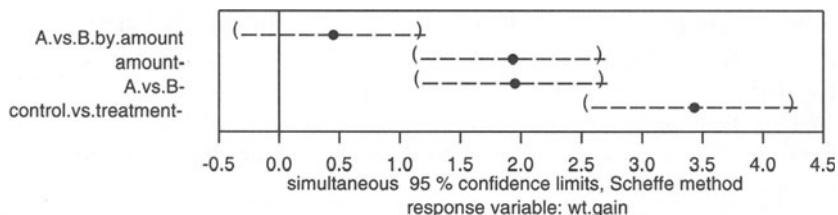


FIGURE 7.7. MMC: Orthogonal basis set of contrasts for Turkey data. Overprinting of contrasts at the same height are separated in Figure 7.8 by a standard multiple comparisons plot.  
`(mcomp/code/turkey-mmcc.s), (mcomp/figure/turkey-lmat.eps.gz)`



**FIGURE 7.8.** Tiebreaker `multicomp` plot for Turkey data. The overprinting in Figure 7.7 is resolved here with a standard multiple comparisons plot ordered to match the order of the MMC plot. These are the same contrasts that appear in Figure 7.4, but negative estimates have been reversed. During the reversal a “-” was appended to contrast names for which it was not possible to figure out how to reverse the contrast name.

#### 7.1.4.2 The Extended Tukey Procedure

The Tukey procedure can be extended to cover the family of all possible contrasts when the samples are of the same size  $n$ . Generalizing Equation (6.11) to any contrast vector  $(c_j)$  in the equal  $n$  case, we get

$$\text{CI} \left( \sum_{j=1}^a c_j \mu_j \right) = \sum_{j=1}^a c_j \bar{y}_j \pm \frac{q_\alpha}{2} \frac{s}{\sqrt{n}} \sum_{j=1}^a |c_j| \quad (7.2)$$

as the set of  $100(1 - \alpha)\%$  simultaneous confidence intervals for all possible contrasts among the population means.

The  $q_\alpha$  here is the same value used in Equation (6.11). Except for very simple contrasts, such as between pairs of means, these generalized Tukey intervals will be even wider than the analogous Scheffé intervals, (Hochberg and Tamhane, 1987). The generalized Tukey intervals (7.2) may be considered for use when interest lies in a family consisting of the union of all pairwise contrasts with a small number of more complicated contrasts.

As discussed in (Hochberg and Tamhane, 1987), the family encompassed by the generalized Tukey intervals also includes the set of individual intervals on each population mean,

$$\text{CI}(\mu_j) = \bar{y}_j \pm q_\alpha \frac{s}{\sqrt{n_j}} \quad (7.3)$$

These intervals are illustrated for the artificial data in Figure 7.10.

## 7.2 The Mean–Mean Multiple Comparisons Display (MMC Plot)

### 7.2.1 Difficulties with Standard Displays

The conclusions from the application of the Tukey procedure to the catalyst data are not well-conveyed by the standard tabular and graphical output provided by SAS and S-PLUS in Tables 6.4 and 6.5 and Figure 6.2. In all three displays the magnitudes of the sample means themselves are obscured. These displays have no capability to depict the relative distances between adjacent sorted sample means.

Another standard display of results of a Tukey test, shown here in Table 7.5, is often used to communicate results when sample sizes are equal. The sample means are listed in ascending magnitude. Straight-line segments are used to indicate significance according to the following rules. If two sample means are not covered by the same line segment, the corresponding population means are declared significantly different. If two sample means

TABLE 7.5. Underlining of means that are not significantly different. It is produced in SAS with the LINES option appended to calls to the Tukey procedure in either PROC GLM or PROC ANOVA. Compare to the means section at the bottom of Table 6.2.  
 (mcomp/transcript/catalystm.aov2a.lst)

---

SAS (mcomp/code/catalystm.aov2a.sas):

```
proc anova data=catalystm;
  class cat;
  model concent = cat;
  means cat ;
  means cat / tukey ;
  means cat / tukey lines;
run;
```

---

SAS (mcomp/transcript/catalystm.aov2ab.lst):  
 Means with the same letter are not significantly different.

| Tukey<br>Grouping | Mean   | N | cat |
|-------------------|--------|---|-----|
| A                 | 56.900 | 5 | A   |
| A                 |        |   |     |
| B A               | 55.775 | 4 | B   |
| B                 |        |   |     |
| B C               | 53.233 | 3 | C   |
| C                 |        |   |     |
| C                 | 51.125 | 4 | D   |

---

are covered by a common line segment, the corresponding population means are declared not significantly different.

With this procedure it is difficult to depict correctly the relative distances between adjacent sorted sample means because the table is constrained by the limited resolution of a fixed-width typewriter font rather than the high resolution of a graphical display.

Further, the procedure cannot be used when sample sizes are unequal. Figures 7.9 and 7.10 illustrate this limitation using artificial data:

| Group | N   | Mean |
|-------|-----|------|
| A     | 5   | 2.0  |
| B     | 100 | 2.1  |
| C     | 100 | 2.8  |
| D     | 5   | 3.0  |

The Tukey procedure uncovers a significant difference between the means of populations B and C, for which the sample sizes are large, but no significant difference between the means of populations A and D, from which the sample sizes are small. Using the `lines` option in SAS (we show a simulated plot in Table 7.6), the nonsignificant difference between the means of A and D requires that a common line covers the range from 2.0 to 3.0, including the location of the means of groups B and C. The presence of this line contradicts the existence of a significant difference between the means of groups B and C. The mean-mean display described in Section 7.2.2 and shown in Figure 7.11 overcomes these deficiencies.

TABLE 7.6. Underlining of means that are not significantly different. Simulated plot for artificial data. The single line, which is valid for the comparison of catalysts A and D (in this example based on the low precision test for small sample sizes), masks the significant difference between catalysts B and C (based on a much higher precision test for much larger sample sizes).

---

SAS (`mcomp/transcript/inconsist.tukey.lines.lst`):  
Means with the same letter are not significantly different.

---

| Tukey<br>Grouping | Mean  | N   | cat |
|-------------------|-------|-----|-----|
| A                 | 3.000 | 5   | D   |
| A                 |       |     |     |
| A                 | 2.800 | 100 | C   |
| A                 |       |     |     |
| A                 | 2.100 | 100 | B   |
| A                 |       |     |     |
| A                 | 2.000 | 5   | A   |

---

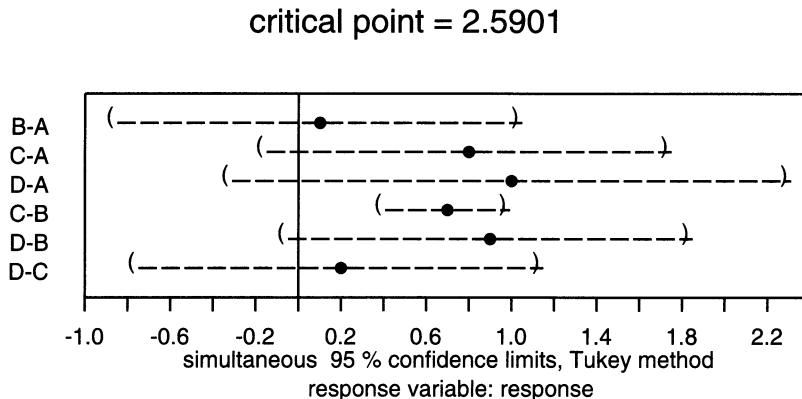


FIGURE 7.9. Simultaneous confidence intervals on all pairs of mean differences. The means of samples B and C both lie between the means of samples A and D. Sample sizes were 5 from populations B and C and 100 from populations A and D. The Tukey procedure finds a significant difference between the means of populations B and C but no significant difference between the means of populations A and D.

(mcomp/code/inconsistent.s), (mcomp/figure/unequal1.eps.gz)

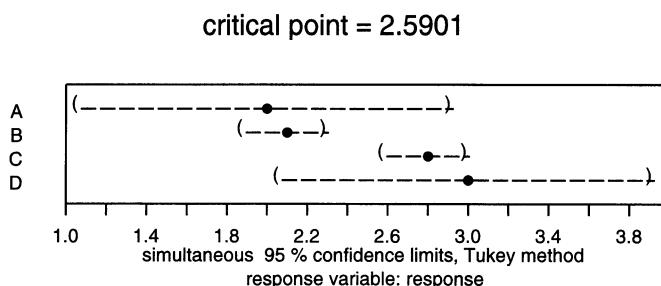


FIGURE 7.10. The means of samples B and C both lie between the means of samples A and D. Sample sizes were 5 from populations B and C and 100 from populations A and D. The Tukey procedure finds a significant difference between the means of populations B and C but no significant difference between the means of populations A and D. The underlying formula for these intervals appears in Equation (7.3).

(mcomp/code/inconsistent.s), (mcomp/figure/unequal2.eps.gz)

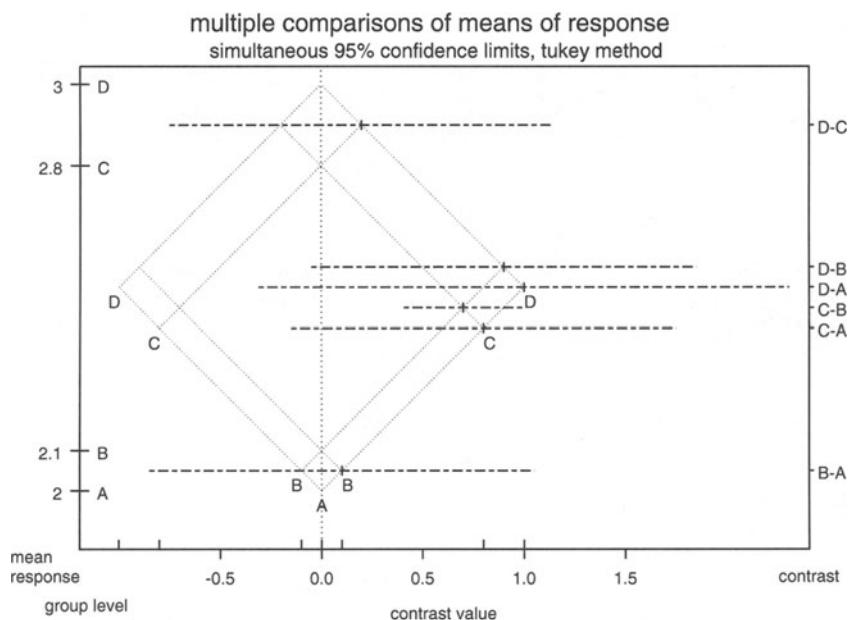


FIGURE 7.11. A mean–mean display (described in Section 7.2.2) of simultaneous confidences on the means from populations A, B, C, D in the artificial data. Each confidence interval on a mean difference is represented by a horizontal dashed line. If and only if a dashed line crosses the contrast value = 0 line, the corresponding population mean difference is declared nonsignificant. This display shows the relative differences between sample means and allows for unequal sample sizes.  
 (mcomp/code/inconsistent.s), (mcomp/figure/unequal3.eps.gz)

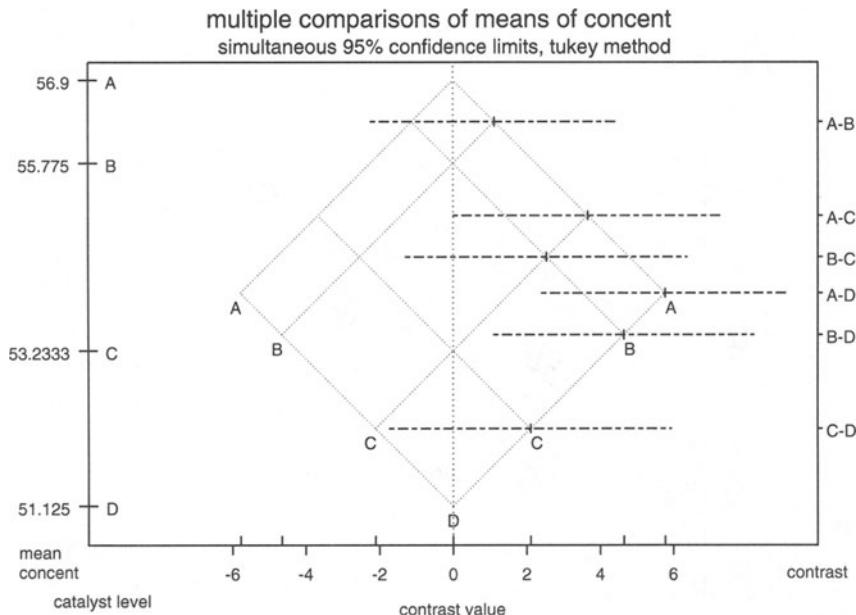


FIGURE 7.12. Multiple comparisons of all pairwise comparisons of catalyst means with the MMC display. This is a repeat of Figure 6.3.

(`mcomp/code/catalystm-mmcc3.s`), (`mcomp/figure/catalystm-mmcc-mca.eps.gz`)

## 7.2.2 Hsu and Peruggia's Mean–Mean Scatterplot

(Hsu and Peruggia, 1994) address the deficiencies in standard displays of multiple comparison procedures with their innovative graphical display of the Tukey procedure for all pairwise comparisons. In Section 7.2.2.1 we show the details of the construction of the MMC plot in Figure 7.12. We postpone interpretation of Figure 7.12 until Section 7.2.2.2.

In Section 7.2.3 we extend their display to show other multiple comparison procedures for arbitrary sets of contrasts. Software for our extension is in the S-PLUS functions in files (`splus.library/mmcc.multicomp.s`) and (`splus.library/multicomp.mmcc.s`).

### 7.2.2.1 Construction of the Mean–Mean Scatterplot

We begin with data-oriented orthogonal  $h$ - and  $v$ -axes in Figures 7.13 and 7.14 and then move to rotated difference ( $h - v$ ) and mean  $(h + v)/2$  axes in Figure 7.15. The rotations by  $45^\circ$  introduce factors of  $\sqrt{2}$  that are there to maintain the orthogonality of  $h$  and  $v$  in the rotated coordinates.

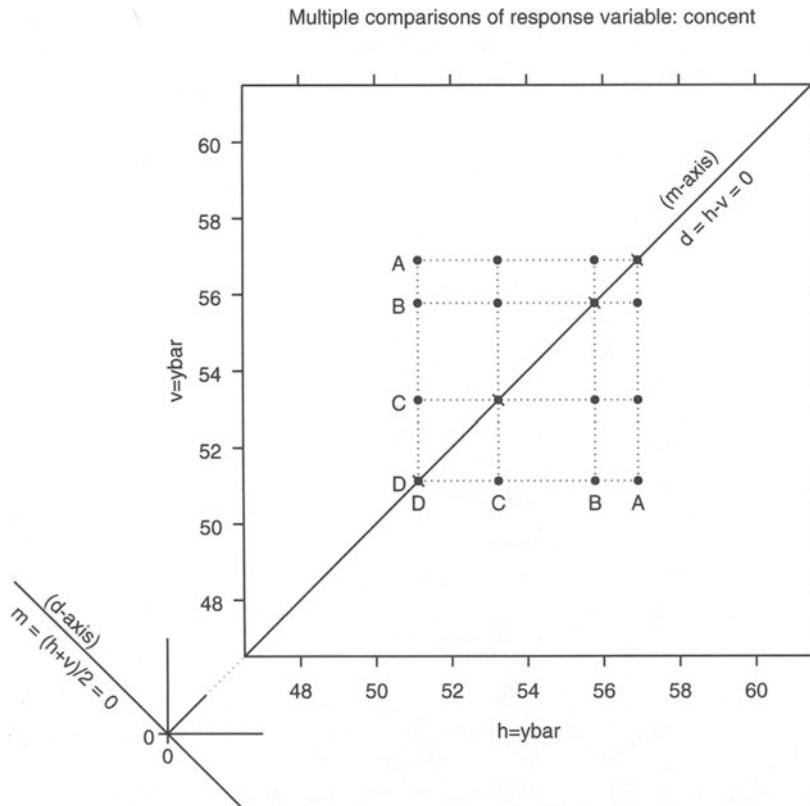


FIGURE 7.13. Construction of mean–mean multiple comparisons plot for the catalyst data.

Data-oriented axes, step 1. Please see the discussion in Section 7.2.

([mcomp/code/mmc.explain.s](#)), ([mcomp/figure/mmc1-a.eps.gz](#))

1. Draw a square plot in Figure 7.13 on  $(h, v)$ -axes. Define  $(d, m)$ -axes at  $\pm 45^\circ$ .
2. Plot each  $\bar{y}_i$  against  $\bar{y}_j$ .
3. Connect the points with  $h = \bar{y}_i$  and  $v = \bar{y}_j$  lines. The lines are labeled with the level names of the group means.
4. Draw the  $45^\circ$  line  $h = v$ . Define the value  $d = h - v$ , where the letter  $d$  indicates differences between group means. The line we just drew corresponds to  $d = 0$ . We will call this line the  $m$ -axis, where the name  $m = (h + v)/2$  indicates the group means.
5. Place tick marks on the  $m$ -axis at the points  $(\bar{y}_i, \bar{y}_i)$ .
6. Draw the  $-45^\circ$  line through the origin  $(h = 0, v = 0)$ . The line we just drew corresponds to  $m = 0$ . We will call this line the  $d$ -axis.

Multiple comparisons of response variable: concert

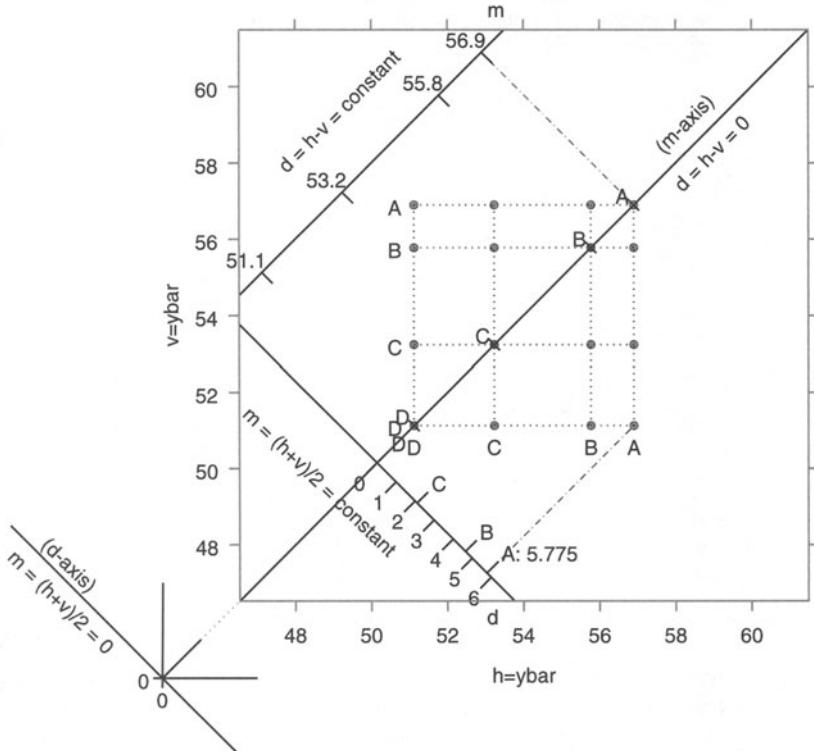


FIGURE 7.14. Construction of mean-mean multiple comparisons plot for the catalyst data. Data-oriented axes, step 2. Please see the discussion in Section 7.2.  
`(mcomp/code/mmc.explain.s), (mcomp/figure/mmc1-b.eps.gz)`

7. Copy Figure 7.13 as gray lines to Figure 7.14.
  8. Draw another  $m$ -axis parallel to the  $d = 0$  line. Drop a perpendicular from the  $(\bar{y}_A, \bar{y}_A)$  intersection on the  $d = 0$  line to the new  $m$ -axis. Place a tick at that point and label it with the  $m = \bar{y}_A$  value. Place similar tick marks at the heights  $m = \bar{y}_i$ . (The actual distances from the  $m = 0$  line to the tick marks are  $\bar{y}_i\sqrt{2}$ .)
  9. Draw another  $d$ -axis parallel to the line  $m = 0$ . Drop a perpendicular from the  $(\bar{y}_A, \bar{y}_D)$  intersection to the new  $d$ -axis. Place a tick at that point and label it with the level name  $A$  and the value  $\bar{y}_A - \bar{y}_D$ . Place similar ticks at the distances  $\bar{y}_i - \bar{y}_D$ . (The actual distances from the  $d = 0$  line to the tick marks are  $(\bar{y}_i - \bar{y}_D)/\sqrt{2}$ .) Place ticks below the  $d$ -axis at the distances  $(0, 1, 2, 3, 4, 5, 6)/\sqrt{2}$  and label them  $(0, 1, 2, 3, 4, 5, 6)$ .

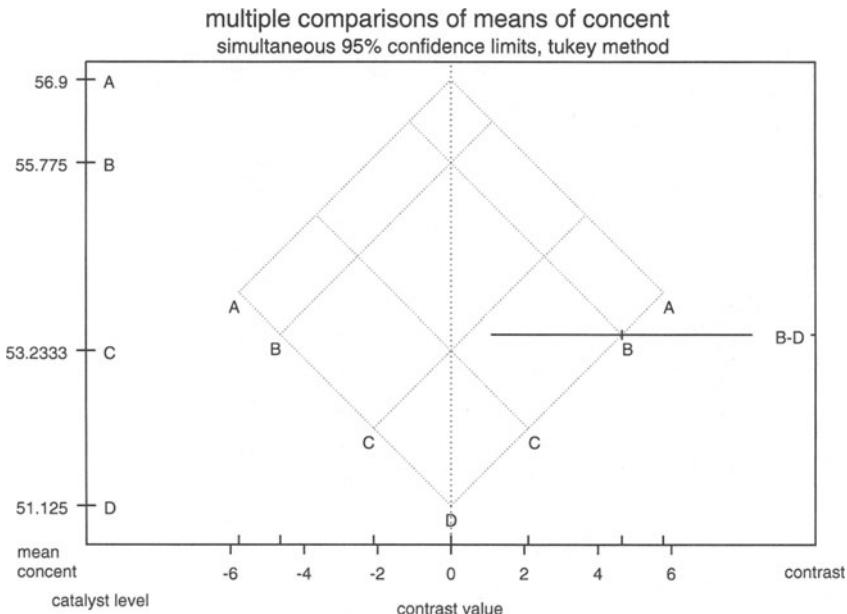


FIGURE 7.15. Construction of mean–mean multiple comparisons plot for the catalyst data. Difference and mean-oriented axes. This figure is essentially the same as Figure 7.13 rotated  $45^\circ$  counterclockwise. Please see the discussion in Section 7.2. This figure shows only one of the six pairwise contrasts. All six contrasts are shown in Figure 7.12.

(`mcomp/code/mmc.explain.s`), (`mcomp/figure/mmc2.eps.gz`)

10. Rotate Figure 7.14 counterclockwise by  $45^\circ$  to get Figure 7.15.
11. Construct the confidence intervals. We show just one pairwise interval, the one centered on the point  $(d = \bar{y}_B - \bar{y}_D, m = (\bar{y}_B + \bar{y}_D)/2)$ . The confidence interval line is parallel to the  $d$ -axis at a height equal to the average of the two observed means. The interval is on the  $d$ -scale and covers all points  $(\bar{y}_B - \bar{y}_D) \pm \hat{\sigma} q \sqrt{1/n_B + 1/n_D}$ , where  $\hat{\sigma}$  is the standard deviation from the ANOVA table and  $q$  is the critical value used for the comparison. In this example we use the critical value  $q = q_{0.05, 4, 12}/\sqrt{2} = 2.969141$  from the Studentized range distribution.
12. We show all  $\binom{4}{2} = 6$  pairwise differences  $\bar{y}_i - \bar{y}_j$  with their confidence intervals in Figure 7.12.

### 7.2.2.2 Interpretation of the Mean–Mean Scatterplot

We construct the background of Figures 7.15 and 7.12 by rotating Figure 7.14 counterclockwise by  $45^\circ$  and suppressing the  $h$ - and  $v$ -axes. The horizontal  $d$ -axis shows the values of the contrasts and the vertical  $m$ -axis shows the average values of the two means being contrasted.

In Figure 7.12, each mean pair  $(\bar{y}_i, \bar{y}_j)$  is plotted on the now-diagonal  $(h, v)$ -axes and can also be identified with its  $(d, m)$ -coordinates. In Figure 7.15, we focus on the pair of means  $\bar{y}_B$  and  $\bar{y}_D$ . We begin with the  $(h, v)$ -system and identify the point as

$$(h, v) = (\bar{y}_B, \bar{y}_D) = (55.8, 51.1)$$

The coordinates of the same pair of means  $(\bar{y}_B, \bar{y}_D)$  in the  $(d, m)$ -system are

$$\begin{aligned} (d, m) &= (\bar{y}_B - \bar{y}_D, (\bar{y}_B + \bar{y}_D)/2) \\ &= ((55.8 - 51.1), (55.8 + 51.1)/2) = (4.65, 53.45) \end{aligned}$$

We choose to label the ticks on the  $m$ -axis by the means because they are more easily interpreted: The confidence interval on  $\bar{y}_B - \bar{y}_D$  is at the mean height  $m = (\bar{y}_B + \bar{y}_D)/2$  in Figure 7.15. Hsu and Peruggia label the ticks on the  $m$ -axis by the sum  $\bar{y}_B + \bar{y}_D = 2m$  because one unit on the  $2m$ -scale takes exactly the same number of inches as one unit on the  $d$ -scale.

Figure 7.12 is constructed from Figure 7.15 by including all of the  $\binom{4}{2} = 6$  pairwise differences  $\bar{y}_i - \bar{y}_j$ , not just the single difference we use for the illustration.

Each of the confidence intervals for the  $\binom{4}{2} = 6$  pairwise differences  $\bar{y}_i - \bar{y}_j$  in Figure 7.12 is centered at a point whose height on the vertical  $m$ -axis is equal to the average of the corresponding means  $\bar{y}_i$  and  $\bar{y}_j$  and whose location along the horizontal  $d$ -axis is at distance  $\bar{y}_i - \bar{y}_j$  from the vertical line  $d = 0$ . Horizontal lines are drawn at these heights so that the midpoints of these lines intersect their  $(h = \bar{y}_i, v = \bar{y}_j)$  intersection. The width of each horizontal line is the width of a confidence interval estimating the difference  $\bar{y}_i - \bar{y}_j$ . By default the endpoints of the line are chosen to be the endpoints of the 95% two-sided confidence interval chosen by the Tukey procedure for all  $\binom{4}{2}$  possible pairs.

If a horizontal confidence interval line crosses the vertical  $d = 0$  line, the mean difference is declared not significant. Otherwise the mean difference is declared significant. If an end of a horizontal line is close to the vertical  $d = 0$ , this says that the declaration of significance was a close call.

When the critical value  $q$  is chosen by one of the standard multiple comparisons procedures (we illustrate with and default to the Tukey procedure),

the widths of the horizontal dashed confidence interval lines are the simultaneous confidence intervals for the six pairs of population mean differences. This depiction is not restricted to the case of equal sample sizes and hence equal interval widths.

The display in Figure 7.12 has several advantages over traditional presentations of Tukey procedure results. In a single graph we see

1. The means themselves, with correct relative distances,
2. The point and interval estimates of the  $\binom{4}{2}$  pairwise differences,
3. Declarations of significance,
4. Confidence interval widths that are correct when the sample sizes are unequal.

### 7.2.3 Extensions of the Mean–Mean Display to Arbitrary Contrasts

(Heiberger and Holland, 2004a) extend the mean–mean multiple comparisons plot to arbitrary contrasts, that is, contrasts that are not limited to the set of pairwise comparisons.

Two critical issues needed to be addressed. The first is the scaling of the contrast and the second is the set of contrasts selected for consideration.

#### 7.2.3.1 Scaling

The standard definition of a contrast in Equation (6.18) requires that it satisfy the zero-sum constraint Equation (6.16). The variance of the contrast is calculated with Equation (6.19).

When we calculate sums of squares and  $F$ -tests, this definition is sufficient. When we wish to plot arbitrary contrasts on the mean–mean multiple comparisons plot described in Section 7.2.2, the contrasts must be comparably scaled. The heights must be in the range of the observed  $\bar{y}_j$ , and all confidence intervals must fall inside the range of the  $d$ -axis. To satisfy this additional requirement, we need to require the absolute-sum-2 scaling introduced in Section 6.9.2.1 and made explicit in Equation (6.20). Any other scaling makes it impossible to fit these values on the mean–mean plot.

With the absolute-sum-2 scaling we can think of any contrast as the comparison of two weighted averages of  $\bar{y}_j$ . Let us call them  $\bar{y}_+ = \sum c_j^+ \bar{y}_j$  and  $\bar{y}_- = \sum c_j^- \bar{y}_j$ , where we use the superscript notation  $a^+ = \max(a, 0)$  and  $a^- = \max(-a, 0)$ . We illustrate with the contrast comparing the average of means  $\bar{y}_A$  and  $\bar{y}_B$  with the mean  $\bar{y}_D$ .

|                | <i>A</i>     | <i>B</i>     | <i>C</i> | <i>D</i>      | $\bar{y}_+$                 | $\bar{y}_-$ |
|----------------|--------------|--------------|----------|---------------|-----------------------------|-------------|
| absolute-sum-2 | .5           | .5           | 0        | -1            | $(\bar{y}_A + \bar{y}_B)/2$ | $\bar{y}_D$ |
| integer        | 1            | 1            | 0        | -2            |                             |             |
| normalized     | $1/\sqrt{6}$ | $1/\sqrt{6}$ | 0        | $-2/\sqrt{6}$ |                             |             |

We plot the contrast centered at the  $(h, v)$ -location  $(\bar{y}_-, \bar{y}_+)$ , where each term is at the correctly weighted average of the observed  $\bar{y}_j$ -values. The height on the  $m$ -axis of the MMC plot is  $(\bar{y}_+ + \bar{y}_-)/2$  and the difference on the  $d$ -axis is  $\bar{y}_+ - \bar{y}_-$ . The confidence interval widths are proportional to the standard error of  $\bar{y}_+ - \bar{y}_-$ , which, from (6.19), is proportional to  $\sqrt{\sum c_j^2/n_j}$ .

### 7.2.3.2 Contrasts

The simplest set of contrasts is the set of all pairwise comparisons  $\bar{y}_i - \bar{y}_j$  (as in Figure 6.3). Others sets include comparisons  $\bar{y}_j - \bar{y}_0$  of all treatment values to a control (as in Figure 7.3) and a basis set of orthogonal contrasts that span all possible contrasts (as in Figure 7.16).

### 7.2.3.3 Labeling

Our presentation of the MMC plot, for example in Figure 6.3, has improved labeling compared to the Hsu and Peruggia presentation.

The left-axis ticks are the  $\bar{y}_i$ -values themselves, at the heights of the intersections of the  $45^\circ$   $h$ - and  $v$ -lines with the vertical  $d = 0$  line. The labels on the outside of the left axis are the  $\bar{y}_i$ -values. The labels on the inside of the left axis are the names of the factor levels.

The right-axis labels belong to the horizontal CI lines for the contrasts. The labels outside the right axis are the automatically generated contrasts, either pairwise  $\bar{y}_i - \bar{y}_j$  or comparisons  $\bar{y}_j - \bar{y}_0$  of all treatment values to a control. The labels inside the right axis are the requested contrasts from the explicitly specified `lmat` matrix. Each CI line is at the height corresponding to the average of the two  $\bar{y}_*$  ( $(\bar{y}_i + \bar{y}_j)/2$  or  $(\bar{y}_+ + \bar{y}_-)/2$ ) values they are comparing. Each CI line is centered at the observed difference ( $(\bar{y}_i - \bar{y}_j)$  or  $(\bar{y}_+ - \bar{y}_-)$ ). The half-width of the (two-sided) CI line is  $qs_{\bar{y}_i - \bar{y}_j}$ , where  $q$  is calculated according to the specified multiple comparisons criterion.

The bottom axis is in the difference  $\bar{y}_i - \bar{y}_j$   $d$ -scale. The ticks and labels outside the bottom axis are regularly spaced values on the difference scale.

The ticks inside the bottom axis, at distances  $\pm |\bar{y}_j - \min_j \bar{y}_j|$ , correspond to the horizontal  $d$ -axis positions of the foot of the  $45^\circ$   $h$ - and  $v$ -lines. The names of the factor levels appear at the foot of each  $45^\circ$  line.

#### 7.2.3.4 $q$ Multipliers

Hypothesis test and confidence interval formulas, introduced in Chapter 3, depend on a multiple of the standard deviation. The multiplier is a quantile chosen from an appropriate distribution. When only one hypothesis is tested or only one interval is constructed, the multiplier is denoted  $z$  when the test statistic is normally distributed and  $t$  when the test statistic is from a  $t$  distribution. Multipliers denoted  $q$ , sometimes with a subscript, are used in many of this chapter's formulas for confidence intervals and rules for rejecting null hypotheses. In both Sections 7.1.2 and 7.1.4.2 discussing Tukey procedures, and in plots in Section 7.2 displaying results from these procedures,  $q$  refers to the Studentized range distribution. The multiplier used in the Dunnett procedure of Section 7.1.3 is a percentile of a marginal distribution of a multivariate  $t$  distribution. The multiplier for the Scheffé procedure is the square root of a percentile of an  $F$  distribution. For details, see (Hochberg and Tamhane, 1987).

### 7.2.4 Display of an Orthogonal Basis Set of Contrasts

The sum of squares associated with the factor  $A$  with  $a$  levels has  $a - 1$  degrees of freedom. The missing degree of freedom is associated with the grand mean and is normally suppressed from the ANOVA table.

In Section 6.8.1 we note that it is always possible to construct an orthogonal set of contrasts that decompose the  $a - 1$  df sum of squares for an effect into  $a - 1$  independent single-df sums of squares. In this section we illustrate the mathematics for constructing an orthogonal basis set by constructing one from the set of pairwise contrasts. From this basis set, we show that we can construct any other set of contrasts. We also show that an orthogonal basis set, augmented with an additional contrast for the grand mean (not actually a contrast since it doesn't sum to 0), can be used to construct any linear combination of the group means.

This discussion uses all the matrix algebra results summarized in Appendix Section F.4. This section is placed here in Chapter 7 because it belongs to the discussion of the MMC plots. It might be more easily read after Section 10.3.

We illustrate the discussion with the (`datasets/catalystm.dat`). We begin with the set of pairwise contrasts in Figure 7.12. Figure 7.16 illustrates an orthogonal basis set of contrasts for the catalyst data. This examination

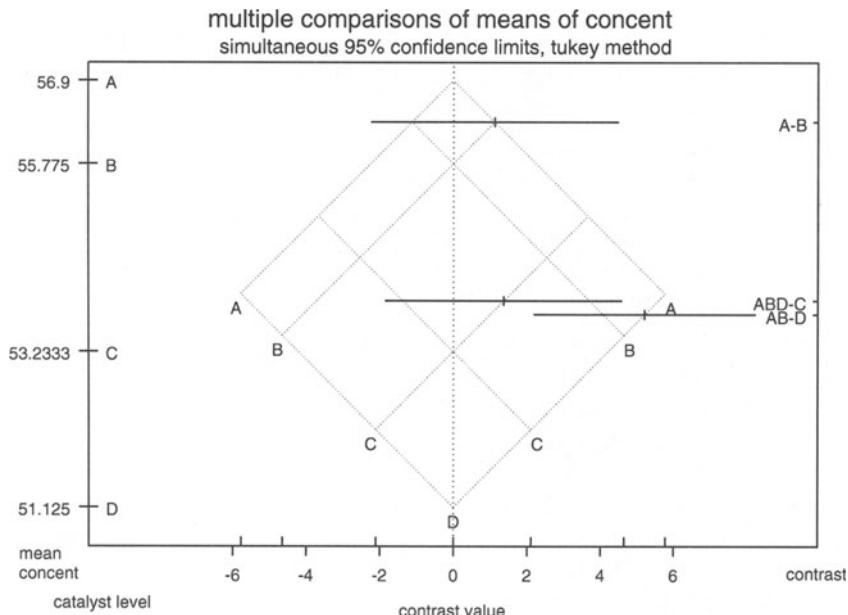


FIGURE 7.16. An orthogonal set of contrasts based on the pairwise set in Figures 7.12 and 6.3. The comparison between the average of  $\bar{y}_A$  and  $\bar{y}_B$  with the mean  $\bar{y}_D$  is the only significant comparison. The other two confidence intervals include 0.

(`mcomp/code/catalystm-mmc3.s`), (`mcomp/figure/catalystm-mmc-lmat.eps.gz`)

of 3 linearly independent contrasts succinctly summarizes the information contained in the 3 degrees of freedom for comparing the means of the 4 levels of the fixed factor `catalyst`. The principal conclusion from Figure 7.12 is that the means of both catalysts A and B significantly exceed the mean of catalyst D. Figure 7.16 reforces this conclusion with the finding that the average of the means of catalysts A and B significantly exceeds the mean of catalyst D because the confidence interval for this contrast lies entirely above 0. A second new conclusion from Figure 7.16 is that the average of the means of catalysts A, B, and D is not significantly different from the mean of catalyst C because the confidence interval for this contrast includes 0.

### 7.2.5 Hsu and Peruggia's Pulmonary Example

This is the example that (Hsu and Peruggia, 1994) use to introduce the mean-mean multiple comparisons plots. The response variable is FVC, forced vital capacity.

|    |  |
|----|--|
| NS | nonsmokers   |
| PS | passive smokers  |
| NI | noninhaling smokers  |
| LS | light smokers (1–10 cigarettes per day for at least the last 20 years)       |
| MS | moderate smokers (11–39 cigarettes per day for at least the last 20 years)   |
| HS | heavy smokers ( $\geq 40$ cigarettes per day for at least the last 20 years) |

There are six levels of the **smoker** factor, hence 5 df for comparing them. Figure 7.17 shows that the three levels {PS, NI, LS} are indistinguishable; we call this the low-smoker cluster. This comparison of three levels uses 2 df. There are only 3 df left. From the SW–NE HS line, we see that the MS–HS contrast is significant, that the comparisons between each of the three levels in the low-smoker cluster with MS is significant, and that the comparison of NS with HS and with MS are each significant. All three comparisons of NS with the low-smoker cluster have lower bounds close to zero, and one of the three comparisons is significant.

We can summarize these visual impressions by constructing an orthogonal set of contrasts that reflect them exactly. Figure 7.18 shows a basis set of five orthogonal contrasts. In the center, the  $p-n_1$  and  $n-1$  contrasts show that the three levels in the low-smoker cluster are indistinguishable. The other three lines show that the nonsmoker group is significantly different from the low-smoker cluster ( $n-pn_1$ ), that the moderate- and heavy-smoker groups are significantly different ( $m-h$ ), and that the combined nonsmoker group and low-smoker cluster are significantly different from the combined moderate- and heavy-smoker groups ( $npn_1-mh$ ).

The center of the interval for each of the contrasts in Figure 7.18 is constructed by the linear combination of the means for the levels. For example, the  $n-pn_1$  interval is on the NW–SE NS line and on the average of the NE–SW PS, NI, and LS lines. The width of the interval is calculated from the algebra of the contrast. A simultaneous 95% coverage probability applies to the five confidence intervals in Figure 7.18 because they are constructed using the extended Tukey procedure. This procedure guarantees the coverage probability over the set of all possible contrasts. In exchange for this guarantee, these extended Tukey intervals are fairly wide. Having used the Tukey procedure to construct the intervals in Figure 7.17, it would be incorrect to switch to the narrower Scheffé procedure simultaneous intervals for the basis set of contrasts. With such a switch we would have two competing analyses, and this would distort the claimed coverage probabilities for the now distinct analyses in the two figures.

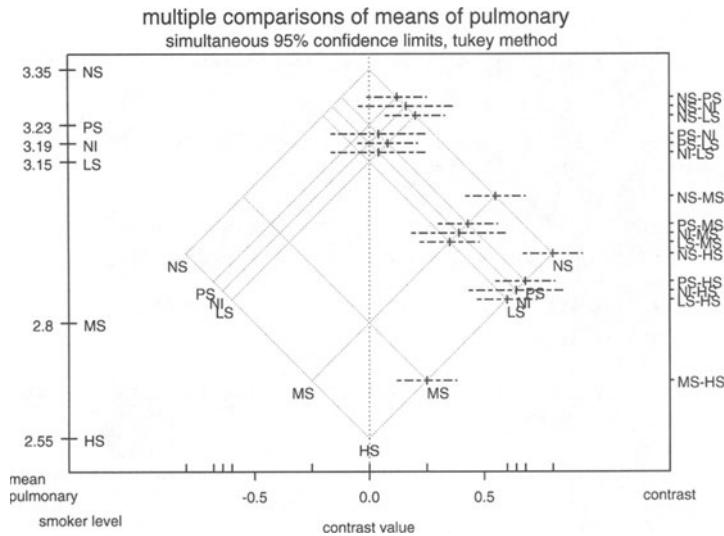


FIGURE 7.17. Hsu and Peruggia's pulmonary example.  
(`mcomp/code/pulmonary.s`), (`mcomp/figure/pulmonary-mmc-mca.eps.gz`)

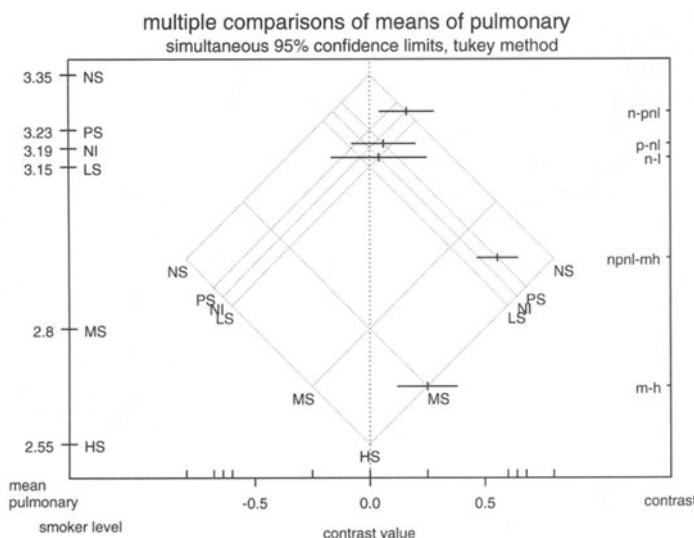


FIGURE 7.18. Hsu and Peruggia's pulmonary example: An orthogonal set of contrasts. The ability to display an arbitrary orthogonal set of contrasts is one of our enhancements to the mean-mean plot.  
(`mcomp/code/pulmonary.s`), (`mcomp/figure/pulmonary-mmc-lmat.eps.gz`)

### 7.3 Exercises

- 7.1.** Use an MMC plot to display the results of the Tukey procedure in Exercise 6.2.
- 7.2.** Use an MMC plot to display the results of the Tukey procedure in Exercise 6.4.
- 7.3.** Use an MMC plot to display the results of the Tukey procedure applied to the log-transformed data discussed in Exercise 6.5.
- 7.4.** Use an MMC plot to display the results of the Tukey procedure in Exercise 6.6.
- 7.5.** Use an MMC plot to display the results of the Tukey procedure in Exercise 6.8.
- 7.6.** Use an MMC plot to display the results of the Tukey procedure in Exercise 6.9.
- 7.7.** The relative rotation angle between tangents to cervical vertebrae C3 and C4 is a standard musculoskeletal measurement. Figure 7.19 illustrates the measurement of relative rotation angles. (Harrison et al., 2004) hypothesize that the value of this angle, C3–C4, in persons complaining of neck pain tends to differ from that in healthy individuals. The file (`datasets/c3c4.dat`) contains the C3–C4 measurements of a random sample of 194 patients of which 72 had no complaints of neck pain, 52 complained of acute neck pain of recent origin, and 70 have had chronic neck pain. The pain **condition** is coded 0 for none, 1 for acute, and 2 for chronic. There is no implied ordering in this coding scheme. Perform an analysis of variance followed by Dunnett's procedure to determine if the mean C3–C4 value of persons with acute or chronic neck pain differs from the mean C3–C4 value of persons without neck pain.

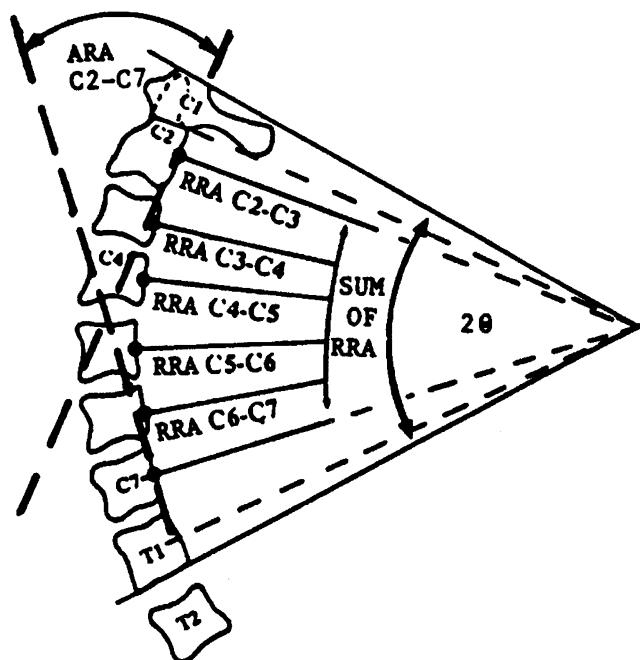


FIGURE 7.19. Illustration of the relative rotation angles between the cervical vertebrae (neck area). Exercise 7.7 uses the C3–C4 angle.  
(mcomp/figure/RRA.ps.gz)

# Linear Regression by Least Squares

## 8.1 Introduction

We usually study more than one variable at a time. When the variables are continuous, and one is clearly a response variable and the others are predictor variables, we usually plot the variables and then attempt to fit a model to the plotted points. With one continuous predictor, the first model we attempt is a straight line; with two or more continuous predictors, we attempt a plane. We plot the model, the residuals from the model, and various diagnostics of the quality of the fit.

In this chapter we are primarily concerned with modeling a straight-line relationship between two variables using  $n$  pairs of observations on these variables, a common and fundamental task. One of these variables, conventionally denoted  $y$ , is a response or output variable. The other variable, often denoted  $x$ , is known as an explanatory or input or predictor variable. Usually, but not always, it is clear from the context which of the two variables is the response and which is the predictor. For example, if the two variables are personal **income** and **consumption** spending, then **consumption** is the response variable because the amount that is spent depends on how much **income** is available to be spent.

The relationship between  $y$  and  $x$  is almost never perfectly linear. When the  $n$  points are plotted in two dimensions, they appear as a random scatter about some unknown straight line. We model this line as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{for } i = 1, \dots, n \tag{8.1}$$

where

$$\epsilon_i \sim N(0, \sigma^2) \quad (8.2)$$

that is, the  $\epsilon_i$  are assumed normally independently distributed with constant mean 0 and common variance  $\sigma^2$  [abbreviated as  $\epsilon_i \sim NID(0, \sigma^2)$ ]. In other words, we assume that the response variable is linearly related to the predictor variables, plus a normally distributed random component. Here the intercept  $\beta_0$  and slope  $\beta_1$  are unknown *regression coefficients* that must be estimated from the data. The variance  $\sigma^2$  is a third unknown parameter, introduced along with the assumption of a normally distributed error term, which must also be estimated.

A commonly used procedure for estimating  $\beta_0$  and  $\beta_1$  is the method of *least squares* because, as we will see in Section 8.3.2, this mathematical criterion leads to simple “closed-form” formulas for the estimates. Under the stated normality assumptions in Equation (8.2) about the residuals  $\epsilon_i$  of Model (8.1), the least-squares estimates of the regression coefficients are also the maximum likelihood estimates of these coefficients.

## 8.2 Example—Body Fat Data

### Study Objectives

The example is taken from (Johnson, 1996). A group of subjects is gathered, and various body measurements and an accurate estimate of the percentage of body fat are recorded for each. Then body fat can be fit to the other body measurements using multiple regression, giving, we hope, a useful predictive equation for people similar to the subjects. The various measurements other than body fat recorded on the subjects are, implicitly, ones that are easy to obtain and serve as proxies for body fat, which is not so easily obtained.

Percentage of body fat, age, weight, height, and ten body circumference measurements (e.g., abdomen) are recorded for 252 men. Body fat, a measure of health, is estimated through an underwater weighing technique. Fitting body fat to the other measurements using multiple regression provides a convenient way of estimating body fat for men using only a scale and a measuring tape.

### Data Description

We will initially use only 47 observations and only five of the measurements that have been recorded.

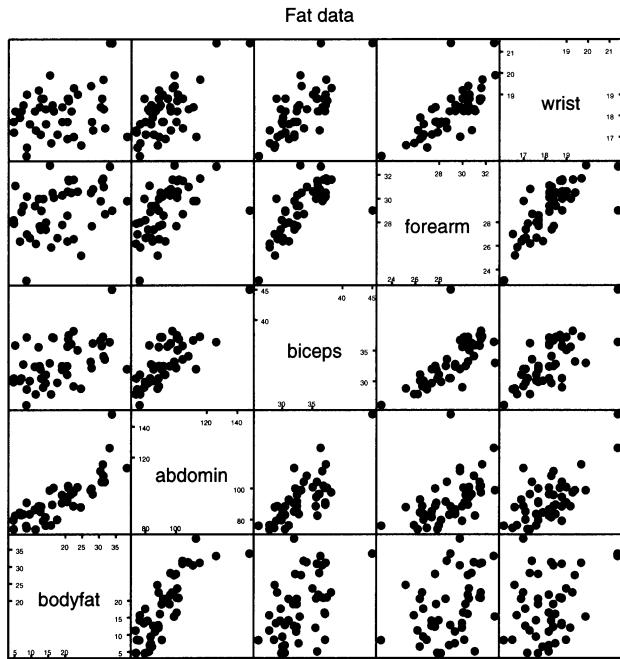


FIGURE 8.1. Body Fat Data  
 (rega/code/rega.f1.s), (rega/code/rega.f1.sas),  
 (rega/figure/f1.EPS.gz)

**bodyfat:** Percent body fat using Siri's equation,  $495/\text{Density} - 450$

**abdomin:** Abdomen circumference (cm) "at the umbilicus and level with the iliac crest"

**biceps:** Extended biceps circumference (cm)

**wrist:** Wrist circumference (cm) "distal to the styloid processes"

**forearm:** Forearm circumference (cm)

## Data Input

We read the data from (`datasets/fat.data`) into S-PLUS or SAS with (`rega/code/readfat.s`) or (`rega/code/readfat.sas`) and then look at the data with the scatterplot matrix in Figure 8.1.

The response variable **bodyfat** is in the bottom row of the plot. We can see that a linear fit makes sense against **abdomin**. A linear relationship between

`bodyfat` and the other predictor variables is also visible in the plot, but is weaker. All the predictor variables show correlation with each other.

## One-X Analysis

The initial analysis will look at just `bodyfat` and `abdomin`. We will come back to the other variables later. We expand the `bodyfat` by `abdomin` panel of Figure 8.1 in the left column of Figure 8.2 and place two straight lines on the graph in the two rightmost columns. The line in column 3 is visibly not a good fit. It is too shallow and is far above the points in the lower left. The line in column 2, labeled “least-squares fit”, is just right. The criterion we use is *least squares*, which means that the sum of the squared differences from the fitted to observed points is to be minimized. The *least-squares* line is the straight line that achieves the minimum.

The top row of Figure 8.2 displays the vertical differences from the fitted to observed points. The bottom row displays the squares of the differences from the fitted to observed points. The least-squares line minimizes the sum of the areas of these squares. It is evident that the sum of the squared areas in column 2 is smaller than the sum of squared areas for the badly fitting line in column 3.

From any of these panels it is apparent that on average, body fat is directly related to abdominal circumference. As will be explained in Section 8.3.5, the least-squares line in Figure 8.2 can be used to predict `bodyfat` from `abdomin`. Note that although it is mathematically correct to say that `abdomin` increases with `bodyfat`, this is a misleading statement because it implies an unlikely direction of causality among these variables.

## 8.3 Simple Linear Regression

### 8.3.1 Algebra

Figure 8.2 illustrates the least-squares line that best fits `bodyfat` to `abdomin`. Now that we see from the bottom row of the figure that the least-squares line actually does minimize the sum of squares, let us review the mathematics behind the calculation of the least-squares line. The standard notation we use for the least-squares straight line is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{8.3}$$

where  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are called the *regression coefficients*. We define the residuals by

$$e_i = y_i - \hat{y}_i \tag{8.4}$$

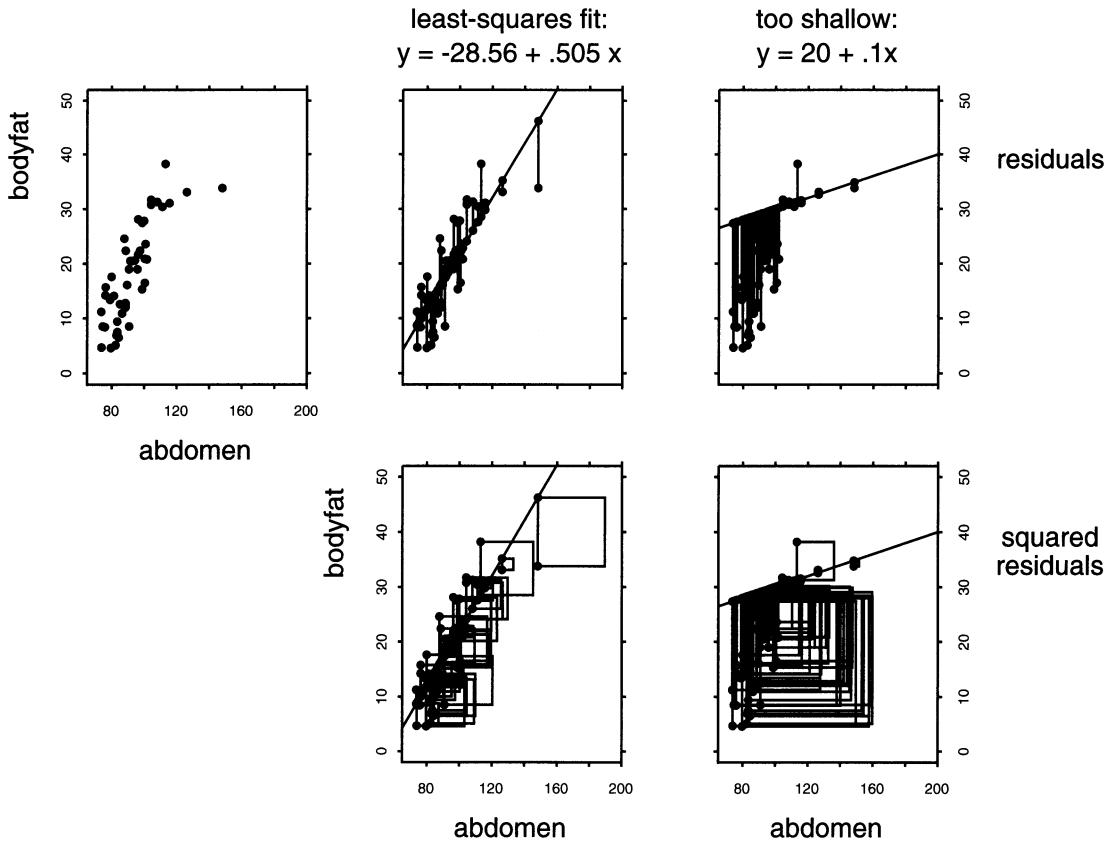


FIGURE 8.2. One  $X$ -variable and two straight lines. The second column is the least-squares line, the third is too shallow. Row 1 shows the residuals. Row 2 shows the squared residuals. The least-squares line minimizes the sum of the squared residuals.

(rega/code/residuals.s), (rega/figure/resid2x2.eps.gz)

We wish to find  $\hat{\beta}_0$  and  $\hat{\beta}_1$  that minimize the expression for the sum of squares of the calculated residuals:

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \quad (8.5)$$

We minimize by differentiation with respect to the parameters  $\beta_0$  and  $\beta_1$ , setting the derivatives to 0 (thus getting what are called the *normal*

equations)

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 = \sum_{i=1}^n 2(y_i - (\beta_0 + \beta_1 x_i))(-1) = 0 \quad (8.6)$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 = \sum_{i=1}^n 2(y_i - (\beta_0 + \beta_1 x_i))(-x_i) = 0$$

and then solving simultaneously for the regression coefficients

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned} \quad (8.7)$$

In addition to minimizing the sum of squares of the calculated residuals,  $\hat{\beta}_0$  and  $\hat{\beta}_1$  have the property that the sum of the calculated residuals is zero, i.e.,

$$\sum_{i=1}^n e_i = 0 \quad (8.8)$$

We request a proof of this assertion in Exercise 8.9.

For two or more predictor variables, the procedure (equating derivatives to zero) is identical but the algebra is more complex. We postpone details until Section 9.2.

### 8.3.2 Normal Distribution Theory

Under the normality assumption (8.2) for the residuals of Model (8.1), the least-squares estimates are also maximum likelihood estimates. This is true because if the residuals are normally distributed, their likelihood function is maximized when Equation (8.5) is minimized.

In Model (8.1), the unknown population variance of the  $\epsilon_i$ ,  $\sigma^2$ , is estimated by the sample variance

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2} \quad (8.9)$$

Because the sample variance is proportional to the residual sum of squares in Equation (8.5), minimizing the sample variance also leads us to the least-squares estimates  $\hat{\beta}_0$  and  $\hat{\beta}_1$  in Equations (8.7). The square root  $s$  of the sample variance in Equation (8.9), variously termed the *standard error of estimate*, the *standard error*, or the *root mean square error*, indicates the

size of a typical vertical deviation of a point from the calculated regression line.

### 8.3.3 Calculations

The results of the statistical analysis are displayed in several tables, primarily the *ANOVA* (analysis of variance) table, the table of regression coefficients, and the table of other statistics shown in Table 8.1. These tables are fundamental to our interpretation of the analysis. The formulas for each number in these tables appear in Tables 8.2, 8.3, and 8.4.

For one- $x$  regression (this example), there is usually only one null and alternative hypothesis of interest:

$$H_0: \beta_1 = 0 \quad \text{vs} \quad H_1: \beta_1 \neq 0 \quad (8.10)$$

Both  $t = 9.297$  in the table of coefficients and  $F = 86.427 = 9.297^2 = t^2$  in the ANOVA table are tests between those hypotheses. The associated  $p$ -value ( $p = .510^{-12}$ , which we report as  $< 0.0001$ ) is smaller than any reasonable  $\alpha$  (the traditional .05 or .01, for example). Therefore, we are justified in rejecting the null hypothesis in favor of the alternative. Inference on  $\beta_0$  frequently makes no sense. In this example, for example,  $\beta_0$  is the expected **bodyfat** of an individual having the impossible **abdomin** with zero circumference.

The **Total** line in the ANOVA table shows the sum of squares and degrees of freedom for the response variable **bodyfat** around its mean. When we divide these two numbers we recognize the formula  $\sum_{i=1}^n (y_i - \bar{y})^2 / (n-1) = 80.678$  as Equation (3.6) for the sample variance of the response variable. The goal of the analysis is to *explain* as much of the variance in the response variable as possible with a model that relates the response to the predictors. When we have explained the variance, the *residual* (or leftover) mean square  $s^2$  is much smaller than the sample variance of the response variable.

The *coefficient of determination*, also known as *Multiple R*<sup>2</sup>, usually accompanies ANOVA tables. This measure, generally denoted  $R^2$ , is the proportion of variation in the response variable that is accounted for by the predictor variable(s). It is desirable that  $R^2$  be as close to 1 as possible. Models with  $R^2$  considerably below 1 may be acceptable in some disciplines. The defining formula for  $R^2$  is

$$R^2 = \frac{SS_{\text{Reg}}}{SS_{\text{Total}}} \quad (8.11)$$

In regression models with only one predictor, an alternative notation is  $r^2$ . This notation is motivated by the fact that  $r^2$  is the square of the sample

TABLE 8.1. ANOVA table and table of regression coefficients for the simple linear regression model with  $y=\text{bodyfat}$  and  $x=\text{abdomin}$ . The tables were typeset from the program output.  
 (rega/transcript/fat.lm.st), (rega/transcript/fat.lst)

```
S-PLUS (rega/code/ls.s):
  ## least-squares fit
  fat.lm <- lm(bodyfat ~ abdomen, data=fat)
  summary(fat.lm, corr=F)
  anova(fat.lm)
```

```
SAS (rega/code/ls.sas):
proc reg data = fat;
    model bodyfat = abdomen ;
run;
```

#### ANOVA Table

| Source    | df | Sum of Sq | Mean Sq  | F-Value | Pr(> F)  |
|-----------|----|-----------|----------|---------|----------|
| abdomin   | 1  | 2440.500  | 2440.500 | 86.427  | < 0.0001 |
| Residuals | 45 | 1270.699  | 28.238   |         |          |
| Total     | 46 | 3711.199  |          |         |          |

#### Table of Regression Coefficients

| Predictor   | Value   | Std. Error | t-value | Pr(>  t ) |
|-------------|---------|------------|---------|-----------|
| (Intercept) | -28.560 | 5.110      | -5.589  | < 0.0001  |
| abdomin     | 0.505   | 0.054      | 9.297   | < 0.0001  |

#### Other Statistics

| Statistic                | Value   |
|--------------------------|---------|
| Multiple $R^2$           | 0.6576  |
| Adjusted $R^2$           | 0.6500  |
| Dependent Mean           | 18.3957 |
| Residual Standard Error  | 5.3139  |
| Coefficient of Variation | 28.8867 |

correlation coefficient  $r$  between the response and predictor variable.  $r$  is the usual estimate of the population correlation coefficient defined and interpreted in Equation 3.12. A formula for the sample correlation  $r$  is

$$r = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum(y_i - \bar{y})^2} \sum(x_i - \bar{x})} \quad (8.12)$$

It can be shown that  $-1 \leq r \leq 1$ . If  $r = \pm 1$ , then  $x$  and  $y$  are perfectly linearly related, directly so if  $r = 1$  and inversely so if  $r = -1$ . The arithmetic sign of  $r$  matches the arithmetic sign of  $\hat{\beta}_1$ .

TABLE 8.2. Interpretation of items in “ANOVA Table” from Table 8.1. The symbols in the **abdomin** section are subscripted **Reg**, short for “Regression”. In this setting, “Regression” refers to the group of all model predictors. In this example there is only one predictor, **abdomin**.

| Name               | Notation            | Formula  | Value in Table 8.1 |
|--------------------|---------------------|--|--------------------|
| <b>Total</b>       |                     |  |                    |
| Sum of Squares     | $SS_{\text{Total}}$ | $\sum_{i=1}^n (y_i - \bar{y})^2$   | 3711.199           |
| Degrees of Freedom | $df_{\text{Total}}$ | $n - 2$  | 46                 |
| <b>Residual</b>    |                     |  |                    |
| Sum of Squares     | $SS_{\text{Res}}$   | $\sum_{i=1}^n (y_i - \hat{y}_i)^2$   | 1270.699           |
| Degrees of Freedom | $df_{\text{Res}}$   | $n - 2$  | 45                 |
| Mean Square        | $MS_{\text{Res}}$   | $\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}$                  | 28.238             |
| <b>abdomin</b>     |                     |  |                    |
| Sum of Squares     | $SS_{\text{Reg}}$   | $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$   | 2440.500           |
| Degrees of Freedom | $df_{\text{Reg}}$   | number of predictor variables  | 1                  |
| Mean Square        | $MS_{\text{Reg}}$   | variability in $\hat{y}$ attributable to $\hat{\beta}_1$                                 |                    |
|                    |                     | $\left( \frac{\text{abdomin Sum of Squares}}{\text{abdomin Degrees of Freedom}} \right)$ | 2440.500           |
| $F$ -Value         | $F_{\text{Reg}}$    | $\left( \frac{\text{abdomin Mean Square}}{\text{Residual Mean Square}} \right)$          | 86.427             |
| $\Pr(> F)$         | $p_{\text{Reg}}$    | $P(F_{1,45} > 86.427) = 1 - \mathcal{F}_{1,45}(86.427)$                                  | < 0.0001           |

TABLE 8.3. Interpretation of items in “Table of Regression Coefficients” from Table 8.1.

| Name           | Notation                 | Formula  | Value in Table 8.1 |
|----------------|--------------------------|--|--------------------|
| (Intercept)    |                          |  |                    |
| Value          | $\hat{\beta}_0$          | $\bar{y} - \hat{\beta}_1 \bar{x}$  | -28.560            |
| Standard Error | $\hat{\sigma}_{\beta_0}$ | $\sigma \sqrt{\frac{1}{n} + \sum_{(x_i - \bar{x})^2}}$                   | 5.110              |
| t-value        | $t_{\beta_0}$            | $\hat{\beta}_0 / \hat{\sigma}_{\beta_0}$                                 | -5.589             |
| $\Pr(>  t )$   | $p_{\beta_0}$            | $P(t_{45} >  -5.589 )$   | < 0.0001           |
| abdomin        |                          |  |                    |
| Value          | $\hat{\beta}_1$          | $\frac{\sum_{(y_i - \bar{y})(x_i - \bar{x})}}{\sum_{(x_i - \bar{x})^2}}$ | 0.505              |
| Standard Error | $\hat{\sigma}_{\beta_1}$ | $\hat{\sigma} / \sqrt{\sum_{(x_i - \bar{x})^2}}$                         | 0.054              |
| t-value        | $t_{\beta_1}$            | $\hat{\beta}_1 / \hat{\sigma}_{\beta_1}$                                 | 9.297              |
| $\Pr(>  t )$   | $p_{\beta_1}$            | $P(t_{45} >  9.297 )$  | < 0.0001           |

TABLE 8.4. Interpretation of items in table of “Other Statistics” from Table 8.1.

| Name   | Notation           | Formula  | Value in Table 8.1 |
|--|--------------------|--|--------------------|
| Coefficient of Determination<br>Multiple $R^2$ | $R^2$              | $\left( \frac{\text{abdomin Sum of Squares}}{\text{Total Sum of Squares}} \right)$ | 0.6576             |
| Adjusted $R^2$                                 | $R_{\text{adj}}^2$ | $1 - \left( \frac{n-1}{n-p} \right) (1 - R^2)$                                     | 0.6500             |
| Dependent Mean                                 | $\bar{Y}$          | $\frac{\sum Y_i}{n}$   | 18.3957            |
| Residual Standard Error                        | $\hat{\sigma} = s$ | $\sqrt{s^2}$   | 5.3139             |
| Coefficient of Variation                       | $cv$               | $s/\bar{Y}$  | 28.8867            |

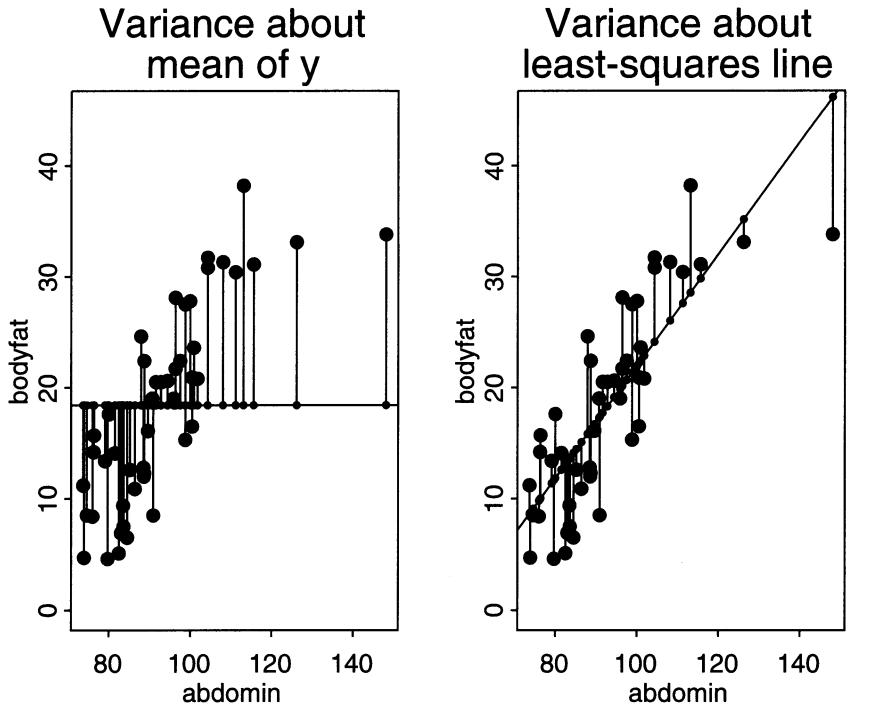


FIGURE 8.3. Variance about mean and about least-squares line.

(rega/code/rega.f5.s), (rega/code/rega.f5.sas),  
 (rega/figure/f5.EPS.gz)

In the present body fat example, we find  $r = 0.811$  and  $r^2 = 0.658$ . This value of  $r$  is consistent with the moderately strong positive linear relationship between **bodyfat** and **abdomin** in the least-squares fit shown in Figure 8.2. Continuing with this example, the estimated response variance ignoring the predictor is 80.678 and the estimated response variance paying attention to the predictor **abdomin**, the **Residuals Mean Square**, is 28.238. Graphically, we see in Figure 8.3 that the variance estimate 80.678 about the mean belongs to Figure 8.3a and the variance estimate about the regression line belongs to Figure 8.3b.

While these two estimates of response variance are intuitive, they are not actually the statistically correct numbers to compare because they are not independent. The **Total Sum of Squares** is the sum of the **Residuals Sum of Squares** and the **abdomin Sum of Squares**. These two components of the **Total Sum of Squares** are independent and are therefore the base for the

correct quantities to compare. The **abdomin mean square** is an unbiased estimate of  $\sigma^2$  if  $H_0$  is true but an overestimate of  $\sigma^2$  if  $H_0$  is false. The **Residuals Mean Square** is unbiased for  $\sigma^2$  in either case. Therefore, the ratio of these two mean squares will tend to be close to 1 if  $H_0$  is true but greater than 1 otherwise. With the assumption of independent normally distributed  $\epsilon_i$ , the ratio, given as the **F-Value** = 86.427 in the table, follows a (central) **F** distribution with 1 and 45 degrees of freedom if  $H_0$  is true, but not otherwise. Appeal to this distribution tells us whether the ratio is significantly greater than 1. When the observed  $\text{Pr}(> F)$  value in the table (in this case < 0.0001) is small, we interpret that as evidence that  $H_0$  is false.

The formal statement of the test is: Under the null hypothesis that  $\beta_1 = 0$  (that is, that information about  $x=\text{abdomin}$  gives no information about  $y=\text{bodyfat}$ ), the probability of observing an **F-value** as large as the one we actually saw (in this case 86.427) is very small (in this case the probability is less than 0.0001). This very small **p-value** (assuming  $H_0$  is true) is very strong evidence that  $H_0$  is not true, that is, it is evidence that  $\beta_1 \neq 0$ . We will therefore act as if  $H_0$  is false and take further actions as if the relationship of the fitted regression model actually explains what is going on.

The estimate  $\hat{\beta}_1$  from Equation (8.7) can be algebraically rewritten as a linear combination of  $y_i$ -values

$$\hat{\beta}_1 = \sum (y_i - \bar{y}) \left( \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right) \quad (8.13)$$

The variance of  $\hat{\beta}_1$

$$\sigma_{\hat{\beta}_1}^2 = \text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \quad (8.14)$$

is constructed from the sum in Equation (8.13) with formulas based on Equation (3.8) (see Exercise 8.7). The sample estimate of the standard error of  $\hat{\beta}_1$  is

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}} \quad (8.15)$$

Under  $H_0$ , and with the assumption of independent normally distributed  $\epsilon_i$ , the **t-ratio**  $t_{\hat{\beta}_1} = \hat{\beta}_1 / \hat{\sigma}_{\hat{\beta}_1}$  has a  $t_{45}$  distribution and we have the right to use the **t table** in our tests.

Similarly, we can show (see Exercise 8.8)

$$\sigma_{\hat{\beta}_0}^2 = \text{var}(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right) \quad (8.16)$$

TABLE 8.5. Residual Mean Square in Regression Printout. The “Residual Mean Square” and “Error Mean Square” are two names for the same concept.

---

For each observation  $i$  the standard regression printout shows

$$\begin{array}{lllll}
 \widehat{\text{var}}(\hat{\mu}_i) & + & \widehat{\text{var}}(e_i) & = & \widehat{\text{var}}(y_i) = \hat{\sigma}^2 \\
 h_i \hat{\sigma}^2 & + & (1 - h_i) \hat{\sigma}^2 & = & \hat{\sigma}^2 \\
 \text{Std Err Predict}^2 & + & \text{Std Err Residual}^2 & = & \text{Error Mean Square (SAS)} \\
 h[i] * (\text{Residuals Mean Sq}) & + & (1-h[i]) * (\text{Residuals Mean Sq}) & = & \text{Residuals Mean Sq (S-PLUS)}
 \end{array}$$


---

The S-PLUS formulas assume the following:

```

my.lm <- lm(y ~ Z[,1] + Z[,2]) ## Z is a matrix
anova(my.lm)
h <- hat(Z)

```

---

### 8.3.4 Residual Mean Square in Regression Printout

The *residual mean square* is also called the *error mean square*. It is called *residual* because it is left over after fitting the model. It is called *error* because it is a measure of the difference between the model and the data. We prefer the term “residual” and discourage the term “error” because the term “error” suggests a mistake, and that is not the intent of this component of the analysis. See Table 8.5 for a comparison of several notations.

### 8.3.5 New Observations

One of the uses of a fitted regression equation is to make inferences about new observations. A new observation  $y_0$  at  $x_0$  has the model

$$y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0 = \mu_0 + \epsilon_0$$

where

- $y_0$  is a single unobserved value
- $x_0$  is the value of the predictor  $x$  at the new observation
- $\beta_0$  and  $\beta_1$  are the regression coefficients.

The concepts that we introduce here extend, almost without change, to the multiple regression setting of Chapter 9. We therefore preview the slightly more elaborate notation of Chapter 9. Model (8.1) can be rewritten in

matrix notation as

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad (8.17)$$

$$\begin{matrix} Y \\ n \times 1 \end{matrix} = \begin{matrix} X \\ n \times (1+p) \end{matrix} \begin{matrix} \beta \\ (1+p) \times 1 \end{matrix} + \begin{matrix} \epsilon \\ n \times 1 \end{matrix}$$

We restrict  $p = 1$  in Chapter 8. More generally, beginning in Chapter 9,  $p$  is a positive integer.

In the extended notation, a new observation  $y_0$  at  $x_{0+}$  has the model

$$y_0 = x_{0+}\beta + \epsilon_0 = \mu_0 + \epsilon_0$$

where

- $y_0$  is a single unobserved value
- $x_{0+}$  is a  $1 \times (1 + p)$  row of predictors [ $(1 x_0)$  in Chapter 8]
- $\beta$  is a  $(1 + p)$ -vector of regression coefficients [ $(\beta_0 \beta_1)'$  in Chapter 8].

There are two related questions to ask about the new observation:

1. Estimate the parameter  $\mu_0 = E(y_0) = x_{0+}\beta$ .
2. Predict a specific observation  $y_0 = \mu_0 + \epsilon_0$ .

Estimation intervals for new  $\mu_0$  and prediction intervals for new  $y_0$  based on a new value  $x_{0+}$  depend on the quantity  $h_0$  defined as

$$h_0 = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (8.18)$$

The formula for  $h_0$  is similar to the leverage formula for  $h_i$  in Equations (9.14) or (9.15), where the new value  $x_{0+}$  replaces one of the observed values  $X_{i+}$ . The notation  $i$  specifically means one of the original  $n$  observations and the notation 0 means an additional observation that need not be one of the original ones. Equation (8.18) is specifically for simple linear regression ( $p = 1$ ). The more complex formula in Equations (9.14) or (9.15) is needed when  $p > 1$ .

Answering the questions requires information about estimated variances:

1. Estimate the
  - a. parameter  $\mu_0 = E(y_0) = x_{0+}\beta$  with

- b. estimator  $\hat{\mu}_0 = x_0 + \hat{\beta}$ ,
  - c. variance of the estimator  $\text{var}(\hat{\mu}_0) = h_0 \sigma^2$ , and
  - d. estimated variance of the estimator  $\widehat{\text{var}}(\hat{\mu}_0) = h_0 \hat{\sigma}^2$ .
2. Predict a
- a. specific observation  $y_0 = \mu_0 + \epsilon_0$  with
  - b. predictor  $\hat{y}_0 = \hat{\mu}_0 = x_0 + \hat{\beta}$  (the same as the parameter estimate),
  - c. variance of the predictor  $\text{var}(\hat{y}_0) = \text{var}(\hat{\mu}_0 + \epsilon_0) = \text{var}(\hat{\mu}_0) + \text{var}(\epsilon_0)$ , and
  - d. estimated variance of the predictor  $\widehat{\text{var}}(\hat{y}_0) = \widehat{\text{var}}(\hat{\mu}_0) + \widehat{\text{var}}(\epsilon_0) = h_0 \hat{\sigma}^2 + \hat{\sigma}^2 = \hat{\sigma}^2(h_0 + 1)$ .

In the special case that  $x_{0+} = x_{i+}$  (one of the observed points), we have

$$\widehat{\text{var}}(\hat{y}_i) = (1 + h_i) \hat{\sigma}^2 = \hat{\sigma}^2 + \widehat{\text{var}}(\hat{\mu}_i)$$

Note that the (standard error)<sup>2</sup> for prediction  $\hat{\sigma}^2(h_0 + 1)$  is larger than the (standard error)<sup>2</sup> for estimation  $\hat{\sigma}^2 h_0$ . A prediction interval for individual observations  $\hat{y}_0$  estimates the range of observations that we might see. A confidence interval for the estimated mean of the new observations estimates the center point of the predicted range.

Most regression programs print the standard error for estimation of the mean:  $\hat{\sigma} \sqrt{h_0}$ , the confidence interval for estimating  $\mu_0 = E(y_0|x)$ :  $\hat{y}_0 \pm t_{df, \frac{\alpha}{2}} \hat{\sigma} \sqrt{h_0}$ , [also shown in Equation (9.25)], and the prediction interval for a new observation ( $y_0|x$ ):  $\hat{y}_0 \pm t_{df, \frac{\alpha}{2}} \hat{\sigma} \sqrt{1 + h_0}$  [also shown in Equation (9.26)]. These items are discussed in detail in Section 9.10.

The commands that construct the confidence and prediction intervals in S-PLUS and SAS, and their interpretation, are shown in Table 8.6. To see the standard error for prediction of a new observation, we must manually do the arithmetic

$$\hat{\sigma}^2 h_0 + \hat{\sigma}^2 = (1 + h_0) \hat{\sigma}^2 \quad (8.19)$$

The two questions about a new observation are actually familiar questions in a new guise. They are the same questions addressed in Section 3.6 about the location parameter  $\mu$  of a sample from a single variable. We elaborate on the comparison in Table 8.7.

In both the confidence interval and the prediction interval of the regression problem in Table 8.7, the magnitude of (Standard Deviation)<sup>2</sup> increases as the new value  $x$  moves further from the mean  $\bar{x}$  of the existing  $x_i$ 's. This indicates that we have more confidence in a prediction for an  $x$  in the vicinity of the  $x_i$ 's of the existing data than in an  $x$  far from the  $x_i$ 's of the existing data. The lesson is that extrapolations of the fitted regression relationship for remote values of  $x$  are likely to be unreliable.

TABLE 8.6. Construction of the confidence and prediction intervals for new observations in S-PLUS and SAS. See also the discussion surrounding Equations (9.25) and (9.26).

| S-PLUS  |   |
|---|---|
| <code>example.lm &lt;- lm(y ~ x1 + x2 + x3, data=old.data)</code> | fit linear model  |
| <code>predict(example.lm,</code>                                  | linear model object   |
| <code>newdata=data.frame(x1=3, x2=2, x3=45),</code>               | new data  |
| <code>se.fit=T,</code>  | $\sqrt{\text{var}(\hat{\mu}_0)} = \hat{\sigma}\sqrt{h_0}$           |
| <code>ci.fit=T,</code>  | $\hat{y}_0 \pm t_{df, \frac{\alpha}{2}} \hat{\sigma}\sqrt{h_0},$    |
| <code>pi.fit=T)</code>  | $\hat{y}_0 \pm t_{df, \frac{\alpha}{2}} \hat{\sigma}\sqrt{1 + h_0}$ |
| SAS   |   |
| <code>data newdata;</code>  | new data  |
| <code>x1=3;</code>  | $x$ -values only, no $y$  |
| <code>x2=2;</code>  |   |
| <code>x3=45;</code>   |   |
| <code>run;</code>   |   |
| <code>proc reg data=olddata newdata;</code>                       | linear model on old and new data                                    |
| <code>model y = x1 x2 x3 /</code>                                 | fit model   |
| <code>P</code>  | $\hat{\mu}_0 = \hat{y}$   |
| <code>CLM</code>  | $\hat{y}_0 \pm t_{df, \frac{\alpha}{2}} \hat{\sigma}\sqrt{h_0},$    |
| <code>CLI ;</code>  | $\hat{y}_0 \pm t_{df, \frac{\alpha}{2}} \hat{\sigma}\sqrt{1 + h_0}$ |
| <code>run;</code>   |   |

Confidence and prediction intervals for a particular new observation at  $x_0$  are shown in Table 8.6. These intervals can be extended to confidence and prediction *bands* by letting  $x_0$  vary over the entire range of  $x$ . Figure 8.4 illustrates such 95% bands for `fat.lm`, the modeling of `bodyfat` as a function of `abdomin`, displayed in Table 8.1. The 0.95 probability statement applies to each particular value of  $x = x_0$ . It does not apply to statements that the bands enclose the infinite set of all possible means or predictions as  $x$  varies over its range.

TABLE 8.7. Comparison of confidence and prediction intervals in the one-sample problem ( $t$ -test) and in the regression problem.

|   | One Sample   | Regression   |
|---|--|--|
| <b>Model Parameters:</b>  |  |  |
| Model Parameter   | $y = \mu_Y + \epsilon$                             | $y_x = \beta_0 + \beta_1 x + \epsilon$   |
| Variance of $\epsilon$  | $\text{var}(\epsilon) = \sigma_Y^2$                | $\text{var}(\epsilon) = \sigma_{YX}^2$   |
| <b>Sample Statistics:</b>   |  |  |
| Estimate  | $\hat{\mu}_Y = \bar{y}$                            | $\hat{\mu}_{yx} = b_0 + b_1 x$   |
| Variance  | $s_Y^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / (n - 1)$ | $s_{YX}^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - 2)$  |
| <b>Estimate Parameter:</b>  |  |  |
| $(\text{Standard Deviation})^2$ for Confidence Interval Estimate                        |  |  |
| What is the average height $\mu_Y$ of everyone?   |  | What is the average height $\mu_{YX}$ of those people who are $x = 10$ years old?  |
| $s_{\mu_Y}^2 = s_Y^2 = \frac{s_Y^2}{n} = s_Y^2 \left( \frac{1}{n} \right)$              |  | $s_{\mu_{yx}}^2 = s_{YX}^2 h_x = s_{YX}^2 \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$ |
| <b>Prediction Interval:</b>   |  |  |
| $(\text{Standard Deviation})^2$ for Prediction Interval for an Individual Response      |  |  |
| How tall is the next person?<br>$\hat{y} = \hat{\mu}_Y + \epsilon = \bar{y} + \epsilon$ |  | How tall is the next 10-year-old?<br>$\hat{y}_x = \hat{\mu}_{yx} + \epsilon = (b_0 + b_1 x) + \epsilon$                        |
| $s_{\hat{y}}^2 = \frac{s_Y^2}{n} + s_Y^2 = s_Y^2 \left( \frac{1}{n} + 1 \right)$        |  | $s_{\hat{y}_x}^2 = s_{YX}^2 h_x + s_{YX}^2 = s_{YX}^2 (1 + h_x)$   |
|   |  | $= s_{YX}^2 \left( 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$                           |

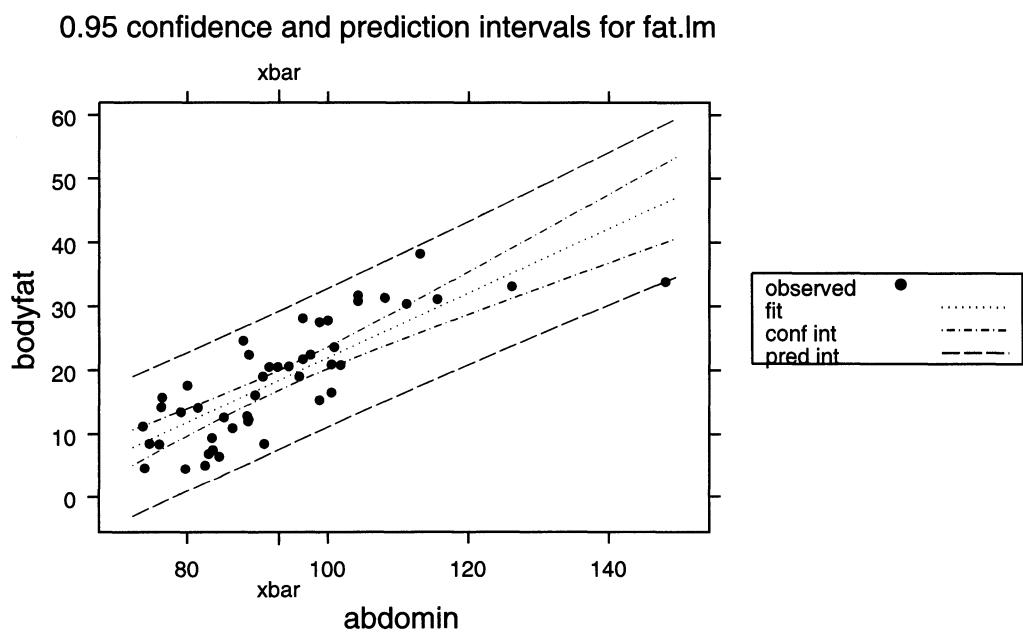


FIGURE 8.4. Confidence and prediction bands for modeling  $\text{bodyfat} \sim \text{abdomin}$ , body fat data. The widths of these bands are minimized at  $x = \bar{x}$  because  $h_0$  is minimized at  $x = \bar{x}$ .  
(rega/code/fat-ci.s), (rega/figure/fat-ci.eps.gz)

## 8.4 Diagnostics

There are two steps to a statistical analysis. The first step is to construct a model and estimate its parameters. Sections 8.3.2 and 9.3 discuss estimation of the parameters of linear models with one and two predictor variables. The second step is to study the quality of the fit of the data to that model and determine if the model adequately describes the data. This section introduces the diagnostics. They are investigated more thoroughly in Section 11.3.

The choice of diagnostic techniques are connected directly to the model and assumptions. If the assumption (8.2) that the error term  $\epsilon_i$  are normally independently distributed with constant mean 0 and variance  $\sigma^2$  is valid, then the residuals  $e_i = (y_i - \hat{y}_i)$  will be approximately normally distributed. More precisely, the  $n$  values  $e_i$  will behave exactly like  $n$  numbers independently chosen from the normal distribution and subjected to  $p + 1$  linear constraints. In the simplest case, when  $p = 0$  (one-sample  $t$ -test in Chapter 5), the residuals  $e_i$  behave like  $n$  independent normals centered on their observed mean  $\bar{x}$ . For simple linear regression ( $p = 1$ ), the residuals behave like  $n$  independent normals vertically centered on a straight line specified by the two estimated parameters  $\hat{\beta}_1$  and  $\hat{\beta}_0$ .

The diagnostic techniques are various procedures for looking at approximately normal numbers and seeing if they display any systematic behavior. If we see systematic behavior, then we conclude that the model did not capture all the interesting features of the data. We iterate the analysis steps by trying to model the systematic behavior we just detected and then looking at the residuals from the newer model.

Figure 8.5 shows the diagnostics from the simple regression model of Section 8.1. These are standard plots constructed in S-PLUS with the single statement `plot(fat.lm)`.

The single most evident feature in Figure 8.5 is the point labeled 39. It has the largest  $\hat{y} = 46.22$ , the largest absolute residual  $e_{39} = -12.42$ , the biggest gap in the normal probability plot between it and the next residual, and at 2.11 the largest Cook's distance (a measure of the unusualness of the point; values of Cook's distance larger than 1 call for investigation. Cook's distance and other diagnostic measures of individual data points will be discussed in Chapter 11).

The six panels from S-PLUS `plot.lm` look best when printed on one page with the two-line idiom

```
par(mfrow=c(2,3))
plot(fat.lm)
```

We will discuss each panel in turn, with the numbering sequence (456)<sup>(123)</sup>.

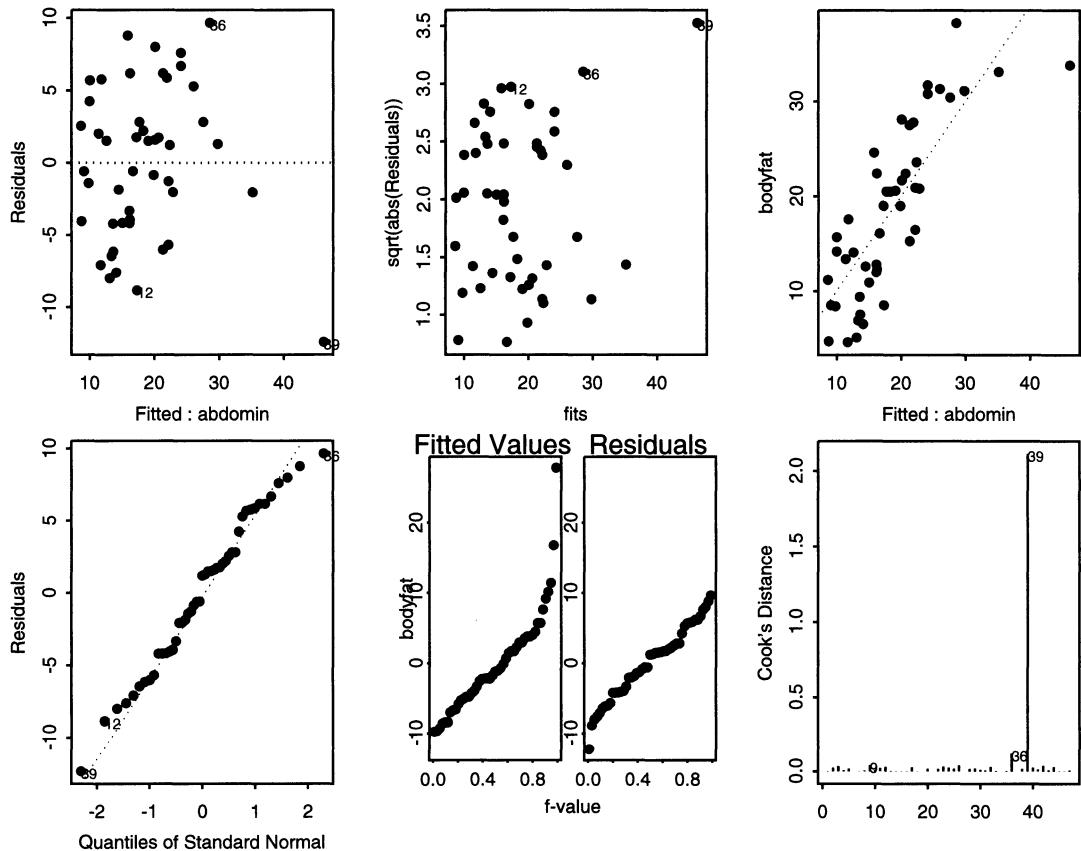
**diagnostics for  $\text{lm}(\text{bodyfat} \sim \text{abdomin}, \text{data}=\text{fat})$** 


FIGURE 8.5. Diagnostics for  $\text{lm}(\text{bodyfat} \sim \text{abdomin}, \text{data}=\text{fat})$ . See Section 8.4 for a discussion of each of the six panels in this display.

(rega/code/rega.f6.s), (rega/figure/f6.EPS.gz)

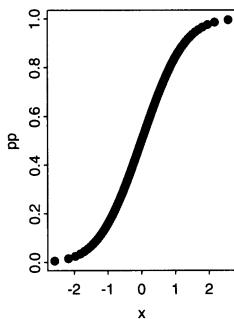
1. This is a plot of the residuals  $e = y - \hat{y}$  against the fitted values  $\hat{y}$  along with a horizontal line at  $e = 0$ . The horizontal line corresponds to the least-squares fit of the residuals against the predicted values. There is, by construction, no linear effect in this panel. There may be quadratic (or higher-order polynomial) effects visible. The marginal distribution of the predicted values may show patterns that need further investigation. When there is only one  $x$ -variable, as in the example in Figure 8.5, the predicted values are a linear transformation of the  $x$ -variable. In this example, we see that  $x_{39}$ , the point with the largest residual, is noticeably larger than any of the other  $x$ -values.

2. This panel plots  $\sqrt{|e|}$  against the fitted values  $\hat{y}$ . It shows much of the same information as panel 1. The absolute value folds the negative residuals onto the positive direction in order to emphasize magnitude of departure from the model at the expense of not showing direction. The square root transformation brings in the larger residuals and spreads out the smaller ones. See the discussion of the ladder of powers in Section 4.8 for more information on the effects of transformations.
3. This is a plot of  $y$  against  $\hat{y}$ . When there is only one  $x$ -variable, this is a rescaled version of the original plot of  $y$  against  $x$ . The regression line is also shown.
4. This is a normal probability plot with the residuals on the vertical axis and the normal quantiles on the horizontal axis. The diagonal line has the standard deviation  $s$  for its slope. When the residuals are approximately normal, the points will be close to the diagonal line. Asymmetries in the residuals will be visible. Short tails in the distribution of the residuals will be visible as an “S”-shaped display, and long tails in the distribution of the residuals (seen as vertical outliers in panels 1 and 2) will be visible as mirror-image “S” shapes. See Section 5.8 for further discussion of probability plots.
5. This panel is subdivided into two transposed empirical distributions (see immediately below). The left panel shows the centered fitted values  $\hat{y} - \bar{y}$  and the right panel shows the residuals  $e$ . The relative vertical ranges of these two panels gives some information on the multiple correlation coefficient  $R^2$ .

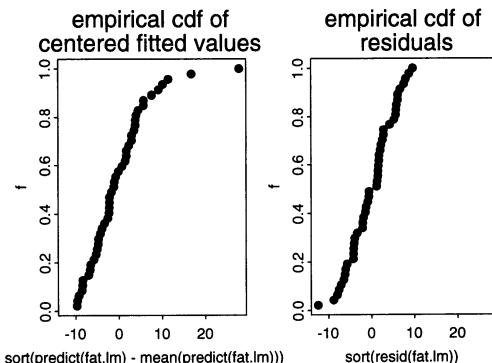
We develop the construction of Figure 8.5’s panel 5 in Figure 8.6. The empirical distribution of  $S(x)$  is defined in Section 5.7 as the fraction of the data that is less than or equal to  $x$ . The empirical distribution is defined analogously to the cumulative distribution  $F(x) = P(X \leq x)$  of a theoretical distribution.

- a. The plot of the cumulative distribution is a plot of  $F(x)$  against  $x$ .
- b. The empirical cumulative distribution of an observed set of data is a plot of proportion( $X \leq x$ ) against  $x$ . If there are  $n$  observations in the dataset, we plot  $i/n$  against  $x_{[i]}$ . We use the convention here that subscripts in square brackets mean that the data have been sorted. For example, let us look at the fitted values  $\hat{y}$  and residuals  $e = y - \hat{y}$  from the regression analysis in Table 8.1. The left side of Figure 8.6b is the cumulative distribution of the fitted values. The right side is the cumulative distribution of the residuals. Note that these plots are on very different scales for the abscissa and therefore cannot easily be compared visually.
- c. We construct Figure 8.6c by making two adjustments to Figure 8.6b. First, we center the fitted values on their mean. Second, we plot

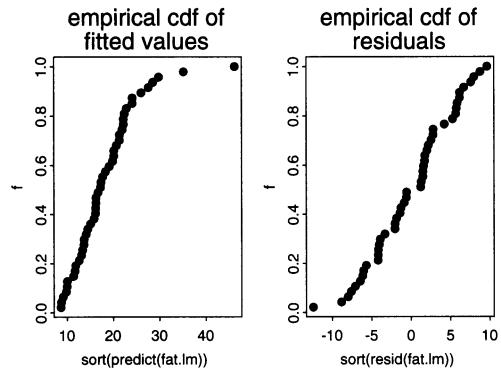
a. Cumulative distribution of the standard normal  $\Phi(x)$  for  $x \sim N(0, 1)$ .

Cumulative Distribution of  $N(0,1)$ 

c. Empirical distributions of fitted values and residuals with common range for the abscissa.



b. Empirical distributions of fitted values and residuals with independent ranges for the abscissa.



d. Transposed empirical distributions of fitted values and residuals with common range for the abscissa. This is panel 5 of Figure 8.5.

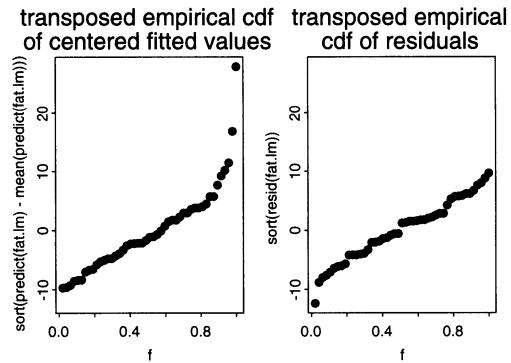


FIGURE 8.6. Explanation of panel 5 of Figure 8.5. Panels a,b,c are empirical distribution plots and panel d is the transposed empirical distribution plot of the fitted values and residuals from the linear regression

```
(rega/code/diagnostic.s), (rega/code/diag1.s), (rega/code/diag2f.s),
(rega/code/diag2r.s), (rega/code/diag3.s), (rega/code/diag4.s),
(rega/figure/diag1.eps.gz), (rega/figure/diag2.eps.gz),
(rega/figure/diag3.eps.gz), (rega/figure/diag4.eps.gz),
```

both graphs on the same abscissa scale by forcing them to have the same  $x$ -axis constructed as the range of the union of their individual abscissas.

- d. Figure 8.6d is the transpose of the pair of graphs in Figure 8.6c. We interchange the axes, putting the proportions on the abscissa and the data (centered fitted values in the left panel and residuals in the right panel) on the ordinate. We therefore force the  $y$ -axes to have a common limits. S-PLUS uses Figure 8.6d as the fifth diagnostic plot of Figure 8.5. The vertical axis now uses the same  $y$  units as panels 1, 3, and 4.

If our model explains the data well, then we would anticipate that the residuals have less variability than the fitted values.

The multiple correlation  $R^2$  can be written as

$$R^2 = \frac{SS_{\text{Reg}}}{SS_{\text{Total}}} = \frac{SS_{\text{Reg}}}{SS_{\text{Reg}} + SS_{\text{Res}}} \quad (8.20)$$

We can use the squared range of the fitted values as a surrogate for the  $SS_{\text{Reg}}$  and the squared range of the residuals as a surrogate for the  $SS_{\text{Res}}$ . This leads to the interpretation of panel 5 as an indicator of  $R^2$ . If the ranges of the left and right panels are similar, then  $R^2 \approx \frac{1}{2}$ . If the range of the fitted values is larger, then the  $R^2$  is closer to 1, and if the range of the fitted values is smaller, then the  $R^2$  is closer to 0.

- 6. Panel 6 is Cook's distance, to be discussed in detail in Section 11.3.4. Cook's distance measures the unusualness of an observation by looking at the values of both its response variable  $y$  and its predictor variables  $x$ . An observation with a large Cook's distance has either a large residual, a large leverage (see Section 9.2.1), or some combination of these two. A case having a Cook's distance greater than 1 is deemed unusual. Such a case is referred to as *influential* relative to the other cases. It is very important to read the vertical scale in the plot of Cook's distances. A point larger than the others is not important unless its numerical value is high.

## 8.5 Graphics

The figures in this chapter represent five different types of plots.

Figure 8.1 is a scatterplot matrix, constructed in S-PLUS with `splom()` and in SAS with `%scatmat`.

All the panels of Figure 8.2 are variations of the same simple scatterplot, all have been forced to the same  $x$ - and  $y$ -ranges for comparability, and all have a superimposed straight line.

Columns 2 and 3 of Figure 8.2 and both panels of Figure 8.3 use `regr1.plot`, a function in file (`splus.library/regr1.plot.s`) that is constructed from the S-PLUS-supplied functions `plot`, `abline`, `points`, and the function `resid.squares` in file (`splus.library/resid.squares.s`). Our function `resid.squares` constructs the squares that represent the squared residuals with real squares on the plotting surface. The heights of the squares are in  $y$ -coordinates. The widths of the squares are the same number of inches on the plotting surface as the heights.

Figures 8.5 and 9.2 use the S-PLUS function `plot.lm` to display six standard plots of residuals and potential outliers from a regression analysis.

Figure 8.6 is a collection of scatterplots.

## 8.6 Exercises

- 8.1.** (Hand et al., 1994) report on a study by (Lea, 1965) that investigated the relationship between mean annual `temperature` (degrees F) in regions of Britain, Norway, and Sweden, and the rate of `mortality` from a type of breast cancer in women. The data appear in (`datasets/breast.dat`).
  - a. Plot the data. Does it appear that the relationship can be adequately modeled by a linear function?
  - b. Estimate the regression line and add this to your plot.
  - c. Calculate and interpret  $R^2$ .
  - d. Calculate and interpret the standard error of estimate.
  - e. Interpret the estimated slope coefficient in terms of `mortality` and `temperature`.
  - f. Find a 95% confidence interval on the population slope coefficient.
  - g. Find a 95% prediction interval for a region having mean annual temperature 45.
  - h. One of these 16 data points is unusual compared to the others. Describe how.
- 8.2.** (Shaw, 1942), later in (Mosteller and Tukey, 1977), shows the `level` of Lake Victoria Nyanza relative to a standard level and the number of `sunspots` in each of 20 consecutive years. The data are located in the file (`datasets/lake.dat`). Use linear regression to model the lake level as a function of the number of sunspots in the same year.

- 8.3.** Does muscle mass decrease with age? The `age` in years and muscle `mass` were obtained from 16 women. The data come from (Neter et al., 1996) and appear in the file (`datasets/muscle.dat`).
- Plot `mass` vs `age` and overlay the fitted regression line.
  - Interpret the slope coefficient in terms of the model variables.
  - Predict with 90% confidence the muscle mass of a 66-year-old woman.
  - Interpret the calculated standard error of estimate.
  - Interpret  $R^2$  in terms of the model variables.
- 8.4.** The dataset (`datasets/girlht.dat`) contains the heights (in cm) at ages 2, 9, and 18 of 70 girls born in Berkeley, California in 1928 or 1929. The variables are named `h2`, `h9`, and `h18`, respectively. The data come from a larger file of physical information on these girls in (Cook and Weisberg, 1999).
- Regress `h18` on `h9` and also `h18` on `h2`.
  - Discuss the comparative strengths of these two regression relationships.
  - Interpret the slope coefficients of both regressions.
- 8.5.** We would expect that the price of a diamond ring would be closely related to the price of the diamond the ring contains. (Chu, 1996) presents data on the `price` (Singapore dollars) of ladies' diamond rings and the number of `carats` in the ring's diamond. The data are contained in the file (`datasets/diamond.dat`).
- Regress `price` on `carats`.
  - Notice that the estimated intercept coefficient is significantly less than 0. Therefore, this model is questionable, although the range of the predictor variables excludes 0. Instead fit a model without an intercept term.
  - Compare the goodness of fits of the two models. Which is preferable?
- 8.6.** The file (`datasets/income.dat`), from (Bureau of the Census, 2001), contains year 2000 data on the percentage of college graduates and per capita personal `income` for each of the 50 states and District of Columbia. Regress `income` on `college`. Interpret the meaning of  $R^2$  for these data. Discuss which states have unusually low or high per capita income in relation to their percentage of college graduates.

**8.7.** Prove Equation (8.14)

$$\sigma_{\hat{\beta}_1}^2 = \text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

The proof is primarily algebraic manipulation. Rewrite (8.13) as a weighted sum of the independent  $y_i$ , that is as

$$\hat{\beta}_1 = \sum(y_i - \bar{y}) \left( \frac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \right) = \sum y_i k_i \quad (8.21)$$

then write

$$\text{var}(\hat{\beta}_1) = \sigma^2 \sum k_i^2 \quad (8.22)$$

and simplify.

**8.8.** Prove Equation (8.16) that the variance of the estimate of the intercept  $\hat{\beta}_0$  has variance

$$\sigma_{\hat{\beta}_0}^2 = \text{var}(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right)$$

**8.9.** Algebraically prove the assertion in Equation (8.8) that in simple regression, the sum of the calculated residuals is zero.**8.10.** In Figure 8.2 we construct the actual squares of the residuals and show that the sum of the areas of the squared residuals is smallest for the least-squares line. We do the construction in the simplest way, placing the other three sides on the side that is already there representing the residual. Other possibilities are

- a. Place the left-right center of the square on the residual line. Use `resid.squares` in file (`splus.library/resid.squares.s`) as the model for your function.
- b. Place a circle (a real circle in inches of graph surface) on the points. Base your function on the functions `resid.squares` in file (`splus.library/resid.squares.s`) and the S-PLUS `symbols` function with argument `add=T` and with a numerical value for argument `inches`.

Option 1: Keep the existing residual line and center the circle on the observed point.

Option 2: Use the existing residual line as the diameter of the circle.

## 8.A Appendix: Computation for Regression Analysis

### 8.A.1 S-PLUS Functions

These functions are included in the library provided in the online files accompanying the book.

#### `regr1.plot`

The `regr1.plot` function in (`splus.library/regr1.plot.s`) is based on the explanation of least-squares regression in (Smith and Gonick, 1993). The function plots  $y$  against  $x$ , draws a straight line (by default the least-squares line), and identifies the fitted values  $\hat{y}$  on the straight line corresponding to the  $x$ -values. With optional arguments it will jitter the  $x$ -values (to make overplotting evident), and plot the residuals either as vertical lines or as squares.

#### `resid.squares`

The plots of the squared residuals are handled by the separate function `resid.squares` in (`splus.library/resid.squares.s`). The squares have the vertical dimensions of the residuals in  $y$  units and horizontal dimensions that are the same number of plotting inches as the vertical dimensions. The statement that calculates the rectangle width `rect.width` with the `par()` values does the arithmetic with aspect ratios and plotting units to make the square a real square.

#### `regr2.plot`

The `regr2.plot` function in (`splus.library/regr2.plot.s`) does the same type of plot for bivariate regression, one  $y$ -variable and two  $x$ -variables. The function is based on the `persp` perspective plotting function in S. We designed the `regr2.plot` function with options to display grids for the base plane and the two back planes in addition to the observed points and the regression plane and the fitted points. We turned off the default plot of the 3-dimensional box. The function `regr2.plot` uses the functions defined in (`splus.library/persp.hh.s`).

### 8.A.2 SAS Macros and Procs

These are either functions that we provide or functions that are distributed with SAS for which we wish to discuss specific details.

# Multiple Regression—More Than One Predictor

In Chapter 8 we introduce the algebra and geometry behind the fitting of a linear model relating a response variable to one or more explanatory (predictor) variables using the criterion of least squares. In this chapter we consider in more detail situations where there are two or more predictors.

The two linear modeling techniques we have studied so far, regression in Chapter 8 and analysis of variance in Chapter 6, have much of their mathematics interpretation in common. In this chapter we explore the common mathematical features, with some examples of how they apply. In the following chapters we use this common structure.

We begin by extending the Chapter 8 discussion of regression with a single predictor (simple regression) to allow for two or more predictors. *Multiple* regression refers to regression analysis with at least two predictors.

## 9.1 Regression with Two Predictors—Least-Squares Geometry

The graphics for least squares with two  $x$ -variables, and in general for more than two  $x$ -variables, are similar to the graphics in Figure 8.2. We will work with `abdomin` and `biceps`. The basic 3-dimensional plot is in Figure 9.1, where `bodyfat` is plotted as  $y$  against the other two variables as  $x_1$  and  $x_2$ .

We think of this plot as a point cloud in 3-space floating over the surface defined by the  $x$ -variables. Any plane other than the least-squares plane will show a larger sum of squared areas than the least-squares plane illustrated here.

### Least-squares with two X-variables

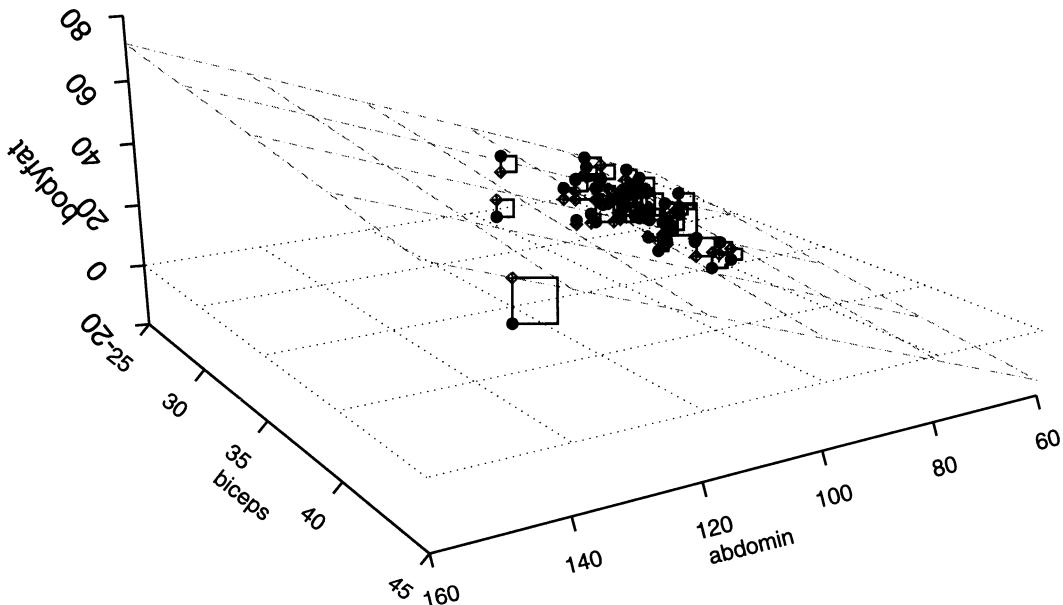


FIGURE 9.1. Body fat data with two  $X$ -variables. The best-fitting plane minimizes the sum of the squared areas. See Table 9.1.

(regb/code/rega.f4.s), (regb/figure/f4.EPS.gz)

Figure 9.2 shows the diagnostics from the two- $X$  regression model of Section 9.3. Compare this to the similar plot for one- $X$  regression in Figure 8.5.

A similar construction is in principle possible for more  $X$ -variables. Illustrating the projection of four or more dimensions onto a two-dimensional graph is difficult at best.

### diagnostics for `lm(bodyfat ~ abdomen + biceps, data=fat)`

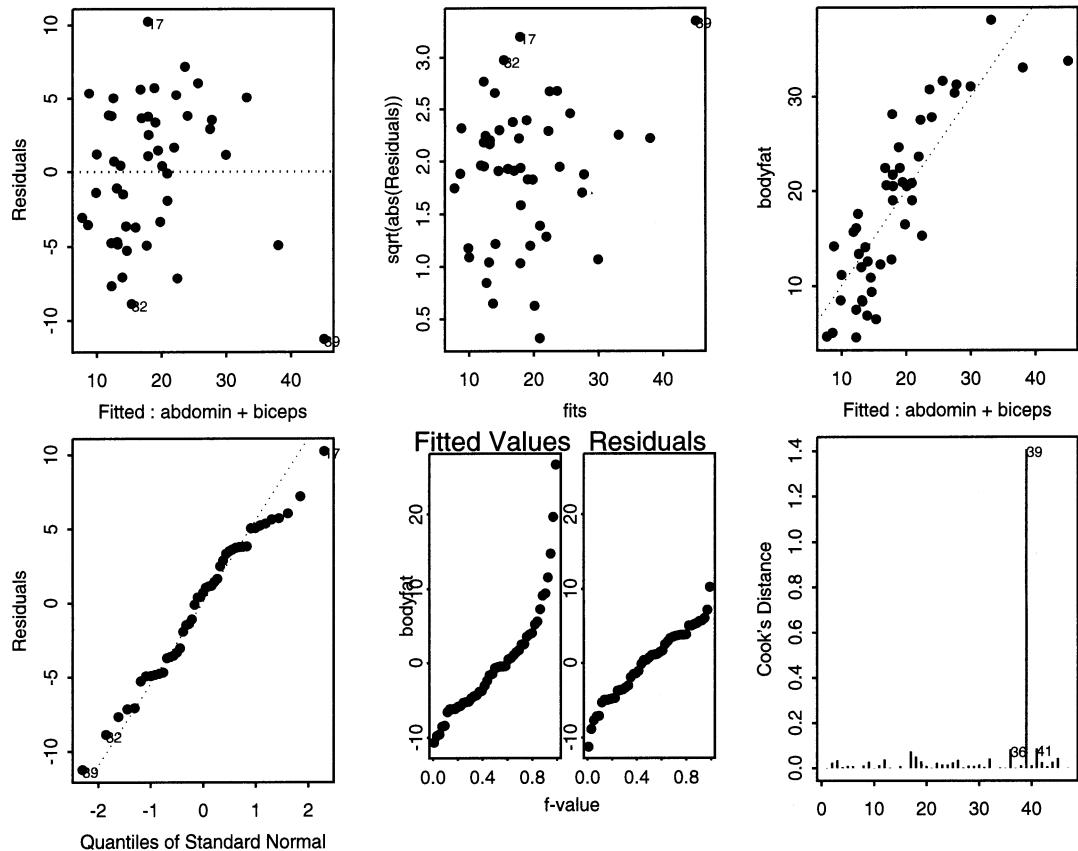


FIGURE 9.2. Diagnostics for `lm(bodyfat ~ abdomen + biceps, data=fat)`. Compare this to the similar plot for one- $X$  regression in Figure 8.5.

(`regb/code/rega.f7.s`), (`regb/figure/f7.EPS.gz`)

## 9.2 Multiple Regression—Algebra

Everything in simple regression analysis carries over to multiple regression. There are additional issues that arise because we must also study the relations among the predictor variables. The algebra for multiple regression is most easily expressed in matrix form. (A brief introduction to matrix algebra appears in Appendix F.) The formulas for simple regression can be derived as the special case of multiple regression with  $p = 1$ .

Assume

$$\underset{n \times 1}{Y} = \underset{n \times 1}{X} \underset{(1+p) \times 1}{\beta} + \underset{n \times 1}{\epsilon} \quad (9.1)$$

or equivalently

$$y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i \quad \text{for } i = 1, \dots, n \quad (9.2)$$

where

- $\underset{n \times 1}{Y}$  are observed values,
- $\underset{n \times (1+p)}{X} = [\mathbf{1} \ X_1 \ X_2 \ \dots \ X_p]$  are observed values with  $\underset{n \times 1}{\mathbf{1}}$  representing the constant column with 1 in each row and  $\underset{n \times 1}{X_j}$  indicating the column with  $X_{ij}$  in the  $i^{\text{th}}$  row,
- $\underset{(1+p) \times 1}{\beta}$  are unknown constants,
- $\underset{n \times 1}{\epsilon} \sim N(0, \sigma^2 I)$  are independent.

Then the least-squares estimate  $\hat{\beta}$  is obtained by minimizing the sum of squared deviations

$$S = (Y - X\beta)'(Y - X\beta) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}))^2$$

by taking the derivatives ( $\partial S / \partial \beta_j$ ) with respect to all the  $\beta_j$  and setting them to 0. The resulting set of equations, called the *Normal Equations* and generalizing Equation 8.6,

$$(X'X)\hat{\beta} = (X'Y) \quad (9.3)$$

are solved for  $\hat{\beta}$ . The solution [equivalent to Equation 8.7] is equal to

$$\hat{\beta} = (X'X)^{-1}(X'Y) = ((X'X)^{-1}X')Y = (X^+Y) \quad (9.4)$$

The symbol  $X^+$   $\stackrel{\text{def}}{=} (X'X)^{-1}X'$  is the notation for the Moore–Penrose *generalized inverse* of any rectangular matrix. In the special case of square invertible matrices the generalized inverse becomes the familiar matrix inverse. We introduce this notation here because it simplifies the appearance of the equations. We start with the model  $Y = X\beta + \epsilon$  in Equation (9.1) and conclude with the estimate  $\hat{\beta} = X^+Y$  in Equation (9.4). We effectively moved the  $X$  to the other side and replaced the  $\epsilon$  with the hat on the  $\beta$ . Note that Equation (9.4) is an identity, but not an efficient computing algorithm. An efficient algorithm uses Gaussian elimination to solve the equations directly.

We construct the fitted values with

$$\hat{Y} = X\hat{\beta} = (X(X'X)^{-1}X')Y = HY \quad (9.5)$$

where the matrix

$$H \stackrel{\text{def}}{=} X(X'X)^{-1}X' \quad (9.6)$$

is a projection matrix. The sum of squares (SS) for the regression is  $\text{SS}_{\text{Reg}} = Y'HY$ . The projection matrix  $H$  is called the *hat matrix* because multiplying  $H$  by  $Y$  places a hat ‘ $\hat{\cdot}$ ’ on  $Y$ . We can see that  $H_{ij} = \partial\hat{Y}_i/\partial Y_j$ . We discuss the hat matrix in Section 9.2.1.

The *residuals* are defined as the difference

$$e = Y - \hat{Y} = (I - H)Y \quad (9.7)$$

between the observed values  $Y$  and the fitted values  $\hat{Y}$ . With least-squares fitting, the residuals are orthogonal to the observed  $x$ -values

$$e'X = 0 \quad (9.8)$$

and therefore to the fitted values

$$e'\hat{Y} = e'X\hat{\beta} = 0 \quad (9.9)$$

The variance-covariance matrix of the residuals  $e$  is  $\sigma^2(I - H)$ . Note in particular that  $\text{var}(e_i) = \sigma^2(1 - H_{ii})$  is not constant for all  $i$ .

An unbiased estimator of  $\sigma^2$  is

$$s^2 = \frac{Y'(I - H)Y}{n - p - 1} = \text{MS}_{\text{Res}} = \text{SS}_{\text{Res}}/\text{df}_{\text{Res}} \quad (9.10)$$

Its square root,  $s$ , sometimes called the standard error of estimate, is an asymptotically unbiased estimator of  $\sigma$ . As in the case of simple regression, the sum of the residuals is zero, that is,

$$\sum_{i=1}^n e_i = \mathbf{1}'e = 0 \quad (9.11)$$

where  $\mathbf{1}'$  is a row vector of ones. The proof of this assertion is requested in Exercise 9.1.

Both  $\hat{\beta}$  and  $\hat{Y}$  are linear combinations of  $y_i$ . The  $y_i$  are independent because the  $\epsilon_i$  are independent. Hence the elementary theorems in Equations (3.7) and (3.8),

$$E(a_1y_1 \pm a_2y_2) = a_1E(y_1) \pm a_2E(y_2)$$

and

$$\text{var}(a_1y_1 \pm a_2y_2) = a_1^2 \text{var}(y_1) + a_2^2 \text{var}(y_2)$$

are applicable. These are where we get Equation (8.14), the standard error for  $\beta_1$ , the corresponding formula

$$\text{var}(\hat{\beta}) = \sigma^2(X'X)^{-1} \quad (9.12)$$

for the estimator of  $\beta$  in Equation (9.4), and formulas (9.25) and (9.26) for tests and confidence intervals about  $E(Y|X)$  and for prediction intervals about  $Y$  for new values of  $X$ .

### 9.2.1 The Hat Matrix and Leverage

The hat matrix in Equation (9.6) is called that because premultiplication by  $H$  places a hat ‘ $\hat{\cdot}$ ’ on  $Y$ :  $\hat{Y} = HY$ . The  $i^{\text{th}}$  diagonal of  $H$  is called the leverage of the  $i^{\text{th}}$  case because it tells how changes in  $Y_i$  affect the location of the fitted regression line, specifically:

$$\frac{\partial \hat{Y}_i}{\partial Y_i} = H_{ii} \quad (9.13)$$

If ( $H_{ii} > 2(p+1)/n$ ), then the  $i^{\text{th}}$  point is called a high leverage point. See Section 11.3.1. Equation (9.13) shows that changes in the observed  $Y_i$ -value of high leverage points have a large effect on the predicted value  $\hat{Y}_i$ , that is, they have a large effect on the location of the fitted regression plane.

The hat matrix is used in regression diagnostics, that is, techniques for evaluating how the individual data points affect the regression analysis. Many diagnostics are discussed in Section 11.3.

Frequently these diagonals of  $H$  are denoted by  $h_i = H_{ii}$ . They are calculated in S-PLUS with the command `hat(x)` and in SAS with the `influence` option in `proc reg`.

A specific formula for the leverage  $h_i$  itself is almost simple:

$$h_i = X_{i\cdot}(X'X)^{-1}X'_{i\cdot} \quad \text{where } X_{i\cdot} \text{ is the } i^{\text{th}} \text{ row of } X \quad (9.14)$$

In an alternate but common notation, the predictor matrix does not include the column **1**. To avoid excessive confusion, define  $Z$  to be all the columns of  $X$  except the initial column **1**:

$$Z = [X_1 X_2 \dots X_p]_{n \times p}$$

and let

$$\bar{Z} = (\bar{X}_1 \bar{X}_2 \dots \bar{X}_p)$$

In this notation the formula for leverage looks worse:

$$h_i = \frac{1}{n} + (Z_{i\cdot} - \bar{Z})((Z - \bar{Z})(Z - \bar{Z})')^{-1}(Z_{i\cdot} - \bar{Z})' \quad (9.15)$$

The term  $\frac{1}{n}$  in Equation (9.15), with the  $Z$  matrix which excludes the column **1**, is not needed in Equation (9.14), with the  $X$  matrix which includes the column **1**. In simple regression, with  $Z = X_1 = x$ , formula (9.15) simplifies to

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (9.16)$$

### 9.3 Multiple Regression—Two- $X$ Analysis

The specification of the analysis for two  $x$ -variables is similar to that for one  $x$ -variable. The sequential ANOVA table and the table of coefficients for a two  $x$ -variable analysis of the body fat data (**datasets/fat.data**) are in Table 9.1.

Since both predictors are significantly different from 0, the arithmetic justifies the illustration in Figure 9.1, where we see that  $\hat{y}$  changes linearly with changes in either  $x_1$  and  $x_2$ . The table of coefficients tells us that on average for this population, percent body fat increases by 0.683 if abdomen circumference increases by one cm and biceps is unchanged, and percent body fat decreases by .922 if biceps increases by one cm while abdomen is unchanged.

The  $t$ -value for **biceps** (the second variable in the ANOVA table) is related to the  $F$ -value for **biceps**:  $t^2 = (-2.946)^2 = 8.677 = F$ . The  $t$ -value (8.693) for **abdomin** (the first variable in the ANOVA table) is not simply related to the correspondingly labeled  $F$ -value (101.172). We investigate this relationship in the discussion of Table 13.27.

TABLE 9.1. Sequential ANOVA table and table of regression coefficients from the two- $x$  model with  $y=\text{bodyfat}$ ,  $x_1=\text{abdomin}$ , and  $x_2=\text{biceps}$ . These tables were typeset from the program output. See Figure 9.1.

(regb/transcript/fat2a.st), (regb/transcript/fat.lst)

---

S-PLUS (regb/code/ls2.s):

```
fat2.lm <- lm(bodyfat ~ abdomin + biceps, data=fat)
summary(fat2.lm, corr=F)
anova(fat2.lm)
```

---

SAS (regb/code/fat.sas):

```
proc glm data = fat;
model bodyfat = abdomin biceps;
run;
```

---

**ANOVA Table**

| Source    | df | Sum of Sq | Mean Sq  | F Value | Pr(> F)  |
|-----------|----|-----------|----------|---------|----------|
| abdomin   | 1  | 2440.500  | 2440.500 | 101.172 | < 0.0001 |
| biceps    | 1  | 209.317   | 209.317  | 8.677   | 0.0051   |
| Residuals | 44 | 1061.382  | 24.122   |         |          |
| Total     | 46 | 3711.199  |          |         |          |

---

**Table of Regression Coefficients**

| Predictor   | Value   | Std. Error | t value | Pr(>  t ) |
|-------------|---------|------------|---------|-----------|
| (Intercept) | -14.594 | 6.692      | -2.181  | 0.0346    |
| abdomin     | 0.683   | 0.079      | 8.693   | < 0.0001  |
| biceps      | -0.922  | 0.313      | -2.946  | 0.0051    |

---

## 9.4 Geometry of Multiple Regression

Several types of pictures go along with multiple regression. The one we have looked at already is the scatterplot matrix, drawn with the S-PLUS command `splom(~data.matrix)`; for example, see Figure 8.1 for the splom of the body fat dataset.

The picture that goes best with the defining least-squares equations is the multidimensional point cloud. It is easiest to illustrate this with  $Y$  and two  $X$ -variables. See Figures 8.2 and 9.1 for one- $X$  and two- $X$  examples that use our function `resid.squares`.

## 9.5 Programming

### 9.5.1 Model Specification

We use several notations for the specification of a regression model to a computer program. How are the statements constructed in each notation, and what are their syntax and their semantics?

For specificity, let us look at a linear regression model with a response variable  $y$  and two predictor variables  $x_1$  and  $x_2$ . We express this model in several equivalent notations. In the algebraic notation of Section 9.2, we have

$$\begin{matrix} Y & = & X & \beta & + & \epsilon \\ n \times 1 & & n \times (1+2) & (1+2) \times 1 & & n \times 1 \end{matrix} \quad (9.17)$$

or equivalently

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \quad \text{for } i = 1, \dots, n \quad (9.18)$$

In S-PLUS model formula notation, we have

$$y \sim x1 + x2 \quad (9.19)$$

In SAS model statement notation, we have

$$y = x1 x2 \quad (9.20)$$

In both computer languages the statement is read, “ $y$  is modeled as a linear function of  $x_1$  and  $x_2$ .”

The four statements (9.17)–(9.20) are equivalent. Both computational specifications remove the redundancy in notation used by the traditional scalar algebra notation. The program knows that the variables ( $y$ ,  $x1$ , and  $x2$ ) have length  $n$ ; there is no need to repeat that information. All linear model specifications have regression coefficients, and most have a constant term

(we discuss models without a constant term in Section 9.9.); there is no need to specify the obvious. There is always an error term because the model does not fit the data exactly; there is no need to specify the error term explicitly. The two pieces of information unknown to the program are

- Which variable is the response and which are the predictors. This is indicated positionally—the response is on the left, and notationally—the “~” or “=” separates the response from the predictors. A separation symbol is needed because the same notation can be generalized to express multiple response variables.
- The relationship between the predictors. S-PLUS indicates summation explicitly with the “+” and SAS indicates it implicitly by leaving a space between the predictor variable names. Other relationships, for example crossing or nesting (to be discussed beginning in Section 13.5), are indicated by other algebraic symbols as indicated in Table 13.22.

The interpretation of operator symbols in the model specification notation is related to, but not identical to, the interpretation of the same symbols in an ordinary algebra statement. The model formulas (9.19) and (9.20) mean:

find the coefficients  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$  that best fit

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \epsilon_i \quad (9.21)$$

for the observed values  $(y_i, x_{i1}, x_{i2})$  for all  $i: 1 \leq i \leq n$ .

The “+” and space “ ” in formulas (9.19) and (9.20) do not have the ordinary arithmetic sense of  $x_{i1} + x_{i2}$ .

### 9.5.2 Printout Idiosyncrasies

The S-PLUS **summary** and **anova** functions do not print the **Total** line in their ANOVA tables.

SAS PROC GLM uses the name “Type I Sum of Squares” for the sequential ANOVA table. See the discussion of sums of squares types in Section 13.6.1.

Figure 9.1 uses our function **regr2.plot**, a function based on the S-PLUS function **persp** for three-dimensional perspective plots, and our function **resid.squares**. Our functions are in files (**splus.library/regr2.plot.s**) and (**splus.library/resid.squares.s**).

## 9.6 Example—Albuquerque Home Price Data

### Study Objectives

Realtors can use a multiple regression model to justify a house selling price based on a list of desirable features the house possesses. Such data are commonly compiled by local boards of realtors. We consider a data file containing a random sample of 117 home sales in Albuquerque, New Mexico during the period February 15 through April 30, 1993, taken from (Albuquerque Board of Realtors, 1993).

### Data Description

We use a subset of five of the eight variables for which data are provided, and 107 of the 117 houses that have information on all five of these variables.

**price:** Selling price in \$100's

**sqft:** Square feet of living space

**custom:** Whether the house was built with custom features (1) or not (0)

**corner:** Whether the house sits on a corner lot (1) or not (0)

**taxes:** Annual taxes in \$

We investigate models of **price** as a function of some or all of the candidate predictors **sqft**, **custom**, **corner** and **taxes**. This example assumes that **taxes** potentially determine **price**. In some real estate contexts the causality could work in the opposite direction: selling prices can affect subsequent home appraisals and hence tax burden.

### Data Input

We read the data from (`datasets/houseprice.dat`) into S-PLUS with the commands in file (`regb/code/housepriceread.s`) or into SAS with (`regb/code/housepriceread.sas`) and then look at the data with the scatterplot matrices in Figures 9.3 and 9.4. Two of the four candidate predictors, **custom** and **corner**, are dichotomous variables, and the panels involving them in Figure 9.3 are wasteful of space and not very informative. Figure 9.4, with separate portions for the two values of **corner** and separate plot symbols for the two values of **custom**, displays the information much more efficiently. We learn from this figure that custom houses tend to have higher prices than regular houses, and corner houses have different patterns of relationships between **price** and the continuous predictors than middle houses.

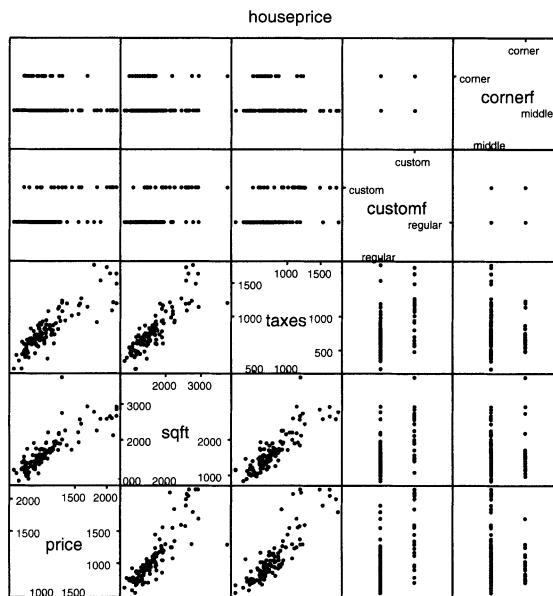


FIGURE 9.3. House-price data. The discreteness of variables `customf` and `cornerf` decreases the informativeness of this splom, particularly the panel for this pair of variables. Figure 9.4 is a preferred splom presentation of these data.

(`regb/code/houseprice.f1.s`), (`regb/figure/regb.f.hpr.eps.gz`)

Figure 9.4 suggests that price is directly related to all four candidate predictors. We proceed with the analysis by regressing `price` on the four variables in Table 9.2. In this Table we examine the signs of the regression coefficients and the magnitudes of their *p*-values. We see that `price` is strongly positively associated with `sqft`, `taxes` and `custom` (as opposed to `regular`) houses. Such conclusions are consistent with common knowledge of house valuation. The predictor `corner` has a marginally significant negative coefficient. Hence there is moderate evidence that, on average, corner houses tend to be lower priced than middle houses.

The magnitudes of the regression coefficients also convey useful information. For example, on average, each additional square foot of living space corresponds to a  $0.2076 \times \$100 = \$20.76$  increase in price, and on average custom houses sell for  $15.681481 \times \$100 = \$1,568.15$  more than regular houses. The  $R^2 = 0.8280$  says that in the population of houses from which (`datasets/houseprice.dat`) is a random sample, 82.8% of the variability in price is accounted for by these four predictors.

TABLE 9.2. Analysis of variance table for house-price data.

SAS (regb/code/houseprice3.sas):  
 proc reg data = houseprice;  
 model price = sqft custom corner taxes;  
 run;

SAS (regb/transcript/houseprice3.lst):  
 Dependent Variable: price

| Analysis of Variance |            |                    |                |         |         |
|----------------------|------------|--------------------|----------------|---------|---------|
| Source               | DF         | Sum of Squares     | Mean Square    | F Value | Pr > F  |
| Model                | 4          | 12941849           | 3235462        | 122.79  | <.0001  |
| Error                | 102        | 2687729            | 26350          |         |         |
| Corrected Total      | 106        | 15629578           |                |         |         |
| Root MSE             | 162.32771  | R-Square           | 0.8280         |         |         |
| Dependent Mean       | 1077.34579 | Adj R-Sq           | 0.8213         |         |         |
| Coeff Var            | 15.06737   |                    |                |         |         |
| Parameter Estimates  |            |                    |                |         |         |
| Variable             | DF         | Parameter Estimate | Standard Error | t Value | Pr >  t |
| Intercept            | 1          | 175.16603          | 56.31157       | 3.11    | 0.0024  |
| sqft                 | 1          | 0.20760            | 0.06105        | 3.40    | 0.0010  |
| custom               | 1          | 156.81481          | 44.49454       | 3.52    | 0.0006  |
| corner               | 1          | -83.40126          | 40.05934       | -2.08   | 0.0399  |
| taxes                | 1          | 0.67707            | 0.10101        | 6.70    | <.0001  |

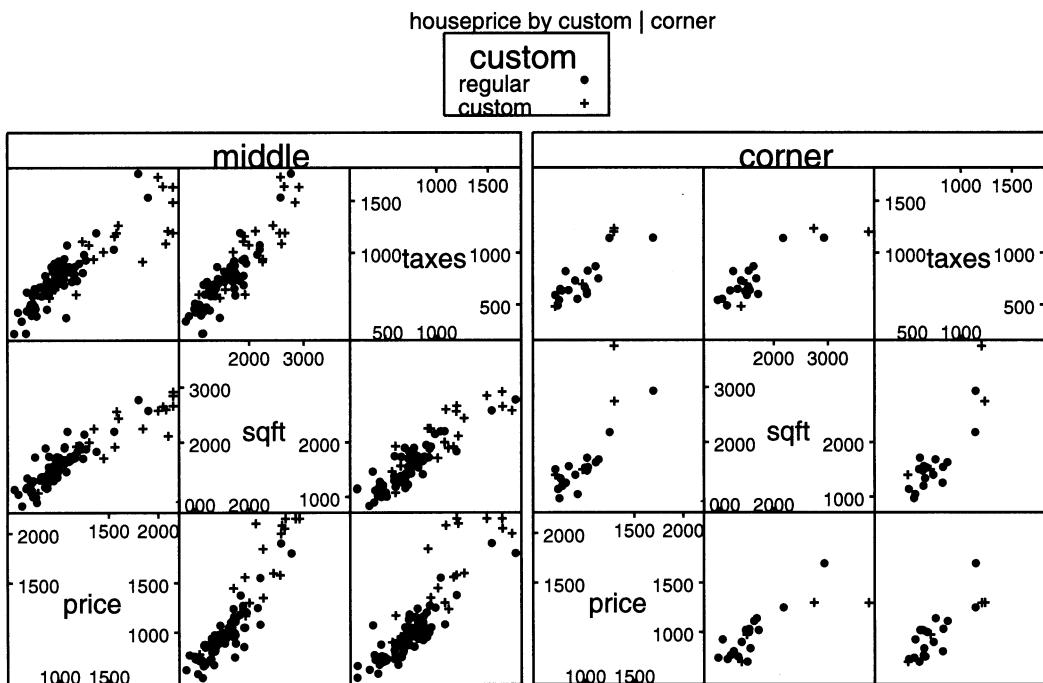


FIGURE 9.4. Albuquerque house-price data. Custom houses go for higher prices than regular houses. Corner houses have a different pattern than middle houses.

(regb/code/houseprice.f1.s), (regb/figure/regb.f.hprcc.eps.gz)

## 9.7 Partial F-Tests

Sometimes we wish to examine whether two or more predictor variables *acting together* have a significant impact on the response variable. For example, suppose we consider the house-price data of Section 9.6 with four candidate predictors, `sqft`, `custom`, `corner` and `taxes`, and wish to examine if `custom` and `corner` together have a significant impact on `price`, above and beyond the impacts of `sqft` and `taxes`. S-PLUS (in Table 9.3) approaches this by direct comparison of two models. The *full model* contains all predictors under consideration. The *reduced model* contains all predictors apart from the ones we test in order to see if then can be eliminated from the model. *Partial F* refers to the fact that we are simultaneously testing *part* of the model's predictors, not all predictors but perhaps more than just one of them. SAS (in Table 9.3) uses an explicit TEST statement to

TABLE 9.3. Partial  $F$ -tests of  $H_0: \beta_{\text{custom}} = \beta_{\text{corner}} = 0$  using the S-PLUS `anova()` command and using the SAS `test` statement.  
 (regb/code/houseprice2.s)

S-PLUS (regb/transcript/houseprice2.st):  
 > houseprice.lm1 <- lm(price ~ sqft + taxes, data=houseprice)  
 > houseprice.lm2 <- lm(price ~ sqft + custom + corner + taxes,  
 + data=houseprice)  
 > anova(houseprice.lm1, houseprice.lm2)  
 Analysis of Variance Table

Response: price

|   | Terms                    | Resid. | Df  | RSS     | Test           | Df | Sum of Sq | F Value | Pr(F)   |
|---|--------------------------|--------|-----|---------|----------------|----|-----------|---------|---------|
| 1 | sqft+taxes               |        | 104 | 3152660 |                |    |           |         |         |
| 2 | sqft+custom+corner+taxes |        | 102 | 2687729 | +custom+corner | 2  | 464931    | 8.82212 | 0.00029 |

SAS (regb/code/houseprice2.sas):  
 proc reg data = houseprice;  
 model price = sqft custom corner taxes;  
 custcorn: test custom = corner = 0;  
 run;

SAS (regb/transcript/houseprice4.lst):  
 Dependent Variable: price

Test custcorn Results for Dependent Variable price

| Source      | DF  | Mean Square | F Value | Pr > F |
|-------------|-----|-------------|---------|--------|
| Numerator   | 2   | 232465      | 8.82    | 0.0003 |
| Denominator | 102 | 26350       |         |        |

state the hypothesis. The idea behind this test is apparent from Table 9.3. The  $F$ -test examines whether the reduction in residual sum of squares as a result of fitting the more elaborate model is a significant reduction. This assessment is performed by measuring the *extra sum of squares*, defined as

$$( \text{residual SS from reduced model} ) - ( \text{residual SS from full model} ) \quad (9.22)$$

against the residual sum of squares from the full model. The degrees of freedom associated with the extra sum of squares equals the number of parameters being tested for possible elimination.

The general form of the test is

$$F = \frac{(\text{extra SS})/(\text{df associated with extra SS})}{(\text{full model residual SS})/(\text{df associated with full model residual SS})} \quad (9.23)$$

The strategy of this approach is used whenever one wishes to compare the fits of two linear models, one of which has the same terms as the other plus at least one more term.

For testing the hypothesis that the population regression coefficients of `custom` and `corner` are both equal to 0, we see that the  $F$ -statistic is 8.82 on 2 and 102 degrees of freedom. There are two numerator degrees of freedom because the null hypothesis involves constraints on two model parameters. The very small  $p$ -value strongly suggests that this null hypothesis is false. We conclude that at least one of `custom` and `corner` is needed in the model.

The preceding discussion assumes that `sqft` and `taxes` were already in the model. It is also possible to test the combined effect on `price` of `custom` and `corner` compared with no other predictors, or exactly one of the predictors `sqft` and `taxes`. However, we do not pursue these possibilities here.

## 9.8 Polynomial Models

If the relationship between a response  $Y$  and an explanatory variable  $X$  is believed to be nonlinear, it is sometimes possible to model the relationship by adding an  $X^2$ -term to the model in addition to an  $X$ -term. For example, if  $Y$  is product demand and  $X$  is advertising expenditure on the product, an analyst might feel that beyond some value of  $X$  there is “diminishing marginal returns” on this expenditure. Then the analyst would model  $Y$  as a function of,  $X$ ,  $X^2$  and possibly other predictors, and anticipate a significant negative coefficient for  $X^2$ . Occasionally a need is encountered for higher-order polynomial terms.

An example from (Hand et al., 1994), original reference (Williams, 1959), is (`datasets/hardness.dat`) which we first encountered in Exercise 4.5. In this section we investigate the modeling of `hardness` as a quadratic function of `density`. We pursue this analysis in Exercise 11.2 from another angle, a transformation of the response variable `hardness`.

Hardness of wood is more difficult to measure than density. Modeling hardness in terms of density is therefore desirable. These data come from a

TABLE 9.4. Quadratic regression of hardness data. The quadratic term, with  $p=.0027$ , is very important in explaining the curvature of the observations. See Figure 9.5 to compare this fit with the linear fit.

(regb/code/hardness.s), (regb/transcript/hardness.st)

---

```
S-PLUS (regb/transcript/hardness-lm.st):
> hardness.quad.lm <- lm(hardness ~ density + density^2, data=hardness)
>
> anova(hardness.quad.lm)
Analysis of Variance Table

Response: hardness

Terms added sequentially (first to last)
  Df Sum of Sq Mean Sq F Value    Pr(F)
density   1  21345674 21345674 815.9232 0.000000000
I(density^2) 1   276041   276041 10.5515 0.002669045
Residuals 33   863325   26161
>
> coef(summary.lm(hardness.quad.lm, corr=F))
            Value Std. Error   t value   Pr(>|t|)
(Intercept) -118.0073759 334.966905 -0.3522956 0.726856611
density      9.4340214 14.935620  0.6316458 0.531969926
I(density^2)  0.5090775  0.156721  3.2483031 0.002669045
```

---

sample of Australian Janka timbers. A quadratic model fits these data better than a linear model. An additional virtue of the quadratic model is that its intercept term differs insignificantly from zero; this is not true of a model for these data containing only a linear term. (If wood has zero hardness, it certainly has zero density.)

The fitted quadratic model model in Table 9.4 is

$$\text{density} = -118.007 + 9.4340 \text{ hardness} + 0.5091 \text{ hardness}^2$$

The regression coefficient for the quadratic term is significantly greater than zero, indicating that the plot is a parabola opening upwards as shown in Figure 9.5. The  $p$ -value for the quadratic regression coefficient is identical to the  $p$ -value for the quadratic term in the ANOVA table because both tests are for the marginal effect of the quadratic term assuming the linear term is already in the model. The two  $p$ -values for the linear term differ because they are testing the linear coefficient in two different models. The  $p$ -value for linear regression coefficient assumes the presence of a quadratic

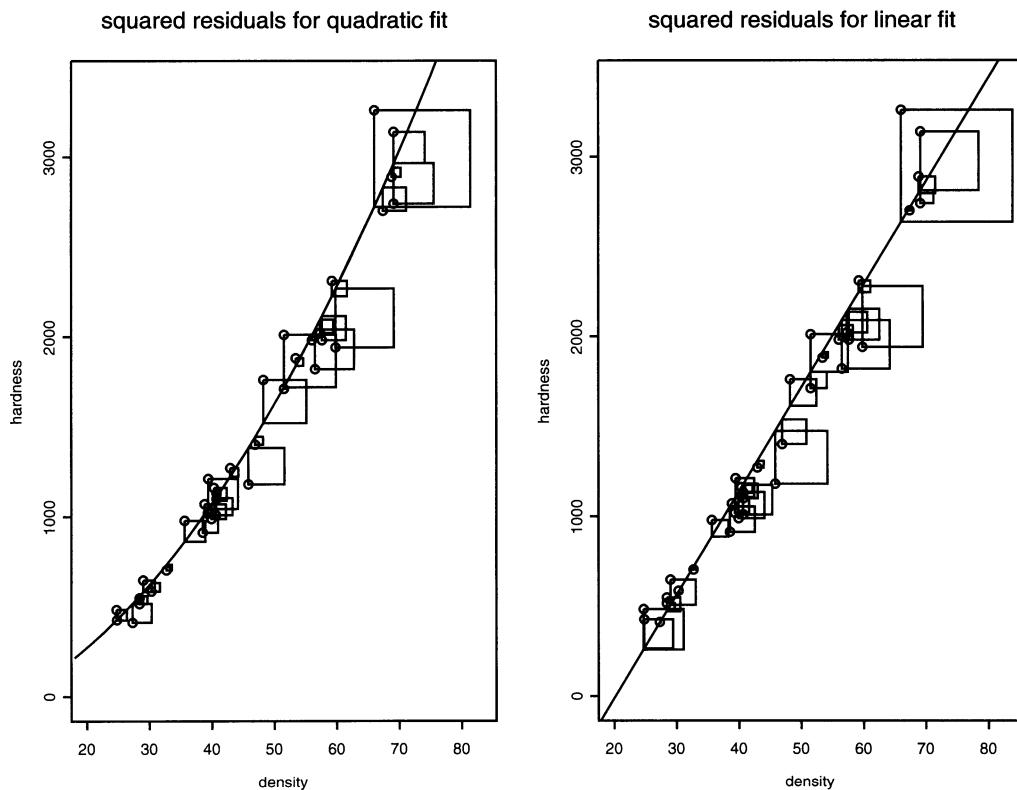


FIGURE 9.5. Linear  $y \sim x$  and quadratic  $y \sim x + x^2$  fits of  $y = \text{hardness}$  to  $x = \text{density}$ . The quadratic curve fits much better as can be seen from the much smaller squares (leading to smaller residual sum of squares) at the left and right ends of the density range in the quadratic fit. See Table 9.4 for the numerical comparison.

(`regb/code/hardness.s`), (`regb/figure/hardness-ls.eps.gz`)

term in the model, but the linear  $p$ -value in the sequential ANOVA table addresses a model with only a linear component.

When fitting a truly quadratic model, it is appropriate to include the linear term in the model even if its coefficient does not significantly differ from zero unless there is subject area theory stating that the relationship between the response and predictor lacks a linear component.

## 9.9 Models Without a Constant Term

Sometimes it is desired that the statistical model for a response not contain a constant (i.e., vertical intercept) term because the response is necessarily equal to zero if all predictors are zero. An example is the modeling of the body fat data discussed in Section 9.1. Obviously, if a “subject” has zero measurements for `abdomin` and `biceps`, then the response `bodyfat` is necessarily zero also. Similarly, if we wish to model the volume of trees in a forest as a function of trees’ diameters and heights, a “tree” having zero diameter and height must have no volume.

An advantage to explicitly recognizing the zero intercept constraint is that a degree of freedom that would be used to estimate the intercept is instead used to estimate the model residual. This results in slightly increased power of tests and decreased sizes of interval estimates of model parameters.

Figure 9.6 and Table 9.5 are for regressions of `bodyfat` on `biceps`, both with and without a constraint that the regression pass through the origin. Note the appreciably smaller slope of the no-intercept regression and that the no-intercept model has 46 df for residual as compared with 45 df for the unconstrained model.

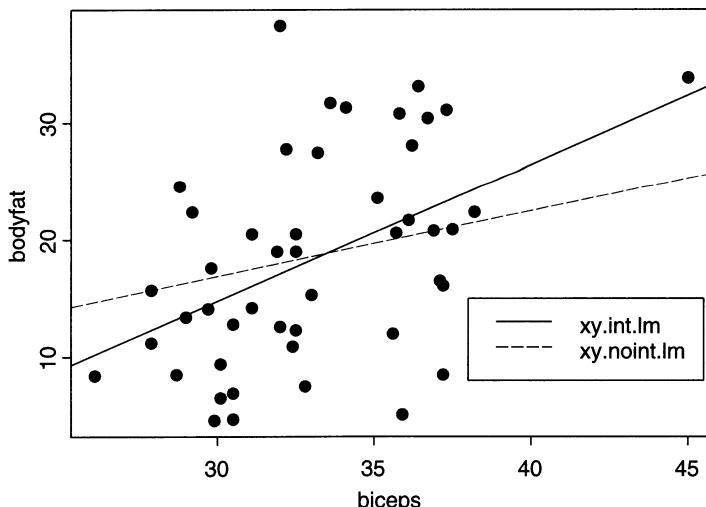


FIGURE 9.6. Regressions with and without a constant term for a portion of the body fat data. See Table 9.5.

(`regb/code/nointercept.s`), (`regb/figure/nointercept.eps.gz`)

TABLE 9.5. Body fat data: Regressions of `bodyfat` on `biceps`, with and without an intercept term. See Figure 9.6. As compared with the intercept model, the no-intercept model has larger values of both the regression sum of squares and the total sum of squares, and hence also a larger value of  $R^2$ .  
(regb/code/nointercept.s), (regb/transcript/nointercept.st)  
(regb/code/nointercept.sas), (regb/transcript/nointercept.lst)

---

```
S-PLUS (regb/transcript/nointercept.st):
> ## usual model with intercept
> xy.int.lm <- lm(bodyfat ~ biceps, data=fat)
> summary(xy.int.lm, corr=F)

Call: lm(formula = bodyfat ~ biceps, data = fat)
Residuals:
    Min      1Q  Median      3Q     Max 
-16.58 -5.443 -0.8457  5.255 21.09 

Coefficients:
            Value Std. Error t value Pr(>|t|)    
(Intercept) -20.3644   10.8551   -1.8760   0.0671  
biceps      1.1712    0.3261    3.5915   0.0008  
                                                        
Residual standard error: 8.006 on 45 degrees of freedom
Multiple R-Squared: 0.2228
F-statistic: 12.9 on 1 and 45 degrees of freedom, the p-value is 0.0008096
>
> ## model without a constant term
> xy.noint.lm <- lm(bodyfat ~ biceps - 1, data=fat)
> summary(xy.noint.lm, corr=F)

Call: lm(formula = bodyfat ~ biceps - 1, data = fat)
Residuals:
    Min      1Q  Median      3Q     Max 
-15.11 -6.145 -0.006463  6.841 20.19 

Coefficients:
            Value Std. Error t value Pr(>|t|)    
biceps      0.5630   0.0360    15.6251   0.0000  
                                                        
Residual standard error: 8.222 on 46 degrees of freedom
Multiple R-Squared: 0.8415
F-statistic: 244.1 on 1 and 46 degrees of freedom, the p-value is 0
```

---

### Specification in S-PLUS and SAS

S-PLUS suppresses the intercept in the model formula by symbolically subtracting the “1” that stands for the constant dummy variable.

```
bodyfat ~ biceps - 1
```

SAS suppresses the intercept in the model statement with an explicit no-intercept option.

```
model bodyfat = biceps / noint ;
```

## 9.10 Prediction

Generalizing the discussion in Section 8.3.5 for simple regression, the multiple regression model equation, with regression coefficients estimated by the least-squares analysis, is commonly used for two distinct but related problems.

1. Find a confidence interval on the conditional mean of the population of  $Y|x$ . That is, find the set of  $Y$ -values that the model could randomly generate from the specified values of the predictors  $x$ .
2. Find a prediction interval for the a new observed response  $Y_0$  from these values of the predictors  $x$ ; i.e., an interval within which a particular new observation will fall with a certain probability.

We continue the analysis of (*datasets/fat.data*) to illustrate the distinction between these two problems. Using S-PLUS we continue with **fat2.lm** calculated in (*regb/code/ls2.s*) and with output displayed in Table 9.1. For specificity, we work with  $x_1 = \text{abdomin} = 93$  and  $x_2 = \text{biceps} = 33$ .

The algebraic setup begins from the model in Equation (9.1), from which it follows that

$$s_e^2 = \frac{Y'Y - \hat{\beta}'X'Y}{n - p - 1} = \frac{(Y - \hat{Y})'(Y - \hat{Y})}{n - p - 1}$$

Let  $x_0 = (93 \ 33)$  denote the vector of predictor values for which we wish to construct these two intervals. Define

$$h_0 = x_0(X'X)^{-1}x_0' \tag{9.24}$$

Let  $t_{\frac{\alpha}{2}, n-p-1}$  denote the  $100(1 - \frac{\alpha}{2})$  percentage point of the  $t$  distribution with  $n - p - 1$  degrees of freedom. The expected response  $E(y|x_0)$  (the center of the confidence interval) and the predicted response  $\hat{y}_{x_0}$  for a new

observation (the center of the prediction interval) are both equal to  $x'_0 \hat{\beta}$ . Then the  $100(1 - \alpha)\%$  confidence interval is

$$x'_0 \hat{\beta} \pm t_{\frac{\alpha}{2}, n-p-1} s_e \sqrt{h_0} \quad (9.25)$$

and the  $100(1 - \alpha)\%$  prediction interval is

$$x'_0 \hat{\beta} \pm t_{\frac{\alpha}{2}, n-p-1} s_e \sqrt{1 + h_0} \quad (9.26)$$

The prediction interval is wider than the confidence interval because we are predicting one particular  $y$  corresponding to  $x_0$ , but estimating with confidence the mean  $E(y|x_0)$  of all possible  $y$ 's that could arise from  $x_0$ . A particular  $y$  could be much smaller or larger than the mean, and hence there is more uncertainty about  $y$  than about the mean. This is captured in the distinction between the two preceding formulas: the “1+” inside the square root. The “1+” arises from the fact that we must predict the  $\epsilon_0$  part of the model, but in the estimation problem, we estimate that  $\epsilon_0$  is zero. As a result, the prediction interval for a given set of explanatory variables is always wider than the corresponding confidence interval.

The confidence and prediction intervals for this example are shown in Table 9.6. The confidence interval (17.0, 19.9) is for the mean percentage body fat of a population of individuals each having `abdomin` circumference 93 cm and `biceps` circumference 33 cm. The prediction interval (8.5, 28.5) is for one particular individual with this combination of `abdomin` and `biceps`. Observe that the prediction interval is wider than the confidence interval. This is because a single person can have atypically low or high body fat, but “many” people includes those with both atypically low and high body-fat percentages in comparison to their `abdomin` and `biceps`, and the lows and highs tend to cancel out when averaging. See Table 8.7 for an illustration of this in the more familiar setting of estimation of a sample mean.

## 9.11 Example—Longley Data

### Study Objectives

The Longley data is a classic small set containing 16 years of annual macroeconomic data that (Longley, 1967) used to illustrate difficulties arising in computations involving highly intercorrelated variables. Both S-PLUS and SAS accurately calculate the regression coefficients for these data. Less numerically sophisticated statistical software packages, including most in existence at the time Longley wrote his article, produce incorrect analyses because the high intercorrelation, or ill-conditioning of the data, is a computational challenge for the numerical solution of linear equations and related matrix operations.

TABLE 9.6. 95% Confidence and prediction intervals for the body-fat example. See Tables 9.1 and 13.27 for the ANOVA table and the regression coefficients. The `predict` function produces  $s_e\sqrt{h_0}=se.fit$ ,  $s_e=residual.scale$  and the confidence and prediction intervals.

(regb/code/fat2.s)

---

```
S-PLUS (regb/transcript/fat2.st):
> predict(fat2.lm,
+         newdata=data.frame(abdomin=93, biceps=33),
+         se.fit=T, pi=T, ci=T)
$fit:
 1
18.4884

$se.fit:
 1
0.7170514

$residual.scale:
[1] 4.911449

$df:
[1] 44

$ci.fit:
      lower   upper
1 17.04327 19.93352
attr(, "conf.level"):
[1] 0.95

$pi.fit:
      lower   upper
1 8.485086 28.4917
attr(, "conf.level"):
[1] 0.95
```

---

We use (`datasets/longley.dat`), distributed with S-PLUS, a subset of all variables in Longley's original data set. Our intent here is to develop a parsimonious model to explain the response variable `Employed` as a function of the remaining variables as candidate predictors. The extreme collinearity arises in this data set because all of its economic variables tend to increase as time progresses. We acknowledge that these are really time series data, and if more than 16 years were involved, it would be appropriate to use time series techniques such as those in Chapter 18 for

a proper analysis. Our intention in this section is to analyze these data using multiple regression, demonstrating ways to bypass or confront the difficulties collinearity presents for regression modeling. In contrast, time series analyses specifically seek to model the interdependence caused by time.

## Data Description

**GNP.deflator:** GNP adjusted for inflation based on year 1954 = 100  
**GNP:** Gross National Product, 1964 Economic Report of the President  
**Unemployed:** 1964 Economic Report of the President  
**Armed.Forces:** Number serving in the U.S. Armed Forces  
**Population:** Noninstitutional, aged at least 14  
**Year:** 1947 through 1962  
**Employed:** Total employment, U.S. Department of Labor, March 1963

## Discussion

Figure 9.7 contains a splom of the Longley data. Here the response variable **Employed** appears in the last row and last column. (In general, for ease of interpretation, response variables should appear in this way or in the first row and first column.)

We see that **Employed** is highly positively correlated with four of the six predictors and mildly positively correlated with the others. In addition, the predictors that are highly correlated with **Employed** are also highly correlated with one another. This suggests that these four predictors carry redundant information and therefore some of them are unnecessary for modeling the response.

Consider the listing in Table 9.7 for a model containing all six candidate predictors. The proportion of variability in the response **Employed** that is collectively explained by all six predictors is given by  $R^2$ , the proportion of the **Sum of Squares** column *not* in the **Residuals** row: more than 0.99. So the predictors can be used to adequately explain **Employed**. In this model, three predictors that seem to be closely correlated with the response **Employed** in Figure 9.7, **Population**, **GNP**, and **GNP.deflator**, are not statistically significant in Table 9.7. We continue to discuss the Longley data, focusing on the selection of an appropriate subset of the predictors, in Sections 9.12 and 9.13.

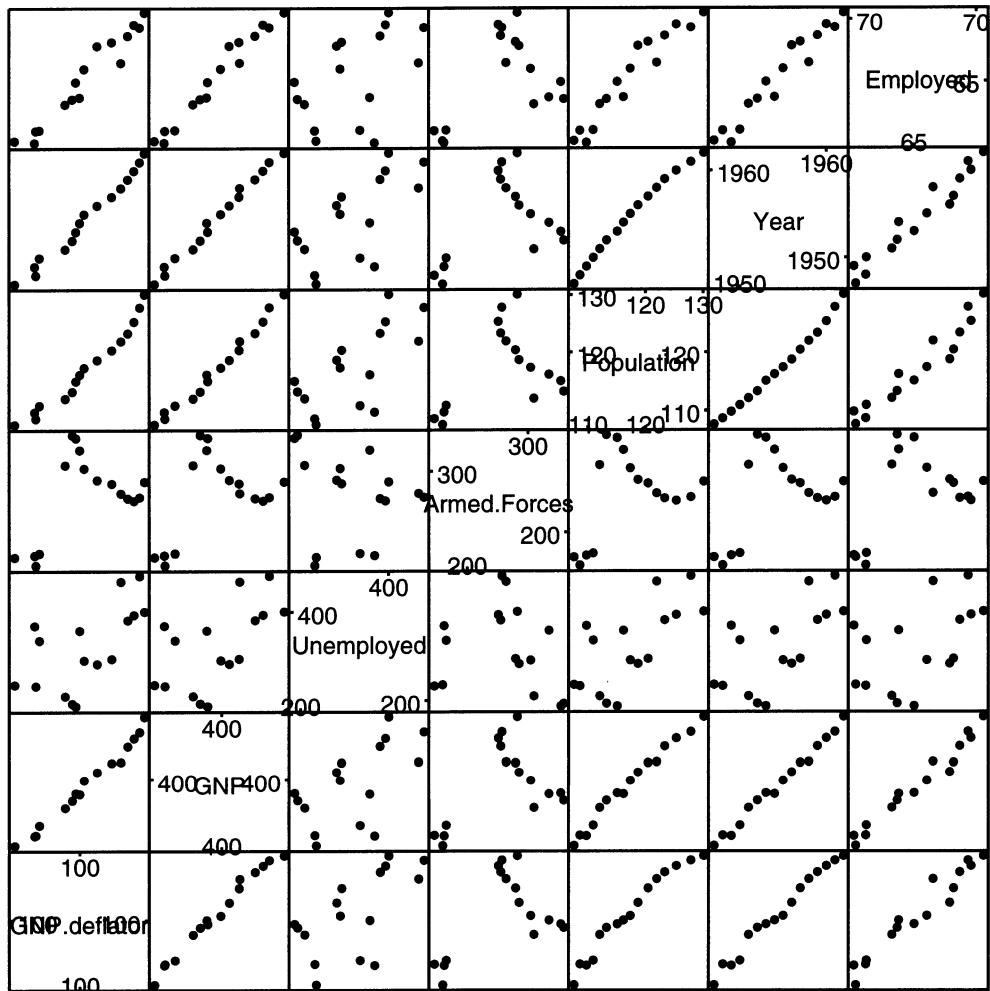


FIGURE 9.7. Longley data splom. Notice the high positive correlations of four predictors with one another and with the response variable **Employed**.  
(regb/code/longley.s), (regb/figure/longley.eps.gz)

TABLE 9.7. Longley data regression using all six original predictors.  
(regb/code/longley.s)

---

```

S-PLUS (regb/transcript/longley.st):
> longley <- data.frame(longley.x, Employed = longley.y)
>
> longley.lm <- lm( Employed ~ . , data=longley)
> summary(longley.lm, corr=F)

Call: lm(formula = Employed ~ . , data = longley)
Residuals:
    Min      1Q  Median      3Q     Max 
-0.4101 -0.1577 -0.02816 0.1016 0.4554 

Coefficients:
            Value Std. Error   t value   Pr(>|t|)    
(Intercept) -3482.2586   890.4204  -3.9108  0.0036    
GNP.deflator   0.0151    0.0849    0.1774  0.8631    
GNP           -0.0358    0.0335   -1.0695  0.3127    
Unemployed    -0.0202    0.0049   -4.1364  0.0025    
Armed.Forces   -0.0103    0.0021   -4.8220  0.0009    
Population    -0.0511    0.2261   -0.2261  0.8262    
Year          1.8292    0.4555    4.0159  0.0030    

Residual standard error: 0.3049 on 9 degrees of freedom
Multiple R-Squared:  0.9955 
F-statistic: 330.3 on 6 and 9 degrees of freedom, the p-value is 4.984e-010
>
> anova(longley.lm)
Analysis of Variance Table

Response: Employed

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq F Value    Pr(F)    
GNP.deflator 1  174.3974 174.3974 1876.533 0.000000000
          GNP 1    4.7872  4.7872  51.511 0.00005211
          Unemployed 1   2.2640  2.2640  24.361 0.00080706
          Armed.Forces 1   0.8764  0.8764   9.430 0.01333568
          Population 1   0.3486  0.3486   3.751 0.08475523
          Year 1    1.4988  1.4988  16.127 0.00303680
          Residuals 9    0.8364  0.0929
>
> vif( Employed ~ . , data=longley)
          GNP Unemployed Armed.Forces Population      Year
135.5324 1788.513   33.61889      3.58893   399.151 758.9806

```

---

## 9.12 Collinearity

Collinearity, also called multicollinearity, is a condition where the model's predictors variables are highly intercorrelated. A consequence of this situation is the inability to estimate the model's regression coefficients with acceptable precision. Therefore, models with this problem are not considered useful. It is unacceptable to reach a final model that has this condition to an appreciable extent.

Collinearity arises when investigators include predictors carrying redundant information in the model. A symptom is a model with a high  $R^2$ , showing that collectively the predictors bear heavily on the response, but paradoxically, few or none of the predictors have regression coefficients significantly different from zero.

Consider the case of a single response  $Y$  and two predictors  $X_1$  and  $X_2$ . The fitted model plots as a plane in the 3-dimensional space of  $(Y, X_1, X_2)$ . A near-collinear situation exists if the correlation between  $X_1$  and  $X_2$  is close to  $\pm 1$ . Geometrically, this occurs when the data points congregate close to a (2-dimensional) straight line when plotted in the 3-dimensional space. When this happens, the points can be fitted fairly well by any plane containing this straight line. Since each of these many planes is a candidate for the best model, the model decided upon as being *the* best will be similar to other model candidates. Therefore, declaring any model to be best will be a tentative decision. This tentativeness is expressed by large standard errors of the estimated regression coefficients that comprise the coefficients of the plane corresponding to the best model.

Figure 9.8, based on a portion of the Longley data introduced in Section 9.11, illustrates these ideas. Here the variables **GNP** and **Year** are almost perfectly correlated and so the scattering of points falls close to a line in 3-dimensional space. Many planes fit this line approximately equally well. The uncertainty about the best fitting of these many planes causes the coefficients of the estimated plane, the regression coefficients, to have large standard errors.

When there are more than two predictors, the geometric argument extends to discussions of hyperplanes and the consequence is again unacceptably large standard errors of regression coefficients.

Although collinearity limits our ability to accurately model the relationship between the predictors and the response, it does not necessarily impede our ability to use the predictors to predict the response. In the context of the example associated with Figure 9.8, if we want to predict the response for values of the predictors near the straight line in 3-dimensional space, many

### Least squares with two highly collinear X-variables

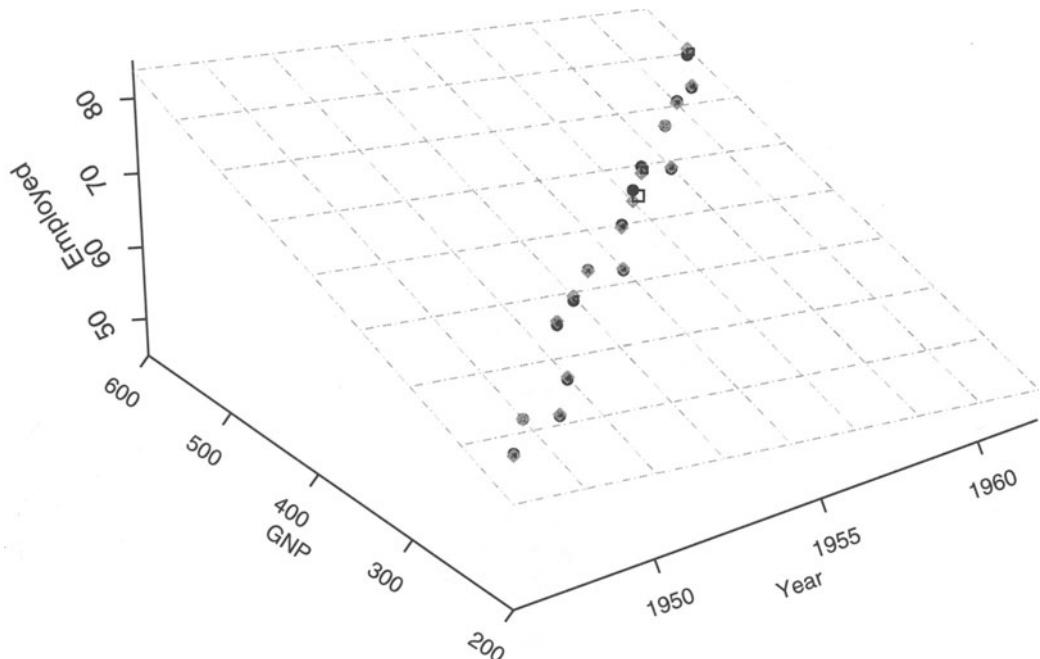


FIGURE 9.8. The two  $X$ -variables, **Year** and **GNP**, are highly collinear. The response variable **Employed** is essentially on a straight line in the three-dimensional space of the figure. The specific plane displayed is almost arbitrary. Any plane that goes through the straight line of the observed points on the plane we see would work just as well.

(regb/code/collinear.s), (regb/figure/longley.collinear.eps.gz)

planes that are good fits to this straight line will yield roughly the same prediction.

A simple diagnostic of collinearity is the *variance inflation factor*, VIF, one for each regression coefficient (other than the intercept). Since the condition of collinearity involves the predictors but not the response, this measure is a function of the  $X$ 's but not of  $Y$ . The VIF for predictor  $i$  is  $1/(1 - R_i^2)$ , where  $R_i^2$  is the  $R^2$  from a regression of predictor  $i$  against the remaining predictors. If  $R_i^2$  is close to 1, this means that predictor  $i$  is well explained by

a linear function of the remaining predictors, and, therefore, the presence of predictor  $i$  in the model is redundant. Values of VIF exceeding 5 are considered evidence of collinearity: The information carried by a predictor having such a VIF is contained in a subset of the remaining predictors. If, however, all of a model's regression coefficients differ significantly from 0 ( $p$ -value  $< .05$ ), a somewhat larger VIF may be tolerable.

VIF is an imperfect measure of collinearity. Occasionally the condition can be attributable to more complicated relationships among the predictors than VIF can detect.

The best approach for alleviating collinearity is to reduce the set of predictors to a noncollinear subset. Methods for accomplishing this are presented in Section 9.13. An ad hoc (manual) procedure, presented in Section 9.13.1, involves eliminating predictors one at a time, at each stage deleting the predictor having the highest VIF. If two predictors are almost tied for highest, then subject area information should be used to choose between them. Proceed until all remaining predictors have  $\text{VIF} \leq 5$ . Other approaches (not discussed in this book) include ridge regression and regression on principal components (Gunst and Mason, 1980).

For the regression analysis of the Longley data, evidence of collinearity appears in Table 9.7 in the variance inflation factors (VIF) for the six predictors. Five of these exceed 33. The next section discusses an approach for dealing with multicollinearity.

Collinearity often arises in polynomial regression models discussed in Section 9.8 because polynomials can be approximated by linear functions within a restricted domain. To avoid both collinearity in polynomial models and numerical instability caused by working with variables of greatly differing orders of magnitude, it is recommended to recenter the response variable to have mean = 0 prior to initiating a polynomial modeling.

## 9.13 Variable Selection

In building a regression model the analyst should consider for use any explanatory variable that is likely to bear upon the response while avoiding the use of two explanatory variables that carry essentially the same information. For example, in modeling the monthly cost of energy needed to heat a 2000-square-foot home, one should avoid using both the mean monthly exterior temperature and the heating degree days (a measure used by heating fuel suppliers) in the same model. The use of redundant explanatory variables is likely to lead to a model with unacceptable collinearity having large standard errors for the predictor regression coefficients.

When subject area theory does not suggest a parsimonious model (i.e., one with relatively few predictors), it is tempting to construct a model using all possibly relevant predictors for which data are available. However, doing so is again likely to result in a collinearity problem. In such circumstances, how can the analyst decide on an appropriate subset of the candidate predictors for a regression model?

Stepwise regression is a tool for answering this question. But this mechanical technique should not be used in order to avoid careful thought about potentially useful predictor variables. Careless use of stepwise regression can, to some extent, distort the significance and confidence levels of inferences in the ultimately specified model, potentially leading to erroneous conclusions. In addition, a model that makes reasonable subject area sense to the client is much preferred to an equally well-fitting one that is less intuitive and harder to understand and explain.

In our experience, a careful systematic approach can often be used to develop a more interpretable model than one produced by a mechanical stepwise algorithm. The starting point is a scatterplot matrix that, along with examination of variance inflation factors, can be used to identify redundant predictors. If two predictors are seen to be highly correlated, we prefer to avoid using the one that has a less obvious subject matter connection to the response variable. An algorithm cannot make such a judgment. Inspection of sploms invite the analyst to consider whether an original variable should be transformed before inclusion in the model. Nevertheless, stepwise approaches to model selection continue to be commonly used, particularly when there are a large number of potential predictors and the analyst has minimal feel for which variables should be or need not be included in the model.

We discuss in turn two systematic methods for model selection, a manual approach and an automated approach, and apply both methods to the Longley data.

### 9.13.1 Manual Use of the Stepwise Philosophy

The first approach involves manual inspections of the VIFs, the  $p$ -values associated with the  $t$ -tests on the regression coefficients, and any available subject matter information to eliminate variables one at a time until a final model is reached with all predictors significant and all VIFs under 5. This approach is viable if the number of predictors is small as in this example. It would be too cumbersome in a situation with more than 12 to 15 predictors.

The three largest VIFs belong to **GNP**, **Year**, and **Population**. The splom implies that they carry almost identical information. We begin by removing

TABLE 9.8. Longley data regression. Best five-predictor model after eliminating one predictor using the manual stepwise approach.  
 (regb/code/longley.back.s)

---

```

S-PLUS (regb/transcript/longley3.st):
> longley3.lm <- lm( Employed ~ GNP.deflator + GNP + Unemployed
+                               + Armed.Forces + Year, data=longley)
> summary(longley3.lm, corr=F)

Call: lm(formula = Employed ~ GNP.deflator + GNP + Unemployed
+       + Armed.Forces + Year, data = longley)

Residuals:
    Min      1Q  Median      3Q     Max
-0.39 -0.143 -0.0356  0.0973  0.461

Coefficients:
            Value Std. Error   t value Pr(>|t|)
(Intercept) -3564.922    772.386   -4.615  0.001
GNP.deflator    0.028     0.061    0.456  0.658
GNP          -0.042     0.018   -2.391  0.038
Unemployed    -0.021     0.003   -6.945  0.000
Armed.Forces   -0.010     0.002   -5.207  0.000
Year          1.869     0.399    4.680  0.001

Residual standard error: 0.29 on 10 degrees of freedom
Multiple R-Squared: 0.995
F-statistic: 438 on 5 and 10 degrees of freedom, the p-value is 2.27e-011

> vif( Employed ~ GNP.deflator + GNP + Unemployed
+           + Armed.Forces + Year, data=longley)

  GNP.deflator    GNP Unemployed Armed.Forces   Year
    76.64  546.9        14.29      3.461  644.6

```

---

one of them from the model. We choose to eliminate **Population** because the *t*-test that its regression coefficient is zero has a larger *p*-value than the tests for either **GNP** or **Year**.

The analysis with all predictors except population appears in Table 9.8.

The outstanding feature of this model is the high *p*-value associated with **GNP.deflator**. Its VIF is well in excess of 5. We proceed with an analysis eliminating **GNP.deflator** in Table 9.9.

TABLE 9.9. Longley data regression. Best four-predictor model after eliminating two predictors using the manual stepwise approach.  
 (regb/code/longley.back.s)

```
S-PLUS (regb/transcript/longley4.st):
> longley4.lm <- lm(Employed ~ GNP + Unemployed + Armed.Forces
+ Year, data=longley)
> summary(longley4.lm, corr=F)

Call: lm(formula = Employed ~ GNP + Unemployed + Armed.Forces + Year, data=longley)
Residuals:
    Min      1Q  Median      3Q     Max 
-0.422 -0.125 -0.0242  0.0837  0.453 

Coefficients:
            Value Std. Error   t value Pr(>|t|)    
(Intercept) -3598.729    740.633   -4.859   0.001    
GNP          -0.040     0.016    -2.440   0.033    
Unemployed   -0.021     0.003    -7.202   0.000    
Armed.Forces -0.010     0.002    -5.522   0.000    
Year         1.887     0.383     4.931   0.000    

Residual standard error: 0.279 on 11 degrees of freedom
Multiple R-Squared:  0.995 
F-statistic: 590 on 4 and 11 degrees of freedom, the p-value is 9.5e-013

> vif( Employed ~ GNP + Unemployed + Armed.Forces + Year, data=longley)
      GNP Unemployed Armed.Forces Year
    515.1      14.11       3.142  638.1
```

All four predictors in this model have significant regression coefficients. However, two of the VIFs are still large, and one of the predictors corresponding to them must be eliminated. We choose to eliminate GNP because its *p*-value, while small, is larger than those of the three other remaining predictors.

The results of the analysis with the remaining predictors **Unemployed**, **Armed.Forces**, and **Year** are in Table 9.10.

This is our tentative final model. The collinearity has been eliminated (all VIFs are below 5), and all regression coefficients differ significantly from zero. In addition,  $R^2 = 0.993$ , so these three predictors account for virtually all of the variability in **Employed**.

TABLE 9.10. Longley data regression. Best three-predictor model after eliminating three predictors using the manual stepwise approach.  
 (regb/code/longley.back.s)

---

```

S-PLUS (regb/transcript/longley5.st):
> longley5.lm <- lm( Employed ~ Unemployed + Armed.Forces + Year, data=longley)
>
> summary(longley5.lm, corr=F)

Call: lm(formula = Employed ~ Unemployed + Armed.Forces + Year, data = longley)
Residuals:
    Min      1Q Median      3Q     Max
-0.573 -0.12  0.0409  0.14  0.753

Coefficients:
            Value Std. Error   t value Pr(>|t|)    
(Intercept) -1797.221    68.642  -26.183   0.000    
Unemployed    -0.015     0.002   -8.793   0.000    
Armed.Forces   -0.008     0.002   -4.204   0.001    
Year          0.956     0.036   26.921   0.000    
                                                        
Residual standard error: 0.332 on 12 degrees of freedom
Multiple R-Squared:  0.993 
F-statistic: 555 on 3 and 12 degrees of freedom, the p-value is 3.92e-013

> vif( Employed ~ Unemployed + Armed.Forces + Year, data=longley)
Unemployed Armed.Forces Year
      3.318        2.223  3.89
  
```

---

### 9.13.2 Automated Stepwise Regression

The second approach to model selection is stepwise regression. This automated approach is recommended when the number of predictors is so large that the manual approach becomes unacceptably laborious. We illustrate here how it is used to reach the same model that we found with the manual procedure. A stepwise approach that examines all subsets of predictors is viable if the number of predictors  $p$  is less than 10 to 12. If  $p > 12$ , then forward selection or backward elimination is preferred.

The three basic methods for automated stepwise regression are

**forward selection:** Predictors are added to the model one at a time until a stopping rule is satisfied.

**backward elimination:** All predictors are initially placed in the model.

Predictors are removed from the model one at a time until a stopping rule is satisfied.

**all subsets:** All  $2^p - 1$  possible models, where  $p$  is the number of predictors, are attempted and the best is identified. This method is viable only for “small” values of  $p$ . Efficient algorithms exist that avoid actually examining every such model.

The literature contains many hybrids and refinements of these basic methods.

Each of the automated stepwise methods uses a criterion for choosing the next step or stopping the algorithm.

Such criteria may relate to appreciable  $R_{\text{adj}}^2$  or  $F$ -statistic improvement or detriment, substantial mean square error decrease or increase, or size of change in Mallows'  $C_p$  statistic discussed below. Another possibility is to look, at each step, at the  $p$ -value for the variables already in the model and for the potential next variable to be brought in to the model. If the largest  $p$ -value of the variables already in the model is larger than the threshold, then remove it. If the smallest  $p$ -value of the potential variables is larger than the threshold, then stop. Otherwise, bring in a new variable and repeat the process.

Computer algorithms allow the option of accepting or overriding default criterion values or thresholds for appreciable change.

Each of the automated stepwise methods uses one or more criteria for choosing among competing models. Here is a list of possible criteria.

$p$  Models containing fewer predictors are easier to interpret and understand. It is desirable that the number of predictors  $p$  be as small as possible.

$\hat{\sigma}^2$  We also require that the predictors account for most of the variability in the response. Equivalently, we wish that the residual mean square,  $MSE = \hat{\sigma}^2$ , be as small as possible, preferably not much larger than for the model containing all candidate predictors. This criterion is easier to meet with more predictors rather than few; hence it asks that the number of predictors  $p$  be as large as possible and competes with the goal of minimizing  $p$ .

The above criteria address one of the two competing objectives at a time. Other criteria jointly address the two objectives.

$R_{\text{adj}}^2$  Unadjusted  $R^2$  is not used as a model selection criterion because it necessarily increases as the number of predictors increases. A model can have  $R^2$  close to 1 but be unacceptable due to severe collinearity.

Instead we use  $R_{\text{adj}}^2$ , which is  $R^2$  adjusted downward for the number of predictors,

$$R_{\text{adj}}^2 = 1 - \left( \frac{n-1}{n-p} \right) (1 - R^2) \quad (9.27)$$

which increases as  $R^2$  increases but provides a penalty for an excessive number of predictors  $p$ . Models with higher  $R_{\text{adj}}^2$  are preferred to ones with lower  $R_{\text{adj}}^2$ .

$C_p$  Mallows'  $C_p$  statistic is another criterion that addresses both the fit of the model and the number of predictors used. Consistent with customary notation, in the context of the  $C_p$  statistic but nowhere else in this chapter,  $p$  is the number of regression coefficient *parameters*, equal to the number of predictors *plus 1*. The original definition is

$$C_p = (\text{SS}_{\text{Res}}/\hat{\sigma}_{\text{full}}^2) + 2p - n \quad (9.28)$$

where  $\text{SS}_{\text{Res}}$  is the residual sum of squares for the reduced model under discussion (fewer  $X$ -variables than the full model) and the  $\hat{\sigma}_{\text{full}}^2$  is the error mean square for the full model containing all candidate predictors. If the extra  $X$ -variables are noise, rather than useful, then the ratio  $\text{SS}_{\text{Res}}/\hat{\sigma}_{\text{full}}^2 \approx ((n-p)\sigma^2)/\sigma_{\text{full}}^2 \approx n-p$ . If the extra  $X$ -variables are useful, then the numerator  $\sigma^2 \gg \sigma_{\text{full}}^2$  and the ratio will be much larger than  $n-p$ . The extra terms  $2p-n$  make the entire  $C_p$  approximate  $p$  when the extra  $X$ -variables are not needed.

A desirable model has  $C_p \approx p$  for a small number of parameters  $p$ . (If  $p_{\max}$  denotes  $p$  for a model containing all candidate predictors, then necessarily  $C_{p_{\max}} = p_{\max}$ , but such a model is almost never acceptable.)  $C_p$  results are often conveyed with a  $C_p$  plot, that is, a plot of  $C_p$  vs  $p$ , with each point labeled with an identifier for its model and the diagonal line having equation  $C_p = p$  added to the plot. Desirable models are those close to this diagonal line.

*AIC* The Akaike information criterion is proportional to the  $C_p$  statistic. The AIC is scaled in sum of squares units.

The S-PLUS function `step` uses the name  $C_p$  to describe one version of the AIC, the Akaike information criterion. At each step it uses the *current* model in the AIC formula in the role usually reserved for the *full* model. We think that labeling their statistic with the  $C_p$  name was a bad choice of terminology. The  $C_p$  was designed by Colin Mallows and Cuthbert Daniel (Mallows, 1973) specifically to have small integer values. Substituting the AIC defeats that purpose.

*F* At each step we can look at the  $p$ -value for the variables already in the model and for the potential next variable to be brought in to the model. If the largest  $p$ -value of the variables already in the model is larger than the threshhold, then remove it. If the smallest  $p$ -value of the potential

variables is larger than the threshold, then stop. Otherwise, bring in a new variable and repeat the process.

SAS implements stepwise regression in PROC REG by specifying the MODEL statement option SELECTION=name, where name is the particular stepwise method to be used.

S-PLUS uses the **stepwise** and **step** functions.

### 9.13.3 Automated Stepwise Modeling of the Longley Data

Table 9.11 contains the results of an S-PLUS stepwise regression analysis considering all subsets of the predictors, with printouts of the properties of two models of each size having smallest residual sum of squares among models having  $C_p < 10$ . Figure 9.9 is a plot of the  $C_p$ -values for all models with  $C_p < 10$ . The acronymic plot symbols in Figure 9.9 are decoded in Table 9.11. According to Table 9.11, the best parsimonious model is the one with the four predictors GNP, Unemployed, Armed.F Forces, and Year displayed in Table 9.9. This model has  $C_p$  close to  $p$ , and a smaller AIC and larger adjusted  $R^2$  than any of the other models in Table 9.11. Unlike the model we selected with our manual approach, this one includes the predictor GNP. The algorithm underlying Table 9.11 suggests inclusion of GNP despite its high correlation with Year and high VIF shown in Table 9.9.

Which model is preferred, the one in Table 9.9 containing four predictors including GNP or the three predictor model in Table 9.10 that excludes GNP? Our answer to this question demonstrates our preference for the manual approach. The coefficient of GNP in Table 9.9 is negative. This model says that holding Unemployed, Armed.F Forces and Year constant, GNP and Employed are *negatively* associated. This statement conflicts with our expectation that this association is positive, and is a strong argument against the four-predictor model in Table 9.9.

Table 9.12 contains an excerpt from a SAS stepwise analysis of the Longley data, produced by requesting a selection considering the best 12 models according to the  $C_p$  criterion alone.

TABLE 9.11. Longley data regression. The model with the three predictors `Unemployed`, `Armed.Forces` and `Year` is competitive with respect to  $C_p$  and other criteria. This model uses fewer predictors than other models shown while having adjusted  $R^2$  almost as large as all other models.  
 (regb/code/longley6.s)

---

```

S-PLUS (regb/transcript/longley6.st):
> longley.step <- stepwise(y=longley$Employed,
+                               x=longley[,c(1:6)],
+                               method="exhaustive",
+                               plot=F, nbest=2)
> ## longley.step ## no need to print this. longley.cp is more legible.
>
> longley.cp <- cp.calc(longley.step, longley, "Employed")
> tmp <- (longley.cp$cp <= 10)
> longley.cp[tmp,]
response variable = Employed
total sum of squares = 185.0088
number of observations = 16
full model is row GGUAPY
      p    cp    aic    rss    r2 r2.adj
..UA.Y 4 6.239 2.067 1.3234 0.9928 0.9911
.GUA.Y 5 3.239 1.788 0.8587 0.9954 0.9937
..UAPY 5 4.606 1.915 0.9857 0.9947 0.9927
.GUAPY 6 5.031 1.955 0.8393 0.9955 0.9932
GGUA.Y 6 5.051 1.956 0.8412 0.9955 0.9932
GGUAPY 7 7.000 2.138 0.8364 0.9955 0.9925

          xvars sw.names
..UA.Y           Unemployed,Armed.Forces,Year 3(#1)
.GUA.Y           GNP,Unemployed,Armed.Forces,Year 4(#1)
..UAPY           Unemployed,Armed.Forces,Population,Year 4(#2)
.GUAPY           GNP,Unemployed,Armed.Forces,Population,Year 5(#1)
GGUA.Y           GNP.deflator,GNP,Unemployed,Armed.Forces,Year 5(#2)
GGUAPY GNP.deflator,GNP,Unemployed,Armed.Forces,Population,Year 6(#1)
>
> old.par <- par(mar=par()$mar+c(0,1,0,0))
> plot(cp ~ p, data=longley.cp[tmp,], ylim=c(0,10), type="n", cex=1.3)
> abline(b=1)
> text(x=longley.cp$p[tmp], y=longley.cp$cp[tmp],
+       row.names(longley.cp)[tmp], cex=1.3)
> title(main="Cp plot for longley.dat, Cp<10")
> par(old.par)
> ## export.eps(hh("regb/figure/regb.f4.longley.eps"))

```

---

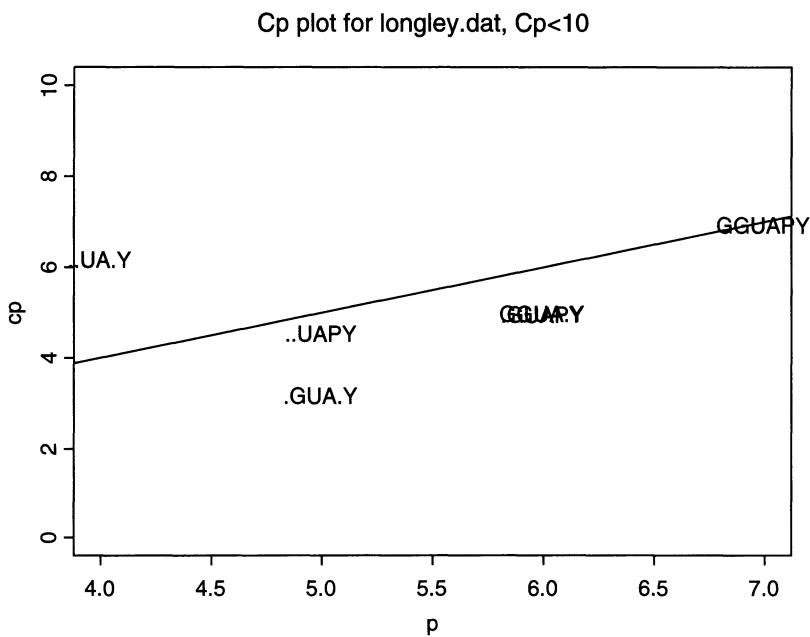


FIGURE 9.9.  $C_p$  Plot for Longley data. See Table 9.11 for interpretations of the acronyms used to label points. The overplotting occurs because, as seen in Table 9.11, two models have identical values of  $C_p$ .

(regb/code/longley6.s), (regb/figure/regb.f4.longley.eps.gz)

TABLE 9.12. Stepwise analysis of the Longley data using the  $C_p$  criterion.  
 (regb/code/longley.sas)

---

SAS (regb/code/longley2.sas):

```
proc reg data=longley;
  model Employed = Deflator GNP Unemployed ArmedForces Population Year /
    VIF selection=cp best=12;
run;
```

---

SAS (regb/transcript/longley2.lst):  
 Dependent Variable: Employed

C(p) Selection Method

| Number in Model | C(p)    | R-Square | Variables in Model                                  |
|-----------------|---------|----------|---|
| 4               | 3.2395  | 0.9954   | GNP Unemployed ArmedForces Year                     |
| 4               | 4.6064  | 0.9947   | Unemployed ArmedForces Population Year              |
| 5               | 5.0315  | 0.9955   | GNP Unemployed ArmedForces Population Year          |
| 5               | 5.0511  | 0.9955   | Deflator GNP Unemployed ArmedForces Year            |
| 5               | 6.1439  | 0.9949   | Deflator Unemployed ArmedForces Population Year     |
| 3               | 6.2395  | 0.9928   | Unemployed ArmedForces Year                         |
| 6               | 7.0000  | 0.9955   | Deflator GNP Unemployed ArmedForces Population Year |
| 4               | 8.2257  | 0.9929   | Deflator Unemployed ArmedForces Year                |
| 4               | 19.4648 | 0.9872   | GNP Unemployed ArmedForces Population               |
| 5               | 21.1274 | 0.9874   | Deflator GNP Unemployed ArmedForces Population      |
| 4               | 21.2586 | 0.9863   | Deflator GNP ArmedForces Population                 |
| 3               | 21.6625 | 0.9851   | GNP Unemployed ArmedForces                          |

---

## 9.14 Residual Plots

Partial residual plots and added variable plots are visual aids for interpreting relationships between variables used in regression and can serve as additional components of our manual approach for variable selection.

Figure 9.10 shows four different types of plots.

- Row 1 shows the response variable  $Y = \text{Employed}$  against each of the six predictors  $X_j$ .
- Row 2 shows the ordinary residuals  $e = Y - \hat{Y}$  from the regression on all six variables against each of the six predictors.
- Row 3 shows the “partial residual plots”, the partial residuals  $e^j$  for each predictor against that predictor. See Section 9.14.1 for construction of the partial residuals and Section 9.14.2 for construction of the partial residual plots.
- Row 4 shows the “added variable plots”, the partial residuals  $e^j$  against the partial residuals  $X_{j|1,2,\dots,j-1,j+1,\dots,p}$  of  $X_j$  regressed on the other five predictors. See Section 9.14.3 for the definition of partial correlation, and Section 9.14.4 for construction of the  $X_{j|1,2,\dots,j-1,j+1,\dots,p}$  and the added variable plots.

We discuss the interpretation of the all four types of plots in Section 9.14.5. We recommend the discussions of partial residual plots and added variable plots in (Weisberg, 1985) and (Hamilton, 1992). See also the help page in S-PLUS for `partial.plot`.

### 9.14.1 Partial Residuals

The partial residuals  $e^j$  for variable  $X_j$  in a model with  $p$  predictor variables  $X_j$  are defined

$$e^j = Y - \hat{Y}_{1,2,\dots,j-1,j+1,\dots,p} \quad (9.29)$$

and calculated with

$$e^j = X_j \hat{\beta}_j + e \quad (9.30)$$

or equivalently

$$e_i^j = X_{ij} \hat{\beta}_j + e_i \quad \text{for } i = 1, \dots, n \quad (9.31)$$

where  $e = (e_i)$  are the ordinary residuals from the model with all  $p$  predictors

$$e = Y - \hat{Y}_{1,2,\dots,p} \quad (9.32)$$

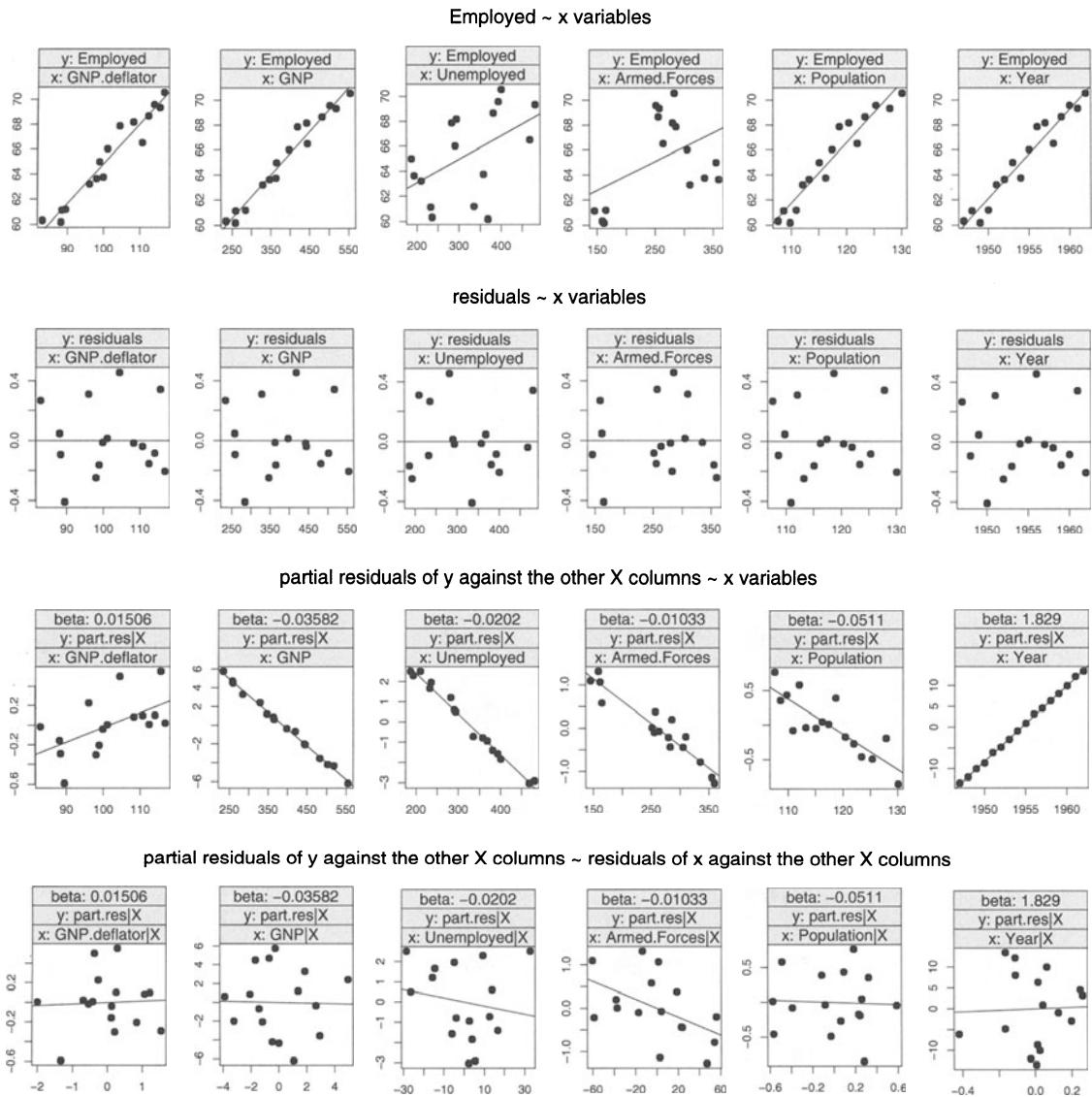


FIGURE 9.10. Four types of plots for the regression of the Longley data against all six potential predictors: Response variable  $Y$  against each  $X$ , residuals  $e$  against each  $X$ , partial residual plots, added variable plots.

(`regb/code/longley.s`), (`splus.library/residual.plots.s`),  
 (`regb/figure/longley.resid.eps.gz`)

The partial residuals are interpreted as the additional information available for  $X_j$  to pick up after all  $X$  except  $X_j$  have been included in the model.

### 9.14.2 Partial Residual Plots

Partial residual plots are the set of plots of  $e^j$  against  $X_j$  for all  $j$ .

We show the partial residual plots for the Longley data in Row 3 of Figure 9.10.

### 9.14.3 Partial Correlation

The partial correlation  $r(X_1, X_2|X_3, X_4, X_5)$  between  $X_1$  and  $X_2$ , after correction for the effect of  $X_3, X_4, X_5$ , is the correlation coefficient between  $X_1$  and  $X_2$  after the (linear) effects of  $X_3, X_4, X_5$  have been removed from both  $X_1$  and  $X_2$ . When  $X_1$  through  $X_5$  are multivariate data, we can compute the sample partial correlation coefficient as follows:

- Regress  $X_1$  on  $X_3, X_4, X_5$ . Get the residuals  $E_1$ .
- Regress  $X_2$  on  $X_3, X_4, X_5$ . Get the residuals  $E_2$ .
- Find the (usual) correlation coefficient between  $E_1$  and  $E_2$ . This turns out to be  $r(X_1, X_2|X_3, X_4, X_5)$ .

In SAS, we can obtain this partial correlation via

```
proc corr;
  var X1 X2;
  partial X3 X4 X5;
```

In S-PLUS, we use

```
partial.corr(cbind(X1,X2),
             cbind(X3,X4,X5))
```

and the function `partial.corr` defined in (`splus.library/partial.corr.s`).

### 9.14.4 Added Variable Plots

The added variable plots are the set of plots of  $E_1 = e^j$  against  $E_2 = X_{j|1,2,\dots,j-1,j+1,\dots,p}$  for all  $j$ . We define  $\hat{X}_{1,2,\dots,j-1,j+1,\dots,p}$  to be the predicted value of  $X_j$  after regressing  $X_j$  against all the other  $X$ -variables in the model. We define the residual

$$X_{j|1,2,\dots,j-1,j+1,\dots,p} = X_j - \hat{X}_{1,2,\dots,j-1,j+1,\dots,p} \quad (9.33)$$

to be the additional information in  $X_j$  after removing the information provided by all the other  $X$  in the model. Thus the added variable plots are the plots of the  $E_1$  and  $E_2$  defined by regressing  $Y$  and  $X_j$  against all the other  $X$ -variables.

We show the added variable plots for the Longley data in Row 4 of Figure 9.10.

### 9.14.5 Interpretation of Residual Plots

#### 9.14.5.1 Response Variable Against Each of the Predictors

Row 1 of Figure 9.10, the plots of the response variable  $Y = \text{Employed}$  against each of the six predictors  $X_j$ , is almost identical to the top row of the splom in Figure 9.7. The only difference is the explicit one- $x$  regression line in Figure 9.10. If there is no visible slope in any of these panels, then we can effectively eliminate that  $x$ -variable from further consideration as a potential explanatory variable. This row is essentially the same as the first step of a stepwise-forward procedure. In this example, we cannot eliminate any of the potential predictors at this stage.

#### 9.14.5.2 Residuals Against Each of the Predictors

Row 2 of Figure 9.10, the plots of the ordinary residuals  $e = Y - \hat{Y}$  from the complete regression of the response against all six potential predictors  $X_j$ , shows horizontal slopes. This is by construction, as the least-squares residuals are orthogonal to all  $X$ -variables. In this example, we see no structure in the plots. The types of structure we look for are

**Curvature.** Plot the residuals from the quadratic fit in the left side of Figure 9.5 against the predictor `density` and note that the residuals are predominantly above the  $y = 0$  axis at the left and right ends of the range and predominantly below the axis in the middle of the range. Curvature in the residual plots often suggests that additional predictors, possibly powers of existing predictors, are needed in the model.

**Nonuniformity of variance.** See the `life.exp ~ ppl.per.tv` panel of Figure 4.13 where we see high variability in `life.exp` for low values of `ppl.per.tv` and very low variability for high values of `ppl.per.tv`. Nonuniformity of variance in the residual plots often suggests power transformations of one or more of the variables. Transformations of both the response and predictor variables need to be considered.

**Bunching or granularity.** See the `residuals ~ lime` panel of Figure 11.11 where we see that `lime` has only two levels and there are different variances for each.

#### 9.14.5.3 Partial Residuals

Both Rows 3 and 4 use the partial residuals of the response as the  $y$ -variable of each plot. Since “partial” means “adjusted for all the other  $x$ -variables”, each column of Rows 3 and 4 is different. Column 1 is adjusted for  $X_2, X_3, \dots, X_6$ . Column 2 is adjusted for  $X_1, X_3, \dots, X_6$ . Similarly through Column 6, which is adjusted for  $X_1, \dots, X_5$ .

In Row 3, the *partial residual plots*, the  $x$ -variables are the observed  $x$ -variables  $X_j$ .

In Row 4, the *added residual plots*, the  $x$ -variables are the adjusted- $x$  variables, that is, “adjusted for all the other  $x$ -variables”. Thus the  $x$ -variable in Column 1 of Row 4 is  $X_{1|2,\dots,6}$ , that is,  $X_1$  adjusted for  $X_2, \dots, X_6$ .

In both Rows 3 and 4 the slope of the two-dimensional least-squares line in panel  $j$  is exactly the value of the regression coefficient  $\beta_j$  for the complete regression of  $Y$  on all the  $X$ -variables in the model.

#### 9.14.5.4 Partial Residual Plots

In Row 3, the partial residuals  $e^j$  are plotted against the observed  $x$ -variables  $X_j$ . Since the partial residuals  $e^j$  are specific to each  $X_j$ , the values for the  $y$ -range are unique to each panel. The  $x$ -range of the  $x$ -variables in Row 3 is the same as it is in Rows 1 and 2 of this display.

We look for the tightness of the points in each plot around their least-squares line. High variability around the two-dimensional least-squares line indicates low significance for the corresponding regression coefficient. Low variability around the least-squares line indicates a significant regression coefficient.

In Row 3 of Figure 9.10, we see that Columns 1 (**GNP.deflator**) and 5 (**Population**) have high variability around their least-squares lines. This is a reflection of the high  $p$ -value that we see for those regression coefficients in Table 9.7. The remaining four columns all look like their points are tightly placed against their least-squares lines, an indication of possible significance. Note that Column 2 (**GNP**) looks tight, even though its  $p$ -value is the nonsignificant 0.3127. We really do need the tabular results to completely understand what the graph is showing us.

#### 9.14.5.5 Added Variable Plots

In Row 4, the partial residuals  $e^j$  are plotted against the adjusted  $x$ -variables  $X_{j|1,2,\dots,j-1,j+1,\dots,p}$ . In Row 4, both the  $x$ - and  $y$ -variables in each column have been adjusted for all the other  $X$ -variables. Therefore, both the  $x$ - and  $y$ -ranges are unique to each panel. The partial residuals, the

$y$ -variables in the added variable plots, are identical to the  $y$ -variables in the partial residual plots; hence the  $y$ -ranges are identical for corresponding columns of Rows 3 and 4.

We look at the slope of the two-dimensional least-squares line in each plot. A nearly horizontal line indicates low significance for the corresponding regression coefficient. A nonzero slope indicates a significant regression coefficient.

The three  $x$ -variables with significant regression coefficients in Table 9.7 have visible nonzero slopes to their least-squares lines in Row 4 of Figure 9.10. The three  $x$ -variables with nonsignificant regression coefficients have almost horizontal least-squares lines.

## 9.15 Example—U.S. Air Pollution Data

Exercise 4.2 introduces the data set (`datasets/usair.dat`), which concerns causes of air pollution in U.S. cities. A scatterplot matrix of these data appears in Figure 9.11a. Here we seek to develop a model to explain the response `S02`,  $\text{SO}_2$  content of air, using a subset of six available explanatory variables.

In Figure 9.11b we see that the three variables `S02`, `mfgfirms`, and `popn` are all pushed against their minimum value with a long tail toward the maximum value. This pattern suggests a log transformation to bring these three distributions close to symmetry. Following these transformations, Figure 9.11b shows the new response variable `lnS02` and the revised list of six potential explanatory variables.

For pedagogical purposes we approach this problem in two different ways. We first use the automated stepwise regression approach and then consider the manual approach.

We illustrate the automated approach with the `stepwise` command in S-PLUS, requesting the `exhaustive` method that considers all subsets. In this problem there are only a small number,  $2^6 - 1 = 31$ , of subsets to consider, so this method is viable. We request the best two subsets for each possible value of the number of included explanatory variables. The tabular and graphical results of the stepwise analysis are displayed in Table 9.13 and Figure 9.12. The model with the four predictors `temp`, `lnmfg`, `wind`, and `precip` seems best. It has  $C_p \approx p$ , the smallest AIC of contenders, the largest  $R_{\text{adj}}^2$ , and one of the smallest values of  $\text{SS}_{\text{Res}}$ .

In Table 9.14 we look at the detail for the selected model. We observe that all VIFs are small and the  $p$ -values are below 0.01 for all model coefficients. The signs of the estimated coefficients are reasonable or defensible.

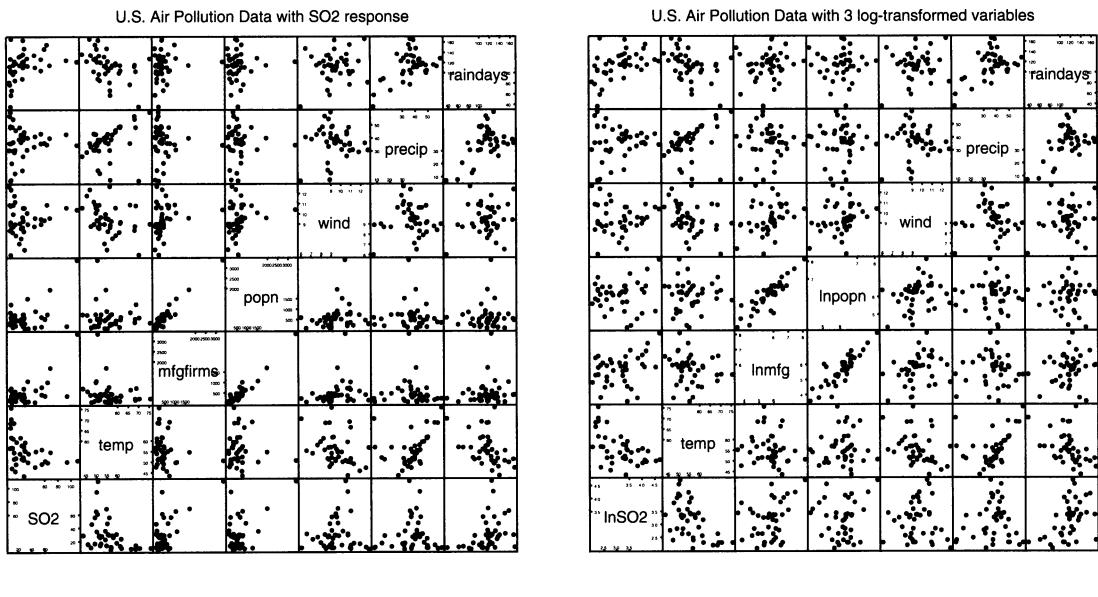


FIGURE 9.11. Scatterplot matrices for air pollution data (a) with original scaling and (b) after a log transformation of three variables that succeeds in improving the symmetry of the distributions.  
`(regb/code/usair.s)`, `(regb/figure/regb.f1.usair.eps.gz)`,  
`(regb/figure/regb.f2.usair.eps.gz)`.

United States cities with high average annual temperature are located in the Sunbelt and tend to have less pollution-causing heavy industry than colder temperature cities well north of the Sunbelt. We are not surprised that greater amounts of manufacturing are associated with more pollution or that wind dissipates pollution.

We can arrive at the same model without a formal stepwise approach. We notice from Figure 9.11b that `lnmfg` and `lnpopn` are highly correlated, so it would be redundant to include both in the model. The variables `precip` and `raindays` seem quite similar, so again, it is unlikely that both are needed. Inspection of the  $C_p$  plot in Figure 9.12 indicates that the model with `temp`, `lnmfg`, `wind`, and `precip` has  $C_p$  close to  $p$  and only one member of each pair of similar predictors.

TABLE 9.13. S-PLUS stepwise regression analysis of U.S. air pollution data. See also Figure 9.12.  
(regb/code/usair.s)

---

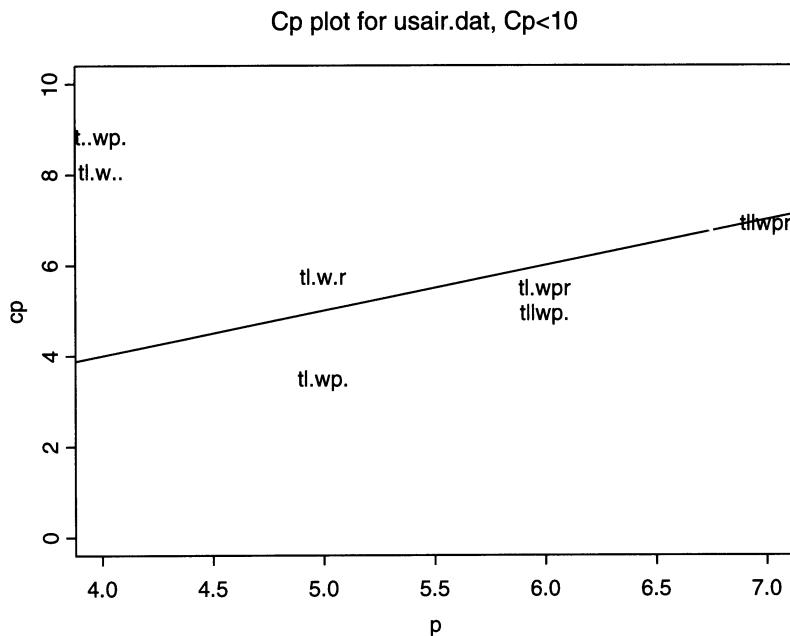
S-PLUS (regb/transcript/usair2.st):

```

> usair.step <- stepwise(y=usair$lnSO2,
+                         x=usair[, c(2,9,10,5:7)],
+                         method="exhaustive",
+                         plot=F, nbest=2)
> usair.cp <- cp.calc(usair.step, usair, "lnSO2")
> tmp <- (usair.cp$cp <= 10)
> usair.cp[tmp,-8]
response variable = lnSO2
total sum of squares = 19.72893
number of observations = 41
full model is row tllwpr
      p    cp     aic     rss     r2 r2.adj      xvars
tl.w.. 4 8.148 12.82 10.736 0.4558 0.4117 temp,lnmfg,wind
t..wp. 4 8.925 13.03 10.939 0.4455 0.4006 temp,wind,precip
tl.wp. 5 3.577 11.63  9.022 0.5427 0.4919 temp,lnmfg,wind,precip
tl.w.r 5 5.823 12.22  9.608 0.5130 0.4589 temp,lnmfg,wind,raindays
tllwp. 6 5.025 12.01  8.878 0.5500 0.4857 temp,lnmfg,lnpopn,wind,precip
tl.wpr 6 5.575 12.15  9.021 0.5427 0.4774 temp,lnmfg,wind,precip,raindays
tllwpr 7 7.000 12.52  8.871 0.5503 0.4710 temp,lnmfg,lnpopn,wind,precip,raindays

```

---



```
S-PLUS (regb/transcript/usair3.st):
> tmp <- (usair.cp$cp <= 10)
> plot(cp ~ p, data=usair.cp[tmp,], ylim=c(0,10), type="n", cex=1.3)
> abline(b=1)
> text(x=usair.cp$p[tmp], y=usair.cp$cp[tmp],
+       row.names(usair.cp)[tmp], cex=1.3)
> title(main="Cp plot for usair.dat, Cp<10")
```

FIGURE 9.12.  $C_p$  plot. Model “tl.wp.” (`temp,lnmfg,wind,precip`) has the smallest  $C_p$  and the largest  $R_{adj}^2$ . See also Table 9.13.  
`(regb/code/usair.s)`, `(regb/figure/regb.f3.usair.eps.gz)`

TABLE 9.14. Fit of recommended model for U.S. air pollution data.  
 (regb/code/usair.s)

---

```
S-PLUS (regb/transcript/usair5.st):
> usair.lm <- lm(lnS02 ~ temp + lnmfg + wind + precip, data=usair)
> summary(usair.lm)

Call: lm(formula = lnS02 ~ temp + lnmfg + wind + precip, data = usair)
Residuals:
    Min      1Q  Median      3Q     Max 
-0.8965 -0.3405 -0.0854  0.2963  1.032 

Coefficients:
            Value Std. Error t value Pr(>|t|)    
(Intercept) 6.8914   1.0701    6.4400  0.0000    
temp        -0.0730   0.0128   -5.6911  0.0000    
lnmfg       0.2400   0.0868    2.7659  0.0089    
wind        -0.1844   0.0620   -2.9725  0.0052    
precip      0.0193   0.0074    2.6156  0.0129    
                                                        
Residual standard error: 0.5006 on 36 degrees of freedom
Multiple R-Squared:  0.5427
F-statistic: 10.68 on 4 and 36 degrees of freedom, the p-value is 8.233e-006

> vif(lm(lnS02 ~ temp + lnmfg + wind + precip, data=usair)
      temp     lnmfg      wind      precip
      1.373342 1.115383 1.253309 1.204027
```

---

## 9.16 Exercises

We recommend that for all exercises involving a data set, you begin by examining a scatterplot matrix of the variables.

- 9.1.** Use matrix algebra to prove the assertion in Equation (9.11) that the sum of the calculated residuals is also zero in multiple regression. We proved the assertion for simple linear regression in Exercise 8.9.

Hint: Write the vector of residuals as  $e = (I - H)Y$ , verify that  $X = HX$ , and use the fact that in a model with a nonzero intercept coefficient, as in Equation (9.1) and following, the first column of  $X$  is a column of ones.

- 9.2.** (Davies and Goldsmith, 1972), reprinted in (Hand et al., 1994), investigated the relationship between the **abrasion** loss of samples of rubber (in grams per hour) as a function of **hardness** and tensile **strength** ( $\text{kg}/\text{cm}^2$ ). Higher values of **hardness** indicate harder rubber. The data appear in the file (**datasets/abrasion.dat**).

- a. Produce a scatterplot matrix of these data. Based on this plot, does it appear that **strength** would be helpful in explaining **abrasion**?
- b. Calculate the fitted regression equation.
- c. Find a 95% prediction interval for the abrasion corresponding to a new rubber sample having hardness 60 and strength 200.

- 9.3.** (Narula and Wellington, 1977) provide data on the sale price of 28 houses in Erie, Pennsylvania, in the early 1970s, along with 11 possible predictors of these prices. See (**datasets/houseprice-erie.dat**). The variables are:

**price**: price in \$100's  
**taxes**: taxes in dollars  
**bathrm**: number of bathrooms  
**lotsize**: lot size in square feet  
**sqfeet**: square footage of living space  
**garage**: number of cars for which there is garage space  
**rooms**: number of rooms  
**bedrm**: number of bedrooms  
**age**: age in years

**type:** type of house  
1 = brick, 2 = brick and frame, 3 = aluminum and frame, 4 = frame  
**style:** 1 = 2 story, 2 = 1.5 story, 3 = ranch  
**fireplac:** number of fireplaces

In parts a–d, exclude factors **type** and **style** from the analysis.

- a. Produce a scatterplot matrix for these data. Notice that two houses had a sale price much higher than the others.
  - b. Use a stepwise regression technique to formulate a parsimonious model for sale price. Do the arithmetic signs of your model's regression coefficients make economic sense?
  - c. Redo part a with the two large-priced houses excluded. Compare your answer with that of part a.
  - d. Add a new variable **sqfeetsq** (defined as the square of **sqfeet**) to the list of variables. Perform the stepwise regression allowing for this new variable. Does its presence change the preferred model?
- 9.4.** (World Almanac and Book of Facts, 2001) lists the winning **times** for the men's 1500-meter sprint event for the Olympics from **years** 1900 through 2000. The data are in the file (**datasets/sprint.dat**).
- a. Plot the data.
  - b. Use linear regression to fit the winning times to the year, producing a plot of the residuals vs the fitted values.
  - c. The residual plot suggests that an additional predictor should be added to the model. Refit this expanded model and compare it with the model you found in part b.
  - d. Interpret the sign of the coefficient of this additional predictor.
- 9.5.** A company wished to model the number of **minutes** required to unload shipments of drums of chemicals at its warehouse as a function of the number of **drums** and the total shipment **weight** in hundreds of pounds.

The data from 20 consecutive shipments, from (Neter et al., 1996), are in the file (`datasets/shipment.dat`). 2.

- a. Regress `minutes` on `drums` and `weight`, storing the residuals.
  - b. Interpret the regression coefficients of `drum` and `weight`.
  - c. Provide and discuss plots of the residuals against the fitted values and both predictors, and a normality plot.
  - d. Provide a 90% prediction interval for the time it would take to unload a new shipment of 10 drums weighing 1000 pounds.
- 9.6.** The dataset (`datasets/uscrime.dat`) is introduced in Exercise 4.3. Use a stepwise regression approach to develop a model to explain `R`. Your solution should not have a collinearity problem, all predictor regression coefficients should be significantly different from zero and have an arithmetic sign consistent with common knowledge of the model variables, and no standard residual plots should display a problem.
- 9.7.** It is desired to model the `manhours` needed to operate living quarters for U.S. Navy bachelor officers. Candidate explanatory variables are listed below. The data in the file (`datasets/manhours.dat`) are from (Freund and Littell, 1991). Perform a thorough regression analysis, including relevant plots. Note that at least initially, there is a minor collinearity problem to be addressed. Show that, no matter how the collinearity is addressed, the predictions are similar. Only the interpretation of the effects of the  $x$ -variables is affected.

`manhours`: monthly manhours needed to operate the establishment

`occupanc`: average daily occupancy

`checkins`: average monthly number of check-ins

`svcdesk`: weekly hours of service desk operation

`sqfeet`: square footage of living space

`garage`: number of cars for which there is garage space

`common`: common use area, in square feet

`wings`: number of building wings

`berthing`: operational berthing capacity

`rooms`: number of rooms

# Multiple Regression—Dummy Variables and Contrasts

Any analysis of variance model (for example, anything in Chapters 6, 12, 13, or 14) can be expressed as a regression with dummy variables. Many software procedures and functions make explicit use of this form of expression. Here we explore this equivalence. The notation in Chapter 10 is that used in Sections F.4.2, 9.2, and 9.5.1.

Section 10.1 introduces dummy variables. Section 10.3 looks at the equivalence of different sets of dummy variable codings for factors. Section 13.5 shows how the S-PLUS and SAS languages express the dummy variable coding schemes. Table 13.22 shows the notation for applying them to describe models with two or more factors.

## 10.1 Dummy (Indicator) Variables

Dummy variables, also called indicator variables, are a way to incorporate qualitative predictors into a regression model. If we have a qualitative predictor  $A$  with  $a$  distinct values, we will need  $a - 1$  distinct dummy variables to code it. For example, suppose we believe that the gender of the subject may impact the response. We could define  $X_{\text{female}} = 1$  if the subject is female and  $X_{\text{female}} = 0$  if the subject is male. Then we interpret the estimated regression coefficient  $\hat{\beta}_{\text{female}}$  as the estimated average amount by which responses for females exceed responses for males, assuming the values of all other predictors are unchanged. If  $\hat{\beta}_{\text{female}} > 0$ , then on average females will tend to have a higher response than males; if  $\hat{\beta}_{\text{female}} < 0$ , then the average male response will exceed the average female response.

There are  $g = 2$  levels to the classification variable `gender`, hence we defined  $g - 1 = 1$  dummy variable to code that information. We pursue this example in Section 10.2.

As another example, suppose one of the predictor variables in a model is the nominal variable `ResidenceLocation`, which can take one of  $r = 3$  values: `urban`, `suburban`, or `rural`. If a qualitative predictor has  $r$  categories, we must assign  $r - 1$  dummy variables to represent it adequately. Otherwise, we may be imposing an unwarranted implicit constraint. It would be incorrect to code this with a single numeric variable  $X_{RL} = 0$  for `urban`, 1 for `suburban`, and 2 for `rural`, as that would imply that the difference between average `urban` and `suburban` responses must equal the difference between average `suburban` and `rural` responses, which is probably not justifiable.

One correct coding is to let  $X_{RLu} = 1$  if `urban` and 0 otherwise and let  $X_{RLs} = 1$  if `suburban` and 0 otherwise. Then the coefficient  $\hat{\beta}_{RLu}$  of  $X_{RLu}$  is interpreted as the average difference between `urban` and `rural` response, and the coefficient  $\hat{\beta}_{RLs}$  of  $X_{RLs}$  is interpreted as the average difference between `suburban` and `rural` response. The difference between the coefficients  $\hat{\beta}_{RLu}$  and  $\hat{\beta}_{RLs}$  is the average difference between the `urban` and `suburban` response. Here we used `rural` as the reference response. The results of the analysis would have been the same had we correctly set up the coding with either `urban` or `suburban` as the reference response. See Section 10.3 for the justification of this statement. See Exercise 10.3 to apply the justification to this example.

This type of coding is done automatically in SAS's PROC ANOVA and PROC GLM with the CLASSES command, and in S-PLUS via the `factor()` function. The coding must be entered explicitly in the DATA step for use with SAS's PROC REG.

Any pair of independent linear combinations of  $X_{RLu}$  and  $X_{RLs}$  would be equally as valid. SAS gives the user choice with the `estimate` and `test` statements on the PROC ANOVA and PROC GLM commands. S-PLUS gives the user choice with the `contrasts()` command.

## 10.2 Example—Height and Weight

### Study Objectives

In the fall of 1998, one of us (RMH) collected the height, weight, and age of the 39 students in one of his classes. The data appear in file (`datasets/htwt.dat`). While this example does give information on the comparative height distributions of men and women, the primary intent then, and now, is to use this example to illustrate how the techniques

of statistics give us terminology and notation for discussing ordinary observations.

## Data Description

**feet**: height in feet rounded down to an integer

**inches**: inches to be added to the height in feet

**lbs**: weight in pounds

**months**: age in months

**sex**: m or f

**meters**: height in meters

## Analysis

### *Data Problems*

From the stem-and-leaf in Table 10.1 we see that even in this small dataset, collected with some amount of care, there are data problems. There are 39 observations, yet only 38 made it to the stem-and-leaf and one of those has a missing value. Further investigation of the data file shows that one student reported her height in meters and another didn't indicate sex. In the S-PLUS code file (`regbb/code/htwt.s`) we converted meters to inches for the one. For the other we had the good fortune to have access to the sample population at the next class meeting and were able to fill in the missing value (m in this case) by checking the data forms directly with the students. We were lucky in this example that the data file was investigated soon enough after collection that the data anomalies could be resolved. That is not always possible, and there are techniques for dealing with missing data; see Section 2.4.

We show a splom of the completed data in Figure 10.1. The age range in our class was 18–28 for women and 19–24 for men. There is no visible relation between age and either height or weight. There is a clear difference in height ranges between men and women and a visible but less strong difference in weight ranges. We investigate this further by showing an expanded `lbs ~ ht` panel in Figure 10.2.

### *Three Variants on the Analysis*

Table 10.2 uses the techniques of Chapter 6 to compare the means of two distributions. The specific features that we will look at are the various values in the ANOVA table and the mean heights for each of the groups.

TABLE 10.1. Stem-and-leaf of Heights from class observation. We used this display to detect the two missing values. Note that this is an edited version of the output. We placed the two distributions adjacent to each other and added additional lines to the high end of the female distribution and to the low end of the male distribution to make the two stem-and-leaf displays align correctly.  
(regbb/code/htwt.s)

---

```
S-PLUS (regbb/transcript/htwt-stem.st):
> length(htwt$ht)
[1] 39
> for (i in tapply(htwt$ht, htwt$sex, c)) stem(i, scale=-1, nl=2)
Removed 1 NAs

female                      male
N = 22  Median = 64.5      N = 15  Median = 72
Quartiles = 63, 66          Quartiles = 66, 72

Decimal point is 1 place to the right of the colon

 5 : 9                      5 :
 6 :                      6 :
 6 : 22223                  6 : 2
 6 : 444445555              6 : 5
 6 : 66677                  6 : 666
 6 : 88                     6 : 89
 7 :                      7 :
 7 :                      7 : 222223
 7 :                      7 : 45
```

---

We will follow by using regression on two different sets of dummy variables to duplicate those numbers.

We initially use the  $g - 1 = 1$  dummy variable  $X_{\text{female}}$  with the  $(1, 0)$  coding scheme suggested above, with value 1 for females and value 0 for males. We display the results of an ordinary linear regression of height on the dummy variable  $X_{\text{female}}$  in Table 10.3. The estimated intercept  $\hat{\beta}_0 = 69.9375$  is the mean height for males. The estimated regression coefficient for the  $X_{\text{female}}$  predictor,  $\hat{\beta}_{\text{female}} = -5.4701$ , is the increment to the intercept that produces the mean height for females. The ANOVA table is identical.

There are many other dummy variable coding schemes that we could use to get exactly the same ANOVA table and the same estimated mean heights for the two groups. We show another in Table 10.4. In this coding, the dummy variable  $X_{\text{treat}}$  has the value 1 for females and the value  $-1$  for males. The estimated intercept  $\hat{\beta}_0 = 67.2024$  is the average of the

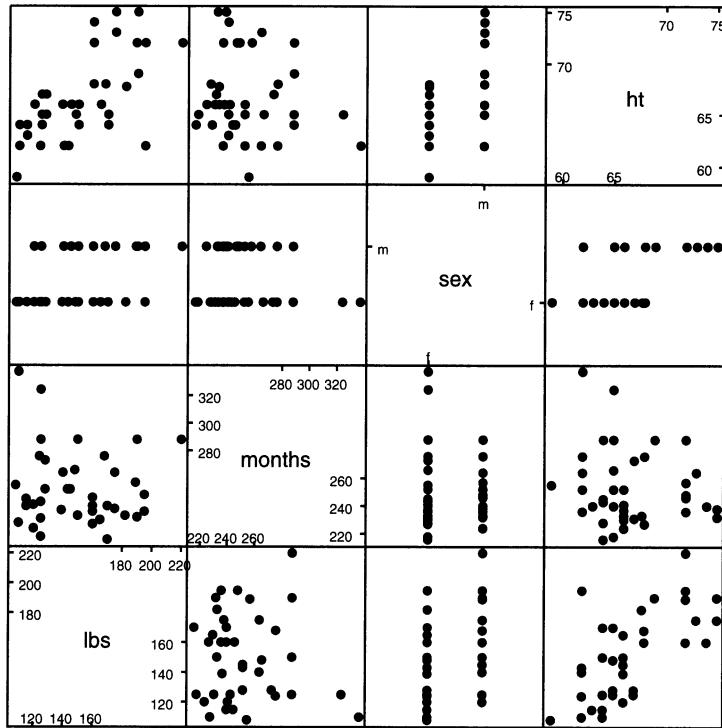


FIGURE 10.1. Scatterplot matrix of completed height and weight data from example collected in class.

(regbb/code/htwt.s), (regbb/figure/htwt.splom.eps.gz)

mean heights for females and males. The estimated regression coefficient for the  $X_{\text{treat}}$  predictor,  $\hat{\beta}_{\text{treat}} = -2.7351$ , is the amount that added to the intercept produces the mean height for females and subtracted from the intercept produces the mean height for males. The ANOVA table is identical.

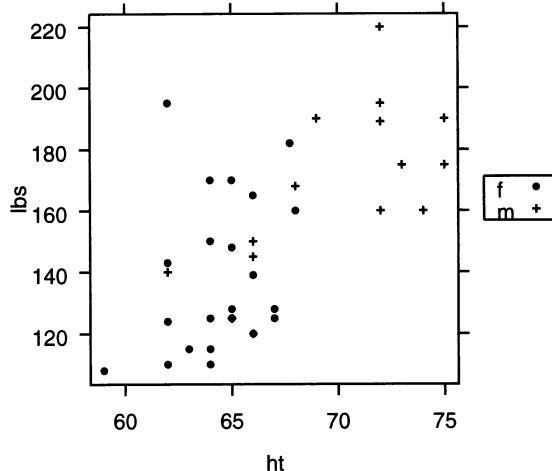


FIGURE 10.2. Expansion of  $\text{lbs} \sim \text{ht}$  panel of Figure 10.1. There is visibly less overlap in the range for the heights of men and women than for their weights.

(regbb/code/htwt.s), (regbb/figure/htwt.xy.eps.gz)

TABLE 10.2. One-way analysis of variance of heights from class observation.  
(regbb/code/htwt.s)

---

```
S-PLUS (regbb/transcript/htwt-aov.st):
> ## one-way analysis of variance
> htwt.aov <- aov(ht ~ sex, data=htwt)
> summary(htwt.aov)
    Df Sum of Sq  Mean Sq  F Value      Pr(F)
sex   1  282.3418 282.3418 30.81829 2.54377e-006
Residuals 37  338.9755   9.1615
> model.tables(htwt.aov, type="means")
Warning messages:
Model was refit to allow projection in: model.tables(htwt.aov, type = "means")

Tables of means
Grand mean

66.712

sex
  f      m
64.467 69.938
rep 23.000 16.000
```

---

TABLE 10.3. Regression analysis of heights from class observation on the dummy variable coding sex as `female=1` for female and `female=0` for male.

(regbb/code/htwt.s)

```
S-PLUS (regbb/transcript/htwt-lm.st):
> htwt$female <- as.numeric(htwt$sex == "f")
> htwt[c(1,4,7),]
   feet inch lbs months sex meters ht female
1     5     4 115    245   f     NA 64      1
4    NA    NA 128    252   f    1.65 65      1
7     6     0 220    288   m     NA 72      0
> htwt.lm <- lm(ht ~ female, data=htwt)
> summary(htwt.lm, corr=F)

Call: lm(formula = ht ~ female, data = htwt)
Residuals:
    Min      1Q Median      3Q      Max
-7.937 -2.202  0.5326  2.063  5.063

Coefficients:
            Value Std. Error t value Pr(>|t|)
(Intercept) 69.9375   0.7567   92.4244  0.0000
female     -5.4701   0.9854   -5.5514  0.0000

Residual standard error: 3.027 on 37 degrees of freedom
Multiple R-Squared: 0.4544
F-statistic: 30.82 on 1 and 37 degrees of freedom, the p-value is 2.544e-006
> anova(htwt.lm)

Analysis of Variance Table

Response: ht

Terms added sequentially (first to last)
          Df Sum of Sq Mean Sq F Value    Pr(F)
female   1  282.3418 282.3418 30.81829 2.54377e-006
Residuals 37  338.9755   9.1615
```

TABLE 10.4. Regression analysis of heights from class observation on the dummy variable coding sex as `treat=1` for female and `treat=-1` for male.  
(regbb/code/htwt.s)

---

```
S-PLUS (regbb/transcript/htwtb-lm.st):
> ## dummy variable
> htwt$treat <- (htwt$sex == "f") - (htwt$sex == "m")
> htwtb.lm <- lm(ht ~ treat, data=htwt)
> summary(htwtb.lm, corr=F)

Call: lm(formula = ht ~ treat, data = htwt)
Residuals:
    Min      1Q Median      3Q     Max
-7.937 -2.202  0.5326  2.063  5.063

Coefficients:
            Value Std. Error   t value Pr(>|t|)
(Intercept) 67.2024    0.4927 136.4028 0.0000
treat       -2.7351    0.4927  -5.5514 0.0000

Residual standard error: 3.027 on 37 degrees of freedom
Multiple R-Squared: 0.4544
F-statistic: 30.82 on 1 and 37 degrees of freedom, the p-value is 2.544e-006
> anova(htwtb.lm)
Analysis of Variance Table

Response: ht

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value    Pr(F)
treat     1  282.3418 282.3418 30.81829 2.54377e-006
Residuals 37  338.9755  9.1615
```

---

## 10.3 Equivalence of Linear Independent $X$ -Variables for Regression

It is not an accident that the ANOVA tables in Tables 10.2, 10.3, and 10.4 are identical. We explore here why that is the case.

Please review the definition of linear dependence in Section F.4.2.

The  $X$  matrix in the linear regression presentation of the one-way analysis of variance model with one factor with  $a$  categories must have a leading column of ones  $X_0 = \mathbf{1}$  for the intercept and at least  $a - 1$  additional columns, for a total of  $c \geq a$  columns. The entire  $X$  matrix can be summarized by a *contrast matrix*  $W$  consisting of the  $a$  unique rows, one for each level of the factor.

We explore the relationship between several different  $W$  matrices in the case  $a = 4$ . The principles work for any value  $a$ . The matrix  $X$  itself consists of  $n_i$  copies of the  $i^{\text{th}}$  row of  $W$  (where  $n = \sum_{i=1}^a n_i$ ):

$$\underset{n \times c}{X} = \begin{pmatrix} n_1 \{(1\ 0\ 0\ 0) \\ n_2 \{(0\ 1\ 0\ 0) \\ n_3 \{(0\ 0\ 1\ 0) \\ n_4 \{(0\ 0\ 0\ 1) \end{pmatrix}_{4 \times c} W = \underset{n \times 4}{N} \underset{4 \times c}{W}$$

Any  $W$  matrix with  $a = 4$  rows and with rank 4 (which means it must have at least 4 columns) is equivalent for linear regression in the senses that

1. Any two such matrices  $W_1$  and  $W_2$  with dimensions  $(4 \times c_1)$  and  $(4 \times c_2)$  where  $c_i \geq 4$  are related by postmultiplication of the first matrix by a full-rank matrix  $\underset{c_1 \times c_2}{A}$ , that is,

$$\underset{4 \times c_1}{W_1} \underset{c_1 \times c_2}{A} = \underset{4 \times c_2}{W_2}$$

Equivalently, any two such matrices  $X_1$  and  $X_2$  with dimensions  $(n \times c_1)$  and  $(n \times c_2)$  are similarly related by

$$\underset{n \times c_1}{X_1} \underset{c_1 \times c_2}{A} = \underset{n \times c_2}{X_2}$$

Examples:

- 1a. SAS parameterization (5 columns with rank=4):

$$\underset{4 \times (1+4)}{W_{\text{SAS}}} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

**1b.** S contr.treatment (4 columns with rank=4):

$$\begin{array}{ccc}
 W_{\text{SAS}} & A & W_{\text{treatment}} \\
 4 \times (1+4) & (1+4) \times (1+3) & 4 \times (1+3)
 \end{array} = \begin{pmatrix}
 1 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 \\
 0 & 1 & 0 & 0 \\
 0 & 0 & 1 & 0 \\
 0 & 0 & 0 & 1
 \end{pmatrix} = \begin{pmatrix}
 1 & 0 & 0 & 0 \\
 1 & 1 & 0 & 0 \\
 1 & 0 & 1 & 0 \\
 1 & 0 & 0 & 1
 \end{pmatrix}$$

**1c.** S cont.helmert (4 columns with rank=4):

$$\begin{array}{ccc}
 W_{\text{SAS}} & A & W_{\text{helmert}} \\
 4 \times (1+4) & (1+4) \times (1+3) & 4 \times (1+3)
 \end{array} = \begin{pmatrix}
 1 & 1 & 1 & 1 \\
 0 & -2 & -2 & -2 \\
 0 & 0 & -2 & -2 \\
 0 & -1 & 1 & -2 \\
 0 & -1 & -1 & 2
 \end{pmatrix} = \begin{pmatrix}
 1 & -1 & -1 & -1 \\
 1 & 1 & -1 & -1 \\
 1 & 0 & 2 & -1 \\
 1 & 0 & 0 & 3
 \end{pmatrix}$$

**2.** The hat matrices are the same.

$$H_1 = (X_1(X_1'X_1)^{-1}X_1') = (X_2(X_2'X_2)^{-1}X_2') = H_2$$

An equivalent statement is that both  $X$  matrices span the same column space.

PROOF.

For the special case that  $c = a$ , hence the  $X'X$  and  $A$  matrices are invertible:

$$\begin{aligned}
 H_2 &= \\
 X_2(X_2'X_2)^{-1}X_2' &= \\
 (X_1A)((X_1A)'(X_1A))^{-1}(X_1A)' &= \\
 (X_1A)(A'X_1'X_1A)^{-1}(A'X_1') &= \\
 X_1(X_1'X_1)^{-1}X_1' &= \\
 H_1 &
 \end{aligned}$$

When  $c > a$ , the step from line 4 to line 5 involves matrix algebra manipulations that we do not discuss here. Effectively, we are dropping any redundant columns.

3. The predicted values are the same.

$$\hat{Y} = H_1 Y = H_2 Y$$

4. The regression coefficients are related by premultiplication of the second set of coefficients by the same matrix  $A$ ,

$$\beta_1 = A\beta_2$$

PROOF.

$$E(Y) = X_2\beta_2 = (X_1A)\beta_2 = X_1(A\beta_2) = X_1\beta_1$$

5. The ANOVA (analysis of variance) table is the same:

| Source     | Sum of Squares |   |                |
|------------|----------------|---|----------------|
| Regression | $SS_{Reg}$     | = | $Y'H_1Y$       |
| Residual   | $SS_{Res}$     | = | $Y'(I - H_1)Y$ |
|            |                | = | $Y'H_2Y$       |
|            |                | = | $Y'(I - H_2)Y$ |

Exercise 10.1 gives you the opportunity to explore the equivalence of the two coding schemes in Section 10.2.

As a consequence of the equivalence up to multiplication by a matrix  $A$ , the regression coefficients in regression analyses with factors (which means most experiments) are uninterpretable unless the definitions of the dummy variables have been provided.

## 10.4 Polynomial Contrasts and Orthogonal Polynomials

(Ott, 1993) reports an experiment that uses an abrasives testing machine to test the wear of a new experimental fabric. The machine was run at six different speeds (measured in revolutions per minute). Forty-eight identical square pieces of fabric were prepared, 8 of which were randomly assigned to each of the 6 machine speeds. Each square was tested for a three-minute period at the appropriate machine setting. The order of testing was appropriately randomized. For each square, the amount of wear was measured and recorded. The data from file (`datasets/fabricwear.dat`) are displayed in Figure 10.3. The initial ANOVA is in Table 10.5.

From Figure 10.3 we see that the assumption in Equation (6.3) of approximately constant variance across groups is satisfied by this dataset, hence

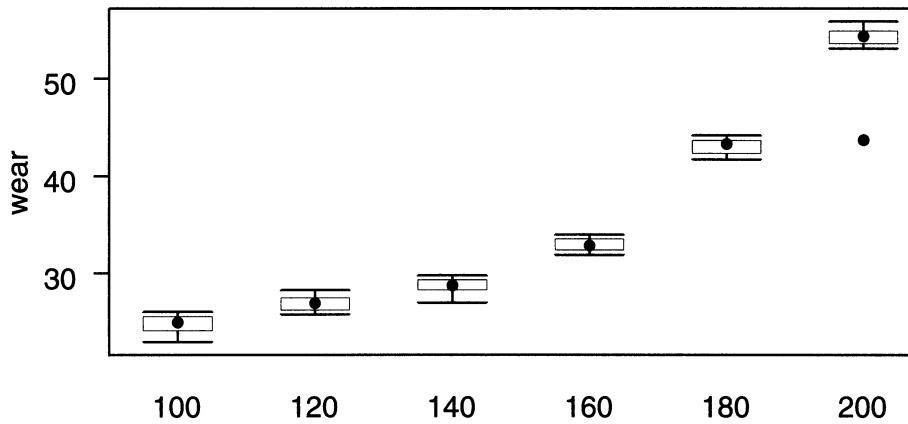


FIGURE 10.3. Fabric wear as a function of speed.  
 (regbb/code/fabricwear.s), (regbb/figure/fabricwear.eps.gz)

TABLE 10.5. ANOVA and means for wear as a function of speed.  
 (regbb/code/fabricwear.s), (regbb/transcript/fabricwear.st)

---



---

```
S-PLUS (regbb/transcript/fabricwear1.st):
> fabricwear$speed <- ordered(fabricwear$speed)
>
> fabricwear.aov <- aov(wear ~ speed, data=fabricwear, x=T)
> summary(fabricwear.aov)
   Df Sum of Sq Mean Sq F Value Pr(F)
speed  5  4872.167  974.4333 297.7046    0
Residuals 42   137.472    3.2732
> model.tables(fabricwear.aov, "mean")
Warning messages:
Model was refit to allow projection in: model.tables(fabricwear.aov, "mean")

Tables of means
Grand mean

34.929

speed
 100     120     140     160     180     200
24.775 26.963 28.675 32.925 43.050 53.188
```

---

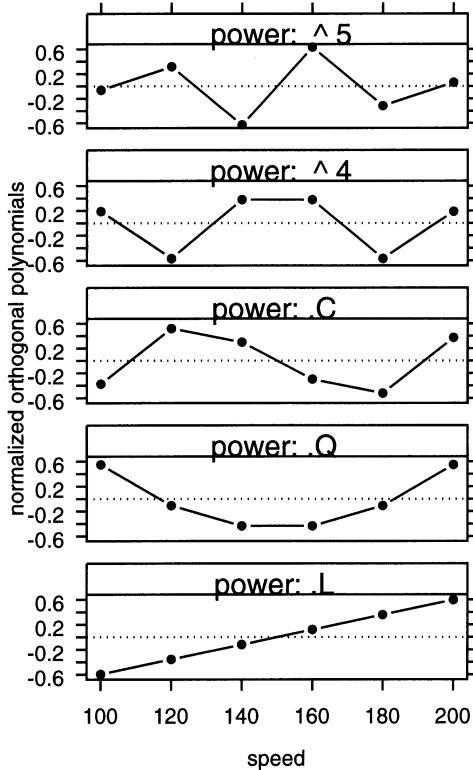


FIGURE 10.4. Orthogonal polynomials for speed.  
 (`regbb/code/fabricwear.s`), (`regbb/figure/orthpoly.eps.gz`)

ANOVA is an appropriate technique for investigating the data. We also note one outlier at speed=200. We will return to that data point later.

The ANOVA table in Table 10.5 shows that speed is significant. From the table of means we see that the means increase with speed and the increase is also faster as speed increases. Figure 10.3 shows the same and suggests that the means are increasing as a quadratic polynomial in speed.

There are several essentially identical ways to check this supposition. We start with the easiest to do and then expand by illustrating the arithmetic behind it. When we defined `speed` as a factor in Table 10.5, we actually did something more specific, we declared it to be an *ordered factor*. This means that the dummy variables are the orthogonal polynomials for six levels. We display the orthogonal polynomials in Figure 10.4 and Table 10.6. See the

TABLE 10.6. Orthogonal polynomials for speed.  
(regbb/code/fabricwear.s), (regbb/transcript/fabricwear.st)

---

S-PLUS (regbb/transcript/fabricwear2.st):

```
> tmp.c <- contrasts(fabricwear$speed)
> tmp.c
      .L       .Q       .C     ^ 4     ^ 5
100 -0.5976143 0.5455447 -0.3726780 0.1889822 -0.06299408
120 -0.3585686 -0.1091089 0.5217492 -0.5669467 0.31497039
140 -0.1195229 -0.4364358 0.2981424 0.3779645 -0.62994079
160  0.1195229 -0.4364358 -0.2981424 0.3779645 0.62994079
180  0.3585686 -0.1091089 -0.5217492 -0.5669467 -0.31497039
200  0.5976143 0.5455447 0.3726780 0.1889822 0.06299408
> zapsmall(crossprod(tmp.c))
      .L .Q .C ^ 4 ^ 5
      .L 1 0 0 0 0
      .Q 0 1 0 0 0
      .C 0 0 1 0 0
      ^ 4 0 0 0 1 0
      ^ 5 0 0 0 0 1
> tmp.min <- apply(abs(tmp), 2, min)
> sweep(tmp.c, 2, tmp.min, "/")
      .L .Q .C ^ 4 ^ 5
100 -5 5 -1.25 1 -1
120 -3 -1 1.75 -3 5
140 -1 -4 1.00 2 -10
160 1 -4 -1.00 2 10
180 3 -1 -1.75 -3 -5
200 5 5 1.25 1 1
```

---

discussion in Section F.4 for an overview of orthogonal polynomials and their construction.

From the panels in Figure 10.4 we see that the linear polynomial plots as a straight line against the speed. The quadratic polynomial plots as a discretization of a parabola. The higher-order polynomials are rougher discretizations of their functions. In Table 10.6 we see that the orthogonal polynomials are scaled so their cross product is the identity matrix, that is, it is a diagonal matrix with 1s on the diagonal. Compare this (in Exercise 10.2) to a matrix of the simple powers of the integers (1, 2, 3, 4, 5, 6). The columns of the simple powers span the same linear space as the orthogonal properties. Because they are not orthogonal (their cross product is not diagonal), their plots are harder to interpret and they may show numerical difficulties when used as predictor variables in a regression.

TABLE 10.7. Regression coefficients on dummy variables, and partitioned ANOVA table.  
 (regbb/code/fabricwear.s), (regbb/transcript/fabricwear.st)

```

S-PLUS (regbb/transcript/fabricwear3.st):
> summary.lm(fabricwear.aov, corr=F)

Call: aov(formula = wear ~ speed, data = fabricwear, x = T)
Residuals:
    Min      1Q Median      3Q     Max
-9.488 -0.6531  0.1813  0.825  2.712

Coefficients:
            Value Std. Error   t value Pr(>|t|)
(Intercept) 34.9292    0.2611 133.7598 0.0000
speed.L     23.2562    0.6396   36.3580 0.0000
speed.Q      8.0086    0.6396   12.5204 0.0000
speed.C      0.9280    0.6396    1.4508 0.1543
speed ^ 4   -1.6772    0.6396   -2.6221 0.0121
speed ^ 5   -0.6000    0.6396   -0.9381 0.3536

Residual standard error: 1.809 on 42 degrees of freedom
Multiple R-Squared: 0.9726
F-statistic: 297.7 on 5 and 42 degrees of freedom, the p-value is 0

> summary(fabricwear.aov,
+           split=list(speed=list(speed.L=1, speed.Q=2, speed.C=3, rest=4:5)))
      Df Sum of Sq Mean Sq F Value    Pr(F)
speed    5  4872.167  974.433 297.705 0.0000000
speed: speed.L 1  4326.792  4326.792 1321.903 0.0000000
speed: speed.Q 1   513.101   513.101  156.760 0.0000000
speed: speed.C 1     6.889     6.889    2.105 0.1542755
speed: rest    2    25.385    12.692    3.878 0.0284832
Residuals 42   137.472     3.273
  
```

In Table 10.7 we show two variants of an expanded display of the ANOVA from Table 10.5. The top of the table shows the regression coefficients for the regression against the orthogonal polynomials used as the dummy variables. Here we see that the linear and quadratic terms are highly significant. The cubic term is not significant. Based on our reading of the graph, and the comparison of the *p*-value for the quartic term to that of the quadratic term, we will interpret the quartic term as not significant and do all continuing work with the quadratic model.

In the bottom of Table 10.7 we show the partitioned ANOVA table with the linear, quadratic, and cubic terms isolated. By dint of the orthogonality the  $F$ -values are the square of the  $t$ -values for the coefficients ( $36.3580^2 = 1321.903$ ) and the  $p$ -values are identical.

What happens when we redo the analysis without the outlier noted in Figure 10.3? The residual mean square goes down by a factor of 4; consequently, all the  $t$ -values go up. While the  $p$ -values for the cubic and quartic terms now show significance at .0001, we will continue to exclude them from our recommended model because the  $p$ -values for the linear and quadratic terms are orders of magnitude smaller ( $< 10^{-16}$ ). See Exercise 10.8.

#### 10.4.1 Specification and Interpretation of Interaction Terms

Example—consider a model

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{34} X_3 X_4$$

to “explain” determinants of annual salary  $Y$  in dollars for workers in some population. Here  $X_1$  is age in years,  $X_2$  is gender (1 if female, 0 if male),  $X_3$  is race (1 if white, 0 if nonwhite), and  $X_4$  is number of years of schooling. (Discussion: What other variables might such a model include to explain salary?)

The existence of the interaction terms allows for the possibility that the degree of enhancement of education on schooling differs for whites and nonwhites.

Consider a white and a nonwhite of the same age and gender and having the same amount of schooling. Then:

- $\beta_4$  is the expected increase in annual salary for a nonwhite attributable to an additional year of schooling.
- $\beta_4 + \beta_{34}$  is the expected increase in annual salary for a white attributable to an additional year of schooling.
- $\beta_{34}$  is the expected amount by which a white’s salary increase as a result of an additional year of schooling exceeds a nonwhite’s salary increase as a result of an additional year of schooling.

Also, still assuming the same age and gender,

- $\beta_3 + \beta_{34}X_4$  is the difference between white and nonwhite expected salary.
- $\beta_3$  is the component of this difference that does not depend on years of schooling and is attributable only to difference in race.

We examine this model further in Exercise 10.7.

## 10.5 Analysis Using a Concomitant Variable (Analysis of Covariance)

In some situations where we seek to compare the differences in the means of a continuous response variable across levels of a factor  $A$ , we have available a second continuous variable that can be used to improve our ability to distinguish among the levels. Historically this extended model has been called the *analysis of covariance* model because the second variable varies along with the first. To avoid confusion with the concept of covariance introduced in Chapter 3, we prefer to call this approach *analysis using a concomitant variable*. Nevertheless we will retain use of the term covariate as a shorthand term for concomitant variable and the acronym ANCOVA as an abbreviation for this method.

If  $X_{ij}$  denotes the  $j^{\text{th}}$  observation of the covariate at the  $i^{\text{th}}$  level of factor  $A$ , our original ANOVA model in Equation (6.1) generalizes to

$$Y_{ij} = \mu + \alpha_i + \beta(X_{ij} - \bar{X}) + \epsilon_{ij} \quad (10.1)$$

for  $i = 1, \dots, a$  and  $j = 1, \dots, n_i$

where  $\bar{X}$  is the grand mean of the  $X_{ij}$ 's and all other terms are as defined in Equation (6.1). The model in Equation (10.1) has separate intercepts  $\alpha_i$  for each level of  $A$  but retains a common slope. The differences between the intercepts  $\alpha_i$  are identical to the vertical differences between the parallel lines (to be illustrated in Figure 10.8). Equation (10.1) is the classical ANCOVA model.

The logic of this approach is that if  $X_{ij}$  is related to  $Y_{ij}$  then the  $\epsilon$ 's of the model in Equation (10.1) will be measured from a different regression line for each level of  $A$  rather than from a different horizontal line as in model (6.1). This will give the  $\epsilon$ 's less variability than those of Equation (6.1), thereby sharpening our inferences on the  $\alpha_i$ 's. The  $\alpha_i$ 's estimated from Equation (10.1) are said to be *adjusted* for the covariate. Quite frequently the range of observed  $X_{ij}$  differs for each level of  $A_i$  and therefore the  $\bar{Y}_i$  means from Equation (6.1) reflect the difference in the  $X$ -values more than the differences attributable to the change in levels of  $A$ .

The next level of generalization allows the slopes to differ, i.e., replace the common  $\beta$  in Equation (10.1) with  $\beta_i$ :

$$Y_{ij} = \mu + \alpha_i + \beta_i(X_{ij} - \bar{X}) + \epsilon_{ij} \quad (10.2)$$

for  $i = 1, \dots, a$  and  $j = 1, \dots, n_i$

We illustrate models Equations (10.1) and (10.2) in Section 10.6. In Section 10.6.3 we will use the model in Equation (10.2) to test the assumption that

the lines are parallel. Formally, we will test whether the lines have the same slope

$$H_0: \beta_1 = \beta_2 = \beta_3 \quad (10.3)$$

$$H_1: \text{Not all } \beta_i \text{ are identical}$$

or the same intercept

$$H_0: \alpha_1 = \alpha_2 = \alpha_3 \quad (10.4)$$

$$H_1: \text{Not all } \alpha_i \text{ are identical}$$

or both (in which case the lines coincide).

These ideas can be extended to situations with more than one covariate variable and to more complicated experimental designs such as those discussed in Chapters 12 through 14.

## 10.6 Example—Hot Dog Data

### Study Objectives

Hot dogs based on poultry are said to be healthier than ones made from either meat (beef and pork) or all beef. A basis for this claim may be the lower-calorie (fat) content of poultry hot dogs. Is this advantage of poultry hot dogs offset by a higher sodium content than meat hot dogs?

Researchers for *Consumer Reports* analyzed three types of hot dog: beef, poultry, and meat (mostly pork and beef, but up to 15% poultry meat). The data in file (`datasets/hotdog.dat`) come from (Consumer Reports, 1986) and were later used by (Moore and McCabe, 1989).

### Data Description

**Type:** Type of hot dog (beef, meat, or poultry)

**Calories:** Calories per hot dog

**Sodium:** Milligrams of sodium per hot dog

#### 10.6.1 One-Way ANOVA

We start by comparing the **Sodium** content of the three hot dog **Types** in Figure 10.5 and in Table 10.8. We see that the three **Types** have similar **Sodium** content.

Figure 10.6 shows the response **Sodium** plotted against the covariate **Calories** by **Type**. Within each panel we plot a horizontal line at the mean

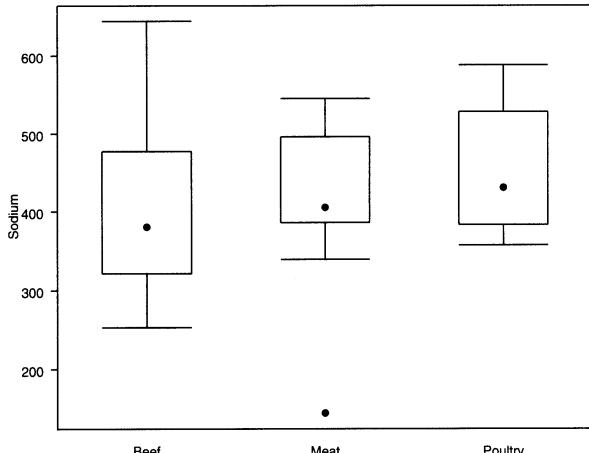


FIGURE 10.5. Boxplots comparing the Sodium content of three Types of hot dogs. See Table 10.8.  
(regbb/code/hotdog.s),(regbb/figure/hotdog1.eps.gz)

TABLE 10.8. Hot dog ANOVA and means. See Figures 10.5 and 10.6.  
(regbb/code/hotdog.s)

```
S-PLUS (regbb/transcript/hotdog-anova1.st):
> ## horizontal lines: zero slope and separate intercepts
> hotdog.aov <- aov(Sodium ~ Type, data=hotdog, x=Calories,
+                         par.strip.text=list(cex=1.2))
> print.trellis(position=c(0,0, 1,.6),
+                 attr(hotdog.aov,"trellis"))
> ## export.eps(hh("regbb/figure/hotdog.f0.eps"))
> summary(hotdog.aov)
    Df Sum of Sq Mean Sq F Value     Pr(F)
Type    2   31738.7 15869.36 1.777791 0.1793247
Residuals 51   455248.8   8926.45
>
> model.tables(hotdog.aov, type="means")

Tables of means
Grand mean

424.83

Type
  Beef   Meat Poultry
  401.15 418.53 459.00
rep  20.00 17.00 17.00
```

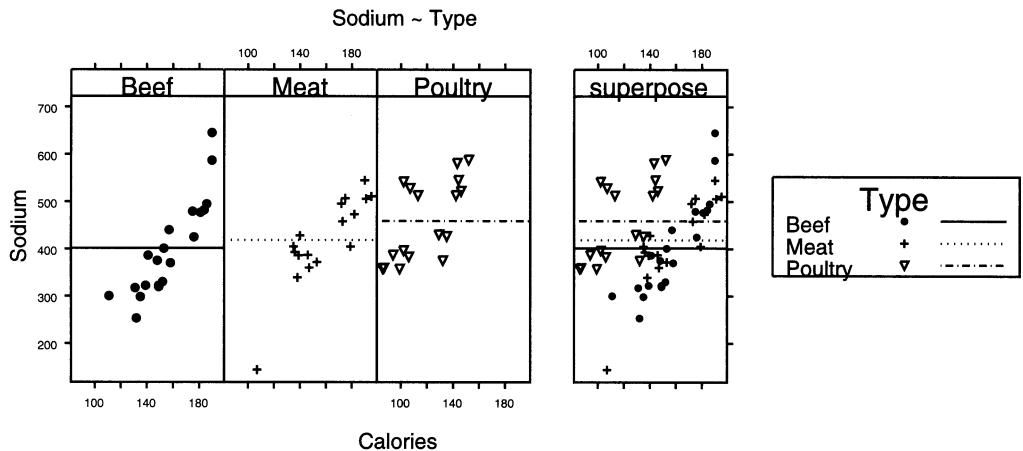


FIGURE 10.6.  $\text{Sodium} \sim \text{Calories} | \text{Type}$ . Horizontal lines at Sodium means for each Type.

$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ . See Table 10.8.

(regbb/code/hotdog.s),

(regbb/figure/hotdog.f0.eps.gz), (regbb/figure/hotdog.f0-color.eps.gz)

of the Sodium values for that Type. The analysis of variance in Table 10.8 compares the vertical distance between these horizontal lines. It ignores the most evident feature of this plot, that the three Types have very different fat contents with Poultry low, Beef intermediate, and Meat high. We wish to see if knowledge about Calories affects our understanding about Sodium.

### 10.6.2 Concomitant Explanatory Variable

It is possible that our finding of similar Sodium content is attributable in part to a need to add sodium to enhance the flavor of higher-fat hot dogs. The Calories information can be incorporated into the analysis by adding Calories to the model as a concomitant explanatory variable. Then in this revised model, comparisons between the mean Sodium contents of the three Types will have been *adjusted for* differing Calories contents. In this way, comparisons between the three Types will be made on the basis that each Type has the mean Calories content of all Types.

We illustrate this revised analysis in two steps. Initially, in Figure 10.7 and Table 10.9, we show the regression of Sodium on Calories ignoring the Types. The common regression line makes some sense in the Superpose panel but very clearly has the wrong slope and wrong intercept in all three of the individual panels.

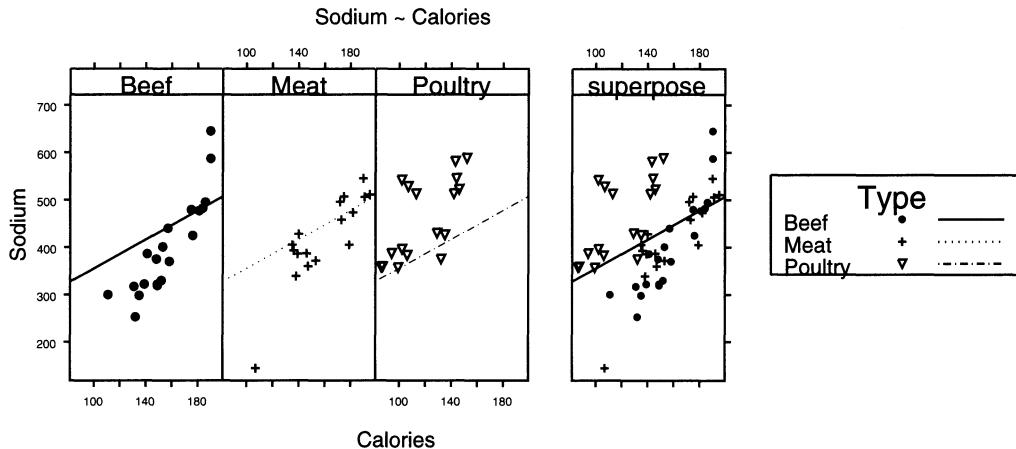


FIGURE 10.7. Sodium ~ Calories. Common regression line that ignores Type.

$$Y_{ij} = \mu + \beta(X_{ij} - \bar{X}) + \epsilon_{ij}. \text{ See Table 10.9.}$$

(regbb/code/hotdog.s),

(regbb/figure/hotdog.f3.eps.gz), (regbb/figure/hotdog.f3-color.eps.gz)

TABLE 10.9. Hot dog ANCOVA with a common regression line that ignores Type. See Figure 10.7.  
(regbb/code/hotdog.s)

---

```
S-PLUS (regbb/transcript/hotdog-ancova.f3.st):
> ## regression
> ## same line: common intercept and common slope
> hC.aov <- ancova(Sodium ~ Calories, groups=Type, data=hotdog,
+                     par.strip.text=list(cex=1.2))
> print.trellis(position=c(0,0, 1,.6),
+                 attr(hC.aov,"trellis"))
> ## export.eps(hh("regbb/figure/hotdog.f3.eps"))
> summary(hC.aov)
   Df Sum of Sq Mean Sq F Value      Pr(F)
Calories  1 106269.7 106269.7 14.51475 0.0003693124
Residuals 52  380717.8    7321.5
```

---

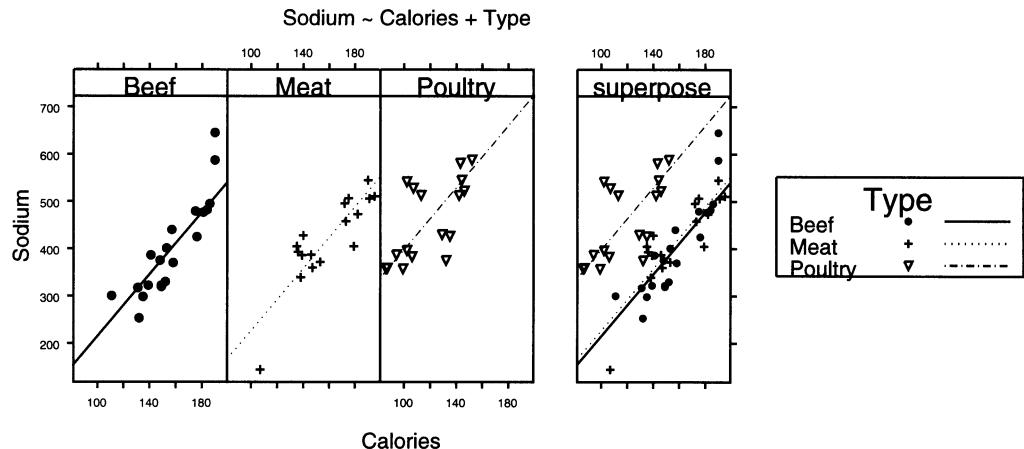


FIGURE 10.8.  $\text{Sodium} \sim \text{Calories} + \text{Type}$ . Parallel lines.  $Y_{ij} = \mu + \alpha_i + \beta(X_{ij} - \bar{X}) + \epsilon_{ij}$ . See Table 10.10.

(regbb/code/hotdog.s),

(regbb/figure/hotdog.f1.eps.gz), (regbb/figure/hotdog.f1-color.eps.gz)

TABLE 10.10. Hot dog ANCOVA with parallel lines and separate intercepts. See Figure 10.8.  
 (regbb/code/hotdog.s), (regbb/code/hotdog.sas), (regbb/transcript/hotdog.lst)

---

S-PLUS (regbb/transcript/hotdog-ancova2.st):

```

> ## analysis of covariance
> ## analysis with a concomitant explanatory variable
> ## parallel lines: separate intercepts and common slope
> hCT.aov <- aov(Sodium ~ Calories + Type, data=hotdog,
+                   par.strip.text=list(cex=1.2))
> print.trellis(position=c(0,0, 1,.6),
+                attr(hCT.aov,"trellis"))
> ## export.eps(hh("regbb/figure/hotdog.f1.eps"))
> summary(hCT.aov)
    Df Sum of Sq  Mean Sq  F Value      Pr(F)
Calories   1 106269.7 106269.7 34.65360 3.281021e-007
Type       2 227386.4 113693.2 37.07433 1.336000e-010
Residuals 50 153331.4   3066.6
  
```

---

Figure 10.8 and Table 10.10 show parallel regression lines for each Type. They have separate intercepts and a common slope. This model is the standard *analysis of covariance* model. We are interested in the vertical distance between the parallel lines. Equivalently, we are interested in the distance

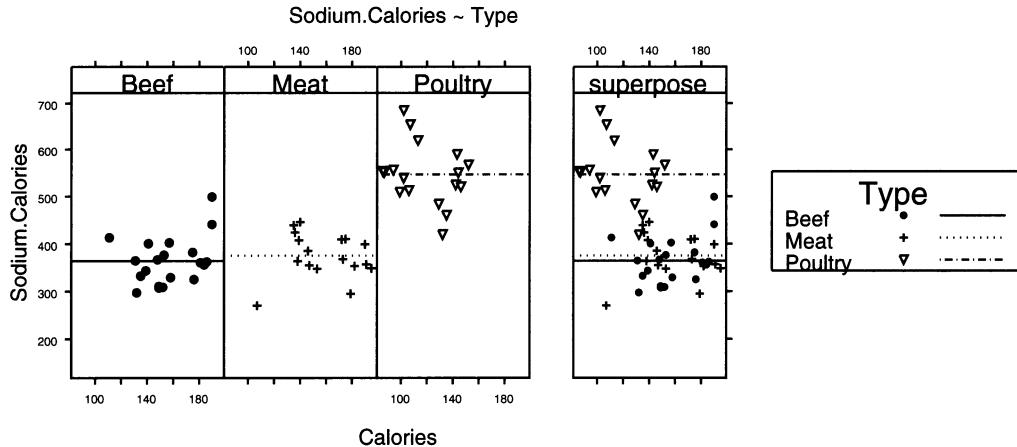


FIGURE 10.9.  $\text{Sodium} \sim \text{Type}$ . Horizontal lines after adjustment for the covariate.  $(Y_{ij}|X_{ij}) = \mu + \alpha_i + \epsilon_{ij}$ . See Table 10.11.

`(regbb/code/hotdog.s),`

`(regbb/figure/hotdog.f4.eps.gz), (regbb/figure/hotdog.f4-color.eps.gz)`

between the intercepts. We see from the  $F = 37.07433$  with  $p = 1.310^{-10}$  in the first part of Table 10.10 that the vertical distance is significant.

The original preliminary conclusion based on Table 10.8 was misleading because it left out the critical dependence of  $y=\text{Sodium}$  on the  $x=\text{Calories}$  variable.

It is possible (see Exercise 10.5 for an example) for the covariate to be significant and not the grouping factor. In this example both are significant.

We construct Figure 10.9 and Table 10.11 to show the means for the response **Sodium** adjusted for the covariate **Calories**. The adjustment maintains the same vertical distance between the fitted lines that we observe in Figure 10.8. From the ANOVA table in Table 10.11 we see that the adjusted means have the same residual sum of squares as the unadjusted means. The residual degrees of freedom are wrong because the analysis in **hSCT.aov** doesn't know that the effect of the **Calories** variable has already been removed. The **Type** sum of squares is not what we anticipated because we did not adjust the **Type** dummy variables for the covariate; we only adjusted the response variable.

TABLE 10.11. Horizontal lines after adjustment for the covariate. See Figure 10.9.  
 (regbb/code/hotdog.s)

---

```
S-PLUS (regbb/transcript/hotdog-ancovaf4.st):
> hotdog <- cbind(hotdog, Sodium.Calories=hotdog$Sodium -
+                         predict.lm(hCT.aov, type="terms", terms="Calories")[,1])
> hSCT.aov <- ancova(Sodium.Calories ~ Type, x=Calories, data=hotdog,
+                         par.strip.text=list(cex=1.2), ylim=c(140,700))
> summary(hSCT.aov)
   Df Sum of Sq Mean Sq F Value    Pr(F)
Type    2     368463  184232   61.28 2.742e-014
Residuals 51     153331      3006
> summary(hCT.aov)
   Df Sum of Sq Mean Sq F Value    Pr(F)
Calories  1     106270  106270   34.65 3.281e-007
          Type  2     227386  113693   37.07 1.000e-010
Residuals 50     153331      3067
>
> model.tables(hSCT.aov, type="means")

Tables of means
Grand mean

424.83

Type
  Beef   Meat Poultry
  363.74 375.03 546.50
rep 20.00 17.00 17.00
Warning messages:
  Model was refit to allow projection in:
  model.tables(hSCT.aov, type = "means")
```

---

TABLE 10.12. Multiple comparisons by Tukey's method of the ANCOVA model in Figure 10.8 and Table 10.10 comparing the mean **Sodium** content of three **Types** of hot dogs adjusted for **Calories**. See also Figure 10.10.

(regbb/code/hotdog.s)

---

```
S-PLUS (regbb/transcript/hotdog-ancova2b.st):
> ## multiple comparisons of ANCOVA
> hCT.mca <- multicomp(hCT.aov, focus="Type")
> hCT.mca <- multicomp.reverse(hCT.mca)  ## positive differences
> print(hCT.mca)

95 % simultaneous confidence intervals for specified
linear combinations, by the Tukey method

critical point: 2.4154
response variable: Sodium

intervals excluding 0 are flagged by '****'

      Estimate Std.Error Lower Bound Upper Bound
Meat-Beef     11.3      18.3      -32.9      55.4
Poultry-Beef   183.0      22.2      129.0     236.0 ****
Poultry-Meat   171.0      23.1      116.0     227.0 ****

> plot(hCT.mca)
> ## export.eps(hh("regbb/figure/hotdog3.eps"))
```

---

Now that we have shown the factor **Type** to be important, we show in Table 10.12 and Figure 10.10 the results of multiple comparisons analysis using the Tukey procedure. These show that **Meat** and **Beef** are indistinguishable and that **Poultry** differs from both **Meat** and **Beef**.

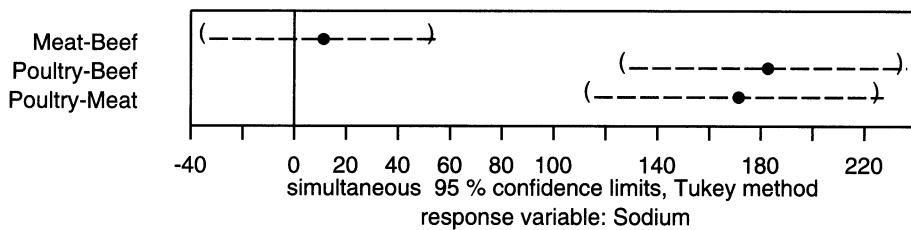


FIGURE 10.10. Multiple comparisons by Tukey's method of the ANCOVA model in Figure 10.8 and Table 10.10 comparing the mean **Sodium** content of three **Types** of hot dogs adjusted for **Calories**. See also Figures 10.8 and 10.9 and Table 10.12.

(regbb/code/hotdog.s), (regbb/figure/hotdog3.eps.gz)

The MMC figures and table are available on the book's online files  
 (regbb/code/hotdog-mmc2.s),  
 (regbb/figure/hotdog-mmc-mca.eps.gz),  
 (regbb/figure/hotdog-mmc-lmat.eps.gz), and  
 (regbb/transcript/hotdog-mmc2b.st).

### 10.6.3 Tests of Equality of Regression Lines

In Section 10.6.2 we assume the constant slope model (10.1) and test whether the intercepts differed by testing (10.4) about  $\alpha_i$ . We can also work with the separate slope model (10.2) and test (10.3) about  $\beta_i$ .

Figure 10.11 and Table 10.13 show separate regression lines for each group. These have separate intercepts and slopes. The *F*-test of **Calories**:**Type** in Table 10.13 having *p*-value = .185 addresses the null hypothesis that the regression lines for predicting **Sodium** from **Calories** are parallel.

Observe in Figure 10.11 that the slopes of the lines for the regressions of **Sodium** on **Calories** appear to differ for the three **Types** of hot dog. This null hypothesis is expressed as two equalities in Equation (10.3) and is tested in Table 10.13 using the two degree-of-freedom sum of squares for the interaction **Calories**:**Type**. The *p*-value for this test, 0.185, implies that the null hypothesis cannot be rejected and therefore that the three slopes are homogeneous. Any difference among them is too small to detect with the sample sizes in this data set.

Conditional on the homogeneity of the three slopes, the two degree-of-freedom sum of squares for **Type** in Table 10.10 tests the hypothesis that the three regression lines have a common intercept, a null hypothesis expressed in Equation (10.4). The zero *p*-value for this test implies that the intercepts are not identical.

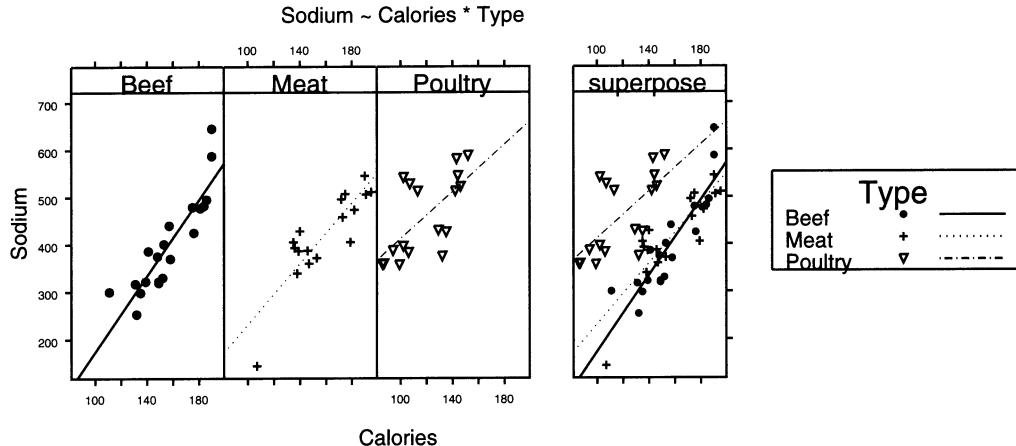


FIGURE 10.11.  $\text{Sodium} \sim \text{Calories} * \text{Type}$ . Separate regression lines.  $Y_{ij} = \mu + \alpha_i + \beta_i(X_{ij} - \bar{X}) + \epsilon_{ij}$ . See Table 10.13.

(regbb/code/hotdog.s),

(regbb/figure/hotdog.f2.eps.gz), (regbb/figure/hotdog.f2-color.eps.gz)

TABLE 10.13. Hot dog ANCOVA with separate regression lines (slopes and intercepts). See Figure 10.11.

(regbb/code/hotdog.s)

```
S-PLUS (regbb/transcript/hotdog-ancova3.st):
> ## interaction: separate intercepts and slopes
> hCTi.aov <- ancova(Sodium ~ Calories * Type, data=hotdog,
+                      par.strip.text=list(cex=1.2))
> print.trellis(position=c(0,0, 1,.6),
+                attr(hCTi.aov,"trellis"))
> ## export.eps(hh("regbb/figure/hotdog.f2.eps"))
> summary(hCTi.aov)
   Df Sum of Sq  Mean Sq F Value    Pr(F)
Calories  1 106269.7 106269.7 35.68848 0.0000003
      Type  2 227386.4 113693.2 38.18150 0.0000000
Calories:Type 2 10401.6   5200.8  1.74659 0.1852666
Residuals 48 142929.8   2977.7
```

TABLE 10.14. Four ways to use the `ancova` function.

---

```

S-PLUS (regbb/code/hotdog-ancova.s):
hotdog <- read.table(hh("datasets/hotdog.dat"), header=T)

## y ~ x                      ## constant line across all groups
ancova(Sodium ~ Calories,      data=hotdog, groups=Type)

## y ~ a                      ## different horizontal line in each group
ancova(Sodium ~               Type, data=hotdog, x=Calories)

## y ~ x + a or y ~ a + x    ## constant slope, different intercepts
ancova(Sodium ~ Calories + Type, data=hotdog)
ancova(Sodium ~ Type + Calories, data=hotdog)

## y ~ x * a or y ~ a * x    ## different slopes, and different intercepts
ancova(Sodium ~ Calories * Type, data=hotdog)
ancova(Sodium ~ Type * Calories, data=hotdog)

```

---

## 10.7 `ancova` Function

The ANCOVA has been calculated with the `ancova` function, one of the functions that we provide in (`splus.library/ancova.s`). The `ancova` function combines the S-PLUS `aov` function and the appropriate `trellis` graphics commands. The result of the function is an "ancova" object, which is essentially an "`aov`" object with a "`trellis`" attribute.

The four basic options are shown in Table 10.14. Output from each is shown in Figures 10.7, 10.6, 10.8, and 10.11 and Tables 10.9, 10.8, 10.10, and 10.13.

## 10.8 Exercises

We recommend that for all exercises involving a data set, you begin by examining a scatterplot matrix of the variables.

### 10.1. Demonstrate that the two coding schemes

$$W_{\text{female}} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad W_{\text{treat}} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

in Section 10.2 are equivalent for regression in the sense of Section 10.3 by finding the  $A$  matrix that relates them.

- 10.2.** Demonstrate that the orthogonal polynomials in Table 10.6 span the same column space as the matrix whose columns are the simple polynomials  $x = (1, 2, 3, 4, 5, 6), x^2, x^3, x^4, x^5$ .
- 10.3.** Demonstrate that the two coding schemes in Section 10.1 for the `ResidenceLocation` example are equivalent by defining the corresponding  $W$  variables and finding the  $A$  matrix that relates them.
- 10.4.** We first investigated the dataset (`datasets/water.dat`) in Exercise 4.4.
- Plot `mortality` vs `calcium`, using separate plot symbols for each value of `derbynor`. Does it appear from this plot that `derbynor` would contribute to explaining the variation in mortality?
  - Perform separate regressions of `mortality` on `calcium` for each value of `derbynor`. Compare these to the estimated coefficients in a multiple regression of `mortality` on both `calcium` and `derbynor`.
  - Interpret the regression coefficients in the multiple regression in terms of the model variables.
  - Suggest the public health conclusions of your analysis.

- 10.5.** Do an analysis of covariance with model (10.1) of the simple dataset

| <i>y</i> | <i>x</i> | <i>a</i> |
|----------|----------|----------|
| 1        | 1        | 1        |
| 2        | 2        | 1        |
| 3        | 3        | 2        |
| 4        | 4        | 2        |
| 5        | 5        | 3        |
| 6        | 6        | 3        |

Show that covariate  $x$  is significant and the grouping factor  $a$  is not.

- 10.6.** The Erie house-price data (`datasets/houseprice-erie.dat`) is introduced in Exercise 9.3. That exercise invites examination of the impact of two high-priced houses by comparing analyses with these houses included or omitted. Revisit these data, adding a dummy variable `highprice` defined as 1 if one of the two high-priced houses and 0 otherwise. Perform a stepwise regression analysis including this new variable and compare your results with those in Exercise 9.3.
- 10.7.** Reconsider the salary model in Section 10.4.1.
- Interpret, in terms of the model variables salary, age, gender, etc., the finding that  $\beta_2$  is significantly less than zero.
  - Write the null hypothesis in terms of the  $\beta_j$ 's:

$E(Y)$  for whites with 12 years of schooling is the same as  $E(Y)$  for nonwhites with 16 years of schooling.

- c. Write the null hypothesis in terms of the  $\beta_j$ 's:

$E(Y)$  increases at the rate of \$2,000 per year of schooling for whites and at the rate of \$2,500 per year of schooling for nonwhites.

- d. If the gender and race are interpreted as factors, rather than as arbitrarily coded dummy variables, then the generated dummy variables differ from the 0 and 1 coding used in Section 10.4.1. Therefore, the estimated  $\hat{\beta}_j$  will differ. Explain why the  $t$ -tests and the  $F$ -test will remain the same.

- 10.8.** Rerun the polynomial contrasts for the (`datasets/fabriwear.dat`) example in Table 10.7 without the outlier noted in Figure 10.3.

# Multiple Regression—Regression Diagnostics

In Chapter 9 we show how to set up and produce an initial analysis of a regression model with several predictors. In the present chapter we discuss ways to investigate whether the model assumptions are met and, when the assumptions are not met, ways to revise the model to better conform with the assumptions. We also examine ways to assess the effect on model performance of individual predictors or individual cases (observations).

## 11.1 Example—Rent Data

### Study Objectives

Alfalfa is a high-protein crop that is suitable as food for dairy cows. There are two research questions to ask the data in file (`datasets/rent.dat`) (from file (`alr162`) in (Weisberg, 1985)). It is thought that rent for land planted to alfalfa relative to rent for other agricultural purposes would be higher in areas with a high density of dairy cows and rents would be lower in counties where liming is required, since that would mean additional expense.

### Data Description

The data displayed in the scatterplot matrices (`sploms`) in Figure 11.1 were collected to study the variation in rent paid in 1977 for agricultural land planted to alfalfa. The unit of analysis is a county in Minnesota; the 67 counties with appreciable rented farmland are included. Note that we

automatically conditioned the splom on the factor `lime`. The original data include:

`rnt.alf`: average rent per acre planted to alfalfa  
`rnt.till`: average rent paid for all tillable land  
`cow.dens`: density of dairy cows (number per square mile)  
`prop.past`: proportion of farmland used as pasture  
`lime`: “lime” if liming is required to grow alfalfa; “no.lime” otherwise  
 (Lime is a calcium oxide compound that is spread on a field as a fertilizer.)

We added one more variable

`alf.till`: the ratio of `rnt.alf` to `rnt.till`

to investigate the relative rent question.

### 11.1.1 Rent Levels

It is immediately clear from the sploms in Figure 11.1 that `lime` is very important in the distribution of `cow.dens` and `prop.past` as neither has any large values in the `lime` splom. The ratio `alf.till` is slightly higher in the `no.lime` splom.

`lime` does not seem to have an effect on either of the rent variables `rnt.alf` or `rnt.till`, as their panels have similar distributions in both sploms. The regression analysis of `rnt.alf` in Table 11.1 supports that impression as `lime` has a very low *t*-value. `prop.past` also has a very low *t*-value.

We therefore look at a simpler model, without the `prop.past` predictor but with the `cow.dens:lime` interaction, in Figure 11.2 and Table 11.2. Although the regression analysis shows the `lime` coefficient as not significant, it shows the interaction of `lime` with cow density to be on the edge of significance ( $p = .055$ ). We left both in the model because there appears to be much higher variability in the residuals for high values of `rnt.till` and lower variability in the residuals for low values of `cow.dens` in the `no.lime` counties as indicated in Figure 11.3.

Our conclusion from this portion of the analysis is that rent for alfalfa is related to rent for tillage and to cow density. The relationship with cow density may depend on the need for lime. We need to investigate the variability of the residuals.

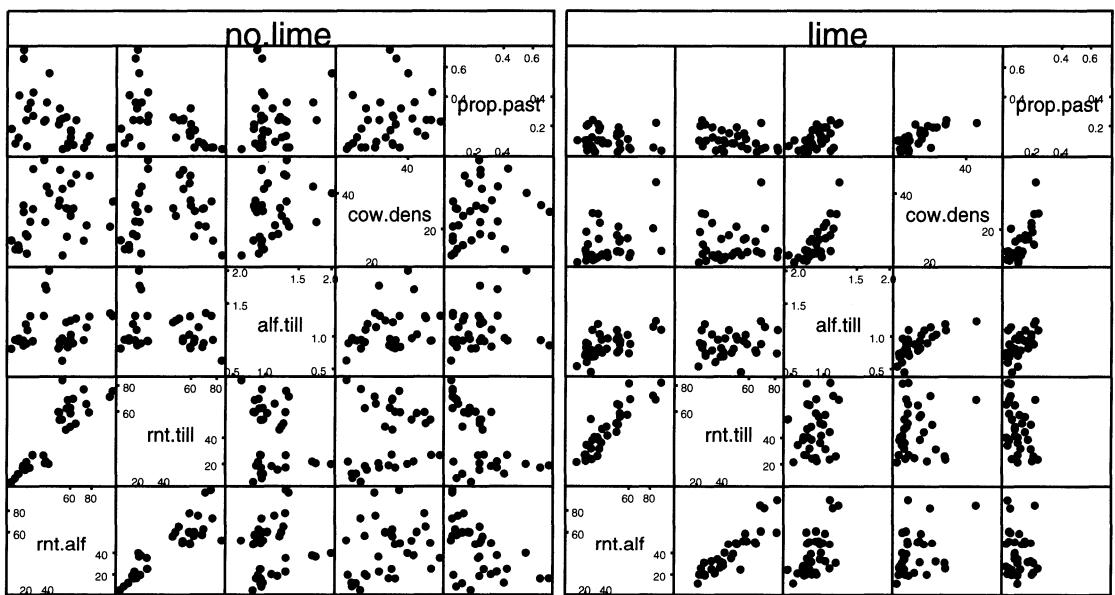


FIGURE 11.1. Scatterplot matrices of all variables conditioned on `lime`.  
(`regc/code/rent4.s`), (`regc/figure/rent1.eps.gz`)

TABLE 11.1. rent.alf regressed against all other observed variables.  
(regc/code/rent4.s)

---

```
S-PLUS (regc/transcript/rent.lm31.st):
> rent.lm31 <- lm(rnt.alf ~ rnt.till + cow.dens + prop.past + lime,
+                     data=rent)
> summary(rent.lm31, corr=F)

Call: lm(formula = rnt.alf ~ rnt.till + cow.dens + prop.past + lime,
       data = rent)
Residuals:
    Min      1Q  Median      3Q     Max 
 -21.23 -4.869 -0.02874  4.755 27.77 

Coefficients:
            Value Std. Error t value Pr(>|t|)    
(Intercept) -3.3341   4.1008   -0.8130  0.4193    
rnt.till     0.8833   0.0690   12.8007  0.0000    
cow.dens     0.4318   0.1080    3.9989  0.0002    
prop.past   -11.3805  11.8937   -0.9568  0.3424    
lime        -0.5059   1.4245   -0.3551  0.7237    

Residual standard error: 9.311 on 62 degrees of freedom
Multiple R-Squared: 0.8404
F-statistic: 81.6 on 4 and 62 degrees of freedom, the p-value is 0
> anova(rent.lm31)
Analysis of Variance Table

Response: rnt.alf

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value    Pr(F)    
rnt.till  1  25824.25 25824.25 297.8904 0.0000000  
cow.dens  1   2386.32  2386.32  27.5269 0.0000020  
prop.past 1     73.86   73.86   0.8520 0.3595658  
lime     1     10.93   10.93   0.1261 0.7237064  
Residuals 62   5374.81   86.69
```

---

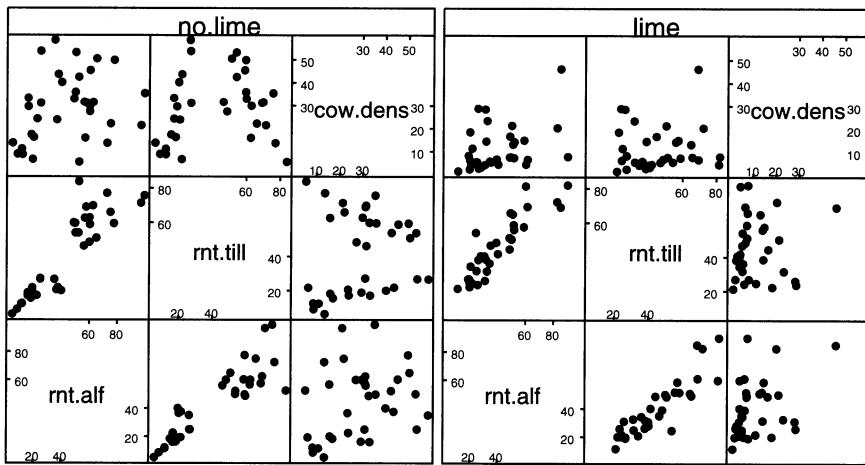


FIGURE 11.2. Scatterplot matrices of `rnt.alf` with 2  $X$ -variables conditioned on `lime`.  
 (regc/code/rent4.s), (regc/figure/rent2.eps.gz)

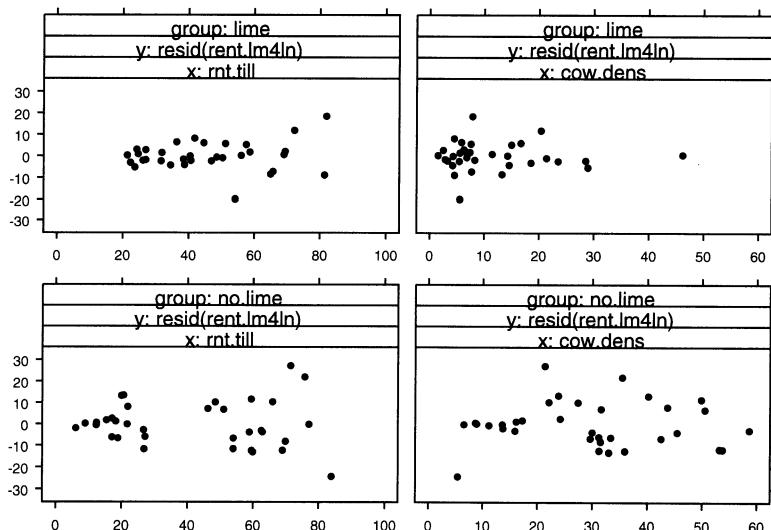


FIGURE 11.3. Residuals from  $\text{rnt.alf} \sim \text{rnt.till} + \text{cow.dens} * \text{lime}$  (in Table 11.2 and Figure 11.2) plotted against the  $X$ -variables conditioned on `lime`.  
 (regc/code/rent4.s), (regc/figure/rent4lnres.eps.gz)

TABLE 11.2. *rent.alf* regressed against all variables except *prop.past*, and including the interaction of cow density with lime.  
(regc/code/rent4.s)

---

```
S-PLUS (regc/transcript/rent.lm4ln.st):
> rent.lm4ln <- lm(rnt.alf ~ rnt.till + cow.dens +
+                      lime + cow.dens:lime, data=rent)
> summary(rent.lm4ln, corr=F)

Call: lm(formula = rnt.alf ~ rnt.till + cow.dens + lime + cow.dens:lime,
       data = rent)
Residuals:
    Min      1Q  Median      3Q     Max 
-24.35 -4.251 -0.1938  4.151  27.19 

Coefficients:
            Value Std. Error t value Pr(>|t|)    
(Intercept) -5.9584   3.0117   -1.9784  0.0523  
rnt.till     0.9269   0.0536   17.2801  0.0000  
cow.dens     0.4567   0.0991    4.6110  0.0000  
lime        -3.6034   2.1642   -1.6650  0.1010  
cow.dens:lime 0.1926   0.0986    1.9530  0.0553  

Residual standard error: 9.103 on 62 degrees of freedom
Multiple R-Squared: 0.8474
F-statistic: 86.07 on 4 and 62 degrees of freedom, the p-value is 0
> anova(rent.lm4ln)
Analysis of Variance Table

Response: rnt.alf

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value    Pr(F)    
rnt.till   1  25824.25 25824.25 311.6144 0.00000000
cow.dens   1   2386.32  2386.32  28.7951 0.00000013
lime       1      5.42     5.42   0.0654 0.7989607
cow.dens:lime 1   316.09   316.09   3.8141 0.0553397
Residuals  62  5138.09    82.87
```

---

TABLE 11.3. alf.till ratio regressed against cow density | lime and proportion in pasture.  
(regc/code/rent4.s)

---

```

S-PLUS (regc/transcript/rent.lm12p.st):
> rent.lm12p <- lm(alf.till ~ lime * cow.dens + prop.past, data=rent)
> summary(rent.lm12p, corr=F)

Call: lm(formula = alf.till ~ lime * cow.dens + prop.past, data = rent)
Residuals:
    Min      1Q  Median      3Q     Max 
-0.3342 -0.1247 -0.02033  0.1045  0.7853 

Coefficients:
            Value Std. Error t value Pr(>|t|)    
(Intercept) 0.7896   0.0564   14.0080  0.0000    
lime        -0.0969   0.0533   -1.8161  0.0742    
cow.dens     0.0094   0.0026    3.6383  0.0006    
prop.past    0.1899   0.2267    0.8376  0.4055    
lime:cow.dens 0.0039   0.0024    1.6185  0.1106    

Residual standard error: 0.2228 on 62 degrees of freedom
Multiple R-Squared:  0.3657
F-statistic: 8.937 on 4 and 62 degrees of freedom, the p-value is 9.167e-006
> anova(rent.lm12p)
Analysis of Variance Table

Response: alf.till

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq F Value    Pr(F)    
lime       1  0.845552 0.8455516 17.03109 0.0001116  
cow.dens   1  0.754167 0.7541675 15.19043 0.0002412  
prop.past  1  0.044950 0.0449505  0.90539 0.3450349  
lime:cow.dens 1  0.130055 0.1300552  2.61957 0.1106295  
Residuals  62  3.078147 0.0496475

```

---

### 11.1.2 Alfalfa Rent Relative to Other Rent

Returning to the sploms in Figure 11.1, we see that that `lime` puts an upper bound on the `alf.till` ratio. The ratio does seem to go up with cow density and seems to have a variance relation with proportion in pasture. In Table 11.3, a regression of the `alf.till` ratio against the nonrent variables, we see that we can drop the `prop.past` variable.

TABLE 11.4. `alf.till` ratio regressed against cow density | lime. See Figure 11.4.  
 (regc/code/rent4.s)

```
S-PLUS (regc/transcript/rent.lm12.st):
> rent.lm12m <- ancova(alf.till ~ lime * cow.dens, data=rent,
+                         par.strip.text=list(cex=1.2))
> print.trellis(position=c(0,0, 1,.6),
+                 attr(rent.lm12m,"trellis"))
> anova(rent.lm12m$aov)
Analysis of Variance Table

Response: alf.till

Terms added sequentially (first to last)
  Df Sum of Sq   Mean Sq F Value    Pr(F)
lime     1  0.845552 0.8455516 17.11214 0.00010631
cow.dens 1  0.754167 0.7541675 15.26272 0.00023094
lime:cow.dens 1  0.140172 0.1401718  2.83677 0.09707772
Residuals 63  3.112981 0.0494124
```

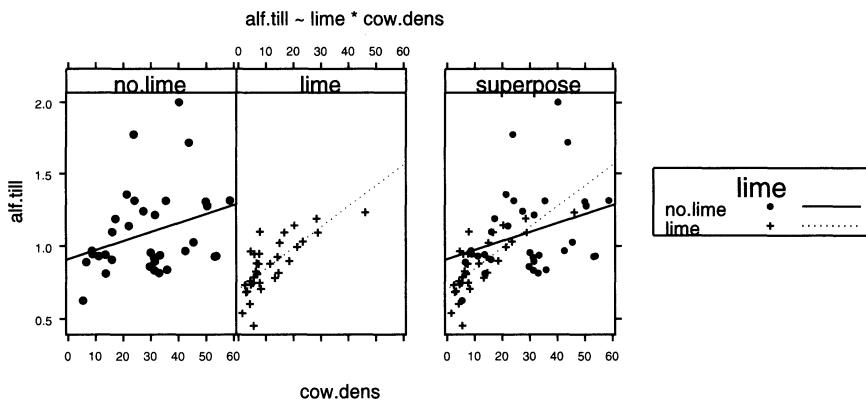


FIGURE 11.4. ANCOVA `rnt.alf/rnt.till ~ cow.dens | lime`. See Table 11.4.  
 (regc/code/rent4.s), (regc/figure/rent.lm12m.eps.gz)

We continue with Table 11.4 and Figure 11.4, which show an ordinary analysis of covariance with model

$$\text{alf.till} \sim \text{cow.dens} * \text{lime} \quad (11.1)$$

The ANOVA table in Table 11.4 shows the interaction is not quite significant.

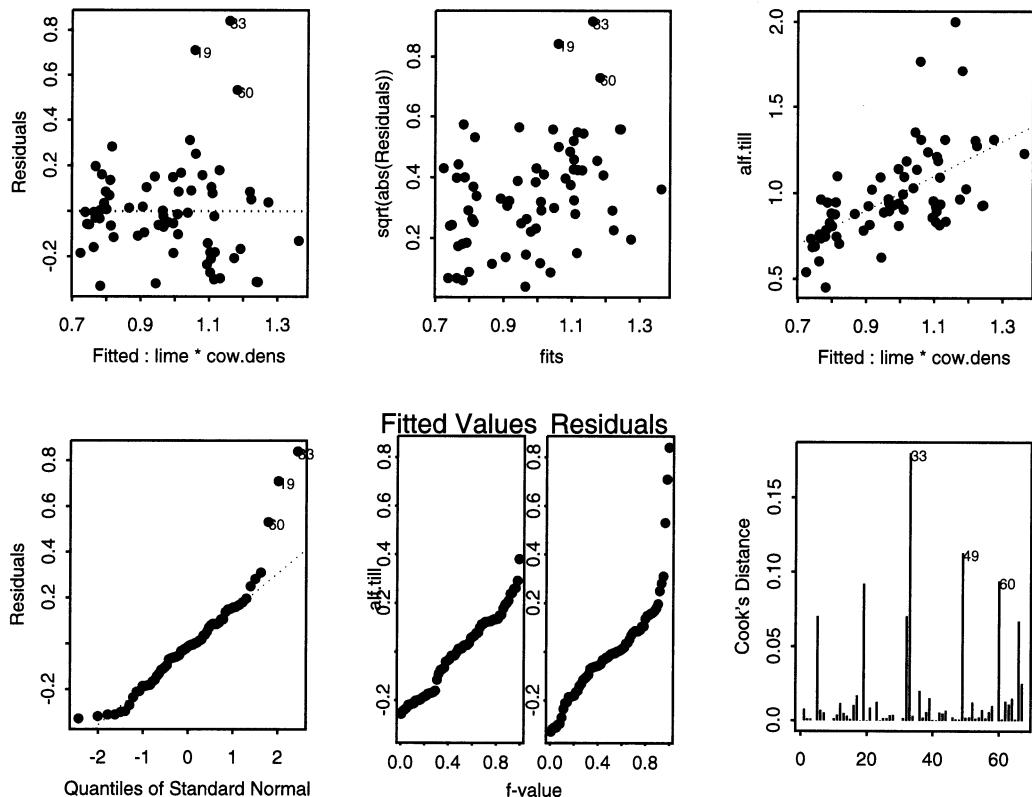


FIGURE 11.5. Residuals from ANCOVA (`rnt.alf/rnt.till`) ~ `cow.dens | lime`. See Table 11.4 and Figure 11.4.  
`(regc/code/rent4.s)`, `(regc/figure/rent.plot.lm12m.eps.gz)`

We choose to investigate individual points by looking at the standard plots of the residuals in Figure 11.5 and the regression diagnostics in Figure 11.6. These show the three points (19, 33, 60) in the `no.lime` group and the single point (49) in the `lime` group as being potentially influential. Figure 11.6, produced with our functions `lm.case.s` and `plot.case.s`, includes boundaries for the standard recommended thresholds for the various diagnostic measures discussed in Section 11.3.

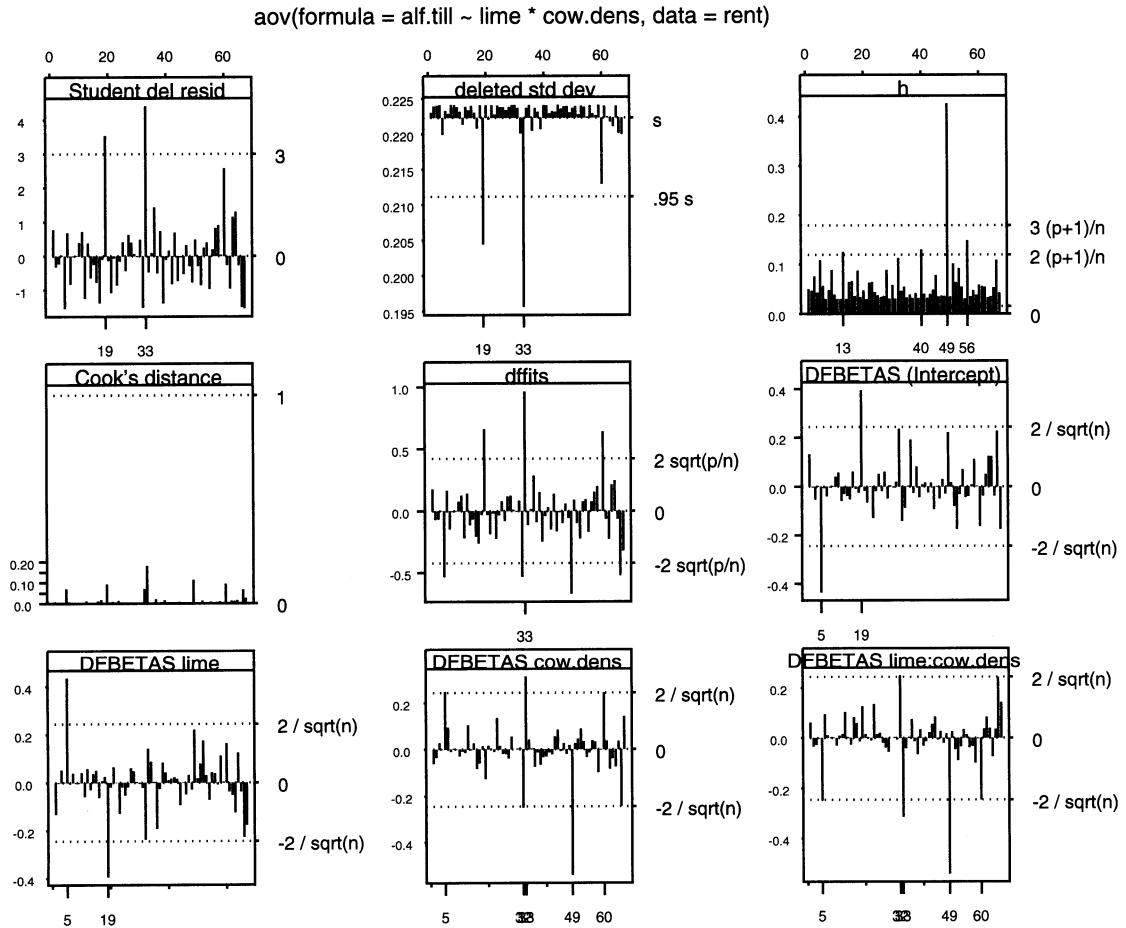


FIGURE 11.6. Diagnostics from ANCOVA (`rnt.alf/rnt.till`) ~ `cow.dens` | `lime`. See Table 11.4 and Figure 11.4.

(`regc/code/rent4.s`), (`regc/figure/rent.diag.lm12m.eps.gz`)

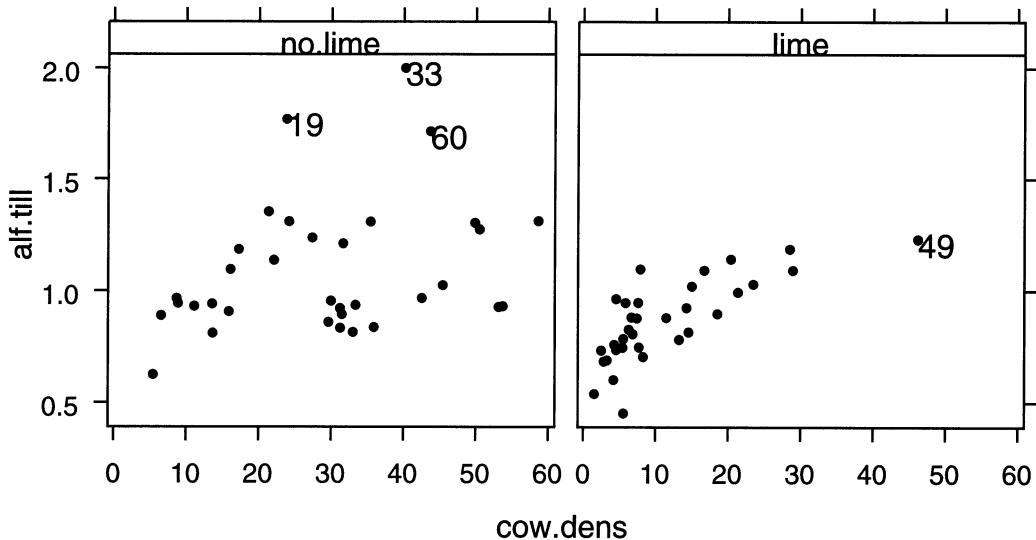


FIGURE 11.7. Identified points in ANCOVA ( $rnt.alf/rnt.till \sim \text{cow.dens} | \text{lime}$ ).  
 (regc/code/rent4.s), (regc/figure/rent.text.lm12m.eps.gz)

We locate the potentially influential points in Figure 11.7 and see them as the three counties with the highest ratios and the one `lime` county with an unusually high cow density. In Section 11.3 we will discuss the statistics displayed in Figures 11.5 and 11.6 as well as their interpretation.

We redo the analysis without these four points in Table 11.5 and Figure 11.8. After isolating these four counties we see significantly different slopes in the `no.lime` and `lime` counties.

Our conclusion at this step is that for most counties, there is a linear relationship of the rent ratio to the cow density, with the slope depending on the need for lime. The three `no.lime` counties and the one `lime` county need additional investigation.

TABLE 11.5. ANCOVA of `alf.till` ratio regressed against cow density and lime with four removed observations. See Figure 11.8. Compare to Table 11.4.  
 (regc/code/rent4.s)

```
S-PLUS (regc/transcript/rent.lm12ms.st):
> rent.lm12ms <- ancova(alf.till ~ lime * cow.dens,
+                         data=rent[-c(19, 33, 60, 49),],
+                         ylim=range(rent$alf.till),
+                         par.strip.text=list(cex=1.2))
> anova(rent.lm12ms)
Analysis of Variance Table

Response: alf.till

Terms added sequentially (first to last)
          Df Sum of Sq   Mean Sq F Value    Pr(F)
lime      1  0.428301 0.4283011 17.81112 0.000085326
cow.dens  1  0.395004 0.3950043 16.42645 0.000150029
lime:cow.dens 1  0.232643 0.2326434  9.67459 0.002876753
Residuals 59  1.418763 0.0240468
```

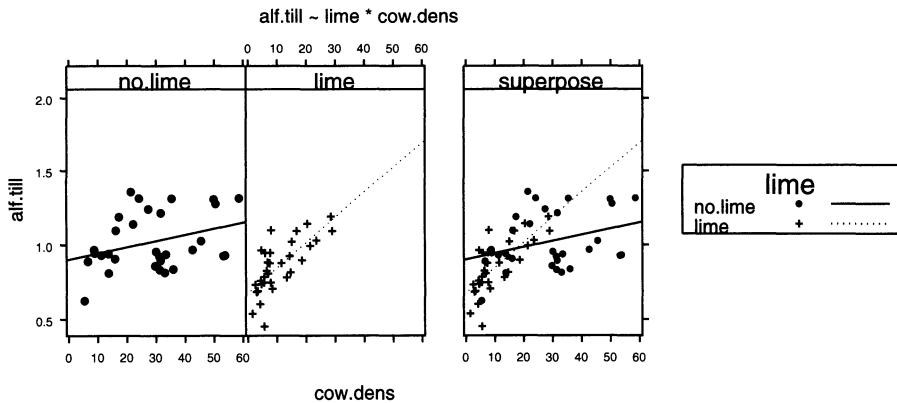


FIGURE 11.8. Removed observations from ANCOVA `rnt.alf/rnt.till ~ cow.dens | lime`. See Table 11.5. Compare to Figure 11.4.  
 (regc/code/rent4.s), (regc/figure/rent.lm12ms.eps.gz)

## 11.2 Checks on Model Assumptions

We assume in Section 9.2 that the model error terms  $\epsilon_i \sim \text{NID}(0, \sigma^2)$  (Normal Independently Distributed), that is that they have the same variance  $\sigma^2$  for all cases, are mutually uncorrelated or independent, and are normally distributed. In order for the conclusions from our analyses to be valid, these assumptions must be true. Therefore, we discuss ways to verify the assumptions and then suggest some remedies when assumptions are not met.

### 11.2.1 Scatterplot Matrix

We previously mentioned the importance of routinely producing scatterplot matrices as part of analyses involving several variables. We produced many such plots in our discussion in Section 11.1. Here we focus on the rows of the scatterplot matrix that correspond to the response variables. The panels in these rows, the plots of the response  $y$  vs each of the explanatory variables  $x_j$ , should each be approximately linear. In Section 11.1.1 the response is shown in the `rnt.alf` row in Figure 11.1 and redisplayed in Figure 11.2. In Section 11.1.2 the response is the `alf.till` row in Figure 11.1 and redisplayed in Figure 11.4. If the plot of  $y$  against any explanatory variable suggests curvature in the relationship, the analyst should consider transforming either the response variable or that explanatory variable so that following transformation the plot of  $y$  vs the transformed  $x_j$  is close to linear. A successful transformation suggests the use of this transformed predictor rather than the original in the regression model. Exercise 11.5 explores this idea.

### 11.2.2 Residual Plots

Before a model can be accepted for use in explanation or prediction, the analyst should produce and examine plots involving the residuals calculated from the fit of the model to the data. The residuals  $e_i$  should be plotted vs each of the following, one plot point per case:

- the fitted values of the response  $\hat{y}_i$
- each of the model's explanatory variables  $x_j$
- possibly other variables under consideration for the model but not yet a part of it
- time, if the data are time-ordered

In addition, the partial residuals (see Section 9.14.1) should be plotted against the corresponding predictors and against the residuals from re-

gressing each predictor against the other predictors (added variable plots; see Section 9.14.4). Ideally, each of these plots should exhibit no systematic character and have random scatter about the horizontal line at 0, the mean of the  $e_i$ .

In addition, in order to check for normality, the analyst should produce a normal probability plot of the residuals. If there is doubt that this plot confirms normality, the analyst can request the  $p$ -value from an all-purpose test of normality having good power against a variety of alternatives, such as the Shapiro–Wilk test mentioned in Section 5.7.

If a residual plot suggests that an assumption is not met, the analyst must seek a remedy following which the assumption is met.

We show in Figure 11.9 the normal probability plot for the rent ratio `alf.till` analysis in Table 11.4 and Figure 11.4. It does not look normal. Compare this plot to Figure 11.10, which shows probability plots of six normal and six non-normal variables.

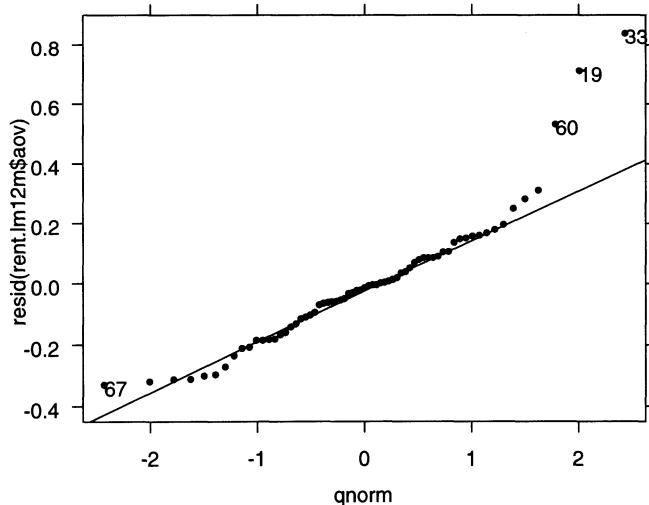


FIGURE 11.9. Normal plot of residuals from ANCOVA `rnt.alf/rnt.till ~ cow.dens | lime`. See Table 11.4 and Figure 11.4. Because the results do not look normal, we identified the four most extreme points. Three of them are the three `no.lime` counties that we had previously identified.  
`(regc/code/rent4b.s)`, `(regc/figure/rent.residn.eps.gz)`

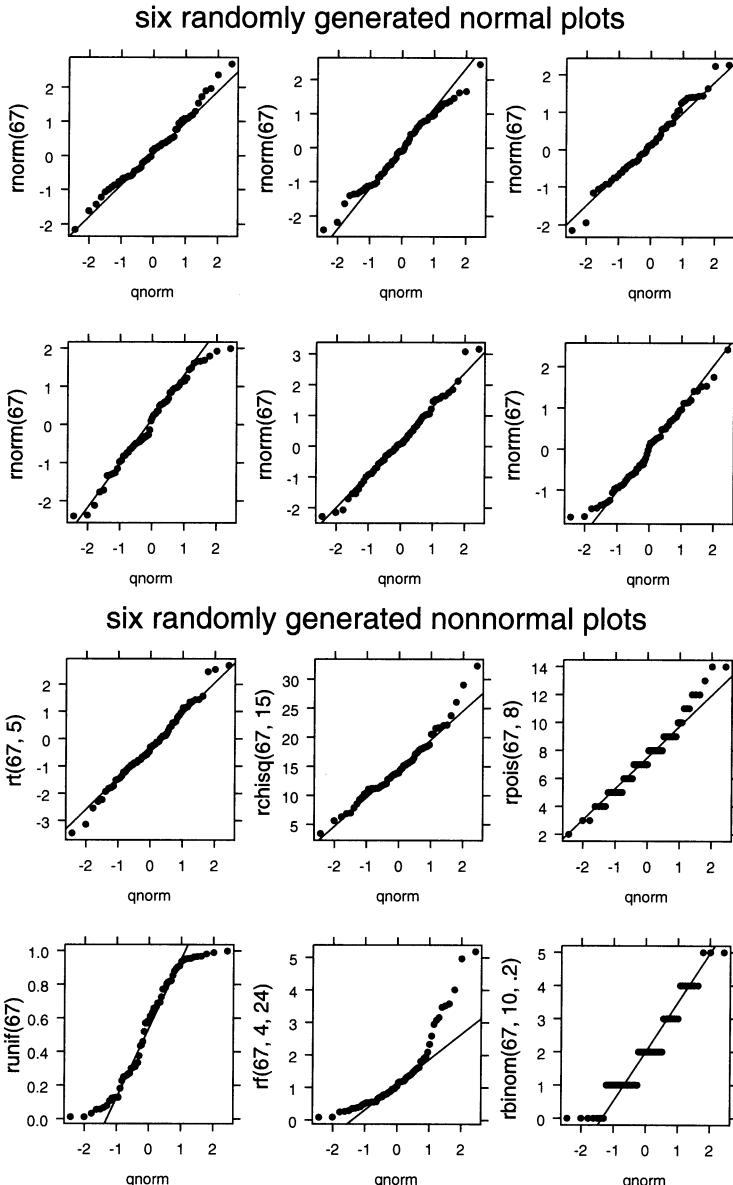


FIGURE 11.10. Normal plot of six randomly generated normal variables and six randomly generated nonnormal variables. These plots are placed here to help you calibrate your eye to what normal and nonnormal distributions look like when plotted against the normal quantiles.

(regc/code/norm.prob.plot.s),

(regc/figure/norm.prob.plot.eps.gz), (regc/figure/nonnorm.prob.plot.eps.gz)

Figure 11.11 shows several plots of the residuals and partial residuals from the model in Table 11.4 and Figure 11.4. From the `lime` column, we see that the ratio `alf.till` is higher for `lime=-1` (no lime) than for `lime=1` (lime). The pattern is similar in the observed variable plots in Row 1 and the partial residuals plots in Row 3, suggesting that the `lime` effect is independent of the other variables.

From the `cow.dens` column, we again see similar behavior in Rows 1 and 3. We also note the higher variability in  $Y$  for the higher densities. We get a sense of why we see that difference in variability from the interaction `lime:cow.dens` column. Here we see, most clearly in the partial residuals plot in Row 3, that the high variability is observed when the interaction variable is negative, corresponding to the `no.lime` counties.

### 11.3 Case Statistics

Many of the diagnostics discussed in this chapter fall under the heading *case statistics*, i.e., they have a value for each of the  $n$  cases in the data set. If a case statistic has a value that is unusual, based on thresholds we discuss, the analyst should *scrutinize* the case. One action the analyst might take is to delete the case. This is justified if the analyst determines the case is not a member of the same population as the other cases in the data set. But deletion is just one possibility. Another is to determine that the flagged case is unusual in ways apart from those available in its information in the present data set, and this may suggest a need to add one or more additional predictors to the model.

There are many case statistics used in regression diagnostics. The concepts are complex and the notation more so. We summarize the notation in Table 11.6. We discuss each of the formulas and illustrate them with the diagnostic plots for the rent data that we originally showed in Figure 11.6. We reproduce each of the panels in that figure as a standalone plot here as part of the discussion.

We focus on five distinct case statistics, each having a different function and interpretation. (One of these, `DFBETAS`, is a vector with a distinct value for each regression coefficient including the intercept coefficient.) For small data sets the analyst may choose to display each of these case statistics for all cases. For larger data sets we suggest that the analyst display only those values of the case statistics that exceed a threshold, or flag, indicating that the case is unusual in some way. Recommended thresholds are mentioned in the following sections.

**Leverage** measures how unusual a case is with respect to the values of its predictors, i.e., whether the values of a case's predictors are an outlying

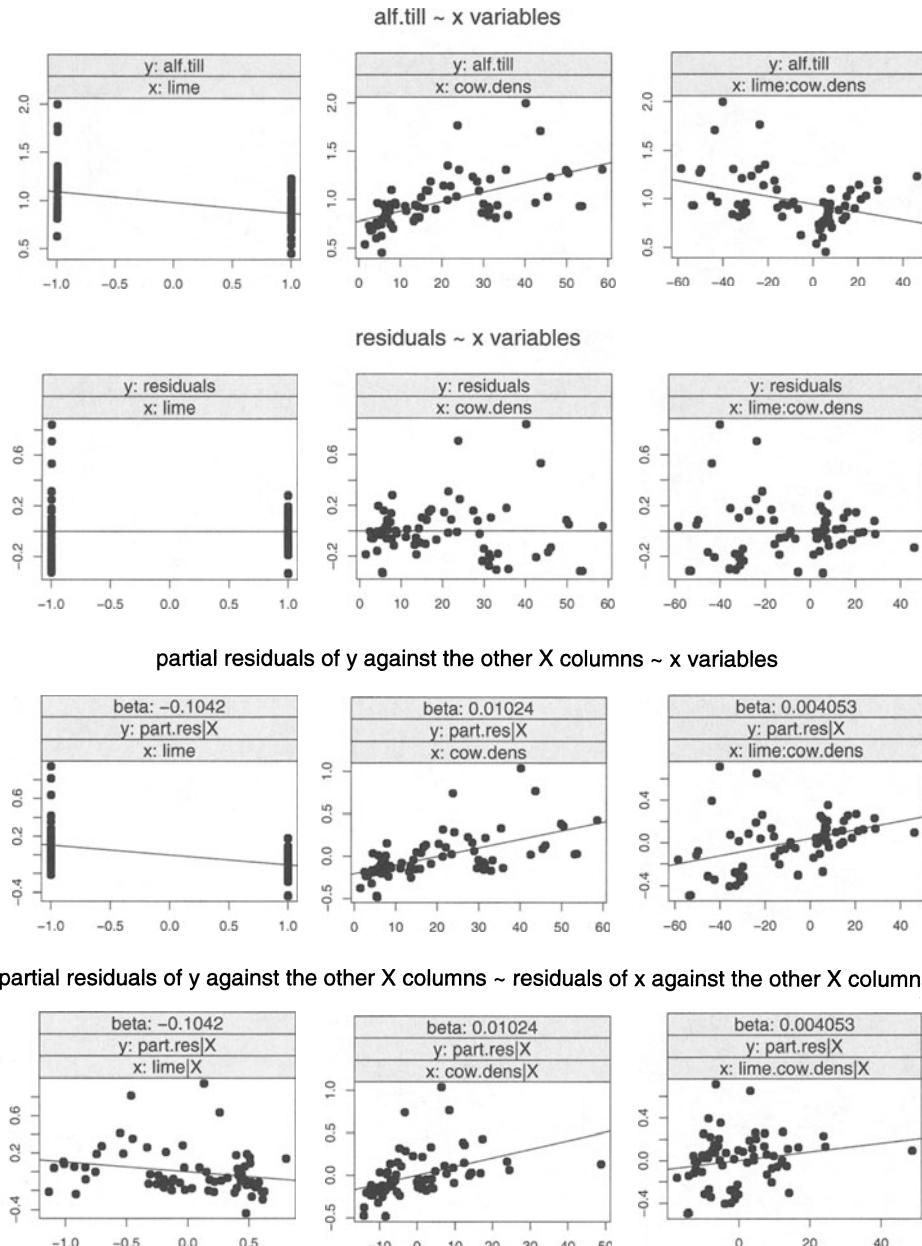


FIGURE 11.11. Row 1 shows the response variable *alf.till* against each of the three predictors. Row 2 shows the ordinary residuals  $e = Y - \hat{Y}$  from the regression on all three variables against each of the three predictors. Row 3 shows the “partial residuals plots”, the partial residuals for each predictor against that predictor. Row 4 shows the “added variable plots”, the partial residuals against the residuals of  $X_j$  regressed on the other two predictors.

(`regc/code/rent4b.s`), (`regc/figure/rent.resid.plots.eps.gz`)

Table 11.6: Regression Diagnostics Formulas

| Name                                  | Notation and definition  | Sequenced calculation formulas  | Description  |
|---------------------------------------|--|---|--|
| observed response variable            | $\mathbf{Y}_{n \times 1}$                                      |   |  |
| observed predictor variables          | $\mathbf{X}_{n \times (1+p)}$                                  | $\mathbf{X} = [\mathbf{1} \ X_1 \ X_2 \ \dots \ X_p]$   |  |
| fitted value                          | $\hat{\mathbf{Y}} = (\hat{Y}_i)$                               | $X\hat{\beta}$  |  |
| residual                              | $e_i$  | $Y_i - \hat{Y}_i$   | All $n$ observations   |
| standard deviation                    | $s = \sqrt{\text{MSE}} = \sqrt{\text{var}(Y_i   X)}$           | $\sqrt{\sum e_i^2 / (n - p - 1)}$   |  |
| leverage                              | $h_{ii} = h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i}$    | $\text{diag}(X(X'X)^{-1}X')$  |  |
| variance of $e_i$                     | $\text{var}(e_i)$  | $s^2(1 - h_{ii})$   |  |
| variance of $\hat{Y}_i$               | $\text{var}(\hat{Y}_i)$  | $s^2 h_{ii}$  |  |
| standardized residual                 | $e_i^*$  | $e_i / (s\sqrt{1 - h_{ii}})$  | $e_i / (\sigma\sqrt{1 - h_{ii}}) \sim N(0, 1)$ when $H_0$ is true  |
|                                       |  |   |  |
| data with $i^{\text{th}}$ row deleted |  | $X_{(1, 2, \dots, i-1, i+1, \dots, n), (0, 1, \dots, p)}$   |  |
| deleted regression coefficients       | $\hat{\beta}_{(i)} = (X'_{(i)} X_{(i)})^{-1} X'_{(i)} Y_{(i)}$ | See description in Section 11.3.6.  |  |
| deleted standard deviation            | $s_{(i)} = \sqrt{\text{MSE}_{(i)}}$                            | $\sqrt{(n-p)s^2 - e_i^2/(1-h_{ii})}/(n-p-1)$  | Estimation of $\beta$ based on $n-1$ observations, all except $i$ . The definition isn't efficient. Use the algorithm in Section 11.3.6. |
| deleted predicted value               | $\hat{Y}_{(i)}$  | $X_{(1, 2, \dots, i-1, i+1, \dots, n), (0, 1, \dots, p)} \hat{\beta}_{(i)}$   | $n-1$ observations, all except $i$ .   |
| Studentized deleted residual          | $t_i$  | $e_i / (s_{(i)}\sqrt{1 - h_{ii}})$  | Prediction $\hat{Y}_i$ based on the remaining $n-1$ observations   |
| Cook's distance                       |  |   | $t_i \sim t_{n-p-2}$ when $H_0$ is true  |
| covariance of coefficients            |  |   |  |
| DFBETAS                               |  | $D_i = (\hat{Y} - \hat{Y}_{(i)})' (\hat{Y} - \hat{Y}_{(i)}) / (p s^2)$<br>$= (\hat{\beta} - \hat{\beta}_{(i)})' X'X (\hat{\beta} - \hat{\beta}_{(i)}) / (p s^2)$  |  |
| inverse of crossproduct of $X$        | $C = (c_{ij})$   | $(X'X)^{-1}$  | Estimated covariance matrix of regression coefficients   |
| covariance of coefficients            | $s^2 C$  | $s^2 (X'X)^{-1}$  |  |
| DFFITS                                |  | $\text{DFBETAS}_{i,k} = (\hat{\beta}_k - \hat{\beta}_{k(i)}) / (s_{(i)} \sqrt{c_{kk}})$<br>$\text{DFFITS}_i = (\hat{Y}_i - \hat{Y}_{(i)}) / (s_{(i)} \sqrt{h_i})$ | Standardized $\Delta \hat{\beta}_k$ when observation $i$ is deleted<br>Standardized $\Delta \hat{Y}_i$ when observation $i$ is deleted   |

point in the  $p$ -dimensional space of predictors. Unlike the other case statistics, leverage does not involve the response variable.

**Studentized deleted residuals** suggest how unusual cases are with respect to the case's value of the response variable.

**Cook's distance** is a combined measure of the unusualness of a case's predictors and response. It sometimes happens that a case is flagged by Cook's distance but not quite flagged by leverage or Studentized deleted residuals.

**DFFITS** indicates the extent to which deletion of the case impacts predictions made by the model.

**DFBETAS**, one for each regression coefficient, show the extent to which deletion of a case would perturb that regression coefficient.

In the following sections we discuss these statistics in turn, presenting two formulas for each of them. The first, the definitional formula, is intended to be intuitive. It is used to explain to the reader what the formula measures and why it is helpful to view it in an analysis. It is also inefficient and should not be used as a computational formula. The second formula, the computational formula, is an order of magnitude more efficient for computation. It is not intuitive. We leave for Exercise 11.8 the proofs that the two sets of formulas are equivalent.

### 11.3.1 Leverage

The calculation of leverages is briefly addressed in 9.2.1. Leverages measure how unusual a case is with respect to its set of predictors. Unlike other measures in this chapter, leverages do not involve the response variable. The leverage  $h_{ii}$  of case  $i$ , usually abbreviated to  $h_i$ , is the  $i^{\text{th}}$  diagonal entry of the *hat matrix*  $H = X(X'X)^{-1}X'$ . This matrix has come to be called the hat matrix because in matrix notation the predicted response is  $\hat{Y} = X(X'X)^{-1}X'Y = HY$ , i.e.,  $H$  transforms  $Y$  to  $\hat{Y}$  by placing a “hat” on the  $Y$ . It can be shown (see Exercise 11.9) that all leverages satisfy  $\frac{1}{n} \leq h_i \leq 1$ . If a model contains  $p$  predictors, an excessively large leverage is one for which

$$h_i > \frac{2(p+1)}{n} \quad \text{or} \quad h_i > \frac{3(p+1)}{n} \tag{11.2}$$

These suggested rules derive from the fact that the average of all  $n$  leverages is  $\frac{p+1}{n}$ , so they are based on exceeding 2 or 3 times this average. A case that is flagged because its leverage exceeds one or both of these thresholds has a value for at least one predictor that is unusual compared to values of

such predictors for other cases. We can show that

$$h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i} \quad \text{and} \quad h_{ij} = \frac{\partial \hat{y}_i}{\partial y_j}$$

The leverage  $h_i$  of case  $i$  is geometrically interpreted as the generalized (Mahalanobis) distance of  $X_i$  (the  $i^{\text{th}}$  row of  $X$ ) from the  $(p+1)$ -dimensional centroid of all  $n$  rows of  $X$ .

More complicated forms of leverage have been devised to diagnose a group of cases that when considered together are unusual but when considered individually are not unusual.

Figure 11.12 displays the leverages for each case of the fit of the rent data using Model (11.1). This figure includes horizontal dotted lines demarking the two leverage thresholds given above. We observe that county 49 exceeds both thresholds, telling us that this county (requiring lime) has an unusually large cow.dens.

### 11.3.2 Deleted Standard Deviation

The deleted standard deviation  $s_{(i)}$  is defined to be the value of  $s$  calculated from the same regression model using all cases *except* case  $i$ . Because the primary use of the  $s_{(i)}$  is in the definition of the Studentized deleted residuals, there are no standard rules for interpreting these values themselves.

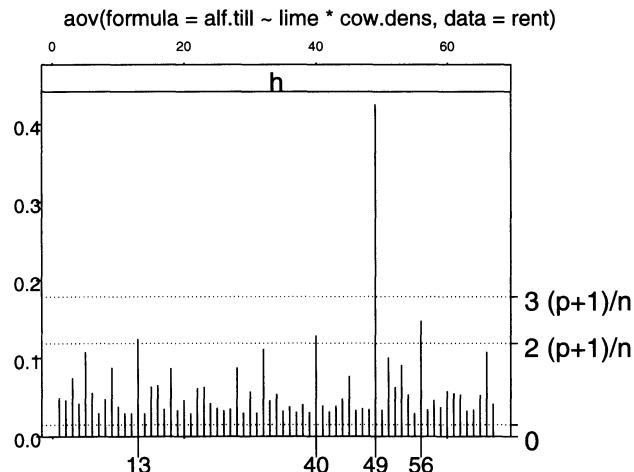


FIGURE 11.12. Leverage for Model (11.1) for rent data.  
(regc/code/rent4b.s), (regc/figure/h.eps.gz)

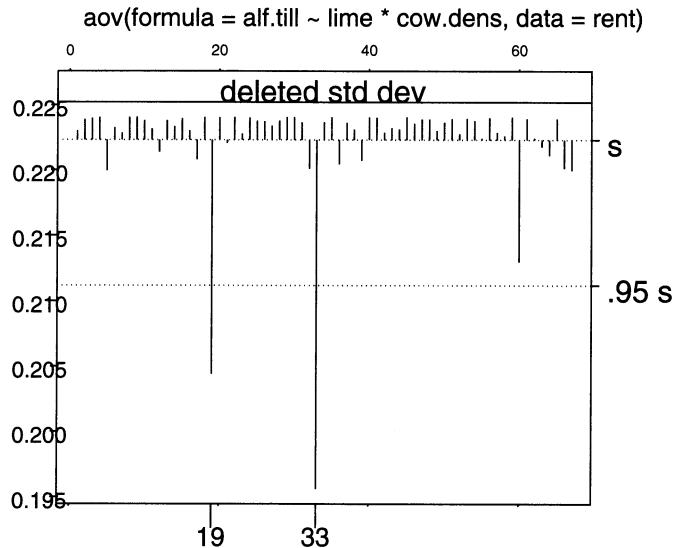


FIGURE 11.13. Deleted standard deviations for Model (11.1) for rent data.  
`(regc/code/rent4b.s), (regc/figure/si.eps.gz)`

We compare the  $s_{(i)}$  values to two thresholds,  $.95s$  and  $1.05s$ . If deletion of an observation shifts the estimated standard deviation by 5% in either direction, we note it on the graph and choose to investigate the observation.

Figure 11.13 shows the deleted standard deviations for the rent data. We see two observations, 19 and 33, that are below our lower threshold.

### 11.3.3 Standardized and Studentized Deleted Residuals

The standardized and Studentized residuals help to assess the effect of each individual case on the calculated regression relationship. For case  $i$  the standardized residual

$$e_i^* = e_i / \sqrt{\text{var}(e_i)} \quad (11.3)$$

is the calculated residual,  $e_i$ , standardized by dividing by its estimated standard error

$$\sqrt{\text{var}(e_i)} = s \sqrt{1 - h_i} \quad (11.4)$$

Note that because this standard error depends on  $i$ , it differs slightly from case to case. The standardized residual is also called the *internally standardized residual* because the calculation of  $s$  includes case  $i$ .

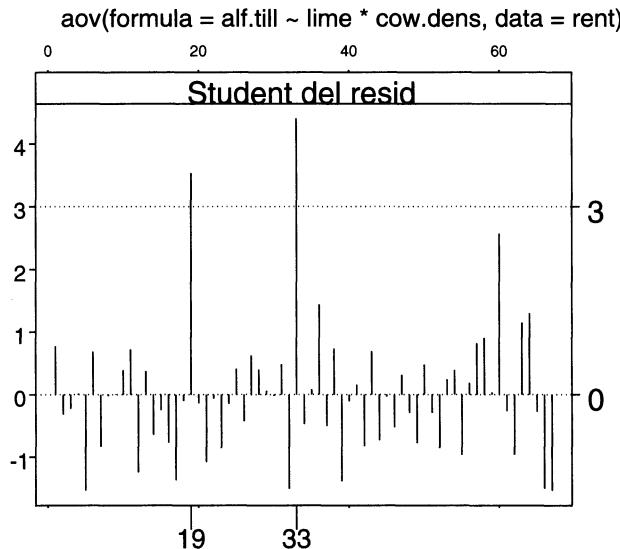


FIGURE 11.14. Studentized deleted residuals for Model (11.1) for rent data.  
 (regc/code/rent4b.s), (regc/figure/stu.res.eps.gz)

The *Studentized deleted residual*, also called the *externally standardized residual*, for case  $i$  is calculated from the regular residuals, the deleted standard deviations, and the hat diagonals.

$$t_i = \frac{e_i}{s_{(i)} \sqrt{1 - h_i}} \quad (11.5)$$

As implied by this notation,  $t_i$  has a Student's  $t$  distribution with  $n - p - 1$  degrees of freedom. Considering the  $t$  distribution with moderate degrees of freedom, we say that case  $i$ 's response value is "unusual" (the actual response differs "appreciably" from the predicted response) if its absolute Studentized deleted residual exceeds 2 or 3. Such a case may be termed an *outlier*. We recommend a threshold of 2 for small data sets and 3 for large data. The reason for this recommendation is that for a large data set, 2 is the approximate 97.5<sup>th</sup> percentile of the  $t$  distribution so that when the model assumptions are satisfied for all cases, approximately 5% of these residuals will exceed 2 by chance alone.

We prefer the use of Studentized deleted residuals rather than standardized residuals because the former are interpretable as  $t$  statistics but the latter are not. A reason is that the numerator and denominator of  $t_i$  are statistically independent, but the numerator and denominator of the standardized residuals  $e_i^*$  are not independent.

It can be shown (see Exercise 11.8c) that the Studentized deleted residual defined intuitively in Equation (11.5) can be calculated more efficiently by the computational formula

$$t_i = e_i \left( \frac{n - p - 1}{\text{SSE} (1 - h_i) - e_i^2} \right)^{\frac{1}{2}} \quad (11.6)$$

where SSE is the error sum of squares under the full model having  $n$  cases. All terms in this expression are available from a single fitting with the  $n$  cases. Therefore, in calculating the  $n$   $t_i$ 's it is not necessary to refit the model  $n$  times corresponding to deleting each case in turn.

For our modeling of the rent data in Table 11.4, Figure 11.14 displays the Studentized (deleted) residuals for each case. We see that counties 19 and 33 both exceed the threshold 3, indicating that these counties have unusually large values of `alf.till`.

#### 11.3.4 Cook's Distance

While leverage addresses the unusualness of a case's predictor variables, and Studentized deleted residuals address (primarily) the unusualness of a case's response variable, the Cook's distance  $D_i$  of a case assesses the unusualness of both its response and predictors. The Cook's distance  $D_i$  for case  $i$  can be interpreted in two ways.

Let  $\hat{Y}$  be the  $n$ -vector of fitted values using all  $n$  cases and  $\hat{Y}_{(i)}$  be the  $n$ -vector of fitted values when case  $i$  is not used in fitting. Then

$$D_i = \frac{(\hat{Y} - \hat{Y}_{(i)})' (\hat{Y} - \hat{Y}_{(i)})}{p \text{MSE}} \quad (11.7)$$

This illustrates the interpretation that Cook's distance for case  $i$  measures the change in the vector of predicted values when case  $i$  is omitted.

Let  $\hat{\beta}_{(i)}$  be the vector of estimated regression coefficients estimated without case  $i$ . Then

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})' X' X (\hat{\beta} - \hat{\beta}_{(i)})}{p \text{MSE}} \quad (11.8)$$

This representation shows that  $D_i$  measures the change in the vector of estimated regression coefficients when case  $i$  is omitted.

As with the Studentized deleted residual, the  $n$  Cook's distances can be calculated without running  $n$  regressions omitting each case in turn. It can be shown that

$$D_i = \frac{e_i^2}{p \text{MSE}} \left( \frac{h_i}{(1 - h_i)^2} \right) \quad (11.9)$$

From this formula it is apparent that a case with a large Cook's distance has either a large residual, a large leverage, or some combination of these two.

We recommend that a case be regarded as unusual if its Cook's distance exceeds 1. This threshold for what constitutes an unusually large value of Cook's distance  $D_i$  follows the recommendation of (Weisberg, 1985) (page 120).

Since for most  $F$  distributions the 50% point is near 1, a value of  $D_i = 1$  will move the estimate to the edge of about a 50% confidence region, a potentially important change. If the largest  $D_i$  is substantially less than 1, deletion of a case will not change the estimate of  $\beta$  by much. To investigate the influence of a case more closely, the analyst should delete the large  $D_i$  case and recompute the analysis to see exactly what aspects of it have changed.

There are also arguments, for example in (Fox, 1991), for a much smaller threshold  $4/(n - p - 1)$  or  $4/n$  that decreases with increasing sample size. We are unconvinced by these arguments.

Figure 11.15 displays the Cook's distances for the rent data. Counties 5, 19, 32, 33, 49, 60, and 66 have much larger Cook's distances than the other

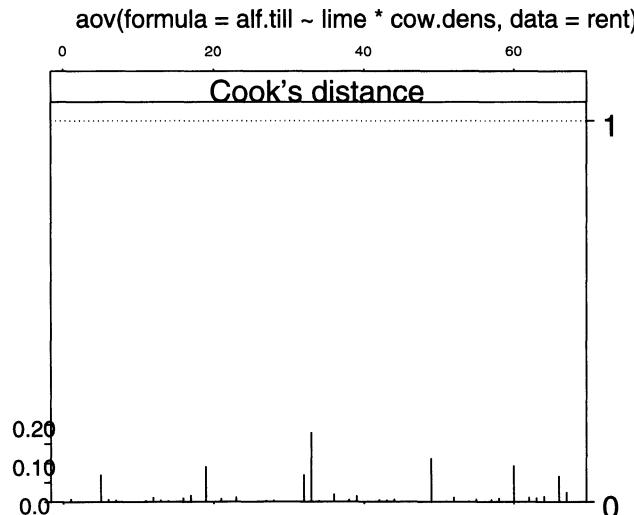


FIGURE 11.15. Cook's distances for Model (11.1) for rent data.  
(regc/code/rent4b.s), (regc/figure/cook.eps.gz)

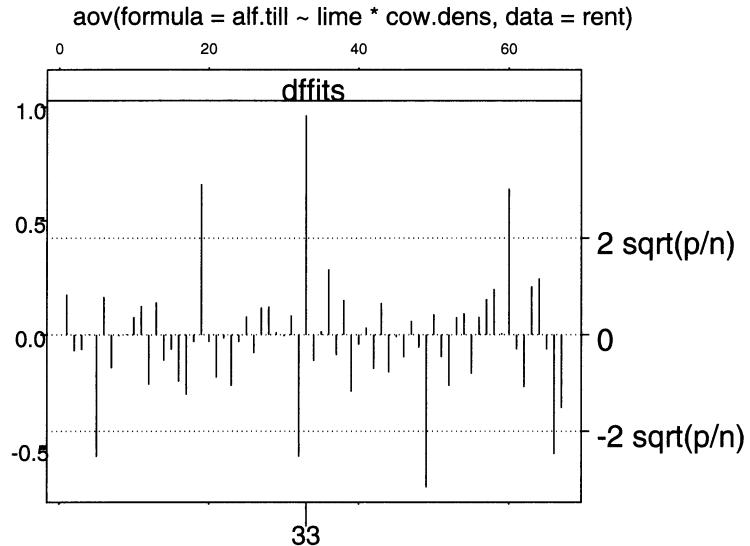


FIGURE 11.16. DFFITS for Model (11.1) for rent data.  
(regc/code/rent4b.s), (regc/figure/dffits.eps.gz)

counties, but none of these 7 counties approaches the threshold of 1 that would flag a county as unusual. Therefore, Cook's distance flags no data points fitted by `alf.till ~ lime*cow.dens`.

### 11.3.5 DFFITS

DFFITS is an abbreviation for “difference in fits”.  $\text{DFFITS}_i$  is a standardized measure of the amount by which predicted value  $\hat{Y}_i$  for case  $i$  changes when the data on this case is deleted from the data set. A flag for a case with large DFFITS is one having absolute value greater than  $2\sqrt{p/n}$ .

The interpretation of  $\text{DFFITS}_i$  is apparent from the formula

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(i)} h_i}} \quad (11.10)$$

where, as before, an  $(i)$  in a subscript means that the quantity is calculated with case  $i$  omitted from the data. As is seen from

$$\text{DFFITS}_i = \left( \frac{n-p-1}{\text{SSE}(1-h_i) - e_i^2} \right)^{\frac{1}{2}} \left( \frac{h_i}{1-h_i} \right)^{\frac{1}{2}} \quad (11.11)$$

$\text{DFFITS}_i$  can be calculated from the output of the regression using all  $n$  cases.

The list of cases shown in Figure 11.16 that exceed the  $\text{DFFITS}$  flag are the same as those exceeding the Cook's flag for the analysis in Table 11.4.

### 11.3.6 DFBETAS

$\text{DFBETAS}_{ik}$  is a standardized measure of the amount by which the  $k^{\text{th}}$  regression coefficient changes if the  $i^{\text{th}}$  observation is omitted from the data set. A case is considered to have a large such measure if its absolute  $\text{DFBETAS}$  is greater than  $2/\sqrt{n}$ . Since a regression analysis has  $np$   $\text{DFBETAS}$  in all, a request for  $\text{DFBETAS}$  in a large complicated regression analysis will generate a lot of output.

$\text{DFBETAS}_{ik}$  is defined by

$$\text{DFBETAS}_{ik} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{\text{MSE}_{(i)} c_{kk}}}$$

for  $k = 0, 1, \dots, p$ , where  $c_{kk}$  is the  $k^{\text{th}}$  diagonal entry in  $(X'X)^{-1}$ .

An efficient calculation algorithm is

1. Let  $\hat{\beta}$  be the regression coefficients from regressing  $y$  on  $x$ .
2. Let  $X$  be the matrix of predictors including the column **1**.
3. Factor  $X = QR$ . See Section F.4.7 for details.
4. Multiply the  $i^{\text{th}}$  row of  $Q$  by  $z_i = e_i/(1 - h_i)$ . Call the result  $Q_z$ .
5. Solve  $R \Delta b = Q'_z$  for  $\Delta b$ .
6. Then  $\hat{\beta}_{(i)} = \hat{\beta} - \Delta b_i$ , where  $\Delta b_i$  is the  $i^{\text{th}}$  column of  $\Delta b$ .

This algorithm is efficient because it does the hard work of solving a linear system only once, when it factors  $X = QR$  to construct the orthogonal matrix  $Q$  and the triangular matrix  $R$ . The backsolve in step 5 is not hard work because it is working with a triangular system. All the remaining steps are simple linear adjustments to the original solution. A simple executable version of this algorithm is in file (`regc/code/dfbeta.s`). A more complete version (with protection against near-singularity) is in the S-PLUS function `lm.influence`.

Figure 11.17 gives one  $\text{DFBETAS}$  plot for each predictor in the model in Table 11.4. We do not ordinarily interpret  $\text{DFBETAS}$  for the intercept term. Figure 11.6 shows that cases 5 and 19 impact the regression coefficient of `lime`, cases 33 and 49 impact the regression coefficient of `cow.dens`, and

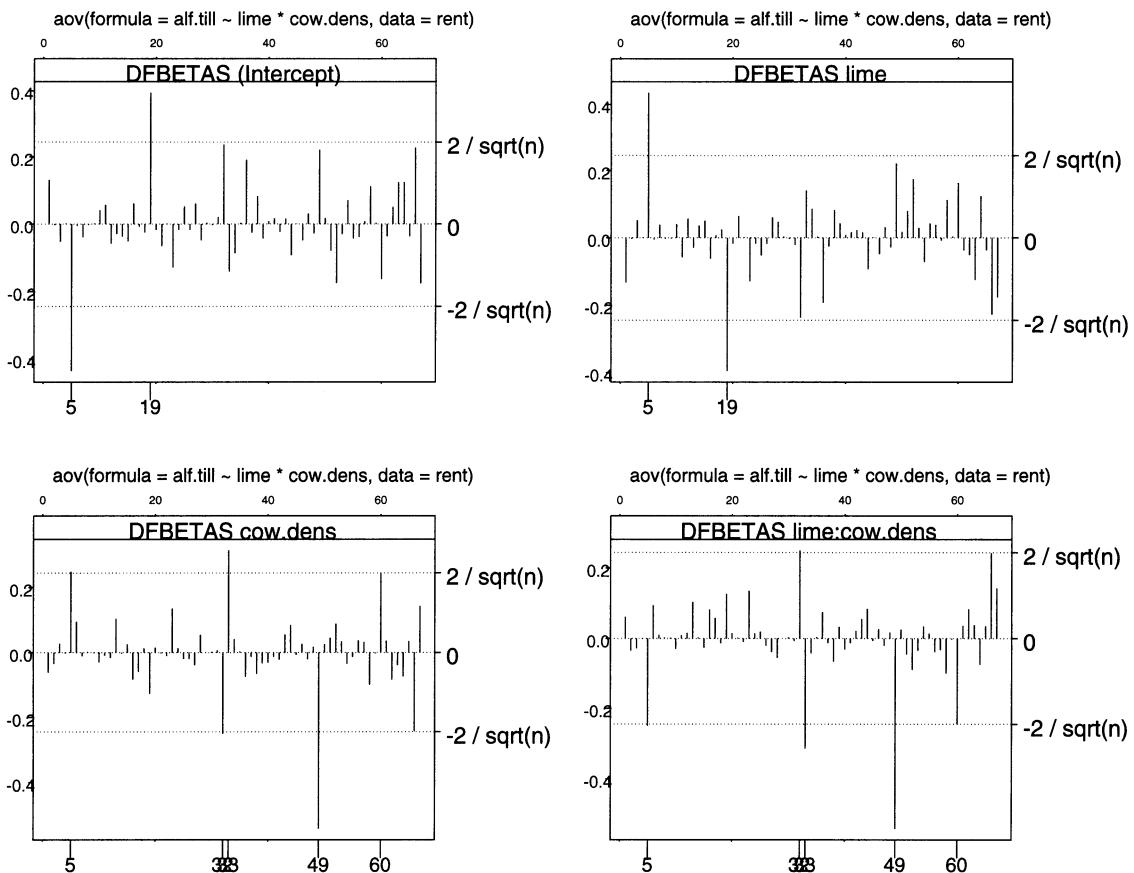


FIGURE 11.17. DFBETAS for all four predictors in Model (11.1) for the rent data: the column of 1s for the intercept, the factor lime, the covariate cow.dens, and the interaction lime:cow.dens.  
 (regc/code/rent4b.s), (regc/figure/DFBETAS..Intercept..eps.gz),  
 (regc/figure/DFBETAS.lime.eps.gz), (regc/figure/DFBETAS.cow.dens.eps.gz),  
 (regc/figure/DFBETAS.lime.cow.dens.eps.gz)

that these four counties plus county 32 are primarily responsible for the difference in slopes of the two regression lines in Figure 11.4.

### 11.3.7 Calculation of Regression Diagnostics

Regression diagnostics are calculated from the matrix formulation of the equations in Table 11.6.

In S-PLUS see the documentation for the functions `ls.diag`, `lm.influence`, and `plot.lm`. See also our functions `lm.case` and `plot.case` in files (`splus.library/lm.case.s`) and (`splus.library/plotcasediagtrellis.s`).

Regression diagnostics in SAS are computed by adding the option `INFLUENCE` to the `MODEL` statement in `PROC REG`.

## 11.4 Exercises

We recommend that for all exercises involving a data set, you begin by examining a scatterplot matrix of the variables.

- 11.1.** Data from (Brooks et al., 1988), reprinted in (Hand et al., 1994), relate the number of monthly **man-hours** associated with the anesthesiology service for 12 U.S. Naval hospitals to the number of surgical **cases**, the eligible population per thousand, and the number of operating **rooms**. The data appear in the file (`datasets/hospital.dat`).
  - a.** Construct and examine a scatterplot matrix of these data. Does it appear that multicollinearity will be a problem?
  - b.** Fit the response to all three predictors, calculating the VIFs. Based on the analysis thus far, which predictor is the best candidate for removal? Why?
  - c.** Fit the response with the predictor in part (b) removed.
  - d.** Calculate the Studentized residuals, leverages, and Cook's distances for the model in part (c). Based on these calculations, what action would you recommend?
- 11.2.** We previously encountered the dataset (`datasets/hardness.dat`) in Section 9.8 and Exercise 4.5. Since density is easily measured but hardness is not, it is desired to model hardness as a function of density.
  - a.** Construct a histogram of **hardness** and confirm that a transformation is required in order to use this chapter's regression modeling procedures.
  - b.** Regress the transformation of **hardness** you chose based on either part (a) or Exercise 4.5. For this regression, produce a scatterplot of the residuals vs the fitted values and of the residuals vs **density**. Conclude from these plots that a quadratic regression is appropriate.
  - c.** We illustrate a linear and a quadratic fit of the hardness data in Figure 9.5 and Table 9.4. Produce residual plots and regression diagnostics for both models.

- 11.3.** The dataset (`datasets/concord.dat`) is described in Exercise 4.6. Use multiple regression analysis to model `water81` as a function of a subset of the five candidate predictors. Consider transforming variables to assure that the assumption of regression analysis are well satisfied. Carefully interpret, in terms of the original model variables, all regression coefficients in your final model.
- 11.4.** Creatine clearance is an important but difficult to measure indicator of kidney function. It is desired to estimate `clearance` from more readily measured variables. (Neter et al., 1996) discuss data, originally from (Shih and Weisberg, 1986), relating `clearance` to serum clearance `concentration`, `age` and `weight`. The datafile is (`datasets/kidney.dat`).
- Regress `clearance` on each of the three individual predictors. Investigate the adequacy of this model.
  - Improve on the model in part (a) by adding to the set of candidate predictors the squares and pairwise products of the three original predictors. Conclude that the addition of one of these six new candidates improves the original model.
  - Investigate the adequacy of this model.
  - Carefully interpret each of the four estimated regression coefficients in terms of the model variables.
- 11.5.** (Heavenrich et al., 1991) provide data on the gasoline mileage (MPG) of 82 makes and models of automobiles as well as 4 potential predictors of MPG. The data appear in (`datasets/mileage.dat`). The potential predictors are
- WT: vehicle weight in 100 lbs
- HP: engine horsepower
- SP: top speed in mph
- VOL: cubic feet of cab space
- We wish to use them to model MPG.
- Produce a scatterplot matrix and comment on the plots of MPG vs HP and of HP vs SP.
  - Regress MPG on WT, HP, and SP. Are the signs of the estimated regression coefficients as expected? Explain what is causing the anomaly.

- c. First regress MPG on WT and SP and then regress MPG on WT and HP. Which of these two regressions is preferred?
  - d. For the model you prefer in part (c), produce a normal plot of the residuals and a plot of the residuals vs the fitted values. What do you conclude?
  - e. Regress the log of MPG on WT and SP and also the log of MPG on the log of WT and SP. Produce residual plots and normal probability plots from both of these runs. Based on the numerical output and plots, explain which model is preferred.
  - f. For the preferred model, produce case diagnostics. For each flagged case, indicate what is unusual about it.
- 11.6.** (Neter et al., 1996) discuss a dataset relating the amount of life insurance carried in thousands of dollars (`lifeins`) to average annual income in thousands of dollars (`anninc`) and risk aversion score (`riskaver`), for 18 managers, where higher scores connote greater risk aversion. The data are contained in the file (`datasets/lifeins.dat`).
- a. Produce a scatterplot matrix. Which of `anninc` and `riskaver` appears to be more closely related to `lifeins`?
  - b. Regress `lifeins` on `anninc` and `riskaver`, storing the residuals.
  - c. From a scatterplot of these residuals vs `anninc`, conclude that the relationship between `lifeins` and `anninc` is nonlinear. Define the square of average annual income, `annincsq` = `anninc`<sup>2</sup>. Regress `lifeins` on the three predictors `anninc`, `annincsq`, and `riskaver`. Plot the residuals from this run against `anninc`. Based on this plot, discuss whether addition of the curvature term seems worthwhile.
  - d. Identify cases (managers) whose values indicate either high influence or high leverage. Also note whether these cases have high values of any of the measures Cook's distance, DFFITS, or DFBETAS. If so, interpret such high values in terms of the model variables.
- 11.7.** Refer to (`datasets/houseprice-erie.dat`), previously considered in Exercise 9.3.
- a. Rerun the regression for the final model you found in Exercise 9.3b, this time requesting a complete set of regression diagnostics.
  - b. Closely examine the values of the diagnostics for the two high-priced houses that are the focus of Exercise 9.3c. Would you recommend both of these houses or just one of them for special scrutiny?

**11.8.** Prove the equivalence of the intuitive and computational formulas for the following case statistics:

- a. DFFITS in Equations (11.10) and (11.11)
- b. Cook's distance in either intuitive Equation (11.7) or (11.8), and computational Equation (11.9)
- c. Studentized deleted residual in Equations (11.5) and (11.6)

**11.9.** Explore the diagonals of the hat matrix  $H = X(X'X)^{-1}X'$ .

- a. Prove that all leverages satisfy  $\frac{1}{n} \leq h_i \leq 1$ . Since  $H$  is a projection matrix, show that the upper bound on the diagonals is 1. Since the column  $X_0 = 1$  is included in the  $X$  matrix, show that the lower bound on the diagonals is  $\frac{1}{n}$ .
- b. Show that the average leverage

$$\frac{\sum_i h_i}{n} \equiv (p+1)/n$$

# Two-Way Analysis of Variance

In Chapter 6 we consider situations where a response variable is measured on groups of observations classified by a single factor and look at ways to compare the changes in the mean of the response variable attributable to the various levels of this factor. Here we extend this to situations where there are two factors. In Chapters 13 and 14 we will discuss instances where there are more than two factors.

## 12.1 Example—Display Panel Data

### Study Objectives

An air traffic controller must be able to respond quickly to an emergency condition indicated on her display panel. It was desired to compare three types of display panel. Each panel was tested under four simulated emergency situations. Two well-trained controllers were assigned to each of the 12 combinations of emergency condition and display panel type; 24 controllers in all. The data in ([datasets/display.dat](#)) are from (Bowerman and O'Connell, 1990). It is clear that the type of display panel is a fixed factor, but unclear from this reference whether emergency situation is a fixed or random factor (review these concepts in Sections 6.2 and 6.4). That is, do these four situations represent the totality of incidents to which air traffic controllers might be exposed, or are they four of far more situations? In the former case, emergency situation is a fixed factor; in the latter case, emergency situation is a random factor.

## Data Description

The response variable is `time` in seconds, and the two factors are `panel` and `emergenc`.

## Data Input

We use file (`tway/code/display.sas`) as the control file. It reads the data into SAS with file (`tway/code/display1.sas`).

## Analysis Goals

We seek to determine whether the three panels afford significantly different display times and whether such conclusions are consistent across different types of emergency.

Exhibited here are graphs and tables that will aid in answering these questions. Discussion of this output is deferred until Section 12.12.

Figure 12.1 shows plots for assessing interaction between `panel` and `emergenc` as well as boxplots for examining the main effects of these factors. The concept of *interaction* is introduced in Section 12.2. Table 12.1 shows the PROC GLM statements from code file (`tway/code/display4.sas`) and the portion of the listing assuming that `emergenc` is a fixed factor. Table 12.2 contains the portion of the listing assuming that `emergenc` is a random factor. Table 12.3 shows the `panel` means and the results of the multiple comparisons by the Tukey method. The structure of the interaction plot in Figure 12.1 is discussed in Section 12.4.

Table 12.3 and Figures 12.2 and 12.3 show the `panel` means and the results of the multiple comparisons by the Tukey method. As will be explained in Section 12.12, the conclusion derived from this table is that there is a significant difference in response times for the three panels. Panel 3 affords a significantly longer response time than panels 1 or 2; response times for panels 1 and 2 do not differ significantly.

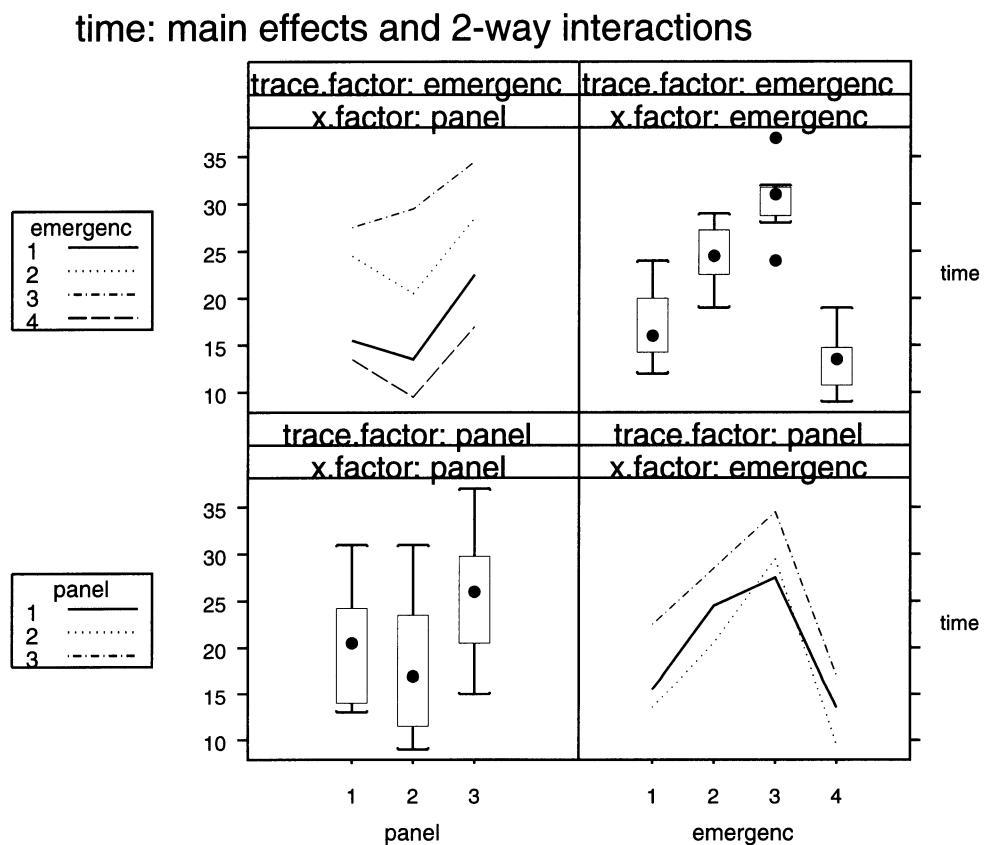


FIGURE 12.1. Interaction plot for display panel experiment. The nearly parallel traces suggest the absence of interaction between panel and emergenc.  
`(tway/code/display.s)`, `(tway/figure/display.eps.gz)`

TABLE 12.1. SAS display panel data: ANOVA table with test of **panel** appropriate if **emergenc** is fixed. The default test, in the **Type III SS** section of the listing, is from the “both factors fixed” column of Table 12.8. That is, all sums of squares are compared to the **Error** line of the ANOVA table. The listing is continued in Table 12.3.

---

SAS (tway/code/display4.sas):

```
proc glm;
  class panel emergenc;
  model time = panel | emergenc / ss3 ;
  means panel / Tukey;
  test h=panel e=panel*emergenc;
run;
```

---

SAS (tway/transcript/display2.lst):

The GLM Procedure

Dependent Variable: time

| Source          | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|-----------------|----|----------------|-------------|---------|--------|
| Model           | 11 | 1314.125000    | 119.465909  | 20.63   | <.0001 |
| Error           | 12 | 69.500000      | 5.791667    |         |        |
| Corrected Total | 23 | 1383.625000    |             |         |        |

| R-Square | Coeff Var | Root MSE | time Mean |
|----------|-----------|----------|-----------|
| 0.949770 | 11.25889  | 2.406588 | 21.37500  |

| Source         | DF | Type III SS | Mean Square | F Value | Pr > F |
|----------------|----|-------------|-------------|---------|--------|
| panel          | 2  | 232.750000  | 116.375000  | 20.09   | 0.0001 |
| emergenc       | 3  | 1052.458333 | 350.819444  | 60.57   | <.0001 |
| panel*emergenc | 6  | 28.916667   | 4.819444    | 0.83    | 0.5675 |

---

TABLE 12.2. SAS display panel data: ANOVA table with test of `panel` appropriate if `emergenc` is random. The `test` statement with the `h` (for “hypothesis”) and `e` (for “error”) arguments is the idiom SAS uses to specify which column of Table 12.8 should be used. In this example, the test is from the “A fixed, B random” column of Table 12.8 with `panel` taking the role of A. That is, the sum of squares for `panel` is compared to the `panel*emergenc` interaction line of the ANOVA table.

SAS (tway/code/display4.sas):

```
proc glm;
  class panel emergenc;
  model time = panel | emergenc / ss3 ;
  means panel / Tukey;
  test h=panel e=panel*emergenc;
run;
```

SAS (tway/transcript/display2b.lst):

Dependent Variable: time

Tests of Hypotheses Using the Type III  
MS for `panel*emergenc` as an Error Term

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|--------|----|-------------|-------------|---------|--------|
| panel  | 2  | 232.7500000 | 116.3750000 | 24.15   | 0.0013 |

TABLE 12.3. SAS Display panel data: ANOVA table with test of panel appropriate if emergenc is fixed. This is the output from just the means statement. Multiple comparisons of panel by Tukey method. The standard deviation for the comparison is based on the Error line of the ANOVA table.

SAS (tway/code/display4.sas):

```
proc glm;
  class panel emergenc;
  model time = panel | emergenc / ss3 ;
  means panel / Tukey;
  test h=panel e=panel*emergenc;
run;
```

SAS (tway/transcript/display2a.lst):  
Tukey's Studentized Range (HSD) Test for time

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

|                                     |          |
|-------------------------------------|----------|
| Alpha                               | 0.05     |
| Error Degrees of Freedom            | 12       |
| Error Mean Square                   | 5.791667 |
| Critical Value of Studentized Range | 3.77278  |
| Minimum Significant Difference      | 3.2101   |

Means with the same letter are not significantly different.

T u k e y  
G r o u p i n g

|   | Mean   | N | panel |
|---|--------|---|-------|
| A | 25.625 | 8 | 3     |
| B | 20.250 | 8 | 1     |
| B |        |   |       |
| B | 18.250 | 8 | 2     |

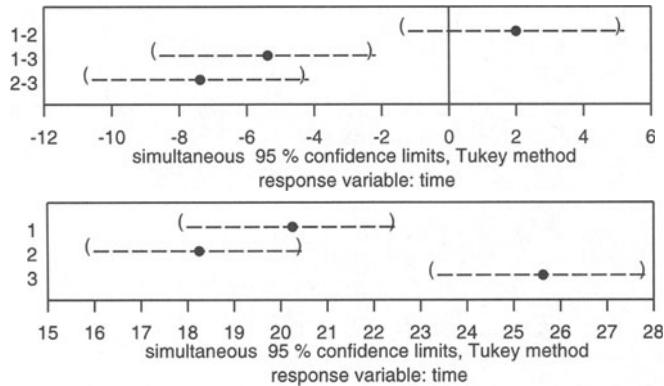


FIGURE 12.2. Tukey multiple comparisons plots by S-PLUS.  
`(tway/code/display.s)`, `(tway/figure/display-mcdiff.eps.gz)`,  
`(tway/figure/display-mcmean.eps.gz)`

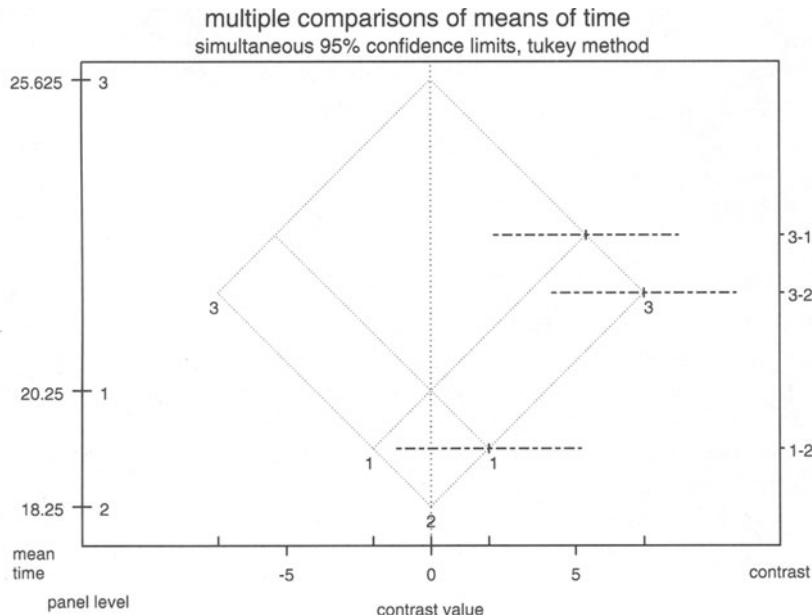


FIGURE 12.3. MMC plot of Tukey multiple comparisons plots by S-PLUS.  
`(tway/code/display.s)`, `(tway/figure/display-mmcdiff.eps.gz)`

## 12.2 Statistical Model

To model an experiment with two factors, we begin by calling the factors A and B, where A has  $a$  levels and B has  $b$  levels. We use  $n_{ij}$  to denote the number of observations taken from cell  $(i, j)$ , i.e., the treatment combination corresponding to level  $i$  of A and level  $j$  of B,  $i = 1, \dots, a$  and  $j = 1, \dots, b$ . Our discussion in this chapter is confined to the case where the  $n_{ij}$  are equal for all  $i, j$ , and sometimes  $n_{ij} = 1$ . We extend the notation of Equation (6.1) by replacing the singly-indexed symbol  $\alpha_i$  with a doubly-indexed set of symbols  $\alpha_i + \beta_j + (\alpha\beta)_{ij}$  and model the  $k^{\text{th}}$  observation at the  $i^{\text{th}}$  level of A,  $j^{\text{th}}$  level of B, as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} = \mu_{ij} + \epsilon_{ijk} \quad (12.1)$$

for  $1 \leq i \leq a$ ,  $1 \leq j \leq b$ , and  $1 \leq k \leq n_{ij}$ . The expectations for the cell means are denoted

$$E(Y_{ijk}) = \mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \quad (12.2)$$

We assume the errors  $\epsilon_{ijk} \sim \text{NID}(0, \sigma^2)$ , that is they are assumed to be normally independently distributed with a common variance  $\sigma^2$ . The parameter  $\mu$  represents the grand mean of all  $ab$  populations.

Each of the factors A and B can be either fixed or random. If A is fixed, then we assume that  $\sum_i \alpha_i = 0$ . If A is random, we assume that each  $\alpha_i \sim N(0, \sigma_A^2)$ . Similarly, if B is fixed, then we assume that  $\sum_j \beta_j = 0$  and if B is random, we assume that each  $\beta_j \sim N(0, \sigma_B^2)$ .

The term  $(\alpha\beta)_{ij}$  models the possibility of interaction between the two factors. If A and B are both fixed factors, then the sum of  $(\alpha\beta)_{ij}$  over either  $i$  or  $j$  is zero. If both factors are random, then  $(\alpha\beta)_{ij} \sim N(0, \sigma_{AB}^2)$ . In the case of a mixed model, where for concreteness we have A fixed and B random,  $(\alpha\beta)_{ij} \sim N(0, \frac{a-1}{a} \sigma_{AB}^2)$  subject to  $\sum_i (\alpha\beta)_{ij} = 0$  for each  $j = 1, \dots, b$ .

Factors A and B are said to *interact* if the difference in response between two levels of A differs according to the level of B. Equivalently, there is *interaction* between factors A and B if the difference in response between two levels of B differs according to the level of A. Graphically, the traces for each level of factor A across levels of B are parallel if there is no interaction and are not parallel when there is interaction. Equivalently, the traces for each level of B across levels of A are parallel if there is no interaction.

## 12.3 Main Effects and Interactions

As in one-way ANOVA, we are interested in comparing the means of observations in each *cell*, that is for each *treatment combination* (combination

of factor levels), in the design, and for combinations of cells. We work with the *cell means*

$$\bar{Y}_{ij} = \sum_k Y_{ijk} / n_{ij} \quad (12.3)$$

and the *marginal means*. The marginal means for the rows are calculated by averaging the cell means in each row over the columns and averaging the cell means in each column over the rows:

$$\bar{Y}_{i\cdot} = \sum_j \bar{Y}_{ij} / b \quad (12.4)$$

$$\bar{Y}_{\cdot j} = \sum_i \bar{Y}_{ij} / a \quad (12.5)$$

where  $n_{i\cdot} = \sum_j n_{ij}$ ,  $n_{\cdot j} = \sum_i n_{ij}$ , and  $n_{\cdot\cdot} = \sum_{ij} n_{ij}$ . Marginal means get their name because they are often displayed on the margins of a two-way table of cell means, as in Table 12.4. We also use the *grand mean*:

$$\bar{Y}_{\cdot\cdot} = \sum_i n_{i\cdot} \bar{Y}_{i\cdot} / n_{\cdot\cdot} = \sum_j n_{\cdot j} \bar{Y}_{\cdot j} / n_{\cdot\cdot} = \sum_{ijk} Y_{ijk} / n_{\cdot\cdot} \quad (12.6)$$

When more than one factor is present, there are three principal types of comparisons that we will investigate.

**Main effects** are comparisons of the marginal means for one of the factors, for example,  $\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}$ . It is usually valid to compare main effects only when there is no interaction.

TABLE 12.4. Table of means for the rhizobium clover experiment of Section 12.13. Means from Table 12.14 have been arranged in a two-way table to display the cell means in the body of the table, the marginal means on the margins of the table, and the grand mean as the margin of the marginal means. `clover` and `clover+alfalfa` are the two levels of the factor `comb`.

(`tway/code/rhiz-read.s`), (`tway/code/rhiz-clov-aov.s`)

| strain      | clover              | clover+alfalfa      | mean                   | strain      | clover | clover+alfalfa | mean  |       |
|-------------|---------------------|---------------------|------------------------|-------------|--------|----------------|-------|-------|
| 3DOk1       | $\bar{Y}_{11}$      | $\bar{Y}_{12}$      | $\bar{Y}_{1\cdot}$     | 3DOk1       | 29.04  | 28.41          | 28.72 |       |
| 3DOk5       | $\bar{Y}_{21}$      | $\bar{Y}_{22}$      | $\bar{Y}_{2\cdot}$     | 3DOk5       | 36.29  | 27.44          | 31.86 |       |
| 3DOk4       | $\bar{Y}_{31}$      | $\bar{Y}_{32}$      | $\bar{Y}_{3\cdot}$     | =           | 3DOk4  | 21.35          | 23.98 | 22.66 |
| 3DOk7       | $\bar{Y}_{41}$      | $\bar{Y}_{42}$      | $\bar{Y}_{4\cdot}$     | 3DOk7       | 22.93  | 24.96          | 23.95 |       |
| 3DOk13      | $\bar{Y}_{51}$      | $\bar{Y}_{52}$      | $\bar{Y}_{5\cdot}$     | 3DOk13      | 22.49  | 24.30          | 23.39 |       |
| k.composite | $\bar{Y}_{61}$      | $\bar{Y}_{62}$      | $\bar{Y}_{6\cdot}$     | k.composite | 25.97  | 24.92          | 25.45 |       |
| mean        | $\bar{Y}_{\cdot 1}$ | $\bar{Y}_{\cdot 2}$ | $\bar{Y}_{\cdot\cdot}$ | mean        | 26.35  | 25.67          | 26.01 |       |

**Interactions** (or interaction effects) are comparisons of the cell means across levels of both factors, for example,  $(\bar{Y}_{13} - \bar{Y}_{23}) - (\bar{Y}_{14} - \bar{Y}_{24})$ . When interaction is present, that is when differences in the cell means across rows depend on the column or equivalently, when comparisons of the form indicated here are significantly different from 0, we usually must use simple effects, not main effects, to discuss the factors.

**Simple effects** are separate comparisons of the cell means across levels of one factor for some or all levels of the other factor, for example,  $\bar{Y}_{13} - \bar{Y}_{23}$ . See Section 13.3.

The analyst should be alert to the possibility that interaction is present. The nature of the analysis when interaction exists is different from that when interaction is absent.

Without interaction, the analysis proceeds similarly to the procedures for one-way analysis. The marginal means are calculated and compared, perhaps by using one of the multiple comparisons techniques discussed in Sections 6.3, 7.1.3, or 7.1.4.1. The advantage of the two-way analysis in this case is in the efficiency, hence increased power, of the comparisons. Because we use the same residual sum of squares for the denominator of both  $F$ -tests (for the rows and for the columns), we can run the combined experiment to test the effect of both factors for less expense than if we were to run two separate experiments.

When interaction between two factors is present, it is not appropriate to compare the main effects, the levels of one of these factors averaged over the levels of the other factor. It is possible, for example, that the mean of  $Y$  increases over factor B for level 1 of factor A and decreases over factor B for level 2 of factor A. Averaging over the levels of factor A would mask that behavior of the response.

We explore main effects, interactions, and simple effects with the rhizobium data in Section 12.13.

## 12.4 Two-Way Interaction Plot

The two-way interaction plot, first shown in Figure 12.1 and used throughout the remainder of this book, shows all main effects and two-way interactions for designs with two or more factors. We construct it by analogy with the splom (scatterplot matrix). The rows and columns of the two-way interaction plot are defined by the Cartesian product of the factors.

1. Each main diagonal panel shows a boxplot for the marginal effect of a factor.

2. Each off-diagonal panel is a standard interaction plot of the factors defining its position in the array. Each point in the panel is the mean of the response variable conditional on the values of the two factors. Each line in the panel connects the cell means for a constant level of the *trace* factor. Each vertically aligned set of points in the panel shows the cell means for a constant value of the *x*-factor.
3. Panels in mirror-image positions interchange the trace- and *x*-factors. This duplication is helpful rather than redundant because one of the orientations is frequently much easier to interpret than the other.
4. The rows are labeled with a key that shows the line type and color for the trace factor by which the row is defined.
5. Each box in the boxplot panels has the same color, and optionally the same line type, as the corresponding traces in its row.
6. The columns are labeled by the *x*-factor.

## 12.5 Sums of Squares in the Two-Way ANOVA Table

Table 12.5 presents the structure of the analysis of variance table for a balanced two-way ANOVA with  $a$  levels of the A factor,  $b$  levels of the B factor, and  $n$  observations at each of the  $ab$  AB-treatment combinations, analogous to Table 6.3 for one-way ANOVA. If the test  $F_{AB}$  shows that AB interaction is present, the  $F$ -tests on A and B are not interpretable.

If the AB interaction is not significant, then the form of the tests for the main effects A and B depends on whether the factors A and B are fixed or random factors. See the discussion in Section 12.10 where Table 12.8 lists the expected mean squares and  $F$ -tests under various assumptions.

Table 12.5 shows the  $F$ -statistics and their  $p$ -values for tests on the main effects A and B under the assumption that both factors represent fixed effects. Most ANOVA programs calculate these values by default whether or not they are appropriate.

## 12.6 Treatment and Blocking Factors

Treatment factors are those for which we wish to determine if there is an effect. Blocking factors are those for which we believe there is an effect. We wish to prevent a presumed blocking effect from interfering with our measurement of the treatment effect.

An experiment with two factors may have either two *treatment* factors or one treatment factor and one *blocking factor*. The primary objective of

TABLE 12.5. Two-way ANOVA structure with both factors representing fixed effects.

| Analysis of Variance of Dependent Variable $y$ |                    |                |             |          |          |
|--|--------------------|----------------|-------------|----------|----------|
| Source   | Degrees of Freedom | Sum of Squares | Mean Square | F        | p-value  |
| Treatment A                                    | $df_A$             | $SS_A$         | $MS_A$      | $F_A$    | $p_A$    |
| Treatment B                                    | $df_B$             | $SS_B$         | $MS_B$      | $F_B$    | $p_B$    |
| AB Interaction                                 | $df_{AB}$          | $SS_{AB}$      | $MS_{AB}$   | $F_{AB}$ | $p_{AB}$ |
| Residual                                       | $df_{Res}$         | $SS_{Res}$     | $MS_{Res}$  |          |          |
| Total  | $df_{Total}$       | $SS_{Total}$   |             |          |          |

Terms of the table are defined by:

| Treatment A |   | Treatment AB |  |
|-------------|---|--------------|--|
| $df_A$      | $a - 1$                                       | $df_{AB}$    | $(a - 1)(b - 1)$                                       |
| $SS_A$      | $bn \sum (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2$ | $SS_{AB}$    | $n \sum (\bar{Y}_{ij} - \bar{Y}_{..})^2 - SS_A - SS_B$ |
| $MS_A$      | $SS_A / df_A$                                 | $MS_{AB}$    | $SS_{AB} / df_{AB}$                                    |
| $F_A$       | $MS_A / MS_{Res}$                             | $F_{AB}$     | $MS_{AB} / MS_{Res}$                                   |
| $p_A$       | $1 - \mathcal{F}_F(F_A   df_A, df_{Res})$     | $p_{AB}$     | $1 - \mathcal{F}_F(F_{AB}   df_{AB}, df_{Res})$        |

| Treatment B |  | Residual   |  |
|-------------|--|------------|--|
| $df_B$      | $b - 1$  | $df_{Res}$ | $ab(n - 1)$                                |
| $SS_B$      | $an \sum (\bar{Y}_{\cdot j} - \bar{Y}_{..})^2$ | $SS_{Res}$ | $\sum_i \sum_j (Y_{ijk} - \bar{Y}_{ij})^2$ |
| $MS_B$      | $SS_B / df_B$                                  | $MS_{Res}$ | $SS_{Res} / df_{Res}$                      |
| $F_B$       | $MS_B / MS_{Res}$                              |            |  |
| $p_B$       | $1 - \mathcal{F}_F(F_B   df_B, df_{Res})$      |            |  |

| Total        |           |   |
|--------------|-----------|---|
| $df_{Total}$ | $abn - 1$ |   |
| $SS_{Total}$ |           | $\sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{..})^2$ |

a factorial experiment is comparisons of the levels of treatment factors. By contrast, a blocking factor is set up in order to enhance one's ability to distinguish between the levels of treatment factors. The term *block* was chosen by analogy to two of the dictionary definitions: a rectangular section of land bounded on each side by consecutive streets; or a set of similar items sold or handled as a unit, such as shares of stock.

We are not interested in comparing the *blocks*, i.e., the levels of a blocking factor. In a well-designed experiment, we anticipate that the response differs across the levels of a blocking factor because if the levels of this factor cover a variety of experimental conditions, this broadens the scope of our inferences about treatment differences. Multiple comparisons across blocks are not meaningful because we know in advance that the blocks are different. In general, blocking is advisable and successful as an experimental and analytical technique if the experimental units can reasonably be grouped into blocks such that the units within every block are homogeneous, while the units in any given block are different from those in any other block. By homogeneous units, we mean that they will tend to respond alike if treated alike. Usually, there is no interaction between blocking and treatment factors; otherwise blocking will not have accomplished its objective and the analysis will be much less able to detect significant differences than if blocking were properly done.

Blocking is the natural extension to three or more treatments of the matched pairs design introduced in Section 5.5. The *F*-test of the treatment effect against the residual is the generalization of the paired *t*-test. It is exactly true that a blocked design with two levels of the treatment factor and with many blocks of size two is identical to the matched pairs design.

For example, in an experiment on tire wear, the location of the tire on the car (say, Right Front) is a treatment effect and the specific car (of the many used in the experiment) is a blocking effect.

## 12.7 Fixed and Random Effects

As mentioned in Sections 6.2 and 6.4, treatment factors may be regarded as either fixed or random. The levels of a fixed factor are the only levels of interest in the experiment, and we wish to see if the response is homogeneous across these levels. The levels of a random factor are a random sample from some large population of levels, and we are interested in assessing whether the variance of responses over this population of levels is essentially zero. Block factors are almost always regarded as random.

The levels of a treatment factor can be either categorical or quantitative. For example, in an experiment where the **fertilizer** treatment has four

levels, the experimental levels of fertilizer could be four different fertilizer compounds, or four different applications per acre of one fertilizer compound. When the levels are quantitative, it is usually preferable to regard the factor as a single degree-of-freedom predictor variable.

## 12.8 Randomized Complete Block Designs

A randomized complete block design (RCBD) has one treatment factor involving  $t$  treatment levels and one blocking factor having  $b$  levels. The  $b$  blocks each contain experimental units arranged according to the principles discussed in Section 12.6. That is, experimental units in the same block are expected to respond alike if treated alike, while the blocks should reflect a variety of experimental conditions to broaden the scope of conclusions to be drawn from inferences about the treatments. It is assumed that blocks and treatments do not interact. This assumption permits us to compare the treatment levels when each block contains exactly  $t$  experimental units, i.e., there is no replication of treatments within any block. If there are  $n > 1$  observations on each treatment within each block, then additional degrees of freedom are available for comparing treatments.

The model for the RCBD with one observation on each treatment in each block is

$$y_{ij} = \mu + \tau_i + \rho_j + \epsilon_{ij} \quad (12.7)$$

where  $\mu$  represents the overall mean,  $\tau_i$  is the differential effect of treatment level  $i$ ,  $\rho_j$  is the differential effect of block  $j$ , and the  $\epsilon$ 's are random  $N(0, \sigma^2)$  residuals. We further define

$$\bar{y}_i = \sum_i y_{ij}/b, \quad \bar{y}_j = \sum_j y_{ij}/t, \quad \text{and} \quad \bar{y}_{..} = \sum_i \sum_j y_{ij}/bt \quad (12.8)$$

The setup of the ANOVA table for an RCBD with  $n = 1$  is shown in Table 12.6. Some ANOVA programs also display an  $F$ -statistic and  $p$ -value for blocks, but it is inappropriate to interpret these since the experiment is designed in such a way that responses differ across blocks.

TABLE 12.6. ANOVA table structure for a randomized complete block design with no replication.

| Analysis of Variance of Dependent Variable $y$ |                    |                |             |          |          |
|--|--------------------|----------------|-------------|----------|----------|
| Source   | Degrees of Freedom | Sum of Squares | Mean Square | F        | p-value  |
| Blocks   | $df_{Blk}$         | $SS_{Blk}$     | $MS_{Blk}$  |          |          |
| Treatments                                     | $df_{Tr}$          | $SS_{Tr}$      | $MS_{Tr}$   | $F_{Tr}$ | $p_{Tr}$ |
| Residuals                                      | $df_{Res}$         | $SS_{Res}$     | $MS_{Res}$  |          |          |
| Total  | $df_{Total}$       | $SS_{Total}$   |             |          |          |

The terms of the table are defined by:

| Blocks     |   | Residual     |   |
|------------|---|--------------|---|
| $df_{Blk}$ | $b - 1$   | $df_{Res}$   | $(b - 1)(t - 1)$  |
| $SS_{Blk}$ | $\sum_i \sum_j (\bar{y}_j - \bar{y}_{..})^2$    | $SS_{Res}$   | $\sum_i \sum_j (y_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}_{..})^2$ |
| $MS_{Blk}$ | $SS_{Blk}/df_{Blk}$                             | $MS_{Res}$   | $SS_{Res}/df_{Res}$   |
| Treatments |   | Total        |   |
| $df_{Tr}$  | $t - 1$   | $df_{Total}$ | $bt - 1$  |
| $SS_{Tr}$  | $\sum_i \sum_j (y_i - \bar{y}_{..})^2$          | $SS_{Total}$ | $\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2$                         |
| $MS_{Tr}$  | $SS_{Tr}/df_{Tr}$                               |              |   |
| $F_{Tr}$   | $MS_{Tr}/MS_{Res}$                              |              |   |
| $p_{Tr}$   | $1 - \mathcal{F}_F(F_{Tr}   df_{Tr}, df_{Res})$ |              |   |

## 12.9 Example—The Blood Plasma Data

### Study Objectives

The dataset (`datasets/plasma.dat`) comes from (Anderson et al., 1981) and is reproduced in (Hand et al., 1994). The data are measurements on plasma citrate concentrations in micromols/liter obtained from 10 subjects at 8am, 11am, 2pm, 5pm, and 8pm. To what extent is there a normal profile for the level in the human body during the day?

This experiment is viewed as an RCBD with treatment factor `time` and blocking factor `subject`. It is desirable here that the subjects (blocks) be as unlike as possible in order to broaden the scope of the conclusion about normal profiles as much as possible. The no-interaction assumption amounts to assuming that the daily response profile is constant across subjects.

### Data Description

We restructure the data in (`datasets/plasma.dat`) into 50 rows with three variables.

**plasma:** the response variable, plasma citrate concentrations in micro-mols/liter

**time:** factor with five values: 8am, 11am, 2pm, 5pm, and 8pm

**id:** factor with 10 levels, one per subject

### Analysis

We begin our analysis in (`tway/code/plasma.s`) with the interaction plots in Figure 12.4. There seem to be anomalies for id=3 at 8pm and for id=6 at 11am, but otherwise both sets of traces look reasonably parallel.

We proceed with an additive model in Table 12.7 and discover that the `id` has a high  $F$ -value, confirming our decision to block on patients. This is not a hypothesis test, because we know at the beginning of our analysis that patients are different from each other.

The test of differences due to `time` rejects the null hypothesis that the response at all times is the same. Since there appears to be no interaction, we can act as if there is a single pattern that applies to everyone.

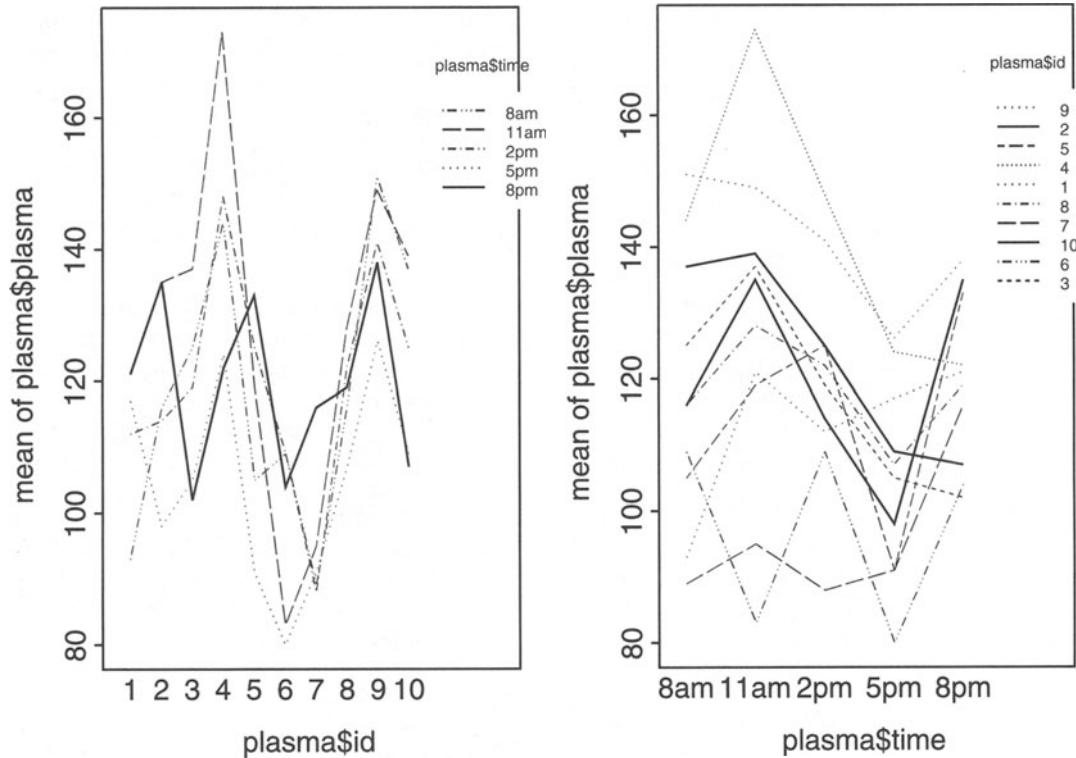


FIGURE 12.4. Interaction Plot for Plasma Citrate  
(tway/code/plasma.s), (tway/figure/plasmaint.eps.gz)

TABLE 12.7. ANOVA Table for Plasma Citrate Experiment  
(tway/code/plasma.s)

S-PLUS (tway/transcript/plasma.st):

```
> plasma.aov <- aov(plasma ~ id + time, data=plasma)
> summary(plasma.aov)
```

|           | Df | Sum of Sq | Mean Sq  | F Value  | Pr(F)       |
|-----------|----|-----------|----------|----------|-------------|
| id        | 9  | 10592.72  | 1176.969 | 7.981735 | 0.000002249 |
| time      | 4  | 2803.92   | 700.980  | 4.753768 | 0.003494595 |
| Residuals | 36 | 5308.48   | 147.458  |          |             |

## 12.10 Random Effects Models and Mixed Models

In Section 6.4, we compare two analyses of the same data assuming the single factor is fixed or random. There we indicate that a table of expected mean squares may be used to formulate the correct mean square ratio to test the hypothesis of interest. We also show that in the single factor case, while the same ratio is used in both the fixed and random cases, the hypothesis tested about the factor differs in the two cases.

When we have two or more factors and interactions, the test statistics as well as the hypotheses depend on whether the factors are fixed or random. The formulas for standard errors for comparing the levels of fixed factors also depend on whether the other factor(s) are fixed or random.

Table 12.8 is an algebraically derived table of expected mean squares for an experiment with two possibly interacting factors A and B and equal sample sizes  $n_{ij} = n \geq 2$  at each of the  $ab$  treatment combinations under each of three assumptions: the fixed model where both factors are fixed, the mixed model where one factor is fixed and the other factor is random, and the random model where both factors are random.

From the lineups of the expected mean squares, we see that for testing the A main effect, the appropriate denominator mean square is the error mean square if factor B is fixed, but the AB-interaction mean square if B is random. The conclusions for testing the B main effect follow from interchanging “A” and “B” in the previous sentence.

TABLE 12.8. Expected mean squares in two-way analysis of variance. Compare to Tables 6.6, 12.5, and 13.11. See Section 12.10 for the discussion on when to use each of the columns.

| Source         | df               | Both factors fixed   | A fixed, B random  | Both factors random                        |
|----------------|------------------|--|--|--|
| Treatment A    | $a - 1$          | $\sigma^2 + nb \frac{\sum_i \alpha_i^2}{a - 1}$                          | $\sigma^2 + n\sigma_{AB}^2 + nb \frac{\sum_i \alpha_i^2}{a - 1}$ | $\sigma^2 + n\sigma_{AB}^2 + nb\sigma_A^2$ |
| Treatment B    | $b - 1$          | $\sigma^2 + na \frac{\sum_j \beta_j^2}{b - 1}$                           | $\sigma^2 + nao_B^2$   | $\sigma^2 + n\sigma_{AB}^2 + nao_B^2$      |
| AB Interaction | $(a - 1)(b - 1)$ | $\sigma^2 + n \frac{\sum_i \sum_j (\alpha\beta)_{ij}^2}{(a - 1)(b - 1)}$ | $\sigma^2 + n\sigma_{AB}^2$                                      | $\sigma^2 + n\sigma_{AB}^2$                |
| Residual       | $ab(n - 1)$      | $\sigma^2$   | $\sigma^2$   | $\sigma^2$                                 |
| Total          | $abn - 1$        |  |  |  |

## 12.11 Introduction to Nesting

In the previous examples the two factors have a *crossed* relationship. Saying that factors  $A$  and  $B$  are crossed indicates that each level of  $A$  may be observed in a treatment combination with any level of  $B$ . Alternatively, two factors may have a *nested* or hierarchical relationship. When  $B$  is nested within  $A$ , the levels of  $B$  are similar but not identical for different levels of  $A$ .

### 12.11.1 Example—Workstation Data

A small electronics firm wishes to compare three methods for assembling an electronic device. For this purpose, the plant has available six different workstations. The study is conducted by randomly assigning  $s = 2$  workstations to each of the  $m = 3$  assembly methods, and at each workstation  $w = 5$  randomly selected production workers will assemble the device for one hour using the appropriate assembly method. The response is the number of devices produced in one hour. The data from (Bowerman and O'Connell, 1990) (p. 890) are in file (`datasets/workstation.dat`) and are displayed in Figure 12.5.

Note that the workstations assigned to any assembly method are different from those assigned to any other method. As a consequence, the factors (which we'll call `station` and `method`) are not crossed with one another, and an analysis using a model we've previously studied would be incorrect. The factor `station` is said to be *nested* within the factor `method` because each workstation is associated with exactly one of the methods.

Our analysis assumes that `station` is a fixed factor. If instead `station` were assumed to be a random factor, the S-PLUS or SAS code would have to be modified to force `station` to be tested against the `station` within `method` mean square instead of against the `Residual` mean square. The procedures for doing so are demonstrated in the data analysis in Section 13.4.

The basic structure of the ANOVA table is in Table 12.9. In SAS the `Model` statement indicating that `station` is nested within `method` is

```
model devices = method station(method);
```

In S-PLUS, we use the formula

```
devices ~ method / station
```

The S-PLUS analysis is in Table 12.10 and the SAS analysis is in Tables 12.11 and 12.12.

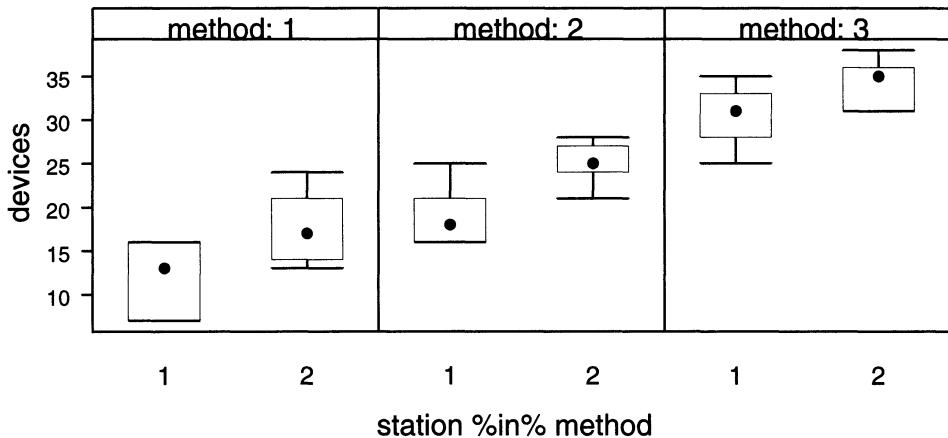


FIGURE 12.5. Boxplot of workstation data. The significance of `method` and `station` within `method` are confirmed in Tables 12.10 and 12.11.  
`(tway/code/workstation.s)`, `(tway/figure/workstation.eps.gz)`

TABLE 12.9. Basic structure of the ANOVA table for a nested design with  $m = 3$ ,  $s = 2$ , and  $w = 5$ .

| Source                           | df          |                                    | MS     | F                   |
|----------------------------------|-------------|------------------------------------|--------|---------------------|
|                                  | Algebra     | Example                            |        |                     |
| Method                           | $m - 1$     | $= 3 - 1 = 2$                      | $MS_m$ | $\frac{MS_m}{MS_w}$ |
| Station within Method            | $m(s - 1)$  | $= 3 \times (2 - 1) = 3$           | $MS_s$ | $\frac{MS_s}{MS_w}$ |
| Worker within Station (Residual) | $ms(w - 1)$ | $= 3 \times 2 \times (5 - 1) = 24$ | $MS_w$ |                     |
| Total                            | $msw - 1$   | $= 3 \times 2 \times 5 - 1 = 29$   |        |                     |

We conclude that when using at least one of the three methods, the two workstations for that method produced a significantly different number of devices. We also conclude that the three methods produced significantly different numbers of devices.

In this example there is balanced sampling. That is, each method has the same number of workstations and each workstation has the same number of workers. Without much additional difficulty, the above nested factorial analysis can be extended to situations with unbalanced sampling. (In con-

TABLE 12.10. Workstation data. S-PLUS ANOVA table and means.  
(tway/code/workstation.s)

---

```
S-PLUS (tway/transcript/workstation.st):
> workstation.aov <- aov(devices ~ method / station,
+                               data=workstation, qr=T)
> summary(workstation.aov)
      Df Sum of Sq  Mean Sq   F Value    Pr(F)
method     2 1545.267 772.6333 51.45172 0.00000000
station %in% method  3   210.200  70.0667  4.66593 0.01047493
Residuals 24   360.400  15.0167
> model.tables(workstation.aov, se=T)

Tables of effects

method
      1      2      3
-8.2667 -0.9667  9.2333

station %in% method
Dim 1 : method
Dim 2 : station
      1      2
1 -3.0  3.0
2 -2.9  2.9
3 -1.9  1.9

Standard errors of effects
method station %in% method
      1.2254          1.733
replic. 10.0000        5.000
```

---

trast, when one has unbalanced sampling and crossed factors, the analysis is considerably more difficult than with balanced sampling.)

## 12.12 Example—Display Panel Data—Continued

In Section 12.1 we introduced the display panel example illustrating a two-way analysis of variance. We continue here with the analysis by discussing Figure 12.1 and Tables 12.1 through 12.3.

In Figure 12.1 we display two-way interaction plots and boxplots for the factors `panel` and `emergenc`. The two interaction plots contain equivalent

TABLE 12.11. Workstation data. SAS ANOVA table.

```
SAS (tway/code/workstation.sas):
title 'Analysis of Workstation Assembly Data';
data one;
  infile "&hh/datasets/workstation.dat" firstobs=2;
  input method station devices;

proc anova;
  class method station;
  model devices =
    method station(method);
  means method station(method);
run;
```

SAS (tway/transcript/workstation.lst):

Dependent Variable: DEVICES

| Source          | DF | Sum of Squares | Mean Square | F Value      | Pr > F |
|-----------------|----|----------------|-------------|--------------|--------|
| Model           | 5  | 1755.4666667   | 351.0933333 | 23.38        | 0.0001 |
| Error           | 24 | 360.4000000    | 15.01666667 |              |        |
| Corrected Total | 29 | 2115.8666667   |             |              |        |
| R-Square        |    | C.V.           | Root MSE    | DEVICES Mean |        |
| 0.829668        |    | 16.79972       | 3.8751344   | 23.066667    |        |

| Source          | DF | Anova SS     | Mean Square | F Value | Pr > F |
|-----------------|----|--------------|-------------|---------|--------|
| METHOD          | 2  | 1545.2666667 | 772.6333333 | 51.45   | 0.0001 |
| STATION(METHOD) | 3  | 210.2000000  | 70.0666667  | 4.67    | 0.0105 |

information, but in general, one of them is more readily interpretable than the other. In this instance, the close-to-parallel traces suggest the absence of interaction between `panel` and `emergenc`. This is confirmed by the large *p*-value for the interaction test in Table 12.1. One set of boxplots in Figure 12.1 evinces a greater response time with panel 3 than with either panel 1 or panel 2. The other set of boxplots shows substantial differences in the

TABLE 12.12. Workstation data. SAS means.  
(tway/code/workstation.sas)

---

```
means method station(method);
```

---

| SAS (tway/transcript/workstationb.lst):<br>-----devices----- |  |    |            |            |
|--|--|----|------------|------------|
| Level of<br>method   |  | N  | Mean       | Std Dev    |
| 1  |  | 10 | 14.8000000 | 5.37070242 |
| 2  |  | 10 | 22.1000000 | 4.38304815 |
| 3  |  | 10 | 32.3000000 | 3.91719855 |

| Level of<br>station | Level of<br>method | -----devices----- |            |            |
|---------------------|--------------------|-------------------|------------|------------|
|                     |                    | N                 | Mean       | Std Dev    |
| 1                   | 1                  | 5                 | 11.8000000 | 4.54972527 |
| 2                   | 1                  | 5                 | 17.8000000 | 4.65832588 |
| 1                   | 2                  | 5                 | 19.2000000 | 3.83405790 |
| 2                   | 2                  | 5                 | 25.0000000 | 2.73861279 |
| 1                   | 3                  | 5                 | 30.4000000 | 3.97492138 |
| 2                   | 3                  | 5                 | 34.2000000 | 3.11448230 |

response times of the four emergencies; this is anticipated since `emergenc` is regarded as a blocking factor and differences in response across blocks are expected by design.

By default, SAS PROC GLM assumes all factors are fixed. From Table 12.1, we see that when `emergenc` is a fixed factor, the  $F$ -statistic for `panel` is 20.09 on 2 and 12 degrees of freedom. The small corresponding  $p$ -value suggests that response time varies with the type of `panel`.

If `emergenc` is a random factor, the pattern of expected mean squares in Table 12.8 indicates that the appropriate denominator mean square for testing `panel` is the interaction mean square. This test of `panel` is produced with an explicitly requested `test` statement in (tway/code/display4.sas). We see that `panel` is tested with  $F = 24.15$  on 2 and 6 degrees of freedom.

The  $F$ -statistic for `panel` corresponds to a small  $p$ -value under either assumption on `emergenc`. Therefore, in this example, we reach the same conclusion under both assumptions: that response time differs across `panels`. However, since in general the  $F$ -statistic differs in the two cases, the ultimate conclusion concerning a fixed factor may depend crucially on

our assumption concerning the other factor. If `emergenc` is a fixed factor, the conclusions regarding panels applies to these four emergencies only. If `emergenc` is a random factor, the panel conclusions apply to the entire population of emergencies from which these four emergencies are assumed to be a random sample.

The  $F$ -test for interaction between `panel` and `emergenc` when `emergenc` is a random factor is the same test as when `emergenc` is a fixed factor.

Since `panel` is a fixed factor, an appropriate followup is a Tukey test to compare the response time for each display panel. This is shown in Table 12.3 for the case where `emergenc` is fixed. We find that both display panel 1 and display panel 2 have significantly shorter response times than display panel 3, but panels 1 and 2 are not significantly different. Therefore, we conclude that display panel 3 can safely be eliminated from further consideration. The absence of interaction tells us that these conclusions are consistent over emergencies. If interaction had existed in this experiment, one would have concluded that the optimal panel differs according to the type of emergency. Then one would need to make separate panel recommendations for each emergency type. Since we will normally select just one panel type for the entire facility, and since we have no control over emergencies, the decision process would become more difficult.

The groupings of levels of `panel` in Table 12.3 and the confidence intervals in Figure 12.2 are different but equivalent ways of displaying the differences between all pairs of `panel` means using the two-sided Tukey multiple comparisons procedure introduced in Section 6.3.

Table 12.3 displays the results of the  $\binom{3}{2} = 3$  pairwise tests. The common letter B encompassing the sample means 18.250 and 20.250 for panels 2 and 1, respectively, indicates that the difference between the corresponding population means is not significant. There is no letter bonding the sample mean 25.625, calculated from panel 3, to any other sample mean. This indicates that the population mean of panel 3 is significantly different from all other population means.

Figure 12.2 provides two confidence interval displays for pairwise comparisons of the population means of the three panels. The top display contains confidence intervals on each pairwise difference. A pairwise difference of means is significantly different from zero; equivalently, the two means differ significantly if the confidence interval for the pairwise difference excludes zero. If this confidence interval includes zero, then conclude that the two population means do not significantly differ. Thus the “1–2” interval says that these two panel means are indistinguishable. The “1–3” and “2–3” intervals says that the mean of panel 3 differs from the means of the other two panels.

The bottom display contains simultaneous confidence intervals for the three population means, where the confidence coefficient, here 95%, is the probability that each interval contains its respective mean. If two of these confidence intervals overlap, then the corresponding population means are not significantly different. Since panels 1 and 2 have overlapping intervals, these two panel means are not distinguishable. If a pair of these confidence intervals does not overlap, then the corresponding population means are declared to differ significantly. Since the panel 3 interval does not overlap the other two, we conclude that the mean of panel 3 differs from the means of the other two panels.

The formula for the simultaneous 95% confidence intervals on pairwise mean differences shown in Figure 12.2 is

$$\mu_i - \mu_j: \quad \bar{Y}_i - \bar{Y}_{i'} \pm q_{.05} \sqrt{\frac{MS_{Res}}{m}} \quad (12.9)$$

where  $q_{.05}$  is the 97.5<sup>th</sup> percentile of the Studentized range distribution and  $m$  is the common sample size used in calculating each sample mean.

In Table 12.1,  $q_{.05} = 3.77278$ ,  $MS_{Res} = 5.791667$  and  $m = 8$ . The “minimum significant difference” in this table is the “±” part of formula (12.9),

$$q_{.05} \sqrt{\frac{MS_{Res}}{m}} = 3.2101 \quad (12.10)$$

The S-PLUS listing that accompanies Figure 12.2 shows a critical point of 2.6679. This is not the Studentized range tabular value  $q_{.05}$  but instead  $q_{.05}/\sqrt{2}$ . This  $\sqrt{2}$  adjustment is used by S-PLUS to produce the simultaneous confidence intervals on the individual population means in the top part of Figure 12.2.

## 12.13 Example—The *Rhizobium* Data

### Study Objectives

Erdman (Erdman, 1946) discusses experiments to determine if antibiosis occurs between *Rhizobium Meliloti* and *Rhizobium Trifolii*. Rhizobium is a bacteria, growing on the roots of clover and alfalfa, that fixes nitrogen from the atmosphere into a chemical form the plants can use. The research question for Erdman was whether there was an interaction between the two types of bacteria, one specialized for alfalfa plants and the other for clover plants. If there were an interaction, it would indicate that clover bacteria mixed with alfalfa bacteria changed the nitrogen fixing response of alfalfa to alfalfa bacteria or of clover to clover bacteria. The biology of the experiment says that interaction indicates antibiosis or antagonism of

the two types of rhizobium. That is, the goal was to test whether the two types of rhizobium kill each other off. If they do, then there will be less functioning bacteria in the root nodules and consequently nitrogen fixation will be slower.

Erdman ran two sets of experiments in parallel. In one the response variable was the nitrogen content in clover plants, in the other the nitrogen content in alfalfa plants. The treatments were combinations of bacterial cultures in which the plants were grown. As a historical note, beginning with (Steel and Torrie, 1960), the one-way analysis of the clover plus alfalfa combination of the Clover experiment has been frequently used as an example to illustrate multiple comparisons procedures. Here we examine the complete data from two related two-way experiments.

### Data Description

Both experiments are two-way factorial experiments with two treatment factors:

**strain:** one of six rhizobium cultures, five pure strains and one a mixture of all five strains. Five strains of alfalfa rhizobium were used for the alfalfa plants and five strains of clover rhizobium were used for the clover plants.

**comb:** a factor at two levels. At one level the rhizobium cultures consisted of only strains specialized for the host plant. At the other level each of the six cultures was combined with a mixture of rhizobium strains specialized for the other plant.

#### 12.13.1 First *Rhizobium* Experiment: Alfalfa Plants

Five observations on the response variable, nitrogen content, were taken at each of the 12 **trt** treatment combinations. Primary interest was in the differences in responses to the six rhizobium treatments. Erdman originally analyzed the response variable “milligrams of nitrogen per 20 plants”. After studying his analysis and his discussion we choose to analyze a related response variable, “milligrams of nitrogen per gram of dry plant weight”. We give the original analysis as Exercise 12.1.

#### 12.13.2 Second *Rhizobium* Experiment: Clover Plants

Five observations on the response variable, nitrogen content, were taken at each of the 12 treatment combinations. Primary interest was in the differences in responses to the six rhizobium treatments. Erdman originally

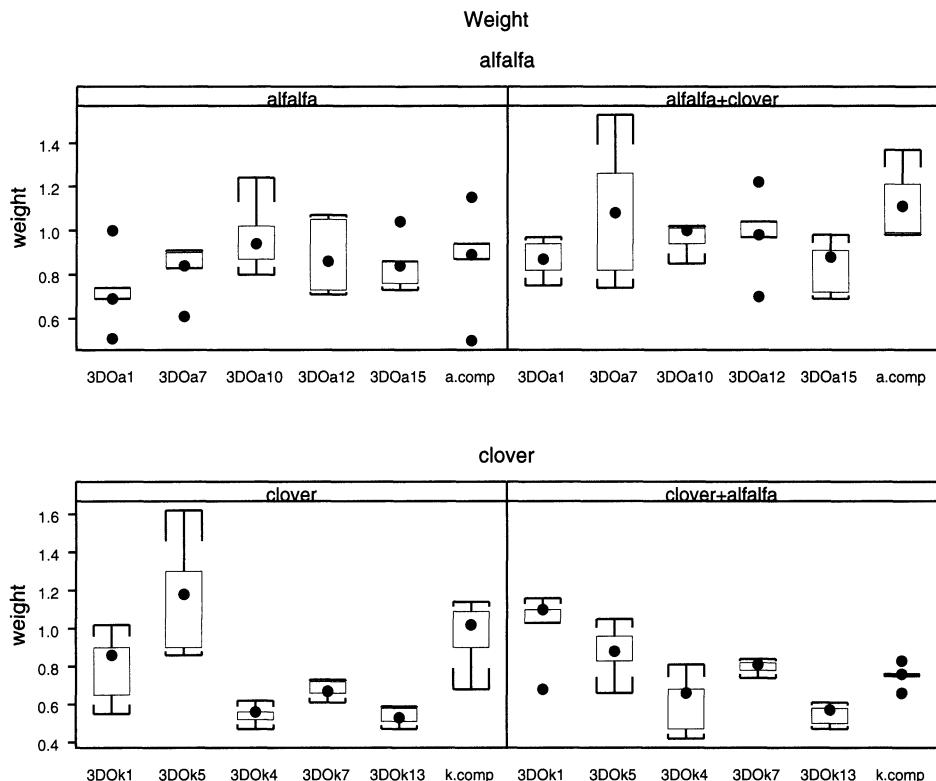


FIGURE 12.6. Rhizobium: Plant weight per pot.  
(tway/code/rhiz-bwplot.ti.s), (tway/figure/rhw.ti.eps.gz)

analyzed the response variable “milligrams of nitrogen per 10 plants”. After studying his analysis and his discussion, we choose to analyze a related response variable, “milligrams of nitrogen per gram of dry plant weight”. We give the original analysis as Exercise 12.2.

### 12.13.3 Initial Plots

Files (`datasets/rhiz1-alfalfa.dat`) and (`datasets/rhiz3-clover.dat`) contain the complete data for both experiments. They are read into S-PLUS with file (`tway/code/rhiz-read.s`) and into SAS with files (`tway/code/rhiz1.sas`) and (`tway/code/rhiz3b.sas`). The data are plotted in Figures 12.6, 12.7, and 12.8. The alfalfa plots are on the top and the clover plots on the bottom. Erdman’s response variable is in Figure 12.7. Our response variable is in Figure 12.8. The single most evident feature

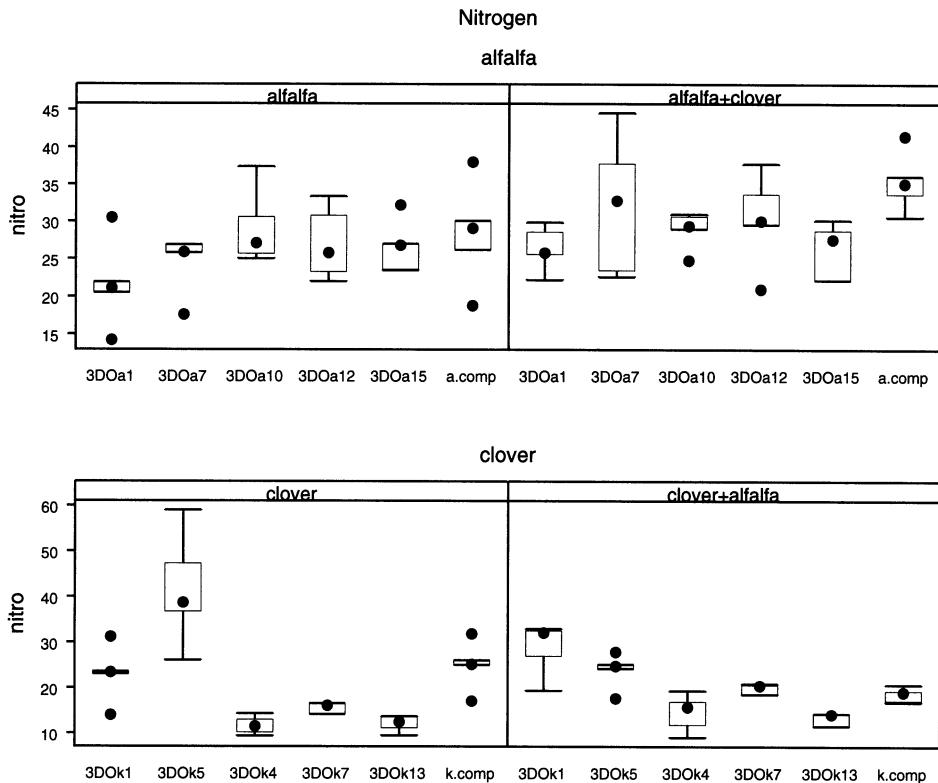


FIGURE 12.7. Rhizobium: Nitrogen weight per pot.  
 (tway/code/rhiz-bwplot.ti.s), (tway/figure/rhn.ti.eps.gz)

from the clover boxplots in Figures 12.6, 12.7, and 12.8 is the large response to the pure culture 3DOK5. This observation is the one that caused us to consider the alternate response variable. There were fewer plants, hence larger plants, for this strain. We posit that the reported values were scaled up, that is reported as grams per 10 plants. We hope that analyzing the ratio, milligrams of nitrogen per gram of plants, rather than the reported rate, milligrams per 10 plants, will adjust for the outliers. Nothing in the alfalfa plots is as clear.

As a graphical aside, we looked at four different layouts for these plots. In Figures 12.6, 12.7, and 12.8 we show vertical boxplots by strain conditioned on combination. We also looked at vertical boxplots by combination conditioned on strain and horizontal boxplots with both conditionings. We chose this one because we have a preference for the response variable on the vertical axis and because we believe the patterns are easier to see when

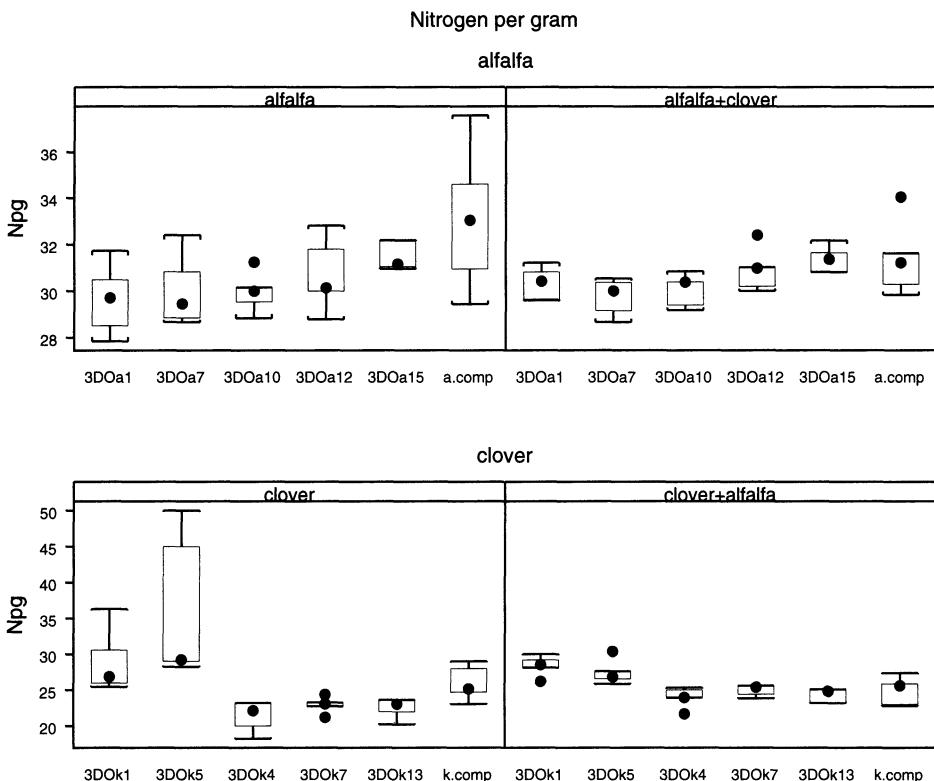


FIGURE 12.8. Rhizobium: Nitrogen weight per milligram. See Tables 12.13 and 12.14.  
 (tway/code/rhiz-bwplot.ti.s), (tway/figure/rhnpg.ti.eps.gz)

this example is conditioned on combination. The other three layouts can be viewed by running the programs in files (tway/code/rhiz-bwplot.t.s), (tway/code/rhiz-bwplot.i.s), (tway/code/rhiz-bwplot.s). Also see the discussion in Section 13.A.

#### 12.13.4 Alfalfa Analysis

The ANOVA table and table of means for the alfalfa experiment are in Table 12.13. Since there was no interaction with the combination of clover strains of bacteria (**strain:comb** interaction  $p$ -value = .53 in Table 12.13), there is no evidence of antibiosis or antagonism.

Since only the **strain** main effect is significant, we confine our investigation to differences among the means for **strain**. Figures 12.9 and 12.10 display the results of the Tukey multiple comparison procedure for compar-

TABLE 12.13. ANOVA table and table of means for alfalfa experiment. See Figure 12.8.  
(tway/code/rhiz-alf-aov.s)

```
S-PLUS (tway/transcript/rhiz-alf-aov.st):
> rhiz.alfalfa.aov <- aov(Npg ~ strain * comb, data=rhiz.alfalfa)
> summary(rhiz.alfalfa.aov)
   Df Sum of Sq  Mean Sq  F Value    Pr(F)
strain      5  46.22159  9.244318 4.564551 0.0017424
comb        1   0.57332  0.573320 0.283087 0.5971392
strain:comb  5   8.43515  1.687030 0.833002 0.5327529
Residuals   48  97.21160  2.025242
>
> alf.means <- model.tables(rhiz.alfalfa.aov, type="means", se=T)
Warning messages:
  Model was refit to allow projection in:
  model.tables.aov(rhiz.alfalfa.aov, type = "means", se = T)
> alf.means$tables$strain
  3D0a1    3D0a7  3D0a10   3D0a12   3D0a15 a.composite
30.00309 29.89324 29.9997 30.81709 31.43472      32.265
> alf.means$n["strain"]
strain
10
> alf.means$se$strain
10
10 0.6364341
```

ing **strain** mean differences. Since **strain** has 6 levels, we simultaneously examine all  $\binom{6}{2} = 15$  pairwise mean differences. Any mean difference having a confidence interval in Figure 12.10 that doesn't include 0 is declared a significantly differing pair. There are three such confidence intervals, therefore we conclude that **a.composite** has a significantly higher mean response than each of **3D0a1**, **3D0a7**, and **3D0a10**; these were the only significant differences detected. The inference is that any of the three treatments with high response (**3D0a12**, **3D0a15**, or **a.composite**) should be used.

Equivalent information is contained in Figure 12.9, where two population means are declared significantly different if their corresponding sample means are *not* underlined by the same line. (Such an underlining display may be used only when all samples have the same size, as is the case here.)

Figure 12.10 is very busy. We provide a graphical summary of our conclusions in Figure 12.11 by constructing an orthogonal basis for the contrasts. We see that the single comparison between **a.composite** and the average

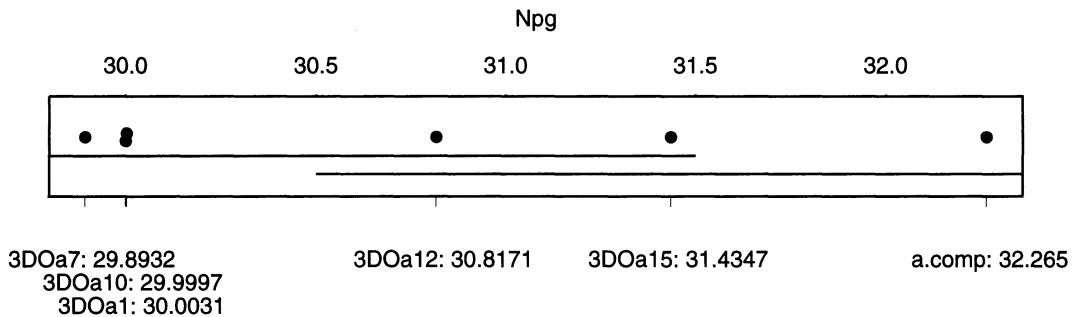
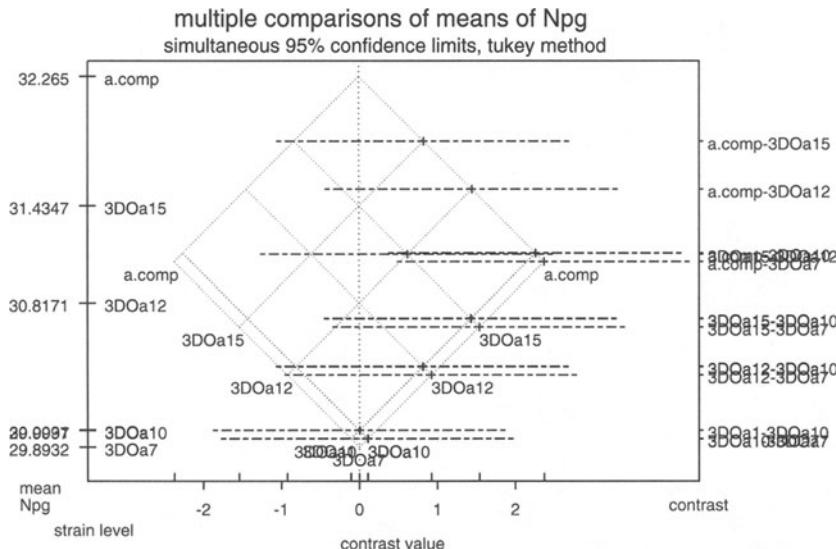


FIGURE 12.9. Means for alfalfa experiment. Dots that appear over a common horizontal line correspond to population means that do not differ significantly according to the Tukey multiple comparisons procedure with simultaneous 95% confidence intervals. Compare this figure to Figure 12.10.  
 (`tway/code/rhiz-alf-aov.s`), (`tway/figure/alfmeans.eps.gz`)

of the three strains with low means (3DOa1, 3DOa7, and 3DOa10) is the only significant effect.

Figure 12.10 and 12.11 each have two parts. Part a is the MMC (mean-mean multiple comparisons) plot discussed in Chapter 7. There is severe overprinting of the confidence intervals and their labeling because so many of the means and estimates of their differences have similar values. The overprinting is itself information of similarity of level means. Nonetheless we need a tiebreaker that will return legibility to the plot. We provide the tiebreaker in part b, an ordinary multiple comparisons plot of the individual contrasts sorted to be in the same order as the contrasts appear in part a. This sort order is based on the values of the level means. The standard sort order used by both S-PLUS in Table 6.4 and SAS in Table 6.5 is based on the names of the levels.

a. Mean–mean multiple comparisons plot. Note that 3DOa1 and 3DOa10 have almost identical means and 3DOa7 is close to both, hence their heights overprint on the left axis and their differences overprint on the right axis.



b. S-PLUS multicomp plot ordered by heights of the mean–mean multiple comparisons plot. We use panel b to break the ties and allow all comparisons to be visible.

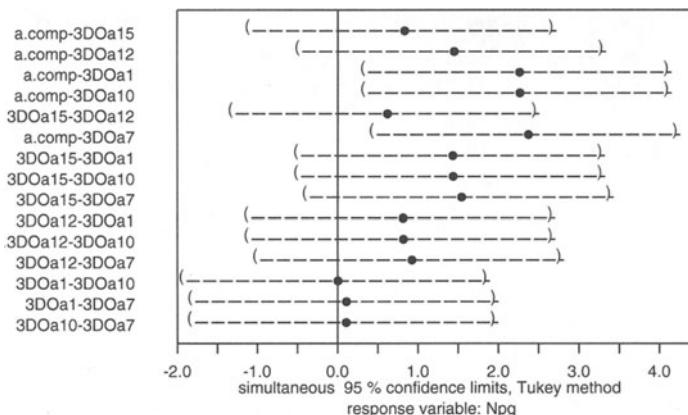
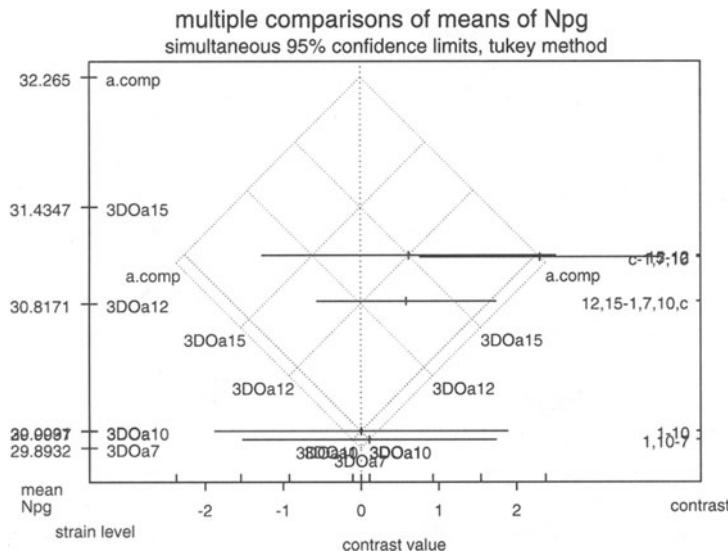


FIGURE 12.10. Tukey contrasts for alfalfa data. Panel a shows all pairwise comparisons, with overlap. Panel b breaks the ties by placing the confidence intervals at equally spaced intervals instead of using the heights used in panel a. The intervals are ordered in panel b to match the ordering of the heights in panel a. See also Figure 12.11 where we have constructed a set of orthogonal contrasts to capture and illustrate the relationships among the levels.

([tway/code/rhiz-alf-mmcc.s](#)),

([tway/figure/alfalfa.mmcc.eps.gz](#)), ([tway/figure/alfalfa.mmcc-tiebreak.eps.gz](#))

a. Mean–mean multiple comparisons plot with an orthogonal basis set of contrasts.



b. S-PLUS multicomp plot of an orthogonal basis set of contrasts ordered by heights of the mean–mean multiple comparisons plot.

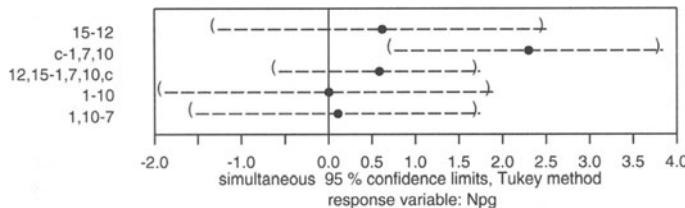


FIGURE 12.11. Orthogonal set of Tukey contrasts for alfalfa data. There are six strains, hence five independent comparisons. This orthogonal set has been chosen to summarize the information in Figure 12.10 and show that only one comparison is significantly different from 0. We see now that the three strains with low means are indistinguishable, that the two intermediate strains are indistinguishable, and that these two clusters are not significantly different from each other. The only significant comparison is from the composite to the cluster of three strains with low means.

(`tway/code/rhiz-alf-mmcc.s`), (`tway/figure/alfalfa.lmat.mmcc.gz`),  
`(tway/figure/alfalfa.lmat.mmcc-tiebreak.gz)`

### 12.13.5 Clover Analysis

The code in (`tway/code/rhiz-clov-aov.s`) produces the ANOVA table and table of means in Table 12.14. In this experiment the `strain:comb` interaction effect is significant and the `comb` main effect is not significant.

The significance of the `strain:comb` in Table 12.14 ( $p$ -value  $< .01$ ) implies that we can't immediately, if at all, interpret the main effect of `strain`. Main effect comparisons of the levels of `comb` and `strain` are inappropriate because the difference in response to two levels of `strain` will differ according to the level of `comb`. From the table of means and the interaction plots in Figure 12.12 we discover, again, that strain 3DOk5 is the anomaly. The interaction is made evident by the lack of parallel profiles in both interaction plot panels. The three points marked as outliers in both boxplot panels are the points that drive much of the remaining analysis.

We therefore repartition the sums of squares in Table 12.15 and look separately at the simple effect of `strain` within each of the levels of `comb`. The differences in the clover strains alone are significant. The differences with the combination clover and alfalfa strains are not. Therefore, we examine only the *simple effects* within the clover strains. These simple effects are the differences between pairs of means of `strain` within the `clover` level of the factor `comb`. We examine and report on those such differences that are statistically significant. Since the simple effect for `strain` within the `clover+alfalfa` level of `comb` is not significant, we do not look further at those means.

Erdman's interpretation of the analysis shows that bacteria strain 3DOk5 showed antibiosis with the alfalfa bacteria strains. With 3DOk5 the response was strong alone and suppressed when combined with the alfalfa bacteria culture.

Table 12.16 shows the contrasts (see Section 6.9) and dummy variables and Table 12.17 shows regression coefficients for the simple effects of `strain` in the clover experiment displayed in Table 12.15. The names for the columns of the dummy variables generated by the program are excessively long and will force the matrix of dummy variables to occupy many pages just to accommodate the column names. Therefore, we abbreviated them. We see the nesting structure in the dummy variables as the `cmbn` columns for pure strains and the `cm+n` columns for combination strains are identical in structure. Only the `cmbn` regression coefficients are significant. We see now that the individual degrees of freedom were constructed from the Helmert contrasts.

TABLE 12.14. ANOVA table and table of means for clover experiment. See Figure 12.8.  
 (tway/code/rhiz-clov-aov.s)

---

```

S-PLUS (tway/transcript/rhiz-clov-aov.st):
> rhiz.clover.aov <- aov(Npg ~ strain * comb, data=rhiz.clover)
> summary(rhiz.clover.aov)
   Df Sum of Sq Mean Sq F Value    Pr(F)
strain      5  642.2538 128.4508 9.915964 0.0000015
comb        1    6.8840   6.8840 0.531420 0.4695524
strain:comb 5  228.2393  45.6479 3.523861 0.0085697
Residuals  48  621.7889 12.9539
> clov.means <- model.tables(rhiz.clover.aov, type="means", se=T)
Warning messages:
  Model was refit to allow projection in:
  model.tables.aov(rhiz.clover.aov, type = "means", se = T)
> clov.means

Tables of means
Grand mean

26.007

strain
 3D0k1 3D0k5 3D0k4 3D0k7 3D0k13 k.composite
28.724 31.863 22.665 23.948 23.393      25.448

comb
clover clover+alfalfa
26.345      25.668

strain:comb
Dim 1 : strain
Dim 2 : comb
  clover clover+alfalfa
  3D0k1 29.042 28.406
  3D0k5 36.286 27.440
  3D0k4 21.354 23.975
  3D0k7 22.932 24.964
  3D0k13 22.486 24.300
  k.composite 25.973 24.923

Standard errors for differences of means
strain    comb strain:comb
1.6096  0.9293    2.2763
replic. 10.0000 30.0000    5.0000

```

---

### Npg: main effects and 2-way interactions

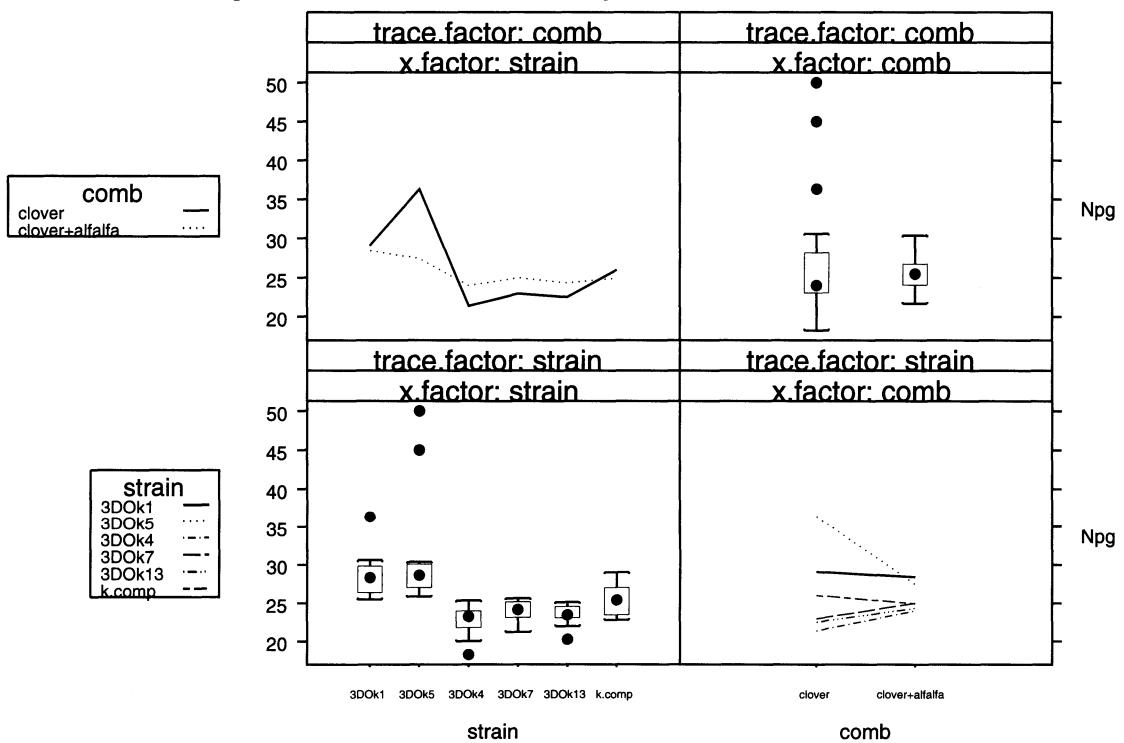


FIGURE 12.12. Interaction plot for clover experiment. The three points marked as outliers are the points that drive much of the remaining analysis.

(`tway/code/rhiz-clov-aov.s`), (`tway/figure/clovint2wt.eps.gz`),  
 (`tway/figure/clovint2wt.color.eps.gz`)

TABLE 12.15. ANOVA table showing simple effects for strain in clover experiment. We partitioned the sums of squares for the nesting with the `split` argument to the `summary` function. We needed to display the names of the individual regression coefficients in order to determine which belonged to each of the levels of `comb`. In this example the `comb` and `strain` effects are orthogonal, hence the partitioning is valid. The individual degrees of freedom are usually not interpretable.

(tway/code/rhiz-clov-aov.s)

```
S-PLUS (tway/transcript/rhiz-clov-nest-aov.st):
> rhiz.clover.nest.aov <- aov(Npg ~ comb/strain, data=rhiz.clover)
> summary(rhiz.clover.nest.aov)
   Df Sum of Sq Mean Sq F Value    Pr(F)
comb  1   6.8840  6.88399 0.531420 0.4695524
strain %in% comb 10  870.4931 87.04931 6.719912 0.0000020
      Residuals 48  621.7889 12.95394
> names(coef(rhiz.clover.nest.aov))
[1] "(Intercept)"                  "comb"
[3] "combcloverstrain1"           "combclover+alfalfastrain1"
[5] "combcloverstrain2"           "combclover+alfalfastrain2"
[7] "combcloverstrain3"           "combclover+alfalfastrain3"
[9] "combcloverstrain4"           "combclover+alfalfastrain4"
[11] "combcloverstrain5"          "combclover+alfalfastrain5"
> summary(rhiz.clover.nest.aov,
+         split=list("strain %in% comb"=
+                     list(clover=c(1,3,5,7,9),
+                           "clover+alfalfa"=c(2,4,6,8,10))))
   Df Sum of Sq Mean Sq F Value    Pr(F)
comb  1   6.8840  6.8840  0.53142 0.4695524
strain %in% comb 10  870.4931 87.0493  6.71991 0.0000020
strain %in% comb: clover  5  788.3930 157.6786 12.17225 0.0000001
strain %in% comb: clover+alfalfa 5  82.1001 16.4200  1.26757 0.2933153
      Residuals 48  621.7889 12.9539
```

TABLE 12.16. Contrasts and dummy variables for simple effects of strain in clover experiment. The sums of squares from these dummy variables are displayed in Table 12.15. The regression coefficients are in Table 12.17. The display has been edited. The unedited listing is also available.

(tway/code/rhiz-clov-aov.s), (tway/transcript/rhiz-clov-nest-aov-x.st),  
 (tway/transcript/rhiz-clov-nest-aov-x-edited.st)

---

```
S-PLUS (tway/transcript/rhiz-clov-nest-aov-x-edited1.st):
> tmp <- abbreviate(names(coef(rhiz.clover.nest.aov)))
>
> contrasts(rhiz.clover$comb)
[1]
clover    -1
clover+alfalfa    1
> contrasts(rhiz.clover$strain)
[1] [2] [3] [4] [5]
3D0k1   -1   -1   -1   -1   -1
3D0k5    1   -1   -1   -1   -1
3D0k4    0    2   -1   -1   -1
3D0k7    0    0    3   -1   -1
3D0k13   0    0    0    4   -1
k.composite 0    0    0    0    5
>
> cnx <- aov(Npg ~ comb:strain, data=rhiz.clover, x=T)$x
> dimnames(cnx)[[2]] <- tmp
> cnx
  (In) comb cmb1 cm+1 cmb2 cm+2 cmb3 cm+3 cmb4 cm+4 cmb5 cm+5
  1    1   -1   -1    0   -1    0   -1    0   -1    0   -1    0
  2    1   -1   -1    0   -1    0   -1    0   -1    0   -1    0
  3    1   -1   -1    0   -1    0   -1    0   -1    0   -1    0
  4    1   -1   -1    0   -1    0   -1    0   -1    0   -1    0
  5    1   -1   -1    0   -1    0   -1    0   -1    0   -1    0

  6    1   -1    1    0   -1    0   -1    0   -1    0   -1    0
  7    1   -1    1    0   -1    0   -1    0   -1    0   -1    0
  8    1   -1    1    0   -1    0   -1    0   -1    0   -1    0
  9    1   -1    1    0   -1    0   -1    0   -1    0   -1    0
 10   1   -1    1    0   -1    0   -1    0   -1    0   -1    0

 11   1   -1    0    0    2    0   -1    0   -1    0   -1    0
...
 20   1   -1    0    0    0    0    3    0   -1    0   -1    0
.....
 51   1    1    0    0    0    0    0    0    0    4    0   -1
...
 60   1    1    0    0    0    0    0    0    0    0    0    5
```

---

TABLE 12.17. Regression coefficients for simple effects of strain in clover experiment. The contrasts and dummy variables are displayed in Table 12.16. The display has been edited. The unedited listing is also available.

(tway/code/rhiz-clov-aov.s), (tway/transcript/rhiz-clov-nest-aov-x.st),  
 (tway/transcript/rhiz-clov-nest-aov-x-edited.st)

S-PLUS (tway/transcript/rhiz-clov-nest-aov-x-edited2.st):

```
> cnxb <- round(coef(summary.lm(rhiz.clover.nest.aov)), 3)
> dimnames(cnxb)[[1]] <- tmp
> cnxb
   Value Std. Error t value Pr(>|t|)
(In) 26.007    0.465 55.971  0.000
comb -0.339    0.465 -0.729  0.470
cmb1  3.622    1.138  3.182  0.003
cm+1 -0.483    1.138 -0.424  0.673
cmb2 -3.770    0.657 -5.737  0.000
cm+2 -1.316    0.657 -2.002  0.051
cmb3 -1.490    0.465 -3.208  0.002
cm+3 -0.411    0.465 -0.884  0.381
cmb4 -0.984    0.360 -2.733  0.009
cm+4 -0.379    0.360 -1.054  0.297
cmb5 -0.075    0.294 -0.254  0.801
cm+5 -0.149    0.294 -0.507  0.615
```

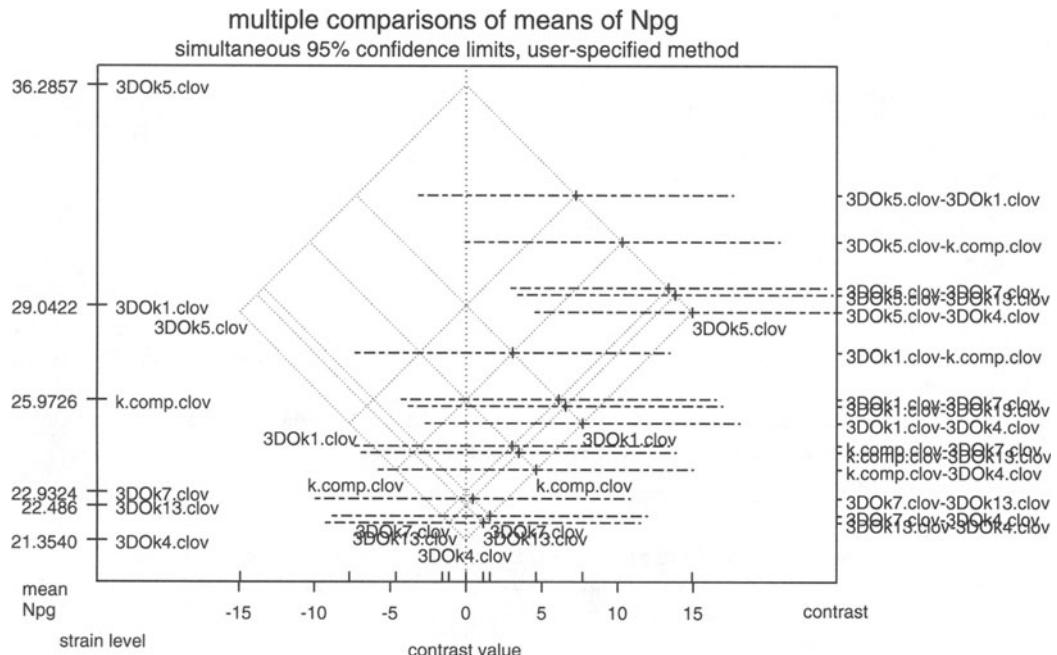


FIGURE 12.13. Tukey simple effect contrasts for `comb="clover"` data. We chose not to break the ties here since it is quite clear that the strain 3DOk5 differs from the rest (significantly for the bottom three strains and marginally for the middle two strains.) Instead, we illustrate this observation in Figure 12.14 with an appropriately chosen set of orthogonal contrasts.

(`tway/code/rhiz-clov-mmc.s`), (`tway/figure/clover.str.clov.mmc.eps.gz`)

Since there is interaction in the clover experiment, we must look at the multiple comparisons for the simple effects of `strain` at each value of `comb`.

Figure 12.13 shows the simple effects for `comb="clover"`. The only significant contrasts are the ones comparing 3DOk5 to the rest of the strains. Since there is pretty clearly nothing else significant, we didn't bother visually breaking the ties. Instead we went directly to a set of orthogonal contrasts in Figure 12.14 where we see that the single contrast comparing 3DOk5 to the others carries all the significance in Figure 12.14.

Figure 12.15 shows that there are no significant contrasts in the simple effects for `comb="clover+alfalfa"`. We forced Figure 12.15 to be on the same scale as Figure 12.13.

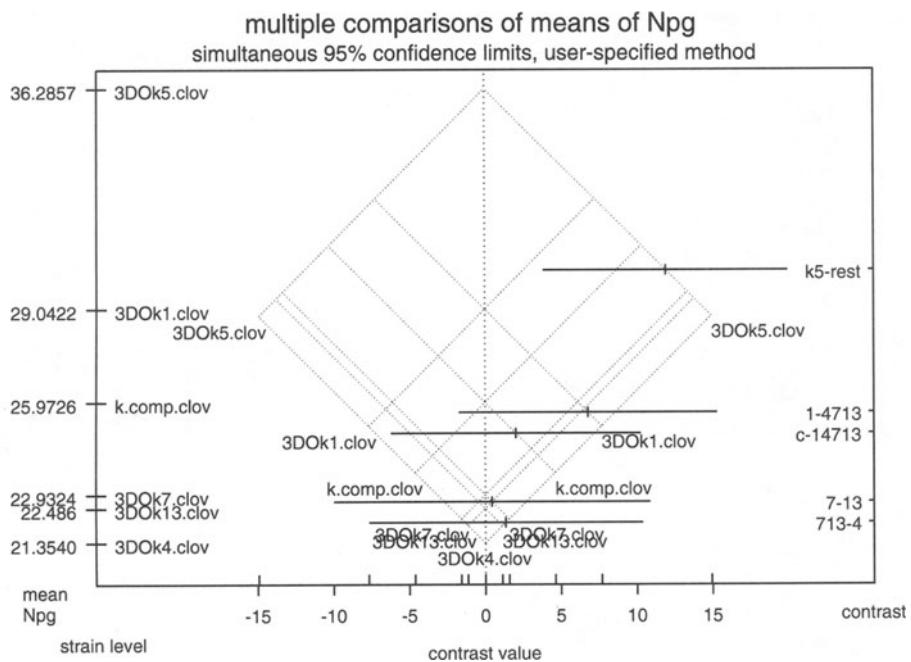


FIGURE 12.14. Orthogonal basis set of Tukey simple effect contrasts for `comb="clover"` data. We summarize the conclusions from Figure 12.13. Four contrasts show that the five strains with the lowest means are indistinguishable. The remaining contrast shows that 3DOk5 differs from the others.  
`(tway/code/rhiz-clov-mmcc.s), (tway/figure/clover.str.clov.lmat.mmcc.eps.gz)`

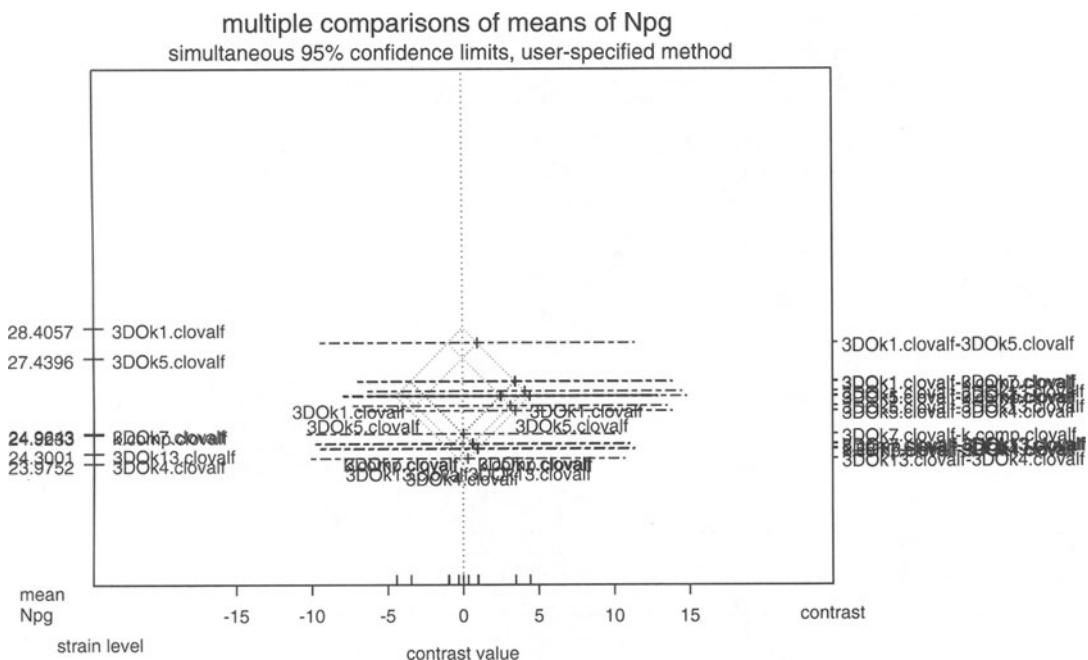


FIGURE 12.15. Tukey simple effect contrasts for `comb="clover+alfalfa"` data. This plot is on the same scale as Figure 12.13. This common scale emphasizes the disparity between 3DOk5 in `comb="clover"` and any values of `strain` in `comb="clover+alfalfa"`. Since none of the simple effects for `strain` within the `clover+alfalfa` level of `comb` are significant, we do not bother breaking the ties of the overprinting.

(`tway/code/rhiz-clov-mmc.s`), (`tway/figure/clover.str.clovalf.mmc.gz`)

## 12.14 Models Without Interaction

Experiments with two factors are normally designed with a large enough sample size to investigate the possibility that the factors interact. When the analyst has previous experience with these factors, or subject area knowledge that the factors are unlikely to interact, it is possible to set up the model without an interaction term:

$$\begin{array}{rcl} Y_{ijk} & = & \mu + \alpha_i + \beta_j + \epsilon_{ijk} \\ \text{SAS} & Y & = A \quad B \\ \text{S-PLUS} & Y & \sim A + B \end{array}$$

The residual portion of this no-interaction model includes the  $(a - 1)(b - 1)$  degrees of freedom that would otherwise have been attributable to the  $AB$  interaction. If the no-interaction assumption is correct, the no-interaction model provides a more precise estimate of the residual than a model incorporating interaction and this in turn implies more power for tests involving the individual main effects or the means of their levels. With this model, comparisons among the levels of factor  $A$  or among the levels of factor  $B$  are undertaken in much the same way as in a one-way experiment, but using this model's residual sum of squares and degrees of freedom.

When we initially posit a model containing the two-factor interaction, it may happen that the analysis of variance test for interaction leads to acceptance of the no-interaction hypothesis. If the evidence for no interaction is sufficiently strong (a large  $p$ -value for this test and/or no strong subject area feeling about the existence of interaction), the analyst may feel comfortable about reverting to the no-interaction model and proceeding with the analysis as above. This amounts to pooling a nonsignificant interaction sum of squares with the previous residual sum of squares (calculated under the now rejected assumption of an interaction) to produce a revised residual mean square (under the assumption of no interaction). This combined or pooled estimate is justified because in the absence of interaction, the interaction mean square estimates the same quantity, the residual variance, as does the residual mean square. The pooled estimate of the residual variance is an improvement over the individual estimates because it is constructed with additional degrees of freedom. Therefore, the pooled estimate provides more powerful inferences on the level means of the two factors than would a residual mean square in a model including interaction. See Section 5.4.2 for further discussion of pooling.

## 12.15 Example—Animal Feed Data

### Study Objectives

A manufacturer of animal feed investigated the influence on the amount of vitamin A retained in feed. The manufacturer considered 15 treatment combinations formed from 5 amounts of feed supplement and 3 levels of temperature at which the supplements were added to the feed. Two samples were selected at each treatment combination. The data in file (`datasets/feed.dat`), taken from (Anderson and McLean, 1974), are said to be on transformed scales that this reference does not specify.

### Data Description

The response variable is `retained` and the two factors are `temp` and `supp`.

### Data Input

We analyzed the data with both SAS and S-PLUS. We read the control file (`tway/code/feed.sas`) into SAS. It reads the data into SAS with file (`tway/code/feed1.sas`).

### Analysis

The PROC GLM sections of the listing are included in Tables 12.18 and 12.19. The S-PLUS file (`tway/code/feed.s`) produced the output file (`tway/transcript/feed.st`) and the interaction plot in Figure 12.16.

The profiles in the interaction plot are sufficiently close to parallel to suggest that there is no interaction between `temp` and `supp`.

Initially we fit an interaction model leading to an interaction *p*-value of 0.43, confirming our impression from the interaction plot that `temp` and `supp` do not interact. It is not unreasonable to conclude that temperature affects each concentration of feed supplement in roughly the same way. Therefore, we abandon the assumption of interaction and move to a no-interaction model.

The fit of the no-interaction model shown in Table 12.18 suggests that both `temp` and `supp` impact significantly on `retained`. Since both of these factors have quantitative levels, our analysis of the nature of the mean differences involves modeling the response `retained` as polynomial functions of both temperature and the amount of supplement. The method for accomplishing such modeling was introduced in Section 10.4.

## retained: main effects and 2-way interactions

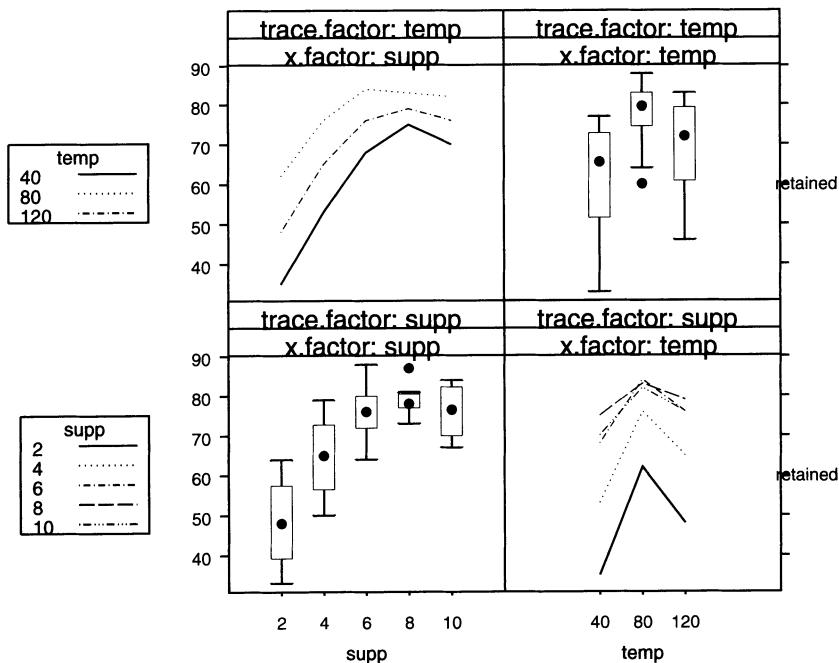


FIGURE 12.16. Feed data interaction plots.  
`(tway/code/feed.s)`, `(tway/figure/feed-i2wt.eps.gz)`,  
`(tway/figure/feed-i2wt-color.eps.gz)`

The level means of `temp` and `supp` are provided in file `(tway/transcript/feed2a.lst)`. The interaction plot in Figure 12.16 suggests that the response to changes in the level of `supp` is quadratic in nature and that possibly the response to changes in the level of `temp` is quadratic as well. Therefore, for both of these factors we performed one degree-of-freedom tests on the linear and quadratic contrasts among the factor levels, with the SAS output shown in Table 12.19. Since the *p*-values for both quadratic contrasts are close to 0, there is strong evidence that the response of vitamin A retention is a quadratic function of both temperature and amount of feed supplement. This finding implies that for maximum vitamin A retention we should recommend intermediate amounts of `temp` and `supp`, perhaps in the vicinity of `temp=80` and `supp=6`. An enlargement of this experiment could more accurately determine the optimal values.

If the analyst had been told, prior to the design of the experiment, that the primary goal was to determine the optimizing combination of the inputs

TABLE 12.18. Feed data: ANOVA and means.

SAS (tway/code/feed2.sas):

```
proc glm data=feed order=data;
  class temp supp;
  model retained = temp|supp / ss3;
  means temp supp;
run;
```

SAS (tway/transcript/feed2.lst):

Dependent Variable: retained

| Source          | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|-----------------|----|----------------|-------------|---------|--------|
| Model           | 14 | 5584.800000    | 398.914286  | 14.05   | <.0001 |
| Error           | 15 | 426.000000     | 28.400000   |         |        |
| Corrected Total | 29 | 6010.800000    |             |         |        |

| R-Square | Coeff Var | Root MSE | retained Mean |
|----------|-----------|----------|---------------|
| 0.929128 | 7.745879  | 5.329165 | 68.80000      |

| Source    | DF | Type III SS | Mean Square | F Value | Pr > F |
|-----------|----|-------------|-------------|---------|--------|
| temp      | 2  | 1479.200000 | 739.600000  | 26.04   | <.0001 |
| supp      | 4  | 3862.133333 | 965.533333  | 34.00   | <.0001 |
| temp*supp | 8  | 243.466667  | 30.433333   | 1.07    | 0.4313 |

temp and supp, the analyst would have considered using a *response surface design*, the most efficient design for this purpose. A brief introduction to such designs is contained in (Montgomery, 2001).

## 12.16 Exercises

- 12.1.** Do the original Erdman alfalfa analysis of Section 12.13.1 with nitro as the response variable. Use data file ([datasets/rhiz1-alfalfa.dat](#)).

TABLE 12.19. Feed data: ANOVA with contrasts.

SAS (tway/code/feed3.sas):

```
proc glm;
  class temp supp;
  model retained = temp supp / ss3;
  contrast 'temp linear' temp -1 0 1;
  contrast 'temp quadratic' temp -1 2 -1;
  contrast 'supp linear' supp -2 -1 0 1 2;
  contrast 'supp quadratic' supp 2 -1 -2 -1 2;
run;
```

SAS (tway/transcript/feed3.lst):

Dependent Variable: retained

| Source          | DF | Sum of Squares |  | Mean Square | F Value | Pr > F |
|-----------------|----|----------------|--|-------------|---------|--------|
|                 |    |                |  |             |         |        |
| Model           | 6  | 5341.333333    |  | 890.222222  | 30.58   | <.0001 |
| Error           | 23 | 669.466667     |  | 29.107246   |         |        |
| Corrected Total |    | 6010.800000    |  |             |         |        |

| R-Square | Coeff Var | Root MSE | retained Mean |
|----------|-----------|----------|---------------|
| 0.888623 | 7.841734  | 5.395113 | 68.80000      |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|--------|----|-------------|-------------|---------|--------|
| temp   | 2  | 1479.200000 | 739.600000  | 25.41   | <.0001 |
| supp   | 4  | 3862.133333 | 965.533333  | 33.17   | <.0001 |

| Contrast       | DF | Contrast SS | Mean Square | F Value | Pr > F |
|----------------|----|-------------|-------------|---------|--------|
| temp linear    | 1  | 369.800000  | 369.800000  | 12.70   | 0.0016 |
| temp quadratic | 1  | 1109.400000 | 1109.400000 | 38.11   | <.0001 |
| supp linear    | 1  | 680.066667  | 680.066667  | 23.36   | <.0001 |
| supp quadratic | 1  | 1360.047619 | 1360.047619 | 46.73   | <.0001 |

- 12.2.** Do the original Erdman clover analysis of Section 12.13.2 with nitro as the response variable. Use data file (*datasets/rhiz3-clover.dat*).

**12.3.** (`datasets/testing.dat`) is  $3 \times 3$  design with 4 observations per treatment combination. The factors are breaker at levels 1 to 3 and Gauger at levels 1 to 3. The observations are strengths of cement. The cement is “gauged” or mixed with water and worked by three different gaugers before casting it into cubes. Three testers or “breakers” later tested the cubes for compressive strength, measured in pounds per square inch. Each gauger gauged 12 cubes, which were divided into 3 sets of 4, and each breaker tested one set of 4 cubes from each gauger. Breakers and gaugers are fixed in this experiment. Breakers and gaugers are people, not machines. Are there differences in the strength of the cement that depend on the handling by the breakers and gaugers?

We got the data from (Hand et al., 1994). The data originally appeared in (Davies and Goldsmith, 1972). There the data were coded by  $.1(X - 1000)$  before analysis. In (`datasets/testing.dat`), columns 1 and 2 are for breaker 1, columns 3 and 4 are for breaker 2, and columns 5 and 6 are for breaker 3; rows 1 and 2 are for gauger 1, rows 3 and 4 are for gauger 2, and rows 5 and 6 are for gauger 3.

The term *coded data* means that they have been centered and scaled to make the numerical values easier to work with by hand. The *F*-tests in the ANOVA table and the t-tests for regression coefficients with coded data are identical to the tests for the original data.

**12.4.** An agronomist compared five different sources of nitrogen fertilizer and a control (no fertilization) on the yield of barley. A randomized block design was used with four types of soil as the blocks. Each soil type was randomly divided into six plots, and the six treatments were randomly assigned to plots within each type. The treatments were, respectively,  $(\text{NH}_4)\text{SO}_4$ ,  $\text{NH}_4\text{NO}_3$ ,  $\text{CO}(\text{NH}_2)_2$ ,  $\text{Ca}(\text{NO}_3)_2$ ,  $\text{NaNO}_3$ , and control. The data, taken from (Peterson, 1985), are contained in the file (`datasets/barleyp.dat`).

- a. Plot the data. Does it appear from the plot that yield is related to treatment? Does it appear from the plot that blocking was successful?
- b. Set up the two-way analysis of variance table for these data and explain what you conclude from it.
- c. Use the Dunnett procedure, introduced in Section 7.1.3, to compare the five fertilizers with the control. Report your findings to the agronomist.

**12.5.** A chemist compared the abilities of three chemicals used on plywood panels to retard the spread of fire. Twelve panels were obtained, and each chemical was randomly sprayed on four of these twelve panels. Two

pieces were cut from each panel and the time was measured for each piece to be completely consumed by a standard flame. (Thus Panel is nested within Chemical and Sample is nested within Panel.) The data, from (Peterson, 1985), appear in the file (`datasets/retard.dat`). Carefully noting the relationship between the factors `chemical` and `panel`, and considering whether these factors are fixed or random, set up an analysis of variance and followup analysis of `chemical` means in order to make a recommendation of the best chemical for retarding the spread of plywood fires.

- 12.6.** The judging of the skating events at the 2002 Winter Olympics in Salt Lake City was very controversial. The file (`datasets/skateslc.dat`) taken from (Olympic Committee, 2001) contains the scores on both `technique` and `presentation` of the five leading `skaters`, assigned by each of nine judges. We have recoded the data with  $10(X - 5)$ . Perform a two-way analysis of variance where the response is the total of both scores. Do further analysis and comment on the consistency of the nine judges across skaters.
- 12.7.** (Box and Cox, 1964), reprinted in (Hand et al., 1994), present the results of a  $3 \times 4$  factorial experiment with four replications per treatment combination to illustrate the importance of investigating the normality assumption underlying analyses of variance. The original response variable is the survival time, `survtime` of each of four antidotes, `treatment` to each of three poisons. The data appear in the file (`datasets/animal.dat`).
- Perform a two-way analysis of variance using `survtime` as the response, taking care to save the calculated cell residuals.
  - Produce a normal probability plot (described in Chapter 5) of the cell residuals and use it to conclude that the residuals are not normally distributed.
  - Redo the two-way analysis of variance with a reciprocal transformation of the response variable `survtime`, and again save the cell residuals. From a normal probability plot of these cell residuals, conclude that these new residuals are normally distributed and hence the transformation was successful.
  - Report your findings to the antidote researchers.
- 12.8.** An experiment was constructed to compare the `effects` on etchings of wafers of four etching `compounds` and heat treatment by each of three `furnaces`. The experiment was reported in (Johnson and Leone, 1967) and the data are in the file (`datasets/furnace.dat`). Viewing `furnaces`

as a random factor and allowing for the possibility of interaction, provide a thorough analyses of these data.

- 12.9.** Anemia, caused by iron deficiency in the blood, is common in some countries. It was hypothesized that food cooked in iron pots would gain iron content from the pot and, hence when eaten, contribute to alleviation of iron deficiency. Research performed by (Adish et al., 1999) compares the iron content (mg/100g) of three types of (traditional Ethiopian) food when cooked in pots of aluminum, clay, or iron. The data, contained in the file (`datasets/ironpot.dat`), give the Iron content in mg/100g of food, the type of Pot, and the type of Food. Perform a two-way analysis of variance and provide interaction plots. Based on your analysis, is the hypothesis supported? Does your conclusion apply to all Foods studied?
- 12.10.** To check the consistency of new car fuel efficiency, the miles per gallon of gasoline consumed was recorded for each of 5 cars of the same year and brand, on each of 10 randomly selected days. The investigation was reported in (Johnson and Leone, 1967) and the data appear in (`datasets/mpg.dat`). Viewing `car` as a random treatment factor and `day` as a random blocking factor, analyze the data and carefully state your conclusions. Suggest ways to elaborate on and improve this experiment.
- 12.11.** (Williams, 1959), originally in (Sulzberger, 1953), examined the effects of temperature on the strength of wood, plywood, and glued joints. The data file is (`datasets/hooppine.dat`) since the studied wood came from hoop pine trees. The response is compressive `strength` parallel to the grain, and the treatment factor is `temperature` in degrees C. An available covariate is the `moisture` content of the wood, and `tree` is a blocking factor.
- a. Fit a full model where both `strength` and `moisture` are adjusted for the blocking factor `tree`, allowing for the possibility that `temp` interacts with `moisture`.
  - b. Conclude that the interaction term can be deleted from this model. Reanalyze without this term. Carefully state your conclusions.
  - c. Investigate the nature of the relationship between `strength` and `temp`. Conclude that a linear fit will suffice.
  - d. Provide plots illustrating the conclusion from part 12.11a and the final model in parts 12.11b and 12.11c.

## 12.A Appendix: Computation for the Analysis of Variance

When there is more than a single factor, the discussion in this chapter is usually limited to the case where the sample size  $n_{ij}$  is the same for each cell. The programs we use for the computation do not usually have this limitation. We will discuss more general cases in Chapters 13 and 14.

With S-PLUS we will be using `aov` for the calculations and `anova` and related commands for the display of the results. `aov` can be used with equal or unequal cell sizes. Model (12.1) is denoted in S-PLUS either by the formula

$$Y \sim A + B + A:B$$

which uses the operator `+` to indicate the sum of two terms and the operator `:` to indicate the interaction of two factors, or by the formula

$$Y \sim A * B$$

which uses the operator `*` to denote the crossing of two factors. The operator `~` is read as “is modeled by”. The second formula is syntactically expanded by the program to the first formula before it is evaluated. We usually prefer the more compact notation  $Y \sim A * B$  because it more closely captures the English statement, “ $Y$  is modeled by the crossing of factors  $A$  and  $B$ .”

With SAS we use `PROC ANOVA` and `PROC GLM`. `PROC ANOVA` is limited to the equal sample size cases (actually, top balanced designs; see the SAS documentation for details). Where there are at least two factors and unequal cell sizes [that is, the  $n_{ij}$  are not constrained to be equal and some cells may be empty (with  $n_{ij} = 0$ )] `PROC GLM` should be used. `PROC ANOVA` may not give sensible answers in such cases. Model (12.1) is denoted in SAS either by the expression

$$Y = A \quad B \quad A*B$$

which uses a space to indicate the sum of two terms and the operator `*` to indicate the interaction term, or by the expression

$$Y = A \mid B$$

which uses the operator `|` to denote the crossing of two factors. The operator `=` is read as “is modeled by”. The second expression is syntactically expanded by the program to the first expression before it is evaluated. We usually prefer the more compact notation  $Y = A \mid B$  because it more closely captures the English statement, “ $Y$  is modeled by the crossing of factors  $A$  and  $B$ .”

The intercept term  $\mu$  and the error term  $\epsilon_{ijk}$  are assumed in both statistical languages. The existence of the subscripts is implied and the actual values are specified by the data values.

# Design of Experiments—Factorial Designs

Designs are often described by the number of factors. Chapter 6, “One-Way Analysis of Variance”, discusses designs with one factor. Chapter 12, “Two-Way Analysis of Variance”, discusses designs with two factors. More generally, we speak of “three-way” or “higher-way” designs and talk about main effects (one factor), two-way interactions (two factors), three-way interactions, four-way interactions, and so forth. Factors can have crossed or nested relationships. A factor can be fixed or random. When interaction is significant, its nature must be carefully investigated. If higher-order interactions, meaning those involving more than two factors, can be assumed to be negligible, it is often possible to design experiments that require observations on only a fraction of all possible treatment combinations.

Section 13.1 discusses a three-way ANOVA design with a covariate and polynomial contrasts. Section 13.2 introduces Latin squares. Section 13.3 introduces simple effects for interaction analyses. Section 13.4 discusses a nested factorial experiment with both crossed and nesting relationships among the factors. Section 13.6.1 discusses SAS types of sums of squares and sequential and conditional tests. Related topics are discussed in Chapter 14.

## 13.1 A Three-Way ANOVA—Muscle Data

(Cochran and Cox, 1957) report on an experiment to assess the effect of electrical stimulation to prevent the atrophy of muscle tissue in rats. The dataset is in file (`datasets/cc176.dat`). The response `wt.d` is the weight of

the treated muscle. There were three fixed factors: the number of treatments daily, `n.treat`, 1, 3, or 6; the duration of treatments in `minutes`, 1, 2, 3, or 5; and the four types of `current` used. A concomitant variable, the weight of corresponding muscle on the opposite untreated side of the rat, `wt.n`, was also made available. There were two replications of the entire experiment.

The analysis is constructed with (`dsgn/code/cc176.s`). The data are plotted in Figure 13.1. The ANCOVA and adjusted means are in Table 13.1. Also included in Table 13.1 is a partitioning of the 2 degrees of freedom for `n.treats` into linear and quadratic components, taking account of the unequal spacing of the quantitative levels of `n.treats`.

Table 13.1 suggests that after adjusting for the concomitant variable `wt.n`, there are no significant interactions and the effect of `minutes` is not significant. This table shows that `n.treat` contributes significantly to explaining `wt.d`,  $p$ -value=.0000036. The visible upward trend in all panels of Figure 13.1 suggests that response `wt.d` increases linearly with `n.treats` and differs according to the type of `current` used. The response variable `y.adj` in Figure 13.2 is constructed by adjusting the response `wt.d` for the covariate `wt.n`. We see in both Figures 13.1 and 13.2 a larger response when `current` is `25.cycle` than when `current` is at one of its other three levels. Inclusion of `wt.n` reinforces these conclusions. The parallel traces in Figure 13.2 correspond to the absence of interaction between `n.treat` and `current`, a finding also suggested by the large  $p$ -value for this interaction in Table 13.1.

A display comparable to Figure 13.3 could be used to determine the nature of a 3-way interaction. Such an interaction does not exist in this example.

Figure 13.4 shows four different models for the relationship between the response `wt.d` and the covariate `wt.n`. The overall conclusion is that the relation between `wt.d` and `wt.n` differs according to the levels of `current` and `n.treat`. More detail appears in the caption of this figure.

Figure 13.5 is a Tukey procedure mean-mean plot examining the six pairwise differences among the four levels of `current`. As summarized in the caption of this figure, four of these six differences are declared statistically significant. Inspection of Figure 13.5 and the means of the levels of `current` in Table 13.1 reveals that `25.cycle` and `60.cycle` current, indistinguishable from each other, correspond to significantly greater treated muscle weight `wt.d` than either `galvanic` or `faradic` current.

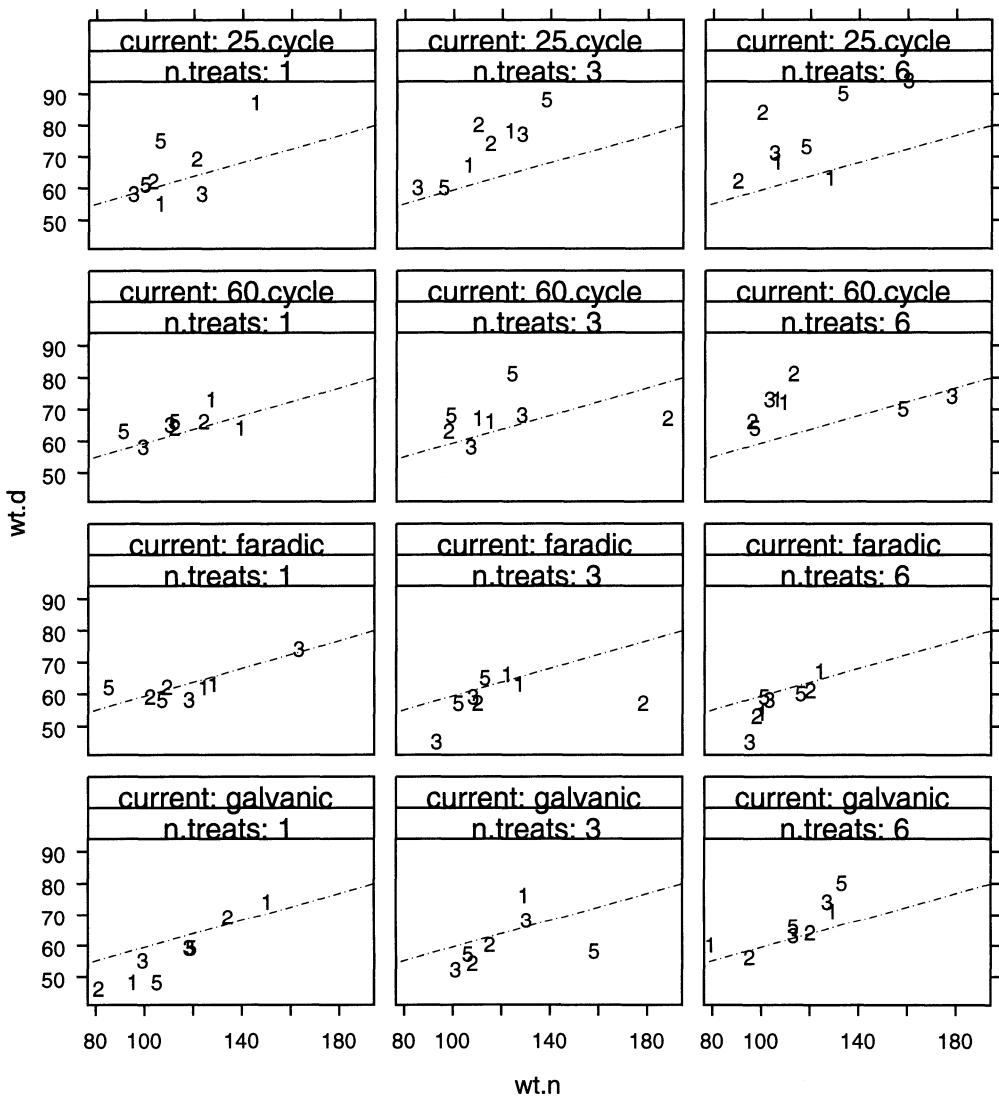


FIGURE 13.1. Muscle data. The response variable **wt.d** is plotted against the covariate **wt.n** within each **current**  $\times$  **n.treats** experimental condition. The plotting symbol is the duration of the treatment in minutes. The ANCOVA and adjusted means are in Table 13.1. We also plotted the common regression line (ignoring experimental conditions) of the response against the covariate. The presence of a covariate **wt.n** effect is evident from the graph by noting that the points approximate the uphill slope of the regression slope. The absence of a **minutes** effect is evident since there is no systematic pattern among the plotting symbols.

(dsgn/code/cc176.s), (dsgn/figure/cc176-1.eps.gz)

TABLE 13.1. Muscle data. ANCOVA and adjusted means. The covariate `wt.n`, the linear effect of `n.treats`, and the `current` are the significant treatment effects. We show the calculation of `y.adj`, the response variable adjusted for the covariate, and the adjusted means.  
 (dsgn/code/cc176.s), (dsgn/transcript/cc176.st)

S-PLUS (dsgn/transcript/cc176-1.st):

```
> cc176.aov <- aov(wt.d ~ rep + wt.n + n.treats*minutes*current, data=cc176)
> summary(cc176.aov,
+           split=list(n.treats=list(n.treats.L=1, n.treats.Q=2)),
+           expand.split=F)
      Df Sum of Sq   Mean Sq   F Value    Pr(F)
      rep   1   605.010  605.010 12.58179 0.0009091
      wt.n  1   1334.085 1334.085 27.74362 0.0000036
      n.treats 2   438.850  219.425  4.56316 0.0155658
      n.treats: n.treats.L 1   428.840  428.840  8.91815 0.0045141
      n.treats: n.treats.Q 1    10.010   10.010  0.20817 0.6503532
      minutes   3   183.933   61.311  1.27502 0.2940919
      current   3   2114.389  704.796 14.65694 0.0000008
      n.treats:minutes 6   198.404   33.067  0.68767 0.6605068
      n.treats:current 6   491.507   81.918  1.70356 0.1416319
      minutes:current 9   382.644   42.516  0.88416 0.5462661
      n.treats:minutes:current 18  1021.618   56.757  1.18031 0.3154209
      Residuals 46  2211.965   48.086
>
> ## adjust y for x
> cc176$y.adj <- cc176$wt.d -
+  (cc176$wt.n - mean(cc176$wt.n))*coef(cc176.aov)[["wt.n"]]
> ## duplicate CC Table 5.17
> cc176.means <- tapply(cc176$y.adj, cc176[,c("current","n.treats")], mean)
> cc176.means
      1       3       6
galvanic 56.03380 59.08068 65.28547
faradic  59.94734 55.79464 57.27463
60.cycle 63.25526 63.92297 68.57735
25.cycle 64.47088 71.78380 73.19818
> apply(cc176.means, 1, mean)
galvanic faradic 60.cycle 25.cycle
60.13332 57.6722 65.25186 69.81762
```

## y.adj: main effects and 2-way interactions

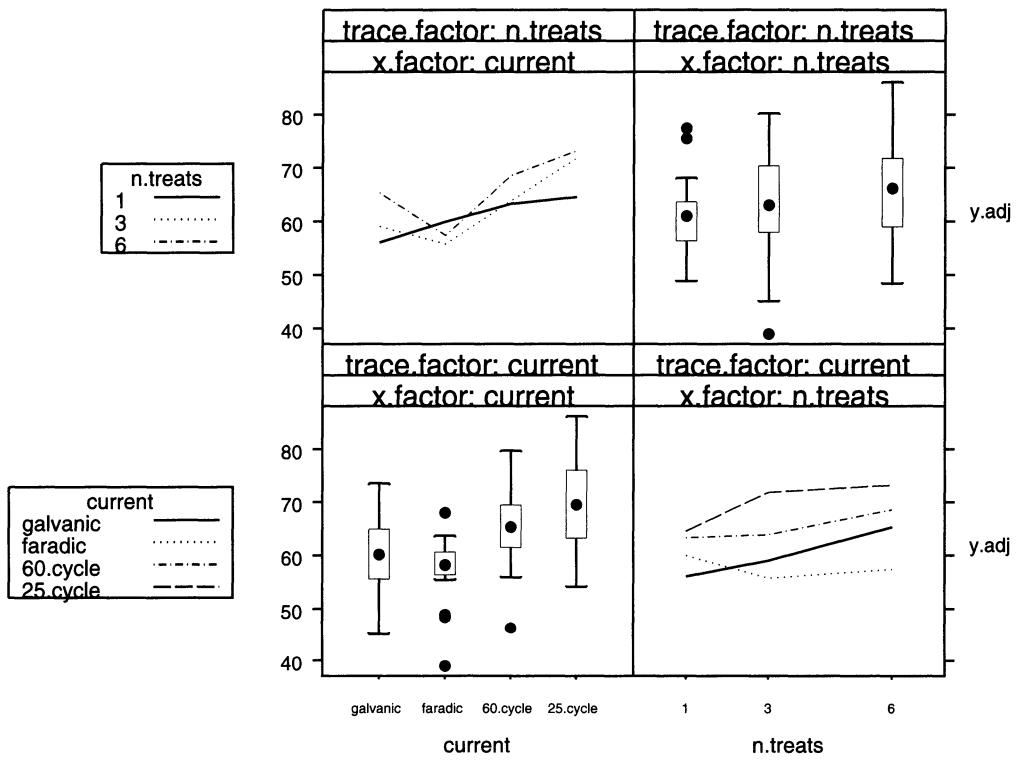


FIGURE 13.2. Muscle data. Two-way interactions of significant main effects from the ANCOVA in Table 13.1. The adjusted response *y.adj* increases linearly with *n.treats* and differs according to the type of current used.

(dsgn/code/cc176.s), (dsgn/figure/cc176-2.eps.gz)

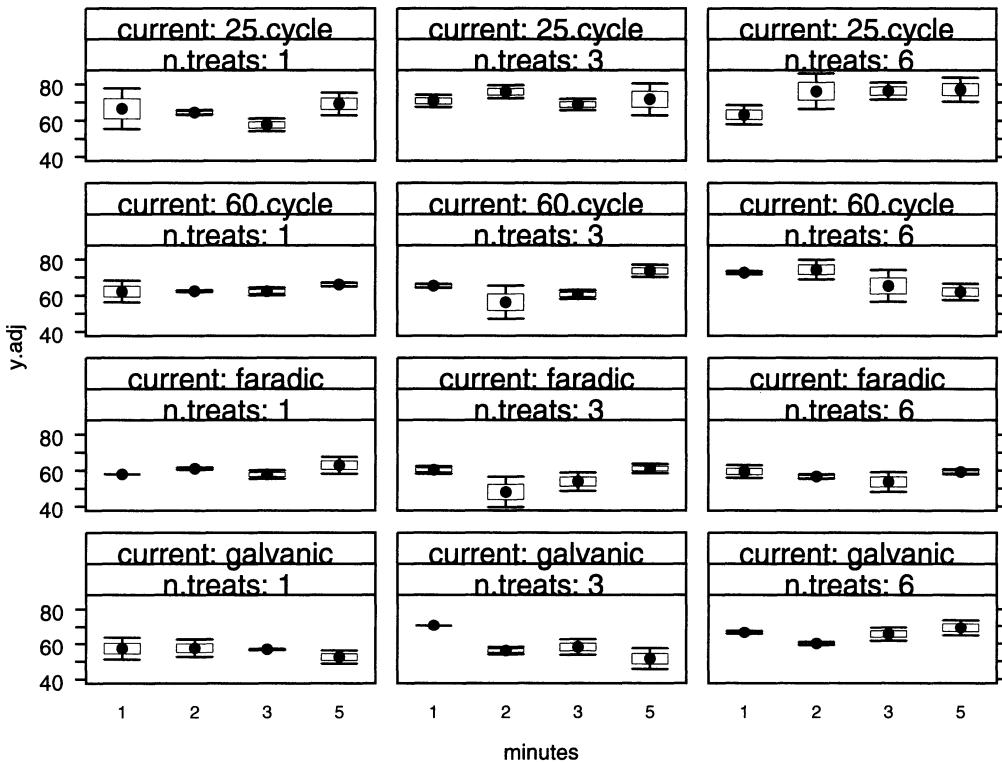


FIGURE 13.3. Muscle data. Three-way interactions of all effects. One of the  $(3!) = 6$  possible orderings. The three-way interaction is not significant in this example. If there were a significant three-way interaction, the patterns in boxplots in adjacent rows and columns would not be the same. For example, we note a hint of a difference in the  $y.\text{adj} \sim \text{minutes}$  behavior across panels. It has a negative slope in the galvanic  $\sim 3$  panel and a positive slope in the faradic  $\sim 3$  panel, but a positive slope in the galvanic  $\sim 6$  panel and a negative slope in the faradic  $\sim 6$  panel. The ANOVA table tells us these differences in slope are not significant. These boxplots are all based on samples of size 2. Such boxplots are a well-defined but uncustomary way to display such a sample.  
 (dsgn/code/cc176.s), (dsgn/figure/cc176-4.eps.gz)

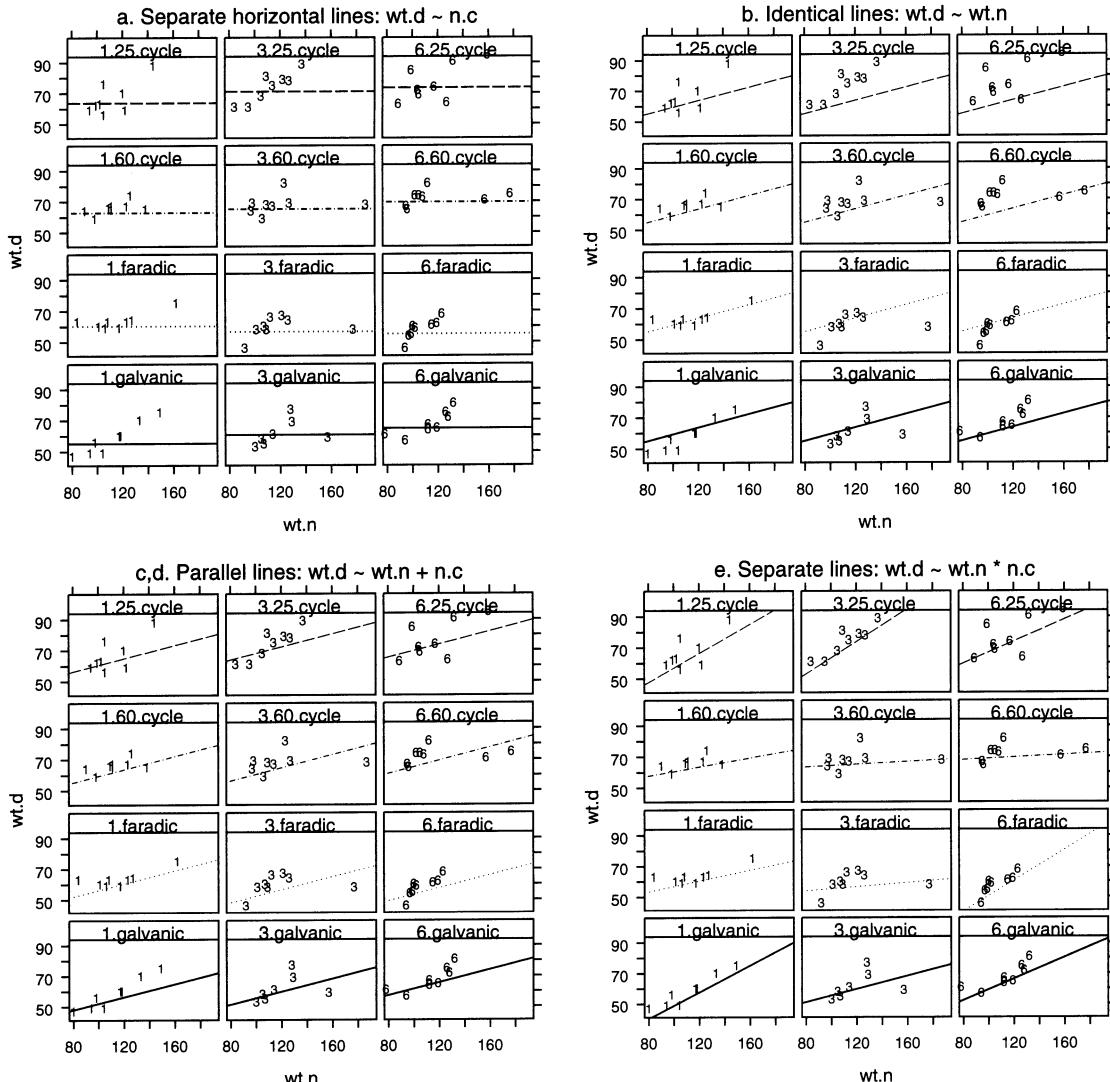


FIGURE 13.4. Muscle data. ANCOVA plots with four different models. Panel a ignores  $wt.n$  and shows the average value of  $wt.d$ . Panel b fits a common regression of  $wt.d$  on  $wt.n$  on all combinations of  $n.treat$  and  $current$  and differs from Figure 13.1 only in its choice of plot symbol. Panel c,d allows for different intercepts but forces common slopes. The difference in intercepts corresponds to the small  $p$ -value for  $wt.n$  in Table 13.1. Panel e shows distinct regressions of  $wt.d$  on  $wt.n$  for each combination of  $current$  and  $n.treat$ . It suggests that the relationship between  $wt.d$  and  $wt.n$  differs according to the levels of  $n.treat$  and  $current$ .

(dsgn/code/cc176-ancova-plot.s),

(dsgn/figure/cc176-5a.eps.gz), (dsgn/figure/cc176-5b.eps.gz),

(dsgn/figure/cc176-5d.eps.gz), (dsgn/figure/cc176-5e.eps.gz)

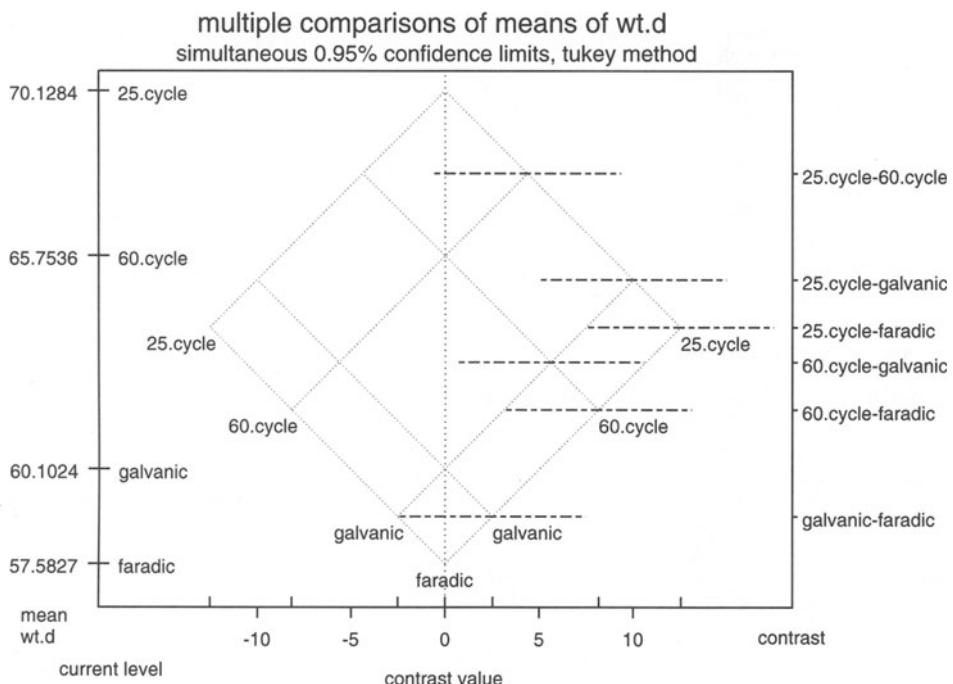


FIGURE 13.5. Muscle data. The Tukey procedure leads to these conclusions about the mean of `wt.d` adjusted for `wt.n` at the four levels of `current`: both of `25.cycle` and `60.cycle` exceed both of `galvanic` and `faradic`; `25.cycle` is indistinguishable from `60.cycle`; and `galvanic` is indistinguishable from `faradic`. The portion of code used to produce this figure uses the argument `valid.check=F`. This override is required because the Tukey procedure is not always valid in situations involving a concomitant variable. It is valid in this balanced design situation with a random concomitant variable (Hochberg and Tamhane, 1987).

(dsgn/code/cc176.s), (dsgn/figure/cc176-7.eps.gz)

## 13.2 Latin Square Designs

This design is useful when we have three factors having the same number, say  $r$ , of levels and the factors do not interact. Although there are  $r^3$  treatment combinations, it permits us to run the experiment with a carefully chosen subset of  $r^2$  of them while retaining the ability to conduct tests on all main effects. A Latin square is a square array of  $r$  Latin letters  $A, B, C, \dots$  such that each letter appears exactly once in each row and once in each column. Typically, the treatment factor is associated with these letters, and both *row* and *column* are blocking type factors. For example, if an experiment is run at  $r$  selected times of day on each of  $r$  days, then each row could represent one of the days, and each column one of the selected times. As another example, displayed in Table 13.2, if we have four cars available to compare the wear of four brands of tire, the rows of the square could represent the wheel position on the car, the columns represent the selected car, and the letters the tire brands.

The basic structure of the ANOVA table is in Table 13.3. Since we are using the Row and Columns factors as blocks, there is no test for those terms in the table. The only test we are justified in making is the test of the Treatment. The purpose of including the Row and Column factors is to pick up some of the Total Sum of Squares and thereby reduce the size of the residual mean square.

The arithmetic of the Latin square design depends on the assumption of no interaction between Row, Column, and Treatment. The arithmetic of the interaction of Row and Column gives  $(r - 1)^2$  df to the interaction and 0 df for an error term. By assuming no interaction, we gain the ability to split the  $(r - 1)^2$  df into two components: Treatment with  $r - 1$  df and Error with  $(r - 1)^2 - (r - 1) = (r - 1)(r - 2)$  df. If the no-interaction assumptions hold, this is a very efficient design.

TABLE 13.2. Sample  $4 \times 4$  Latin square design. The rows represent tire positions: LF is Left-Front, RF is Right-Front, LR is Left-Rear, and RR is Right-Rear.

| Position | Car |   |   |   |
|----------|-----|---|---|---|
|          | 1   | 2 | 3 | 4 |
| LF       | C   | D | A | B |
| RF       | B   | C | D | A |
| LR       | A   | B | C | D |
| RR       | D   | A | B | C |

TABLE 13.3. Sample ANOVA for  $4 \times 4$  Latin square design.

| Source    | df               | Sum of Sq    | Mean Sq    | F Value             | Pr(> F)                                   |
|-----------|------------------|--------------|------------|---------------------|---|
| Row       | $r - 1$          | $SS_{Row}$   | $MS_{Row}$ |                     |   |
| Column    | $r - 1$          | $SS_{Col}$   | $MS_{Col}$ |                     |   |
| Treatment | $r - 1$          | $SS_{Trt}$   | $MS_{Trt}$ | $MS_{Trt}/MS_{Res}$ | $1 - \mathcal{F}_{df_{Trt}, df_{Res}}(F)$ |
| Residual  | $(r - 1)(r - 2)$ | $SS_{Res}$   | $MS_{Res}$ |                     |   |
| Total     | $r^2 - 1$        | $SS_{Total}$ |            |                     |   |

Almost always,  $5 \leq r \leq 8$ , for if  $r < 5$  there are too few df for error, and one is unlikely to encounter situations where one has three factors each having  $r > 8$  levels, two of which are blocking factors. However, it is possible to run an experiment containing several  $3 \times 3$  squares or several  $4 \times 4$  squares, each of which is considered a block, in order to achieve sufficient error df.

Catalogues of Latin squares appear in (Cochran and Cox, 1957) and elsewhere. In practice, one selects a square from a catalogue and randomizes it by randomly assigning levels of one of the blocking factors to the rows of the square, randomly assigning levels of the other blocking factor to the columns of the square, and then randomly assigning treatment levels to the letters.

### 13.2.1 Example—Latin Square

The dataset in file (`datasets/tires.dat`), from (Hicks, 1964) (pages 46 and 67–68), is displayed in Table 13.4 alongside the original Latin square.

An initial ANOVA run, Tables 13.5 and 13.8, revealed significant differences among cars and brands, but not among positions. Here  $r = 4$ , allowing just 6 df for estimating error. Hence the denominator df of the  $F$ -tests is also 6, which as discussed in Section 5.4.4 implies that these tests have little

TABLE 13.4. Latin square of tire wear experiment. The Latin square from Table 13.2 is repeated here.

| Position | Car |   |   |   | Position | Car |    |    |    |
|----------|-----|---|---|---|----------|-----|----|----|----|
|          | 1   | 2 | 3 | 4 |          | 1   | 2  | 3  | 4  |
| LF       | C   | D | A | B | LF       | 12  | 11 | 13 | 8  |
| RF       | B   | C | D | A | RF       | 14  | 12 | 11 | 13 |
| LR       | A   | B | C | D | LR       | 17  | 14 | 10 | 9  |
| RR       | D   | A | B | C | RR       | 13  | 14 | 13 | 9  |

power. Nevertheless, the differences in this example among **cars** (Table 13.6) and **brands** are large enough for the *F*-tests to detect them.

To learn about the nature of the brand differences, we reran with a request for Tukey multiple comparisons tests on the brand means (Tables 13.7 and 13.8). We find that brand 1 had significantly greater wear than brands 3 and 4, but the improvement in wear of brand 1 over brand 2 was not significant. We also see that cars 1, 2, and 3 all had significantly greater wear than car 4; no significant difference in tire wear was detected among cars 1–3.

In this example the primary interest is studying the differences between brands of tires. Both car and position are blocking factors. We assume different cars will have different effects on tires because each person who owns a car drives different routes and puts the car through different wear patterns. We know there are differences in position on the car. Front tires are used for steering, rear tires just follow. The goal of the Latin square experiment is to reduce the residual sum of squares by absorbing some of the variation into known blocking factors. This makes the comparisons of interest, those on brand, more precise because they can be made with a smaller standard deviation (based on the residual mean square).

The results of the *F*-test on a blocking factor are not ordinarily presented in the discussion because block differences are expected, and multiple comparisons on block means are not usually performed. Nevertheless, when blocks are significant, it is an indication that the blocking was worthwhile. Most experimental design texts contain formulas for the efficiency attributable to blocking in Latin square, randomized complete block, and other experimental designs; see, for example, (Cochran and Cox, 1957) (Section 4.37).

We continue with this example in Exercise 13.6 where we use dummy variables to illustrate the linear dependence of the treatment (**brand**) sum of squares on the interaction of the two blocking factors.

TABLE 13.5. Latin square design. SAS analysis for *tires* data. The ANOVA table from the *model* statement is here. The means are in Table 13.6. The significant *wear* difference between *brands* is apparent in Tables 13.6 and 13.7.

```
SAS (dsgn/code/tires.sas):
title 'Analysis of Tires Latin Square Data';
data one;
  infile "&hh/datasets/tires.dat" firstobs=2;
  input car position brand wear;

proc anova;
  class car position brand;
  model wear = car position brand;
  means car;
  means brand / tukey;
run;
```

SAS (dsgn/transcript/tires-1.lst):  
Dependent Variable: WEAR

| Source          | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|-----------------|----|----------------|-------------|---------|--------|
| Model           | 9  | 75.5625000     | 8.3958333   | 9.37    | 0.0066 |
| Error           | 6  | 5.3750000      | 0.8958333   |         |        |
| Corrected Total | 15 | 80.9375000     |             |         |        |

| R-Square | C.V.     | Root MSE | WEAR Mean |
|----------|----------|----------|-----------|
| 0.933591 | 7.846505 | 0.94648  | 12.0625   |

| Source   | DF | Anova SS   | Mean Square | F Value | Pr > F |
|----------|----|------------|-------------|---------|--------|
| CAR      | 3  | 38.6875000 | 12.8958333  | 14.40   | 0.0038 |
| POSITION | 3  | 6.1875000  | 2.0625000   | 2.30    | 0.1769 |
| BRAND    | 3  | 30.6875000 | 10.2291667  | 11.42   | 0.0068 |

TABLE 13.6. Latin square design. The means and standard deviations for the significant block factor car are from the first means statement.  
(dsgn/code/tires.sas)

---

```
means car;
```

---

| SAS (dsgn/transcript/tires-2.lst): |   |                |            |
|------------------------------------|---|----------------|------------|
| Level of<br>car                    | N | -----wear----- |            |
|                                    |   | Mean           | Std Dev    |
| 1                                  | 4 | 14.000000      | 2.16024690 |
| 2                                  | 4 | 12.750000      | 1.50000000 |
| 3                                  | 4 | 11.750000      | 1.50000000 |
| 4                                  | 4 | 9.750000       | 2.21735578 |

---

TABLE 13.7. Latin square design. The Tukey multiple comparisons output on the treatment factor brand is from the second **means** statement. Brand 1 shows significantly greater mean **wear** than brand 3 or brand 4.

(dsgn/code/tires.sas)

---

---

means brand / tukey;

---

SAS (dsgn/transcript/tires-3.lst):  
The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for wear

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

|                                     |          |
|-------------------------------------|----------|
| Alpha                               | 0.05     |
| Error Degrees of Freedom            | 6        |
| Error Mean Square                   | 0.895833 |
| Critical Value of Studentized Range | 4.89559  |
| Minimum Significant Difference      | 2.3168   |

Means with the same letter are not significantly different.

| Tukey<br>Groupi<br>ng | Mean    | N | brand |
|-----------------------|---------|---|-------|
| A                     | 14.2500 | 4 | 1     |
| A                     |         |   |       |
| B A                   | 12.2500 | 4 | 2     |
| B                     |         |   |       |
| B                     | 11.0000 | 4 | 4     |
| B                     |         |   |       |
| B                     | 10.7500 | 4 | 3     |

---

TABLE 13.8. Latin square design. S-PLUS analysis for *tires* data. Brand 1 shows significantly greater mean *wear* than brand 3 or brand 4.  
 (dsgn/code/tires.s)

---

```

S-PLUS (dsgn/transcript/tires.st):
> summary(tires.aov)
      Df Sum of Sq Mean Sq F Value    Pr(F)
car       3   38.6875 12.89583 14.39535 0.0037845
position  3     6.1875  2.06250  2.30233 0.1769470
brand     3   30.6875 10.22917 11.41860 0.0068252
Residuals 6     5.3750  0.89583
>
> tapply(tires$wear, tires$car, "mean")
  1   2   3   4
14 12.75 11.75 9.75
> tapply(tires$wear, tires$position, "mean")
  1   2   3   4
11 12.5 12.5 12.25
> tapply(tires$wear, tires$brand, "mean")
  1   2   3   4
14.25 12.25 10.75 11

> multicomp(tires.aov, method="tukey", focus="brand")

95 % simultaneous confidence intervals for specified
linear combinations, by the Tukey method

critical point: 3.462
response variable: wear

intervals excluding 0 are flagged by '****'

      Estimate Std.Error Lower Bound Upper Bound
1-2      2.00     0.669    -0.317     4.32
1-3      3.50     0.669     1.180     5.82 ****
1-4      3.25     0.669     0.933     5.57 ****
2-3      1.50     0.669    -0.817     3.82
2-4      1.25     0.669    -1.070     3.57
3-4     -0.25     0.669    -2.570     2.07

```

---

### 13.3 Simple Effects for Interaction Analyses

When the analyst believes that interaction exists because of a low  $p$ -value for an interaction term in an ANOVA table, she must study the nature of the interaction.

We re-emphasize that in this situation, tests of the main effects of the factors comprising the interaction are inappropriate. Instead we seek to analyze the *simple effects*, which we now define. An analysis of simple effects, along with interaction plots of cell means which we've previously discussed, are the correct tools for investigating interaction. We note which simple effects are appreciable, either by examining confidence intervals on the simple effects or by testing whether the simple effects are zero. Since such activities involve simultaneous inferences, it is desirable to give attention to use either simultaneous confidence levels or a familywise error rate for simultaneous tests. The importance of studying individual simple effects rather than the overall interaction effect is comparable to the ecological fallacy introduced in Section 4.2 and the cautions resulting from Simpson's paradox, discussed in Section 15.3.

We confine attention here to the case of two factors, say  $A$  and  $B$ , at  $a$  and  $b$  levels, respectively. Continuing with the notation of Equation (12.1), let  $\mu_{ij}$  denote the mean response of the treatment combination where  $A$  is at its  $i^{\text{th}}$  level and  $B$  is at its  $j^{\text{th}}$  level. Then a simple effect for  $B$  is a pairwise comparison of levels of  $B$  at a particular level of  $A$ , for example,  $\mu_{12} - \mu_{13}$ . Similarly, a simple effect for  $A$  is a pairwise comparison of levels of  $A$  at a particular level of  $B$ , such as  $\mu_{32} - \mu_{12}$ .

This assumes that the levels of a factor for which we are calculating simple effects are qualitative in nature. If instead the levels of factor  $B$  are quantitative, then a different analysis is called for, namely, a comparison of comparable polynomial contrasts of the cell means at each level of factor  $A$ . This analysis is superior to performing a separate one-way analyses comparing the levels of  $B$  because we are pooling the information from all levels of  $A$  to estimate the common error variance. This enables us to compare the levels of  $B$  with maximum available power.

In experiments with three or more factors, there is a potential for 3-factor interaction. If there are three factors ( $A, B, C$ ) that interact, this may be interpreted as saying that the nature of the  $AB$  interaction varies according to the particular level of factor  $C$ . An analysis of such an interaction is more complicated than is the case for two interacting factors.

### 13.3.1 Example—The *filmcoat* Data

We illustrate the use of simple effects in analyzing interaction with the dataset (`datasets/filmcoat.dat`) from (Iman, 1994) (pp. 768–778).

#### Study Objectives

Chemical vapor deposition is a process used in the semiconductor industry to deposit thin films of silicon dioxide and photoresist on substrates of wafers as they are manufactured. The films must be as thin as possible and have a uniform thickness, which is measured by a process called infrared interference.

A process engineer evaluated a low-pressure chemical vapor deposition (LPCVD) process that reduces costs and increases productivity. The engineer set up an experiment to study the effect of chamber temperature and pressure on film thickness. Three temperatures and three pressures were selected to represent the low, medium, and high levels of operating conditions for both factors. The experiment was conducted by randomly selecting one of the temperature–pressure combinations and determining the thickness of the film coating after processing is completed. This experiment was repeated three times with each temperature–pressure combination. The engineer wanted to determine the joint effect of temperature and pressure on the mean film thickness. The response was thickness (in Angstroms) of film coatings applied to wafers.

#### Data Description

`temp`: temperature: low, medium, and high levels

`pressure`: pressure: low, medium, and high levels

`coat`: thickness of film coat

The data are displayed in Table 13.9. The table of means and ANOVA table are in Table 13.10. The plots of the means and interactions are in Figure 13.6.

We observe that the `temp × pressure` interaction is moderately significant. Therefore, conclusions about which level of `temp` tends to minimize `coat` depend on the level of `pressure`. This statement is supported by Figure 13.6, which suggests that for low and high `pressure`, the response `coat` is minimized at medium `temp` while for medium `pressure`, `coat` is minimized at high `temp`. In addition, it is suggested that for low and medium `temp`, `coat` is minimized at low `pressure` while for high `temp`, `coat` is minimized at medium `pressure`.

TABLE 13.9. *filmcoat* data. Thickness of film coat at various settings of temperature and pressure.  
 (datasets/filmcoat.dat)

| Temperature | Pressure   |            |            |
|-------------|------------|------------|------------|
|             | Low        | Medium     | High       |
| Low         | 42, 43, 39 | 45, 43, 45 | 45, 44, 47 |
| Medium      | 36, 34, 37 | 39, 39, 37 | 40, 42, 38 |
| High        | 38, 37, 37 | 35, 36, 33 | 40, 41, 42 |

TABLE 13.10. Means and ANOVA table for *filmcoat* data. The moderately significant interaction between pressure and temperature requires that we examine simple effects rather than main effects.  
 (dsgn/code/filmcoat.s)

```
S-PLUS (dsgn/transcript/filmcoat.ma.st):
> tapply(filmcoat$coat, filmcoat[, "temppt"], mean)
  t.low t.med  t.high
43.66667   38 37.66667

> tapply(filmcoat$coat, filmcoat[, "pressure"], mean)
  p.low  p.med  p.high
38.11111 39.11111 42.11111

> tapply(filmcoat$coat, filmcoat[, c("temppt", "pressure")], mean)
  p.low  p.med  p.high
t.low 41.33333 44.33333 45.33333
t.med 35.66667 38.33333 40.00000
t.high 37.33333 34.66667 41.00000

> film.aov1 <- aov(coat ~ temppt*pressure, data=filmcoat)
> summary(film.aov1)
  Df Sum of Sq Mean Sq F Value    Pr(F)
temppt    2  204.6667 102.3333 47.63793 0.00000006
pressure   2    78.0000  39.0000 18.15517 0.00004825
temppt:pressure 4   37.3333  9.3333  4.34483 0.01238203
Residuals 18   38.6667   2.1481
```

These informal visual impressions are formally investigated by examining the simultaneous confidence intervals on the simple effects displayed in Figures 13.7 and 13.8. Control of the simultaneous confidence level at 95% within each of the two sets of nine intervals is maintained by using simulation-generated critical points for this procedure as recommended

## coat: main effects and 2-way interactions

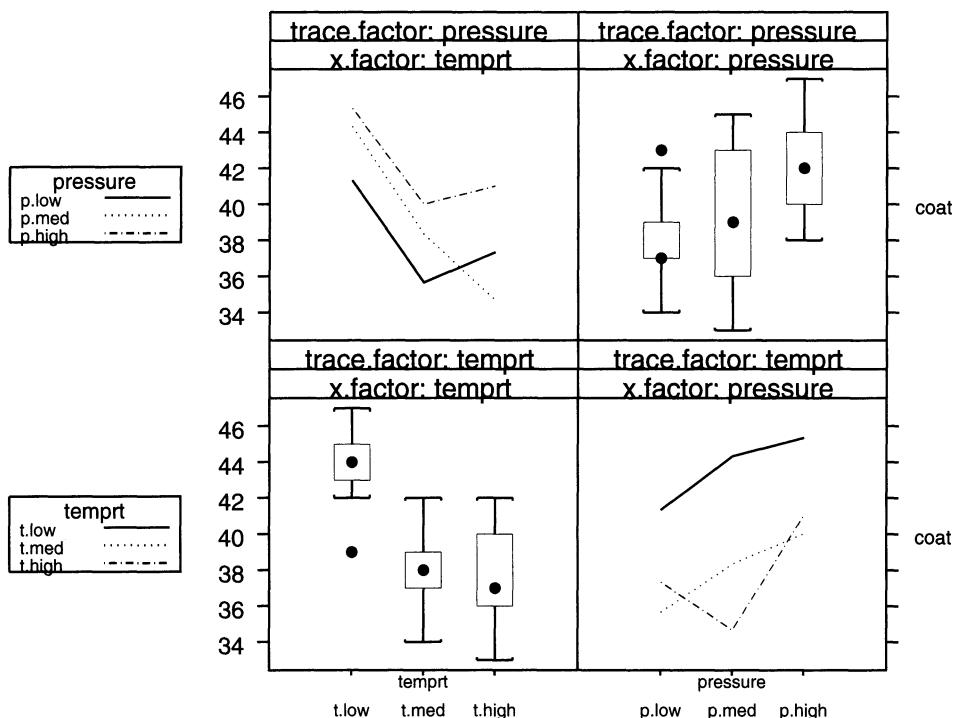


FIGURE 13.6. Main effect and interaction plots for *filmcoat* data. At medium **pressure**, mean **coat** is minimized at high **temp**. No other firm conclusions can be drawn because many of the simultaneous confidence intervals in Figures 13.7 and 13.8 overlap zero.

(dsgn/code/filmcoat.s), (dsgn/figure/filmcoat.int.eps.gz)

by (Edwards and Berry, 1987). These simultaneous confidence intervals are produced in S-PLUS with the command `multicomp`. The simultaneous confidence of the collection of all 18 confidence intervals is closer to 90%.

We examine which of these intervals excludes zero and for those that do, whether the interval lies above or below zero. In Figure 13.7, we see that at medium **pressure**,

- the confidence interval on mean **coat** at low **temp** minus mean **coat** at high **temp** lies entirely above zero,
- the confidence interval on mean **coat** at medium **temp** minus mean **coat** at high **temp** comes very close to lying entirely above zero.

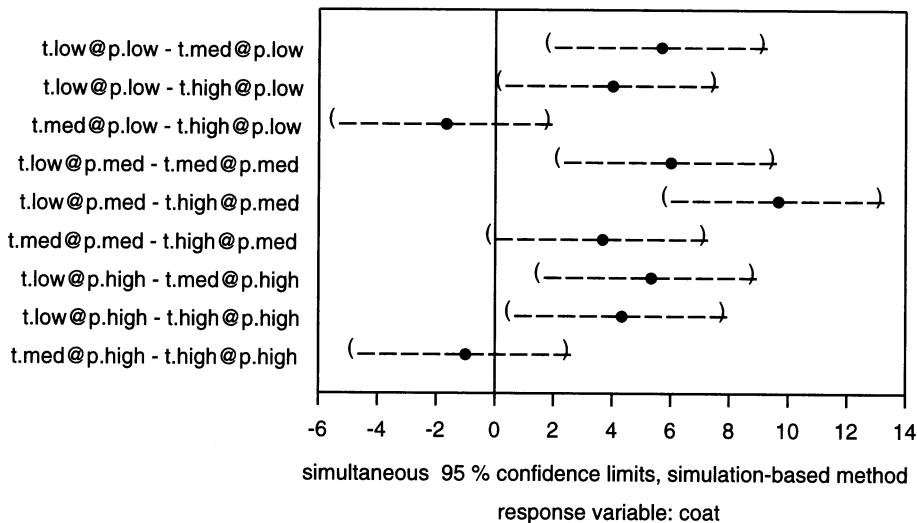


FIGURE 13.7. Simultaneous 95% simulation-based confidence intervals for the `filmcoat` data: Simple effects of temperature at each level of pressure.

(`dsgn/code/filmcoat.s`), (`dsgn/figure/mcout5.eps.gz`)

See also (`dsgn/code/filmcoat2.s`), (`dsgn/figure/filmcoatb.t.p.high.eps.gz`),

(`dsgn/figure/filmcoatb.t.p.med.eps.gz`), (`dsgn/figure/filmcoatb.t.p.low.eps.gz`)

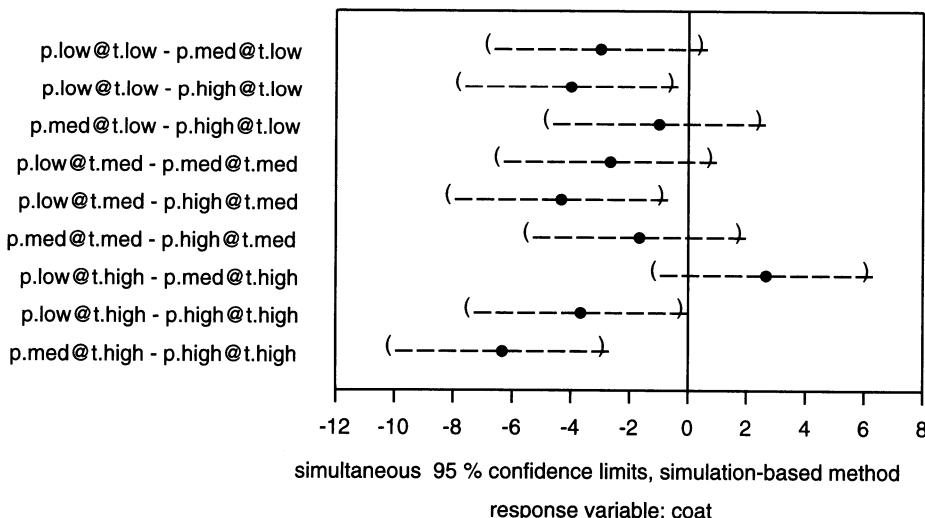


FIGURE 13.8. Simultaneous 95% simulation-based confidence intervals for the `filmcoat` data: Simple effects of pressure at each level of temperature.

(`dsgn/code/filmcoat.s`), (`dsgn/figure/mcout6.eps.gz`)

See also (`dsgn/code/filmcoat2.s`), (`dsgn/figure/filmcoatb.p.t.high.eps.gz`),

(`dsgn/figure/filmcoatb.p.t.med.eps.gz`), (`dsgn/figure/filmcoatb.p.t.low.eps.gz`)

From these two statements we can conclude that at medium **pressure**, mean **coat** is minimized at high **temp**. This is the only firm conclusion we can draw about **coat** minimization because many of the other intervals in these two figures overlap zero, indicating nonsignificant differences of means. We are unable to formally confirm our other graphical impressions that for low and high **pressure**, the response **coat** is minimized at medium **temp**. Nor are we able to confirm our initial graphical impressions about levels of **pressure** that minimize **coat** at each level of **temp**.

In summary, while we can make some confident assertions about differences in coating between some of the combinations of temperature and pressure, it is not possible to infer from these data an overall recommendation of the optimal combination of temperature and pressure. It is possible that a larger experiment would have led to such a conclusion.

## 13.4 Nested Factorial Experiment

Thus far we have considered situations where the relationships among the factors are either completely crossed or completely nested. It is also possible to have an experiment with three or more factors having both crossed and nesting relationships. Such an arrangement is called a nested factorial experiment.

### 13.4.1 Example—Gunload Data

We illustrate one possible arrangement with an example taken from (Hicks, 1964). It was desired to improve the number of rounds per minute that could be fired from a large naval gun. There are two levels of loading **method**, 1=new and 2=old, and three **groups** defining the physiques of the loaders, 1=slight, 2=average, 3=heavy. From each of these groups the experimenter selected three equal sized **teams** of men. Thus there are three teams of men having slight build, three teams of men having average build, and three teams of men having heavy build. Using both of the two methods, each of the nine teams fired the gun on two separate occasions. It is seen that **team** is nested within **group**, and that **method** is crossed with both **group** and **team** within **group**. The factors **method** and **group** are fixed, while **team** is a random factor. The data are contained in the file ([datasets/gunload.dat](#)). We display the data in Figure 13.9.

If all three factors were fixed factors, then the residual mean square would serve as the denominator for all analysis of variance table *F*-tests on main effects and interactions. When at least one factor is random, the *F*-test denominators are sometimes another mean square in the ANOVA table.

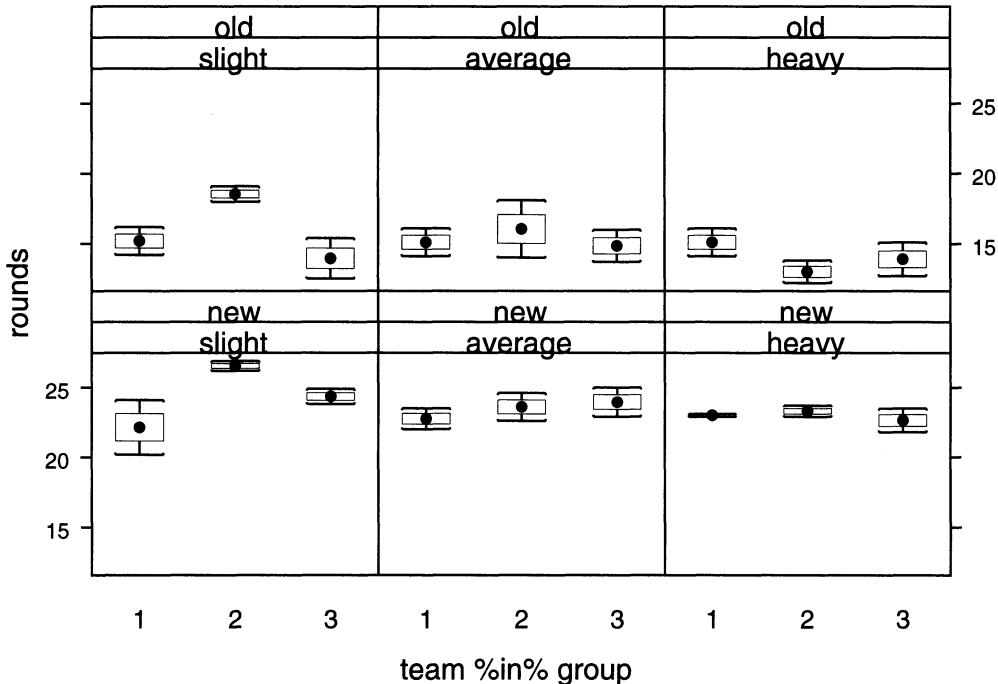


FIGURE 13.9. Boxplot of gunload data. The response **rounds** is higher for the new **method** than the old **method** and does not appear to differ across the three physique **groups**. These findings are consistent with the small *p*-value for **method** and the large *p*-value for **group** in Tables 13.13 and 13.16.  
(`dsgn/code/gunload.s`), (`dsgn/figure/dsgn.gunload.eps.gz`)

As with our use of Table 12.8 to determine the correct denominator for an analysis with two crossed factors where one or both could be random factors, we construct Table 13.11 to aid in our analysis of the **gunload** data. The table is constructed by writing the sums of squares as quadratic forms in the  $Y_{ijkl}$  defined in Equation (13.1), and using Equation (F.6) in Appendix F for finding the expected values of these quadratic forms. Then as with Table 12.8, the ANOVA *F*-test of any effect uses as the denominator the mean square having expectation identical to the expected mean square of the effect, apart from the term for the effect itself.

In Table 13.11, we have three factors, which we call M, G, and T for easy association with **method**, **group**, and **team** in the **gunload** example. We use the corresponding Greek letters  $\theta$  (for *meTHod*),  $\gamma$ , and  $\tau$  for the population effects. Here T is nested within G, and M is crossed with both G and T. In this table, the factors M, G, and T have *m*, *g*, and *t* levels, respectively,

and the common number of replications of each treatment combination is  $n$ . In the `gunload` example, there are  $n = 2$  occasions.

The statistical model associated with this analysis may be written as

$$Y_{ijkl} = \mu + \gamma_i + \tau_{j(i)} + \theta_k + (\gamma\theta)_{ik} + (\tau\theta)_{jk(i)} + \epsilon_{ijkl} \quad (13.1)$$

Here  $\gamma_i$  is the effect of `group` level  $i$ ,  $\tau_{j(i)}$  is the effect of level  $j$  of `team` nested within level  $i$  of `group`,  $\theta_k$  represents level  $k$  of `method`,  $(\gamma\theta)_{ik}$  represents the interaction of `group` and `method`,  $(\tau\theta)_{jk(i)}$  represents the interaction of `team` and `method` within `group` level  $i$ , and  $\epsilon_{ijkl}$  is the residual error.

For ease of presentation, we use the convention that

$$\sigma_A^2 = \frac{\sum_i \alpha_i^2}{a - 1} \quad \text{if A is a fixed factor}$$

and

$$\sigma_{AB}^2 = \frac{\sum_{ij} (\alpha\beta)_{ij}^2}{(a - 1)(b - 1)} \quad \text{if A and B are both fixed factors.}$$

We use the indicator function  $I_A$  defined as 1 if A is a random factor or 0 if A is a fixed factor.

We illustrate the use of Table 13.11 by considering the test of the main effect for `group`, factor G. Since in this example `method` is fixed and `team` is random, we have  $I_M = 0$  and  $I_T = 1$ . Therefore, the expected value of

TABLE 13.11. Expected mean squares for a three-factor nested factorial ANOVA. See also Tables 6.6 and 12.8.

| Source      | df                | Expected Mean Square  |
|-------------|-------------------|---|
| G           | $g - 1$           | $\sigma^2 + nI_M\sigma_{TM}^2 + ntI_M\sigma_{GM}^2 + nmI_T\sigma_T^2 + nmI_G\sigma_G^2$ |
| T within G  | $g(t - 1)$        | $\sigma^2 + nI_M\sigma_{TM}^2 + nmI_T\sigma_T^2$  |
| M           | $m - 1$           | $\sigma^2 + nI_T\sigma_{TM}^2 + ntI_G\sigma_{GM}^2 + ngt\sigma_M^2$                     |
| GM          | $(m - 1)(g - 1)$  | $\sigma^2 + nI_T\sigma_{TM}^2 + nt\sigma_{GM}^2$  |
| TM within G | $g(m - 1)(t - 1)$ | $\sigma^2 + n\sigma_{TM}^2$   |
| Residual    | $mgt(n - 1)$      | $\sigma^2$  |
| Total       | $mgtn - 1$        |   |

the mean square for G is

$$\sigma^2 + mn\sigma_T^2 + nmt\sigma_G^2 \quad (13.2)$$

and the expected value of the mean square for T nested in G is

$$\sigma^2 + nm\sigma_T^2 \quad (13.3)$$

The ratio of these mean squares is appropriate for testing equality of the levels of group, factor G, because the corresponding ratio of these expected mean squares exceeds one if and only if  $\sigma_G^2 > 0$ . Use of the residual mean square as the denominator of the F-test would be inappropriate because such a ratio would exceed one if there is a G effect, a T (team) effect, or both effects.

Note that if instead `method` were a random factor and `team` were a fixed factor, the pattern of expected mean squares would be quite different from those in Table 13.11, with different denominators appropriate for some of the ANOVA F-tests.

The model specifications for the sums of squares in the gunload example are shown for both SAS and S-Plus in Table 13.12. Discussion of the operators in this table appears in Section 13.5 and Table 13.22.

With both SAS and S-Plus we overrode the default choices of the denominator mean squares for the F-tests for `method`, `group`, and `method*group`. These new choices are necessitated by the facts that one of the factors is random and there are both mixing and crossing of factors. Our conclusions here are that after correcting for both loaders' physiques and other person-to-person differences, the two methods have significantly different loading speeds. The new method averaged 23.59 rounds per minute compared with 15.08 rounds per minute for the old method. The analysis also shows a secondary finding that loading times do not differ significantly across physique groups.

TABLE 13.12. Nested factorial model specifications in SAS and S-Plus. Specifications for the sums of squares for the gunload example are shown here. The tests are specified separately. In Table 13.13 we show use of the `Error` function in S-Plus. In Table 13.16 we show the `TEST` statements in SAS.

|         |   |
|---------|---|
| Algebra | $Y_{ijkl} = \mu + \gamma_i + \tau_{j(i)} + \theta_k + (\gamma\theta)_{ik} + (\tau\theta)_{jk(i)} + \epsilon_{ijkl}$ |
| SAS     | $Y = G T(G) M G*M T*M(G)$   |
| S-PLUS  | $Y \sim G + T%in%G + M + G:M + T:M%in%G$  |

TABLE 13.13. Gunload data. S-PLUS display. The *F*-tests that appear without the **Error** function are incorrect. Here we produce correct tests by overriding the default choice of denominators of *F*-tests.  
 (dsgn/code/gunload.s)

---

```
S-PLUS (dsgn/transcript/gunloada.st):
> gunload.aov <- aov(rounds ~ method*group + Error((team %in% group)/method),
+                         data=gunload, qr=T)
> summary(gunload.aov)
Error: team %in% group
      Df Sum of Sq Mean Sq F Value    Pr(F)
group     2 16.05167 8.025833 1.226619 0.3575894
Residuals  6 39.25833 6.543056

Error: method %in% (team %in% group)
      Df Sum of Sq Mean Sq F Value    Pr(F)
method    1 651.9511 651.9511 364.8413 0.0000013
method:group  2   1.1872  0.5936  0.3322 0.7297484
Residuals   6 10.7217  1.7869

Error: Within
      Df Sum of Sq Mean Sq F Value Pr(F)
Residuals 18    41.59  2.310556
```

---

TABLE 13.14. Gunload data. Table of means.  
(dsgn/code/gunload.s)

---

S-PLUS (dsgn/transcript/gunloadb.st):

```
> ## model.tables for models that have an Error() term require a bug fix
> ## in S-Plus 6.1 and earlier. The hh("splus.library") fixes that bug.
> ## Another bug requires you to set se=F for models with an Error() term.
>
> model.tables(gunload.aov, type="means", se=F)

Tables of means
Grand mean

19.333

method
[,1]
new 23.589
old 15.078

group
[,1]
slight 20.125
average 19.383
heavy 18.492

method:group
Dim 1 : method
Dim 2 : group
    slight average heavy
new 24.350 23.433 22.983
old 15.900 15.333 14.000
```

---

TABLE 13.15. Incorrect default SAS analysis of gunload data. The correct analysis appears in Table 13.16. We display this table because it always appears and you must be aware that it is not to be used when an explicit `test` statement is included.

```
SAS (dsgn/code/gunload.sas):
title 'Analysis of Gunload Data';
data gunload;
  infile "&hh/datasets/gunload.dat" firstobs=2;
  input method group team rounds;
run;

proc glm data=gunload;
  classes method group team;
  model rounds = method group method*group
    team(group) method*team(group) /ss3 ;
  test h=group e=team(group);
  test h=method method*group e=method*team(group);
  means method | group ;
run;
```

SAS (dsgn/transcript/gunload.lst):  
Dependent Variable: rounds

| Source          | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|-----------------|----|----------------|-------------|---------|--------|
| Model           | 17 | 719.1700000    | 42.3041176  | 18.31   | <.0001 |
| Error           | 18 | 41.5900000     | 2.3105556   |         |        |
| Corrected Total | 35 | 760.7600000    |             |         |        |

| R-Square | Coeff Var | Root MSE | rounds Mean |
|----------|-----------|----------|-------------|
| 0.945331 | 7.862334  | 1.520051 | 19.33333    |

| Source             | DF | Type III SS | Mean Square | F Value | Pr > F |
|--------------------|----|-------------|-------------|---------|--------|
| method             | 1  | 651.9511111 | 651.9511111 | 282.16  | <.0001 |
| group              | 2  | 16.0516667  | 8.0258333   | 3.47    | 0.0530 |
| method*group       | 2  | 1.1872222   | 0.5936111   | 0.26    | 0.7762 |
| team(group)        | 6  | 39.2583333  | 6.5430556   | 2.83    | 0.0403 |
| method*team(group) | 6  | 10.7216667  | 1.7869444   | 0.77    | 0.6009 |

TABLE 13.16. Correct SAS analysis of gunload data achieved by overriding the default choice of denominators of *F*-tests.

(dsgn/code/gunload.sas)

---

```
test h=group e=team(group);  
test h=method method*group e=method*team(group);
```

---

SAS (dsgn/transcript/gunloada.lst):

Tests of Hypotheses Using the Type III MS for team(group) as an Error Term

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|--------|----|-------------|-------------|---------|--------|
| group  | 2  | 16.05166667 | 8.02583333  | 1.23    | 0.3576 |

Tests of Hypotheses Using the Type III MS  
for method\*team(group) as an Error Term

| Source       | DF | Type III SS | Mean Square | F Value | Pr > F |
|--------------|----|-------------|-------------|---------|--------|
| method       | 1  | 651.9511111 | 651.9511111 | 364.84  | <.0001 |
| method*group | 2  | 1.1872222   | 0.5936111   | 0.33    | 0.7297 |

---

TABLE 13.17. Gunload data. Table of means.  
(dsgn/code/gunload.sas)

---

```
means method | group ;
```

---

| SAS (dsgn/transcript/gunloadb.lst): |    |            |            |
|-------------------------------------|----|------------|------------|
| -----rounds-----                    |    |            |            |
| Level of<br>method                  | N  | Mean       | Std Dev    |
| 1                                   | 18 | 23.5888889 | 1.58481355 |
| 2                                   | 18 | 15.0777778 | 1.97202659 |

| -----rounds-----  |    |            |            |
|-------------------|----|------------|------------|
| Level of<br>group | N  | Mean       | Std Dev    |
| 1                 | 12 | 20.1250000 | 4.96773682 |
| 2                 | 12 | 19.3833333 | 4.45519785 |
| 3                 | 12 | 18.4916667 | 4.81389246 |

| Level of<br>method | Level of<br>group | N | -----rounds----- |            |
|--------------------|-------------------|---|------------------|------------|
|                    |                   |   | Mean             | Std Dev    |
| 1                  | 1                 | 6 | 24.3500000       | 2.35860128 |
| 1                  | 2                 | 6 | 23.4333333       | 1.17075474 |
| 1                  | 3                 | 6 | 22.9833333       | 0.66458007 |
| 2                  | 1                 | 6 | 15.9000000       | 2.42652014 |
| 2                  | 2                 | 6 | 15.3333333       | 1.71191900 |
| 2                  | 3                 | 6 | 14.0000000       | 1.45602198 |

---

### 13.4.2 Example—Turkey Data (continued)

We continue the discussion of the turkey data (`datasets/turkey.dat`) from Section 6.8.

Contrasts of the form used in Table 6.9 are so important in the design of experiments and in their analysis that we have a simple terminology and notation to describe them. In this experiment there are three distinct factors, each at three levels:

`trt.vs.control`: with levels `control` and `treatment`

`additive`: with levels `control`, A, and B

`amount`: with levels 0, 1, and 2

occurring in the pattern illustrated in Table 13.18. The formula describing the model is

$$Y_{mijk} = \mu + \tau_m + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{mijk} = \mu_{ij} + \epsilon_{mijk} \quad (13.4)$$

Several issues are raised here to be discussed. First, how do these factors relate to each other? Second, how does describing a design in terms of the factors specify the analysis?

Factors can be related in several ways (see Table 13.22). In the turkey example, we illustrate two relations: crossing and nesting.

**crossing:** Every level of `additive` appears at every level of `amount`. In this example, Additive A appears at Amounts 1 and 2, as does Additive B.

**nesting:** Some Additive–Amount combinations (A1, A2, B1, B2) appear in only the treatment level of `trt.vs.control`. Other Additive–Amount combinations (control-0) appear in only the control level of `trt.vs.control`. The factors `additive` and `amount` are then said to be *nested* within the factor `trt.vs.control`.

When we add these factors to the dataset, for example with commands in Table 13.19, we can write a much simpler model formula that automatically produces the easily readable ANOVA table in Table 13.20. Notice that the

TABLE 13.18. Factor structure for turkey data.

| Treatment | Level |   |   | Trt.vs.Cont | Treatment | Level |   |   |
|-----------|-------|---|---|-------------|-----------|-------|---|---|
|           | 0     | 1 | 2 |             |           | 0     | 1 | 2 |
| control   | x     |   |   | C           | control   | x     |   |   |
| A         |       | x | x | T           | A         |       | x | x |
| B         | x     | x |   | T           | B         |       | x | x |

TABLE 13.19. S-PLUS commands to create factors for turkey data. The data were read in Section 6.8.

---

```

S-PLUS (dsgn/code/turkey.factors.s):
## follows oway/code/turkey-oway.s

turkey[c(1,7,13,19,25),]

turkey$trt.vs.control <- factor(rep(c("control","treatment"), c(6,24)))
contrasts(turkey$trt.vs.control) <- c(4,-1)

turkey$additive <- factor(rep(c("control","A","B"), c(6,12,12)),
                           levels=c("control","A","B"))
contrasts(turkey$additive) <- c(0,1,-1)

turkey$amount <- factor(rep(c(0,1,2,1,2), c(6,6,6,6,6)))
contrasts(turkey$amount) <- c(0,1,-1)

turkey[c(1,7,13,19,25),]

```

---

TABLE 13.20. ANOVA for turkey data with nested and crossed factors. Interaction is borderline nonsignificant. The main effects of **additive** and **amount** are significant.

(dsgn/code/turkey.aov2.s)

---

```

S-PLUS (dsgn/transcript/turkey.aov2.st):
> turkey3.aov <- aov(wt.gain ~ trt.vs.control / (additive*amount),
+                      data=turkey, x=T)
> summary(turkey3.aov)

```

|                                     | Df | Sum of Sq | Mean Sq  | F Value  | Pr(F)      |
|-------------------------------------|----|-----------|----------|----------|------------|
| trt.vs.control                      | 1  | 56.58133  | 56.58133 | 179.3955 | 0.00000000 |
| additive %in% trt.vs.control        | 1  | 22.81500  | 22.81500 | 72.3367  | 0.00000001 |
| amount %in% trt.vs.control          | 1  | 22.42667  | 22.42667 | 71.1055  | 0.00000001 |
| additive:amount %in% trt.vs.control | 1  | 1.21500   | 1.21500  | 3.8523   | 0.06089888 |
| Residuals                           | 25 | 7.88500   | 0.31540  |          |            |

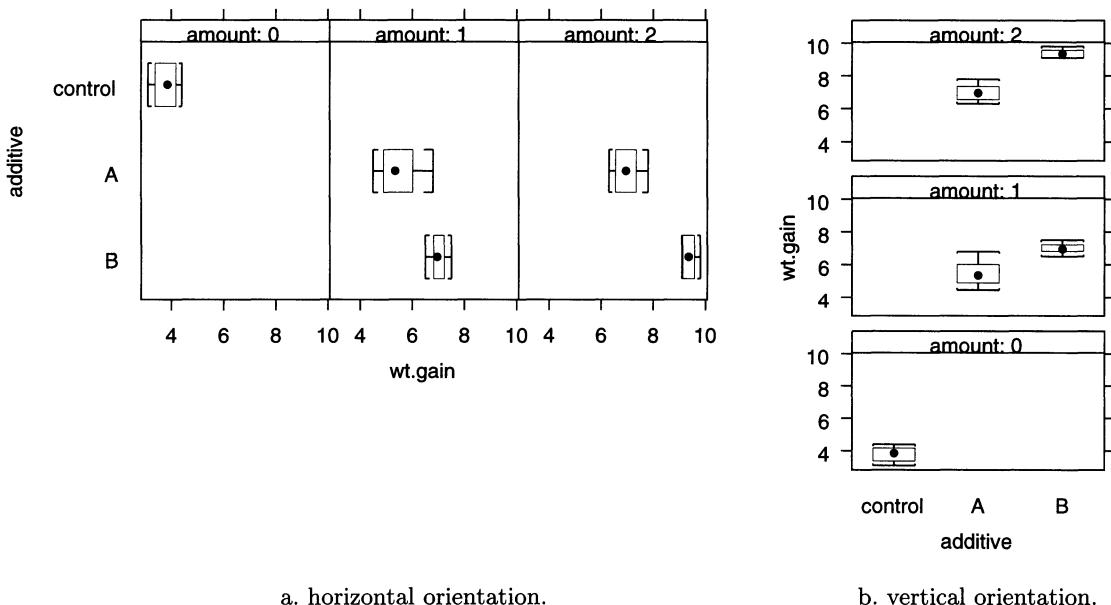
---

four 1-degree-of-freedom sums of squares in Table 13.20 are a decomposition of the 4-degree-of-freedom sum of squares in Table 6.8. The significance of the corresponding *F*-test in Table 6.8 is a rationale for producing and interpreting Table 13.20. We illustrate the structure with the table of means in Table 13.21 and the boxplots in Figure 13.10. See the discussion on orientation of boxplots in Section 13.A.

TABLE 13.21. Means for turkey data.  
(dsgn/code/turkey.means.s)

S-PLUS (dsgn/transcript/turkey.means.st):

```
> tapply(turkey$wt.gain,
+        turkey[,c("additive","amount")],
+        mean)
      0   1   2
control 3.783333 NA  NA
          A     NA 5.5 6.983333
          B     NA 7.0 9.383333
```



a. horizontal orientation.

b. vertical orientation.

FIGURE 13.10. Turkey data with factor structure. Interaction is borderline nonsignificant. The main effects of additive and amount are significant. Panel a shows the boxplots in horizontal orientation and panel b shows the same information in vertical orientation. See the discussion on orientation of boxplots in Section 13.A.

(dsgn/code/turkey.f2.s),

(dsgn/figure/turkey.f2h.eps.gz), (dsgn/figure/turkey.f2v.eps.gz)

Two additional orientations are in

(dsgn/figure/turkey.f2h-v.eps.gz), (dsgn/figure/turkey.f2v-h.eps.gz)

TABLE 13.22. Model specification operators.

|              | S-PLUS |            | SAS    |          | Algebra                 |
|--------------|--------|------------|--------|----------|-------------------------|
|              | abbrev | expanded   | abbrev | expanded |                         |
| Double index | a:b    |            | a*b    |          | $(ab)_{ij}$             |
| Sum          | a+b    |            | a b    |          | $a_i + b_j$             |
| Cross        | a*b    | a+b+a:b    | a b    | a b a*b  | $a_i + b_j + (ab)_{ij}$ |
| Nested       | b%in%a |            | b(a)   | a*b      | $b_{j(i)}$              |
| Nest         | a/b    | a + b%in%a | a b(a) | a a*b    | $a_i + b_{j(i)}$        |

In the turkey example there does not seem to be serious interaction ( $p \approx .06$ ). In other situations the interaction dominates the analysis. An example with prominent interaction is the analysis of the *Rhizobium* clover data in Section 12.13.5.

## 13.5 Specification of Model Formulas

Dummy variables (discussed in Section 10.1), and the contrasts they code for, are so important that all statistical languages have constructs for describing them and the relations between them. The model specification operators in S-PLUS and SAS are detailed in Table 13.22.

Let us explore the meaning of the concepts of crossed and nested factors with a set of simple examples using the data in Table 13.23. The dataset **abc** has two factors, **A** with three levels and **B** with four levels. Files (**dsgn/code/contrasts.s**) and (**dsgn/transcript/contrasts.st**) contain the complete examples. These files show the model formula specifications, the generated dummy variables, the ANOVA tables, and the estimated  $a_i$ ,  $b_j$ ,  $(ab)_{ij}$ , and  $b_{j(i)}$  values. We recommend that you read these files closely and experiment with them using your software.

TABLE 13.23. Sample data used to explore concepts of crossed and nested factors. Factor **A** has three levels and factor **B** has four levels. The interaction **AB** has 12 levels named by crossing the level names of **A** and **B**. The nested factor **BwA** (**B** within **A**) has 12 levels named without reference to factor **A**. (**dsgn/code/contrasts.s**), (**dsgn/transcript/contrasts.st**)

| obs | A | B | AB  | BwA | y     | obs | A | B | AB  | BwA | y     | obs | A | B | AB  | BwA | y     |
|-----|---|---|-----|-----|-------|-----|---|---|-----|-----|-------|-----|---|---|-----|-----|-------|
| 1   | r | w | r.w | c   | 0.17  | 5   | s | w | s.w | g   | -0.24 | 9   | t | w | t.w | k   | 0.34  |
| 2   | r | x | r.x | d   | 2.25  | 6   | s | x | s.x | h   | 1.71  | 10  | t | x | t.x | l   | -0.15 |
| 3   | r | y | r.y | e   | -1.57 | 7   | s | y | s.y | i   | 0.38  | 11  | t | y | t.y | m   | -1.70 |
| 4   | r | z | r.z | f   | -1.55 | 8   | s | z | s.z | j   | -1.26 | 12  | t | z | t.z | n   | -1.93 |

TABLE 13.24. These dummy variables are constructed for a fit of the form  $\hat{y}_i = m + a_i$  to a model of the form  $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ . With treatment contrasts,  $m$  is an estimate of  $\mu + \alpha_r$ ,  $a_s$  is an estimate of  $\alpha_s - \alpha_r$ , and  $a_t$  is an estimate of  $\alpha_t - \alpha_r$ . With sum contrasts,  $m$  is an estimate of  $\mu$ ,  $a_1$  is an estimate of  $\alpha_r$ , and  $a_2$  is an estimate of  $\alpha_s$ . There is no need for an  $a_3$  because of the constraint on the parameters  $\alpha_t = -(\alpha_r + \alpha_s)$ .

(dsgn/code/contrasts.s), (dsgn/transcript/contrasts.st)

| S-PLUS (dsgn/transcript/contrasts-contr.st): |   |   |                                |     |   |    |    |  |  |
|--|---|---|--------------------------------|-----|---|----|----|--|--|
| > model.matrix(~A, data=abc,                 |   |   | > model.matrix(~A, data=abc,   |     |   |    |    |  |  |
| + contrasts=list(A=contr.treatment))         |   |   | + contrasts=list(A=contr.sum)) |     |   |    |    |  |  |
| (Intercept) As At                            |   |   | (Intercept) A1 A2              |     |   |    |    |  |  |
| r.w  | 1 | 0 | 0                              | r.w | 1 | 1  | 0  |  |  |
| r.x  | 1 | 0 | 0                              | r.x | 1 | 1  | 0  |  |  |
| r.y  | 1 | 0 | 0                              | r.y | 1 | 1  | 0  |  |  |
| r.z  | 1 | 0 | 0                              | r.z | 1 | 1  | 0  |  |  |
| s.w  | 1 | 1 | 0                              | s.w | 1 | 0  | 1  |  |  |
| s.x  | 1 | 1 | 0                              | s.x | 1 | 0  | 1  |  |  |
| s.y  | 1 | 1 | 0                              | s.y | 1 | 0  | 1  |  |  |
| s.z  | 1 | 1 | 0                              | s.z | 1 | 0  | 1  |  |  |
| t.w  | 1 | 0 | 1                              | t.w | 1 | -1 | -1 |  |  |
| t.x  | 1 | 0 | 1                              | t.x | 1 | -1 | -1 |  |  |
| t.y  | 1 | 0 | 1                              | t.y | 1 | -1 | -1 |  |  |
| t.z  | 1 | 0 | 1                              | t.z | 1 | -1 | -1 |  |  |

We provide a detailed discussion here of just one example, the crossing of two factors. The simplest set of dummy variables (that is, easiest to understand) is the set of treatment contrasts. The most frequently used is the set of sum contrasts. We show both in Table 13.24.

The dummy variables for interaction in the crossing model

$$\text{Algebra: } y_i = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ij}$$

$$\text{SAS: } Y = A \quad B \quad A*B$$

$$\text{S-PLUS: } Y \sim A + B + A:B$$

are constructed as the outer product of the rows of the dummy variables for each of the main effects. We illustrate in Table 13.25 with the sum contrasts.

TABLE 13.25. Dummy variables for the interaction  $(\alpha\beta)_{ij}$  constructed as the outer product of the rows of the dummy variables for the two main effects A and B. We continue with the data of Table 13.23 and the sum contrasts defined in the right side of Table 13.24. For example, in row `rz` the value in column `A1B1` is the product of the 1 in column `A1` and the  $-1$  in column `B1`. There are two degrees of freedom, hence two dummy variables for the A effect. There are three dummy variables for the B effect. Therefore, there are  $2 \times 3$  dummy variables for the A:B interaction.

(dsgn/code/contrasts.s), (dsgn/transcript/contrasts.st)

| S-PLUS (dsgn/transcript/contrasts-AB.st): |             |    |    |    |    |    |      |      |      |      |      |      |
|---|-------------|----|----|----|----|----|------|------|------|------|------|------|
|   | (Intercept) | A1 | A2 | B1 | B2 | B3 | A1B1 | A2B1 | A1B2 | A2B2 | A1B3 | A2B3 |
| r.w                                       | 1           | 1  | 0  | 1  | 0  | 0  | 1    | 0    | 0    | 0    | 0    | 0    |
| r.x                                       | 1           | 1  | 0  | 0  | 1  | 0  | 0    | 0    | 1    | 0    | 0    | 0    |
| r.y                                       | 1           | 1  | 0  | 0  | 0  | 1  | 0    | 0    | 0    | 0    | 1    | 0    |
| r.z                                       | 1           | 1  | 0  | -1 | -1 | -1 | -1   | 0    | -1   | 0    | -1   | 0    |
| s.w                                       | 1           | 0  | 1  | 1  | 0  | 0  | 0    | 1    | 0    | 0    | 0    | 0    |
| s.x                                       | 1           | 0  | 1  | 0  | 1  | 0  | 0    | 0    | 0    | 1    | 0    | 0    |
| s.y                                       | 1           | 0  | 1  | 0  | 0  | 1  | 0    | 0    | 0    | 0    | 0    | 1    |
| s.z                                       | 1           | 0  | 1  | -1 | -1 | -1 | 0    | -1   | 0    | -1   | 0    | -1   |
| t.w                                       | 1           | -1 | -1 | 1  | 0  | 0  | -1   | -1   | 0    | 0    | 0    | 0    |
| t.x                                       | 1           | -1 | -1 | 0  | 1  | 0  | 0    | 0    | -1   | -1   | 0    | 0    |
| t.y                                       | 1           | -1 | -1 | 0  | 0  | 1  | 0    | 0    | 0    | 0    | -1   | -1   |
| t.z                                       | 1           | -1 | -1 | -1 | -1 | -1 | 1    | 1    | 1    | 1    | 1    | 1    |

### Example—Dummy Variables for Crossed Factors Nested Within Another Factor

A model formula specifies a set of dummy variables. Just as in one-way analysis of variance, we control the structure of the dummy variables with the contrast matrix assigned to each factor. Let us look at the dummy variables generated for us by the model formula

```
wt.gain ~ trt.vs.control / (additive*amount)
```

We do so in S-PLUS by adding the argument `x=T` to the `aov` statement in Table 13.20 and then displaying the `x` component of the resulting `aov` object in Table 13.26.

There are some complications in the display in Table 13.26. The generation of the `x` matrix of dummy variables doesn't know about the actual degrees of freedom for each effect. It assumes the maximum possible if all implied cells were observed (in this example there are  $2 \times 3 \times 3 = 18$  cells implied by the complete crossing of `trt.vs.control`, `additive`, and `amount`). Only five of those implied cells actually have observations. The `match` and the

TABLE 13.26. Regression coefficients and dummy variables for turkey data.  
 (dsgn/code/turkey.aov3.s)

```
S-PLUS (dsgn/transcript/turkey.aov3.st):
> match(dimnames(coef(summary.lm(turkey3.aov)))[[1]],
+       dimnames(turkey3.aov$x)[[2]])
[1] 1 2 4 8 12
>
> tmp1 <- coef(summary.lm(turkey3.aov))
> dimnames(tmp1)[[1]][3:5]
[1] "trt.vs.controltreatmentadditive1"
[2] "trt.vs.controltreatmentamount1"
[3] "trt.vs.controltreatmentadditive1amount1"
> dimnames(tmp1)[[1]][3:5] <- c("additive", "amount", "additive:amount")
> dimnames(tmp1)[[1]][3:5]
[1] "additive"           "amount"            "additive:amount"
>
> tmp2 <- turkey3.aov$x[,c(1,2,4,8,12)]
> dimnames(tmp2)[[2]][3:5] <- c("additive", "amount", "additive:amount")
>
> round(tmp1, 4)
      Value Std. Error t value Pr(>|t|)
(Intercept) 6.5300    0.1025 63.6859  0.0000
trt.vs.control -0.6867   0.0513 -13.3939  0.0000
  additive -0.9750   0.1146 -8.5051  0.0000
  amount -0.9667   0.1146 -8.4324  0.0000
additive:amount  0.2250   0.1146  1.9627  0.0609
> tmp2[c(1,7,13,19,25),]
  (Intercept) trt.vs.control additive amount additive:amount
  1             1                 4        0       0          0
  7             1                -1        1       1          1
 13            1                -1        1      -1         -1
 19            1                -1       -1        1         -1
 25            1                -1       -1      -1          1
```

relabeling are used to find just the ones that matter in this example and to give them more reasonable names. See Exercise 13.10 for guidance on discovering how the `match` function is used.

The predicted value for an observation  $i$  is calculated, as with any linear model, as the inner product of the regression coefficients with the dummy variables in row  $i$ . In this example, we predict the weight gain

for observation 7 as

$$\hat{y}_7 = (1 \quad -1 \quad 1 \quad 1 \quad 1) \begin{pmatrix} 6.5300 \\ -0.6867 \\ -0.9750 \\ -0.9667 \\ 0.2250 \end{pmatrix} = 5.5$$

## 13.6 Sequential and Conditional Tests

When there are two or more predictors in a model, they are usually not orthogonal to each other. Therefore, the interpretation given to the relative importance of each predictor depends on the order in which they enter the model. One of the important goals of designed experiments is the choice of combinations of levels for factors that will make the dummy variables for each factor or interaction orthogonal to the others. Most of the examples in this book in Chapters 12, 13, and 14 have orthogonal effects.

When the data for an example have continuous predictors or covariates, or are classified by factors with unequal numbers of observations per cell, the effects are usually not orthogonal. Most of the examples in Chapters 9, 10, and 11 have continuous predictor variables and therefore do not have orthogonal effects.

When effects are not orthogonal, the sequence in which they are entered into the model affects the interpretation of the effects. See Sections 9.7 (Partial *F*-Tests) and 9.14 (Residual Plots) for techniques used to investigate the relative importance of the predictors.

The sequential ANOVA table depends on the order in which the effects are entered into the model. Each row of the table is calculated under the assumption that all effects in higher rows have already been included and that all effects in lower rows have not. S-PLUS normally prints the sequential ANOVA table. SAS calls the sequential ANOVA table the table of Type I sums of squares.

There are several types of conditional ANOVA tables. One of the most frequently used is Yates' weighted squares of mean, what SAS calls Type III sums of squares, in which each row of the table is calculated under the assumption that all other rows—both higher and lower—have already been included.

Another method, Yates' Method of Fitting Constants, what SAS calls Type II sums of squares, makes different assumptions for each class of effect. ANOVA rows for main effects assume all other main effects are already

included in the model. ANOVA rows for two-way interactions assume all main effects and other two-way interactions are in the model. Higher-order interactions assume all lower-order effects and interactions are already in the model.

### 13.6.1 SAS Types of Sums of Squares

In Section 9.5.1 we mention that SAS uses various *types* of partitionings of the total sum of squares in presenting analysis of variance tables. The terminologies Type I, Type II, and Type III sum of squares originated by SAS have become so widely known that they are used nowadays even outside the context of interpreting SAS listing files. In order that readers be able to request and interpret SAS analysis of variance presentations, we provide here more detail on these types in the context of designed experiments having two factors.

Suppose the response is  $Y$ , the two factors are  $A$  and  $B$ , and the SAS model statement reads

Model  $Y = A \ B \ A*B$

Since no particular types of sums of squares were requested, SAS provides by default Types I and III. If the user wishes to override the default, particular types can be requested as illustrated here:

Model  $Y = A \ B \ A*B /ss1 \ ss2$

The Type I sum of squares for each effect is the portion of model sum of squares attributable to that effect above and beyond what is attributable to all effects listed prior to it in the expanded model statement. Thus in the illustration, the Type I sum of squares for  $B$  is the marginal contribution of factor  $B$  assuming that factor  $A$  is already in the model. Use of this sum of squares is appropriate if a model containing factor  $A$  without factor  $B$  makes sense, but a model containing factor  $B$  makes no sense unless the model already includes factor  $A$ .

For each main effect in the model statement, the Type II sum of squares is the marginal contribution of that effect beyond the sum of squares attributable to all other main effects in the model statement. The Type II sum of squares for  $A*B$  is the portion of model sum of squares attributable to this interaction after the main effects  $A$  and  $B$  are already in the model. (Yates, 1934) gave the name *method of fitting constants* to what is now called Type II sums of squares. In the absence of interaction, this method produces the maximum power tests for the main effects.

Note that while the Type I sums of squares for  $A$ ,  $B$  and  $A*B$  add up to the model sum of squares, the Type II sums of squares for these three

effects are not in general an orthogonal partitioning of the model sum of squares and hence do not in general sum to the model sum of squares. An exception occurs when the data are balanced (each of the  $ab$  cells contain the same number of observations); in this case the Type I and Type II sums of squares coincide.

Type III sums of squares can be used with the above model statement provided that each of the  $ab$  cells contains at least one observation. If the sampling is balanced, Type III coincides with both Type I and Type II. Otherwise, the Type III sum of squares for any effect, including  $A^*B$ , is adjusted for *all* other effects in the model. The Type III partitioning provides what is known as the Yates' weighted squares of means analysis of unbalanced data; see (Searle, 1971).

Type IV sums of squares coincide with Type III sums of squares when all cells contain observations. This partitioning is used when some cells are empty, a situation we do not pursue in this text. The Type IV sum of squares partitioning is not unique, a feature that makes many analysts uncomfortable with their use.

The nomenclature Type I and Type III (as well as Type II and Type IV) was originated by SAS in (Goodnight, 1978) and summarized in (SAS Institute, Inc., 1999).

### 13.6.2 Example—Application to Body Fat Data

We revisit in Table 13.27 the analysis begun in Section 9.3 of a portion of the body fat data (`datasets/fat.data`) using the two predictors `abdomin` and `biceps` of the response `bodyfat`.

The  $F$ -value 54.92 applies to the composite hypothesis  $H_0: \beta_1 = \beta_2 = 0$  against the alternative that at least one  $\beta_i$  is nonzero, where the  $\beta_i$  are the coefficients of the two predictors. The corresponding small  $p$ -value indicates that either `abdomin` or `biceps` or both are linearly related to `bodyfat`. The  $R^2 = 0.714$  tells us that 71.4% of the variability in these subjects' `bodyfat` is accounted for by their `abdomin` and `biceps` measurements. The remaining 28.6% of `bodyfat` variability is explained by other measurable variables not presently in the model as well as the random error component of the model in Equation (9.1).

The SAS Type I sums of squares are sequential in that the sum of squares 2440.5 for the first listed predictor, `abdomin`, is calculated assuming that this is the only predictor in the model, while the sum of squares 209.32 for the second listed predictor, `biceps`, is calculated assuming that the first listed predictor is already in the model. In general, the top to bottom

TABLE 13.27. SAS display for two- $X$  regression of `bodyfat`. The “Type I SS” corresponds to the display in Table 9.1. The Type I sums of squares are an orthogonal partitioning of the model sum of squares. The sum of the two Type III sums of squares is unequal to the model sum of squares.

---

```
SAS (regb/code/fat.sas):
proc glm data = fat;
    model bodyfat = abdomen biceps;
run;
```

---

```
SAS (regb/transcript/fat.lst):
The GLM Procedure
Dependent Variable: bodyfat

   Sum of
Source      DF      Squares      Mean Square      F Value      Pr > F
Model        2      2649.816730     1324.908365      54.92      <.0001
Error       44      1061.382419      24.122328
Corrected Total  46      3711.199149

   R-Square      Coeff Var      Root MSE      bodyfat Mean
          0.714006      26.69883      4.911449      18.39574

   Source      DF      Type I SS      Mean Square      F Value      Pr > F
abdomin      1      2440.499986     2440.499986      101.17      <.0001
biceps      1      209.316744      209.316744       8.68      0.0051

   Source      DF      Type III SS      Mean Square      F Value      Pr > F
abdomin      1      1823.019877     1823.019877      75.57      <.0001
biceps      1      209.316744      209.316744       8.68      0.0051

   Standard
Parameter      Estimate      Error      t Value      Pr > |t|
Intercept     -14.59373632     6.69222199      -2.18      0.0346
abdomin       0.68293791      0.07855885       8.69      <.0001
biceps      -0.92215436      0.31304822      -2.95      0.0051
```

---

ordering of sources of variation in the Type I sum of squares table is the same as the ordering of these sources in the `Model` statement.

Each predictor's SAS Type III sum of squares is calculated assuming that all other predictors are already in the model. Thus the Type III sum of squares for `abdomin`, 1823.02, is *conditional* in the sense that it is calculated under the assumption that the model already contains the other predictor `biceps`. In general, any entry in a Type III sum of squares table is conditioned on the existence in the model of all sources above it in this table.

The parameter estimates in both Table 9.1 and Table 13.27 are based on this same “last-in” rule, corresponding to the Type III sums of squares. The Type III  $F$ -value for `abdomin`, 75.57, is the square of the  $t$ -value for `abdomin`, 8.69, in the `Parameter` section of Table 13.27 and in Table 9.1.

In this context, it is preferable to work with the Type I analysis if the investigator believes that a model containing `biceps` makes no sense unless `abdomin` is already in the model. Otherwise, the Type III approach is preferred. In this example, each predictor has a statistically significant impact on `bodyfat` after the other predictor has already been included in the model. In general, it is possible for one predictor to have an insignificant additional impact on the response when other more prominent predictors are already in the model. See Section 9.12 on collinearity for a discussion of this issue.

## 13.7 Exercises

- 13.1. Consider an experiment to determine which of four types of `valve` used in an artificial heart maximizes blood pressure control as measured by maximum `flow` gradient (mm Hg). Flow was maintained at each of the same six `pulse` rates for each valve type. Two `runs` were made for each valve type. The order of the eight runs at the four valve types was randomized. Note that `run` is a random factor, nested within `valve`. The data, contained in (`datasets/heartvalve.dat`), come from (Anderson and McLean, 1974). Perform a thorough analysis including plots of the data.
- 13.2. An experiment reported in (Lewin and Shakun, 1976) investigated whether an Octel filter (`type=1`) or a standard filter (`type=2`) provided superior suppression of `noise` produced by automobile exhaust systems. The experiment considered three vehicle `sizes` coded 1 small, 2 medium, 3 large; and both the right 1 and left 2 `side` of cars. The data appear in the file (`datasets/filter.dat`). Perform a thorough

analysis leading to a recommendation of which filter to use under the various experimental conditions.

- 13.3. An experiment explored the abilities of six commercial laboratories to accurately measure the percentage **fat** content in samples of powdered eggs. A pair of samples from a single can was sent to each lab. The labs were told that the samples were of two **types**, but in fact they were from the same can. Each lab assigned two **technicians** to analyze each type. The data, from (Bliss, 1967), are in the file (**datasets/eggs.dat**). Analyze the data in order to recommend which lab(s) have superior or inferior abilities to ascertain the fat content of powdered eggs.
- 13.4. (Box and Cox, 1964), reprinted in (Hand et al., 1994) reported the results of a  $3^3$  factorial experiment. The data are in (**datasets/wool.dat**). The response is the **cycles** under tension to failure of worsted yarn. The three factors are **length** of test specimen (250, 300, 350 mm), **amplitude** of loading cycle (8, 9, 10 mm), and **load** (40, 45, 50 g). The levels of all three factors are coded as -1, 0, 1 in the data file. The authors recommend a preliminary log transformation of the response. Perform an analysis to determine the influences of the factors on the response.
- 13.5. A  $5 \times 3 \times 4$  factorial experiment is designed to compare the wear resistance of vulcanized rubber (Davies, 1954) (p. 192). The three factors are

**filler:** 5 qualities

**pretreatment:** 3 methods

**raw:** 4 qualities

There is only one replicate; thus the assumption must be made that the three-factor interaction is negligible and the three-factor sum of squares can be used for the error term.

The data are in the file (**datasets/vulcan.dat**). The S-PLUS code is in the file (**dsgn/code/vulcan.s**) and the S-PLUS transcript is in the file (**dsgn/transcript/vulcan.st**). The graph of the main effects and two-way interactions are in Figure 13.11.

- a. Determine from the ANOVA table whether any of the main effects or two-way interactions are significant.
- b. Why can't we test the three-way interaction?
- c. From the figures and tables of means, determine if any levels of any factors can be eliminated from further consideration. Assume that we are looking for big numbers for the best wear resistance.

### wear: main effects and 2-way interactions

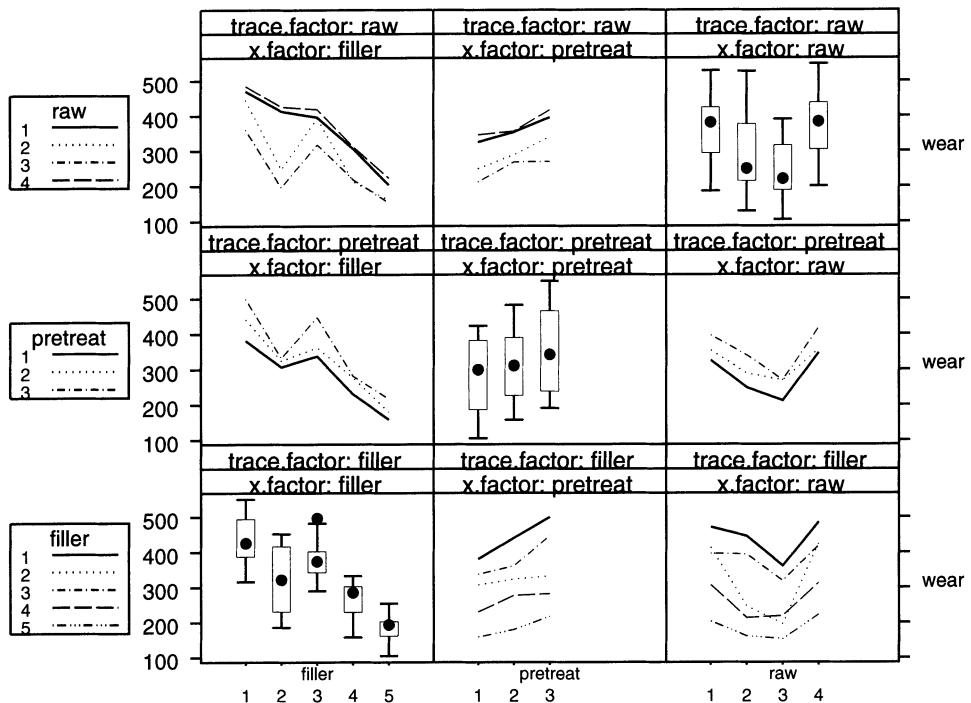


FIGURE 13.11. Main effects and two-way interactions for wear resistance of vulcanized rubber.  
 (dsgn/code/vulcan.s), (dsgn/figure/interaction.trellis.eps.gz),  
 (dsgn/figure/interaction.trellis.color.eps.gz)

- d. Of the treatment combinations that are left for consideration, are any clearly dominant? Would we need to make a conditional recommendation to the client?
- 13.6.** Continue the Latin square example using the file (`datasets/tires.dat`) in Section 13.2. The treatment sum of squares with 3 degrees of freedom is linearly dependent on the Row×Column interaction with  $(r - 1) \times (c - 1)$  degrees of freedom. Demonstrate the dependency by showing that each of the dummy variables for brand has a zero residual when regressed on the dummy variables for car\*position.
- You may use the file (`dsgn/code/tires.latin.s`) as a starting point. Explain why the residual sum of squares in `tr1.lm` (and the analogous `tr2.lm` and `tr3.lm`) is 0.

**13.7.** (Peterson, 1985) discusses an experiment to assess the effects on strengths of spot welds ( $\text{psi} \times 10^{-3}$ ) created by robots on automobile assembly lines. On each of two assembly lines (blocks) there were three fixed treatment factors: **maker** at two levels, rod **diameters** at three levels (30 mm, 60 mm, 90 mm), and **chromium** content at three levels (1%, 1.5%, 2%). The 18 treatment combinations were randomly assigned to 18 robots on each assembly line. The data appear in the file (**datasets/weld.dat**). Analyze the data, including

- a. a discussion of interaction among the treatment factors.
- b. a recommendation of the combination of the treatment factors for maximizing **strength**. Explain how you know that your recommendation for **diameter** is distinctly better than the next-best choice of **diameter**.

**13.8.** (Peterson, 1985) describes an investigation to compare the abilities of seven washday **products** to remove dirt in cloth:

- liquid detergent
- granular detergent
- detergent flakes
- liquid detergent plus phosphate
- granular detergent plus phosphate
- detergent flakes with phosphate
- soap

Each of these seven **products** was assigned to three bedsheets soiled in a standard way, and the amount of **dirt** removed (mg) from each bedsheet was recorded. The data are in the file (**datasets/washday.dat**). Analyze these data.

- a. Perform a one-way analysis of variance to assess whether the mean amount of **dirt** removed is the same for all seven **products**.
- b. Partition the 6 degrees of freedom sum of squares for **product** into 6 mutually orthogonal 1 degree-of-freedom sums of squares, each of which has an interpretation based on the similarities and differences among the **products**.
- c. Estimate each of the six corresponding contrasts.
- d. The six levels (hence five contrasts) within detergent can be specified as the crossing of three levels of form (liquid, granular, flakes) and

two levels of ingredient (none, phosphate). Rewrite the model as a crossing of form and phosphate nested within soap.vs.detergent.

- e. Assuming that the costs per wash are roughly the same for all seven products, provide recommendations for consumers.

**13.9.** (Neter et al., 1996) describe an experiment to compare the work of market research firms. The data file is (`datasets/market.dat`). It was desired to evaluate the effects on quality of work performed by 48 firms of the factors of the three crossed factors fee level (`feelevel`), scope, and supervision. Fee level has three levels (1 = high, 2 = average, 3 = low), scope has two levels (1 = all performed in-house; 2 = some contracted out), and supervision has two levels (1 = local supervisors, 2 = traveling supervisors). Construct an analysis of variance table. Produce and interpret interaction plots for any interaction found significant in the table. Compare the means of the levels of any factors not involved in a significant interaction.

**13.10.** In Table 13.26 we use the S-PLUS `match` function to identify which of the implied dummy variables in a nested design are actually used. The complete command using the `match` function is

TABLE 13.28. Isolated code fragments to be run one line at a time to help learn what the complete statement is doing. See Exercise 13.10 for more detail.

(`dsgn/code/turkey.aov3.s`)

S-PLUS (`dsgn/code/turkey.aov3-match.s`):  
## follows `dsgn/code/turkey.aov3.s`

```

summary.lm(turkey3.aov)
coef(summary.lm(turkey3.aov))
dimnames(coef(summary.lm(turkey3.aov)))
dimnames(coef(summary.lm(turkey3.aov)))[[1]]

turkey3.aov$x
dimnames(turkey3.aov$x)
dimnames(turkey3.aov$x)[[2]]

match(dimnames(coef(summary.lm(turkey3.aov)))[[1]],
      dimnames(turkey3.aov$x)[[2]])

```

```
match(dimnames(coef(summary.lm(turkey3.aov))))[[1]],  
      dimnames(turkey3.aov$x)[[2]])
```

Study the command by picking up pieces of it and dropping them into the Commands window. For example, assuming you have already defined all the variables by running the S files leading up to (*dsgn/code/turkey.aov3.s*), open file (*dsgn/code/turkey.aov3.s*) in your editor and highlight and run the pieces of code corresponding to the lines in Table 13.28.

- 13.11. It is desired to compare a response variable **dimvar**, dimensional variability, of a component produced by each of three **machines**. Each machine is comprised of two **spindles**, and four components are selected at random from each spindle. This example is attributable to (Montgomery, 2001), and the data file is (*datasets/spindle.dat*). Perform an analysis to determine the effects of **spindle** and **machine** on **dimvar**, assuming that both factors are fixed.
- 13.12. In an experiment reported by (Montgomery, 2001) having the data file (*datasets/surface.dat*), the response variable is a measure of **surface** finish of a metal part. Each part is produced by one of four **machines**, a fixed factor. Three **operators** are assigned to produce parts on each **machine**. The operators are selected at random and a total of 12 different operators are chosen for the 4 machines. Analyze the data to determine the effects on **surface** of **machine** and **operator**.

## 13.A Appendix: Orientation for Boxplots

We display the boxplots for the turkey data in two orientations in Figure 13.10. The horizontal orientation takes less vertical space on the page. The vertical orientation places the response variable in the vertical direction and accords with how we have been trained to think of functions—levels of the independent variable along the abscissa and the response variable along the ordinate. Most of the graphs in this book are oriented with the response variable in the vertical direction. With Figure 13.10 we couldn't decide which orientation we prefer and therefore elected to present both along with this discussion. We give code for two additional orientations in the file (`dsgn/code/turkey.f2.s`).

We chose to display Figures 12.6, 12.7, and 12.8 in vertical orientation. We investigated three other options and make available programs to generate them, varying the horizontal and vertical orientation and varying the conditioning variable. One of our concerns was legibility of the labels when there are too many long labels on the abscissa.

# Design of Experiments—Complex Designs

In this chapter we introduce some additional topics in experimental design beyond those discussed in Chapters 6, 12, and 13. The principle of confounding is used to design efficient experiments having many factors but using only a small subset of all possible treatment combinations. Split plot designs involve placing a restriction on the randomization of treatments to experimental units in order to achieve more precision for comparisons involving levels of one factor in exchange for reduced precision for comparisons involving levels of another factor. We illustrate crossover designs that allow for the estimation of treatment effects that can linger across time periods. We show how to test for interaction in two-way designs having exactly one observation at each treatment combination.

## 14.1 Confounding

In order to understand the following sections on fractional factorial designs and split plot designs, one must become familiar with the concept of *confounding* of factors.

Two effects (main effects or interactions) are said to be *confounded* if they cannot be independently estimated. (English language equivalents of the statistics term *confounded* include intermixed, intertwined, and confused.) Two completely confounded effects are said to be *aliases* of one another. Each such effect is referred to as an alias of the other effect or is said to be aliased with the other effect. The whole plot effect (indexing on the physical location of the plot) and the treatment effect (indexing on the level of the

treatment assigned to the whole plot) in a split plot design (see Sections 14.2 and 14.3) are completely confounded. Effects can also be partially confounded. See a design of experiments text ((Cochran and Cox, 1957), for example) for a more complete discussion of confounding.

If the analyst must be able to estimate separately the effects of both of the two factors or interactions, it is essential that these factors not be confounded. On the other hand, if the effects of some interactions can be *assumed to be negligible*, an effective design strategy may be to *purposely* confound such negligible interactions with nonnegligible factors or interactions. By doing so, the analyst strives to be able to estimate all effects of interest with a much smaller experiment than would be required without using confounding. The fractional factorial designs in Section 14.4 illustrate confounding of interactions with blocks.

To illustrate the importance of avoiding the confounding of nonnegligible factors, let's return to the turkey data (`datasets/turkey.dat`) analysis in Section 13.4.2. In that experiment there were five groups of six turkeys per group. The turkeys in each group were fed one of five diets, a control diet and four experimental diets A1 A2 B1 B2. The naming convention for the experimental diets refers to four combinations of the two factors **additive**, with levels A or B, and **amount**, with levels 1 or 2. Both of these factors were a priori believed likely to impact on the response, `wt.gain`. Suppose instead that a novice investigator without training in statistics fed 12 turkeys diet A1 and 12 other turkeys diet B2. If the novice then subtracts the mean weight gain on A1 from the mean weight gain on B2, there is no way to tell whether the result is attributable to the difference between amounts 2 and 1, the difference between diets B and A, or some combination of these two factors. In this poorly designed experiment, **additive** and **amount** are confounded. With the correctly designed experiment, it is possible to separately estimate the effects of **additive** and **amount**, as well as the interaction between these factors.

As another illustration of confounding, consider an experiment involving three factors  $A, B, C$  each having two levels, where blocks of homogeneous experimental units are of size at most 4, so that we can examine just four treatment combinations (t.c.'s) in any block. Also suppose that we are able to assume that all interactions are negligible, i.e., the response is additive with respect to the three factors.

A word on notation. For each factor we arbitrarily designate one of the levels as the upper, or 1 level, and the other level as the lower, or 0 level. A t.c. may be written by listing the lowercase letters of all factors observed at their 1 level. Thus if there are four factors  $A, B, C, D$ , the t.c.  $bd$  is that with factors  $A$  and  $C$  at their lower levels and factors  $B$  and  $D$  at their

upper levels. The t.c. where all factors are at their lower level is denoted (1).

Returning to our three-factor experiment, suppose we run just the four t.c.'s  $a$ ,  $b$ ,  $c$ , and  $abc$  in any block. Then we can estimate the  $A$  main effect as  $\frac{1}{2}(abc + a - b - c)$ , and the other two main effects similarly. In this setup, we say that factor  $A$  is confounded with the  $BC$  interaction because  $BC$  would be estimated with this same estimate. Since we are assuming that this interaction is negligible, we are estimating only the  $A$  main effect. Similarly, the estimate of the  $B$  main effect would be confounded with the negligible  $AC$  interaction.

In this way we can estimate all main effects with just four observations. However, no degrees of freedom are available to estimate the error needed to produce confidence intervals and conduct tests. This can be handled by replicating runs of the above four t.c.'s or their mirror image  $ab$ ,  $bc$ ,  $ac$ , and (1), in additional blocks. The set of four t.c.'s run in each block is called a *fractional replicate* of the set of all possible treatment combinations. In Section 14.4 we study fractional factorial designs, where an entire experiment consists of one large fractional replicate, usually arranged in blocks consisting of smaller fractional replicates.

In more complicated designs it is common for main effects and interactions to have several aliases. A design may be described by providing an equation that specifies its aliasing structure.

## 14.2 Split Plot Designs

This design involves placing a *restriction* on the randomization of treatments to experimental units. Sometimes it is easiest to administer an experiment by applying one treatment factor to groups of experimental units, called *plots*, and another treatment factor to the individual experimental units, referred to as *subplots*. Designs with such a restriction on the randomization are called *split plot designs*. This design strategy is especially useful if the experimenter wants to gain greater precision for inferences involving the treatment applied to subplots, the subplot treatment, at the expense of lower precision for inferences on the treatment applied to whole plots, the whole plot treatment. We confine our attention to the simple case of one blocking factor and two fixed treatment factors. However, the principles behind this restricted randomization approach can be applied to other design types such as ones without blocks, or Latin squares, factors that are random rather than fixed, and situations involving more than two treatment factors. The terminology *split plot* refers to the possibility of fur-

ther splitting the plot into sub-subplots within subplots to accommodate three or more treatment factors.

We model this experiment as follows, where  $Y_{ijk}$  is the yield of the observation in block  $i$  receiving level  $j$  of fixed treatment A, and level  $k$  of fixed treatment B. Let the number of blocks be  $r$ , the number of levels of A be  $a$ , and the number of levels of B be  $b$ .

$$Y_{ijk} = \mu + \rho_i + \alpha_j + \eta_{ij} + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk} = \mu_{ijk} + \epsilon_{ijk} \quad (14.1)$$

where  $1 \leq i \leq r$ ,  $1 \leq j \leq a$ , and  $1 \leq k \leq b$ . Note that this model contains two random error components,  $\eta_{ij} \sim N(0, \sigma_\eta^2)$  and  $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$ , the first associated with the plots and the second associated with the subplots. If (as expected)  $\sigma_\epsilon^2 < \sigma_\eta^2$ , then comparisons of the levels of B will be performed with greater precision at the cost of less precision for comparisons of the levels of A. Also note that within each block, treatment A is completely confounded with whole plots.

Although the example below is an agricultural experiment, split plot designs are widely used in other application areas including industry, clinical trials, and the social sciences. The agricultural terminologies plot and subplot are usually retained when working in other application areas.

### 14.3 Example—Yates Oat Data

This example comes from (Yates, 1937). We are interested in examining the effects of nitrogen fertilizer and seed variety on the yield of an oat crop. A total of 72 experimental units are arranged in 6 blocks of size 12. Each block is randomly subdivided into three plots, and each plot is further randomly subdivided into four subplots. In each block, the three varieties of seed are randomly assigned to the three plots. Then within each plot, the four levels of nitrogen are randomly assigned to the four subplots. Thus the randomization proceeds in two stages. By assigning varieties to plots and nitrogen to subplots it is implicit that there is more interest in the comparison of nitrogen levels than the comparison of varieties. The nitrogen levels are equally spaced amounts of a single fertilizer, 0, .2, .4, and .6.

The data are contained in the file (`datasets/oats.dat`). Note that this file contains six variables: those for `yield`, `blocks`, `variety`, `plots`, `nitrogen`, and `subplots`.

The design layout for this example is in Table 14.1. The physical locations of the blocks, plots, and subplots are indicated positionally. The random assignment of variety to whole plots is visible since each column within a

TABLE 14.1. Experimental layout for oat yield data with display of randomization scheme. Within each block, the **variety** factor is randomly assigned to an entire plot. Within each **block/plot**, the **nitrogen** factor is randomly assigned to the subplots.

(dsgntwo/code/yatesppl-layout.s), (dsgntwo/transcript/yatesppl-layout.st)

S-PLUS (dsgntwo/transcript/yatesppl-layout.ste):

| , , B1<br><table style="margin-left: 20px; border-collapse: collapse;"> <thead> <tr><th>P1</th><th>P2</th><th>P3</th></tr> </thead> <tbody> <tr><td>S1 V3:N.6 V1:N.0 V2:N.0</td><td></td><td></td></tr> <tr><td>S2 V3:N.4 V1:N.2 V2:N.2</td><td></td><td></td></tr> <tr><td>S3 V3:N.2 V1:N.6 V2:N.4</td><td></td><td></td></tr> <tr><td>S4 V3:N.0 V1:N.4 V2:N.6</td><td></td><td></td></tr> </tbody> </table><br>, , B2<br><table style="margin-left: 20px; border-collapse: collapse;"> <thead> <tr><th>P1</th><th>P2</th><th>P3</th></tr> </thead> <tbody> <tr><td>S1 V3:N.4 V1:N.6 V2:N.2</td><td></td><td></td></tr> <tr><td>S2 V3:N.0 V1:N.0 V2:N.0</td><td></td><td></td></tr> <tr><td>S3 V3:N.2 V1:N.2 V2:N.4</td><td></td><td></td></tr> <tr><td>S4 V3:N.6 V1:N.4 V2:N.6</td><td></td><td></td></tr> </tbody> </table><br>, , B3<br><table style="margin-left: 20px; border-collapse: collapse;"> <thead> <tr><th>P1</th><th>P2</th><th>P3</th></tr> </thead> <tbody> <tr><td>S1 V2:N.2 V3:N.6 V1:N.0</td><td></td><td></td></tr> <tr><td>S2 V2:N.4 V3:N.2 V1:N.6</td><td></td><td></td></tr> <tr><td>S3 V2:N.6 V3:N.4 V1:N.2</td><td></td><td></td></tr> <tr><td>S4 V2:N.0 V3:N.0 V1:N.4</td><td></td><td></td></tr> </tbody> </table> | P1 | P2 | P3 | S1 V3:N.6 V1:N.0 V2:N.0 |  |  | S2 V3:N.4 V1:N.2 V2:N.2 |  |  | S3 V3:N.2 V1:N.6 V2:N.4 |  |  | S4 V3:N.0 V1:N.4 V2:N.6 |  |  | P1 | P2 | P3 | S1 V3:N.4 V1:N.6 V2:N.2 |  |  | S2 V3:N.0 V1:N.0 V2:N.0 |  |  | S3 V3:N.2 V1:N.2 V2:N.4 |  |  | S4 V3:N.6 V1:N.4 V2:N.6 |  |  | P1 | P2 | P3 | S1 V2:N.2 V3:N.6 V1:N.0 |  |  | S2 V2:N.4 V3:N.2 V1:N.6 |  |  | S3 V2:N.6 V3:N.4 V1:N.2 |  |  | S4 V2:N.0 V3:N.0 V1:N.4 |  |  | , , B4<br><table style="margin-left: 20px; border-collapse: collapse;"> <thead> <tr><th>P1</th><th>P2</th><th>P3</th></tr> </thead> <tbody> <tr><td>S1 V3:N.4 V2:N.0 V1:N.2</td><td></td><td></td></tr> <tr><td>S2 V3:N.6 V2:N.4 V1:N.4</td><td></td><td></td></tr> <tr><td>S3 V3:N.0 V2:N.6 V1:N.6</td><td></td><td></td></tr> <tr><td>S4 V3:N.2 V2:N.2 V1:N.0</td><td></td><td></td></tr> </tbody> </table><br>, , B5<br><table style="margin-left: 20px; border-collapse: collapse;"> <thead> <tr><th>P1</th><th>P2</th><th>P3</th></tr> </thead> <tbody> <tr><td>S1 V2:N.6 V1:N.4 V3:N.4</td><td></td><td></td></tr> <tr><td>S2 V2:N.0 V1:N.6 V3:N.6</td><td></td><td></td></tr> <tr><td>S3 V2:N.4 V1:N.0 V3:N.2</td><td></td><td></td></tr> <tr><td>S4 V2:N.2 V1:N.2 V3:N.0</td><td></td><td></td></tr> </tbody> </table><br>, , B6<br><table style="margin-left: 20px; border-collapse: collapse;"> <thead> <tr><th>P1</th><th>P2</th><th>P3</th></tr> </thead> <tbody> <tr><td>S1 V1:N.4 V2:N.6 V3:N.0</td><td></td><td></td></tr> <tr><td>S2 V1:N.0 V2:N.4 V3:N.2</td><td></td><td></td></tr> <tr><td>S3 V1:N.6 V2:N.0 V3:N.4</td><td></td><td></td></tr> <tr><td>S4 V1:N.2 V2:N.2 V3:N.6</td><td></td><td></td></tr> </tbody> </table> | P1 | P2 | P3 | S1 V3:N.4 V2:N.0 V1:N.2 |  |  | S2 V3:N.6 V2:N.4 V1:N.4 |  |  | S3 V3:N.0 V2:N.6 V1:N.6 |  |  | S4 V3:N.2 V2:N.2 V1:N.0 |  |  | P1 | P2 | P3 | S1 V2:N.6 V1:N.4 V3:N.4 |  |  | S2 V2:N.0 V1:N.6 V3:N.6 |  |  | S3 V2:N.4 V1:N.0 V3:N.2 |  |  | S4 V2:N.2 V1:N.2 V3:N.0 |  |  | P1 | P2 | P3 | S1 V1:N.4 V2:N.6 V3:N.0 |  |  | S2 V1:N.0 V2:N.4 V3:N.2 |  |  | S3 V1:N.6 V2:N.0 V3:N.4 |  |  | S4 V1:N.2 V2:N.2 V3:N.6 |  |  |
|---|----|----|----|-------------------------|--|--|-------------------------|--|--|-------------------------|--|--|-------------------------|--|--|----|----|----|-------------------------|--|--|-------------------------|--|--|-------------------------|--|--|-------------------------|--|--|----|----|----|-------------------------|--|--|-------------------------|--|--|-------------------------|--|--|-------------------------|--|--|---|----|----|----|-------------------------|--|--|-------------------------|--|--|-------------------------|--|--|-------------------------|--|--|----|----|----|-------------------------|--|--|-------------------------|--|--|-------------------------|--|--|-------------------------|--|--|----|----|----|-------------------------|--|--|-------------------------|--|--|-------------------------|--|--|-------------------------|--|--|
| P1  | P2 | P3 |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S1 V3:N.6 V1:N.0 V2:N.0   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S2 V3:N.4 V1:N.2 V2:N.2   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S3 V3:N.2 V1:N.6 V2:N.4   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S4 V3:N.0 V1:N.4 V2:N.6   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| P1  | P2 | P3 |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S1 V3:N.4 V1:N.6 V2:N.2   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S2 V3:N.0 V1:N.0 V2:N.0   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S3 V3:N.2 V1:N.2 V2:N.4   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S4 V3:N.6 V1:N.4 V2:N.6   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| P1  | P2 | P3 |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S1 V2:N.2 V3:N.6 V1:N.0   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S2 V2:N.4 V3:N.2 V1:N.6   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S3 V2:N.6 V3:N.4 V1:N.2   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S4 V2:N.0 V3:N.0 V1:N.4   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| P1  | P2 | P3 |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S1 V3:N.4 V2:N.0 V1:N.2   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S2 V3:N.6 V2:N.4 V1:N.4   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S3 V3:N.0 V2:N.6 V1:N.6   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S4 V3:N.2 V2:N.2 V1:N.0   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| P1  | P2 | P3 |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S1 V2:N.6 V1:N.4 V3:N.4   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S2 V2:N.0 V1:N.6 V3:N.6   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S3 V2:N.4 V1:N.0 V3:N.2   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S4 V2:N.2 V1:N.2 V3:N.0   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| P1  | P2 | P3 |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S1 V1:N.4 V2:N.6 V3:N.0   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S2 V1:N.0 V2:N.4 V3:N.2   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S3 V1:N.6 V2:N.0 V3:N.4   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |
| S4 V1:N.2 V2:N.2 V3:N.6   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |   |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |    |    |    |                         |  |  |                         |  |  |                         |  |  |                         |  |  |

block contains only one level of variety. The random assignment of nitrogen to subplots is made visible since each column within a block contains all four levels of nitrogen.

Here we have two fixed factors bearing a crossed relationship. In this situation the restricted randomization requires that the plot factor **variety** must be tested with denominator mean square for **plots(blocks)** or “whole plot error” as in Table 14.3 by SAS and Tables 14.4 and 14.6 by S-PLUS. In both SAS and S-PLUS this specification requires a statement to override of the default choice of the *F*-test.

Testing **variety** against the residual mean square is incorrect in this example because that test assumes an unrestricted randomization. Tables 14.2 and Table 14.5 are therefore not correct.

The interaction plot of the two treatment variables is in Figure 14.1. The correct tabular analyses support the visual impressions from this figure:

### y: main effects and 2-way interactions

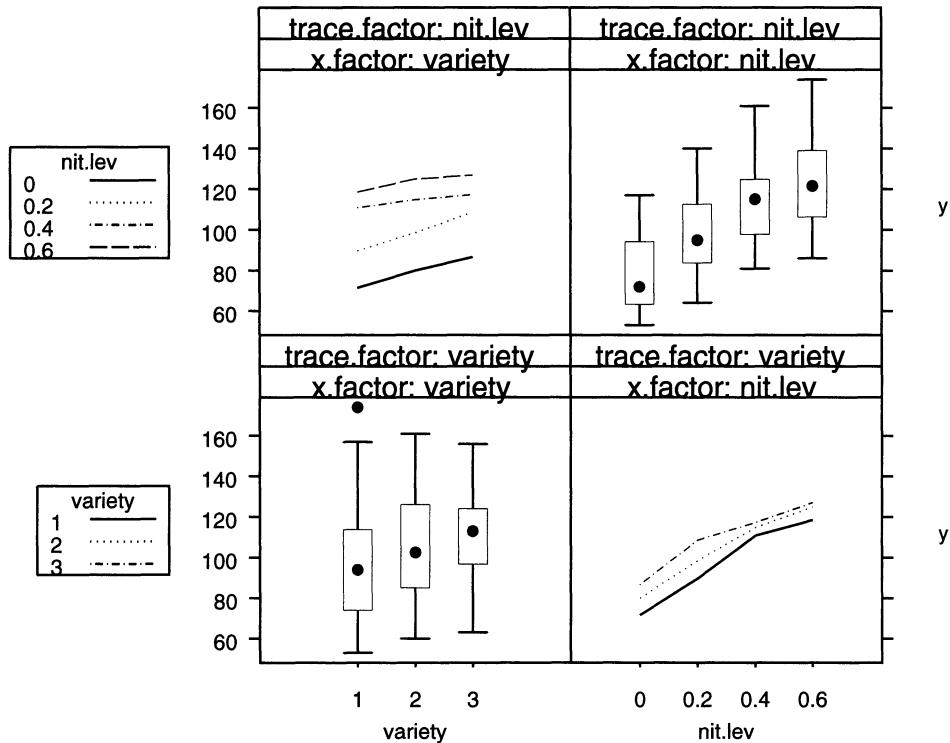


FIGURE 14.1. Interaction plot for Yates split plot on oats. The linear effect of nitrogen level is clearly visible in both right-hand panels. The lack of effect of variety, confirmed in Tables 14.3 and 14.4, is also visible.

([dsgntwo/code/yatesppl-bwplot.s](#)), ([dsgntwo/figure/yatesppl.eps.gz](#))

- The factors **variety** and **nitrogen** do not interact.
- The mean **Yield** increases linearly as the amount of **nitrogen** increases. **nitrogen** linear has a small *p*-value; the *p*-value of **nitrogen** quadratic is large.
- The mean **Yield** does not differ significantly across the three levels of **variety**.

TABLE 14.2. Default SAS display for oats split plot with incorrect *F*-test for variety. The *model* statement always generates tests appropriate for a fixed effects model. The correct output from the *contrast* and *test* statements is in Table 14.3.

(dsgntwo/code/oats.sas), (dsgntwo/transcript/oats.lst)

SAS (dsgntwo/code/oats.glm.sas):

```
proc glm data=oats;
  classes blocks plots subplots variety nitrogen;
  model yield = blocks
    variety plots(blocks)
    nitrogen variety*nitrogen /ss1;
  test h=variety e=plots(blocks);
  contrast 'nitrogen linear' nitrogen -3 -1 1 3;
  contrast 'nitrogen quadratic' nitrogen 1 -1 -1 1;
  contrast 'nitrogen cubic' nitrogen -1 3 -3 1;
run;
```

SAS (dsgntwo/transcript/oats-a.lst):  
Dependent Variable: yield

| Source           | DF        | Sum of Squares | Mean Square | F Value | Pr > F |
|------------------|-----------|----------------|-------------|---------|--------|
| Model            | 26        | 44017.19444    | 1692.96902  | 9.56    | <.0001 |
| Error            | 45        | 7968.75000     | 177.08333   |         |        |
| Corrected Total  | 71        | 51985.94444    |             |         |        |
| R-Square         | Coeff Var | Root MSE       | yield Mean  |         |        |
| 0.846713         | 12.79887  | 13.30727       | 103.9722    |         |        |
| Source           | DF        | Type I SS      | Mean Square | F Value | Pr > F |
| blocks           | 5         | 15875.27778    | 3175.05556  | 17.93   | <.0001 |
| variety          | 2         | 1786.36111     | 893.18056   | 5.04    | 0.0106 |
| plots(blocks)    | 10        | 6013.30556     | 601.33056   | 3.40    | 0.0023 |
| nitrogen         | 3         | 20020.50000    | 6673.50000  | 37.69   | <.0001 |
| variety*nitrogen | 6         | 321.75000      | 53.62500    | 0.30    | 0.9322 |

TABLE 14.3. Correct tests for split plot design specified with the SAS **contrast** and **test** statements.  
 (dsgntwo/code/oats.sas), (dsgntwo/transcript/oats.lst)

| SAS (dsgntwo/transcript/oats-b.lst): |    |             |             |         |        |  |
|--------------------------------------|----|-------------|-------------|---------|--------|--|
| Contrast                             | DF | Contrast SS | Mean Square | F Value | Pr > F |  |
| nitrogen linear                      | 1  | 19536.40000 | 19536.40000 | 110.32  | <.0001 |  |
| nitrogen quadratic                   | 1  | 480.50000   | 480.50000   | 2.71    | 0.1065 |  |
| nitrogen cubic                       | 1  | 3.60000     | 3.60000     | 0.02    | 0.8873 |  |

| Tests of Hypotheses Using the Type I<br>MS for plots(blocks) as an Error Term |    |             |             |         |        |  |
|---|----|-------------|-------------|---------|--------|--|
| Source  | DF | Type I SS   | Mean Square | F Value | Pr > F |  |
| variety   | 2  | 1786.361111 | 893.180556  | 1.49    | 0.2724 |  |
| Contrast  | DF | Contrast SS | Mean Square | F Value | Pr > F |  |
| nitrogen linear   | 1  | 19536.40000 | 19536.40000 | 110.32  | <.0001 |  |
| nitrogen quadratic  | 1  | 480.50000   | 480.50000   | 2.71    | 0.1065 |  |
| nitrogen cubic  | 1  | 3.60000     | 3.60000     | 0.02    | 0.8873 |  |

| Tests of Hypotheses Using the Type I<br>MS for plots(blocks) as an Error Term |    |             |             |         |        |  |
|---|----|-------------|-------------|---------|--------|--|
| Source  | DF | Type I SS   | Mean Square | F Value | Pr > F |  |
| variety   | 2  | 1786.361111 | 893.180556  | 1.49    | 0.2724 |  |

TABLE 14.4. Correct S-PLUS display for oats split plot design. The ANOVA table was constructed by specifying the denominators for the appropriate  $F$ -tests in the `Error` function.  
 (dsgntwo/code/yatesppl.s), (dsgntwo/transcript/yatesppl.st)

S-PLUS (dsgntwo/transcript/yatesppl-1.st):

```
> yatesppl.anova <- aov(y ~ variety*nitrogen + Error(blocks/plots/subplots),
+                           data=yatesppl)
> summary(yatesppl.anova)

Error: blocks
      Df Sum of Sq Mean Sq F Value Pr(F)
Residuals 5   15875.28 3175.056

Error: plots %in% blocks
      Df Sum of Sq Mean Sq F Value     Pr(F)
variety  2    1786.361 893.1806 1.48534 0.2723869
Residuals 10   6013.306 601.3306

Error: subplots %in% (blocks/plots)
      Df Sum of Sq Mean Sq F Value     Pr(F)
nitrogen 3   20020.50 6673.500 37.68565 0.0000000
variety:nitrogen 6    321.75   53.625  0.30282 0.9321988
Residuals 45   7968.75 177.083
```

TABLE 14.5. Incorrect S-PLUS specification for oats split plot design that ignores the split plot. The test for `variety` is incorrectly against the 45-df Residual and incorrectly shows as significant.  
 (dsgntwo/code/yatesppl.s), (dsgntwo/transcript/yatesppl.st)

S-PLUS (dsgntwo/transcript/yatesppl-3.st):

```
> ## incorrect analysis that ignores split plot
> yatesppl.wrong.anova <- aov(y ~ (blocks*variety)+(nitrogen*variety),
+                                 data=yatesppl)
> summary(yatesppl.wrong.anova)

      Df Sum of Sq Mean Sq F Value     Pr(F)
blocks 5   15875.28 3175.056 17.92973 0.0000000
variety 2    1786.36 893.181  5.04384 0.0105573
nitrogen 3   20020.50 6673.500 37.68565 0.0000000
blocks:variety 10   6013.31 601.331  3.39575 0.0022511
nitrogen:variety 6    321.75   53.625  0.30282 0.9321988
Residuals 45   7968.75 177.083
```

TABLE 14.6. S-PLUS display for oats split plot. Table of means.  
(dsgntwo/code/yatesppl.s), (dsgntwo/transcript/yatesppl.st)

---

```
S-PLUS (dsgntwo/transcript/yatesppl-2.st):
> model.tables(yatesppl.anova, type="means", se=T)
Refitting model to allow projection

Tables of means
Grand mean

103.97

variety
 1   2   3
97.625 104.5 109.79

nitrogen
 1   2   3   4
79.389 98.889 114.22 123.39

variety:nitrogen
Dim 1 : variety
Dim 2 : nitrogen
 1   2   3   4
1 71.50 89.67 110.83 118.50
2 80.00 98.50 114.67 124.83
3 86.67 108.50 117.17 126.83

Standard errors for differences of means

variety 7.0789
replic. 24.0000

nitrogen 4.4358
replic. 18.0000

variety:nitrogen
When comparing means with same levels of:
variety otherwise
 7.683    9.715
replic. 6
```

---

### 14.3.1 Alternate Specification

Our presentation of the Yates oat data in Table 14.1 and our emphasis in the analysis in Tables 14.3 and 14.4 show five distinct factors in the split plot design. We believe this is the best way to illustrate the concepts of restricted randomization, of different precisions for different comparisons, and the logistics and practical details of running an experiment. Many texts and examples show only three factors by suppressing the explicit identification of the **plots** and **subplots**.

The arithmetic of the analysis and the interpretation of the results are identical whether or not the structure is made explicit with the extra factors. The two ANOVA tables generated by the statements in Table 14.7 are identical. The second specification leads the reader to ask nonsense questions like “how can nitrogen be crossed with variety and nested within variety at the same time?” The first specification, by explicitly naming the random **plots** and **subplots** factors and explicitly showing the random assignment of the fixed **variety** and **nitrogen** factors to the random factors, makes the distinction clear. The plot structure shows the **subplots** nested in the **plots**. The treatment structure shows the **variety** crossed with the **nitrogen**. They are different factors. We are therefore not surprised that they have different relationships.

The *ANOVA table* and *interpretation* of the analysis are identical with either specification. The *logic* behind the EMS (expected mean squares) calculations is displayed when the **plots** and **subplots** factors are visible. The statistical justification for the appropriate *F*-tests is cryptic at best

TABLE 14.7. S-PLUS Alternate specifications of design.

---

```
S-PLUS (dsgntwo/code/yatesppl-alt.s):
## statement in book and recommended way to think
## about split plot designs
yatesppl.anova <- aov(y ~ variety*nitrogen +
                         Error(blocks/plots/subplots),
                         data=yatesppl)
summary(yatesppl.anova)

## alternate specification
yatesppl2.anova <- aov(y ~ variety*nitrogen +
                         Error(blocks/variety/nitrogen),
                         data=yatesppl)
summary(yatesppl2.anova)
```

---

when the `plots` and `subplots` factors are suppressed. We further explore the equivalence of these two formulations in Exercise 14.7.

The common practice of suppressing the structure is a legitimate response to the computing technology at the time (1930s) when the split plot design was invented. The calculation of the analysis with 3 explicit factors costs  $O(26^3)$  multiplications. It costs  $O(72^3)$  multiplications [that is,  $(72/26)^3 \approx 21$  times as many] with 5 explicit factors. (The “big  $O$ ” notation is defined in Appendix Section F.4.1 in the “operation count” discussion.) When ANOVA analyses were routinely performed with handcrank-operated calculating equipment, the time savings was well worth the ambiguity in notation.

### 14.3.2 Polynomial Effects for Nitrogen

Note that, as anticipated, the mean square for comparing the levels of the whole plot factor `variety`, 601, is greater than the mean square for comparing the levels of the levels of the subplot factor `nitrogen`, 177. There is no evidence of interaction between `variety` and `nitrogen`. The large  $p$ -value for `variety` suggests that the three varieties do not differ significantly, but the small  $p$ -value for the subplot factor, `nitrogen`, tells us `yield` is significantly affected by the amount of `nitrogen` used.

We further investigate the nature of the relationship between `nitrogen` and `yield` by decomposing the 3-df sum of squares for `nitrogen` into orthogonal contrasts for linear, quadratic, and cubic effects. In SAS we explicitly define the contrasts in `contrast` statements in the SAS file. In S-PLUS we use the polynomial contrast function `cont.poly` to assign the contrasts to the `nitrogen` factor. See Tables 14.3 and 14.8 for details. Since only the linear contrast is significant ( $p$ -value  $< .01$ ) we conclude that `yield` increases linearly with `nitrogen`. This finding suggests a need for further experimentation to determine the amount of nitrogen that should be used to maximize `yield`.

Response surface methodology is the experimental design technique used to determine the combination of inputs that maximizes or minimizes output. We do not pursue this further but recommend the interested reader consult either (Montgomery, 2001) or (Box et al., 1978).

Multiple comparisons plots for `nitrogen` drawn by the code in file (`dsgntwo/code/yatesppl-mmc2.s`) are available in the online files. File (`dsgntwo/figure/yatesppl-mmc-pairs.eps.gz`) shows the pairwise comparisons and file (`dsgntwo/figure/yatesppl-mmc.eps.gz`) shows the polynomial contrasts. The linear contrast appears to be carrying all the significance.

TABLE 14.8. Continuation of the split plot analysis from Table 14.4. The 3-df nitrogen effect is partitioned into polynomial contrasts. The linear effect carries almost the entire sum of squares and is the only significant contrast.

(dsgntwo/code/yatesppl.s), (dsgntwo/transcript/yatesppl.st6)

```

S-PLUS (dsgntwo/transcript/yatesppl.st6):
> ## polynomial contrasts in nitrogen
> contrasts(yatesppl$nitrogen)
 [,1] [,2] [,3]
 1   -1   -1   -1
 2    1   -1   -1
 3    0    2   -1
 4    0    0    3
> contrasts(yatesppl$nitrogen) <- contr.poly(4)
> contrasts(yatesppl$nitrogen)
.L .Q .C
1 -0.6708204 0.5 -0.2236068
2 -0.2236068 -0.5  0.6708204
3  0.2236068 -0.5 -0.6708204
4  0.6708204 0.5  0.2236068
>
> ## split plot analysis with polynomial contrasts
> yatespplp.anova <- aov(y ~ variety*nitrogen +
+                         Error(blocks/plots/subplots),
+                         data=yatesppl)
> summary(yatespplp.anova,
+          split=list(nitrogen=list(linear=1, quad=2, cub=3)),
+          expand.split=F)
Error: blocks
      Df Sum of Sq Mean Sq F Value Pr(F)
Residuals 5  15875.28 3175.056

Error: plots %in% blocks
      Df Sum of Sq Mean Sq F Value     Pr(F)
variety  2  1786.361 893.1806 1.48534 0.2723869
Residuals 10  6013.306 601.3306

Error: subplots %in% (blocks/plots)
      Df Sum of Sq Mean Sq F Value     Pr(F)
nitrogen  3  20020.50 6673.50 37.6856 0.0000000
nitrogen: linear 1  19536.40 19536.40 110.3232 0.0000000
nitrogen: quad  1    480.50   480.50   2.7134 0.1064745
nitrogen: cub   1     3.60     3.60   0.0203 0.8872574
variety:nitrogen 6    321.75    53.62   0.3028 0.9321988
Residuals 45  7968.75   177.08

```

## 14.4 Introduction to Fractional Factorial Designs

The idea behind fractional factorial designs is to substantially reduce the number of *experimental units* (e.u.'s) required for the experiment by purposely confounding (see Section 14.1) all effects of interest with only effects that are not of interest and that can be assumed negligible. One is almost always able to assume that high-order interactions are negligible. This strategy permits estimation of main effects and low-order interactions while experimenting on only a small proportion of all possible *treatment combinations* (t.c.'s). The resulting experimental plan is called a *fractional replicate*. Implementation of this class of designs involves carefully selecting a fractional replicate subset of the possible t.c.'s so as to purposely confound high-order factor interactions with one another (and with blocks, if any), while maintaining the unconfoundedness of main effects and of lower-order interactions of interest.

Since it is frequently the case that there are more t.c.'s to be run than there are homogeneous experimental units (e.u.'s), a blocking scheme is usually part of this type of design. A fractional replicate of the complete experiment is run within each homogeneous block. The assignment of t.c.'s to e.u.'s purposely confounds higher-order interactions of the treatment factors with the block effects.

Note that the  $r \times r$  Latin square design in Section 13.2 is a special case of fractional replication, a  $\frac{1}{r}$  replicate of an  $r^3$  experiment.

Our discussion will be limited to the situation where there are  $n$  factors each at 2 levels and only  $2^k$  e.u.'s are available,  $k < n$ ; this is referred to as a  $1/(2^{n-k})$  fractional replication. The design is called a  $2^{n-(n-k)}$  design. Fractional factorial designs exist when all factors have a common number of levels greater than 2, or the factors have varying numbers of levels, for example three factors with 2 levels each and four factors each having 3 levels. The need to consider a situation with many 2-level factors is not uncommon, for the 2 levels can be the presence or absence of a particular condition.

### 14.4.1 Example— $2^{8-2}$ Design

Suppose we have 8 factors (denoted by the letters A through H), each with 2 levels, and we have enough experimental units to run  $2^6$  of the  $2^8$  possible t.c.'s. Further assume that the maximum sized set of homogeneous experimental units is  $2^4$ , so that the 64 selected t.c.'s will be arranged in 4 blocks, each containing 16 e.u.'s. Table 14.9 is an experimental layout for the  $2^{8-2} = (2^8)/4$  design, prior to randomization.

TABLE 14.9. Experimental layout for the  $2^{8-2} = (2^8)/4$  design, prior to randomization. This design follows from a permutation of the factor labels in Plan 6A.16 of (Cochran and Cox, 1957)

|         | block 1  | block 2 | block 3    | block 4 |
|---------|----------|---------|------------|---------|
| (1)     | ab       | ce      | de         |         |
| ach     | bch      | aeh     | acdeh      |         |
| aef     | bef      | acf     | adf        |         |
| cefh    | abcef h  | fh      | cdfh       |         |
| bdh     | adh      | bcdeh   | beh        |         |
| abcd    | cd       | abde    | abce       |         |
| abdefh  | defh     | abcdh   | abfh       |         |
| bcd e f | acdef    | bdf     | bcf        |         |
| beg     | aeg      | bcg     | bdg        |         |
| abcegh  | cegh     | abgh    | abcdgh     |         |
| abfg    | fg       | abcefg  | abdefg     |         |
| bcfgh   | acfgh    | befgh   | bcd e f gh |         |
| degh    | abdegh   | cdgh    | gh         |         |
| acdeg   | bcdeg    | adg     | acg        |         |
| adfg h  | bdfgh    | acdefgh | aefgh      |         |
| cdfg    | adcd f g | defg    | cefg       |         |

The treatment combinations have been very carefully chosen so that, provided one can assume that all 3-factor and higher-order interactions are negligible, one can “cleanly” estimate all main effects and 2-factor interactions with a sufficient number of error df to assure tests of reasonable power. Blocks, all main effects, and all two-factor interactions are confounded only with three-factor and higher interactions. The basic form of the ANOVA table for the  $2^{8-2}$  design in Table 14.9 is in Table 14.10. Estimates are not available for interaction effects that are confounded with blocks.

If instead we had been required to maintain a maximum block size of 8, it would have been necessary to completely confound 2 of the 28 2-factor interactions with blocks, and error df would have decreased to 22.

TABLE 14.10. ANOVA table for the  $2^{8-2} = (2^8)/4$  design from Table 14.9. The main effects and 2-factor interactions are unconfounded with blocks.

| Source                | df | comments                                      |
|-----------------------|----|---|
| Blocks                | 3  | blocks are aliased with abc, def and abcdef   |
| Main effects          | 8  | unconfounded                                  |
| 2-factor interactions | 28 | $\binom{8}{2}$ terms, unconfounded            |
| Error                 | 24 | aliased with 3-factor and higher interactions |
| Total                 | 63 |   |

TABLE 14.11. Model formula (S-PLUS) and model statement (SAS) for  $2^{8-2} = (2^8)/4$  fractional factorial design. The S-PLUS notation includes syntax for generating the complete set of 2-factor interactions. In our setting of the SAS statements we aligned the factor names to make it easy to verify that we wrote them all down.

---

```
S-PLUS (dsgntwo/code/2.8-2.s):
y ~ blocks + (a+b+c+d+e+f+g+h)^2

SAS (dsgntwo/code/2.8-2.sas):
model y = blocks
    a b c d e f g h
    a*b a*c a*d a*e a*f a*g a*h
    b*c b*d b*e b*f b*g b*h
    c*d c*e c*f c*g c*h
    d*e d*f d*g d*h
    e*f e*g e*h
    f*g f*h
    g*h ;
```

---

The analysis of such data in either S-PLUS or SAS is very straightforward. Both statements are shown in Table 14.11. We only need ensure that the model statement declare only blocks, the 8 main effects, and the 28 interactions.

#### 14.4.2 Example— $2^{5-1}$ Design

Five factors involved in a manufacturing process for an integrated circuit were investigated. For brevity we refer to the five factors as A, B, C, D, E. Resources were available to examine only 16 of the  $2^5$  treatment combinations. A particular half-replicate of the complete experiment was used such that all main effects are confounded with four-factor interactions and all two-factor interactions are confounded with three-factor interactions. Based on experience with these factors, the investigator was very confident that only factors A, B, C, and the AB interaction were likely to have an appreciable effect on the process yield. This is confirmed by examining the interaction plot in Figure 14.2 or the means in Table 14.12. The S-PLUS output in Table 14.12 also contains two tables of means demonstrating that the D main effect and AC interaction are not significant. The data from (Montgomery, 2001) is in (datasets/circuit.dat). The SAS analysis is

### yield: main effects and 2-way interactions

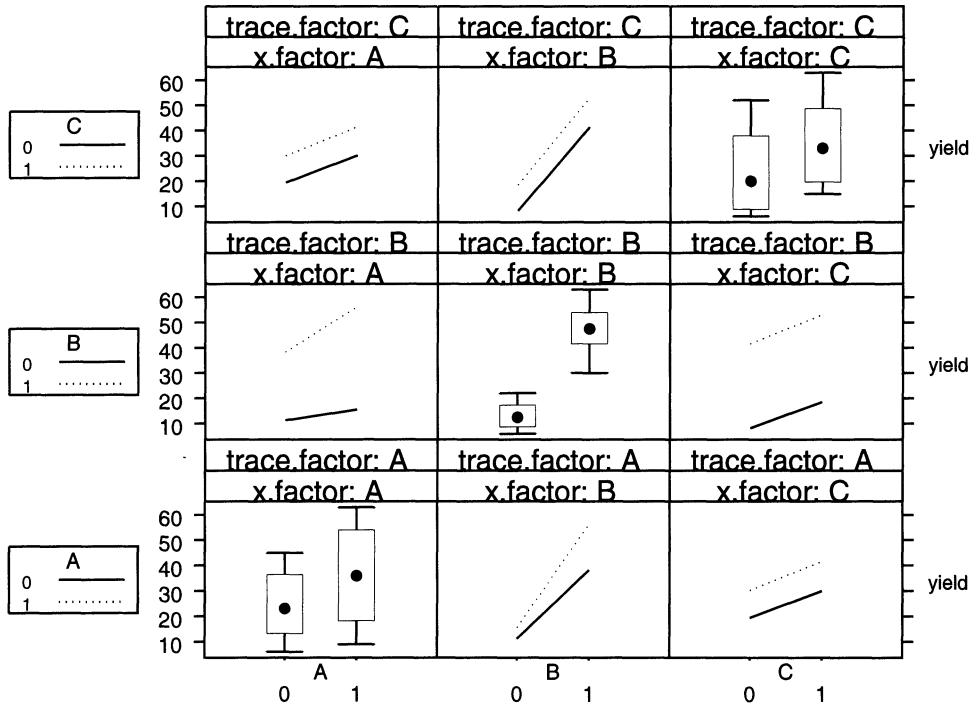


FIGURE 14.2. Interaction plot for circuit  $2^{5-1}$  design. All three main effects are visibly significant, as is the  $A \times B$  interaction. The traces in the  $A \times C$  and  $B \times C$  panels are parallel, hence not significant. (dsgntwo/code/circuit.s), (dsgntwo/figure/circuit.eps.gz)

in Table 14.13. We find that there is a significantly higher yield at the higher level of each of  $A, B, C$  than at their respective lower levels, and that the simple effect of  $B$  at the higher level of  $A$  is significantly greater than the simple effect of  $B$  at the lower level of  $A$ .

TABLE 14.12. S-PLUS Analysis of  $2^{5-1} = (2^5)/2$  fractional factorial design.  
 (dsgntwo/code/circuit.s)

---

```
S-PLUS (dsgntwo/transcript/circuit.st):
> ## Analysis of Integrated Circuit Data
>
> circuit.aov <- aov( yield ~ a + b + c + a:b, data=circuit)
>
> summary(circuit.aov)
      Df Sum of Sq  Mean Sq   F Value      Pr(F)
a          1    495.062  495.062 193.195 2.534800e-008
b          1   4590.062 4590.062 1791.244 0.000000e+000
c          1    473.062  473.062 184.610 3.213600e-008
a:b        1    189.063  189.063  73.780 3.301648e-006
Residuals 11    28.188    2.563
>
> tapply(circuit[,"yield"], circuit[, "a"], mean)
  0      1
24.75 35.875
> tapply(circuit[,"yield"], circuit[, "b"], mean)
  0      1
13.375 47.25
> tapply(circuit[,"yield"], circuit[, "c"], mean)
  0      1
24.875 35.75
> tapply(circuit[,"yield"], circuit[,c("a","b")], mean)
  0      1
0 11.25 38.25
1 15.50 56.25
> tapply(circuit[,"yield"], circuit[, "d"], mean)
  0      1
30.75 29.875
> tapply(circuit[,"yield"], circuit[,c("a","c")], mean)
  0      1
0 19.50 30.0
1 30.25 41.5
```

---

TABLE 14.13. SAS Analysis of  $2^{5-1} = (2^5)/2$  fractional factorial design.

```

SAS (dsgntwo/code/circuit.sas):
title 'Analysis of Integrated Circuit Data';
data circuit;
  infile "&hh/datasets/circuit.dat" firstobs=2;
  input A B C D E yield;

proc anova;
  class A B C D E;
  model yield = A B C A*B;
run;

```

SAS (dsgntwo/transcript/circuit.lst):  
 Analysis of Integrated Circuit Data

Dependent Variable: YIELD

| Source          | DF | Sum of Squares | Mean Square  | F Value   | Pr > F     |
|-----------------|----|----------------|--------------|-----------|------------|
| Model           | 4  | 5747.250000    | 1436.8125000 | 560.71    | 0.0001     |
| Error           | 11 | 28.1875000     | 2.5625000    |           |            |
| Corrected Total | 15 | 5775.4375000   |              |           |            |
|                 |    |                |              |           |            |
| R-Square        |    | C.V.           | Root MSE     |           | YIELD Mean |
|                 |    | 0.995119       | 5.280927     | 1.6007811 | 30.312500  |

| Source | DF | Anova SS     | Mean Square  | F Value | Pr > F |
|--------|----|--------------|--------------|---------|--------|
| A      | 1  | 495.0625000  | 495.0625000  | 193.20  | 0.0001 |
| B      | 1  | 4590.0625000 | 4590.0625000 | 1791.24 | 0.0001 |
| C      | 1  | 473.0625000  | 473.0625000  | 184.61  | 0.0001 |
| A*B    | 1  | 189.0625000  | 189.0625000  | 73.78   | 0.0001 |

## 14.5 Introduction to Crossover Designs

This is a subclass of *repeated measures designs*, used when one applies two or more treatments to each of several subjects over the course of two or more periods, and needs to account for the possibility that a *carryover* or *residual* effect of a treatment lingers into the following period (and possibly beyond it). Thus a subject's response may be attributable to both the treatment given in the period and the treatment administered in the preceding period. One seeks to be able to provide unconfounded estimates of both the direct and residual effects. These designs are also referred to as changeover or residual effects designs.

The possible existence of residual effects is easy to imagine in medical or agricultural experiments. It also must be accounted for in meteorological experiments involving cloud seeding intended to induce precipitation.

Intuitively, a “good” design is one in which

1. Each treatment occurs equally often on each subject.
2. Each treatment occurs equally often in each period.
3. Each treatment follows each other treatment the same number of times.

But the available numbers of treatments, subjects, and periods often make it impossible to satisfy all three criteria. As a simple example, suppose we have two treatments, say A and B, to compare in three periods on two experimental animals. Consider the following two designs:

| Period | Design 1 |          | Design 2 |          |
|--------|----------|----------|----------|----------|
|        | Animal 1 | Animal 2 | Animal 1 | Animal 2 |
| 1      | A        | B        | A        | B        |
| 2      | B        | A        | B        | A        |
| 3      | A        | B        | B        | A        |

Which design is preferred?

It turns out that both the direct and residual treatment effects can be estimated much more precisely in Design 2 than in Design 1. Design 1 has the deficiency that each treatment is always preceded by the other treatment, never by itself. Only Design 2 satisfies the third of the above intuitive criteria. This design is a member of a class of crossover designs constructed as a Latin square with the last row repeated once. This class has the property that the estimation of direct and residual treatment effects are orthogonal to one another.

TABLE 14.14. Two  $3 \times 3$  Latin squares for crossover design with display of the residual effect. The factor **nores** is an indicator for observations that do not have a residual effect because there is no preceding treatment. The residual treatment factor **restreat** has the value of the treatment **treat** for the preceding period with the same **square** and **sequence**.  
 (datasets/cc135.dat)

a. Design arranged to show the Latin square structure.

| cow    | square | 1 |   |   | 2 |   |   |
|--------|--------|---|---|---|---|---|---|
|        |        | 1 | 2 | 3 | 4 | 5 | 6 |
| period | 1      | A | B | C | A | B | C |
|        | 2      | B | C | A | C | A | B |
|        | 3      | C | A | B | B | C | A |

b. Design and data arranged by observation.

| period | square | sequence | cow | treat | yield | nores | restreat |
|--------|--------|----------|-----|-------|-------|-------|----------|
| 1      | 1      | 1        | 1   | A     | 38    | 0     | 0        |
|        | 1      | 2        | 2   | B     | 109   | 0     | 0        |
|        | 1      | 3        | 3   | C     | 124   | 0     | 0        |
|        | 2      | 1        | 4   | A     | 86    | 0     | 0        |
|        | 2      | 2        | 5   | B     | 75    | 0     | 0        |
|        | 2      | 3        | 6   | C     | 101   | 0     | 0        |
| 2      | 1      | 1        | 1   | B     | 25    | 1     | A        |
|        | 1      | 2        | 2   | C     | 86    | 1     | B        |
|        | 1      | 3        | 3   | A     | 72    | 1     | C        |
|        | 2      | 1        | 4   | C     | 76    | 1     | A        |
|        | 2      | 2        | 5   | A     | 35    | 1     | B        |
|        | 2      | 3        | 6   | B     | 63    | 1     | C        |
| 3      | 1      | 1        | 1   | C     | 15    | 1     | B        |
|        | 1      | 2        | 2   | A     | 39    | 1     | C        |
|        | 1      | 3        | 3   | B     | 27    | 1     | A        |
|        | 2      | 1        | 4   | B     | 46    | 1     | C        |
|        | 2      | 2        | 5   | C     | 34    | 1     | A        |
|        | 2      | 3        | 6   | A     | 1     | 1     | B        |

### Example—Two Latin Squares

This design, from (Cochran and Cox, 1957) and (Heiberger, 1989) uses two  $3 \times 3$  Latin squares to estimate the residual effects as well as the direct effects of milk yields resulting from three treatments to dairy cows. The design and data are in Table 14.14.

We specify the design as if the residual effects are attributable to another factor. The residual effects `restreat` are explicitly set to the value of the direct effects `treat` in the preceding `period`. A new dummy variable `nores` is set to 0 for the first period (in which there are no residual treatments), and to 1 for the remaining periods (in which it is meaningful to speak about residual effects from the previous period). The new variable `nores` is confounded with one degree of freedom of `period`. `restreat` is nested in `nores`. The arithmetic in both programs does not need the identification of the `nores` dummy variable. The analysis is easier to follow if `nores` is explicitly identified.

The analysis specification statements for S-PLUS and SAS are in files (`dsgntwo/code/cc135.s`) and (`dsgntwo/code/cc135.sas`). There are two sets of `aov` or `PROC GLM` statements, each with a different ordering for the direct effects `treat` and residual effects `restreat`. The ANOVA table edited from the output of either package is in Table 14.15.

The direct and residual effects are not orthogonal to each other. Their sum is partitioned in two different ways in Table 14.15. The additional sum of squares for the residual effects after accounting for the direct effects, for which we use the first model statement and the results of which are reflected in the first set of braced lines in the ANOVA table, tells us whether there are longer-term differences that need to be accounted for. The additional sum

TABLE 14.15. Latin Square for Residual Effects design cc135. The braced expressions are two different partitionings of the same 4 df for the combined `treat + res.treat` effects. This ANOVA table is manually constructed from the output of the programs.

(`dsgntwo/code/cc135.s`), (`dsgntwo/transcript/cc135.st`), (`dsgntwo/code/cc135.sas`), (`dsgntwo/transcript/cc135.lst`)

| Source                 | Df | Sum of Sq | Mean Sq  | F Value | Pr(F) |
|------------------------|----|-----------|----------|---------|-------|
| cow                    | 5  | 5781.111  | 1156.222 | 23.211  | 0.005 |
| period in square       | 4  | 11489.111 | 2872.278 | 57.662  | 0.001 |
| treat+res.treat        | 4  | 2892.972  |          |         |       |
| {treat                 | 2  | 2276.778  | 1138.389 | 22.853  | 0.006 |
| {res.treat after treat | 2  | 616.194   | 308.097  | 6.185   | 0.060 |
| {res.treat             | 2  | 38.422    | 19.211   | 0.386   | 0.703 |
| {treat after res.treat | 2  | 2854.550  | 1427.275 | 28.653  | 0.004 |
| Residuals              | 4  | 199.250   | 49.812   |         |       |
| Total                  | 17 | 20362.444 |          |         |       |

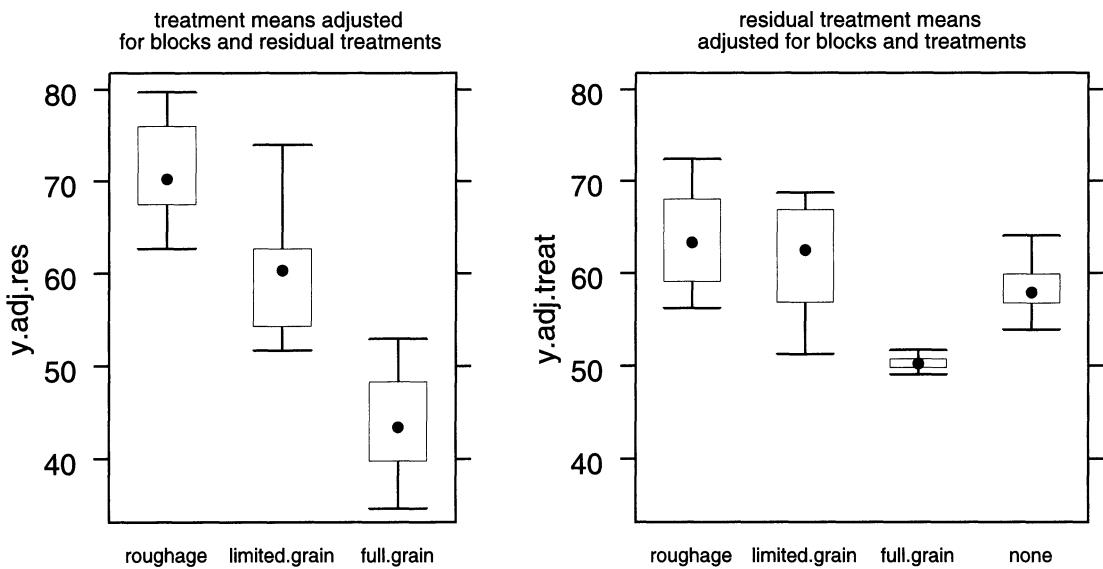


FIGURE 14.3. Boxplots for residual effects design, treatments and residual treatments, each adjusted for the blocking factors and for each other.

(dsgntwo/code/cc135-bwplot.s), (dsgntwo/figure/cc135.f.bwplot.eps.gz)

of squares for the direct effects after accounting for the residual effects, for which we use the second model statement and the results of which are reflected in the second set of braced lines in the ANOVA table, tells us whether the single period effects are themselves important. Since the residual effects after the direct effects are borderline significant with  $p = .06$ , we conclude that isolating the residual effect is important. Had we not done so the power for detecting the direct effects would have been reduced because the `res.treat` sum of squares would have been left in the residual. It would have inflated the residual, hence decreased the  $F$ -value and increased the  $p$ -value for the direct effects.

The boxplots of the adjusted treatments and adjusted residual treatments are in Figure 14.3.

## 14.6 Example—Apple Tree Data

In Section 10.5 we study the analysis using a concomitant variable (covariate) in an experiment having only one factor. In this example we demonstrate the use of this technique when there are two factors.

### Study Objectives

(Pearce, 1983), later reprinted in (Hand et al., 1994), describes a randomized block experiment to determine the effects of six ground cover treatments on the yield of apple trees. A concomitant variable, the *volume of crop* during the four years prior to treatment, is available. Perform the analysis both using and ignoring the concomitant variable. Provide recommendations as to treatment.

### Data Description

The data appear in the file (`datasets/apple.dat`). The variables are

**treat:** ground cover treatments. Treatments 1–5 are experimental treatments; Treatment 6 is a control

**block:** four randomized blocks

**yield:** pounds over a four-year period following treatment

**pre:** volume of crop over a four-year period prior to treatment

### Data Analysis

Our complete analysis specification is in file (`dsgntwo/code/apple3.s`). The complete output is in Figures 14.4, 14.5, 14.6, 14.7 and in file (`dsgntwo/transcript/apple3.st`). The ANOVA tables are displayed in Tables 14.17, 14.19, 14.20, 14.21, and 14.22. Multiple comparisons are in Figures 14.6 and 14.7 and Table 14.18.

The strategy is to begin with the most complex model and then progress toward simpler ones. This systematic approach assures that the ultimately selected model will include all significant effects but be no more complex than necessary. In Table 14.16 we list all the models used in this example, along with the names of the S-PLUS `ancova` objects containing their results and the table number in which the ANOVA is displayed.

TABLE 14.16. Index to analysis of covariance models for the apple data in Section 14.6. The parentheses for the rows in Figure 14.5 indicate that the sums of squares for the terms of models 1, 2, and 2b are the same as the sums of squares for models 3, 4, and 4b.

(dsgntwo/code/apple3.s), (dsgntwo/transcript/apple3.st)

| Model           | S-PLUS<br>object name | ANOVA<br>in Table | Row in<br>Figure 14.5 |
|-----------------|-----------------------|-------------------|-----------------------|
| yield           | ~ block + pre * treat | apple.ancova.1    | 14.17 (2)             |
| yield           | ~ block + pre + treat | apple.ancova.2    | 14.17 (5)             |
| yield           | ~ block + treat + pre | apple.ancova.2b   | 14.17 (5)             |
| yield.block     | ~ pre.block * treat   | apple.ancova.3    | 14.20 2               |
| yield.block     | ~ treat * pre.block   | apple.ancova.3b   | 14.20 2               |
| yield.block     | ~ pre.block + treat   | apple.ancova.4    | 14.22 5               |
| yield.block     | ~ treat + pre.block   | apple.ancova.4b   | 14.22 5               |
| yield.block     | ~ treat               | apple.ancova.6    | 14.21 4               |
| yield.block.pre | ~ treat               | apple.ancova.5    | 14.21 6               |
| yield.block     | ~ pre.block           | apple.ancova.7    | 14.21 3               |
| yield           | ~ pre * treat         | apple.ancova.8    | 14.19 1               |

### 14.6.1 Models in Table 14.17

Model 1, `yield ~ block + pre*treat`, allows for the possibility that the covariate `pre` and the treatment factor `treat` interact. This model allows for differing slopes in the simple regressions of `yield` on `pre` across the levels of `treat`. We draw two conclusions from the ANOVA of model 1.

1. The ANOVA table for model 1 shows that most of the sum of squares for the data is attributable to the blocking factor ( $SS_{block}/SS_{Total} = 47852/72034 = 0.66$ ).

The top panels of Figure 14.4 strongly suggest that the means of both `yield` and `pre` are heterogeneous across blocks. Therefore, in subsequent analysis we adjust both `yield` and `pre` for blocks. From the bottom panels of Figure 14.4 we see that the adjusted values are more homogeneous. The adjusted response variables have the same sums of squares for the treatment and covariate effects in their ANOVA tables as the unadjusted response variables.

For example, the quantity `yield.block` in Table 14.16 represents `yield` adjusted for `blocks`. This is the original data vector of yields minus the vector of least-squares estimates of `block` effects. We can write the no-interaction model 2b as

$$Y_{ij} = \mu + \beta_i + \tau_j + \gamma(X_{ij} - \bar{X}) + \epsilon_{ij}$$

where  $Y_{ij}$  is the value of `yield` for `block`  $i$  and `treat`  $j$ ,  $\beta_i$  is the effect of `block`  $i$ ,  $\tau_j$  is the effect of `treat`  $j$ , and  $X_{ij}$  is the value of `pre` for

TABLE 14.17. Analyses of variance for apple data: Models 1, 2, and 2b.  
 (dsgntwo/code/apple3.s)

---

```

S-PLUS (dsgntwo/transcript/apple3.ste):
> apple.ancova.1 <- aov(yield ~ block + pre*treat, data=apple)
> anova(apple.ancova.1)
Analysis of Variance Table

Response: yield

Terms added sequentially (first to last)
  Df Sum of Sq Mean Sq F Value    Pr(F)
block   3  47852.83 15950.94 80.81854 0.0000008
      pre  1  15943.57 15943.57 80.78119 0.0000086
      treat  5   4352.89   870.58  4.41095 0.0262163
pre:treat  5   2108.90   421.78  2.13703 0.1520702
Residuals  9   1776.31    197.37

> apple.ancova.2 <- aov(yield ~ block + pre + treat, data=apple)
> anova(apple.ancova.2)
Analysis of Variance Table

Response: yield

Terms added sequentially (first to last)
  Df Sum of Sq Mean Sq F Value    Pr(F)
block   3  47852.83 15950.94 57.47786 0.0000004
      pre  1  15943.57 15943.57 57.45129 0.00000255
      treat  5   4352.89   870.58  3.13705 0.04170982
Residuals 14   3885.20    277.51

> apple.ancova.2b <- aov(yield ~ block + treat + pre, data=apple)
> anova(apple.ancova.2b)
Analysis of Variance Table

Response: yield

Terms added sequentially (first to last)
  Df Sum of Sq Mean Sq F Value    Pr(F)
block   3  47852.83 15950.94 57.47786 0.0000000
      treat  5     749.50   149.90  0.54015 0.7430149
      pre   1  19546.96 19546.96 70.43581 0.0000008
Residuals 14   3885.20    277.51

```

---

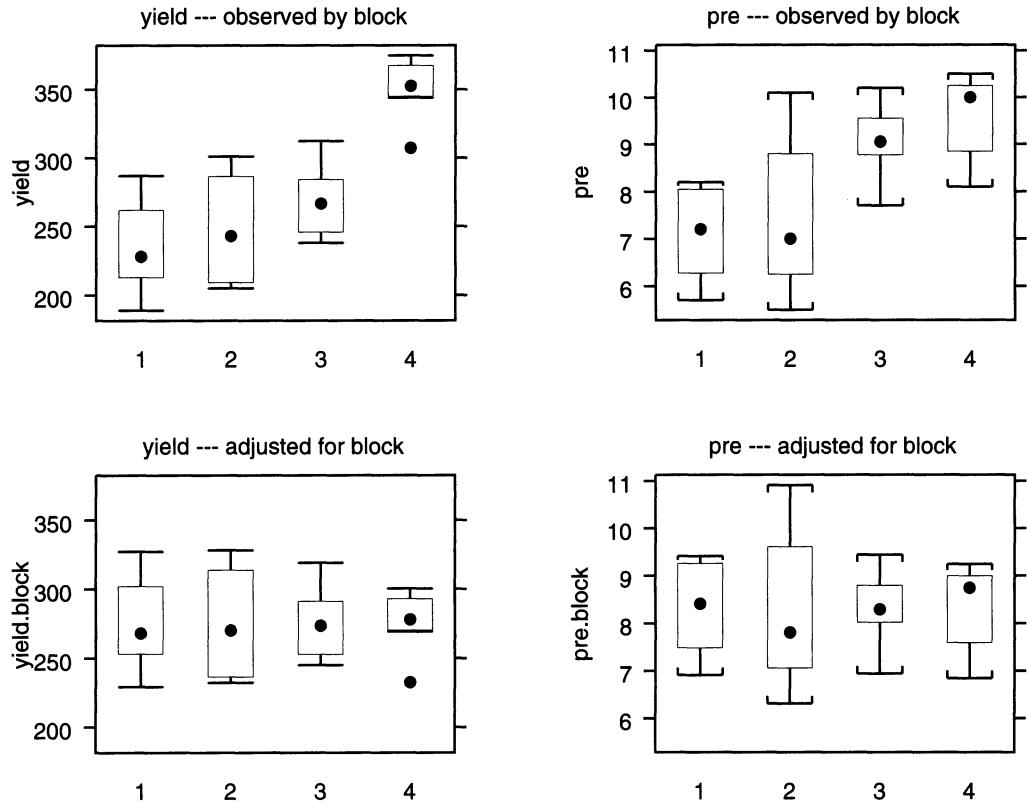


FIGURE 14.4. Apple `yield` and `pre` by `block`. In the top row the block effect is visible. In the bottom row, in which we have adjusted the data for the block effect, we see the data appear more homogeneous. (dsgntwo/code/apple3.s) (dsgntwo/figure/apple.y.p.eps.gz)

level  $i$  of `block` and level  $j$  of `treat`. Then

$$\text{yield.block}_{ij} = Y_{ij} - \hat{\beta}_i,$$

where  $\hat{\beta}_i$  is the least-squares estimate of the effect of `block`  $i$ .

Comparable definitions apply to `pre.block` and `yield.block.pre`. These adjustments are in the same sense used in several places in Section 9.14, and allow us to focus our attention on the remaining variables in the model.

The heterogeneity of responses across blocks is also visible in Row 1 of Figure 14.5. Here we display separate regressions of `yield` on `pre` for each of the 6 levels of `treat`, where the plot labels a,b,c,d represent

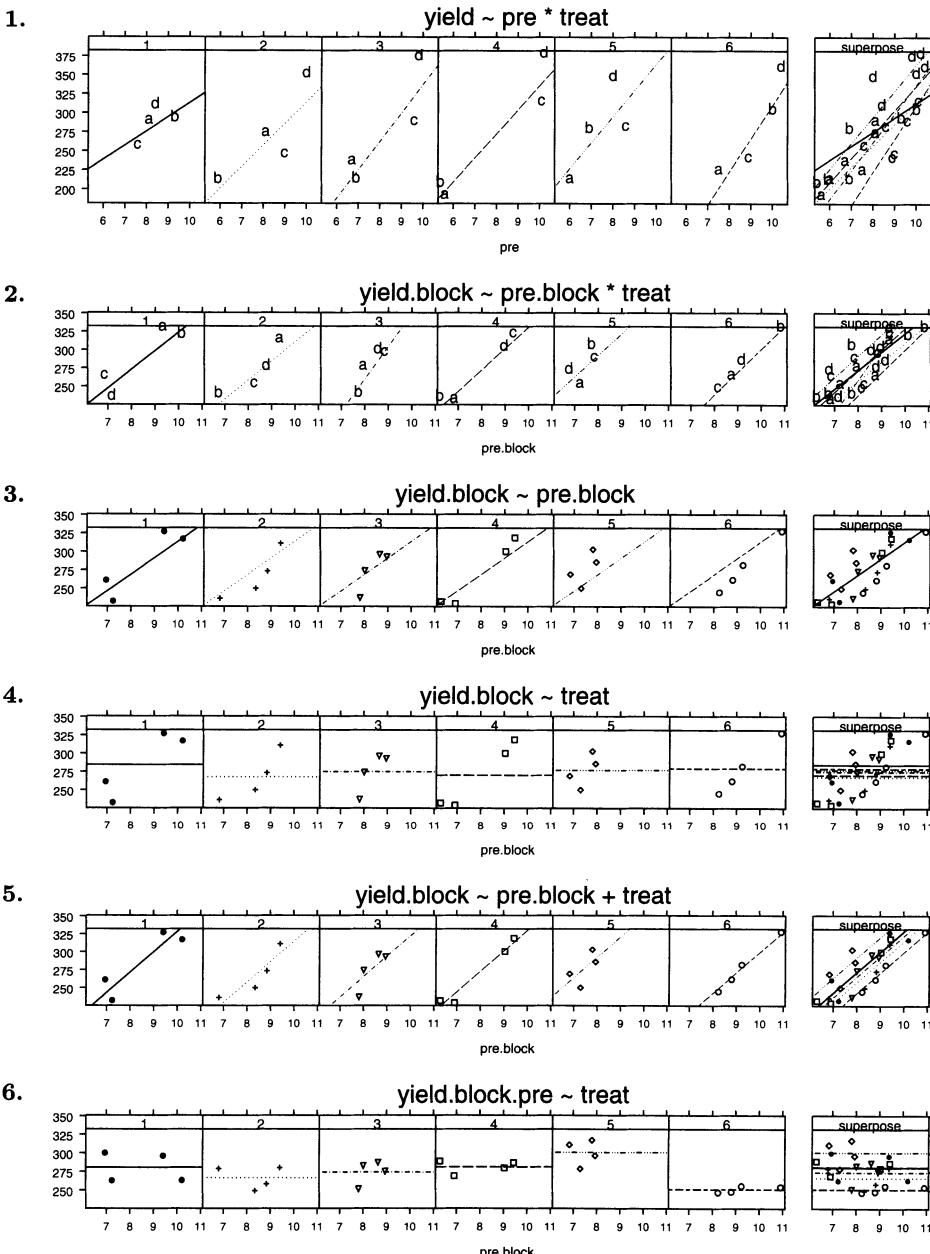


FIGURE 14.5. Several models for *yield*. See the text for the discussion of the models in each row. The plotting symbol in the top two rows indicates the block. The plotting symbol in the bottom four rows indicates the treatment.

(dsgntwo/code/apple3.s) (dsgntwo/figure/apple.ancova.eps.gz)

blocks 1–4. We see that block 4 has a consistently higher yield than the other 3 blocks.

[Note that we can discuss the proportion of total sum of squares for the blocking factor `block`, but not the  $F$ - and  $p$ -values—even though the program calculated and printed them in the ANOVA table (most ANOVA programs print these values). We assumed that blocks are important for this study and designed the study by stratifying the sample within blocks. Because the study was designed under the assumption that blocks are important, there is no testable hypothesis about blocks. We do have the right to calculate an estimate of the efficiency of the blocking. Most experimental design texts contain formulas for the efficiency attributable to blocking in Latin Square, Randomized Complete Block, and other experimental designs. See, for example, (Cochran and Cox, 1957) (Section 4.37).]

2. Since the ANOVA table shows  $p = .152$  for the interaction `pre:treat` we conclude that the slopes do not significantly differ. We do not illustrate this model here, but proceed immediately to the next model. We will come back to the interaction term when we discuss Figure 14.5.

Model 2, `yield ~ block + pre + treat`, forces parallel regression lines with possibly differing intercepts. The value  $p = 0.0417$  for `treat` tells us that there is a borderline significant difference in the `treat` adjusted means. Stepping forward for a moment, to be justified by the time we get there, the intercepts of the parallel regression lines in the last two rows of Figure 14.5 are borderline significantly different.

Model 2b, `yield ~ block + treat + pre`, gives the same regression coefficients, fitted values, and residuals as model 2. Models 2 and 2b have different sequential sums of squares for the `pre` and `treat` terms. In model 2, the sum of squares for `treat` after adjustment for the covariate `pre` is deemed significant. In model `apple.ancova.2b`, the effect of `treat` is assessed prior to consideration of `pre` and found to be not significant. Therefore, taking account of the available covariate `pre` has enabled us to detect differences in the adjusted mean `yield` at the levels of `treat`. Without the presence of this covariate, we would not have detected `treat` differences.

If the covariate `pre` had not been available for this analysis, the sums of squares for `block` and for `treat` would be identical to those in the anova table for model `apple.ancova.2b`. But without `pre` in the model, the corresponding terms in the anova table would have lower  $F$  statistics and higher  $p$ -values than those in Table 14.17 because the sum of squares for `pre` would instead be a large component of the sum of squares for `Residual`.

TABLE 14.18. Multiple comparisons for `apple.ancova.2`.  
 (dsgntwo/code/apple3.s)

---

```
S-PLUS (dsgntwo/transcript/apple2.2mc.ste):
> apple.mmc <-
+ multicomp.mmc(apple.ancova.2,
+                 comparisons="mcc", method="dunnett", valid.check=F,
+                 focus="treat", lmat.rows=7:12)
> ## export.eps(hh("tway/figure/apple.mmc.eps"))
> apple.mmc$mca

95 % simultaneous confidence intervals for specified
linear combinations, by the Dunnett method

critical point: 2.8197
response variable: yield

intervals excluding 0 are flagged by '****'

  Estimate Std.Error Lower Bound Upper Bound
5-6      49.6     13.3     12.10     87.1 ****
4-6      29.8     12.7     -5.91     65.5
1-6      29.1     12.1     -5.05     63.3
3-6      22.7     12.2    -11.70     57.2
2-6      15.2     12.2    -19.20     49.7

> plot(apple.mmc$mca)
> ## export.eps(hh("tway/figure/apple.multicomp.mca.eps"))
```

---

### 14.6.2 Multiple Comparisons

Now that we have detected `treat` differences, we wish to investigate their nature. Since level 6 of `treat` has been designated as a control, we follow up the finding of significant `treat` with Dunnett's multiple comparison procedure to simultaneously compare the mean adjusted `yield` of `treat` levels 1 through 5 with the mean adjusted `yield` of `treat` level 6. We illustrate the multiple comparisons in Figures 14.6 with the mean-mean display, in Figure 14.7 with a conventional display sorted by the `treat` means, and in Table 14.18.

These displays tell us that the adjusted mean `yield` of `treat` level 5 is significantly greater than the adjusted mean `yield` of the control (`treat` level 6). No other significant differences with `treat` level 6 are uncovered.

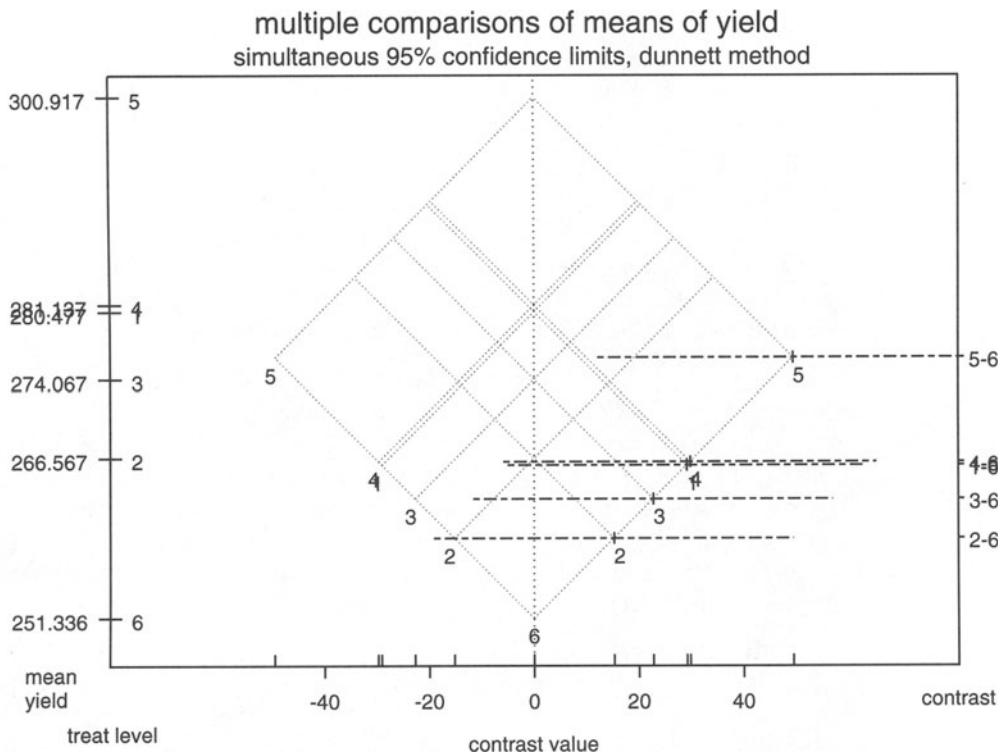


FIGURE 14.6. Apple multiple comparisons, mean–mean display.  
(dsgntwo/code/apple3.s) (dsgntwo/figure/apple mmc .eps .gz)

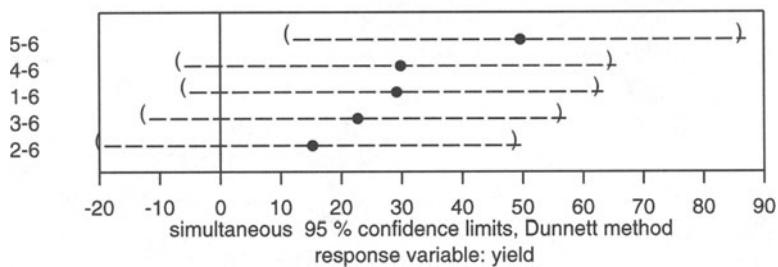


FIGURE 14.7. Apple multiple comparisons, standard S-PLUS display ordered by levels of treat.  
(dsgntwo/code/apple3.s) (dsgntwo/figure/apple multicomp.mca.eps.gz)

### 14.6.3 Models in Figure 14.5

The discussion of the progression from model 1 to models 2 and 2b in Section 14.6.1 has been based primarily on the ANOVA tables in Table 14.17. In this section we study Figure 14.5 to see the graphical presentation of each line of the ANOVA table for model 2.

In summary, the comparison of Rows 1 and 2 of Figure 14.5 shows the block effect. Row 5 shows the covariate and treatment effects adjusted for the blocking factor. Row 6 shows the treatment effect adjusted for the block and covariate. Row 4 shows that the treatment effect is not significant if we neglect to adjust for the covariate. Row 3 shows that the covariate alone, without the treatment effect, is not sufficient to model the data.

Model 2, `yield ~ block + pre + treat`, the model that fits the apple data well, is an ANCOVA model with a blocking factor. In Figure 14.5 we illustrate each of the lines of model 2's ANOVA table with a figure based on the ANCOVA display developed in Section 10.5. All six rows of Figure 14.5 are constructed from model specifications that are related to the recommended model 2. All six rows are on the same  $y$ -scale. The top two rows use the block to define the plotting symbol. The bottom four rows use the treatment to define the plotting symbol.

Row 1 of Figure 14.5 shows all four variables. There are six panels, one for each level of `treat`. Each panel displays  $y=\text{yield}$  by  $x=\text{pre}$ . The plotting symbol represents the blocking factor `block`. The “d” for block 4 is uniformly much higher than the symbols for the other blocks. Within each panel of Row 1 we show the least-squares line of  $\text{yield} \sim \text{pre}$  with the `block` factor completely ignored. These lines correspond to model 8 in Table 14.19.

Because the `block` effect so dominates Row 1 (and accounts for 0.66 of the total sum of squares in models 1 and 2), we must remove the effect of `block` from the variables before we can proceed. Rows 2 through 6 of Figure 14.5 use the adjusted variables  $y=\text{yield.block}$  by  $x=\text{pre.block}$  calculated in Table 14.20.

Row 2 is similar to Row 1. Each panel displays the adjusted variables  $y=\text{yield.block}$  by  $x=\text{pre.block}$  for a different treatment level. Each panel shows a separate least-squares line  $\text{yield.block} \sim \text{pre.block}$ . The range of the `yield` means across levels of `treat` has been halved. Row 2 displays model 3,  $\text{yield.block} \sim \text{pre.block*treat}$ , and again we mark the points in each panel with the level of the `block` factor. We see that the adjustment has made all blocks similar. The “d” is evenly distributed among the four adjusted means.

TABLE 14.19. Analysis of variance for apple data: Model 8.  
(dsgntwo/code/apple3.s)

---

```
S-PLUS (dsgntwo/transcript/apple3d.ste):
> apple.ancova.8 <- ancova(yield ~ pre * treat, data=apple)
> tmp8 <- anova(apple.ancova.8)[,1:2]
> tmp8
Analysis of Variance Table

Response: yield

Terms added sequentially (first to last)
  Df Sum of Sq
  pre   1  48413.19
  treat  5   6478.05
  pre:treat 5   1285.97
  Residuals 12  15857.30
```

---

Figure 14.4 shows the ranges of `yield` and `pre` before and after adjustment for `block`. The top row of Figure 14.5 corresponds to the top row of Figure 14.4. Not until the bottom five rows of Figure 14.5, corresponding to the bottom row of Figure 14.4 and to components of models `apple.ancova.2b` and `apple.ancova.2b`, do we see any significant effect for `treat`.

Row 2 shows the least-squares lines constructed with model 3. The regression coefficients and sums of squares for model 3 are identical to those for model 1. Since we concluded from model 1 that the interaction is not significant, equivalently that the slopes in the panels are not significantly different from each other, we continue our discussion with Rows 3, 4, and 5.

Row 3 of Figure 14.5 displays model 7, `yield.block ~ pre.block`, and shows the least-squares fit of ‘`yield` adjusted for `block`’ on ‘`pre` adjusted for `block`’ using the entire data set. Row 3 and model 7 draw a common line in each of the panels. This line is not a good fit to the points in most of these panels, suggesting that `treat` cannot be ignored as an explanatory variable.

The second conclusion we found with model 1 was the lack of significance for the `pre:treat` interaction term. We illustrate that lack in Rows 2 and 5 of Figure 14.5. Row 2 shows model 3, `yield.block ~ pre.block * treat` and Row 5 shows model 4, `yield.block ~ pre.block + treat`. The fitted least-squares lines in the individual panels of Row 2 are unique to each treatment. The fitted least-squares lines in the individual panels of Row 5

TABLE 14.20. Analyses of variance for apple data: Models 3 and 3b.  
 (dsgntwo/code/apple3.s)

---

```

S-PLUS (dsgntwo/transcript/apple3b.ste):
> ## find and remove block effect from response variable and covariate
> yield.block.effect <- fitted(lm(yield ~ block, data=apple))-mean(apple$yield)
> pre.block.effect <- fitted(lm(pre ~ block, data=apple))-mean(apple$pre)
> yield.block <- apple$yield-yield.block.effect
> pre.block <- apple$pre-pre.block.effect
> apple <- cbind(apple, yield.block=yield.block, pre.block=pre.block)

> ## Same sums of squares as apple.ancova.1 and apple.ancova.2
> ## for pre and treat adjusted for block
> ## The sum of the pre:treat and residual sum of squares is correct.
> ## The residual df includes the block df and is therefore wrong.
> ## Therefore we suppress the residual Means Square and the F tests

> apple.ancova.3 <- ancova(yield.block ~ pre.block*treat, data=apple,
+                                blocks=apple$block)
> tmp3 <- anova(apple.ancova.3)[,1:3]
> tmp3[4,3] <- NA ; tmp3
Analysis of Variance Table

Response: yield.block

Terms added sequentially (first to last)
  Df Sum of Sq Mean Sq
pre.block 1 15943.57 15943.57
  treat 5 4352.89 870.58
pre.block:treat 5 346.21 69.24
  Residuals 12 3538.99
> apple.ancova.3b <- ancova(yield.block ~ treat*pre.block, data=apple,
+                                blocks=apple$block)
> tmp3b <- anova(apple.ancova.3b)[,1:3]
> tmp3b[4,3] <- NA ; tmp3b
Analysis of Variance Table

Response: yield.block

Terms added sequentially (first to last)
  Df Sum of Sq Mean Sq
  treat 5 749.50 149.90
pre.block 1 19546.96 19546.96
treat:pre.block 5 346.21 69.24
  Residuals 12 3538.99

```

---

have a common slope and unique intercepts. The `superpose` panel in both rows look similar, consistent with the conclusion from the ANOVA table for model 1 that the interaction term (the carrier of different slopes) is not needed to fit the data.

The ANOVA table for model 7, in Table 14.21, has the same sums of squares and degrees of freedom for covariate `pre` adjusted for `block` as model 2 shows for `pre`. In model 2, the adjustment is implicit because we show the sequential ANOVA table. In model 7, the adjustment is explicit because we removed (in Table 14.21) the block effect from `yield` and `pre`, to get `yield.block` and `pre.block`. The `Residuals` sum of squares for model 7 is equal to the sum of the `treat` and `Residuals` sums of squares for model 2. The `Residuals` degrees of freedom for model 7 includes all degrees of freedom that aren't assigned to the covariate. Therefore, we suppressed the mean square, the  $F$ , and the  $p$ -value columns from the printed ANOVA table.

Row 4 (model 6, `yield.block ~ treat`) of Figure 14.5 compares the mean `yield` adjusted for blocks across the six levels of `treat` but ignoring the covariate `pre`. The heights of the horizontal lines drawn at each mean do not differ appreciably. Referring to model 2b in Table 14.17, we see that without `pre` in the model, the `treat` means are not significantly different; the  $p$ -value for `treat` is 0.74. The ANOVA table for model 6 is in Table 14.21.

Row 5 (model 4, `yield.block ~ pre.block + treat`) of Figure 14.5 portrays separate regressions of `yield` adjusted for `block` on `pre` adjusted for blocks for each of the six levels of `treat`. The effect of `treat` is demonstrated by the differences in intercept across the six panels. The ANOVA table for model 4 in Table 14.22 has the same sums of squares for covariate, treatment, and residuals as does model 2. The ANOVA table for model 2 shows that the `treat` differences are significant. Graphically, we state that the vertical distances between the parallel least-squares lines are significantly different.

Most people have trouble seeing vertical distances between nonhorizontal parallel lines. Therefore, we use Row 6 to make the lines horizontal and retain the same vertical differences. We do so by making an additional adjustment to the response variable, subtracting out the effect of the covariate `pre.block` from the response variable `yield.block`. Details are in Table 14.21.

Row 6 (model 5, `yield.block.pre ~ treat`) of Figure 14.5 adjusts `yield` for both `treat` and `pre`. The differences in the heights of the lines drawn at adjusted `yield` means for each `treat` level displays the extent of the effect of `treat` on `yield` after accounting for both `block` and `pre`. The difference

TABLE 14.21. Analyses of variance for apple data: Models 6, 5, and 7.  
 (dsgntwo/code/apple3.s)

---

```

S-PLUS (dsgntwo/transcript/apple3c.ste):
> apple.ancova.6 <- ancova(yield.block ~ treat, x=pre.block, data=apple)
> tmp6 <- anova(apple.ancova.6)[,1:3]
> tmp6[2,3] <- NA ; tmp6
Analysis of Variance Table

Response: yield.block

Terms added sequentially (first to last)
  Df Sum of Sq Mean Sq
treat 5    749.50   149.9
Residuals 18  23432.17

> yield.block.pre <-
+   apple$yield.block -
+   predict.lm(apple.ancova.4, type="terms", terms="pre.block")
>
> apple <- cbind(apple, yield.block.pre=yield.block.pre)
> apple.ancova.5 <- ancova(yield.block.pre ~ treat, x=pre.block, data=apple)
> tmp5 <- anova(apple.ancova.5)[,1:2] ; tmp5
Analysis of Variance Table

Response: yield.block.pre

Terms added sequentially (first to last)
  Df Sum of Sq
treat 5  5471.949
Residuals 18  3885.204
> attr(apple.ancova.5, "trellis")$ylim <- attr(apple.ancova.3,"trellis")$ylim

> apple.ancova.7 <- ancova(yield.block ~ pre.block, groups=treat, data=apple)
> tmp7 <- anova(apple.ancova.7)[,1:3]
> tmp7[2,3] <- NA ; tmp7
Analysis of Variance Table

Response: yield.block

Terms added sequentially (first to last)
  Df Sum of Sq Mean Sq
pre.block 1  15943.57 15943.57
Residuals 22  8238.10

```

---

TABLE 14.22. Analyses of variance for apple data: Models 4 and 4b.  
 (dsgntwo/code/apple3.s)

---

```

S-PLUS (dsgntwo/transcript/apple3e.ste):
> ## The residual Df includes the block df and is therefore wrong.
> ## Therefore we suppress the residual Means Square and the F tests

> apple.ancova.4 <- ancova(yield.block ~ pre.block + treat, data=apple)
> tmp4 <- anova(apple.ancova.4)[,1:3]
> tmp4[3,3] <- NA
> tmp4
Analysis of Variance Table

Response: yield.block

Terms added sequentially (first to last)
  Df Sum of Sq Mean Sq
pre.block  1 15943.57 15943.57
  treat     5   4352.89   870.58
Residuals 17   3885.20

> apple.ancova.4b <- ancova(yield.block ~ treat + pre.block, data=apple)
> tmp4b <- anova(apple.ancova.4b)[,1:3]
> tmp4b[3,3] <- NA
> tmp4b
Analysis of Variance Table

Response: yield.block

Terms added sequentially (first to last)
  Df Sum of Sq Mean Sq
  treat    5    749.50   149.90
pre.block  1 19546.96 19546.96
Residuals 17   3885.20

```

---

in intercepts of the horizontal regression lines in Row 6 is identical to the difference in intercepts in Row 5. The ANOVA table for model 5 is in Table 14.21.

In all six rows of Figure 14.5 the discrepancy in lines across the six panels is summarized in the superpose panels in the rightmost column of this figure.

Tables 14.20, 14.21, and 14.22 contain calculations used to construct the adjusted variables plotted in Rows 2 to 6 of Figure 14.5. The analyses in

these tables show correct sums of squares and incorrect degrees of freedom. Therefore, they produce incorrect  $F$ -statistics and  $p$ -values. Consequently, the  $F$ -statistics and  $p$ -values are suppressed from these ANOVA tables.

## 14.7 Example—`testscore.dat`

The example in Section 14.6 was modeled with one covariate. This example requires two covariates.

### Study Objectives

(Johnson and Tsao, 1945), also reprinted in (Anderson and Bancroft, 1952), discuss the modeling of a final test score using four factors and two concomitant variables. Their goal was to explain the response variable `final` score by using the information in the factors and continuous variables. As a secondary goal they used this paper to illustrate the advantages of using concomitant variables (covariates) to increase the sensitivity of their experiment and the precision with which estimates and predictions can be made.

### Data Description

`final`: response variable: final test score  
`sex`: male=1, female=0  
`scholastic standing`: good=1, average=2, poor=3  
`order`: order on a battery of standardized tests: high=1, medium=2, low=3  
`grade`: 10, 11, 12  
`initial`: initial test score  
`mental.age`: mental age

Note that this was not a longitudinal study: Different students were assessed at each grade level.

The study has no replication. The original authors used the four-way interaction as their initial error term. They found that all of the three-way interactions and some of the two-way interactions were not significant so they continued by pooling all the nonsignificant interactions. Our analysis begins by assuming that the four-way and all three-way interactions are negligible. We proceed to investigate the two-way interactions and main effects, both with and without adjustment for the continuous variables.

## Analysis—Plots

The goal of the analysis is to explain the variability in the `final` scores based on the initial score, the covariates, and the factors. We start the analysis by looking at the scatterplot matrix of the data in Figure 14.8. Two things are immediately visible. First, the plots of the factor levels against each other form simple lattices. This reflects the design of the study in which the factors were chosen to be balanced. Second, we see that the response variable `final` is positively correlated with the two continuous variables `initial` and `mental.age` and negatively correlated with two of the factors `standing` and `order`. When we look closer at the factors, we discover that they are coded with high scores first. Therefore, the visual negative correlation is an artifact of the coding. We recode both `standing` and `order` to place the low scores first and thereby have all variables coded in a consistent direction.

We redraw the plot with the recoded variables in Figure 14.9. While here, we also changed the order of the variables, placing all the continuous variables together and all the factors together. There is now a clear display of the positive correlations among all continuous variables and ordered factors. We would display Figure 14.9 rather than Figure 14.8 in our final report.

## Analysis—ANOVA

In an analysis with both factors and continuous concomitant variables, it is necessary to determine the contribution of each. We therefore fit four different models,

1. factors only,
2. continuous first: fitting factors after continuous predictors,
3. continuous second: fitting continuous predictors after factors,
4. continuous only.

Details for all four are in the online files in (`dsgntwo/transcript/testscore.st`). We summarize all four in Table 14.23. The first thing to notice is the residual mean square  $\hat{s}^2$  for the models. The reduction in sum of squares from “factors only” to both factors and continuous (observed in either column “continuous first” or “continuous second”) has an  $F_{2,26} = ((97.704 - 28.057)/2)/1.0791 = 32.27$  with  $p = 10^{-7}$ . The reduction from continuous only to both has  $F_{25,26} = ((131.291 - 28.057)/25)/1.0791 = 3.827$  with  $p = 0.0006$ . The details are in the file (`dsgntwo/transcript/testscore-test.st`).

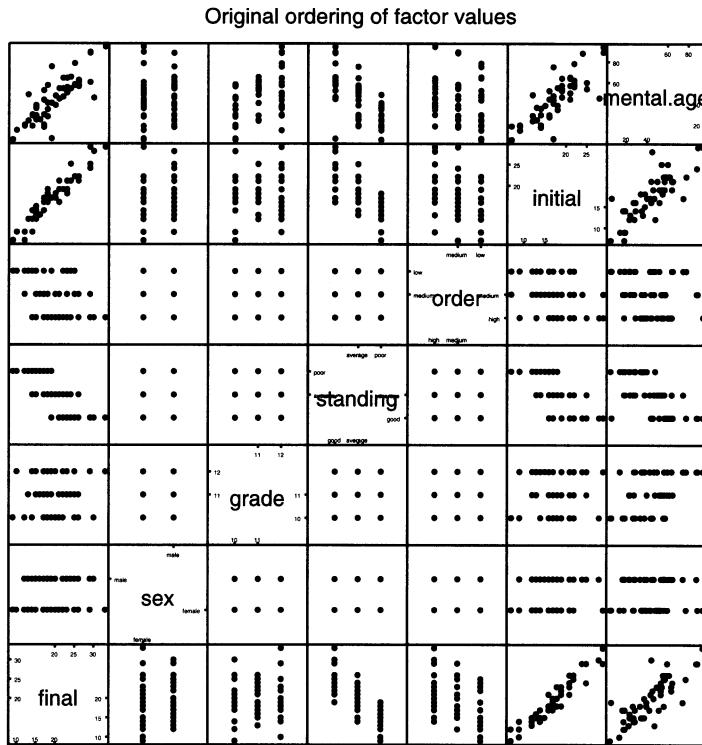


FIGURE 14.8. `testscore` data, original ordering of variables and factor levels.  
 (dsgntwo/code/testscore.s), (dsgntwo/figure/testscore.f.spalom1.eps.gz)

We look further at the listing for the continuous-variables-first analysis in Table 14.24. We see that the `initial` score all by itself has  $R^2 = 1468.423/1607.333 \approx 0.9$ . We determined from the results of `testscore2.aov` in Table 14.23 that none of the two-way interactions (with the possible exception of `sex:order`) is significant; therefore, we dropped them. We also see the main effect `sex` is not significant as the first factor; therefore, we moved it last. In the ANOVA `testscore5.aov` we see that `grade` is the least important of the main effects, so we moved it to next-to-last. In the ANOVA `testscore6.aov` we see that we don't need `grade` at all. This brings us to the ANOVA `testscore7.aov` with three significant main effects after adjusting for the covariates.

Let us plot the information in ANOVA table `testscore7.aov`. As is usual with analysis of covariance, we are interested in the difference between the

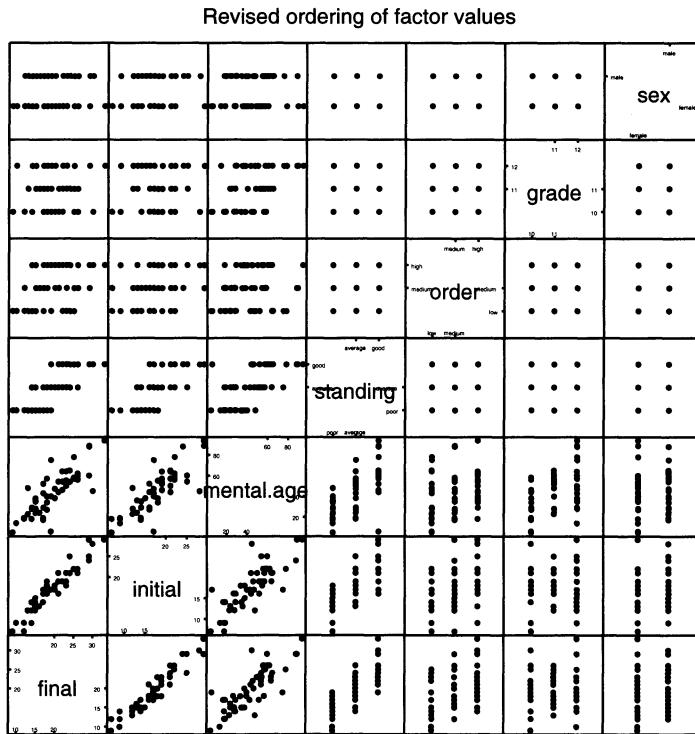


FIGURE 14.9. `testscore` data, revised ordering of variables and factor levels.  
`(dsgntwo/code/testscore.s)`, `(dsgntwo/figure/testscore.f.splover2.eps.gz)`

intercepts for each of the groups assuming parallel planes defined by the coefficients of the covariates. In the `testscore7.aov` setting we have 2 covariates and 18 groups with model

$$Y_{ijkl} = \hat{\mu}_{ijk} + \hat{\beta}_1 X_{1,ijkl} + \hat{\beta}_2 X_{2,ijkl} + \hat{\epsilon}_{ijkl}$$

We move the  $X$  terms to the other side to get `final.adj`, the final scores adjusted for the covariates

$$\hat{Y}_{\text{adj}} = \hat{Y}_{ijkl} - \hat{\beta}_1 X_{1,ijkl} - \hat{\beta}_2 X_{2,ijkl} = \hat{\mu}_{ijk} + \hat{\epsilon}_{ijkl}$$

and plot the  $\hat{Y}_{\text{adj}}$  in Figure 14.10. The standard error of the difference of two observations in the plot is approximately  $\sqrt{(2)(1.324)/3} \approx 0.94$ . Therefore, the visible differences in the plot are significant. We can easily see all three main effects. We can also see the hint of interaction between `sex` and `order` in the reversal of direction in the `female:medium` panel.

TABLE 14.23. Comparison of sequential *p*-values for four different models of the `testscore` data. In each of the models, the response variable is the `final` score.  
`(dsgntwo/code/testscore.s)`, `(dsgntwo/transcript/testscore.st)`,  
`(dsgntwo/transcript/testscore-comp.st)`

|                   | Factors<br>only         |          | Continuous<br>first     |          | Continuous<br>second     |          | Continuous<br>only      |          |
|-------------------|-------------------------|----------|-------------------------|----------|--------------------------|----------|-------------------------|----------|
|                   | <code>testscore1</code> |          | <code>testscore2</code> |          | <code>testscore3s</code> |          | <code>testscore4</code> |          |
|                   | Df                      | Pr(F)    | Df                      | Pr(F)    | Df                       | Pr(F)    | Df                      | Pr(F)    |
| initial           |                         |          | 1                       | 0.0000   | 1                        | 0.0000   | 1                       | 0.0000   |
| mental.age        |                         |          | 1                       | 0.0132   |                          |          | 1                       | 0.0914   |
| sex               | 1                       | 0.0069   | 1                       | 0.2522   | 1                        | 0.0000   |                         |          |
| grade             | 2                       | 0.0010   | 2                       | 0.0029   | 2                        | 0.0000   |                         |          |
| standing          | 2                       | 0.0000   | 2                       | 0.0005   | 2                        | 0.0000   |                         |          |
| order             | 2                       | 0.0000   | 2                       | 0.0000   | 2                        | 0.0000   |                         |          |
| sex:grade         | 2                       | 0.0184   | 2                       | 0.6567   | 2                        | 0.0000   |                         |          |
| sex:standing      | 2                       | 0.7728   | 2                       | 0.7506   | 2                        | 0.4427   |                         |          |
| sex:order         | 2                       | 0.6719   | 2                       | 0.0792   | 2                        | 0.2885   |                         |          |
| grade:standing    | 4                       | 0.0519   | 4                       | 0.1519   | 4                        | 0.0001   |                         |          |
| grade:order       | 4                       | 0.8250   | 4                       | 0.1319   | 4                        | 0.3304   |                         |          |
| standing:order    | 4                       | 0.6085   | 4                       | 0.2426   | 4                        | 0.0952   |                         |          |
| initial           |                         |          |                         |          | 1                        | 0.0000   |                         |          |
| mental.age        |                         |          |                         |          | 1                        | 0.1066   |                         |          |
| Residuals Df      | 28                      |          | 26                      |          | 26                       |          | 51                      |          |
| Residuals Mean Sq |                         | 3.4894   |                         | 1.0791   |                          | 1.0791   |                         | 2.5743   |
| Residuals Sum Sq  |                         | 97.704   |                         | 28.057   |                          | 28.057   |                         | 131.291  |
| Total Sum Sq      | 53                      | 1607.333 | 53                      | 1607.333 | 53                       | 1607.333 | 53                      | 1607.333 |

## Summary of ANOVA

We conclude the search for a model with our final model, `testscore7.aov`. We find three main effects: `standing`, `order`, and `sex` after accounting for the two covariates: `initial` and `mental.age`. These are the same variables the original authors found.

The original investigation of this data used the analysis to predict `final` scores given the variables in the data description.

Table 14.25 contains predictions for the `final` test score for the average values of the two covariates and for all possible values of the three factors we retained. We continue in file `(dsgntwo/transcript/testscore-pred.st)` to construct marginal predictions.

TABLE 14.24. Continuous variables first, and further refinements.  
 (dsgntwo/code/testscore.s)

---

```
S-PLUS (dsgntwo/transcript/testscore-5.st):
> ## after looking at all of above
> var(testscore$final, S=T)
[1] 1607.333
> testscore5.aov <- aov(final ~ initial + mental.age +
+                         grade + standing + order + sex
+                         + sex:order, data=testscore)
> summary(testscore5.aov)
    Df Sum of Sq Mean Sq F Value    Pr(F)
initial   1 1468.423 1468.423 1176.018 0.0000000
mental.age 1    7.619    7.619   6.102 0.0176491
grade     2   15.827   7.913   6.338 0.0039321
standing   2   20.556  10.278   8.231 0.0009635
order     2   29.506  14.753  11.815 0.0000849
sex       1    7.100    7.100   5.686 0.0216920
sex:order  2    5.860    2.930   2.347 0.1081102
Residuals 42   52.443   1.249

> testscore6.aov <- aov(final ~ initial + mental.age +
+                         standing + order + grade + sex,
+                         data=testscore)
> summary(testscore6.aov)
    Df Sum of Sq Mean Sq F Value    Pr(F)
initial   1 1468.423 1468.423 1108.182 0.0000000
mental.age 1    7.619    7.619   5.750 0.0207992
standing   2   29.053  14.526  10.963 0.0001370
order     2   33.830  16.915  12.765 0.0000425
grade     2    3.006   1.503   1.134 0.3309351
sex       1    7.100    7.100   5.358 0.0253536
Residuals 44   58.303   1.325

> testscore7.aov <- aov(final ~ initial + mental.age +
+                         standing + order + sex,
+                         data=testscore)
> summary(testscore7.aov)
    Df Sum of Sq Mean Sq F Value    Pr(F)
initial   1 1468.423 1468.423 1108.958 0.0000000
mental.age 1    7.619    7.619   5.754 0.02056336
standing   2   29.053  14.526  10.970 0.00012719
order     2   33.830  16.915  12.774 0.00003869
sex       1    7.498    7.498   5.662 0.02153500
Residuals 46   60.911   1.324
```

---

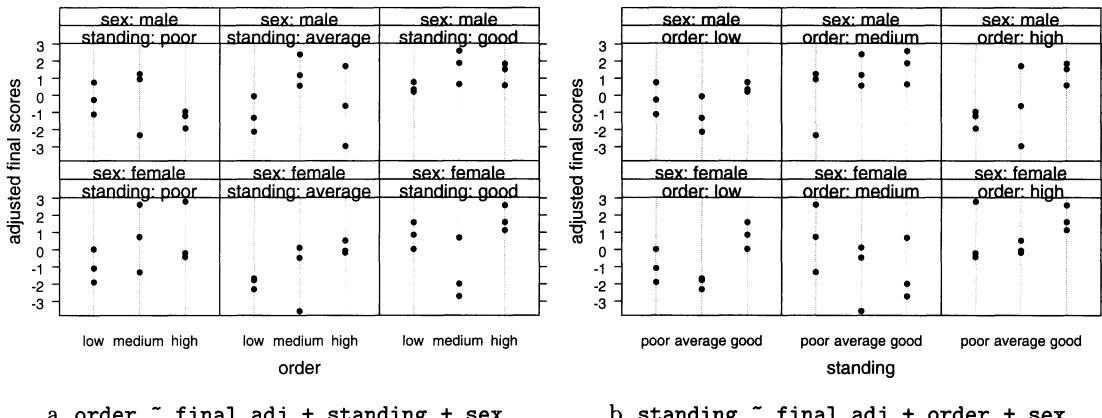


FIGURE 14.10. Final scores adjusted for the covariates using model `testscore7.aov`. Both subfigures show exactly the same information. The only difference is the interchange of the roles of the factors `order` and `standing` from  $x$ -variable to conditioning variable.

(`dsgntwo/code/testscore.s`), (`dsgntwo/figure/final.adjo.eps.gz`),  
 (`dsgntwo/figure/final.adjs.eps.gz`)

## 14.8 The Tukey One Degree of Freedom for Nonadditivity

### 14.8.1 Example—Crash Data

#### Study Objectives

Does the tendency of teenage drivers to be involved in automobile accidents increase dramatically with the number of passengers in the car? Believing this to be so, legislators in several states require that junior license holders be prohibited from driving with more than one passenger. (Williams, 2001) presents the data (`datasets/crash.dat`).

#### Data Description

`crashrate`: crashes per 10,000 trips

`agerange`: driver age: 16–17, 18–19, 30–59

`passengers`: number of passengers: 0, 1, 2, 3+

The original author displayed the data in the barplot format of Figure 14.11. We prefer the interaction plot format of Figure 14.12 to display the data and the interactions between the two factors. The main problem with barplots is that most of the space of the figure is used to display places where the data isn't. As usual, when we present an interaction plot we

TABLE 14.25. Predictions of **final** score for each of the categories using model **testscore7.aov**.  
 (dsgntwo/code/testscore.s), (dsgntwo/transcript/testscore-pred.st)

---

```
S-PLUS (dsgntwo/transcript/testscore-pred-edit.st):
> ## prediction
> newdata <- cbind(initial=mean(testscore$initial),
+                     mental.age=mean(testscore$mental.age),
+                     testscore[c(1:9,28:36), c("standing","order","sex")])
> final.pred <- predict(testscore7.aov, newdata=newdata)
> final.pred <- tapply(final.pred, newdata[,3:5], c)
> class(final.pred) <- "table"
> final.pred
Dim 1 : standing
Dim 2 : order
Dim 3 : sex

, , female
      low   medium    high
poor 16.04599 17.69223 18.68615
average 17.47542 19.12165 20.11557
good 20.37340 22.01964 23.01356

, , male
      low   medium    high
poor 16.81409 18.46032 19.45424
average 18.24351 19.88974 20.88366
good 21.14149 22.78773 23.78165
```

---

show two versions, interchanging the trace- and  $x$ -factors between them. In this example the traces in Figure 14.12b match the heights of the bars in each panel of Figure 14.11. We find the lack of parallelism in the traces in the same panel to be more compelling than the same information spread across several panels.

Table 14.26 contains an initial analysis of variance. It suggests that the number of **passengers** does not impact significantly on **crashrate**. The figures show completely different means and variances for different levels of **passengers**, implying that some type of interaction is present. The usual way of investigating interaction can't be used for this example because there is only one observation for each combination of **agerange** and **passengers**. The entire residual sum of squares has only 6 degrees of freedom. If all of them are used to estimate the interaction, then none are left for the residual, and testing becomes impossible.

### Crash Rates by Driver Age and Passenger Presence per 10,000 Trips

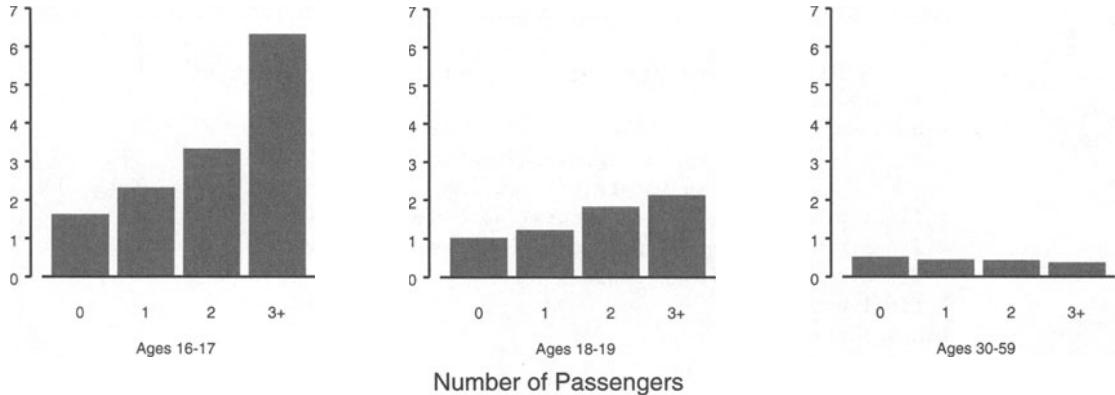


FIGURE 14.11. Barplot of crash data.  
(dsgntwo/code/crash.s), (dsgntwo/figure/crash-bar.eps.gz)

### Interactions for Crash Rates by Driver Age and Passenger Presence

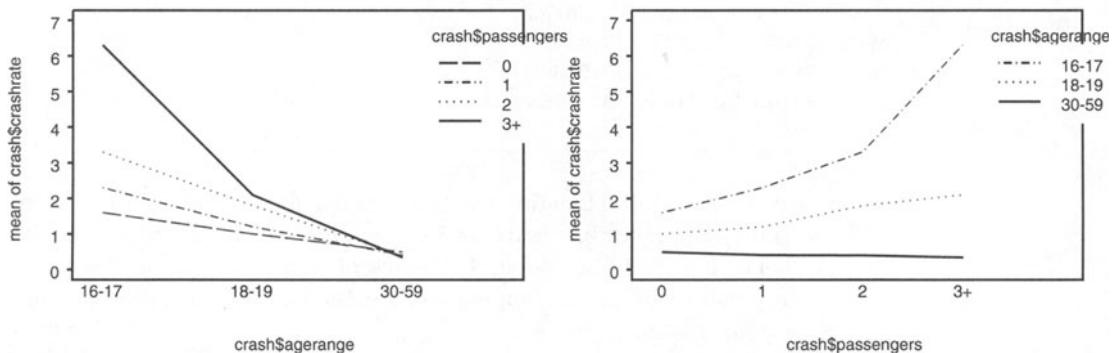


FIGURE 14.12. Interaction plot of crash data. a. Each line traces the number of passengers. b. Each line traces the age of the driver.

(dsgntwo/code/crash.s), (dsgntwo/figure/crash-interaction.eps.gz)

The Tukey *One Degree of Freedom for Nonadditivity* (ODOFFNA) (Tukey, 1949) and (Hoaglin et al., 1983) is the proper response to this impasse. Tukey constructs the *comparison values*, the linear-by-linear component of the interaction. He identifies it as one of the degrees of freedom in the residual and tests it for significance against the remaining degrees of

TABLE 14.26. ANOVA for crash data.

(dsgntwo/code/crash.s)

S-PLUS (dsgntwo/transcript/crash.aov.st):

```
> crash.aov <- aov(crashrate ~ agerange + passengers, data=crash, qr=T)
> summary(crash.aov)
   Df Sum of Sq Mean Sq F Value    Pr(F)
agerange   2    17.9426 8.971300 7.238227 0.0251588
passengers 3     6.2298 2.076600 1.675443 0.2701424
Residuals   6     7.4366 1.239433
```

freedom in the residual. There are several possible options for dealing with a significant ODOFFNA: Find an outlier, transform the data, regress on the comparison value.

We construct the comparison value by partitioning the data into four components. We augment the  $r \times c$  table with a column of row margins  $r_i^0 = 0$  and a row of column margins  $c_j^0 = 0$  and a summary number in the last row-last column  $\hat{\mu}^0 = 0$ . The table at the right symbolically indicates the four sections of the augmented table. Note that each value of the table has been partitioned into four pieces, one in each section of the augmented table. We can trivially reconstruct the original table with  $Y_{ij} = \hat{\mu}^0 + r_i^0 + c_j^0 + Y_{ij}$ .

|       | 0    | 1    | 2    | 3+   | row  |       | 0     | 1 | 2        | 3+ | row           |
|-------|------|------|------|------|------|-------|-------|---|----------|----|---------------|
| 16–17 | 1.60 | 2.30 | 3.30 | 6.30 | 0.00 | 16–17 |       |   |          |    |               |
| 18–19 | 1.00 | 1.20 | 1.80 | 2.10 | 0.00 | =     | 18–19 |   | $Y_{ij}$ |    | $r_i^0$       |
| 30–59 | 0.49 | 0.41 | 0.40 | 0.34 | 0.00 |       | 30–59 |   |          |    |               |
| col   | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |       | col   |   | $c_j^0$  |    | $\hat{\mu}^0$ |

In the second display of the table, we find the row means  $r_i^1$  for each of the rows of the first table and place them in the column of row margins. We also subtract the row means from each value of the original table. Now the reconstruction of the original table with  $Y_{ij} = \hat{\mu}^0 + r_i^1 + c_j^0 + e_{ij}^1$  is still possible, but is no longer trivial.

|       | 0      | 1      | 2      | 3+     | row   |       | 0     | 1                           | 2                   | 3+ | row           |
|-------|--------|--------|--------|--------|-------|-------|-------|-----------------------------|---------------------|----|---------------|
| 16–17 | -1.775 | -1.075 | -0.075 | 2.925  | 3.375 | 16–17 |       |                             |                     |    |               |
| 18–19 | -0.525 | -0.325 | 0.275  | 0.575  | 1.525 | =     | 18–19 | $e_{ij}^1 = Y_{ij} - r_i^1$ | $r_i^1 = \bar{Y}_i$ |    |               |
| 30–59 | 0.080  | 0.000  | -0.010 | -0.070 | 0.410 |       | 30–59 |                             |                     |    |               |
| col   | 0.0000 | 0.0000 | 0.0000 | 0.000  | 0.000 |       | col   |                             | $c_j^0$             |    | $\hat{\mu}^0$ |

In the third display of the table, we find the column means  $c_j$  for each of the columns of the second table and place them in the row of column margins. We also subtract these column means from each value of the body of the second table. Now the reconstruction of the original table with  $Y_{ij} = \hat{\mu} + r_i + c_j + e_{ij}$  is meaningful. For example, the entry 1.60 in the upper-left corner of the original table is calculated with components in the final table:  $1.770 + 1.605 - .740 - 1.035$ .

|       | 0      | 1       | 2       | 3+      | row    |         | 0                           | 1 | 2 | 3+ | row |
|-------|--------|---------|---------|---------|--------|---------|-----------------------------|---|---|----|-----|
| 16–17 | −1.035 | −0.6083 | −0.1383 | 1.7817  | 1.605  | 16–17   |                             |   |   |    |     |
| 18–19 | 0.215  | 0.1417  | 0.2117  | −0.5683 | −0.245 | = 18–19 | $e_{ij} = e_{ij}^1 - c_j$   |   |   |    |     |
| 30–59 | 0.820  | 0.4667  | −0.0733 | −1.2133 | −1.360 | 30–59   |                             |   |   |    |     |
| col   | −0.740 | −0.4667 | 0.06333 | 1.1433  | 1.770  | col     | $c_j = \bar{e}_{\cdot j}^1$ |   |   |    |     |

This process is called row and column polishing by means. The notation was introduced by Tukey with medians and is described in the “Median Polish” chapter of (Hoaglin et al., 1983). Median polish is an iterative technique that usually takes at least two cycles. The analogous mean polish illustrated here uniquely converges on just the one cycle illustrated here. Medians are resistant to outliers whereas means are not. The advantage of means is that they correspond to least-squares fits.

We construct the comparison value by recombining the components

$$\text{comparison value} = \frac{r_i \times c_j}{\hat{\mu}} = x_{cv}$$

|       | 0       | 1       | 2        | 3+      |
|-------|---------|---------|----------|---------|
| 16–17 | −0.6710 | −0.4232 | 0.05743  | 1.0368  |
| 18–19 | 0.1024  | 0.0646  | −0.00877 | −0.1583 |
| 30–59 | 0.5685  | 0.3586  | −0.04866 | −0.8785 |

We then compute the ANOVA table (in Table 14.27) with the comparison value as a covariate.

The regression coefficient of the comparison value, in this example  $\hat{\beta}_{cv} = 1.5516$ , is identical to the regression of the residuals from the first ANOVA against the comparison value as illustrated in Figure 14.13. Tukey calls this the *diagnostic plot*.

There are several things to do with this plot. If most of the points lie in a horizontal band, then the ones that don't are potential outliers and need to be carefully investigated. If there is a clear nonhorizontal line, then the slope of the line gives a hint as to an appropriate transformation. When the slope is  $\hat{\beta}_{cv}$ , then an appropriate power transformation that will stabilize the variance of the groups is given by  $k = 1 - \hat{\beta}_{cv}$ . In this example, the suggested

TABLE 14.27. ANOVA with the comparison value as a covariate. See also Figure 14.13. The regression coefficient for the comparison value is 1.5516.

(dsgntwo/code/crash.s)

```
S-PLUS (dsgntwo/transcript/crash2.aov.st):
> crash2.aov <- aov(crashrate ~ agerange + passengers + as.vector(cv),
+                      data=crash, qr=T)
> summary(crash2.aov)
   Df Sum of Sq  Mean Sq F Value    Pr(F)
agerange  2 17.94260 8.971300 161.2531 0.0000287989
passengers 3  6.22980 2.076600 37.3255 0.0007576544
as.vector(cv) 1  7.15843 7.158425 128.6679 0.0000931382
Residuals  5  0.27817 0.055635

> coef(crash2.aov)
(Intercept) agerange1 agerange2 passengers.L passengers.Q passengers.C as.vector(cv)
      1.77      -0.925     -0.68      1.38189     0.4033333    0.06559133      1.551647
```

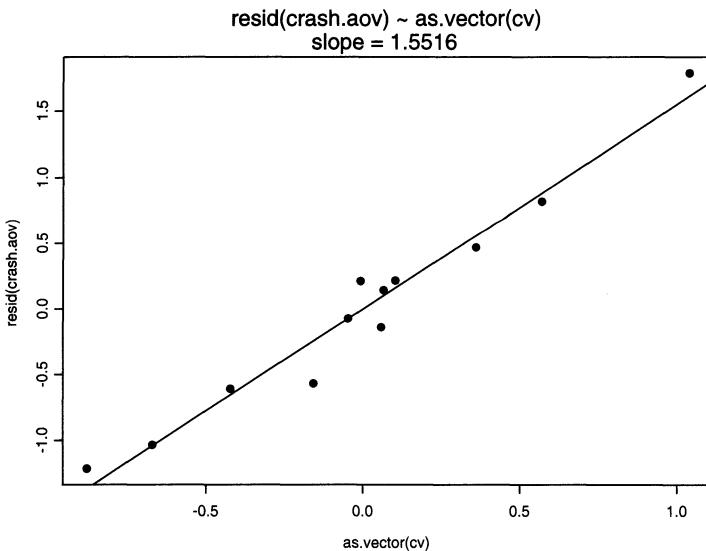


FIGURE 14.13. Diagnostic plot for crash data. See also Table 14.27. The data follow the regression line. Therefore, the slope of 1.5516 suggests a power transformation of  $1 - 1.5516 = -.5516$ .

(dsgntwo/code/crash.s), (dsgntwo/figure/crash-diag.eps.gz)

TABLE 14.28. ANOVA of the reciprocal of crash rate. The **passengers** effect is not shown as significant and the single degree of freedom for the linear by linear component of the interaction isn't shown at all.

(dsgntwo/code/crash.s)

```
S-PLUS (dsgntwo/transcript/crashi.aov.st):
> crashi.aov <- aov(1/crashrate ~ agerange + passengers, data=crash, qr=T)
> summary(crashi.aov)
   Df Sum of Sq Mean Sq F Value    Pr(F)
agerange   2  10.17854 5.089271 44.96802 0.0002446
passengers 3   0.02418 0.008059  0.07121 0.9732476
Residuals  6   0.67905 0.113175
```

power is  $1 - 1.5516 = -.5516$ . As with the Box–Cox transformations in Section 4.7, we usually look at graphs of the transformed values to help make the decision. We show in Figure 14.14 the data distributions and interaction plots for four power transformations: the original scale ( $k = 1$ ), and powers  $k = (0, -0.5, -1)$ .

We see in Figure 14.14 that our transformation has accomplished several of its purposes. Moving from panel a to b to c to d, we see that the variance has been stabilized. In panel a, the boxplots show strong changes in height; in panels c and d, the boxes are almost the same height. In panel a, the lines in the subpanel tracing **passengers** are not parallel and the lines in the subpanel tracing **agerange** are uphill or nearly horizontal. In panels c and d, the lines in the subpanel tracing **passengers** are parallel and downward sloping. The lines tracing **agerange** are uphill and parallel for the two teenage classifications and downhill for the adults.

The simplistic ANOVA table for the reciprocal displayed in Table 14.28 correctly shows **agerange** as significant, but it also shows **passengers** as not significant. We still need more work to make an ANOVA table of the reciprocal **crashrate** show the interaction between **agerange** and **passengers**, an interaction that we can see in Figures 14.13 and 14.14d and in Table 14.27.

We choose the reciprocal over the reciprocal square root for the presentation of the analysis to the client even though both produce similar graphs. The reciprocal is measured in units of trips per crash, units that are easily interpreted. The reciprocal square root, in units of square root of trips per crash, is not easily interpreted. We use the reciprocal, not the negative reciprocal, when we present the results to the client. We do not need to maintain for the client the monotonicity of the series of transformations

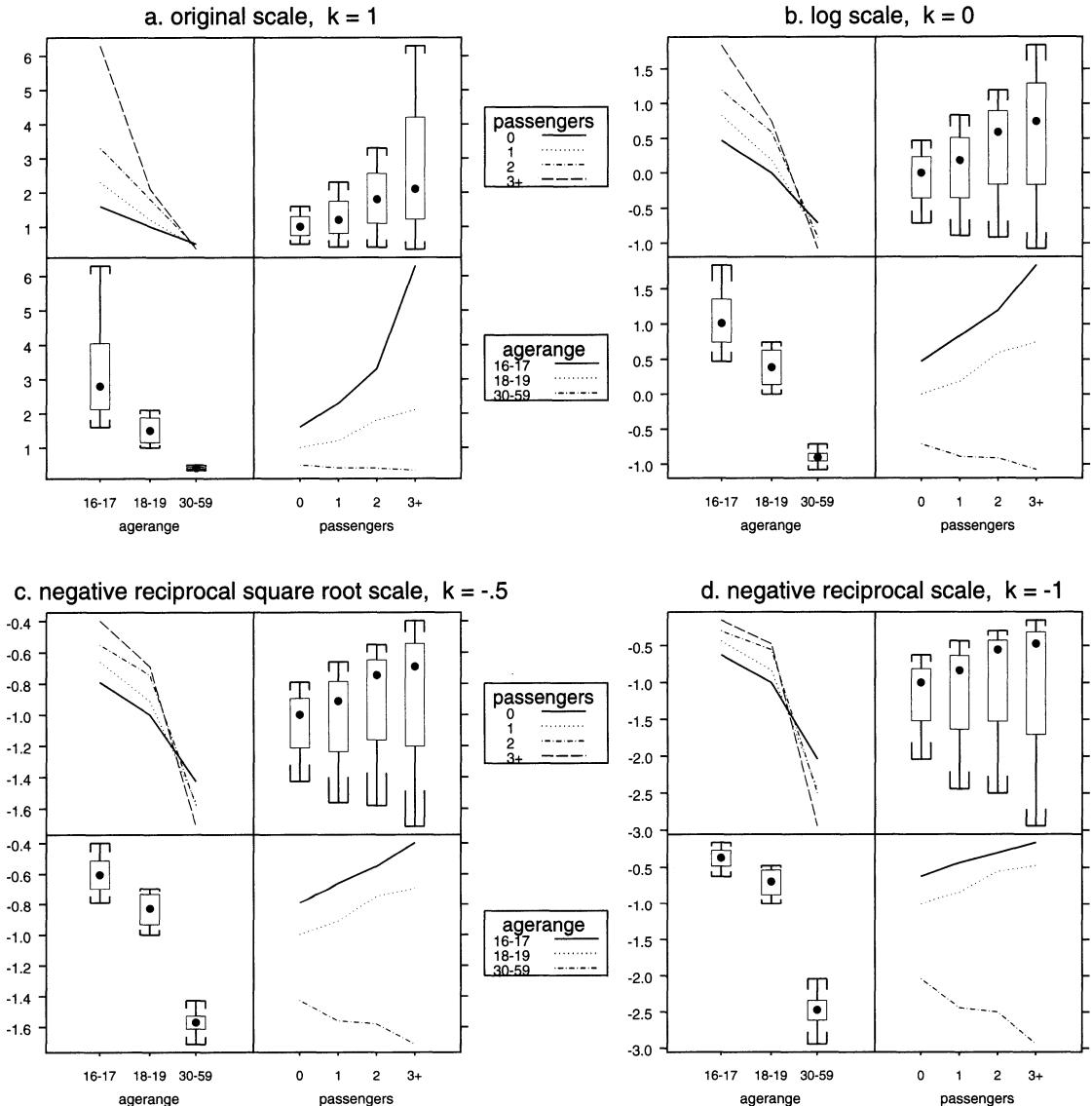


FIGURE 14.14. Main effect and interaction plots of the crash data with several different power transformations. The  $k = -0.5$  and  $k = -1$  plots look similar. Since it is easy to explain the reciprocal—it is measured in units of trips per crash—we select the reciprocal for further analysis and display in Table 14.28. Compare panel a to Figure 14.12. The cover illustration is based on panel d.  
`(dsgntwo/code/crash.s), (dsgntwo/figure/crash-original.eps.gz),`  
`(dsgntwo/figure/crash-log.eps.gz), (dsgntwo/figure/crash-neg-rec-sqrt.eps.gz),`  
`(dsgntwo/figure/crash-neg-rec.eps.gz),`  
`(dsgntwo/code/crash-cover.s), (dsgntwo/figure/crash-neg-rec-cover.eps.gz)`

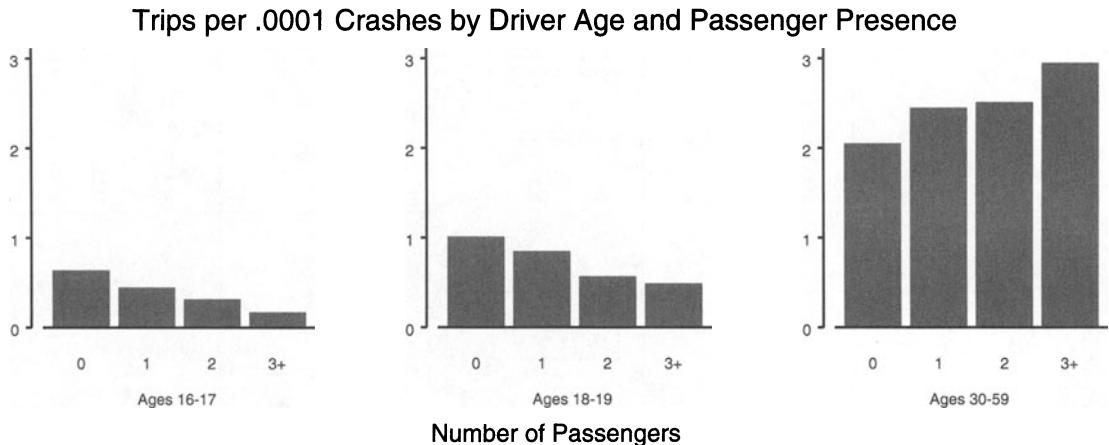


FIGURE 14.15. Barplot of reciprocal crash data.  
`(dsgntwo/code/crash.s)`, `(dsgntwo/figure/crash-bar-rec.eps.gz)`

that we need when we compare several powers chosen from the ladder of powers.

The transformation of the crash rate to the reciprocal scale of trips per crash also works in the barplot format of Figure 14.15. The downhill and parallel pattern for the teens is evident. This contrasts sharply with the larger values and uphill pattern of the adults.

Now that we can see the different behavior for the **passengers** conditional on the **agerange**, let us make the ANOVA table show it. We do so by isolating the linear contrast for **passengers** and nesting it in the **agerange**. We combine the two teenage groups, which show a parallel upward trend over **passengers**. The resulting ANOVA table is in Table 14.29. The *p*-values in Table 14.29 are comparable to those in Table 14.27.

What actually did we do in Table 14.29 that we were unable to do in Table 14.28? The basic idea of the ODOFFNA is to look at the linear by linear component of the interaction. There aren't enough degrees of freedom with only one observation per cell to do more than that. From the graph in Figure 14.14d we see that the two teenage groups show a clear uphill trend and the adults a clear downhill trend. When we combine those two effects by averaging over the ages, we lose the distinction between age groups. We therefore isolated just one degree of freedom of the **passengers** effect, the linear component. We nested the linear effect within the **ageranges**

TABLE 14.29. ANOVA of the reciprocal of crash rate with linear effect of number of passengers nested within age range.  
 (dsgntwo/code/crash.s), (dsgntwo/transcript/crashin.aov.st)

---

S-PLUS (dsgntwo/transcript/crashinlin.aov.st):

```

> pass <- as.numeric(crash$passengers)
> crashinlin.aov <- aov(1/crashrate ~ agerange/pass,
+                         data=crash, qr=T)
> coef(summary.lm(crashinlin.aov))[4:6,]
      Value Std. Error   t value    Pr(>|t|)
agerange16-17pass -0.1530562 0.03350223 -4.568537 0.0038163656
agerange18-19pass -0.1849206 0.03350223 -5.519652 0.0014874046
agerange30-59pass  0.2762056 0.03350223  8.244395 0.0001721117
> summary(crashinlin.aov,
+           split=list("pass %in% agerange"=
+                         list(teens=1:2, adults=3)))
      Df Sum of Sq  Mean Sq   F Value    Pr(F)
      agerange  2  10.17854 5.089271 906.8557 0.000000036
      pass %in% agerange  3   0.66956 0.223186 39.7694 0.000235786
      pass %in% agerange: teens  2   0.28811 0.144055 25.6690 0.001145840
      pass %in% agerange: adults  1   0.38145 0.381448 67.9700 0.000172112
      Residuals  6   0.03367 0.005612
  
```

---

and easily see in Table 14.29 that the teens and adults have strong, and opposite, linear components of the **passengers** effect.

In summary, adults have many more trips between crashes than teenagers. Adults have fewer crashes as the number of passengers increases. Teens have more crashes as the number of passengers increases.

### 14.8.2 Theory

The heart of the analysis in Section 14.8.1 involves partitioning the residual sum of squares in Table 14.26 into two components in Table 14.27. The sum of squares for one of these components involves the vector of comparison values, the linear-by-linear effect constructed in Section 14.8.1. By construction, the vector of comparison values is orthogonal to the dummy variables for the row and column effects. Hence, it must be in the space of the interaction of the row and column effects. In this section we outline the theory justifying that the statistic for testing nonadditivity in Table 14.27, the ratio of nonadditivity mean square to remaining residual mean square, has an *F*-distribution under the null hypothesis of nonadditivity.

If in the model in Equation (12.1) we have  $n_{ij} = 1$  for all  $i$  and  $j$ , that is, there is exactly one observation at each treatment combination, then no degrees of freedom are available to estimate the interactions  $(\alpha\beta)_{ij}$  as we would have if there were at least two observations at each treatment combination. However, for this situation (Tukey, 1949) developed a test for the special case of multiplicative interaction, that is, a test of

$$H_0: \theta = 0 \quad \text{vs} \quad H_1: \theta \neq 0 \quad (14.2)$$

in the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \theta\alpha_i\beta_j + \epsilon_{ij} \quad \text{for } i = 1, \dots, a \quad \text{and } j = 1, \dots, n_i$$

where  $\sum_i \alpha_i = 0$ , and  $\sum_j \beta_j = 0$ . In this situation, unbiased estimators are

$$\begin{aligned}\hat{\alpha}_i &= \bar{Y}_{i\cdot} - \bar{Y}_{..} \\ \hat{\beta}_j &= \bar{Y}_{\cdot j} - \bar{Y}_{..} \\ \hat{\theta}\alpha_i\beta_j &= Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{..}\end{aligned}$$

Therefore a reasonable estimator of  $\theta$  is

$$\hat{\theta} = \frac{\sum_{i,j} (\bar{Y}_{i\cdot} - \bar{Y}_{..})(\bar{Y}_{\cdot j} - \bar{Y}_{..})(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{..})}{\sum_{i,j} (\bar{Y}_{\cdot j} - \bar{Y}_{..})^2(\bar{Y}_{\cdot j} - \bar{Y}_{..})^2}$$

Since the numerator estimates

$$\sum_{i,j} \alpha_i\beta_j (\hat{\theta}\alpha_i\beta_j) = \theta \sum_{i,j} \alpha_i^2 \beta_j^2$$

and the denominator estimates

$$\sum_{i,j} \alpha_i^2 \beta_j^2$$

it follows that

$$\hat{\theta} = \frac{\sum_{i,j} (\bar{Y}_{i\cdot} - \bar{Y}_{..})(\bar{Y}_{\cdot j} - \bar{Y}_{..})(Y_{ij})}{\sum_{i,j} (\bar{Y}_{\cdot j} - \bar{Y}_{..})^2(\bar{Y}_{\cdot j} - \bar{Y}_{..})^2}$$

From this it can be shown that if  $\theta = 0$ ,

$$S_N = \frac{\sum_{i,j} (\bar{Y}_{i\cdot} - \bar{Y}_{..})(\bar{Y}_{\cdot j} - \bar{Y}_{..})(Y_{ij})^2}{\sum_{i,j} (\bar{Y}_{\cdot j} - \bar{Y}_{..})^2(\bar{Y}_{\cdot j} - \bar{Y}_{..})^2} \quad (14.3)$$

has a  $\chi^2$  distribution with 1 df, and if  $\theta \neq 0$ , (14.3) has a noncentral  $\chi^2$  distribution. (Tukey, 1949) showed that if  $SS_{\text{Res}}$  denotes the usual residual Sum of Squares with  $(a-1)(b-1)$  df, then

$$\frac{S_N}{\left( \frac{SS_{\text{Res}} - S_N}{ab - a - b} \right)} \quad (14.4)$$

has an  $F$  distribution with  $1$  and  $ab - a - b$  degrees of freedom if the null hypothesis in Equation (14.2) is true and a noncentral  $F$  distribution otherwise. Therefore, the test of the no multiplicative interaction hypothesis is given by Equation (14.4).

Experience with this test has shown that it is sensitive to the presence of other forms of interaction.

## 14.9 Exercises

**14.1.** (Peterson, 1985) reports an experiment to determine the effect of the annealing (heating to be followed by slow cooling) **temperature** on the **strength** of three metal **alloys**. Four temperatures were used, 675, 700, 725, and 750 degrees F. It was only possible to make one run per day in each of the four ovens available for the experiment, but all three alloys were accommodated in each oven run. The experimenter regarded this as a split plot design with days as blocks and ovens as plots. The data are contained in (**datasets/anneal.dat**). Perform a complete analysis including data plots and an investigation of the borderline significant interaction.

**14.2.** (Peterson, 1985) discusses an experiment to compare the **yield** in kg/plot of four varieties of beans (New Era, Big Green, Little Gem, and Red Lake), also taking into account 3 **spacings** between rows, 20, 40, and 60 cm. A randomized block design was used, with four blocks. The data appear in the file (**datasets/bean.dat**). Analyze the data including an investigation of the simple effects of **variety** at each level of spacing.

**14.3.** (Barnett and Mead, 1956), also in (Johnson and Leone, 1967), discuss an experiment to determine the efficiency of radioactivity decontamination. The response is a measure (alpha activity) of remaining contamination, modeled by these four 2-level factors:

**B:** added barium chloride

**A:** added aluminum sulfate

**C:** added carbon

**P:** final pH

The maximum possible block size consisted of eight plots, but four blocks were used. The four-factor interaction was confounded with blocks, so that an even number of the four factors were at their upper levels in blocks 1 and 3, and an odd number of the four factors

were at their upper levels in blocks 2 and 4. The experimental layout is implicit in the data file (`datasets/radioact.dat`). Assuming that all three-factor interactions are negligible and that treatments do not interact with blocks, perform an analysis of variance and produce estimated contrasts to determine which factors and two-factor interactions contribute most to the remaining contamination. Produce a written recommendation for the client.

- 14.4.** (Neter et al., 1996) describe an experiment to examine the effect of two **irrigation** methods and two **fertilizers** on the **yield** of wheat. The data appear in (`datasets/wheat.dat`). Five fields were available to conduct this experiment. Each field was divided into two plots to which the two **irrigation** methods were randomly assigned. Each plot was subdivided into two subplots to which the fertilizers were randomly assigned. Perform a complete analysis. Produce a written recommendation for the farmer.
- 14.5.** In Section 14.3 it is stated that use of the split plot design resulted in increased precision for inferences involving the subplot factor **nitrogen** at the cost of reduced precision for the whole plot factor **variety**. Demonstrate this by reanalyzing the data under the assumption that the experimental design was a completely random one, with no randomization restrictions. That is, compare the *F*-tests for **nitrogen** and for **variety** under both design assumptions.
- 14.6.** (Fouts, 1973) reports on an experiment in which times in minutes were recorded for each of 4 chimpanzees to learn each of 10 signs of American Sign Language. The data appear in (`datasets/chimp.dat`).
  - a.** Use the Tukey one-degree-of-freedom test to check whether there is interaction between sign and chimpanzee.
  - b.** Is there evidence that the signs differ in the time required for the chimps to securely learn them? Discuss.
- 14.7.** In Section 14.3.1 we note that the three-factor representation of the split plot design gives the same ANOVA table as the more complete five-factor representation. In this exercise we ask you to verify that statement by looking at the column space spanned by the various dummy variables implicitly defined by the model formula. Review the definitions of column space and orthogonal bases in Appendix Section F.4. We continue with the example in Section 14.3. The S-PLUS code for this exercise is in the file (`dsgntwo/code/yatesppl.ex.s`).
  - a.** The whole plot column space is defined by the

```
plots %in% blocks
```

dummy variables generated by the

```

## alternate residuals formula
yatesppl.resida.aov <- aov(y ~ blocks/plots,
                               data=yatesppl, x=T)
summary(yatesppl.resida.aov)
t(yatesppl.resida.aov$x)

```

- b. This is the same column space defined by the  
`variety + blocks:variety`  
dummy variables generated by the

```

## computational shortcut
yatesppl.short.aov <-
  aov(terms(y ~ blocks + variety + blocks*variety +
            nitrogen + variety*nitrogen,
            keep.order=T), ## try it without keep.order=T
      data=yatesppl, x=T)
summary(yatesppl.short.aov)
t(yatesppl.short.aov$x)

```

- c. We illustrate the equivalence by regressing the response variable `y` on the `variety + blocks:variety` dummy variables:

```

## project y onto blocks/plots dummy variables
plots.aov <-
  lm(y ~ yatesppl.resida.aov$x[,7:18], data=yatesppl)
summary.aov(plots.aov)
y.bp <- predict(plots.aov)
variety.aov <- aov(y.bp ~ blocks*variety, data=yatesppl)
summary(variety.aov)

```

and seeing that we reproduce the `plots %in% blocks` stratum of the ANOVA table

| Error: plots %in% blocks |           |          |          |         |           |
|--------------------------|-----------|----------|----------|---------|-----------|
| Df                       | Sum of Sq | Mean Sq  | F Value  | Pr(F)   |           |
| variety                  | 2         | 1786.361 | 893.1806 | 1.48534 | 0.2723869 |
| Residuals                | 10        | 6013.306 | 601.3306 |         |           |

obtained from the complete five-factor specification:

```

## split plot analysis
yatesppl.anova <- aov(y ~ variety*nitrogen +
                        Error(blocks/plots/subplots),
                        data=yatesppl)
summary(yatesppl.anova)

```

# Bivariate Statistics—Discrete Data

In this chapter we discuss bivariate discrete distributions. Bivariate means that there are two factors (categorical variables) defining cells. The response values are frequencies, that is, counts or instances of observations, at each cell.

It is convenient to arrange such data in a contingency table, that is, a table with  $r$  rows representing the possible values of one categorical variable and  $c$  columns representing the possible values of the other categorical variable. Each of the  $rc$  cells of the table contains an integer, the number of observations having the levels of the two variables specified by the cell location. We give extra attention to the special case where  $r = c = 2$ , that is, a  $2 \times 2$  contingency table.

This data type is different from situations with two categorical variables (factors) described in Chapters 12 through 14. In those chapters the response variables are one or more continuous measurements at each cell.

## 15.1 Two-Dimensional Contingency Tables—Chi-Square Analysis

### 15.1.1 Example—Drunkenness Data

Table 15.1 shows the number of persons convicted of drunkenness in two London courts during the first six months of 1970. The data come from (Cook, 1971), later reprinted in (Hand et al., 1994). There are two rows, males and females. The five columns are five age categories. The question of interest is whether the age distribution of convicted offenders is the

TABLE 15.1. Persons convicted of drunkenness in two London courts during the first six months of 1970.  
 (datasets/drunk.dat)

| Age group         | 0-29 | 30-39 | 40-49 | 50-59 | $\geq 60$ |
|-------------------|------|-------|-------|-------|-----------|
| Number of males   | 185  | 207   | 260   | 180   | 71        |
| Number of females | 4    | 13    | 10    | 7     | 10        |

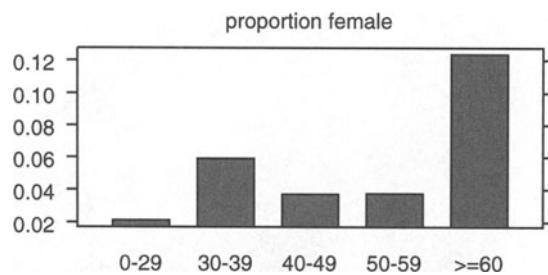


FIGURE 15.1. Proportion female for drunkenness data.  
 (twtb/code/drunk.s), (twtb/figure/drunk.prop.fem.eps.gz)

same for both genders, or equivalently, as illustrated in Figure 15.1, if the proportion of female offenders is the same for all age categories.

The tabular display from the chi-square analysis (to be described in Section 15.1.2) is in Table 15.2.

The  $p$ -value .0042 for the chi-square test strongly suggests an association between age and gender of convicted offenders, that is, the proportion of females in each age group is not identical. From the “Cell Chi-square” values in Table 15.2 and their square roots displayed in Figure 15.2, it seems that only 1 of the 10 cells contributes appreciably to total chi-square. The cell for female offenders aged at least 60 contributes 10.34, or 68%, of the total chi-square value 15.25. We observe 10 female offenders aged at least 60, but under the null hypothesis of independence we expect only 3.8 offenders. No other cell is suggestive of dependence.

We must be careful not to overinterpret this finding as meaning that older females have a greater tendency toward this crime than older males. We believe that the finding may be an artifact of the demographic distribution. The population proportion of females under the age of 60 is roughly 50%. This proportion tends to increase after age 60 because of higher male mortality beginning at approximately that age. Therefore, it is possible that

TABLE 15.2. SAS Chi-square analysis of Drunkenness Data.  
 (twtb/code/drunk.sas), (twtb/code/drunk1.sas)

SAS (twtb/code/drunk2.sas):

```
proc freq order=data;
  weight n;
  tables gender*age / chisq nopercent nocol norow expected cellchi2;
run;
```

SAS (twtb/transcript/drunk2.lst):

Table of gender by age

|         | gender    | age |                 |        |        |        |        |        |  |  |  |  |  | Total |
|---------|-----------|-----|-----------------|--------|--------|--------|--------|--------|--|--|--|--|--|-------|
|         | Frequency |     | Cell Chi-Square | 0-29   | 30-39  | 40-49  | 50-59  | >=60   |  |  |  |  |  |       |
|         | Expected  |     |                 |        |        |        |        |        |  |  |  |  |  |       |
| males   |           |     |                 | 185    | 207    | 260    | 180    | 71     |  |  |  |  |  | 903   |
|         |           |     |                 | 180.22 | 209.78 | 257.46 | 178.31 | 77.237 |  |  |  |  |  |       |
|         |           |     |                 | 0.1269 | 0.0368 | 0.0252 | 0.016  | 0.5036 |  |  |  |  |  |       |
| females |           |     |                 | 4      | 13     | 10     | 7      | 10     |  |  |  |  |  | 44    |
|         |           |     |                 | 8.7814 | 10.222 | 12.545 | 8.6885 | 3.7635 |  |  |  |  |  |       |
|         |           |     |                 | 2.6034 | 0.7551 | 0.5163 | 0.3281 | 10.335 |  |  |  |  |  |       |
| Total   |           |     |                 | 189    | 220    | 270    | 187    | 81     |  |  |  |  |  | 947   |

Statistics for Table of gender by age

| Statistic                   | DF | Value   | Prob   |
|-----------------------------|----|---------|--------|
| Chi-Square                  | 4  | 15.2461 | 0.0042 |
| Likelihood Ratio Chi-Square | 4  | 12.6670 | 0.0130 |
| Mantel-Haenszel Chi-Square  | 1  | 4.8961  | 0.0269 |
| Phi Coefficient             |    | 0.1269  |        |
| Contingency Coefficient     |    | 0.1259  |        |
| Cramer's V                  |    | 0.1269  |        |

Sample Size = 947

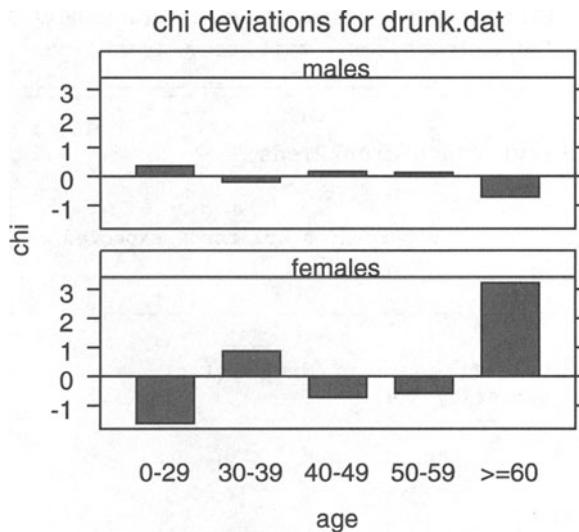


FIGURE 15.2. Cell chi-deviations (signed square root of the cell chi-square values from Table 15.2) for drunkenness data.  
 (twtb/code/drunk.s), (twtb/figure/drunk.chi.eps.gz)

this study could have been improved by adjusting the responses to a per capita basis.

### 15.1.2 Chi-Square Analysis

When we work with two-dimensional tables, such as this example, we often want to test whether the row and column classifications are independent, that is, whether the probability of an entry's being in a particular row is independent of the entry's column.

When  $r = 2$ , the test is essentially asking whether the proportion of data in Row 1 is homogeneous across the  $c$  columns, i.e., whether  $c$  binomial populations have the same (unspecified) proportion parameter. Therefore, this is a generalization of the inferences comparing  $c = 2$  population proportions discussed in Chapter 5 to  $c \geq 2$  population proportions.

If the total number of observations  $n$  is sufficiently large and if none of the  $rc$  expected cell frequencies is less than somewhere between 3 and 5, the chi-square distribution may be used to test the hypothesis of independence. The logical idea behind this test is to compare, in each of the cells,

$n_{ij}$  the observed frequency in the cell in row  $i$  and column  $j$   
 with

$e_{ij}$  the expected frequency calculated under the assumption  
 that the independence null hypothesis is true.

The test statistic is a function of the aggregate discrepancy between the  $n_{ij}$ 's and  $e_{ij}$ 's.

Define

$$\left\{ \begin{array}{lcl} n_{i\cdot} & = & \sum_j n_{ij} \quad \text{the row totals} \\ n_{\cdot j} & = & \sum_i n_{ij} \quad \text{the column totals} \\ n = n_{..} & = & \sum_{i,j} n_{ij} \quad \text{the grand total} \end{array} \right. \quad (15.1)$$

Under the null hypothesis of independence between rows and columns, the expected frequency for the cell in row  $i$  and column  $j$  is

$$e_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n} \quad (15.2)$$

The cell chi-deviations, the scaled deviations

$$\frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}} \quad (15.3)$$

are displayed in Figure 15.2 and their squares, the cell chi-square values, are displayed in Table 15.2.

The test statistic is the sum of the scaled deviations squared

$$\hat{\chi}^2 = \sum_{ij} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \quad (15.4)$$

where the sum is taken over all rows and columns. If the null hypothesis is true,  $\hat{\chi}^2$  has, approximately, a chi-square distribution with  $(r - 1)(c - 1)$  degrees of freedom, and the  $p$ -value is  $1 - \mathcal{F}_{\chi^2}(\hat{\chi}^2 \mid (r - 1)(c - 1))$ , the chi-square tail probability associated with  $\hat{\chi}^2$ . Most authorities agree that the chi-square approximation is good if almost all  $e_{ij}$  are at least 5 and none is less than 3. The degrees-of-freedom formula derives from the fact that if all marginal totals are given, knowledge of  $(r - 1)(c - 1)$  interior values uniquely determines the remaining  $r + c - 1$  interior values.

Apart from the degrees-of-freedom calculation, the form of this test, comparing observed and expected frequencies, is identical to the goodness-of-fit test described in Section 5.7.1. This methodology can be extended to contingency tables having more than two dimensions.

Further analysis is required to assess the nature of any lack of independence that is suggested by the chi-square test. One approach to this is discussed in the next paragraph. Another is a multivariate display technique called *correspondence analysis*. See (Greenacre, 1984) for an introduction to this topic.

The chi-square test of independence is printed by SAS PROC FREQ when the CHISQ option to the TABLES command is specified. When the cellchi2 option to the TABLES command is specified, each cell's contribution to the chi-square statistic is also displayed. The test is printed by S-PLUS with the chisq.test command. We calculated the cell chi-square values in the file (twtb/code/drunk.s).

Cells with a sizeable cell chi-square value have an appreciable discrepancy between their observed and expected frequency. Scrutiny of such cells leads to interpretation of the nature of the dependence between rows and columns. A cell chi-square is calculated as  $(n - e)^2/e$ , where  $n$  and  $e$  are, respectively, the cell's observed frequency and expected frequency under the null hypothesis. Under the model that observations are randomly assigned to cells with a Poisson distribution, the variance of  $n$  is also  $e$ . Hence  $(n - e)^2/e$  has the form

$$\frac{(n - E(n))^2}{\text{var}(n)}$$

and, using the normal approximation, we interpret it as approximately a one-df chi-square statistic. Since the 95<sup>th</sup> and 99<sup>th</sup> percentiles of this chi-square distribution are 3.84 and 6.63, we recommend reporting the discrepancy between the observed and expected frequency for all cells with cell chi-square exceeding the higher value. Also, consideration should be given to reporting cells having chi-square between 3.84 and 6.63 when the discrepancy between the cell's observed and expected frequency can be meaningfully interpreted.

## 15.2 Two-Dimensional Contingency Tables—Fisher's Exact Test

An alternative to the approximate chi-square statistic discussed above is Fisher's exact test, which uses the exact hypergeometric distribution probabilities calculated for all tables at least as extreme in the alternative hypothesis direction as the existing table. Since it is exact, this procedure, when available, is preferable to the chi-square test, but it is extremely computer intensive for all but the smallest tables, even by contemporary standards.

Fisher's exact test is available as the `exact` option to the `tables` command under SAS's `PROC FREQ`. Documentation for PC-SAS Version 8.2 cautions

For some large problems, computation of exact tests may require a large amount of time and memory.

The two-sided test is available in S-PLUS as `fisher.test(x)`, where `x` is a two-dimensional contingency table in matrix form. The only data restriction stated in the S-PLUS documentation is  $n \leq 200$ .

### 15.2.1 Example—Do Juvenile Delinquents Eschew Wearing Eyeglasses?

(Weindling et al., 1986) discuss a small study of juvenile delinquent boys and a control group of nondelinquents. The data in Table 15.3 also appear in (Hand et al., 1994). All of these subjects failed a vision test. The boys were also classified according to whether or not they wore glasses.

Clearly, the data set is much too small to use the chi-square analysis of the previous section. Therefore, we request an analysis using Fisher's exact test, shown in Table 15.4. We are interested in the two-sided  $p$ -value, .0350. Since this falls between the two thresholds .01 and .05, we can say there is suggestive but inconclusive evidence that a smaller proportion of delinquents than nondelinquents wear glasses.

The calculation of the  $p$ -values for Fisher's exact test utilizes the hypergeometric probability distribution discussed in Appendix D to calculate the probability of obtaining the observed table and “more extreme” tables assuming that the table's marginal totals are fixed. We illustrate the calculations in Table 15.5. The observed table is shown in Column 1. All possible tables with the same row and column margins are also shown, indexed by the `[glasses,del]` cell count. The probability of observing the counts in Table 15.3 (identical to Column 1, marked “\*”, in Table 15.5) given this

TABLE 15.3. Wearing prescribed glasses and juvenile delinquency.  
(`datasets/glasses.dat`)

|                      | Juvenile delinquents | Nondelinquents |
|----------------------|----------------------|----------------|
| Wears glasses        | 1                    | 5              |
| Doesn't wear glasses | 8                    | 2              |

TABLE 15.4. SAS analysis of glasses data by PROC FREQ. Fisher's exact test for glasses data. The  $p$ -value for the two-sided exact test is .035. Compare this to  $p = .0134$  from the chi-square approximation that SAS warns might be invalid.

(twtb/code/glasses.sas), (twtb/code/glasses1.sas), (twtb/transcript/glasses2.lst)

```
SAS (twtb/code/glasses2.sas):
proc freq order=data;
  weight n;
  table wearer*delinquent / exact nopercent nocol norow;
run;
```

SAS (twtb/transcript/glasses2b.lst):  
Fisher's Exact Test

|                          |        |
|--------------------------|--------|
| Cell (1,1) Frequency (F) | 1      |
| Left-sided Pr <= F       | 0.0245 |
| Right-sided Pr >= F      | 0.9991 |
| Table Probability (P)    | 0.0236 |
| Two-sided Pr <= P        | 0.0350 |

Sample Size = 16

TABLE 15.5. All possible  $2 \times 2$  tables with the same margins as the observed glasses table. In these tables `glasses` denotes “wears glasses” and `del` denotes “delinquent.” We show the probabilities of each under the assumption of a hypergeometric distribution for the `[glasses,del]` position and the common row and column cell margins. The observed table **1** is marked with “\*” and the more extreme tables in the same tail **0** and opposite tail **6** are marked “<”.

(twtb/code/glasses.exact.s)

| [glasses,del] cell count |        |        |        |        |        |        |        |        |        |        |        |        |        |        |
|--------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| glasses                  | 0      |        | 1      |        | 2      |        | 3      |        | 4      |        | 5      |        | 6      |        |
|                          | del    | no.del |
| glasses                  | 0      | 6      | 1      | 5      | 2      | 4      | 3      | 3      | 4      | 2      | 5      | 1      | 6      | 0      |
| no.glasses               | 9      | 1      | 8      | 2      | 7      | 3      | 6      | 4      | 5      | 5      | 4      | 6      | 3      | 7      |
| probability              | 0.0009 |        | 0.0236 |        | 0.1573 |        | 0.3671 |        | 0.3304 |        | 0.1101 |        | 0.0105 |        |
| which                    | <      |        | *      |        |        |        |        |        |        |        |        |        | <      |        |

table's marginal totals is

$$\frac{\binom{9}{1} \binom{7}{5}}{\binom{16}{6}} = 0.0236$$

The one-sided  $p$ -value is the sum of this probability and the probability, 0.0009, of the more extreme table **0** on the same tail of the distribution.

Table **6**, the more extreme on the opposite tail of the distribution, has probability

$$\frac{\binom{9}{6} \binom{7}{0}}{\binom{16}{6}} = 0.0105$$

The two-sided  $p$ -value is the sum of the probabilities of the observed table and both more extreme tables,  $0.0236 + 0.0009 + 0.0105 = 0.0350$ .

## 15.3 Simpson's Paradox

Simpson's paradox, a counterintuitive situation, occurs when the presence of a third variable is unexpectedly responsible for a change, or even reversal, of the relationship between two categorical variables. The following example taken from (Blyth, 1972), with dataset (`datasets/blyth.dat`), illustrates this phenomenon.

The data, including the margin summed over location, are shown in Table 15.6 and Figure 15.3. A medical researcher selected 11,000 human subjects at location A and 10,100 subjects at location B. At A, 1,000 of the subjects were randomly assigned to the standard treatment (`standard`) and the remaining 10,000 subjects were assigned to a new treatment (`new`). At B, 10,000 of the subjects were randomly assigned to `standard` and the remaining 100 subjects were assigned to `new`. Eventually, each subject was classified as not-survived (`not`) or survived (`survive`).

Figure 15.3 displays the actual counts seen in Table 15.6. The intent is to show for each location that the percentage surviving with the new treatments is larger than with the standard treatments, but that the reverse is true when the two locations are combined. This portrayal fails for these data because it is not possible to distinguish the very small counts 5, 50, and 95 in three of the eight cells in Table 15.6 when they are displayed on a common numerical scale with counts ranging from 950 to 9,005 in the table's other cells.

TABLE 15.6. Blyth's data illustrating Simpson's Paradox. Within each location, **new** has a higher survival rate than **standard**. Summed over locations, **new** has a lower survival rate than **standard**.  
 (twtb/code/blyth.s)

```
S-PLUS (twtb/transcript/blyth.st):
> ## append the location margin
> blyth.t
      A.standard A.new B.standard B.new standard   new
not          950  9000      5000     5    5950  9005
survive       50  1000      5000    95    5050 1095

> ## proportion
> round(blyth.t[2,] / apply(blyth.t, 2, sum))
      A.standard A.new B.standard B.new standard   new
            0.05   0.1      0.5   0.95    0.459  0.108
```

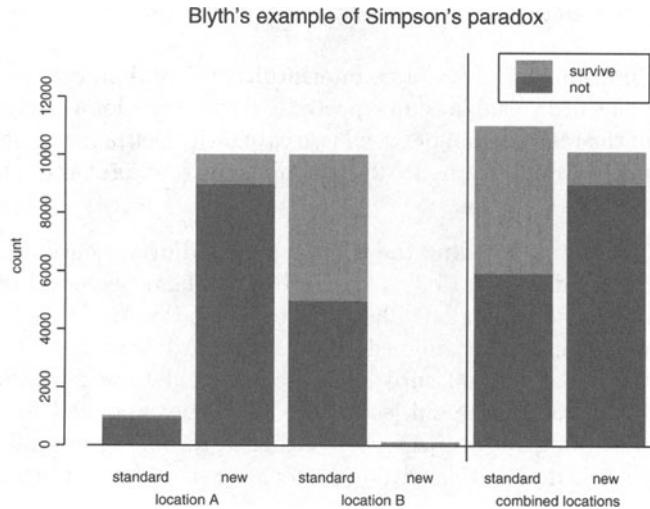


FIGURE 15.3. Blyth's data illustrating Simpson's paradox. Within each location, **new** has a higher survival rate than **standard**. Summed over locations, **new** has a lower survival rate than **standard**. The great disparity in counts among the four groups makes it almost impossible to see the survival rates in this graph.

(twtb/code/blyth.s), (twtb/figure/blyth.count.eps.gz)

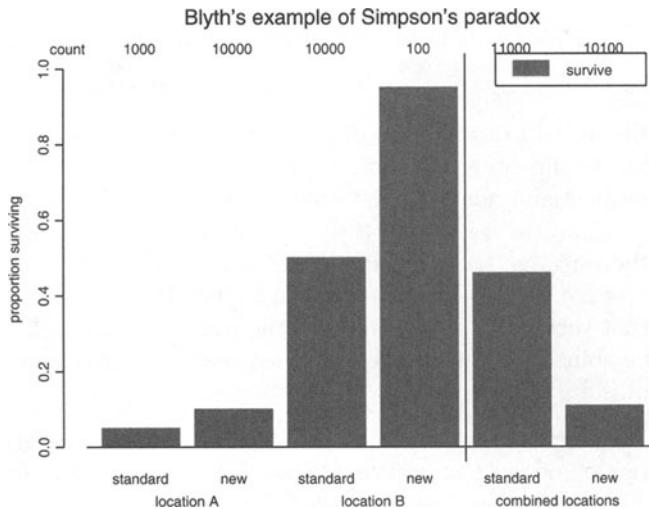


FIGURE 15.4. Blyth's data illustrating Simpson's paradox. Within each location, **new** has a higher survival rate than **standard**. Summed over locations, **new** has a lower survival rate than **standard**. We give the observation counts for the location–treatment combinations above the bars. We note that the combined location information is almost the same as the **B.standard** and **A.new** information.  
`(twtb/code/blyth.s)`, `(twtb/figure/blyth.proportion.eps.gz)`

The paradox is better communicated by Figure 15.4, which graphs the percentages themselves. We observe that in location A, the percent surviving following the new treatment was 10% as compared to 5% with the standard treatment. In location B, the new treatment improved the survival percentage to 95% from 50% for the standard treatment. It seems that the new treatment was very successful. Now look at the summary results for both locations combined. The survival rate for the standard treatment is 46%, but it is only 11% for the new treatment. The combined results suggests that the new treatment is a disaster!

The substantive reason for this finding is that the subjects in location A were much less healthy than those in B and the new treatment was given mostly to subjects in A, where it could not be expected to fare as well as with subjects in B. That is, the factors treatment and location are not independent. When this is so, it can happen as here that

$$P(\text{survive} \mid \text{new}) < P(\text{survive} \mid \text{standard})$$

while both

$$P(\text{survive} \mid \text{new} \cap A) \geq P(\text{survive} \mid \text{standard} \cap A)$$

and

$$P(\text{survive} \mid \text{new} \cap B) \geq P(\text{survive} \mid \text{standard} \cap B)$$

corresponding to  $.11 < .46$  but  $.10 > .05$  and  $.95 > .50$  in this example.

Figure 15.4 displays the disparity better than Figure 15.3 because it transforms the observed data from the scale reported by the client to the proportion scale, a scale in which the reversal is visible. In the proportion scale we can easily see that the combined location information is almost the same as the `B.standard` and `A.new` information. In retrospect we can also see the same information in Figure 15.3. We can explain it by noting that there is almost no data in the `A.standard` and `B.new` cells, hence the combined information really is just the `B.standard` and `A.new` information plus a little noise.

File (`twtb/code/blyth.s`) gives code for analogous figures for the relative risk or odds ( $P(\text{survive})/P(\text{not})$ ) and logit ( $\log(P(\text{survive})/P(\text{not}))$ ). The figures are on the files (`twtb/figure/blyth.odds.eps.gz`) and (`twtb/figure/blyth.logit.eps.gz`).

When analyzing contingency table data, we should be alert to the possibility illustrated in this example that results for tables individually can differ from those when these tables are combined.

What is the resolution in situations such as this where individual results contradict combined results? Almost always, the individual results have more credence because combining such individuals cannot be adequately justified. In Blyth's example, the disparity between individual and combined results could have been attributable to different baseline health status of the patients at the two locations. Or it could be an artifact of the radically different treatment allocation patterns at the two locations.

Simpson's paradox is the discrete analogue of the ecological fallacy discussed in Section 4.2. It is also related to the need to examine simple effects in the presence of interaction of qualitative factors, discussed in Chapters 12 to 14, since both problems refer to the importance of distinguishing overall conclusions from conclusions for subgroups.

## 15.4 Relative Risk and Odds Ratios

Analysis of data arranged in a  $2 \times 2$  table is equivalent to comparing two proportions. However, analyzing the difference  $p_1 - p_2$  via a CI or test as outlined in Chapter 5 is often not an appropriate way to compare them. Instead a measure of *relative* difference is appropriate. Consider two cases, the first with  $p_1 = .02$  and  $p_2 = .07$  and the second with  $p_1 = .50$  and

$p_2 = .55$ . In both cases,  $p_2 - p_1 = .05$ . However, in the first case,  $p_2$  is 250% more than  $p_1$ , but in the second case  $p_2$  is only 10% more than  $p_1$ , and from this point of view it is inadequate to merely consider differences of proportions, particularly proportions close to either 0 or 1.

We discuss two additional measures for comparing two proportions. The first, the *relative risk*, is simply the ratio of the two proportions,  $\hat{p}_1/\hat{p}_2$ .

The *odds ratio* is a widely used measure of relative difference. It is more informative than a chi-square test for a  $2 \times 2$  table because it measures the magnitude of difference between two proportions. Unlike the chi-square test, the odds ratio is minimally affected by the size of the sample.

Based on the definition of odds in Equation (3.2), if  $\hat{p}$  is an estimated probability of success, the estimated *odds in favor* of success are  $\hat{\omega} = \hat{p}/(1 - \hat{p})$ . For comparing two estimated proportions,  $\hat{p}_1$  and  $\hat{p}_2$  in a  $2 \times 2$  contingency table, the estimated ratio of two odds, the odds ratio, is

$$\hat{\Psi} = \hat{\omega}_2/\hat{\omega}_1 \quad (15.5)$$

A quick way to hand-calculate the estimated odds ratio is  $\hat{\Psi} = (n_{11}n_{22})/(n_{21}n_{12})$ , and for this reason the odds ratio is also known as the *cross-product ratio*.

If the odds ratio exceeds 1 so does the relative risk, and conversely.

### 15.4.1 Glasses (Again)

For example, reconsider the data of Section 15.2.1. The relative risk is

$$\frac{\left(\frac{8}{10}\right)}{\left(\frac{1}{6}\right)} = 4.8$$

This says that based on these data, nonwearers of glasses are four times more likely to become delinquent than wearers of glasses.

The odds ratio is

$$\hat{\Psi} = \frac{\left(\frac{\frac{5}{7}}{1 - \frac{5}{7}}\right)}{\left(\frac{\frac{1}{9}}{1 - \frac{1}{9}}\right)} = 20$$

This means that the odds that a nondelinquent wears glasses are estimated to be 20 times the odds that a delinquent wears glasses. Alternatively, the odds that a delinquent wears glasses are estimated as 1/20 times the odds that a nondelinquent wears glasses. If this ratio had been maintained for a larger sample, an implication might have been that police needn't pay much attention to boys wearing glasses.

Often investigators report the log of the odds ratio since the change in the reference group simply reverses the sign of the log odds:  $\ln(20) = 2.996$  and  $\ln(\frac{1}{20}) = -2.996$ .

### 15.4.2 Large Sample Approximations

A useful property of the odds ratio is that for large sample sizes, the log of the estimated odds ratio is approximately normally distributed

$$\ln(\hat{\Psi}) \sim N\left(\ln(\Psi), \sigma_{\ln(\hat{\Psi})}^2\right) \quad (15.6)$$

with mean equal to the log of the population odds ratio and estimated variance

$$\hat{\sigma}_{\ln(\hat{\Psi})}^2 = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \quad (15.7)$$

These facts lead to large sample confidence intervals and hypothesis tests for odds ratios.

A test of  $H_0: \ln(\Psi) = 0$ , or equivalently,  $\Psi = 1$ , is based on

$$z_{\text{calc}} = \frac{\ln(\hat{\Psi})}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}}$$

An approximate  $100(1 - \alpha)\%$  confidence interval on  $\ln(\Psi)$  is

$$\text{CI}(\ln(\Psi)) = \ln(\hat{\Psi}) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = (L, U) \quad (15.8)$$

If we denote this interval by  $(L, U)$ , the approximate confidence interval on the odds ratio  $\Psi_{21}$  is  $(e^L, e^U)$ .

### 15.4.3 Example—Treating Cardiac Arrest with Therapeutic Hypothermia

(Holzer, 2002) supervised a multicenter trial of patients who were randomly assigned to receive or not receive therapeutic hypothermia (lowered body temperature) to assist in recovery following resuscitation from cardiac arrest. Six months after cardiac arrest, patients were classified as having a favorable neurologic outcome or not. Of 136 patients treated with hypothermia, 75 had a favorable neurological outcome. Of 137 patients not treated with hypothermia, 54 had a favorable neurological outcome (see Table 15.7). All patients received standard treatment for cardiac arrest apart from hypothermia and the treatment was blinded from the assessors of the outcome.

TABLE 15.7. Results of therapeutic hypothermia investigation (Holzer, 2002).

|         |    | Favorable neurological outcome |     | Total |
|---------|----|--------------------------------|-----|-------|
|         |    | Yes                            | No  |       |
| Treated | 75 | 61                             | 136 |       |
|         | 54 | 83                             | 137 |       |

For a patient who receives the therapeutic hypothermia treatment, the estimated odds in favor of a favorable neurologic outcome are

$$\hat{\omega}_2 = \left( \frac{\frac{75}{136}}{1 - \frac{75}{136}} \right) \approx 1.23$$

For a patient who does not receive this treatment, the estimated odds in favor of a favorable neurological outcome are

$$\hat{\omega}_1 = \left( \frac{\frac{54}{137}}{1 - \frac{54}{137}} \right) \approx 0.65$$

The estimated odds ratio is

$$\hat{\Psi} = \frac{\hat{\omega}_2}{\hat{\omega}_1} = \frac{(75)(83)}{(54)(61)} \approx 1.8898$$

The reader can verify that the estimated standard deviation of the log of the odds ratio is 0.246, and that this leads to an approximate 95% confidence interval (1.168, 3.058) for the population odds ratio. This means that the odds of a favorable neurological outcome for a patient receiving the therapeutic hypothermia treatment are estimated to be between 1.17 and 3.06 times the odds of a favorable neurological outcome for a patient not receiving this particular treatment. Further, the calculated  $z$ -statistic for a test of the one-sided alternative hypothesis that the population odds ratio exceeds 1 is 2.592, with corresponding  $p$ -value less than 0.01. Therefore, there is strong evidence that the therapeutic hypothermia treatment improves probability of a successful neurological outcome.

For the hypothermia example with  $\alpha = .05$ , we have

$$\text{CI}(\ln(\Psi)) = \ln(1.8898) \pm 1.96 \sqrt{\frac{1}{75} + \frac{1}{54} + \frac{1}{61} + \frac{1}{83}} \approx (0.1552, 1.1177)$$

We therefore have an approximate confidence interval on the odds ratio of

$$\text{CI}(\Psi) \approx (1.168, 3.058)$$

For fixed probability of favorable outcome for control patients of  $54/137 = .3942$ , corresponding to fixed odds of favorable outcome of  $.3942/.6058 =$

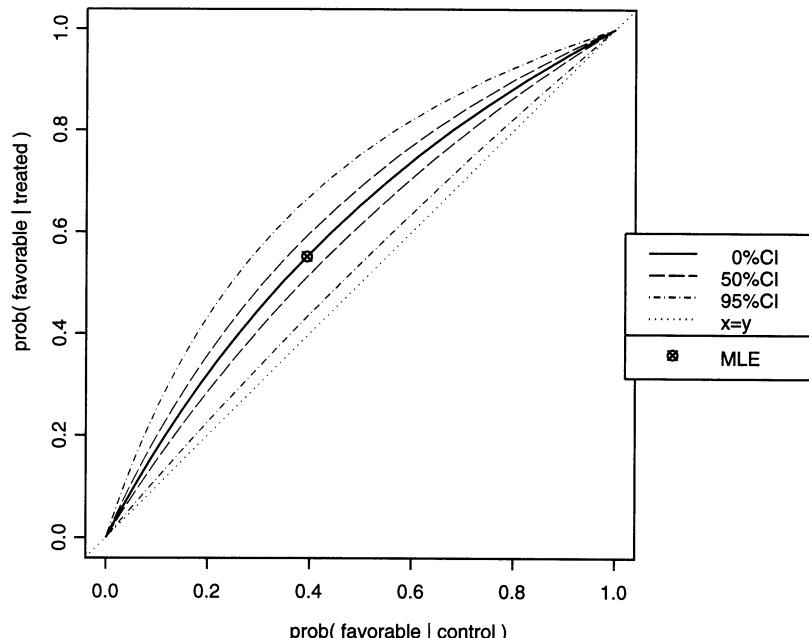


FIGURE 15.5. Confidence intervals for  $P(\text{favorable} \mid \text{treated})$  in the hypothermia example given the odds ratio. The confidence intervals are calculated from an assumed  $P(\text{favorable} \mid \text{control})$  and the given odds ratio using the odds ratio formula in Equation 15.8. Details for  $P(\text{favorable} \mid \text{control}) = 54/137 = .3942$ , corresponding to given odds of favorable outcome of  $.3942/.6058 = .6506$ , are shown in Section 15.4.3. Symmetrically, we can assume any fixed probability of favorable outcome for control and then calculate the confidence interval for the probability of favorable outcome for treatment. ([twtb/code/hypothermia.s](#)), ([twtb/figure/hypothermia.ci.eps.gz](#))

.6506, the point estimate for the odds of favorable outcome for treated patients is given by  $1.8898 \times .6506 = 1.2295$ , and an estimated confidence interval for the odds by  $(1.168 \times .6506, 3.058 \times .6506) = (0.7599, 1.9895)$ . The corresponding point and interval estimates for the probabilities of favorable outcome for treated are  $1.2295/2.2295 = 0.5515$  and  $(0.7599/1.7599, 1.9895/2.9895) = (0.4318, 0.6655)$ . The point estimate of the probability of favorable outcome for treated is exactly the observed proportion  $75/136 = 0.5515$ , and the confidence interval of the proportion of favorable outcomes excludes the observed proportion for the control group.

We can extend this discussion by assuming any fixed probability of favorable outcome for treatment and then calculating the confidence interval for the probability of favorable outcome for control. We do so in Figure 15.5 for the set of fixed probabilities  $p_1 = (0, .05, \dots, 1)$ .

## 15.5 Retrospective and Prospective Studies

Consider two possible experiments to assess whether vitamin C supplementation prevents occurrences of the common cold.

In the first experiment we select 100 people who have had a cold during the past two months and 100 people who have not had a cold during the past two months. We then ask these people whether or not they have taken a daily vitamin C supplement during this period. In the second experiment we select 200 volunteers, assigning them to take no vitamin C supplementation apart from that offered by the study. We randomly assign 100 of these subjects to receive the study's vitamin C supplement and the other 100 subjects to receive a placebo, indistinguishable by these subjects from vitamin C. Then, two months later, we ask the subjects whether or not they have had a cold since the experiment began.

The first experiment is an example of a retrospective study, also called a case-control study. Subjects having a condition are called cases, subjects not having a condition are termed controls, and subjects are cross-classified with a risk factor (present or absent). In the above example, the risk factor is presence or absence of vitamin C supplementation. In retrospective studies, the subjects are selected after the events in question have already occurred, in this case having contracted one or more colds. Such studies are common in medical research because they generally assure a larger number of subjects than prospective studies.

The second experiment is an example of a prospective study, also known as a cohort study. Samples are taken from a population of subjects classified according to two risk factors (events) defined prior to initiating sampling, in this case assignment to vitamin C or placebo. Such studies often require that subjects be followed for a period of time until the subjects are determined to have a condition or not.

In the cold example, the analysis of the retrospective study can be done immediately, but analysis of the prospective study must wait two months to see if colds develop. Prospective studies often run over a long time period; 5 to 10 years is not unusual. It is not uncommon for subjects to withdraw or be lost to the study. For this reason, it is more difficult to obtain sizeable samples from prospective studies than from retrospective ones. Prospective studies are more informative than retrospective studies. Investigators have more control over the risk factor in prospective studies than in retrospective studies. In prospective studies investigators are often able to obtain information on important confounding variables that bear on the response. Such information is usually unavailable in retrospective studies. The experiment discussed in Section 15.4.3 is an example of a prospective study.

Odds ratios are particularly important in the analysis of experiments involving retrospective studies. In a retrospective study it is unlikely that the cases can be considered a random sample of all persons afflicted with the condition. In the context of our example, we cannot be sure that the 100 selected people with colds are representative of all people with colds. Therefore, in such a study we cannot estimate the proportion of people having the risk factor who have the condition, or the proportion of people without the risk factor who have the condition. Nevertheless, in a retrospective study we are able to measure the odds ratio and we can claim that the sample odds ratio estimates the population odds ratio.

## 15.6 Mantel–Haenszel Test

Analysts are often called on to interpret  $k$   $2 \times 2$  contingency tables, related to one another by the fact that each table has the same row and column categories. The  $k$  tables usually represent the  $k$  levels of a third (categorical) factor in addition to the two-level factors specified by rows and columns. For example, we look in Table 15.8 at data studying the effectiveness of the Salk vaccine for polio protection for  $k = 6$  different age groups. Each of the  $k = 6$   $2 \times 2$  tables in the “Observed” column shows the response (paralysis or no.paralysis) for subjects who were or were not vaccinated (vac or no.vac). The complete discussion of the dataset and the table are in Section 15.7.

In earlier sections of this chapter we consider procedures for testing independence of the row and column categories for individual tables. We are now interested in testing the hypothesis that all  $k$  tables show the same pattern of relation of the rows to the columns: Either all tables show independence of rows and columns or all show the same dependency structure.

The Mantel–Haenszel test, also referred to as the Cochran–Mantel–Haenszel test, tests the hypothesis that row vs column independence holds *simultaneously in each table*. It is designed to be sensitive to an overall consistent pattern. It has low power for detecting association when patterns of association for some strata are in the opposite direction of other strata.

Let us now look at the algebra of the test statistic. Since we now have  $k$   $2 \times 2$  tables, we require a third subscript on the  $n$ ’s. Let the  $k^{\text{th}}$  table be

|           |           |            |
|-----------|-----------|------------|
| $n_{11k}$ | $n_{12k}$ | $n_{1..k}$ |
| $n_{21k}$ | $n_{22k}$ | $n_{2..k}$ |
| $n_{.1k}$ | $n_{.2k}$ | $n_{..k}$  |

TABLE 15.8. Detail for calculation of the Cochran–Mantel–Haenszel test of the polio example. See the discussion in Section 15.7, where we find the that the Mantel–Haenszel chi-square test without the continuity correction is 16.54.

(twtb/code/salk-mh.s), (twtb/code/salk.s), (twtb/transcript/salk.st),  
 (twtb/code/salk2.sas), (twtb/transcript/salk2.1st)

| Age Group    | Observed |     | Expected |       | prob<br>no.par | Chi-Square |         | [1,1] position for<br>Mantel–Haenszel test |       |       |      |    |       |      |  |
|--------------|----------|-----|----------|-------|----------------|------------|---------|--|-------|-------|------|----|-------|------|--|
|              |          |     |          |       |                | chisq      | p.chisq | O  | E     | O–E   | var  | n  | dev   | mh   |  |
|              | no.par   | par | no.par   | par   |                |            |         |  |       |       |      |    |       |      |  |
| <b>0–4</b>   |          |     |          |       |                |            |         |  |       |       |      |    |       |      |  |
| no.vac       | 10       | 24  | 15.00    | 19.00 | 0.294          | 5.965      | 0.015   | 10   | 15.00 | -5.00 | 4.25 | 68 | -2.42 | 5.88 |  |
| vac          | 20       | 14  | 15.00    | 19.00 | 0.588          |            |         |  |       |       |      |    |       |      |  |
| <b>5–9</b>   |          |     |          |       |                |            |         |  |       |       |      |    |       |      |  |
| no.vac       | 3        | 15  | 7.20     | 10.80 | 0.167          | 6.806      | 0.009   | 3  | 7.20  | -4.20 | 2.65 | 45 | -2.58 | 6.65 |  |
| vac          | 15       | 12  | 10.80    | 16.20 | 0.556          |            |         |  |       |       |      |    |       |      |  |
| <b>10–14</b> |          |     |          |       |                |            |         |  |       |       |      |    |       |      |  |
| no.vac       | 3        | 2   | 3.00     | 2.00  | 0.600          | 0.000      | 1.000   | 3  | 3.00  | 0.00  | 0.67 | 10 | 0.00  | 0.00 |  |
| vac          | 3        | 2   | 3.00     | 2.00  | 0.600          |            |         |  |       |       |      |    |       |      |  |
| <b>15–19</b> |          |     |          |       |                |            |         |  |       |       |      |    |       |      |  |
| no.vac       | 1        | 6   | 3.11     | 3.89  | 0.143          | 4.219      | 0.040   | 1  | 3.11  | -2.11 | 1.12 | 18 | -2.00 | 3.99 |  |
| vac          | 7        | 4   | 4.89     | 6.11  | 0.636          |            |         |  |       |       |      |    |       |      |  |
| <b>20–39</b> |          |     |          |       |                |            |         |  |       |       |      |    |       |      |  |
| no.vac       | 7        | 5   | 8.44     | 3.56  | 0.583          | 1.501      | 0.221   | 7  | 8.44  | -1.44 | 1.44 | 27 | -1.20 | 1.45 |  |
| vac          | 12       | 3   | 10.56    | 4.44  | 0.800          |            |         |  |       |       |      |    |       |      |  |
| <b>40+</b>   |          |     |          |       |                |            |         |  |       |       |      |    |       |      |  |
| no.vac       | 3        | 2   | 3.33     | 1.67  | 0.600          | 0.600      | 0.439   | 3  | 3.33  | -0.33 | 0.22 | 6  | -0.71 | 0.50 |  |
| vac          | 1        | 0   | 0.67     | 0.73  | 1.000          |            |         |  |       |       |      |    |       |      |  |

Also define

$$e_{ijk} = \frac{(n_{i..})(n_{..j})}{n_{..k}}$$

to be the expected  $(i, j)$  cell count under independence in table  $k$ , and

$$V(n_{11k}) = \frac{n_{1..k} n_{2..k} n_{1..k} n_{2..k}}{n_{..k}^2 (n_{..k} - 1)} = \frac{e_{11k} e_{22k}}{(n_{..k} - 1)}$$

to be the estimated variance of  $n_{11k}$  under the assumption of a hypergeometric distribution of a  $2 \times 2$  table with fixed margins. Then we make a normal approximation and work with

$$n_{11k} \sim N(e_{11k}, V(n_{11k}))$$

The sum  $\sum_k n_{11k}$  is also approximately normal with mean  $\sum_k e_{11k}$  and variance  $\sum_k V(n_{11k})$ . We therefore use as the test statistic the quantity

$$M^2 = \frac{\left[ \sum_k n_{11k} - \sum_k e_{11k} \right]^2}{\sum_k V(n_{11k})} \quad (15.9)$$

and the *p*-value of the test is  $1 - \mathcal{F}_{\chi^2}(M^2 \mid 1)$ , the corresponding tail percentage of the chi-square distribution with 1 df.

Sometimes we will wish to use a variant of  $M^2$  with a continuity correction and then we use

$$M^2 = \frac{\left[ \left| \sum_k n_{11k} - \sum_k e_{11k} \right| - .5 \right]^2}{\sum_k V(n_{11k})} \quad (15.10)$$

Inspecting the form of  $M^2$  tells us that significance can occur under either of two conditions:

1. Most or all tables must have the observed (1,1) cell count at least as large as expected under the null hypothesis.
2. Most or all tables must have the observed (1,1) cell count at most as small as expected under the null hypothesis.

Equivalently, most or all of the tables must have an odds ratio either

1. at least 1, or
2. at most 1.

Note that  $M^2$  is **not** the same as the chi-square statistic one gets from the  $2 \times 2$  table formed as the sum of the  $k$  tables.

## 15.7 Example—Salk Polio Vaccine

(Chin et al., 1961), also in (Agresti, 1990), discuss 174 polio cases classified by age of subject, whether or not the subject received the Salk polio vaccine, and whether the subject was ultimately paralyzed by polio. The data file is (`datasets/salk.dat`). We wish to learn if symptom status (paralysis or not) is independent of vaccination status after controlling for age.

Each of the  $k=6$  “Observed” subtables in Table 15.8, one for each of  $k=6$  age ranges, shows two estimated probabilities of no paralysis, for subjects

without vaccine and subjects with vaccine. In the “0–4” subtable, for example, we see  $p_{\text{no.vac}}(\text{no.par})=.294$  and  $p_{\text{vac}}(\text{no.par})=.588$ . In all cases the observed proportion with vaccine is higher. The “chi-square” column shows the ordinary contingency table chi-square for each subtable. The four subtables with older subjects do not have many observations and do not strongly support the conclusion that vaccine is better. The Cochran–Mantel–Haenszel test provides a way of combining the information, properly weighted, from all six subtables to get a stronger conclusion. The “O”, “E”, and “O–E” columns show the [1,1] or [no.vac,no.par] position from the “Observed” and “Expected” tables. “O–E” is the weighted difference of the row probabilities  $O_i - E_i = w_i(p_{\text{no.vac}}(\text{no.par}) - p_{\text{vac}}(\text{no.par}))$  [with weights  $w_i = 1/(1/n_1 + 1/n_2)$  for the  $i^{\text{th}}$  subtable, where  $n_j$  is the total count on the  $j^{\text{th}}$  row]. While we choose to focus on the counts of the [1,1] cells, an identical conclusion would be reached if the focus were on any of the three other cells of the  $2 \times 2$  table.

The “var” column shows the variance of “O–E” under the assumption of the hypergeometric distribution for  $O_i$  assuming both row and column margins of the  $i^{\text{th}}$  table are fixed. The “dev” column is a standardized deviation,  $(O-E)/\sqrt{\text{var}}$ , and the “mh” column is the squared standardized deviation. We plot the standardized deviations in Figure 15.6. The squared standardized deviation is the Mantel–Haenszel statistic for the subtable. The MH statistic for a subtable is very close to the chi-square statistic.

The Cochran–Mantel–Haenszel (CMH) test for the set of all  $k=6$  subtables is constructed as a weighted combination of the same components used for the subtable statistics. Since each “O–E” is a random variable with mean and variance, we use Equations (3.7) and (3.8) to combine them. The

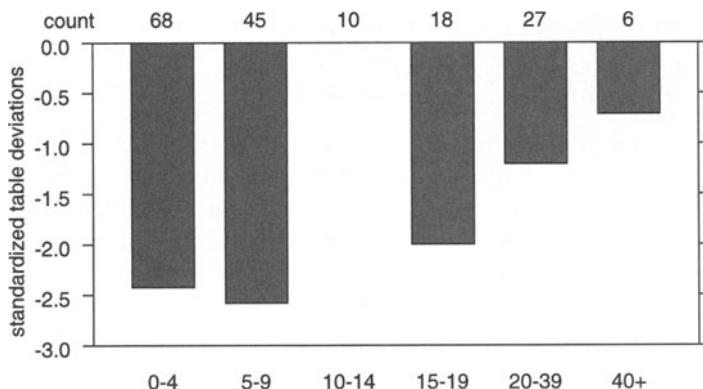


FIGURE 15.6. Standardized deviations for individual table Mantel–Haenszel values.  
(twtb/code/salk-mh.s), (twtb/figure/salk.dev.eps.gz)

CMH statistic is constructed from the sum of the “O–E” for the subtables, divided by the standard deviation of the sum, which is the square root of the sum of the variances:  $\sum(O-E)/\sqrt{\sum(var)}$ . Then the whole is squared. Thus the CMH statistic for this example is

$$\frac{(\sum(O-E))^2}{\sum(var)} = \frac{(-5 - 4.20 - 0 - 2.11 - 1.44 - .33)^2}{(4.25 + 2.65 + .67 + 1.12 + 1.44 + .22)} = 16.54$$

For each of the age ranges with a sufficiently large sample, Fisher’s exact test [see file (`twtb/transcript/salk2.1st`)] performed on the  $2 \times 2$  tables detects a positive association between symptom and vaccination status: Persons vaccinated had a significantly lower incidence of paralysis than persons not vaccinated. In addition, each of the six  $2 \times 2$  tables has an odds ratio of at least 1. Therefore, the Mantel–Haenszel test can be used, and the test statistic has value 16.54, which is highly significant. This means that the relationship between symptom and vaccination status is consistent over all age ranges.

## 15.8 Exercises

- 15.1. (Hand et al., 1994) revisit a dataset attributed to Karl Pearson, (`datasets/crime.dat`), that examines the relationship between type of crime committed and whether the perpetrator was a drinker or abstainer. Investigate whether these two classifications are independent, and if they are not, discuss the nature of the dependence.
- 15.2. (Senie et al., 1981), also in (Hand et al., 1994), investigate whether the frequency of breast self-examination is related to age group. The data appear in the file (`datasets/selfexam.dat`). Do you agree that there is a relationship? If so, describe it.
- 15.3. (Sokal and Rohlf, 1981), later in (Hand et al., 1994), concern an experiment to determine the preference of invading ant colonies on two species of acacia tree. A total of 28 trees was made available for the study, 15 of species A and 13 of species B. Initially each tree was treated with insecticide to remove all existing colonies. Then 16 ant colonies were invited to invade any of the trees they chose. By construction, the  $2 \times 2$  data, in (`datasets/acacia.dat`), have both margins fixed. Use Fisher’s exact test to determine if the ants have a significant preference for one species over the other.
- 15.4. (Fleiss, 1981) presents the data in (`datasets/mortality.dat`) concerning mortality following 37,840 live births to nonwhite mothers in New York City in 1974. In the file, rows are birth weights,  $\leq 2500$  grams or

$>2500$  grams, and columns are outcomes one year after birth, dead or alive. Construct and carefully interpret the sample odds ratio for these data and construct a 95% confidence interval for the population odds ratio.

- 15.5. (Wynder et al., 1958), later in (Fleiss, 1981), report a retrospective study of factors associated with cancer of the oral cavity. In this study there were 34 women with this cancer and 214 women, matched by age, without it. It was found that 24% of the cases but 66% of the controls were nonsmokers. Construct and carefully interpret the sample odds ratio for these data and then construct a 95% confidence interval on the population odds ratio.
- 15.6. (Braungart, 1971) refers to ([datasets/political.dat](#)), also found in (Bishop et al., 1975), in which 271 1960s college students who admitted to extreme political leanings were cross classified according to the style of parental decision making they received, authoritarian or democratic, and their political leaning, left or right. Construct and carefully interpret the sample odds ratio for these data and also construct a 95% confidence interval on the population odds ratio.
- 15.7. (Westbrooke, 1998) discusses claims that Maori are underrepresented on juries in districts in New Zealand. Jury pool composition data for the Rotorua and Nelson districts are shown in Table 15.9, along with totals for these two districts combined.

From this table it is easy to verify the following:

- The population of Rotorua is 27.0% Maori, but this district's jury pool is only 23.4% Maori.
- The population of Nelson is 3.9% Maori, but this district's jury pool is only 1.7% Maori.
- However, the combined population of these two districts is 15.3% Maori, but the combined jury pools of these districts is 20.3% Maori.

Discuss whether Maori are indeed underrepresented on the juries of these two districts.

TABLE 15.9. Jury pool composition data for the Rotorua and Nelson districts.

|                   | Rotorua |           | Nelson |           | Combined |           |
|-------------------|---------|-----------|--------|-----------|----------|-----------|
|                   | Maori   | Non-Maori | Maori  | Non-Maori | Maori    | Non-Maori |
| [1.5ex] Jury pool | 79      | 258       | 1      | 56        | 80       | 314       |
| Others            | 8,810   | 23,751    | 1,328  | 32,602    | 10,138   | 56,353    |

TABLE 15.10. Chronic Obstructive Pulmonary Disease.  
(twtb/transcript/intubate.s)

| Ventilation therapy | Center |     |       |     |       |     |       |     |       |     |
|---------------------|--------|-----|-------|-----|-------|-----|-------|-----|-------|-----|
|                     | 1      |     | 2     |     | 3     |     | 4     |     | 5     |     |
|                     | invas  | not | invas | not | invas | not | invas | not | invas | not |
| vent.ther           | 3      | 6   | 2     | 3   | 1     | 7   | 0     | 5   | 5     | 11  |
| not.vent            | 9      | 0   | 5     | 1   | 4     | 5   | 3     | 1   | 10    | 4   |

**15.8.** (Brochard et al., 1995) report a prospective study of patients with chronic obstructive pulmonary disease, assessing the effect of noninvasive ventilation therapy on reducing the need for subsequent invasive intubation. A total of 85 patients were recruited at five centers. The data are in Table 15.10 and file ([datasets/intubate.dat](#)). The data in the file are arranged differently than in the table: The first column is the center number; the second column entries are yes if received ventilation therapy and no if didn't receive this therapy; the third column entries are yes if required invasive intubation and no if didn't require invasive intubation; and the entries in column 4 are the number of patients in the categories specified in columns 1–3.

Compute the odds ratio at each center. Use the Mantel-Haenszel test to produce a carefully stated conclusion of the combined data.

# Nonparametrics

## 16.1 Introduction

Most of the statistical procedures we've introduced in the previous 15 chapters require an assumption about the form of a probability distribution, often the Normal distribution. When such assumptions are unjustified, the consequences of the procedure are dubious at best. In situations where distributional assumptions cannot be justified, even after a well-chosen data transformation, the analyst should consider another approach.

*Nonparametric* statistical procedures are ones that do not require the assumption of a specific probability distribution. (In contrast, procedures that do make distributional assumptions are referred to as parametric procedures.) In exchange for not requiring a detailed distributional assumption, nonparametric testing procedures have less power, and nonparametric confidence intervals are wider than their parametric analogues in the same problem situation using the same error control. In addition, the parameter(s) of interest in a nonparametric procedure may not be identical to those of corresponding parametric procedures. For example, we may use a nonparametric procedure for comparing two population medians as a substitute for a parametric procedure for comparing two means. For these reasons, parametric procedures are preferred if their assumptions are reasonably well met.

In previous chapters we have discussed a number of procedures for inferring about one or more population means. Those procedures required assumptions that underlying populations are normally distributed, at least approximately, and that when inferring about the means of two

or more populations, these populations have a common variance. Often transformations such as those discussed in Chapter 4 can be used to make such assumptions tenable once the transformations have been applied. Sometimes this is not possible, for example, when the data contain outliers whose elimination cannot be justified. In such instances, a nonparametric approach may be considered. Our discussion here is limited to hypothesis tests. Analogous confidence interval estimates cannot be described succinctly and are discussed in (Lehmann, 1998) and (Desu and Raghavarao, 2003).

Many nonparametric procedures use statistics based on *ranks* of the data rather than the data themselves. Such procedures require only that the data be on at least an ordered scale. In contrast, parametric procedures generally carry the more stringent requirement that the data be measured on an interval or ratio scale. The various scale types are defined in Section 2.1.

We've actually encountered some nonparametric procedures in earlier chapters. For example, the two-sample Kolmogorov–Smirnov goodness-of-fit test discussed in Section 5.9 does not require that we specify the natures of the two populations being sampled. In this chapter we introduce some additional commonly used nonparametric procedures. Our examples include checks of the assumptions of competing parametric procedures and comparisons of the nonparametric and parametric results.

## 16.2 Sign Test for the Location of a Single Population

Example 5.2 discusses a parametric approach to a problem involving ([datasets/vocab.dat](#)), concerning whether  $\mu = 10$  is consistent with a random sample of 54 test scores. Since Figure 5.1 shows that the sample had one high outlier, there was at least some doubt that the assumptions underlying the parametric analysis were correct.

The nonparametric approach here is the *sign test* to assess the hypothesis that the population median equals 10. If the population is not symmetric so that its mean and median are not identical, then the analogy between the nonparametric and parametric inferences is imperfect.

The logic of this nonparametric test stems from the insight that if the population median  $\eta$  is indeed 10, we would expect half of the sample values not exactly equal to 10 to fall on either side of 10. If appreciably more than half exceed 10, this suggests that the median exceeds 10. Let  $n$  be the number of sample items different from 10 and  $m$  be the number of

these exceeding 10. The formal test of

$$H_0: \eta = 10$$

vs

$$H_1: \eta > 10$$

is based on the distribution of a binomial random variable  $X$  with  $n$  trials and success probability .5. The test has  $p\text{-value} = P(X \geq m)$ . The two-sided test, with the same null hypothesis but having alternative hypothesis

$$H_1: \eta \neq 10$$

has  $p\text{-value} = 2P(X \geq \max(n - m, m))$ .

This test is called the sign test because it is based on the arithmetic signs of the differences between the data and the null hypothesized median. Some authors refer to it as the binomial test.

In ([datasets/vocab.dat](#)), 4 of the 54 scores equalled the null value  $\eta=10$ , and  $n=50$  scores were not exactly equal to the null value of  $\eta=10$ . We observed  $m=49$  scores that exceeded 10 and  $n - m=1$  score less than 10. If the null were true, then  $X$  comes from the distribution  $\text{Bin}(n = 50, p = .5)$

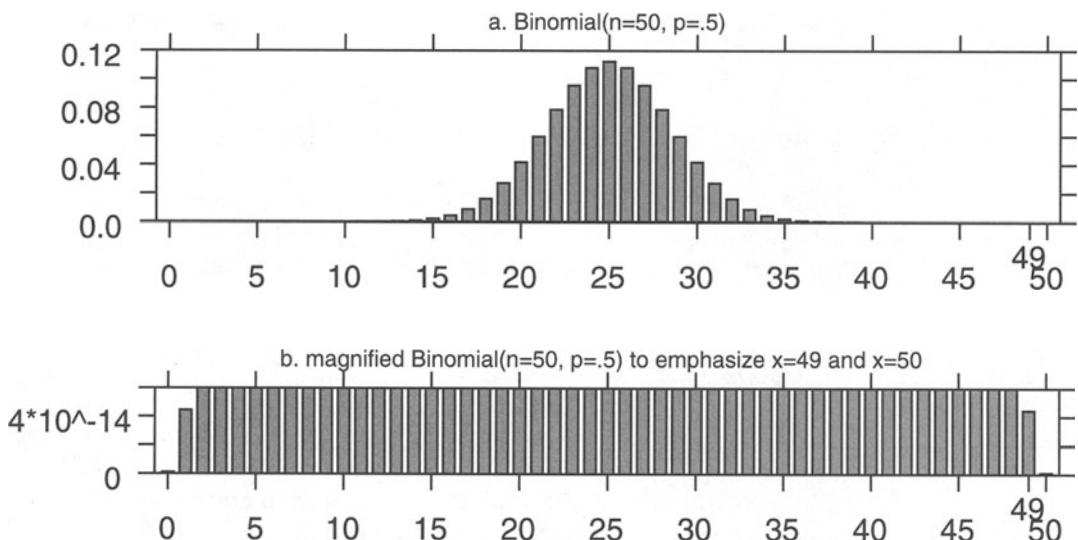


FIGURE 16.1. Panel a shows the  $\text{Bin}(n = 50, p = .5)$  discrete density. Panel b magnifies (and truncates) the probability scale to emphasize that there are nonzero values at  $x = 49$  and  $x = 50$ . The  $p\text{-value}$  for the sign test is the probability that a random selection from this distribution would yield the observed  $x = 49$  or the larger  $x = 50$ . From the graph we see that this is the very small number  $p \approx 4 \times 10^{-14}$ .

([npar/code/vocab2.s](#)), ([npar/figure/vocab.sign.eps.gz](#))

with discrete density shown in Figure 16.1. The one-sided  $p$ -value is the probability of observing 49 or more (that is, 49 or 50) larger scores from this distribution. This probability is the sum of the probabilities for the bars at 49 and 50. Therefore, the one-sided  $p$ -value is

$$P(X \geq 49) = 1 - \mathcal{F}_{Bi}(48 | 50, .5)$$

using the binomial distribution with  $n = 50$  and  $p = .5$ . This number  $p \approx 4 \cdot 10^{-14}$  is calculated in SAS by

`1 - PROBBNML(.5, 48, 50)`

or in S-PLUS by

`1 - pbinom(48, 50, .5)`

$p \approx 4 \cdot 10^{-14}$  is overwhelming evidence that the population median  $\eta$  exceeds 10, consistent with the strong parametrically based conclusion in Example 5.2 that the population mean  $\mu$  exceeds 10.

The procedure we just developed is an example of an *exact test* or of a *randomization test*. In this form of testing procedure, a discrete model for the distribution of the observed statistic under the null hypothesis is postulated. Then the probability of obtaining the observed value (or larger) from that model is calculated and used as the  $p$ -value of the test.

## 16.3 Comparing the Locations of Paired Populations

Just as the parametric paired  $t$ -test is equivalent to a one-sample  $t$ -test on the pair differences, the locations of paired populations can be compared by inferring about the median of the population consisting of the differences (in a consistent direction) between the two individual populations.

### 16.3.1 Sign Test

Exercise 5.13 requests a comparison of the population means of pre- and posttreatment measurements based on a random sample of 48 patients. However, the stem-and-leaf display of the posttreatment measurements in Table 16.1 shows a number of both low and high outliers, suggesting that the post population may not be close to normally distributed, and calling into question the validity of an analysis based on a two-sample  $t$ -test. Therefore, a nonparametric approach should be considered.

We show the sign test for paired observations in Table 16.2. Analogous to Section 16.2, let  $n$  be the number of pairs of observations, excluding tied pairs. For testing against the one-sided alternative hypothesis that the

TABLE 16.1. The stem-and-leaf display shows a number of both low and high outliers, suggesting that the post population may not be close to normally distributed. The outliers are well outside the vicinity of the bulk of the data. To avoid having these observations consume most of the vertical dimension of this diagram, S-PLUS lists the outliers following the descriptors **low:** and **high:** respectively.  
 (npar/code/har1.s)

---

```
S-PLUS (npar/transcript/har1.st):
> har1 <- read.table(hh("datasets/har1.dat"), header=T)
> stem(har1$Post)

N = 48 Median = -6.1
Quartiles = -7.55, -4.4

Decimal point is at the colon

Low: -14.2

-10 : 7
-9 : 321
-8 : 77330
-7 : 865400
-6 : 8755432200
-5 : 9631
-4 : 9988544
-3 : 54
-2 : 886650
-1 :
-0 :
 0 : 6
 1 : 2

High: 4.6
```

---

posttreatment median is less than the pretreatment median, let  $m$  be the number of sample pairs where the posttreatment measurement is less than the pretreatment measurement. Then for binomially distributed  $X$  with  $n$  trials and success probability 0.5, the  $p$ -value is  $P(X \leq m)$ . For these data,  $n = 46, m = 15$  and we use S-PLUS to calculate  $p = pbinom(15, 46, .5) \approx .0129$ . This  $p$ -value indicates moderate evidence that in the population of patients from which this sample was selected, the median posttreatment angle is less than the median pretreatment angle. For the analogous parametric paired  $t$ -test requested in Exercise 5.13, the  $p$ -value is considerably

TABLE 16.2. Sign test applied to relative rotation angle data. The conclusion of the test is that the median posttreatment angle is less than the median pretreatment angle.  
 (npar/code/har1.s)

---

```

S-PLUS (npar/transcript/har1a.st):
> ## sign test
> tmp <- table(sign(har1$Post - har1$Pre))
> tmp
-1 0  1
31 2 15
>
> pbinom(q    = tmp["1"],
+         size = sum(tmp[c("-1","1")]),
+         prob = .5)
      1
0.01294804

```

---

less than 0.01. Because we question the validity of the  $t$ -test, we believe the  $p$ -value from the  $t$ -test is way too small.

### 16.3.2 Wilcoxon Signed-Ranks Test

The sign test in Section 16.3.1 uses the arithmetic sign of differences between members of each pair but ignores the magnitude of these differences. Such information can be crucial, particularly if the samples contain outlying values. The Wilcoxon signed-ranks test uses information on the magnitude of the differences and therefore is more sensitive than the sign test. The Wilcoxon test assumes an interval or ratio measurement scale involving continuous variables. If the data are ordinal but not interval, the Wilcoxon test cannot be used.

The construction of the signed-rank test for (datasets/har1.dat) is depicted in Table 16.3. For testing hypotheses involving the difference between the medians of two populations, say the  $X = \text{har1$Pre}$  population and  $Y = \text{har1$Post}$  population, let  $D_i = X_i - Y_i = \text{har$diff}$  be the observed difference for the  $i^{\text{th}}$  pair. Let  $n = 46$  be the number of such differences that are not zero; hereafter ignore any zero differences. Rank the  $n$  nonzero absolute differences ( $|D_i|, i = 1, \dots, n$ ) =  $\text{har$abs}$  in ascending order from 1 to  $n$  and store the ranks in  $\text{har$rank}$ . (If there are ties, assign the average rank. For example, if the first and second largest absolute differences are both equal to .3, assign rank 1.5 to both absolute differences.) Copy the ranks for the positive differences into  $\text{har$prnk}$ . Then the test statistic is

TABLE 16.3. Detail for construction of the Wilcoxon signed-ranks test applied to relative rotation angle data. The **diff** column contains the differences Pre - Post. The **abs** column contains the absolute value of the differences. The **rank** column contains the ranks of the absolute differences. The **prnk** column contains the same ranks, but only for rows that have positive differences. The test statistic is the sum of the **prnk** column.  
(npar/code/har2.s)

---

S-PLUS (npar/transcript/har2.st):

```
> har[order(har$abs),]
   diff abs rank prnk      diff abs rank prnk      diff abs rank prnk      diff abs rank prnk
  4  0.3 0.3  1.5  1.5    12 -0.7 0.7 13.0  0.0    11 -1.9 1.9 25.0  0.0    47  4.1 4.1 37.0 37.0
 22  0.3 0.3  1.5  1.5    15 -0.8 0.8 14.0  0.0    33  2.1 2.1 26.0 26.0    40  4.2 4.2 38.0 38.0
 13 -0.4 0.4  3.5  0.0    42  0.9 0.9 15.0 15.0    18  2.2 2.2 27.0 27.0    38  4.3 4.3 39.0 39.0
 35  0.4 0.4  3.5  3.5    37  0.9 0.9 16.0 16.0    36  2.2 2.2 28.5 28.5    14  4.5 4.5 41.0 41.0
  1  0.4 0.4  5.0  5.0    20 -1.0 1.0 17.0  0.0    44 -2.2 2.2 28.5  0.0    24  4.5 4.5 41.0 41.0
 17 -0.4 0.4  7.0  0.0    28 -1.1 1.1 18.0  0.0    46  2.3 2.3 30.0 30.0    30  4.5 4.5 41.0 41.0
 21 -0.4 0.4  7.0  0.0    27 -1.1 1.1 19.0  0.0    25  2.3 2.3 31.0 31.0    34 -4.6 4.6 43.0  0.0
 39  0.4 0.4  7.0  7.0    48 -1.1 1.1 20.0  0.0     5  2.9 2.9 32.0 32.0    29  5.0 5.0 44.0 44.0
 19  0.5 0.5  9.5  9.5    10  1.4 1.4 21.0 21.0    43  3.2 3.2 33.0 33.0     8  7.5 7.5 45.0 45.0
 45  0.5 0.5  9.5  9.5    32 -1.5 1.5 22.0  0.0     6  3.2 3.2 34.0 34.0    41  9.0 9.0 46.0 46.0
 16  0.7 0.7 11.5 11.5     2  1.7 1.7 23.0 23.0    23  3.8 3.8 35.0 35.0     3  0.0 NA 0.0  0.0
 26 -0.7 0.7 11.5  0.0     7 -1.7 1.7 24.0  0.0    31  3.9 3.9 36.0 36.0     9  0.0 NA 0.0  0.0
```

---

the sum of the ranks in **har\$prnk**. The idea here is that if the population medians are equal, we expect half of the differences to be positive and so the test statistic should be approximately one half the sum of the ranks of the nonzero differences, or  $n(n + 1)/4$ , where  $n$  is now the number of nonzero differences. A test statistic much different from this value suggests that the difference between the population medians is nonzero.

Figure 16.2 show the differences and the ranks of the absolute values of these differences used in a Wilcoxon signed-ranks test for the data in (datasets/har1.dat). The positive differences (between the pre- and post-treatment angles) in this example generally have larger absolute values than the negative differences. The magnitudes of these differences is relevant to the test result—a large positive difference is greater evidence that pre exceeds post than a small positive difference. These magnitudes are ignored by the sign test but accounted for by the Wilcoxon signed-ranks test.

Assuming one-sided alternative hypothesis as in Section 16.2, this test is performed in S-Plus with

```
wilcox.test(X, Y, alternative="greater", paired=T, exact=T)
```

If very restrictive conditions are met, S-Plus calculates the  $p$ -value based on the exact distribution of the test statistic. These conditions are  $n \leq 25$ ,

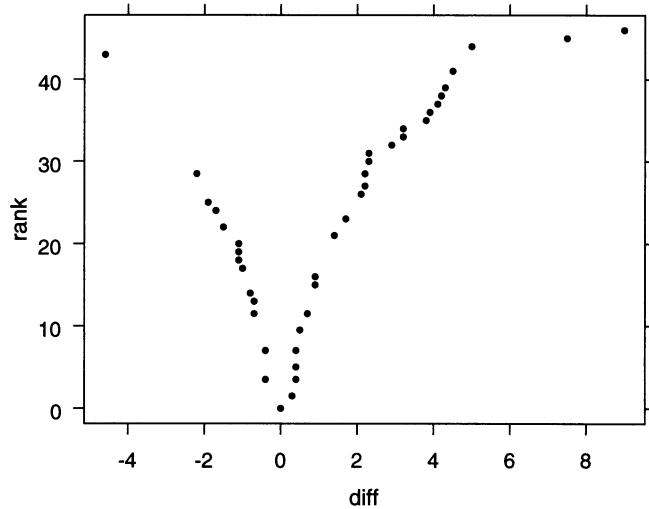


FIGURE 16.2. The graph depicts the differences and ranks from Table 16.3 used in the Wilcoxon test of (*datasets/har1.dat*). Observe that the positive differences between pre- and posttreatment angles tend to have larger magnitudes than the negative differences. The Wilcoxon test, which takes account of these magnitudes, has more power than the sign test, which ignores the magnitudes. The test statistic is the sum of the ranks for positive differences.

(*npar/code/har2.s*), (*npar/figure/har1.rank.diff.eps.gz*)

TABLE 16.4. Wilcoxon signed-ranks test applied to relative rotation angle data. The *exact=F* argument to *wilcox.test* suppresses warning messages generated by S-PLUS when *n* is too large for its exact-test algorithm. The conclusion of the test is that the median posttreatment angle is less than the median pretreatment angle.

(*npar/code/har1.s*)

---

S-PLUS (*npar/transcript/har1b.st*):

```
> wilcox.test(har1$Pre, har1$Post, alternative="greater", paired=T, exact=F)

Wilcoxon signed-rank test

data: har1$Pre and har1$Post
signed-rank normal statistic with correction Z = 2.9082, p-value = 0.0018
alternative hypothesis: true mu is greater than 0
```

---

no zero differences, and no tied ranks. Otherwise, as in the analysis of the (*datasets/har1.dat*) dataset, the *p*-value is based on a normal approximation. The results of the *wilcox.test* command applied to (*datasets/har1.dat*) are presented in Table 16.4.

The Wilcoxon signed-ranks test is performed in SAS by applying *proc univariate* to the difference  $D = X - Y$ . The SAS code and edited output are shown in Table 16.5. *proc univariate* calculates an exact *p*-value if  $n \leq 20$  and a Student's *t* approximation, different from the approximation used by S-PLUS when, as in this example,  $n > 20$ . With the approximation used by S-PLUS, the two-sided *p*-value is .0018. With the approximation used by SAS, the *p*-value is .0021. The small difference in the *p*-values arises from slightly different normal approximations used by the programs.

The fact that the *p*-value for the Wilcoxon test is much less than that of the sign test for the same data demonstrates that the signed-ranks test is more powerful than the sign test.

TABLE 16.5. Wilcoxon signed-ranks test applied to relative rotation angle data. The test conclusion is that the median posttreatment angle is less than the median pretreatment angle.  
(*npar/transcript/har1.lst*)

---

SAS (*npar/code/har1.sas*):

```
data har1;
    infile "&hh/datasets/har1.dat" firstobs=2;
    input Pre Post;
    Diff = Pre - Post;
    run;

    proc univariate;
        var Diff;
    run;
```

---

SAS (*npar/transcript/har1.ed.lst*):

Tests for Location: Mu0=0

| Test        | -Statistic- |          | ----p Value----- |        |
|-------------|-------------|----------|------------------|--------|
| Student's t | t           | 3.556374 | Pr >  t          | 0.0009 |
| Sign        | M           | 8        | Pr >=  M         | 0.0259 |
| Signed Rank | S           | 271.5    | Pr >=  S         | 0.0021 |

---

## 16.4 Mann–Whitney Test for Two Independent Samples

Some authors refer to this test as the Wilcoxon–Mann–Whitney test because Frank Wilcoxon initiated its development. It is analogous to the parametric two-sample  $t$ -test but compares medians rather than means. We wish to compare the medians of two populations based on independent random samples from each. It is assumed that the measurement scale is at least ordinal. Combine the two samples and then rank the resulting observations in ascending order. As in Section 16.3.2, if there are tied observations, each should be assigned the average rank. The test statistic  $T$  is the sum of the ranks of the smaller sample. Assume the smaller sample is assigned index 1 and the larger sample index 2. Denoting the observed value of  $T$  as  $t_{\text{calc}}$ , the  $p$ -value depends on the form of the alternative hypothesis. If the alternative hypothesis is  $\eta_1 > \eta_2$ , then  $p\text{-value} = P(T > t_{\text{calc}})$ . For the alternative  $\eta_1 < \eta_2$ , the  $p$ -value is  $= P(T < t_{\text{calc}})$ . For the two-sided alternative,  $p\text{-value} = 2 \min(P(T > t_{\text{calc}}), P(T < t_{\text{calc}}))$ .

The data in (`datasets/balance.dat`), taken from (Teasdale et al., 1993), illustrate this test. These authors sought to compare the forward/backward sway in mm of 9 elderly and 8 young subjects who took part in an investigation of their reaction times. We see in Figure 16.3 that one of the `elderly` measurements is an extreme outlier, and if it is retained in the analysis the two-sample  $t$ -test is invalid. The alternative hypothesis is that the median sway of elderly subjects exceeds the median sway of young subjects. The S-PLUS analysis uses `wilcox.test` as in Section 16.3.2.

In Table 16.6 we list the two samples and their ranks after they are combined. The analysis with S-PLUS is in Table 16.7 and with SAS in Table

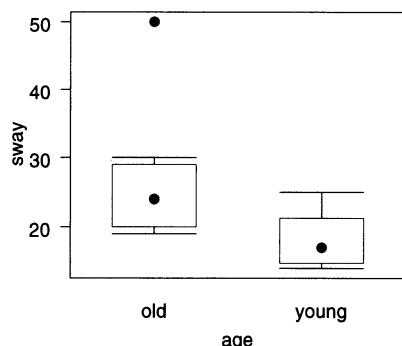


FIGURE 16.3. Balance data. The extreme outlier in the `age=="old"` group makes the two-sample  $t$ -test invalid. See Table 16.6.  
`(npar/code/balance.s)`, `(npar/figure/balance.eps.gz)`

TABLE 16.6. Mann-Whitney test applied to balance data. See Figure 16.3.  
(*npar/code/balance.s*)

| old  |       | young |       |
|------|-------|-------|-------|
| sway | ranks | sway  | ranks |
| 19   | 6.5   | 25    | 13.5  |
| 30   | 16.0  | 21    | 9.5   |
| 20   | 8.0   | 17    | 4.5   |
| 19   | 6.5   | 15    | 3.0   |
| 29   | 15.0  | 14    | 1.5   |
| 25   | 13.5  | 14    | 1.5   |
| 21   | 9.5   | 22    | 11.0  |
| 24   | 12.0  | 17    | 4.5   |
| 50   | 17.0  |       |       |

49.0 = rank sum, smaller sample

TABLE 16.7. Mann-Whitney test applied to balance data by S-PLUS. S-PLUS does not compute the exact test when some of the ranks are tied.

(*npar/code/balance.s*)

```
S-PLUS (npar/transcript/balance.st):
> wilcox.test(balance$sway[balance$age=="old"],
+               balance$sway[balance$age=="young"],
+               alternative="greater", exact=F)

Wilcoxon rank-sum test

data: balance$sway[balance$age == "old"] and
balance$sway[balance$age == "young"]
rank-sum normal statistic with correction Z = 2.1717, p-value = 0.0149
alternative hypothesis: true mu is greater than 0
```

16.8. The SAS analysis invokes the *wilcoxon* option in *proc npar1way*. S-PLUS and SAS agree that the one-sided *p*-value based on a normal approximation is .0149. SAS also provides the exact *p*-value .0119. We conclude that there is moderate evidence that, on average, elderly subjects have greater sway than young subjects.

TABLE 16.8. Mann-Whitney test applied to balance data by SAS.  
(npar/transcript/balance.lst)

---

SAS (npar/code/balance.sas):

```
data balance;
    infile "&hh/datasets/balance.dat";
    input sway age $;
    run;

proc npar1way wilcoxon;
    class age;
    exact;
    var sway;
    run;
```

---

S-PLUS (npar/transcript/balance.ed.lst):  
Wilcoxon Two-Sample Test

Statistic (S) 49.0000

Normal Approximation

|                   |         |
|-------------------|---------|
| Z                 | -2.1717 |
| One-Sided Pr < Z  | 0.0149  |
| Two-Sided Pr >  Z | 0.0299  |

t Approximation

|                   |        |
|-------------------|--------|
| One-Sided Pr < Z  | 0.0226 |
| Two-Sided Pr >  Z | 0.0453 |

Exact Test

|                           |        |
|---------------------------|--------|
| One-Sided Pr <= S         | 0.0119 |
| Two-Sided Pr >=  S - Mean | 0.0246 |

Z includes a continuity correction of 0.5.

---

## 16.5 Kruskal–Wallis Test for Comparing the Locations of at Least Three Populations

This test is the nonparametric analogue of the  $F$ -test for equality of the means of several populations. A natural generalization of the Mann–Whitney test, it tests the null hypothesis that all  $k$  populations have a common median. It is assumed that the measurement scale is at least ordered and that the  $k$  random samples are mutually independent.

Suppose  $n_i$  is the size of the sample from population  $i$  and  $n_+ = \sum_i n_i$ . Rank all  $n_+$  observations from 1 to  $n_+$  and let  $R_i$  be the sum of the ranks from sample  $i$ . As usual, in case of ties, assign average ranks to the tied values. The test statistic is

$$T = \frac{12}{n_+(n_+ + 1)} \sum_{i=1}^k \frac{[R_i - n_i(n_+ + 1)/2]^2}{n_i} \quad (16.1)$$

The idea behind this formula is that if the null hypothesis is exactly true, the expected sum of the ranks of sample  $i$  is  $E(R_i) = n_i(n_+ + 1)/2$ .

This test is conducted in S-PLUS with the command `kruskal.test(a, b)`, where `a` is the vector of sample observations and `b` is the same sized vector indicating the population number from which the observation came. The test is performed in SAS using `proc npar1way`, again invoking the `wilcoxon` option, which produces the Kruskal–Wallis analysis when  $k > 2$ .

Exercise 6.2 refers to an experiment comparing the pulse rates of workers while performing six different tasks (`datasets/pulse.dat`). From Figure 16.4 we see that the normality assumption required for a standard one-way analysis of variance is somewhat questionable as most tasks seem to show a uniform distribution of pulses. We therefore investigate the data analysis via the Kruskal–Wallis test.

SAS in Table 16.9 and S-PLUS in Table 16.10 agree that the chi-square approximation to the distribution of the statistic for this test yields a  $p$ -value of .0068. This is not too dissimilar from the one-way ANOVA  $p$ -value of .0015. We conclude, unsurprisingly, that the pulse rate of workers differs according to the task performed immediately prior to the pulse reading.

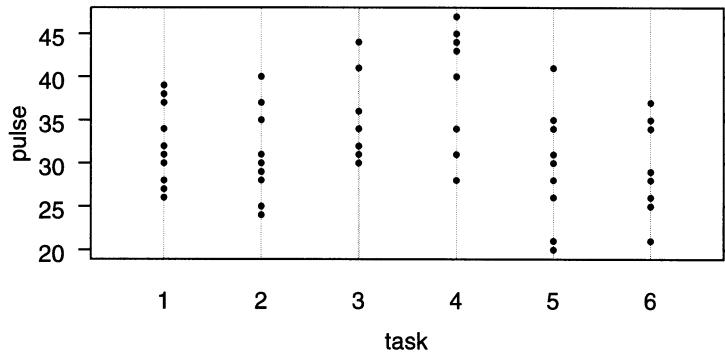


FIGURE 16.4. Pulse data. The uniform distributions within the groups makes the one-way ANOVA inappropriate. See Tables 16.9 and 16.10.

(*npar/code/pulse.s*), (*npar/figure/pulse.eps.gz*)

TABLE 16.9. Kruskal-Wallis rank sum test applied to pulse rate data by S-PLUS.  
(*npar/code/pulse.s*)

---



---

```
S-PLUS (npar/transcript/pulse.st):
> pulserate <- read.table(hh("datasets/pulse.dat"), header=T)
> names(pulserate) <- c("task", "pulse")
> kruskal.test(pulserate$pulse, pulserate$task)

Kruskal-Wallis rank sum test

data: pulserate$pulse and pulserate$task
Kruskal-Wallis chi-square = 15.9995, df = 5, p-value = 0.0068
alternative hypothesis: two.sided
```

---

TABLE 16.10. Kruskal-Wallis rank sum test applied to pulse rate data by SAS.

```
SAS (npar/code/pulse.sas):
data pulse;
    infile "&hh/datasets/pulse.dat" firstobs=2;
    input task pulse;
run;

proc npar1way wilcoxon;
    class task;
    var pulse;
run;
```

S-PLUS (npar/transcript/pulse.lst):  
 Wilcoxon Scores (Rank Sums) for Variable pulse  
 Classified by Variable task

| task | N  | Sum of Scores | Expected Under H0 | Std Dev Under H0 | Mean Score |
|------|----|---------------|-------------------|------------------|------------|
| 1    | 13 | 434.00        | 448.50            | 63.941948        | 33.384615  |
| 2    | 12 | 367.00        | 414.00            | 61.989407        | 30.583333  |
| 3    | 10 | 464.00        | 345.00            | 57.589969        | 46.400000  |
| 4    | 10 | 503.50        | 345.00            | 57.589969        | 50.350000  |
| 5    | 12 | 320.50        | 414.00            | 61.989407        | 26.708333  |
| 6    | 11 | 257.00        | 379.50            | 59.877907        | 23.363636  |

Average scores were used for ties.

#### Kruskal-Wallis Test

|                 |         |
|-----------------|---------|
| Chi-Square      | 15.9995 |
| DF              | 5       |
| Pr > Chi-Square | 0.0068  |

## 16.6 Exercises

- 16.1.** A study of (Darwin, 1876), discussed in (Hand et al., 1994) with the data in ([datasets/darwin.dat](#)), compared the ultimate heights of plants grown from otherwise comparable seedlings that were either cross-fertilized or self-fertilized. Compare the heights using the Wilcoxon signed-ranks procedure. Would a paired  $t$ -test have been appropriate?
- 16.2.** High levels of carbon monoxide transfer are a risk factor for contracting pneumonia. (Ellis et al., 1987), also in (Hand et al., 1994), studied the levels of carbon monoxide transfer in 7 chicken pox patients who were smokers. They were measured upon hospital admission and one week later. The data are in the file ([datasets/pox.dat](#)).
- Verify the inappropriateness of a paired  $t$ -test for these data. Discuss your reasoning.
  - Analyze using the Wilcoxon signed-ranks procedure.
- 16.3.** (Simpson et al., 1975), also in (Chambers et al., 1983), studied the amount of rainfall (measured in acre-feet) following cloudseeding. They seeded 26 clouds with silver nitrate and 26 clouds with a placebo seeding (that is the airplane went up but didn't release the silver nitrate). The data are contained in the file ([datasets/seeding.dat](#)). Assume these data represent independent random samples.
- Use the Mann–Whitney test to assess whether cloud seeding impacted rainfall.
  - Construct the ladder of powers graph (see Figure 4.18) to find a power transformation that makes the histograms of posttransformed samples symmetric and bell-shaped. Redo part 16.3a using the two-sample  $t$ -test instead of the Mann–Whitney test.
  - Discuss which of these procedures is preferred.
- 16.4.** (VanVliet and Gupta, 1973), later cited by (Hand et al., 1994), compared the birthweights in kg of 50 infants having severe idiopathic respiratory distress syndrome. Twenty-seven (27) of these infants subsequently died and the remainder lived. The data appear in the file ([datasets/distress.dat](#)). Perform a nonparametric test to address whether deaths from this cause are associated with low birthweight.
- 16.5.** Exercise 6.8 requested a comparison of the mean disintegration times of four types of pharmaceutical tablets. The data are in file ([datasets/tablet1.dat](#)). Compare the results of a Kruskal–Wallis test with the conclusion of the  $F$ -test for that exercise.

## Logistic Regression

Logistic regression is a technique similar to multiple regression with the new feature that the predicted response is a probability. Logistic regression is appropriate in the often-encountered situation where we wish to model a dependent variable which is either

**dichotomous:** The dependent variable can assume only the two possible values 0 and 1 (often as a coding of a two-valued categorical variable such as Male/Female or Treatment/Control).

**sample proportion:** The dependent variable is a probability and hence confined to the interval (0, 1).

The methodology of ordinary multiple regression analysis cannot cope with these situations because ordinary regression assumes that the dependent variable is continuous on the infinite interval  $(-\infty, \infty)$ . When this assumption on the dependent variable is not met, we must employ a suitable transformation—a *link* function—to change its range. One such transformation (shown in Equation (17.1) and Figure 17.1) from the closed interval  $[0, 1]$  to the set of all real numbers is the logarithm of the odds, known as the logit transformation

$$y = \text{logit}(p) = \ln\left(\frac{p}{1-p}\right) \quad (17.1)$$

The logit transformation is the key to logistic regression. S-PLUS functions for the logit and its inverse

$$p = \text{antilogit}(y) = \frac{e^y}{1 + e^y} \quad (17.2)$$

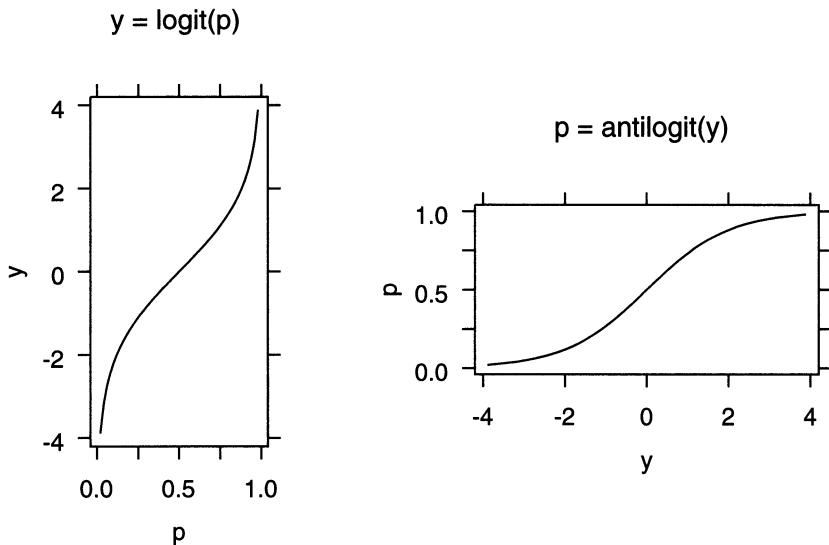


FIGURE 17.1. Panel a.  $y = \text{logit}(p)$ . Panel b.  $p = \text{antilogit}(y)$ .  
`(logi/code/logit-plot.s)`, `(logi/figure/logit-plot.eps.gz)`

are defined in `(splus.library/logit.s)`.

The model for logistic regression is

$$\text{logit}(p) = X\beta + \epsilon \quad (17.3)$$

where

$p$  The response is either a binary 0/1 variable indicating failure/success or a number in the range [0, 1] indicating an observed proportion of successes.

$X$  matrix of predictor variables.

$\beta$  vector of regression coefficients.

$\epsilon$  vector of discrepancies assumed to have a binomial distribution.

In the special case of a single continuous predictor, model (17.3) specializes to

$$\text{logit}(p) = \beta_0 + \beta_1 x + \epsilon \quad (17.4)$$

Logistic regression is a special case of a *generalized linear model*. There are two components to the generalization.

1. In ordinary linear models, the response variable is assumed to be linearly related to the predictor models. In generalized linear models a function (the link function) of the response variable is assumed to be linearly

related to the predictor models. In logistic models we usually use the logit function.

2. In ordinary linear models, the variance of the residuals is assumed to be normal. In *glm* (generalized linear models), the variance function is usually something else. With logistic models the variance function is assumed to be binomial.

Probit regression is another type of generalized linear model. Instead of using the logit link function, probit regression uses the inverse of the normal c.d.f.,  $\Phi^{-1}$ , defined in Section 3.9, to map from  $(0, 1)$  to the set of all real numbers. As with logistic regression, the variance function for probit models is usually assumed to be binomial. Probit regression and logistic regression give very similar results unless many of the estimated probabilities are close to either 0 or 1.

The predictor variables in logistic regression are the same types of continuous variables and factors that we use in ordinary multiple regression. All the familiar operations on the predictor variables (nesting, crossing, polynomial powers) are also appropriate with generalized linear models.

## 17.1 Example—The Space Shuttle Challenger Disaster

### Study Objectives

The NASA space shuttle has two booster rockets, each of which has three joints sealed with O-rings. A warm O-ring quickly recovers its shape after a compression is removed, but a cold one will not. An inability of an O-ring to recover its shape can allow a gas leak, which may lead to disaster. On January 28, 1986, the Space Shuttle Challenger exploded during the launch.

The coldest previous launch temperature was 53 degrees F. The temperature forecast for time of launch of Challenger on the morning of January 28, 1986, was 31 degrees Fahrenheit. On the evening of January 27, a teleconference was held among people at Morton Thiokol, Marshall Space Flight Center, and Kennedy Space Center. There was a substantial discussion among engineers over whether the flight should be cancelled. No statistician was present for any of these discussions.

### Data Description

The input dataset in (`datasets/spacshu.dat`) from (Dalal et al., 1989) contains two columns.

`tempF`: temperature in degrees Fahrenheit at time of launch

`damage`: 1 if an O-ring was damaged and 0 otherwise

Each launch has six cases, one for each O-ring. There are a total of  $23 \times 6 = 138$  cases (the O-rings for one flight were lost at sea).

### 17.1.1 Graphical Display

The five panels in Figures 17.2 and 17.3 show the relationship between number of damaged O-rings and launch temperature for space shuttle flights prior to the Challenger disaster. They clearly suggest a temperature effect. Logit regression can be used to model the probability of O-ring damage as a function of launch temperature, and hence estimate the probability that any one particular O-ring is damaged at launch temperature 31°F.

The logistic curves in Figures 17.3d and 17.3e decrease as the predictor variable `tempF` increases. This behavior differs from that displayed in Figure 17.1b because the logistic regression coefficient of `tempF` is negative. Models with a single predictor and a positive logistic regression coefficient have logit fits resembling Figure 17.1b.

The dataset (`datasets/spacshu.dat`) from (Dalal et al., 1989) includes one observation for each O-ring. Figure 17.2a shows the observed data, jittered (by adding random noise to break ties) so multiple O-rings at the same launch (hence same temperature) are visible. Note that the `tempF` scale includes only the observed temperatures. Figure 17.2b shows a simplistic model fit to the data. Within each 5-degree interval we have calculated the proportion of O-rings that were damaged. We see a strong indication that the proportion of damaged rings goes up as the temperature goes down. The graph suggests that the probability of damage will be high when the temperature reaches 31°F. There are several limitations to this inference. First, the graph doesn't extend to 31°F. Second, the model doesn't extend to 31°F.

Figure 17.3 extends the axes to include the temperature of the launch day in question. It is easier to see the suggested inference of high probability of damage even in panel c. We need a different model than simply averaging over a 5-degree range to clarify the impression. Panels d and e show the prediction bands for estimating probabilities of damage to individual O-rings (• in panel d) and for estimating the number of O-rings damaged per flight (× in panel e). Note this is an extrapolation. The fitted response values and their standard errors are calculated in Table 17.1

Figure 17.3d shows confidence bands for the number of damaged O-rings. The shape of the prediction bands in panels d and e are the same because

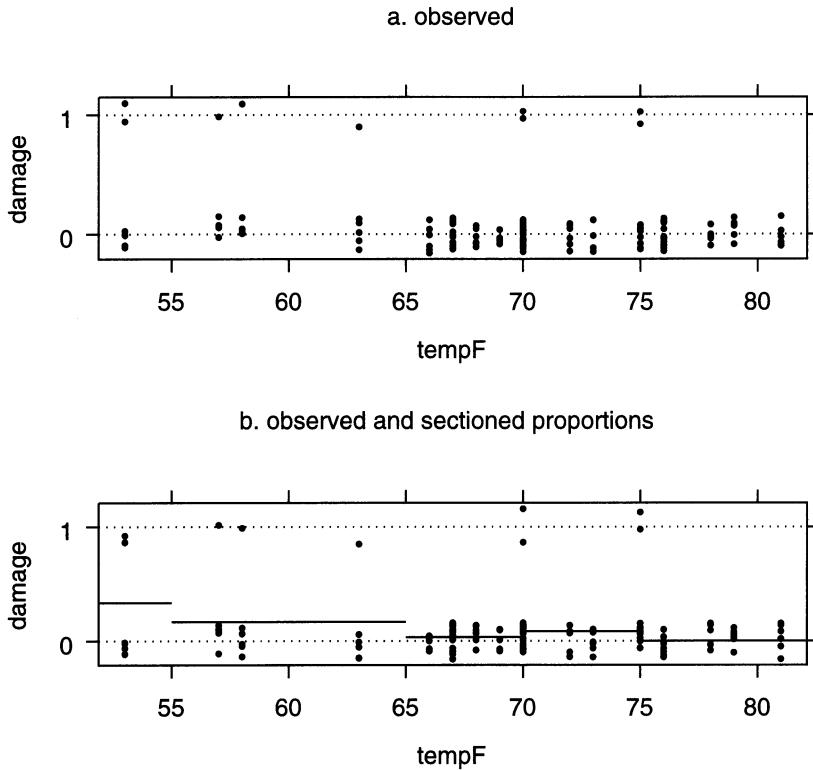


FIGURE 17.2. Panel a. Original data. Panel b. Sectioned fit.  
`(logi/code/spaceshuttle.s)`, `(logi/figure/spaceshuttle.ab.eps.gz)`

the expected number of damaged rings is the constant 6 (the number of O-rings on a shuttle) times the probability of damage to an individual ring. The file `(logi/figure/spaceshuttle.all.eps.gz)` superimposes Figures 17.3d and 17.3e.

At the January 28, 1986, teleconference, they displayed not Figure 17.3, but Figure 17.4, a figure showing only those launches with at least one damaged O-ring. Figure 17.4 is essentially just the top of our Figure 17.2a, the portion with damage=1. One cannot tell from such a chart that there is a temperature/damage relationship. Information about temperatures for launches without damage is highly relevant. Tragically, the assembled engineers did not realize the vital importance of seeing the number of launches without damage. They okayed the launch on January 28.

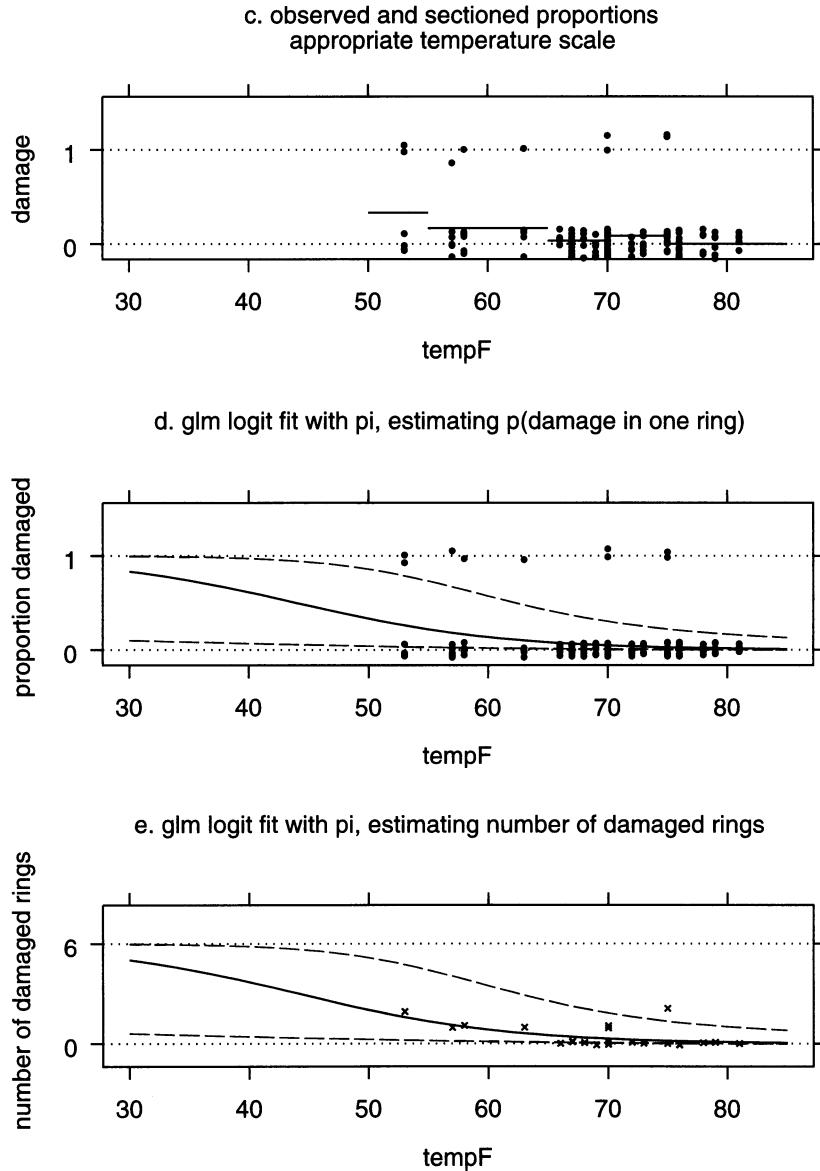


FIGURE 17.3. Panel c. Sectioned fit with appropriate temperature scale. Panel d. Logit fit with 95% prediction band focusing on estimating probability of damage in one ring on the next launch. Panel e. Logit fit with 95% prediction band focusing on estimating number of damaged rings in the next launch.

(logi/code/spaceshuttle.s), (logi/figure/spaceshuttle.cde.eps.gz)

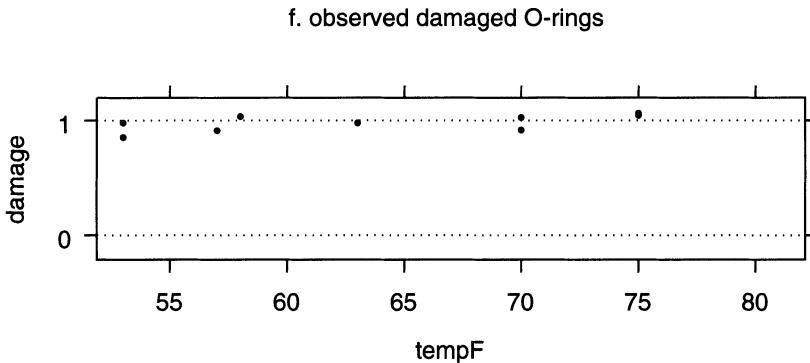


FIGURE 17.4. Panel f. Observed damaged O-rings, without the information about total number of rings.

(logi/code/spaceshuttle.s), (logi/figure/spaceshuttle.damaged.eps.gz)

### 17.1.2 Numerical Display

We show in Table 17.1 a logistic regression analysis of these data to the model

$$\text{logit}(p) = \beta_0 + \beta_{\text{tempF}} \text{tempF} + \epsilon \quad (17.5)$$

where  $p = P(\text{damage})$ . Discussion of the terminology used in this Table appears in Section 17.2.

The fitted equation is  $\text{logit}(\hat{p}) = 5.085 - .1156 \text{ tempF}$ . A test of hypothesis that the coefficient of `tempF` is zero is the 1 d.f.  $\chi^2$  statistic 6.144 with corresponding  $p$ -value .0132. This  $p$ -value suggests that there is a moderately significant relationship between `temp` and `damage`. Inserting `tempF=31` gives  $\text{logit}(\hat{p}) = 1.5014 \Rightarrow \hat{p} = .8178$ . If the six O-rings on a Space Shuttle fail independently of one another (a roughly true assumption), we could have expected  $6 \times .8178 = 4.9$  failures of the six O-rings for the launch! This analysis could have been performed prior to launch!

The interpretation of logistic regression coefficients is less straightforward than interpretations of linear regression coefficients. In this problem, an increase in launch temperature of  $1^\circ\text{F}$  multiplies the expected odds in favor of O-ring failure by  $e^{-0.1156} = 0.8908$ . Equivalently, each  $1^\circ\text{F}$  decrease in launch temperature corresponds to multiplication of the expected odds of O-ring failure by  $e^{0.1156} = 1.1225$ . In this problem the intercept coefficient 5.085 is not readily interpretable because `tempF = 0` is not a feasible launch temperature.

TABLE 17.1. Logistic regression of Challenger data.  
(logi/code/spaceshuttle.s)

---

```
S-PLUS (logi/transcript/spaceshuttle.glm.st):
> spacshu.bin.glm <- glm(damage ~ tempF, data=spacshu, family=binomial)
>
> anova(spacshu.bin.glm, test="Chi")
Analysis of Deviance Table

Binomial model

Response: damage

Terms added sequentially (first to last)
  Df Deviance Resid. Df Resid. Dev   Pr(Chi)
NULL             137   66.54037
tempF    1 6.144035      136   60.39634 0.01318561
>
> coef(summary(spacshu.bin.glm))
            Value Std. Error   t value
(Intercept) 5.0848265 3.04711489 1.668735
tempF     -0.1155985 0.04691852 -2.463814
>
> spacshu.pred <-
+   predict(spacshu.bin.glm, data.frame(tempF=30:85), type="response",
+           se.fit=T, ci.fit=T, pi.fit=T)
```

---

We must study Table 17.1 together with the `logit.p ~ tempF` plot in Figure 17.5a. The model says that  $\text{logit}.p = \text{logit}(p)$  is linearly related to temperature `tempF`. The slope  $\beta_{\text{tempF}} = -0.1156$  and intercept  $\beta_0 = 5.0848$  describe the straight line in Figure 17.5a.

Table 17.2 compares the three scales used in logistic regression analysis. Figure 17.5 shows the predicted probability of failure on each scale. Logits are hard to interpret as they are not in a scale that we are comfortable thinking about. Two alternate transformations are the odds ratio in the `odds scale` panel, and the probability in the `probability scale` panel.

TABLE 17.2. Three scales used in logistic regression.

|               |  |  |
|---------------|--|--|
| probabilities | $\hat{p}$  | > <code>p.hat &lt;- predict.glm(spacshu.bin.glm, type="response")</code> |
| odds          | $\frac{\hat{p}}{1 - \hat{p}}$  | > <code>odds.hat &lt;- p.hat/(1-p.hat)</code>                            |
| logit         | $\text{logit } \hat{p} = \log\left(\frac{\hat{p}}{1 - \hat{p}}\right)$ | > <code>logit.p.hat &lt;- log(odds.hat)</code>                           |

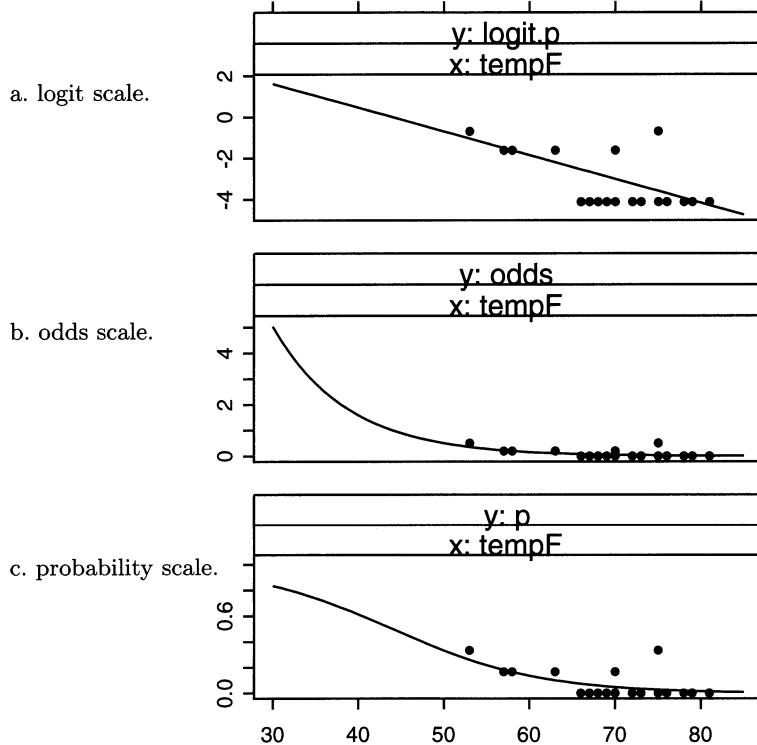


FIGURE 17.5. Observed damage and estimated proportion of damaged O-rings in three scales plotted against temperature. Panel a shows the logit scale, where we see the linear relationship that  $\text{logit}(p)$  is proportional to temperature. Panel b shows the odds scale, where we see the tendency for the odds in favor of damage to go up as temperature goes down. Panel c shows the probability scale, where we see the tendency for the probability of damage to go up as temperature goes down.

(logi/code/spaceshuttle.s), (logi/figure/spaceshuttle.logit-xysplom.eps.gz)

TABLE 17.3. Details of logistic regression of Challenger data.  
(logi/code/spaceshuttle.s)

---

```

S-PLUS (logi/transcript/spaceshuttle.pred.st):
> ## prediction on response scale, in this case (0,1).
> ## leading to Figure logi/figure/spaceshuttle.cde.eps Panel d
> spacshu.pred <-
+   predict(spacshu.bin.glm, data.frame(tempF=30:85), type="response",
+           se.fit=T, ci.fit=T, pi.fit=T)
> cbind(tempF=30:85,
+       round(as.data.frame(spacshu.pred[-(3:4)]), digits=2))[c(1:3,54:56),]
      tempF fit se.fit ci.fit.lower ci.fit.upper pi.fit.lower pi.fit.upper
1     30 0.83  0.23      0.16      0.99     0.10      1.00
2     31 0.82  0.24      0.16      0.99     0.10      0.99
3     32 0.80  0.25      0.15      0.99     0.09      0.99
54    83 0.01  0.01      0.00      0.07     0.00      0.14
55    84 0.01  0.01      0.00      0.06     0.00      0.14
56    85 0.01  0.01      0.00      0.06     0.00      0.13

> ## prediction on link scale, in this case $(-\infty,\infty)
> ## leading to Figure logi/figure/spaceshuttle.pred.logit.f.eps Panel g
> spacshu.pred.link <-
+   predict(spacshu.bin.glm, data.frame(tempF=30:85), type="link",
+           se.fit=T, ci.fit=T, pi.fit=T)
> cbind(tempF=30:85,
+       round(as.data.frame(spacshu.pred.link[-(3:4)]), digits=2))[c(1:3,54:56),]
      tempF fit se.fit ci.fit.lower ci.fit.upper pi.fit.lower pi.fit.upper
1     30 1.62  1.66     -1.66      4.89     -2.21      5.44
2     31 1.50  1.61     -1.69      4.69     -2.25      5.25
3     32 1.39  1.57     -1.71      4.48     -2.29      5.06
54    83 -4.51  0.94     -6.36     -2.66     -7.22     -1.80
55    84 -4.63  0.98     -6.57     -2.69     -7.40     -1.86
56    85 -4.74  1.02     -6.77     -2.71     -7.57     -1.91

```

---

The calculations for the 95% prediction bands based on the linear model for the  $\text{logit}(p)$  are shown in the proportion scale and the logit scale in Table 17.3. The bands are displayed in the proportion scale in Figure 17.3d and in the logit scale in Figure 17.6g.

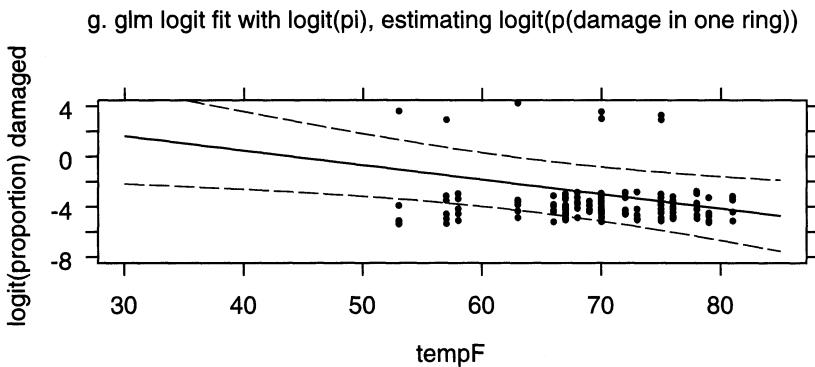


FIGURE 17.6. Panel g. Logit fit on logit scale with 95% prediction band.  
`(logi/code/spaceshuttle.s), (logi/figure/spaceshuttle.pred.logit.f.eps.gz)`

(Lavine, 1991) disagreed with the analysis in (Dalal et al., 1989), stating that the 31°F temperature at Challenger's launch is too far below the lowest previously observed launch temperature, 53°F, to assume that the logistic relationship persisted for the Challenger launch. He claimed that without additional engineering input, one can only say that the probability of O-ring failure at 31°F is less than or equal to the probability of O-ring failure at 53°F. We agree that the extrapolation from 53°F to 31°F is too great for anyone to be highly confident in our  $\hat{p} = .8178$  estimate of the failure probability at 31°F.

## 17.2 Estimation

The computations for calculating the parameter estimates  $\hat{\beta}$  are usually done by the method of maximum likelihood with an iterative computer program. The likelihood is defined algebraically as the joint probability of the observations viewed as a function of the parameters conditional on the data.

Let us use the notation that the  $i^{\text{th}}$  case, for  $i: 1, \dots, n$ , consists of a single response value  $y_i$  and a single predictor variable  $x_i$ . The likelihood with a single predictor variable is written as

$$L(\beta|y; x) = L\left((\beta_0, \beta_1) \mid (y_1, \dots, y_n); (x_1, \dots, x_n)\right)$$

$$\begin{aligned}
 &= \prod_{i=1}^n f(y_i \mid x_i; \beta_0, \beta_1) \\
 &= \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}
 \end{aligned} \tag{17.6}$$

where

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = \text{antilogit}(\beta_0 + \beta_1 x_i) \tag{17.7}$$

$$(17.8)$$

or equivalently,

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i \tag{17.9}$$

The logistic link function appears in the likelihood equation with the expression of  $p$  as the antilogit of the linear function of the predictor  $x$ -variables.

In general, the terminology *link* function refers to the transformation of the response variable such that the transformed response achieves an approximate linear relationship with explanatory variables. After selecting a link function, it is also necessary to specify a variance function, usually by accepting the software's default choice. A binomial variance is the usual choice with logistic or probit links. A Poisson variance is associated with a logarithmic link function. See (McCullagh and Nelder, 1983) and the software help files for more detail.

The binomial error function appears in the likelihood expression (17.6) as the product of Bernoulli (binomial with  $n = 1$ ) terms, one for each of the  $n$  observations. The binomial distribution is thus involved in the likelihood equations for estimating the  $\beta$  coefficients.

The log of the likelihood, called the *loglikelihood*, is written as

$$\begin{aligned}
 \ell(\beta|y; x) &= \log(L(\beta|y; x)) \\
 &= \sum_{i=1}^n \left( y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right)
 \end{aligned} \tag{17.10}$$

We estimate the parameters  $\beta$  by differentiating the log of the likelihood with respect to the vector of parameters  $\beta$ , setting the derivatives to 0, and solving for  $\beta$ . We can rewrite these equations in terms of the observed probabilities  $p_i$  and fitted values  $\hat{p}_i$  or in terms of their complements  $q_i =$

$1 - p_i$  and  $\hat{q}_i = 1 - \hat{p}_i$ :

$$\begin{aligned}\frac{\partial}{\partial \beta_0} \ell(\beta|y; X) &= \sum_{i=1}^n (p_i - \hat{p}_i) &= \sum_{i=1}^n (q_i - \hat{q}_i) &= 0 \\ \frac{\partial}{\partial \beta_1} \ell(\beta|y; X) &= \sum_{i=1}^n x_i(p_i - \hat{p}_i) &= \sum_{i=1}^n x_i(q_i - \hat{q}_i) &= 0\end{aligned}\tag{17.11}$$

Exercise 17.1 gives you an opportunity to follow the steps in detail.

Logistic regression is a special case of a *generalized linear model*, hence the model specification is easily interpreted.

**glm:** The statement in Table 17.1

```
S-PLUS (logi/code/spac.glm.s):
spacshu.bin.glm <- glm(damage ~ tempF, data=spacshu,
family=binomial)
```

says that the logit of the expected value of the response variable `damage` is to be modeled by a linear function of the predictor variable `tempF`, that is,

$$\text{logit}(E(\hat{p}_j)) = \log \left( \frac{E(\hat{p}_j)}{1 - E(\hat{p}_j)} \right) = \beta_0 + \beta_x x$$

The `family=binomial` argument says that the error term has a binomial distribution and that the link function is the logit. The logit is the default link for the binomial family; we could have specified the link explicitly with

```
S-PLUS (logi/code/spac.glm.s):
spacshu.bin.glm <- glm(damage ~ tempF, data=spacshu,
family=binomial(link=logit))
```

S-PLUS places the results of the fitting process into an object called `spacshu.bin.glm` that we can look at.

Two principal functions are used to display textual information about the object.

**anova:** The `anova` function

```
S-PLUS (logi/code/spac.glm.s):
anova(spacshu.bin.glm, test="Chi")
```

displays the *analysis of deviance table*, an analogue to the *analysis of*

*variance table* in ordinary linear models. Table 17.1 is consistent with what we see in Figure 17.5. There is a clearly visible effect in the figure. The ANOVA table shows the linear dependence on `tempF` has  $p$ -value 0.013, highly significant at  $\alpha = .05$ . The table is interpreted similarly to an ordinary analysis of variance table. The *deviance* for `tempF` is twice the drop in the loglikelihood from the model without the `tempF` term to the model with the `tempF` term,  $6.144035 = 66.54037 - 60.39634$ , with degrees of freedom the difference in the number of parameters in the two models,  $1 = 137 - 136$ . We use the  $\chi^2$  table, hence  $1 - \mathcal{F}_{\chi^2}(6.144035 | 1) = 0.01318561$ .

Deviance is very general concept. The formula for deviance specializes to variance for the normal error function with the identity link function. In this, and in many other ways, least squares is a special case of maximum likelihood.

**summary:** We also show the table of regression coefficients taken from the `summary` function.

---

S-PLUS (`logi/code/spac.glmc.s`):  
`anova(spacshu.bin.glm, test="Chi")`

---

It is less helpful than the deviance table with only a single predictor variable. In this example it says that the  $t$ -value for `tempF` is  $-0.1155985$ . This number does not have an exact  $t$ -distribution, which is why no  $p$ -value is associated with it. Since  $| -2.46 | > 2$ , it appears to be significant, but we must look at the chi-square value in the analysis of deviance table to make a valid inference.

### 17.3 Example—Budworm Data

An experiment discussed in (Collett, 1991) and (Venables and Ripley, 1997) was performed to model the dose response of a particular insecticide required to kill or incapacitate budworms, insects that attack tobacco plants. A potential additional factor was the budworm's sex. In particular, it was desired to estimate the lethal dose proportions LD25, LD50, LD75. LD50, lethal dose 50, is an abbreviation for the dose that is lethal for 50% of the budworms. Twenty moths of each sex were exposed to each of seven experimental doses. The data, in file (`datasets/budworm.dat`), present `ldose`, the  $\log_2$  of dose, along with the number of moths of each sex that were killed after three days of exposure.

Figure 17.7 shows the relationship of predicted probability of kill, odds, and logit to the observed data. The `logit ~ log.dose` panel shows the

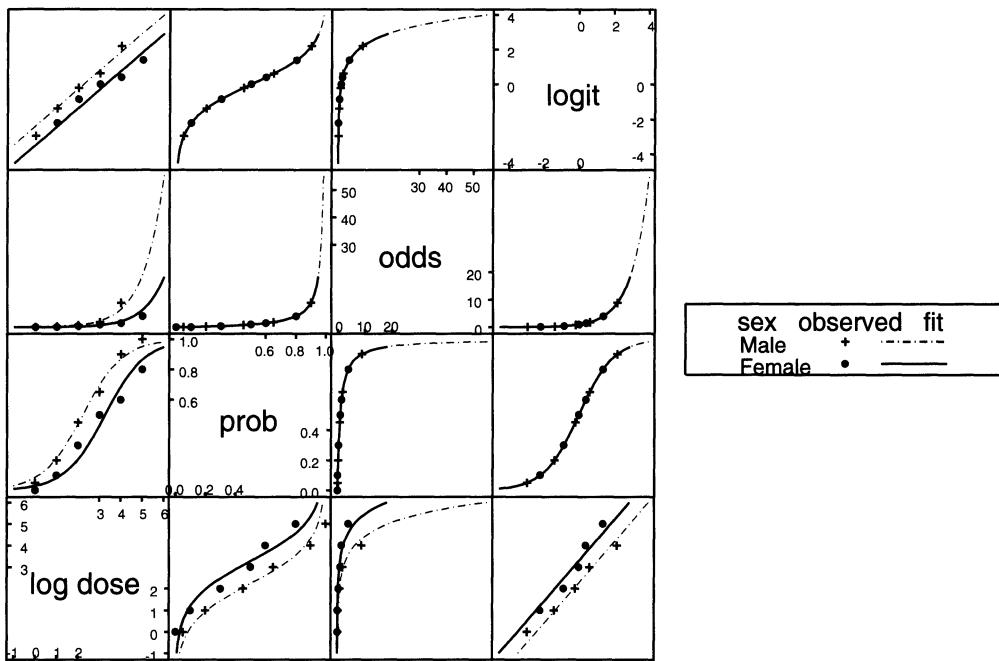


FIGURE 17.7. Predicted probabilities of kill, odds, and  $\log(\text{odds})$  as a function of  $\log(\text{dose})$ . The upper-right  $3 \times 3$  submatrix shows the functional relation among these three scales. A constant difference in the  $\log(\text{odds})$ , seen as parallel lines in the  $\text{logit} \sim \log(\text{dose})$  panel, corresponds to a constant ratio in the odds.

(logi/code/budworm.s),

(logi/figure/budworm-7.eps.gz), (logi/figure/budworm-7-color.eps.gz)

assumed linear relationship of  $\text{logit}(p) = \log(p/(1-p))$ . The upper  $3 \times 3$  panels show mathematics and therefore do not display a sex difference.

For each dose of insecticide, male budworms have higher mortality than females. We read the predicted LD<sub>50</sub> (on  $\log_2$  scale) by placing a horizontal line at  $\text{prob}=.5$  on the  $\text{prob} \sim \log(\text{dose})$  panel of Figure 17.7. We look at the  $\log(\text{dose})$  coordinate of the intersection of the horizontal line with the fitted lines and find that the LD<sub>50</sub> is slightly over 2 for males and slightly over 3 for females. Precise LD<sub>50</sub> predictions as well as ones for LD<sub>25</sub> and LD<sub>75</sub> are available in the file (logi/transcript/budworm.st) in the online files.

## 17.4 Example—Lymph Nodes

These data come from (Brown, 1980). The problem is outlined in this reference as follows:

When a patient is diagnosed as having cancer of the prostate, an important question in deciding on treatment strategy for the patient is whether or not the cancer has spread to the neighboring lymph nodes. The question is so critical in prognosis and treatment that it is customary to operate on the patient for the sole purpose of examining the nodes and removing tissue samples to examine under the microscope for evidence of cancer. However, certain variables that can be measured without surgery are predictive of the nodal involvement; and the purpose of the study presented here was to examine the data for 53 prostate cancer patients receiving surgery, to determine which of five preoperative variables are predictive of nodal involvement, and how accurately the prediction can be made.

In particular, the medical investigator was interested in whether or not an elevated level of acid phosphatase in the blood serum would be of added value in the prediction of whether or not the lymph nodes were affected, given the other four more generally used variables.

### 17.4.1 Data

The variables in data file (`datasets/lymph.dat`) are described in Table 17.4. They are read into an S-PLUS data frame `lymph` with the code (`logi/code/lymph.s`) or read with (`logi/code/logi.c1.sas`) into a SAS dataset. The data are displayed in Figure 17.8.

TABLE 17.4. Lymph Data from Brown (Brown, 1980)

---

Response Variable

`nodes` 1 indicates nodal involvement found at surgery,  
0 no nodal involvement

Predictor Variables

|                      |   |
|----------------------|---|
| <code>X.ray</code>   | 1 is serious, 0 is less serious   |
| <code>stage</code>   | measure of size and location of tumor,<br>1 is serious, 0 is less serious |
| <code>grade</code>   | pathology reading of a biopsy, 1 is serious, 0 is less serious            |
| <code>age</code>     | at diagnosis  |
| <code>acid.ph</code> | level of serum acid phosphatase   |

---

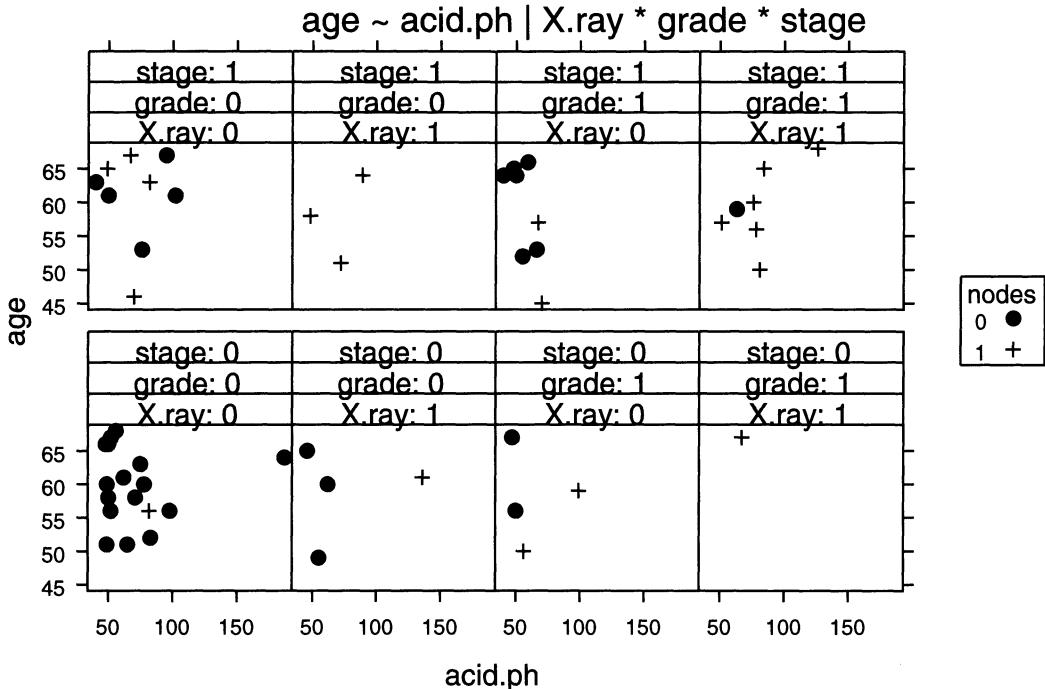


FIGURE 17.8.  $\text{age} \sim \text{acid.ph} | \text{X.ray} * \text{grade} * \text{stage}$   
 $(\text{logi}/\text{code}/\text{lymph.s}), (\text{logi}/\text{figure}/\text{a2.eps.gz})$

In terms of this picture, the investigator wants to know if the predicted proportion of “+” in each panel (that is, conditional on grade, stage, X-ray, and age) is affected by the additional information on acid phosphatase.

#### 17.4.2 Data Analysis

Figure 17.8 is constructed as a plot of the two continuous  $x$ -variables conditioned on the three 2-level factors and with the plotting symbol chosen to reflect the 2-level response variable. Our first impression from Figure 17.8 is that age seems not to make a difference. Therefore, our subsequent analysis considers only the three remaining factors and one continuous predictor.

In Table 17.5 we model the logit transformation of **nodes** as a function of **X.ray**, **stage**, **grade**, and **acid.ph**. Figure 17.9 contains plots of jittered **nodes** vs **acid.ph** along with the model's predicted probability of nodal involvement partitioned according to the  $8 = 2^3$  combinations of the factors **X.ray**, **stage**, and **grade**. The predicted relationship is plausible

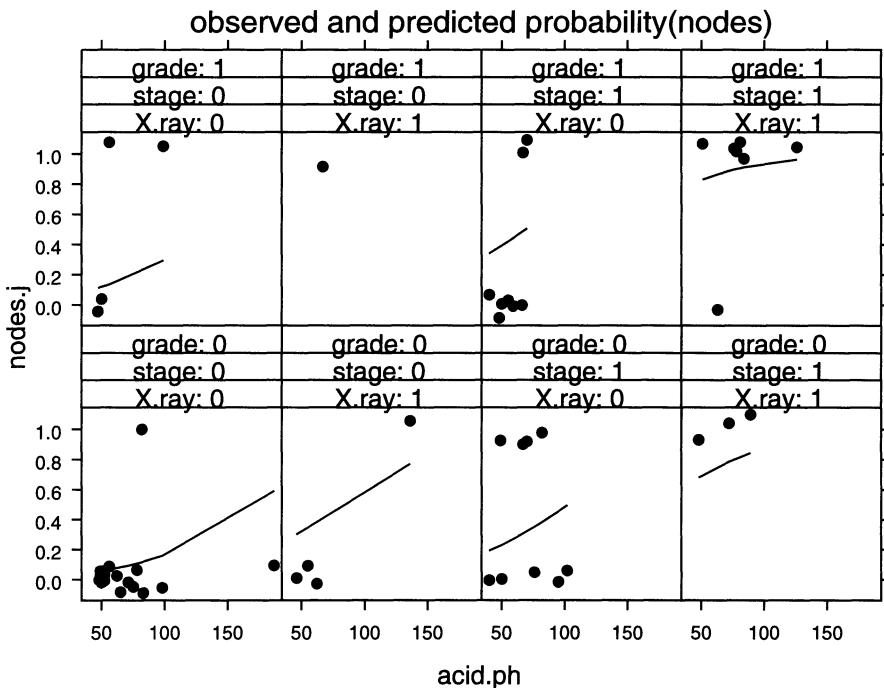


FIGURE 17.9. Jittered and predicted `nodes` vs `acid.ph` for each combination of values of `stage`, `grade`, and `X.ray`  
`(logi/code/lymph2.s)`, `(logi/figure/p8.eps.gz)`

in 5 of these 8 panels, where there is a visible tendency for the `nodes=1` points to be farther to the right, that is to have higher `acid.ph` scores, than the `nodes=0` cases. Two of the panels have only `nodes=1` cases and can't be used for this comparison. In only one panel (`X.ray=0`, `grade=0`, `stage=1`) is there a balanced overlap of ranges. This suggests that `acid.ph` is positively associated with `nodes` and therefore may be a useful additional predictor.

The *p*-value for `acid.ph` in Table 17.5, 0.075, indicates that `acid.ph` is a borderline predictor of the presence of `nodes`. A tentative interpretation is that `acid.ph` is a potential predictor for some but not all combinations of `X.ray`, `stage`, and `grade`. In this same table, the *p*-value for `grade` is 0.45, suggesting that this factor can be dropped from the model. However, we choose to retain `stage` in because a rerun of the model without `stage` actually increases the *p*-value of `acid.ph`. Investigation of this issue is requested in Exercise 17.10.

TABLE 17.5. Logistic regression with three factors and one continuous predictor. See also Figure 17.9.  
(logi/code/lymph.s)

```
S-PLUS (logi/transcript/logit-j.st):
> lymph3.glm <- glm(nodes ~ X.ray + stage + grade + acid.ph,
+                      data=lymph, family=binomial)
> anova(lymph3.glm, test="Chisq")

Analysis of Deviance Table

Binomial model

Response: nodes

Terms added sequentially (first to last)
  Df Deviance Resid. Df Resid. Dev   Pr(Chi)
NULL              52  70.25215
X.ray      1 11.25135    51  59.00080 0.0007957
stage       1  5.64738    50  53.35342 0.0174814
grade       1  0.57341    49  52.78001 0.4489072
acid.ph     1  3.16455    48  49.61546 0.0752533

> summary(lymph3.glm)$coef

          Value Std. Error   t value
(Intercept) -1.7339495 1.00180899 -1.730818
X.ray        1.0060907 0.39664532  2.536500
stage        0.7750349 0.38367115  2.020050
grade        0.3861685 0.37719069  1.023802
acid.ph      0.0228317 0.01298434  1.758403
```

Figures 17.10 and 17.11 depict the jittered `nodes`, `acid.ph`, `X.ray`, the predicted probability of nodal involvement, and the transformation of these predicted probabilities to predicted odds and predicted logits for the additive model in Table 17.5. Figure 17.10 uses different plot symbols for the two levels of `X.ray` and Figure 17.11 provides separate sploms for the two levels of `X.ray`. In the row and column for `acid.ph`, separate predictions are provided for each of the  $8 = 2^3$  combinations of the three factors `X.ray`, `stage`, and `grade`. In the `logit.p.hat ~ acid.ph` panels there are 8 parallel lines, one for each value of `acid.ph*X.ray*nodes`. The `p.hat ~ acid.ph` panels in these figures suggest that there is a slight positive association between nodal involvement and `acid.ph`.

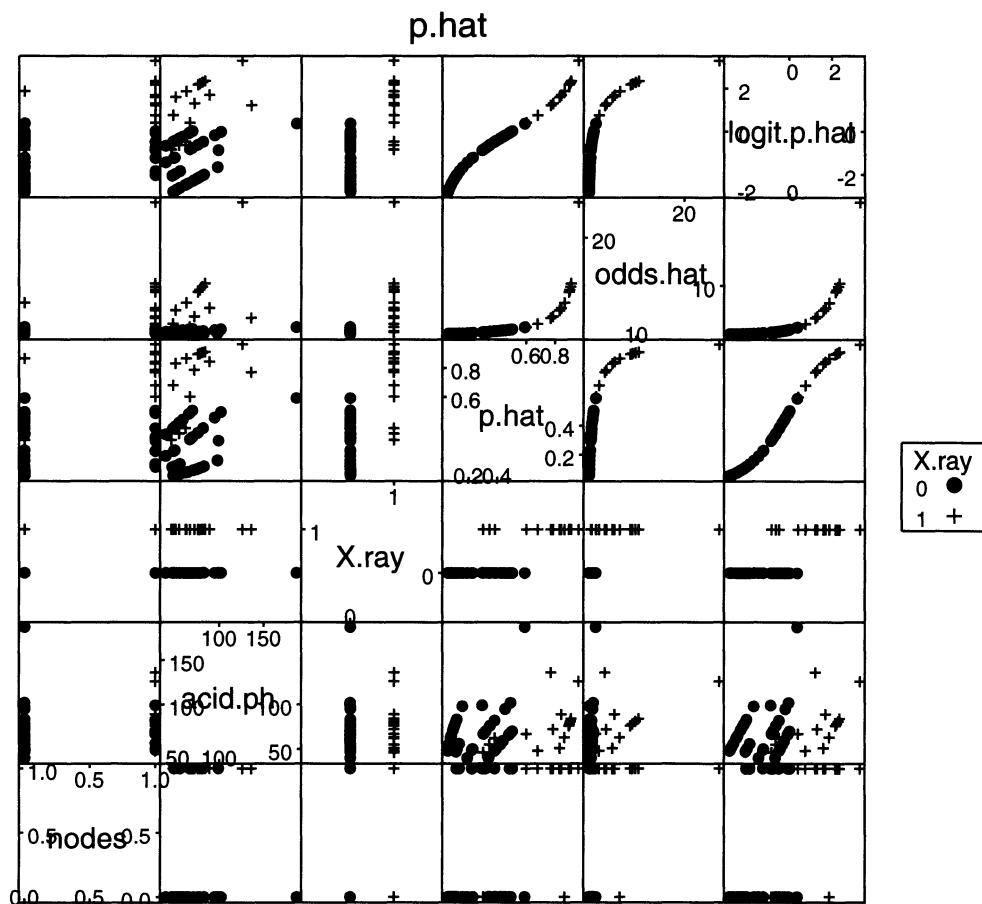


FIGURE 17.10. Scatterplot matrix of `nodes ~ acid.ph | stage + grade + X.ray`  
`(logi/code/lymph2.s), (logi/figure/m.eps.gz)`

### 17.4.3 Additional Techniques

In the remainder of this section we consider a simpler logistic regression model for `nodes` primarily to further illustrate the variety of modeling and plotting techniques available to analysts. We model the probability of nodal involvement as a function of just `acid.ph` and `X.ray`, allowing for the possibility that the slope in logistic relationship between `nodes` and `acid.ph` differs according to whether `X.ray` = 0 or 1.

Figure 17.12 shows `nodes` as a function of `acid.ph` separately for each value of `X.ray`. Except for the single point with `acid.ph=187`, to be investigated

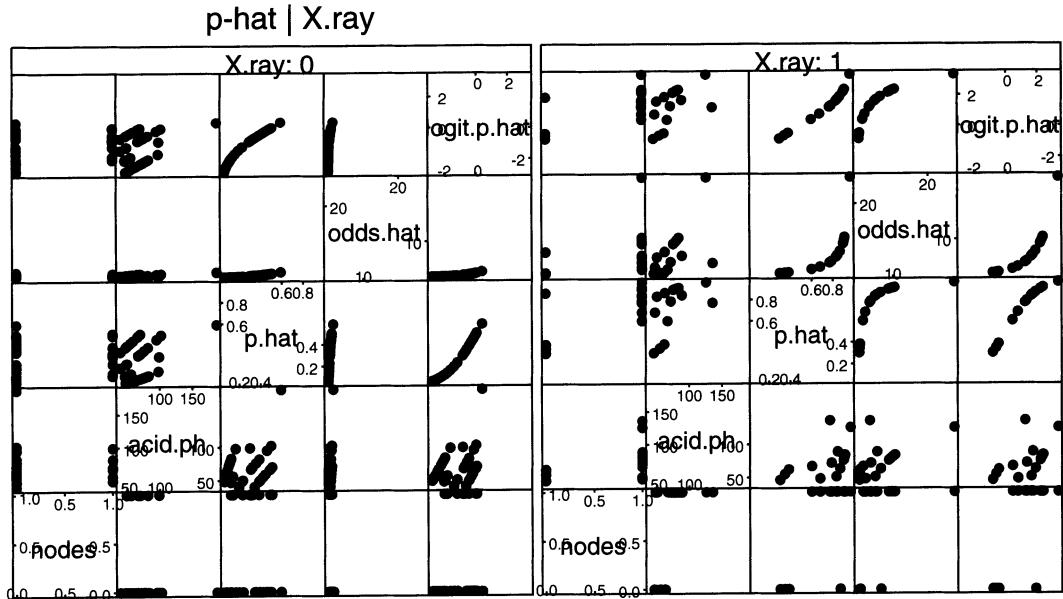


FIGURE 17.11. Scatterplot matrices of `nodes ~ acid.ph | stage + grade + X.ray` conditioned on the value of `X.ray`  
`(logi/code/lymph2.s)`, `(logi/figure/n.eps.gz)`

later, the illustration shows clearly that high `acid.ph` predicts `nodes=1` for `X.ray=1`, and that high `acid.ph` doesn't help very much for `X.ray=0`.

We can get a sense of this interpretation by partitioning the  $x$ -axis (`acid.ph`) into sections 30 units wide and plotting the proportion of 1's in each section. We do so in Figure 17.13. The line segments plotted in each  $x$ -section mostly represent increasing proportions of "1"s as `acid.ph` goes up. The logistic regression technique uses a continuous  $x$ -axis, thus no segmentation, and it forces the fit to be monotone increasing.

We model Figure 17.12 by fitting a linear model with  $x=\text{acid.ph}$  to the logit of the response variable

$$\text{logit}(E(\hat{p}_j)) = \log \left( \frac{E(\hat{p}_j)}{1 - E(\hat{p}_j)} \right) = \beta_0 + \beta_j + \beta_x x$$

with the error distribution assumed to be binomial with

$$p_j = P(\text{nodes} = 1 | X.\text{ray} = j) \quad (17.12)$$

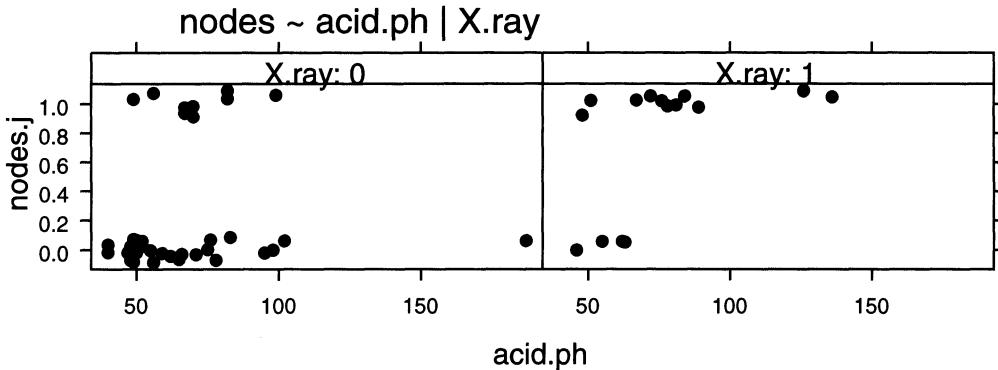


FIGURE 17.12.  $\text{nodes} \sim \text{acid.ph} | \text{X.ray}$   
`(logi/code/lymph.s), (logi/figure/c.eps.gz)`

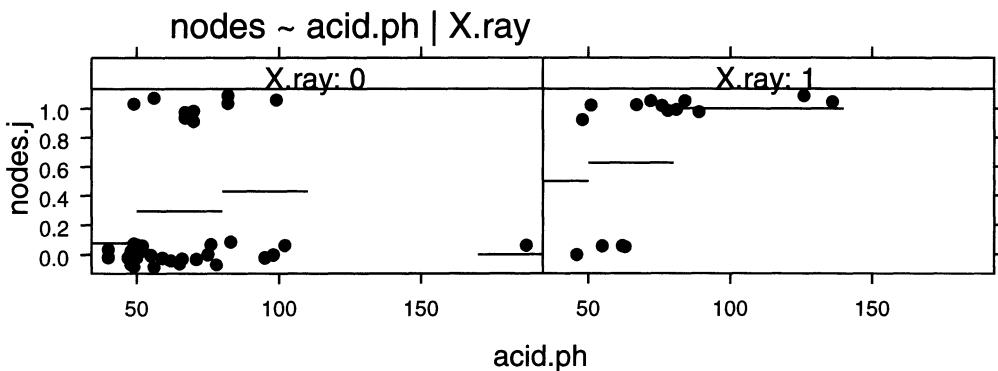


FIGURE 17.13.  $\text{nodes} \sim \text{acid.ph} | \text{X.ray}$  with sectioned proportions.  
`(logi/code/lymph.s), (logi/figure/cg.eps.gz)`

The model in Equation (17.12) is structurally similar to an ANCOVA model, with common slope  $\beta_x$  and with different intercepts.  $\beta_0$  is the reference intercept and  $\beta_j$  is the offset of the intercept for group  $j$ . We see the appropriateness of this model in the `logit.p.hat ~ acid.ph` panel of Figure 17.14 and, even more strikingly, in Figure 17.10.

The results of the model are displayed in Figures 17.14 (all together) and 17.15 (conditioned on `X.ray`). The `nodes ~ acid.ph | x.ray` subpanels of Figures 17.14 and 17.15 repeat the contents of Figure 17.12. The top three rows of the plots show three different transformations of the predicted  $\hat{p}_j$ .

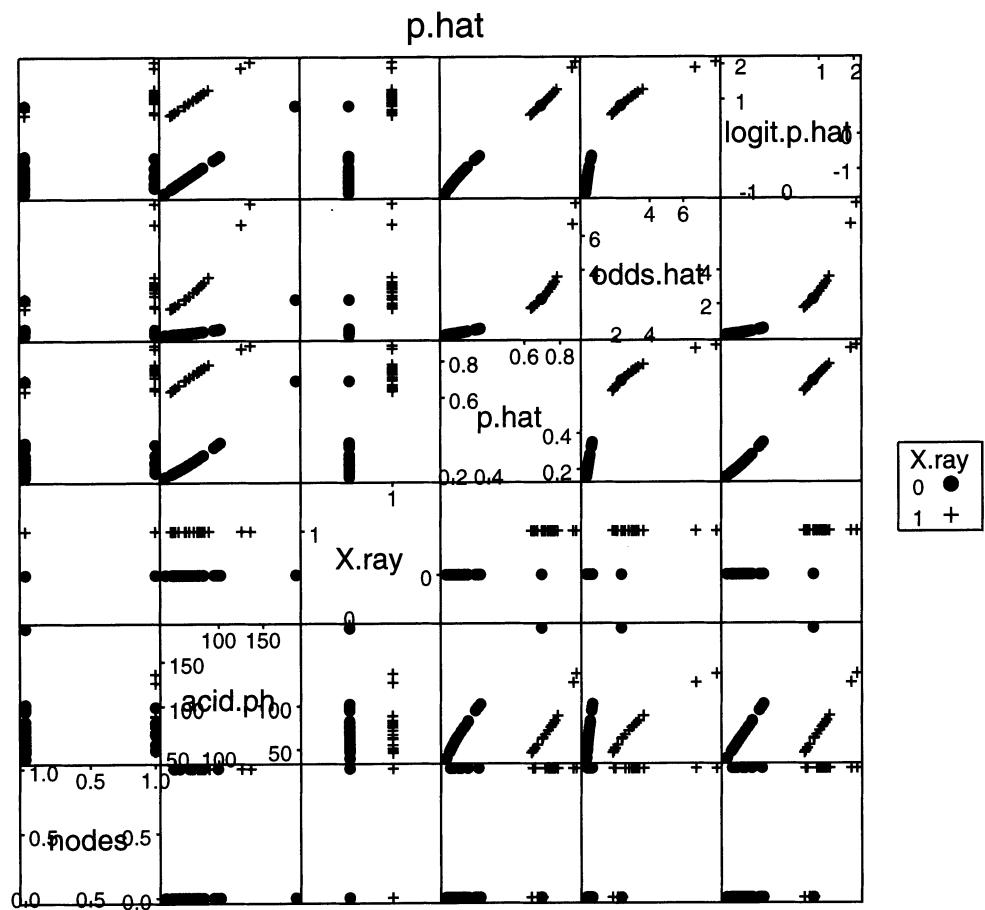


FIGURE 17.14. splom of fit to  $nodes \sim acid.ph | X.ray$  with the responses at the different values of  $X.ray$  superposed on the same plot.  
 (logi/code/lymph.s), (logi/figure/d.eps.gz)

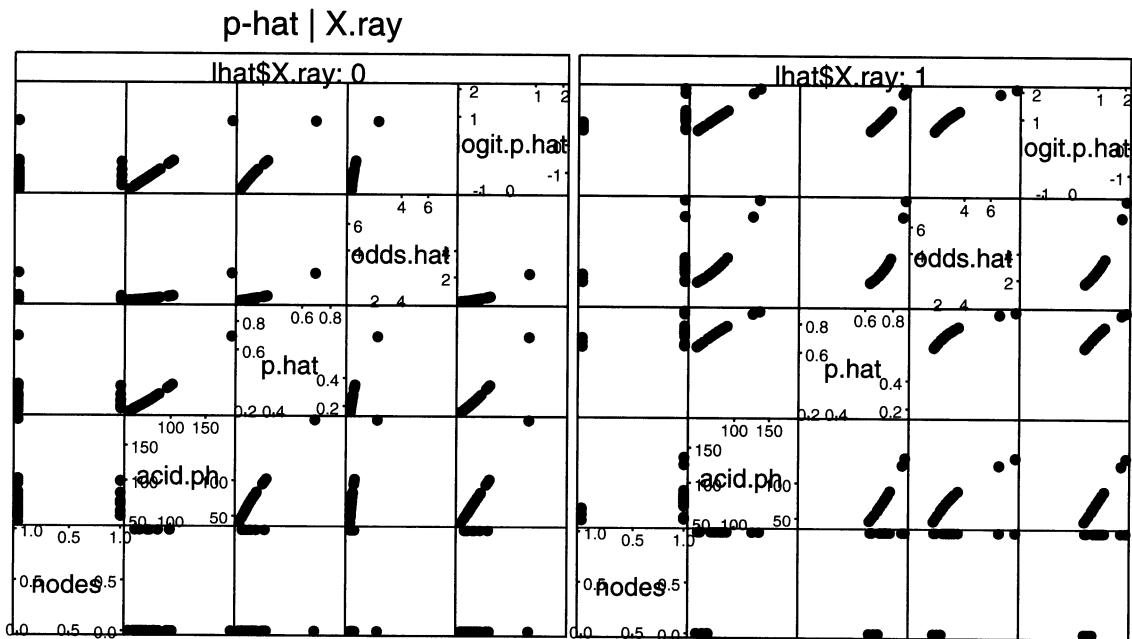


FIGURE 17.15. Splom of fit to  $\text{nodes} \sim \text{acid.ph} | \text{X.ray}$  with the responses at the different values of  $\text{X.ray}$  displayed in adjacent panels. The  $\text{nodes} \sim \text{acid.ph}$  subpanels of these plots are repeated in Figure 17.12 and modeled in Figures 17.13 and 17.16.  
 $(\text{logi}/\text{code}/\text{lymph.s}), (\text{logi}/\text{figure}/\text{e.eps.gz})$

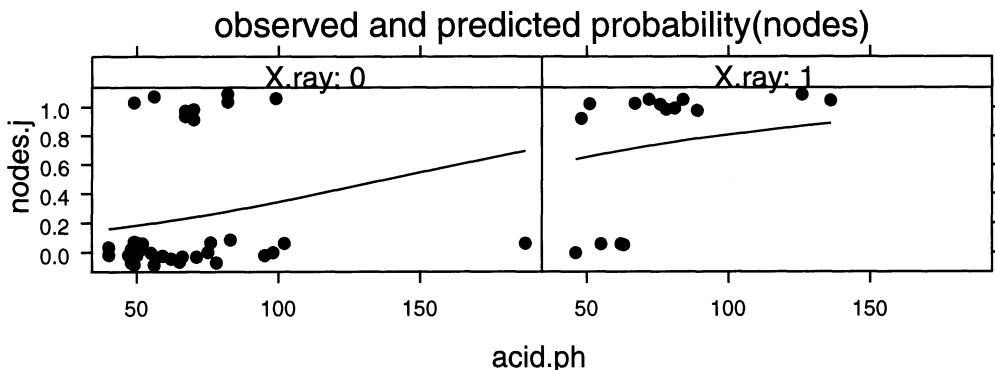


FIGURE 17.16. Observed and predicted probability(nodes) for model `nodes ~ acid.ph | X.ray`. The points here are the jittered points from the `nodes ~ acid.ph` panels of Figure 17.15.  
`(logi/code/lymph.s)`, `(logi/figure/f.eps.gz)`

Figure 17.16 displays partitioned fits by value of `X.ray`, suggesting that forcing a common slope (in the logit scale) might not be the best model for these data. Let us try nesting the slope within the level of `X.ray`. We see in Table 17.6 that the `acid.ph %in% X.ray` has gone from a *p*-value of .16 in the first analysis `lymph1a.glm` to one of .065 in the second analysis `lymph1b.glm`. From the coefficients we see that most of the significance is in the `X.ray=1` panel.

Figure 17.16 displays the overlay of the  
`(nodes ~ acid.ph | x.ray)` and `(p.hat ~ acid.ph | x.ray)`  
subpanels of Figures 17.15. (Figure 17.16 actually displays the entire fitted line for the `p.hat` panels, not just the specific points corresponding to the observed data.) Compare this to Figure 17.12. Note that the farther right we move in each panel, the higher the proportion of dots that appear in `nodes=1` and therefore the higher the predicted probability that a case will have `nodes=1`.

Each panel of Figure 17.16 is constructed as the superposition of the subpanels for observed data (`nodes ~ acid.ph | X.ray`) and for predicted probability (`p.hat ~ acid.ph | X.ray`) from the corresponding panel of Figure 17.15. We see that the probability of nodal involvement as a function of `acid.ph` depends on the value of `X.ray`.

When we trace Figure 17.16 backward to Figure 17.14, we see that the `logit.p.hat ~ acid.ph` panel shows two distinct parallel lines. These parallel lines correspond to the predicted probabilities for the two levels of `X.ray`. Remembering that  $\text{logit}(\hat{p}) = \log(\text{odds}(\hat{p}))$ , we see that the parallel lines have common slope (0.01685342 `logit.p.hat` units per `acid.ph`) and different intercepts 2.116 (=2×1.05796265) units apart.

TABLE 17.6. Logistic regression with slopes nested within X.ray. See also Figure 17.16.  
(logi/code/lymph.s)

---

```

S-PLUS (logi/transcript/logit-k.st):
> lymph1a.glm <- glm(nodes ~ X.ray + acid.ph, data=lymph, family=binomial)
> anova(lymph1a.glm, test="Chisq")
Analysis of Deviance Table

Binomial model

Response: nodes

Terms added sequentially (first to last)
  Df Deviance Resid. Df Resid. Dev  Pr(Chi)
NULL                   52   70.25215
X.ray    1 11.25135      51   59.00080 0.0007957
acid.ph  1  1.94210      50   57.05871 0.1634410
> summary(lymph1a.glm)$coef
            Value Std. Error   t value
(Intercept) -1.27760476 0.95349929 -1.339912
          X.ray  1.05796265 0.35395386  2.988985
        acid.ph 0.01685342 0.01257315  1.340430

> lymph1b.glm <- glm(nodes ~ X.ray/acid.ph, data=lymph,
+ family=binomial)
> anova(lymph1b.glm, test="Chisq")

Analysis of Deviance Table

Binomial model

Response: nodes

Terms added sequentially (first to last)
  Df Deviance Resid. Df Resid. Dev Pr(Chi)
NULL                   52   70.25215
          X.ray  1 11.25135      51   59.00080 0.00080
acid.ph %in% X.ray  2  5.46609      49   53.53471 0.06502
> summary(lymph1b.glm)$coef
            Value Std. Error   t value
(Intercept) -3.681491430 1.95589795 -1.8822513
          X.ray -1.987642914 1.95589795 -1.0162304
X.ray0acid.ph  0.007677856 0.01350193  0.5686488
X.ray1acid.ph  0.101628027 0.06076561  1.6724596

```

---

Translating the difference in intercepts back to odds ratios, we find the odds ratio attributable to `X.ray` is  $e^{2.116} = 8.297879$ , and we can see from the (`odds.hat ~ acid.ph`) panel of Figure 17.14 that the `X.ray=1` line is about 8 times higher than the `X.ray=0` line.

#### 17.4.4 Diagnostics

All the usual diagnostic calculations, statistics, and graphs developed for linear regression are used with logit regression to help determine goodness of fit. Details are discussed in (Hosmer and Lemeshow, 2000).

The most important diagnostic in this dataset is the graph in Figure 17.16 where we note the outlier for `acid.ph=187`. Without that single point, a new fit in the `X.ray=0` panel would be almost horizontal.

### 17.5 Numerical Printout

We calculate the fitted values in three scales in Table 17.2 and plot them. The fit in `lymph1a.glm` is illustrated in Figures 17.14 and 17.16. The `logit.p.hat ~ acid.ph` panel in Figure 17.14 shows the straight-line fit of  $\text{logit } \hat{p}$  to `acid.ph`. There are two parallel lines in the panel, one for each value of the factor `X.ray`. The slope is given by the coefficient  $\beta_{\text{acid.ph}}$ . The difference in intercepts is given by twice the coefficient  $\beta_{X.\text{ray}}$  (note the dependence on the  $-1, 1$  dummy variable coding scheme). The `p.hat ~ acid.ph` panel is the same thing, transformed to the probability scale. Figure 17.15 is the same as Figure 17.14 except that it is now conditioned on `X.ray`. Figure 17.16 overlays the `logit.p.hat ~ acid.ph` panels on the `nodes ~ acid.ph` panels of Figure 17.15.

The more complete model with all three factors in Figure 17.9 is calculated in Table 17.5.

## 17.6 Graphics

The data for this example are displayed with several types of graphs. Most of the graphs are constructed with multiple panels to display several variables in a coordinated fashion.

#### 17.6.1 Conditioned Scatterplots

The initial display in Figure 17.8 is a scatterplot of the two continuous variables `age` on the  $y$ -axis and `acid.ph` on the  $x$ -axis, conditioned on the

three binary-valued  $x$ -factors and using the binary-valued `nodes`-variable as the plotting character. All eight panels of the display are scaled alike in order to help the eye distinguish the important features. Common scaling is critical for making comparisons.

Figure 17.9 is similarly constructed, although in this case with `age` suppressed and with `nodes` on the  $y$ -axis. This graph presents the same view of the data, with important differences. We have jittered the `nodes` variable to counter the overprinting. Also, in Figure 17.9, we have overlaid the observed data with the predicted probabilities from the logistic model. By doing so we get an immediate sense of how the predictions are affected by the values of the conditioning factors and can see directly in each panel that higher `acid.ph` values correspond to higher predicted probabilities and to a higher proportion of points with observed response `nodes=1`.

The displays of the simplified model in Figures 17.12 and 17.16 are constructed similarly to Figure 17.9. In this simpler model there are only two panels rather than eight.

### 17.6.2 Scatterplot Matrix

Figures 17.14 and 17.15 are **Scatterplot Matrices** (*splom*). The splom is one of the most important plots and is often used as the initial view of the dataset. We did not begin with the splom here because most of the variables are 2-level factors and the panels would therefore consist of noninformative  $2 \times 2$  lattices. With refinements the splom is used for later views of the raw data and (as is done here) for functions of the predicted values. Figures 17.14 and 17.15 are sploms. Figure 17.14 does not condition on the value of the factor `X.ray`. Figure 17.15 does condition on `X.ray`.

### 17.6.3 Common Scaling in Comparable Plots

Let us illustrate the importance of common scaling by looking in Figure 17.17 at the predictions of Figure 17.9 in the logit scale both with common scaling and with idiosyncratic scaling. In the logit scale all eight lines have identical slope. With common scaling we see the parallelism. We also see the outlier at `X.ray=0` and `acid.ph=187`. With idiosyncratic scaling we see two unrelated parallel structures, and worse, we think the distant points at `X.ray=1` and `acid.ph≈130` are as far out as the real outlier at `X.ray=0` and `acid.ph=187`. Only when the viewer looks very closely at the printed scale, a requirement that indicates a poorly designed graphic, is it possible to see the major message in this idiosyncratically scaled pair of graphs.

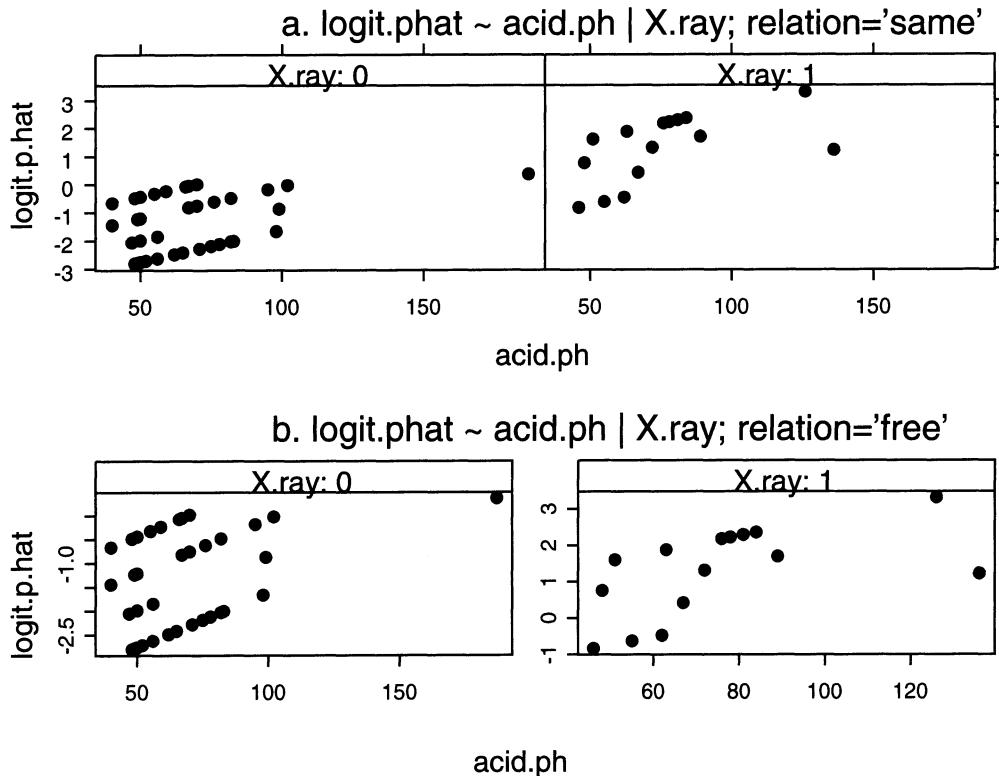


FIGURE 17.17.  $\text{logit}(\hat{p}) \sim \text{acid.ph} | \text{X.ray}$  with common scale (a) and separate scales (b). With a common scale we see that the  $\text{X.ray}=0$  observations are mostly to the left of and below the  $\text{X.ray}=1$  observations. With separate scales, each set of points is individually centered in its plotting window and the ability to compare them visually is lost.

(logi/code/lymph.s), (logi/figure/k.eps.gz)

#### 17.6.4 Functions of Predicted Values

Figures 17.14, 17.15, and 17.16 show the predicted values for the simpler of the two logistic regression models studied so far. The observed  $x=\text{acid.ph}$  and  $y=\text{nodes}$  are the same in all three graphs. Figure 17.14 (without conditioning on  $\text{X.ray}$ ) superposes the two entire scatterplot matrices in each panel of Figure 17.15 onto a single scatterplot matrix. We are primarily interested here in the  $\text{nodes} \sim \text{acid.ph}$  panel. Figure 17.16 shows the predicted probability of success subpanels  $\text{p.hat} \sim \text{acid.ph}$  superposed on the observed success or failure subpanels  $\text{nodes} \sim \text{acid.ph}$  from each panel of Figure 17.15.

Focus on Figure 17.16. The variable  $\hat{p} = p.\hat{}$  is in the same scale (0 to 1) as the observed data. In this scale, the S-shaped logistic curve is apparent.

Now look again at Figure 17.15. The odds transformation  $\frac{\hat{p}}{1-\hat{p}} = \text{odds}.\hat{}$  shows that the predicted odds for the two groups have a constant ratio. Constant ratios are difficult to see in a graph. The third transformation, the logit transformation  $\text{logit}(\hat{p}) = \log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \text{logit}.\hat{}$ , most clearly shows the model assumptions: The two lines for the two groups are parallel. A constant difference in the  $\text{log}(\text{odds})$  corresponds to a constant ratio in the odds.

## 17.7 Model Specification

Logistic regression is a special case of generalized linear modeling, hence the programming constructs developed for ordinary linear models can be used.

SAS and S-PLUS use different default factor coding schemes. The fitted values are the same from both programs, but the regression coefficients and their standard deviations frequently differ by a factor of 2.

### 17.7.1 S-PLUS

The basic statements for general linear models are

---

S-PLUS (logi/code/logit-da.s):

```
lymph1.glm <- glm(nodes ~ acid.ph + X.ray,
                     data=lymph, family=binomial)
anova(lymph1.glm, test="Chisq")
summary(lymph1.glm)$coef
```

---

and

---

S-PLUS (logi/code/logit-ga.s):

```
lymph2.glm <- glm(nodes ~ X.ray + stage + grade + age + acid.ph,
                     data=lymph, family=binomial)
anova(lymph2.glm, test="Chisq")
summary(lymph2.glm)$coef
```

---

The user needs to use ordinary data management statements to arrange for the appropriate subjects and variables to appear in the data.frames. Graphical display of the results also requires data management statements to select columns of the data and to construct the transformations of the predicted values.

### 17.7.2 SAS

The basic statements for logistic models are illustrated for the model in Figure 17.9 and Table 17.5 in

---

SAS (logi/code/logi.lymph3.sas):

```
proc logistic data=lymph descending;
    title "lymph 3" ;
    model Y = X_ray stage grade acid_ph ;
        output out=probs3 predicted=prob xbeta=logit;
run;
```

---

The user needs to use ordinary DATA step data management statements to arrange for the appropriate subjects and variables to appear in the datasets. Graphical display of the results also requires data management statements to select columns of the data and to construct the transformations of the predicted values. SAS PROC LOGISTIC doesn't accept the nesting and crossing statements of PROC GLM; therefore, we need to construct the variables `acidXry0` and `acidXry1` in an extra data step. The `descending` option is needed to make PROC LOGISTIC model  $P(\text{nodes} = 1)$ . The default is to model  $P(\text{nodes} = 0)$ . The complete set of SAS models is calculated in file (logi/code/logi.c2.sas).

## 17.8 Fitting Models When the Response Is a Sample Proportion

The examples in Sections 17.1 and 17.4 both have dichotomous (0 or 1) responses. As noted in the introductory part of this chapter, responses for logistic regression models can also be proportions. Models with a proportion-valued response variable are requested in Exercises 17.8 and 17.9. Specifying such models in SAS and S-PLUS differs slightly from specifications of models with dichotomous responses.

Suppose the total number of observations is in a column named `n.total` for S-PLUS or `n_total` for SAS, and the number of observations having the attribute under study is in a column named `n.attribute` or `n_attribute`. Further assume there are two explanatory variables `X1` and `X2`; extension to situations with more than two explanatory variables will be obvious.

S-PLUS and SAS use different syntax for this type of problem.

SAS PROC GENMOD uses the *proportion*, expressed as a fraction "`n_attribute/n_total`" for the response in the model statement. Here is the entire model statement

```
MODEL n_attribute/n_total = X1 X2 / dist=binomial
```

S-PLUS `glm()` uses the *odds*, expressed as the numerator and denominator of the odds ratio “`cbind(n.attribute, n.total-n.attribute)`”, for the response in the model formula. Here is the entire model formula

```
glm(cbind(n.attribute, n.total-n.attribute) ~ X1 + X2,
    family=binomial)
```

## 17.9 LogXact

The model estimates shown in Section 17.2 and confidence limits and *p*-values in computer listings throughout this chapter are based on maximum likelihood estimation, and asymptotic (large-sample) standard errors and normality. LogXact<sup>©</sup> (Cytel Software Corporation, 2004) is a standalone software package that uses algorithms from (Mehta et al., 2000) to perform exact, rather than asymptotic, estimation and inference for logistic regression models. Documentation for LogXact states that its exact results sometimes differ appreciably from the asymptotic results provided by SAS and S-PLUS. We recommend that readers who work extensively with logistic regression models become familiar with this package.

## 17.10 Exercises

- 17.1.** In Equations (17.6) and (17.10) we show the likelihood and loglikelihood equations for the logistic regression model (17.4). For the dataset

| x | y |
|---|---|
| 1 | 1 |
| 2 | 0 |
| 3 | 1 |
| 4 | 0 |

estimate the parameters  $\beta_0$  and  $\beta_1$ . This problem is almost doable by hand. The derivatives in Equation (17.11) are tedious (remember to substitute the data values for  $x$  and  $y$  and to take the derivatives with respect to  $\beta_j$ ). Once you have simplified the derivatives (an easy task), you have two nasty fractions in  $\beta_0$  and  $\beta_1$  to set equal to zero and solve simultaneously (a hard task). Verify that you have done the algebra and arithmetic correctly by comparing the results to a computer program. You can also substitute these answers into your simultaneous equations and verify that they are both satisfied.

Computer programs usually use iterative techniques rather than the brute-force technique suggested here.

- 17.2.** Complete the analysis of the data in Exercise 17.1. Plot the data and the analysis and interpret the numerical values and the graph.
- 17.3.** (Mendenhall et al., 1989), subsequently discussed in several other places, studied the effect of radiotherapy on the absence (`response = 1`) or presence (`response = 0`) of tongue carcinoma three years after the completion of radiotherapy. The explanatory variable `days` is the number of days across which the fixed total dose of radiotherapy was administered. The data file is (`datasets/tongue.dat`).
- Use logistic regression to model `response` as a function of `days`. Is the coefficient of `days` significantly different from 0? Estimate the odds `response` resulting from one additional day of radiotherapy.
  - Produce a plot of the fitted equation with 95% prediction bands analogous to Figure 17.3d.
  - The negative arithmetic sign of the coefficient of `days` may at first glance seem counterintuitive. Offer a possible explanation for this result.
- 17.4.** Data file (`datasets/icu.dat`), from (Hosmer and Lemeshow, 2000), presents the ICU data, a selection of cases from a larger study of survival in an intensive care unit. The major goal of the study was to develop a logistic regression model to predict the probability of survival to hospital discharge of these patients. The code sheet for the data is in Table 17.7. You may use our files (`logi/code/icu.s`) or (`logi/code/icu.sas`) to read the data. We will initially look at just two variables, `AGE` and `SEX`. There are a few young males and no young females who did not survive, suggesting that `SEX` might be an important indicator.
- Plot the response variable `STA` (survival status) against the continuous variable `AGE` conditioned on `SEX`.
  - Fit a logistic regression model to `STA` with predictor variables `AGE` and `SEX`.
  - Plot the logistic fit of `STA` against `AGE` conditioned on `SEX`, comparable in style to Figure 17.16.
  - Interpret your results.
- 17.5.** For the same dataset (`datasets/icu.dat`) used in Exercise 17.4, we will also look at CPR status. Did the patient receive CPR prior to being admitted?
- Plot the response variable `STA` (survival status) against the continuous variable `AGE` conditioned on CPR.

TABLE 17.7. Code Sheet for the ICU Study. This is Table 1.5 from (Hosmer and Lemeshow, 2000).

| Variable | Description   | Codes/Values   | Name |
|----------|---|--|------|
| 1        | Identification Code   | ID Number  | ID   |
| 2        | Vital Status  | 0 = Lived, 1 = Died                                      | STA  |
| 3        | Age   | Years  | AGE  |
| 4        | Sex   | 0 = Male, 1 = Female                                     | SEX  |
| 5        | Race  | 1 = White, 2 = Black, 3 = Other                          | RACE |
| 6        | Service at ICU<br>Admission                                   | 0 = Medical, 1 = Surgical                                | SER  |
| 7        | Cancer Part of<br>Present Problem                             | 0 = No, 1 = Yes  | CAN  |
| 8        | History of Chronic<br>Renal Failure                           | 0 = No, 1 = Yes  | CRN  |
| 9        | Infection Probable<br>at ICU Admission                        | 0 = No, 1 = Yes  | INF  |
| 10       | CPR Prior to ICU<br>Admission                                 | 0 = No, 1 = Yes  | CPR  |
| 11       | Systolic Blood Pressure<br>at ICU Admission                   | mm Hg  | SYS  |
| 12       | Heart Rate<br>at ICU Admission                                | Beats/min  | HRA  |
| 13       | Previous Admission to<br>an ICU Within 6 Months               | 0 = No, 1 = Yes  | PRE  |
| 14       | Type of Admission   | 0 = Elective, 1 = Emergency                              | TYP  |
| 15       | Long Bone, Multiple,<br>Neck, Single Area, or<br>Hip Fracture | 0 = No, 1 = Yes  | FRA  |
| 16       | PO2 from Initial<br>Blood Gases                               | 0 = $> 60$ , 1 = $\leq 60$                               | PO2  |
| 17       | PH from Initial<br>Blood Gases                                | 0 = $\geq 7.25$ , 1 = $< 7.25$                           | PH   |
| 18       | PCO2 from Initial<br>Blood Gases                              | 0 = $\leq 45$ , 1 = $> 45$                               | PCO  |
| 19       | Bicarbonate from<br>Initial Blood Gases                       | 0 = $\geq 18$ , 1 = $< 18$                               | BIC  |
| 20       | Creatinine from<br>Initial Blood Gases                        | 0 = $\leq 2.0$ , 1 = $> 2.0$                             | CRE  |
| 21       | Level of Consciousness<br>at ICU Admission                    | 0 = No Coma or Deep Stupor,<br>1 = Deep Stupor, 2 = Coma | LOC  |

- b. Fit a logistic regression model to STA with predictor variables AGE and CPR.
- c. Plot the logistic fit of STA against AGE conditioned on CPR, comparable in style to Figure 17.16.

d. Interpret your results.

- 17.6. The data file `esr.dat` from (Collett, 1991) deals with the relationship between erythrocyte sedimentation rate (`ESR`) and two other blood chemistry measures, `fibrin`, the level of plasma fibrinogen (g/liter), and gamma `globulin` level (g/liter). `ESR` is the settlement rate of red blood cells from suspension in blood. Healthy individuals have `ESR` below 20 mm/hr. In this analysis the variable `ESR` is dichotomized to be 1 if `ESR` < 20 and 0 if `ESR`  $\geq$  20. Use logistic regression to determine if `fibrin` or `globulin` impact on `ESR`.
- 17.7. (Lee, 1980) describes an investigation to determine if any of six prognostic variables can be used to predict whether a patient will respond to a treatment for acute myeloblastic leukemia. The data file (`datasets/leukemia.dat`) contains the prognostic variables in columns 1–6 and `response` = 1 if responds to treatment and 0 if doesn't respond to treatment in column 7. (This data file also contains variables in columns 8–9 that are to be ignored in this exercise.) The prognostic variables are

`age`: in years

`smear`: smear differential (%)

`infiltrate`: absolute infiltrate (%)

`labeling`: labeling index (%)

`blasts`: absolute blasts ( $\times 10^3$ )

`temp`: temperature ( $\times 10$  degrees F)

a. Construct scatterplot matrices of the prognostic variables conditioned on the two values of `response`. Which prognostic variables seem most closely associated with `response`?

b. Justify a log transformation of `blasts`.

c. Find a good-fitting model to explain `response`. Interpret all estimated logistic regression coefficients.

- 17.8. (Higgins and Koch, 1977), later reprinted in (Hand et al., 1994), discuss a study of workers in the U.S. cotton industry to discover factors that relate to contraction of the lung disease byssinosis. In the data file (`datasets/byss.dat`) the columns are

`yes`: number suffering from byssinosis

`no`: number not suffering from byssinosis

**dust**: dustiness of workplace, 1 = high, 2 = medium, 3 = low

**race**: 1 = white, 2 = other

**sex**: 1 = male, 2 = female

**smoker**: 1 = yes, 2 = no

**emp.length**: length of employment: 1 = “<10 years”, 2 = “10–20 years”, 3 = “>20 years”

- a. Fit a logistic model to explain what factors affect the probability of contracting byssinosis. Since the response variable involves the counts in the two response categories rather than a dichotomous indicator, you will need to use model syntax similar to that described in Section 17.8. Look at the main effects and the two-way interactions.
- b. Produce a plot showing all the significant main effects and interactions. One such plot is a multipanel display of the observed proportion of byssinosis sufferers against each of the significant effects.
- c. Carefully state all conclusions.

- 17.9.** (Murray et al., 1981), also in (Hand et al., 1994), report on a survey that explored factors affecting the pattern of psychotropic drug consumption. The columns of the file **psycho.dat** are

**sex**: 0 = male, 1 = female

**agegroup**

**mean.age**

**GHQ**: General Health Questionnaire, 0 = low, 1 = high

**taking**: number taking psychotropic drugs

**total**: total number

- a. Use logistic regression to model the proportion taking psychotropic drugs as a function of those explanatory variables among **sex**, **agegroup**, and **GHQ** and their interactions that are significant. Since the response variable is a sample proportion, not a dichotomous indicator, you will have to fit models using the syntax described in Section 17.8.
- b. Produce a plot showing all the significant main effects and interactions. One such plot is a multipanel display of the observed proportion of byssinosis sufferers against the significant effects.

c. Interpret the logistic regression coefficients of all main effects.

**17.10.** Table 17.5 shows a  $p$ -value of 0.45 for the test that the coefficient of `grade` is zero.

a. Fit the model `logit.p.hat ~ X.ray + stage + acid.ph`.

b. Account for the difference between the  $p$ -values in Table 17.5 and in part 17.10a by fitting and interpreting a model containing the term `acid.ph %in% grade`.

# Time Series Analysis

## 18.1 Introduction

Time series analysis is the technique used to study observations that are measured over time. Examples include natural phenomena (temperature, humidity, wind speed) and business variables (price of commodities, stock market indices) that are measured at regular intervals (hourly, daily).

Like regression analysis, time series analysis seeks to model a response variable as a function of one or more explanatory variables. Time series analysis differs from other forms of regression analysis in one fundamental way. Previously we have assumed the observations are uncorrelated with each other, except perhaps through their dependency on the explanatory variables. Thus, in regression, we would model

$$Y = X\beta + \epsilon$$

and make the assumption that  $\text{corr}(\epsilon_i, \epsilon_j) = 0$ . In time series analysis we do not make the assumption of independence. Instead we make an explicit assumption of **dependence** and the task of the analysis is to model the dependence.

Conventionally, the notation used to denote a time series is  $X_t, t = 1, 2, \dots, n$ . In this chapter we assume that the successive times at which observations are taken are equally spaced apart, for example, monthly, quarterly, or annual observations. Initially, we also assume that  $X_t$  is a *stationary* zero-mean time series. (A time series is said to be stationary if the distribution of  $\{X_t, \dots, X_{t+n}\}$  is the same as that of  $\{X_{t+k}, \dots, X_{t+n+k}\}$  for any choice of  $t, n$ , and  $k$ .) This means that as time passes, the series

does not drift away from its mean value. Often when stationarity is absent it can be achieved by analyzing differences between successive terms of the time series rather than the original time series itself.

The term *lag* is used to describe earlier observations in a sequence. We indicate lagged observations with the *backshift* operator  $B$ , defined to mean

$$BX_t = B^1 X_t = X_{t-1}$$

and by extension

$$B^2 X_t = B(BX_t) = BX_{t-1} = X_{t-2}$$

Often  $B$  is used as the argument of a polynomial function. For example, if

$$\psi(B) = 3B^2 - 4B + 2$$

then

$$\psi(B)X_t = 3X_{t-2} - 4X_{t-1} + 2X_t$$

There is a distinction between the residual error  $\epsilon$  in regression analysis and  $\varepsilon$  components of time series models. A time series  $X_t, t = 1, 2, \dots, n$ , is a special case of a large class of models known as *stochastic processes*. Random variables  $\varepsilon_t, \varepsilon_{t-1}, \dots$  are *random shocks* to the process. This shock concept is distinct from regression residuals  $\epsilon$  that represent the inability of model predictors to completely explain the response. In some time series models a linear combination of lagged  $\varepsilon$ 's,  $\sum \theta_k \varepsilon_{t-k}$ , can be viewed as an error concept analogous to  $\epsilon$  in regression.

The ARIMA class of models discussed in this chapter can be fit to most regular time series that exhibit systematic behavior with random perturbations that are small compared to the systematic components. They are not appropriate for modeling time series having irregular cyclical behavior (such as the business cycle when modeling Gross National Product) or irregular sizeable shocks (such as federal spending for relief from natural disasters such as major hurricanes, floods, earthquakes, etc.).

Standard time-related data manipulations are easier when the time parameter is built into the data object. Fundamental operations like comparisons of two series or merging two series (`ts.union` or `ts.intersect`) are easily specified and the program automatically aligns the time parameter. Table 18.1 illustrates the two alignment options. The S-PLUS command `rts` stands for regular time series, that is, all  $\Delta t_i$  are equal. Other possible commands, not shown here, include `cts` for calendar dates associated with the observations and `its` for series sampled at irregular intervals.

TABLE 18.1. Alignment of time series with the (`ts.union` or `ts.intersect`) functions.

---

|   |  |
|---|--|
| <pre>&gt; x &lt;- rts(sample(10), start=1978) &gt; y &lt;- rts(sample(6), start=1980) &gt; x 1978:  8  4  3  2 10  7  9  5  6  1       start   deltat   frequency       1978        1            1 &gt; y 1980: 3 6 2 4 1 5       start   deltat   frequency       1980        1            1 &gt; ts.union(x,y)       x   y 1978 8 NA 1979 4 NA 1980 3 3 1981 2 6 1982 10 2 1983 7 4 1984 9 1 1985 5 5 1986 6 NA 1987 1 NA       start   deltat   frequency       1978        1            1</pre> | <pre>&gt; ts.intersect(x,y)       x   y 1980 3 3 1981 2 6 1982 10 2 1983 7 4 1984 9 1 1985 5 5       start   deltat   frequency       1980        1            1</pre> |
|---|--|

---

## 18.2 The ARIMA Approach to Time Series Modeling

In this chapter we introduce the Box–Jenkins ARIMA approach to time series modeling (Box and Jenkins, 1976). This methodology involves two primary types of dependence structures, autoregression and moving averages, as well as the concept of differencing. We assume throughout that the independent random shocks  $\varepsilon_t$  are distributed with mean 0 and a common variance  $\sigma^2$ .

### AutoRegression (AR):

The equation describing the first-order autoregression model AR(1) is

$$X_t = \phi X_{t-1} + \varepsilon_t \quad (18.1)$$

Each observation  $X_t$  is correlated with the preceding observation (at lag=1)  $X_{t-1}$  and, to a lesser extent, with all earlier observations. In the AR(1) model, each observation  $X_t$  has correlation  $\phi$  with the preceding

(at lag=1) observation  $X_{t-1}$ . The correlation of  $X_t$  with  $X_{t-k}$  is

$$\text{corr}(X_t, X_{t-k}) = \phi^k, \quad k = 1, 2, \dots \quad (18.2)$$

That is, the correlation decreases exponentially with the length of lag. For example,

$$\begin{aligned} \text{corr}(X_t, X_{t-1}) &= \text{corr}(\phi X_{t-1} + \varepsilon_t, X_{t-1}) \\ &= \phi \end{aligned}$$

and

$$\begin{aligned} \text{corr}(X_t, X_{t-2}) &= \text{corr}(\phi X_{t-1} + \varepsilon_t, X_{t-2}) \\ &= \text{corr}(\phi(\phi X_{t-2} + \varepsilon_{t-1}) + \varepsilon_t, X_{t-2}) \\ &= \phi^2 \text{corr}(X_{t-2}, X_{t-2}) + \text{corr}(\phi \varepsilon_{t-1} + \varepsilon_t, X_{t-2}) \\ &= \phi^2 + 0 \end{aligned}$$

The AR(1) model is further discussed in Section 18.5.2.

With  $p$ -order lags, the autoregression equation is written as

$$\Phi_p(B)X_t = \varepsilon_t \quad (18.3)$$

where

$$\Phi_p(B) = \phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$$

is a  $p^{\text{th}}$ -degree polynomial. This model is referred to as AR( $p$ ). The AR(1) model in Equation (18.1) is the special case where  $\Phi_p(B) = \Phi_1(B) = 1 - \phi B$ .

### Moving Average (MA):

The equation describing the first-order moving average model MA(1) is

$$X_t = \varepsilon_t - \theta \varepsilon_{t-1} \quad (18.4)$$

This model is called “moving average” because the right-hand side is a weighted moving average of the independent random shock  $\varepsilon_t$  at two adjacent time periods.

Each observation  $X_t$  in the MA(1) model is correlated with the preceding observation  $X_{t-1}$  and is uncorrelated with earlier observations. For example,

$$\text{corr}(X_t, X_{t-1}) = -\theta/(1 + \theta^2)$$

and

$$\begin{aligned} \text{corr}(X_t, X_{t-2}) &= \text{corr}(\varepsilon_t - \theta \varepsilon_{t-1}, \varepsilon_{t-2} - \theta \varepsilon_{t-3}) \\ &= 0 \end{aligned}$$

With  $q$ -order lags, the equation is written as

$$X_t = \Theta_q(B)\varepsilon_t \quad (18.5)$$

where

$$\Theta_q(B) = \theta(B) = 1 - \theta_1 B - \dots - \theta_q B^q$$

is a  $q^{\text{th}}$ -degree polynomial. This model is denoted MA( $q$ ). The MA(1) model in Equation (18.4) is the special case where  $\Theta_q(B) = \Theta_1(B) = 1 - \theta B$ . The MA(1) model is further discussed in Section 18.5.3.

### Differencing:

Differencing of order 1 is defined by

$$\nabla X_t = (1 - B)X_t = X_t - X_{t-1} \quad (18.6)$$

Simple models are written for the differenced data, for example,

$$\nabla X_t = \varepsilon_t - \theta \varepsilon_{t-1} \quad (18.7)$$

or, equivalently

$$X_t - X_{t-1} = \varepsilon_t - \theta \varepsilon_{t-1}$$

Model (18.7) is structurally the same as Model (18.4) in that it has the same right-hand side. The left-hand sides differ. Model (18.7) uses the differenced time series  $\nabla X_t$  as its response variable where Model (18.4) used the observed variable  $X_t$ . Differencing removes nonstationarity in the mean. More complicated models involving higher-order differencing are denoted by a polynomial

$$\nabla^d(B) = (1 - B)^d$$

The interpretation is

$$\nabla^1(B)X_t = X_t - X_{t-1}$$

### ARIMA (Autoregressive Integrated Moving Average) Models:

We work with both AR( $p$ ) and MA( $q$ ) with lags greater than or equal to 1, and with a combined situation called ARIMA( $p, d, q$ ) (autoregressive integrated moving average). The term *integrated* means that we use the AR and MA techniques on *differenced* data. The general form of the ARIMA( $p, d, q$ ) model is

$$\Phi_p(B) \nabla^d X_t = \Theta_q(B) \varepsilon_t \quad (18.8)$$

where  $\varepsilon_t$  is a random shock with mean zero and  $\text{var}(\varepsilon_t) = \sigma_\varepsilon^2$ .

There are many important special cases.

**ARIMA(1,0,0) = AR(1)** model is in Equation (18.1).

**ARIMA(0,0,1) = MA(1)** model is in Equation (18.4).

**ARIMA(0,1,0)** is the first difference model in Equation (18.6).

**ARIMA(0,1,1)** model is shown in Equation (18.7).

**ARIMA(1,1,1)** model looks like

$$\begin{aligned}\Phi_1(B)\nabla X_t &= \Theta_1(B)\varepsilon_t \\ (1 - \phi B)(1 - B)X_t &= (1 - \theta B)\varepsilon_t \\ (1 - (1 + \phi)B + \phi B^2)X_t &= (1 - \theta B)\varepsilon_t \\ X_t - (1 + \phi)X_{t-1} + \phi X_{t-2} &= \varepsilon_t - \theta \varepsilon_{t-1}\end{aligned}$$

**ARIMA( $p, 0, q$ )** with  $d = 0$ , hence no differencing, is also called an ARMA( $p, q$ ) model (autoregressive moving average)).

## 18.3 Autocorrelation

Two principal tools for studying time series are the autocorrelation function (ACF) and the partial autocorrelation function (PACF). The ACF assists in the diagnosis of MA models. The PACF is used in the diagnosis of AR models.

### 18.3.1 Autocorrelation Function (ACF)

The defining equation for the lag- $k$  autocorrelation coefficient  $\rho_k$  is

$$\rho_k = \text{corr}(X_t, X_{t-k})$$

The discrete function  $\{\rho_k\}$  indexed by the lag  $k$  is called the autocorrelation function of the series  $Z$ . The sample estimators  $\{r_k\}$  are defined by

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum_{t=1}^n X_t \\ c_k &= \frac{1}{n} \sum_{t=k+1}^n (X_t - \bar{X})(X_{t-k} - \bar{X}) \quad \text{autocovariance} \\ r_k &= c_k / c_0 \quad \text{autocorrelation}\end{aligned}$$

Note that the division is always by  $n$ .

The ACF for a time series  $z = (x_1, \dots, x_n)$  is calculated in S-PLUS by

`acf(z)`

### 18.3.2 Partial Autocorrelation Function (PACF)

The defining equation for the PACF is

$$\text{pacf}(k) = \text{corr}(X_t, X_{t-k} | X_{t-1}, X_{t-2}, \dots, X_{t-(k-1)})$$

TABLE 18.2. Illustrative definition of the PACF. Do NOT use in actual calculations!

```
S-PLUS (tser/code/my.pacf.s):
## This function illustrates the definition of the pacf.
## Do NOT use in actual calculations!
##
## my.pacf requires a detrended series, otherwise the answer is
## nonsense as it starts losing precision after the first few lags.
##
## The argument z must have a class of "rts", "cts", or "its".
## The function will not work with class "ts".

my.pacf <- function(z, k=2) {
  z <- z - mean(z)
  x <- ts.intersect(z, lag(z,-1))
  if (k==1) return(cor(x[,1], x[,2]))
  for (kk in 2:k) x <- ts.intersect(x, lag(z,-kk))
  nr <- nrow(x)
  nc <- ncol(x)
  r1 <- lm(x[,1] ~ -1 + x[,-c(1,nc)])$resid
  r2 <- lm(x[,nc] ~ -1 + x[,-c(1,nc)])$resid
  cor(r1,r2)
}
```

The sample estimators are defined by solving the Yule–Walker equations that hold for an AR( $p$ ) process (see (Box and Jenkins, 1976) for details). An illustrative (but not practical) estimator is shown in the S-PLUS function in Table 18.2. The PACF is calculated in S-PLUS by

```
acf(z, type="partial")
```

Note that  $\text{pacf}(1) = \text{acf}(1)$ .

## 18.4 Analysis Steps

There are three main steps in time series analysis using the ARIMA models of the Box–Jenkins approach.

**Identification:** choice of the proper transformations to apply to the time series, consisting of variance-stabilizing transformations and of differencing. Determining the number of model parameters:  $d$ , the order of differencing;  $p$ , the number of autoregressive parameters; and  $q$ , the number of moving average parameters.

**Estimation:** estimation of the parameters of the identified model, usually by maximum likelihood.

**Diagnostics:** verification that the estimated model and parameters do indeed capture the essence of the behavior of the data.

We offer these recommendations for interpreting sequence plots, ACF plots, and PACF plots. They are based on the Box–Jenkins methodology described in the texts by (Box and Jenkins, 1976) and (Wei, 1990).

1. Trends in the sequence plot must be removed by differencing. This is required before attempting to interpret the ACF and PACF plots. The interpretation below of ACF and PACF plots depends on stationarity.
2. No correlation—white noise  
The ACF and PACF are negligible at all lags.
3. AR( $p$ )  
The ACF decays slowly.  
The PACF cuts off at lag  $p$ .
4. MA( $q$ )  
The ACF cuts off at lag  $q$ .  
The PACF decays slowly.
5. ARMA( $p, q$ )  
The ACF decays slowly from lag  $\max(q - p, 0)$  on.  
The PACF decays slowly from lag  $\max(p - q, 0)$  on.  
The orders  $p$  and  $q$  usually can't be read directly from these plots. Looking at the ACF and PACF plots for models with larger values of  $p$  and  $q$  can be helpful. The ESACF (extended sample autocorrelation function) (see, for example, (Wei, 1990), p. 128) can also be helpful.

Several additional tools are used to identify well-fitting models.

- The *Akaike information criterion* (AIC) for a particular model is defined as  $-2(\ln L) + 2m$ , where  $L$  is the model's loglikelihood and  $m$  is the number of parameters needed to estimate the model. Like the  $C_p$  statistic used to decide among multiple regression models, introduced in Equation (9.28), the AIC is the sum of a goodness-of-fit component and a penalty for lack of simplicity. Low values of AIC are preferred to large values.
- The *portmanteau goodness-of-fit test* for a particular model at lag  $\ell$  is actually a collection of tests, one for each  $k = 1, 2, \dots, \ell$ . Each of these individual tests is a test of the negligibility of the autocorrelations of the model residuals up to and including lag  $\ell$ . In a well-fitting model, these hypotheses should be retained because such a model should have negligible autocorrelations. Therefore, for well-fitting models the  $p$ -values of these tests should not be small.

- The highest-order AR and MA parameters of well-fitting models are significantly different from zero, indicating that the corresponding model terms and terms of lower order are needed. Such significance is suggested by a corresponding  $t$  statistic that exceeds 2 in absolute value. For example, a model with  $p = 2$  must have  $\phi_2$  significantly different from zero, but it is not essential that  $\phi_1$  be significantly nonzero.
- A well-fitting model has an estimated residual variance  $\hat{\sigma}^2$  at least as small as those of competing models. The estimated residual variance and the AIC carry similar information. The AIC is usually preferred to the residual variance because the AIC includes a penalty for lack of simplicity and the residual variance does not.

In most situations these diagnostics will point to the same uniquely best model or subset of equivalently well-fitting models.

## 18.5 Some Algebraic Development, Including Forecasting

Usually, the ultimate purpose of finding a well-fitting time series model is the production of forecasts and forecast intervals  $h$  periods beyond the final observation of the existing series. While we have shown how to produce forecasts and intervals in S-PLUS, here we provide a brief introduction to the algebra behind such forecasts. The algebra is intractable by hand for all but a few special cases.

### 18.5.1 The General ARIMA Model

The time series model for a 0-mean time series  $X_t$ , with  $E(X_t) = 0$  and  $\text{var}(X_t) = \sigma^2$ , is

$$\phi(B)\nabla^d X_t = \theta(B)\varepsilon_t \quad (18.9)$$

One way to rewrite (18.9) is

$$\varepsilon_t = \theta^{-1}(B)\phi(B)\nabla^d X_t$$

Once the coefficients of  $\phi$  and  $\theta$  have been estimated by maximum likelihood, the fitted model is expressed in terms of the calculated residuals as

$$\hat{\varepsilon}_t = \hat{\theta}^{-1}(B)\hat{\phi}(B)\nabla^d X_t$$

where  $\hat{\theta}(\cdot)$  and  $\hat{\phi}(\cdot)$  are the polynomials in  $B$  after the coefficient estimates have been substituted into  $\theta(\cdot)$  and  $\phi(\cdot)$ .

The calculated residuals  $\hat{\varepsilon}_t$  will be used in many subsequent calculations.

The model (18.9) can also be rewritten as

$$\begin{aligned} X_t &= \phi^{-1}(B)\nabla^{-d}\theta(B)\varepsilon_t \\ &\stackrel{\text{def}}{=} \psi(B)\varepsilon_t \\ &= (1 + \psi_1 B + \dots)\varepsilon_t \\ &= \varepsilon_t + \psi_1\varepsilon_{t-1} + \dots + \psi_k\varepsilon_{t-k} + \dots \end{aligned}$$

where  $\psi(B)$  may have an infinite number of terms. The number of terms is finite with purely MA models (where  $\psi = \theta$ ) and infinite when there are AR or differencing factors. In order that the model be stationary and invertible (that is, explicitly solvable for  $X_t$ ), it is required that the roots of both polynomials  $\phi(B)$  and  $\psi(B)$  lie outside the unit circle. In addition  $\phi(B)$  and  $\theta(B)$  must have no roots in common. If the polynomials have common roots, these roots can be factored out.

The nonzero-mean case is essentially the same. Let the nonzero-mean time series be  $Y_t = X_t + \mu$ . We can subtract the mean from the observed  $Y_t$ -values to construct a 0-mean times series  $X_t = Y_t - \mu$  and then proceed.

When we use the model for forecasting  $h$  steps ahead, we use the equation

$$\hat{X}_{t+h} = E(X_{t+h}|X_t, X_{t-1}, \dots) = (\psi_h + \psi_{h+1}B + \dots)\varepsilon_t$$

with forecast error

$$e_{t+h} = X_{t+h} - \hat{X}_{t+h} = \varepsilon_{t+h} + \psi_1\varepsilon_{t+h-1} + \dots + \psi_{h-1}\varepsilon_{t+1}$$

and with variance of the forecast error

$$\text{var}(e_{t+h}) = \sigma^2(1 + \psi_1^2 + \psi_2^2 + \dots + \psi_{h-1}^2)$$

The forecast error for  $h$ -step ahead forecasts, and its variance, have exactly  $h$  terms. The  $\varepsilon_t$  are uncorrelated. The forecast errors are correlated.

Probability limits for the forecasts are calculated as

$$\hat{X}_{t+h} \pm z_{\alpha/2} \hat{\sigma} \sqrt{1 + \psi_1^2 + \psi_2^2 + \dots + \psi_{h-1}^2}$$

### 18.5.2 Special case—The AR(1) model

Starting from  $X_2 = \phi X_1 + \varepsilon_2$ , incrementing the subscripts on  $X_t$ , and then back-substituting [for example,  $X_3 = \phi(\phi X_1 + \varepsilon_2) + \varepsilon_3$ ], we eventually get

$$X_{t+h} = \phi^h X_t + \phi^{h-1}\varepsilon_{t+1} + \dots + \phi\varepsilon_{t+h-1} + \varepsilon_{t+h}$$

As a consequence, we take  $\hat{X}_{t+h} = \hat{\phi}^h X_t$ . Further,

$$\begin{aligned}\text{var}(X_{t+h}) &= \sigma^2(1 + \phi^2 + \phi^4 + \dots + \phi^{2(h-1)}) \\ &= \sigma^2 \left( \frac{1 - \phi^{2h}}{1 - \phi^2} \right)\end{aligned}$$

A  $100(1 - \alpha)\%$  prediction interval for  $X_{t+h}$  is

$$\hat{X}_{t+h} \pm z_{\alpha/2} \hat{\sigma} \sqrt{\frac{1 - \hat{\phi}^{2h}}{1 - \hat{\phi}^2}}$$

### 18.5.3 Special Case—The MA(1) Model

Here we have  $X_{t+1} = \varepsilon_{t+1} - \theta_1 \varepsilon_t$ , and the general formulas simplify to

$$\begin{array}{lll}\hat{X}_{t+1} &= -\theta_1 \hat{\varepsilon}_t & \text{for } h = 1 \\ \hat{X}_{t+h} &= 0 & \text{for } h > 1 \\ \text{var}(\hat{X}_{t+1}) &= \sigma^2 & \text{for } h = 1 \\ \text{var}(\hat{X}_{t+h}) &= \sigma^2(1 + \theta_1^2) & \text{for } h > 1\end{array}$$

In the MA( $q$ ) models the  $\hat{\varepsilon}_{t+j}$ -values are known for past observations (those for which  $j \leq 0$ ), hence they appear in the prediction equations. A  $100(1 - \alpha)\%$  prediction interval for  $X_{t+h}$  is

$$\begin{array}{ll}\hat{X}_{t+1} & \pm z_{\alpha/2} \hat{\sigma} & \text{for } h = 1 \\ \hat{X}_{t+h} & \pm z_{\alpha/2} \hat{\sigma} \sqrt{1 + \theta_1^2} & \text{for } h > 1\end{array}$$

## 18.6 Graphical Displays for Time Series Analysis

We present a number of graphical displays to facilitate the identification and model checking steps of ARIMA( $p, d, q$ ) modeling. Much of this material previously appeared in (Heiberger and Teles, 2002). We discuss an extension of these displays to model time series with seasonal components in Section 18.8. A general discussion of the features of these graphs appears in Section 18.A of this chapter's appendix.

Table 18.3 summarizes the nine achievable models formed by possible combinations of the number of AR parameters ( $p = 0, 1, 2$ ) and MA parameters ( $q = 0, 1, 2$ ). The appearance of the left-hand and right-hand side of the model equations is shown for each value of  $(p, 0, q)$ .

TABLE 18.3.  $3 \times 3$  layout for the ARIMA( $p, 0, q$ ) models. All the time series diagnostic plots and summary tabular data are constructed on this pattern. The rows give the number of AR parameters ( $p = 0, 1, 2$ ) and the corresponding left-hand side of the model equation. The columns give the number of MA parameters ( $q = 0, 1, 2$ ) and the corresponding right-hand side of the model equation. For example, the (1, 1) cell of the array shows the information for the ARIMA(1, 0, 1) model:

$$X_t - \phi_1 X_{t-1} = \varepsilon_t - \theta_1 \varepsilon_{t-1}$$

In all the displays we show, the differencing parameter  $d$  ( $d = 0$  in this example) and the seasonal parameters ( $P, D, Q_s$ , if any) are held constant.

| Autoregression model |   | Moving average model — Right-hand side |  |   |
|----------------------|---|--|--|---|
|                      |   | $q = 0$                                | $q = 1$                                      | $q = 2$   |
| $p$                  | Left-hand side                          | $\varepsilon_t$                        | $\varepsilon_t - \theta_1 \varepsilon_{t-1}$ | $\varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2}$ |
| $p = 0$              | $X_t$                                   | (0, 0, 0)                              | (0, 0, 1)                                    | (0, 0, 2)   |
| $p = 1$              | $X_t - \phi_1 X_{t-1}$                  | (1, 0, 0)                              | (1, 0, 1)                                    | (1, 0, 2)   |
| $p = 2$              | $X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2}$ | (2, 0, 0)                              | (2, 0, 1)                                    | (2, 0, 2)   |

Figures 18.1 and 18.3 are examples of coordinated plots useful for identifying an ARIMA time series model.

Figure 18.1 contains a plot of the original time series along with its autocorrelation function and partial autocorrelation function. (Figure 18.2 is comparable to Figure 18.1 but for a differenced time series.)

The set of plots in Figure 18.3 consists of the residual ACF and PACF, the portmanteau goodness-of-fit test statistic (GOF), the standardized residuals, and the Akaike information criterion (AIC). The panels in the first four sets of plots are indexed by the number of ARMA parameters  $p$  and  $q$ . The AIC plot uses  $p$  and  $q$  as plotting variables. The orders of differencing and the orders of the autoregressive and moving average operators have been limited to  $0 \leq p, d, q, \leq 2$ . While this limitation is usually reasonable in practice, it is not inherent in the software.

Each set of nine panels is systematically structured in a  $3 \times 3$  array indexed by the number of AR parameters  $p$  and MA parameters  $q$ . All nine panels in a set are scaled identically. Thus the reader can scan a row or column of the array of panels and see the effect of adding one more parameter to either the AR or MA side of the model.

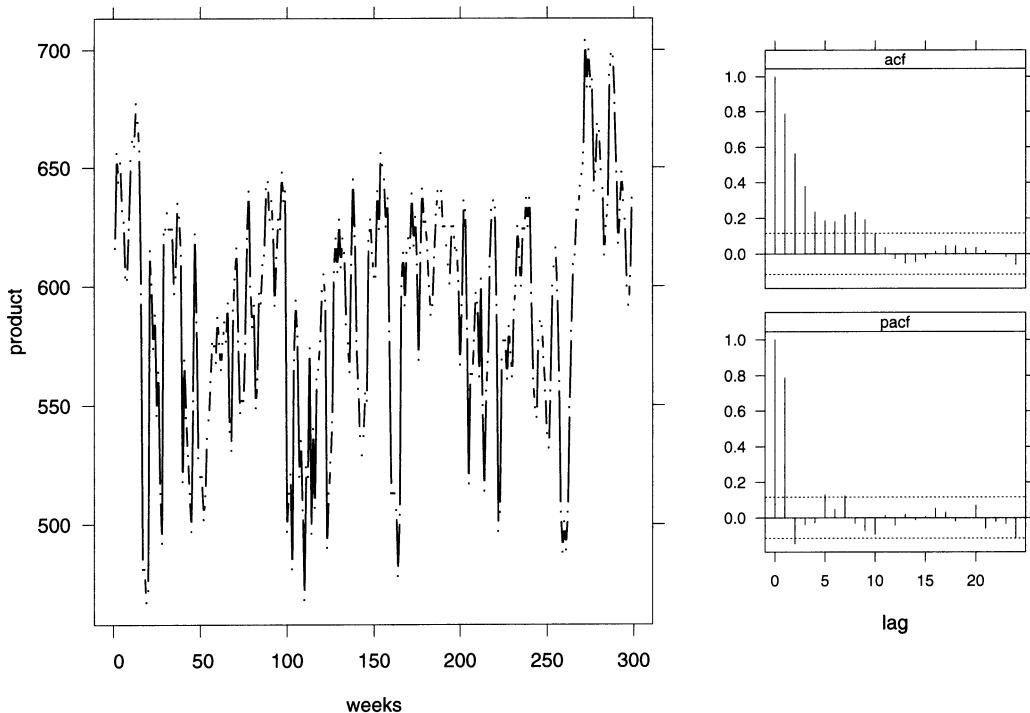


FIGURE 18.1. Coordinated time series plot and ACF/PACF plots for the `product` time series:  $y_t$ . The response variable on the time series plot is weekly sales of the product.  
`(tser/code/product.s), (tser/figure/prodfig1.eps.gz)`

The graphics are used to analyze the product data in (Hand et al., 1994), reprinted from the original in (Nicholls, 1979). These data are the weekly sales of a plastic container used for the packaging of drugs in the United States. Figures 18.1 and 18.2 show the reported data  $y_t = \text{product}$  and the first differences  $\nabla y_t \stackrel{\text{def}}{=} y_t - y_{t-1}$ . (The data file (`datasets/product.dat`) consists of the *cumulative* weekly sales. First differences were taken to produce the  $y_t$ .) The horizontal dashed lines on the ACF and PACF plots are the critical values for  $\alpha = .05$  tests of the hypothesis, at each individual lag  $k$ , that the correlation coefficient is zero. Spikes on these plots that fall outside these horizontal boundaries suggest the possibility of a nonzero correlation.

Figure 18.1 suggests that successive weeks' sales are positively associated:  $\text{corr}(y_t, y_{t-1}) \approx 0.8$ . To address the positive association between successive

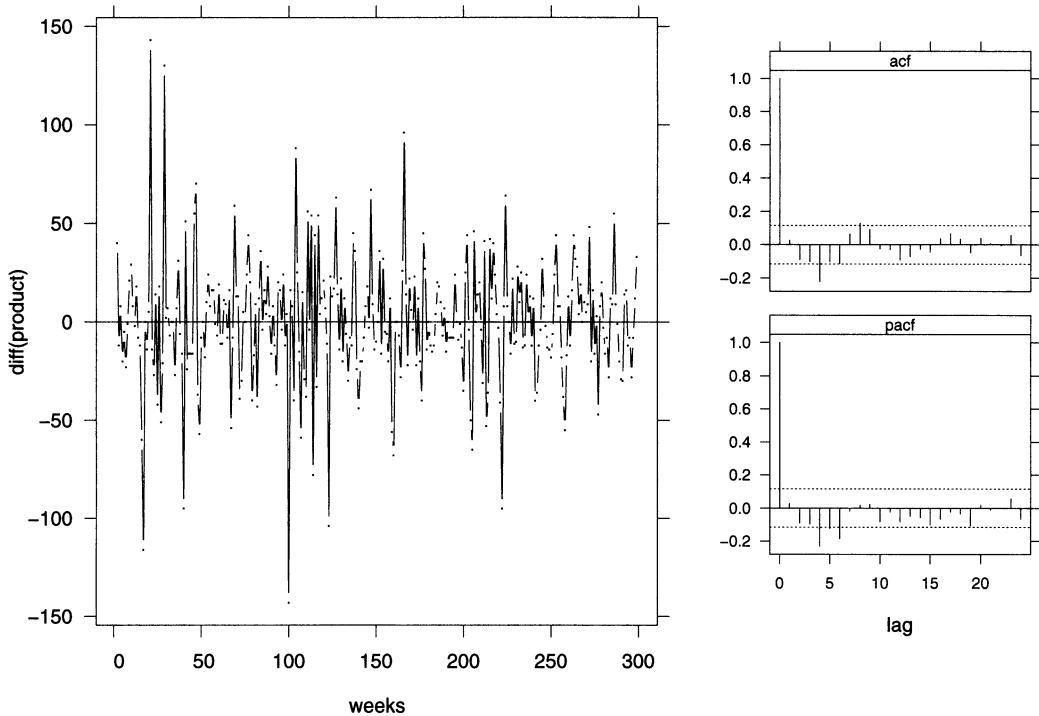


FIGURE 18.2. Coordinated time series and ACF/PACF plots for the differenced product time series:  
 $\nabla_t y_t$ .  
(tser/code/product.s), (tser/figure/prodfig2.eps.gz)

weeks we analyze first differences in Figure 18.2. The spikes at lag of 4 on both the ACF and PACF plots suggest the slight possibility of a monthly seasonal effect ( $4 \text{ weeks} \approx 1 \text{ month}$ ). We defer discussion of models with seasonal effects to Section 18.7 and of this example's seasonal effect to Exercise 18.4. No additional differencing is suggested by Figure 18.1. Since the ACF and PACF show systematic behavior, we proceed to Figure 18.3, a collection of five sets of coordinated plots on a single page designed to facilitate identifying the best ARIMA( $p, 1, q$ ) model based on fits of the nine models with  $0 \leq p, q \leq 2$ . For each of these nine models, Figure 18.3 shows the ACF and PACF plots of the standardized residuals, a plot of the  $p$ -values of portmanteau goodness-of-fit tests at various lags, the Akaike information criterion arranged in the form of a pair of interaction plots, and the standardized residuals themselves. While in theory it is possible

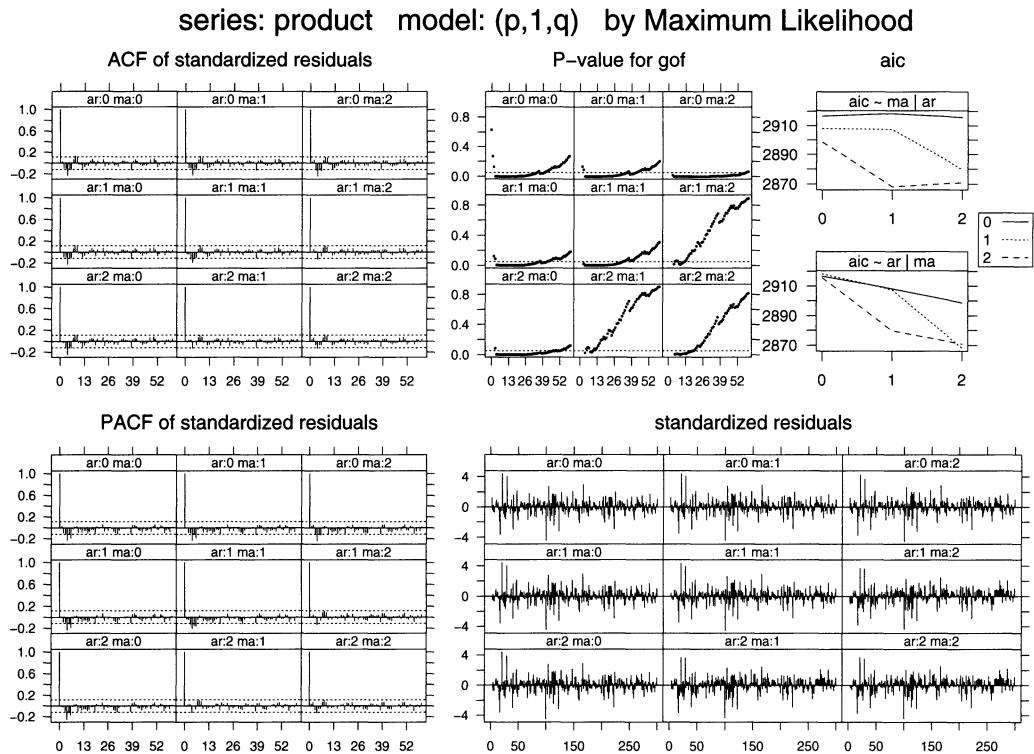


FIGURE 18.3. Diagnostic plots for the set of models ARIMA( $p,1,q$ ) fit to the Product data by maximum likelihood. Each set of nine panels is systematically structured in a  $3 \times 3$  array with rows indexed by the number of AR parameters  $p$  and columns by the number of MA parameters  $q$ . All nine panels in a set are scaled identically. The AIC has been plotted as a pair of interaction plots: AIC plotted against  $q$  using line types defined by  $p$ ; and AIC plotted against  $p$ , using line types defined by  $q$ .

(tser/code/product.s), (tser/figure/prodfig4.eps.gz)

for  $p$  or  $q$  to exceed 2 (and the software permits larger values of  $p$  and/or  $q$ ), this is unlikely to occur in practice provided that the data have been properly differenced and that any seasonal effects have been addressed.

Using Figure 18.3 we can immediately eliminate the six models with  $pq < 2$ . For all such models an ACF or PACF spike crosses the threshold of significance, some  $p$ -values for the goodness-of-fit test are below 0.05, and the Akaike criteria are higher than for the remaining models. We are unable to distinguish between the ordinary residual plots of the nine models.

This leads us to further examine the three remaining models having  $(p, q) = (1, 2), (2, 1)$ , and  $(2, 2)$ . As is clearly seen from Figure 18.3, the best of these models appears to be  $p = 2$  and  $q = 1$  because this model has the largest  $p$ -values of the goodness-of-fit test for small lags, and also the smallest value of the Akaike information criterion.

## 18.7 Models with Seasonal Components

One of the strengths of the Box–Jenkins method is its handling of seasonal parameters. Time series frequently show seasonal patterns in their correlation structure. Most economic series have annual, quarterly, monthly, or weekly patterns as well as daily patterns. For example, retail sales figures often shows a surge in activity in December; power consumption figures show seasonal patterns as heating is used in the winter months and air conditioning in the summer months, and a weekly pattern where weekday consumption differs systematically from weekend consumption.

### 18.7.1 Multiplicative Seasonal ARIMA Models

Seasonal parameters are handled similarly to the nonseasonal parameters, with the subscripts varying by increments of the season. With monthly data, an annual season  $s = 12$  is denoted by using 12-month lags, that is,  $X_t$  and  $X_{t-12}$ . We use uppercase Greek letters  $\Phi$  and  $\Theta$  to denote autoregressive and moving average polynomials, respectively, in the seasonal backshift operator  $B^s$ . The polynomials in the backshift  $B^s$  are denoted  $\Phi(B^s)$  and  $\Theta(B^s)$ , and the differences are  $\nabla_s = 1 - B^s$ .

The seasonal portion of a seasonal model is denoted  $\text{ARIMA}(P, D, Q)_s$  (with, for example, the seasonal  $s = 12$  used for annual seasons when the underlying data is monthly, and  $s = 7$  for weekly seasons when the underlying data is daily), where

$P$  is the number of lags in the seasonal AR portion of the model, equivalently the order of the polynomial

$$\Phi(B^s) = 1 - \Phi_1 B^s - \dots - \Phi_P B^{sP}$$

$Q$  is the number of lags in the seasonal MA portion of the model, equivalently the order of the polynomial

$$\Theta(B^s) = 1 - \Theta_1 B^s - \dots - \Theta_Q B^{sQ}$$

$D$  is the number of seasonal differences prior to the AR and MA modelling, equivalently the power of the differencing binomial

$$\nabla_s^D = (1 - B^s)^D$$

The general multiplicative seasonal model, denoted

$$\text{ARIMA}(p, d, q) \times (P, D, Q)_s$$

is given by

$$\Phi(B^s) \phi(B) \nabla_s^D \nabla^d X_t = \Theta(B^s) \theta(B) \varepsilon_t \quad (18.10)$$

For various technical reasons, the roots of the seasonal polynomials  $\Phi(B)$  and  $\Psi(B)$  must satisfy certain conditions that parallel the restrictions on  $\phi(B)$  and  $\psi(B)$  mentioned in Section 18.5.1. The roots of  $\Phi(B)$  and  $\Psi(B)$  must lie outside the unit circle to assure that the model is stationary and invertible (solvable for  $X(t)$ ). In addition,  $\Phi(B)$  and  $\Psi(B)$  must have no common roots. If these polynomials have common roots, these roots can be factored out.

### 18.7.2 Example—co2 ARIMA(0, 1, 1) $\times$ (0, 1, 1)<sub>12</sub> Model

The final model for the co2 example discussed in Section 18.8 is written as ARIMA(0, 1, 1)  $\times$  (0, 1, 1)<sub>12</sub> model for  $X_t$ :

$$\nabla_{12} \nabla X_t = (1 - \theta_1 B) (1 - \Theta_1 B^{12}) \varepsilon_t \quad (18.11)$$

which expands to

$$X_t - X_{t-1} - X_{t-12} + X_{t-13} = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \Theta_1 \varepsilon_{t-12} + \theta_1 \Theta_1 \varepsilon_{t-13}$$

### 18.7.3 Determining the Seasonal AR and MA Parameters

The procedure for determining the order  $P$  and  $Q$  of the seasonal parameters is comparable to the recommendations given in Section 18.4 for determining the order  $p$  and  $q$  of the nonseasonal parameters. As before, we work with the ACF and PACF for an appropriately differenced model. The distinction is that in examining the behavior of the ACF and PACF for seasonality, we examine only the values at seasonal intervals. For example, for monthly data with annual season ( $s = 12$ ), these plots are examined at  $t = 12, 24, 36, \dots = 12 \times (1, 2, 3, \dots)$ , ignoring values at other times. We then visualize the cutoff or decay behavior where *lag* now refers to seasonal intervals. If the ACF decays slowly at  $t = 12, 24, 36, \dots = 12 \times (1, 2, 3, \dots)$  and the PACF cuts off at  $t = 24 = 12 \times 2$ , then  $P = 2$  and  $Q = 0$ . If the PACF decays slowly at  $t = 12, 24, 36, \dots = 12 \times (1, 2, 3, \dots)$  and the ACF cuts off at  $t = 12 \times 1$ , then  $P = 0$  and  $Q = 1$ .

## 18.8 Example of a Seasonal Model—The Monthly $\text{CO}_2$ Data

We extend the graphical displays discussed in Section 18.6 to the identification and model checking steps of  $\text{ARIMA}(p, d, q) \times (P, D, Q)_s$  modeling. These graphs also first appeared in (Heiberger and Teles, 2002). A general discussion of the features of these graphs is deferred to Section 18.A.

The graphics are illustrated with one of the time series datasets distributed as part of S-PLUS, the Mauna Loa Carbon Dioxide Concentration series collected by the Scripps Institute of Oceanography, in La Jolla, California. The source is the climatology database maintained by the Oak Ridge National Laboratory (Peterson, 1990). These data represent monthly  $\text{CO}_2$  concentrations in parts per million (ppm) from January 1959 to December 1990. Missing values have been filled in by linear interpolation.

Figures 18.4 through 18.6 are structured presentations of the plots of the series itself, of the ACF, and of the PACF. Figure 18.4 displays the raw data series while Figure 18.5 displays the differenced series (monthly) and Figure 18.6 displays the twice-differenced series (monthly and annually).

### 18.8.1 Identification of the Model

Figure 18.4 is the plot of the observed data. The plot of the series itself shows a strong upward trend and a systematic labeling, with peaks occurring in the spring months and troughs in the autumn months. It is clear that the mean of this series is not constant over time. Both the ACF and PACF show systematic behavior. The ACF exhibits large values and a very slow decay with an annual periodicity. The PACF has large values and an annual periodicity. The conclusion is that the series is nonstationary, that is, it does not have a constant mean, and its autocorrelation function is time-dependent, implying that it shows nonrandom time-dependent behavior. Monthly differencing is required to model the nonstationarity, and annual differencing is necessary to remove the periodicity.

The time series and ACF and PACF plots for the differenced series  $\nabla X_t$  in Figure 18.5 also show systematic annual behavior. The time series plot shows August/September troughs. The ACF exhibits a very slow decay at the seasonal lags, lags that are multiples of the seasonal period  $s = 12$  months. This confirms that seasonal differencing with period 12 is required.

Figure 18.6 shows the time series (and the ACF and PACF) after non-seasonal and seasonal differencing  $\nabla_{12} \nabla X_t$ . There are no longer systematic components visible in the plot of the differences. The differenced series is stationary and it becomes possible to identify a model for the series, that is, to look for the AR and MA parameters that best fit the twice-differenced data.

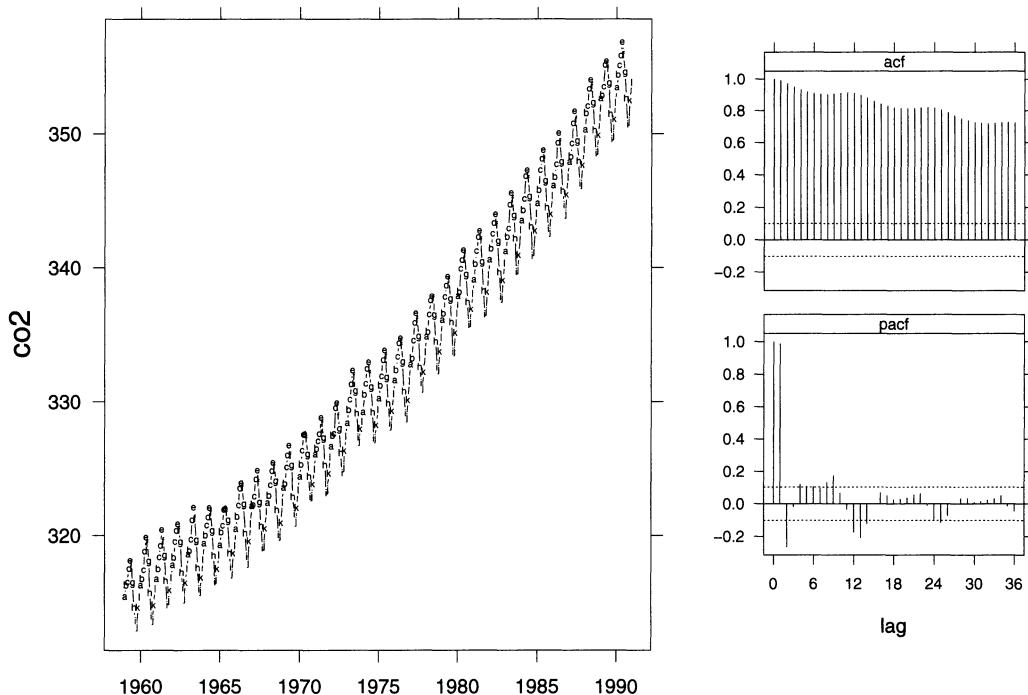


FIGURE 18.4. Coordinated time series plot and ACF/PACF plots for the Mauna Loa CO<sub>2</sub> time series:  $X_t$ . The response variable on the time series plot is concentration in parts per million.  
 (tser/code/tsamstat.s), (tser/figure/tsamsta1.ps.gz),  
 (tser/figure/tsamsta1.color.ps.gz)

The nonseasonal component of the model of  $X_t$  is identified by looking at the first few monthly lags of the sample ACF and PACF of  $\nabla_{12}\nabla X_t$  in Figure 18.6. The ACF seems to cut off after lag 1 and the PACF shows an exponential decay. The same type of behavior is seen at the seasonal lags, i.e., the ACF cuts off after lag 12 and the PACF shows an exponential decay at lags 12, 24, 36, .... These characteristics of the ACF and PACF suggest the ARIMA(0, 1, 1)  $\times$  (0, 1, 1)<sub>12</sub> model for  $X_t$ :

$$\nabla_{12}\nabla X_t = (1 - \theta_1 B)(1 - \Theta_1 B^{12}) \varepsilon_t \quad (18.12)$$

A closer look at the ACF in Figure 18.6 indicates that it too may show an exponential decay in the first lags, suggesting that the ARIMA(1, 1, 1)  $\times$

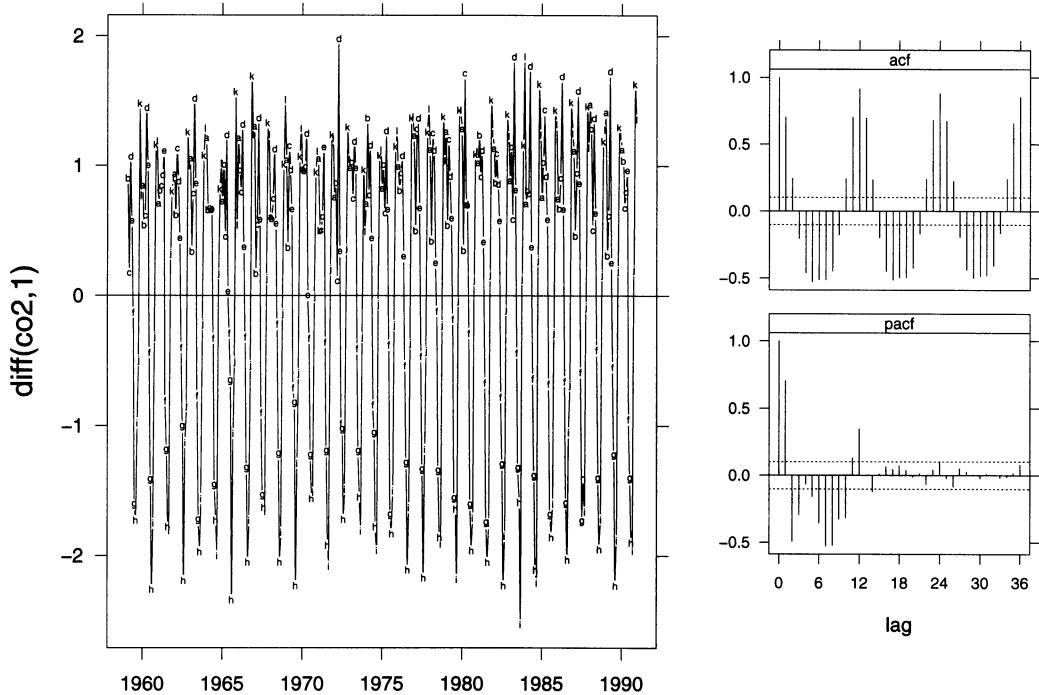


FIGURE 18.5. Coordinated time series plot and ACF/PACF plots for the differenced Mauna Loa CO<sub>2</sub> time series:  $\nabla X_t$ .  
 (tser/code/tsamstat.s), (tser/figure/tsamsta1a.ps.gz),  
 (tser/figure/tsamsta1a.color.ps.gz)

$(0, 1, 1)_{12}$  model

$$(1 - \phi_1 B) \nabla_{12} \nabla X_t = (1 - \theta_1 B) (1 - \Theta_1 B^{12}) \varepsilon_t \quad (18.13)$$

might also be appropriate.

### 18.8.2 Parameter Estimation and Diagnostic Checking

In general, when analyzing seasonal time series data, initial guesses of at least some of the parameters  $p, q, P, Q$  may be provided from inspections of coordinated plots of original and differenced data such as Figures 18.4 to 18.6. Figures 18.3 and 18.7 each simultaneously consider nine models produced with the user function `arma.loop` described in Section 18.A.4.

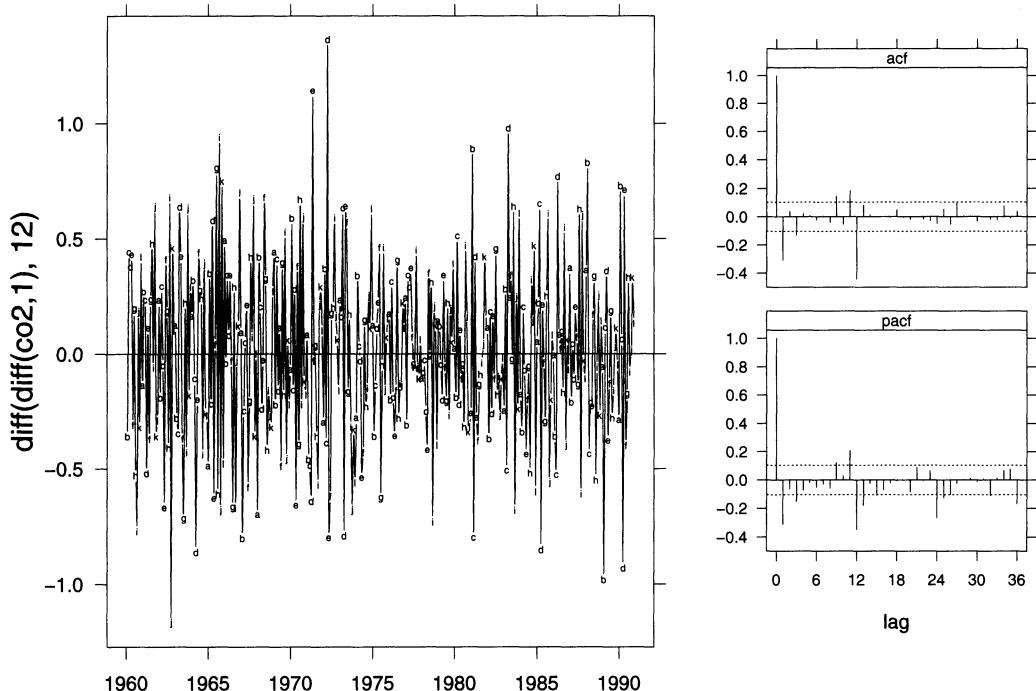


FIGURE 18.6. Coordinated time series plot and ACF/PACF plots for the twice-differenced Mauna Loa CO<sub>2</sub> time series:  $\nabla_{12}\nabla X_t$ .  
`(tser/code/tsamstat.s)`, `(tser/figure/tsamsta2.ps.gz)`,  
`(tser/figure/tsamsta2.color.ps.gz)`

Figures in this class can be used to suggest seasonal parameters  $P$  and  $Q$  for a given set of nonseasonal and differencing parameters  $p, q, d, D$ , or to suggest nonseasonal parameters  $p$  and  $q$  for a given set of seasonal and differencing parameters  $P, Q, d, D$ . Alternating consideration of figures of both of these types can be used to settle on a final model.

Continuing with the `co2` data, Figure 18.7 displays a set of diagnostic plots, comparable to Figure 18.3, for several models without a seasonal component, the ARIMA( $p, 1, q$ )  $\times$   $(0, 1, 0)_{12}$  models with  $0 \leq p, q \leq 2$ , that have been fit to the series  $\nabla_{12}\nabla X_t$ .

Since the `co2` data exhibit a seasonal behavior, Figure 18.7 is expected to confirm that seasonal parameters are required in the model of  $\nabla_{12}\nabla X_t$ . All

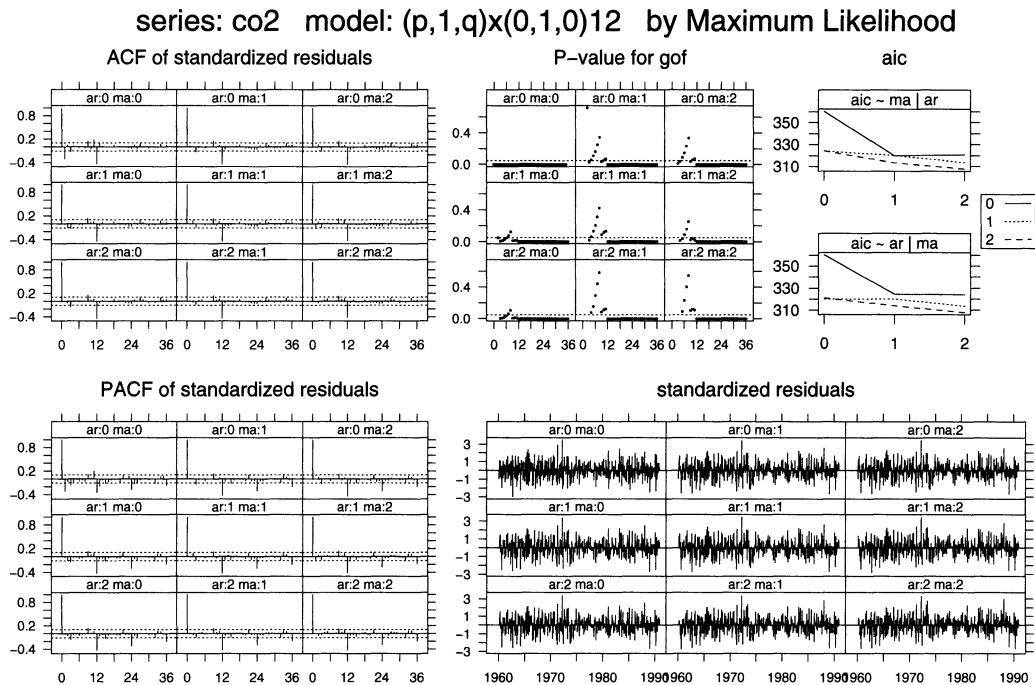


FIGURE 18.7. Diagnostic plots for the set of models  $\text{ARIMA}(p, 1, q) \times (0, 1, 0)_{12}$  fit to the CO<sub>2</sub> data by maximum likelihood. Each set of nine panels is systematically structured in a  $3 \times 3$  array with rows indexed by the number of AR parameters  $p$  and columns by the number of MA parameters  $q$ . All nine panels in a set are scaled identically. The AIC is plotted as a pair of interaction plots: AIC plotted against  $q$  using line types defined by  $p$ ; and AIC plotted against  $p$ , using line types defined by  $q$ . (tser/code/tsamstat.s), (tser/figure/tsamsta3.ps.gz)

the residual ACF plots show a significant spike at  $\text{lag}=12$  months, and all the GOF plots show a break at the same  $\text{lag}=12$  months.

The residual ACF, PACF, and GOF plots in Figure 18.7 clearly confirm that seasonal parameters are necessary. The cutoff after the spike at  $\text{lag}=12$  of the residual ACF, and the exponential decay of the residual PACF at the seasonal lags (those that are multiples of 12 months), show that a seasonal MA parameter is necessary. This agrees with the identification of candidate models (18.11) and (18.13).

Next consider the  $\text{ARIMA}(p, 1, q) \times (0, 1, 1)_{12}$  models with  $0 \leq p, q \leq 2$ . The diagnostic plots for models including the seasonal MA parameter are

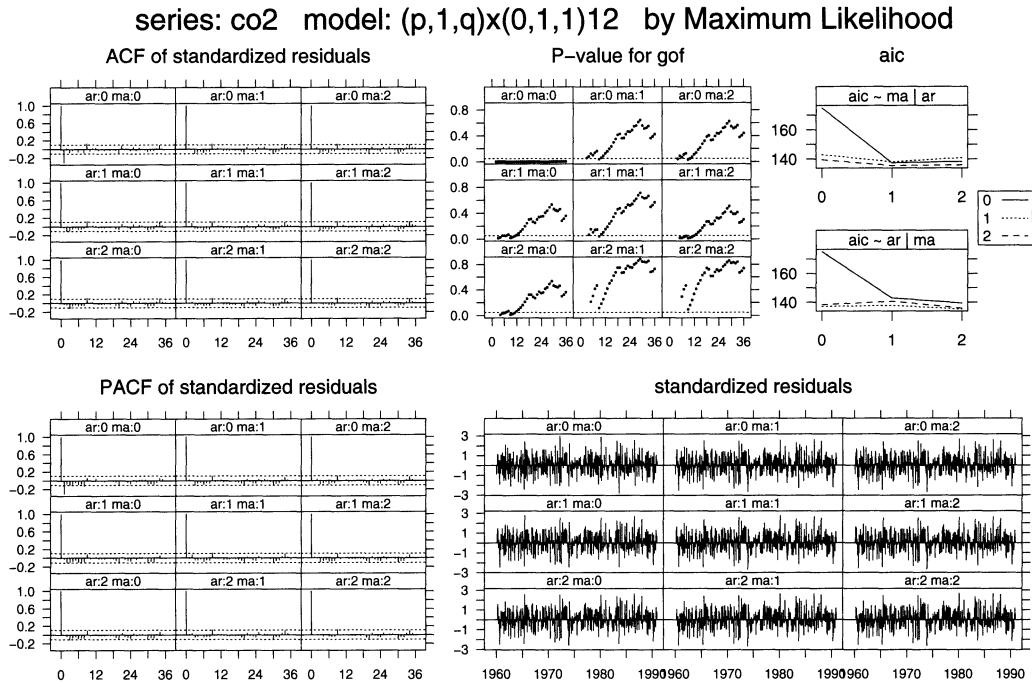


FIGURE 18.8. Diagnostic plots for the set of models  $\text{ARIMA}(p, 1, q) \times (0, 1, 1)_{12}$  fit to the CO<sub>2</sub> data by maximum likelihood.

(tser/code/tsamstat.s), (tser/figure/tsamsta4.ps.gz)

in Figure 18.8. The  $q = 0$  column of the residual ACF and GOF plots shows poor fits. The  $q = 1$  column appears better than the  $q = 2$  column. All three GOF plots for  $q = 1$  are similar. The AIC plots show almost identical values when  $q = 1$ . This is seen as three almost coincident points at  $q = 1$  in the “aic ~ ma | ar” plot and as a horizontal line for  $q = 1$  over all three values of  $p$  in the “aic ~ ar | ma” plot. The conclusion is that one nonseasonal MA parameter is necessary.

Table 18.4 shows the AIC, the estimates of  $\sigma^2_\varepsilon$ , and the estimates of the ARMA parameters with their  $t$ -statistics for the set of  $\text{ARIMA}(p, 1, q) \times (0, 1, 1)_{12}$  models. From the t.coef section of Table 18.4, the  $t$  statistics for both AR parameters in the  $\text{ARIMA}(2, 1, 1) \times (0, 1, 1)_{12}$  model are not significant and this model can be rejected. The  $t$ -statistic for the AR(1) parameter in the  $\text{ARIMA}(1, 1, 1) \times (0, 1, 1)_{12}$  model is marginally signifi-

TABLE 18.4. Estimation results for ARIMA( $p, 1, q) \times (0, 1, 1)_{12}$  models fit to the CO<sub>2</sub> data.

---

```

$model:
[1] "(p,1,q)x(0,1,1)12"

$sigma2:
      0       1       2
0 0.08842 0.08030 0.08010
1 0.08145 0.08003 0.08024
2 0.08035 0.07908 0.07881

$aic:
      0       1       2
0 175.2 137.3 138.3
1 143.0 138.0 140.7
2 139.5 135.5 136.0

$coef:
            ar(1)    ar(2)    ma(1)    ma(2)    ma(12)
(0,1,0)x(0,1,1)12     NA      NA      NA      NA  0.9027
(1,1,0)x(0,1,1)12 -0.30103     NA      NA      NA  0.8691
(2,1,0)x(0,1,1)12 -0.32345 -0.08429     NA      NA  0.8666
(0,1,1)x(0,1,1)12     NA      NA  0.3634     NA  0.8581
(1,1,1)x(0,1,1)12  0.24966     NA  0.5935     NA  0.8573
(2,1,1)x(0,1,1)12  0.27872  0.06425  0.6386     NA  0.8581
(0,1,2)x(0,1,1)12     NA      NA  0.3563  0.05386  0.8563
(1,1,2)x(0,1,1)12 -0.65823     NA -0.3003  0.23691  0.8561
(2,1,2)x(0,1,1)12 -0.08159  0.21781  0.2738  0.27101  0.8585

$t.coef:
            ar(1)    ar(2)    ma(1)    ma(2)    ma(12)
(0,1,0)x(0,1,1)12     NA      NA      NA      NA  40.41
(1,1,0)x(0,1,1)12 -6.07218     NA      NA      NA  33.80
(2,1,0)x(0,1,1)12 -6.23543 -1.6250     NA      NA  33.37
(0,1,1)x(0,1,1)12     NA      NA  7.51267     NA  32.18
(1,1,1)x(0,1,1)12  2.00190     NA  5.72599     NA  32.03
(2,1,1)x(0,1,1)12  1.37981  0.6611  3.31520     NA  32.09
(0,1,2)x(0,1,1)12     NA      NA  6.87252  1.03890  31.94
(1,1,2)x(0,1,1)12 -0.03071     NA -0.01402  0.03095  31.86
(2,1,2)x(0,1,1)12 -0.12107  1.3575  0.40428  0.73630  32.12

```

---

cant, leading us to consider both the models ARIMA(0, 1, 1)  $\times$  (0, 1, 1)<sub>12</sub> and ARIMA(1, 1, 1)  $\times$  (0, 1, 1)<sub>12</sub> for  $X_t$ . A different criterion is needed to distinguish between them. Both models are consistent with the analysis at the identification stage.

The detailed display of the estimation results for the ARIMA(0, 1, 1)  $\times$  (0, 1, 1)<sub>12</sub> model is shown in Table 18.5 (as displayed by the new print method for `arima` objects). A similar display for the ARIMA(1, 1, 1)  $\times$  (0, 1, 1)<sub>12</sub> model (not shown here) shows that the AR(1) and MA(1) parameters are highly correlated ( $r = .91$ ). The ARIMA(1, 1, 1)  $\times$  (0, 1, 1)<sub>12</sub> models can be discarded from further consideration.

The final step is the verification of the adequacy of the ARIMA(0, 1, 1)  $\times$  (0, 1, 1)<sub>12</sub> model. The residual ACF and PACF plots exhibit no significant spikes and all the GOF  $p$ -values are also not significant, showing that the residuals are approximately white noise. The AIC values have dropped from 310 in Figure 18.7 to 136 in Figure 18.8. The standardized residuals in Figure 18.8 are not inconsistent with the normal distribution. The ARIMA(0, 1, 1)  $\times$  (0, 1, 1)<sub>12</sub> model seems to be appropriate for  $X_t$  and the estimated model is

$$\begin{aligned}\nabla_{12} \nabla X_t &= (1 - \hat{\theta}_1 B) (1 - \hat{\Theta}_1 B^{12}) \hat{\varepsilon}_t \\ &= (1 - 0.36338 B)(1 - 0.85806 B^{12}) \hat{\varepsilon}_t \\ \hat{\sigma}_{\varepsilon}^2 &= 0.080299\end{aligned}\tag{18.14}$$

### 18.8.3 Forecasting

The final plot in Figure 18.9 shows the last year of observed data and the forecasts, with their 95% forecast limits, obtained from the fitted model for the following year, i.e., for the months January through December 1991.

## 18.9 Exercises

Many of the time series exercises ask you to construct and/or interpret plots of the time series itself, of the ACF and PACF, and of the diagnostics from a  $3 \times 3$  set of ARIMA models. For Exercises 18.1, 18.2, and 18.3, go through this set of steps:

- Describe the plot of the data and the ACF and PACF plots. Comment on whether you see anything systematic in the plot of the data. Are there spikes in the ACF and PACF plots. At which lags do they appear and what do they suggest? Do the ACF and PACF plots show any indication of a seasonal effect?
- We chose to investigate a family of ARIMA( $p, 0, q$ ) models, with  $0 \leq p, q \leq 2$ . Study the figures showing the diagnostic plots and the tables listing the parameter estimates. Describe each of the four sections of the diagnostic plot. What characteristics of each suggest a final model? Does

TABLE 18.5. Estimation results for ARIMA(0, 1, 1)  $\times$  (0, 1, 1)<sub>12</sub> models fit to the CO<sub>2</sub> data.

---

```
Method: Maximum Likelihood
Model : (0,1,1)x(0,1,1)12

Coefficients:
    ma(1)   ma(12)
coef 0.36338  0.85806
      t 7.51267 32.18304

Variance-Covariance Matrix of Coefficients:
            ma(1)       ma(12)
ma(1)  2.3395e-003 -1.3076e-008
ma(12) -1.3076e-008  7.1086e-004

Correlation Matrix of Coefficients:
            ma(1)       ma(12)
ma(1)  1e+000 -1e-005
ma(12) -1e-005  1e+000

Optimizer has converged
Convergence Type: relative function convergence
AIC: 137.32474
sigma2: 0.0803
```

---

the  $\sigma^2$  (sigma2) section in the tables also suggest the same final model? Note for these three exercises, all of which are stationary and have zero mean, that the (0, 0) panels of the ACF, PACF, and standardized residuals plots are essentially the same as the ACF, PACF, and time series plot of the data.

- c. We printed the detail for the ARIMA(1, 0, 1) model. Compare the ARIMA(1, 0, 1) model to a simpler model with the closest  $\sigma^2$ . How do the AIC and the  $\sigma^2$  compare? Would you recommend the simpler model? Why or why not?

- 18.1.** Figure 18.10 shows the sequence, ACF, and PACF plots for a mystery time series  $X$  from the data file (`datasets/tser.mystery.X.dat`). Figure 18.11 and Table 18.6 show the diagnostics and estimated coefficients

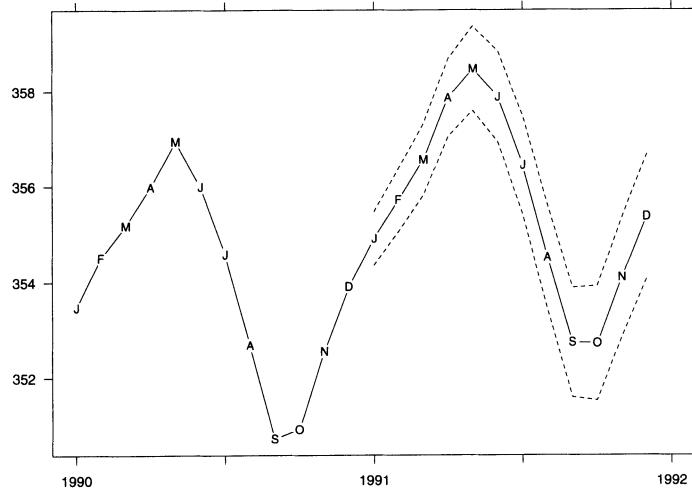


FIGURE 18.9. CO<sub>2</sub>—1990 observed, 1991 forecast + 95% CI.  
(tser/code/tsamstat.s), (tser/figure/tsamsta5.ps.gz)

obtained by fitting the  $3 \times 3$  set of ARIMA( $p, 0, q$ ) models to the series. Table 18.7 shows the detail for the ARIMA(1, 0, 1) model. Study the graphs and tables and explain why and how they indicate that one of these models seems better suited to explain the data than the others.

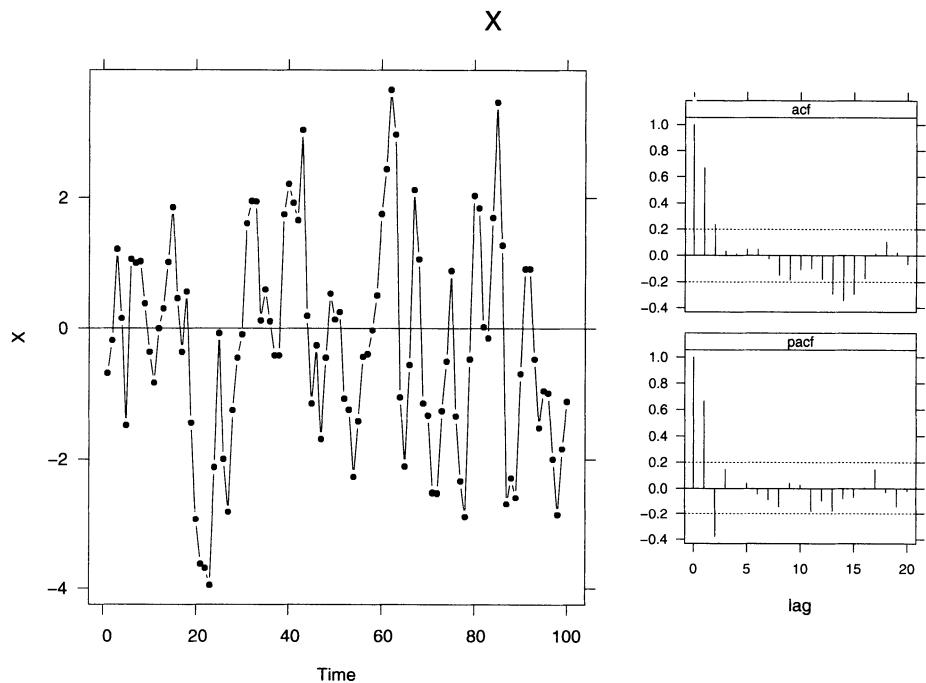


FIGURE 18.10. Mystery time series  $X$ .  
(tser/code/arima.sim.XYZ.s), (tser/figure/arima.sim.X.eps.gz)

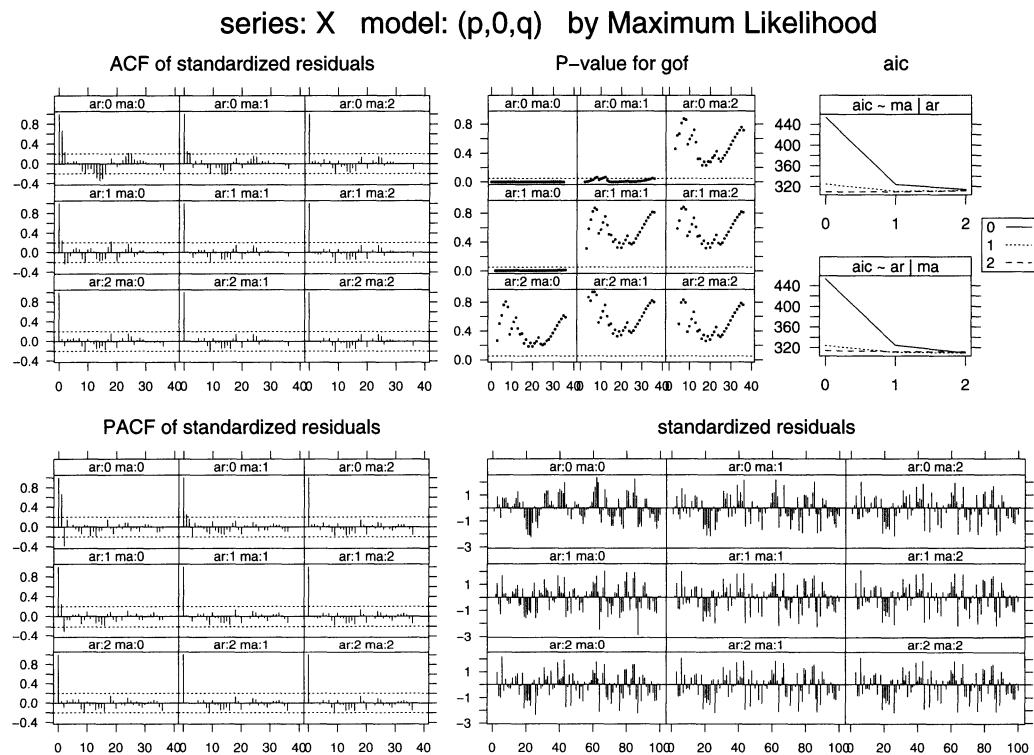


FIGURE 18.11. Mystery time series  $X$ .  
 $(\text{tser}/\text{code}/\text{arima.sim.XYZ.s}), (\text{tser}/\text{figure}/\text{arima.sim.Xd.eps.gz})$

TABLE 18.6. Mystery time series  $X$ .  
 (tser/code/arima.sim.XYZ.s)

---

```

S-PLUS (tser/transcript/arima.sim.X.st):
> X.loop
$series:
[1] "X"

$model:
[1] "(p,0,q)"

$sigma2:
      0         1         2
0 2.742391 1.456903 1.289263
1 1.521076 1.292067 1.284827
2 1.323008 1.294617 1.292542

$aic:
      0         1         2
0 453.4466 324.1757 314.1018
1 324.4722 310.6271 312.1996
2 309.5429 309.5445 311.5073

$coef:
      ar(1)     ar(2)     ma(1)     ma(2)
(0,0,0)    NA        NA        NA        NA
(1,0,0)  0.6768806    NA        NA        NA
(2,0,0)  0.9284674 -0.3741113    NA        NA
(0,0,1)    NA        NA -0.7285542    NA
(1,0,1)  0.4510401    NA -0.5156809    NA
(2,0,1)  0.6387896 -0.1797642 -0.3459361    NA
(0,0,2)    NA        NA -0.9323331 -0.33569277
(1,0,2)  0.3284451    NA -0.6519414 -0.13252817
(2,0,2)  0.4889538 -0.1224599 -0.4954896 -0.09278172

$t.coef:
      ar(1)     ar(2)     ma(1)     ma(2)
(0,0,0)    NA        NA        NA        NA
(1,0,0)  9.1494934    NA        NA        NA
(2,0,0)  9.9110631 -3.9935069    NA        NA
(0,0,1)    NA        NA -10.6360108    NA
(1,0,1)  3.9437041    NA -4.6968908    NA
(2,0,1)  2.7576595 -0.9701832 -1.5287748    NA
(0,0,2)    NA        NA -9.8976768 -3.5637248
(1,0,2)  1.2802168    NA -2.4739029 -0.6294127
(2,0,2)  0.5124973 -0.2725873 -0.5187577 -0.1841019

```

---

TABLE 18.7. Mystery time series  $X$ .  
(tser/code/arima.sim.XYZ.s)

---

```
S-PLUS (tser/transcript/arima.sim.X11.st):
> print.arima(X.loop[["1","1"]])
Call: arima.mle(x = x, model = model[sapply(model, npar.arima) != 0])
Method: Maximum Likelihood
Model : (1,0,1)

Coefficients:
      ar(1)    ma(1)
coef 0.45104 -0.51568
      t 3.94370 -4.69689

Variance-Covariance Matrix of Coefficients:
      ar(1)    ma(1)
ar(1) 0.013080422 0.007790073
ma(1) 0.007790073 0.012054276

Correlation Matrix of Coefficients:
      ar(1)    ma(1)
ar(1) 1.00000 0.62038
ma(1) 0.62038 1.00000

Optimizer has converged
Convergence Type: relative function convergence
AIC: 310.62706
sigma2: 1.29207
```

---

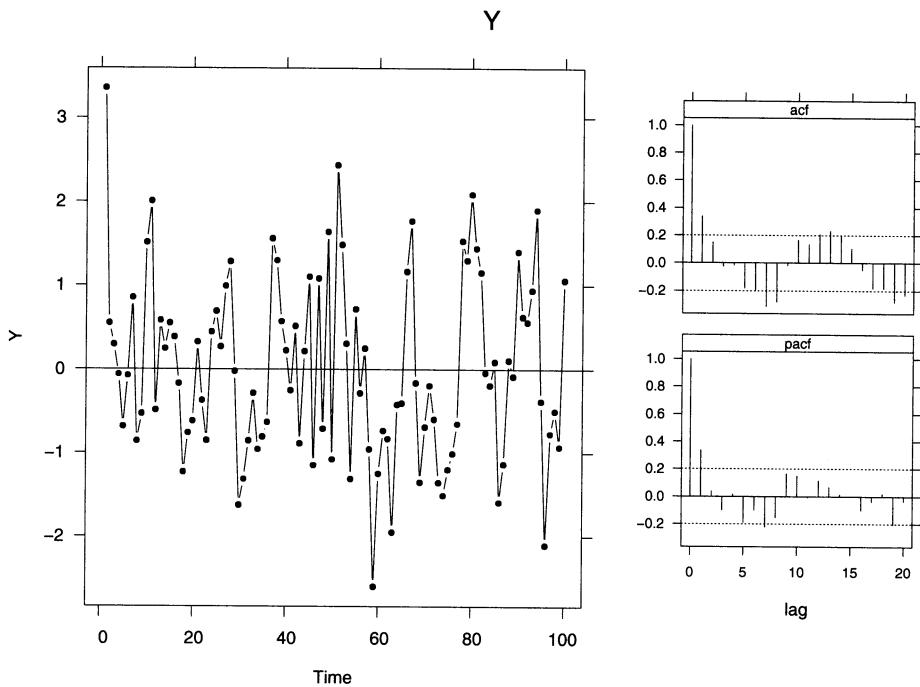


FIGURE 18.12. Mystery time series  $Y$ .  
`(tser/code/arima.sim.XYZ.s)`, `(tser/figure/arima.sim.Y.eps.gz)`

- 18.2.** Figure 18.12 shows the sequence, ACF, and PACF plots for a mystery time series  $Y$  from the data file (`datasets/tser.mystery.Y.dat`). Figure 18.13 and Table 18.8 show the diagnostics and estimated coefficients obtained by fitting the  $3 \times 3$  set of ARIMA( $p, 0, q$ ) models to the series. Table 18.9 shows the detail for the ARIMA(1, 0, 1) model. Study the graphs and tables and explain why and how they indicate that one of these models seems better suited to explain the data than the others.

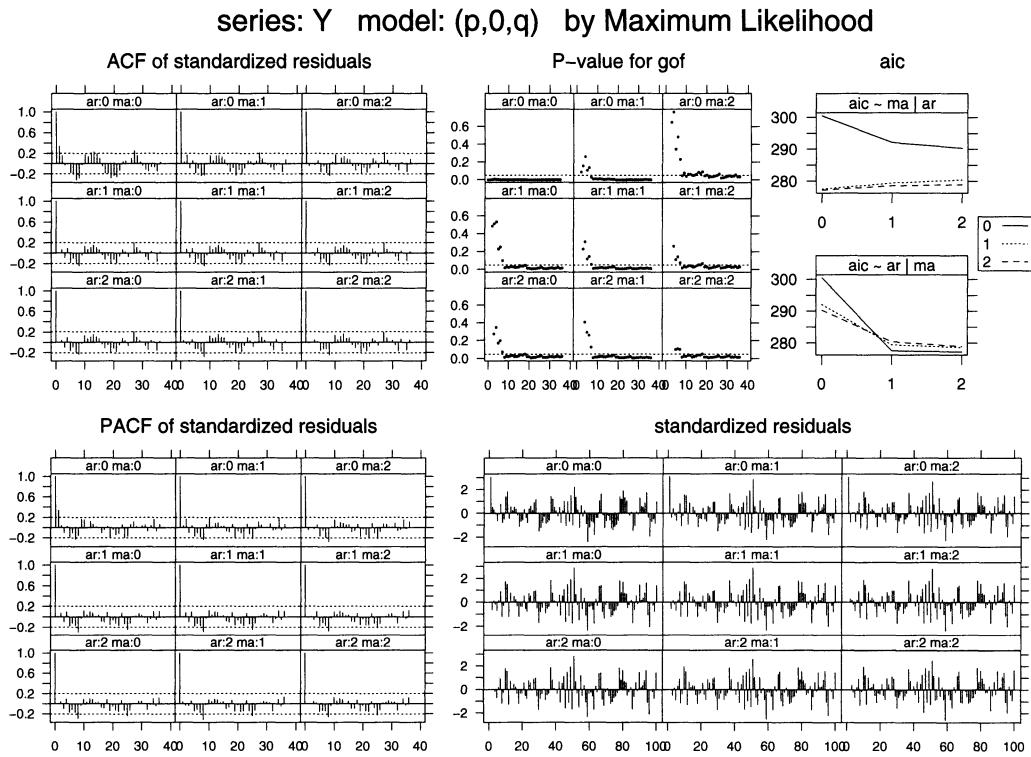


FIGURE 18.13. Mystery time series  $Y$ .  
 $(tser/code/arima.sim.XYZ.s), (tser/figure/arima.sim.Yd.eps.gz)$

TABLE 18.8. Mystery time series  $Y$ .  
 (tser/code/arima.sim.XYZ.s)

---

S-PLUS (tser/transcript/arima.sim.Y.st):

```

> Y.loop
$series:
[1] "Y"

$model:
[1] "(p,0,q)"

$sigma2:
      0         1         2
0 1.1982644 1.0646497 1.0246476
1 0.9462449 0.9460288 0.9345788
2 0.9510754 0.9389170 0.9204628

$aic:
      0         1         2
0 300.5780 292.1444 290.3996
1 277.4797 279.4578 280.3647
2 277.1961 278.6206 278.9613

$coef:
      ar(1)     ar(2)     ma(1)     ma(2)
(0,0,0)    NA        NA        NA        NA
(1,0,0)  0.34131334    NA        NA        NA
(2,0,0)  0.34904067 0.02856166    NA        NA
(0,0,1)    NA        NA -0.29662865    NA
(1,0,1)  0.31857025    NA -0.02755428    NA
(2,0,1) -0.37820395 0.32524879 -0.70434648    NA
(0,0,2)    NA        NA -0.37803116 -0.1850425
(1,0,2)  0.07362672    NA -0.29222867 -0.1516626
(2,0,2) -0.57306605 0.03132960 -0.96070407 -0.3709060

$t.coef:
      ar(1)     ar(2)     ma(1)     ma(2)
(0,0,0)    NA        NA        NA        NA
(1,0,0)  3.6129862    NA        NA        NA
(2,0,0)  3.4567366 0.2828614    NA        NA
(0,0,1)    NA        NA -3.10608228    NA
(1,0,1)  1.1473539    NA -0.09410405    NA
(2,0,1) -0.9063212 2.3948159 -1.63977696    NA
(0,0,2)    NA        NA -3.84674239 -1.882943
(1,0,2)  0.1283272    NA -0.51511621 -0.739265
(2,0,2) -2.0611370 0.1226034 -3.67419539 -1.666354

```

---

TABLE 18.9. Mystery time series  $Y$ .  
(tser/code/arima.sim.XYZ.s)

---

```
S-PLUS (tser/transcript/arima.sim.Y11.st):
> print.arima(Y.loop[["1","1"]])
Call: arima.mle(x = x, model = model[sapply(model, npar.arima) != 0])
Method: Maximum Likelihood
Model : (1,0,1)

Coefficients:
      ar(1)     ma(1)
coef 0.31857 -0.027554
      t 1.14735 -0.094104

Variance-Covariance Matrix of Coefficients:
      ar(1)     ma(1)
ar(1) 0.07709312 0.07636426
ma(1) 0.07636426 0.08573564

Correlation Matrix of Coefficients:
      ar(1)     ma(1)
ar(1) 1.00000 0.93929
ma(1) 0.93929 1.00000

Optimizer has converged
Convergence Type: relative function convergence
AIC: 279.45784
sigma2: 0.94603
```

---

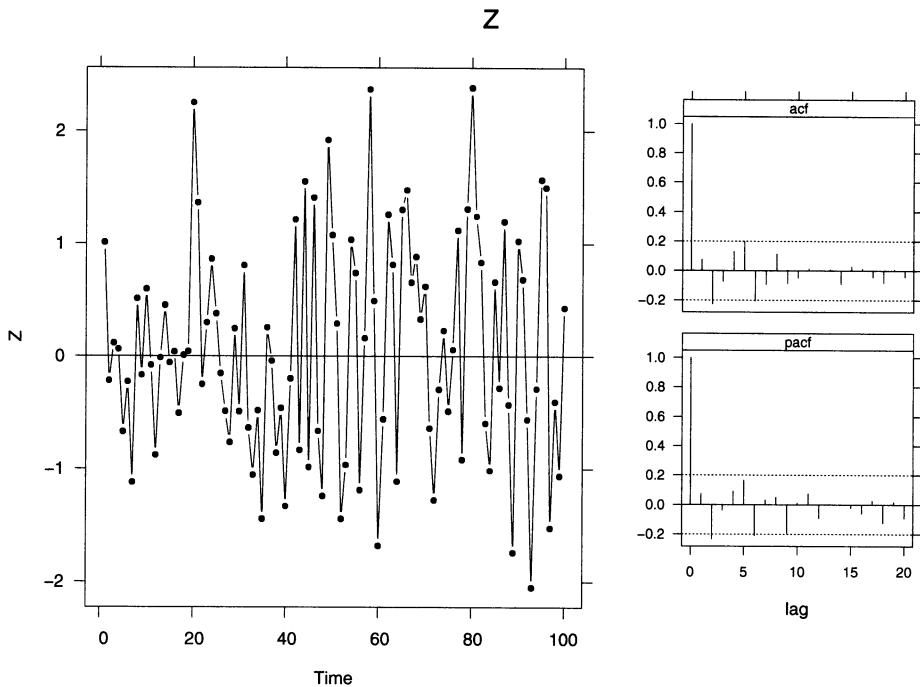


FIGURE 18.14. Mystery time series  $Z$ .  
`(tser/code/arima.sim.XYZ.s)`, `(tser/figure/arima.sim.Z.eps.gz)`

- 18.3.** Figure 18.14 shows the sequence, ACF, and PACF plots for a mystery time series  $Z$  from data file (`datasets/tser.mystery.Z.dat`). Figure 18.15 and Table 18.10 show the diagnostics and estimated coefficients obtained by fitting the  $3 \times 3$  set of ARIMA( $p, 0, q$ ) models to the series. Table 18.11 shows the detail for the ARIMA(1, 0, 1) model. Study the graphs and tables and explain why and how they indicate that one of these models seems better suited to explain the data than the others.

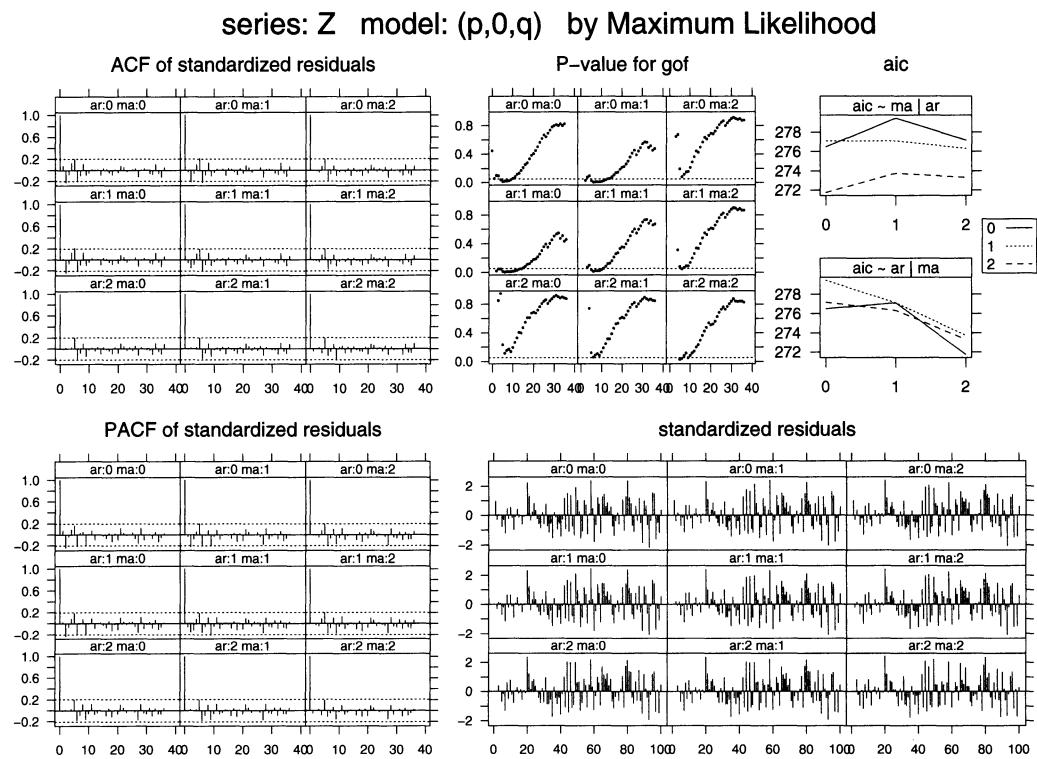


FIGURE 18.15. Mystery time series  $Z$ .  
 (tser/code/arima.sim.XYZ.s), (tser/figure/arima.sim.Zd.eps.gz)

TABLE 18.10. Mystery time series  $Z$ .  
 (tser/code/arima.sim.XYZ.s)

---

```

S-PLUS (tser/transcript/arima.sim.Z.st):
> Z.loop
$series:
[1] "Z"

$model:
[1] "(p,0,q)"

$sigma2:
      0         1         2
0 0.9549937 0.9386643 0.8983640
1 0.9426700 0.9198630 0.8972181
2 0.8997078 0.8994223 0.8696178

$aic:
      0         1         2
0 276.4942 279.4771 277.1765
1 277.1050 277.1089 276.3165
2 271.7548 273.7262 273.3261

$coef:
      ar(1)     ar(2)     ma(1)     ma(2)
(0,0,0)    NA        NA        NA        NA
(1,0,0)  0.079238355    NA        NA        NA
(2,0,0)  0.102257295 -0.2334210    NA        NA
(0,0,1)    NA        NA -0.13755216    NA
(1,0,1) -0.420271774    NA -0.59038930    NA
(2,0,1)  0.149600648 -0.2378387  0.05027415    NA
(0,0,2)    NA        NA -0.10003125  0.2104630
(1,0,2) -0.002689034    NA -0.10338262  0.2060427
(2,0,2)  0.003236101 -0.7631525 -0.13353451 -0.6014245

$t.coef:
      ar(1)     ar(2)     ma(1)     ma(2)
(0,0,0)    NA        NA        NA        NA
(1,0,0)  0.790898506    NA        NA        NA
(2,0,0)  1.041053924 -2.376397    NA        NA
(0,0,1)    NA        NA -1.3887221    NA
(1,0,1) -1.042683908    NA -1.6467266    NA
(2,0,1)  0.357923791 -2.325579  0.1168396    NA
(0,0,2)    NA        NA -1.0232310  2.152849
(1,0,2) -0.005521595    NA -0.2169380  1.792656
(2,0,2)  0.016579343 -4.556603 -0.5670478 -2.767573

```

---

TABLE 18.11. Mystery time series  $Z$ .  
(tser/code/arima.sim.XYZ.s)

---

```
S-PLUS (tser/transcript/arima.sim.Z11.st):
> print.arima(Z.loop[["1","1"]])
Call: arima.mle(x = x, model = model[sapply(model, npar.arima) != 0])
Method: Maximum Likelihood
Model : (1,0,1)

Coefficients:
      ar(1)     ma(1)
coef -0.42027 -0.59039
      t -1.04268 -1.64673

Variance-Covariance Matrix of Coefficients:
      ar(1)     ma(1)
ar(1) 0.1624632 0.1407614
ma(1) 0.1407614 0.1285387

Correlation Matrix of Coefficients:
      ar(1)     ma(1)
ar(1) 1.00000 0.97407
ma(1) 0.97407 1.00000

Optimizer has converged
Convergence Type: relative function convergence
AIC: 277.10888
sigma2: 0.91986
```

---

- 18.4.** The tables in file (`tser/transcript/product.st`) were produced during the analysis of the product data (`datasets/product.dat`) of Section 18.6. The AR(1) model converged with AIC=2918, a larger number than for nonconverging models with more terms. The nonconverging models showed high correlation between the estimates of the AR and MA coefficients. The peaks in the ACF and PACF plots of Figure 18.2 at 4 and 8 weeks suggest that there might be a monthly effect in this data. Examine the set of  $\text{ARIMA}(p, 1, q) \times (1, 0, 0)_4$  models for this data.
- 18.5.** Figures 18.16 and 18.17 and Tables 18.12 and 18.13 show the mean monthly air temperature in degrees Fahrenheit from January 1920–December 1939 at Nottingham Castle. The data are in the file (`datasets/nottem.dat`). S-PLUS and R users can use the `nottem` data in `library(MASS)`. We got the data from (Venables and Ripley, 1997). The original source is “Meteorology of Nottingham” in *City Engineer and Surveyor*. We show the original series, the seasonally differenced series, the diagnostic display from the series of models  $\text{ARIMA}(p, 0, q) \times (2, 1, 0)_{12}$ , and numerical results from the set of all nine models table and detail on the recommended model  $\text{ARIMA}(1, 0, 0) \times (2, 1, 0)_{12}$ .
- a. What are the most evident features of the plot of the original data?
  - b. Compare the plot of the seasonally differenced data to the original plot. What structure was captured by the differencing? What remains?
  - c. Compare the recommended model  $\text{ARIMA}(1, 0, 0) \times (2, 1, 0)_{12}$  to the next most likely model  $\text{ARIMA}(2, 0, 0) \times (2, 1, 0)_{12}$ . Do you agree that the `ar(2)` term is not needed? Why?

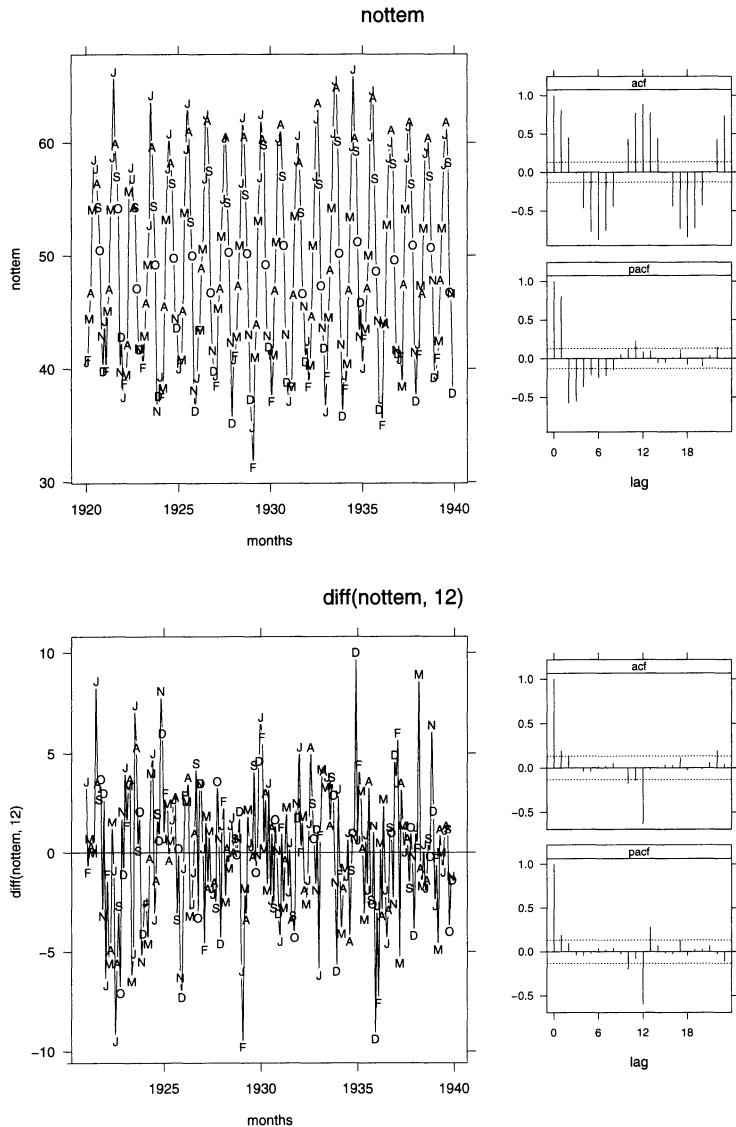


FIGURE 18.16. Mean monthly air temperature in degrees Fahrenheit from January 1920–December 1939 at Nottingham Castle. Top shows the original data, bottom shows the seasonal differences.  
`(tser/code/nottem.s), (tser/figure/nottem.a.ps.gz), (tser/figure/nottemb.ps.gz)`

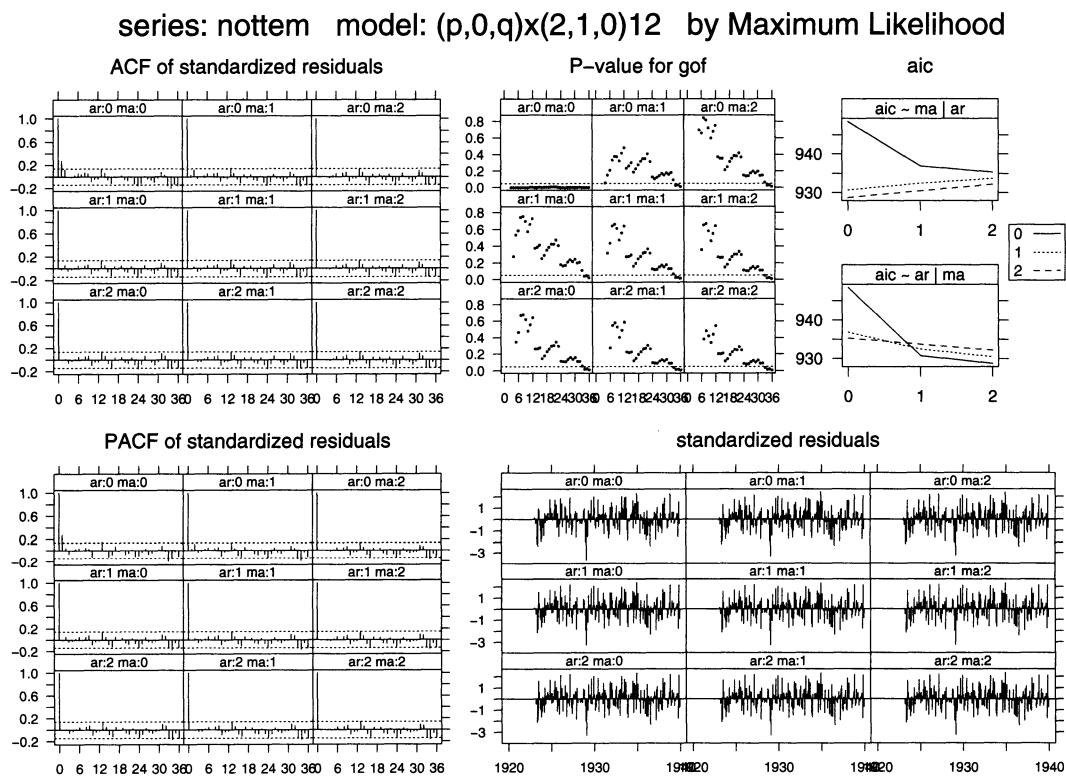


FIGURE 18.17. Mean monthly air temperature in degrees Fahrenheit from January 1920–December 1939 at Nottingham Castle.

(tser/code/nottem.s), (tser/figure/nottemc.ps.gz)

TABLE 18.12. Nottingham temperature—models ARIMA( $p, 0, q) \times (2, 1, 0)_{12}$ .  
 (tser/code/nottem.s)

---

```

S-PLUS (tser/transcript/nottem-3x3.st):
> nottem.loop <- arma.loop(nottem, list(list(order=c(2,0,2)),
+                                         list(order=c(2,1,0), period=12)))
> nottem.loop
$series:
[1] "nottem"

$model:
[1] "(p,0,q)x(2,1,0)12"

$sigma2:
      0     1     2
0 6.001 5.612 5.515
1 5.571 5.562 5.542
2 5.586 5.574 5.568

$aic:
      0     1     2
0 948.5 936.9 935.3
1 930.7 932.4 933.8
2 928.7 930.5 932.1

$coef:
          ar(1)    ar(2)    ma(1)    ma(2)   ar(12)   ar(24)
(2,1,0)12      NA      NA      NA      NA -0.7908 -0.2828
(1,0,0)x(2,1,0)12 0.2809      NA      NA      NA -0.8317 -0.2879
(2,0,0)x(2,1,0)12 0.2671  0.04699      NA      NA -0.8337 -0.2949
(0,0,1)x(2,1,0)12      NA      NA -0.2391      NA -0.8202 -0.2770
(1,0,1)x(2,1,0)12  0.3859      NA  0.1136      NA -0.8332 -0.2925
(2,0,1)x(2,1,0)12 -0.1403  0.17426 -0.4072      NA -0.8370 -0.2991
(0,0,2)x(2,1,0)12      NA      NA -0.2659 -0.1393 -0.8339 -0.3010
(1,0,2)x(2,1,0)12  0.0616      NA -0.2064 -0.1246 -0.8347 -0.3010
(2,0,2)x(2,1,0)12  0.1273 -0.04600 -0.1438 -0.1483 -0.8360 -0.2990

$t.coef:
          ar(1)    ar(2)    ma(1)    ma(2)   ar(12)   ar(24)
(2,1,0)12      NA      NA      NA      NA -11.78 -4.212
(1,0,0)x(2,1,0)12 4.1702      NA      NA      NA -12.37 -4.283
(2,0,0)x(2,1,0)12 3.8010  0.66861      NA      NA -12.40 -4.386
(0,0,1)x(2,1,0)12      NA      NA -3.5170      NA -12.19 -4.117
(1,0,1)x(2,1,0)12  1.6972      NA  0.4640      NA -12.41 -4.358
(2,0,1)x(2,1,0)12 -0.1581  0.73833 -0.4543      NA -12.47 -4.450
(0,0,2)x(2,1,0)12      NA      NA -3.8350 -2.0091 -12.49 -4.508
(1,0,2)x(2,1,0)12  0.1225      NA -0.4134 -0.8991 -12.47 -4.496
(2,0,2)x(2,1,0)12  0.1102 -0.08296 -0.1247 -0.4916 -12.45 -4.449

```

---

TABLE 18.13. Nottingham temperature—recommended model ARIMA(1, 0, 0)  $\times$  (2, 1, 0)<sub>12</sub>.  
 (tser/code/nottem.s)

---

```
S-PLUS (tser/transcript/nottem-100x210.st):
> print.arima(nottem.loop[["1","0"]])
series: nottem    model: (1,0,0)x(2,1,0)12   by Maximum Likelihood

ar(1)      ar(12)     ar(24)
coef 0.28091 -0.8317 -0.28788
t 4.17024 -12.3738 -4.28296

var.coef
ar(1)      ar(12)     ar(24)
ar(1) 0.00454 0.00000 0.00000
ar(12) 0.00000 0.00452 0.00292
ar(24) 0.00000 0.00292 0.00452

corr.coef
ar(1)      ar(12)     ar(24)
ar(1)      1 0.00000 0.00000
ar(12)      0 1.00000 0.64579
ar(24)      0 0.64579 1.00000

aic loglik sigma2 n.used n.cond
930.74 924.74 5.5705     203     37

converged           conv.type
TRUE      relative function convergence
```

---

- 18.6.** We have a time series of size  $n = 100$  for which we have determined that we have an ARIMA(1,0,0) model and have estimated  $\hat{\mu} = 15$ ,  $\hat{\phi} = .2$ , and  $\hat{\sigma}^2 = 3$ . The last few observations in the series are

| $t$   | 97 | 98 | 99 | 100 |
|-------|----|----|----|-----|
| $X_t$ | 13 | 15 | 18 | 17  |

Forecast, with 95% forecast intervals, the values  $\hat{X}_{101}$  and  $\hat{X}_{102}$ .

- 18.7.** We have a nonseasonal time series in the data file (`datasets/tsq.dat`) covering 100 periods. The time series and its ACF and PACF plots are displayed in Figure 18.18. Table 18.14 contains the S-PLUS output from a  $3 \times 3$  set of ARIMA models fit to the data. The `tsdiagplot` for these data is in Figure 18.19. Use this information to answer the following questions:

- a. Recommend the  $(p, 0, q)$  order for an ARIMA modeling of these data.
- b. Write out the equation for the best-fitting model following your recommendation in part (a).
- c. Use your model and the coefficient information in the S output to produce forecasts and 95% forecast intervals for the value of this series in periods 101 and 102.

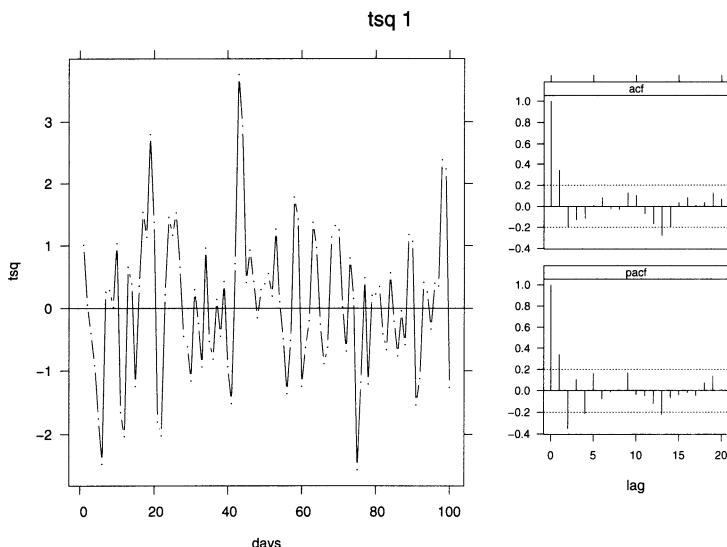


FIGURE 18.18. Time series and its ACF and PACF plots for Exercise 18.7.  
(tser/code/tsq.s), (tser/figure/tsq1.ps.gz)

TABLE 18.14. Three by three set of ARIMA models for Exercise 18.7. Listing is continued in Table 18.14a.

(tsr/code/tsq.s)

---

S-PLUS (tsr/transcript/tsqa.st):

```
> tsq[96:100] ## last 5 of n=100 observations
96: 0.4011052 0.3557547 2.3804020 2.2301364 -1.2674070
start deltat frequency
    96      1      1
>
> tsq.plot <- tsacfplots(tsq, main="tsq 1")
> tsq.plot
>
> tmp <- tsq.plot$acf.plots
> acfs <- t(matrix(tmp$y, ncol=2,
+                     dimnames=list(matrix(tmp$x, ncol=2)[,1], c("acf","pacf"))))
> round(acfs[,1:6], digits=3)
      0      1      2      3      4      5
acf 1 0.341 -0.195 -0.130 -0.116 0.008
pacf 1 0.341 -0.352  0.103 -0.214 0.163
>
> tsq.loop <- arma.loop(tsq, list(list(order=c(2,0,2))))
> tsq.loop
$series:
[1] "tsq"

$model:
[1] "(p,0,q)"

$sigma2:
      0      1      2
0 1.343776 0.9494927 0.9306199
1 1.178872 0.9326204 0.9189261
2 1.020505 0.9287361 0.9110090

$aic:
      0      1      2
0 314.9836 281.4417 281.3979
1 299.2411 279.0785 279.2216
2 284.1012 277.6789 277.9802
```

---

TABLE 18.14a. Continuation of Table 18.14.  
 (tser/code/tsq.s)

---

S-PLUS (tser/transcript/tsqb.st):

\$coef:

|         | ar(1)       | ar(2)       | ma(1)       | ma(2)     |
|---------|-------------|-------------|-------------|-----------|
| (0,0,0) | NA          | NA          | NA          | NA        |
| (1,0,0) | 0.34947642  | NA          | NA          | NA        |
| (2,0,0) | 0.49767579  | -0.39041383 | NA          | NA        |
| (0,0,1) | NA          | NA          | -0.75291230 | NA        |
| (1,0,1) | -0.13903845 | NA          | -0.80289404 | NA        |
| (2,0,1) | -0.09824895 | -0.15152875 | -0.74570660 | NA        |
| (0,0,2) | NA          | NA          | -0.61136452 | 0.1583023 |
| (1,0,2) | 0.36046254  | NA          | -0.24673359 | 0.4386951 |
| (2,0,2) | 0.65250183  | -0.01139369 | 0.01774904  | 0.6272346 |

\$t.coef:

|         | ar(1)      | ar(2)       | ma(1)        | ma(2)    |
|---------|------------|-------------|--------------|----------|
| (0,0,0) | NA         | NA          | NA           | NA       |
| (1,0,0) | 3.7112594  | NA          | NA           | NA       |
| (2,0,0) | 5.3514317  | -4.19806023 | NA           | NA       |
| (0,0,1) | NA         | NA          | -11.44033476 | NA       |
| (1,0,1) | -1.0439332 | NA          | -10.01433489 | NA       |
| (2,0,1) | -0.7055301 | -1.25164848 | -6.95186944  | NA       |
| (0,0,2) | NA         | NA          | -6.19171852  | 1.603239 |
| (1,0,2) | 0.8190730  | NA          | -0.60317999  | 1.513004 |
| (2,0,2) | 2.6282710  | -0.07196816 | 0.07800037   | 3.014682 |

---

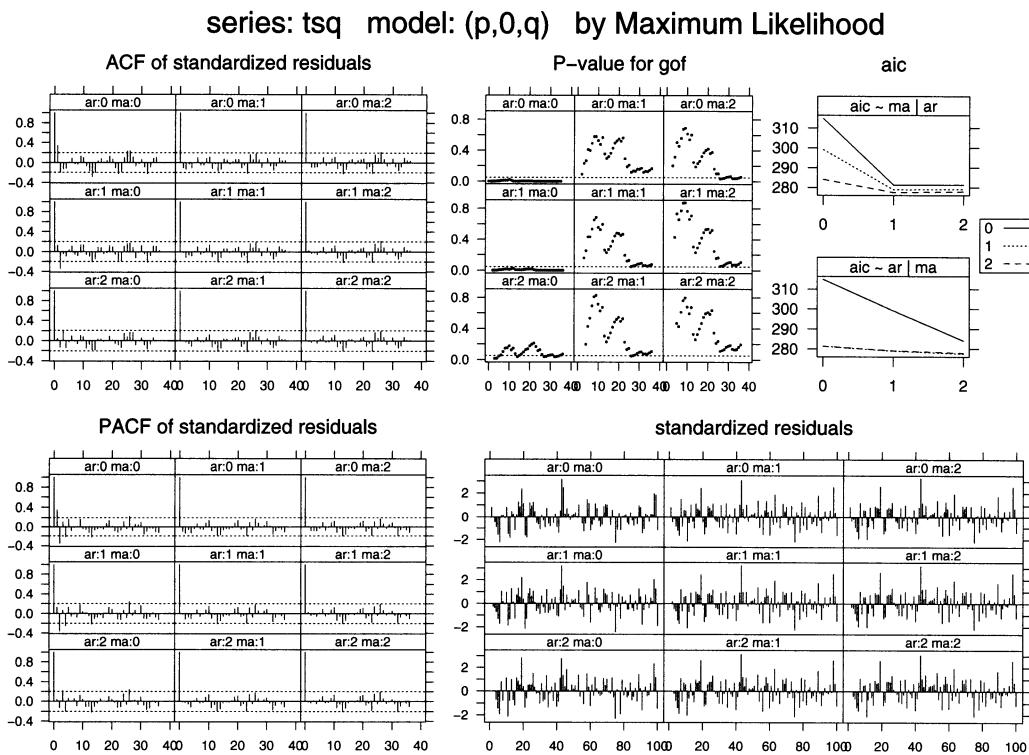


FIGURE 18.19. Diagnostic plots for the  $3 \times 3$  set of ARIMA models in Exercise 18.7.  
`(tser/code/tsq.s), (tser/figure/tsq2.ps.gz)`

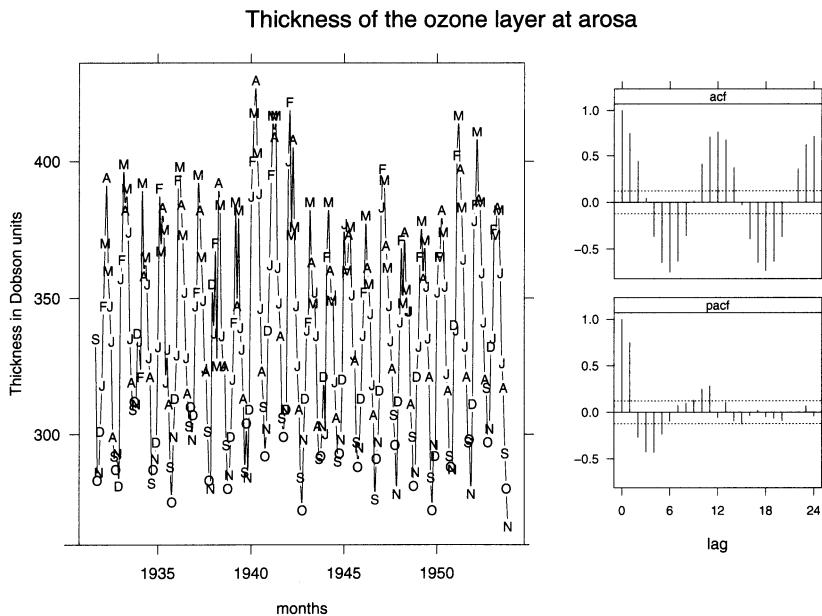


FIGURE 18.20. Thickness of the ozone layer (measured in Dobson units) at Arosa, Switzerland, from September 1931 through November 1953.

(tser/code/ozone.s), (tser/figure/ex0304-4.ps.gz)

**18.8.** Figure 18.20 contains time series, ACF, and PACF plots for monthly data on the thickness of the ozone layer (measured in Dobson units) at Arosa, Switzerland, from September 1931 through November 1953. Note the labeling of the months of the year (J = January, June, or July, F = February, etc.) at the plot points. The data in the file (`datasets/ozone.dat`) are from (Andrews and Herzberg, 1985), Table 12.1.

- Comment on the seasonal nature of the time series plot and discuss how this is consistent with what you see in the ACF plot.
- Notice that the variability of the series appears to increase, at least temporarily, in the early 1940s and around 1952 to 1953. For each of these periods, *identify historical events* that potentially impacted on the atmosphere to produce this increased variability.

|               | lag |      |      |       |      |       |      |      |      |      |
|---------------|-----|------|------|-------|------|-------|------|------|------|------|
|               | 1   | 2    | 3    | 4     | 5    | 6     | 7    | 8    | 9    | 10   |
| acf( $Z_t$ )  | 0.8 | 0.61 | 0.47 | 0.40  | 0.31 | 0.21  | 0.18 | 0.11 | 0.06 | 0.01 |
| pacf( $Z_t$ ) | 0.8 | 0.08 | 0.00 | -0.11 | 0.00 | -0.12 | 0.07 | 0.05 | 0.01 | 0.02 |

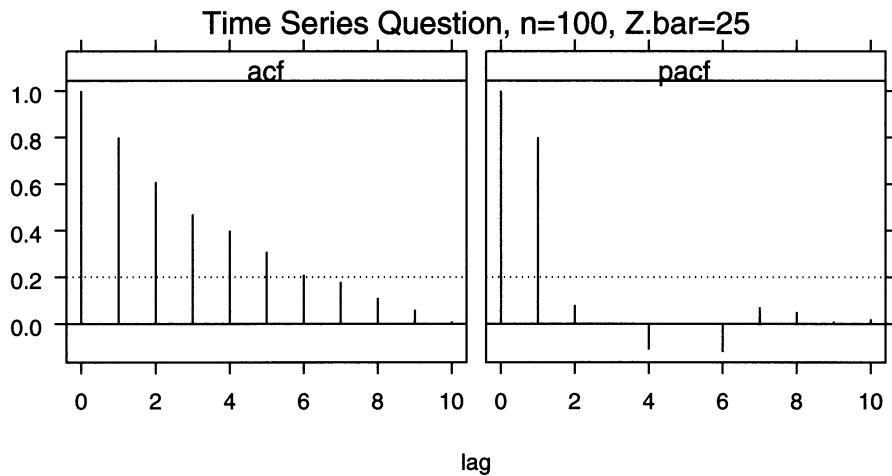


FIGURE 18.21. ACF and PACF for Exercise 18.9,  $n = 100$  and  $\bar{Z} = 25$ . The same information is presented in both tabular and graphical form.  
 (tser/code/co697ts.s), (tser/figure/co697ts1.eps.gz)

**18.9.**  $n = 100$  and  $\bar{Z} = 25$ . See Figure 18.21.

- a. Identify a tentative underlying model in explicit form and justify your model.
- b. Propose possible preliminary parameter estimates for your model.
- c. Assume the residual sum of squares from the fitting of your model is 256, and  $Z_{98} = 24$ ,  $Z_{99} = 26$ ,  $Z_{100} = 25$ . Compute your forecasts for  $Z_{101}$  and  $Z_{102}$  and their 95% forecast intervals.

|                          | lag   |       |      |      |       |      |      |       |      |      |
|--------------------------|-------|-------|------|------|-------|------|------|-------|------|------|
|                          | 1     | 2     | 3    | 4    | 5     | 6    | 7    | 8     | 9    | 10   |
| $\text{acf}(Z_t)$        | 0.93  | 0.92  | 0.90 | 0.90 | 0.87  | 0.86 | 0.85 | 0.84  | 0.82 | 0.80 |
| $\text{acf}(\nabla Z_t)$ | -0.57 | -0.10 | 0.12 | 0.06 | -0.12 | 0.09 | 0.05 | -0.01 | 0.02 | 0.03 |

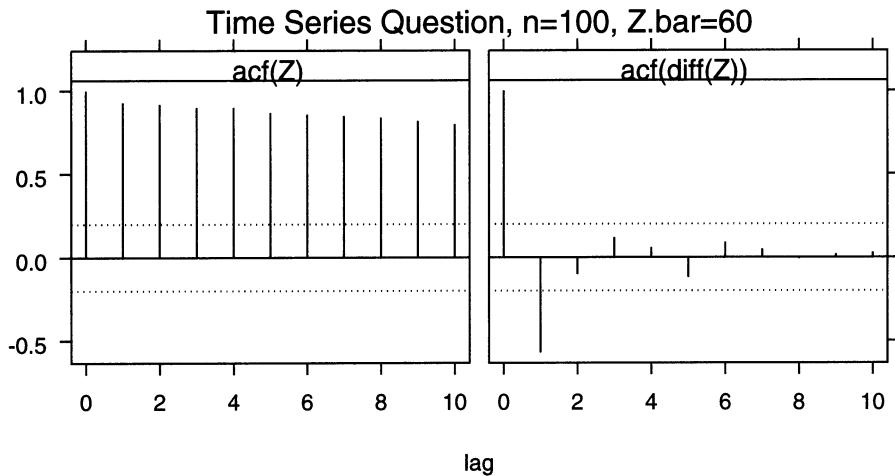


FIGURE 18.22. ACF and PACF for Exercise 18.10,  $n = 100$  and  $\bar{Z} = 60$ . The same information is presented in both tabular and graphical form.

(tser/code/co697ts.s), (tser/figure/co697ts2.eps.gz)

- 18.10.**  $n = 100$  and  $\bar{Z} = 60$ . Identify a tentative underlying model in explicit form and justify your model. See Figure 18.22.

|                              | lag  |      |      |      |      |      |       |       |      |      |
|------------------------------|------|------|------|------|------|------|-------|-------|------|------|
|                              | 1    | 2    | 3    | 4    | 5    | 6    | 7     | 8     | 9    | 10   |
| acf( $Z_t$ )                 | 0.99 | 0.94 | 0.87 | 0.81 | 0.75 | 0.65 | 0.55  | 0.53  | 0.43 | 0.40 |
| acf( $\nabla Z_t$ )          | 0.43 | 0.28 | 0.51 | 0.80 | 0.65 | 0.44 | 0.31  | 0.77  | 0.30 | 0.20 |
| acf( $\nabla_4 Z_t$ )        | 0.72 | 0.67 | 0.55 | 0.32 | 0.38 | 0.23 | 0.24  | 0.23  | 0.18 | 0.13 |
| acf( $\nabla \nabla_4 Z_t$ ) | 0.30 | 0.07 | 0.32 | 0.50 | 0.20 | 0.01 | -0.05 | -0.01 | 0.02 | 0.03 |

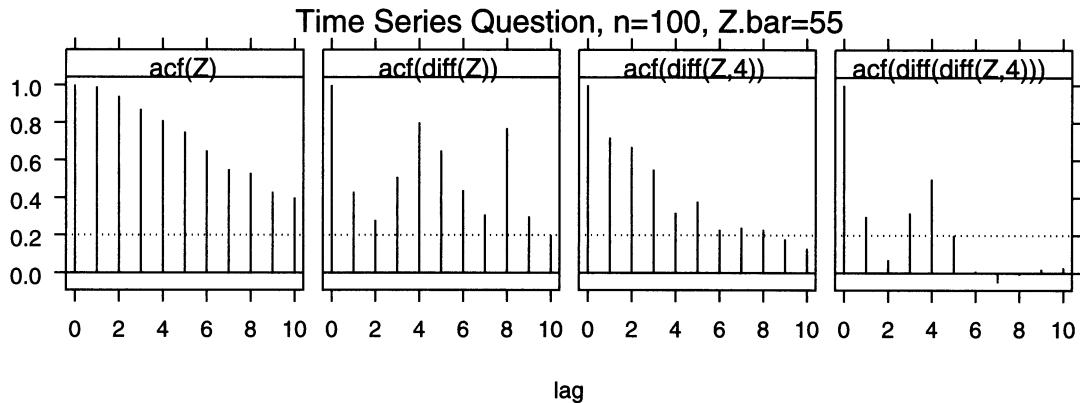


FIGURE 18.23. ACF and PACF for Exercise 18.11,  $n = 100$  and  $\bar{Z} = 55$ . The same information is presented in both tabular and graphical form.

(tser/code/co697ts.s), (tser/figure/co697ts3.eps.gz)

**18.11.**  $n = 100$  and  $\bar{Z} = 55$ . Identify a tentative underlying model in explicit form and justify your model. See Figure 18.23.

- 18.12.** Time series data differs from any other data type we have discussed in one important characteristic: The observations are not independent. What are the implications of that difference for modeling time series data? Be sure to discuss implications for each of
- a. Modeling
  - b. Estimation
  - c. Prediction
- 18.13.** Figure 18.24 shows the “United States of America Monthly Employment Figures for Males Aged 16–19 Years from 1948–1981”. The data in the file (`datasets/employM16.dat`) are Table T.65.1 from (Andrews and Herzberg, 1985). What are the features of this plot that you would try to capture in a time series model? Comment on
- a. Seasonality
  - b. Trend
  - c. Aberrations

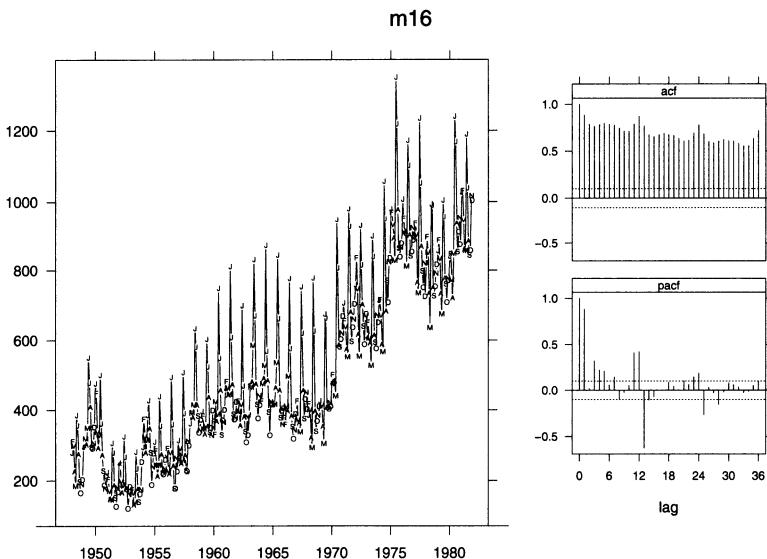


FIGURE 18.24. United States of America Monthly Employment Figures for Males Aged 16-19 Years from 1948-1981.  
 (`tser/code/employM16.s`), (`tser/figure/employM16.eps.gz`)

## 18.A Appendix: Graphical Displays for Time Series Analysis

This section discusses the technical aspects of the construction of the set of plots used to check the validity of the proposed model. The interpretation of the plots, and the discussion of how to use them to help identify the model that best fits the data, appear in Sections 18.6 and 18.8.

The graphical display techniques demonstrated in Sections 18.6 and 18.8 were developed by (Heiberger and Teles, 2002). The S-PLUS functions used to produce these displays are in file (`splus.library/ARIMA-trellis.s`) in the book's downloaded online files and in the HH library described in Appendix B.

The set of plots in Figure 18.7 consists of the residual ACF and PACF, the portmanteau goodness-of-fit test statistic (GOF), the standardized residuals, and the Akaike information criterion (AIC). The panels in the first four sets of plots are indexed by the number of nonseasonal ARMA parameters  $p$  and  $q$  for fixed values of the seasonal parameters  $P$  and  $Q$ . The AIC plot uses  $p$  and  $q$  as plotting variables. The orders of differencing and the orders of the autoregressive and moving average operators (both seasonal and nonseasonal) have been limited to  $0 \leq p, d, q, P, D, Q \leq 2$ . While this limitation is usually reasonable in practice, it is not inherent in the software. AS Each set of nine panels is systematically structured in a  $3 \times 3$  array indexed by the number of AR parameters and MA parameters. All nine panels in a set are scaled identically. Thus the reader can scan a row or column of the array of panels and see the effect of adding one more parameter to either the AR or MA side of the model.

Traditionally (that is, as constructed by the standard S-PLUS `arima.diag`), the plots coordinated in Figure 18.7 are shown on nine separate pages, one page for each model. The standard display shows the standardized residuals, the residual ACF and PACF plots, and the portmanteau goodness-of-fit test. The nine sets of plots, each associated with a different model, will not necessarily be scaled alike. Even the GOF and ACF/PACF plots for the same model may have different lag scales.

Labeling the axis in months and putting the residual ACF and PACF plots and the GOF plot on the same set of lags make it easy to compare the plots for different models. In this example it is easy to see that something is happening at `lag=12` months. The AIC plots for all the models in Figure 18.7 are similar, with  $AIC \approx 315$ . The AIC has been plotted as a pair of interaction plots: AIC plotted against  $q$ , the number of nonseasonal MA parameters, using line types defined by  $p$ , the number of nonseasonal AR parameters; and AIC plotted against  $p$ , using line types defined by  $q$ . These plots enable us to study the magnitudes of the differences in AIC of competing models.

### 18.A.1 Characteristics of This Presentation of the Time Series Plot

- Individual points are identified with a letter indicating the position of each observation according to the frequency of collection of the data. The user can control the choice of plotting characters. The default characters used are dependent on the frequency of collection of the data. For example, when the frequency is 12 and units are not explicitly defined, the default plotting characters are the letters A,B,C,D,E,F,G,H,I,J,K,L.

In this example, the data were collected monthly (`frequency=12`) and the units are explicitly defined as `months` so the plotting characters are the month abbreviations J,F,M,A,M,J,J,A,S,O,N,D.

When `frequency=7` and `units=days`, the characters are the initial letters of the days of the week S,M,T,W,T,F,S.

- The plotting characters are an explicit argument and can be chosen by the user (with `pch.seq`), or suppressed entirely with `type="l"`.
- Color is often very helpful with the time series plots. Color versions of Figures 18.4 to 18.6 are included on the in the online files. The color plots show the seasonal pattern more strongly than the black and white. The color version of Figure 18.4 has a clear pattern of blue-May and orange-April along its top and green-October along the bottom. The first differences in the color version of Figure 18.5 show a clear pattern of green-August along the bottom and color-coded June–July–August–September below the axis. The color version of Figure 18.6 shows random behavior in colors.

### 18.A.2 Characteristics of This Presentation of the Sample ACF and PACF Plots

- The axes are coordinated and have the same scale.
- Lags are indicated in appropriate units (for example, months for monthly series).
- The ACF and PACF plots consistently both show, or do not show (at the user's option), the spike for correlation=1 at `lag=0`.
- The default tick marks are related to the frequency of collection of the data. The user has control over tick mark location.
- Most of the plotting surface is occupied by the body of the plot, and the amount of surface used for labeling is minimized.

We point out that the individual plots are accessible to the user. They can be placed on their own pages or displayed with other relative

spacings. For details, see the function `print.tsacfplots` in the file (`splus.library/ARIMA-trellis.s`).

### 18.A.3 Construction of Graphical Displays

This section shows how to construct the two display types presented in this chapter. For brevity, only Figures 18.6 and 18.8 are described.

Figure 18.6, a single display with subgraphs, is constructed with the single command:

```
tsacfplots(diff(diff(co2,1), 12))
```

The figure uses the majority of the plotting surface to display the time series itself and a minority of the plotting surface to display the ACF and PACF plots drawn to the same scale.

All models in the family of ARIMA models under investigation are fit with a single command specified in standard S-PLUS time series model notation:

```
co2.loop <-
  arma.loop(co2,
            model=list(list(order=c(2,1,2)),
                      list(order=c(0,1,1), period=12)))
```

Figure 18.8 plots the family of models, again as a single display with coordinated sets of subgraphs, with another single command:

```
tsdiagplot(armas=co2.loop)
```

The series of plots in each set of subgraphs is displayed in the same systematic order. All plots of the same form are displayed to the same scale.

Fine control of plotting options and labeling is possible with optional arguments to the `tsacfplots` and `tsdiagplot` functions. Each of the individual subgraphs is also directly accessible to the user.

Formal and systematic display of a series of models makes it easy to recognize the structural differences in the series of models and to compare them.

### 18.A.4 User Functions Written for S-PLUS

Several user functions are provided and described here in terms of their role in the modeling. In addition to these user functions, there are secondary functions that the user functions call to do much of the work.

### Primary User Functions

**tsacfplots:** Provides a single display (of the form of Figure 18.6) with the times series plot central and both the ACF and PACF plots on the same scale. It does so by calling **seqplot** (equivalent to **ts.plot** but with much finer control of labeling options) for the time series plot and then **acf.pacf.plot** for the coordinated ACF and PACF plots. These in turn are constructed by the S-PLUS routine **acf**.

**arma.loop:** Takes a time series and a model statement of the form

$$(p_{\max}, d, q_{\max}) \times (P, D, Q)_{\text{period}}$$

It then loops through the family of models indexed by the model parameters  $1:p_{\max}$  and  $1:q_{\max}$ , with  $d, P, D, Q$  held constant. Results are stored in a list indexed by the values of  $p$  and  $q$ .

**arma.loop** also permits the model statement (note that order matters)

$$(P_{\max}, D, Q_{\max})_{\text{period}} \times (p, d, q)$$

and then loops through the family of models indexed by the model parameters  $1:P_{\max}$  and  $1:Q_{\max}$ , with  $p, d, q, D$  held constant. Results are stored in a list indexed by the values of  $P$  and  $Q$ .

**diag.arma.loop:** Produces an indexed list of the **arima.diag** results for each model in the result of the **arma.loop**. Diagnostics are calculated on the boundary values of the parameters  $p$  and  $q$ , and in particular for those functions defined in the special case  $(p, d, q) = (0, 0, 0)$ .

**tsdiagplot:** Takes a time series and a model statement and calls all the diagnostic plot routines. It makes sensible default choices for all the arguments and produces a graph similar to Figure 18.8. For printouts of any of the numerical tables, or finer control over the layout and labeling of the plots, the user should study the more detailed illustrations of function use in the **statlib** example.

### Print Methods

**coef.arima:** S-PLUS does not provide this function.

**print.arima:** This version gives more statistics than the S-PLUS version.

**print.arma.loop and summary.arma.loop:** Produce tables similar to Table 18.4 from the nested list result of the **arma.loop** function.

### Individual Plot Functions

Each of the subgraphs in `tsacfplots` and `tsdiagplot` is directly accessible to the users. Each is fully parameterized.

`tsacfplots`: Figures 18.4 and 18.6.

`seqplot` Time series

`acf.pacf.plot` Coordinated ACF and PACF

`tsdiagplot`: Figures 18.7 and 18.8.

`acfplot` ACF and PACF of residuals

`residplot` Standardized residuals

`gofplot` Portmanteau goodness-of-fit statistic (GOF)

`aicsigplot` Interaction plot of AIC or  $\sigma^2$

`plot.forecast`: Figure 18.9. Data, forecasts, and confidence bands.

### Additional functions

Not for direct use by users.

`rearrange.diag.arma.loop`: Rearranges the list of diagnostics indexed by model into a list of matrices of diagnostics, each matrix indexed by the models. The sole purpose of this rearrangement is for plotting.

# Appendix A

---

## Software

We discuss and recommend statistical software, text editing and word processing software, graphics display software, and our practice and preferences for using the software. The urls for all software recommended here are in the file (`sftw/code/url.htm`).

### A.1 Statistical Software

We have chosen to work with both of what we believe are the two leading statistical software languages available today: S and SAS. We have produced almost all analyses using both languages. The code to both languages is referenced throughout our text, even when we display output from just one of the languages. Most of our graphics are developed only in S because we believe that S provides a more powerful and flexible environment for graphics.

S is an exceptionally well-developed tool for statistical research and analysis, that is for exploring and designing new techniques of analysis, as well as for analysis. S is especially strong for statistical graphics, the output of data analysis through which both the raw data and the results are displayed for the analyst and the client. The S language was originally developed at Bell Labs in the 1970s. The Association for Computing Machinery (ACM) awarded John M. Chambers of Bell Labs the 1998 Software System Award for developing the S system.

There are two implementations of the S language. S-PLUS is developed and distributed by Insightful Corporation (Insightful Corp., 2002), origi-

nally under license from Lucent Technologies. In 2004 Insightful purchased the code base from Lucent. The S-PLUS student edition is available for download to qualified students and faculty.

R (R Development Core Team, 2004) is the Open Software version available by download. The developers are The R Development Core Team, an international group that includes John Chambers and other Bell Labs researchers.

SAS (SAS Institute, 2000) is the most widely used package for extensive statistical analysis and data management. A site-licensed student edition is available through many university computer centers. SAS was originally developed at North Carolina State University under a grant from the U.S. Department of Agriculture. It has been developed and distributed by the SAS Institute since the 1970s.

## A.2 Text Editing Software

We distinguish between the concepts of text editing (discussed in this section) and word processing (discussed in Section A.3). Text editing is moving characters around on the screen with the expectation that they will stay where you put them. This is critical when writing computer programs where the physical placement of lines and characters on the page is part of what the computer program interprets. In the S language the two layouts of the same characters in Table A.1 have completely different interpretations.

An excellent text editor is an indispensable tool for the statistical analyst. The editor is the single program in which we spend most of our time. We use it for looking at raw data, for writing commands in the statistical languages we use, for reading the output tables produced by our statistical programs, for writing reports, for reading and writing correspondence about our studies to our clients, consultants, supervisors, and subordinates. See Appendix E for a detailed discussion of editors.

### A.2.1 Emacs

We recommend Emacs (Free Software Foundation, 2003), a mature, powerful, and easily extensible text editing system freely available under the GNU General Public License for a large number of platforms, including Unix, Mac, and Windows. See Section E.3 for a detailed discussion of Emacs. We use Emacs as the access to our statistical software by using ESS (Rossini et al., 2004a) and (Rossini et al., 2004b), a package that extends Emacs to provide a functional, easily extensible, and uniform interface for multiple statistical packages.

TABLE A.1. Two different interpretations of the same characters that depend on their placement on separate lines. If the first set of input lines were reformatted according to English language paragraph formatting rules it would be interpreted as if it were the second set, which has a completely different meaning. This is the simplest possible example of why we make a distinction between text editing and word processing.

| Two lines            | One line |
|----------------------|----------|
| <b>S-PLUS Input</b>  |          |
| 3                    | 3 + 4    |
| + 4                  |          |
| <b>S-PLUS Output</b> |          |
| > 3                  | > 3 + 4  |
| [1] 3                | [1] 7    |
| > +4                 |          |
| [1] 4                |          |

We recommend that users of this text seriously consider becoming Emacs users. However, Emacs usage is not essential for successfully learning from and working with this book.

### A.2.2 Microsoft Word

Microsoft Word is configured by default as a word processor. It can be used as a text editor by changing those default settings. The most critical features are the font and the paragraph reflow. Courier (or another mono-width font) should be used for program writing or for summary reports in which your displayed output from S-PLUS or SAS are included. The software output from these packages is designed to look right (alignment and spacing) with mono-width fonts. It looks terrible, to the point of illegibility, in proportional fonts. See examples in Sections B.6 and E.2. Other features to turn off are spell checking and syntax checking, both of which are designed to make sense with English (or another natural language) but not with programming languages.

## A.3 Word Processing Software

Word processing is moving sentences, paragraphs, sections, figures, and cross-references around. L<sup>A</sup>T<sub>E</sub>X and Microsoft Word are word processors.

Word can be used as a text editor by manually turning off many of the word processing features.

### A.3.1 L<sup>A</sup>T<sub>E</sub>X

We recommend L<sup>A</sup>T<sub>E</sub>X (the standard required by many statistics and mathematics journals, and the typesetting package with which we wrote this book). L<sup>A</sup>T<sub>E</sub>X has macros that completely adapt to any standard style (or write your own) and knows all the intricacies of mathematical typesetting. It is very easy to include figures and tables into the manuscript.

The `latex` function in S-PLUS (Heiberger and Harrell, 1994) may be used to prepare formatted typeset tables for a L<sup>A</sup>T<sub>E</sub>X document. Several tables in this book (8.1, 9.1, 12.4, 15.5, and 15.8) were prepared in this way.

The T<sub>E</sub>X system (Knuth, 1984) was developed by Donald E. Knuth at Stanford University. It is now maintained by CTAN, the Comprehensive T<sub>E</sub>X Archiving Network (Comprehensive T<sub>E</sub>X Archiving Network, 2002). L<sup>A</sup>T<sub>E</sub>X (Lamport, 1986) is a documentation preparation system built on T<sub>E</sub>X. There are several distributions available. We use MikT<sub>E</sub>X (Schenk, 2001).

### A.3.2 Microsoft Word

Microsoft Word is often used as a technical editor or word processing system. Neither of us has used Microsoft Word in that way. From the experience of our colleagues and students we have come to believe that the *MathType* fonts (Design Science, Inc., 2000) are helpful in producing quality mathematical material.

## A.4 Graphics Display Software

Adobe PostScript is the standard display format in the printing industry. We use the Ghostscript (Aladdin Enterprises Software Pty Ltd., 2001) and Ghostview (Ghostgum Software Pty Ltd., 2001) Open Source licensed implementations of the PostScript language.

All graphs in this book, and the S or SAS code from which they were constructed, are included in the online files. All figure captions include the filenames of the code and of the graph.

## A.5 Operating Systems

The standard operating system for scientific computing is the Unix operating system. Complete Unix functionality is available for Microsoft Windows computers with the Cygwin (Red Hat, Inc., 2002) package, a complete Open Source implementation of the Unix operating system that runs under Windows.

We recommend, but do not require, Unix or Cygwin. Everything in the book can be done on Microsoft Windows with either point-and-click or from the MS-DOS prompt window.

## A.6 Mathematical Fonts

All mathematical fonts, and the typographical knowledge to use them well, are built-in to  $\text{\TeX}$  and  $\text{\LaTeX}$ .

Should you find a need to write mathematics in Microsoft Word, you might need the additional *MathType* mathematical fonts. A *MathType* reader is freely available for download from (Design Science, Inc., 2000). Software to create new documents with the fonts requires purchase of a license.

## A.7 Directory Structure

There are several sets of computer directories associated with this book. These are

1. Your working directories, normally subdirectories of your **HOME** directory.
2. The files we provide. These all under the **hh** directory.
3. Software. Installed wherever the program installation put them. On Microsoft Windows, this is usually a subdirectory of **c:\progra~1\** or directly under the root of the hard disk as a subdirectory of **c:\**. (The directory **c:\progra~1\** is also known as **c:\Program Files\**. We avoid using filenames with embedded blanks and use the 8.3 equivalent instead. You can find the 8.3 name with **dir /X** in Windows XP or with **dir** in older Windows systems.)

### A.7.1 HOME Directory

You need a home directory, the one where you keep all your personal subdirectories and files. On Microsoft Windows machines rmh uses

c:\HOME\rmh\ and on Unix machines rmh uses /usr/users/rmh/. The home directory is where you keep all your subdirectories (for example, c:\HOME\rmh\hh-user\, the directory in which we recommend work with this text be kept, or c:\HOME\rmh\503.f03\ and c:\HOME\rmh\504.s04\, the directories for the course we teach for which this book is designed).

On Windows machines you will probably have to create both the c:\HOME\ and c:\HOME\rmh\ directories and define the HOME environment variable.

On Windows 95/98/ME, set the HOME environment variable to the pathname of your home directory in c:\autoexec.bat with, for example (for rmh, use your name for you), the line

```
set HOME=c:\HOME\rmh
```

On Windows NT/XP set the HOME environment variable from the Control Panel by clicking */System/Advanced/Environment Variables*.

Windows NT/XP defines a directory for each user of the system, c:\Documents and Settings\rmh for me, that seems to be intended as a HOME directory. Windows also puts lots of system stuff there, so I am happier creating my own c:\HOME\rmh directory that is fully under my control.

On Unix, the home directory and HOME environment variable are set up by the system administrator when the account is created.

The home directory can be referred to directly, on Unix or with Cygwin bash on Windows, by referencing the environment variable \$HOME, and with the MS-DOS prompt on Windows by referencing the environment variable %HOME%. The environment variable has the values (for rmh)

| Operating system          | Environment variable | Value          |
|---------------------------|----------------------|----------------|
| Windows and MS-DOS prompt | %HOME%               | c:\HOME\rmh    |
| Cygwin under Windows      | \$HOME               | c:/HOME/rmh    |
| Unix                      | \$HOME               | /usr/users/rmh |

In Unix shells (including Cygwin) and in Emacs, the symbol ~ is a synonym for \$HOME.

We recommend that you create a subdirectory \$HOME/hh-user/ (for rmh on Windows that expands to c:\HOME\rmh\hh-user\ ) for your work with this book. This will be a directory under your home directory and parallel to your course directory and other directories that you develop for other projects.

### A.7.2 HH Book Online Files

See Preface Section 3 for a discussion of the directory structure of the unzipped online files. In summary, we have a directory for each chapter, each with a subdirectories for `code` (containing programs written in the S-PLUS and SAS languages), `transcript` (containing the output listings from the files in `code`), and `figure` (containing PostScript files for all the figures in the book, mostly constructed by statements in the `code` files).

We provide many new graphical functions written in the S language. These are available in attachable libraries for both S-PLUS and R, and in ASCII code in files `splus.library/*.s`. The `splus.library` is required for many of our examples.

We recommend that you copy the unzipped online files to your hard disk in its own directory, parallel to your `$HOME` directory. Thus it would appear, for example, in Windows as `c:\HOME\hh\` and in Unix as `/usr/users/hh/`. When this is done, and when the S-PLUS and SAS initializations recommended in Appendices B and C are completed, then all code in the HH online files will work as written with no changes.

# Appendix B

---

## S-PLUS and R

We provide many functions for S-PLUS that are designed to extend capabilities and simplify many tasks. This appendix describes how to attach the library containing our functions. We also give our recommendations on working style when using S-PLUS and R.

The source files `*.s` are included in the online files for this book in directory `splus.library`. We provide the executable library for S-PLUS 6 (`.Data`), and R (`.RData`). We also describe in Section B.9 how to construct the executable library from the source files.

There are two major current dialects of S: S-PLUS (version 6 for Windows and Unix) and R (version 1.9 for Windows and Unix).

Our examples and functions work with both. There are small (well, we hope small) differences in getting started with each. The `.Data` library will work on all versions of S-PLUS 6. The `.RData` image will work on all versions of R.

### A Note on Notation

Some of the commands in this chapter are to be executed at the shell command prompt. We normally use the Cygwin `bash` shell (Red Hat, Inc., 2002) on Windows. This gives us a complete Unix environment on Windows machines. We also provide the equivalent MS-DOS prompt commands. In Unix and in Cygwin `bash`, pathnames are written with forward slashes “`/`”. In the Windows MS-DOS prompt shell and in Windows icon `Properties` windows, pathnames are written with backslashes “`\`”. In this chapter, when

we are talking about shells (Unix prompt or MS-DOS prompt), we will indicate both forms. When we are inside the S language, we will use only the forward slashes “/”.

## B.1 Create Your Working Directory and Make the HH Library Available

Getting started is always the hardest part. The steps are simple. There are too many of them, and they must be done in the right order. The details are slightly different for each dialect and platform. The paths in these examples are appropriate for the default locations of S-PLUS 6.1 and R 1.9.0 on Windows. For different releases of S-PLUS and R, or for a nondefault location, use the paths appropriate for your computer.

### B.1.1 Windows—Both S-PLUS and R

The steps are

1. We recommend that you copy the entire **hh** directory tree from the online files to your hard disk. We recommend that you use the **c:\HOME\hh** (**c:/HOME/hh**) directory on Windows. We are recommending that you create a new user **hh** with home directory parallel to your own directory. After you create this directory it will be made read-only.

On Windows, open two Windows Explorers, one to **c:\HOME\** and one to the unzipped online files. Pick up the **hh** directory from the online files and drop it into the **c:\HOME\**.

2. Create your working directory on your hard disk for the HH book. We recommend that you create a separate project for this book. We suggest the Windows directory name **c:\HOME\yourname\hh-user** (**c:/HOME/yourname/hh-user**). One of us uses **c:/HOME/rmh/hh-user** and the other **c:/HOME/burt/hh-user**.

Please note that this is your directory and that it is distinct from the **hh** directory that contains the book’s files.

### B.1.2 Windows and S-PLUS

1. Initialize S-PLUS in the `hh-user` directory.

**S-Plus 6 on Windows, MS-DOS prompt:** Execute the commands

```
cd c:\HOME\yourname\hh-user
C:\Progra~1\Insightful\splus61\cmd\CHAPTER
```

**S-Plus 6 on Windows, bash shell:** Execute the commands

```
cd c:/HOME/yourname/hh-user
C:/Progra~1/Insightful/splus61/cmd/CHAPTER
```

2. Open S-PLUS in the `hh-user` directory. As usual, there are several ways to do this, depending on dialect, platform, and preference. Use only one of the following:

**S-Plus for Windows, by clicking on an icon.** Copy the existing S-PLUS icon to the Desktop. Rename the copy to S-PLUS `hh-user`. Right-click on the icon and select **Properties**. Select the **Shortcut** tab. Modify the **Target:** by adding the argument

```
S_PROJ=c:\HOME\yourname\hh-user
```

at the end. Click OK. Now that the new icon is defined, click on it.

The first time you use the new `hh-user` directory, S-PLUS will give a warning that the “`.Data` and/or `.Prefs`” directory doesn’t exist. Actually, `.Data` does exist. We created it with the `CHAPTER` command in the previous step. Allow S-PLUS to create `.Prefs`.

**S-Plus for Windows, from ESS in Emacs.** Enter `M-x S`. You will be prompted for a directory. Enter `~/hh-user`

3. Tell S-PLUS where the HH library sits on your computer. Do this by entering a version of the `.First` function that has been tailored to your computer. Table B.1 contains illustrations of the needed tailoring. This is the only S-PLUS function in this book that *must* be tailored to your individual computer. The illustration in Table B.1 shows a sample (`splus.library/First-w-6.s`) for S-PLUS 6 on Windows assuming you take our recommendations on the directory pathname for the HH library. The caption of Table B.1 describes the changes needed if you place the HH library elsewhere.

When you have defined `.First`, confirm that it works by executing `.First()` at the S prompt. Then do `search()`. The HH library `HH` should be the second item on the `search` list. We need to be the second item because we replaced several of the S-PLUS-supplied functions.

Once the workspace has been saved with `.First` defined, `.First` will be executed automatically on all future starts of S-PLUS.

TABLE B.1. Sample `.First` function to tell S-PLUS or R where the HH library sits on your computer. The displayed sample for S-PLUS 6 for Windows assumes you have placed the HH library at our recommended location of `c:/HOME/hh`. The `options` line is the only line in any code in the HH library that might need to be customized for your computer.

We have similar samples for the following platforms and dialects:

|                             |                   |
|-----------------------------|-------------------|
| (splus.library/First-w-6.s) | Windows, S-PLUS 6 |
| (splus.library/First-w-r.s) | Windows, R        |
| (splus.library/First-u-6.s) | Unix, S-PLUS 6    |
| (splus.library/First-u-r.s) | Unix, R           |

If you need to place the files elsewhere (or even work directly from the CD of the unzipped downloaded online files (see Preface Section 3), then modify the `options` line in one of these files to define a `.First` function specific to your installation.

---

```
S-PLUS (splus.library/First-w-6.s):
.First <- function() {
  options(HH.ROOT.DIR="c:/HOME/hh")  ## This is the only line
                                         ## you need to customize.
  library("HH",
         lib.loc = paste(options()$HH.ROOT.DIR, "splus.library", sep="/"),
         first = T)
}
```

---

### B.1.3 Windows and R

#### 1. Initialize R in the `hh-user` directory.

**R on Windows, MS-DOS prompt:** Execute the commands

```
cd c:\HOME\yourname\hh-user      ## shell command
C:\Progra^1\R\rw1090\bin\Rterm   ## shell command (use correct pathname)
q()                            ## R command
yes                           ## response to R query
```

**R on Windows, bash shell:** Execute the commands

```
cd c:/HOME/yourname/hh-user      ## shell command
C:/Progra^1/R/rw1090/bin/Rterm   ## shell command (use correct pathname)
q()                            ## R command
yes                           ## response to R query
```

#### 2. Open R in the `hh-user` directory. As usual, there are several ways to do this, depending on dialect, platform, and preference. Use only one of the following:

**R for Windows, by clicking on an icon.** Copy the existing R icon to the Desktop. Rename the copy to R hh-user. Right-click on the icon and select Properties. Select the Shortcut tab. Change the value in the Start in: to

```
c:\HOME\yourname\hh-user
```

at the end. Click OK. Now that the new icon is defined, click on it.

**R for Windows from ESS in Emacs.** Enter M-x R. You will be prompted for a directory. Enter ~/hh-user

3. Tell R where the HH library sits on your computer. Do this by entering a version of the .First function that has been tailored to your computer. This is the only R function in this book that *must* be tailored to your individual computer. You may use (*splus.library/First-w-r.s*) for R on Windows assuming you take our recommendations on the directory pathname for the HH library. See Table B.1 for the S-PLUS version, for the caption, and for additional comments. The caption of Table B.1 describes the changes needed if you place the HH library elsewhere.

When you have defined .First, confirm that it works by executing .First() at the R prompt. Then do search(). The HH library *splus.library/HH* should be the eighth item on the search list. We need to be the eighth item because we have replacement function definitions for several of the R-supplied functions in the lattice and grid packages.

Once the workspace has been saved with .First defined, .First will be executed automatically on all future starts of R.

#### B.1.4 Unix—Both S-PLUS and R

The steps are

1. We recommend that you copy the entire hh directory tree from the online files to your hard disk. We recommend that you use the /usr/users/hh directory on Unix. We are recommending that you create a new user hh with home directory parallel to your own directory. After you create this directory it will be made read-only.

On Unix at the shell prompt,

```
cd /usr/users/hh  
(cd /CDROM/hh ; tar cf - *) | tar xf -
```

2. Create your working directory on your hard disk for the HH book. We recommend that you create a separate project for this book. On Unix we recommend the directory name /usr/users/yourname/hh-user (thus,

one of us uses the name `/usr/users/rmh/hh-user` and the other uses `/usr/users/burt/hh-user`).

Please note that this is your directory and that it is distinct from the `hh` directory that contains the book's files.

### B.1.5 Unix and S-PLUS

1. Initialize S-PLUS in the `hh-user` directory.

Execute the commands

```
cd /usr/users/yourname/hh-user  
S CHAPTER
```

2. Open S-PLUS in the `hh-user` directory. Use only one of the following:

**S-Plus for Unix from the shell prompt.**

```
cd /usr/users/yourname/hh-user  
S
```

**S-Plus for Unix, from ESS in Emacs.** Enter `M-x S`. You will be prompted for a directory. Enter `~/hh-user`

3. Tell S-PLUS where the HH library sits on your computer. Do this by entering a version of the `.First` function that has been tailored to your computer. This is the only S-PLUS function that must be tailored to your individual computer. You may use (`splus.library/First-u-6.s`) for S-PLUS 6 on Unix assuming you take our recommendations on the directory pathname for the HH library. See Table B.1 for the S-PLUS for Windows version, for the caption, and for additional comments. The caption of Table B.1 describes the changes needed if you place the HH library elsewhere.

When you have defined `.First`, confirm that it works by executing `.First()` at the S-PLUS prompt. Then do `search()`. The HH library `HH` should be the second item on the search list. We need to be the second item because we have replacement function definitions for several of the S-PLUS-supplied functions.

Once the workspace has been saved with `.First` defined, `.First` will be executed automatically on all future starts of S-PLUS.

### B.1.6 Unix and R

1. Initialize R in the `hh-user` directory.

Execute the commands

```

cd c:/HOME/yourname/hh-user ## shell command
R ## shell command
q() ## R command
yes ## response to R query

```

2. Open R in the `hh-user` directory. Use only one of the following:

**R for Unix from the shell prompt.**

```

cd /usr/users/yourname/hh-user
R

```

**R for Unix, from ESS in Emacs.** Enter `M-x R`. Respond to the prompt for a directory with `~/hh-user`

3. Tell R where the HH library sits on your computer. Do this by entering a version of the `.First` function that has been tailored to your computer. This is the only R function in this book that *must* be tailored to your individual computer. You may use (`splus.library/First-w-r.s`) for R on Windows assuming you take our recommendations on the directory pathname for the HH library. See Table B.1 for the S-PLUS version, for the caption, and for additional comments. The caption of Table B.1 describes the changes needed if you place the HH library elsewhere.

When you have defined `.First`, confirm that it works by executing `.First()` at the R prompt. Then do `search()`. The HH library `splus.library/HH` should be the second item on the `search` list. We need to be the eighth item because we have replacement function definitions for several of the R-supplied functions in the `lattice` and `grid` packages.

Once the workspace has been saved with `.First` defined, `.First` will be executed automatically on all future starts of R.

## B.2 Using S-PLUS and R with HH

After the initialization described in Section B.1, all future starts of S-PLUS or R in the `hh-user` directory will automatically execute `.First()` and therefore automatically attach the HH library. Manual intervention will not be needed. All the HH library functions listed in Table B.3 will be available for your use.

All file references in the `*/code/*.s` files are relative to the `hh` directory that was defined in Table B.1. Datasets, for example `fat.data`, are written `hh("datasets/fat.data")` in S code. The `hh` function is one of the functions in the HH library. It uses the value of the `HH.ROOT.DIR` option to locate the code. Once the `.First` function has been defined, then all code

in the `*/code/*.s` files will work as written. You can **source** them, for example `source(hh("grap/code/grap.read.le.s"))`, or open them in your editor and execute one line at a time (for example, by using **C-c C-n** in Emacs, or by copy and paste from an editor that doesn't work smoothly with S). NO changes are needed for ANY code files in the book once `.First` has been correctly defined.

## B.3 S-PLUS for Windows—Recommended Options

1. We recommend that you change the S-PLUS 6.1 default options.
  - a. Click the Restore Down button on the S-PLUS frame, then set the top of the frame to the top of the screen and drag the bottom of the frame to the bottom of the screen. The left margin of the S-PLUS frame should be on the left margin of the monitor and the right margin about  $\frac{2}{3}$  across the screen. The Commands window should start in the upper-left corner, and go to  $\frac{1}{2}$  inch above the bottom of the frame. The Commands window should be 80 columns wide. See Figure B.1.
  - b. Options/General settings.../Data  
Default Text Col.: character (efficient data storage)  
Date Format (input and output): MM/dd/yyyy (four digit year)  
Time Format (input and output): hh:mm:ss tt
  - c. Options/General settings.../Startup  
Command Line: check
  - d. Options/Command Line.../Options  
Key Scroll: Page keys
  - e. Options/Graph Options...  
Auto Pages: Every Graph
  - f. Options/"Save Window Size/Properties as Default"
2. When you use S-PLUS for Windows through Emacs (the recommended way), there is a short delay (about a minute) during startup where Emacs looks frozen. Wait until Emacs comes back.
3. When you start S-PLUS or R from within Emacs, you may need to tell Emacs the pathname to S-PLUS or R by adding the lines in file (`Sp1s/code/sample.el`) to your `~/.emacs` file.

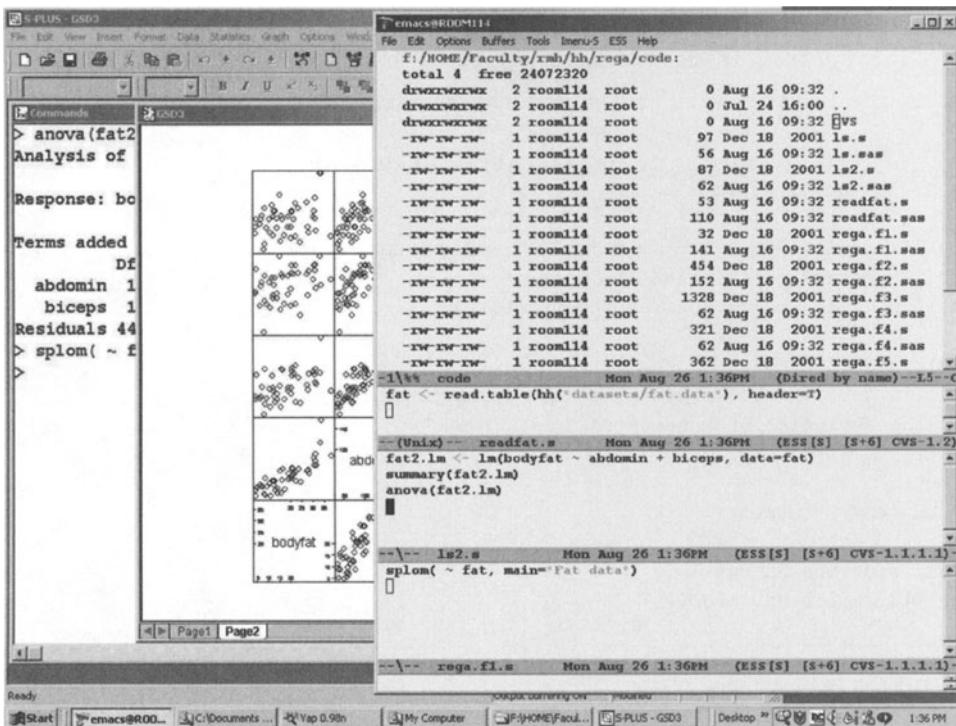


FIGURE B.1. Recommended screen layout for S-PLUS for Windows and Emacs (or other editor). The S-PLUS window is on the left, with the Commands window in its upper left, and Emacs on the right.

(Spls/figure/screenshot270.ps.gz)

4. When you use S-PLUS for Windows through Emacs, we recommend that the S-PLUS window be on the left side of the monitor screen, with the Commands window in the upper-left corner, and Emacs on the right. See Figure B.1. This arrangement allows you to see the left side of the Commands window when the cursor is in the Emacs buffer `myfile.s`. The left side is where you monitor lines that have been sent over from the Emacs editing buffer with the ESS commands `C-c C-n` and relatives.
5. The colors that S-PLUS trellis uses by default are difficult to see on the overhead projector and when printed on a black-and-white printer. We list several alternate black-and-white and color options in Table B.2. If you want to permanently set one of the color options, you may modify the `.First` function by adding a `trellis.device` line.

TABLE B.2. Optional color schemes for S-PLUS trellis.  
(splus.library/trellis.device.hh.s), (splus.library/trellis.device.hh.r)

---

```
S-PLUS (Spls/code/trellis.device.s):
## S-Plus color schemes

trellis.device()          ## default trellis color.scheme
trellis.device(color.scheme="standard")
trellis.device(color.scheme="topographical")

trellis.device(color=F)
trellis.device(color.scheme="Trellis Black on White")
trellis.device(color.scheme="white on black")

## HH color schemes

trellis.device.hh.bw()
trellis.device.hh.color()
```

---

## B.4 HH Library Functions

The functions in Table B.3 are available in the HH library. They are currently described in the chapters where they are first used.

## B.5 Learning the S Language

There are several excellent sources for learning the S language.

S-PLUS includes several manuals as part of the distribution. All of them are online, accessible from the S-PLUS window by clicking Help/Online Manuals. Full-price editions of S-PLUS (i.e., not the downloadable Student Edition) come with printed manuals. For learning the language, we recommend the *User's Guide*, Chapter 10 “Using the Commands Window” for an introduction to the language, and *The Programmer's Guide*, Chapters 1 to 4 for more extensive discussion.

R includes several manuals as part of the distribution. All of them are online, accessible from the Rgui window by clicking Help/Manuals or directly from the directory by double-clicking on R/rw1090/doc/manual/\*.pdf. We recommend beginning with R-intro.pdf.

TABLE B.3. Functions available in `splus.library`. The illustration assumes that your `.First` function takes the values illustrated in our recommendation in Table B.1. This table continues on page 642.

---

```
S-PLUS (Spls/transcript/splus.library.st):
> ## The value depends on your .First(). This is our recommendation.
> search()[2]
[1] "HH"
> options(width=81)
> print(objects(2),q=F)
[1] .arima.info.names.not.ordered      .pearson.x2.hh
[3] GSremove                          X.residuals
[5] X.residuals.default              X.residuals.formula
[7] X.residuals.lm                  [.arma.loop
[9] [.cp.object                      [.diag.arma.loop
[11] [.mmc.mmc.multicomp            abind
[13] acf.pacf.plot                  acfplot
[15] aicsigplot                     ancova
[17] anova.ancova                  anova.mean
[19] antilogit                       arma.loop
[21] as.character.arima.model       bwplott
[23] c.rts                           chisq.test.hh
[25] ci.plot                          coef.ancova
[27] coef.arima                      coefficients.ancova
[29] cp.calc                          diag.arma.loop
[31] diag.maybe.null                 do.formula.trellis.xysplom
[33] export.eps                      gof.calculation
[35] gofplot                         hh
[37] hov                             hov.bf
[39] if.R                            interaction.plot
[41] interaction2wt                 interaction2wt.default
[43] interaction2wt.formula         ladder
[45] ladder.f                         ladder.fstar
[47] ladder3                         list.hh
[49] lm.case                         lm.fit.qr
[51] logit                           model.tables.aovlist
[53] multicomp.label.change          multicomp.label.change.mmc.mmc.multicomp
[55] multicomp.label.change.multicomp multicomp.lm
[57] multicomp.mean                  multicomp.mmc
[59] multicomp.mmc.mean              multicomp.order
[61] multicomp.reverse              na.keep
[63] norm.curve                      norm.setup
[65] npar.arima                      npar.arma
[67] objip                           odds.ratio
[69] orthog.complete                orthog.construct
```

---

TABLE B.3 continued. Functions available in `splus.library`.

---

| S-PLUS (Spls/transcript/splus.libraryb.st): |                            |
|---|----------------------------|
| [71] panel.acf                              | panel.ancova               |
| [73] panel.barchart                         | panel.barchartt            |
| [75] panel.bwplot.intermediate.hh           | panel.bwplott              |
| [77] panel.cartesian                        | panel.case                 |
| [79] panel.ci.plot                          | panel.dotplott             |
| [81] panel.gof                              | panel.hov                  |
| [83] panel.interaction2wt                   | panel.pairs.hh             |
| [85] panel.parallel                         | panel.std.resid            |
| [87] panel.stripplot                        | panel.xysplom              |
| [89] partial.corr                           | persp.back.wall.x          |
| [91] persp.back.wall.y                      | persp.floor                |
| [93] persp.plane                            | plot.ancova                |
| [95] plot.case                              | plot.forecast              |
| [97] plot.hov                               | plot.hov.bf                |
| [99] plot.lm                                | plot.mlm                   |
| [101] plot.mmc.multicomp                    | plot.multicomp             |
| [103] plot.odds.ratio                       | predict.ancova             |
| [105] pretty.base                           | print.ancova               |
| [107] print.arima                           | print.arma.loop            |
| [109] print.arma.loop.list                  | print.cp.object            |
| [111] print.multicomp                       | print.tsacfplots           |
| [113] print.tsdiagplot                      | print1.tsdiagplot          |
| [115] print2.tsdiagplot                     | proj.aoalist               |
| [117] rearrange.diag.arma.loop              | regri.plot                 |
| [119] regr2.plot                            | resid.squares              |
| [121] residplot                             | residual.plots             |
| [123] rfplot                                | seqplot                    |
| [125] seqplot.cts                           | seqplot.default            |
| [127] seqplot.its                           | seqplot.rts                |
| [129] strip.1TT                             | strip.background0          |
| [131] strip.interaction2wt                  | strip.ladder               |
| [133] strip.xysplom                         | summary.ancova             |
| [135] summary.arma.loop                     | summary.arma.loop.list     |
| [137] superpose.symbol.size                 | t.trellis                  |
| [139] trellis.device.hh.bw                  | trellis.device.hh.color    |
| [141] trellis.device.hh.ps.bw               | trellis.device.hh.ps.color |
| [143] tsacfplots                            | tsdiagplot                 |
| [145] vif                                   | vif.default                |
| [147] vif.formula                           | vif.lm                     |
| [149] xysplom                               | xysplom.default            |
| [151] xysplom.formula                       |                            |

---

TABLE B.4. Correct and incorrect alignment of computer listings. The columns in the correct example are properly aligned. The concept of columns isn't even visible in the incorrect example.

| Courier (with correct alignment of columns) |          |            | Times Roman (alignment is lost) |
|---|----------|------------|---------------------------------|
| > tv[1:5, 1:3]                              |          |            | > tv[1:5, 1:3]                  |
|   | life.exp | ppl.per.tv | life.exp                        |
| Argentina                                   | 70.5     | 4.0        | ppl.per.phys                    |
| Bangladesh                                  | 53.5     | 315.0      | Argentina 70.5 4.0 370          |
| Brazil                                      | 65.0     | 4.0        | Bangladesh 53.5 315.0 6166      |
| Canada                                      | 76.5     | 1.7        | Brazil 65.0 4.0 684             |
| China                                       | 70.0     | 8.0        | Canada 76.5 1.7 449             |
|   |          |            | China 70.0 8.0 643              |

## B.6 S Language Style

S is a language. Languages have standard styles in which they are written. When a language is displayed without paying attention to the style, it looks unattractive and may be illegible. It may also give valid statements that are not what the author intended.

The basic style conventions are simple. They are also self-evident after they have been pointed out. Look at the examples in files `*/code/*.s` and in the S-PLUS manuals. Read what you turn in before turning it in.

1. Use the courier font for computer listings. This is **courier**. This is Times Roman. Notice that displaying program output in a font other than one for which it was designed destroys the alignment and makes the output illegible. We illustrate the illegibility of improper font choice in Table B.4 by displaying the first few lines of the (`datasets/tv.dat`) data from Chapter 4 [read into S-PLUS with file (`grap/code/grap.read.le.s`)] in both correct and incorrect fonts.
2. Use sensible spacing to distinguish the words and symbols visually. This convention allows people to read the program.

---

bad      abc<-def      no space surrounding the <-

---

good    abc <- def

---

3. Use sensible indentation to display the structure of long statements. Additional arguments on continuation lines are most easily parsed by people when they are aligned with the parentheses that define their depth in the set of nested parentheses.

---

```
bad  names(tv) <- c("life.exp", "ppl.per.tv", "ppl.per.phys",
  "fem.life.exp", "male.life.exp")
```

---

```
good  names(tv) <- c("life.exp",
  "ppl.per.tv",
  "ppl.per.phys",
  "fem.life.exp",
  "male.life.exp")
```

---

Use Emacs to help with indentation. For example, open up a new file `tv.s` in Emacs and type the above bad example—in two lines with the indentation exactly as displayed. Emacs in ESS [S] mode will automatically indent it correctly.

4. Use a page width in the Commands window that your word processor and printer supports. We recommend

```
options(width=80)
```

if you work with the natural width of  $8\frac{1}{2} \times 11$  paper with 10-pt type. If you use a word processor that insists on folding lines at some shorter width (72 characters is a common—and inappropriate—default folding width), you must either take control of your word processor, or tell S-PLUS to use a shorter width. Table B.5 shows a fragment from the file (`rega/transcript/fat2.lm.st`) with two different width settings for the word processor.

5. Read all the `hh("*/code/*.s")` files. Note that some read their data with the `read.table` command with `header=T` and others with `header=F`. Look at the `hh("datasets/*.dat")` files to see why. Find out what the arguments to `read.table` are by asking S-PLUS, by typing `?read.table` in the Commands window.
6. Reserved names. S-PLUS has functions with the single-letter names `c`, `s`, and `t`. S-PLUS also has many functions whose names are commonly used statistical terms, for example: `mean`, `median`, `resid`, `fitted`. If you inadvertently name an object with one of the names used by the system, your object might mask the system object and strange errors would ensue. Do not use system names for your variables. See Section B.7 for information on how to detect name conflicts.

TABLE B.5. Legible and illegible printings of the same table. The illegible table was inappropriately folded by an out-of-control word processor. You, the user, must take control of folding widths.

---

Legible:

---

```
> anova(fat2.lm)
Analysis of Variance Table

Response: bodyfat

Terms added sequentially (first to last)
Df Sum of Sq Mean Sq F Value    Pr(F)
abdomin   1  2440.500 2440.500 101.1718 0.00000000
biceps    1   209.317  209.317   8.6773 0.00513392
Residuals 44  1061.382   24.122
```

---

Illegible (folded at 31 characters):

---

```
> anova(fat2.lm)
Analysis of Variance Table

Response: bodyfat

Terms added sequentially (first
to last)
Df Sum of Sq Mean Sq
F Value    Pr(F)
abdomin   1  2440.500 2440.500
101.1718 0.00000000
biceps    1   209.317  209.317
8.6773 0.00513392
Residuals 44  1061.382   24.122
```

---

## B.7 S-PLUS Inexplicable Error Messages

In general, weird and inexplicable errors mean that there are masked function names. That's the easy part. The trick is to find which name. The name conflict is frequently inside a function that has been called by the function that you called directly. The general method, which we usually won't need, is to trace the action of the function you called, and all the functions it called in turn. See `?trace`, `?browser`, and `?debugger` for help on using these functions.

The method we will use is to find all occurrences of our names that might mask system functions. S-PLUS provides two functions that help us. See **?find** and **?conflicts** for further detail.

**find:** Returns a vector of names, or positions of databases and/or frames that contain an object.

- This example is not a problem because we extended the definition of **plot.multicomp** to give the user control over the vertical spacing in the plot.

```
> find(plot.multicomp)
[1] "HH" "splus"
```

- This example is a problem because our working directory is listed.

```
> find(s)
[1] "c:\\HOME\\hh-user" "splus"
```

**conflicts:** This function checks a specified portion of the search list for items that appear more than once.

- The only items we need to worry about are the ones that appear in our working directory.

```
> conflicts(detail=T)
$"c:\\HOME\\hh-user":
[1] "s"

$HH:
[1] "plot.multicomp"

$splus:
[1] "plot.multicomp" "s"
```

Once we have found those names we must assign their value to some other variable name and then remove them from the working directory. In the above example, we have used the system name "s" for one of our variable names. We must assign the value to a name without conflict, and then remove the conflicting name.

```
> find(s)
[1] "c:\\HOME\\hh-user" "splus"
> s
[1] 1.697097
> find(s.myproject)
Problem: Object "s.myproject" not found
> s.myproject <- s
```

---

```
> find(s.myproject)
[1] "c:\\HOME\\hh-user"
> rm(s)
> find(s)
[1] "splus"
```

## B.8 Using S-PLUS with Emacs and ESS

We recommend using ESS and Emacs as your primary interface and editor for S-PLUS and R. See Section E.3 for details.

If you are using Emacs as your editor (strongly recommended), then the syntax highlighting will catch many typographical errors and therefore make your task much easier.

If you are using Emacs as your editor, remember that `*.st` files come up read-only. If you need to make them writable, use `C-x C-q`.

If you are using an editor that thinks it knows more than you, be very sure that the `*.s` and `*.st` files are displayed in Courier and fit on the page.

## B.9 Constructing the HH Library with S-PLUS and R

The executable libraries for S-PLUS 6 (`splus.library/HH/.Data`) and R (`splus.library/HH/.RData`) work on both Windows and Unix releases of S-PLUS and R. This section outlines how we constructed the libraries. Users of S-PLUS and R will not normally need this section.

Details for creating the library differ by dialect and operating system. We give sample setup scripts for S-PLUS in file (`splus.library/HH-setup.s`) and for R in file (`splus.library/HH-setup.r`). Much of our source code is common for the S-PLUS and R dialects of the S language. The primary differences between the dialects is in the details of how the trellis functions are implemented. S-PLUS builds trellis on its earlier graphics primitives. R constructed an entirely new, and much more powerful, set of graphics primitives in the `grid` library. The setup scripts are hard-wired for rmh's Windows computer, with comments at the places that need changing for any other machine. The S-PLUS and R libraries will each work with Unix releases of their respective systems. The Windows scripts are written for the Cygwin `bash` shell, not the MS-DOS prompt. We assume in these scripts that the `hh` directory tree has already been copied to `c:/HOME/hh` or `/usr/users/hh`. Although the scripts have been given `.s` extensions, they are actually designed to be pasted into the `bash` shell one line at a time.

The first few lines are commented-out shell commands that initialize the **.Data** or **.RData** directory for the library. The remaining lines are S commands that read the **splus.library/\*.s** and **splus.library/\*.r** files into the library. The directory is then detached from the current session and is ready to be attached as a library.

# Appendix C

---

## SAS

SAS works identically on all platforms. Startup procedures differ slightly as they depend on details of the operating system. We give details for operating SAS with the HH examples, exercises, and code on the Windows and Unix platforms.

### C.1 Make the HH Library Available

Getting started is always the hardest part. The steps are simple. There are too many of them, and they must be done in the right order. The details are slightly different for each platform.

#### C.1.1 Windows

The steps are

1. We recommend that you copy the entire unzipped `hh` directory tree (see Preface Section 3) from the online files to your hard disk. We recommend that you use the `c:\HOME\hh` (`c:/HOME/hh`) directory on Windows. We are recommending that you create a new user `hh` with home directory parallel to your own directory. After you create this directory it will be made read-only.

On Windows, open two Windows Explorers, one to `c:\HOME\` and one to the unzipped online files. Pick up the `hh` directory from the online files and drop it into the `c:\HOME\`.

2. Create your working directory on your hard disk for the HH book. We recommend that you create a separate project for this book. We suggest the Windows directory name `c:\HOME\yourname\hh-user` (`c:/HOME/yourname/hh-user`) (one of us uses `c:/HOME/rmh/hh-user` and the other uses `c:/HOME/burt/hh-user`).  
Please note that `hh-user` is your directory and that it is distinct from the `hh` directory that contains the book's files.
3. Define the file `hh-user/hh.sas`. Construct your `hh-user/hh.sas` by editing one of the sample files in `sas.library/code`. We illustrate our Windows version (`sas.library/code/hh-pslmono-w.sas`) in Table C.1.
4. Open SAS in the `hh-user` directory. On Windows we recommend opening SAS with the SAS Windows interface.

#### SAS for Windows, by clicking on an icon.

Copy the existing SAS icon to the Desktop. Rename the copy to SAS `hh-user`. Right-click on the icon and select **Properties**. Select the **Shortcut** tab. Modify the **Target:** by adding the argument

`-autoexec c:\HOME\yourname\hh-user\hh.sas`

at the end. Modify the **startin:** to `c:\HOME\yourname\hh-user`  
Click OK.

We refer to this target as `hh-user/hh.sas`. It is the file defined above.

### C.1.2 Unix

The steps are

1. We recommend that you copy the entire unzipped `hh` directory tree (see Preface Section 3) from the online files to your hard disk. We recommend that you use the `/usr/users/hh` directory on Unix. We are recommending that you create a new user `hh` with home directory parallel to your own directory. After you create this directory it will be made read-only.

On Unix at the shell prompt,

```
cd /usr/users/hh  
(cd /CDROM/hh ; tar cf - *) | tar xf -
```

2. Create your working directory on your hard disk for the HH book. We recommend that you create a separate project for this book. On Unix we recommend the directory name `/usr/users/yourname/hh-user`

TABLE C.1. Sample version of the file **hh.sas** to tell SAS where the HH library sits on your computer.

We display the sample Windows version in the file (**sas.library/code/hh-pslmono-w.sas**). The sample Unix version is in file (**sas.library/code/hh-pslmono-u.sas**). The only difference between the Windows and Unix versions is the syntax of the **hh** pathname. SAS accepts forward slashes “/” in pathnames on both Unix and Windows.

The first three lines of the file (**%let; filename; options sasautos;**) are the critical ones that tell SAS where the HH files reside on your computer. The first line of the file gives the path of the directory where you copied the files from the HH online files. All file references in our code and transcript subdirectories book are relative to this path. This line must be changed to reflect the location where you copied the disk. The next two lines define and use the **filename hh\_lib**. This **filename** tells SAS how to find the macros we have provided. These two lines must not be changed.

The remaining **options** and **goptions** lines give the choices we made for the book. These choices are described in Section C.3.

---

```
SAS (sas.library/code/hh-pslmono-w.sas):
%let hh=c:/HOME/hh;                                /* this line must be customized */

filename hh_lib "&hh/sas.library/code"; /* this line must not be changed */
options sasautos=hh_lib;                            /* this line must not be changed */

/* everything after here is optional */
options noovp ;          /* prevents error messages in triplicate */
options formdlim=' ' ;   /* suppresses form feeds (physical page breaks)*/

options ls=73 ;      /* makes tables fit on letter paper in portrait*/
options ps=32767 ;    /* minimizes page breaks */
options nocenter ;   /* puts tables on left margin of page */
* options nodate ;   /* suppresses time stamp */

/*
  targetdevice specifies the printer.
  Graphs will appear on the screen in the
  standard screen format for your setup.
*/
goptions targetdevice=pslmono /* pslmono is PostScript in black and white*/
  gsfname=grafout
  gsfmode=replace
  gaccess=sasgastd
  csymbol=black ; /* csymbol makes 'symbol v=' work correctly */

run;
```

---

(one of us therefore uses `/usr/users/rmh/hh-user` and the other uses `/usr/users/burt/hh-user`).

Please note that this is your directory and that it is distinct from the `hh` directory that contains the book's files.

3. Define the file `hh-user/hh.sas`. Construct your `hh-user/hh.sas` by editing the sample files in (`sas.library/code/hh-pslmono-w.sas`). See Table C.1 for details.
4. Open SAS in the `hh-user` directory. On Unix we recommend either opening SAS with the SAS windowing interface or using SAS batch.

**SAS for Unix, by defining autoexec.sas.**

Rename the file `hh-user/hh.sas` defined above to the new name `$HOME/autoexec.sas`.

**SAS for Unix, by creating a shell script.**

Create a shell script that starts SAS with the `hh-user/hh.sas` file defined above. For example, see Table C.2.

## C.2 Using SAS with HH

All file references in our SAS code examples are relative to the directory "`&hh`". The user MUST modify the first line of one of the files

`(sas.library/code/hh-pslmono-w.sas)` (Windows)

`(sas.library/code/hh-pslmono-u.sas)` (Unix)

as described in Section C.1, to point to the directory where the files are installed on the user's machine and save the modified file as `hh-user/hh.sas`.

### C.2.1 Reading HH Datasets

Use commands like

TABLE C.2. Shell script to start SAS on a Unix machine in the `hh-user` directory. This script will start an interactive SAS session if you are running on an X-Windows terminal or a batch session if you give a `.sas` file as argument. Remember to make the shell script executable.

---

SAS (`sas.library/code/sas-hh.sh`):  
## shell script to start SAS on Unix and initialize with the hh.sas file.  
`sas -autoexec $HOME/hh-user/hh.sas $*`

---

```

data gunload;
  infile "&hh/datasets/gunload.dat" firstobs=2;
  input method group team rounds;
run;

```

### C.2.2 Any Other Data Files

You can access any file on the computer by giving the complete pathname of the file in the **infile** statement. If the file is in your working directory, you can give just the filename.

### C.2.3 ASCII Data Files with TAB Characters

Data files are normally ASCII with columns neatly aligned and separated by spaces. The **infile** statement works without difficulty in that case.

When the file has TABS, a normal occurrence with many editors, the DATA step gives relatively uninterpretable messages, for example:

```

NOTE: Invalid data for b in line 1 3-5.
RULE:      -----1-----2-----3-----4-----+
           2   CHAR  4 5 . 6 5
           ZONE  32303
           NUMR  40596
           a=1  b=.  c=4 _ERROR_=1 _N_=1

```

The data in the file consists of

```

1 2TAB3
4 5TAB6

```

The three characters “2TAB3” in line 1 columns 3–5 are identified as a single item and labeled as “invalid data” and shown as missing (in the statement **b=.**). The complete line 2 (as seen by SAS) is displayed in hex code. The “4 5.6” in columns 1–5 of the line 2 are uninterpretable by SAS. The ZONE–NUMR 09 corresponding to the “.” is the hex code for CTRL-I, the TAB character.

The fix is to tell SAS that the TABs are there, with the **expandtabs** option to the **infile** statement, for example

```

data test;
  infile "datatabs.dat" expandtabs;
  input a b c;
run;

```

#### C.2.4 Windows and Unix EOL (End-of-Line) Conventions

SAS runs on both Unix computers and Microsoft Windows computers. The most important difference between the Windows and Unix operating systems is the end-of-line (EOL) convention. Unix uses the only the line-feed character (LF). Windows uses both the carriage-return and line-feed characters (CR LF). When you initially save files (.sas or .dat files usually) on a Windows machine and then attempt to use them on a Unix machine, you will get uninterpretable messages from SAS. Here is an example:

The original data line has 5 characters

1 2 3

The SAS log complains about the 6<sup>th</sup>:

The “.” in position 6 is the CR. The ZONE-NUMR value OD is the hex code for the CTRL-M, the CR character.

The fix is to convert the files to Unix EOL before running SAS.

You can convert them on the PC before sending the files if you have a smart editor.

Or you can convert them on the Unix machine before running SAS on your file using the appropriate utility. For example, on “Compaq Tru64 UNIX” you can run the program

```
/usr/bin/mtools/dos2unix myfile.sas
```

at the Unix prompt before you run

sas mvfile

If you are using a smart editor, you can read the `myfile.log` and `myfile.1st` files produced by a Unix machine on your PC without converting to DOS line endings. If you need to convert, use the inverse functions on the Unix machine before bringing the files back to the PC:

/usr/bin/mtools/unix2dos hw0206b.log

/usr/bin/mt800ls/unix2dos hw0206b.lst

Don't touch the `sasgraph.ps` file. It works correctly as is.

### C.3 Macros

(`sas.library/code/hh-pslmono-w.sas`) (Windows):

(`sas.library/code/hh-pslmono-u.sas`) (Unix):

The primary purpose of these initialization files is to define the pathname of the root directory for the `hh` files. See the discussion in Table C.1 and in Section C.2.

We also set several options and goptions in `hh-user/hh.sas`:

`ls=73`: 73 is the minimum linesize that allows an ANOVA table to be printed without folding.

`ps=32767`: This is the maximum allowable `pagesize`. We wish to minimize the number of page breaks that SAS inserts in the output.

`nocenter`: SAS by default centers output on the page. We want it left-justified because it works better with the margins of our book.

`nodate`: For the text we suppress the dates. For your work with the text and for clients, you probably want to display the dates.

`targetdevice=pslmono`: We want our figures in black and white PostScript for the book. The `targetdevice`, as opposed to `device`, tells the interactive SAS graphics device that we will be saving many of the figures in PostScript format.

(`sas.library/code/ischeffe.sas`): Macro for the Scheffé multiple comparison procedure. See the discussion in Section 7.1.4.1.

## C.4 Learning the SAS Language

There are several excellent sources for learning the SAS language. The most immediate is the online documentation that comes with the SAS system on the SAS OnlineDoc CD. With a full installation, it is found from the SAS window by clicking Help/Books and Training/Online Doc. Then we recommend looking at the Contents tab in the left frame of the browser and going to the Base SAS Software section for a discussion of the language. You might also look at the Overview section, particularly for the list of the many hardcopy books available from SAS Institute Publications.

## C.5 SAS Coding Conventions

We have adopted two coding conventions for our SAS code.

1. All file references in our SAS code examples are relative to the directory "&hh". You must construct and use the file `hh-user/hh.sas`, as described in Section C.1, to resolve this reference. Once the "&hh" macro has been defined in `hh-user/hh.sas`, then all code in the `*/code/*.sas` files will work as written. NO changes are needed for ANY code files in the book once "&hh" has been correctly defined.
2. All of our SAS code samples include the `run;` statement after each `PROC` and `DATA` step. This allows us to submit a section of code that will run immediately when we highlight it and click `RUN`.

# Appendix D

---

## Probability Distributions

### D.1 Common Probability Distributions with S-PLUS and SAS Commands

We list, with some discussion, several common probability distributions. We include S-PLUS and SAS commands for working with these distributions. We use **lowercase** for S-PLUS and **UPPERCASE** for SAS.

**binomial:** This distribution was introduced in Section 3.4.1. If  $X$  has a binomial distribution with parameters  $n$  and  $p$ , then

$$P[X \leq x | n, p] = \text{pbinom}(x, n, p) = \text{PROBBNML}(p, n, x) = \mathcal{F}_{Bi}(x | n, p)$$
$$\text{pbinom}(4, 5, .5) = 0.96875$$
$$\text{qbinom}(.96875, 5, .5) = 4$$

**hypergeometric:** This distribution is used in Chapter 15. We sample  $n$  items *without* replacement from a population of  $N$  items comprised of  $M$  successes and  $N - M$  failures. Then the number of successes  $X$  observed in the population is said to have a hypergeometric distribution with parameters  $N, M$ , and  $n$ .

$$P(X = x | N, M, n) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}$$
$$P(X = x | N, M, n) = \text{dhyper}(x, M, N-M, n) = \text{PROBHYP}(N, M, n, x)$$
$$P(X \leq x | N, M, n) = \text{phyper}(x, M, N-M, n)$$
$$\text{dhyper}(1, 9, 7, 6) = \text{PROBHYP}(16, 9, 6, 1) = 0.0236$$

**Poisson:** Let the random variable  $X$  be the number of occurrences of some event that are observed in a unit of time, volume, area, etc., and let  $\lambda$  be the mean number of occurrences of the event per unit, assumed to be constant throughout the process that generates the occurrences. Suppose that the occurrence(s) of the event in any one unit are independent of the occurrence(s) of the event in any other non-overlapping unit. Then  $X$  has a Poisson distribution with parameter  $\lambda$ .

$$P[X \leq x | \lambda] = \text{ppois}(x, \lambda) = \text{POISSON}(\lambda, x) = \mathcal{F}_{\text{Poi}}(x | \lambda)$$

`ppois(3, 5) = 0.2650259`

`qpois(0.2650259, 5) = 3`

**beta:** This two-parameter distribution is often used to model phenomena restricted to the range  $(0, 1)$ , for example sample proportions. It is used in Section 5.1.2 to construct alternative one-sided confidence intervals on a population proportion.

$$P[X \leq x | a, b] = \text{pbeta}(x, a, b) = \text{CDF}('BETA', x, a, b) = \mathcal{F}_{\text{Be}}(x | a, b)$$

`pbeta(0.6, .5, 2) = 0.9295`

`qbta(.9295, .5, 2) = 0.6`

**exponential:**  $\mu$  is both the mean and s.d. of this distribution. S-PLUS parameterizes the exponential distribution with the rate  $1/\mu$ , the reciprocal of the mean  $\mu$ . Times between successive Poisson events with mean rate of occurrence  $\mu$  have this distribution. The exponential distribution is the only distribution with the “lack of memory” or “lack of deterioration” property, which states that the probability that an exponential random variable exceeds  $t_1 + t_2$  given that it exceeds  $t_1$  equals the probability that it exceeds  $t_2$ .

$$P[X \leq x | \mu] = 1 - e^{-x/\mu} = \text{pexp}(x, 1/\mu) = \mathcal{F}_{\text{exp}}(x | \mu).$$

`pexp(1, .6) = 0.4511884`

`qexp(0.4511884, .6) = 1`

SAS seems not to have the exponential distribution.

**normal:** This distribution was introduced in Section 3.4.2. If  $Z$  is standard normal  $N(0, 1)$ ,

---

`pnorm(z) = PROBNORM(z) =  $P[Z \leq z] = \Phi(z) = \mathcal{F}_{\text{norm}}(z | 0, 1)$`   
`pnorm(1.645) = PROBNORM(1.645) =  $P[Z \leq 1.645] = 0.95 = \Phi(1.645)$`   
`qnorm(.95) = PROBIT(0.95) = 1.645 = \Phi^{-1}(0.95) = z_{.05}`  
 We can generate a random  $N(\mu = 6, \sigma^2 = 16)$  r.v. with  
`rnorm(1, mean=6, sd=4) = 6+4*RANNOR(seed)`

(Student's)  $t$ : This distribution was introduced in Section 3.4.3. If  $t$  has a Student's  $t$  distribution with 23 degrees of freedom,

`pt(1.714, 23) = PROBT(1.714, 23) = 0.95 = \mathcal{F}_t(1.714 | 23)`  
`qt(0.95, 23) = INVT(0.95, 23) = 1.714 = t_{.05, 23}`

**Studentized Range Distribution:** This distribution is used in the Tukey multiple comparisons procedure discussed in Section 6.3. Let  $\bar{y}_{(1)}$  and  $\bar{y}_{(a)}$  denote the smallest and largest means of samples of size  $n$  drawn from  $a$  populations having a common variance  $\sigma^2$ , and let  $s = \sqrt{\text{MS}_{\text{Res}}}$  be the estimate of  $\sigma$  calculated from the ANOVA table, for example, Table 6.3. Then the random variable

$$Q = \frac{\bar{y}_{(a)} - \bar{y}_{(1)}}{s/\sqrt{n}}$$

has a Studentized range distribution with parameters  $a$  and  $\text{df}_{\text{Res}} = a(n - 1)$ . Note that the Studentized range distribution is defined on the domain  $0 \leq q < \infty$ .

$\mathcal{F}_Q^{-1}(0.95 | 4, 12) = 4.199$   
`=qtukey(.95, 4, 12) = PROBMC("RANGE", . , .95, 12, 4)`  
  
 $\mathcal{F}_Q(4.199 | 4, 12) = 0.95$   
`=ptukey(4.199, 4, 12) = PROBMC("RANGE", 4.199, . , 12, 4)`

The command `ptukey` is available in R but not in S-PLUS.

(Central)  $\chi^2$  (chi-square): The central  $\chi^2$  distribution with  $k$  degrees of freedom is the distribution of the sum of squares of  $k$  independent standard normal r.v.'s. If  $k > 2$ , a  $\chi^2$  r.v. has a unimodal, positively skewed PDF starting at zero and asymptotically tapering to the horizontal axis for large values. The mean of this distribution is  $k$ , and  $k$  is also approximately its median if  $k$  is large. The r.v.  $[(n - 1)s^2]/\sigma^2$ , where  $s^2$  is the variance of a normal sample, has a  $\chi^2$  distribution with  $n - 1$  degrees of freedom.

This distribution is used in inferences about the variance (or s.d.) of a single population and as the approximate distribution of many non-

parametric test statistics, including goodness-of-fit tests and tests for association in contingency tables.

$$\begin{aligned} \text{pchisq}(18.31, 10) &= \text{PROBCHI}(18.31, 10) = 0.95 = \mathcal{F}_{\chi^2}(18.31 | 10) \\ \text{qchisq}(0.95, 10) &= \text{CINV}(0.95, 10) = 18.31 = \chi^2_{0.05, 10} \end{aligned}$$

**F:** The  $F$  distribution is related to the  $\chi^2$  distribution. If  $U_i$ , for  $i = \{1, 2\}$ , is a  $\chi^2$  r.v. with  $\nu_i$  degrees of freedom, and if  $U_1$  and  $U_2$  are independent, then  $F = U_1/U_2$  has an  $F$  distribution with  $\nu_1$  and  $\nu_2$  df. This distribution is extensively used in problems involving the comparison of variances of two normal populations or comparisons of means of two or more normal populations.

$$\begin{aligned} P(F \leq f) &= \text{pf}(f, \nu_1, \nu_2) = \text{PROBF}(f, \nu_1, \nu_2) = \mathcal{F}_F(f | \nu_1, \nu_2) \\ f &= \text{qf}(1-\alpha, \nu_1, \nu_2) = \text{FINV}(1-\alpha, \nu_1, \nu_2) = F_{\alpha, \nu_1, \nu_2} \end{aligned}$$

**lognormal:** A r.v.  $X$  is said to have a *lognormal* distribution with parameters  $\mu$  and  $\sigma$  if  $Y = \ln(X)$  is  $N(\mu, \sigma^2)$ ; i.e., if  $Y$  is normal, then  $e^Y$  is lognormal. This is a positively skewed unimodal distribution defined for  $x > 0$ . It is commonly used as a good approximation for positively skewed data, such as a distribution of income.

$$\begin{aligned} P(X \leq x) &= \text{plnorm}(q, \text{mean}=\mu, \text{sd}=\sigma) = \mathcal{F}_{\text{lognormal}}(z|\mu, \sigma) \\ \text{plnorm}(5.180252, 0, 1) &= 0.95 \\ \text{qlnorm}(.95, 0, 1) &= 5.180252 \end{aligned}$$

**multinomial:** The (discrete) *multinomial* distribution is a generalization of the binomial distribution to the case of  $k > 2$  categories. Suppose there are  $n$  independent trials, each of which can result in just one of  $k$  possible categories such that  $p_j$  is the probability of resulting in the  $j^{\text{th}}$  of these  $k$  categories. (Hence  $p_1 + p_2 + \dots + p_k = 1$ .) Let  $X_j$  be the number of occurrences in category  $j$ . Then the vector  $(X_1, X_2, \dots, X_k)$  is said to have a multinomial distribution with parameters  $n, p_1, p_2, \dots, p_k$ . Its PMF is

$$P(X_j = x_j | j = 1, \dots, k) = \frac{n! p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}}{x_1! x_2! \dots x_k!}, \quad x_1 + x_2 + \dots + x_k = n$$

If a proportion  $p_j$  of a population of customers prefers product number  $j, j = 1, \dots, k$ , among  $k$  products, then the multinomial distribution provides the probability of observing any particular configuration of preferences among a random sample of  $n$  customers.

### D.1.1 An Example Involving Calculations with the Binomial Distribution

Suppose that 10% of the residents of a certain county have been exposed to a particular disease. If the 8,000 residents of a city in this county can be viewed as a random sample from this county, find the probability that the number of exposed city residents is between 770 and 820, inclusive. This may be written as

$$P(770 \leq X \leq 820) = P(X \leq 820) - P(X \leq 769)$$

*Solution in SAS:*

`PROBBNML(p,n,x)` produces the probability of  $x$  or fewer successes. Then with a few interactive lines of code, we find

$$\text{PROBBNML(.1, 8000, 820)} - \text{PROBBNML(.1, 8000, 769)} = 0.6506956$$

*Solution in S-PLUS:*

`pbinom(x,n,p)` produces the probability of  $x$  or fewer successes. Then

$$\text{pbinom}(820, 8000, .1) - \text{pbinom}(769, 8000, .1) = 0.6506956.$$

## D.2 Noncentral Probability Distributions

In hypothesis testing, except in special cases such as testing with the normal distribution, one deals with a central distribution when the null hypothesis is true and an analogous noncentral distribution when the null hypothesis is false. Thus calculations of probabilities under the alternative hypothesis, as are required when doing Type II error analysis and when constructing O.C. and power curves, necessitate the use of noncentral distributions.

The forms of the  $t$ , chi-square, and  $F$  distributions we've considered thus far have all been central distributions. For example, if  $\bar{X}$  is the mean and  $s$  the standard deviation of a random sample of size  $n$  from a normal population with mean  $\mu$ ,  $t = (\bar{x} - \mu)/(s/\sqrt{n})$  has a central  $t$  distribution with  $n - 1$  df. Suppose, however, that the population mean is instead  $\mu_1$ , different from  $\mu$ . Then  $t$  above is said to have a noncentral  $t$  distribution with  $n - 1$  df and a *noncentrality parameter* proportional to  $((\mu - \mu_1)/\sigma)^2$ . (If  $\mu = \mu_1$ , so that the noncentrality parameter is zero, the noncentral  $t$  distribution reduces to the central  $t$  distribution.)

A noncentral chi-square ( $\chi^2$ ) r.v. is a sum of squares of independent normal r.v.'s each with s.d. 1 but at least some of which have a nonzero mean. A

noncentral  $F$  r.v. is the ratio of a noncentral chi-square r.v. to a central chi-square r.v., where the two chi-squares are independent.

For tests using the  $t$ , chi-square, or  $F$  distribution, the power of the test (protection against Type II errors) is an increasing function of the noncentrality parameter.

Noncentral distributions are specified with one more parameter than their corresponding central distribution. Consequently, tabulations of their cumulative distribution function appear much less frequently than those for central distributions. Fewer statistical software packages include them. There is no noncentral normal distribution. The distribution under the alternative hypothesis is just an ordinary normal distribution with a shifted mean.

The SAS functions for noncentral  $t$ , chi-square, and  $F$  CDFs and the S-PLUS functions for noncentral chi-square and  $F$  CDFs are the same as those for the corresponding central distribution with the addition of an argument for the noncentrality parameter `nc`, for example,

SAS : `PROBCHI(x, df, nc)`  
S-PLUS : `pchisq(x, df, nc)`

The noncentrality parameter defaults to zero (hence to a central distribution) if it is not specified.

# Appendix E

---

## Editors

An excellent text editor is an indispensable tool for the statistical analyst. The editor is the single program in which we spend most of our time. We use it for looking at raw data, for writing commands in the statistical languages we use, for reading the output tables produced by our statistical programs, for writing reports, and for reading and writing correspondence about our studies to our clients, consultants, supervisors, and subordinates.

As we indicated in the Preface, our goal in teaching statistical languages is to make the student aware of the capabilities of the language for describing data and their analyses. The language is approached through the text editor.

### Requirements

These are the requirements we place on any text editor that is to be used for interacting with a computing language:

1. Permit easy modification of computing instructions and facilitate their resubmission for processing
2. Be able to operate on output as well as input
3. Be able to cut, paste, and rearrange text; to search documents for strings of text; to work with rectangular regions of text
4. Be aware of the language, for example, to illustrate the syntax with appropriate indentation and with color and font highlighting, to detect

syntactic errors in input code, and to check the spelling of statistical keywords

5. Handle multiple files simultaneously
6. Interact cleanly with the presentation documents (reports and correspondence) based on the statistical results
7. Check spelling of words in natural languages
8. Permit placement of graphics within text.
9. Permit placement of mathematical expressions in the text.

We discuss two contemporary possibilities, Emacs in Section E.3 and MS Word in Section E.4. We believe Emacs is best, but we acknowledge that MS Word is most prevalent.

## E.1 Working Style

Our personal working style (and the one we recommend to our readers) is to write a file of commands that specify the analysis, run the commands and review the graphs and tables so produced, and then correct and augment the analysis specifications. We construct the command file interactively. Initially we write the code to read the data into the program and to prepare several graphs and the initial tables. If there are errors (in typography or in programming logic), we correct them in the file and rerun the commands. As we progress and gain insight into what the data say, we add to the code to produce additional graphs and tables. When we have finished, we have a file of commands that could be run as a batch job to produce the complete output. We also have a collection of graphs and tables that are stored in additional files.

Several file structures are possible for maintaining the commands, the tables, and the graphs. We prefer a system as described above, and as reflected in the online files accompanying this book. We have command language files (with names `myfile.s` or `myfile.sas`), transcript files (`myfile.st` for S-PLUS and `myfile.log` and `myfile.lst` for SAS), and graph files (usually `myfile.n.ps.gz` or `myfile.n.eps.gz` for PostScript). The base name (`myfile` in this example) is chosen to reflect the content of the file. Thus the files for the first regression chapter are called `rega.*`.

Our organizational structure uses the directory structure provided in either the Unix or MS Windows operating systems. We have a directory for each chapter, and subdirectories under each chapter for code, figures, transcripts (and several additional subdirectories for different categories of working

files). The standard file extensions `.sas`, `.s`, etc. are used to distinguish the different types of files within each subdirectory.

This book is written in L<sup>A</sup>T<sub>E</sub>X, with one file `chapter.tex` for the text of each chapter. The file `chapter.tex` references the figures in the PostScript files `figure/chapter.*.eps.gz`, the code in files `code/chapter.*`, and the transcripts in files `transcript/chapter.*` and, after processing, produces a “device-independent file” `chapter.dvi` that incorporates the text and figures. We would write a smaller report the same way.

A differently appearing working style, which has some important features in common with our preferred style, is to use a single file for organization. A single file `myfile.doc` for a complex editor such as MS Word can maintain within itself all the distinct pieces discussed above. The author would use the file, and embedded sections within the file, in the same way that we use the directory structure.

## E.2 Typography

Reports based on computer printout must be typed correctly. We recommend L<sup>A</sup>T<sub>E</sub>X (the standard required by many statistics and mathematics journals, and the typesetting package with which we wrote this book). We do accept other word processing software. Whichever software you use, you must use it correctly. Specific style issues that you must be aware of are

**Fonts.** Computer listings in S-PLUS and SAS are designed for monowidth fonts (such as Courier) and are unreadable in Times Roman. English text looks best in a proportional font such as Times Roman. **In our classes, we return UNREAD any papers that use a proportional font for computer listings.** In L<sup>A</sup>T<sub>E</sub>X, programs and transcripts must be placed in a `verbatim` environment. In MS Word, programs and transcripts must be highlighted and then explicitly placed into Courier.

Here is an example of the issue. The Courier rendition is consistent with the design of the output by the program designer. The Times Roman is exactly the same text dropped into an environment that is incorrectly attempting to space it in accordance with English language typesetting rules.

TABLE E.1. Correct usage of all four dashlike symbols (---) and the keys to generate them in L<sup>A</sup>T<sub>E</sub>X and MS Word.

| Symbol | Use     | Example       | L <sup>A</sup> T <sub>E</sub> X | MS Word                         |
|--------|---------|---------------|---------------------------------|---------------------------------|
| -      | hyphen  | compound word | <i>t-test</i>                   | -                               |
| -      | en dash | range         | 100–120                         | -- ctrl-num -                   |
| -      | minus   | negation      | -12                             | \$-\$ Insert-menu/symbol... / - |
| —      | em dash | apposition    | punctuation—like this           | --- alt-crtl-num -              |

| Courier (correct spacing)   | Times Roman (incorrect spacing)   |
|---|---|
| <pre>&gt; summary(ex0221)       weight      code       Min.:23.20  1:35       1st Qu.:24.75 2:24       Median:25.70       Mean:25.79       3rd Qu.:26.50       Max.:31.10</pre> | <pre>&gt; summary(ex0221) weight code Min.:23.20 1:35 1st Qu.:24.75 2:24 Median:25.70 Mean:25.79 3rd Qu.:26.50 Max.:31.10</pre> |

**Alignment.** Numbers in a column are to be aligned on decimal points. Alignment makes it possible to visually compare printed numbers in columns. There are two reasons for getting it wrong. One is carelessness. The other is blind copying from a source that gets it wrong. The most typical poor source is Excel spreadsheets that use the default formatting. If you must use Excel, make sure that the columns are formatted with aligned decimal points. See Section E.5 for further comments on Excel.

| Correct | Wrong  |
|---------|--------|
| 123.45  | 123.45 |
| 12.34   | 12.34  |
| -4.32   | -4.32  |
| 0.12    | 0.12   |

**Minus signs and dashes.** There are four distinct concepts that have four different typographical symbols in well-designed fonts. On typewriters all four are usually displayed with the symbol “-” that appears on the hyphen key (next to the number 0). You are expected to know the difference and to use the symbols correctly.

Table E.1 shows an example of the correct usage of all four symbols and the keys in L<sup>A</sup>T<sub>E</sub>X and MS Word.

The misuse of dashes that touches my (rmh) hot button the most is misuse of hyphen when minus is meant, for example

| correct | WRONG |
|---------|-------|
| +12.2   | +12.2 |
| -12.2   | -12.2 |

**Right margins and folding.** Table E.2 intentionally misuses formatting to illustrate how bad it can look. This usually occurs when the S-PLUS window width or SAS linesize is wider than your word processor is willing to work with. Verify that you picked a width consistent with your word processor. You can make the width narrower in S-PLUS either by manually grabbing the right margin of the Commands window and making it narrower, or by using the S-PLUS command

```
options(width=72)
```

You can make the SAS listing file narrower by using the command

```
options linesize=73;
```

We suggest 73 for SAS because 72 folds the ANOVA table.

## E.3 Emacs and ESS

Emacs (Stallman, 2000) is a mature, powerful, and easily extensible text editing system freely available under the GNU General Public License for a large number of platforms, including Unix, Mac, and Windows. Emacs shares some features with word processors and, more importantly, shares many characteristics with operating systems. Most importantly, Emacs can interact with and control other programs either as subprocesses or as cooperating processes.

The name “Emacs” is an acronym for *Editing MACroS*. It comes from the Free Software Foundation (for which Richard Stallman got a MacArthur genius award) also known as the gnu project (Gnu is Not Unix).

Emacs provides facilities that go beyond simple insertion and deletion: viewing two or more files at once; editing formatted text; visual comparison of two similar files; and navigation in units of characters, words, lines, sentences, paragraphs, and pages. Emacs knows the syntax of each programming language. It can provide automatic indentation of programs and highlight with fonts or colors specified syntactic characteristics. Emacs is extensible using a dialect of Lisp (Chassell, 1999, Graham, 1996). This means that new functions, with user interaction, can be written for common and repeated text editing tasks.

TABLE E.2. Intentional misuse of formatting. Never turn in anything that looks like either of these.

Do not allow the right margins of your work to run off the edge of the page. It is hard to read text that isn't visible.  
 Do not allow lines to be arbitrarily folded in a way that destroys the formatting. This is particularly a problem if you use the default printing from the S-PLUS Commands window through Notepad. Use S-mode in Emacs and get it right. Your other option is to cut and paste into Word or Wordpad and force the font to courier.

Folding makes this table impossible to read.

| Source                 | DF        | Squares            | Mean Squ  |
|------------------------|-----------|--------------------|-----------|
| re                     | F Value   | Pr > F             |           |
| Model                  | 2         | 2649.816730        | 1324.9083 |
| 65                     | 54.92     | <.0001             |           |
| Error                  | 44        | 1061.382419        | 24.1223   |
| 28                     |           |                    |           |
| <b>Corrected Total</b> | <b>46</b> | <b>3711.199149</b> |           |

### Buffers

Emacs can work with many multiple files simultaneously. It brings each into a *buffer*. A buffer is copy of a file within the Emacs editor. Any editing changes made to the contents of a buffer are temporary until the buffer is saved back into the file system. A buffer can hold a file, a directory listing on the local computer, a directory listing on a remote computer, an interactive session with the operating system, an interactive instance of another running program, a *telnet* session to a remote computer, or an *ftp* session to a remote computer.

Emacs allows you to open and edit an unlimited number of files and login sessions simultaneously, running each in its own buffer. The files or login sessions can be on another computer, anywhere in the world. You can run simultaneous multiple sessions even over dial-up lines (rmh often does). The size of a buffer is limited only by the size of the computer. You can have as few as one buffer, or as many as your monitor permits, visible at the same time. One of us (rmh) normally has several buffers visible and

frequently has hundreds of open buffers (several chapters, their code files, their transcript files, the control buffer for S-PLUS, `telnet` sessions to two or three remote computers, directory listings on the remote computers, and a listing of the currently open buffers).

### *Shell Mode*

Emacs includes a *shell mode* in which a terminal interaction runs inside an Emacs buffer. The Unix terminology for the program that runs an interactive command line session is a “shell”. There are several commonly used shell programs: `sh` is the original and most fundamental shell program. Other shell programs are `csh` and `bash`. We usually use the Cygwin version of `bash` as our shell under MS Windows. The MS-DOS prompt window is the native shell program in MS Windows.

A terminal interaction running inside an Emacs buffer is much more powerful than one run in an ordinary terminal emulator window. The entire live login session inside an Emacs buffer is just another editable buffer (with full search capability). The only distinction is that both you and the computer program you are working with can write to the buffer. This is exceedingly important because it means nothing ever rolls off the top of the screen and gets lost. Just roll it back. The session can be saved to a file and then is subject to automatic backup to protect you from system crash or loss of connection to a remote machine.

ESS (see Section E.3.1) builds on shell mode to provide modes for interacting with statistical processes. The terminal interaction can be local (on the same computer on which Emacs is running) or remote (anywhere else, through `telnet` or a secure transport mechanism such as `ssh`).

### *Text Editing Features*

Most programming and documentation tasks fall under the realm of text editing. This work is enhanced by features such as contextual highlighting and recognition of special reserved words appropriate to the programming language in use. In addition, editor behaviors such as folding, outlining, and bookmarks can assist with maneuvering around a file. Typesetting and word processing, which focus on the presentation of a document, are tasks that are not pure text editing. Emacs shares many features with word processing programs and cooperates with document preparation systems such as L<sup>A</sup>T<sub>E</sub>X, HTML, and XML).

We strongly recommend that students in our graduate statistics classes use Emacs as their primary text editor. The primary reason for this recommendation is that Emacs is the only general editor we know of that understands the syntax and formatting rules for the statistical languages S-PLUS and

SAS that we use in our courses. Emacs has many other advantages (listed above), as evidenced by Richard Stallman having won a MacArthur award for developing Emacs.

### E.3.1 ESS

ESS ((Rossini et al., 2004b), (Heiberger, 2001), (Rossini et al., 2004a)) extends Emacs to provide a functional, easily extensible, and uniform interface for multiple statistical packages. Currently ESS works with S-PLUS, R, SAS, STATA, and XLISPSTAT. The online documentation includes an introduction in file `ESS/ess/doc/intro.tex` and two introductory talks in files `ESS/ess/doc/rmh-talk.tex` and `ESS/ess/doc/ajr-talk.tex`. Online help is available from within Emacs by entering `C-h i` and paging down to `ESS`, pressing `ENTER` and then following the info items.

The discussion here is based on (Rossini et al., 2004a). ESS provides:

#### *Syntactic Indentation and Color/Font-Based Source Code Highlighting*

The ESS interface includes a description of the syntax and grammar of each statistical language it knows about. This gives ESS the ability to edit the programming language code, often more smoothly than with editors distributed with the languages. The process of programming code is enhanced as ESS provides the user with a clear presentation of the code with syntax highlighting to denote assignment, reserved words, strings, and comments.

#### *Partial Code Evaluation*

Emacs can send individual lines, entire function definitions, marked regions, and whole edited buffers from the window in which the code is displayed for editing to the statistical language/program for execution. Emacs sends the code directly to the running program and receives the printed output back from the program in an editable Emacs buffer. This is a major improvement over cut-and-paste as it does not require switching buffers or windows.

#### *Object Name Completion*

In addition, for languages in the S family (S developed at Bell Labs, S-PLUS, and R) ESS provides object-name completion of both user- and system-defined functions and data.

#### *Source Code Checking*

ESS facilitates the editing of source code by providing a means for loading and error-checking of small sections of code. This allows for source-level debugging of batch files.

### *Process Interaction*

Emacs has historically referred to processes under its control as “inferior”, accounting for the name inferior ESS (**iESS**) to denote the mode for interfacing with the statistical package. The output of the package goes directly to an editable text buffer in Emacs. This mode allows for command-line editing and saving history, as well as recalling and searching for previously entered commands. Filename completion is available. In addition (currently only for S languages), there exists object-name and function-name completion. Transcripts are easily recorded and can be edited into an ideal activity log, which can then be saved. There is a good interface for handling and intercepting calls to the internal help systems for S-PLUS, R, XLISPSTAT, and STATA.

### *Interacting with Statistical Programs on Remote Computers*

ESS provides the facility to edit and run programs on remote machines in the same session and with the same simplicity as if they were running on the local machine. The remote machine could be a very different platform than the local machine.

### *Transcript Editing and Reuse*

Once a transcript log is generated, perhaps by saving an **iESS** buffer, transcript-mode assists with reuse of part or all of the entered commands. It permits editing and re-evaluating the commands directly from the saved transcript. This is useful for demonstration of techniques as well as for reconstruction of data analyses. There currently exist functions within ESS for cleaning transcripts from the S-PLUS and SAS languages back to source code by finding all input lines and isolating them into an input file.

### *Help File Editing (R)*

ESS also provides an interface for writing help files for R functions and packages. It provides the ability to view and execute embedded R source code directly from the help file in the same manner as ESS normally handles code from a source file. **Rd** mode provides syntax highlighting and the ability to submit code to a running ESS process, either R or S-PLUS.

## E.3.2 Mouse and Keyboard

Recent versions of Emacs provide a GUI (graphical user interface) or terminal interface that can respond to both keyboard and mouse commands. We recommend the keyboard as it is faster (once learned) and healthier (the user is less likely to suffer from repetitive motion injuries). The mouse-based interface, through menus and toolbars (toolbars available currently

with XEmacs), tries to facilitate the learning of keystroke-based shortcuts. Additional user-defined menus and toolbars can be constructed by the user as needed.

### E.3.3 Learning Emacs

There are several ways to learn Emacs, most are online and therefore require that you open Emacs to use them.

1. Tutorial. Enter **C-h t**. Then read the file and follow the instructions. You are working with your own private copy of the **TUTORIAL** file, so you can practice the keystrokes as suggested.
2. Manual. The manual is online in the hyperlinked Info system. Enter **C-h i** to bring up the **Info:** buffer. Move the cursor to the **\* Emacs:** line (by mouse or arrow keys) and press **RET**. Continue by placing the cursor on highlighted topics and pressing **RET**. A hard copy of the manual can be ordered from the Free Software Foundation. See the file **/emacs/emacs-\*/etc/ORDERS** for details.
3. Help. To find help apropos a topic, for example to answer the question “How do I save my current editing buffer to a file?”, enter **C-h a save RET** and get a list of all commands with **save** as part of their name. You probably want the command **save-buffer** and will see that you can use that command by typing **C-x C-s** or by using the FILES pull-down menu.
4. Reference Card. The Emacs reference card is in the file **/emacs/emacs-\*/etc/refcard.ps**. You can view it on-screen (with Ghostview) or on paper. The refcard can be made to fit on a single sheet of paper by executing the Unix shell script (**edit/code/refcard.sh**) or Windows batch file (**edit/code/refcard.bat**). In both files, verify both the version of Emacs and the path to the destination file.

### E.3.4 Requirements

Emacs satisfies the requirements detailed above.

1. ESS provides full interaction between the commands file, the statistical process, and the transcript.
2. The statistical process runs in an Emacs buffer and is therefore fully searchable and editable.
3. These are standard editing features for any buffer.

4. Emacs comes with modes specialized for all standard computing languages, ESS provides the modes for S-PLUS and SAS and by default highlights all keywords in the language (so if a word isn't highlighted it probably isn't a keyword).
5. Emacs handles multiple files as part of its basic design.
6. Emacs has a L<sup>A</sup>T<sub>E</sub>X mode, and therefore provides the best mathematical typesetting system currently available anywhere.
7. Emacs has several text modes.
8. Several spell-check programs works with Emacs; we use `ispell`.
9. Graphics in PostScript and bitmap formats can be embedded into L<sup>A</sup>T<sub>E</sub>X documents. Emacs 21 on Unix computers permits embedding of graphics directly into the Emacs buffer. The Windows version of Emacs 21 does not yet (April 2004) include graphics.

## E.4 Microsoft Word

When we use MS Word for programming, it is critical to use a monowidth font (a font such as Courier in which all letters are equally wide) and not a proportional font (such as Times Roman). The reason for this imperative is that the primary output for both S-PLUS and SAS is designed to look right in a monowidth font. See Section E.2 for an illustration of correct and incorrect font choice.

### E.4.1 Learning Word

Word is probably the most prevalent text editor and word processor today. There is no need to teach it here.

### E.4.2 Requirements

MS Word satisfies some of the requirements detailed above.

1. MS Word can edit the commands file. It does not interact directly with the running statistical process; manual cut-and-paste is required.
2. The output from the statistical process is in a window independent of MS Word. The output can be picked up and pasted into an MS Word window.
3. These are standard editing features for any window.

4. When checking is on, MS Word will inappropriately check computer programs for English syntax and spelling.
5. MS Word handles multiple files as part of its basic design.
6. Reports can be written and output text can be embedded into them;
7. MS Word has syntax and spell-checking facilities limited to the natural language (English in our case).
8. Graphics can be pasted directly into an MS Word document.
9. Mathematical formulas can be entered in an MS Word document.

## E.5 Microsoft Excel

MS Excel is well-suited for two tasks: as a database management system and as a way of organizing calculations. We do not recommend Excel for the actual calculations.

### E.5.1 Database Management

S-PLUS, R, and SAS can read and write Excel files. Since many people within an organization collect and distribute their data in Excel spreadsheets, this is a very important feature.

### E.5.2 Organizing Calculations

R can be connected to Excel via DCOM, Microsoft's protocol for exchanging information across programs. Used in this way, Excel can be used similarly to the ways that Emacs and Word are used. Excel can be used to control R, for example by putting R functions inside Excel cells and making them subject to automatic recalculation. Or R can be in control, and use Excel as one of its subroutines. We recommend (Baier, 2003, Baier and Neuwirth, 2003) for further details and for download information.

### E.5.3 Excel as a Statistical Calculator

Excel is usually a poor choice for statistical computations because:

1. As of this writing the built-in statistical functions are naively written, not including even basic numerical protection. See Table E.3 to compare

how Excel, S-PLUS, and SAS fare on calculating the variance of three numbers. This test is based on (Chan et al., 1983) who discuss various strategies needed to make sure that the fundamental goal of numerical analysis is achieved:

If the input values can be represented by the computer, and if the answer can be represented by the computer, then the calculation should get the right answer.

There are three commonly used algorithms for the variance:

**numerically unstable one-pass algorithm:**

$$(\sum(x^2) - n\bar{x}^2)/(n - 1) \quad (\text{E.1})$$

**numerically stable two-pass algorithm:**

$$\sum(x - \bar{x})^2/(n - 1) \quad (\text{E.2})$$

**even better corrected two-pass algorithm:**

$$\begin{aligned} y &= x - \bar{x} \\ \sum(y - \bar{y})^2/(n - 1) \end{aligned} \quad (\text{E.3})$$

Since MS Excel gets the right answer for  $10^7$  and the wrong answer for  $10^8$ , it looks like it is using the numerically unstable one-pass algorithm.

2. Most add-in packages are not standard and are not powerful. If add-ins are used along with an introductory textbook, they will most likely be limited in capability to the level of the text. They are unlikely to be available on computers in a work situation.

TABLE E.3. Comparison of the built-in variance function in MS Excel, S-PLUS, and SAS for the variance of the three adjacent numbers  $(1,2,3) + 10^k$ . In all cases the correct answer is 1. Excel can represent the data up to  $10^{14}$  but gets the correct answer only up to  $10^7$ . It fails disastrously at  $10^{16}$ , showing  $3.610^{16}$  when the correct answer for those input values is 0. S-PLUS and SAS can represent the data up to  $10^{15}$  and get the right answer at that value. At  $10^{16}$  S-PLUS and SAS give the correct answer for their representation of the data. We do not know why Excel is unable to represent the data at  $10^{15}$  as we are running Excel, S-PLUS, and SAS on the same machine with the same underlying representation of numbers.

(edit/code/var.xls), (edit/code/var.s), (edit/code/var.sas)

| $(1 : 3) + 10^k$                     |           |                  |                  |                  |                              |
|--------------------------------------|-----------|------------------|------------------|------------------|------------------------------|
| $k$                                  | $10^k$    | $1 + 10^k$       | $2 + 10^k$       | $3 + 10^k$       | $\text{var}((1 : 3) + 10^k)$ |
| Microsoft® Excel 2002 (10.2614.2625) |           |                  |                  |                  |                              |
| 6                                    | $10^6$    | 1000001          | 1000002          | 1000003          | 1                            |
| 7                                    | $10^7$    | 10000001         | 10000002         | 10000003         | 1                            |
| 8                                    | $10^8$    | 100000001        | 100000002        | 100000003        | 0                            |
| 14                                   | $10^{14}$ | 1000000000000001 | 1000000000000002 | 1000000000000003 | 0                            |
| 15                                   | $10^{15}$ | 1000000000000000 | 1000000000000000 | 1000000000000000 | 0                            |
| 16                                   | $10^{16}$ | 1000000000000000 | 1000000000000000 | 1000000000000000 | $3.610^{16}$                 |

S-PLUS Professional Edition Version 6.1.2 Release 1 for Microsoft Windows : 2002

SAS® Proprietary Software Release 8.2 (TS2M0)

|    |           |                  |                  |                  |   |
|----|-----------|------------------|------------------|------------------|---|
| 6  | $10^6$    | 1000001          | 1000002          | 1000003          | 1 |
| 7  | $10^7$    | 10000001         | 10000002         | 10000003         | 1 |
| 8  | $10^8$    | 100000001        | 100000002        | 100000003        | 1 |
| 14 | $10^{14}$ | 1000000000000001 | 1000000000000002 | 1000000000000003 | 1 |
| 15 | $10^{15}$ | 1000000000000001 | 1000000000000002 | 1000000000000003 | 1 |
| 16 | $10^{16}$ | 1000000000000000 | 1000000000000002 | 1000000000000004 | 4 |

## E.6 Exhortations, Some of Which Are Writing Style

This section is a compendium of exhortations that we give our classes.

Within categories these comments are in arbitrary order.

### E.6.1 Writing Style

#### 1. Check spelling (in Emacs: M-x `ispell-buffer`).

- Select the right homophone: brake vs break.
- Learn to spell technical words correctly. The following words seem to be particularly liable to misspelling:
  - separate: The fourth letter is “a”.
  - correlation: The letter “r” is doubled.
  - collinear: The letter “l” is doubled.
  - stationary: not moving
  - stationery: writing paper and envelopes
  - symmetric: The letter “m” is doubled.
  - asymmetric: The letter “s” is single.
  - Tukey: John W. Tukey
  - turkey: a bird
- “*p*-value” is preferred (with *p* in math italic). “P-value” is not OK (with “P” in uppercase roman).
- Spell people’s names correctly and with proper capitalization (John W. Tukey, Dennis Cook).

#### 2. Punctuation.

“.” “:” “,” “;” always touch the preceding character. They always have a space after them.

## E.6.2 Programming Style and Common Errors

1. Align decimal points in tables.

| Right |        | Wrong |        |
|-------|--------|-------|--------|
| 12.34 | 567.89 | 12.34 | 567.89 |
| 43.2  | 98.76  | 43.2  | 98.76  |

2. Data entry: Use the data disk and use real variable names. The default variable names “ $X_1$ ” and “ $X_2$ ” carry no information. Variable names like “height” and “weight” carry information.
3. The S-PLUS `splom()` gives easy to read plots with a single axis of symmetry over the entire set of square panels. The S-PLUS `pairs()` and the SAS PROC INSIGHT `scatter` statement give many conflicting axes of symmetry and rectangular panels. See Figure 4.11 and the accompanying discussion.
4. Analyze the experiment given you. Don’t ignore the block factor. Usually the block factor is placed first in the model formula, for example, in Section 12.9 we use `aov(plasma ~ id + time)` so the sequential analysis of variance table will read

```
id  
time  
Residuals
```

This way, in nonbalanced designs, the sequential sum of squares for the treatment factor (`time` in this example) is properly adjusted for the blocking factor `id`.

5. Please use the S-PLUS command language, not the GUI. You will get
  - much better-looking output
  - more control
  - the ability to reproduce what you did

When GUI point-and-click operations have been used, for example, to construct preliminary graphical views of the data, the commands corresponding to these operations can be viewed by clicking the S-PLUS menu `Window/History/Display`. These commands can then be used as components in the construction of more complex commands needed to produce highly customized graphs.

6. Store results of an S-PLUS function in a variable to permit easy extraction of various displays from the results. For example,
-

---

```
> my.lm <- lm( y ~ x , data=mydata)
> summary(my.lm, corr=F)
> plot(my.lm)
> coef(my.lm)
> anova(my.lm)
```

---

7. Analysis of Variance requires that the classification factor be declared as a factor. Otherwise you will get a nonsense analysis. The wrong degrees of freedom for the treatment effect is usually the indicator that you forgot the **factor(treatment)** command in S-PLUS or **CLASS treatment;** statement in SAS.
8. The degrees of freedom for a problem always comes from the **Residual** or **ERROR** line of the ANOVA table.
9. Please use **par(mfrow=c(2,3))** for plotting the results of a **lm()** or **aov()**. That way the plot uses only one piece of paper, not six.
10. All comparable graphs must be on the same scale—on the same axes is often better. See Section 17.6.3, especially Figure 17.17, for an example of comparable scaling in Panel a and noncomparable scaling in Panel b. See Figure 4.8 for comparable scaling in Panels a and b and noncomparable scaling in Panels c, d, and e.

### E.6.3 Presentation of Results

1. Use the minus sign “-4” in numbers. We do not accept hyphens “-4”. See Table E.1.
2. For multiple comparisons we can use the **multicomp** default of Tukey comparisons.
3. S-PLUS **multicomp** transcript listings are almost unreadable. Please turn in the plot rather than the transcript.
4. We don’t do multiple comparisons of blocks.
5. Write an experiment description that tells the reader how to reproduce the experiment.
6. Distinguish between “bigger than the others” and “big”. In Figure 8.5, observation 39 has a big Cook’s distance (about 2). The remaining observations have Cook’s distances about .1 and are not big. In Figure 11.15, where the largest Cook’s distance is about .18, none of the Cook’s distances are big.

Similarly for residuals. The `plot.lm()` labels the three biggest points. It doesn't care if they are significantly big.

7. `summary.lm(...)` doesn't usually provide interesting information in designed experiments. `summary.aov(..., split=list())` is frequently interesting. The ANOVA table in Table 12.15, for example, shows the partition of the 10-df "strain nested within combination" into two easily interpretable 5-df sums of squares, "strain within clover" and "strain within clover+alfalfa". It is easy to interpret these partitioned sums of squares along with the interaction means at the bottom of Table 12.14 and in the upper-left panel of Figure 12.12.

Had we used `summary.lm` we would have gotten information in Table 12.17 on the regression coefficients for the dummy variables, and we would need to see the dummy variables in Table 12.16 for the coefficients themselves to make any sense.

8. Always state the conclusions in the context of the problem.
  9. Please do not turn in impossible or illegal formatting. For example, the following line breaks in the middle of words or character strings are unacceptable:
- 
- 

```
Residual standard error: 0.4811 on 10 degrees of freedom

plot(y ~ x, main="abc      ### wrong
def")                                    

plot(y ~ x,                      ##### correct
     main="abc def")
```

---

10. English language text is normally in Times Roman. Computer output is in Courier.
  11. Do not turn in lists of data values. We know the data values. In the classroom situation we gave them to you. You may show a few observations to verify that you have read the data correctly. For example:
- 

S-PLUS:  
`ex0324[1:4,]`

SAS:

```
PROC PRINT data=ex0324 (obs=4);
```

---

**12.** Plots of the data are very interesting. We usually expect to see appropriate plots of the data (scatterplot matrix, interaction plots, sets of boxplots) and of the analysis.

**13.** We do not want a cover page for homework.

**14.** Use spacing for legibility, for example:

---

|               |                  |
|---------------|------------------|
| abc<--5+3     | is hard to read. |
| abc <- -5 + 3 | is easy to read. |

---

**15.** When you copy output, particularly by mouse from a document in a monowidth font to one with a default proportionally spaced font, make sure you keep the original spacing and indentation.

**16.** Short, complete answers are best.

# Appendix F

---

## Mathematics Preliminaries

A certain degree of mathematical maturity is a prerequisite for understanding the material in this book. Many chapters in this book require a basic understanding of these areas of mathematics:

- algebra
- differential calculus
- elementary matrix algebra, with special attention devoted to quadratic forms, eigenvalues and eigenvectors, transformations of coordinate systems, and ellipsoids in matrix notation
- combinations and permutations

This appendix provides a brief review of these topics at a level comparable to the book's exposition of statistics.

### F.1 Algebra Review

We begin with some topics in algebra, focusing on the case of two dimensions. The labels  $x$  and  $y$  are given to the horizontal and vertical dimensions, respectively.

The general equation of a straight line is given by  $ax + by = c$ , where  $a, b, c$  are constants with  $b \neq 0$ . This line intersects the  $x$ - and  $y$ -axes at  $c/a$  and  $c/b$ , respectively, and has slope  $-a/b$ .

The general equation of a *parabola* having a vertical axis is the quadratic equation  $y = ax^2 + bx + c$  for constants  $a, b, c$  with  $a \neq 0$ . The graph of the parabola opens upward if  $a > 0$  and attains a minimum when  $x = -b/2a$ . The graph opens downward if  $a < 0$  and attains a maximum when  $x = -b/2a$ . The quantity  $d = b^2 - 4ac$  is called the *discriminant*. The parabola intersects the horizontal axis at

$$x = \frac{-b \pm \sqrt{d}}{2a} \quad (\text{F.1})$$

The number of intersections, or real roots, is 2, 1, or 0 according to whether  $d >, =, < 0$ . The parabola intersects the  $y$ -axis at  $y = c$ . Equation (F.1) is referred to as the quadratic formula for solving the equation  $ax^2 + bx + c = 0$ .

The equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

represents an *ellipse* centered at  $(0, 0)$  with major and minor axes parallel to the coordinate system. When  $a > b$ , the semimajor axis has length  $a$  and the semiminor axis has length  $b$ . We graph a sample ellipse, with  $a = 3$  and  $b = 2$ , in Figure F.1.

An ellipse centered at  $(\mu_x, \mu_y)$  is obtained by replacing  $x$  and  $y$  in the above with  $x - \mu_x$  and  $y - \mu_y$ , respectively. Ellipses having axes nonparallel to the coordinate system are important in statistics and will be discussed in Section F.4.12 as an example of the use of matrix notation.

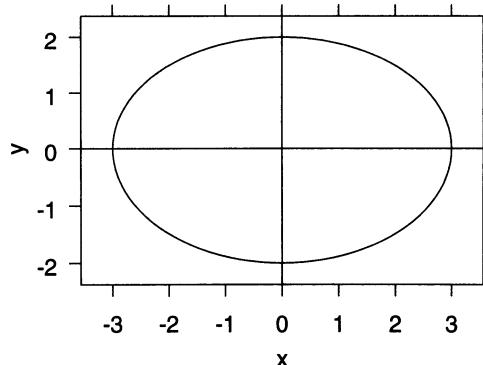


FIGURE F.1. Ellipse  $\frac{x^2}{3^2} + \frac{y^2}{2^2} = 1$

(mthp/code/ellipse32.s), (mthp/figure/ellipse32.eps.gz)

A common algebraic problem is the determination of the solution to two (or more) simultaneous equations. In the case of two linear equations, the number of solutions may be 0, 1, or  $\infty$ . There are no solutions if the equations are contradictory, such as  $x + y = 8$  and  $x + y = 9$ ; there are an infinite number of solutions if one equation is indistinct from the other, for example  $x + y = 8$  and  $2x + 2y = 16$ . When there is a unique solution, several approaches exist for finding it. One of these is adding a carefully chosen multiple of one equation to the other equation so as to result in an easily solved new linear equation involving just one variable. For example, suppose the two equations are  $x + y = 8$  and  $2x - 3y = 1$ . Adding three times the first equation to the second yields  $5x + 0y = 25$ , which implies  $x = 5$  and then  $y = 3$ .

Two additional elementary functions are the exponential function and logarithmic function. The exponential function  $y = c_1 e^{c_2 x}$  (where  $c_1, c_2$  are nonzero constants) has the property that the rate of change in  $y$  in response to a change in  $x$  is proportional to the current value of  $y$ . The logarithmic function  $y = c \ln(x)$ ,  $x > 0$  is useful in situations when successive changes in  $x$  are geometrical (i.e., proportional to the current value of  $x$ ).

An asymptote is a straight line that is gradually approached by a curved line. This concept is used to describe the ultimate behavior of the curved line. For example, the graph of  $y = \frac{1}{x}$  has the horizontal axis as its asymptote as  $x \rightarrow \infty$  and the vertical axis as its asymptote as  $x \rightarrow 0^+$ .

## F.2 Elementary Differential Calculus

If  $y = f(x)$  expresses the functional relationship between  $x$  and  $y$ , the derivative of  $y$  with respect to  $x$ , denoted  $\frac{dy}{dx}$  or  $D_{xy}$  or  $f'(x)$ , is a new function of  $x$ . For each value of  $x$ ,  $f'(x)$  gives the relative amount that  $y$  changes in response to a small change in  $x$ . For example, if  $y = f(x) = x^2$ , then it can be shown that  $f'(x) = 2x$ . When  $x = 3$ , a small increase in  $x$  will beget a sixfold increase in  $y$  because  $f'(3) = 2(3) = 6$ . Graphically,  $f'(x_0)$  is the slope of the straight line tangent to  $f(x)$  at  $x = x_0$ .

If  $f(x) = a_0 + a_1x + a_2x^2 + \dots + a_mx^m$ , an  $m^{\text{th}}$ -degree polynomial, then  $f'(x) = a_1 + 2a_2x + 3a_3x^2 + \dots + ma_mx^{m-1}$ . This rule can be used to differentiate (i.e., find the derivative of) many common functions. If  $f(x)$  can be expressed as the product of two functions, say  $f(x) = g(x) h(x)$ , then its derivative is given by the *product rule*  $f'(x) = g(x) h'(x) + g'(x) h(x)$ . This can be used, for example, to find the derivative of  $(3x^2+4x+5)(-8x^2+7x-6)$  without first multiplying the quadratics.

The most important application of  $f'(x)$  is in finding relative extrema (i.e., maxima or minima) of  $f(x)$ . A *necessary* condition for  $x_0$  to be an ex-

tremum of  $f(x)$  is that  $f'(x_0) = 0$ . This follows from the interpretation of the derivative as a tangent slope. Additional investigation is needed to confirm that such an  $x_0$  corresponds to either a maximum or minimum, and then to determine which of the two it is. For example, if  $f(x) = x^3 - 3x$ , then  $f'(x) = 3x^2 - 3$ . Setting  $f'(x) = 0$ , we find  $x = \pm 1$ .  $x = 1$  corresponds to a local minimum and  $x = -1$  corresponds to a local maximum. As another example, consider  $f(x) = x^3$ . While for this function,  $f''(0) = 0$ ,  $x = 0$  is neither a relative minimum nor relative maximum of  $f(x)$ .

### Example of an optimization problem

A rectangular cardboard poster is to have 150 square inches for printed matter. It is to have a 3-inch margin at the top and bottom and a 2-inch margin on each side. Find the dimensions of the poster so that the amount of cardboard used is minimized.

**Solution:** Let the vertical dimension of the printed matter be  $x$ . Then for the printed area to be 150, the horizontal dimension of the printed matter is  $\frac{150}{x}$ . Then allowing for the margins, the vertical and horizontal dimensions of the poster are  $x + 6$  and  $\frac{150}{x} + 4$ . The product of these dimensions is the area of the poster that we seek to minimize:  $a(x) = 174 + 4x + \frac{900}{x}$ . Taking the derivative and setting it equal to zero gives  $a'(x) = 4 - \frac{900}{x^2} = 0$ , which leads to the positive solution  $x = 15$ . Thus the minimum-sized poster with required printed area is 21 inches high by 14 inches wide, and its printed area is 15 inches high by 10 inches wide.

## F.3 An Application of Differential Calculus

We introduce Newton's method for solutions of an equation of the form  $f(x) = 0$ . A common application is the need to solve  $f'(x) = 0$  to find extrema, as discussed in Section F.2. Many equations of this type are readily solvable by successively moving toward isolation of a lone  $x$  on one side of the equation, or via a specialized technique such as the quadratic formula. In other situations one must employ one of a number of numerical techniques designed for this purpose, one of which is Newton's method.

Newton's method has the disadvantage of requiring knowledge of the derivative  $f'(x)$ , but it will often converge to a solution within a small number of iterations. As with all procedures for dealing with this problem, one must start with a first approximation  $x_0$ , and if this is "too far" from the solution  $x^*$ , the procedure may fail to converge.

The idea behind Newton's method is not difficult to understand. It is based on the equation of the tangent line to  $f(x)$  at  $x = x_0$ . If this tangent line intersects the  $x$ -axis at  $x = x_1$ , then

$$f'(x_0) = \frac{f(x_0)}{x_0 - x_1} \rightarrow x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

The iteration then proceeds with

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

and so on.

### An illustration of Newton's Method

Consider  $f(x) = x^3 - x - 5 = 0$ ,  $f'(x) = 3x^2 - 1$ , and let  $x_0 = 2$ . You can verify the following sequence:

| $i$ | $x_i$      | $f(x_i)$             |
|-----|------------|----------------------|
| 0   | 2.00000000 | $1.00 \cdot 10^0$    |
| 1   | 1.90909091 | $4.88 \cdot 10^{-2}$ |
| 2   | 1.90417486 | $1.38 \cdot 10^{-4}$ |
| 3   | 1.90416086 | $1.14 \cdot 10^{-9}$ |

## F.4 Topics in Matrix Algebra

We next provide an overview of selected topics from matrix algebra that are useful in applied statistics. Not only do matrices (the plural of matrix) allow for a more concise notation than scalar algebra, but they are an indispensable tool for communicating statistical findings. Additional material on matrix algebra is contained in the appendices of most books dealing with regression analysis, linear models, or multivariate analysis.

A *matrix* is a rectangular array consisting of  $r$  rows and  $c$  columns. A *vector* is a special type of matrix, having either  $r$  or  $c$  equal to 1. Data files are often arranged as a matrix, such that each variable is one column and each observation is one row. Systems of linear equations may be succinctly written in matrix notation.

Multivariate analysis involves probability distributions of random vectors rather than scalar random variables. Each component of a random vector is a scalar random variable. The variances and covariances of the components of a random vector are arranged in a variance–covariance matrix.

(Such a symmetric matrix  $V$ , also called covariance matrix or dispersion matrix, has the variances on the main diagonal and the covariance between variables  $i$  and  $j$  in its row  $i$  column  $j$  position. See Section 3.3.5.) The multivariate normal distribution of the random vector  $x$  with mean vector  $\mu$  and covariance matrix  $V$ , with notation

$$x \sim N(\mu, V) \quad (\text{F.2})$$

and probability density function

$$f(x) = \frac{1}{(2\pi)^{k/2} |V|^{1/2}} e^{-(x-\mu)' V^{-1} (x-\mu)/2} \quad (\text{F.3})$$

is the most important distribution used in multivariate analysis.

Matrices may be used to translate from one coordinate system to another. Orthogonal matrices perform rigid rotations in order to facilitate meaningful interpretation of results.

### F.4.1 Elementary Operations

**sum:** If two matrices are of the same size, their sum is the new matrix of this size comprised of the elementwise sums. The difference of two matrices of the same size is defined similarly.

**transpose:** If  $A$  is an  $r \times c$  matrix then its *transpose*, denoted  $A'$ , is the  $c \times r$  matrix having columns identical to the rows of  $A$  and conversely.

**inner product:** The *inner product* of two vectors of the same size, say  $u = (u_1, u_2, \dots, u_k)'$  and  $v = (v_1, v_2, \dots, v_k)'$ , is written  $u'v = v'u = \sum_{i=1}^k u_i v_i$ , i.e., the sum of the products of the corresponding elements of the two vectors. The inner product of a vector with itself yields the sum of squares of the elements of this vector. A vector  $u$  is said to be normalized if  $u'u = 1$ , i.e., if its (Euclidean) length equals 1. Two vectors  $u$  and  $v$  are said to be orthogonal if their inner product is zero:  $u'v = 0$ .

**matrix product:** The matrix product  $AB$  of an  $r \times c$  matrix  $A$  with an  $m \times n$  matrix  $B$  is defined when  $c = m$ . The element in row  $i$  and column  $j$  of  $AB$  is calculated as the inner product of the  $i^{\text{th}}$  row of  $A$  and the  $j^{\text{th}}$  column of  $B$ . The condition  $c = m$  assures that the vectors forming the inner products are of the same size. Matrix addition has the mathematical properties of commutativity and associativity. Matrix multiplication has only associativity. When  $AB$  is defined,  $BA$  may have different dimensions from  $AB$  or even be undefined.

Matrix multiplication is used when expressing systems of linear equations in matrix notation. Earlier we considered the system  $x + y = 8$  and  $2x - 3y = 1$ . If we define the  $2 \times 1$  vectors  $X = \begin{pmatrix} x \\ y \end{pmatrix}$  and  $c = \begin{pmatrix} 8 \\ 1 \end{pmatrix}$  and

the  $2 \times 2$  matrix  $A = \begin{pmatrix} 1 & 1 \\ 2 & -3 \end{pmatrix}$ , then this system can be written  $AX = c$ . In this context, the matrix  $A$  is referred to as the *coefficient* matrix.

**transpose:** The transpose of the product of two matrices is the product of their transposes in reverse order:  $(AB)' = B'A'$ .

**square:** A matrix is said to be square if it has the same number of rows as columns.

**identity:** The  $n \times n$  identity matrix has ones on its main diagonal positions (those where the row number equals the column number) and zeros everywhere else. Thus, for example, the  $3 \times 3$  identity matrix is

$$I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The identity matrix plays the same role in matrix notation as does the number 1 in scalar notation: If  $I$  denotes an identity matrix such that the indicated multiplication is defined, then  $AI = A$  and  $IA = A$ .

**$J_n$ :** We define  $J_n$  to be the  $n \times n$  matrix having all entries equal to 1.

**symmetric:** A square matrix  $A$  is said to be a symmetric matrix if  $A = A'$ . This means that row number  $i$  corresponds to column number  $i$  for all  $i$ . Let  $a_{ij}$  denote the element in row  $i$  and column  $j$  of matrix  $A$ . Then if  $A$  is square, its *trace* is the sum of its main diagonal elements:  $\text{trace}(A) = \sum_i a_{ii}$ .

**operation count:** Numerical analysts count the number of multiplications in an algorithm as an indicator of the costliness of the algorithm. A vector inner product  $\sum_{i=1}^n a_i b_i$ , for example, takes  $n$  multiplications to complete. There are other operations (indexing, adding) that must also be performed. Rather than report them we say instead that the amount of computation is proportional to the number of multiplications and indicate it by saying the operation count is  $O(n)$  (read as “big ‘O’ of  $n$ ”). Similarly, the operation count for matrix multiplication is proportional to  $n^3$  and is reported as  $O(n^3)$ .

**determinant:** The *determinant* of a square matrix  $A$ , denoted  $|A|$ , is a scalar calculated from the elements of  $A$ . In the  $2 \times 2$  case where

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

the determinant is  $|A| = a_{11}a_{22} - a_{21}a_{12}$ . If  $A$  is a square coefficient matrix of a system of linear equations (thus implying that the system has the same number of equations as unknowns), then the system has a unique solution if and only if  $|A| \neq 0$ . The determinant has useful math-

ematical properties, but is totally impractical from a computational standpoint. It is almost never needed as an intermediate calculation. There is almost always a cheaper way to calculate the final answer.

**nonsingular:** If  $|A| \neq 0$ , then  $A$  is said to be a nonsingular matrix. There is no vector  $v$  other than  $v = 0$  such that  $Av \equiv 0$ .

**inverse:** A nonsingular matrix has associated with it a unique *inverse*, denoted  $A^{-1}$ . The inverse has the property that  $AA^{-1} = A^{-1}A = I$ . The unique solution to a system of linear equations  $AX = c$  is then  $X = A^{-1}c$ . For example, the inverse of

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \text{is} \quad \frac{1}{|A|} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$$

provided that  $|A| \neq 0$ . Note that this is a mathematical identity. It is not to be interpreted as an algorithm for calculation of the inverse. As an algorithm it is very expensive, requiring  $O(2n^3)$  arithmetic operations for an  $n \times n$  matrix  $A$ . An efficient algorithm requires only  $O(n^3)$  operations.

**singular:** If  $|A| = 0$ , then  $A$  is said to be a singular matrix. There exists at least one vector  $v$  other than  $v = 0$  such that  $Av \equiv 0$ . A singular matrix does not have an inverse.

**idempotent:** A square matrix  $A$  is said to be an idempotent matrix if  $AA = A$ , i.e.,  $A$  equals the product of  $A$  with itself. A simple example is

$$\begin{pmatrix} .5 & -.5 \\ -.5 & .5 \end{pmatrix}$$

#### F.4.2 Linear Independence

A matrix  $X$  consists of a set of column vectors

$$X_{n \times (1+p)} = [\mathbf{1} X_1 X_2 \dots X_p] = [X_0 X_1 X_2 \dots X_p]$$

The columns are numbered  $0, 1, \dots, p$ .

The matrix  $X$  is said to have linearly dependent columns if there exists a nonzero  $(1 + p)$ -vector  $\ell$  such that

$$X\ell = 0$$

or, equivalently,

$$\left( \sum_j \ell_j X_j \right) = \underset{n \times 1}{(0)}$$

The matrix  $X$  is said to have linearly independent columns if no such vector  $\ell$  exists. For example, the matrix

$$X_{4 \times (1+4)} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \quad (\text{F.4})$$

has linearly dependent columns because there exists a vector  $\ell = (-1 \ 1 \ 1 \ 1 \ 1)'$  such that  $X\ell = 0$ .

The matrix

$$X_{4 \times (1+3)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \quad (\text{F.5})$$

has linearly independent columns because there exists no nonzero vector  $\ell$  such that  $X\ell = 0$ .

The *rank* of a matrix is the number of linearly independent columns it contains. Both matrices above,  $X$  and  $X_{(1,-1)}$  in Equations (F.4) and (F.5), have rank 4. For any matrix  $X$ ,  $\text{rank}(X) = \text{rank}(X'X)$ . A *full-rank* matrix is one whose rank is equal to the minimum of its row and column dimensions, that is,

$$\text{rank} \left( \underset{r \times c}{A} \right) = \min(r, c)$$

### F.4.3 Rank

The rank of a matrix  $A$  is the maximum number of linearly independent rows (equivalently, the maximum number of linearly independent columns) the matrix has. For example, the matrix

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 3 & 4 \end{pmatrix}$$

has rank 2 since columns 1 and 2 add to column 3, but any two of the columns are linearly independent of one another (i.e., are not proportional to one another).

#### F.4.4 Quadratic Forms

A *quadratic form* is a scalar resulting from the matrix product  $x'Ax$ , where  $x$  is a  $k \times 1$  vector and  $A$  is a  $k \times k$  symmetric matrix. The matrix  $A$  is termed the matrix of the quadratic form. Complicated sums of squares and products as often occur in an *analysis of variance* can be written as quadratic forms. If  $x$  has a standardized multivariate normal distribution  $x \sim N(0, I)$  [see Equation (F.2)], then  $x'Ax$  has a  $\chi^2$ -distribution with  $\nu$  degrees of freedom if and only if  $A$  is an idempotent matrix with rank  $\nu$ . For example, the numerator of the usual univariate sample variance,  $\sum_{i=1}^n (x_i - \bar{x})^2$ , can be written as  $x'Ax$  where  $x = (x_1, x_2, \dots, x_n)'$  and  $A = I_n - \frac{1}{n}J_n$ . It can be shown that this matrix  $A$  has rank  $n - 1$ , the degrees of freedom associated with the sample variance.

If  $x$  is a random vector with expected value  $\mu$  and covariance matrix  $V$ , then the expected value of the quadratic form  $x'Ax$  is

$$E(x'Ax) = \mu'A\mu + \text{trace}(AV)$$

Result (F.6) does not require that  $x$  have a multivariate normal distribution.

A square symmetric matrix  $A$  is said to be a *positive definite* (abbreviated p.d.) matrix if  $x'Ax > 0$  for all vectors  $x$  other than a vector of zeros. The matrix associated with the quadratic form representation of a sum of squares is always p.d.

#### F.4.5 Orthogonal Transformations

A matrix  $M$  is said to be *orthogonal* if  $M'M = I$ . An example is

$$M = \begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & -1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & 0 & -2/\sqrt{6} \end{pmatrix}$$

A transformation based on an orthogonal matrix is called an *orthogonal transformation*. Such transformations are rotations that preserve relative distances:  $y = Mx \rightarrow y'y = x'M'Mx = x'x$ . Orthogonal transformations are frequently encountered in statistics. A common use of them is to transform from a correlated set of random variables  $x$  to an uncorrelated set  $y$ .

The columns of an orthogonal matrix are said to be *orthonormal*. The columns are orthogonal to each other, that is  $M'_{,j}M_{,j'} = 0$  for  $j \neq j'$ . In addition, the columns have been scaled so that  $M'_{,j}M_{,j} = 1$ .

### F.4.6 Orthogonal Basis

If  $\text{rank}_{r \times c}(A) = p < \min(r, c)$  and  $c \leq r$ , then any set of  $p$  linearly independent columns of  $A$  constitutes a *basis* for  $A$ . The set of all vectors that can be expressed as a linear combination  $Xv$  of the columns of  $A$  is called the column space of  $A$ , denoted  $\mathcal{C}(A)$ . Therefore,  $\mathcal{C}(A)$  is completely specified by any basis of  $A$ . We say that the columns of the basis *span* the column space of  $A$ .

An *orthogonal* basis for  $A$  is a basis for  $A$  with the property that any two vectors comprising it are orthogonal. Starting from an arbitrary basis for  $A$ , algorithms are available for constructing an orthogonal basis for  $A$ . We show one algorithm, the Gram–Schmidt process, in Section F.4.7.

A *basis* set of column vectors for a matrix  $X$  is a set of column vectors  $U_i$  that span the same linear space as the original columns  $X_i$ . An orthogonal basis is a set of column vectors that are mutually orthogonal, that is  $U'_i U_j = 0$  for  $i \neq j$ . An orthonormal basis is an orthogonal basis whose columns have been rescaled to have norm  $\|U_i\| = \sqrt{U'_i U_i} = 1$ .

### F.4.7 Matrix Factorization— $QR$

Any rectangular matrix  $X_{n \times m}$  can be factored into the product of an matrix with orthogonal columns  $Q_{n \times m}$  and an upper triangular  $R_{m \times m}$

$$X_{n \times m} = Q_{n \times m} R_{m \times m}$$

The columns of  $Q$  span the same column space as the columns of the original matrix  $X$ . This means that any linear combination of the columns of  $X$ , say  $Xv$ , can be constructed as a linear combination of the columns of  $Q$ . Specifically,  $Xv = (QR)v = Q(Rv)$ .

There are many algorithms available to construct the matrix factorization. We show one, the Modified Gram–Schmidt (MGS) algorithm (Bjork, 1967). “Modified” means that the entire presentation is in terms of the columns  $Q_i$  of the matrix under construction. The MGS algorithm is numerically stable when calculated in finite precision. The original Gram–Schmidt (GS) algorithm, which constructs  $Q$  in terms of the columns  $X_i$  of the original matrix, is not numerically stable and should not be used for computation.

*Modified Gram–Schmidt (MGS) algorithm*

Let  $X_{n \times m} = [X_1 X_2 \dots X_m]$ . The results of the factorization will be stored in  $Q_{n \times m}$  and  $R_{m \times m}$ . The columns of  $X$  and  $Q$  and both the rows and columns of  $R$  are numbered  $1, \dots, m$ .

We will construct  $Q$  and  $R$  in steps.

1. Initialize  $R$  to 0.

$$R \leftarrow \mathbf{0}$$

2. Initialize  $Q$  to  $X$ .

$$Q \leftarrow X$$

3. Initialize the column counter.

$$i \leftarrow 1$$

4. Normalize column  $Q_i$ .

$$r_{i,i} \leftarrow \sqrt{Q'_i Q_i}$$

$$Q_i \leftarrow Q_i / r_{i,i}$$

If  $i = m$ , we are done.

5. For each of the remaining columns  $Q_j$ ,  $j = i + 1, \dots, m$ , find the component of  $Q_j$  orthogonal to  $Q_i$  by

$$r_{i,j} \leftarrow Q'_i Q_j$$

$$Q_j \leftarrow Q_j - Q_i r_{i,j}$$

6. Update the column counter.

$$i \leftarrow i + 1$$

7. Repeat steps 4–6 until completion.

An expository S-PLUS function illustrating this algorithm is in file (`mthp/code/mgs.s`). An example is in file (`mthp/code/mgs-example.s`).

The numerically efficient S-PLUS function `qr` is the computational heart of the linear models and analysis of variance functions. The intermediate matrices  $Q$  and  $R$  are usually not explicitly produced. If you wish to see them, use the `qr.Q` and `qr.R` functions.

### F.4.8 Matrix Factorization—Cholesky

Any square positive definite matrix  $S_{m \times m}$  can be factored into the product of an upper triangle matrix  $R_{n \times m}$  and its transpose

$$S = R'R$$

When  $S$  has been constructed as the cross product  $S = X'X$  of a rectangular matrix  $X_{n \times m}$ , then the upper triangular matrix  $R$  is the same matrix we get from the  $QR$  factorization.

$$S = X'X = (QR)'(QR) = R'(Q'Q)R = R'R$$

The numerically efficient S-PLUS function `chol` is available.

### F.4.9 Orthogonal Polynomials

Consider the  $k$ -vector  $v = (v_1, v_2, \dots, v_k)'$ , where  $v_1 < v_2 < \dots < v_k$ . Construct a matrix  $V = [v^0, v^1, v^2, \dots, v^{k-1}]$ , where we use the notation  $v^j = (v_1^j, v_2^j, \dots, v_k^j)'$

An orthogonal basis  $Q$  constructed from the matrix  $V$  is called a set of orthogonal polynomials. In the analysis of variance and related techniques, we often construct dummy variables for ordered factors from a set of contrasts that are orthogonal polynomials.

### F.4.10 Projection Matrices

Given any matrix  $X_{n \times m} = Q_{n \times n} R_{n \times m}$  the matrix  $P_X = X(X'X)^{-1}X' = QQ'$  is a *projection matrix* that projects an  $n$ -vector  $y$  onto the space spanned by the columns of  $X$ , that is, the product  $P_X y$  is in the column space  $\mathcal{C}(X)$ . If  $X$  has  $m$  columns and rank  $r \leq m$ , then the eigenvalues of  $P_X$  consist of  $r$  1s and  $m - r$  0s.

### F.4.11 Geometry of Matrices

We provide some detail of the application to two-dimensional geometry. Each two-dimensional vector represents a point; alternatively, a directed line segment from the origin to this point. A  $2 \times 2$  matrix postmultiplied by a vector transforms this point to another point. Consider the orthogonal matrix

$$M = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}$$

and let  $x$  be a  $2 \times 1$  vector representing a point in two dimensions. Then  $w = Mx$  produces a new point  $w$  which is where  $x$  appears in the new coordinate system formed by rotating the old one  $\theta$  degrees counterclockwise around the origin.

If  $x$  and  $y$  are each two-dimensional vectors and  $\theta$  is the angle between them, then  $\cos(\theta) = x'y/\sqrt{x'x\ y'y} = \text{corr}(x, y)$ , the *correlation* between these two vectors. Note that if the vectors are orthogonal so that  $x'y = 0$ , then  $\cos(\theta) = 0$  and  $\theta = 90^\circ$ .

### F.4.12 Eigenvalues and Eigenvectors

Next we study the concepts of eigenvectors and eigenvalues of an  $n \times n$  symmetric matrix  $V$ .

If  $V\xi = \lambda\xi$ , where  $\xi$  is an  $n \times 1$  vector and  $\lambda$  is a scalar, then  $\lambda$  is said to be an *eigenvalue* of  $V$  with corresponding *eigenvector*  $\xi$ . Without loss of generality, we can take the eigenvector to be normalized. Geometrically, the matrix  $V$  transforms its eigenvectors into multiples of themselves. Any two distinct eigenvectors are orthogonal:  $\xi_i'\xi_j = 0$ ,  $i \neq j$ .  $V$  can be written as its *spectral decomposition*  $V = \sum_i \lambda_i \xi_i \xi_i'$ .  $V$  can be written as its *eigenvalue factorization*  $V = \Xi \Lambda \Xi'$ .

A matrix is nonsingular if and only if it has only nonzero eigenvalues. A matrix is positive definite if and only if all of its eigenvalues are positive. The eigenvalues of  $V^{-1}$  are the reciprocals of the eigenvalues of  $V$ . The determinant of a matrix equals the product of its eigenvalues  $|V| = \prod \lambda_i$ , and this is normally the most efficient way to calculate the determinant. The trace of a matrix equals the sum of its eigenvalues  $\text{trace}(V) = \sum \lambda_i$ .

Consider the problem of choosing  $x$  to maximize  $x'Vx$  subject to  $x'x = 1$ . The maximum value is the largest eigenvalue of  $V$  and is attained when  $x$  is the eigenvector corresponding to this eigenvalue. Similarly,  $x'Vx$  is minimized (subject to  $x'x = 1$ ) at the value of the smallest eigenvalue of  $V$ , occurring at the corresponding eigenvector.

Here is an example of hand calculation of eigenvalues and eigenvectors. Consider the  $2 \times 2$  matrix

$$V = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

Its 2 eigenvalues are the 2 scalar solutions  $\lambda$  to the equation  $|V - \lambda I| = 0$ . Taking the determinant leads to  $(2 - \lambda)(1 - \lambda) - 1 = 0 \implies \lambda^2 - 3\lambda + 1 = 0 \implies \lambda = (3 \pm \sqrt{5})/2$ , which expands to  $\approx 2.618$  or  $\approx 0.382$ . (Note that we are explicit about the approximation to 3 decimal digits. Our usual practice is *not* to round answers.) The eigenvector  $\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$  corresponding

TABLE F.1. Eigenvalues and eigenvectors of  $V = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ .  
 (mthp/code/eigen2111.s)

```
S-PLUS (mthp/transcript/eigen2111.st):
> V <- matrix(c(2,1,1,1), 2,2)
> V
     [,1] [,2]
[1,]    2    1
[2,]    1    1
> eV <- eigen(V)
> eV
$values:
[1] 2.618034 0.381966

$vectors:
      [,1]      [,2]
[1,] -0.8506508  0.5257311
[2,] -0.5257311 -0.8506508

> sqrt(eV$val) ## semimajor and semiminor axis lengths
[1] 1.618034 0.618034
> atan(c(eV$vec[2,1]/eV$vec[1,1], eV$vec[2,2]/eV$vec[1,2])) ## angle of axes
[1] 0.5535744 -1.0172220
> diff(atan(c(eV$vec[2,1]/eV$vec[1,1], eV$vec[2,2]/eV$vec[1,2]))) ## right angle
[1] -1.570796
```

to  $\lambda \approx 2.618$  is the solution to the equation  $V\xi \approx 2.618\xi$ . This implies that  $2\xi_1 + \xi_2 \approx 2.618$ , and coupled with the normalization restriction  $\xi_1^2 + \xi_2^2 = 1$  we find that  $\xi_1 \approx .8507$  and  $\xi_2 \approx .5257$ . The eigenvector corresponding to the other eigenvalue is found similarly.

As a geometric application of eigenvalues, consider the ellipse having equation  $(x - \mu)'V^{-1}(x - \mu) = b$ . In statistics, this is the form of a confidence ellipse for  $\mu$ . Let  $\lambda_1 < \lambda_2$  be the eigenvalues of  $V$  with corresponding normalized eigenvectors  $\xi_1 = \begin{pmatrix} \xi_{11} \\ \xi_{21} \end{pmatrix}$ ,  $\xi_2 = \begin{pmatrix} \xi_{12} \\ \xi_{22} \end{pmatrix}$ . Then the semimajor axis of this ellipse has length  $\sqrt{\lambda_2 b}$  and the semiminor axis has length  $\sqrt{\lambda_1 b}$ . The angle between the extension of the semimajor axis and the horizontal axis is  $\arctan(\frac{\xi_{12}}{\xi_{22}})$ .

Continuing the example, we calculate the eigenvalues of  $V = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$  in Table F.1 and graph the ellipse  $x'Vx = x'\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}x = 1$  in Figure F.2.

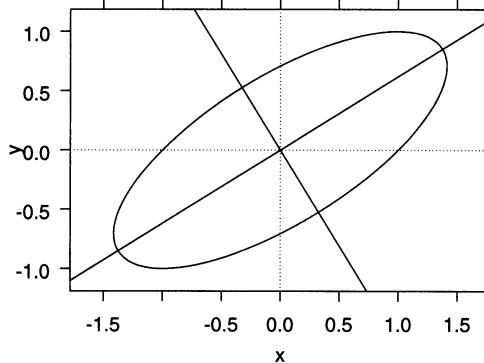


FIGURE F.2. Ellipse  $x'Vx = x'\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}x = 1$ .

(`mthp/code/ellipse2111.s`), (`mthp/figure/ellipse2111.eps.gz`)

### F.4.13 Singular Value Decomposition

Let  $M$  be an arbitrary  $r \times c$  matrix,  $U$  the  $r \times r$  matrix containing the eigenvectors of  $MM'$ , and  $W$  the  $c \times c$  matrix containing the eigenvectors of  $M'M$ . Let  $\Delta$  be the  $r \times c$  matrix having  $\delta_{ij} = 0$  ( $i \neq j$ ). If  $r \geq c$  (the usual case in statistical applications), define  $\delta_{ii}$  = the square root of the eigenvalue of  $M'M$  corresponding to the eigenvector in the  $i^{\text{th}}$  column of  $W$ . If  $r < c$ , define  $\delta_{ii}$  = the square root of the eigenvalue of  $MM'$  corresponding to the eigenvector in the  $i^{\text{th}}$  column of  $U$ . Then the *singular value decomposition* of  $M$  is  $M = U\Delta W$ . Note that the number of nonzero diagonal values in  $\Delta$  is  $\min(r, c)$ .

### F.4.14 Generalized Inverse

For any rectangular matrix  $X_{n \times m} = U\Delta W'$ , the Moore–Penrose generalized inverse is defined as

$$X^- = W\Delta^{-1}U'$$

Since  $\Delta = \text{diag}(\delta_i)$  is a diagonal matrix, its inverse is  $\Delta^{-1} = \text{diag}(\delta_i^{-1})$ . The definition is extended to the situation when  $\text{rank}(X) < \min(n, m)$  by using  $0^{-1} = 0$ .

When  $\text{rank}(X) = m = n$ , hence the inverse exists, the generalized inverse is equal to the inverse.

### F.4.15 Solving Linear Equations

There are three cases.

#### F.4.15.1 $n = m = \text{rank}(X)$

Given a matrix  $X_{n \times m}$  and an  $n$ -vector  $y$ , the solution  $\beta$  of the linear equation

$$y = X\beta$$

is uniquely defined by

$$\beta = X^{-1}y$$

when  $X$  is invertible, that is when  $n = m = \text{rank}(X)$ .

#### F.4.15.2 $n > m = \text{rank}(X)$

When  $n > m = \text{rank}(X)$ , the linear equation is said to be overdetermined. Some form of arbitrary constraint is needed to find a solution. The most frequently used technique is least-squares, a technique in which  $\hat{\beta}$  is chosen to minimize the norm of the residual vector

$$\min_{\beta} \|y - X\beta\|^2 = \sum_{i=1}^n \left( y_i - \sum_{j=1}^m x_{ij}\beta_j \right)^2$$

The solution  $\hat{\beta}$  is found by solving the related linear equations

$$X'y = X'X\hat{\beta}$$

The solution is often expressed as

$$\hat{\beta} = (X'X)^{-1}(X'y) = X^{-1}y$$

This is a definition, not an efficient computing algorithm.

#### F.4.15.3 $m > p = \text{rank}(X)$

When  $m > p = \text{rank}(X)$ , there are an infinite number of solutions to the linear equation. The singular value decomposition  $X = U\Delta W'$  will have  $m - p$  zero values along the diagonal of  $\Delta$ . Let  $\beta_0$  be one solution. Then

$$\beta_{\gamma} = \beta_0 + W \begin{pmatrix} \mathbf{0} \\ \gamma \end{pmatrix}$$

where  $\mathbf{0}$  is a vector of  $p$  zeros and  $\gamma$  is any vector of length  $m - p$ , is also a solution.

## F.5 Combinations and Permutations

### F.5.1 Factorial

For a positive integer  $n$ , the notation  $n!$ , read “ $n$  factorial”, is used to indicate the product of all the integers from 1 through  $n$ :

$$n! = n \times (n - 1) \times \dots \times 1 = n \times (n - 1)!$$

The factorial of zero,  $0!$ , is separately defined to equal 1.

Thus

| $n$      | $n!$     | $=$ | $n!$     | $=$ | $n((n - 1)!)$ |
|----------|----------|-----|----------|-----|---------------|
| 0        | $0!$     | $=$ | 1        | $=$ | 1             |
| 1        | $1!$     | $=$ | 1        | $=$ | $1 \times 1$  |
| 2        | $2!$     | $=$ | 2        | $=$ | $2 \times 1$  |
| 3        | $3!$     | $=$ | 6        | $=$ | $3 \times 2$  |
| 4        | $4!$     | $=$ | 24       | $=$ | $4 \times 6$  |
| 5        | $5!$     | $=$ | 120      | $=$ | $5 \times 24$ |
| $\vdots$ | $\vdots$ | $=$ | $\vdots$ | $=$ | $\vdots$      |

### F.5.2 Permutations

The notation  $nP_p$ , read “ $n$  permute  $p$ ”, indicates the number of ways to select  $p$  distinct items from  $n$  possible items where two different orderings of the same  $p$  items are considered to be distinct. Equivalently,  $nP_p$  is the number of distinct ways of *arranging*  $p$  items from  $n$  possible items:

$$nP_p = \frac{n!}{(n - p)!}$$

For example,

$$5P_3 = \frac{5!}{(5 - 3)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1} = 5 \times 4 \times 3 = 60$$

### F.5.3 Combinations

The notation  $nC_p$  or  $\binom{n}{p}$ , read “ $n$  choose  $p$ ”, indicates the number of ways to select  $p$  distinct items from  $n$  possible items, where two different orderings of the same  $p$  items are considered to be the same selection. Equivalently,  $nC_p$  is the number of distinct ways of *choosing*  $p$  items from  $n$  possible

items:

$$\binom{n}{p} = {}_n C_p = \frac{n!}{p! \times (n-p)!} = \frac{{}_nP_p}{p!}$$

For example,

$$\binom{5}{3} = \frac{{}_5 C_3}{2!} = \frac{5!}{3!(5-3)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1)(2 \times 1)} = \frac{5 \times 4}{2 \times 1} = 10$$

## F.6 Exercises

- F.1.** Start from the matrix in Equation (F.6)

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 3 & 4 \end{pmatrix}$$

Give an example of a basis for  $A$ . Then give an example of a vector in  $\mathcal{C}(A)$  and also a vector not in  $\mathcal{C}(A)$ . Give an example of an orthogonal basis of  $A$ , demonstrating that it is orthogonal.

- F.2.** Verify that Equation (F.6) defines a family of solutions to the set of linear equations with  $p = \text{rank}(X) < m$ .

# Appendix G

---

## Graphs Based on Cartesian Products

We emphasize throughout this book that graphical display is an integral part of data analysis. Superior data analysis almost always benefits from high-quality graphics. Appropriately drawn graphs are, in our opinion, the best way to gain an understanding of what data have to say, and also to convey results of the analysis to others, both other statisticians and persons with minimal training in statistics.

We illustrate many standard graphs. We also illustrate many graphical displays that are not currently standard and some of which are new. The software for our displays is available in files in directories `splus.library` and `sas.library`.

Analysts occasionally require a graph unlike any readily available elsewhere. We recommend that serious data analysts invest time in becoming proficient in writing code rather than using the graphical user interface (GUI). Very few of the graphs in this book can be produced using a GUI. Users of a GUI are limited to the current capabilities of the GUI. While the design of GUIs will continually improve, their capabilities will always remain far behind what skilled programmers can produce. Even less-skilled analysts can take advantage of cutting-edge graphics by accessing libraries of graphing functions such as those included in the online files accompanying this book, or those available at StatLib (Vlachos, 2004), or available elsewhere on the Internet.

## G.1 Structured Sets of Graphs

Several of our examples extend the concept of a structured presentation of plots of different sets of variables, or of different parametric transformations of the same set of variables. Several of our examples extend the interpretation of the model formula, that is, the semantics of the formula, to allow easier exposition of standard statistical techniques.

In this appendix we list these displays in order to comment on their construction. We provide a reference to an example in the book for each type of display. Discussion of the interpretation of the graphs appears in the indicated chapters.

### G.1.1 Cartesian Products

A feature common to many of the displays is the Cartesian product principle behind their construction.

The Cartesian product of two sets  $A$  and  $B$  is the set consisting of all possible ordered pairs  $(a, b)$  where  $a$  is a member of the set  $A$  and  $b$  is a member of the set  $B$ . Many of the innovative graphs in this book are formed as a rectangular set of panels, or subgraphs, where each panel is based on one pair from a Cartesian product. The sets defining the Cartesian product differ for each graph type. For example, a set can be a collection of variables, functions of a single variable, levels of a single factor, functions of a fitted model, different models, etc.

### G.1.2 Trellis Paradigm

Many of the graphs in the book are constructed using the trellis paradigm pioneered by S-PLUS. The trellis system of graphics in S-PLUS is based on the paradigm of repeating the same graphical specifications for each element in a Cartesian product of levels of one or more factors.

The majority of the methods supplied in the S-PLUS **trellis** library are based on a typical formula having the structure

$$y \sim x \mid a * b \tag{G.1}$$

where

- y is either continuous or factor
- x is continuous
- a is factor
- b is factor

and each panel is a plot of  $y \sim x$  for the subset of the data defined by the Cartesian product of the levels of `a` and `b`. Figures 10.4 and 13.1 are examples conditioned, respectively, on one and two factors.

The term “trellis” comes from gardening, where it describes an open structure used as a support for vines. In graphics, a trellis provides a framework in which related graphs can be placed.

## G.2 Scatterplot Matrices: `splom` and `xysplom`

A scatterplot matrix (`splom`) is a trellis display in which the panels are defined by a Cartesian product of variables. In the standard scatterplot matrix constructed by `splom`, Figure 4.5 for example, the same set of variables define both the rows and columns of the matrix. More generally, Figure 4.3 for example, we have different sets of variables defining the rows and columns of the matrix.

1. **`splom`.** A scatterplot matrix (`splom`) does not follow the semantic paradigm of Equation (G.1). It differs from the majority of trellis-based methods in two ways. First, each of the panels is a plot of a different set of variables. Second, each of the panels is based on the entire set of observations.

Sections 4.4 and 4.6 contain extensive discussions of scatterplot matrices. We strongly recommend the use of a `splom`, sometimes conditioned on values of relevant categorical variables, as an initial step in analyzing a set of data.

2. **`xysplom`.** Our function `xysplom` is used to produce a rectangular subset, often an off-diagonal block, of a scatterplot matrix. It involves a Cartesian product of the form `[variables] × [variables]`. An example of an `xysplom` is Figure 4.3, a block taken from the full `splom` in Figure 4.5.

We use an extension

```
u + v ~ w + x + y + z | a * b
```

of the syntax the standard model formula to define the variables to the `xysplom` function. Figure 4.4 is constructed with the formula

```
xysplom(sprice ~ beds + drarea + kitarea | CondoHouse)
```

The rows of the `xysplom` are defined by the crossing of the set of variables on the left-hand side of the formula (only one in this example) and the levels of the factor `CondoHouse`. The columns are defined by the set of variables on the right-hand side of the formula. In this example there

is only one variable in the set of variables on the left-hand side of the formula.

An `xysplom` is useful in situations where the number of variables under study is too large to produce a legible `splom` containing all variables on a single page. In such a circumstance we recommend the use of two or more pages of `sploms` and `xysploms` to display pairwise relationships among variables.

### G.3 Cartesian Products of Sets of Functions

In this group of displays, at least one of the sets included in the Cartesian product is a set of functions.

1. Plots illustrating lack of homogeneity of variance (Figure 6.5) — [functions of data]  $\times$  [levels of factor]. In this graph, one of the sets in the Cartesian product is the function of the data represented (observed data, median-centered data, absolute value of the median-centered data). We discuss in Section 6.10 the Brown–Forsyth test for variance homogeneity.
2. Logistic regression plots (Figure 17.11) — [variables and functions of variables]  $\times$  [variables and functions of variables]. This is an ordinary `splom` of the response variable, selected predictor variables, and the prediction displayed on three scales: the logit scale, the odds scale, and the probability scale.
3. Ladder-of-power plots (Figure 4.16) — [powers of  $y$ ]  $\times$  [powers of  $x$ ]. This plot is useful in a regression context for determining the optimal power transformations of both the response and predictor variables.
4. Regression diagnostics plots (Figure 11.6) — [statistics]. This plot displays all common regression diagnostics on a single page. Included are thresholds for flagging cases as unusual along with identification of such cases.

### G.4 Graphs Requiring Multiple Calls to `xysplom`

When one of the sets in the Cartesian product is a set of functions, the easiest way to construct the product is to make several `xysplom` calls, one for each function in the set.

1. Partial residual plots (Figure 9.10) — [functions of fitted values and residual]  $\times$  [variables]. Response against predictors, residuals against predictors, partial residuals against predictors (partial residual plots),

and partial residuals of  $Y$  against partial residuals of  $X$  (added variable plots). Each row of Figure 9.10 is a different function of fitted values or residual. Each column is either one of the predictor variables or a function of the predictor variables. See the discussion in Section 9.14.

2. Analysis of covariance plots (One example is in the set of Figures 10.6, 10.7, 10.8, and 10.9. Another example is in Figure 14.5) — [models]  $\times$  [levels]. A key feature of this set of plots is its presentation of all points both superposed into one panel and also segregated into individual panels defined by the levels of a factor. In this framework, the superposition of all levels of the factor is itself considered a level.
3. ODOFFNA plots (Figure 14.14) — [transformation power]  $\times$  [factors]  $\times$  [factors], a 3-dimensional Cartesian product. This is a series of interaction plots indexed by a third variable, the transformation power, all on a single page. Figure 14.14 is intended to find a satisfactory power transformation to achieve homogeneity of variance and then assess interaction among the two factors for the chosen power transformation.

## G.5 Asymmetric Roles for the Row and Column Sets

1. Interaction plots (Figure 12.1) — [factors]  $\times$  [factors]. Each off-diagonal panel is a standard interaction plot. Panels in transpose positions interchange the trace- and  $x$ -factors. Rows are labeled by the trace factor. Columns are labeled by the  $x$ -factor. The main diagonal is used for boxplots of the main effects.
2. ARIMA-trellis plots (Figure 18.7) — [number of AR parameters]  $\times$  [number of MA parameters]  $\times$  [type of display]. Each of the  $3 \times 3$  displays contains diagnostic information about each of the 9 models indexed by the numbers of autoregressive and moving average parameters  $p$  and  $q$ . In addition we group several types of display on a single page. This plot displays most commonly used diagnostics for identifying the number of AR and MA parameters in time series models of the ARIMA class.

## G.6 Rotated Plots

1. Mean–mean multiple comparisons plots (MMC plots) (Figure 7.18) — [means at levels]  $\times$  [means at levels]. The plot is designed as a crossing of the means of a response variable at the levels of one factor with itself. It is then rotated  $45^\circ$  so the horizontal axis can be interpreted as the differences in mean levels and the vertical axis can be interpreted as

the weighted averages of the means comprising each comparison. This class of plots is used to display the results of a multiple comparison procedure.

## G.7 Squared Residual Plots

The fundamental concept of “least squares” is difficult to present to introductory classes. We illustrate the squares. The sum of their areas is the sum of squares that is minimized according to the “least-squares” principle.

1. Illustrations of 2D and 3D least-squares fits. (Figures 8.2, 9.1, and 9.5) — [fitted models]  $\times$  [methods of displaying residuals]. The rows of Figure 8.2 are ways of displaying residuals; the first row shows the residuals as vertical lines, the second as squares. The columns show different models: none, least-squares, and a too-shallow fit.

## G.8 Alternate Presentations

We have alternate presentations of existing ideas.

1. Transposed trellis plots (Figure 13.10). S-PLUS offers trellis displays (boxplots, dotplots, and others) with the response variable on the horizontal axis. We added the ability to transpose each panel to get the response variable on the vertical axis. The vertical orientation places the response variable in the vertical direction and accords with how we have been trained to think of functions—levels of the independent variable along the abscissa and the response variable along the ordinate.
2. Odds-ratio CI plot (Figure 15.5). The odds ratio  $(\frac{p_1}{q_1}) / (\frac{p_2}{q_2})$  does not, by construction, give information on both underlying  $p_1$ - and  $p_2$ -values. It is necessary to specify one of them to estimate the other. We backtransform the CI on the odds ratio to a CI on the probability scale and plot the CI of  $p_2$  for all possible values of  $p_1$ . The two axes have the same  $(0, 1)$  probability scale.

---

# References

- (Adish et al., 1999) Adish, A. A., Esrey, S. A., Gyorkos, T. W., Jean-Baptiste, J., and Rojhani, A. (1999). Effect of consumption of food cooked in iron pots on iron status and growth of young children: A randomised trial. *The Lancet*, 353.
- (Agresti, 1990) Agresti, A. (1990). *Categorical Data Analysis*. Wiley.
- (Agresti and Caffo, 2000) Agresti, A. and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *American Statistician*, 54(4):280–288.
- (Aladdin Enterprises Software Pty Ltd., 2001) Aladdin Enterprises Software Pty Ltd. (2001). Ghostscript.  
<ftp://ftp.cs.wisc.edu/ghost/GS700W32.EXE>.
- (Albuquerque Board of Realtors, 1993) Albuquerque Board of Realtors (1993).  
<http://lib.stat.cmu.edu/DASL/Stories/homeprice.html>.
- (American Statistical Association, 2002) American Statistical Association (2002). Careers in statistics.  
<http://www.amstat.org/careers/brochure.html>.
- (Anderson et al., 1981) Anderson, A. H., Jensen, E. B., and Schou, G. (1981). Two-way analysis of variance with correlated errors. *International Statistical Review*, 49:153–167.
- (Anderson and Bancroft, 1952) Anderson, R. L. and Bancroft, T. A. (1952). *Statistical Theory in Research*. McGraw-Hill.
- (Anderson and McLean, 1974) Anderson, V. L. and McLean, R. A. (1974). *Design of Experiments*. Marcel Dekker.
- (Andrews and Herzberg, 1985) Andrews, D. F. and Herzberg, A. M. (1985). *Data: A collection of problems from many fields for the student and research worker*. Springer. The entire data collection is available for download from <http://lib.stat.cmu.edu/datasets/Andrews/>.

- (Anionwu et al., 1981) Anionwu, E., Watford, D., Brozovic, M., and Kirkwood, B. (1981). Sickle cell disease in a British urban community. *British Medical Journal*, 282:283–286.
- (Asabere and Huffman, 1996) Asabere, P. K. and Huffman, F. E. (1996). Negative and positive impacts of golf course proximity on home prices. *The Appraisal Journal*, pages 351–355.
- (Baier, 2003) Baier, T. (2003). R: Windows component services, integrating r and excel on the com layer. In Hornik, K. and Leisch, F., editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. Technische Universität Wien, Vienna, Austria.  
<http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Drafts/Baier.pdf>.
- (Baier and Neuwirth, 2003) Baier, T. and Neuwirth, E. (2003). High-level interface between r and excel. In Hornik, K. and Leisch, F., editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. Technische Universität Wien, Vienna, Austria.  
<http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Drafts/BaierNeuwirth.pdf>.
- (Barnett and Mead, 1956) Barnett, M. K. and Mead, F. C. (1956). A  $2^4$  factorial experiment in four blocks of eight. *Applied Statistics*, 5:122–131.
- (Becker et al., 1988) Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988). *The S Language; A Programming Environment for Data Analysis and Graphics*. Wadsworth & Brooks/Cole, Pacific Grove, CA.
- (Bishop et al., 1975) Bishop, Y. Y. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis*. MIT Press.
- (Bjork, 1967) Bjork, A. (1967). Solving least squares problems by Gram–Schmidt orthogonalization. *BIT*, 7:1–21.
- (Bliss, 1967) Bliss, C. I. (1967). *Statistics in Biology*. McGraw-Hill.
- (Blyth, 1972) Blyth, C. R. (1972). On Simpson’s paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67:364–366.
- (Bowerman and O’Connell, 1990) Bowerman, B. L. and O’Connell, R. T. (1990). *Linear Statistical Models*. Duxbury.
- (Box and Cox, 1964) Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *J. Royal Statist Soc B*, 26:211–252.
- (Box et al., 1978) Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*. Wiley.
- (Box and Jenkins, 1976) Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, revised edition.
- (Braungart, 1971) Braungart, R. G. (1971). Family status, socialization and student politics: A multivariate analysis. *American Journal of Sociology*, 77:108–130.
- (Brochard et al., 1995) Brochard, L., Mancebo, J., Wysocki, M., Lofaso, F., Conti, G., Rauss, A., Simonneau, G., Benito, S., Gasparetto, A., Lemaire, F., Isabey, D., and Harf, A. (1995). Noninvasive ventilation for acute exacerbations of chronic pulmonary disease. *New England Journal of Medicine*, 333(13):817–822.

- (Brooks et al., 1988) Brooks, D. G., Carroll, S. S., and Verdini, W. A. (1988). Characterizing the domain of a regression model. *The American Statistician*, 42:187–190.
- (Brown, 1980) Brown, Jr., B. W. (1980). Prediction analyses for binary data. In Miller, J. R. G., Efron, B., Brown, Jr., B. W., and Moses, L. E., editors, *Biostatistics Casebook*. Wiley.
- (Brown and Forsyth, 1974) Brown, M. B. and Forsyth, A. B. (1974). Robust tests for equality of variances. *Journal of the American Statistical Association*, 69:364–367.
- (Bureau of the Census, 2001) Bureau of the Census (2001). *Statistical Abstract of the United States*. U.S. Department of Commerce.
- (Cai, 2003) Cai, T. T. (2003). One-sided confidence intervals in discrete distributions.  
<http://www-stat.wharton.upenn.edu/~tcai/paper/1sidedCI.pdf>.
- (Cameron and Pauling, 1978) Cameron, E. and Pauling, L. (1978). Supplemental ascorbate in the supportive treatment of cancer: Re-evaluation of prolongation of survival times in terminal human cancer. *Proceedings of the National Academy of Science*, 75:4538–4542.
- (Chambers et al., 1983) Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Wadsworth.
- (Chan et al., 1983) Chan, T. F. C., Golub, G. H., and LeVeque, R. J. (1983). Algorithms for computing the sample variance: Analysis and recommendations. *The American Statistician*, 37(3):242–247.
- (Chassell, 1999) Chassell, R. (1999). *Programming in Emacs Lisp: An Introduction*. Free Software Foundation, 2nd edition.
- (Chin et al., 1961) Chin, T. W., Hall, E., Gravelle, C., and Speers, J. (1961). The influence of Salk vaccination on the epidemic pattern and spread of the virus in the community. *American Journal of Hygiene*, 73:67–94.
- (Chu, 1996) Chu, S. (1996). Diamond ring pricing using linear regression. *Journal of Statistical Education*, 4.  
<http://www.amstat.org/publications/jse/v4n3/datasets.chu.html>.
- (Cochran and Cox, 1957) Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*. Wiley.
- (Collett, 1991) Collett, D. R. (1991). *Modelling Binary Data*. Chapman & Hall.
- (Comprehensive TeX Archiving Network, 2002) Comprehensive TeX Archiving Network (2002). CTAN.  
<ftp://metalab.unc.edu/pub/packages/TeX/index.html>.
- (Conover et al., 1981) Conover, W. J., Johnson, M. E., and Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, 23:351–361.
- (Consumer Reports, 1986) Consumer Reports (1986). Hot dogs. *Consumer Reports*, pages 366–367.  
<http://lib.stat.cmu.edu/DASL/Stories/Hotdogs.html>.
- (Cook and Weisberg, 1999) Cook, R. D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. John Wiley & Sons.

- (Cook, 1971) Cook, T. (1971). Convictions for drunkenness. *New Society*.
- (Cytel Software Corporation, 2004) Cytel Software Corporation (2004). Logxact statistical software: Release 5.  
[http://www.cytel.com/LogXact/logxact\\_brochure.pdf](http://www.cytel.com/LogXact/logxact_brochure.pdf).
- (Dalal et al., 1989) Dalal, S. R., Fowlkes, E. B., and Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-Challenger prediction of failure. *Journal of the American Statistical Association*, 84:945–957.
- (Darwin, 1876) Darwin, C. (1876). *The Effect of Cross- and Self-Fertilization in the Vegetable Kingdom*. John Murray, second edition.
- (Data Archive, 1997) Data Archive (1997). Journal of statistics education.  
[http://www.amstat.org/publications/jse/jse\\_data\\_archive.html](http://www.amstat.org/publications/jse/jse_data_archive.html).
- (Davies, 1954) Davies, O. L. (1954). *Design and Analysis of Industrial Experiments*. Oliver and Boyd.
- (Davies and Goldsmith, 1972) Davies, O. L. and Goldsmith, P. L., editors (1972). *Statistical Methods in Research and Production*. Oliver and Boyd, fourth edition.
- (Design Science, Inc., 2000) Design Science, Inc. (2000). Mathtype.  
<http://www.dessci.com/support/fonts/default.stm>.
- (Desu and Raghavarao, 2003) Desu, M. M. and Raghavarao, D. (2003). *Non-parametric Statistical Methods for Complete and Censored Data*. Chapman & Hall, first edition.
- (Edwards and Berry, 1987) Edwards, D. and Berry, J. J. (1987). The efficiency of simulation-based multiple comparisons. *Biometrics*, 43:913–928.
- (Ellis et al., 1987) Ellis, M. E., Neal, K. R., and Webb, A. K. (1987). Is smoking a risk factor for pneumonia for patients with chickenpox? *British Medical Journal*, 294:1002.
- (Emerson and Stoto, 1983) Emerson, J. D. and Stoto, M. A. (1983). Transforming data. In Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors, *Understanding Robust and Exploratory Data Analysis*. Wiley.
- (Erdman, 1946) Erdman, L. W. (1946). Studies to determine if antibiosis occurs among rhizobia: I. between rhizobium meliloti and rhizobium trifolii. *Journal of the American Society of Agronomy*, 38:251–258.
- (Fleiss, 1981) Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*. Wiley, 2nd edition.
- (Forbes Magazine, 1993) Forbes Magazine (1993).  
<http://lib.stat.cmu.edu/DASL/Datafiles/ceodat.html>.
- (Fouts, 1973) Fouts, R. S. (1973). Acquisition and testing of gestural signs in four young chimpanzees. *Science*, 180:978–980.
- (Fox, 1991) Fox, J. (1991). *Regression Diagnostics: An Introduction*. Sage Publications.
- (Free Software Foundation, 2003) Free Software Foundation (2003). Emacs.  
<ftp://ftp.gnu.org/gnu/windows/emacs/21.3/emacs-21.3-fullbin-i386.tar.gz>.
- (Freund and Littell, 1991) Freund, R. J. and Littell, R. C. (1991). *SAS System for Regression*. SAS Institute, Inc.
- (Friendly, 1991) Friendly, M. (1991). *SAS System for Statistical Graphics*. SAS Institute, Inc.

- (Friendly, 2004) Friendly, M. (2004). Macro programs from "the sas system for statistical graphics".  
<http://www.math.yorku.ca/SCS/sssg>.
- (Gelman et al., 2002) Gelman, A., Pasarica, C., and Dodhia, R. (2002). Let's practice what we teach: Turning tables into graphs. *The American Statistician*, 56:121–130.
- (Ghostgum Software Pty Ltd., 2001) Ghostgum Software Pty Ltd. (2001). Ghostview.  
<ftp://ftp.cs.wisc.edu/ghost/GSV40W32.EXE>.
- (Goodnight, 1978) Goodnight, J. H. (1978). Tests of hypotheses in fixed effects linear models. Technical Report R-101, SAS Institute.
- (Graham, 1996) Graham, P. (1996). *ANSI Common Lisp*. Prentice Hall.
- (Greenacre, 1984) Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press.
- (Gunst and Mason, 1980) Gunst, R. F. and Mason, R. L. (1980). *Regression Analysis and Its Application: A Data-Oriented Approach*. Marcel Dekker.
- (Hamilton, 1983) Hamilton, L. C. (1983). Saving water: A causal model of household conservation. *Sociological Perspectives*, 26(4):355–374.
- (Hamilton, 1992) Hamilton, L. C. (1992). *Regression with Graphics*. Brooks-Cole.
- (Hand et al., 1994) Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., and Ostrowski, E. (1994). *A Handbook of Small Data Sets*. Chapman and Hall.
- (Harrison et al., 2004) Harrison, D. D., Harrison, D. E., Janik, T. J., Cailliet, R., Ferrantelli, J. R., Hass, J. W., and Holland, B. (2004). Modeling of the sagittal cervical spine as a method to discriminate hypo-lordosis: Results of elliptical and circular modeling in 72 asymptomatic subjects, 52 acute neck pain subjects, and 70 chronic neck pain subjects. *Spine*. In press.
- (Harrison et al., 2002) Harrison, D. E., Cailliet, R., Harrison, D. D., Janik, T. J., and Holland, B. (2002). Changes in sagittal lumbar configuration with a new method of extension traction combined with spinal manipulation and its clinical significance: Non-randomized clinical control trial. *Archives of Physical Medicine and Rehabilitation*, 83(11):1585–1591.
- (Heavenrich et al., 1991) Heavenrich, R. M., Murrell, J. D., and Hellman, K. H. (1991). *Light Duty Automotive Technology and Fuel Economy Trends through 1991*. U.S. Environmental Protection Agency.
- (Heiberger, 1989) Heiberger, R. M. (1989). *Computation for the Analysis of Designed Experiments*. Wiley.
- (Heiberger, 1998) Heiberger, R. M. (1998). Design of statistical graphs. Philadelphia Chapter of the American Statistical Association.
- (Heiberger, 2001) Heiberger, R. M. (2001). Emacs Speaks Statistics: One interface — many programs. In Hornik, K. and Leisch, F., editors, *Proceedings of the 2nd International Workshop on Distributed Statistical Computing (DSC 2001)*. Technische Universität Wien, Vienna, Austria.  
<http://www.ci.tuwien.ac.at/Conferences/DSC.html>, ISSN 1609-395X.
- (Heiberger and Harrell, 1994) Heiberger, R. M. and Harrell, Jr., F. E. (1994). Design of object-oriented functions in S for screen display, interface and con-

- trol of other programs (SAS and L<sup>A</sup>T<sub>E</sub>X), and S programming. In *Computing Science and Statistics*, volume 26, pages 367–371. The software is available in both S-PLUS and R in library(hmisc). It is included in the S-PLUS distribution. It may be downloaded for R from the contrib page of the (R Development Core Team, 2004) website.
- (Heiberger and Holland, 2002) Heiberger, R. M. and Holland, B. (2002). New display functions for Trellis. In *2002 Insightful Technology Conference*. Insightful Corp.  
[http://www.insightful.com/events/2002uc/poster/heiberger\\_splus2002.pdf](http://www.insightful.com/events/2002uc/poster/heiberger_splus2002.pdf).
- (Heiberger and Holland, 2003a) Heiberger, R. M. and Holland, B. (2003a). Statistical analysis and data display. Short Course, Insightful 2003 Technology Conference.  
[http://www.insightful.com/events/2003uc/temple\\_course.asp](http://www.insightful.com/events/2003uc/temple_course.asp).
- (Heiberger and Holland, 2003b) Heiberger, R. M. and Holland, B. (2003b). Trellis extensions. In Hornik, K., Leisch, F., and Zeileis, A., editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, ISSN 1609-395X. Technische Universität Wien, Vienna, Austria.  
[www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/HeibergerHolland.pdf](http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/HeibergerHolland.pdf).
- (Heiberger and Holland, 2004a) Heiberger, R. M. and Holland, B. (2004a). Mean-mean multiple comparison plots for arbitrary contrasts. Technical report, Temple University, Department of Statistics.
- (Heiberger and Holland, 2004b) Heiberger, R. M. and Holland, B. (2004b). *Statistical Analysis and Data Display: An Intermediate Course: Accompanying CD*. Springer-Verlag, New York.  
<http://springeronline.com>.
- (Heiberger and Teles, 2002) Heiberger, R. M. and Teles, P. (2002). Displays for direct comparison of ARIMA models. *The American Statistician*, 56:131–138, 258–260.
- (Hicks, 1964) Hicks, C. R. (1964). *Fundamental Concepts in Design of Experiments*. Saunders.
- (Higgins and Koch, 1977) Higgins, J. E. and Koch, G. G. (1977). Variable selection and generalized chi-square analysis of categorical data applied to a large cross-sectional occupational health survey. *International Statistical Review*, 45:51–62.
- (Hoaglin et al., 1983) Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors (1983). *Understanding Robust and Exploratory Data Analysis*. Wiley.
- (Hochberg, 1988) Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–803.
- (Hochberg and Tamhane, 1987) Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. Wiley.
- (Holzer, 2002) Holzer, M. (2002). Mild therapeutic hypothermia to improve the neurologic outcome after cardiac arrest. *New England Journal of Medicine*, 346(8):549–556.
- (Hosmer and Lemeshow, 2000) Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley, second edition.

- (Hsu and Peruggia, 1994) Hsu, J. and Peruggia, M. (1994). Graphical representations of tukey's multiple comparison method. *Journal of Computational and Graphical Statistics*, 3:143–161.
- (Iman, 1994) Iman, R. L. (1994). *A Data-Based Approach to Statistics*. Duxbury.
- (Insightful Corp., 2002) Insightful Corp. (2002). S-PLUS Statistical Software: Release 6.1.  
<http://www.insightful.com>.
- (Johnson and Leone, 1967) Johnson, N. L. and Leone, F. C. (1967). *Statistics and Experimental Design in Engineering and the Physical Sciences*, volume 2. Wiley.
- (Johnson and Tsao, 1945) Johnson, P. O. and Tsao, F. (1945). Factorial design and covariance in the study of individual educational development. *Psychometrika*, 10:133–162.
- (Johnson, 1996) Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, 4(1).  
<http://www.amstat.org/publications/jse/archive.htm>.
- (Knuth, 1984) Knuth, D. E. (1984). *The TeXbook*. Addison-Wesley.
- (Krantz, 1999) Krantz, L. (1999). *1999–2000 Jobs Rated Almanac: The Best and Worst Jobs—250 in All—Ranked by More Than a Dozen Vital Factors Including Salary, Stress, Benefits and More*. St. Martins Press.  
[http://www.hallmaps.com/almanacs\\_yearbooks/29.shtml](http://www.hallmaps.com/almanacs_yearbooks/29.shtml).
- (Lamport, 1986) Lamport, L. (1986). *L<sup>A</sup>T<sub>E</sub>X: A Document Preparation System*. Addison-Wesley.
- (Lavine, 1991) Lavine, M. (1991). Problems in extrapolation illustrated with space shuttle o-ring data. *Journal of the American Statistical Association*, 86:919–921.
- (Lea, 1965) Lea, A. J. (1965). New observations on distribution of neoplasms of female breast in certain European countries. *British Medical Journal*, 1:488–490.
- (Lee, 1980) Lee, E. T. (1980). *Statistical Methods for Survival Data Analysis*. Lifetime Learning Publications.
- (Lehmann, 1998) Lehmann, E. (1998). *Nonparametrics—Statistical Methods Based on Ranks*. Prentice Hall, revised first edition.
- (Lewin and Shakun, 1976) Lewin, A. Y. and Shakun, M. F. (1976). *Policy Sciences, Methodology and Cases*. Pergamon Press.
- (Little and Rubin, 2002) Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, second edition.
- (Longley, 1967) Longley, J. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. *Journal of the American Statistical Association*, 62:819–841.
- (Mallows, 1973) Mallows, C. L. (1973). Some comments on  $C_p$ . *Technometrics*, 15(4):661–675.
- (McCullagh and Nelder, 1983) McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. Chapman and Hall.

- (Mehta et al., 2000) Mehta, C. R., Patel, N. R., and Senchaudhuri, P. (2000). Efficient Monte Carlo methods for conditional logistic regression. *Journal of the American Statistical Association*, 95(449):99–108.
- (Mendenhall et al., 1989) Mendenhall, W. M., Parsons, J. T., Stringer, S. P., Cassissi, N. J., and Million, R. R. (1989). T2 oral tongue carcinoma treated with radiotherapy: Analysis of local control and complications. *Radiotherapy and Oncology*, 16:275–282.
- (Milliken and Johnson, 1984) Milliken, G. A. and Johnson, D. E. (1984). *Analysis of Messy Data*, volume I. Wadsworth.
- (Montgomery, 1997) Montgomery, D. C. (1997). *Design and Analysis of Experiments*. Wiley, 4<sup>th</sup> edition.
- (Montgomery, 2001) Montgomery, D. C. (2001). *Design and Analysis of Experiments*. Wiley, 5<sup>th</sup> edition.
- (Moore and McCabe, 1989) Moore, D. S. and McCabe, G. P. (1989). *Introduction to the Practice of Statistics*. Freeman.
- (Mosteller and Tukey, 1977) Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley.
- (Murray et al., 1981) Murray, J. D., Dunn, G., Williams, P., and Tarnopolksky, A. (1981). Factors affecting the consumption of psychotropic drugs. *Psychological Medicine*, 11:551–560.
- (Narula and Wellington, 1977) Narula, S. C. and Wellington, J. T. (1977). Prediction, linear regression and the minimum sum of errors. *Technometrics*, 19:185–190.
- (Neter et al., 1996) Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied Linear Statistical Models*. Irwin, fourth edition.
- (Nicholls, 1979) Nicholls, D. F. (1979). The analysis of time series — the time domain approach. *The Australian Journal of Statistics*, 21:93–120.
- (NIST, 2002) NIST (2002). National Institute of Standards and Technology, Statistical Engineering Division.  
<http://www.itl.nist.gov/div898/software/dataplot.html/datasets.htm>.
- (Olympic Committee, 2001) Olympic Committee (2001). Salt Lake City 2002 Winter Olympics.  
<http://www.saltlake2002.com>.
- (Ott, 1993) Ott, R. L. (1993). *An Introduction to Statistical Methods and Data Analysis*. Duxbury, fourth edition.
- (Pearce, 1983) Pearce, S. C. (1983). *The agricultural field experiment*. Wiley.
- (Penrose et al., 1985) Penrose, K., Nelson, A., and Fisher, A. (1985). Generalized body composition prediction equation for men using simple measurement techniques (abstract). *Medicine and Science in Sports and Exercise*, 17(2).
- (Peterson, 1990) Peterson, D. H., editor (1990). *Aspects of Climate Variability in the Pacific and the Western Americas*. Number 55 in Geophysical Monograph. American Geophysical Union.
- (Peterson, 1985) Peterson, R. G. (1985). *Design and Analysis of Experiments*. Marcel Dekker.
- (R Development Core Team, 2004) R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical

- Computing, Vienna, Austria. ISBN 3-900051-00-3,  
<http://www.r-project.org/>.
- (Red Hat, Inc., 2002) Red Hat, Inc. (2002). Cygwin: A Unix Environment for Windows.  
<http://sources.redhat.com/cygwin/>.
- (Robinson, 1950) Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15:351–357.
- (Rossini et al., 2004a) Rossini, A. J., Heiberger, R. M., Sparapani, R. A., Mächler, M., and Hornik, K. (2004a). Emacs Speaks Statistics (ESS): A multiplatform, multipackage development environment for statistical analysis. *Journal of Computational and Graphical Statistics*, 13(1):247–261.
- (Rossini et al., 2004b) Rossini, A. J., Mächler, M., Hornik, K., Heiberger, R. M., Sparapani, R., and Eglen, S. (2004b). ESS (Emacs Speaks Statistics).  
<http://www.analytics.washington.edu/downloads/ess/> or  
<http://lib.stat.cmu.edu/general/ESS/>.
- (Rossman, 1994) Rossman, A. J. (1994). Televisions, physicians, and life expectancy. *Journal of Statistics Education*.  
<http://www.amstat.org/publications/jse/archive.htm>.
- (Sarkar, 1998) Sarkar, S. (1998). Some probability inequalities for ordered  $MTP_2$  random variables: A proof of the Simes conjecture. *Annals of Statistics*, 26:494–504.
- (SAS Institute, 2000) SAS Institute (2000). SAS Statistical Software: Release 8.0.  
<http://sas.com>.
- (SAS Institute, Inc., 1999) SAS Institute, Inc. (1999). The four types of estimable functions. In *SAS/STAT User's Guide*. SAS Institute, Inc.
- (Schenk, 2001) Schenk, C. (2001). MikTeX.  
<ftp://metalab.unc.edu/pub/packages/TeX/systems/win32/miktex>.
- (Searle, 1971) Searle, S. R. (1971). *Linear Models*. Wiley.
- (Senie et al., 1981) Senie, R. T., Rosen, P. P., Lesser, M. L., and Kinne, D. W. (1981). Breast self-examinations and medical examination relating to breast cancer stage. *American Journal of Public Health*, 71:583–590.
- (Shaw, 1942) Shaw, N. (1942). *Manual of Meteorology*, volume 1. Cambridge University Press.
- (Shih and Weisberg, 1986) Shih, W. J. and Weisberg, S. (1986). Assessing influence in multiple linear regression with incomplete data. *Technometrics*, 28:231–240.
- (Simpson et al., 1975) Simpson, J., Olsen, A., and Eden, J. C. (1975). A Bayesian analysis of a multiplicative treatment effect in weather modification. *Technometrics*, 17:161–166.
- (Smith and Gonick, 1993) Smith, W. and Gonick, L. (1993). *The Cartoon Guide to Statistics*. HarperCollins.
- (Sokal and Rohlf, 1981) Sokal, R. R. and Rohlf, F. J. (1981). *Biometry*. W.H. Freeman, second edition.
- (Stallman, 2000) Stallman, R. M. (2000). *GNU Emacs Manual, for Version 20.7*. Free Software Foundation, 13th edition.

- (Steel and Torrie, 1960) Steel, R. G. D. and Torrie, J. H. (1960). *Principles and Procedures of Statistics*. McGraw-Hill, first edition.
- (Sulzberger, 1953) Sulzberger, P. H. (1953). The effects of temperature on the strength of wood, plywood and glued joints. Technical report, Aeronautical Research Consultative Committee, Australia, Department of Supply.
- (Teasdale et al., 1993) Teasdale, N., Bard, C., LaRue, J., and Fleury, M. (1993). On the cognitive penetrability of posture control. *Experimental Aging Research*, 19:1–13.
- (Till, 1974) Till, R. (1974). *Statistical methods for the earth scientist*. Macmillan.
- (Tufte, 2001) Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Graphics Press, second edition.
- (Tukey, 1949) Tukey, J. W. (1949). One degree of freedom for nonadditivity. *Biometrics*, 5(3):232–242.
- (Vandaele, 1978) Vandaele, W. (1978). Participation in illegitimate activities: Erlich revisited. In Blumstein, A., Cohen, J., and Nagin, D., editors, *Deterrence and Incapacitation*, pages 270–335. National Academy of Sciences.
- (VanVliet and Gupta, 1973) VanVliet, P. K. and Gupta, J. M. (1973). Thiam-v-sodium bicarbonate in idiopathic respiratory distress syndrome. *Archives of Disease in Childhood*, 48:249–255.
- (Venables and Ripley, 1997) Venables, W. N. and Ripley, B. D. (1997). *Modern Applied Statistics with S-PLUS*. Springer, second edition.
- (Vlachos, 2004) Vlachos, P. (2004). Statlib: Data, software and news from the statistics community.  
<http://lib.stat.cmu.edu>.
- (Wainer, 1997) Wainer, H. (1997). *Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*. Copernicus Books.
- (Wei, 1990) Wei, W. W. S. (1990). *Time Series Analysis, Univariate and Multivariate Methods*. Addison-Wesley.
- (Weindling et al., 1986) Weindling, A. M., Bamford, F. M., and Whittall, R. A. (1986). Health of juvenile delinquents. *British Medical Journal*, 292:447.
- (Weisberg, 1985) Weisberg, S. (1985). *Applied Linear Regression*. Wiley, second edition.
- (Westbrooke, 1998) Westbrooke, I. (1998). Simpson's paradox: An example in a New Zealand survey of jury composition. *Chance*, 11:40–42.
- (Westfall and Rom, 1990) Westfall, P. H. and Rom, D. (1990). Bootstrap step-down testing with multivariate location shift data. Unpublished.
- (Williams, 2001) Williams, A. F. (2001). Teenage passengers in motor vehicle crashes: A summary of current research. Technical report, Insurance Institute for Highway Safety, Arlington, VA.
- (Williams, 1959) Williams, E. J. (1959). *Regression Analysis*. Wiley.
- (Woods et al., 1986) Woods, N., Fletcher, P., and Hughes, A. (1986). *Statistics in Language Studies*. Cambridge University Press.
- (World Almanac and Book of Facts, 2001) World Almanac and Book of Facts (2001). *World Almanac and Book of Facts*. World Almanac Books, 2002 edition.

- (Wynder et al., 1958) Wynder, E. L., Naravvette, A., Arostegui, G. E., and Llambes, J. L. (1958). Study of environmental factors in cancer of the respiratory tract in Cuba. *Journal of the National Cancer Institute*, 20:665–673.
- (Yates, 1934) Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. *Journal of the American Statistical Association*, 29:51–56.
- (Yates, 1937) Yates, F. (1937). *The Design and Analysis of Factorial Experiments*. Imperial Bureau of Soil Science, Harpenden, U.K.

---

# List of Datasets

abrasion.dat, 264  
acacia.dat, 508  
animal.dat, 377  
anneal.dat, 483  
apple.dat, 452  
balance.dat, 520  
barleyp.dat, 376  
batch.dat, 137  
bean.dat, 483  
blood.dat, 151  
blyth.dat, 495  
breast.dat, 210  
budworm.dat, 540  
byss.dat, 561  
c3c4.dat, 184  
catalystm.dat, 124, 148, 180  
cc135.dat, 449  
cc176.dat, 381  
cereals.dat, 61  
chimp.dat, 484  
circuit.dat, 444  
concord.dat, 89, 325  
crash.dat, 472  
crime.dat, 508  
darwin.dat, 526  
diamond.dat, 211  
display.dat, 329  
distress.dat, 526  
draft70mm.dat, 15, 61, 87, 151  
drunk.dat, 488  
eggs.dat, 422  
employM16.dat, 617  
fabricwear.dat, 277  
fabriwear.dat, 296  
fat.data, 189, 221, 235, 419  
feed.dat, 372  
filmcoat.dat, 397, 398  
filter.dat, 421  
furnace.dat, 377  
girlht.dat, 211  
glasses.dat, 493  
gunload.dat, 401  
har1.dat, 121, 516–519  
har2.dat, 121  
hardness.dat, 89, 230, 324  
heartvalve.dat, 421  
hooppine.dat, 378  
hospital.dat, 324  
hotdog.dat, 284  
houseprice-erie.dat, 264, 295, 326  
houseprice.dat, 225, 226  
htwt.dat, 268  
icu.dat, 559  
income.dat, 211  
intubate.dat, 510  
ironpot.dat, 378  
kidney.dat, 325  
lake.dat, 210

leukemia.dat, 561  
lifeins.dat, 326  
longley.dat, 237  
lymph.dat, 542  
manhours.dat, 266  
market.dat, 425  
mice.dat, 152  
mileage.dat, 325  
mortality.dat, 508  
mpg.dat, 378  
muscle.dat, 211  
njgolf.dat, 65  
notch.dat, 150  
nottem.dat, 604  
oats.dat, 432  
operator.dat, 150  
ozone.dat, 613  
patient.dat, 15, 150  
plasma.dat, 344  
political.dat, 509  
potency.dat, 151  
pox.dat, 526  
product.dat, 577, 604  
pulse.dat, 150, 523  
radioact.dat, 484  
rent.dat, 297  
retard.dat, 377  
rhiz1-alfalfa.dat, 355, 374  
rhiz3-clover.dat, 355, 375  
salary.dat, 60  
salinity.dat, 148  
salk.dat, 506  
seeding.dat, 526  
selfexam.dat, 508  
shipment.dat, 266  
sickle.dat, 150  
skateslc.dat, 377  
spacshu.dat, 529, 530  
spindle.dat, 426  
sprint.dat, 265  
surface.dat, 426  
tablet1.dat, 151, 526  
teachers.dat, 102  
testing.dat, 376  
tires.dat, 390, 423  
tongue.dat, 559  
tser.mystery.X.dat, 590  
tser.mystery.Y.dat, 596  
tser.mystery.Z.dat, 600  
tsq.dat, 609  
turkey.dat, 137, 162, 410, 430  
tv.dat, 30, 71, 643  
usair.dat, 87, 259  
uscrime.dat, 88, 266  
vocab.dat, 96, 512, 513  
vulcan.dat, 422  
washday.dat, 424  
water.dat, 88, 295  
weightloss.dat, 16, 157  
weld.dat, 424  
wheat.dat, 484  
wool.dat, 422  
workstation.dat, 347

---

# Index

- $\Phi$ , 49  
 $\alpha$ , 43, 45  
 $\beta$ , 45  
 $\cap$ , 21–22  
 $\chi^2$ , 29  
 $\cup$ , 21–22  
 $\epsilon$ , 236  
 $\varepsilon$ , 566  
 $\eta$ , 28  
 $\mu$ , 27, 34  
 $\bar{X}$ , 28  
 $\rho$ , 35  
 $P$  (uppercase  $\rho$ ), 35  
 $\sigma$ , 27, 35  
 $\sigma^2$ , 27, 219  
 $X^+$ , 218  
.Data, 647  
.RData, 647  
=, 223  
~, 223  
8.3 filename, 627
- Absolute-sum-2 scaling, 146, 178  
accuracy, 676  
ACF, *see* autocorrelation  
acid phosphatase, 542  
ACM, 623  
added residual plots, 258  
added variable plots, 254–259, 310, 313  
AIC, *see* Akaike information criterion  
Akaike information criterion, 249, 261, 572  
Akaike, Hirotugu, 249  
algebra, 683  
algorithm, 690
- alias, 429, 431, 443  
alignment, 13, 643, 666  
time series, 567  
analysis of covariance, *see* ANCOVA  
analysis of deviance table, 539  
analysis of variance, *see* ANOVA  
analysis of variance table, *see* ANOVA  
table  
analysis with concomitant variables, *see* ANCOVA  
ANCOVA, 283–295, 382–384, 452–472  
ANCOVA plot, 304, 308, 383, 387  
ANOVA, 123–153, 284, 286, 329–379  
computation, 379  
ANOVA table, 124–130, 193, 300,  
302–304, 308, 339  
antilogit, 527, 528  
ARIMA, 566, 567, 569  
ARMA, 570  
ASCII files, 653  
ASCII format, 14, 16–18  
asymmetry, 84  
autocorrelation, 570  
autoregression (AR), 567
- B**ackshift operator, 566  
backslash, 631  
barplot, 472, 474, 480  
Bell Labs, 623  
Bernoulli distribution, 538  
beta distribution, 658  
bias, 43, 56  
binomial distribution, 37–38, 513–514,  
657, 661  
binomial test, 513

- bivariate discrete distribution, 487  
 bivariate normal distribution, 35, 36  
 block, 341  
 blocking factor, 339, 341  
 Bonferroni inequality, 156  
 Bonferroni method, 156–157  
 Box, George E. P., 567  
 Box–Cox transformations, 79, 478  
 Box–Jenkins method, 567  
 boxplot, 32–34, 285, 356, 402, 427  
 Brown–Forsyth test, 148–149  
 bunching, 257  
 byssinosis, 561
- Calculus**, 685, 686  
 cancer, 542  
 carryover effect, 448  
 Cartesian product, 338, 703–708  
 case, 297, 503  
 case-control study, 503  
 categorical variable, 487  
 CD, *see* online files  
   burning, ix  
 cell, 336, 487  
 cell means, 337  
 Central Limit Theorem, 40–41  
 Chambers, John M., 623  
 changeover design, 448  
 chi-square analysis, 487–492  
 chi-square distribution, 490, 659  
 Cholesky Factorization, 695  
 class variable, 11  
 cluster random sampling, 54–55  
 Cochran, William G., 504  
 code, computer, ix  
 coded data, 376  
 coding, 11, 163, 267–268, 270, 277, 294, 295, 376  
 coefficient of determination, 193, 194, 196, 207, 209, 248, 261  
 cohort study, 503  
 collinearity, 236–244, 324  
 column space, 276  
 combinations, 700  
 command language, ix  
 comparison value, 474–477  
 computing note, 159  
 concomitant variable, 283  
 concomitant variables, analysis of, *see* ANCOVA  
 conditional probability, 22  
 conditional tests, 417–421  
 confidence bands, 202, 204  
   logistic regression, 530  
 confidence interval, 42–45  
   matched pairs of means, 101–102  
   one-sided, 44  
   population mean, 91, 95–96
- population proportion, 93–94  
 population variance, 96  
 two means, 98–100  
 two proportions, 98  
 variance ratio, 100  
 conflicts, names, 645  
 confounding, 429–444, 503  
 conservative, 157  
 consistency, 43  
 contingency table, 9  
 contingency tables, 487–510  
 contrast matrix, 275  
 contrast vector, 145  
 contrasts, 137, 143–147, 179, 373, 375, 410  
   arbitrary, 178–180  
   orthogonal, 139–143, 180–181, 440, 441  
 controls, 503  
 Cook's distance, 205, 209, 314, 315, 319–321  
 Cook, R. Dennis, 315, 319  
 correlation, 34–37  
 correlation coefficient, 195, 197  
 Courier, 665  
 covariance, 34–37  
 covariance adjustment, 283, 289, 291, 292, 457, 460  
 covariance matrix, 34, 35  
 covariance, analysis of, *see* ANCOVA  
 covariate, 283, 284, 289, 295  
 $C_p$ , 248–250, 253, 261  
 $C_p$  plot, 252, 262  
 cross-product ratio, 499  
 crossing, 224, 379, 410, 433  
 crossover design, 429, 448–451  
 customize, x  
 Cygwin, 627
- Daniel, Cuthbert, 248  
 .Data, 647  
 data  
   categorical, 11  
   continuous, 12  
   count, 11  
   discrete, 12  
   importing, 14  
   interval, 11, 12, 512  
   missing, 14, 16–17, 269  
   multivariate, 12  
   ordered, 11, 12, 512, 523  
   ratio, 11, 12, 512  
   rearrangement, 15  
   types of, 11  
 data display, 2, 191  
 data snooping, 131  
 datasets, *see* List of Datasets  
 datasets for HH, 18  
 degrees of freedom, 39, 230

- deleted predicted value, 314  
 deleted regression coefficients, 314  
 deleted standard deviation, 314, 316–317  
 deviance, 539, 540  
**DFBETAS**, 314, 315, 322–323  
**dffits**, 314, 315, 321–322  
 diagnostic  
     logistic regression, 553  
 diagnostic plot, 476, 477  
     time series, 579  
 diagnostics  
     time series, 572, 584  
 dichotomous, 41, 225, 527, 557, 561  
 differencing, 569  
 direct effect, 448  
 directory structure, x, 664  
 discrete distribution, 487  
 discretization, 280  
 distribution, *see* the name of the  
     particular distribution  
 dummy variable, 267–270, 273, 274, 277,  
     289, 295, 413–417, 481, 484  
**Dunnett procedure**, 157–161, 458, 459  
  
**Ecological correlation**, 64–65, 498  
 editors, 663–681  
 efficiency, 338, 457  
 eigenvalues, 696  
 elementary operations, 688  
**Emacs**, 624, 633, 635–639, 644, 647,  
     667–673  
     buffer, 668  
     shell mode, 669  
 empirical cumulative distribution plot,  
     207, 208  
**EMS**, *see* expected mean squares, 439  
 end-of-line convention, 654–655  
**EOL convention**, *see* end-of-line  
     convention  
 error, *see* residuals  
**ESS**, 647, 667–673  
 estimate, 41  
 estimation, 41–45, 572  
 estimator, 41  
     point, 42–43  
 exact test, 514  
**Excel**, Microsoft, 14–16, 674–675  
 exhortations, 677  
 expectation, 27  
 expected mean squares, 135–136, 339,  
     346, 402–404, 439  
 experimental units, 431, 442  
 exponential distribution, 658  
 externally standardized residuals, *see*  
     Studentized deleted residuals  
 extra sum of squares, 229  
  
 $\mathcal{F}$ , 26
- F*-distribution, 660  
**F-test**, 98, 100, 124, 128, 130, 135, 136,  
     146, 147, 193, 195, 523  
 factor, 11, 123, 487  
 factorial, 700  
 family (of related inferences), 155  
 familywise error rate, 130, 155, 156  
 figure, ix  
 filename, ix  
 files, online, *see* online files  
 Fisher's exact test, 492–495, 508  
 Fisher, Ronald A., 492  
 fitting constants, 418  
 fixed effects, 123, 127, 341–342  
 fixed factor, 127  
 folding, 645, 667  
 font, 643, 665  
 forecasting, 574, 589  
 formatting, 668  
 forward slash, 631  
 fractional factorial design, 429–431,  
     442–447  
 fractional replicate, 431, 442  
 full model, 228  
 function, ix  
**FWE**, *see* familywise error rate  
  
**Gaussian elimination**, 218  
 generalized inverse, 218, 698  
 generalized linear model, 529, 539  
 geometry, 223  
 geometry of matrices, 695  
 Ghostscript, 18, 626  
 Ghostview, 18, 626  
**glm**, *see* generalized linear model  
**GOF**, *see* goodness-of-fit  
 goodness-of-fit test, 106–109, 114–116,  
     118, 491, 512  
     portmanteau, 572  
 Gram–Schmidt algorithm, 694  
 grand mean, 337  
 granularity, 257  
 graph, 356  
 graphical design, 2, 9, 17, 169, 553–556,  
     703–708  
     ACF plot, 619  
     ANCOVA plot, 284, 294, 707  
     ARIMA-trellis plot, 707  
     barplot, 472  
     boxplot, 356–357, 412, 427  
     common scaling, 554  
     construction, 620  
     interaction plot, 338–339, 707  
     logistic regression plot, 706  
     MMC plot, 168–182, 707  
     odds-ratio CI plot, 502, 708  
     ODOFFNA plot, 477–479, 707  
     regression diagnostic plots, 305, 306

- graphical (*cont.*)
  - seasonal time series, 582
  - squared residual plot, 213, 708
  - time series, 575–580, 618–620
  - time series plot, 619
  - transposed trellis plot, 708
- graphical display
  - logistic regression, 530–531
- graphical user interface, 703
- grid** library, 647
- GSview, 18, 626
- GUI, *see* graphical user interface
- gunzip**, 18
- gzip**, 18
- Haenszel**, William, 504
- hat matrix, 219, 220, 315, 327
- Helmert contrasts, 276, 362
- HH library, 632–638, 640, 641, 647, 649–653
- hh** (S-PLUS function), 19
- hh** (SAS macro), 19
- hierarchical factorial relationship, 347
- high leverage point, 220
- higher way designs, 381–426
- histogram, 30–31
- Hochberg procedure, 157
- HOME** directory, 627
- hov**, *see* variance, homogeneity of
- Hsu, Jason, 173
- hypergeometric distribution, 38, 492–494, 657
- hypothesis test
  - matched pairs of means, 101–102
  - one-sided, 48
  - population mean, 92–93
  - population proportion, 94–95
  - population variance, 97
  - two means, 99
  - two variances, 100
  - two-sided, 48
- hypothesis testing, 45–50
- Identification**, 571
- ill-conditioned data, 236
- imputation, 16
- indentation, 643
- independence, 22, 25–27, 490–492
- indicator variable, *see* dummy variable
- inductive inference, 3
- inexplicable error messages, 645
- influence, 209, 305, 307, 320, 324
- Insightful Corporation, 623
- integer scaling, 147
- interaction, 6, 139, 282, 304, 330–339
  - models without, 371–372
- interaction plot, 331, 338–339, 349–350, 364, 373, 378, 385, 386, 399, 423, 434, 445, 474, 479
- internally standardized residuals, *see* standardized residuals
- intersection, *see*  $\cap$
- Jenkins, Gwilym M., 567
- jitter, 530
- Kolmogorov–Smirnov** test, 107
- Kruskal–Wallis test, 523–525
- L'Hôpital's rule**, 82
- ladder of powers, 82, 479, 480
- lag, 566
- language
  - command, ix
  - LAT $\backslash$ EX, 626, 665
- Latin square design, 389–395, 423, 431, 442, 448–450, 457
- lattice, 467
- LD50, 540
- least squares, 188, 190, 191, 213, 218
- least-squares geometry, 215
- least-squares plane, 215
- level, 11, 123
- leverage, 200, 209, 220, 312, 314–316, 319
- likelihood ratio test, 119
- line width, 667
- linear dependence, 275
- linear equations, 699
- linear independence, 690
- linearly independent, 268
- link, 527
- link function, 528, 538
- Linux, *see* Unix
- logistic regression, 527–563
- logistic regression plot, 541
- logit, 527, 535
- logit scale, 535
- lognormal distribution, 660
- LogXact, 558
- longitudinal study, 466
- Lucent Technologies, 624
- lung, 561
- lymph nodes, 542
- Macro**, ix, 655
- MAD, 1, 149
- main effect, 337, 431
- Mallows, Colin, 248, 249
- Mann–Whitney test, 520–523
- Mantel, Nathan, 504
- Mantel–Haenszel test, 504–508, 510
- marginal means, 337
- margins, 667

- MathType, 627  
 matrix  
     geometry, 695  
 matrix algebra, 687  
 matrix factorization, 693  
 maximum likelihood, 188  
 maximum likelihood estimation, 118  
 mean polish, 476  
 mean square  
     error, *see* mean square, residual  
     residual, 193, 197–199  
 mean square, residual, 248  
 mean–mean display, *see* MMC plot  
 median, 28  
 median polish, 476  
 method of fitting constants, 417  
 Microsoft Windows, *see* Windows,  
     Microsoft  
 Microsoft Word, *see* Word, Microsoft  
 minus sign, 666, 679  
 missing data, *see* data, missing  
 mixed model, 346  
 MMC plot, 131, 134, 146, 161, 166–183,  
     335, 360, 361, 368–370, 388, 459  
 model formula, 223  
 model specification, 223–224, 404,  
     413–414, 556  
 monowidth font, 665  
 Moore–Penrose generalized inverse, 218,  
     698  
 moving average (MA), 568  
 MS-DOS, 627, 628, 631  
 MSE, *see* mean square, residual  
 multicollinearity, *see* collinearity  
 multinomial distribution, 660  
 multiple comparison procedures,  
     130–134, 155–185, 458  
 multiplicity, 156, 162–163  
 multivariate distribution, 34, 35  
 multivariate normal distribution, 36, 688
- N**  
 Name conflicts, 645  
 nested factorial experiment, 401–413  
 nesting, 224, 347–353, 362, 365, 401–413  
 new observation, 199–202  
 Newton's method, 687  
 NID, 188, 309, 336  
 no-intercept models, 211, 233–235  
 nominal variable, 11  
 nonadditivity, 474  
 noncentral chi-square distribution, 661,  
     662  
 noncentral distribution, 661–662  
 noncentral  $F$  distribution, 662  
 noncentral  $t$  distribution, 661, 662  
 noncentrality parameter, 661, 662  
 nonparametric methods, 511–526  
 nonparametric procedures, 127
- normal distribution, 38–39, 658  
 normal equations, 192, 218  
 normal probability plot, 110–113, 205,  
     207, 310, 311  
 normalized scaling, 146  
 notch, 34  
 numerical stability, 676
- $O(n)$ , 689  
 O.C. curve, *see* operating characteristic  
     curve  
 odds, 498–502, 527, 535  
 odds ratio, 498–502, 504, 509  
 odds scale, 535  
 odoffna, *see* one degree of freedom for  
     nonadditivity  
 one degree of freedom for nonadditivity,  
     472–484  
 online files, ix, 2, 14, 629, 631, 632, 635,  
     649, 650, 664  
 operating characteristic curve, 45–51,  
     661  
 operator symbols, 224  
 optimization, 686  
 order statistics, 28  
 ordered factor, 279  
 orientation, 427  
 origin, regression through, 211, 233–235  
 orthogonal basis set, 358, 361, 369, 693  
 orthogonal contrasts, *see* contrasts,  
     orthogonal, 360, 361, 368, 369  
 orthogonal matrix, 688  
 orthogonal polynomials, 277–282, 295,  
     695  
 orthogonal transformation, 692  
 outlier, 318, 512, 514–516, 520  
 output, ix
- p*-value, 47  
 PACF, *see* autocorrelation  
 paired  $t$ -test, 100, 515  
 parameter, 3  
 parameterization, 275  
 parsimony, 237, 244  
 partial  $F$ -tests, 228–230  
 partial correlation, 256  
 partial residual plots, 254–258, 312, 313  
 partial residuals, 254, 256, 258  
 permutations, 700  
 Peruggia, Mario, 173  
 placebo, 101, 503  
 plots, 431  
 p.m.f., *see* probability mass function  
 point cloud, 215, 223  
 Poisson distribution, 122, 657  
 polynomial, 277  
 polynomial contrasts, 277–282, 296, 440  
 polynomial function, 566

- polynomial model, 230–232, 243  
 pooling, 98, 371  
 population, 3  
 portmanteau goodness-of-fit test, 572  
*PostScript*, ix, 18, 626  
 power, 45, 136, 156, 511, 662  
 power curve, 45–51, 661  
 power transformations, 79  
 powers, ladder of, 82  
 precision, 13, 676  
 prediction, 235–236  
 prediction bands, 202, 204  
     logistic regression, 530, 532, 536  
 prediction interval, 200–203, 210,  
     235–236  
 predictor matrix, 220  
 predictor variable, 215, 221, 226, 267,  
     268, 280  
 presentation of results, 679  
 probability, 21–23  
 probability distribution, *see the name of*  
     the particular distribution  
 probability distributions, 22–41  
 probability mass function, 24  
 probability scale, 535  
 probit regression, 529  
 programming style, 678  
 projection matrix, 219, 695  
 prospective study, 503–504
- Q** quadratic form, 692  
 quadratic model, 231  
 quantile plot, 110–113, 121  
 quartiles, 32, 33
- R**, viii, 2, 631–648  
 R Development Core Team, 624  
 $R^2$ , *see* coefficient of determination  
     adjusted, 194, 195, 249, 261  
 random effects, 123, 135–137, 341–342  
 random effects model, 346  
 random factor, 135  
 random sample, 3  
 random sampling, 53  
 random shock, 566  
 random variable, 21–28  
 random vector, 34, 687  
 randomization, 52  
 randomization test, 514  
 randomized complete block design,  
     342–344, 457  
 rank, 12, 512, 516–526, 691  
 RCBD, *see* randomized complete block  
     design  
 .RData, 647  
 reduced model, 228  
 regression analysis  
     diagnostics, 297–327  
     multiple linear regression, 215–266  
     simple linear regression, 187–213  
     using dummy variables, 267–294  
 regression coefficients, 188, 190, 194, 196,  
     199, 200, 218, 219, 222, 227, 263,  
     267, 271, 277, 300, 302, 303  
 regression diagnostics, 205–209, 314  
 relative risk, 498–502  
 repeated measures design, 448  
 residual effect, 448  
 residual effects design, 448–451  
 residual mean square, *see* mean square,  
     residual  
 residual plot, 206, 207, 309–312  
 residual sum of squares, *see* sum of  
     squares, residual  
 residuals, 190, 199, 219  
 response surface, 374  
 response surface methodology, 440  
 retrospective study, 503–504, 509  
 rhizobium, 353  
 risk factor, 503  
 root mean square error, 192  
 rounding, 13–14  
 r.v., *see* random variable
- S**, *see* S-PLUS and R  
 S language, 640, 643  
 S-PLUS, viii, 2, 631–648  
 sample, 3  
 sample proportion, 527  
     models, 557  
 sample size, 45  
 sample size determination, 105–106  
 sampling, 52–55  
 sampling distribution, 40–41  
 SAS, viii, 2, 649–656  
 SAS Institute, 624  
 SAS language, 655  
 Satterthwaite option, 100  
 scaled deviation, 491  
 scaling, 178  
 scatterplot, 65–68  
 scatterplot matrix, 67–70, 74–78, 87,  
     189, 209, 223, 225, 226, 228, 244,  
     260, 269, 298, 299, 301, 303, 309,  
     338, 468, 469, 541, 549, 550, 554,  
     705–706  
 Scheffé procedure, 162–167  
 s.d., *see* standard deviation  
 seasonal models, 580–581  
 sequential tests, 417–421  
 Shapiro–Wilk test, 111, 310  
 sign test, 512–516  
 significant, 47  
 significant digits, 676  
 simple effects, 338, 362, 365–370, 396–401  
 simple random sampling, 53

- Simpson's paradox, 495–498, 509  
 simultaneous confidence intervals, 131–132  
 singular value decomposition, 698  
 skewness, 28–29  
 slash, 631  
 software, 16, 19, 623–629  
 space shuttle, 529  
 split plot design, 429–441  
 splom, *see* scatterplot matrix  
 springeronline.com, ix  
 squared residual plot, 191, 216  
 standard deviation, 27  
 standard error, 192, 198, 201  
 standard error of estimate, 192, 219  
 standard error of sample mean, 56  
 standardized residuals, 314, 317–319  
 start value, 80  
 stationarity, 565  
 statistic, 3  
 statistical model, 41–42, 127, 218, 263, 336, 403, 432  
 statistically significant, 47  
 statistics, 3  
 statistics profession, 4  
 StatLib, 703  
 stem-and-leaf display, 31–32, 269, 270, 514, 515  
 stepwise regression, 244, 247–250  
   all subsets, 248  
   backward elimination, 247  
   forward selection, 247  
 stochastic process, 566  
 stratified random sampling, 53–54  
 Student's *t* distribution, 39–40, 659, 661  
 Studentized deleted residuals, 314, 315, 317–319  
 Studentized range distribution, 130, 353, 659  
 subplot, 431, 432  
 sufficiency, 43  
 sum contrasts, 414–415  
 sum of squares, 219  
   regression, 194, 195  
   residual, 192, 194, 195, 197  
   total, 194, 195, 197  
 surveys, 6  
 symmetry, 77  
 systematic random sampling, 55
- t* distribution, 39–40  
*t*-test, 100, 102, 203, 205, 516, 520  
 TAB characters, 653  
 text editors, 624–625, 669  
 time series analysis, 565–622  
 Times Roman, 665  
 total sum of squares, *see* sum of squares, total
- trace factor, 339  
 transcript, ix  
 transformation, 29, 78–86, 127, 259, 309, 512, 527  
 treatment, 339  
 treatment combinations, 336, 442  
 trellis, 704  
 Tukey procedure, 130–132, 157, 168, 171, 291, 330, 334, 335, 352, 357, 359–361, 370, 382, 388, 394, 395  
 Tukey, John, 28, 32, 472  
 Type I error, 45–49  
 Type I Sum of Squares, 224, 417–421  
 Type II error, 45–46, 661, 662  
 Type II Sum of Squares, 417–419  
 Type III Sum of Squares, 332–334, 417–421  
 Type IV Sum of Squares, 419  
 typography, 665
- Unbalanced sampling, 348  
 unbiased, 42  
 union, *see*  $\cup$   
 Unix, xi, 627, 635–637
- Variable selection, 243–250, 253  
 variance, 27  
   accuracy, 676  
   homogeneity of, 127, 147–149  
 variance function, 538  
 variance inflation factor, 242–246, 324  
 variance stabilization, 78, 476, 478  
 variance, analysis of, *see* ANOVA  
 variance-covariance matrix, *see* covariance matrix  
 VIF, *see* variance inflation factor
- Web page, ix  
 weighted squares of means, 419  
 Welch two-sample *t*-test, 100  
 whole plot, 429–431  
 whole plot error, 433  
 Wilcoxon signed-ranks test, 516–519  
 Wilcoxon, Frank, 516, 520  
 Wilk–Shapiro test, 111, 310  
 Windows, Microsoft, xi, 627, 632–635, 638–639  
 word processing software, 625–626, 645  
 Word, Microsoft, 625, 673–674  
 working style, 664  
 writing style, 677, 679
- $X^+$ , 218  
 $x$ -factor, 339  
 xysplom, 535, 540, 705, 706
- Yates, Frank, 417

## Springer Texts in Statistics (*continued from page ii*)

---

|  |  |
|--|--|
| <i>Lehmann</i>                           | Elements of Large-Sample Theory  |
| <i>Lehmann</i>                           | Testing Statistical Hypotheses, Second Edition   |
| <i>Lehmann and Casella</i>               | Theory of Point Estimation, Second Edition   |
| <i>Lindman</i>                           | Analysis of Variance in Experimental Design  |
| <i>Lindsey</i>                           | Applying Generalized Linear Models   |
| <i>Madansky</i>                          | Prescriptions for Working Statisticians  |
| <i>McPherson</i>                         | Applying and Interpreting Statistics: A Comprehensive Guide, Second Edition                              |
| <i>Mueller</i>                           | Basic Principles of Structural Equation Modeling:<br>An Introduction to LISREL and EQS                   |
| <i>Nguyen and Rogers</i>                 | Fundamentals of Mathematical Statistics, Volume I:<br>Probability for Statistics                         |
| <i>Nguyen and Rogers</i>                 | Fundamentals of Mathematical Statistics, Volume II:<br>Statistical Inference                             |
| <i>Noether</i>                           | Introduction to Statistics: The Nonparametric Way  |
| <i>Nolan and Speed</i>                   | Stat Labs: Mathematical Statistics Through Applications  |
| <i>Peters</i>                            | Counting for Something: Statistical Principles and Personalities   |
| <i>Pfeiffer</i>                          | Probability for Applications   |
| <i>Pitman</i>                            | Probability  |
| <i>Rawlings, Pantula,<br/>and Dickey</i> | Applied Regression Analysis  |
| <i>Robert</i>                            | The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, Second Edition |
| <i>Robert and Casella</i>                | Monte Carlo Statistical Methods  |
| <i>Rose and Smith</i>                    | Mathematical Statistics with <i>Mathematica</i>  |
| <i>Ruppert</i>                           | Statistics and Finance: An Introduction  |
| <i>Santner and Duffy</i>                 | The Statistical Analysis of Discrete Data  |
| <i>Saville and Wood</i>                  | Statistical Methods: The Geometric Approach  |
| <i>Sen and Srivastava</i>                | Regressions Analysis: Theory, Methods, and Applications  |
| <i>Shao</i>                              | Mathematical Statistics, Second Edition  |
| <i>Shorack</i>                           | Probability for Statisticians  |
| <i>Shumway and Stoffer</i>               | Time Series Analysis and Its Applications  |
| <i>Simonoff</i>                          | Analyzing Categorical Data   |
| <i>Terrell</i>                           | Mathematical Statistics: A Unified Introduction  |
| <i>Timm</i>                              | Applied Multivariate Analysis  |
| <i>Toutenburg</i>                        | Statistical Analysis of Designed Experiments, Second Edition   |
| <i>Wasserman</i>                         | All of Statistics: A Concise Course in Statistical Inference   |
| <i>Whittle</i>                           | Probability via Expectation, Fourth Edition  |
| <i>Zacks</i>                             | Introduction to Reliability Analysis: Probability Models and Statistical Methods                         |

---