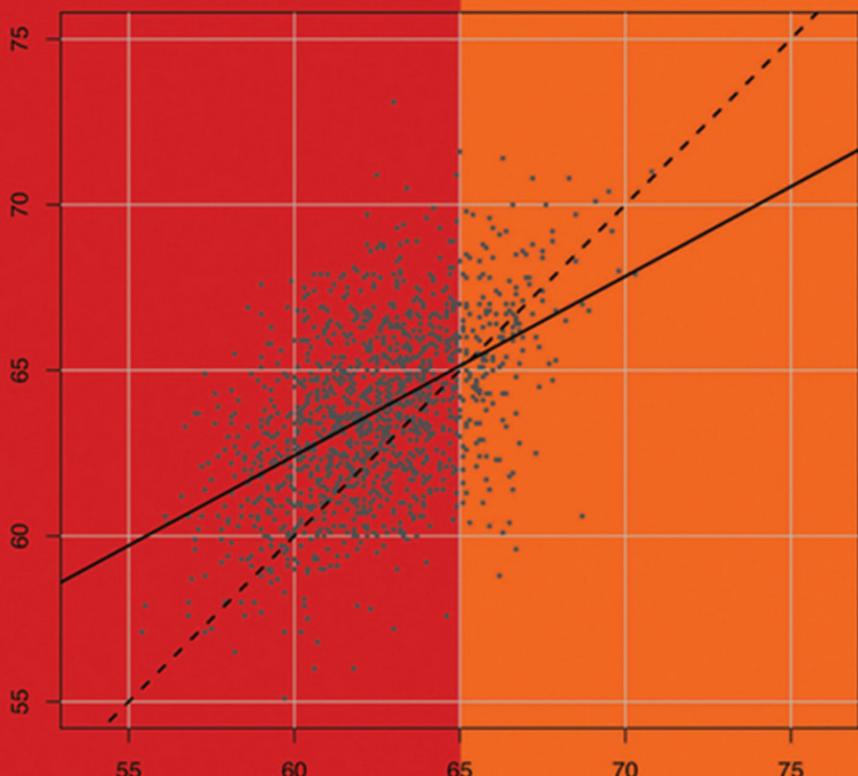


FOURTH EDITION

Applied Linear Regression

SANFORD WEISBERG



WILEY

Applied Linear Regression

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Harvey Goldstein, Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, Joseph B. Kadane, Jozef L. Teugels*

A complete list of the titles in this series appears at the end of this volume.

Applied Linear Regression

Fourth Edition

SANFORD WEISBERG

School of Statistics
University of Minnesota
Minneapolis, MN

WILEY

Copyright © 2014 by John Wiley & Sons, Inc. All rights reserved

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Weisberg, Sanford, 1947-

Applied linear regression / Sanford Weisberg, School of Statistics, University of Minnesota, Minneapolis, MN.—Fourth edition.

pages cm

Includes bibliographical references and index.

ISBN 978-1-118-38608-8 (hardback)

1. Regression analysis. I. Title.

QA278.2.W44 2014

519.5'36—dc23

2014026538

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To Carol, Stephanie,
and
the memory of my parents

Contents

Preface to the Fourth Edition	xv
1 Scatterplots and Regression	1
1.1 Scatterplots, 2	
1.2 Mean Functions, 10	
1.3 Variance Functions, 12	
1.4 Summary Graph, 12	
1.5 Tools for Looking at Scatterplots, 13	
1.5.1 Size, 14	
1.5.2 Transformations, 14	
1.5.3 Smoothers for the Mean Function, 14	
1.6 Scatterplot Matrices, 15	
1.7 Problems, 17	
2 Simple Linear Regression	21
2.1 Ordinary Least Squares Estimation, 22	
2.2 Least Squares Criterion, 24	
2.3 Estimating the Variance σ^2 , 26	
2.4 Properties of Least Squares Estimates, 27	
2.5 Estimated Variances, 29	
2.6 Confidence Intervals and <i>t</i> -Tests, 30	
2.6.1 The Intercept, 30	
2.6.2 Slope, 31	
2.6.3 Prediction, 32	
2.6.4 Fitted Values, 33	
2.7 The Coefficient of Determination, R^2 , 35	
2.8 The Residuals, 36	
2.9 Problems, 38	

3 Multiple Regression	51
3.1 Adding a Regressor to a Simple Linear Regression Model, 51	
3.1.1 Explaining Variability, 53	
3.1.2 Added-Variable Plots, 53	
3.2 The Multiple Linear Regression Model, 55	
3.3 Predictors and Regressors, 55	
3.4 Ordinary Least Squares, 58	
3.4.1 Data and Matrix Notation, 60	
3.4.2 The Errors e , 61	
3.4.3 Ordinary Least Squares Estimators, 61	
3.4.4 Properties of the Estimates, 63	
3.4.5 Simple Regression in Matrix Notation, 63	
3.4.6 The Coefficient of Determination, 66	
3.4.7 Hypotheses Concerning One Coefficient, 67	
3.4.8 t -Tests and Added-Variable Plots, 68	
3.5 Predictions, Fitted Values, and Linear Combinations, 68	
3.6 Problems, 69	
4 Interpretation of Main Effects	73
4.1 Understanding Parameter Estimates, 73	
4.1.1 Rate of Change, 74	
4.1.2 Signs of Estimates, 75	
4.1.3 Interpretation Depends on Other Terms in the Mean Function, 75	
4.1.4 Rank Deficient and Overparameterized Mean Functions, 78	
4.1.5 Collinearity, 79	
4.1.6 Regressors in Logarithmic Scale, 81	
4.1.7 Response in Logarithmic Scale, 82	
4.2 Dropping Regressors, 84	
4.2.1 Parameters, 84	
4.2.2 Variances, 86	
4.3 Experimentation versus Observation, 86	
4.3.1 Feedlots, 87	
4.4 Sampling from a Normal Population, 89	

4.5	More on R^2 , 91	
4.5.1	Simple Linear Regression and R^2 , 91	
4.5.2	Multiple Linear Regression and R^2 , 92	
4.5.3	Regression through the Origin, 93	
4.6	Problems, 93	
5	Complex Regressors	98
5.1	Factors, 98	
5.1.1	One-Factor Models, 99	
5.1.2	Comparison of Level Means, 102	
5.1.3	Adding a Continuous Predictor, 103	
5.1.4	The Main Effects Model, 106	
5.2	Many Factors, 108	
5.3	Polynomial Regression, 109	
5.3.1	Polynomials with Several Predictors, 111	
5.3.2	Numerical Issues with Polynomials, 112	
5.4	Splines, 113	
5.4.1	Choosing a Spline Basis, 115	
5.4.2	Coefficient Estimates, 116	
5.5	Principal Components, 116	
5.5.1	Using Principal Components, 118	
5.5.2	Scaling, 119	
5.6	Missing Data, 119	
5.6.1	Missing at Random, 120	
5.6.2	Imputation, 122	
5.7	Problems, 123	
6	Testing and Analysis of Variance	133
6.1	<i>F</i> -Tests, 134	
6.1.1	General Likelihood Ratio Tests, 138	
6.2	The Analysis of Variance, 138	
6.3	Comparisons of Means, 142	
6.4	Power and Non-Null Distributions, 143	
6.5	Wald Tests, 145	
6.5.1	One Coefficient, 145	
6.5.2	One Linear Combination, 146	
6.5.3	General Linear Hypothesis, 146	
6.5.4	Equivalence of Wald and Likelihood-Ratio Tests, 146	

6.6	Interpreting Tests, 146	
6.6.1	Interpreting <i>p</i> -Values, 146	
6.6.2	Why Most Published Research Findings Are False, 147	
6.6.3	Look at the Data, Not Just the Tests, 148	
6.6.4	Population versus Sample, 149	
6.6.5	Stacking the Deck, 149	
6.6.6	Multiple Testing, 150	
6.6.7	File Drawer Effects, 150	
6.6.8	The Lab Is Not the Real World, 150	
6.7	Problems, 150	
7	Variances	156
7.1	Weighted Least Squares, 156	
7.1.1	Weighting of Group Means, 159	
7.1.2	Sample Surveys, 161	
7.2	Misspecified Variances, 162	
7.2.1	Accommodating Misspecified Variance, 163	
7.2.2	A Test for Constant Variance, 164	
7.3	General Correlation Structures, 168	
7.4	Mixed Models, 169	
7.5	Variance Stabilizing Transformations, 171	
7.6	The Delta Method, 172	
7.7	The Bootstrap, 174	
7.7.1	Regression Inference without Normality, 175	
7.7.2	Nonlinear Functions of Parameters, 178	
7.7.3	Residual Bootstrap, 179	
7.7.4	Bootstrap Tests, 179	
7.8	Problems, 179	
8	Transformations	185
8.1	Transformation Basics, 185	
8.1.1	Power Transformations, 186	
8.1.2	Transforming One Predictor Variable, 188	
8.1.3	The Box–Cox Method, 190	
8.2	A General Approach to Transformations, 191	
8.2.1	The 1D Estimation Result and Linearly Related Regressors, 194	
8.2.2	Automatic Choice of Transformation of Predictors, 195	

8.3	Transforming the Response,	196
8.4	Transformations of Nonpositive Variables,	198
8.5	Additive Models,	199
8.6	Problems,	199
9	Regression Diagnostics	204
9.1	The Residuals,	204
9.1.1	Difference between \hat{e} and e ,	205
9.1.2	The Hat Matrix,	206
9.1.3	Residuals and the Hat Matrix with Weights,	208
9.1.4	Residual Plots When the Model Is Correct,	209
9.1.5	The Residuals When the Model Is Not Correct,	209
9.1.6	Fuel Consumption Data,	211
9.2	Testing for Curvature,	212
9.3	Nonconstant Variance,	213
9.4	Outliers,	214
9.4.1	An Outlier Test,	215
9.4.2	Weighted Least Squares,	216
9.4.3	Significance Levels for the Outlier Test,	217
9.4.4	Additional Comments,	218
9.5	Influence of Cases,	218
9.5.1	Cook's Distance,	220
9.5.2	Magnitude of D_i ,	221
9.5.3	Computing D_i ,	221
9.5.4	Other Measures of Influence,	224
9.6	Normality Assumption,	225
9.7	Problems,	226
10	Variable Selection	234
10.1	Variable Selection and Parameter Assessment,	235
10.2	Variable Selection for Discovery,	237
10.2.1	Information Criteria,	238
10.2.2	Stepwise Regression,	239
10.2.3	Regularized Methods,	244
10.2.4	Subset Selection Overstates Significance,	245
10.3	Model Selection for Prediction,	245
10.3.1	Cross-Validation,	247
10.3.2	Professor Ratings,	247
10.4	Problems,	248

11 Nonlinear Regression	252
11.1 Estimation for Nonlinear Mean Functions,	253
11.2 Inference Assuming Large Samples,	256
11.3 Starting Values,	257
11.4 Bootstrap Inference,	262
11.5 Further Reading,	265
11.6 Problems,	265
12 Binomial and Poisson Regression	270
12.1 Distributions for Counted Data,	270
12.1.1 Bernoulli Distribution,	270
12.1.2 Binomial Distribution,	271
12.1.3 Poisson Distribution,	271
12.2 Regression Models for Counts,	272
12.2.1 Binomial Regression,	272
12.2.2 Deviance,	277
12.3 Poisson Regression,	279
12.3.1 Goodness of Fit Tests,	282
12.4 Transferring What You Know about Linear Models,	283
12.4.1 Scatterplots and Regression,	283
12.4.2 Simple and Multiple Regression,	283
12.4.3 Model Building,	284
12.4.4 Testing and Analysis of Deviance,	284
12.4.5 Variances,	284
12.4.6 Transformations,	284
12.4.7 Regression Diagnostics,	284
12.4.8 Variable Selection,	285
12.5 Generalized Linear Models,	285
12.6 Problems,	285
Appendix	290
A.1 Website,	290
A.2 Means, Variances, Covariances, and Correlations,	290
A.2.1 The Population Mean and E Notation,	290
A.2.2 Variance and Var Notation,	291
A.2.3 Covariance and Correlation,	291
A.2.4 Conditional Moments,	292
A.3 Least Squares for Simple Regression,	293

A.4	Means and Variances of Least Squares Estimates,	294
A.5	Estimating $E(Y X)$ Using a Smoother,	296
A.6	A Brief Introduction to Matrices and Vectors,	298
A.6.1	Addition and Subtraction,	299
A.6.2	Multiplication by a Scalar,	299
A.6.3	Matrix Multiplication,	299
A.6.4	Transpose of a Matrix,	300
A.6.5	Inverse of a Matrix,	301
A.6.6	Orthogonality,	302
A.6.7	Linear Dependence and Rank of a Matrix,	303
A.7	Random Vectors,	303
A.8	Least Squares Using Matrices,	304
A.8.1	Properties of Estimates,	305
A.8.2	The Residual Sum of Squares,	305
A.8.3	Estimate of Variance,	306
A.8.4	Weighted Least Squares,	306
A.9	The QR Factorization,	307
A.10	Spectral Decomposition,	309
A.11	Maximum Likelihood Estimates,	309
A.11.1	Linear Models,	309
A.11.2	Logistic Regression,	311
A.12	The Box–Cox Method for Transformations,	312
A.12.1	Univariate Case,	312
A.12.2	Multivariate Case,	313
A.13	Case Deletion in Linear Regression,	314
References		317
Author Index		329
Subject Index		331

Preface to the Fourth Edition

This is a *textbook* to help you learn about applied linear regression. The book has been in print for more than 30 years, in a period of rapid change in statistical methodology and particularly in statistical computing. This fourth edition is a thorough rewriting of the book to reflect the needs of current students. As in previous editions, the overriding theme of the book is to help you learn to do data analysis using linear regression. Linear regression is a excellent model for learning about data analysis, both because it is important on its own and it provides a framework for understanding other methods of analysis.

This edition of the book includes the majority of the topics in previous editions, although much of the material has been rearranged. New methodology and examples have been added throughout.

- Even more emphasis is placed on graphics. The first two editions stressed graphics for diagnostic methods (Chapter 9) and the third edition added graphics for understanding data before any analysis is done (Chapter 1). In this edition, *effects plots* are stressed to summarize the fit of a model.
- Many applied analyses are based on understanding and interpreting parameters. This edition puts much greater emphasis on parameters, with part of Chapters 2–3 and all of Chapters 4–5 devoted to this important topic.
- Chapter 6 contains a greatly expanded treatment of testing and model comparison using both likelihood ratio and Wald tests. The usefulness and limitations of testing are stressed.
- Chapter 7 is about the variance assumption in linear models. The discussion of weighted least squares has been expanded to cover problems of ecological regressions, sample surveys, and other cases. Alternatives such as the bootstrap and heteroskedasticity corrections have been added or expanded.
- Diagnostic methods using transformations (Chapter 8) and residuals and related quantities (Chapter 9) that were the heart of the earlier editions have been maintained in this new edition.

- The discussion of variable selection in Chapter 10 has been updated from the third edition. It is designed to help you understand the key problems in variable selection. In recent years, this topic has morphed into the area of *machine learning* and the goal of this chapter is to show connections and provide references.
- As in the third edition, brief introductions to nonlinear regression (Chapter 11) and to logistic regression (Chapter 12) are included, with Poisson regression added in Chapter 12.

Using This Book

The website for this book is <http://z.umn.edu/alr4ed>.

As with previous editions, this book is not tied to any particular computer program. A primer for using the free R package (R Core Team, 2013) for the material covered in the book is available from the website. The primer can also be accessed directly from within R as you are working. An optional published companion book about R is Fox and Weisberg (2011).

All the data files used are available from the website and in an R package called `alr4` that you can download for free. Solutions for odd-numbered problems, all using R, are available on the website for the book¹. You cannot learn to do data analysis without working problems.

Some advanced topics are introduced to help you recognize when a problem that looks like linear regression is actually a little different. Detailed methodology is not always presented, but references at the same level as this book are presented. The bibliography, also available with clickable links on the book's website, has been greatly expanded and updated.

Mathematical Level

The mathematical level of this book is roughly the same as the level of previous editions. Matrix representation of data is used, particularly in the derivation of the methodology in Chapters 3–4. Derivations are less frequent in later chapters, and so the necessary mathematics is less. Calculus is generally not required, except for an occasional use of a derivative. The discussions requiring calculus can be skipped without much loss.

ACKNOWLEDGMENTS

Thanks are due to Jeff Witmer, Yuhong Yang, Brad Price, and Brad's Stat 5302 students at the University of Minnesota. New examples were provided by April Bleske-Rechek, Tom Burk, and Steve Taff. Work with John Fox over the last few years has greatly influenced my writing.

For help with previous editions, thanks are due to Charles Anderson, Don Pereira, Christopher Bingham, Morton Brown, Cathy Campbell, Dennis Cook,

¹All solutions are available to instructors using the book in a course; see the website for details.

Stephen Fienberg, James Frane, Seymour Geisser, John Hartigan, David Hinkley, Alan Izenman, Soren Johansen, Kenneth Koehler, David Lane, Michael Lavine, Kinley Larntz, Gary Oehlert, Katherine St. Clair, Keija Shan, John Rice, Donald Rubin, Joe Shih, Pete Stewart, Stephen Stigler, Douglas Tiffany, Carol Weisberg, and Howard Weisberg.

Finally, I am grateful to Stephen Quigley at Wiley for asking me to do a new edition. I have been working on versions of this book since 1976, and each new edition has pleased me more than the one before it. I hope it pleases you, too.

SANFORD WEISBERG

St. Paul, Minnesota
September 2013

C H A P T E R 1

Scatterplots and Regression

Regression is the study of dependence. It is used to answer interesting questions about how one or more predictors influence a response. Here are a few typical questions that may be answered using regression:

- Are daughters taller than their mothers?
- Does changing class size affect success of students?
- Can we predict the time of the next eruption of Old Faithful Geyser from the length of the most recent eruption?
- Do changes in diet result in changes in cholesterol level, and if so, do the results depend on other characteristics such as age, sex, and amount of exercise?
- Do countries with higher per person income have lower birth rates than countries with lower income?
- Are highway design characteristics associated with highway accident rates? Can accident rates be lowered by changing design characteristics?
- Is water usage increasing over time?
- Do conservation easements on agricultural property lower land value?

In most of this book, we study the important instance of regression methodology called *linear regression*. This method is the most commonly used in regression, and virtually all other regression methods build upon an understanding of how linear regression works.

As with most statistical analyses, the goal of regression is to summarize observed data as simply, usefully, and elegantly as possible. A theory may be available in some problems that specifies how the response varies as the values

of the predictors change. If theory is lacking, we may need to use the data to help us decide on how to proceed. In either case, an essential first step in regression analysis is to draw appropriate graphs of the data.

We begin in this chapter with the fundamental graphical tools for studying dependence. In regression problems with one predictor and one response, the *scatterplot* of the response versus the predictor is the starting point for regression analysis. In problems with many predictors, several simple graphs will be required at the beginning of an analysis. A *scatterplot matrix* is a convenient way to organize looking at many scatterplots at once. We will look at several examples to introduce the main tools for looking at scatterplots and scatterplot matrices and extracting information from them. We will also introduce notation that will be used throughout the book.

1.1 SCATTERPLOTS

We begin with a regression problem with one predictor, which we will generically call X , and one response variable, which we will call Y .¹ Data consist of values (x_i, y_i) , $i = 1, \dots, n$, of (X, Y) observed on each of n units or *cases*. In any particular problem, both X and Y will have other names that will be displayed in this book using typewriter font, such as temperature or concentration, that are more descriptive of the data that are to be analyzed. The goal of regression is to understand how the values of Y change as X is varied over its range of possible values. A first look at how Y changes as X is varied is available from a scatterplot.

Inheritance of Height

One of the first uses of regression was to study inheritance of traits from generation to generation. During the period 1893–1898, Karl Pearson (1857–1936) organized the collection of $n = 1375$ heights of mothers in the United Kingdom under the age of 65 and one of their adult daughters over the age of 18. Pearson and Lee (1903) published the data, and we shall use these data to examine inheritance. The data are given in the data file `Heights`.²

Our interest is in inheritance *from* the mother *to* the daughter, so we view the mother's height, called `mheight`, as the predictor variable and the daughter's height, `dheight`, as the response variable. Do taller mothers tend to have taller daughters? Do shorter mothers tend to have shorter daughters?

A scatterplot of `dheight` versus `mheight` helps us answer these questions. The scatterplot is a graph of each of the n points with the response `dheight` on the vertical axis and predictor `mheight` on the horizontal axis. This plot is

¹In some disciplines, predictors are called independent variables, and the response is called a dependent variable, terms not used in this book.

²See Appendix A.1 for instructions for getting data files from the Internet.

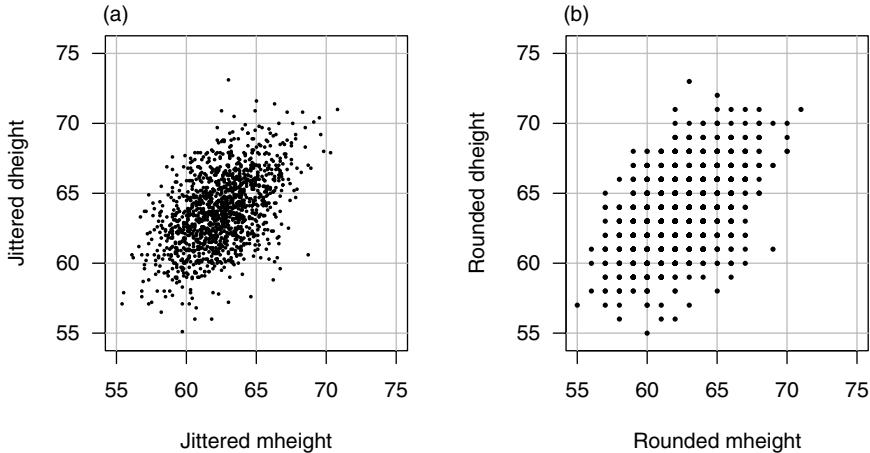


Figure 1.1 Scatterplot of mothers' and daughters' heights in the Pearson and Lee data. The original data have been jittered to avoid overplotting in (a). Plot (b) shows the original data, so each point in the plot refers to one or more mother–daughter pairs.

shown in Figure 1.1a. For regression problems with one predictor X and a response Y , we call the scatterplot of Y versus X a *summary graph*.

Here are some important characteristics of this scatterplot:

1. The range of heights appears to be about the same for mothers and for daughters. Because of this, we draw the plot so that the lengths of the horizontal and vertical axes are the same, and the scales are the same. If all mothers and daughters pairs had *exactly* the same height, then all the points would fall exactly on a 45° -line. Some computer programs for drawing a scatterplot are not smart enough to figure out that the lengths of the axes should be the same, so you might need to resize the plot or to draw it several times.
2. The original data that went into this scatterplot were rounded so each of the heights was given to the nearest inch. The original data are plotted in Figure 1.1b. This plot exhibits substantial *overplotting* with many points at exactly the same location. This is undesirable because one point on the plot can correspond to many cases. The easiest solution is to use *jittering*, in which a small uniform random number is added to each value. In Figure 1.1a, we used a uniform random number on the range from -0.5 to $+0.5$, so the jittered values would round to the numbers given in the original source.
3. One important function of the scatterplot is to decide if we might reasonably assume that the response on the vertical axis is independent of the predictor on the horizontal axis. This is clearly not the case here since as we move across Figure 1.1a from left to right, the scatter of points is

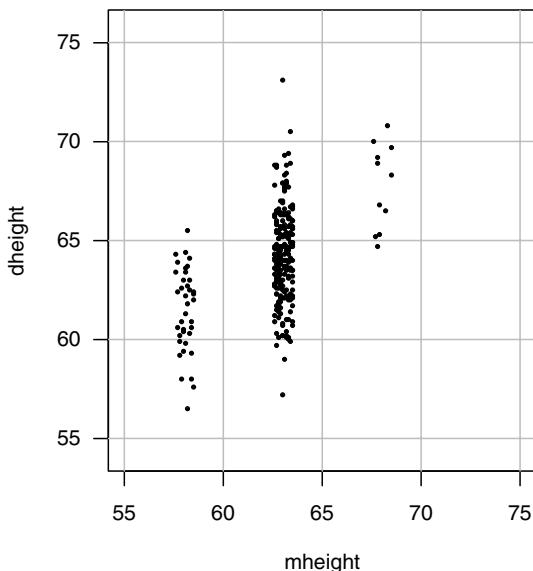


Figure 1.2 Scatterplot showing only pairs with mother's height that rounds to 58, 64, or 68 inches.

different for each value of the predictor. What we mean by this is shown in Figure 1.2, in which we show only points corresponding to mother-daughter pairs with $mheight$ rounding to either 58, 64, or 68 inches. We see that within each of these three strips or *slices*, the number of points is different, and the mean of $dheight$ is increasing from left to right. The vertical variability in $dheight$ seems to be more or less the same for each of the fixed values of $mheight$.

4. In Figure 1.1a the scatter of points appears to be more or less elliptically shaped, with the major axis of the ellipse tilted upward, and with more points near the center of the ellipse rather than on the edges. We will see in Section 1.4 that summary graphs that look like this one suggest the use of the simple linear regression model that will be discussed in Chapter 2.
5. Scatterplots are also important for finding *separated points*. Horizontal separation would occur for a value on the horizontal axis $mheight$ that is either unusually small or unusually large relative to the other values of $mheight$. Vertical separation would occur for a daughter with $dheight$ either relatively large or small compared with the other daughters with about the same value for $mheight$.

These two types of separated points have different names and roles in a regression problem. Extreme values on the left and right of the horizontal axis are points that are likely to be important in fitting regression models and are called *leverage* points. The separated points on the vertical axis, here unusually tall or short daughters give their mother's height, are potentially *outliers*, cases that are somehow different from

the others in the data. Outliers are more easily discovered in residual plots, as illustrated in the next example.

While the data in Figure 1.1a do include a few tall and a few short mothers and a few tall and short daughters, given the height of the mothers, none appears worthy of special treatment, mostly because in a sample size this large, we expect to see some fairly unusual mother–daughter pairs.

Forbes's Data

In an 1857 article, the Scottish physicist James D. Forbes (1809–1868) discussed a series of experiments that he had done concerning the relationship between atmospheric pressure and the boiling point of water. He knew that altitude could be determined from atmospheric pressure, measured with a barometer, with lower pressures corresponding to higher altitudes. Barometers in the middle of the nineteenth century were fragile instruments, and Forbes wondered if a simpler measurement of the boiling point of water could substitute for a direct reading of barometric pressure. Forbes collected data in the Alps and in Scotland. He measured at each location the atmospheric pressure `pres` in inches of mercury with a barometer and boiling point `bp` in degrees Fahrenheit using a thermometer. Boiling point measurements were adjusted for the difference between the ambient air temperature when he took the measurements and a standard temperature. The data for $n = 17$ locales are reproduced in the file `Forbes`.

The scatterplot of `pres` versus `bp` is shown in Figure 1.3a. The general appearance of this plot is very different from the summary graph for the heights data. First, the sample size is only 17, as compared with over 1,300 for the heights data. Second, apart from one point, all the points fall almost exactly on a smooth curve. This means that the variability in pressure for a given boiling point is extremely small.

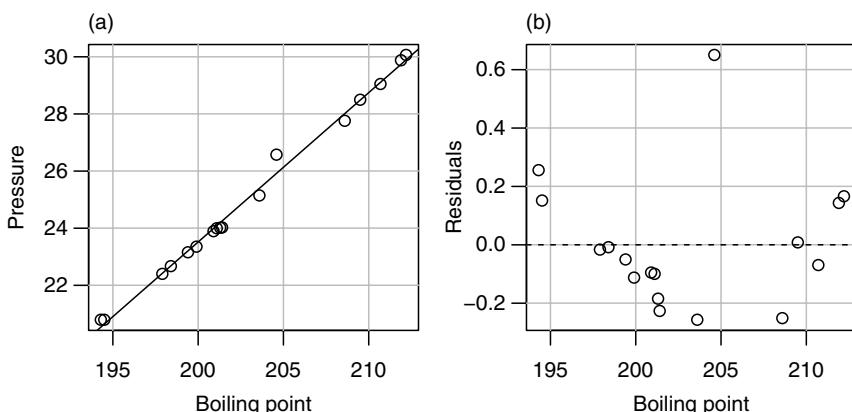


Figure 1.3 Forbes data: (a) `pres` versus `bp`; (b) residuals versus `bp`.

The points in Figure 1.3a appear to fall very close to the straight line shown on the plot, and so we might be encouraged to think that the mean of pressure given boiling point could be modeled by a straight line. Look closely at the graph, and you will see that there is a small systematic deviation from the straight line: apart from the one point that does not fit at all, the points in the middle of the graph fall below the line, and those at the highest and lowest boiling points fall above the line. This is much easier to see in Figure 1.3b, which is obtained by removing the linear trend from Figure 1.3a, so the plotted points on the vertical axis are given for each value of bp by

$$\text{residual} = \text{pres} - \text{point on the line}$$

This allows us to gain resolution in the plot since the range on the vertical axis in Figure 1.3a is about 10 inches of mercury while the range in Figure 1.3b is about 0.8 inches of mercury. To get the same resolution in Figure 1.3a, we would need a graph that is $10/0.8 = 12.5$ as big as Figure 1.3b. Again ignoring the one point that clearly does not match the others, the curvature in the plot is clearly visible in Figure 1.3b.

While there is nothing at all wrong with curvature, the methods we will be studying in this book work best when the plot can be summarized by a straight line. Sometimes we can get a straight line by transforming one or both of the plotted quantities. Forbes had a physical theory that suggested that $\log(\text{pres})$ is linearly related to bp . Forbes (1857) contains what may be the first published summary graph based on his physical model. His figure is redrawn in Figure 1.4. Following Forbes, we use base-ten common logs in this example, although in most of the examples in this book we will use natural logarithms. The choice

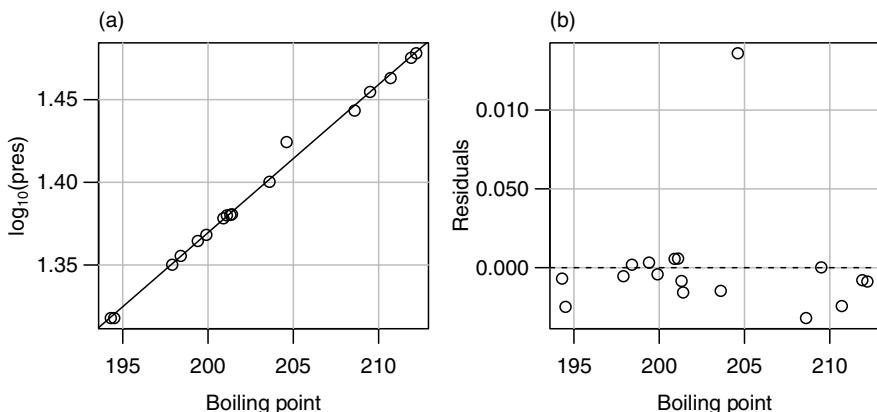


Figure 1.4 (a) Scatterplot of Forbes's data. The line shown is the ols line for the regression of $\log(\text{pres})$ on bp . (b) Residuals versus bp .

of base has no material effect on the appearance of the graph or on fitted regression models, but interpretation of parameters can depend on the choice of base.

The key feature of Figure 1.4a is that apart from one point, the data appear to fall very close to the straight line shown on the figure, and the residual plot in Figure 1.4b confirms that the deviations from the straight line are not systematic the way they were in Figure 1.3b. All this is evidence that the straight line is a reasonable summary of these data.

Length at Age for Smallmouth Bass

The smallmouth bass is a favorite game fish in inland lakes. Many smallmouth bass populations are managed through stocking, fishing regulations, and other means, with a goal to maintain a healthy population.

One tool in the study of fish populations is to understand the growth pattern of fish such as the dependence of a measure of size like fish length on age of the fish. Managers could compare these relationships between different populations that are managed differently to learn how management impacts fish growth.

Figure 1.5 displays the Length at capture in mm versus Age at capture for $n = 439$ smallmouth bass measured in West Bearskin Lake in Northeastern Minnesota in 1991. Only fish of age 8 or less are included in this graph. The data were provided by the Minnesota Department of Natural Resources and are given in the file `wblake`. Similar to trees, the scales of many fish species have annular rings, and these can be counted to determine the age of a fish.

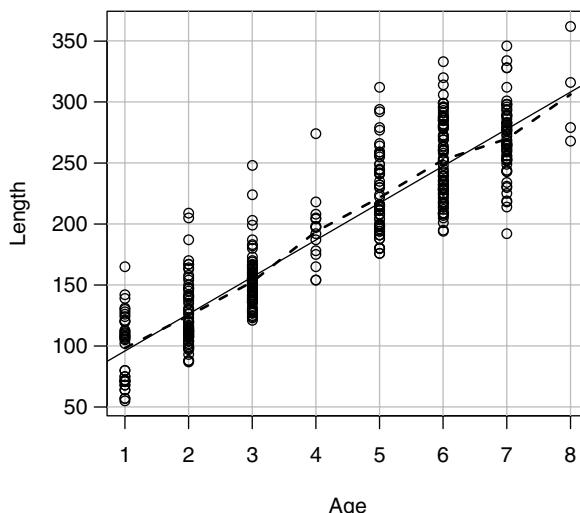


Figure 1.5 Length (mm) versus Age for West Bearskin Lake smallmouth bass. The solid line shown was estimated using ordinary least squares or OLS. The dashed line joins the average observed length at each age.

These data are *cross-sectional*, meaning that all the observations were taken at the same time. In a *longitudinal* study, the same fish would be measured each year, possibly requiring many years of taking measurements.

The appearance of this graph is different from the summary graphs shown for the last two examples. The predictor `Age` can only take on integer values corresponding to the number of annular rings on the scale, so we are really plotting eight distinct populations of fish. As might be expected, length generally increases with age, but the length of the longest fish at age 1 exceeds the length of the shortest fish at age 4, so knowing the age of a fish will not allow us to predict its length exactly; see Problem 2.15.

Predicting the Weather

Can early season snowfall from September 1 until December 31 predict snowfall in the remainder of the year, from January 1 to June 30? Figure 1.6, using data from the data file `ftcollinssnow`, gives a plot of Late season snowfall from January 1 to June 30 versus Early season snowfall for the period September 1 to December 31 of the previous year, both measured in inches at Ft. Collins, Colorado (Colorado Climate Center, 2012). If `Late` is related to `Early`, the relationship is considerably weaker than in the previous examples, and the graph suggests that early winter snowfall and late winter snowfall may be completely unrelated or *uncorrelated*. Interest in this regression problem will therefore be in testing the hypothesis that the two variables are uncorrelated versus the alternative that they are not uncorrelated, essentially

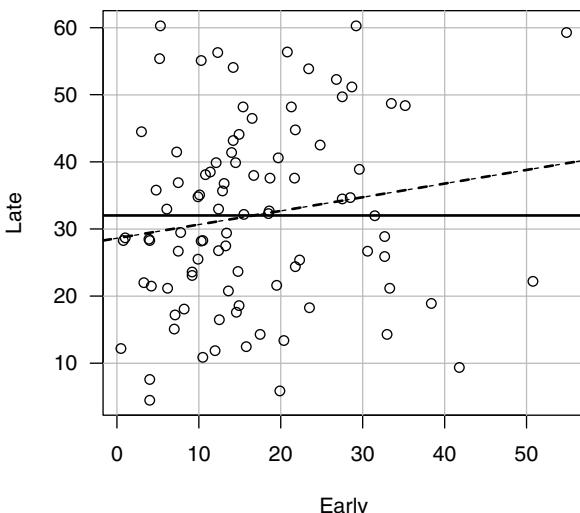


Figure 1.6 Plot of snowfall for 93 years from 1900 to 1992 in inches. The solid horizontal line is drawn at the average late season snowfall. The dashed line is the OLS line.

comparing the fit of the two lines shown in Figure 1.6. Fitting models and performing tests will be helpful here.

Turkey Growth

This example is from an experiment on the growth of turkeys (Noll et al., 1984). Pens of turkeys were grown with an identical diet, except that each pen was supplemented with a `Dose` of the amino acid methionine as a percentage of the total diet of the birds. The methionine was provided using either a standard source or one of two experimental sources. The response is average weight gain in grams of all the turkeys in the pen.

Figure 1.7 provides a summary graph based on the data in the file `turkey`. Except at `Dose` = 0, each point in the graph is the average response of five pens of turkeys; at `Dose` = 0, there were 10 pens of turkeys. Because averages are plotted, the graph does not display the variation between pens treated alike. At each value of `Dose` > 0, there are three points shown, with different symbols corresponding to the three sources of methionine, so the variation between points at a given `Dose` is really the variation between sources. At `Dose` = 0, the point has been arbitrarily labeled with the symbol for the first group, since `Dose` = 0 is the same treatment for all sources.

For now, ignore the three sources and examine Figure 1.7 in the way we have been examining the other summary graphs in this chapter. Weight gain is seen to increase with increasing `Dose`, but the increase does not appear to be linear, meaning that a straight line does not seem to be a reasonable

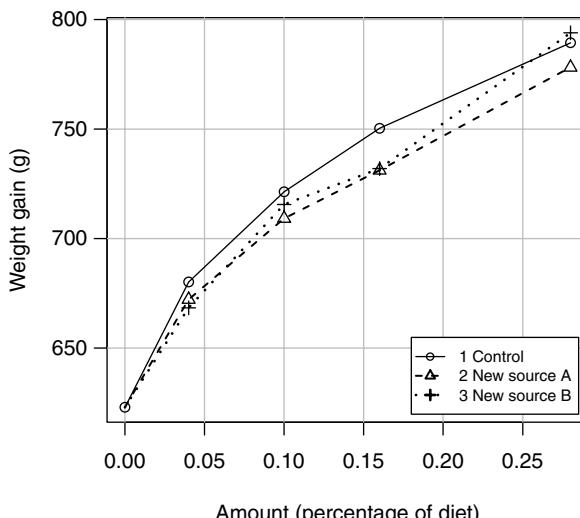


Figure 1.7 Weight gain versus `Dose` of methionine for turkeys. The three symbols for the points refer to sources of methionine. The lines on the plot join the means within a source.

representation of the average dependence of the response on the predictor. This leads to study of mean functions.

1.2 MEAN FUNCTIONS

Imagine a generic summary plot of Y versus X . Our interest centers on how the distribution of Y changes as X is varied. One important aspect of this distribution is the *mean function*, which we define by

$$E(Y|X = x) = \text{a function that depends on the value of } x \quad (1.1)$$

We read the left side of this equation as “the expected value of the response when the predictor is fixed at the value $X = x$ ”; if the notation “ $E()$ ” for expectations and “ $\text{Var}()$ ” for variances is unfamiliar, refer to Appendix A.2. The right side of (1.1) depends on the problem. For example, in the heights data in Example 1.1, we might believe that

$$E(\text{dheight}|\text{mheight} = x) = \beta_0 + \beta_1 x \quad (1.2)$$

that is, the mean function is a straight line. This particular mean function has two *parameters*, an intercept β_0 and a slope β_1 . If we knew the values of the β s, then the mean function would be completely specified, but usually the β s need to be estimated from data. These parameters are discussed more fully in the next chapter.

Figure 1.8 shows two possibilities for the β s in the straight-line mean function (1.2) for the heights data. For the dashed line, $\beta_0 = 0$ and $\beta_1 = 1$. This mean function would suggest that daughters have the same height as their mothers on the average for mothers of any height. The second line is estimated using ordinary least squares, or OLS, the estimation method that will be described in the next chapter. The OLS line has slope less than 1, meaning that tall mothers tend to have daughters who are taller than average because the slope is positive, but shorter than themselves because the slope is less than 1. Similarly, short mothers tend to have short daughters but taller than themselves. This is perhaps a surprising result and is the origin of the term *regression*, since extreme values in one generation tend to revert or regress toward the population mean in the next generation (Galton, 1886).

Two lines are shown in Figure 1.5 for the smallmouth bass data. The dashed line joins the average length at each age. It provides an estimate of the mean function $E(\text{Length}|\text{Age})$ without actually specifying any functional form for the mean function. We will call this a *nonparametric* estimated mean function; sometimes we will call it a *smoother*. The solid line is the OLS estimated straight line (1.1) for the mean function. Perhaps surprisingly, the straight line and the dashed lines that join the within-age means appear to agree very closely, and we might be encouraged to use the straight-line mean function to describe

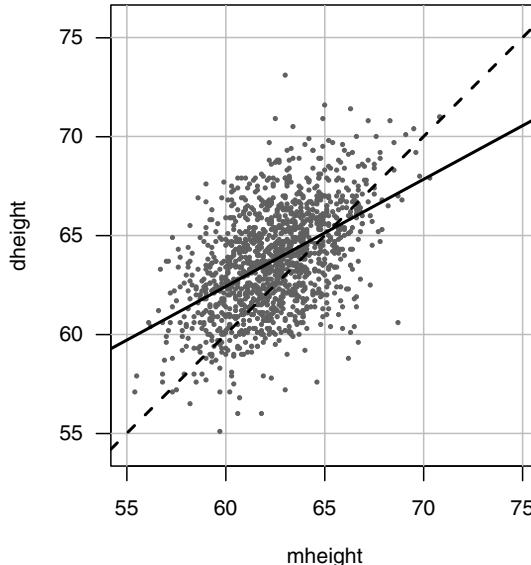


Figure 1.8 The heights data. The dashed line is for $E(dheight|mheight) = mheight$, and the solid line is estimated by ols.

these data. The increase in length per year is modeled to be the same for all ages. We cannot expect this to be true if we were to include older-aged fish because eventually the growth must slow down. For the range of ages here, the approximation seems to be adequate.

For the Ft. Collins weather data, we might expect the straight-line mean function (1.1) to be appropriate but with $\beta_1 = 0$. If the slope is 0, then the mean function is parallel to the horizontal axis, as shown in Figure 1.6. We will eventually test for independence of Early and Late by testing the hypothesis that $\beta_1 = 0$ against the alternative hypothesis that $\beta_1 \neq 0$.

Not all summary graphs will have a straight-line mean function. In Forbes's data, to achieve linearity we have replaced the measured value of *pres* by $\log(\text{pres})$. Transformation of variables will be a key tool in extending the usefulness of linear regression models. In the turkey data and other growth models, a nonlinear mean function might be more appropriate, such as

$$E(Y|\text{Dose} = x) = \beta_0 + \beta_1[1 - \exp(-\beta_2 x)] \quad (1.3)$$

The β s in (1.3) have a useful interpretation, and they can be used to summarize the experiment. When $\text{Dose} = 0$, $E(Y|\text{Dose} = 0) = \beta_0$, so β_0 is the baseline growth without supplementation. Assuming $\beta_2 > 0$, when the *Dose* is large, $\exp(-\beta_2 \text{Dose})$ is small, and so $E(Y|\text{Dose})$ approaches $\beta_0 + \beta_1$ for larger values of *Dose*. We think of $\beta_0 + \beta_1$ as the limit to growth with this additive. The rate

parameter β_2 determines how quickly maximum growth is achieved. This three-parameter mean function will be considered in Chapter 11.

1.3 VARIANCE FUNCTIONS

Another characteristic of the distribution of the response given the predictor is the *variance function*, defined by the symbol $\text{Var}(Y|X = x)$ and in words as the variance of the response given that the predictor is fixed at $X = x$. For example, in Figure 1.2 we can see that the variance function for dheightlmheight is approximately the same for each of the three values of mheight shown in the graph. In the smallmouth bass data in Figure 1.5, an assumption that the variance is constant across the plot is plausible, even if it is not certain (see Problem 1.2). In the turkey data, we cannot say much about the variance function from the summary plot because we have plotted treatment means rather than the actual pen values, so the graph does not display the information about the variability between pens that have a fixed value of Dose.

A frequent assumption in fitting linear regression models is that the variance function is the same for every value of x . This is usually written as

$$\text{Var}(Y|X = x) = \sigma^2 \quad (1.4)$$

where σ^2 (read “sigma squared”) is a generally unknown positive constant. Chapter 7 presents more general variance models.

1.4 SUMMARY GRAPH

In all the examples except the snowfall data, there is a clear dependence of the response on the predictor. In the snowfall example, there might be no dependence at all. The turkey growth example is different from the others because the average value of the response seems to change nonlinearly with the value of the predictor on the horizontal axis.

The scatterplots for these examples are all typical of graphs one might see in problems with one response and one predictor. Examination of the summary graph is a first step in exploring the relationships these graphs portray.

Anscombe (1973) provided the artificial data given in the file anscombe that consists of 11 pairs of points (x_i, y_i) , to which the simple linear regression mean function $E(y|x) = \beta_0 + \beta_1 x$ is fit. Each data set leads to an identical summary analysis with the same estimated slope, intercept, and other summary statistics, but the visual impression of each of the graphs is very different. The first example in Figure 1.9a is as one might expect to observe if the simple linear regression model were appropriate. The graph of the second data set given in Figure 1.9b suggests that the analysis based on simple linear regres-

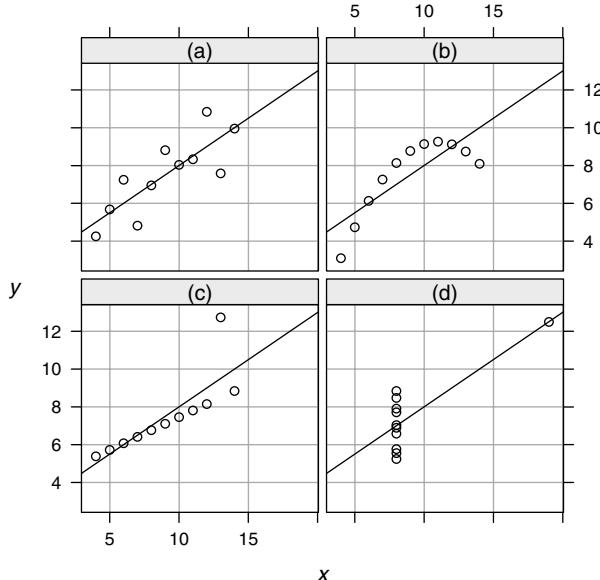


Figure 1.9 Four hypothetical data sets (Anscombe, 1973).

sion is incorrect and that a smooth curve, perhaps a quadratic polynomial, could be fit to the data with little remaining variability. Figure 1.9c suggests that the prescription of simple regression may be correct for most of the data, but one of the cases is too far away from the fitted regression line. This is called the *outlier problem*. Possibly the case that does not match the others should be deleted from the data set, and the regression should be refit from the remaining cases. This will lead to a different fitted line. Without a context for the data, we cannot judge one line “correct” and the other “incorrect.” The final set graphed in Figure 1.9d is different from the others in that there is not enough information to make a judgment concerning the mean function. If the separated point were deleted, we could not even estimate a slope. We must distrust an analysis that is so heavily dependent upon a single case.

1.5 TOOLS FOR LOOKING AT SCATTERPLOTS

Because looking at scatterplots is so important to fitting regression models, we establish some common vocabulary for describing the information in them and some tools to help us extract the information they contain.

The summary graph is of the response Y versus the predictor X . The mean function for the graph is defined by (1.1), and it characterizes how Y changes on the average as the value of X is varied. We may have a parametric model for the mean function and will use data to estimate the parameters. The

variance function also characterizes the graph, and in many problems we will assume at least at first that the variance function is constant. The scatterplot also will highlight separated points that may be of special interest because they do not fit the trend determined by the majority of the points.

A *null plot* has a horizontal straight line as its mean function, constant variance function, and no separated points. The scatterplot for the snowfall data appears to be a null plot.

1.5.1 Size

We may need to interact with a plot to extract all the available information, by changing scales, by resizing, or by removing linear trends. An example of this is given in Problem 1.3.

1.5.2 Transformations

In some problems, either or both of Y and X can be replaced by transformations so the summary graph has desirable properties. Most of the time, we will use power transformations, replacing, for example, X by X^λ for some number λ . Because logarithmic transformations are so frequently used, we will interpret $\lambda = 0$ as corresponding to a log transform.

1.5.3 Smoothers for the Mean Function

In the smallmouth bass data in Figure 1.5, we computed an estimate of $E(\text{Length}|\text{Age})$ using a simple nonparametric smoother obtained by averaging the repeated observations at each value of Age . Smoothers can also be defined when we do not have repeated observations at values of the predictor by averaging the observed data for all values of X *close to*, but not necessarily equal to, x . The literature on using smoothers to estimate mean functions has exploded in recent years, with fairly elementary treatments given by Bowman and Azzalini (1997), Green and Silverman (1994), Härdle (1990), and Simonoff (1996). Although these authors discuss nonparametric regression as an end in itself, we will generally use smoothers as *plot enhancements* to help us understand the information available in a scatterplot and to help calibrate the fit of a parametric mean function to a scatterplot.

For example, Figure 1.10 repeats Figure 1.1a, this time adding the estimated straight-line mean function and smoother called a *loess* smooth (Cleveland, 1979). Roughly speaking, the *loess* smooth estimates $E(Y|X=x)$ at the point x by fitting a straight line to a fraction of the points closest to x ; we used the fraction of 0.20 in this figure because the sample size is so large, but it is more usual to set the fraction to about 2/3. The smoother is obtained by joining the estimated values of $E(Y|X=x)$ for many values of x . The loess smoother and the straight line agree almost perfectly for $mheight$ close to average, but they

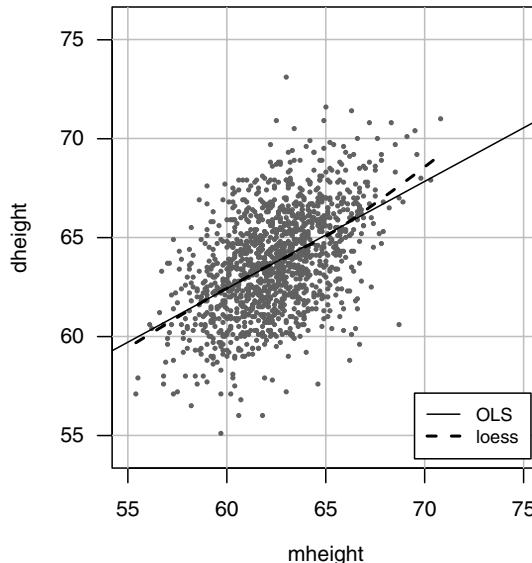


Figure 1.10 Heights data with the ols line and a loess smooth with span = 0.20.

agree less well for larger values of mheight where there is much less data. Smoothers tend to be less reliable at the edges of the plot. We briefly discuss the `loess` smoother in Appendix A.5, but this material is dependent on the results in Chapters 2 and 3.

1.6 SCATTERPLOT MATRICES

With one predictor, a scatterplot provides a summary of the regression relationship between the response and the predictor. With many predictors, we need to look at many scatterplots. A *scatterplot matrix* is a convenient way to organize these plots.

Fuel Consumption

The goal of this example is to understand how fuel consumption varies over the 50 United States and the District of Columbia (Federal Highway Administration, 2001). Table 1.1 describes the variables to be used in this example; the data are given in the file `fuel2001`. The data were collected by the U.S. Federal Highway Administration.

Both `Drivers` and `FuelC` are state totals, so these will be larger in states with more people and smaller in less populous states. `Income` is computed per person. To make all these comparable and to attempt to eliminate the effect of size of the state, we compute rates `Dlic` = `Drivers`/`Pop` and

Table 1.1 Variables in the Fuel Consumption Data^a

Drivers	Number of licensed drivers in the state
FuelC	Gasoline sold for road use, thousands of gallons
Income	Per person personal income for the year 2000, in thousands of dollars
Miles	Miles of Federal-aid highway miles in the state
Pop	2001 population age 16 and over
Tax	Gasoline state tax rate, cents per gallon
Fuel	$1000 \times \text{FuelC}/\text{Pop}$
Dlic	$1000 \times \text{Drivers}/\text{Pop}$
log(Miles)	Natural logarithm of Miles

^aAll data are for 2001, unless otherwise noted. The last three variables do not appear in the data file, but are computed from the previous variables, as described in the text.

$\text{Fuel} = \text{FuelC}/\text{Pop}$. Additionally, we replace Miles by its logarithm before doing any further analysis. Justification for replacing Miles with log(Miles) is deferred to Problem 8.7.

Many problems will require replacing the observable predictors like Drivers and Pop with a function of them like Dlic. We will use the term *predictor* to correspond to the original variables, and the new term *regressor*, described more fully in Section 3.3, to refer to variables that are computed from the predictors. In some instances this distinction is artificial, but in others the distinction can clarify issues.

The scatterplot matrix for the fuel data is shown in Figure 1.11. Except for the diagonal, a scatterplot matrix is a 2D array of scatterplots. The variable names on the diagonal label the axes. In Figure 1.11, the variable log(Miles) appears on the horizontal axis of all the plots in the rightmost column and on the vertical axis of all the plots in the bottom row.³

Each plot in a scatterplot matrix is relevant to a particular one predictor regression of the variable on the vertical axis, given the variable on the horizontal axis. For example, the plot of Fuel versus Tax in the top row and second column of the scatterplot matrix in Figure 1.11 is relevant for the regression of Fuel on Tax. We can interpret this plot as we would a scatterplot for simple regression. We get the overall impression that Fuel decreases on the average as Tax increases, but there is lot of variation. We can make similar qualitative judgments about the each of the regressions of Fuel on the other variables. The overall impression is that Fuel is at best weakly related to each of the variables in the scatterplot matrix.

Does this help us understand how Fuel is related to all four predictors simultaneously? The marginal relationships between the response and each of

³The scatterplot matrix program used to draw Figure 1.11, which is the pairs function in R, has the diagonal running from the top left to the lower right. Other programs, such as the splom function in R, has the diagonal from lower left to upper right. There seems to be no compelling reason to prefer one over the other.

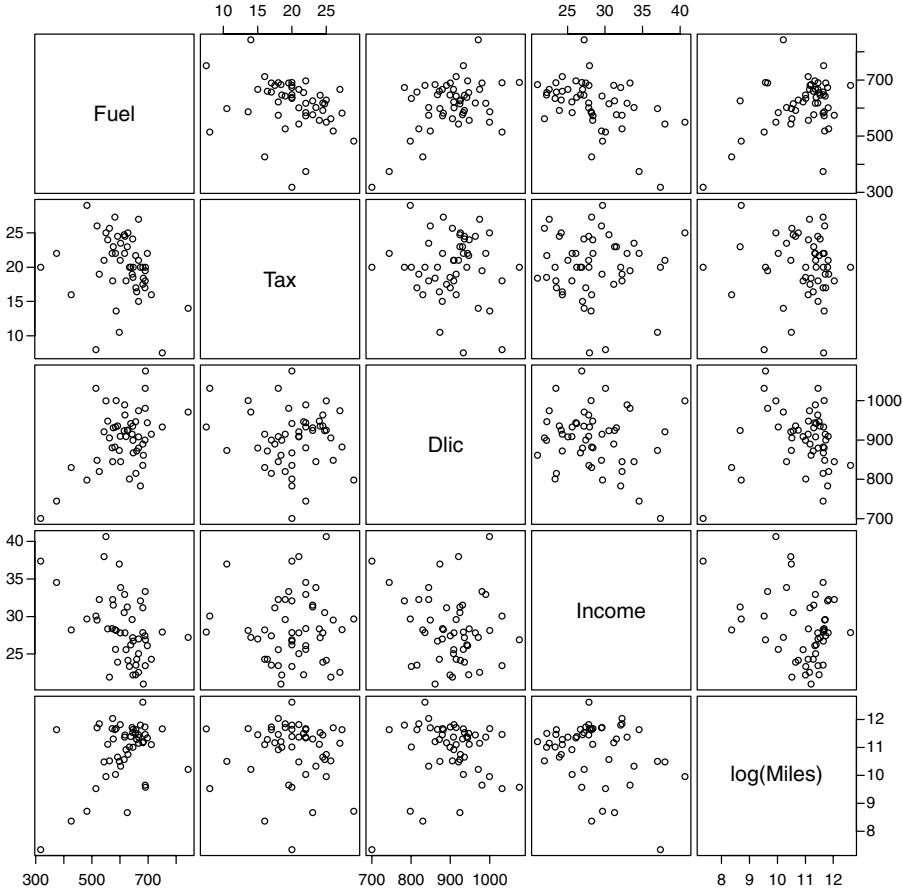


Figure 1.11 Scatterplot matrix for the fuel data.

the variables are *not* sufficient to understand the *joint* relationship between the response and the more than one predictor at a time. The interrelationships among the predictors are also important. The pairwise relationships between the predictors can be viewed in the remaining cells of the scatterplot matrix. In Figure 1.11, the relationships between all pairs of predictors appear to be very weak, suggesting that for this problem, the marginal plots including Fuel are quite informative about the multiple regression problem. General considerations for other scatterplot matrices will be developed in later chapters.

1.7 PROBLEMS

- 1.1 United Nations** (Data file: UN11) The data in the file UN11 contains several variables, including ppgdp, the gross national product per person

in U.S. dollars, and `fertility`, the birth rate per 1000 females, both from the year 2009. The data are for 199 localities, mostly UN member countries, but also other areas such as Hong Kong that are not independent countries. The data were collected from United Nations (2011). We will study the dependence of `fertility` on `ppgdp`.⁴

- 1.1.1** Identify the predictor and the response.
- 1.1.2** Draw the scatterplot of `fertility` on the vertical axis versus `ppgdp` on the horizontal axis and summarize the information in this graph. Does a straight-line mean function seem to be plausible for a summary of this graph?
- 1.1.3** Draw the scatterplot of `log(fertility)` versus `log(ppgdp)` using natural logarithms. Does the simple linear regression model seem plausible for a summary of this graph? If you use a different base of logarithms, the *shape* of the graph won't change, but the *values on the axes* will change.
- 1.2 Smallmouth bass data** (Data file: `wblake`) Compute the means and the variances for each of the eight subpopulations in the smallmouth bass data. Draw a graph of average length versus `Age` and compare with Figure 1.5. Draw a graph of the standard deviations versus age. If the variance function is constant, then the plot of standard deviation versus `Age` should be a null plot. Summarize the information.
- 1.3** (Data file: `Mitchell`) The data shown in Figure 1.12 give average soil temperature in degrees C at 20 cm depth in Mitchell, Nebraska for 17 years beginning January 1976, plotted versus the month number. The data were collected by K. Hubbard (Burnside et al., 1996).
- 1.3.1** Summarize the information in the graph about the dependence of soil temperature on month number.
- 1.3.2** The data used to draw Figure 1.12 are in the file `Mitchell`. Redraw the graph, but this time make the length of the horizontal axis at least 4 times the length of the vertical axis. Repeat Problem 1.3.1.

- 1.4 Old Faithful** (Data file: `oldfaith`) The data file gives information about eruptions of Old Faithful Geyser during October 1980. Variables are the `Duration` in seconds of the current eruption, and the `Interval`, the time in minutes to the next eruption. The data were collected by volunteers and were provided by the late Roderick Hutchinson. Apart from missing data for the period from midnight to 6 a.m., this is a complete record of eruptions for that month.

⁴In the third edition of this book, similar data from 2000 were used in this problem. Those data are still available in the R package that accompanies this book and is called `UN1`.

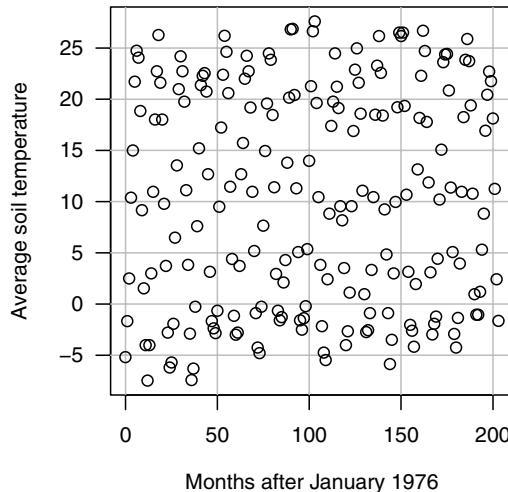


Figure 1.12 Monthly soil temperature data.

Old Faithful Geyser is an important tourist attraction, with up to several thousand people watching it erupt on pleasant summer days. The park service uses data like these to obtain a prediction equation for the time to the next eruption.

Draw the relevant summary graph for predicting interval from duration and summarize your results.

- 1.5 Water runoff in the Sierras** (Data file: `water`) Can Southern California's water supply in future years be predicted from past data? One factor affecting water availability is stream runoff. If runoff could be predicted, engineers, planners, and policy makers could do their jobs more efficiently. The data file contains 43 years' worth of precipitation measurements taken at six sites in the Sierra Nevada mountains (labeled `APMAM`, `APSAB`, `APSLAKE`, `OPBPC`, `OPRC`, and `OPSLAKE`) and stream runoff volume at a site near Bishop, California, labeled `BSAAM`.

Draw the scatterplot matrix for these data and summarize the information available from these plots.

- 1.6 Professor ratings** (Data file: `Ratuprof`) In the website and online forum RateMyProfessors.com, students rate and comment on their instructors. Launched in 1999, the site includes millions of ratings on thousands of instructors. The data file includes the summaries of the ratings of 364 instructors at a large campus in the Midwest (Bleske-Rechek and Fritsch, 2011). Each instructor included in the data had at least 10 ratings over a several year period. Students provided ratings of 1–5 on quality,

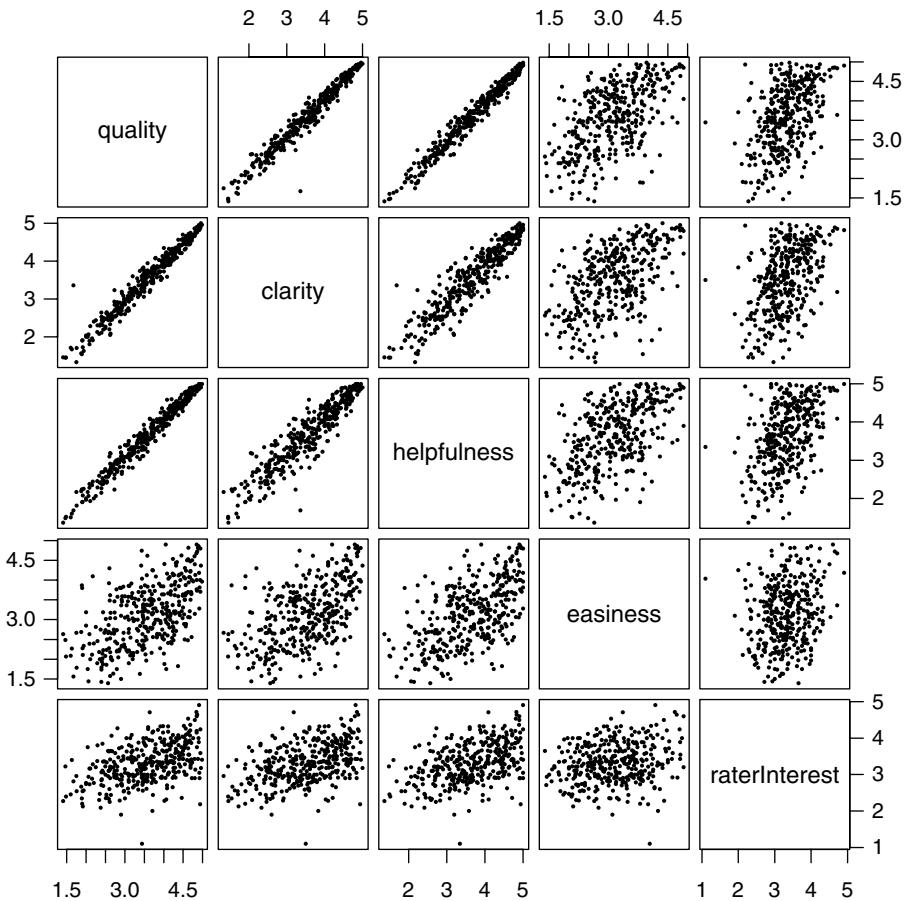


Figure 1.13 Average professor ratings from the file Rateprof.

helpfulness, clarity, easiness of instructor's courses, and rater-Interest in the subject matter covered in the instructor's courses. The data file provides the averages of these five ratings, and these are shown in the scatterplot matrix in Figure 1.13.

Provide a brief description of the relationships between the five ratings.

C H A P T E R 2

Simple Linear Regression

The *simple linear regression model* consists of the mean function and the variance function

$$\begin{aligned}E(Y|X = x) &= \beta_0 + \beta_1 x \\ \text{Var}(Y|X = x) &= \sigma^2\end{aligned}\tag{2.1}$$

The parameters in the mean function are the intercept β_0 , which is the value of $E(Y|X = x)$ when x equals 0, and the slope β_1 , which is the rate of change in $E(Y|X = x)$ for a unit change in X ; see Figure 2.1. We can get all possible straight lines by varying the parameters. The values of the parameters are usually unknown and must be estimated using data. In the simple regression model, the variance function in (2.1) is assumed to be constant, with a positive value σ^2 that is usually unknown.

Because the variance $\sigma^2 > 0$, the observed value of the i th response y_i will typically not equal its expected value $E(Y|X = x_i)$. To account for this difference between the observed data and the expected value, statisticians have invented a quantity called a statistical error, or e_i , for case i defined implicitly by the equation $y_i = E(Y|X = x_i) + e_i$ or explicitly by $e_i = y_i - E(Y|X = x_i)$. The errors e_i depend on unknown parameters in the mean function and so are not observable quantities. They are random variables and correspond to the *vertical distance between the point y_i and the mean function $E(Y|X = x_i)$.* In the heights data, Section 1.1, the errors are the differences between the heights of particular daughters and the average height of all daughters with mothers of a given fixed height.

We make two important assumptions concerning the errors. First, we assume that $E(e_i|x_i) = 0$, so if we could draw a scatterplot of the e_i versus the x_i , we would have a null scatterplot, with no patterns. The second assumption is that the errors are all *independent*, meaning that the value of the error for one case

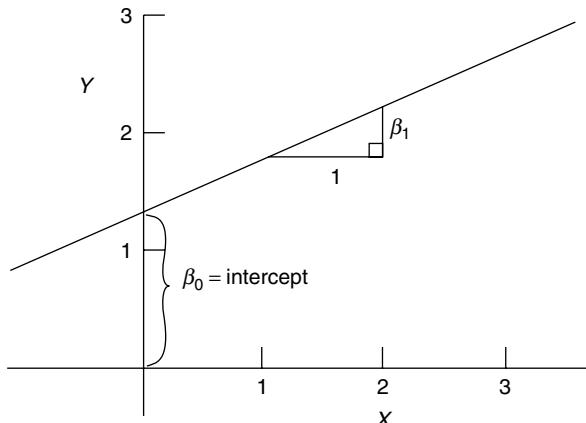


Figure 2.1 Graph of a straight line $E(Y|X=x) = \beta_0 + \beta_1 x$. The *intercept* parameter β_0 is the expected value of the response when the predictor $x = 0$. The *slope* parameter β_1 gives the change in the expected value when the predictor x increases by 1 unit.

gives no information about the value of the error for another case. This is likely to be true in the examples in Chapter 1, although this assumption will not hold in all problems.

Errors are often assumed to be normally distributed, but normality is much stronger than we need. In this book, the normality assumption is used primarily to obtain tests and confidence statements with small samples. If the errors are thought to follow some different distribution, such as the Poisson or the binomial, other methods besides OLS may be more appropriate; we return to this topic in Chapter 12.

2.1 ORDINARY LEAST SQUARES ESTIMATION

Many methods have been suggested for obtaining estimates of parameters in a model. The method discussed here is called *ordinary least squares*, or OLS, in which parameter estimates are chosen to minimize a quantity called the *residual sum of squares*. A formal development of the least squares estimates is given in Appendix A.3.

Parameters are unknown quantities that characterize a model. Estimates of parameters are computable functions of data and are therefore *statistics*. To keep this distinction clear, parameters are denoted by Greek letters like α , β , γ , and σ , and estimates of parameters are denoted by putting a “hat” over the corresponding Greek letter. For example, $\hat{\beta}_1$ (read “beta one hat”) is the estimator of β_1 , and $\hat{\sigma}^2$ is the estimator of σ^2 . The *fitted value* for case i is given by $\hat{E}(Y|X = x_i)$, for which we use the shorthand notation \hat{y}_i ,

$$\hat{y}_i = \hat{E}(Y|X = x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (2.2)$$

Table 2.1 Definitions of Symbols^a

Quantity	Definition	Description
\bar{x}	$\sum x_i/n$	Sample average of x
\bar{y}	$\sum y_i/n$	Sample average of y
S_{XX}	$\sum(x_i - \bar{x})^2 = \sum(x_i - \bar{x})x_i$	Sum of squares for the xs
SD_x^2	$S_{XX}/(n-1)$	Sample variance of the xs
SD_x	$\sqrt{S_{XX}/(n-1)}$	Sample standard deviation of the xs
S_{YY}	$\sum(y_i - \bar{y})^2 = \sum(y_i - \bar{y})y_i$	Sum of squares for the ys
SD_y^2	$S_{YY}/(n-1)$	Sample variance of the ys
SD_y	$\sqrt{S_{YY}/(n-1)}$	Sample standard deviation of the ys
S_{XY}	$\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum(x_i - \bar{x})y_i$	Sum of cross-products
s_{xy}	$S_{XY}/(n-1)$	Sample covariance
r_{xy}	$s_{xy}/(SD_x SD_y)$	Sample correlation

^aIn each equation, the symbol Σ means to add over all n values or pairs of values in the data.

Although the e_i are random variables and not parameters, we shall use the same hat notation to specify the residuals: the residual for the i th case, denoted \hat{e}_i , is given by the equation

$$\hat{e}_i = y_i - \hat{E}(Y|X = x_i) = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \quad i = 1, \dots, n \quad (2.3)$$

which should be compared with the equation for the statistical errors,

$$e_i = y_i - (\beta_0 + \beta_1 x_i) \quad i = 1, \dots, n$$

The computations that are needed for least squares for simple regression depend only on averages of the variables and their sums of squares and sums of cross-products. Definitions of these quantities are given in Table 2.1. Sums of squares and cross-products are centered by subtracting the average from each of the values before squaring or taking cross-products. Appropriate alternative formulas for computing the corrected sums of squares and cross products from uncorrected sums of squares and cross-products that are often given in elementary textbooks are useful for mathematical proofs, but they can be highly inaccurate when used on a computer and should be avoided.

Table 2.1 also lists definitions for the usual univariate and bivariate summary statistics, the sample averages (\bar{x}, \bar{y}), sample variances (SD_x^2, SD_y^2), which are the squares of the sample standard deviations, and the estimated covariance and correlation (s_{xy}, r_{xy}).¹ The “hat” rule described earlier would suggest that different symbols should be used for these quantities; for example, $\hat{\rho}_{xy}$ might be more appropriate for the sample correlation if the population correlation is ρ_{xy} . This inconsistency is deliberate since these sample quantities estimate population values only if the data used are a random sample from a

¹See Appendix A.2.2 for the definitions of the corresponding population quantities.

population. The random sample condition is not required for regression calculations to make sense, and will often not hold in practice.

To illustrate computations, we will use Forbes's data introduced in Section 1.1, for which $n = 17$. The data are given in the file `Forbes`. The response given in the file is the base-ten logarithm of the atmospheric pressure, `lpres` = $100 \times \log_{10}(\text{pres})$ rounded to two decimal digits, and the predictor is the boiling point `bpt`, rounded to the nearest 0.1°F. Neither multiplication by 100 nor the base of the logarithms has important effects on the analysis. Multiplication by 100 avoids using scientific notation for numbers we display in the text, and changing the base of the logarithms merely multiplies the logarithms by a constant. For example, to convert from base-ten logarithms to base-two logarithms, multiply by $\log(10)/\log(2) = 3.3219$. To convert natural logarithms to base-two, multiply by 1.4427.

Forbes's data were collected at 17 selected locations, so the sample variance of boiling points, $SD_x^2 = 33.17$, is not an estimate of any meaningful population variance. Similarly, r_{xy} depends as much on the method of sampling as it does on the population value ρ_{xy} , should such a population value make sense. In the heights example, Section 1.1, if the 1375 mother–daughter pairs can be viewed as a sample from a population, then the sample correlation is an estimate of a population correlation.

The usual sample statistics are often presented and used in place of the corrected sums of squares and cross-products, so alternative formulas are given using both sets of quantities.

2.2 LEAST SQUARES CRITERION

The criterion function for obtaining estimators is based on the residuals, which are the vertical distances between the fitted line and the actual y -values, as illustrated in Figure 2.2. The residuals reflect the inherent asymmetry in the roles of the response and the predictor in regression problems.

The OLS estimators are those values $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the function²

$$\text{RSS}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \quad (2.4)$$

When evaluated at $(\hat{\beta}_0, \hat{\beta}_1)$, we call the quantity $\text{RSS}(\hat{\beta}_0, \hat{\beta}_1)$ the residual sum of squares, or just RSS.

The least squares estimates can be derived in many ways, one of which is outlined in Appendix A.3. They are given by the expressions

$$\begin{aligned} \hat{\beta}_1 &= \frac{\text{SXY}}{\text{SXX}} = r_{xy} \frac{SD_y}{SD_x} = r_{xy} \left(\frac{\text{SYY}}{\text{SXX}} \right)^{1/2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned} \quad (2.5)$$

²We occasionally abuse notation by using the symbol for a fixed though unknown quantity like β_0 or β_1 as if it were a variable argument. Thus, for example, $\text{RSS}(\beta_0, \beta_1)$ is a function of 2 variables to be evaluated as its arguments β_0 and β_1 vary.

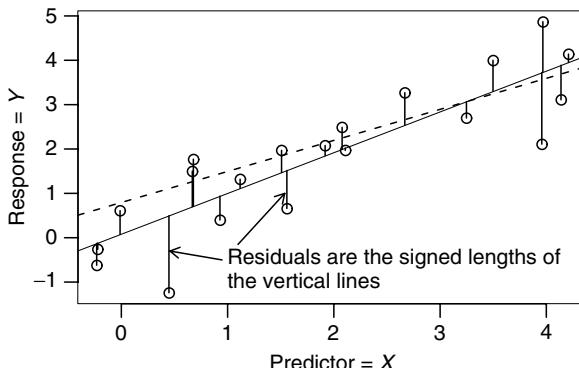


Figure 2.2 A schematic plot for OLS fitting. Each data point is indicated by a small circle. The solid line is the OLS line. The vertical lines between the points and the solid line are the residuals. Points below the line have negative residuals, while points above the line have positive residuals. The true mean function shown as a dashed line for these simulated data is $E(Y|X=x) = 0.7 + .8x$.

The several forms for $\hat{\beta}_1$ are all equivalent.

We emphasize again that OLS produces *estimates* of parameters but not the actual values of the parameters. As a demonstration, the data in Figure 2.2 were created by setting the x_i to be random sample of 20 numbers from a normal distribution with mean 2 and variance 1.5 and then computing $y_i = 0.7 + 0.8x_i + e_i$, where the errors were sampled from a normal distribution with mean 0 and variance 1. The graph of the true mean function is shown in Figure 2.2 as a dashed line, and it seems to match the data poorly compared with OLS, given by the solid line. Since OLS minimizes (2.4), it will always fit at least as well as, and generally better than, the true mean function.

Using Forbes's data to illustrate computations, we will write \bar{x} to be the sample mean of `bpx` and \bar{y} to be the sample mean of `lpres`. The quantities needed for computing the least squares estimators are

$$\begin{aligned}\bar{x} &= 202.9529 & SXX &= 530.7824 & SXY &= 475.3122 \\ \bar{y} &= 139.6053 & SYY &= 427.7942\end{aligned}\tag{2.6}$$

The quantity SYY , although not yet needed, is given for completeness. In the rare instances that regression calculations are not done using statistical software, intermediate calculations such as these should be done as accurately as possible, and rounding should be done only to final results. We will generally display final results with two or three digits beyond the decimal point. Using (2.6), we find

$$\hat{\beta}_1 = \frac{SXY}{SXX} = 0.895$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -42.138$$

The estimated intercept $\hat{\beta}_0 = -42.138^\circ\text{F}$ is the estimated value of lpres when $\text{bp} = 0$. Since the temperatures in the data are in the range from about 194°F to 212°F , this estimate does not have a useful physical interpretation. The estimated slope of $\hat{\beta}_1 = 0.895$ is the change in lpres for a 1°F change in bp .

The estimated line given by

$$\hat{E}(\text{lpres}|\text{bp}) = -42.138 + 0.895\text{bp}$$

was drawn in Figure 1.4a. The fit of this line to the data is excellent.

2.3 ESTIMATING THE VARIANCE σ^2

Since the variance σ^2 is essentially the average squared size of the e_i^2 , we should expect that its estimator $\hat{\sigma}^2$ is obtained by averaging the squared residuals. Under the assumption that the errors are uncorrelated random variables with 0 means and common variance σ^2 , an unbiased estimate of σ^2 is obtained by dividing $\text{RSS} = \sum \hat{e}_i^2$ by its *degrees of freedom* (df), where residual $df = \text{number of cases minus the number of parameters in the mean function}$. For simple regression, residual $df = n - 2$, so the estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-2} \quad (2.7)$$

This quantity is called the *residual mean square*. In general, any sum of squares divided by its df is called a mean square. The residual sum of squares can be computed by squaring the residuals and adding them up. It can also be computed from the formula (Problem 2.18)

$$\text{RSS} = SYY - \frac{SXY^2}{SXX} = SYY - \hat{\beta}_1^2 SXX \quad (2.8)$$

Using the summaries for Forbes's data given at (2.6), we find

$$\begin{aligned} \text{RSS} &= 427.794 - \frac{475.3122^2}{530.7824} \\ &= 2.1549 \end{aligned} \quad (2.9)$$

$$\sigma^2 = \frac{2.1549}{17-2} = 0.1436 \quad (2.10)$$

The square root of $\hat{\sigma}^2$, $\hat{\sigma} = \sqrt{0.1436} = 0.379$ is called the *standard error of regression*. It is in the same units as is the response variable.

If in addition to the assumptions made previously the e_i are drawn from a normal distribution, then the residual mean square will be distributed as a multiple of a chi-squared random variable with $df = n - 2$, or in symbols,

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-2} \chi^2(n-2)$$

This is proved in more advanced books on linear models and is used to obtain the distribution of test statistics and also to make confidence statements concerning σ^2 . In addition, since the mean of a χ^2 random variable with m df is m ,

$$E(\hat{\sigma}^2|X) = \frac{\sigma^2}{n-2} E[\chi^2(n-2)] = \frac{\sigma^2}{n-2}(n-2) = \sigma^2$$

This shows that $\hat{\sigma}^2$ is an unbiased estimate of σ^2 if the errors are normally distributed, although normality is not required for this result to hold. Expectations throughout this chapter condition on X to remind us that X is treated as fixed and the expectation is over the conditional distribution of $Y|X$, or equivalently of the conditional distribution of $e|X$.

2.4 PROPERTIES OF LEAST SQUARES ESTIMATES

The OLS estimates depend on data only through the statistics given in Table 2.1. This is both an advantage, making computing easy, and a disadvantage, since any two data sets for which these are identical give the same fitted regression, even if a straight-line model is appropriate for one but not the other, as we have seen in the example from Anscombe (1973) in Section 1.4. The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ can both be written as linear combinations of y_1, \dots, y_n . Writing $c_i = (x_i - \bar{x})/\text{SXX}$ (see Appendix A.3), then

$$\hat{\beta}_1 = \left(\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\text{SXX}} \right) = \sum \left(\frac{x_i - \bar{x}}{\text{SXX}} \right) y_i = \sum c_i y_i$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \sum \left(\frac{1}{n} - c_i \bar{x} \right) y_i = \sum d_i y_i$$

with $d_i = (1/n - c_i \bar{x})$. A fitted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is equal to $\sum(d_i + c_i x_i)y_i$, also a linear combination of the y_i .

The fitted value at $x = \bar{x}$ is

$$\hat{E}(Y|X = \bar{x}) = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{y}$$

so the fitted line passes through the point (\bar{x}, \bar{y}) , intuitively the center of the data. Finally, as long as the mean function includes an intercept, $\sum \hat{e}_i = 0$. Mean functions without an intercept may have $\sum \hat{e}_i \neq 0$.

Since the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ depend on the random e_i s, the estimates are also random variables. If all the e_i have 0 mean and the mean function is correct, then, as shown in Appendix A.4, the least squares estimates are unbiased,

$$\begin{aligned} E(\hat{\beta}_0|X) &= \beta_0 \\ E(\hat{\beta}_1|X) &= \beta_1 \end{aligned}$$

The variances of the estimators, assuming $\text{Var}(e_i|X) = \sigma^2$, $i = 1, \dots, n$, and $\text{Cov}(e_i, e_j|X) = 0$, $i \neq j$, are from Appendix A.4,

$$\begin{aligned} \text{Var}(\hat{\beta}_1|X) &= \sigma^2 \frac{1}{S_{XX}} \\ \text{Var}(\hat{\beta}_0|X) &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right) \end{aligned} \tag{2.11}$$

From (2.5) we have $\hat{\beta}_0$ depends on $\hat{\beta}_1$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, and so it is no surprise that the estimates are correlated, and

$$\begin{aligned} \text{Cov}(\hat{\beta}_0, \hat{\beta}_1|X) &= \text{Cov}(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1|X) \\ &= \text{Cov}(\bar{y}, \hat{\beta}_1|X) - \bar{x} \text{Var}(\hat{\beta}_1|X) \\ &= -\sigma^2 \frac{\bar{x}}{S_{XX}} \end{aligned} \tag{2.12}$$

The estimated slope and intercept are generally correlated unless the predictor is centered to have $\bar{x} = 0$ (Problem 2.8). The correlation between the intercept and slope estimates is

$$\rho(\hat{\beta}_0, \hat{\beta}_1|X) = \frac{-\bar{x}}{\sqrt{S_{XX}/n + \bar{x}^2}}$$

The correlation will be close to plus or minus 1 if the variation in the predictor reflected in S_{XX} is small relative to \bar{x} .

The *Gauss–Markov theorem* provides an optimality result for ols estimates. Among all estimates that are linear combinations of the ys and

unbiased, the OLS estimates have the smallest variance. These estimates are called the *best linear unbiased estimates*, or BLUE. If one believes the assumptions and is interested in using linear unbiased estimates, the OLS estimates are the ones to use.

The means and variances, and covariances of the estimated regression coefficients do not require a distributional assumption concerning the errors. Since the estimates are linear combinations of the y_i , and hence linear combinations of the errors e_i , the central limit theorem shows that the coefficient estimates will be approximately normally distributed if the sample size is large enough.³ For smaller samples, if the errors $e = y - E(y|X = x)$ are independent and normally distributed, written in symbols as

$$e_i|X \sim NID(0, \sigma^2) \quad i = 1, \dots, n$$

then the regression estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ will have a joint normal distribution with means, variances, and covariances as given before. When the errors are normally distributed, the OLS estimates can be justified using a completely different argument, since they are then also maximum likelihood estimates, as discussed in any mathematical statistics text, for example, Casella and Berger (2001).

2.5 ESTIMATED VARIANCES

Estimates of $\text{Var}(\hat{\beta}_0|X)$ and $\text{Var}(\hat{\beta}_1|X)$ are obtained by substituting $\hat{\sigma}^2$ for σ^2 in (2.11). We use the symbol $\widehat{\text{Var}}(\cdot)$ for an estimated variance. Thus

$$\widehat{\text{Var}}(\hat{\beta}_0|X) = \hat{\sigma}^2 \frac{1}{S_{XX}}$$

$$\widehat{\text{Var}}(\hat{\beta}_1|X) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right)$$

The square root of an estimated variance is called a *standard error*, for which we use the symbol $\text{se}(\cdot)$. The use of this notation is illustrated by

$$\text{se}(\hat{\beta}_1|X) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_1|X)}$$

The terms *standard error* and *standard deviation* are sometimes used interchangeably. In this book, an estimated standard deviation always refers to the variability between values of an observable random variable like the response

³The main requirement for all estimates to be normally distributed in large samples is that $\max_i [(x_i - \bar{x})^2/S_{XX}]$ must get close to 0 as the sample size increases (Huber and Ronchetti, 2009, Proposition 7.1).

y_i or an unobservable random variance like the errors e_i . The term standard error will always refer to the square root of the estimated variance of a statistic like a mean \bar{y} , or a regression coefficient $\hat{\beta}_1$.

2.6 CONFIDENCE INTERVALS AND t -TESTS

Estimates of regression coefficients and fitted values are all subject to uncertainty, and assessing the amount of uncertainty is an important part of most analyses. Confidence intervals result in interval estimates, while tests provide methodology for making decisions concerning the value of a parameter or fitted value.

When the errors are NID(0, σ^2), parameter estimates, fitted values, and predictions will be normally distributed because all of these are linear combinations of the y_i and hence of the e_i . Confidence intervals and tests can be based on a t -distribution, which is the appropriate distribution with normal estimates but using $\hat{\sigma}^2$ to estimate the unknown variance σ^2 . There are many t -distributions, indexed by the number of df associated with $\hat{\sigma}$. Suppose we let $t(\alpha/2, d)$ be the value that cuts off $\alpha/2 \times 100\%$ in the *upper tail* of the t -distribution with d df . These values can be computed in most statistical packages or spreadsheet software.⁴

2.6.1 The Intercept

The intercept is used to illustrate the general form of confidence intervals for normally distributed estimates. The standard error of the intercept is $se(\hat{\beta}_0|X) = \hat{\sigma}(1/n + \bar{x}^2/SXX)^{1/2}$. Hence, a $(1 - \alpha) \times 100\%$ confidence interval for the intercept is the set of points $\hat{\beta}_0$ in the interval

$$\hat{\beta}_0 - t(\alpha/2, n-2)se(\hat{\beta}_0|X) \leq \hat{\beta}_0 \leq \hat{\beta}_0 + t(\alpha/2, n-2)se(\hat{\beta}_0|X)$$

For Forbes's data, $se(\hat{\beta}_0|X) = 0.379(1/17 + (202.953)^2/530.724)^{1/2} = 3.340$. For a 90% confidence interval, $t(0.05, 15) = 1.753$, and the interval is

$$\begin{aligned} -42.138 - 1.753(3.340) &\leq \hat{\beta}_0 \leq -42.138 + 1.753(3.340) \\ -47.99 &\leq \hat{\beta}_0 \leq -36.28 \end{aligned}$$

Ninety percent of such intervals will include the true value.

A hypothesis test of

$$\begin{aligned} NH: \quad \beta_0 &= \beta_0^*, \quad \beta_1 \text{ arbitrary} \\ AH: \quad \beta_0 &\neq \beta_0^*, \quad \beta_1 \text{ arbitrary} \end{aligned}$$

⁴Readily available functions include `tinv` in Microsoft Excel, and the function `pt` in R. Tables of the t distributions can be easily found by googling `t table`.

is obtained by computing the *t*-statistic

$$t = \frac{\hat{\beta}_0 - \beta_0^*}{\text{se}(\hat{\beta}_0|X)} \quad (2.13)$$

and referring this ratio to the *t*-distribution with $df = n - 2$, the number of *df* in the estimate of σ^2 . For example, in Forbes's data, consider testing the NH $\beta_0 = -35$ against the alternative that $\beta_0 \neq -35$. The statistic is

$$t = \frac{-42.138 - (-35)}{3.34} = -2.137$$

Since AH is two-sided, the *p*-value corresponds to the probability that a *t*(15) variable is less than -2.137 or greater than $+2.137$, which gives a *p*-value that rounds to 0.05, providing some evidence against NH. This hypothesis test for these data is not one that would occur to most investigators and is used only as an illustration.

2.6.2 Slope

A 95% confidence interval for the slope, or for any of the partial slopes in multiple regression, is the set of β_1 such that

$$\hat{\beta}_1 - t(\alpha/2, df)\text{se}(\hat{\beta}_1|X) \leq \beta_1 \leq \hat{\beta}_1 + t(\alpha/2, df)\text{se}(\hat{\beta}_1|X) \quad (2.14)$$

For simple regression, $df = n - 2$ and $\text{se}(\hat{\beta}_1|X) = \hat{\sigma}/\sqrt{S_{XX}}$. For Forbes's data, $df = 15$, $\text{se}(\hat{\beta}_1|X) = 0.0165$, and

$$0.895 - 2.131(0.0165) \leq \beta_1 \leq 0.895 + 2.131(0.0165)$$

$$0.86 \leq \beta_1 \leq 0.93$$

As an example of a test for slope equal to 0, consider the Ft. Collins snowfall data in Section 1.1. One can show, Problem 2.5, that $\hat{\beta}_1 = 0.203$, $\text{se}(\hat{\beta}_1|X) = 0.131$. The test of interest is of

$$\begin{aligned} \text{NH: } & \beta_1 = 0 \\ \text{AH: } & \beta_1 \neq 0 \end{aligned} \quad (2.15)$$

and $t = (0.203 - 0)/0.131 = 1.553$. To get a significance level for this test, compare *t* with the *t*(91) distribution; the two-sided *p*-value is 0.124, suggesting no evidence against the NH that Early and Late season snowfalls are independent.

2.6.3 Prediction

The estimated mean function can be used to obtain values of the response for given values of the predictor. The two important variants of this problem are *prediction* and *estimation of fitted values*. Since prediction is more important, we discuss it first.

In prediction we have a new case, possibly a future value, not one used to estimate parameters, with observed value of the predictor x_* . We would like to know the value y_* , the corresponding response, but it has not yet been observed. If we assume that the data used to estimate the mean function are relevant to the new case, then the model fitted to the observed data can be used to predict for the new case. In the heights example, we would probably be willing to apply the fitted mean function to mother–daughter pairs alive in England at the end of the nineteenth century. Whether the prediction would be reasonable for mother–daughter pairs in other countries or in other time periods is much less clear. In Forbes’s problem, we would probably be willing to apply the results for altitudes in the range he studied. Given this additional assumption, a point prediction of y_* , say \tilde{y}_* , is just

$$\tilde{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$$

\tilde{y}_* predicts the as yet unobserved y_* . Assuming the model is correct, then the true value of y_* is

$$y_* = \beta_0 + \beta_1 x_* + e_*$$

where e_* is the random error attached to the future value, presumably with variance σ^2 . Thus, even if β_0 and β_1 were known exactly, predictions would not match true values perfectly, but would be off by a random amount with standard deviation σ . In the more usual case where the coefficients are estimated, the prediction error variability will have a second component that arises from the uncertainty in the estimates of the coefficients. Combining these two sources of variation and using Appendix A.4,

$$\text{Var}(\tilde{y}_*|x_*) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\text{SXX}} \right) \quad (2.16)$$

The first σ^2 on the right of (2.16) corresponds to the variability due to e_* , and the remaining term is the error for estimating coefficients. If x_* is similar to the x_i used to estimate the coefficients, then the second term will generally be much smaller than the first term. If x_* is very different from the x_i used in estimation, the second term can dominate.

Taking square roots of both sides of (2.16) and estimating σ^2 by $\hat{\sigma}^2$, we get the standard error of prediction (sepred) at x_* ,

$$\text{sepred}(\tilde{y}_*|x_*) = \sigma \left(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\text{SXX}} \right)^{1/2} \quad (2.17)$$

A prediction interval uses multipliers from the *t*-distribution with *df* equal to the *df* in estimating σ^2 . For prediction of $100 \times \log(\text{pres})$ for a location with $x_* = 200$, the point prediction is $\tilde{y}_* = -42.138 + +0.895(200) = 136.961$, with standard error of prediction

$$\begin{aligned} \text{sepred}(\tilde{y}_*|x_* = 200) &= 0.379 \left(1 + \frac{1}{17} + \frac{(200 - 202.9529)^2}{530.7824} \right)^{1/2} \\ &= 0.393 \end{aligned}$$

Thus, a 99% predictive interval is the set of all y_* such that

$$136.961 - 2.95(0.393) \leq y_* \leq 136.961 + 2.95(0.393)$$

$$135.803 \leq y_* \leq 138.119$$

More interesting would be a 99% prediction interval for `pres`, rather than for $100 \times \log(\text{pres})$. A point prediction is just $10^{(136.961/100)} = 23.421$ inches of Mercury. The prediction interval is found by exponentiating the end points of the interval in log scale. Dividing by 100 and then exponentiating, we get

$$10^{135.803/100} \leq \text{pres} \leq 10^{138.119/100}$$

$$22.805 \leq \text{pres} \leq 24.054$$

In the original scale, the prediction interval is not symmetric about the point estimate.

For the heights data, Figure 2.3 is a plot of the estimated mean function given by the dashed line for the regression of `dheight` on `mheight` along with curves at

$$\hat{\beta}_0 + \hat{\beta}_1 x_* \pm t(.025, 1373) \text{sepred}(\text{dheight}_*|\text{mheight}_*)$$

The vertical distance between the two solid curves for any value of `mheight` corresponds to a 95% prediction interval for daughter's height given mother's height. Although not obvious from the graph because of the very large sample size, the interval is wider for mothers who were either relatively tall or short, as the curves bend outward from the narrowest point at `mheight` = `mheight`.

2.6.4 Fitted Values

In rare problems, one may be interested in obtaining an estimate of $E(Y|X=x_*)$. In the heights data, this is like asking for the population mean height of all

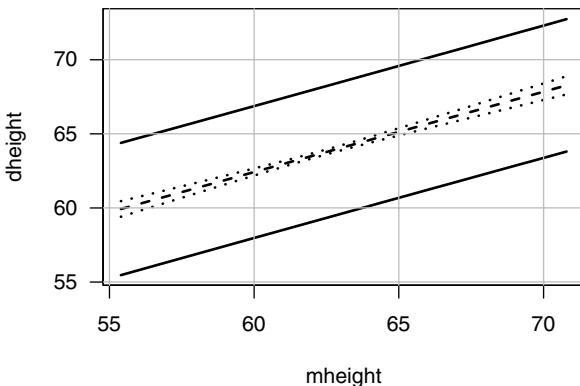


Figure 2.3 Prediction intervals (solid lines) and intervals for fitted values (dashed lines) for the heights data.

daughters of mothers with a particular height. This quantity is estimated by the fitted value $\hat{y} = \beta_0 + \beta_1 x_*$, and its standard error is

$$\text{sefit}(\hat{y}|x_*) = \hat{\sigma} \left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{XX}} \right)^{1/2}$$

To obtain confidence intervals, it is more usual to compute a simultaneous interval for all possible values of x . This is the same as first computing a joint confidence region for β_0 and β_1 , and from these, computing the set of all possible mean functions with slope and intercept in the joint confidence set. The confidence region for the mean function is the set of all y such that

$$\begin{aligned} & (\hat{\beta}_0 + \hat{\beta}_1 x) - \text{sefit}(\hat{y}|x)[2F(\alpha; 2, n-2)]^{1/2} \leq y \\ & \leq (\hat{\beta}_0 + \hat{\beta}_1 x) + \text{sefit}(\hat{y}|x)[2F(\alpha; 2, n-2)]^{1/2} \end{aligned}$$

This formula uses an F -distribution with 2 and $n - 2$ df in place of the t distribution to correct for the simultaneous inference about two estimates rather than just one.⁵ For multiple regression, replace $2F(\alpha; 2, n - 2)$ by $p'F(\alpha; p', n - p')$, where p' is the number of parameters estimated in the mean function including the intercept. The simultaneous band for the fitted line for the heights data is shown in Figure 2.3 as the vertical distances between the two dotted lines. The prediction intervals are much wider than the confidence intervals. Why is this so (Problem 2.13)?

⁵Like the t distributions, tables of F distributions are available using the `finv` function in Microsoft Excel, the function `pf` in R or by googling `F table`.

2.7 THE COEFFICIENT OF DETERMINATION, R^2

Ignoring all possible predictors, the best prediction of a response y would simply be the sample average \bar{y} of the values of the response observed in the data. The *total sum of squares* $SYY = \sum(y_i - \bar{y})^2$ is the observed total variation of the response, ignoring any and all predictors. The total sum of squares is the sum of squared deviations from the horizontal line illustrated in Figure 2.4.

When we include a predictor, the *unexplained* variation is given by RSS , the sum of squared deviations from the fitted line, as shown on Figure 2.4. The difference between these sums of squares is called the *sum of squares due to regression*, $SSreg$, defined by

$$SSreg = SYY - RSS \quad (2.18)$$

We can get a computing formula for $SSreg$ by substituting for RSS from (2.8),

$$SSreg = SYY - \left(SYY - \frac{(SYY)^2}{SXX} \right) = \frac{(SXY)^2}{SXX} \quad (2.19)$$

If both sides of (2.18) are divided by SYY , we get

$$\frac{SSreg}{SYY} = 1 - \frac{RSS}{SYY} \quad (2.20)$$

The left-hand side of (2.20) is the proportion of observed variability in the response explained by regression on the predictor. The right-hand side consists of one minus the remaining unexplained variability. This concept of dividing

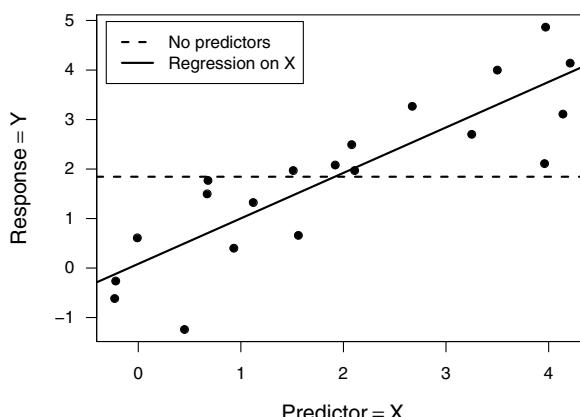


Figure 2.4 Unexplained variation. The sum of squared deviations of the points from the horizontal line represent is the total variation. The sum of squared deviations from the OLS line is the remaining variation unexplained by regression on the predictor.

up the total variability according to whether or not it is explained is of sufficient importance that a special name is given to it. We define R^2 , the coefficient of determination, to be

$$R^2 = \frac{SS_{\text{reg}}}{SYY} = 1 - \frac{RSS}{SYY} \quad (2.21)$$

R^2 is a scale-free one-number summary of the strength of the relationship between the x_i and the y_i in the data. It generalizes nicely to multiple regression, depends only on the sums of squares, and appears to be easy to interpret. For Forbes's data,

$$R^2 = \frac{SS_{\text{reg}}}{SYY} = \frac{425.6391}{427.794} = 0.995$$

and thus about 99.5% of the variability in the observed values of $100 \times \log(\text{pres})$ is explained by boiling point. Since R^2 does not depend on units of measurement, we would get the same value if we had used logarithms with a different base, or if we did not multiply $\log(\text{pres})$ by 100, or if we replaced the response bp by $a_0 + a_1 \text{bp}$ for any a_0 and for any $a_1 \neq 0$.

By appealing to (2.18) and to Table 2.1, we can write

$$R^2 = \frac{SS_{\text{reg}}}{SYY} = \frac{(SXY)^2}{SXX \times SYY} = r_{xy}^2$$

and thus R^2 in simple linear regression is the same as the square of the sample correlation between the predictor and the response.

Many computer packages will also produce an adjusted R^2 , defined by

$$R_{\text{adj}}^2 = 1 - \frac{RSS/df}{SYY/(n-1)}$$

This differs from (2.21) by adding a correction for df of the sums of squares that can facilitate comparing models in multiple regression. R_{adj}^2 is not used in this book because there are better ways of making this comparison discussed in Chapter 10.

2.8 THE RESIDUALS

Plots of residuals versus other quantities are used to find failures of assumptions. The most common plot, especially useful in simple regression, is the plot of residuals versus the fitted values. A null plot would indicate no failure of assumptions. Curvature might indicate that the fitted mean function is

inappropriate. Residuals that seem to increase or decrease in average magnitude with the fitted values might indicate nonconstant residual variance. A few relatively large residuals may be indicative of outliers, cases for which the model is somehow inappropriate.

The plot of residuals versus fitted values for the heights data is shown in Figure 2.5. This is a null plot, indicating no particular problems. The simple linear regression model provides a useful summary for these data.

The fitted values and residuals for Forbes's data are plotted in Figure 2.6. The residuals are generally small compared with the fitted values, and they do not follow any distinct pattern in Figure 2.6. The residual for case number 12 is about 4 times the size of the next largest residual in absolute value. This may suggest that the assumptions concerning the errors are not correct. Either $\text{Var}(100 \times \log(\text{pressure})|\text{bp})$ may not be constant or for case 12, the corresponding error may have a large fixed component. Forbes may have misread or miscopied the results of his calculations for this case, which would suggest

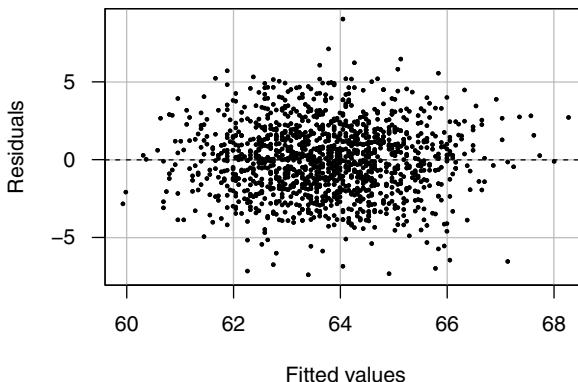


Figure 2.5 Residuals versus fitted values for the heights data.

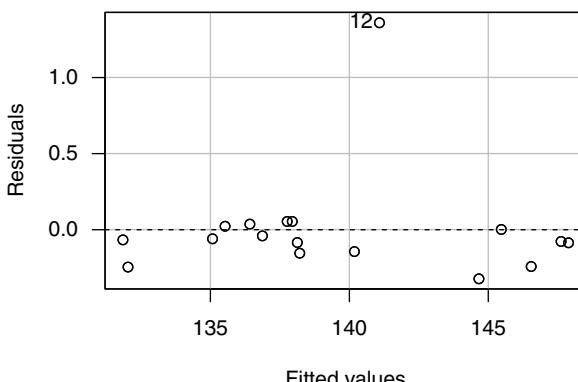


Figure 2.6 Residuals versus fitted values for Forbes's data.

Table 2.2 Summary Statistics for Forbes's Data with All Data and with Case 12 Deleted

Quantity	All Data	Delete Case 12
$\hat{\beta}_0$	-42.138	-41.308
$\hat{\beta}_1$	0.895	0.891
$se(\hat{\beta}_0)$	3.340	1.001
$se(\hat{\beta}_1)$	0.016	0.005
$\hat{\sigma}$	0.379	0.113
R^2	0.995	1.000

that the numbers in the data do not correspond to the actual measurements. Forbes noted this possibility himself, by marking this pair of numbers in his paper as being “evidently a mistake,” presumably because of the large observed residual.

Since we are concerned with the effects of case 12, we could refit the data, this time without case 12, and then examine the changes that occur in the estimates of parameters, fitted values, residual variance, and so on. This is summarized in Table 2.2, giving estimates of parameters, their standard errors, $\hat{\sigma}^2$, and the coefficient of determination R^2 with and without case 12. The estimated intercept is somewhat smaller when case 12 is removed, while the intercept is nearly unchanged. In other regression problems, deletion of a single case can change everything. The effect of case 12 on standard errors is more marked: if case 12 is deleted, standard errors are decreased by a factor of about 3.1, and variances are decreased by a factor of about $3.1^2 \approx 10$. Inclusion of this case gives the appearance of less reliable results than would be suggested on the basis of the other 16 cases. In particular, prediction intervals of *pres* are much wider based on all the data than on the 16-case data, although the point predictions are nearly the same. The residual plot obtained when case 12 is deleted before computing indicates no obvious failures in the remaining 16 cases.

Two competing fits using the same mean function but somewhat different data are available, and they lead to slightly different conclusions, although the results of the two analyses agree more than they disagree. On the basis of the data, there is no real way to choose between the two, and we have no way of deciding which is the correct OLS analysis of the data. A good approach to this problem is to describe both or, in general, all plausible alternatives.

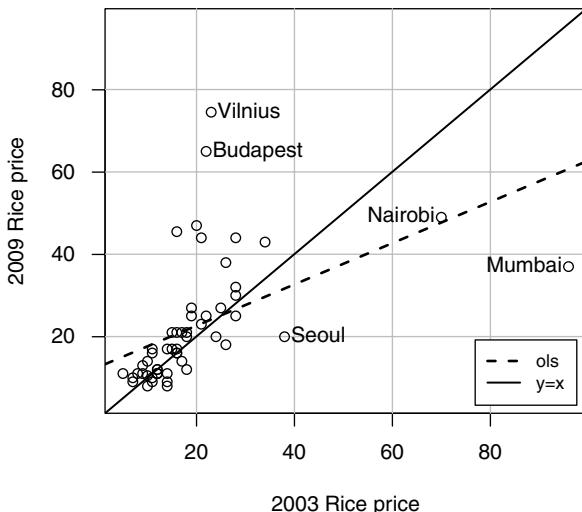
2.9 PROBLEMS

2.1 Height and weight data (Data file: *Htwt*) The table below and the data file give *ht* = height in centimeters and *wt* = weight in kilograms for a

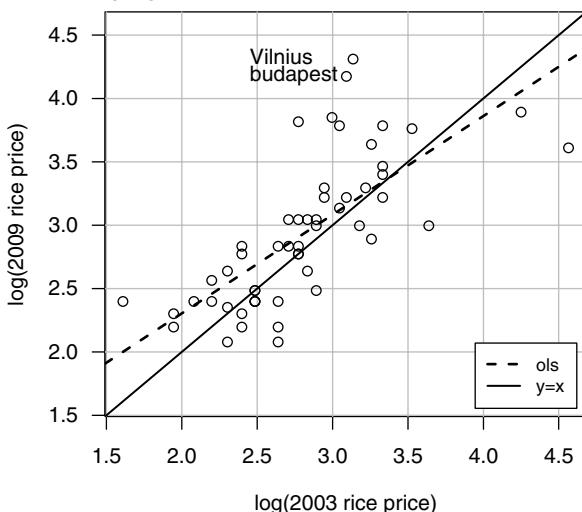
sample of $n = 10$ 18-year-old girls. The data are taken from a larger study described in Problem 3.3. Interest is in predicting weight from height.

ht	wt
169.6	71.2
166.8	58.2
157.1	56.0
181.1	64.5
158.4	53.0
165.6	52.4
166.7	56.8
156.5	49.2
168.1	55.6
165.3	77.8

- 2.1.1** Draw a scatterplot of wt on the vertical axis versus ht on the horizontal axis. On the basis of this plot, does a simple linear regression model make sense for these data? Why or why not?
- 2.1.2** Show that $\bar{x} = 165.52$, $\bar{y} = 59.47$, $S_{XX} = 472.08$, $S_{YY} = 731.96$, and $S_{XY} = 274.79$. Compute estimates of the slope and the intercept for the regression of Y on X . Draw the fitted line on your scatterplot.
- 2.1.3** Obtain the estimate of σ^2 and find the estimated standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$. Also find the estimated covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$. Compute the t -tests for the hypotheses that $\beta_0 = 0$ and that $\beta_1 = 0$ and find the appropriate p -values using two-sided tests.
- 2.2** (Data file: UBSprices) The international bank UBS regularly produces a report (UBS, 2009) on prices and earnings in major cities throughout the world. Three of the measures they include are prices of basic commodities, namely 1kg of rice, a 1kg loaf of bread, and the price of a Big Mac hamburger at McDonalds. An interesting feature of the prices they report is that prices are measured in the minutes of labor required for a “typical” worker in that location to earn enough money to purchase the commodity. Using minutes of labor corrects at least in part for currency fluctuations, prevailing wage rates, and local prices. The data file includes measurements for rice, bread, and Big Mac prices from the 2003 and the 2009 reports. The year 2003 was before the major recession hit much of the world around 2008, and the year 2009 may reflect changes in prices due to the recession.
- The figure below is the plot of $y = \text{rice2009}$ versus $x = \text{rice2003}$, the price of rice in 2009 and 2003, respectively, with the cities corresponding to a few of the points marked.



- 2.2.1** The line with equation $y = x$ is shown on this plot as the solid line. What is the key difference between points above this line and points below the line?
- 2.2.2** Which city had the largest increase in rice price? Which had the largest decrease in rice price?
- 2.2.3** The OLS line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is shown on the figure as a dashed line, and evidently $\hat{\beta}_1 < 1$. Does this suggest that prices are lower in 2009 than in 2003? Explain your answer.
- 2.2.4** Give two reasons why fitting simple linear regression to the figure in this problem is not likely to be appropriate.
- 2.3** (Data file: UBSprices) This is a continuation of Problem 2.2. An alternative representation of the data used in the last problem is to use log scales, as in the following figure:



2.3.1 Explain why this graph and the graph in Problem 2.2 suggests that using log-scale is preferable if fitting simple linear regression is desired.

2.3.2 Suppose we start with a proposed model

$$E(y|x) = \gamma_0 x^{\beta_1}$$

This is a common model in many areas of study. Examples include *allometry* (Gould, 1966), where x could represent the size of one body characteristic such as total weight and y represents some other body characteristic, such as brain weight, *psychophysics* (Stevens, 1966), in which x is a physical stimulus and y is a psychological response to it, or in economics, where x could represent inputs and y outputs, where this relationship is often called a Cobb–Douglas production function (Greene, 2003).

If we take the logs of both sides of the last equation, we get

$$\log(E(y|x)) = \log(\gamma_0) + \beta_1 \log(x)$$

If we approximate $\log(E(y|x)) \approx E(\log(y)|x)$, and write $\beta_0 = \log(\gamma)$, to the extent that the logarithm of the expectation equals the expectation of the logarithm, we have

$$E(\log(y)|x) = \beta_0 + \beta_1 \log(x)$$

Give an interpretation of β_0 and β_1 in this setting, assuming $\beta_1 > 0$.

2.4 (Data file: `UBSprices`) This problem continues with the data file `UBSprices` described in Problem 2.2.

2.4.1 Draw the plot of $y = \text{bigmac2009}$ versus $x = \text{bigmac2003}$, the price of a Big Mac hamburger in 2009 and 2003. On this plot draw (1) the ols fitted line; (2) the line $y = x$. Identify the most unusual cases and describe why they are unusual.

2.4.2 Give two reasons why fitting simple linear regression to the figure in this problem is not likely to be appropriate.

2.4.3 Plot $\log(\text{bigmac2009})$ versus $\log(\text{bigmac2003})$ and explain why this graph is more sensibly summarized with a linear regression.

2.5 Ft. Collins snowfall data (Data file: `ftcollinssnow`) Verify the t -test for the slope in the Ft. Collins snowfall data given in Section 2.6.2.

2.6 Ft. Collins temperature data (Data file: `ftcollinstemp`) The data file gives the mean temperature in the fall of each year, defined as September 1 to November 30, and the mean temperature in the following winter, defined as December 1 to the end of February in the following calendar year, in degrees Fahrenheit, for Ft. Collins, CO (Colorado

Climate Center, 2012). These data cover the time period from 1900 to 2010. The question of interest is: Does the average fall temperature predict the average winter temperature?

- 2.6.1** Draw a scatterplot of the response versus the predictor, and describe any pattern you might see in the plot.
- 2.6.2** Use statistical software to fit the regression of the response on the predictor. Add the fitted line to your graph. Test the slope to be 0 against a two-sided alternative, and summarize your results.
- 2.6.3** Compute or obtain from your computer output the value of the variability in winter explained by fall and explain what this means.
- 2.6.4** Divide the data into 2 time periods, an early period from 1900 to 1989, and a late period from 1990 to 2010. You can do this using the variable `year` in the data file. Are the results different in the two time periods?

- 2.7 More with Forbes's data** (Data files: `Forbes` and `Hooker`) An alternative approach to the analysis of Forbes's experiments comes from the Clausius–Clapeyron formula of classical thermodynamics, which dates to Clausius (1850). According to this theory, we should find that

$$E(\text{pres}|\text{bp}) = \beta_0 + \beta_1 \frac{1}{\text{bpKelvin}} \quad (2.22)$$

where `bpKelvin` is boiling point in kelvin, which equals $255.37 + (5/9) \times \text{bp}$. If we were to graph this mean function on a plot of `pres` versus `bpKelvin`, we would get a curve, not a straight line. However, we can estimate the parameters β_0 and β_1 using simple linear regression methods by defining u_1 to be the inverse of temperature in kelvin,

$$u_1 = \frac{1}{\text{bpKelvin}} = \frac{1}{(5/9)\text{bp} + 255.37}$$

The mean function (2.22) can be rewritten as

$$E(\text{pres}|\text{bp}) = \beta_0 + \beta_1 u_1 \quad (2.23)$$

for which simple linear regression is suitable. The notation we have used in (2.23) is a little different, as the left side of the equation says we are conditioning on `bp`, but the variable `bp` does not appear explicitly on the right side of the equation, although of course the regressor u_1 depends on `bp`.

- 2.7.1** Draw the plot of `pres` versus u_1 , and verify that apart from case 12 the 17 points in Forbes's data fall close to a straight line. Explain

why the apparent slope in this graph is negative when the slope in Figure 1.4a is positive.

- 2.7.2** Compute the linear regression implied by (2.23), and summarize your results.
- 2.7.3** We now have two possible models for the same data based on the regression of `pres` on `bp` used by Forbes, and (2.23) based on the Clausius–Clapeyron formula. To compare these two mean functions, draw the plot of the fitted values from Forbes's mean function fit versus the fitted values from (2.23). On the basis of these and any other computations you think might help, is it possible to prefer one approach over the other? Why?
- 2.7.4** In his original paper, Forbes provided additional data collected by the botanist Joseph D. Hooker (1817–1911) on temperatures and boiling points measured often at higher altitudes in the Himalaya Mountains. The data for $n = 31$ locations is given in the file `Hooker`. Find the estimated mean function (2.23) for Hooker's data.

- 2.8 Deviations from the mean** Sometimes it is convenient to write the simple linear regression model in a different form that is a little easier to manipulate. Taking Equation (2.1), and adding $\beta_1\bar{x} - \beta_1\bar{x}$, which equals 0, to the right-hand side, and combining terms, we can write

$$\begin{aligned} y_i &= \beta_0 + \beta_1\bar{x} + \beta_1x_i - \beta_1\bar{x} + e_i \\ &= (\beta_0 + \beta_1\bar{x}) + \beta_1(x_i - \bar{x}) + e_i \\ &= \alpha + \beta_1(x_i - \bar{x}) + e_i \end{aligned} \tag{2.24}$$

where we have defined $\alpha = \beta_0 + \beta_1\bar{x}$. This is called the deviations from the sample mean form for simple regression.

- 2.8.1** What is the meaning of the parameter α ?
2.8.2 Show that the least squares estimates are

$$\hat{\alpha} = \bar{y}, \quad \hat{\beta}_1 \text{ as given by (2.5)}$$

- 2.8.3** Find expressions for the variances of the estimates and the covariance between them.

2.9 Invariance

- 2.9.1** In the simple regression model (2.1), suppose the value of the predictor X is replaced by $Z = aX + b$, where $a \neq 0$ and b are constants. Thus, we are considering 2 simple regression models,

$$\text{I: } E(Y|X = x) = \beta_0 + \beta_1 x$$

$$\text{II: } E(Y|Z = z) = \gamma_0 + \gamma_1 z = \gamma_0 + \gamma_1(ax + b)$$

Find the relationships between β_0 and γ_0 ; between β_1 and γ_1 ; between the estimates of variance in the 2 regressions, and between the t -tests of $\beta_1 = 0$ and of $\gamma_1 = 0$.

- 2.9.2** Suppose each value of the response Y is replaced by $V = dY$, for some $d \neq 0$, so we consider the two regression models

$$\text{I: } E(Y|X=x) = \beta_0 + \beta_1 x$$

$$\text{III: } E(V|X=x) = \delta_0 + \delta_1 x$$

Find the relationships between β_0 and δ_0 ; between β_1 and δ_1 ; between the estimates of variance in the 2 regressions, and between the t -tests of $\beta_1 = 0$ and of $\delta_1 = 0$.

- 2.10 Two-sample tests** One of the basic problems in elementary statistics is testing for equality of two means. If $\bar{y}_j, j = 0, 1$, are the sample means, the sample sizes are $m_j, j = 0, 1$, and the sample standard deviations are $SD_j, j = 0, 1$, then under the assumption that sample j is $NID(\mu_j, \sigma^2)$, the statistic

$$t = \frac{\bar{y}_1 - \bar{y}_0}{\hat{\sigma} \sqrt{1/m_0 + 1/m_1}} \quad (2.25)$$

with $\hat{\sigma}^2 = [(m_0 - 1)SD_0^2 + (m_1 - 1)SD_1^2]/[m_0 + m_1 - 2]$ is used to test $\mu_0 = \mu_1$ against a general alternative. Under normality and the assumptions of equal variance in each population, the null distribution is $t \sim t(m_0 + m_1 - 2)$.

For simplicity assume $m_0 = m_1 = m$, although the results do not depend on the equal sample sizes. Define a predictor X with values $x_i = 0$ for $i = 1, \dots, m$ and $x_i = 1$ for $i = m+1, \dots, 2m$. Combine the response y_i into a vector of length $2m$, the first m observations corresponding to population 0 and the remaining to population 1. In this problem we will fit the simple linear regression model (2.1) for this X and Y , and show that it is equivalent to the two-sample problem.

- 2.10.1** Show that $\bar{y} = (\bar{y}_0 + \bar{y}_1)/2$, $\bar{x} = 1/2$, $S_{XX} = m/2$, and $S_{XY} = m(\bar{y}_1 - \bar{y}_0)/2$.
- 2.10.2** Give the formulas for the OLS estimates of β_0 and β_1 in the simple linear regression model with the Y and X as specified in this problem. Interpret the estimates.
- 2.10.3** Find the fitted values and the residuals. Give an expression for RSS obtained by squaring and adding up the residuals and then dividing by the df .
- 2.10.4** Show that the t -statistic for testing $\beta_1 = 0$ is exactly the same as the usual two-sample t -test for comparing two groups with an assumption of equal within-group variance.

- 2.10.5** The group indicator is set to $x_i = 0$ for one group and $x_i = 1$ for the other group. Suppose we used as the group indicator $x_i^* = -1$ for the first group and $x_i^* = +1$ for the second group. How will this change the estimates of β_0 and β_1 and the meaning of the test that $\beta_1 = 0$? (*Hint:* Find values a and b such that $x_i^* = a(x_i + b)$, and then apply Problem 2.9.1.)
- 2.10.6** (Data file: cathedral) The datafile contains the Height and Length in feet of 25 cathedrals, nine in the Romanesque style, and 16 in the later Gothic style. Consider only the first 18 rows of this data file, which contain all the Romanesque cathedrals and nine of the Gothic cathedrals, and consider testing the hypothesis that the mean length is the same for Romanesque and Gothic cathedrals against the alterative that they are different. Use these data to verify all the results of the preceding sections of this problem. (*Hint:* In the data file the group indicator Type is a text variable with values Romanesque and Gothic that you may need to convert to zeros and ones. In R, for example, the statement

```
> cathedral$group <-
+   ifelse(cathedral$type=="Romanesque", 0, 1)
```

will do it. Don't forget to remove the last seven rows of the file, although if you do forget the test computed will still be a t -test of the hypothesis that the two types of cathedrals have the same mean height, but based on different data.)

2.11 The slope estimate as an average of pairwise slopes

- 2.11.1** Suppose we have a sample $(x_i, y_i), i = 1, \dots, n$. By completing the square show that

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n (x_i - x_j)^2 &= 2nSXX \\ \sum_{i=1}^n \sum_{j=1}^n (x_i - x_j)(y_i - y_j) &= 2nSXY \end{aligned}$$

- 2.11.2** Given any 2 points (x_i, y_i) and (x_j, y_j) , the slope of the line joining the first point to the second point is:

$$b_{ij} = \frac{y_i - y_j}{x_i - x_j}$$

(To allow for data sets with repeated values of the x_i , define $b_{ii} = 0$ if $x_i = x_j$.) Show that the OLS estimate $\hat{\beta}_1 = SXY/SXX$ is a weighted combination of the b_{ij} ,

$$\hat{\beta}_1 = \sum_{i=1}^n \sum_{j=1}^n w_{ij} b_{ij}$$

where the weights $w_{ij} = (x_i - \bar{x}_j)^2 / (2nS_{XX})$. This weighting scheme gives larger weight to pairs of points that are widely separated, $(x_i - \bar{x}_j)^2$ is large, and less weight to pairs that are close together.

- 2.12 The t -test for slope as a function of the correlation** Show that the t -statistic for testing the slope $\beta_1 = 0$ can be written as a function of sample size and the sample correlation r_{xy} ,

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{XX}}} = \sqrt{n-2} \frac{r_{xy}}{\sqrt{1-r_{xy}^2}} \quad (2.26)$$

2.13 Heights of mothers and daughters (Data file: `Heights`)

- 2.13.1** Compute the regression of `dheight` on `mheight`, and report the estimates, their standard errors, the value of the coefficient of determination, and the estimate of variance. Write a sentence or two that summarizes the results of these computations.
- 2.13.2** Obtain a 99% confidence interval for β_1 from the data.
- 2.13.3** Obtain a prediction and 99% prediction interval for a daughter whose mother is 64 inches tall.

- 2.14 Average prediction error** (Data file: `Heights`) In many problems, the analyst may wish to characterize the average prediction error for a regression model either to describe the accuracy that could be expected for predictions of future values in general, or possibly to help choose between competing regression models, as will be discussed in Chapter 10. If sufficient data are available, a simple *cross-validation scheme* can be used. We divide the data into two parts, a *construction set* to be used to estimate coefficients, and a *validation set* used to test the accuracy of the prediction equation.
- 2.14.1** Using the `Heights` data, create a construction set by selecting approximately 2/3 of the rows of the data file at random. The remaining 1/3 of the rows will comprise the validation set.
- 2.14.2** Obtain predictions from the model fit to the construction set for the values of `Mheight` in the validation set. Compute and report the average squared residual. The square root of this quantity is an estimate of the average prediction error.
- 2.14.3** As an alternative to cross-validation in this problem, use the fitted model based on the construction set to obtain predictions and the standard error of prediction (2.17) for each of the rows in the validation set. Then compute the average squared prediction

error and its square root. Compare with the results of the last subproblem.

2.15 Smallmouth bass (Data file: `wblake`)

- 2.15.1** Using the West Bearskin Lake smallmouth bass data in the file `wblake`, obtain 95% intervals for the mean length at ages 2, 4, and 6 years.
- 2.15.2** Obtain a 95% interval for the mean length at age 9. Explain why this interval is likely to be untrustworthy.

2.16 United Nations data (Data file: `UN11`)

Refer to the UN data in Problem 1.1.

- 2.16.1** Use a software package to compute the simple linear regression model corresponding to the graph in Problem 1.1.3.
- 2.16.2** Draw a graph of $\log(\text{fertility})$ versus $\log(\text{ppgdp})$, and add the fitted line to the graph.
- 2.16.3** Test the hypothesis that the slope is 0 versus the alternative that it is negative (a one-sided test). Give the significance level of the test and a sentence that summarizes the result.
- 2.16.4** Give the value of the coefficient of determination, and explain its meaning.
- 2.16.5** For a locality not in the data with $\text{ppgdp} = 1000$, obtain a point prediction and a 95% prediction interval for $\log(\text{fertility})$. If the interval (a, b) is a 95% prediction interval for $\log(\text{fertility})$, then a 95% prediction interval for fertility is given by $(\exp(a), \exp(b))$. Use this result to get a 95% prediction interval for fertility .
- 2.16.6** Identify (1) the locality with the highest value of fertility ; (2) the locality with the lowest value of fertility ; and (3) the two localities with the largest positive residuals from the regression when both variables are in log scale, and the two countries with the largest negative residuals in log scales.

2.17 Regression through the origin Occasionally, a mean function in which the intercept is known a priori to be 0 may be fit. This mean function is given by

$$E(y|x) = \beta_1 x \quad (2.27)$$

The residual sum of squares for this model, assuming the errors are independent with common variance σ^2 , is $\text{RSS} = \sum (y_i - \hat{\beta}_1 x_i)^2$.

- 2.17.1** Show that the least squares estimate of β_1 is $\hat{\beta}_1 = \sum x_i y_i / \sum x_i^2$. Show that $\hat{\beta}_1$ is unbiased and that $\text{Var}(\hat{\beta}_1 | X) = \sigma^2 / \sum x_i^2$. Find an expression for $\hat{\sigma}^2$. How many df does it have?

2.17.2 (Data file: `snake`) The data file gives X = water content of snow on April 1 and Y = water yield from April to July in inches in the Snake River watershed in Wyoming for $n = 17$ years from 1919 to 1935 (Wilm, 1950). Fit a regression through the origin and find $\hat{\beta}_1$ and σ^2 . Obtain a 95% confidence interval for β_1 . Test the hypothesis that the slope $\beta_1 = 0.49$, against the alternative that $\beta_1 > 0.49$.

2.17.3 Plot the residuals versus the fitted values, and comment on the adequacy of the mean function with 0 intercept. In regression through the origin, $\sum \hat{e}_i \neq 0$.

2.18 Using Appendix A.3, verify Equation (2.8).

2.19 Zipf's law (Data file: `MWwords`) Suppose we counted the number of times each word was used in the written works by Shakespeare, Alexander Hamilton, or some other author with a substantial written record. Can we say anything about the frequencies of the most common words?

Suppose we let f_i be the rate per 1000 words of text for the i th most frequent word used. The linguist George Zipf (1902–1950) observed a law-like relationship between rate f_i and rank i (Zipf, 1949),

$$E(f_i|i) = \alpha/i^\gamma$$

and further observed that the exponent γ is close to 1. Taking logarithms of both sides, we get approximately

$$E(\log(f_i)|\log(i)) = \log(\alpha) - \gamma \log(i) \quad (2.28)$$

Zipf's law has been applied to frequencies of many other classes of objects besides words, such as the frequency of visits to web pages on the Internet and the frequencies of species of insects in an ecosystem.

The data file gives the frequencies of 165 common words like “the,” “of,” “to,” and “which,” in works from four sources: the political writings of eighteenth-century American political figures Alexander Hamilton, James Madison, and John Jay, and the book *Ulysses* by twentieth-century Irish writer James Joyce. The data are from Mosteller and Wallace (1964, table 8.1-1). Several missing values occur in the data; these are really words that were used so infrequently that their count was not reported in Mosteller and Wallace's table.

2.19.1 Using only the 50 most frequent words in Hamilton's work (i.e., using only rows in the data for which `HamiltonRank` ≤ 50), draw the appropriate summary graph, estimate the mean function (2.28), and summarize your results. The response variable should

be Hamilton, the frequency with which Hamilton used a word, and the predictor is `HamiltonRank`, the rank of that word among the words that Hamilton used.

- 2.19.2** Test the hypothesis that $\gamma = 1$ against the two-sided alternative in (2.28) and summarize.
- 2.19.3** Repeat Problem 2.19.1, but for words with rank of 75 or less, and with rank less than 100. For larger number of words, Zipf's law may break down. Does that seem to happen with these data?

2.20 Old Faithful (Data file: `oldfaith`) Use the data from Problem 1.4.

- 2.20.1** Use simple linear regression methodology to obtain a prediction equation for `interval` from `duration`. Summarize your results in a way that might be useful for the nontechnical personnel who staff the Old Faithful Visitor's Center.
- 2.20.2** An individual has just arrived at the end of an eruption that lasted 250 seconds. Give a 95% confidence interval for the time the individual will have to wait for the next eruption.
- 2.20.3** Estimate the 0.90 quantile of the conditional distribution of

$$\text{interval}(\text{duration} = 250)$$

assuming that the population is normally distributed.

2.21 Windmills (Data file: `wm1`) Energy can be produced from wind using windmills. Choosing a site for a wind farm, the location of the windmills, can be a multimillion dollar gamble. If wind is inadequate at the site, then the energy produced over the lifetime of the wind farm can be much less than the cost of building and operation. Prediction of long-term wind speed at a candidate site can be an important component in the decision to build or not to build. Since energy produced varies as the square of the wind speed, even small errors can have serious consequences.

The data in the file `wm1` provides measurements that can be used to help in the prediction process. Data were collected every 6 hours for the year 2002, except that the month of May 2002 is missing. The values `Cspd` are the calculated wind speeds in meters per second at a candidate site for building a wind farm. These values were collected at a tower erected on the site. The values `RSpd` are wind speeds at a *reference site*, which is a nearby location for which wind speeds have been recorded over a very long time period. Airports sometimes serve as reference sites, but in this case, the reference data comes from the National Center for Environmental Modeling (NCAR, 2013). The reference is about 50 km southwest of the candidate site. Both sites are in the northern part of South Dakota.

The data were provided by Mark Ahlstrom and Rolf Miller of WindLogics.

- 2.21.1** Draw the scatterplot of the response $CSpd$ versus the predictor $RSpd$. Is the simple linear regression model plausible for these data?
- 2.21.2** Fit the simple regression of the response on the predictor, and present the appropriate regression summaries.
- 2.21.3** Obtain a 95% prediction interval for $CSpd$ at a time when $RSpd = 7.4285$.
- 2.21.4** Using generic notation, let $x = RSpd$, $y = CSpd$ and let n be the number of cases used in the regression ($n = 1116$ in the data we have used in this problem) and \bar{x} and Sxx defined from these n observations. Suppose we want to make predictions at m time points with values of wind speed x_{*1}, \dots, x_{*m} that are different from the n cases used in constructing the prediction equation. Show that (1) the average of the m predictions is equal to the prediction taken at the average value \bar{x}_* of the m values of the predictor, and (2) using the first result, the standard error of the average of m predictions is

$$\text{se of average prediction} = \sqrt{\frac{\hat{\sigma}^2}{m} + \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(\bar{x}_* - \bar{x})^2}{Sxx} \right)} \quad (2.29)$$

If m is very large, then the first term in the square root is negligible, and the standard error of average prediction is essentially the same as the standard error of a fitted value at \bar{x}_* .

- 2.21.5** For the period from January 1, 1948, to July 31, 2003, a total of $m = 62,039$ wind speed measurements are available at the reference site, excluding the data from the year 2002. For these measurements, the average wind speed was $\bar{x}_* = 7.4285$. Give a 95% prediction interval on the long-term average wind speed at the candidate site. This long-term average of the past is then taken as an estimate of the long-term average of the future and can be used to help decide if the candidate is a suitable site for a wind farm.

C H A P T E R 3

Multiple Regression

Multiple linear regression generalizes the simple linear regression model by allowing for many *regressors* in a mean function. We start with adding just a regressor to the simple regression mean function because the ideas generalize to adding many regressors.

3.1 ADDING A REGRESSOR TO A SIMPLE LINEAR REGRESSION MODEL

We start with a response Y and the simple linear regression mean function

$$E(Y|X_1 = x_1) = \beta_0 + \beta_1 x_1$$

Now suppose we have a second variable X_2 and would like to learn about the simultaneous dependence of Y on X_1 and X_2 . By adding X_2 to the problem, we will get a mean function that depends on both the value of X_1 and the value of X_2 ,

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad (3.1)$$

The main idea in adding X_2 is *to explain the part of Y that has not already been explained by X_1 .*

United Nations Data

We will use the United Nations data discussed in Problem 1.1. To the regression with response `lifeExpF` and regressor `log(ppgdp)` we consider adding `fertility`, the average number of children per woman. Interest therefore centers on the distribution of `log(lifeExpF)` as `log(ppgdp)` and `fertility` both vary. The data are in the file `UN11`.

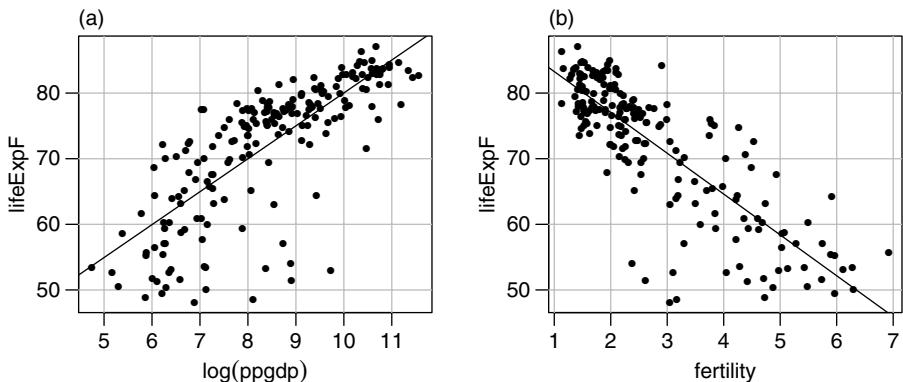


Figure 3.1 United Nations data on 199 localities, mostly nations: (a) `lifeExpF` versus `log(ppgdp)`; (b) `lifeExpF` versus `fertility`.

Figure 3.1a is a summary graph for the simple regression of `lifeExpF` on `log(ppgdp)`. This graph can also be called a *marginal plot* because it ignores all other regressors. The fitted mean function to the marginal plot using OLS is

$$\hat{E}(\text{lifeExpF}|\text{log}(ppgdp)) = 29.815 + 5.019 \text{log}(ppgdp) \quad (3.2)$$

with $R^2 = 0.596$, so about 60% of the variability in `lifeExpF` is explained by `log(ppgdp)`. Expected `lifeExpF` increases as `log(ppgdp)` increases.

Similarly, Figure 3.1b is the marginal plot for the regression of `lifeExpF` on `fertility`. This simple regression has fitted mean function

$$\hat{E}(\text{lifeExpF}|fertility) = 89.481 - 6.224 \text{fertility}$$

with $R^2 = 0.678$, so `fertility` explains about 68% of the variability in `lifeExpF`. Expected `lifeExpF` decreases as `fertility` increases. Thus, from Figure 3.1a, the response `lifeExpF` is related to the regressor `log(ppgdp)` ignoring `fertility`, and from Figure 3.1b, `lifeExpF` is related to `fertility` ignoring `log(ppgdp)`.

If the regressors `log(ppgdp)` and `fertility` were uncorrelated, then the marginal plots shown in Figure 3.1 would provide a complete summary of the dependence of the response on the regressors, as the effect of `fertility` adjusted for `log(ppgdp)` would be the same as the effect of `fertility` ignoring `log(ppgdp)`. Figure 3.2 is a plot of the regressors. Countries with larger `log(ppgdp)` also tend to have lower `fertility` and so these variables are negatively correlated. The regressors will in part be explaining the same variation.¹

¹There are a few localities with relatively large `log(ppgdp)` that have higher values of `fertility` than would be expected by the overall trend in Figure 3.2, and perhaps also have relatively low `lifeExpF` from Figure 3.1a. Can you identify these localities and what they have in common (Problem 3.1)?

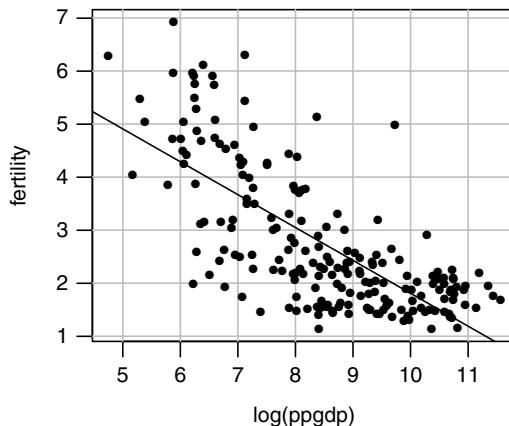


Figure 3.2 Marginal plot of `fertility` versus `log(ppgdp)`.

3.1.1 Explaining Variability

Given these graphs, what can be said about the proportion of variability in `lifeExpF` explained jointly by `log(ppgdp)` and `fertility`? The total explained variation must be at least 67.8%, the larger of the variation explained by each variable separately, since using both `log(ppgdp)` and `fertility` must surely be at least as informative as using just one of them. If the regressors were uncorrelated, then the variation explained by them jointly would equal the sum of the variations explained individually. In this example, the sum of the individual variations explained exceeds 100%, $59.6\% + 67.8\% = 127.4\%$. As confirmed by Figure 3.2, the regressors are correlated so this simple addition formula won't apply. The variation explained by both variables can be smaller than the sum of the individual variation explained if the regressors are in part explaining the same variation. The total can exceed the sum if the variables act jointly so that knowing both gives more information than knowing just one of them. For example, the area of a rectangle may be only poorly determined by either the length or width alone, but if both are considered at the same time, area can be determined exactly. It is precisely this inability to predict the joint relationship from the marginal relationships that makes multiple regression rich and complicated.

3.1.2 Added-Variable Plots

To get the effect of adding `fertility` to the model that already includes `log(ppgdp)`, we need to examine the part of the response `lifeExpF` not explained by `log(ppgdp)` and the part of the new regressor `fertility` not explained by `log(ppgdp)`.

1. Compute the regression of the response `lifeExpF` on the first regressor `log(ppgdp)`, corresponding to the OLS line shown in Figure 3.1a. The

fitted equation is given at (3.2). Keep the residuals from this regression. These residuals are *the part of the response lifeExpF not explained by the regression on log(ppgdp)*.

2. Compute the regression of *fertility* on $\log(\text{ppgdp})$, corresponding to Figure 3.2. Keep the residuals from this regression as well. These residuals are *the part of the new regressor fertility not explained by $\log(\text{ppgdp})$* .
3. The *added-variable plot* is of the unexplained part of the response from (1) on the unexplained part of the added regressor from (2).

The added-variable plot is shown in Figure 3.3a. It summarizes the relationship between *lifeExpF* and *fertility* *adjusting* for $\log(\text{ppgdp})$, while Figure 3.3b, repeated for convenience from Figure 3.1b, shows this relationship but *ignoring* $\log(\text{ppgdp})$. If Figure 3.3a shows a stronger relationship than does Figure 3.3b, meaning that the points in the plot show less variation about the fitted straight line, then the two variables act jointly to explain extra variation. If the two graphs have similar variation, then the total explained variability by both variables is less than the additive amount. The latter is the case here.

If we fit the simple regression mean function to Figure 3.3a, the fitted line has 0 intercept, since the averages of the plotted variables are 0, and the estimated slope via OLS is $\hat{\beta}_2 = -4.199$. It turns out that this is exactly the estimate $\hat{\beta}_2$ that would be obtained using OLS to get the estimates using the mean function (3.1) with both regressors. The proportion of variability explained in this plot is 0.367, which is the *square of the partial correlation between lifeExpF and fertility adjusted for $\log(\text{ppgdp})$* . Thus, adding *fertility* explains 36.7% of the remaining variability in *lifeExpF* after adjusting for $\log(\text{ppgdp})$.

We now have two estimates of the coefficient β_2 for *fertility*:

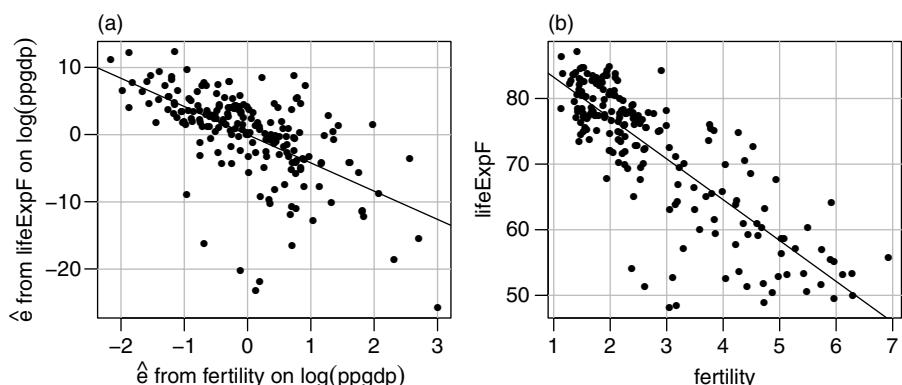


Figure 3.3 (a) Added-variable plot for *fertility* after $\log(\text{ppgdp})$. (b) The marginal plot of *lifeExpF* versus *fertility* ignoring $\log(\text{ppgdp})$, repeated from Figure 3.1b.

$$\hat{\beta}_2 = -6.224 \text{ ignoring } \log(\text{ppgdp})$$

$$\hat{\beta}_2 = -4.199 \text{ adjusting for } \log(\text{ppgdp})$$

The slope in the added-variable plot is about 30% smaller than the slope in the plot that ignores $\log(\text{ppgdp})$, although in this instance, after adjusting for $\log(\text{ppgdp})$, the effect of fertility is still important. The regressor `fertility` is useful after adjusting for $\log(\text{ppgdp})$.

To get the coefficient estimate for $\log(\text{ppgdp})$ in the regression of `lifeExpF` on both regressors, we would use the same procedure we used for `fertility` and consider the problem of adding $\log(\text{ppgdp})$ to a mean function that already includes `fertility`. This would require looking at the graph of the residuals from the regression of `lifeExpF` on `fertility` versus the residuals from the regression of $\log(\text{ppgdp})$ on `fertility` (see Problem 3.2).

3.2 THE MULTIPLE LINEAR REGRESSION MODEL

The general multiple linear regression model with response Y and regressors X_1, \dots, X_p will have the form

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p \quad (3.3)$$

The symbol X in $E(Y|X)$ means that we are conditioning on all the regressors on the right side of the equation. When we are conditioning on specific values for the predictors x_1, \dots, x_p that we will collectively call \mathbf{x} , we write

$$E(Y|X = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (3.4)$$

As in Chapter 2, the β s are unknown parameters to be estimated. When $p = 1$, X has only one element, and we get the simple regression problem discussed in Chapter 2. When $p = 2$, the mean function (3.3) corresponds to a plane in 3 dimensions. When $p > 2$, the fitted mean function is a *hyperplane*, the generalization of a p -dimensional plane in a $(p + 1)$ -dimensional space. We cannot draw a general p -dimensional plane in our three-dimensional world.

3.3 PREDICTORS AND REGRESSORS

Regression problems start with a collection of potential predictors. Some of these may be continuous measurements, like the height or weight of an object. Some may be discrete but ordered, like a doctor's rating of overall health of a patient on a nine-point scale. Other potential predictors can be categorical, like eye color or an indicator of whether a particular unit received a treatment.

All these types of potential predictors can be useful in multiple linear regression.

From the pool of potential predictors, we create a set of *regressors*² that are the X -variables that appear in (3.3). The regressors might include

The intercept Suppose we define $\mathbf{1}$ to be a regressor that is always equal to 1. The mean function (3.3) can be rewritten as

$$E(Y|X) = \beta_0 \mathbf{1} + \beta_1 X_1 + \cdots + \beta_p X_p$$

Mean functions without an intercept would not have this regressor included. In most computer programs, an intercept is included unless it is specifically suppressed.

Predictors The simplest type of regressor is equal to a predictor, for example, the variable `mheight` in the heights data or `fertility` in the UN data.

Transformations of predictors Sometimes the original predictors need to be transformed in some way to make (3.3) hold to a reasonable approximation. This was the case in the UN data in which `ppgdp` was used in log scale. The willingness to replace predictors by transformations of them greatly expands the range of problems that can be summarized with a linear regression model.

Polynomials Problems with curved mean functions can sometimes be accommodated in the multiple linear regression model by including polynomial regressors in the predictor variables. For example, we might include as regressors both a predictor X_1 and its square X_1^2 to fit a quadratic polynomial in that predictor. Complex polynomial surfaces in several predictors can be useful in some problems, as will be discussed in Section 5.3.³

Interactions and other combinations of predictors Combining several predictors is often useful. An example of this is using body mass index, given by weight in kilograms divided by height in meters squared, in place of both height and weight, or using a total test score in place of the separate scores from each of several parts. Products of regressors called *interactions* are often included in a mean function along with the base regressors to allow for joint effects.

Dummy variables and factors A categorical predictor with two or more levels is called a *factor*. Factors are included in multiple linear regression

²In the third edition of this book, the word *terms* was used for the variables called *regressors* in this edition. This change in notation is consistent with Fox and Weisberg (2011).

³This discussion of polynomials might puzzle some readers because in Section 3.2, we said the linear regression mean function was a hyperplane, but here we have said that it might be curved, seemingly a contradiction. However, *both* of these statements are correct. If we fit a mean function like $E(Y|X=x) = \beta_0 + \beta_1 x + \beta_2 x^2$, the mean function is a quadratic curve in the plot of the response versus X but a plane in the three-dimensional plot of the response versus X and X^2 .

using *dummy variables*, which are typically regressors that have only two values, often 0 and 1, indicating which category is present for a particular observation. We will see in Chapter 5 that a categorical predictor with two categories can be represented by one dummy variable, while a categorical predictor with many categories can require several dummy variables.

Regression splines Polynomials represent the effect of a predictor by using a sum of regressors, like $\beta_1x + \beta_2x^2 + \beta_3x^3$. We can view this as a linear combination of basis functions, given in the polynomial case by the functions $\{x, x^2, x^3\}$. Using *splines* is similar to fitting a polynomial, except we use different basis functions that can have useful properties under some circumstances. We return to the use of splines in Section 5.4.

Principal components In some problems we may have a large number of predictors that are thought to be related. For example, we could have predictors that correspond to the amount of a particular drug that is present in repeated samples on the same subject. Suppose X_1, \dots, X_m are m such predictors. For clarity, we may wish to replace these m predictors by a single regressor $Z = \sum a_j X_j$ where Z summarizes the information in the multiple indicators as fully as possible. One way to do this is to set all the $a_j = 1/m$, and then Z is just the average of the X_j . Alternatively, the a_j s can be found that satisfy some criterion, such as maximizing the variance of Z . This leads to the use of principal components as predictors, as described in Section 5.5.

A regression with k predictors may combine into fewer than k regressors or expand to require more than k regressors. The distinction between predictors and regressors can be very helpful in thinking about an appropriate mean function to use in a particular problem, and in using graphs to understand a problem. For example, a regression with 1 predictor can always be studied using the 2D scatterplot of the response versus the predictor, regardless of the number of regressors required in the mean function.

We will use the fuel consumption data introduced in Section 1.6 as the primary example for the rest of this chapter. As discussed earlier, the goal is to understand how fuel consumption varied (in 2001!) as a function of state characteristics. The variables were defined in Table 1.1 and are given in the file `fuel2001`. From the six initial predictors, we define regressors in the regression mean function.

Basic summary statistics for the relevant variables in the fuel data are given in Table 3.1, and these begin to give us a picture of these data. First, there is quite a bit of variation in `Fuel`, with values between a minimum of about 317 gal./year and a maximum of about 843 gal./year. The `gas Tax` varies from only 8 cents/gal. to a high of 29 cents/gal., so unlike much of the world, gasoline taxes account for only a small part of the cost to consumers of gasoline. Also of interest is the range of values in `Dlic`: The number of licensed drivers per 1000 population over the age of 16 is between about 700 and 1075. Some states appear to have more licensed drivers than they have population over age 16. Either these states allow drivers under the age of 16, allow nonresidents to

Table 3.1 Summary Statistics for the Fuel Data

	N	Average	Std Dev	Min	Max
Tax	51	20.15	4.54	7.50	29.00
Dlic	51	903.68	72.86	700.20	1075.29
Income	51	28.40	4.45	20.99	40.64
log(Miles)	51	10.91	1.03	7.34	12.61
Fuel	51	613.13	88.96	317.49	842.79

obtain a driver's license, or the data are in error. For this example, we will assume one of the first two reasons.

Of course, these univariate summaries cannot tell us much about how fuel consumption depends on the other variables. For this, graphs are very helpful. The scatterplot matrix for the fuel data is repeated in Figure 3.4. From our previous discussion, `Fuel` decreases on the average as `Tax` increases, but there is a lot of variation. We can make similar qualitative judgments about each of the regressions of `Fuel` on the other variables. The overall impression is that `Fuel` is at best weakly related to each of the variables in the scatterplot matrix, and in turn, these variables are only weakly related to each other.

Does this help us understand how `Fuel` is related to all four regressors simultaneously? We know from the discussion in Section 3.1 that the marginal relationships between the response and each of the variables is *not* sufficient to understand the *joint* relationship between the response and the regressors. The interrelationships among the regressors are also important. The pairwise relationships between the regressors can be viewed in the remaining cells of the scatterplot matrix. In Figure 3.4, the relationships between all pairs of regressors appear to be very weak, suggesting that for this problem the marginal plots including `Fuel` are quite informative about the multiple regression problem.

A more traditional, and less informative, summary of the two-variable relationships is the matrix of sample correlations, shown in Table 3.2. In this instance, the correlation matrix helps to reinforce the relationships we see in the scatterplot matrix, with fairly small correlations between the predictors and `Fuel`, and essentially no correlation between the predictors themselves.

3.4 ORDINARY LEAST SQUARES

From the initial collection of potential predictors, we have computed a set of $p + 1$ regressors, including an intercept, $X = (1, X_1, \dots, X_p)$. The mean function and variance function for multiple linear regression are

$$\begin{aligned} E(Y|X) &= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \\ \text{Var}(Y|X) &= \sigma^2 \end{aligned} \tag{3.5}$$

Both the β s and σ^2 are unknown parameters that are to be estimated.

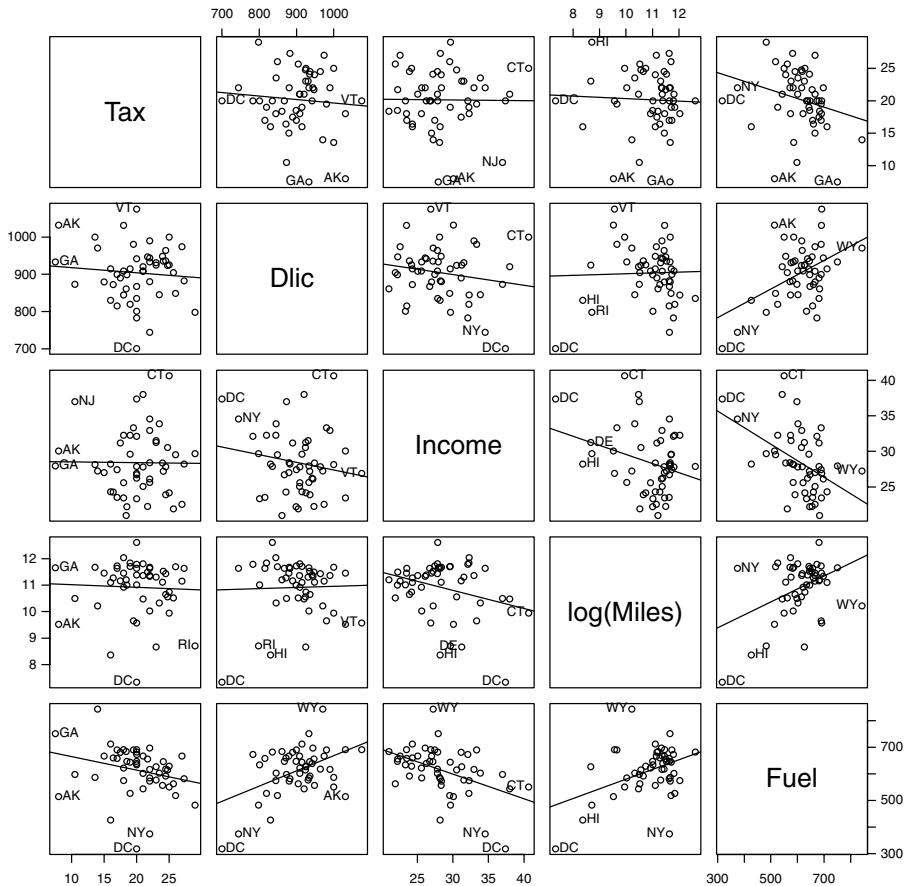


Figure 3.4 Scatterplot matrix for the fuel data.

Table 3.2 Sample Correlations for the Fuel Data

	Tax	Dlic	Income	log (Miles)	Fuel
Tax	1.0000	-0.0858	-0.0107	-0.0437	-0.2594
Dlic	-0.0858	1.0000	-0.1760	0.0306	0.4685
Income	-0.0107	-0.1760	1.0000	-0.2959	-0.4644
log (Miles)	-0.0437	0.0306	-0.2959	1.0000	0.4220
Fuel	-0.2594	0.4685	-0.4644	0.4220	1.0000

3.4.1 Data and Matrix Notation

In this and the next few sections we use matrix notation as a compact way to describe data and perform manipulations of data. Appendix A.6 contains a brief introduction to matrices and linear algebra that some readers may find helpful.

Suppose we have observed data for n cases or units, meaning we have a value of Y and all of the regressors for each of the n cases. We define

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \quad (3.6)$$

so \mathbf{Y} is an $n \times 1$ vector and \mathbf{X} is an $n \times (p + 1)$ matrix. The i th row of \mathbf{X} will be defined by the symbol \mathbf{x}_i' , which is a $(p + 1) \times 1$ vector for mean functions that include an intercept. Even though \mathbf{x}_i is a row of \mathbf{X} , we use the convention that all vectors are column vectors and therefore need to include the transpose on \mathbf{x}_i' to represent a row. The first few and the last few rows of the matrix \mathbf{X} and the vector \mathbf{Y} for the fuel data are

$$\mathbf{X} = \begin{pmatrix} 1 & 18.00 & 1031.38 & 23.471 & 16.5271 \\ 1 & 8.00 & 1031.64 & 30.064 & 13.7343 \\ 1 & 18.00 & 908.597 & 25.578 & 15.7536 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 25.65 & 904.894 & 21.915 & 15.1751 \\ 1 & 27.30 & 882.329 & 28.232 & 16.7817 \\ 1 & 14.00 & 970.753 & 27.230 & 14.7362 \end{pmatrix} \quad \mathbf{Y} = \begin{pmatrix} 690.264 \\ 514.279 \\ 621.475 \\ \vdots \\ 562.411 \\ 581.794 \\ 842.792 \end{pmatrix}$$

The first row of \mathbf{X} is $\mathbf{x}_1' = (1, 18.00, 1031.38, 23.471, 16.5271)'$, and the first row of \mathbf{Y} is $y_1 = 690.264$, an ordinary number or scalar. The regressors in \mathbf{X} are in the order intercept, Tax, Dlic, Income, and finally, log(Miles). The matrix \mathbf{X} is 51×5 and \mathbf{Y} is 51×1 .

Next, define $\boldsymbol{\beta}$ to be a $(p + 1) \times 1$ vector of unknown regression coefficients,

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$$

An equation for the mean function evaluated at \mathbf{x}_i is

$$\begin{aligned} E(Y|X = \mathbf{x}_i) &= \mathbf{x}_i' \boldsymbol{\beta} \\ &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \end{aligned} \quad (3.7)$$

and the mean function in matrix terms is

$$E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \quad (3.8)$$

where \mathbf{Y} is the vector of responses, and \mathbf{X} is the $n \times (p + 1)$ matrix whose i th row is \mathbf{x}'_i .

3.4.2 The Errors \mathbf{e}

Define the unobservable random vector of errors \mathbf{e} elementwise by $e_i = y_i - E(Y|X = \mathbf{x}_i) = y_i - \mathbf{x}'_i\boldsymbol{\beta}$, and $\mathbf{e} = (e_1, \dots, e_n)'$. The assumptions concerning the e_i s given in Chapter 2 are summarized in matrix form as

$$E(\mathbf{e}|X) = \mathbf{0} \quad \text{Var}(\mathbf{e}|X) = \sigma^2 \mathbf{I}_n$$

where $\text{Var}(\mathbf{e}|X)$ means the covariance matrix of \mathbf{e} for a fixed value of X , \mathbf{I}_n is the $n \times n$ matrix with ones on the diagonal and zeroes everywhere else, and $\mathbf{0}$ is a matrix or vector of zeroes of appropriate size. If we add the assumption of normality, we can write

$$(\mathbf{e}|X) \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

3.4.3 Ordinary Least Squares Estimators

The least squares estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is chosen to minimize the residual sum of squares function

$$\text{RSS}(\boldsymbol{\beta}) = \sum (y_i - \mathbf{x}'_i\boldsymbol{\beta})^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \quad (3.9)$$

The OLS estimates can be found from (3.9) by differentiation in a matrix analog to the development of Appendix A.3. The OLS estimate is given by the formula

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (3.10)$$

provided that the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ exists.⁴ The estimator $\hat{\boldsymbol{\beta}}$ depends only on the sufficient statistics $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{Y}$, which are matrices of uncorrected sums of squares and cross products.

Do not compute the least squares estimates using (3.10)! Uncorrected sums of squares and cross products are prone to large rounding error, and so

⁴Practical methods for problems for which this inverse does not exist are discussed in Section 4.1.4; theoretical discussions can be found in any book on linear models such as Christensen (2011).

computations can be highly inaccurate. The preferred computational methods are based on matrix decompositions as briefly outlined in Appendices A.9 and A.10. At the very least, computations should be based on *corrected* sums of squares and cross products. Suppose we define \mathcal{X} to be the $n \times p$ matrix

$$\mathcal{X} = \begin{pmatrix} (x_{11} - \bar{x}_1) & \cdots & (x_{1p} - \bar{x}_p) \\ (x_{21} - \bar{x}_1) & \cdots & (x_{2p} - \bar{x}_p) \\ \vdots & \vdots & \vdots \\ (x_{n1} - \bar{x}_1) & \cdots & (x_{np} - \bar{x}_p) \end{pmatrix}$$

This matrix consists of the original \mathbf{X} matrix, but with the first column removed and the column mean subtracted from each of the remaining columns. Similarly, \mathcal{Y} is the vector with typical elements $y_i - \bar{y}$. Then

$$\mathcal{C} = \frac{1}{n-1} \begin{pmatrix} \mathcal{X}'\mathcal{X} & \mathcal{X}'\mathcal{Y} \\ \mathcal{Y}'\mathcal{X} & \mathcal{Y}'\mathcal{Y} \end{pmatrix} \quad (3.11)$$

is the matrix of sample variances and covariances, and this is the summary of the data that is most often produced in regression software. When $p = 1$, the matrix \mathcal{C} is given by

$$\mathcal{C} = \frac{1}{n-1} \begin{pmatrix} S_{XX} & S_{XY} \\ S_{XY} & S_{YY} \end{pmatrix}$$

The elements of \mathcal{C} are the summary statistics needed for OLS computations in simple linear regression. If we let $\hat{\beta}^*$ be the parameter vector excluding the intercept β_0 , then for $p \geq 1$,

$$\begin{aligned} \hat{\beta}^* &= (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathcal{Y} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}^*\bar{x} \end{aligned} \quad (3.12)$$

where \bar{x} is the vector of sample means for all the regressors except for the intercept.

Once $\hat{\beta}$ is computed, we can define several related quantities. The fitted values are $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ and the residuals are $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}}$. The function (3.9) evaluated at $\hat{\beta}$ is the residual sum of squares, or RSS. Recognizing that $\mathcal{Y}'\mathcal{Y} = S_{YY}$,

$$\begin{aligned} \text{RSS} &= \hat{\mathbf{e}}'\hat{\mathbf{e}} \\ &= (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) \\ &= (\mathcal{Y} - \mathcal{X}\hat{\beta}^*)'(\mathcal{Y} - \mathcal{X}\hat{\beta}^*) \\ &= S_{YY} - \hat{\beta}^{**}(\mathcal{X}'\mathcal{X})\hat{\beta}^* \\ &= S_{YY} - SS_{\text{reg}} \end{aligned} \quad (3.13)$$

The last equation implicitly defines the *regression sum of squares* to be the difference between the total sum of squares SYY and the residual sum of squares RSS . Most regression software will provide two of these three quantities, and the third can be computed by subtraction.

3.4.4 Properties of the Estimates

Additional properties of the OLS estimates are derived in Appendix A.8 and are only summarized here. Assuming that $E(\mathbf{e}|X) = \mathbf{0}$ and $\text{Var}(\mathbf{e}|X) = \sigma^2 \mathbf{I}_n$, then $\hat{\beta}$ is unbiased, $E(\hat{\beta}|X) = \beta$, and

$$\text{Var}(\hat{\beta}|X) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \quad (3.14)$$

Excluding the intercept regressor,

$$\text{Var}(\hat{\beta}^*|X) = \sigma^2 (\mathcal{X}'\mathcal{X})^{-1} \quad (3.15)$$

and so $(\mathcal{X}'\mathcal{X})^{-1}$ is all but the first row and column of $(\mathbf{X}'\mathbf{X})^{-1}$. An estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{RSS}{n-(p+1)} \quad (3.16)$$

If \mathbf{e} is normally distributed, then the residual sum of squares has a chi-squared distribution,

$$\frac{n-(p+1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-(p+1))$$

By substituting $\hat{\sigma}^2$ for σ^2 in (3.14), we find the estimated variance of $\hat{\beta}$ to be

$$\widehat{\text{Var}}(\hat{\beta}|X) = \hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1} \quad (3.17)$$

3.4.5 Simple Regression in Matrix Notation

For simple regression, \mathbf{X} and \mathbf{Y} are given by

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

and thus

$$(\mathbf{X}'\mathbf{X}) = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \quad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

By direct multiplication, $(\mathbf{X}'\mathbf{X})^{-1}$ can be shown to be

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{S_{XX}} \begin{pmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \quad (3.18)$$

so that

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \frac{1}{S_{XX}} \begin{pmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \\ &= \begin{pmatrix} \bar{y} - \hat{\beta}_1 \bar{x} \\ S_{XY}/S_{XX} \end{pmatrix} \end{aligned}$$

as found previously. Also, since $\sum x_i^2/(nS_{XX}) = 1/n + \bar{x}^2/S_{XX}$, the variances and covariances for $\hat{\beta}_0$ and $\hat{\beta}_1$ found in Chapter 2 are identical to those given by $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

The results are simpler in the deviations from the sample mean form, since

$$\mathcal{X}'\mathcal{X} = S_{XX} \quad \mathcal{X}'\mathcal{Y} = S_{XY}$$

and

$$\begin{aligned} \hat{\beta}_1 &= (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathcal{Y} = \frac{S_{XY}}{S_{XX}} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

Fuel Consumption Data

We will generally let p equal the number of regressors in a mean function excluding the intercept, and $p' = p + 1$ equal if the intercept is included; $p' = p$ if the intercept is not included. We shall now fit the mean function with $p' = 5$ regressors, including the intercept for the fuel consumption data. The model can be specified using Wilkinson and Rogers (1973) notation,

$$\text{Fuel} \sim \text{Tax} + \text{Dlic} + \text{Income} + \log(\text{Miles}) \quad (3.19)$$

This is shorthand for using OLS to fit the multiple linear regression model with mean function

$$E(Fuel|X) = \beta_0 + \beta_1 Tax + \beta_2 Dlic + \beta_3 Income + \beta_4 \log(Miles) \quad (3.20)$$

where, as usual, conditioning on X is short for conditioning on all the regressors in the mean function. The intercept is not present in (3.19) but is included in (3.20) unless it is specifically excluded. See Section 5.1 for more discussion of this notation.

All the computations are based on the summary statistics, which are the sample means given in Table 3.1 and the sample covariance matrix \mathcal{C} defined at (3.11) and given by

	Tax	Dlic	Income	$\log(Miles)$	Fuel
Tax	20.6546	-28.4247	-0.2162	-0.2048	-104.8944
Dlic	-28.4247	5308.2591	-57.0705	2.2968	3036.5905
Income	-0.2162	-57.0705	19.8171	-1.3572	-183.9126
$\log(Miles)$	-0.2048	2.2968	-1.3572	1.0620	38.6895
Fuel	-104.8944	3036.5905	-183.9126	38.6895	7913.8812

Most statistical software will give the sample correlations rather than the covariances. The reader can verify that the correlations in Table 3.2 can be obtained from these covariances. For example, the sample correlation between Tax and Income is $-0.2162/\sqrt{(20.6546 \times 19.8171)} = -0.0107$ as in Table 3.2.

The 5×5 matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is given by

	Intercept	Tax	Dlic	Income	$\log(Miles)$
Intercept	9.02e+00	-2.85e-02	-4.08e-03	-5.98e-02	-2.79e-01
Tax	-2.85e-02	9.79e-04	5.60e-06	4.26e-05	2.31e-04
Dlic	-4.08e-03	5.60e-06	3.92e-06	1.19e-05	7.79e-06
Income	-5.98e-02	4.26e-05	1.19e-05	1.14e-03	1.44e-03
$\log(Miles)$	-2.79e-01	2.31e-04	7.79e-06	1.44e-03	2.07e-02

The elements of $(\mathbf{X}'\mathbf{X})^{-1}$ often differ by several orders of magnitude, as is the case here, where the smallest element in absolute value is $3.92 \times 10^{-6} = 0.00000392$, and the largest element is $9.02 \times 10^0 = 9.02$. It is the combining of these numbers of very different magnitude that can lead to numerical inaccuracies in computations. Matrices like this one are often displayed in scientific notation, which can be hard to read.

The lower-right 4×4 submatrix of $(\mathbf{X}'\mathbf{X})^{-1}$ is $(\mathbf{X}'\mathbf{X})^{-1}$. Using the formulas based on corrected sums of squares in this chapter, the estimate $\hat{\beta}^*$ is computed to be

$$\hat{\beta}^* = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{y} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{pmatrix} = \begin{pmatrix} -4.2280 \\ 0.4719 \\ -6.1353 \\ 26.7552 \end{pmatrix}$$

The estimated intercept is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}^{*'} \bar{x} = 154.1928$$

and the residual sum of squares is

$$RSS = \mathcal{Y}'\mathcal{Y} - \hat{\beta}^{*'}(\mathcal{X}'\mathcal{X})\hat{\beta}^* = 193,700$$

so the estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{RSS}{n - (p + 1)} = \frac{193,700}{51 - (4 + 1)} = 4,211$$

Estimated variances and covariances of the $\hat{\beta}_j$ are found by multiplying $\hat{\sigma}^2$ by the elements of $(\mathbf{X}'\mathbf{X})^{-1}$. Estimated standard errors are the square roots of the corresponding estimated variances. For example,

$$se(\hat{\beta}_2 | X) = \hat{\sigma} \sqrt{3.922 \times 10^{-6}} = 0.129$$

Virtually all statistical software packages include higher-level functions that will fit multiple regression models, but getting intermediate results like $(\mathbf{X}'\mathbf{X})^{-1}$ may be a challenge. Table 3.3 shows typical output from a statistical package. This output gives the estimates $\hat{\beta}$ and their standard errors computed based on $\hat{\sigma}^2$ and the diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$. The column marked t value is the ratio of the estimate to its standard error. The column labeled $Pr(>|t|)$ will be discussed shortly. Below the table are the estimated residual standard deviation $\hat{\sigma}$, its df discussed previously, and the coefficient of determination R^2 , also to be discussed shortly.

3.4.6 The Coefficient of Determination

Rearranging (3.13), the total sum of squares SYY can we written as

$$SYY = RSS + SS_{reg} \quad (3.21)$$

Table 3.3 Multiple Linear Regression Summary in the Fuel Data

	Estimate	Std. Error	t -Value	$Pr(> t)$
(Intercept)	154.1928	194.9062	0.79	0.4329
Tax	-4.2280	2.0301	-2.08	0.0429
Dlic	0.4719	0.1285	3.67	0.0006
Income	-6.1353	2.1936	-2.80	0.0075
log(Miles)	26.7552	9.3374	2.87	0.0063

$\hat{\sigma} = 64.8912$ with $46 df$, $R^2 = 0.5105$.

where the residual sum of squares RSS is the unexplained sum of squares, and the regression sum of squares SSreg is the explained sum of squares. As with simple regression, the ratio

$$R^2 = \frac{\text{SSreg}}{\text{SYY}} = 1 - \frac{\text{RSS}}{\text{SYY}} \quad (3.22)$$

gives the proportion of variability in Y explained by regression on the regressors. R^2 can also be shown to be the square of the correlation between the observed values Y and the fitted values \hat{Y} ; we will explore this further in the next chapter. R is also called the *multiple correlation coefficient* because it is the maximum of the correlation between Y and *any* linear combination of the regressors in the mean function.

For the fuel consumption data we have

$$R^2 = 1 - \frac{\text{RSS}}{\text{SYY}} = 1 - \frac{193,700}{395,694} = 1 - 0.490 = 0.510$$

About half the variation in `Fuel` is explained by the regressors. The value of R^2 is given in Table 3.3 and is typically produced by regression software.

3.4.7 Hypotheses Concerning One Coefficient

The multiple regression model has many regression coefficients, and so many tests are possible. In this section we consider only testing of individual coefficients and defer more general testing to Chapter 6.

As in simple regression, an estimated coefficient divided by its standard error provides the basis for a test that the coefficient is equal to 0. In the Fuel data, consider a test concerning β_1 , the coefficient for the regressor `Tax`. The hypothesis tested is

$$\begin{aligned} \text{NH: } & \beta_1 = 0, \quad \beta_0, \beta_2, \beta_3, \beta_4 \text{ arbitrary} \\ \text{AH: } & \beta_1 \neq 0, \quad \beta_0, \beta_2, \beta_3, \beta_4 \text{ arbitrary} \end{aligned} \quad (3.23)$$

This hypothesis explicitly shows that the test concerns β_1 only and that all other coefficients are not effected, so it is essentially testing *the effect of adding Tax* to a mean function that already includes all the other regressors. From Table 3.3, $t = -2.08$. The df associated with the t -statistic is the number of df in the estimate of variance, which is $n - p' = 46$. Most computer programs will find the corresponding significance level for this test for you, and it is given in Table 3.3 as $p = 0.043$, providing some evidence that the effect of `Tax` on fuel consumption, after adjusting for the other predictors, is different from 0. For a one-sided alternative, for example testing $\beta_1 < 0$, the significance level would be $0.043/2 = 0.022$ because $\hat{\beta}_1 < 0$. For the one-sided test that $\beta_1 > 0$, the significance level would be $1 - 0.043/2 = 0.978$.

Tests concerning other regression coefficients have a similar interpretation: each of the tests is computed as if that regressor were added last to a regression model, and the test adjusted for all regressors in the model.

A t -test that β_j has a specific value versus a two-sided or one-sided alternative with all other coefficients arbitrary can be carried out as described in Section 2.6.

3.4.8 t -Tests and Added-Variable Plots

In Section 3.1, we discussed adding a regressor to a simple regression mean function. The same general procedure can be used to add a regressor to *any* linear regression mean function. For the added-variable plot for a regressor, say X_1 , plot the residuals from the regression of Y on all the other X s versus the residuals for the regression of X_1 on all the other X s. One can show (Problem 3.2) that (1) the slope of the regression in the added-variable plot is the estimated coefficient for X_1 in the regression with all the regressors, (2) the t -test for testing the slope to the 0 in the added-variable plot is essentially the same as the t -test for testing $\beta_1 = 0$ in the fit of the larger mean function, the only difference being a correction for df , and (3) the value of R^2 in the added-variable plot is equal to the square of the partial correlation between the response and the regressor, adjusted for the other regressors in the mean function.

3.5 PREDICTIONS, FITTED VALUES, AND LINEAR COMBINATIONS

Suppose we have observed, or will in the future observe, a new case with its own set of predictors that result in a vector of regressors \mathbf{x}_* . We would like to predict the value of the response given \mathbf{x}_* . In exactly the same way as was done in simple regression, the point prediction is $\tilde{y}_* = \mathbf{x}'_* \hat{\boldsymbol{\beta}}$, and the standard error of prediction, $\text{sepred}(\tilde{y}_* | \mathbf{x}_*)$, using Appendix A.8, is

$$\text{sepred}(\tilde{y}_* | \mathbf{x}_*) = \hat{\sigma} \sqrt{1 + \mathbf{x}'_* (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_*} \quad (3.24)$$

Similarly, the estimated average of all possible units with a value \mathbf{x} for the regressors is given by the estimated mean function at \mathbf{x} , $\hat{E}(Y | X = \mathbf{x}) = \hat{y} = \mathbf{x}' \hat{\boldsymbol{\beta}}$ with standard error given by

$$\text{sefit}(\hat{y} | \mathbf{x}) = \hat{\sigma} \sqrt{\mathbf{x}' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}} \quad (3.25)$$

Virtually all software packages will give the user access to the fitted values, but getting the standard error of prediction and of the fitted value may be harder. If a program produces sefit but not sepred , the latter can be computed from the former from the result

$$\text{sepred}(\tilde{y}_*|\mathbf{x}_*) = \sqrt{\hat{\sigma}^2 + \text{sefit}(\tilde{y}_*|\mathbf{x}_*)^2}$$

A minor generalization allows computing an estimate and standard error for any linear combination of estimated coefficients. Suppose \mathbf{a} is a vector of numbers of the same length as $\boldsymbol{\beta}$. Then the linear combination $\ell = \mathbf{a}'\boldsymbol{\beta}$ has estimate and standard error given by

$$\hat{\ell} = \mathbf{a}'\hat{\boldsymbol{\beta}} \quad \text{se}(\hat{\ell}|X) = \hat{\sigma}\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}} \quad (3.26)$$

3.6 PROBLEMS

- 3.1** (Data file: UN11) Identify the localities corresponding to the poorly fitting points in Figure 3.2 and explain what these localities have in common.
- 3.2** **Added-variable plots** (Data file: UN11) This problem uses the United Nations example in Section 3.1 to demonstrate many of the properties of added-variable plots. This problem is based on the mean function $\text{fertility} \sim \log(\text{ppgdp}) + \text{pctUrban}$. There is nothing special about a two-predictor regression mean function, but we are using this case for simplicity.
- 3.2.1** Examine the scatterplot matrix for $(\text{fertility}, \log(\text{ppgdp}), \text{pctUrban})$, and comment on the marginal relationships.
 - 3.2.2** Fit the two simple regressions for $\text{fertility} \sim \log(\text{ppgdp})$ and for $\text{fertility} \sim \text{pctUrban}$, and verify that the slope coefficients are significantly different from 0 at any conventional level of significance.
 - 3.2.3** Obtain the added-variable plots for both predictors. Based on the added-variable plots, is $\log(\text{ppgdp})$ useful after adjusting for pctUrban , and similarly, is pctUrban useful after adjusting for $\log(\text{ppgdp})$? Compute the estimated mean function with both predictors included as regressors, and verify the findings of the added-variable plots.
 - 3.2.4** Show that the estimated coefficient for $\log(\text{ppgdp})$ is the same as the estimated slope in the added-variable plot for $\log(\text{ppgdp})$ after pctUrban . This correctly suggests that *all the estimates in a multiple linear regression model are adjusted for all the other regressors in the mean function*.
 - 3.2.5** Show that the residuals in the added-variable plot are identical to the residuals from the mean function with both predictors.
 - 3.2.6** Show that the t -test for the coefficient for $\log(\text{ppgdp})$ is not quite the same from the added-variable plot and from the regression with both regressors, and explain why they are slightly different.

3.3 Berkeley Guidance Study (Data file: `BGSgirls`) The Berkeley Guidance Study enrolled children born in Berkeley, California, between January 1928 and June 1929, and then measured them periodically until age 18 (Tuddenham and Snyder, 1954). The data we use include heights in centimeters at ages 2, 9, and 18, (`HT2`, `HT9`, and `HT18`), weights in kilogram (`WT2`, `WT9`, and `WT18`), leg circumference in centimeters (`LG2`, `LG9`, and `LG18`), and strength in kilogram (`ST2`, `ST9`, and `ST18`). Two additional measures of body type are also given, `soma`, somatotype, a scale from 1, very thin, to 7, obese, and body mass index, computed as $BMI18 = WT18 / (HT18/100)^2$, weight in kilogram divided by the square of mass in meters, a standard measure of obesity. The data are in the files `BGSgirls` for girls only, `BGSboys` for boys only, and `BGSall` for boys and girls combined (in this last file an additional variable `Sex` has value 0 for boys and 1 for girls). For this problem use only the data on the girls.⁵

- 3.3.1** For the girls only, draw the scatterplot matrix of `HT2`, `HT9`, `WT2`, `WT9`, `ST9`, and `BMI18`. Write a summary of the information in this scatterplot matrix. Also obtain the matrix of sample correlations between the these same variables and compare with the scatterplot matrix.
- 3.3.2** Starting with the mean function $E(BMI18|WT9) = \beta_0 + \beta_1 WT9$, use added-variable plots to explore adding `ST9` to get the mean function $E(BMI18|WT9, ST9) = \beta_0 + \beta_1 WT9 + \beta_2 ST9$. Obtain the marginal plots of `BMI18` versus each of `WT9` and `ST9`, the plot of `ST9` versus `WT9`, and then the added-variable plots for `ST9`. Summarize your results.
- 3.3.3** Fit the multiple linear regression model with mean function

$$E(BMI18|X) = \beta_0 + \beta_1 HT2 + \beta_2 WT2 + \beta_3 HT9 + \beta_4 WT9 + \beta_5 ST9 \quad (3.27)$$

Find $\hat{\sigma}$ and R^2 . Compute the t -statistics to be used to test each of the β_j to be 0 against two-sided alternatives. Explicitly state the hypotheses tested and the conclusions.

3.4 The following questions all refer to the mean function

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad (3.28)$$

- 3.4.1** Suppose we fit (3.28) to data for which $x_1 = 2.2x_2$, with no error. For example, x_1 could be a weight in pounds, and x_2 the weight of the same object in kilogram. Describe the appearance of the added-variable plot for X_2 after X_1 .
- 3.4.2** Again referring to (3.28), suppose now that $Y = 3X_1$ without error, but X_1 and X_2 are not perfectly correlated. Describe the appearance of the added-variable plot for X_2 after X_1 .

⁵The variable `soma` was used in earlier editions of this book but is not used in this problem.

- 3.4.3** Under what conditions will the added-variable plot for X_2 after X_1 have exactly the same shape as the marginal plot of Y versus X_2 ?
- 3.4.4** True or false: The vertical variation of the points in an added-variable plot for X_2 after X_1 is always less than or equal to the vertical variation in a plot of Y versus X_2 . Explain.
- 3.5** Suppose we have a regression in which we want to fit the mean function (3.1). Following the outline in Section 3.1, suppose that the two terms X_1 and X_2 have sample correlation equal to 0. This means that, if x_{ij} , $i = 1, \dots, n$, and $j = 1, 2$ are the observed values of these two terms for the n cases in the data, $SX_1X_2 = \sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = 0$. Define $SX_jX_j = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$ and $SX_jY = \sum_{i=1}^n (x_{ij} - \bar{x}_j)(y_i - \bar{y}_j)$, for $j = 1, 2$.
- 3.5.1** Give the formula for the slope of the regression for Y on X_1 , and for Y on X_2 . Give the value of the slope of the regression for X_2 on X_1 .
- 3.5.2** Give formulas for the residuals for the regressions of Y on X_1 and for X_2 on X_1 . The plot of these two sets of residuals corresponds to the added-variable plot for X_2 .
- 3.5.3** Compute the slope of the regression corresponding to the added-variable plot for the regression of Y on X_2 after X_1 , and show that this slope is exactly the same as the slope for the simple regression of Y on X_2 ignoring X_1 . Also find the intercept for the added-variable plot.
- 3.6** (Data file: water) Refer to the data described in Problem 1.5. For this problem, consider the regression problem with response BSAAM, and three predictors as regressors given by OPBPC, OPRC, and OPSLAKE.
- 3.6.1** Examine the scatterplot matrix drawn for these three regressors and the response. What should the correlation matrix look like (i.e., which correlations are large and positive, which are large and negative, and which are small)? Compute the correlation matrix to verify your results.
- 3.6.2** Get the regression summary for the regression of BSAAM on these three regressors. Explain what the “ t -values” column of your output means.
- 3.7** Suppose that \mathbf{A} is a $p \times p$ symmetric matrix that we write in partitioned form

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}'_{12} & \mathbf{A}_{22} \end{pmatrix}$$

The matrix \mathbf{A}_{11} is $p_1 \times p_1$, so \mathbf{A}_{22} is $(p - p_1) \times (p - p_1)$. One can show that if \mathbf{A}^{-1} exists, it can be written as

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}'_{12} & -\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}'_{12} & \mathbf{A}_{22}^{-1} \end{pmatrix}$$

Using this result, show that, if \mathbf{X} is an $n \times (p + 1)$ data matrix with all 1s in the first column,

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \bar{\mathbf{x}}'(\mathcal{X}'\mathcal{X})^{-1}\bar{\mathbf{x}} & -\bar{\mathbf{x}}'(\mathcal{X}'\mathcal{X})^{-1} \\ -(\mathcal{X}'\mathcal{X})^{-1}\bar{\mathbf{x}} & (\mathcal{X}'\mathcal{X})^{-1} \end{pmatrix}$$

where \mathcal{X} and $\bar{\mathbf{x}}$ are defined in Section 3.4.3.

C H A P T E R 4

Interpretation of Main Effects

The computations that are done in multiple linear regression, including drawing graphs, creation of regressors, fitting models, and performing tests, will be similar in most problems. Interpreting the results, however, may differ by problem, even if the outline of the analysis is the same. Many issues play into drawing conclusions, and some of them are discussed in this chapter, with elaborations in Chapter 5 where more complex regressors like factors, interactions, and polynomials are presented.

4.1 UNDERSTANDING PARAMETER ESTIMATES

We start with the fitted mean function for the fuel consumption data, given by

$$\begin{aligned}\hat{E}(\text{Fuel}|X) = & 154.19 - 4.23 \text{Tax} + 0.47 \text{Dlic} - 6.14 \text{Income} \\ & + 26.76 \log(\text{Miles})\end{aligned}\tag{4.1}$$

This equation represents the estimated conditional mean of Fuel given a fixed value for the regressors collected in X . The β -coefficients, often called slopes or partial slopes, have *units*. Since Fuel is measured in gallons per person, all the quantities on the right of (4.1) must also be in gallons. The intercept is 154.19 gal. It corresponds to the expected fuel consumption in a state with no taxes, no drivers, no income and essentially no roads, and so is not interpretable in this problem because no such state could exist. Since Income is measured in thousands of dollars, the coefficient for Income must be in gallons per person per thousand dollars of income. Similarly, the units for the coefficient for Tax is gallons per person per cent of tax.

4.1.1 Rate of Change

The usual interpretation of an estimated coefficient is as a rate of change: increasing Tax rate by 1 cent, with all the other regressors in the model held fixed, is associated with a change in Fuel of about -4.23 gal per person on the average. We can visualize the effect of Tax by fixing the other regressors in (4.1) at their sample mean values, $\bar{\mathbf{x}}_2 = (\text{Dlic} = 903.68, \text{Income} = 28.4, \log(\text{Miles}) = 10.91)'$, to get

$$\begin{aligned}\hat{E}(\text{Fuel} | X_1 = x_1, X_2 = \bar{\mathbf{x}}_2) &= \hat{\beta}_0 + \hat{\beta}_1 \text{Tax} + \hat{\beta}_2 \overline{\text{Dlic}} + \hat{\beta}_3 \overline{\text{Income}} + \hat{\beta}_4 \overline{\log(\text{Miles})} \\ &= 154.19 - 4.23 \text{Tax} + 0.47(903.68) - 6.14(28.4) \\ &\quad + 26.76(10.91) \\ &= 606.92 - 4.23 \text{Tax}\end{aligned}$$

We can then draw the graph shown in Figure 4.1. This graph is called an *effects plot* (Fox, 2003), as it shows the effect of Tax with all other predictors held fixed at their sample mean values. For a mean function like (4.1), choosing any other fixed value of the remaining predictors X_2 would not change the shape of the curve in the plot, but would only change the intercept. The dotted lines on the graph provide a 95% pointwise confidence interval for the fitted values, as described in Section 3.5, computed at $(x_1, \bar{\mathbf{x}}_2)$ as x_1 is varied, and so the graph can show both the effect and its variability. This graph shows that the expected effect of higher Tax rate is lower Fuel consumption. Some readers will find this graph to be a better summary than a numeric summary of the estimated $\hat{\beta}_1$ and its standard error, although both contain the same information.

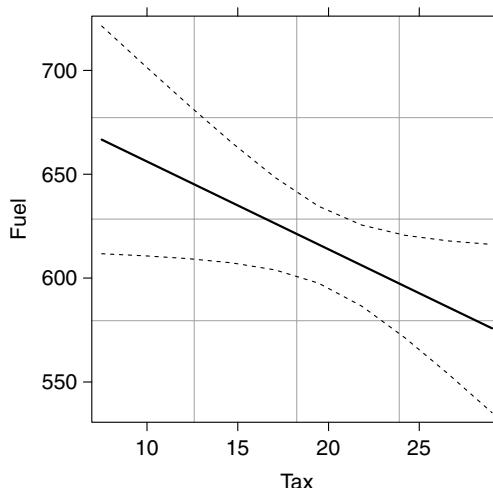


Figure 4.1 Effects plot for Tax in the fuel consumption data.

Interpreting a coefficient or its estimate as a rate of change given that other regressors are fixed assumes that the regressor can in fact be changed without affecting the other regressors in the mean function and that the available data will apply when the predictor is so changed. The fuel data are *observational* since the assignment of values for the predictors was not under the control of the analyst, so whether increasing taxes would *cause* a decrease in fuel consumption cannot be assessed from these data. We can observe *association* but not cause: states with higher tax rates are *observed* to have lower fuel consumption. To draw conclusions concerning the effects of changing tax rates, the rates must in fact be changed and the results observed.

4.1.2 Signs of Estimates

The sign of a parameter estimate indicates the direction of the relationship between the regressor and the response after adjusting for all other regressors in the mean function, and in many studies, the most important finding is the sign, not the magnitude, of an estimated coefficient. If regressors are correlated, both the magnitude and the sign of a coefficient may change depending on the other regressors in the model. While this is mathematically possible and, occasionally, scientifically reasonable, it certainly makes interpretation more difficult. Sometimes this problem can be removed by redefining the regressors into new linear combinations that are easier to interpret.

4.1.3 Interpretation Depends on Other Terms in the Mean Function

The value of a parameter estimate not only depends on the other regressors in a mean function, but it can also change if the other regressors are replaced by linear combinations of the regressors.

Berkeley Guidance Study

Data from the Berkeley Guidance Study on the growth of boys and girls are given in Problem 3.3. We will view body mass index at age 18, BMI_{18} , as the response, and weights in kilogram at ages 2, 9, and 18, WT_2 , WT_9 , and WT_{18} as predictors, for the $n = 70$ girls in the study. The scatterplot matrix for these four variables is given in Figure 4.2.

Look at the first row of this figure, giving the marginal response plots of BMI_{18} versus each of the three potential predictors. BMI_{18} is increasing with each of the potential predictors, although the relationship is strongest at the oldest age, as would be expected because BMI is computed from weight, and weakest at the youngest age.¹ The two-dimensional plots of each pair of

¹One point corresponding to a value of $\text{BMI}_{18} > 35$ is separated from the other points, and a more careful analysis would repeat any analysis with and without that point to see if the analysis is overly dependent on that point.

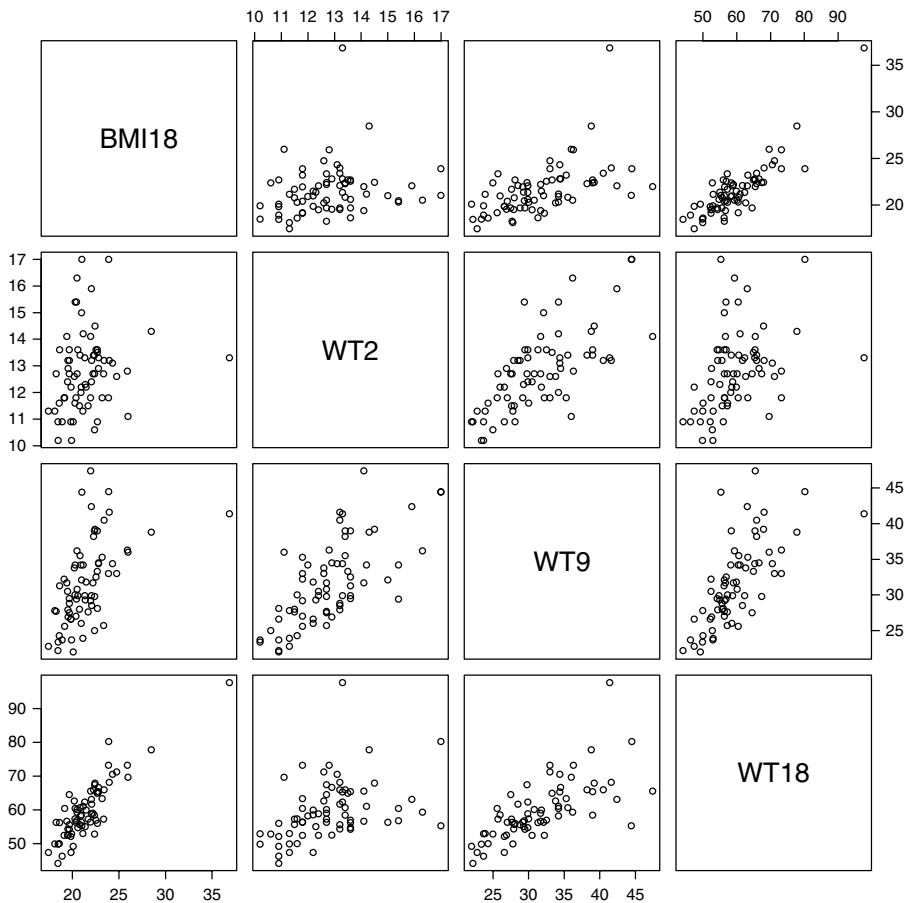


Figure 4.2 Scatterplot matrix for the girls in the Berkeley Guidance Study.

predictors suggest that the predictors are correlated among themselves. Taken together, we have evidence that the regression on all three predictors cannot be viewed as just the sum of the three separate simple regressions because we must account for the correlations between the regressors.

We will proceed with this example using the three original predictors as regressors and BMI18 as the response. We are encouraged to do this because of the appearance of the scatterplot matrix. Since each of the two-dimensional plots appears to be well summarized by a straight-line mean function, we will see later that this suggests transformations are unnecessary and that the regression of the response with regressors given by the original predictors is likely to be appropriate.

The parameter estimates for the regression with regressors WT2, WT9, and WT18 given in the column marked “Model 1” in Table 4.1 leads to the unexpected conclusion that heavier girls at age 2 may tend to be thinner and have

Table 4.1 Regression of BMI18 on Different Combinations of Three Weight Variables for the n = 70 Girls in the Berkeley Guidance Study

Regressor	Model 1	Model 2	Model 3
(Intercept)	8.298*	8.298*	8.298*
WT2	-0.383*	-0.065	-0.383*
WT9	0.032		0.032
WT18	0.287*		0.287*
DW9		0.318*	Aliased
DW18		0.287*	Aliased

*Indicates p -value < 0.05.

lower expected BMI18. We reach this conclusion based on the small p -value for the t -test that the coefficient of WT2 is equal to zero ($t = -2.53$, p -value = 0.01, two-tailed). The unexpected sign may be due to the correlations between the regressors. In place of the preceding variables, consider the following:

$$\text{WT2} = \text{Weight at age 2}$$

$$\text{DW9} = \text{WT9} - \text{WT2} = \text{Weight gain from age 2 to 9}$$

$$\text{DW18} = \text{WT18} - \text{WT9} = \text{Weight gain from age 9 to 18}$$

Since all three original regressors measure weight, combining them in this way is reasonable. If the variables were in different units, then taking linear combinations of them could lead to uninterpretable estimates. The parameter estimates for the regression with regressors WT2, DW9, and DW18 are given in the column marked “Model 2” in Table 4.1. Although not shown in the table, summary statistics for the regression like R^2 and $\hat{\sigma}^2$ are identical for all the mean functions in Table 4.1. In Model 2, the coefficient estimate for WT2 is about one-fifth the size of the estimate in Model 1, and the corresponding t -statistic is much smaller ($t = -0.51$, p -value = 0.61, two-tailed). In Model 1, the “effect” of WT2 seems to be negative and significant, while in the equivalent Model 2, the effect of WT2 would be judged not different from zero. As long as predictors are correlated, interpretation of the effect of a predictor depends not only on the other predictors in a model but also upon which linear transformation of those variables is used.

Another interesting feature of Table 4.1 is that the estimate for WT18 in Model 1 is identical to the estimate for DW18 in Model 2. This is not a coincidence. In Model 1, the estimate for WT18 is the effect on BMI18 of increasing WT18 by 1 kg, with all other regressors held fixed. In Model 2, the estimate for DW18 is the change in BMI18 when DW18 changes by 1 kg, when all other regressors are held fixed. *But the only way DW18 = WT18 - WT9 can be changed by 1 kg with the other variables, including WT9 = DW9 - WT2, held fixed is by changing WT18 by 1 kg.* Consequently, the regressors WT18 in Model 1 and

DW18 in Model 2 play identical roles and therefore we get the same estimates, even though the regressors are different.

4.1.4 Rank Deficient and Overparameterized Mean Functions

In the last example, several regressors derived from the basic predictors WT2 , WT9 , and WT18 were studied. One might naturally ask what would happen if more than three combinations of these predictors were used in the same regression model. As long as we use linear combinations of the predictors, as opposed to nonlinear combinations or transformations of them, we cannot use more than three, the number of linearly independent quantities.

To see why this is true, consider adding DW9 to the mean function, including WT2 , WT9 , and WT18 . As in Chapter 3, we can learn about adding DW9 using an added-variable plot of the residuals from the regression $\text{BMI18} \sim \text{WT2} + \text{WT9} + \text{WT18}$ versus the residuals from the regression $\text{DW9} \sim \text{WT2} + \text{WT9} + \text{WT18}$. Since DW9 can be written as an exact linear combination of the other predictors, $\text{DW9} = \text{WT9} - \text{WT2}$, the residuals from this second regression are all exactly zero. A slope coefficient for DW9 is thus not defined after adjusting for the other three regressors. We would say that the four regressors WT2 , WT9 , WT18 , and DW9 are *linearly dependent*, since one can be determined exactly from the others. The three variables WT2 , WT9 , and WT18 are *linearly independent* because one of them cannot be determined exactly by a linear combination of the others. The maximum number of linearly independent regressors that could be included in a mean function is called the *rank* of the data matrix \mathbf{X} .

Model 3 in Table 4.1 gives the estimates produced in a computer package when we tried to fit $\text{BMI18} \sim \text{WT2} + \text{WT9} + \text{WT18} + \text{DW9} + \text{DW18}$. Some computer packages will silently select three of these five regressors, usually the first three. Others may indicate the remaining coefficient estimates to be `NA` for not available, or as *aliased*, a better choice because it can remind the analyst that the choice of which three coefficients to estimate is arbitrary. The same R^2 , $\hat{\sigma}^2$, fitted values, and residuals would be obtained for all choices of the three coefficients to estimate.

Mean functions that are overparameterized occur most often in designed experiments. The simplest example is the one-way design that will be described more fully in Section 5.1. Suppose that an experimental unit is assigned to one of three treatment groups, and let $X_1 = 1$ if the experimental unit is in group one and 0 otherwise, $X_2 = 1$ if the experimental unit is in group two and 0 otherwise, and $X_3 = 1$ if the experimental unit is in group three and 0 otherwise. For each unit, we must have $X_1 + X_2 + X_3 = 1$ since each unit is in only one of the three groups. We therefore cannot fit the model

$$\text{E}(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

because the sum of the X_i is equal to the column of ones, and so, for example, $X_3 = 1 - X_1 - X_2$. To fit a model, we must do something else. The options are

(1) place a constraint like $\beta_1 + \beta_2 + \beta_3 = 0$ on the parameters; (2) exclude one of the X_j from the model, or (3) leave out an explicit intercept. All of these options will in some sense be equivalent, since the same overall fit result. Of course, some care must be taken in using parameter estimates, since these will surely depend on the parameterization used to get a full rank model. For further reading on matrices and models of less than full rank, see, for example, Christensen (2011), Schott (2005), or Fox and Weisberg (2011, section 4.6.1).

4.1.5 Collinearity

Suppose \mathbf{X} is the data matrix for the set of regressors in a particular regression problem. We say that the set of regressors is *collinear* if we can find a vector of constants \mathbf{a} such that $\mathbf{X}\mathbf{a} \approx \mathbf{0}$. If the “ \approx ” is replaced by an “=” sign, then at least one of the regressors is a linear combination of the others, and we have an overparameterized model as outlined in Section 4.1.4. If \mathbf{X} is collinear, then the R^2 for the regression of one of the regressors on all the remaining regressors, including the intercept, is close to one. Collinearity depends on the sample correlations between the regressors, not on theoretical population quantities.²

The data in the file `MinnWater` provide yearly water usage in Minnesota for the period 1988–2011. For the example we consider here, the response variable is `log(muniUse)`, the logarithm of water used in metropolitan areas, in billions of gallons, and potential predictors are `year` of measurement, `muniPrecip`, growing season precipitation in inches, and `log(muniPop)` the logarithm of the metropolitan state population in census years, and U.S. Census estimates between census years. The data were collected to explore if water usage has changed over the 24 years in the data.

The data are shown in Figure 4.3. The bottom row of this graph shows the marginal relationships between `log(muniUse)` and the regressors. The bottom-left graph shows that usage was clearly increasing over the time period, and the second graph in the bottom row suggests that usage may be somewhat lower when precipitation is lower. The two regressors appear to be nearly uncorrelated because the second graph in the third row appears to be a null plot.

Table 4.2 summarizes three multiple linear regression mean functions fit to model `log(muniUse)`. The first column labeled Model 1 uses only `year` as a regressor. Listed in the table are the values of the estimated intercept and slope. To save space, we have used an asterisk (*) to indicate estimates with corresponding significance levels less than 0.01.³ As expected, `log(muniUse)` is increasing over time. When we add `muniPrecip` to the mean function in the second column, the estimate for `year` hardly changes, as expected from the lack of correlation between `year` and `muniPrecip`.

²The term *multicollinearity* contains more syllables, but no additional information, and is a synonym for collinearity.

³For data collected in time order like these, the standard *t*-tests might be questionable because of lack of independence of consumption from year to year. Several alternative testing strategies are presented in Chapter 7.

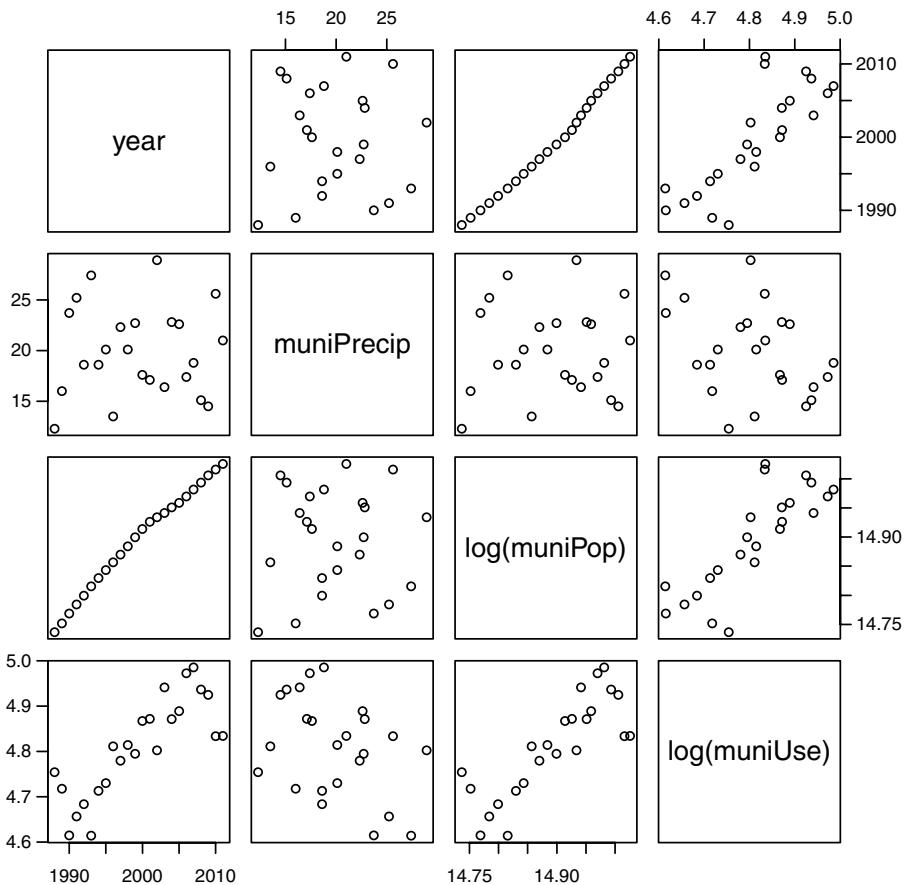


Figure 4.3 Scatterplot matrix for the Minnesota water use data.

Table 4.2 Regression of $\log(\text{muniUse})$ on Different Combinations of Regressors for the Minnesota Water Use Data

Regressor	Model 1	Model 2	Model 3
(Intercept)	-20.0480*	-20.1584*	-1.2784
year	0.0124*	0.0126*	-0.0111
muniPrecip		-0.0099*	-0.0106*
$\log(\text{muniPop})$			1.9174

*Indicates p -value < 0.01 .

Adding $\log(\text{muniPop})$, however, tells a different story: the coefficient for year is much smaller and negative, and has a large corresponding significance level. The cause of this is clear: $\log(\text{muniPop})$ is seen in Figure 4.3 to be very highly correlated with year , so year and $\log(\text{muniPop})$ are possibly explaining the same variation in $\log(\text{muniUse})$. We simply cannot tell from this

regression if the increase in usage is simply a reflection of increased population or increased usage per person per year.

An alternative approach is to use $\log(\text{perCapitaUse}) = \log(10^6 \text{muniUse}/\text{muniPop})$ as a response variable. The multiplier 10^6 is included to rescale to thousands of gallons per person rather than billions of gallons per person. We leave as a homework problem to show that in the per capita scale there is no evidence of increasing municipal water usage.

4.1.6 Regressors in Logarithmic Scale

Logarithms are commonly used both for the response and for regressors. Commonly used logarithmic scales include decibels for loudness, the Richter scale for the intensity of an earthquake, and pH levels for measuring acidity. Varshney and Sun (2013) suggest that in many cases, human perception is logarithmic. As a practical matter, predictors that span several orders of magnitude should be transformed to log scale. Examples of this in this book include miles of roadway in a state, which vary from around 1,500 miles to over 300,000 miles, and per capita gross domestic product in different countries, which vary from about \$100 per person to about \$100,000 per person.

The regressor $\log(\text{Miles})$ in the fuel consumption data summarized in Equation (4.1) uses natural logarithms. The effects plot for $\log(\text{Miles})$ is shown as Figure 4.4a, a straight line with standard error lines similar to Figure 4.1, the effects plot for Tax. Figure 4.4b is a different version of the effects plot for $\log(\text{Miles})$, with the horizontal axis in the original units Miles rather than the transformed units. The transformation changes the straight line for

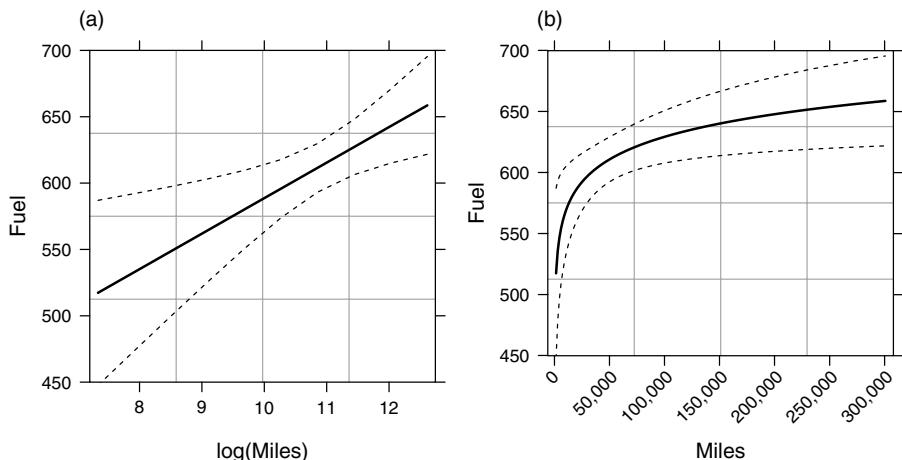


Figure 4.4 (a) Effects plot for $\log(\text{Miles})$ in the fuel consumption data. (b) The horizontal axis is given in the more useful scale of Miles, and thus the fitted effect is a curve rather than a straight line.

the effects plot into a curve. We see that the effect of Miles is increasing fastest in states with fewest miles of roadway, with relatively little change in states with the most roads. This is the usual effect of logarithms: it allows fitted effect that change most rapidly when the predictor is small and less rapidly when the predictor is large.

4.1.7 Response in Logarithmic Scale

As with predictors, transforming a response to log scale is sometimes based on theoretical considerations, but even lacking a theory if the response is strictly positive log scale will generally be desirable when errors are of the form “plus or minus 5%” rather than additive errors of the form “plus or minus 5 units.”

Suppose the response is $\log(Y)$. The interpretation of the regression coefficient for the j th regressor β_j in a regression model is the rate of change in $\log(Y)$ as X_j varies. This is generally not very useful because $\log(Y)$ changes nonlinearly with Y , and an alternative interpretation is often used.

Concentrating on the j th coefficient, we use the subscript (j) to imply excluding the j th element, so $X_{(j)}$, $\mathbf{x}_{(j)}$, $\boldsymbol{\beta}_{(j)}$ are, respectively, the regressors excluding X_j , the observed values of the regressors excluding x_j , and the regression coefficients excluding β_j . The regression model is

$$E[\log(Y)|X_j = x_j, X_{(j)} = \mathbf{x}_{(j)}] = \beta_0 + \beta_j x_j + \boldsymbol{\beta}'_{(j)} \mathbf{x}_{(j)} \quad (4.2)$$

We first approximate the expected value of $\log(Y)$ by the logarithm of the expected value,

$$\log[E(Y|X_j = x_j, X_{(j)} = \mathbf{x}_{(j)})] \approx E[\log(Y)|X_j = x_j, X_{(j)} = \mathbf{x}_{(j)}] \quad (4.3)$$

We use the “ \approx ” sign to indicate approximate equality that is generally sufficiently accurate for the results of this section. Exponentiating both sides of (4.3), using (4.2) we get

$$\begin{aligned} E[Y|X_j = x_j, X_{(j)} = \mathbf{x}_{(j)}] &\approx \exp\{E[\log(Y)|X_j = x_j, X_{(j)} = \mathbf{x}_{(j)}]\} \\ &= \exp(\beta_0 + \beta_j x_j + \boldsymbol{\beta}'_{(j)} \mathbf{x}_{(j)}) \\ &= \exp(\beta_j x_j) \exp(\beta_0 + \boldsymbol{\beta}'_{(j)} \mathbf{x}_{(j)}) \end{aligned}$$

If we increase x_j by 1 while keeping $X_{(j)}$ fixed, we have

$$\begin{aligned} E[Y|X_j = x_j + 1, X_{(j)} = \mathbf{x}_{(j)}] &\approx \exp[\beta_j(x_j + 1)] \exp(\beta_0 + \boldsymbol{\beta}'_{(j)} \mathbf{x}_{(j)}) \\ &= \exp(\beta_1) \exp(\beta_0 + \beta_j x_j + \boldsymbol{\beta}'_{(j)} \mathbf{x}_{(j)}) \\ &= \exp(\beta_1)[E(Y|X_1 = x_1, X_2 = \mathbf{x}_2)] \quad (4.4) \end{aligned}$$

Thus, for any j increasing x_j by 1 will *multiply* the mean of Y by approximately $\exp(\beta_j)$. This is often expressed as a percentage change, and

$$100 \times \frac{\mathbb{E}[Y|X_j = x_j + 1, X_{(j)} = \mathbf{x}_{(j)}] - \mathbb{E}[Y|X_j = x_j, X_{(j)} = \mathbf{x}_{(j)}]}{\mathbb{E}[Y|X_j = x_j, X_{(j)} = \mathbf{x}_{(j)}]} = 100(\exp(\beta_j) - 1)$$

is the approximate percentage increase, or decrease if the value of β_j is negative, in the response when X_j is increased by 1. For example, if $\beta_j = 0.3$, then $100(\exp(\beta_j) - 1) = 34\%$, or a 34% increase in the expected value of the response. If $\beta_j = -0.2$, then increasing X_j yields a percentage increase of $100(\exp(-0.2) - 1) = -18\%$, or an 18% decrease in the expected value of the response.

Using the UN data, fit a regression model with response `log(fertility)` and regressors `log(ppgdp)` and `lifeExpF`. The fitted regression is

$$\widehat{\text{log(fertility)}} = 3.507 - 0.065 \text{log(ppgdp)} - 0.028 \text{lifeExpF}$$

Increasing `lifeExpF` by 1 year is associated with $100(\exp(-0.028) - 1) = -2.8\%$ decrease in `fertility`.

If natural logarithms are used and the value of β_j is close to 0, say $-0.4 \leq \beta_j \leq 0.4$, then to a reasonable approximation $(\exp(\beta_j) - 1) \approx \beta_j$, so in this case β_j can be interpreted as the fractional increase (or decrease if the sign is negative) in the expected value of Y when X_j increases by 1 and the other regressors are fixed. For example, in the Minnesota water use data, the coefficient for `year` from Model 2 in Table 4.2 was $\hat{\beta}_1 = 0.0126$, suggesting an approximate 1.3% increase in water use per year (assuming, of course, that Model 2 is meaningful).

If both the regressor and the response are in log scale, then increasing the regressor by 1 unit as in (4.4) corresponds to multiplying the regressor by $e \approx 2.718 \dots$, and this rarely makes sense. Suppose Z is a predictor and X_j is the regressor representing it in the model, $X_j = \log(Z)$. If the observed value z of Z is replaced by cz , then the regressor becomes $\log(cz) = \log(c) + \log(z)$. Then the result similar to (4.4) is

$$\begin{aligned} \mathbb{E}[Y|X_j = x_j + \log(c), X_{(j)} = \mathbf{x}_{(j)}] &\approx \exp[\beta_j(x_j + \log(c))] \exp(\beta_0 + \boldsymbol{\beta}'_{(j)} \mathbf{x}_{(j)}) \\ &= \exp(\log(c)\beta_j)[\mathbb{E}(Y|X_1 = x_1, X_2 = \mathbf{x}_2)] \end{aligned} \quad (4.5)$$

For example, if Z is increased by 10%, then $c = 1.1$, and the expected response is multiplied by $\exp[\log(1.1)\beta_j] \approx \exp(0.1\beta_j)$ because $\log(1.1) \approx 0.1$. In the UN example, for a 10% increase in `ppgdp`, the expected `fertility` will be multiplied by $\exp(0.1 \times (-0.065)) = 0.994$, corresponding to a change in `fertility` of $100\% (0.994 - 1) = -0.6\%$. A 25% increase in `ppgdp` is associated with a change in `fertility` of -1.4% .

4.2 DROPPING REGRESSORS

The regression parameters are always conditioned on a set of regressors; if the regressors are changed, then usually so are the parameters and their interpretation. If a linear regression model is appropriate for one set of regressors, we shall see that it is not necessarily true that a linear regression model is appropriate for a subset of the regressors. In particular, if

$$E(Y|X_1 = \mathbf{x}_1, X_2 = \mathbf{x}_2) = \beta_0 + \boldsymbol{\beta}'_1 \mathbf{x}_1 + \boldsymbol{\beta}'_2 \mathbf{x}_2 \quad (4.6)$$

is known to hold, what can we say about $E(Y|X_1 = \mathbf{x}_1)$, obtained by dropping the regressors in X_2 ?

An example might be helpful. Suppose we have a sample of n rectangles from which we want to model $\log(\text{area})$ as a function of $\log(\text{length})$, perhaps through the simple regression mean function

$$E(\log(\text{area})|\log(\text{length})) = \eta_0 + \eta_1 \log(\text{length}) \quad (4.7)$$

We know from elementary geometry that $\text{area} = \text{length} \times \text{width}$, and so the “true” mean function for $Y = \log(\text{area})$ is given by (4.6), with $X_1 = \log(\text{length})$ and $X_2 = \log(\text{width})$. In this instance, we also know the parameters $\beta_0 = 0$ and $\beta_1 = \beta_2 = 1$. Is (4.7) appropriate if $\log(\text{width})$ is not used?

4.2.1 Parameters

The results in Appendix A.2.4 provide the answer. The mean function for $Y|X_1$ is

$$\begin{aligned} E(Y|X_1 = \mathbf{x}_1) &= E[E(Y|X_1 = \mathbf{x}_1, X_2) | X_1 = \mathbf{x}_1] \\ &= \beta_0 + \boldsymbol{\beta}'_1 \mathbf{x}_1 + \boldsymbol{\beta}'_2 E(X_2 | X_1 = \mathbf{x}_1) \end{aligned} \quad (4.8)$$

We cannot simply drop a set of regressors from a correct mean function, but we need to substitute the conditional expectation of the regressors dropped given the regressors that remain in the mean function.

In the context of the rectangles example, we get

$$\begin{aligned} E(\log(\text{area})|\log(\text{length})) &= \beta_0 + \beta_1 \log(\text{length}) \\ &\quad + \beta_2 E(\log(\text{width})|\log(\text{length})) \end{aligned} \quad (4.9)$$

The answers to the questions posed depend on the mean function for the regression $\log(\text{width}) \sim \log(\text{length})$. This conditional expectation has little to do with the area of rectangles, but much to do with the way we obtain a sample of rectangles to use in our study. We will consider three cases.

In the first case, imagine that each of the rectangles in the study is formed by sampling a $\log(\text{length})$ and a $\log(\text{width})$ from independent distributions. If the mean of the $\log(\text{width})$ distribution is W , then by independence,

$$E(\log(\text{width})|\log(\text{length})) = E(\log(\text{width})) = W$$

Substituting into (4.9),

$$\begin{aligned} E(\log(\text{area})|\log(\text{length})) &= \beta_0 + \beta_1 \log(\text{length}) + \beta_2 W \\ &= (\beta_0 + \beta_2 W) + \beta_1 \log(\text{length}) \\ &= W + \log(\text{length}) \end{aligned}$$

where the last equation follows by substituting $\beta_0 = 0$, $\beta_1 = \beta_2 = 1$. For this case, the mean function (4.7) would be appropriate for the regression $\log(\text{width}) \sim \log(\text{length})$. The intercept for the mean function (4.7) would be W , and so it depends on the distribution of the widths in the data. The slope for $\log(\text{length})$ is the same in the full model or the model with only one regressor.

In the second case, suppose that

$$E(\log(\text{width})|\log(\text{length})) = \gamma_0 + \gamma_1 \log(\text{length})$$

so the mean function for the regression of $\log(\text{width})$ on $\log(\text{length})$ is a straight line. This could occur, for example, if the rectangles in our study were obtained by sampling from a family of similar rectangles, so the ratio $\gamma = \text{width}/\text{length}$ is the same for all rectangles in the study. Substituting this into (4.9) and simplifying gives

$$\begin{aligned} E(\log(\text{area})|\log(\text{length})) &= \beta_0 + \beta_1 \log(\text{length}) + \beta_2(\gamma_0 + \gamma_1 \log(\text{length})) \\ &= (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \beta_2 \gamma_1) \log(\text{length}) \\ &= \gamma_0 + (1 + \gamma_1) \log(\text{length}) \end{aligned}$$

Once again, fitting using (4.7) will be appropriate, but the values of η_0 and η_1 depend on the parameters of the regression of $\log(\text{width})$ on $\log(\text{length})$. Two experimenters who sample rectangles of different shapes will end up estimating different parameters.

For a final case, suppose that the mean function

$$E(\log(\text{width})|\log(\text{length})) = \gamma_0 + \gamma_1 \log(\text{length}) + \gamma_2 \log(\text{length})^2$$

is quadratic. Substituting into (4.9), setting $\beta_0 = 0$, $\beta_1 = \beta_2 = 1$ and simplifying gives

$$\begin{aligned} E(\log(\text{area})|\log(\text{length})) &= \beta_0 + \beta_1 \log(\text{length}) \\ &\quad + \beta_2 (\gamma_0 + \gamma_1 \log(\text{length}) + \gamma_2 \log(\text{length})^2) \\ &= \gamma_0 + (1 + \gamma_1) \log(\text{length}) + \gamma_2 \log(\text{length})^2 \end{aligned}$$

which is a quadratic function of $\log(\text{length})$. If the mean function is quadratic, or any other function beyond a straight line, then fitting (4.7) is inappropriate.

From the above three cases, we see that both the mean function and the parameters for the response depend on the mean function for the regression of the removed regressors on the remaining regressors. If the mean function for the regression of the removed regressors on the retained regressors is not linear, then a linear mean function will not be appropriate for the regression problem with fewer regressors.

4.2.2 Variances

Variances are also affected when regressors are dropped. Returning to the true mean function given by (4.6), the general result for the regression of Y on X_1 alone is, from Appendix A.2.4,

$$\begin{aligned} \text{Var}(Y|X_1 = \mathbf{x}_1) &= E[\text{Var}(Y|X_1 = \mathbf{x}_1, X_2)|X_1 = \mathbf{x}_1] \\ &\quad + \text{Var}[E(Y|X_1 = \mathbf{x}_1, X_2)|X_1 = \mathbf{x}_1] \\ &= \sigma^2 + \boldsymbol{\beta}_2' \text{Var}(X_2|X_1 = \mathbf{x}_1) \boldsymbol{\beta}_2 \end{aligned} \quad (4.10)$$

In the context of the rectangles example, $\boldsymbol{\beta}_2$ is a scalar, $\boldsymbol{\beta}_2 = 1$, and we get

$$\text{Var}(\log(\text{area})|\log(\text{length})) = \sigma^2 + \text{Var}(\log(\text{width})|\log(\text{length}))$$

Although fitting (4.7) can be appropriate if $\log(\text{width})$ and $\log(\text{length})$ are linearly related, the errors for this mean function can be much larger than those for (4.6) if $\text{Var}(\log(\text{width})|\log(\text{length}))$ is large. If $\text{Var}(\log(\text{width})|\log(\text{length}))$ is small enough, then fitting (4.7) can actually give answers that are nearly as accurate as fitting with the true mean function (4.6).

4.3 EXPERIMENTATION VERSUS OBSERVATION

There are fundamentally two types of predictors that are used in a regression analysis: *experimental* and *observational*. Experimental predictors have values that are under the control of the experimenter. For observational predictors, the values are observed rather than set. Consider, for example, a hypothetical study of factors determining the yield of a certain crop. Experimental variables

might include the amount and type of fertilizers used, the spacing of plants, and the amount of irrigation, since each of these can be assigned by the investigator to the units, which are plots of land. Observational predictors might include characteristics of the plots in the study, such as drainage, exposure, soil fertility, and weather variables. All of these are beyond the control of the experimenter, yet may have important effects on the observed yields.

The primary difference between experimental and observational predictors is in the inferences we can make. From experimental data, we can often infer causation. If we assign the level of fertilizer to plots, usually on the basis of a randomization scheme, we can infer a causal relationship between fertilization and yield. Observational predictors allow weaker inferences. We might say that weather variables are associated with yield, but the causal link is not available for variables that are not under the experimenter's control. Some experimental designs, including those that use randomization, are constructed so that the effects of observational factors can be ignored or used in analysis of covariance (Cox, 1958; Oehlert, 2000).

Purely observational studies that are not under the control of the analyst can only be used to predict or model the events that were observed in the data, as in the fuel consumption example. To apply observational results to predict future values, additional assumptions about the behavior of future values compared with the behavior of the existing data must be made. The goal of inferring causality from data has been studied in depth in a number of subject-matter areas, including statistics. A recent book of essays on this subject is Berzuini et al. (2012).

4.3.1 Feedlots

A *feedlot* is a farming operation that includes large number of cattle, swine, or poultry in a small area. Feedlots are efficient producers of animal products and can provide high-paying skilled jobs in rural areas. They can also cause environmental problems, particularly with odors, groundwater pollution, and noise.

Taff et al. (1996) reported a study on the effect of feedlots on property values. This study was based on all 292 rural residential property sales in two southern Minnesota counties in 1993–1994. Regression analysis was used. The response was the logarithm of sale price. Predictors were derived from house characteristics, such as size, number of bedrooms, age of the property, and so on. Additional predictors described the relationship of the property to existing feedlots, such as distance to the nearest feedlot, number of nearby feedlots, and related features of the feedlots such as their size. The “feedlot effect” could be inferred from the coefficients for the regressors created from the feedlot variables.

In the analysis, the coefficient estimates for feedlot effects were generally positive and judged to be nonzero, meaning that close proximity to feedlots was associated with an *increase* in sale prices. While association of the opposite

sign was expected, the positive sign is plausible if the positive economic impact of the feedlot outweighs the negative environmental impact. The positive effect is estimated to be small, however, and equal to 5% or less of the sale price of the homes in the study.

These data are purely observational, with no experimental predictors. The data collectors had no control over the houses that actually sold, or siting of feedlots. Consequently, any inference that nearby feedlots *cause* increases in sale price is unwarranted from this study. Given that we are limited to association, rather than causation, we might next turn to whether we can generalize the results. Can we infer the same association to houses that were *not* sold in these counties during this period? We have no way of knowing from the data if the same relationship would hold for homes that did not sell. For example, some homeowners may have perceived that they could not get a reasonable price and may have decided not to sell. This would create a bias in favor of a positive effect of feedlots.

Can we generalize geographically, to other Minnesota counties or to other places in the Midwest United States? The answer to this question may depend on the characteristics of the two counties studied. Both are rural counties with populations of about 17,000. Both had very low property values with median sale price in this period of less than \$50,000, a low value even in 1993–1994. Each county had different regulations for operators of feedlots, and these regulations could impact pollution problems. Applying the results to a county with different demographics or regulations cannot be justified by these data alone, and additional information and assumptions are required.

Joiner (1981) coined the picturesque phrase *lurking variable* to describe a predictor variable not included in a mean function that if included would change the interpretation of a fitted model. Suppose we have a regression with regressors X derived from the available predictors and a lurking variable L not included in the study, and that the true regression mean function is

$$E(Y|X = \mathbf{x}, L = \ell) = \beta_0 + \sum_{j=1}^p \beta_j x_j + \delta \ell \quad (4.11)$$

with $\delta \neq 0$. We assume that X and L are correlated and for simplicity we assume further that $E(L|X = \mathbf{x}) = \gamma_0 + \sum \gamma_j x_j$. When we fit the incorrect mean function that ignores the lurking variable, we get, from Section 4.2,

$$\begin{aligned} E(Y|X = \mathbf{x}) &= \beta_0 + \sum_{j=1}^p \beta_j x_j + \delta E(L|X = \mathbf{x}) \\ &= (\beta_0 + \delta \gamma_0) + \sum_{j=1}^p (\beta_j + \delta \gamma_j) x_j \end{aligned} \quad (4.12)$$

Suppose we are particularly interested in inferences about the coefficient for X_1 , and, unknown to us, β_1 in (4.11) is equal to 0. If we were able to fit with

the lurking variable included, we could conclude that X_1 is not important. If we fit the incorrect mean function (4.12), the coefficient for X_1 becomes $(\beta_1 + \delta\gamma)$, which will be nonzero if $\delta \neq 0$. The lurking variable masquerades as the variable of interest to give an incorrect inference. A lurking variable can also hide the effect of an important variable if, for example, $\beta_1 \neq 0$ but $\beta_1 + \delta\gamma = 0$.

All large observational studies like this feedlot study potentially have lurking variables. For this study, a casino had recently opened near these counties, creating many jobs and a demand for housing that might well have overshadowed any effect of feedlots. In experimental data with random assignment, the potential effects of lurking variables are greatly decreased, since the random assignment guarantees that the correlation between the regressors in the mean function and any lurking variable is small or 0.

The interpretation of results from a regression analysis depends on the details of the data design and collection. The feedlot study has extremely limited scope and is but one element to be considered in trying to understand the effect of feedlots on property values. Studies like this feedlot study are easily misused. The study was cited in spring 2004 in an application for a permit to build a feedlot in Starke county, Indiana, claiming that the study supports the positive effect of feedlots on property values, confusing association with causation, and inferring generalizability to other locations without any established foundation for doing so.

4.4 SAMPLING FROM A NORMAL POPULATION

Much of the intuition for the use of least squares estimation is based on the assumption that the observed data are a sample from a multivariate normal population. While the assumption of multivariate normality is only rarely tenable in practical regression problems, it is worthwhile to explore the relevant results for normal data, first assuming random sampling and then removing that assumption.

Suppose that all of the observed variables are normal random variables, and the observations on each case are independent of the observations on each other case. In a two-variable problem, for the i th case observe (x_i, y_i) , and suppose that

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \text{Cov}(x, y) \\ \text{Cov}(x, y) & \sigma_y^2 \end{pmatrix}\right) \quad (4.13)$$

Equation (4.13) says that x_i and y_i are each realizations of normal random variables with means μ_x and μ_y , variances σ_x^2 and σ_y^2 , and the covariance $\text{Cov}(x, y) = \rho_{xy}\sigma_x\sigma_y$. Now, suppose we consider the conditional distribution of y_i given that we have already observed the value of x_i . It can be shown

(Casella and Berger, 2001) that the conditional distribution of y_i given x_i is normal and

$$y_i|x_i \sim N\left(\mu_y + \rho_{xy} \frac{\sigma_y}{\sigma_x} (x_i - \mu_x), \sigma_y^2(1 - \rho_{xy}^2)\right) \quad (4.14)$$

If we define

$$\beta_0 = \mu_y - \beta_1 \mu_x \quad \beta_1 = \rho_{xy} \frac{\sigma_y}{\sigma_x} \quad \sigma^2 = \sigma_y^2(1 - \rho_{xy}^2) \quad (4.15)$$

then the conditional distribution of y_i given x_i is simply

$$y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \quad (4.16)$$

which is essentially the same as the simple regression model with the added assumption of normality. The variance $\sigma^2 = \sigma_y^2(1 - \rho_{xy}^2)$ in (4.16) is often called the *residual variance*, as it is the variance of the part of y that is not explained by x .

Given random sampling, the five parameters in (4.13) are estimated, using the notation of Table 2.1, by

$$\begin{aligned} \hat{\mu}_x &= \bar{x} & \hat{\sigma}_x^2 &= SD_x^2 & \hat{\rho}_{xy} &= r_{xy} \\ \hat{\mu}_y &= \bar{y} & \hat{\sigma}_y^2 &= SD_y^2 \end{aligned} \quad (4.17)$$

Maximum likelihood estimates of the regression parameters β_0 and β_1 are found in Appendix A.11. They are obtained by substituting estimates from (4.17) for parameters in (4.15), so that $\hat{\beta}_1 = r_{xy} SD_y / SD_x$, and so on, as derived in Chapter 2. To get the unbiased estimate of σ^2 we must correct for degrees of freedom, $\hat{\sigma}^2 = [(n-1)/(n-2)]SD_y^2(1 - r_{xy}^2)$.

If the observations on the i th case are y_i and a $p \times 1$ vector \mathbf{x}_i not including a constant, multivariate normality is shown symbolically by

$$\begin{pmatrix} \mathbf{x}_i \\ y_i \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}'_{xy} & \sigma_y^2 \end{pmatrix}\right) \quad (4.18)$$

where $\boldsymbol{\Sigma}_{xx}$ is a $p \times p$ matrix of variances and covariances between the elements of \mathbf{x}_i , and $\boldsymbol{\Sigma}_{xy}$ is a $p \times 1$ vector of covariances between \mathbf{x}_i and y_i . The conditional distribution of y_i given x_i is then

$$y_i|\mathbf{x}_i \sim N((\mu_y - \boldsymbol{\beta}^{*\prime} \boldsymbol{\mu}_x) + \boldsymbol{\beta}^{*\prime} \mathbf{x}_i, \sigma^2) \quad (4.19)$$

If \mathcal{R}^2 is the population squared multiple correlation we can write

$$\boldsymbol{\beta}^* = \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy}; \quad \sigma^2 = \sigma_y^2 - \boldsymbol{\Sigma}'_{xy} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy} = \sigma_y^2 (1 - \mathcal{R}^2) \quad (4.20)$$

The formulas for $\boldsymbol{\beta}^*$ and σ^2 and the formulas for their least squares estimators differ only by the substitution of estimates for parameters, with $(n-1)^{-1}(\mathcal{X}'\mathcal{X})$ estimating $\boldsymbol{\Sigma}_{xx}$, and $(n-1)^{-1}(\mathcal{X}'\mathcal{Y})$ estimating $\boldsymbol{\Sigma}_{xy}$.

The last result generalizes: if \mathbf{z} is a $k \times 1$ random variable with a multivariate normal distribution, for any vector \mathbf{a} , $y = \mathbf{a}'\mathbf{z}$ is a linear combination of the elements of \mathbf{z} , and for any matrix $p \times k$ matrix \mathbf{B} , $\mathbf{X} = \mathbf{B}\mathbf{z}$ is a set of linear combinations of the elements of \mathbf{z} , then the regression of y on \mathbf{z} is always a linear regression model.

4.5 MORE ON R^2

The conditional distribution in (4.14) or in (4.19) does not depend on random sampling, but only on normal distributions, so whenever multivariate normality seems reasonable, a linear regression model is suggested for the conditional distribution of one variable given the others. However, if random sampling is not used, some of the usual summary statistics, including R^2 , lose their connection to population parameters.

Apart from using a different aspect ratio in the plot, Figure 4.5a repeats Figure 1.1, the scatterplot of dheight versus mheight for the heights data. These data closely resemble a bivariate normal sample (see Problem 4.12), and so $R^2 = 0.24$ estimates the population \mathcal{R}^2 for this problem. Figure 4.5b repeats this last figure, except that all cases with mheight between 61 and 64 inches, which are the lower and upper quartiles of the mother's heights rounded to the nearest inch, respectively, have been removed from the data. The OLS regression line appears similar, but the value of $R^2 = 0.37$ is about 50% larger. Removing the middle of the data increased R^2 , and it no longer estimates a population value. Similarly, in Figure 4.5c, we exclude all the cases with mheight outside the quartiles, and get $R^2 = 0.027$, and the relationship between dheight and mheight virtually disappears.

We have seen that we can manipulate the value of R^2 merely by changing our sampling plan for collecting data: if the values of the regressors are widely dispersed, then R^2 will tend to be too large, while if the values are over a very small range, then R^2 will tend to be too small. Because the notion of proportion of variability explained is so useful, a diagnostic method is needed to decide if it is a useful concept in any particular problem.

4.5.1 Simple Linear Regression and R^2

In simple linear regression problems, we can always determine the appropriateness of R^2 as a summary by examining the summary graph of the response versus the regressor. If the plot looks like a sample from a bivariate normal

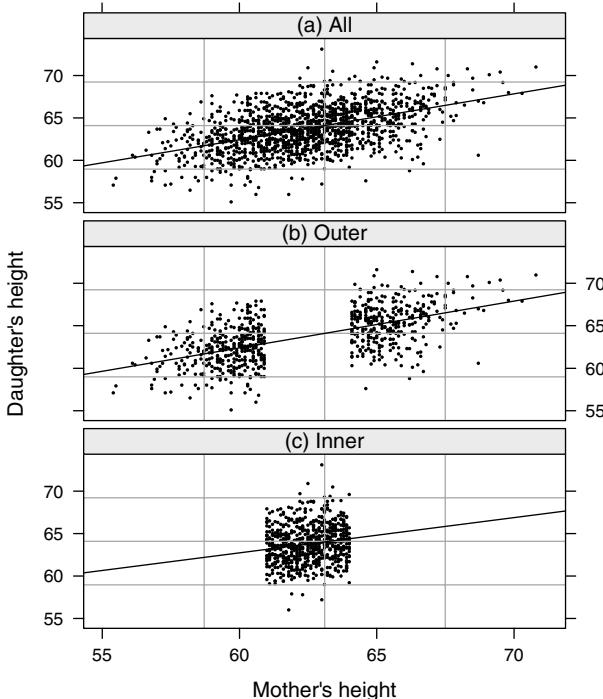


Figure 4.5 Three views of the heights data.

population, as in Figure 4.5a, then R^2 is a useful measure. The less the graph looks like this figure, the less useful is R^2 as a summary measure.

Figure 4.6 shows six summary graphs. Only for the first three of them is R^2 a useful summary of the regression problem. In Figure 4.6d, the mean function appears curved rather than straight so correlation is a poor measure of dependence. In Figure 4.6e the value of R^2 is virtually determined by one point, making R^2 necessarily unreliable. The regular appearance of Figure 4.6f suggests a different type of problem. We may have several identifiable groups of points caused by a lurking variable not included in the mean function, such that the mean function for each group has a negative slope, but when groups are combined the slope becomes positive. Once again, R^2 is not a useful summary of this graph.

4.5.2 Multiple Linear Regression and R^2

The sample multiple correlation coefficient R^2 can be shown to be the correlation between the response Y and the OLS fitted values \hat{Y} . This suggests that a plot of Y on the vertical axis and \hat{Y} on the horizontal axis can be useful for interpreting R^2 paralleling the methodology for simple regression just outlined. When the data are sampled from a multivariate normal, R^2 will estimate

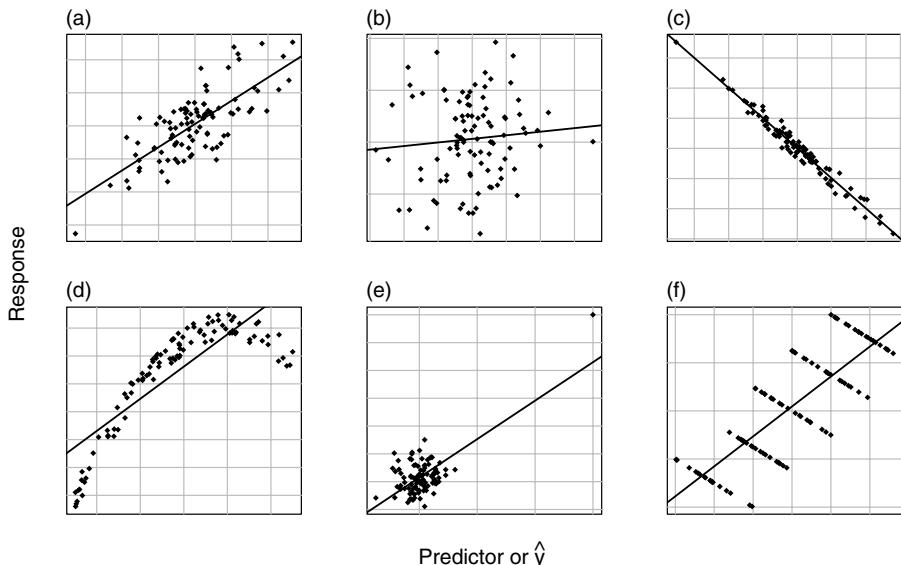


Figure 4.6 Six summary graphs. R^2 is an appropriate measure for a–c, but inappropriate for d–f.

the population multiple correlation \mathcal{R}^2 defined implicitly in (4.19). Without multivariate normality, R^2 estimates a quantity that depends on the sampling plan.

For other regression methods such as nonlinear regression, we can continue to define R^2 to be the square of the correlation between the response and the fitted values, and use this summary graph to decide if R^2 is a useful summary. This extension is not universal, however. In binary regression models described in Chapter 12, R^2 -like measures are not directly analogous to the plot of a response versus fitted values.

4.5.3 Regression through the Origin

With regression through the origin, the proportion of variability explained is given by $1 - \text{RSS}/\sum y_i^2$, using uncorrected sums of squares. This quantity is *not invariant under location change*, so, for example, if units are changed from Fahrenheit to Celsius, the value for the proportion of variability explained changes. For this reason, use of an R^2 -like measure for regression through the origin is not recommended.

4.6 PROBLEMS

- 4.1** (Data file: `BGSgirls`) In the Berkeley Guidance Study data discussed in Section 4.1, another set of linear transformations of the weight variables is

$$\text{ave} = (\text{WT2} + \text{WT9} + \text{WT18})/3$$

$$\text{lin} = \text{WT18} - \text{WT2}$$

$$\text{quad} = \text{WT2} - 2\text{WT9} + \text{WT18}$$

Since the three weight variables are approximately equally spaced in time, these three variables correspond to the average weight, a linear component in time, and a quadratic component in time; see Oehlert (2000) or Kennedy and Gentle (1980), for example, for a discussion of orthogonal polynomials.

Fit with these regressors using the girls in the Berkeley Guidance Study data and compare with the results in Section 4.1.

- 4.2** (Data file: Transact) The data in this example consists of a sample of branches of a large Australian bank (Cunningham and Heathcote, 1989). Each branch makes transactions of two types, and for each of the branches we have recorded the number t_1 of type 1 transactions and the number t_2 of type 2 transactions. The response is `time`, the total minutes of labor used by the branch.

Define $a = (t_1 + t_2)/2$ to be the average transaction time, and $d = t_1 - t_2$, and fit the following four mean functions

$$M1: E(\text{time}|t_1, t_2) = \beta_{01} + \beta_{11}t_1 + \beta_{21}t_2$$

$$M2: E(\text{time}|t_1, t_2) = \beta_{02} + \beta_{32}a + \beta_{42}d$$

$$M3: E(\text{time}|t_1, t_2) = \beta_{03} + \beta_{23}t_2 + \beta_{43}d$$

$$M4: E(\text{time}|t_1, t_2) = \beta_{04} + \beta_{14}t_1 + \beta_{24}t_2 + \beta_{34}a + \beta_{44}d$$

- 4.2.1** In the fit of M4, some of the coefficients estimates are labeled as “aliased” or else they are simply omitted. Explain what this means and why this happens.
- 4.2.2** What aspects of the fitted regressions are the same? What aspects are different?
- 4.2.3** Why is the estimate for t_2 different in M1 and M3?

4.3 Finding a joint distribution

- 4.3.1** Starting with (4.14), we can write

$$y_i = \mu_y + \rho_{xy} \frac{\sigma_y}{\sigma_x} (x_i - \mu_x) + e_i$$

Ignoring the error term e_i , solve this equation for x_i as a function of y_i and the parameters.

- 4.3.2** Find the conditional distribution of $x_i|y_i$. Under what conditions is the equation you obtained in Problem 4.3.1, which is computed by inverting the regression of y on x , the same as the regression of x on y ?
- 4.4** Suppose we have a vector \mathbf{z} which has a multivariate normal distribution,

$$\mathbf{z} \sim N(\boldsymbol{\mu}_z, \boldsymbol{\Sigma})$$

Let $y = \mathbf{a}'\mathbf{z}$ for some $k \times 1$ vector \mathbf{a} , and let $\mathbf{x} = \mathbf{B}\mathbf{z}$ for some $p \times k$ matrix \mathbf{B} . Using (A.17) and (A.18) in Appendix A.7, show that the conditional distribution of $y|\mathbf{x}$ is normal and that the conditional mean is a linear function of \mathbf{x} . Get expressions for the parameters of the conditional distribution.

- 4.5** If you use the response $\log_{10}(Y)$, show that the interpretation of a regression coefficient as a percentage change in Y changes slightly; how does it change?
- 4.6** (Data file: UN11) In the simple linear regression of `log(fertility)` on `pctUrban` using the `UN11` data, the fitted model is

$$\widehat{\log(\text{fertility})} = 1.501 - 0.01\text{pctUrban}$$

Provide an interpretation of the estimated coefficient for `pctUrban`.

- 4.7** (Data file: UN11) Verify that in the regression `log(fertility) ~ log(ppgdp) + lifeExpF` a 25% increase in `ppgdp` is associated with a 1.4% decrease in expected fertility.

- 4.8** Suppose we fit a regression with the true mean function

$$E(Y|X_1 = x_1, X_2 = x_2) = 3 + 4x_1 + 2x_2$$

Provide conditions under which the mean function for $E(Y|X_1 = x_1)$ is linear but has a negative coefficient for x_1 .

- 4.9** In a study of faculty salaries in a small college in the Midwest, a linear regression model was fit, giving the fitted mean function

$$\widehat{E(\text{Salary}|\text{Sex})} = 24697 - 3340\text{Sex} \quad (4.21)$$

where `Sex` equals 1 if the faculty member was female and 0 if male. The response `Salary` is measured in dollars (the data are from the 1970s).

- 4.9.1** Give a sentence that describes the meaning of the two estimated coefficients.
- 4.9.2** An alternative mean function fit to these data with an additional term, *Years*, the number of years employed at this college, gives the estimated mean function

$$\widehat{E(\text{Salary}|\text{Sex}, \text{Years})} = 18065 + 201\text{Sex} + 759\text{Years} \quad (4.22)$$

The important difference between these two mean functions is that the coefficient for *Sex* has changed signs. Using the results of this chapter, explain how this could happen. (Data consistent with these equations are presented in Problem 5.17).

- 4.10** Suppose you are given random variables x and y such that

$$x \sim N(\mu_x, \sigma_x^2)$$

$$y|x \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

so you have the marginal distribution of x and the conditional distribution of y given x . The joint distribution of (x, y) is bivariate normal. Find the 5 parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy})$ of the bivariate normal.

- 4.11** For this problem, you will use normal random deviates. First, generate vectors \mathbf{x} and \mathbf{e} , each of 10,000 standard normal random deviates, and then compute $\mathbf{y} = 2\mathbf{x} + \mathbf{e}$. View \mathbf{x} as 10,000 realizations of a random variable x and \mathbf{y} as 10,000 corresponding realizations of a random variable y . Because a very large sample size is used, estimated statistics are nearly equal to population values.

4.11.1 What is the joint distribution of (x, y) ? Verify that the conditional distribution of $y|x \sim N(2x, 1)$.

4.11.2 Fit the simple regression of y on x . Verify that the estimates of the intercept, slope, variance, and R^2 agree with the theoretical values to at least two decimal places.

4.11.3 Test the hypothesis that the slope is equal to 2 against a two-sided alternative and obtain the significance level. What is the probability of rejecting this hypothesis?

4.11.4 Refit the simple regression model, but to the following subsets of the data: (1) all observations with $|x| < 2/3$; (2) all observations with $|x| > 2/3$, and (3) all observations with $x < 0$. The first two cases are similar to the models used in Figure 4.5, but the last case is different because it is not symmetric about the middle of the distribution of x . In each case, about 50% of the data is used in the estimation. Whether or not we observe a case depends on the predictor but not the response, and we would say the data are

missing at random. Compare the estimates of the intercept, slope, σ^2 , and the value of R^2 for these three cases and for the regression fit to all the data.

- 4.11.5** Repeat Problem 4.11.4, but this time select the subset based on the value of the response y , according to the rules (1) all observations with $|y| < 1.5$; (2) all observations with $|y| > 1.5$, and (3) all observations with $y < 0$; as with the last problem each of these will include about 50% of the data in the estimation. Here the probability of observing a case depends on the value that would have been observed, and the data are not missing at random.

In Problem 4.11.4, the mechanism for observing data depends on the predictor but not on the response. In this problem the mechanism for observing data depends on the response. Compare the estimates of the intercept, slope, σ^2 , and the value of R^2 for these three cases and for the regression fit to all the data.

- 4.12** This problem is for you to see what two-dimensional plots of data will look like when the data are sampled from a variety of distributions. For this problem you will need a computer program that allows you to generate random numbers from given distributions. In each of the cases below, set the number of observations $n = 300$, and draw the indicated graphs. Few programs have easy-to-use functions to generate bivariate random numbers, so in this problem you will generate first the predictor X , then the response Y given X .

- 4.12.1** Generate X and e to be independent standard normal random vectors of length n . Compute $Y = 2 + 3X + \sigma e$, where in this problem we take $\sigma = 1$. Draw the scatterplot of Y versus X , add the true regression line $Y = 2 + 3X$, and the OLS regression line. Verify that the scatter of points is approximately elliptical, and the regression line is similar to, but not exactly the same as, the major axis of the ellipse.

- 4.12.2** Repeat Problem 4.12.1 twice, first set $\sigma = 3$ and then repeat again with $\sigma = 6$. How does the scatter of points change as σ changes?

- 4.12.3** Repeat Problem 4.12.1, but this time set X to have a standard normal distribution and e to have a Cauchy distribution (set $\sigma = 1$). The easy way to generate a Cauchy is to generate two vectors V_1 and V_2 of standard normal random numbers, and then set $e = V_1/V_2$. With this setup, the values you generate are not bivariate normal because the Cauchy does not have a population mean or variance.

- 4.13** (Data file: MinnWater) As suggested in Section 4.1.5, examine the regression with response given by $\log(\text{perCapitaUse}) = \log(10^6 \text{muniUse}/\text{muniPop})$ and the regressors described in the text and summarize your results.

C H A P T E R 5

Complex Regressors

In this chapter we describe methods for including predictors in a regression problem that will require more than one regressor, or regressors that are functions of more than one predictor. The most important of these are *factors*, predictors whose values are typically category labels rather than numeric. A factor with d categories will generally require $d - 1$ regressors in a regression model. We also consider *interactions*, which are formed by taking the products of regressors derived from two or more predictors. We round out the discussion of complex regressors by including polynomial and splines that allow modeling curved relationships, and principal component scores that may be useful to reduce a large number of similar predictors to a more manageable set of regressors. The effects plots introduced in the last chapter provide a useful graphical approach to understanding the effect of predictors on the response, even in cases where direct interpretation of coefficient estimates is rather opaque, and so we emphasize graphical summaries.

5.1 FACTORS

Factors allow the inclusion of qualitative or categorical predictors in the mean function of a multiple linear regression model. Factors can have two levels, such as male or female, treated or untreated, and so on, or they can have more than two levels, such as eye color, location, or type of business.

As an example, we return to the United Nations data described in Section 3.1. This is an observational study of all $n = 199$ localities, mostly countries, for which the United Nations provides data. The factor we use is called `group`, which classified the countries into three categories, `africa` for the 53 countries on the African continent, `oecd` for the 31 countries that are members of the OECD, the Organisation for Economic Co-operation and Development,

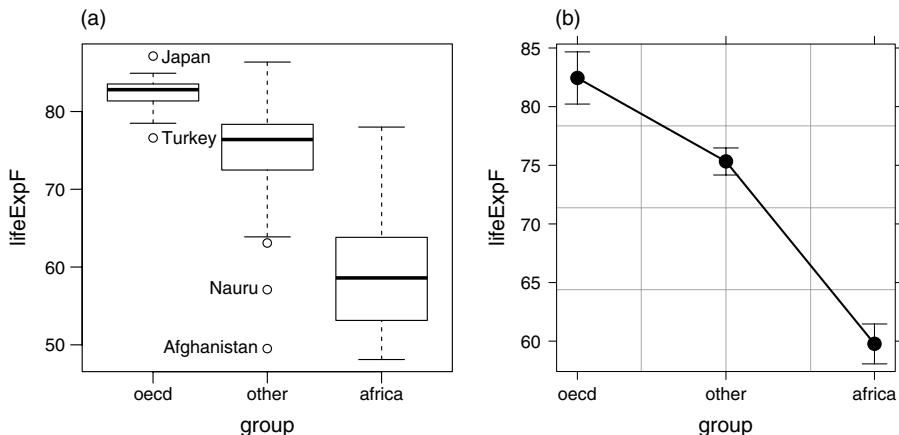


Figure 5.1 UN data. (a) Boxplot of `lifeExpF` separately for each `group` in the UN data. (b) Effect plot for `group` for the one-way model.

an international body¹ whose members are generally the wealthier nations, none of which are in Africa, and `other` for the remaining 115 countries in the data set that are neither in Africa nor in the OECD. The variable `group` is a factor, with these $d = 3$ levels. We will use as a response the variable `lifeExpF`, the expected life span of women in each country, and so the problem at first is to see how `lifeExpF` differs between the three groups of countries.

5.1.1 One-Factor Models

With no predictors beyond `group`, the model we fit returns estimated mean values for `lifeExpF` for each level of `group`. This is called a *one-factor* design or a *one-way* design.

Figure 5.1a provides a boxplot (Mosteller and Tukey, 1977) of `lifeExpF` versus `group` for the example. A boxplot is a useful graphical device for comparing different levels of a factor. Each of the boxes corresponds to one of the levels of `group`. The thick line near the middle of each box is the group median. The “box” extends to the quartiles, and so 50% of the data in the group falls inside the box. The distance between the quartiles is called the *interquartile range* (IQR) and is a measure of variability that is roughly similar to the standard deviation. The “whiskers” generally extend to the observed value closest to the median that is at least 1.5 IQRs from the median. Any points outside this range are shown explicitly. Boxplots give information about location or typical value through the median, scale through the IQR, symmetry by comparing the part of the boxes above the median to the part below it, and possible outliers, the points shown explicitly. In this example the `oecd` countries generally have the highest `lifeExpF` and `africa` has the lowest. There

¹See <http://www.oecd.org>.

is some overlap between `other` and the remaining levels. Two countries have relatively low `lifeExpF` for the `other` group, while in the `oecd` Japan is high and Turkey is low for that group. The levels of the factor have been ordered according to the median value of the response, not alphabetically, which facilitates comparison between groups. The variation in the `oecd` group appears to be smallest and in `africa` it is the largest.

Factor predictors can be included in a multiple linear regression mean function using *dummy variables*. For a factor with two levels, a single dummy variable, a regressor that takes the value 1 for one of the categories and 0 for the other category, can be used. Assignment of labels to the values is generally arbitrary, and will not change the outcome of the analysis. Dummy variables can alternatively be defined with a different set of values, perhaps -1 and 1, or possibly 1 and 2. The important point is the regressor has only two values.

Since `group` has $d = 3$ levels, the j th dummy variable U_j for the factor, $j = 1, \dots, d$ has i th value u_{ij} , for $i = 1, \dots, n$, given by

$$u_{ij} = \begin{cases} 1 & \text{if } \text{group}_i = j\text{th category of group} \\ 0 & \text{otherwise} \end{cases} \quad (5.1)$$

The values of the dummy variables for the first 10 cases in the example are as follows:

	Group	U_1	U_2	U_3
Afghanistan	other	0	1	0
Albania	other	0	1	0
Algeria	africa	0	0	1
Angola	africa	0	0	1
Anguilla	other	0	1	0
Argentina	other	0	1	0
Armenia	other	0	1	0
Aruba	other	0	1	0
Australia	oecd	1	0	0
Austria	oecd	1	0	0

The variable U_1 is the dummy variable for the first level of `group`, which is `oecd`, U_2 is for `other`, and U_3 is for the remaining level `africa`.

If we add an intercept to the mean function, the resulting model would be overparameterized as in Section 4.1.4 because $U_1 + U_2 + U_3 = 1$, a column of 1s, and the column of 1s is the regressor that corresponds to the intercept. This problem can be solved by dropping one of the dummy variables.²

²The statistical program `R` by default deletes the dummy variable for the first level, while `SAS`, `SPSS`, and `Stata` delete the dummy variable for the last level of the factor. All these programs allow reordering the levels of a factor, and so the choice of the deleted level can be changed by the user. The choice of the deleted level effects interpretation of parameters and sometimes tests, but will not generally change fitted values or other summary statistics like R^2 .

$$E(\text{lifeExpF}|\text{group}) = \beta_0 + \beta_2 U_2 + \beta_3 U_3 \quad (5.2)$$

Since the first level of group will be implied when $U_2 = U_3 = 0$,

$$E(\text{lifeExpF}|\text{group} = \text{oecd}) = \beta_0 + \beta_2 0 + \beta_3 0 = \beta_0$$

and so β_0 is the mean for the first level of group. For the second level $U_2 = 1$ and $U_3 = 0$,

$$E(\text{lifeExpF}|\text{group} = \text{other}) = \beta_0 + \beta_2 1 + \beta_3 0 = \beta_0 + \beta_2$$

and $\beta_0 + \beta_2$ is the mean for the second level of group. Similarly, for the third level $U_2 = 0$ and $U_3 = 1$

$$E(\text{lifeExpF}|\text{group} = \text{africa}) = \beta_0 + \beta_2 0 + \beta_3 1 = \beta_0 + \beta_3$$

Most computer programs allow the user to use a factor³ in a mean function without actually computing the dummy variables. For example, the R package uses notation for indicating factors and interactions first suggested by Wilkinson and Rogers (1973). If group has been declared to be a factor, then the mean function (5.2) is be specified by

$$\text{lifeExpF} \sim 1 + \text{group} \quad (5.3)$$

where the “1” specifies fitting the intercept, and group specifies fitting the dummy variable regressors that are created for the factor group. Since most mean functions include an intercept, R assumes it will be included, and the specification

$$\text{lifeExpF} \sim \text{group} \quad (5.4)$$

is equivalent to (5.3).⁴

Table 5.1 summarizes the fit of the one-way model.⁵ For group level oecd the values of the dummy variables are $(U_1, U_2) = (0, 0)$, for other the values

³A factor is called a *class variable* in SAS. Some older programs, such as the “Linear Regression” procedure in SPSS, do not allow symbolic specification of factors and require the user to create dummy variables for them. Programs that allow specifying factors without the user constructing dummy variables, such as the “General Linear Model” program in SPSS, are to be preferred.

⁴The intercept is excluded in R by including -1 in the model, and in other programs by selecting an available option for no intercept.

⁵Using OLS here assumes the variability of the response is the same for each level of group. While Figure 5.1 suggests that the variability may not be constant, we continue for this example assuming constant variance. Methods described in Chapter 7 could provide alternative approaches if non-constant variance were indeed a problem.

Table 5.1 Regression Summary for Model (5.4)

	Estimate	Std. Error	t-Value	Pr(> t)
(Intercept), $\hat{\beta}_0$	82.4465	1.1279	73.09	0.0000
other, $\hat{\beta}_2$	-7.1197	1.2709	-5.60	0.0000
africa, $\hat{\beta}_3$	-22.6742	1.4200	-15.97	0.0000

$\hat{\sigma} = 6.2801$ with 196 df, $R^2 = 0.6191$.

are $(U_1, U_2) = (1, 0)$, and for africa they are $(U_1, U_2) = (0, 1)$. The means for the 3 groups are

$$\begin{aligned}\hat{E}(\text{lifeExpFlgroup} = \text{oecd}) &= \hat{\beta}_0 + \hat{\beta}_2 0 + \hat{\beta}_3 0 = 82.45 \\ \hat{E}(\text{lifeExpFlgroup} = \text{other}) &= \hat{\beta}_0 + \hat{\beta}_2 1 + \hat{\beta}_3 0 = 82.45 - 7.12 \quad (5.5) \\ \hat{E}(\text{lifeExpFlgroup} = \text{africa}) &= \hat{\beta}_0 + \hat{\beta}_2 0 + \hat{\beta}_3 1 = 82.45 - 22.67\end{aligned}$$

The intercept is the sample mean for the omitted level oecd. The estimated coefficient for other is the difference between the sample mean for other and the sample mean for oecd. The coefficient estimate for africa is the difference between the oecd sample mean and the sample mean for africa. The standard errors in the second column of Table 5.1 are, respectively, the standard errors of the estimated mean of oecd followed by the standard errors of the estimated differences between oecd and the other two groups. The standard error for the difference between other and africa is not given in this table but can be computed most easily by changing the baseline level of the factor to be other and refitting or by using the method in Section 5.1.2 to get the standard error and then compute the test.

The column of t-values provide test statistics of, respectively, the mean for oecd is 0; the difference in mean between oecd and other is 0; and the difference in mean between oecd and africa is 0. Using two-sided alternatives, in all three cases, the null hypotheses are clearly rejected since the corresponding significance levels are 0 to the number of digits shown.

A graphical summary of the fitted model is the effects plot introduced in Section 4.1.1. Whereas the boxplot like Figure 5.1a shows the variability in the original data, the effects plot in Figure 5.1b shows the variability in the estimated means. The intervals shown are 95% confidence intervals (without correction for multiple intervals), assuming the variance σ^2 is constant. An important feature of the effects plot is that the summary is in terms of fitted values, in this case fitted means, not in terms of the particular parameterization used for the factor. The baseline level of group is treated differently in the parameterization, but all levels of group are treated equally in the effects plot.

5.1.2 Comparison of Level Means

A common component of the analysis of problems with factors is the comparisons of means for the various levels of a factor adjusted for other factors

and regressors included in the model. To compare the means pairwise in general requires computing the standard error of the difference between each pair of means.

Even in more complicated models, the estimated adjusted difference between means will be given by a linear combination of the estimated regression coefficients. For the example, the estimated difference between means for `other` and `africa` is $\hat{\beta}_2 - \hat{\beta}_3 = -7.12 - (-22.67) = 15.55$, and the standard error of this difference is not given in Table 5.1. It can, however, be computed using the general results given in Section 3.5. Let $\mathbf{a} = (0, 1, -1, 0)'$ so $\ell = \mathbf{a}'\boldsymbol{\beta} = \beta_2 - \beta_3$ is the difference between the group means. Using equation (3.26)

$$\begin{aligned} se(\hat{\ell}|X) &= \hat{\sigma}\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}} \\ &= \hat{\sigma}\sqrt{c_{22} + c_{33} - 2c_{23}} \end{aligned}$$

where c_{ij} is the (i, j) element of $(\mathbf{X}'\mathbf{X})^{-1}$. In R, the function `vcov` applied to a regression model returns $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$, but not all programs provide easy access to this matrix. Often, a higher-level method will be available in the program that will compute all the pairwise comparisons for you. In SAS, SPSS, and others, the term *lsmeans* is often used for the method that does these tests. Table 5.2 presents a summary that would be obtained for comparing differences between level means. The estimate column is the difference in means, and the SE is the standard error of the difference. The *t*-value is the ratio of the estimated difference to its standard error. The *p*-value is the significance level of the test. The *p*-values for these tests are generally adjusted to account for multiple testing, in this case using the Tukey method. Oehlert (2000, chapter 5) provides a useful discussion of multiple comparisons. Bretz et al. (2010) and Lenth (2013) discuss implementations of multiple testing using R. In this example, the additional testing was probably unnecessary because the levels were so obviously different.

5.1.3 Adding a Continuous Predictor

As an additional predictor in the UN example, suppose we add `log(ppgdp)`, the per person gross domestic product in the country, as a measure of relative wealth. The data can now be visualized as in Figure 5.2, which is a plot of the response `lifeExpF` versus the continuous regressor `log(ppgdp)`, with

Table 5.2 Pairwise Comparisons of Level Means for Group

Comparison	Estimate	SE	<i>t</i> -Value	<i>p</i> -Value
<code>oecd</code> – <code>other</code>	7.12	1.27	5.60	0.000
<code>oecd</code> – <code>africa</code>	22.67	1.42	15.97	0.000
<code>other</code> – <code>africa</code>	15.55	1.04	14.92	0.000

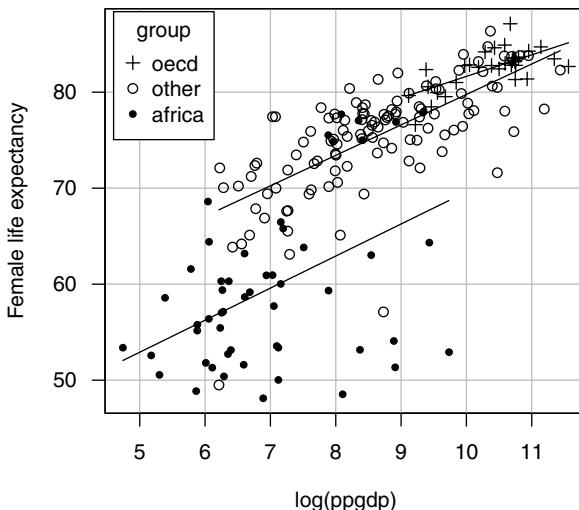


Figure 5.2 Plot of `lifeExpF` versus $\log(\text{ppgdp})$ for the UN data. Points are marked with different symbols for the 3 levels of `group`, and OLS lines are shown for each of the 3 levels of `group`.

separate symbols used for points in the different levels of `group`. Also shown on the graph are OLS lines fit separately to each of the levels, so each level has its own intercept and slope.

The solid circles for countries in the level `africa` generally have lower values for both the response and the regressor, while the points for the remaining levels are similar. The range of values for $\log(\text{ppgdp})$ is considerably smaller in the `oecd` level, reflecting that the countries in this group are generally wealthier. The three lines on the figure seem to be nearly parallel, suggesting that while the intercepts may differ, or at least the intercepts differ between `africa` and the other levels, the slopes, the change in `lifeExpF` as $\log(\text{ppgdp})$ increases, may be the same in each group.

The model fit to obtain the three lines in Figure 5.2 corresponds fitting a separate intercept and slope in each group. Writing $\text{group} = j$ to represent an observation in level j ,

$$E(\text{lifeExpF} | \log(\text{ppgdp}) = x, \text{group} = j) = \eta_{0j} + \eta_{1j}x \quad (5.6)$$

where (η_{0j}, η_{1j}) are the intercept and slope for level $j = 1, \dots, d$, so there $2d = 6$ parameters. This model is generally parameterized differently using *main effects* and *interactions*, as

$$\begin{aligned} E(\text{lifeExpF} | \log(\text{ppgdp}) = x, \text{group}) &= \beta_0 + \beta_{02}U_2 + \beta_{03}U_3 \\ &\quad + \beta_1x + \beta_{22}U_2x + \beta_{13}U_3x \end{aligned} \quad (5.7)$$

As verified in Problem 5.2,

$$\begin{aligned}\eta_{01} &= \beta_0 & \eta_{11} &= \beta_1 \\ \eta_{02} &= \beta_0 + \beta_{02} & \eta_{12} &= \beta_1 + \beta_{12} \\ \eta_{03} &= \beta_0 + \beta_{03} & \eta_{13} &= \beta_1 + \beta_{13}\end{aligned}$$

The parameters (β_0, β_1) are the intercept and slope for the baseline level, while the remaining β s are differences between the other levels and the baseline.

Statistical packages generally allow (5.7) to be fit symbolically. In R one specification is

```
lifeExpF ~ group + log(ppgdp) + group : log(ppgdp)
```

The colon “:” is the indicator for an interaction in R. There is a shorthand for this available in R,

```
lifeExpF ~ group * log(ppgdp)
```

The asterisk “*” in R expands to include all main effects and interactions. In SAS, you cannot transform a variable inside a model specification, so you must create pre-compute `log(ppgdp)`, which we call `lppgdp`. The SAS specification is

```
model lifeExpF = group lppgdp group * lppgdp;
```

Unlike R, SAS uses the asterisk “*” to indicate an interaction. In menu-based programs like SPSS, interactions are specified in dialog boxes.

The regression summary for (5.7) is given in Table 5.3. The estimated intercept is largest for `oecd` because the estimates of both β_{02} and β_{03} are negative. The estimated slope is smallest for `oecd` because both $\hat{\beta}_{12}$ and $\hat{\beta}_{23}$ are positive. The *t*-tests for the coefficients, however, are very confusing, as apart from β_0 none of the coefficients are clearly different from 0, which contradicts intuition from Figure 5.2. We will return to this in the next section and more comprehensively in Section 6.1.

Table 5.3 Regression Summary for Model (5.7)

	Estimate	Std. Error	<i>t</i> -Value	Pr(> <i>t</i>)
(Intercept), $\hat{\beta}_0$	59.2137	15.2203	3.89	0.0001
other, $\hat{\beta}_{02}$	-11.1731	15.5948	-0.72	0.4746
africa, $\hat{\beta}_{03}$	-22.9848	15.7838	-1.46	0.1470
$\log(\text{ppgdp})$, $\hat{\beta}_1$	1.5544	1.0165	1.53	0.1278
other: $\log(\text{ppgdp})$, $\hat{\beta}_{12}$	0.6442	1.0520	0.61	0.5410
africa: $\log(\text{ppgdp})$, $\hat{\beta}_{13}$	0.7590	1.0941	0.69	0.4887

$\hat{\sigma} = 5.1293$ with 193 *df*, $R^2 = 0.7498$.

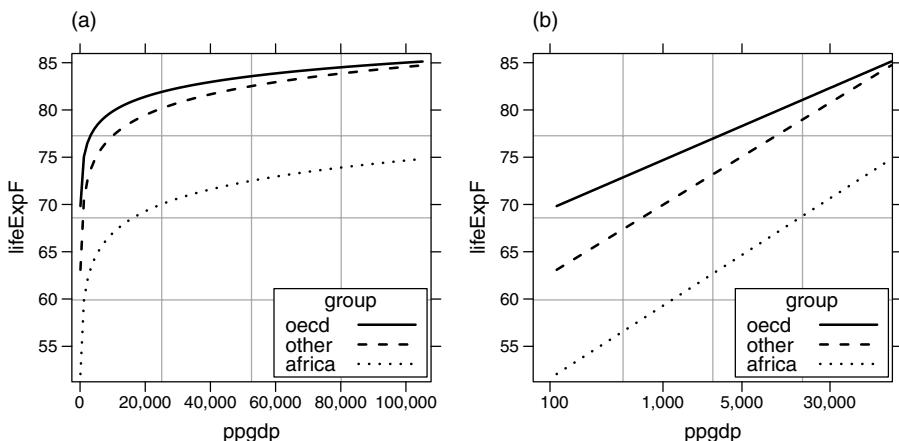


Figure 5.3 Effects plot for the interaction model (5.7) for the UN data. (a) `ppgdp` on the horizontal axis. (b) `ppgdp` in log-scale.

Figure 5.3 provides two variations of effects plots for the fit of the interaction model. In each of the plot curves are shown each level of the factor group. The curves give the fitted value of `lifeExpF` for each given value of `ppgdp`. The lines shown in Figure 5.3a are curves, not straight lines, because the horizontal axis is for `ppgdp`, not its logarithm, while in Figure 5.3b the horizontal axis is in log scale, and the curves become straight lines.

Graphing in the original rather than logarithmic scale emphasizes that the greatest change in fitted `lifeExpF` occurs for increases in smaller values of `ppgdp`, and that in the wealthier countries, changes in `ppgdp` are associated with only small increases in `lifeExpF`. The line shown for `oecd` is really an extrapolation at the lower end of the `ppgdp` scale, because all the countries in this group have high values of `ppgdp`; similarly in `africa`, the curve at the high end is an extrapolation. The `other` and `oecd` groups are clearly very similar to each other, while the curve for `africa` starts lower and stays lower.

5.1.4 The Main Effects Model

Examination of Figure 5.2 suggests that while intercepts might differ for the three levels of `group`, the slopes may be equal. This suggests fitting a model that allows each group to have its own intercept, but all groups have the same slope,

$$E(\text{lifeExpF} | \log(\text{ppgdp}) = x, \text{group}) = \beta_0 + \beta_{02}U_2 + \beta_{03}U_3 + \beta_1 x \quad (5.8)$$

Model (5.8), whose Wilkinson–Rogers representation is `lifeExpF ~ log(ppgdp) + group`, is obtained from (5.7) by dropping the interaction, so we call this a *main effects model*. Main effects models are much simpler

Table 5.4 Regression Summary for Model (5.8)

	Estimate	Std. Error	t-value	Pr(> t)
(Intercept), $\hat{\beta}_0$	49.5292	3.3996	14.57	0.0000
other, $\hat{\beta}_{02}$	-1.5347	1.1737	-1.31	0.1926
africa, $\hat{\beta}_{03}$	-12.1704	1.5574	-7.81	0.0000
log(ppgdp), $\hat{\beta}_1$	2.2024	0.2190	10.06	0.0000

$\hat{\sigma} = 5.1798$ with 195 df, $R^2 = 0.7422$.

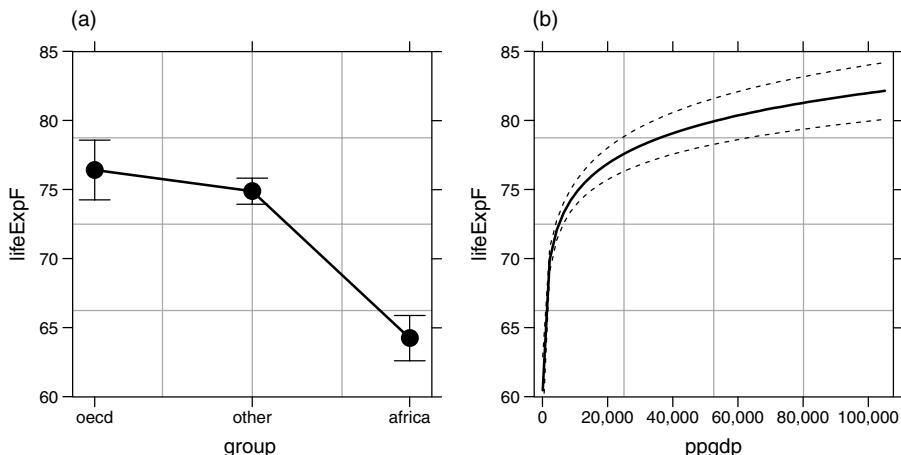


Figure 5.4 Effects plots for the main effects model (5.8) for the UN data: (a) group, (b) ppgdp. Dotted lines are drawn at plus and minus 1 standard error.

than are models with interactions because the effect of the continuous regressor is the same for all levels of the factor. Similarly, the difference between levels of the factor are the same for every fixed value of the continuous regressor. When primary interest is in differences due to the level of the factor, for example, if the factor were the levels of a treatment randomly assigned to subjects, model (5.8) is called the *analysis of covariance*.

The fit of the main effects model (5.8) is given in Table 5.4. The *t*-statistics are less baffling in this fit. The intercept for `africa` differs from the intercept for `oecd` because the coefficient estimate $\hat{\beta}_{03}$ is about 7.8 standard deviations from 0. No test is provided here of the difference in intercept between `other` and `africa`. Multiple testing comparing levels of `group` follow the general prescription given in Section 5.1.2, except that the comparisons are made between the means adjusted for $\log(\text{ppgdp})$.

The effects plots for (5.8) are shown in Figure 5.4. For the more complicated interaction model, there was only one effects plots with separate curves for each level of `group`. In the main effects model, separate plots are drawn for each effect. The effects plot for `group` displays fitted values with the other regressors in the model set to a fixed value. The default behavior will

depend on the software used. The default used in this book is to set the other variables to their mean value. Write \bar{x} for the mean of $\log(\text{ppgdp})$. Then the 3 plotted points in the effects plot for `group` are the fitted values $\hat{E}(\text{lifeExpF} | \log(\text{ppgdp}) = \bar{x}, \text{group} = j)$, for $j \in \{\text{oecd, other, africa}\}$. These are also the adjusted means that would be used in a multiple comparison procedure.

A point on the curve shown in the effect plot for `ppgdp` is a little more complicated because we need to fix the factor `group` at a “typical value.” The procedure used in this book to draw the graph is (1) compute a fitted value for each level of the factor; and (2) use a weighted average of these fitted values, with the weights determined by the sample size in each level of the factor. This is admittedly an arbitrary way of combining the levels of the factor, but using any sensible procedure would change the values on the vertical axis, but not the shape of the curve, and the shape is the important feature of the plot.⁶ As before, the effects plot for `ppgdp` is a curve because the model uses the regressor $\log(\text{ppgdp})$ but we display using the original untransformed predictor to get a curve. The main effects model says that the predictors can be interpreted separately. Level `africa` of `group` has a lower fitted value than the other two levels. The predictor `ppgdp` is associated with increased fitted `lifeExpF`, with the greatest effect for values of `ppgdp` less than about \$15,000.

5.2 MANY FACTORS

Increasing the number of factors or the number of continuous predictors in a mean function can add considerably to complexity but does not really raise new fundamental issues. Consider first a problem with many factors but no continuous predictors. The data in the file `Wool` are from a small experiment to understand the strength of wool as a function of three factors that were under the control of the experimenter (Box and Cox, 1964). The variables are summarized in Table 5.5. Each of the three factors was set to one of three

Table 5.5 The Wool Data

Variable	Definition
<code>len</code>	Length of test specimen (250, 300, 350 mm)
<code>amp</code>	Amplitude of loading cycle (8, 9, 10 mm)
<code>load</code>	Load put on the specimen (40, 45, 50 g)
<code>log(cycles)</code>	Logarithm of the number of cycles until the specimen fails

⁶The plotted quantities in effects plots can be viewed as adjusted means. The effects plots in this book use the default adjustments in the `effects` package in R (Fox, 2003). SAS and many other programs use `lsmeans` (SAS Institute, Inc., 2013), which can produce somewhat different adjustments.

levels, and all $3^3 = 27$ possible combinations of the three factors were used exactly once in the experiment, so we have a single replication of a 3^3 design. The response variable `log(cycles)` is the logarithm of the number of loading cycles to failure of worsted yarn. We will treat each of the three predictors as a factor with 3 levels.

A main effects mean function for these data includes an intercept and two dummy variables for each of the factors, for a total of seven parameters. A full *second-order* mean function adds all the two-factor interactions to the mean function. The interaction between two factors is obtained by multiplying each of the dummy variables for the first factor by each of the dummy variables for the second factor, so in this experiment a two-factor interaction requires $2 \times 2 = 4$ regressors. The second-order model will have $7 + 3 \times 4 = 19$ parameters. The third-order model includes the three-factor interaction with $2 \times 2 \times 2 = 8$ dummy variables for a total of $19 + 8 = 27$ parameters. This latter mean function will fit the data exactly because it has as many parameters as data points.

Wilkinson–Rogers's specification of these three mean functions are, assuming that `len`, `amp`, and `load` have all been declared as factors,

```
log(cycles) ~ len + amp + load
log(cycles) ~ len + amp + load + len:amp + len:load + amp:load
log(cycles) ~ len + amp + load + len:amp + len:load + amp:load
    + len:amp:load
```

Other mean functions can be obtained by dropping some of the two-factor interactions.

Mean functions with only factors and interactions are often called *analysis of variance models* after the type of analysis that is generally applied. These models are discussed more completely in experimental design books such as Oehlert (2000) or Montgomery (2012). Analysis of variance models are really a subset of multiple linear regression models. We discuss the analysis of variance method in Chapter 6. Analysis of the wool data is continued in Problems 5.19 and 8.6.

5.3 POLYNOMIAL REGRESSION

If a mean function with one predictor X is smooth but not straight, integer powers of the predictors can be used to approximate $E(Y|X)$. The simplest example of this is *quadratic regression*, in which the mean function is

$$E(Y|X = x) = \beta_0 + \beta_1 x + \beta_2 x^2 \quad (5.9)$$

Depending on the signs of the β s, a quadratic mean function can look like either of curves shown in Figure 5.5. Quadratic mean functions can therefore

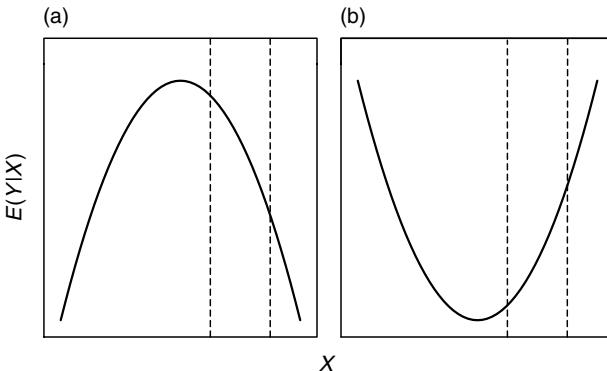


Figure 5.5 Generic quadratic curves. A quadratic is the simplest curve that can approximate a mean function with a minimum or maximum within the range of possible values of the predictor. It can also be used to approximate some nonlinear functions without a minimum or maximum in the range of interest, possibly using the part of the curve between the dashed lines.

be used when the mean is expected to have a minimum or maximum in the range of the predictor. The minimum or maximum will occur for the value of X for which the derivative $dE(Y|X = x)/dx = 0$, which occurs at

$$x_M = -\beta_1/(2\beta_2) \quad (5.10)$$

x_M is estimated by substituting estimates for the β s into (5.10).⁷

Quadratics can also be used when the mean function is curved but does not have a minimum or maximum within the range of the predictor. Referring to Figure 5.5a, if the range of X is between the dashed lines, then the mean function is everywhere decreasing but not linear, while in Figure 5.5b it is increasing but not linear. In these cases, however, using polynomials can lead to nonsensical answers when a fitted model is applied for new values of the predictors outside the range of the observed data.

Quadratic regression is an important special case of *polynomial regression*. The polynomial mean function of degree d with one predictor is

$$E(Y|X = x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_d x^d \quad (5.11)$$

If $d = 2$, the model is quadratic, $d = 3$ is cubic, and so on. Any smooth function can be estimated by a polynomial of high-enough degree. Polynomial mean functions are generally used as approximations and rarely represent a physical model.

⁷The standard error of a nonlinear function of parameters can be computed with the delta method, Section 7.6.

5.3.1 Polynomials with Several Predictors

With more than one predictor, we can contemplate having integer powers and products of all the predictors as regressors in the mean function. For example, for the important special case of two predictors, the *second-order mean function* is given by

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 \quad (5.12)$$

The new regressor in (5.12) is the multiplicative interaction $x_1 x_2$. With k predictors, the second-order model includes an intercept, k linear regressors, k quadratic regressors, and $k(k - 1)/2$ interaction regressors. If $k = 5$, the second-order mean function has 21 regressors, and with $k = 10$, it has 66 regressors. A usual strategy is to view the second-order model as consisting of too many regressors and use testing or other selection strategies such as those to be outlined in Section 10.2.1 to delete regressors for unneeded quadratics and interactions. Without the interaction regressors, the effect of each predictor is the same regardless of the values of the other predictors. With the interaction, the effect of a predictor can change depending on the values of the other predictors.

Cakes

Oehlert (2000, Example 19.3) provides data from a small experiment with $n = 14$ observations on baking packaged cake mixes. Two factors, X_1 = baking time in minutes and X_2 = baking temperature in degrees F, were varied in the experiment. The response Y was the average palatability score of four cakes baked at a given combination of (X_1, X_2) , with higher values desirable.

The estimated mean function based on (5.12) and using the data in the file *cakes* is

$$\begin{aligned} E(Y|X_1 = x_1, X_2 = x_2) &= -2204.485 + 25.9176x_1 + 9.9183x_2 \\ &\quad - 0.1569x_1^2 - 0.012x_2^2 - 0.0416x_1 x_2 \end{aligned} \quad (5.13)$$

Each of the coefficient estimates, including both quadratics and the interaction, has significance level of 0.005 or less, so all regressors are useful in the mean function (see Problem 5.8).

Effects plots provide graphical summary of the fit, as in Figure 5.6. In Figure 5.6a, the horizontal axis is the baking time X_1 , and the vertical axis is the fitted response \hat{Y} . The three curves shown on the graph are obtained by fixing the value of temperature X_2 at either 340, 350, or 360, and substituting into (5.13). For example, when $X_2 = 350$, substitute 350 for X_2 in (5.13), and simplify to get

$$\begin{aligned} E(Y|X_1 = x_1, X_2 = 350) &= \hat{\beta}_0 + \hat{\beta}_2(350) + \hat{\beta}_{22}(350)^2 + \hat{\beta}_1 x_1 + \hat{\beta}_{12}(350)x_1 + \hat{\beta}_{11}x_1^2 \\ &= -203.08 + 11.36x_1 - 0.16x_1^2 \end{aligned} \quad (5.14)$$

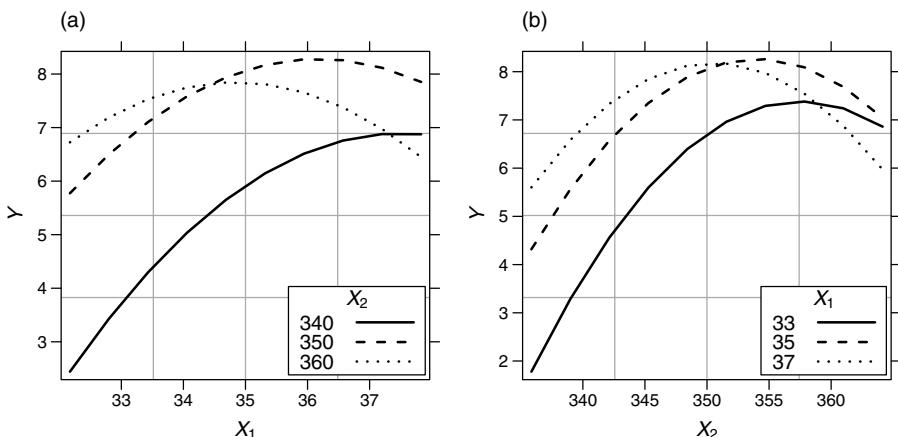


Figure 5.6 Effects plots for the cakes data, based on (5.13). Both plots show the same effects, in (a) with X_1 on the x -axis and levels of X_2 indicated by separate curves, and in (b) with X_2 on the x -axis and levels of X_1 indicated by separate curves.

Equation (5.14) is plotted as the dashed line in Figure 5.6a. Each of the lines shown is a quadratic curve because both X_1^2 and X_2^2 regressors are in the mean function. Each of the curves has a somewhat different shape because the interaction is present. For example, in Figure 5.6a, the baking time X_1 that maximizes the response is lower at $X_2 = 360$ degrees than it is at $X_2 = 340$ degrees. Figure 5.6b contains the same information as Figure 5.6a except that the roles of X_1 and X_2 have been reversed with X_2 on the horizontal axis and fixed values of $X_1 \in \{33, 35, 37\}$ providing the 3 curves. The response curves are about the same for baking time of 35 or 37 minutes, but the response is lower at the shorter baking time. The palatability score is perhaps surprisingly sensitive to changes in temperature of 10 or 15 degrees and baking times of just a few minutes.

5.3.2 Numerical Issues with Polynomials

Numerical problems can arise when using polynomial regressors in a regression. The first problem is that regressors X^d and X^{d+1} can be very highly correlated, and high correlations can cause inaccurate computation of the ols estimator. A second problem is that computers have only a finite number of digits to represent a number, and in some problems X^d can be so large (or, if $|X| < 1$, so small) that significant round-off error occurs.

A solution to these computing problems is to use orthogonal polynomials to define the polynomial regressors. For example, for fitting with a cubic polynomial with regressors X , X^2 , and X^3 , we would fit with regressors $Q_1 = X - \bar{X}$, the residuals Q_2 from the regression of X^2 on Q_1 , and the residuals Q_3 from the regression of X^3 on Q_1 and Q_2 . The Q s are then rescaled to have

unit length. The resulting Q_j are uncorrelated and have elements that by the rescaling are neither too small nor too large, and so replacing (X, X^2, X^3) by the rescaled Q_j avoids numerical problems.⁸

Most statistical packages will automatically orthogonalize before computing, so this need not be a concern of the user. If you write your own software, however, you should take care to avoid numerical problems.

5.4 SPLINES

Figure 5.7 shows polynomial fits of degree $d = 1, 2, 3, 4$ for the Old Faithful Geyser data (Problem 1.4). The predictor in this problem is `Duration`, the length of the current eruption of the geyser in seconds, and the response is the time `Interval` in minutes until the next eruption. The data fall in two clusters, with little data between the two clusters. The $d = 1$ fit in Figure 5.7a predicts `Interval` increasing linearly with `Duration`. The $d = 2$ fit in Figure 5.7b flattens out the predictions in the cluster with larger values of `Duration` but doesn't effect the cluster with smaller values of `Duration`. The predictions for $d = 3$ flatten out in the larger cluster even more than the $d = 2$ fit, but add the undesirable and unlikely feature of decreasing predictions for the largest values of `Duration`. The $d = 4$ fit flattens the predictions in both clusters but has an undesirable increase for the smallest values of `Duration` and an unlikely increase in slope for the largest values of `Duration`. These figures demonstrate that increasing the dimension d of the polynomial can make some aspects of a fitted curve better, but it can also make other aspects of the fitted curve worse.

A polynomial fit is really just a weighted sum of *basis functions*,

$$\mathbb{E}(Y|X = x) = \beta_0 + \sum_{j=1}^d \beta_j x^j$$

The basis functions are the monomials $\{x = x^1, x^2, \dots, x^d\}$, and the weights are the β s. Since the monomials are defined for all possible values of X , they are best for modeling global behavior of a function, but may not be very useful for modeling local behavior, as would be desirable for the data in Figure 5.7.

Splines provide a different set of basis functions, each of which acts locally, so changing the weight for one of the basis functions will mostly effect the fitted curve only for a limited range. Figure 5.8 shows a set of $d = 6$ basis functions $b_1(x), \dots, b_6(x)$ for a hypothetical predictor x with values between -1 and $+1$. In this book we will use cubic B-splines to define the basis, although the literature includes many other options.⁹ The first or leftmost of these basis

⁸The resulting Q_j are the QR-factorization of $(X - \bar{X}, X^2, X^3)$; see Appendix A.9.

⁹The algebraic form for cubic B-splines is not particularly enlightening and is omitted; see de Boor (1978). Many computer programs will include functions for computing the cubic B-spline basis.

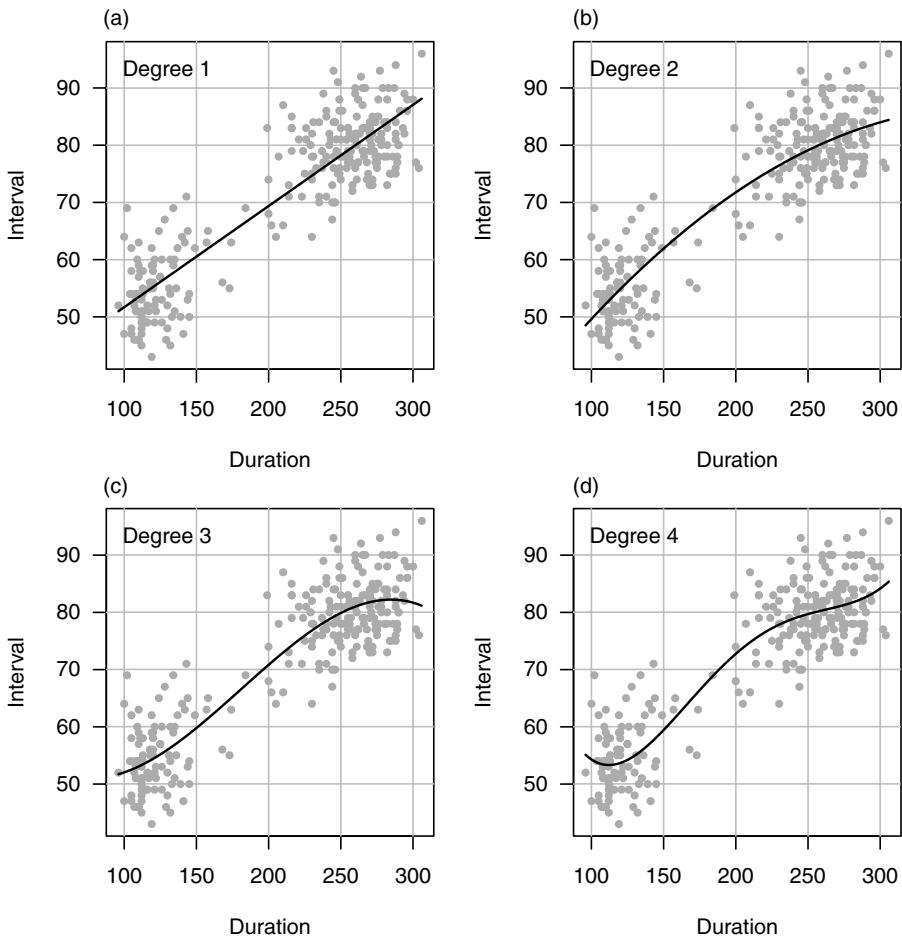


Figure 5.7 Polynomial fits for the Old Faithful Geyser data.

functions $b_1(x)$ is mostly concentrated on $x \in (-1, -0.5)$, and so changing the weight on this basis function will mostly change the fitted curve for small values of x . The third basis function is largest for $x \approx 0$ and decreases symmetrically around 0. The last basis function is concentrated for values of x close to 1. For $x \approx 0.5$, for example, the fit will be mostly determined by the weights for b_4 and b_5 , and to a lesser extent b_3 because these are the only basis functions that are substantially different from 0 at $x = 0.5$.

As long as we can compute the spline basis we can use OLS or other standard methods to fit the mean function

$$\mathbb{E}(Y|X = x) = \beta_0 + \sum_{j=1}^d \beta_j b_j(x) \quad (5.15)$$

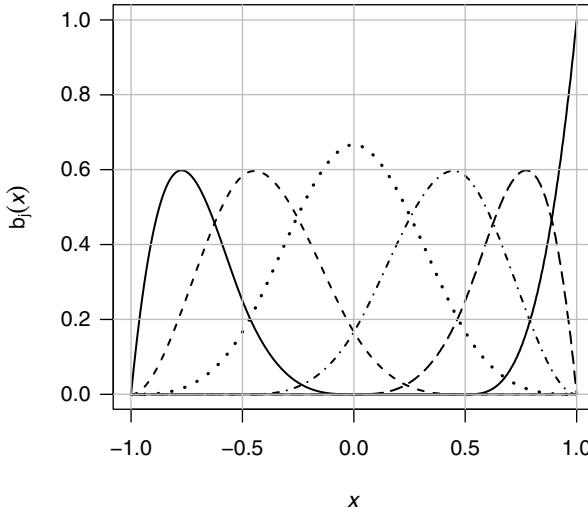


Figure 5.8 The B-spline basis with $d = 6$ basis functions.

where d is the number of splines in the basis. Using the cubic B-spline basis, we can hope to fit local features of a mean function without distorting features for a different part of the range of the predictors.

Figure 5.9 shows four cubic B-spline fits to the Old Faithful Geyser data, with varying number of basis functions. These fits are more nearly consistent with the idea that the within-cluster predictions should be nearly constant, but with between-cluster differences. All the fits have trouble with the extremes of the range of the predictor because there is no information outside the range to temper the very local information at the extremes.

5.4.1 Choosing a Spline Basis

With polynomials, the user can choose the degree d of polynomial as a smoothing parameter. With cubic B-splines, the number of vectors in the basis, which we also call d , is a similar smoothing parameter. B-splines in general have more smoothing parameters, including the relative width and center of each of the basis functions, but for the purposes of this book, selecting d will generally provide adequate flexibility.

In most problems, setting $d = 3$ or $d = 4$ will allow matching many functions that could be encountered in practice. Wood (2006) provides a comprehensive reference for using splines to fit regression models.

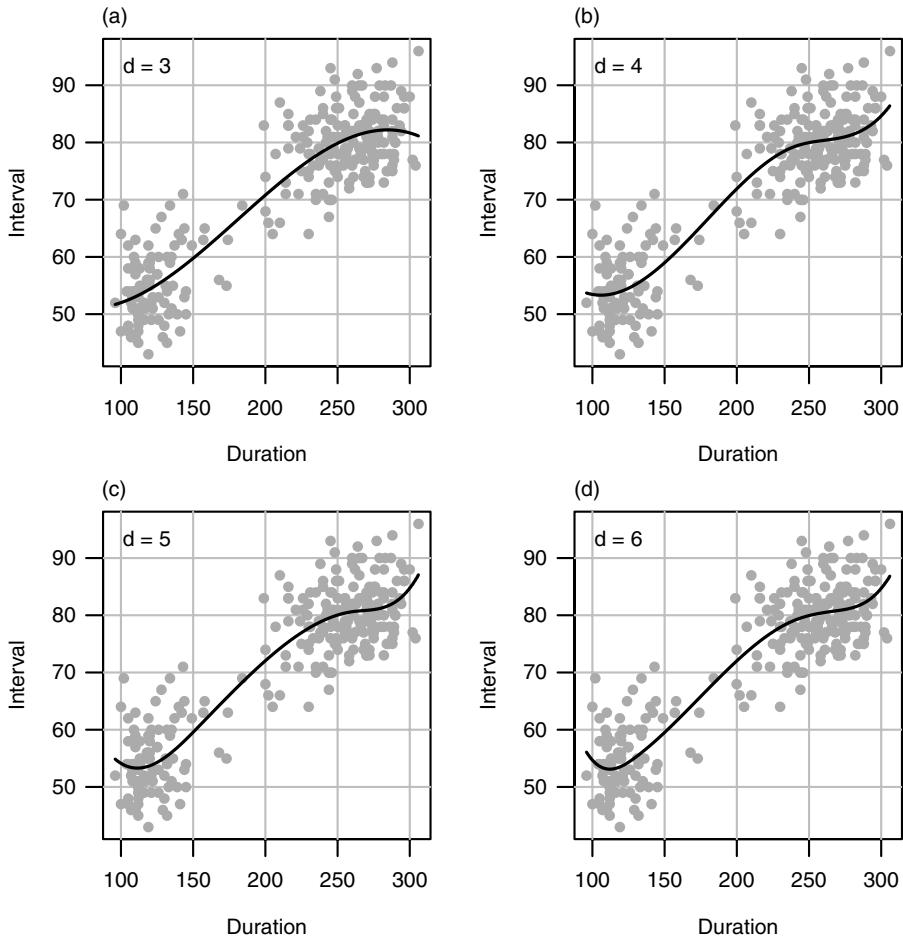


Figure 5.9 Spline fits for the Old Faithful geyser data.

5.4.2 Coefficient Estimates

Coefficient estimates for the β s for the spline basis in (5.15) are rarely of interest, and the useful summaries are necessarily graphical as in Figure 5.9.

5.5 PRINCIPAL COMPONENTS

Suppose we have variables X_1, \dots, X_k with k large, although the same methodology applies for any $k \geq 2$. Our goal is to replace the k variables with $k_0 < k$ linear combinations of them such that the smaller set of variables represents the larger set as closely as possible. We start with $k_0 = 1$, so our goal is to replace

the k predictors with 1 linear combination. Let $\mathbf{X}' = (X_1, \dots, X_k)$ be the variables written as a vector, and let \mathbf{u}_1 be a $p \times 1$ vector of constants, subject to the constraint that $\mathbf{u}_1' \mathbf{u}_1 = 1$. The first principal component will be a linear combination $Z_1 = \mathbf{u}_1' \mathbf{X}$ such that the variance of Z_1 ,

$$\text{Var}(Z_1) = \text{Var}(\mathbf{u}_1' \mathbf{X}) = \mathbf{u}_1' \text{Var}(\mathbf{X}) \mathbf{u}_1 \quad (5.16)$$

is as large as possible to retain as much as the variation in the predictors as possible. If $\text{Var}(\mathbf{X})$ were known, then as sketched in Appendix A.10, this is a standard problem in linear algebra, and the solution is to set \mathbf{u}_1 to be the eigenvector corresponding to the largest eigenvalue of $\text{Var}(\mathbf{X})$. For a solution with k_0 principal components, the linear combinations are the eigenvectors corresponding to the k_0 largest eigenvalues.

In the usual case, $\text{Var}(\mathbf{X})$ is unknown, and the sample covariance matrix is used in place of the unknown variance matrix. We use the notation $\hat{\mathbf{u}}_j$ to refer to the j th eigenvector and $\hat{\lambda}_j$ as the corresponding eigenvalue of the sample covariance matrix. To simplify the presentation, we assume no two eigenvalues are the same.

Professor Ratings

The data used in Problem 1.6 in the file `Rateprof` on professor ratings from the website `RateMyProfessor.com` includes averages of many student ratings for each instructor on five different measures, including quality, helpfulness, clarity, easiness of the course, and raterInterest in the subject matter. All the ratings were on a five-point scale, so the averages are numbers between 1 and 5. The scatterplot matrix of the ratings is shown in Figure 1.13, where we see that the first three ratings very highly correlated, and the remaining two ratings are less highly correlated with each other and with the first three ratings. Using principal components would replace these five ratings by linear combinations of them.

Computing can be done with software designed for principal component analysis or with more general software that finds the eigenvalues and eigenvectors of a matrix. Typical summary output from a special-purpose program is shown in Table 5.6. The upper part of the table labeled “Importance of components” refers to the eigenvalues $\hat{\lambda}_j$. By construction $\hat{\lambda}_j$ is the estimated variance of the j th principal component, and these variances are shown in the first row. The variance for the first principal component is considerably larger than the others. The next two rows of the first part of the table summarize the relative importance of the principal components. The second row is $\hat{\lambda}_j / \sum_{m=1}^k \hat{\lambda}_m$, the estimated fraction of the total variance in the data that is included in the j th principal component. The third row gives the cumulative proportion of variance in the first j principal components. This idea of “proportion of variance” is justified by the fact that the sum of the eigenvalues is equal to the sum of the diagonal elements of estimated variance matrix, or the sum of the estimated variances of the original data.

**Table 5.6 Principal Component Analysis for the Professor Ratings Data
Importance of Components $\hat{\lambda}_j$ (Eigenvalues)**

Component	1	2	3	4	5
Variance = $\hat{\lambda}_j$	2.39	0.39	0.22	0.06	0.00
Proportion of Variance	0.78	0.13	0.07	0.02	0.00
Cumulative Proportion	0.78	0.91	0.98	1.00	1.00
Linear combinations $\hat{\mathbf{u}}_j$ (eigenvectors)					
	$\hat{\mathbf{u}}_1$	$\hat{\mathbf{u}}_2$	$\hat{\mathbf{u}}_3$	$\hat{\mathbf{u}}_4$	$\hat{\mathbf{u}}_5$
quality	-0.535	-0.155	0.150	-0.046	-0.815
helpfulness	-0.529	-0.136	0.136	-0.701	0.438
clarity	-0.537	-0.188	0.167	0.711	0.379
easiness	-0.336	0.916	-0.215	0.037	0.005
raterInterest	-0.181	-0.287	-0.941	0.009	-0.001

In this example, about 78% of the variance in the five ratings is captured by the first principal component. The second principal component increases the variance to nearly 91% of the total, and three components capture nearly all the variation. This finding agrees well with examination of Figure 1.13: the first three ratings are measuring essentially the same thing, and so they can be replaced without much loss by any one of them or by any linear combination of them. The remaining two ratings are distinct, leaving three useful components.

The columns of the second part of Table 5.6 give the $\hat{\mathbf{u}}_j$. The vector $\hat{\mathbf{u}}_1$ gives almost equal weight to the first three scales, and lower weight to the remaining scales. With distinct eigenvalues, the $\hat{\mathbf{u}}_j$ are unique only up to multiplication by ± 1 , so the signs of the weights shown could all be changed to + signs. If we ignore the ratings with the smaller weights, the first principal component is essentially the sum, or average, of the first three ratings. The eigenvector $\hat{\mathbf{u}}_2$ is essentially for the rating easiness because the weight for this rating is close to 1; recall that by construction, the sum of the squares of the weights is always 1. Similarly, the third component is essentially for raterInterest.

5.5.1 Using Principal Components

Principal components are sometimes used in regression problems to replace several variables by just a few linear combinations of them. If we write \mathbf{X} to be the vector of the five ratings in the professor rating data, we might choose to use $Z_j = \mathbf{X}'\hat{\mathbf{u}}_j$ for $j = 1, 2, 3$ as regressors in the model. In this particular problem we might choose to use the three regressors consisting the average of the first three ratings, easiness and raterInterest, because these regressors are much easier to interpret, but not all problems will permit this simple explanation. Similarly, if the ratings were responses, using principal

components has reduced the number of responses from five to just three or less, and this too can simplify a problem.

5.5.2 Scaling

Unlike almost all the methods in this book, principal components will change depending on the scale of each of the predictors that is used. In the professor ratings data, all the ratings were on the range 1 to 5, and all standard deviations of all the ratings are similar. In other problems, the standard deviations of the predictors can be very different. For example, in the Berkeley Guidance Study, variables include heights, weights, and leg strengths, all measured in different units with very different standard deviations. Using principal components on the raw data will generally give more weight to variables with larger variances.

The standard “solution” to the scaling problem with principal components is to replace each of the original predictors by a standardized version obtained by dividing by the sample standard deviations. The sample correlation matrix is then used to find the principal components. This solution is not without problems. If the data at hand are not a random sample from a population, then the sample standard deviations used to standardize the variables will not estimate a population quantity, and so the variables are now measured on some arbitrary scale that depends on the sampling design. Experimenters who collect data in different ways will end up with different standard deviations and eventually different principal components. Thus, reproducibility of results based on principal components can be questionable without strong assumptions about the sampling plan used to collect the data.

5.6 MISSING DATA

In many problems, some variables will be unrecorded for some cases. The methods we study in this book generally assume and require complete data, without any missing values. The literature on analyzing incomplete data problems is very large, and our goal here is more to point out the issues than to provide solutions. Two important books on this topic are by Little and Rubin (2002) and Schafer (1997). Survey articles include Allison (2001) and Schafer and Graham (2002).

Minnesota Agricultural Land Sales

The data file `MinnLand` includes information on nearly every agricultural land sale in the six major agricultural regions of the state of Minnesota for the period 2002–2011, a total of 18,700 sales. The data were collected from the Minnesota Department of Revenue to study the effect of enrollment of land in the U.S. Conservation Reserve Program (CRP) (Taff and Weisberg, 2007). The CRP is a voluntary program in which farmers commit environmentally

sensitive land for conservation usage in exchange for a fixed payment. The period of this agreement, also called an easement, is typically for 10–15 years. The land owner or purchaser of a property with a CRP easement cannot change the use of the land until the easement expires.

The model $\log(\text{acrePrice}) \sim \text{year} * \text{region} + \text{crpPct} + \text{financing}$ was fit, where the variable `crpPct` is the percentage of the total parcel that is committed to a CRP easement at the time of sale, `financing` is an indicator of whether the sale was owner-financed, `region` is a factor with six levels for the six economic regions of the state included in the data, and `year` is a factor for years. The response variable $\log(\text{acrePrice})$ is the logarithm of the sale price per acre of the land adjusted to a common day within the year to account for seasonal and within-year changes in prices.

The row labeled Model 1 of Table 5.7 shows a 95% confidence interval for the coefficient `crpPct`. According to this model, a 1% increase in land committed to CRP is associated with about 0.59–0.51% lower per acre price; a 50% commitment to CRP is associated with lower value about 50 times this interval, from about 29.5% lower to 25.5% lower.

One possible explanation for this very large effect is that farmers with less valuable land could have more to gain from enrollment in CRP, so the apparent CRP effect could really be a land quality effect. Another variable in the database is `productivity`, a score between 1 and 100 based on University of Minnesota soil studies. Higher values should correspond to more valuable land. The variable `productivity` is missing for 9717 of the records in the data, and so Model 2 in the second row in Table 5.7, which fits $\log(\text{acrePrice}) \sim \text{year} * \text{region} + \text{crpPct} + \text{financing} + \text{productivity}$, is based on the 8983 complete cases. The apparent effect of `crpPct` adjusted for `productivity` as well as `year` and `region` is smaller than in Model 1, but still quite large. Does omitting more than half the data make any sense?

5.6.1 Missing at Random

The most common solution to missing data problems is to delete either cases or variables so the resulting data set is complete, as done in Table 5.7. Most software packages delete partially missing cases by default and fit regression models to the remaining, complete, cases. This is a reasonable approach as long as the fraction of cases deleted is small enough, and the cause of values being

Table 5.7 Confidence Intervals for `crpPct`

	2.5%	97.5%
Model 1	-0.0059	-0.0051
Model 2	-0.0046	-0.0036
Model 3	-0.0058	-0.0050

unobserved is unrelated to the relationships under study. This would include data lost through an accident like dropping a test tube, or making an illegible entry in a logbook. If the reason for not observing values depends on the values that would have been observed, then the analysis of data may require modeling the cause of the failure to observe values. For example, if values of a measurement are unrecorded if the value is less than the minimum detection limit of an instrument, then the value is missing because the value that should have been observed is too small. A simple expedient in this case that is sometimes helpful is to substitute a value less than or equal to the detection limit for the unobserved values. This expedient is not always entirely satisfactory because substituting, or imputing, a fixed value for the unobserved quantity can reduce the variation on the filled-in variable and yield misleading inferences.

As a second example, suppose we have a clinical trial that enrolls subjects with a particular medical condition, assigns each subject a treatment, and then the subjects are followed for a period of time to observe their response, which may be time until a particular landmark occurs, such as improvement of the medical condition. Subjects who do not respond well to the treatment may drop out of the study early, while subjects who do well may be more likely to remain in the study. Since the probability of observing a value depends on the value that would have been observed, simply deleting subjects who drop out early can easily lead to incorrect inferences because the successful subjects will be overrepresented among those who complete the study.

In some studies, the response variable is not observed because the study ends, not because of patient characteristics. In this case, we call the response times *censored*, and for each patient we know either the time to the landmark or the time to censoring. This is a different type of missing data problem, and analysis needs to include both the uncensored and censored observations. Many book-length treatments of censored survival data are available, including Hosmer et al. (2008).

As a final example, consider a cross-cultural demographic study. Some demographic variables are harder to measure than others, and some variables, such as the rate of employment for women over the age of 15, may not be available for less-developed countries. Deleting countries that do not have this variable measured could change the population that is studied by excluding less-developed countries.

Rubin (1976) defined data to be *missing at random* (MAR) if the failure to observe a value does not depend on the value that would have been observed. With MAR data, case deletion can be a useful option. Determining whether an assumption of MAR is appropriate for a particular data set is an important step in the analysis of incomplete data.

In the Minnesota agricultural land sales example including the productivity variable reduces the sample size by more than half. The remaining sample is still quite large, and so the expedient of examining only fully observed cases could be reasonable here if the MAR assumption is

reasonable. The percentage of observations with productivity observed was between 20.8% in the Northwest region and 95.4% in the Southwest region. The Northwest region also had the lowest observed average $\log(\text{acrePrice})$. Missingness varies less by year, between 39% in 2004 and 54.8% in 2009.

Productivity scores can be reported only if they are computed in the first place. Counties had to pay the University for the productivity score, and not all counties in some of the regions chose to participate. It is at least plausible that the counties that did not participate have less valuable land, which would violate the MAR assumption. Model 3 in Table 5.7 is $\log(\text{acrePrice}) \sim \text{year} * \text{region} + \text{crpPct} + \text{financing} + \text{hasprod}$, where `hasprod` is a dummy indicator of 0 for observations for which productivity is missing and 1 if productivity is observed. The coefficient estimate for `crpPct` is essentially the same as the estimate in Model 1. The coefficient estimate for `hasprod` is 0.123, suggesting that sales with a productivity score reported were on average 12% higher priced. These analyses suggest that additional use of CRP is associated with lower per acre sales price, but quantifying the amount of change is not completely clear.

What exactly to do about missing data depends on the problem. There are many problems for which a textbook prescription is likely to be inadequate.

5.6.2 Imputation

An alternative to deleting cases with missing values that may be appropriate in some problems is to “fill in” the missing data with plausible values. For example, the web page of the U.S. Census (undated) explains methods the census uses to fill in missing values in the Current Population Survey. One of the methods they use is called the *hot deck*, in which a missing entry is filled in with a value from another individual with a similar record on other variables. This will permit standard estimation methods and standard computer programs to be used to process the data. As long as the fraction of missing values is relatively small, this procedure is likely to work well.

In regression problems, an attractive procedure is to fill in the missing values by fitting regression models. For example, to impute missing values for a particular predictor X_1 based on a set of other predictors X_2 , one could build a regression model for $E(X_1|X_2)$ based on complete data. The fitted model can be used to estimate a predicted value for X_1 based on X_2 for the cases for which X_1 is unobserved. In general this can give fill in values that are “too good” in the sense that the imputed values will be less variable than would the unobserved “true” values. A solution to this is using *multiple imputation*, in which several filled in data sets are created, a complete data analysis is performed for each data set, the results are averaged to get an overall analysis. Carpenter and Kenward (2012) provide many examples and a very useful companion website.

5.7 PROBLEMS

5.1 For a factor X with d categories, the one-factor mean function is

$$E(Y|U_2, \dots, U_d) = \beta_0 + \beta_2 U_2 + \dots + \beta_d U_d \quad (5.17)$$

where U_j is a dummy variable equal to 1 for the j th level of the factor and 0 otherwise.

5.1.1 Show that $\mu_1 = \beta_0$ is the mean for the first level of X and that $\mu_j = \beta_0 + \beta_j$ is the mean for all the remaining levels, $j = 2, \dots, d$.

5.1.2 It is convenient to use two subscripts to index the observations, so y_{ji} is the i th observation in level j of the factor, $j = 1, \dots, d$ and $i = 1, \dots, n_j$. The total sample size is $n = \sum n_j$. The residual sum of squares function can then be written as

$$\text{RSS}(\boldsymbol{\beta}) = \sum_{j=1}^d \sum_{i=1}^{n_j} (y_{ji} - \beta_0 - \beta_2 U_2 - \dots - \beta_d U_d)^2$$

Find the OLS estimates of the β s, and then show that the OLS estimates of the group means are $\hat{\mu}_j = \bar{y}_j$, $j = 1, \dots, d$, where \bar{y}_j is the average of the ys for the j th level of X .

5.1.3 Show that the residual sum of squares can be written

$$\text{RSS} = \sum_{j=1}^d (n_j - 1) \text{SD}_j^2$$

where SD_j is the standard deviation of the responses for the j th level of X . What is the df for RSS?

5.1.4 If all the n_j are equal, show that (1) the standard errors of $\hat{\beta}_2, \dots, \hat{\beta}_d$ are all equal, and (2) the standard error of $\hat{\beta}_0$ is equal to the standard error of each of $\hat{\beta}_0 + \hat{\beta}_j$, $j = 2, \dots, d$.

5.2 Verify the relationships between the η -parameters in (5.6) and the β -parameters in (5.7).

5.3 (Data file: UN11)

5.3.1 In the fit of `lifeExpF ~ group`, verify the results of Table 5.2.

5.3.2 Compare all adjusted mean differences in the levels of `group` in the model `lifeExpF ~ group + log(ppgpd)` with the results in Table 5.2.

5.4 (Data file: MinnLand) The data file includes information on nearly every agricultural land sale in the six major agricultural regions of Minnesota for the period 2002–2011. The data are from the Minnesota Department

Table 5.8 Minnesota Agricultural Land Sales

Variable	Definition
acrePrice	Sale price in dollars per acre, adjusted to a common date within year
year	Year of sale
acres	Size of property, acres
tillable	Percentage of farm rated arable
improvements	Percentage of property value due to buildings and other improvements
financing	Type of financing either title transfer or seller finance
crp	Enrolled of any part of the acreage is enrolled in the U.S. Conservation Reserve Program (CRP), and none otherwise
crpPct	Percentage of land in CRP
productivity	A numeric score between 1 and 100 with larger values indicating more productive land, calculated by the University of Minnesota

of Revenue and were provided by Steven Taff. Two of the variables in the data are `acrePrice`, the selling price per acre adjusted to a common date within a year, and `year`, the year of the sale. All the variables are described in Table 5.8.

- 5.4.1** Draw boxplots of $\log(\text{acrePrice})$ versus `year`, and summarize the information in the boxplots. In particular, housing sales prices in the United States were generally increasing from about 2002–2006, and then began to fall beginning in 2007 or so. Is that pattern apparently repeated in Minnesota farm sales?
- 5.4.2** Fit a regression model with $\log(\text{acrePrice})$ as the response and a factor representing the year. Provide an interpretation of the estimated parameters. Interpret the t -statistics. (*Hint:* Since `year` is numeric, you may need to turn it into a factor.)
- 5.4.3** Fit the regression model as in the last subproblem, but this time omit the intercept. Show that the parameter estimates are the means of $\log(\text{acrePrice})$ for each year. The standard error of the sample mean in year j is $SD_j / \sqrt{n_j}$, where SD_j and n_j are the sample standard deviation and sample size of the for the j th year. Show that the standard errors of the regression coefficients are not the same as these standard errors and explain why they are different.

- 5.5 Interpreting parameters with factors and interactions** Suppose we have a regression problem with a factor A with two levels (a_1, a_2) and a factor B with three levels (b_1, b_2, b_3), so there are six treatment combinations.

Suppose the response is Y , and further that $E(Y|A = a_i, B = b_j) = \mu_{ij}$. The estimated μ_{ij} are the quantities that are used in effects plots. The purpose of this problem is to relate the μ_{ij} to the parameters that are actually fit in models with factors and interactions.

- 5.5.1** Suppose the dummy regressors (see Section 5.1.1) for factor A are named (A_1, A_2) and the dummy regressors for factor B are named (B_1, B_2, B_3) . Write the mean function

$$E(Y|A = a_i, B = b_j) = \beta_0 + \beta_1 A_2 + \beta_2 B_2 + \beta_3 B_3 + \beta_4 A_2 B_2 + \beta_5 A_2 B_3$$

in Wilkinson–Rogers notation (e.g., (3.19) in Chapter 3).

- 5.5.2** The model in Problem 5.5.1 has six regression coefficients, including an intercept. Express the β s as functions of the μ_{ij} .

- 5.5.3** Repeat Problem 5.5.2, but start with $Y \sim A + B$.

- 5.5.4** We write $\mu_{ij} = (\mu_{1j} + \mu_{2j})/2$ to be the “main effect” of the j th level of factor B , obtained by averaging over the levels of factor A . For the model of Problem 5.5.2, show that the main effects of B depend on all six β -parameters. Show how the answer simplifies for the model of Problem 5.5.3.

- 5.5.5** Start with the model of Section 5.5.1. Suppose the combination (a_2, b_3) is not observed, so we have only five unique cell means. How are the β s related to the μ_{ij} ? What can be said about the main effects of factor B ?

- 5.6** The coding of factors into dummy variables described in the text is used by default in most regression software. Older sources, and sources that are primarily concerned with designed experiments, may use *effects coding* for the dummy variables. For a factor X with d levels $\{1, 2, \dots, d\}$ define $V_j, j = 1, \dots, d - 1$ with elements v_{ji} are given by:

$$v_{ji} = \begin{cases} 1 & i = j \\ -1 & i = d \\ 0 & \text{otherwise} \end{cases}$$

The mean function for the one-factor model is then

$$E(Y|V_1, \dots, V_{d-1}) = \eta_0 + \eta_1 V_1 + \dots + \eta_{d-1} V_{d-1} \quad (5.18)$$

- 5.6.1** Show that the mean for the j th level of the factor is $\eta_0 + \alpha_j$, where

$$\alpha_j = \begin{cases} \eta_j & j \neq d \\ -(\eta_1 + \eta_2 + \dots + \eta_{d-1}) & j = d \end{cases}$$

By taking the mean of the level means show that η_0 is the mean of the response ignoring the factor. Thus, we can interpret α_j , the difference between the overall mean and the level mean, as the effect of level j , and $\sum \alpha_j = 0$.

- 5.7** Suppose X_1 were a continuous predictor, and F is a factor with three levels, represented by two dummy variables X_2 with values equal to 1 for the second level of F and X_3 with values equal to 1 for the third level of F . The response is Y . Consider three mean functions:

$$E(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad (5.19)$$

$$E(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 \quad (5.20)$$

$$E(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1(x_1 - \delta) + \beta_{12}(x_1 - \delta)x_2 + \beta_{13}(x_1 - \delta)x_3 \quad (5.21)$$

Equation (5.21) includes an additional unknown parameter δ that may need to be estimated.

All of these mean functions specify that for a given level of F the plot of $E(Y|X_1, F)$ is a straight line, but in each the slope and the intercept changes. For each of these three mean functions, determine the slope(s) and intercept(s), and on a plot of Y on the vertical axis and X_1 on the horizontal axis, sketch the three fitted lines.

The model (5.21) is a generalization of (5.20). Because of the extra parameter δ that multiplies some of the β s, this is a nonlinear model; see Saw (1966) for a discussion.

5.8 Cake data (Data file: cakes)

5.8.1 Fit (5.12) and verify that the significance levels for the quadratic terms and the interaction are all less than 0.005. When fitting polynomials, tests concerning main effects in models that include a quadratic are generally not of much interest.

5.8.2 The cake experiment was carried out in two blocks of seven observations each. It is possible that the response might differ by block. For example, if the blocks were different days, then differences in air temperature or humidity when the cakes were mixed might have some effect on Y . We can allow for block effects by adding a factor for block to the mean function and possibly allowing for block by regressor interactions. Add block effects to the mean function fit in Section 5.3.1 and summarize results. The blocking is indicated by the variable `Block` in the data file.

5.9 (Data file: salarygov) The data file gives the maximum monthly salary for 495 nonunionized job classes in a midwestern governmental unit in 1986. The variables are described in Table 5.9.

Table 5.9 The Governmental Salary Data

Variable	Description
MaxSalary	Maximum salary in dollars for employees in this job class, the response
NE	Total number of employees currently employed in this job class
NW	Number of women employees in the job class
Score	Score for job class based on difficulty, skill level, training requirements and level of responsibility as determined by a consultant to the governmental unit. This value for these data is in the range between 82 and 1017.
JobClass	Name of the job class; a few names were illegible or partly illegible

- 5.9.1** Examine the scatterplot of `MaxSalary` versus `Score`, and verify that simple regression provides a poor description of this figure.
- 5.9.2** Fit the regression with response `MaxSalary` and regressors given by B-splines, with d given by 4, 5, and 10. Draw the fitted curves on a figure with the data and comment.
- 5.9.3** According to Minnesota statutes, and probably laws in other states as well, a job class is considered to be female dominated if 70% of the employees or more in the job class are female. These data were collected to examine whether female-dominated positions are compensated at a lower level, adjusting for `Score`, than are other positions. Create a factor with two levels that divides the job classes into female dominated or not. Then, fit a model that allows for a separate B-spline for `Score` for each of the two groups. Since the coefficient estimates for the B-splines are uninterpretable, summarize the results using an effects plot. If your program does not allow you to use B-splines, use quadratic polynomials.
- 5.10** (Data file: `MinnLand`) Refer to Problem 5.4. Another variable in this data file is the `region`, a factor with six levels that are geographic identifiers.
- 5.10.1** Assuming both `year` and `region` are factors, consider the two mean functions given in Wilkinson–Rogers notation as:
- (a) $\log(\text{acrePrice}) \sim \text{year} + \text{region}$
 - (b) $\log(\text{acrePrice}) \sim \text{year} + \text{region} + \text{year:region}$
- Explain the difference between these two models (no fitting is required for this problem).
- 5.10.2** Fit model (b). Examining the coefficients of this model is unpleasant because there are so many of them, and summaries either

using graphs or using tests are required. We defer tests until the next chapter. Draw an effects plot for the year by region interaction and summarize the graph or graphs.

5.11 (Data file: MinnLand) This is a continuation of Problem 5.10. Another variable in the `MinnLand` data is the type of financing for the sale, a factor with levels `seller_financed` for sales in which the seller provides a loan to the buyer, and `title_transfer` in which financing of the sale does not involve the seller.

5.11.1 Add the variable `financing` to model (b) in Problem 5.10, and obtain and interpret a 95% confidence interval for the effect of financing.

5.11.2 Comment on each of the following statements:

1. Seller financing lowers sale prices.
2. Seller financing is more likely on lower-priced property transactions.

5.12 (Data file: lathe1) The data in the file `lathe1` are the results of an experiment on characterizing the life of a drill bit in cutting steel on a lathe. Two factors were varied in the experiment, Speed and Feed rate. The response is `Life`, the total time until the drill bit fails, in minutes. The values of `Speed` and `Feed` in the data have been coded by computing

$$\text{Speed} = \frac{(\text{Actual speed in feet per minute} - 900)}{300}$$

$$\text{Feed} = \frac{(\text{Actual feed rate in thousandths of an inch per revolution} - 13)}{6}$$

The coded variables are centered at zero. Coding has no material effect on the analysis but can be convenient in interpreting coefficient estimates.

5.12.1 Draw a scatterplot matrix of `Speed`, `Feed`, `Life`, and `log(Life)`, the logarithm of tool life. Add a little jittering to `Speed` and `Feed` to reveal overplotting. The plot of `Speed` versus `Feed` gives a picture of the experimental design, which is called a *central composite design*. It is useful when we are trying to find a value of the factors that maximizes or minimizes the response. Also, several of the experimental conditions were replicated, allowing for an estimate of variance and lack-of-fit testing. Comment on the scatterplot matrix.

5.12.2 For experiments in which the response is a time to failure or time to event, the response often needs to be transformed to a more

useful scale, typically by taking the log of the response, or sometimes by taking the inverse. For this experiment, log scale can be shown to be appropriate (Problem 9.15). Fit the full second-order mean function (5.12) to these data using $\log(\text{Life})$ as the response. Find the fitted equation, and obtain tests for the quadratic and interaction regressors.

- 5.12.3** Draw appropriate summary graphs for the fitted model. If either of the quadratics or the interaction is unnecessary, drop it and refit before drawing graphs.
- 5.13** (Data files: `Forbes` and `Hooker`) Refer to the data in Problem 2.7. Assuming equal intercepts, obtain tests of equality of slopes for the two sources of observations. How does the test change if the suspected outlier in Forbes's data, case 12, is removed?
- 5.14** (Data file: `BGSall`) Refer to the Berkeley Guidance study described in Problem 3.3. Using the data file `BGSall`, consider the regression of `HT18` on `HT9` and the grouping factor `Sex`.
- 5.14.1** Draw the scatterplot of `HT18` versus `HT9`, using a different symbol for males and females. Comment on the information in the graph about an appropriate mean function for these data.
- 5.14.2** Obtain the appropriate test for a parallel regression model.
- 5.14.3** Assuming the parallel regression model is adequate, estimate a 95% confidence interval for the difference between males and females. For the parallel regression model, this is the difference in the intercepts of the two groups.
- 5.15** (Data file: `BGSall`) Continuing with Problem 5.14, consider the response `HT18` and the continuous predictors `HT2` and `HT9` and the factor `Sex`. Explain the meaning of each of the following models, written in Wilkinson–Rogers notation:
- $HT18 \sim 1 + HT2 + HT9 + Sex$
 - $HT18 \sim 1 + HT2 + HT9 + Sex + Sex:HT2 + Sex:HT9$
 - $HT18 \sim 1 + HT2 + HT9 + HT2:HT9 + Sex + Sex:HT2 + Sex:HT9 + Sex:HT2:HT9$
- 5.16 Gothic and Romanesque cathedrals** (Data file: `cathedral`) The data file gives `Height` = nave height and `Length` = total length, both in feet, for medieval English cathedrals. The cathedrals can be classified according to their architectural style, either Romanesque or the later Gothic style. Some cathedrals have both a Gothic and a Romanesque part, each of differing height; these cathedrals are included twice. Names of the cathedrals are also provided in the file. The data were provided by Stephen Jay Gould based on plans given by Clapham (1934).

- 5.16.1** For these data, it is useful to draw *separate* plots of Length versus Height for each architectural style. Summarize the differences apparent in the graphs in the regressions of Length on Height for the two styles. Include in your graph the fitted simple and quadratic regressions.
- 5.16.2** Use the data to obtain tests that verify the visual results from the graphs.
- 5.17** **Sex discrimination** (Data file: salary) The data file concerns salary and other characteristics of all faculty in a small Midwestern college collected in the early 1980s for presentation in legal proceedings for which discrimination against women in salary was at issue. All persons in the data hold tenured or tenure track positions; temporary faculty are not included. The variables include degree, a factor with levels PhD and MS; rank, a factor with levels Asst, Assoc, and Prof; sex, a factor with levels Male and Female; Year, years in current rank; ysdeg, years since highest degree, and salary, academic year salary in dollars.
- 5.17.1** Get appropriate graphical summaries of the data and discuss the graphs.
- 5.17.2** Test the hypothesis that the mean salary for men and women is the same. What alternative hypothesis do you think is appropriate?
- 5.17.3** Assuming no interactions between sex and the other predictors, obtain a 95% confidence interval for the difference in salary between males and females.
- 5.17.4** Finkelstein (1980), in a discussion of the use of regression in discrimination cases, wrote, “[a] variable may reflect a position or status bestowed by the employer, in which case if there is discrimination in the award of the position or status, the variable may be ‘tainted.’” Thus, for example, if discrimination is at work in promotion of faculty to higher ranks, using rank to adjust salaries before comparing the sexes may not be acceptable to the courts.
Exclude the variable rank, refit, and summarize.
- 5.18** (Data file: salary) Using the salary data in Problem 5.17, one fitted mean function is

$$E(\text{salary}|\text{sex}, \text{year}) = 18223 - 571 \text{ sex} + 741 \text{ year} + 169 \text{ sex} \times \text{year}$$

- 5.18.1** Give the coefficients in the estimated mean function if Sex were coded so males had the value 2 and females had the value 1 (the coding given to get the above mean function was 0 for males and 1 for females).

- 5.18.2** Give the estimated coefficients if `sex` were coded as -1 for males and $+1$ for females.
- 5.19** (Data file: `Wool`) Refer to the Wool Data discussed in Section 5.2.
- 5.19.1** Write out in full the main effects and the second-order mean functions, assuming that the three predictors will be turned into factors, each with three levels. This will require you to define appropriate dummy variables and parameters.
- 5.19.2** For the two mean functions in Problem 5.19.1, write out the expected change in the response when `len` and `amp` are fixed at their middle levels, but `load` is increased from its middle level to its high level.

- 5.20** (Data file: `domedata`) Until 2010, the Minnesota Twins professional baseball team played its games in the Metrodome, an indoor stadium with a fabric roof.¹⁰ In addition to the large air fans required to keep the roof from collapsing, the baseball field is surrounded by ventilation fans that blow heated or cooled air into the stadium. Air is normally blown into the center of the field equally from all directions.

According to a retired supervisor in the Metrodome, in the late innings of some games, the fans would be modified so that the ventilation air would blow out from home plate toward the outfield. The idea is that the air flow might increase the length of a fly ball. For example, if this were done in the middle of the eighth inning, then the air-flow advantage would be in favor of the home team for six outs, three in each of the eighth and ninth innings, and in favor of the visitor for three outs in the ninth inning, resulting in a slight advantage for the home team.

To see if manipulating the fans could possibly make any difference, a group of students at the University of Minnesota and their professor built a “cannon” that used compressed air to shoot baseballs. They then did the following experiment in the Metrodome in March 2003:

1. A fixed angle of 50 degrees and velocity of 150 ft/s was selected. In the actual experiment, neither the velocity nor the angle could be controlled exactly, so the actual angle and velocity varied from shot to shot.
2. The ventilation fans were set so that to the extent possible all the air was blowing in from the outfield toward home plate, providing a headwind. After waiting about 20 minutes for the air flows to stabilize, 20 balls were shot into the outfield, and their distances were recorded. Additional variables recorded on each shot include the weight (in

¹⁰The Metrodome is scheduled to be replaced by a football-only stadium in 2014.

grams) and diameter (in centimeters) of the ball used on that shot, and the actual velocity and angle.

3. The ventilation fans were then reversed, so as much as possible air was blowing out toward the outfield, giving a tailwind. After waiting 20 minutes for air currents to stabilize, 15 balls were shot into the outfield, again measuring the ball weight and diameter, and the actual velocity and angle on each shot.

The data from this experiment are in the file `domedata`, courtesy of Ivan Marusic. The variable names are `Cond`, the condition, head or tail wind; `Velocity`, the actual velocity in feet per second; `Angle`, the actual angle; `BallWt`, the weight of the ball in grams used on that particular test; `BallDia`, the diameter in inches of the ball used on that test; `Dist`, distance in feet of the flight of the ball.

- 5.20.1** Summarize any evidence that manipulating the fans can change the distance that a baseball travels. Be sure to explain how you reached your conclusions, and provide appropriate summary statistics that might be useful for a newspaper reporter (a report of this experiment is given in the Minneapolis *StarTribune* of July 27, 2003).

- 5.20.2** One could argue that this experiment by itself cannot provide adequate information to decide if the fans can affect length of a fly ball. The treatment is *manipulating the fans*; each condition was set up only once and then repeatedly observed. Resetting the fans after each shot is not practical because of the need to wait at least 20 minutes for the air flows to stabilize.

A second experiment was carried out in May 2003, using a similar experimental protocol. As before, the fans were first set to provide a headwind, and then, after several trials, the fans were switched to a tailwind. Unlike the first experiment, however, the nominal `Angle` and `Velocity` were varied according to a 3×2 factorial design. The data file `domedata1` contains the results from both the first experiment and the second experiment, with an additional column called `Date` indicating which sample is which. Analyze these data, and write a brief report of your findings.

C H A P T E R 6

Testing and Analysis of Variance

Hypothesis testing is a regular part of regression analysis. The tests we have encountered so far concerned either a single regression coefficient (Sections 2.6 and 3.4.7), or a linear combination of them (Section 3.5). In either case, suppose $\hat{\theta}$ is the estimator of a parameter θ , and its standard error is $\text{se}(\hat{\theta})$. To test the simple null hypothesis $\text{NH} : \theta = \theta_0$ versus the alternative hypothesis $\text{AH} : \theta \neq \theta_0$, compute the statistic

$$t = \frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})} \quad (6.1)$$

Large values of $|t|$ suggest evidence that the unknown θ is different from θ_0 , while small values of $|t|$ support the NH. To get a significance level for the test, we generally refer the value of $|t|$ to a tabled distribution.¹ In most linear regression situations, the appropriate tabled distribution is a t -distribution with df given by the df in the estimate of σ^2 used in the standard error. In some instances, for example, if σ^2 is known, the standard normal distribution is used. The p -value is the area under the standard curve that is either greater than $|t|$ or less than $-|t|$. One-sided tests (Section 3.4.7), for example, with $\text{AH} : \theta > \theta_0$, would use only the area under the curve that is greater than t . Tests that are based on comparing the difference between an estimate and a hypothesized value, standardized by an estimate of error, are called *Wald tests*, in honor of Abraham Wald (1902–1950).

In this chapter, we present a different approach to testing based on comparing the fit of mean functions rather than comparing parameter estimates to hypothesized values. In linear regression, this leads to F -tests. These are also called *analysis of variance* tests, named after the way the tests are often

¹The bootstrap based tests introduced in Section 7.7.4 provide an alternative to using a standard distribution for tests.

summarized rather than the way they are computed, or *likelihood-ratio tests*, because of the general approach to testing in mathematical statistics that justifies these tests.

6.1 F-TESTS

Suppose we have a response Y and a vector of p' regressors $\mathbf{X}' = (\mathbf{X}'_1, \mathbf{X}'_2)$ that we partition into two parts so that \mathbf{X}_2 has q regressors and \mathbf{X}_1 has the remaining $p' - q$ regressors. The intercept, if present, is generally included in \mathbf{X}_1 , but this is not required. The general hypothesis test we consider is

$$\begin{aligned} \text{NH: } & E(Y|\mathbf{X}_1 = \mathbf{x}_1, \mathbf{X}_2 = \mathbf{x}_2) = \mathbf{x}'_1 \boldsymbol{\beta}_1 \\ \text{AH: } & E(Y|\mathbf{X}_1 = \mathbf{x}_1, \mathbf{X}_2 = \mathbf{x}_2) = \mathbf{x}'_1 \boldsymbol{\beta}_1 + \mathbf{x}'_2 \boldsymbol{\beta}_2 \end{aligned} \quad (6.2)$$

This is a different approach to hypothesis testing, since the null and alternative models refer to specification of mean functions, rather than to restrictions on parameters. A necessary condition for the methodology of this section to apply is that the model under NH must be a special case of the model under AH. In (6.2), the NH is obtained by setting $\boldsymbol{\beta}_2 = \mathbf{0}$.

For any linear regression model, the residual sum of squares measures the amount of variation in the response not explained by the regressors. If the NH were false, then the residual sum of squares RSS_{AH} under the alternative model would be considerably smaller than the residual sum of squares RSS_{NH} under the null model. This provides the basis of a test, and we will have evidence against the NH if the difference $(\text{RSS}_{NH} - \text{RSS}_{AH})$ is large enough.

The general formula for the test is

$$F = \frac{(\text{RSS}_{NH} - \text{RSS}_{AH})/(df_{NH} - df_{AH})}{\text{RSS}_{AH}/df_{AH}} \quad (6.3)$$

$$= \frac{\text{SSreg}/df_{Reg}}{\hat{\sigma}^2} \quad (6.4)$$

In this equation, df_{NH} and df_{AH} are the *df* for residual under NH and AH, $\text{SSreg} = \text{RSS}_{NH} - \text{RSS}_{AH}$ is the *sum of squares for regression*, and $df_{Reg} = df_{NH} - df_{AH}$ is its *df*. The denominator of the statistic is generally the estimate of σ^2 computed assuming that AH is true, $\hat{\sigma}^2 = \text{RSS}_{AH}/df_{AH}$, but as we will see later, other choices are possible. A sum of squares divided by its *df* is called a *mean square*, and so the *F*-test is the mean square for regression divided by the mean square for error under AH.

The *F*-test as described here appears to require fitting the model under both NH and AH, getting the residual sums of squares, and then applying (6.3); viewing the test in this way explains exactly what the test is doing. Computer packages generally take advantage of the elegant structure of the linear regres-

sion model to compute the test while fitting only under the AH by computing SS_{reg} in (6.4) directly.

If we assume that the errors are NID(0, σ^2) random variables, then if NH is true, (6.3) has an $F(df_{Reg}, df_{AH})$ -distribution, and large values of F provide evidence against the NH. The letter “F” is used in honor of R. A. Fisher (1890–1962) who is generally credited with the first use of this type of testing (Fisher and Mackenzie, 1923). The theory behind these tests is very beautiful and worthy of study; see Christensen (2011, chapter 3), among others, for general results.

Overall Test, Simple Regression

If we have a simple linear regression with the mean function $E(Y|X = x) = \beta_0 + \beta_1 x$, the *overall F-test* is of the hypotheses

$$\begin{aligned} \text{NH: } E(Y|X = x) &= \beta_0 \\ \text{AH: } E(Y|X = x) &= \beta_0 + \beta_1 x \end{aligned} \tag{6.5}$$

Under NH the response depends on none of the regressors apart from the intercept, and under AH it depends on the regressor X . The model for the NH including only the intercept is called a *null model*.

Under the null model, the residual sum of squares function is

$$\text{RSS}_{NH}(\beta_0) = \sum (y_i - \beta_0)^2$$

This function is minimized at $\hat{\beta}_0 = \bar{y}$, and so $\text{RSS}_{NH} = \sum(y_i - \hat{\beta}_0)^2 = \sum(y_i - \bar{y})^2 = SYY$, the total sum of squares. The df is the number n of observations minus the number of estimated parameters in the mean function which is equal to 1 in this situation, so $df_{NH} = n - 1$.

The AH is just the simple linear regression mean function, so its residual sum of squares and df are, respectively, RSS given at (2.8) and $df = n - 2$. Substituting into (6.2), the overall test is

$$\begin{aligned} F &= \frac{(SYY - \text{RSS})/[(n-1) - (n-2)]}{\hat{\sigma}^2} \\ &= \frac{\text{SSreg}}{\hat{\sigma}^2} \end{aligned}$$

where $\text{SSreg} = SYY - \text{RSS}$ is the sum of squares for regression defined at (2.7), and $\hat{\sigma}^2$ is the estimated variance from simple linear regression. This statistic is compared with the $F(1, n - 2)$ distribution to get significance levels.

For Forbes’s data, from Section 2.3, we have $n = 17$, $SYY = 427.794$, $\text{RSS} = 2.155$, and $\hat{\sigma}^2 = 0.144$. We can compute

$$F = \frac{427.794 - 2.155}{0.144} = 2962.79$$

which is compared with the $F(1, 15)$ distribution. F is so large that the p -value is effectively zero, and the evidence is very strong against NH. This is no surprise in light of Figure 1.3b.

For the Ft. Collins snowfall data, Section 2.6.2, the overall F -statistic is $F = 2.41$, with $(1, 91)$ df . Comparing to the $F(1, 91)$ distribution, the significance level is 0.12, providing very weak evidence against the null hypothesis.

Overall Test, Multiple Regression

For the fuel consumption example discussed in Section 3.3, the overall F -test compares the null model with no regressors except for the intercept as NH with the AH fitting all the regressors. As with simple regression, the fit under the null model is $\beta_0 = \bar{y}$, and so the residual sum of squares is SYY with $n - 1$ df . Under AH, the residual sum of squares and df are from the fit when all the regressors are used. We get $F = 11.99$ which is compared with the $F(4, 46)$ distribution to get a p -value that rounds to 0. This provides strong evidence against NH.

Wool Data

With this example, we show that this testing paradigm can be used in more complex situations beyond the overall test. For the wool data, Section 5.2, the predictors `len`, `amp`, and `load` are factors, each with 3 levels. We can consider testing

NH: $\log(\text{cycles}) \sim \text{len} + \text{amp} + \text{load} + \text{len} : \text{amp} + \text{len} : \text{load}$

AH: $\log(\text{cycles}) \sim \text{len} + \text{amp} + \text{load} + \text{len} : \text{amp}$
 $+ \text{len} : \text{load} + \text{amp} : \text{load}$

The statement of these hypotheses use the Wilkinson and Rogers (1973) notation. The NH model includes three main effects and two interactions. The AH includes all these regressors plus the `amp:load` interaction. Under NH this last interaction is zero, and under AH it is nonzero. The regressors that are common to NH and AH are estimated under both models. Thus, the desired test is for adding `amp:load` to a model that includes other regressors.

Hypothesis	df	RSS
NH	12	0.181
AH	8	0.166

For the F -test we estimate σ^2 under AH as $\hat{\sigma}^2 = 0.166/8 = 0.0208$, and

$$F = \frac{(0.181 - 0.166)/(12 - 8)}{0.0208} = 0.18$$

When compared with the $F(4, 8)$ distribution, we get a p -value of 0.94, suggesting no evidence against NH.

For this problem we could contemplate tests for each of the interactions and possibly for each of the main effects. We return to this example in Section 6.2.

UN Data

The UN data discussed in Section 5.1 considered a sequence of mean functions given in Wilkinson–Rogers notation as

Mean function	df	RSS	
lifeExpF ~ 1	198	20293.2	(6.6)
lifeExpF ~ group	196	7730.2	(6.7)
lifeExpF ~ log(ppgdp)	197	8190.7	(6.8)
lifeExpF ~ group + log(ppgdp)	195	5090.4	(6.9)
lifeExpF ~ group + log(ppgdp) + group:log(ppgdp)	193	5077.7	(6.10)

The first of these models (6.6) is the null model, so it has residual sum of squares equal to SY^2 , and $df = n - 1$. Mean function (6.7) has a separate mean for each level of group but ignores $\log(\text{ppgdp})$. Mean function (6.8) has a common slope and intercept for each level of group; (6.9) has separate intercepts but a common slope. The most general (6.10) has separate slopes and intercepts.

Tests can be derived to compare most of these mean functions. A reasonable procedure is to start with the most general, comparing NH: mean function (6.9) to AH: mean function (6.10),

$$F = \frac{(5090.4 - 5077.7)/(195 - 193)}{5077.7/193} = 0.24 \quad (6.11)$$

When compared with the $F(2, 193)$ distribution, we get a p -value of 0.79, providing no evidence of the need for separate slopes, confirming the visual impression of Figure 5.2.

If this first test had suggested that separate slopes and intercepts were needed, then further testing would not be needed.² Since the interaction is probably unnecessary, we can consider further testing using the first-order mean function (6.9) as AH, and either (6.7) or (6.8) as NH. For the test for (6.8) versus (6.9), we get

$$F = \frac{(8190.7 - 5090.4)/(197 - 195)}{5090.4/195} = 59.38$$

²A model not considered here would have a common intercept but separate slopes, and a test of this model versus model (6.10) could be reasonable; see Problem 6.4.

When compared with the $F(2, 195)$ distribution, we get a p -value of essentially 0, providing strong evidence that intercepts for the three levels of group are not all equal. The remaining test is left as a homework problem (Problem 6.3).

The two tests illustrated above used different denominators for the F -tests, as suggested by the general formula (6.3). When testing is summarized in an analysis of variance table, as to be discussed shortly, the largest model (6.10) would be used to provide the denominator for all tests. In this example, changing the denominator would change the value of F to 58.92, a change of no practical importance, but in other problems, pooling dropped regressors into the estimate of σ^2 can change the outcome of the test.

Cakes Data

For the cakes data in Section 5.3.1, we fit the full second-order model,

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 \quad (6.12)$$

Several hypothesis tests are of interest here:

$$\text{NH: } \beta_5 = 0 \text{ vs. AH: } \beta_5 \neq 0 \quad (6.13)$$

$$\text{NH: } \beta_2 = 0 \text{ vs. AH: } \beta_2 \neq 0 \quad (6.14)$$

$$\text{NH: } \beta_1 = \beta_2 = \beta_5 = 0 \text{ vs. AH: Not all 0} \quad (6.15)$$

These hypotheses are presented in terms of parameters rather than mean functions, but are equivalent to comparing mean functions. The notation is again a shorthand; for example, the test (6.13) implies that all coefficients not explicitly shown in the hypothesis statement are included in both NH and AH. In test (6.13), AH is given by (6.12), and the NH requires that the interaction regressor is dropped. The test in (6.14) specifies that the quadratic regressor in X_2 has a zero coefficient, while the third test (6.15) is somewhat nonstandard and tests to see if all regressors that involve X_1 can be dropped. You are asked to do these tests in Problem 6.9.

6.1.1 General Likelihood Ratio Tests

The F -tests described here are applications of likelihood ratio tests to linear models with normal errors. Any textbook on mathematical statistics, such as Casella and Berger (2001, section 8.4), will provide the general formulation of these tests.

6.2 THE ANALYSIS OF VARIANCE

In any given regression problem, many tests are possible. Which tests are appropriate, and the order in which the testing should be done, is not always clear, and many approaches have been proposed.

Suppose we fit the following model in Wilkinson–Rogers notation:

$$Y \sim A + B + C + A:B + A:C + B:C + A:B:C \quad (6.16)$$

In this model, each of A , B , or C could represent a continuous predictor with a single df , or a factor, polynomial, or spline basis with more than 1 df . An interaction like $A:B$ can have many df .

The approach to testing we adopt in this book follows from the *marginality principle* suggested by Nelder (1977). A lower-order term, such as the A main effect, is *never* tested in models that include any of its higher-order relatives like $A:B$, $A:C$, or $A:B:C$. All regressors that are *not* higher-order relatives of the regressor of interest, such as B , C , and $B:C$, are *always* included in both NH and AH.

Based on the marginality principle, testing should begin with the highest-order interaction first:

$$\text{NH: } Y \sim A + B + C + A:B + A:C + B:C$$

$$\text{AH: } Y \sim A + B + C + A:B + A:C + B:C + A:B:C$$

If the $A:B:C$ interaction is judged to be nonzero, no further testing is called for, since $A:B:C$ is a higher-order relative of all remaining regressors in the mean function.

If the $A:B:C$ interaction is judged nonsignificant, then proceed to examine the two-factor interactions, such as

$$\text{NH: } Y \sim A + B + C + A:C + B:C$$

$$\text{AH: } Y \sim A + B + C + A:B + A:C + B:C$$

which tests the $A:B$ interaction. There are similar tests for $A:C$ and $B:C$.

Tests for a main-effect like A would be carried out only if all its higher-order relatives, $A:B$, $A:C$, and $A:B:C$, are judged to be unimportant. One would then test

$$\text{NH: } Y \sim B + C + B:C$$

$$\text{AH: } Y \sim A + B + C + B:C$$

The $B:C$ interaction is included in both the NH and the AH.

All tests that satisfy the marginality principle can be collected into an analysis of variance or ANOVA table, as in Table 6.1 for the UN data. Apart from the last row of the table, the column marked df are the degrees of freedom for the numerator of the test. The next column is the sum of squares for regression for the numerator of the tests, and the third column is the corresponding mean square, the sum of squares divided by its df . The last row of the table gives the df for the residual, RSS, and the estimate of variance $\hat{\sigma}^2$ for fitting a model with all regressors. The F -values are the ratio of the regression mean squares to $\hat{\sigma}^2$, and the final column gives p -values for these tests, obtained essentially by looking up the F -value in the appropriate table of critical values of F .

Table 6.1 Analysis of Variance for the UN Data

	<i>df</i>	Sum Sq	Mean Sq	F-Value	Pr(>F)
Group	2	3100.31	1550.15	58.92	0.00
$\log(\text{ppgdp})$	1	2639.81	2639.81	100.34	0.00
group: $\log(\text{ppgdp})$	2	12.68	6.34	0.24	0.79
Residuals	193	5077.70	26.31		

Analysis of variance tables should be read from bottom to top to conform to the marginality principle. The bottom test is for the group: $\log(\text{ppgdp})$ interaction, and is identical to the test given at (6.11) for the test NH given by (6.9) versus AH (6.10). In this problem the test for the interaction has a large *p*-value, so testing lower-order effects is reasonable. Both of the main-effect tests have tiny *p*-values, suggesting that separate intercepts and a single nonzero slope are required. An effects plot was shown in Figure 5.3. The parallel lines of the fitted model have become curves because the horizontal axis is ppgdp rather than its logarithm. The africa group appears to be different from oecd and other, which suggests further testing of equality for these latter two groups.

An analysis of variance table derived under the marginality principle has the unfortunate name of *Type II analysis of variance*. At least two other types of analysis of variance are commonly available in software packages:

- Type I analysis of variance, also called *sequential analysis of variance*, fits models according to the order that the regressors are entered into in the mean function. For example, if (6.16) were fit, the sequence of models that would be represented in the ANOVA table would have regressors {A}, {A, B}, {A, B, C}, {A, B, C, A:B}, {A, B, C, A:B, A:C}, {A, B, C, A:B, A:C, B:C}, and {A, B, C, A:B, A:C, B:C, A:B:C}. One result of this is that one of the interactions, A:B is adjusted for none of the other interactions, another, A:C is adjusted for A:B, and A:C is adjusted for both A:C and A:B. If the terms were written in a different order, then the analysis would have different conditioning. Except in the special case of orthogonal regressors to be described shortly, when all the types described here are equivalent, Type I ANOVA generally has only pedagogical interest and should not be used, even though it may be the default ANOVA in some computer programs.³
- Type III analysis of variance violates the marginality principle. It computes the test for every regressor adjusted for every other regressor; so, for example, the test for the A main effect would include the interactions A:B, A:C, and A:B:C in both NH and AH. There is a justification for

³One such program is R, but to be fair, Type I ANOVA can be appropriate for other problems beyond those discussed here. Type II ANOVA is available with the Anova function in the car package in R.

this testing paradigm, called the *marginal means* method by Hocking (1985, 2003), but some of these tests depend on the parameterization used for the regressors and so they are not recommended for general use (McCullagh, 2002). In particular, the standard coding used for factors of omitting a baseline level (Section 5.1.1) is not appropriate for computing Type III sums of squares. The packages SAS and SPSS use Type III by default but have Type II as an available option.

The analysis of variance was originally formulated for problems in which all the regressors are orthogonal, or equivalently are uncorrelated with each other. Many designed experiments (Oehlert, 2000) will have this property. Now ANOVA is used more generally in problems with continuous regressors, polynomials, complex interactions, and nonorthogonal factors. In this situation, ANOVA is more complicated.

The wool data, Section 6.1, is from a designed experiment in which all the factors are orthogonal (Appendix A.6.6) to each other. Table 6.2 is the ANOVA table for the full second-order model, including all main effects and two-factor interactions. Because the regressors are orthogonal, Type I, Type II, and Type III tests are identical. We prefer to interpret the tests in every instance using the hypotheses formulated under the marginality principle.

Once again, the table is read from bottom to top. The `amp:load` and `len:load` interactions both have large p -values, and so both of these interactions can be neglected. The `len:amp` interaction has p -value of 0.028, which we may treat as evidence that this interaction is nonzero. In the presence of the `len:amp` interaction, the main effects for `len` and `load` become relatively uninteresting, but the test for the main effect of `load`, which is not part of the interaction, is of interest, and has a tiny p -value.

This analysis suggests refitting without the interactions that appear to be unimportant. The summarizing effects plot is shown in Figure 6.1. The small p -value for the main effect of `load` corresponds to the line in the plot for `load` differing from a horizontal line: larger loads result in smaller number of cycles. None of the interactions with `load` are included in the model, so this plot provides a complete summary of the relationship between `load` and the response. The small p -value for the `amp:len` interaction suggests the need to

Table 6.2 Analysis of Variance for the Second-Order Model for the Wool Data

	<i>df</i>	Sum Sq	Mean Sq	<i>F</i> -Value	Pr(> <i>F</i>)
len	2	12.516	6.258	301.74	0.000
amp	2	7.167	3.584	172.80	0.000
load	2	2.802	1.401	67.55	0.000
len:amp	4	0.401	0.100	4.84	0.028
len:load	4	0.136	0.034	1.64	0.256
amp:load	4	0.015	0.004	0.18	0.945
Residuals	8	0.166	0.021		

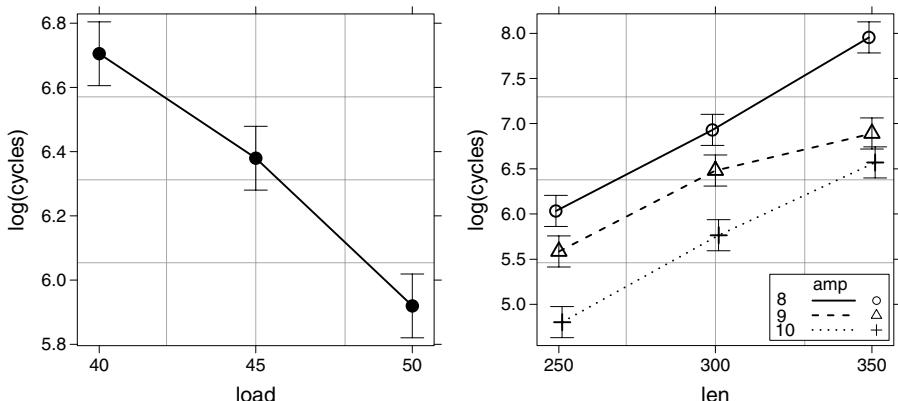


Figure 6.1 Effects plots for the wool data after deleting unimportant interactions.

consider amp and len simultaneously, as in the second graph in Figure 6.1. The lines are not parallel because of the interaction. When $(\text{amp}, \text{len}) = (9, 350)$, the fitted response is lower than would be expected if no interaction were present. Given the graph and small sample sizes, many experimenters might judge this interaction to be of little practical importance without verification from further experimentation.

6.3 COMPARISONS OF MEANS

The comparisons of adjusted means for levels of a factor, or for levels of an interaction, proceed as outlined in Section 5.1.2. There are two apparent impediments. First, the combinations of the parameters corresponding to the adjusted means can be complicated. Second, because many comparisons are possible, adjustment of significance levels of tests to account for multiple testing can be critical. From a practical point of view, both of these complications are nearly ignorable because software is generally available to construct the correct linear combinations and also to adjust the tests.

Comparisons of means can be made for any effect that satisfies the marginality principle. For example, in the fit of $\log(\text{cycles}) \sim \text{load} + \text{len} : \text{amp}$ in the wool data, comparisons of levels of load adjusted for $\text{len} : \text{amp}$ make sense, as do comparisons of the nine levels of $\text{len} : \text{amp}$ given load . Comparisons of the levels of len would generally not be recommended as these violate the marginality principle. The appropriate comparisons correspond to the effects plots in Figure 6.1.

For the $\text{amp} : \text{len}$ interaction, there are nine means so there are 36 possible paired comparisons. The interesting comparisons are likely to be between levels of amp for each level of len , for which there are only nine comparisons. We leave this for homework (Problem 6.13).

6.4 POWER AND NON-NULL DISTRIBUTIONS

For a fixed significance level, the probability of rejecting an NH is called the *power* of the test (Casella and Berger, 2001, section 8.3.1). Suppose that f^* is the critical value for a test at level α obtained from an F -table, meaning that $\text{Prob}(F > f^* | \text{NH is true}) = \alpha$. The power is

$$\begin{aligned}\text{Power} &= \text{Prob}(\text{detect a false NH}) \\ &= \text{Prob}(F > f^* | \text{AH is true})\end{aligned}$$

When the AH is true, the numerator and denominator of the test statistic (6.3) remain independent. The denominator estimates σ^2 under both the NH and the AH. The distribution of the numerator sum of squares is different under the NH and the AH. Apart from df , the numerator under the AH is distributed as σ^2 times a *noncentral χ^2* . In particular, the expected value of the numerator of (6.3) will be

$$E(\text{numerator of (6.3)}) = \sigma^2(1 + \text{noncentrality parameter}) \quad (6.17)$$

The larger the value of the noncentrality parameter, the greater the power of the test. For hypothesis (6.2), and now interpreting \mathbf{X}_1 as an $n \times (p' - q)$ matrix and \mathbf{X}_2 as an $n \times q$ matrix, the noncentrality parameter λ is given by the expression

$$\lambda = \frac{\boldsymbol{\beta}'_2 \mathbf{X}'_2 (\mathbf{I} - \mathbf{X}_1 (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1) \mathbf{X}_2 \boldsymbol{\beta}_2}{q\sigma^2} \quad (6.18)$$

This formidable equation can be simplified in special cases. For $q = 1$, $\boldsymbol{\beta}_2$ is a scalar and the test is for a single regressor. Write R_{X_2, X_1}^2 to be the value of R^2 for the OLS regression with response X_2 and regressors X_1 . As is usual, we write SD_2 to be the standard deviation of the regressor X_2 . Then, the noncentrality parameter λ is

$$\lambda = (n-1) \left(\frac{\boldsymbol{\beta}_2}{\sigma} \right)^2 [SD_2^2 (1 - R_{X_2, X_1}^2)] \quad (6.19)$$

Power increases with λ , so it increases with sample size n , the “size” of the parameter relative to the error standard deviation $(\boldsymbol{\beta}_2/\sigma)^2$, and it increases with the unexplained variability in X_2 after X_1 . In the special case that X_2 and X_1 are uncorrelated, or in simple regression, $R_{X_2, X_1}^2 = 0$. If X_2 is an indicator of a treatment that will be allocated to half of the cases at random, $\lambda = n\boldsymbol{\beta}_2^2/(4\sigma^2)$.

In most designed experiments, interesting tests concern effects that are orthogonal, and in this case (6.18) becomes for $q \geq 1$

$$\lambda = (n-1) \frac{\beta_2' S_2 \beta_2}{q \sigma^2} \quad (6.20)$$

where S_2 is the sample covariance matrix for \mathbf{X}_2 . General results on F -tests are presented in advanced linear model texts such as Christensen (2011).

Many computer programs include power calculators that can help you decide on necessary sample size to detect a difference of interest in a number of problems. It is typical of these calculators that the user specifies the type of problem, such as linear regression or a some other problem. Lenth (2006–2009) provides a Java applet for computing power and discussion of how to use it.

Minnesota Farm Sales

Problem 5.10 presents models of log price per acre, $\log(\text{acrePrice})$ as a function of the factors `year` of sale and `region`, for $n = 18,700$ sales in Minnesota for the years 2002–2011. Table 6.3 gives the Type II analysis of variance for the model $\log(\text{acrePrice}) \sim \text{year} + \text{region} + \text{year:region}$. The F -test for the interaction has a p -value that rounds to zero to five digits, suggesting a `year` by `region` interaction for these data. The effects plot, however (Problem 5.10) suggests relatively unimpressive differences between the response curves for the regions.

The very large sample size here implies that the test for an interaction is very powerful and likely to detect even very small differences. If the sample size were smaller, the test might not have given a significant result. To demonstrate this a simulation was done. The model with the interaction was fit to a subset of the data selected at random, and the p -value for the test for interaction was recorded. This was repeated 100 times for each several sample sizes varying from 935, corresponding to 5% of the original sample size, to $n = 5610$, corresponding to 30% of the original sample size. The average and standard deviation of the p -values are shown in Table 6.4. Also shown in the table is the empirical power, the fraction of times the test had a p -value less than 0.05.

When the sample is the smallest, $n = 935$, the average significance level is about 0.40 with standard deviation of 0.23, and the empirical power is only 12%. For comparison, if the NH is true, then the p -value is uniformly distrib-

Table 6.3 Analysis of Variance for the Minnesota Farm Sales

	<i>df</i>	Sum Sq	Mean Sq	<i>F</i> -Value	Pr(> <i>F</i>)
fyear	9	153.01	17.00	73.02	0.00000
region	5	4200.29	840.06	3608.17	0.00000
region:year	5	29.87	5.97	25.66	0.00000
Residuals	18680	4349.09	0.23		

Table 6.4 Average p -Value in Simulation

$n =$	935	1870	2805	3740	4675	5610
Average	0.40	0.22	0.09	0.02	0.01	0.00
SD	0.29	0.23	0.14	0.04	0.02	0.00
Power	0.12	0.32	0.61	0.87	0.97	1.00

uted with mean 0.5 and standard deviation of about 0.29. With this sample size, the outcome of the test is close to the outcome that would be expected if the NH were true. A sample size of 935 is hardly a small sample. A sample size of about 2800 is required for about 60% power. A sample consisting of 30% of the original data of 5610 has empirical power of 1. This simulation varied only the sample size to show that power can change. The power also depends on the size of the differences, but those were not changed in this simulation. Large sample sizes can find small differences, and at least in this instance, statistical significance may not translate into practical significance.

6.5 WALD TESTS

Wald tests about coefficients in regression are based on the distribution of the estimator $\hat{\beta}$. In most regression problems, the estimator is at least approximately normally distributed,

$$\hat{\beta} \sim N(\beta, \mathbf{V})$$

Generally, the $\text{Cov}(\beta) = \mathbf{V}$ is unknown, but an estimate $\hat{\mathbf{V}}$ is available. From Appendix 3.4.4, for OLS estimators we have

$$\hat{\mathbf{V}} = \hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}$$

With other estimation methods like weighted least squares and logistic regression encountered later in this book, the form of $\hat{\mathbf{V}}$ changes, but the results given here still apply.

6.5.1 One Coefficient

To test a hypothesis concerning a particular coefficient estimate, say NH : $\beta_j = \beta_{j0}$ versus AH : $\beta_j \neq \beta_{j0}$, compute, as in Section 2.6, $t = (\hat{\beta}_j - \beta_{j0}) / \sqrt{\hat{v}_{jj}}$, where \hat{v}_{jj} is the (j, j) element of $\hat{\mathbf{V}}$. This test is compared with the t -distribution with df equal to the df in estimating σ^2 to get p -values. In problems like logistic regression in which there is no σ^2 to estimate, the Wald test is compared with the standard normal distribution. One-sided tests use the same test statistic, but only a one-sided tail-area to get the p -value.

6.5.2 One Linear Combination

Suppose \mathbf{a} is a vector of numbers of the same length as $\boldsymbol{\beta}$. Then the linear combination $\ell = \mathbf{a}'\boldsymbol{\beta}$ has estimate $\hat{\ell} = \mathbf{a}'\hat{\boldsymbol{\beta}}$ and from (3.26)

$$\hat{\ell} \sim N(\ell, \mathbf{a}'\mathbf{V}\mathbf{a})$$

The standard error is $se(\hat{\ell}) = \sqrt{\mathbf{a}'\hat{\mathbf{V}}\mathbf{a}}$. Thus, for $NH : \ell = \ell_0$, the statistic is $t = (\hat{\ell} - \ell_0) / se(\hat{\ell})$, which is compared with the t -distribution with df given by the df for $\hat{\sigma}^2$. The same statistic is used for two-sided or one-sided alternatives, but tail areas are different (Section 3.4.7). If ℓ consists of all zeros except for a single one for the j th coefficient, then this t -test is identical to the t -test in Section 6.5.1.

6.5.3 General Linear Hypothesis

More generally, let \mathbf{L} be any $q \times p'$ matrix of constants which we take to be of full row rank q . Suppose we wish to test $NH : \mathbf{L}\boldsymbol{\beta} = \mathbf{c}$ versus the alternative $AH : \mathbf{L}\boldsymbol{\beta} \neq \mathbf{c}$. This generalizes from one linear combination to q linear combinations. The test statistic is

$$F = \frac{(\mathbf{L} - \mathbf{c})'(\mathbf{L}'\hat{\mathbf{V}}\mathbf{L})^{-1}(\mathbf{L} - \mathbf{c})}{q} \quad (6.21)$$

Under NH and normality this statistic can be compared with an $F(q, n - p')$ distribution to get significance levels.

6.5.4 Equivalence of Wald and Likelihood-Ratio Tests

For linear models, the Wald tests and the likelihood ratio tests give the same results for any fixed hypothesis test. Thus, for example, the square of the Wald t -test for a single coefficient is numerically identical to the likelihood ratio F -test for the same coefficient. As long as the hypothesis matrix \mathbf{L} is correctly formulated, (6.21) and (6.3) will be numerically identical. This equality does not carryover to other regression settings like logistic regression. Wald and likelihood ratio tests for logistic regression are equivalent, in the sense that for large enough samples they will give the same inference, but not equal, as the computed statistics generally have different values. Likelihood ratio tests are generally preferable.

6.6 INTERPRETING TESTS

6.6.1 Interpreting p -Values

Under the appropriate assumptions, the p -value is the conditional probability of observing a value of the computed statistic, here the value of F , as extreme

or more extreme, here as large or larger, than the observed value, given that the NH is true. A small p -value provides evidence against the NH.

In many research areas it has become traditional to adopt a *fixed significance level* when examining p -values. For example, if a fixed significance level of α is adopted, then we would say that an NH is rejected at level α if the p -value is less than α . The most common choice for α is 0.05, which would mean that, were the NH to be true, we would incorrectly find evidence against it about 5% of the time, or about one test in 20. Accept-reject rules like this are generally unnecessary for reasonable scientific inquiry, although they may be mandated by some research journals. Simply reporting p -values and allowing readers to decide on significance seems a better approach.

There is an important distinction between statistical significance, the observation of a sufficiently small p -value, and scientific significance, observing an effect of sufficient magnitude to be meaningful. Judgment of the latter usually will require examination of more than just the p -value.

6.6.2 Why Most Published Research Findings Are False

A widely circulated article, Ioannidis (2005), has the same title as this section. While this section title is intended as hyperbole, there are several reasons to doubt findings or question interpretation of results based on a single hypothesis test. For simplicity in this discussion, suppose that all tests are done at level α , and that all have the same power or probability of detecting a false NH of γ .

Following Ioannidis (2005),

- Suppose that a fraction f of hypothesis tests are *potential discoveries*. This is the fraction of tests for which rejecting the NH is the correct decision.
- A *true discovery* will occur if we correctly reject an NH when it is false. This will occur with probability $f\gamma$.
- A *false discovery* will occur if NH is rejected but NH is actually true, and this will occur with probability $(1 - f)\alpha$.
- The probability of a *discovery* is the sum of these, $f\gamma + (1 - f)\alpha$.

From these, we can compute the conditional probability of a true discovery given a discovery,

$$\text{Prob}(\text{true discovery}|\text{discovery}) = \frac{f\gamma}{f\gamma + (1 - f)\alpha}$$

Figure 6.2 gives a graph of this probability as a function of the fraction of potential discoveries f with both variables in log scale, for three values of $\gamma \in \{0.50, 0.75, 0.99\}$ and for $\alpha = 0.05$. The power γ has relatively limited effect on these curves, so we will discuss only the case of $\gamma = 0.75$. If the fraction of potential discoveries is high, say $f = 0.90$, then $\text{Prob}(\text{true discovery}|\text{discovery})$

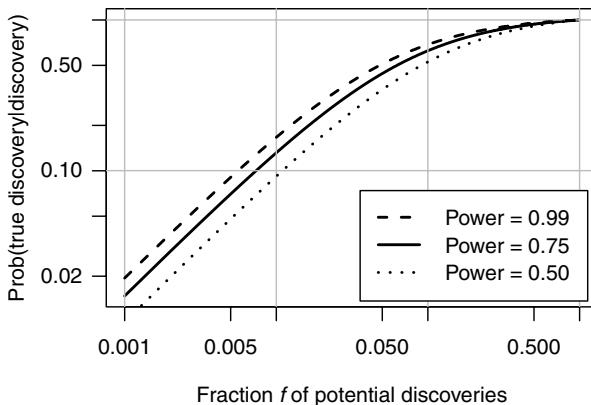


Figure 6.2 The probability of a true discovery as a function of the fraction f of false NH and the power of the test.

= 0.99, so we can reliably believe that a rejected NH will correspond to a true discovery. When $f = 1/16$, $\text{Prob}(\text{true discovery}|\text{discovery}) = 0.5$, so a rejected NH is equally likely to be a true or false discovery. If $f = 0.01$, then $\text{Prob}(\text{true discovery}|\text{discovery}) = 0.13$ and the vast majority of discoveries will be false discoveries.

The reliability of tests in a particular situation requires an assessment of the relevant value of f . In testing situations where data are collected based on a well-established theory, we might hope for $f > 0.5$, with more theoretical knowledge corresponding to larger values of f , and most discoveries will be true discoveries. Not all experiments fit this paradigm, however. Ioannidis presents as an example a genome association study in which 100,000 gene polymorphisms will each be tested to find the 10 or so genes that are associated with a particular disease. For this study $f = 10/100,000 = 0.001$, and nearly all discoveries will be false discoveries. Lehrer (2010) presents a nontechnical review of many other findings that were eventually not supported by later data, or findings that seem to have weakened over time.

6.6.3 Look at the Data, Not Just the Tests

Tests can be computed for any data set, whether the test is appropriate or not. The scatterplots in Figure 1.9 provide an example. If simple regression is fit to any of the graphs shown, the t -statistic for testing the slope equal to zero is $t = 4.24$, with a corresponding two-sided p -value = 0.002, but only for Figure 1.9a is the test meaningful because in the other graphs, either simple regression is clearly inappropriate, or the inference is effectively determined by only one data point.

Many analysts skip the step of actually looking at the data, and they do so at their own peril. The graphical methods throughout this book, and the diagnostic methods to be presented in Chapter 9, can help avoid this type of pitfall.

6.6.4 Population versus Sample

Tests are designed to infer from observations on a sample to a larger population. In the Berkeley Guidance Study (Problem 3.3), for example, tests could be inferences to the population of other children born near Berkeley in the same era, or with somewhat less justification to children born in other areas in California or even the United States during that era. Inference to the population of children in a different era is more of a stretch, as too much else may have changed to make children born in 1928–1929 representative of children from other times or other places.

Applicability is also at issue with the UN data used extensively in this book. The unit of analysis is a locality, generally a country, for which the UN provides statistics and for which the variables described are measured. The countries/localities represented in the data include more than 99% of the world's people, and so the data in the UN examples form a population, not a sample from a population. The only variation in the estimates is due to measurement errors in the variables, but not to sampling from a population. If the variables were measured without error, then the “estimates” in the data would be the true parameter values.

Freedman and Lane (1983) proposed an alternative interpretation of summaries of tests that they call *reported significance levels*. Using an argument related to the bootstrap, they suggest that a small reported significance levels characterizes an unusual data set relative to hypothetical data sets that could have arisen if NH were true. For the UN data, but ignoring the additional problems outlined in the next paragraph, this would suggest that the significance levels of tests remain helpful summaries of the analysis.

6.6.5 Stacking the Deck

The UN data have been used to explore the dependence of female life expectancy on national per capita income, separately for three groups of countries/localities. Perhaps you found the grouping puzzling, since countries were divided geographically into Africa and not Africa, and the not Africa group was subdivided according to membership in the OECD. This seems like a very strange way to divide up the world.

It happened like this: Figure 3.1c, discussed further in Problem 3.1, suggested that the African countries had a different relationship between `life-ExpF` and `log(ppgdp)` than did other countries, and so this became the basis of the example. To make the problem more interesting for presentation in this book, the “not Africa” nations were divided again. OECD membership provided a convenient way to divide this group roughly into richer and poorer countries.

Any test to compare the groups is almost certain to show significance because the `group` variable was defined to match the different groups seen in Figure 3.1. We “stacked the deck.” The data, not theory, guided the test, and

this renders the test concerning differences between Africa and non-Africa at least suspect.

6.6.6 Multiple Testing

Multiple testing is one of the most important problems with interpreting tests. If 100 independent tests are done, each at level $\alpha = 0.05$, even if NH is true in all 100 tests, then about $0.05 \times 100 = 5$ of the tests are expected to be “significant at the 5% level” and therefore false discoveries. Traditional methods of surviving multiple testing are to control the *family-wise error rate* rather than the *per-test error rate* (Miller, 1981), but recent methodology is based on controlling the *false discovery rate*, as proposed by Benjamini and Hochberg (1995); see Bretz et al. (2010) for current methodology. Except for testing for outliers in Section 9.4.3, we leave discussion and application of multiple testing methods to other sources.

6.6.7 File Drawer Effects

Similar to the multiple testing problem is the *file drawer problem*. If 100 investigators set out to do the same experiment to learn about a treatment effect, about 5% of them will get significant results even if there are no real effects. The 95% who find no difference may put their experiment aside in a file drawer and move on; the remaining 5% seek to publish results. Consequently, published results can appear significant only because the reader of them is unaware of the unpublished results.

6.6.8 The Lab Is Not the Real World

Observing a phenomenon can change its outcome, and effects that are observed in a study or in a laboratory setting may not persist in a natural setting with no one watching or interfering. People, animals, and even plants can behave differently when they are being studied than when they are acting independently. This is called a *Hawthorne effect*, after a set of experiments with lighting in work spaces at the Western Electric Hawthorne Works in the 1920s (Hart, 1943). Similarly, in a medical trial, patients in a controlled setting may have different outcomes than they would if they were responsible for their own care, perhaps due to failure to understand or to comply with a protocol. As another example, a variety of a crop may be more successfully planted at an experimental farm with the latest in agronomic methodology than it would be planted elsewhere.

6.7 PROBLEMS

- 6.1** (Data file: UN11) With the UN data, perform a test of NH : (6.6) against the alternative AH : (6.7), and summarize results. This is an overall *F*-test for a model with one factor and no additional regressors.

6.2 (Data file: UN11) With the UN data, explain why there is no *F*-test comparing models (6.7) and (6.8).

6.3 (Data file: UN11) In the UN data, compute the *F*-test

$$\begin{aligned} \text{NH: } & \text{Model (6.7)} \\ \text{AH: } & \text{Model (6.9)} \end{aligned}$$

and summarize results.

6.4 (Data file: UN11) With the UN data, consider testing

$$\begin{aligned} \text{NH: } & \text{lifeExpF} \sim \log(\text{ppgdp}) + \text{group} : \log(\text{ppgdp}) \\ \text{AH: } & \text{lifeExpF} \sim \text{group} + \log(\text{ppgdp}) + \text{group} : \log(\text{ppgdp}) \end{aligned}$$

The AH model is the most general model given at (6.10), but the NH is was not given previously.

6.4.1 Explain in a sentence or two the meaning of the NH model.

6.4.2 Perform the test and summarize results.

6.5 (Data file: UN11) In the UN data, start with the parallel regression model (6.9).

6.5.1 Test for equality of intercepts for the `oecd` and `other` levels of the factor `group`.

6.5.2 Test for equality of the intercepts for group `other` and `africa`.

6.6 (Data file: fuel2001) State the null and alternative hypotheses for the overall *F*-test for the fuel consumption data (Section 3.3). Perform the test and summarize results.

6.7 (Data file: fuel2001) With the fuel consumption data, consider the following two models in Wilkinson–Rogers notation:

$$\text{fuel} \sim \text{Tax} + \text{Dlic} + \text{Income} + \log(\text{Miles}) \quad (6.22)$$

$$\text{fuel} \sim \log(\text{Miles}) + \text{Income} + \text{Dlic} + \text{Tax} \quad (6.23)$$

These models are of course the same, as they only differ by the order in which the regressors are written.

6.7.1 Show that the Type I ANOVA for (6.22) and (6.23) are different. Provide an interpretation of each of the tests.

6.7.2 Show that the Type II ANOVA is the same for the two models. Which of the Type II tests are equivalent to Type I tests?

- 6.8** Show that the overall F -test for multiple regression with an intercept can be written as

$$F = \left(\frac{n-p'}{p} \right) \frac{R^2}{1-R^2}$$

where R^2 is the proportion of variability explained by the regression. Thus, the F -statistic is just a transformation of R^2 .

- 6.9** (Data file: cakes) For the cakes data in Section 5.3.1, we fit the full second-order model,

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$$

Compute and summarize the following three hypothesis tests.

$$\text{NH: } \beta_5 = 0 \text{ vs. AH: } \beta_5 \neq 0$$

$$\text{NH: } \beta_2 = 0 \text{ vs. AH: } \beta_2 \neq 0$$

$$\text{NH: } \beta_1 = \beta_2 = \beta_5 = 0 \text{ vs. AH: Not all 0}$$

- 6.10** RateMyProfessor.com (Data file: Rateprof) In the professor ratings data introduced in Problem 1.6, suppose we were interested in modeling the quality rating. We take as potential predictors characteristics of the instructor, including gender of the professor, the number of years numYears in which the instructor had ratings, between 1999 and 2009, a factor discipline, with levels for humanities, social science, pre-professional, and stem for science technology, engineering, and mathematics. Additional potential predictors are easiness, average rating of the easiness of the course, raterInterest in the course material. A final predictor is pepper, a factor with levels no and yes. A value of yes means that the consensus is that the instructor is physically attractive. The variables helpfulness and clarity have been excluded, since these are essentially the same as quality (Section 5.5). Data are included for $n = 366$ professors.

- 6.10.1** Fit the first-order regression model `quality~gender+numYears+pepper+discipline+easiness+raterInterest`, and print the summary table of coefficient estimates. Suppose that β_2 is the coefficient for `numYears`. Provide a test and significance level for the following three hypothesis tests: (1) NH : $\beta_2 = 0$ versus AH : $\beta_2 \neq 0$; (2) NH : $\beta_2 = 0$ versus AH : $\beta_2 \leq 0$; (3) NH : $\beta_2 = 0$ versus AH : $\beta_2 \geq 0$.

- 6.10.2** Obtain the Type II analysis of variance table. Verify that the F -tests in the table are the squares of the t -tests in the regression coefficient table, with the exception of the tests for the dummy regressors for `discipline`. Summarize the results of the tests.

6.10.3 Draw the effects plot for `discipline`. It will suggest that the adjusted `quality` varies by discipline, in agreement with the test for `discipline`. Describe as carefully you can the `discipline` effect. You may want to report further tests.

6.10.4 Summarize the dependence of `quality` on the predictors.

6.11 (Data file: `salarygov`) For the government salary data described in Problem 5.9, use the model of Problem 5.9.3, obtain tests for the interaction between the indicator for female-dominated occupations and the spline basis for `Score`. Obtain a 95% confidence interval for the difference between female-dominated job classes and all other job classes.

6.12 (Data file: `twins`) The data in the file `twins` give the IQ scores of identical twins, one raised in a foster home, `IQf`, and the other raised by birth parents, `IQb`. The data were published by Burt (1966), and their authenticity has been questioned. For purposes of this example, the twin pairs can be divided into three social classes `C`, low, middle, or high, coded in the data file 1, 2, and 3, respectively, according to the social class of the birth parents. Treat `IQf` as the response and `IQb` as the predictor, with `C` as a factor.

Describe the dependence of the response on the predictors, using appropriate graphs, models discussed in the last chapter, and tests described in this chapter.

6.13 (Data file: `wool`) With the wool data, fit the model `log(cycles) ~ load + len:amp`. Use computer software to obtain (1) the estimates and standard errors of the adjusted means of each level of `load` and for each combination of `len:amp`; (2) obtain tests to compare the levels of `load` and to compare the levels of `amp` for each level of `len`. (*Hints:* The levels of `len`, `amp`, and `load` are numeric, and so you may need to tell your computer program to treat them as factors; for example, in R, you would use the “factor” function. Many programs use the keyword `lsmeans` to describe the means you need to compute. In R there is a package called `lsmeans` (Lenth, 2013) that will do all the computations you need. The estimated means computed in this package can differ slightly from the means plotted by the `effects` package because they have slightly different defaults for values for conditioning.)

6.14 Testing for lack-of-fit (Data file: `MinnLand`) Refer to the Minnesota farm sales data introduced in Problem 5.4.

6.14.1 Fit the regression model `log(acrePrice) ~ year` via `ols`, where `year` is not a factor, but treated as a continuous predictor. What does this model say about the change in price per acre over time? Call this model A.

- 6.14.2** Fit the regression model via $\log(\text{acrePrice}) \sim 1 + \text{fyear}$ via ols, where fyear is a factor with as many levels are there are years in the data, including the intercept in the model. What does this model say about the change in price per acre over time? Call this model B. (*Hint:* fyear is not included in the data file. You need to create it from the variable year.)
- 6.14.3** Show that model A is a special case of model B, and so a hypothesis test of $\text{NH} : \text{model A}$ versus $\text{AH} : \text{model B}$ is reasonable.
- 6.14.4** A question of interest is whether or not model A provides an adequate description of the change in $\log(\text{acrePrice})$ over time. The hypothesis test of $\text{NH} : \text{model A}$ versus $\text{AH} : \text{model B}$ addresses this question, and it can be called a *lack-of-fit test* for model A. Perform the test and summarize results.
- 6.15** (Data file: MinnLand) Continuing with the last problem, suppose you fit the model $\log(\text{acrePrice}) \sim \text{year} + \text{fyear}$, including year both as a continuous predictor and as a factor. What do you think will happen? Try it and find out if you were right!
- 6.16** (Data file: MinnLand) Repeat the simulation of Section 6.4, but for the lack-of-fit test of Problem 6.14. In the simulation, use the fraction of data used in the test $f \in \{0.01, 0.05, 0.10, 0.15, 0.20\}$. Comment on the results.
- 6.17** An experiment is planned in which we have a set of regressors \mathbf{X}_1 and one addition regressor X_2 will be created with values 0 for subjects that get a control treatment and 1 for subjects that get the experimental treatment. Treatment assignment will be done at random, with half of the subjects getting the experimental treatment. The test for a treatment effect will be done at level $\alpha = 0.05$, and it desired to make the experiment large enough to have 90% power. An estimate of σ is required, and suppose that setting $\sigma = 0.5$ is reasonable.
Use a sample size calculator such as that of Lenth (2006–2009) to determine the sample size if the smallest meaningful treatment effect is equal to 1.0. Repeat for effect sizes of 0.5 and 2.0.
- 6.18 Windmill data** (Data file: `wm2`) In Problem 2.21 we considered data to predict wind speed `CSpd` at a candidate site based on wind speed `RSpd` at a nearby reference site where long-term data are available. In addition to `RSpd`, we also have available the wind direction, `RDir`, measured in degrees. A standard method to include the direction data in the prediction is to divide the directions into several bins and then fit a separate mean function for `CSpd` on `RSpd` in each bin. In the wind farm literature, this is called the *measure, correlate, predict* method (Derrick, 1992). The data file contains values of `CSpd`, `RSpd`, `RDir`, and `Bin` for 2002 for

the same candidate and reference sites considered in Problem 2.21. Sixteen bins are used, the first bin for cases with `RDir` between 0 and 22.5 degrees, the second for cases with `RDir` between 22.5 and 45 degrees, . . . , and the last bin between 337.5 and 360 degrees. Both the number of bins and their starting points are arbitrary.

6.18.1 Obtain an appropriate graphical summary of the data.

6.18.2 Obtain tests that compare fitting the four mean functions discussed in Section 5.1.3 with the 16 bins. How many parameters are in each of the mean functions?

6.19 Land valuation (Data file: `prodscore`) Taxes on farmland enrolled in a “Green Acres” program in metropolitan Minneapolis-St. Paul are valued only with respect to the land’s value as productive farmland; the fact that a shopping center or industrial park has been built nearby cannot enter into the valuation. This creates difficulties because almost all sales, which are the basis for setting assessed values, are priced according to the development potential of the land, not its value as farmland. A method of equalizing valuation of land of comparable quality was needed.

One method of equalization is based on a soil productivity score P , a number between 1, for very poor land, and 100, for the highest quality agricultural land. The data in the file `prodscore`, provided by Douglas Tiffany, give P along with `Value`, the average assessed value, the `Year`, either 1981 or 1982, and the `County` name for four counties in Minnesota, Le Sueur, Meeker, McLeod, and Sibley, where development pressures had little effect on assessed value of land in 1981–1982. The unit of analysis is a township, roughly 6 miles square.

The goal of analysis is to decide if soil productivity score is a good predictor of assessed value of farmland. Be sure to examine county and year differences, and write a short summary that would be of use to decision makers who need to determine if this method can be used to set property taxes.

Variances

In this chapter we consider a variety of extensions to the linear model that allow for more general variance structures than the independent, identically distributed errors assumed in earlier chapters. This greatly extends the problems to which linear regression can be applied. Some of the extensions require only minor adaptation of earlier results, while others add considerable complexity. Most of these latter extensions are only briefly outlined here with references to other sources for more details.

7.1 WEIGHTED LEAST SQUARES

The assumption that the variance function $\text{Var}(Y|X)$ is the same for all values of X can be relaxed in a number of ways. In an important generalization, suppose we have the multiple regression mean function given for the i th case by

$$\mathbb{E}(Y|X = \mathbf{x}_i) = \boldsymbol{\beta}'\mathbf{x}_i \quad (7.1)$$

but rather than assume that errors are constant, we assume that

$$\text{Var}(Y|X = \mathbf{x}_i) = \text{Var}(e_i) = \sigma^2/w_i \quad (7.2)$$

where w_1, \dots, w_n are *known positive numbers*. The variance function is still characterized by only one unknown positive number σ^2 , but the variances can be different for each case. This will lead to the use of *weighted least squares*, or *wls*, in place of *OLS*, to get estimates.

The *wls* estimator $\hat{\boldsymbol{\beta}}$ is chosen to minimize the weighted residual sum of squares function,

$$\text{RSS}(\boldsymbol{\beta}) = \sum w_i (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 \quad (7.3)$$

Squared differences $(y_i - \mathbf{x}'_i \boldsymbol{\beta})^2$ in (7.3) with relatively larger values of w_i are more influential in the weighted RSS, and so observations with smaller variance, that is σ^2/w_i smaller, are more important.

We will generally use the symbol $\hat{\boldsymbol{\beta}}$ for both the OLS and WLS estimators because OLS is a special case of WLS with $w_i = 1$ for all i . Writing \mathbf{W} as the $n \times n$ matrix with the w_i on the diagonal and zeroes elsewhere, the WLS estimator is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W} \mathbf{Y} \quad (7.4)$$

The properties of the WLS estimator are very similar to the properties of the OLS estimator, and are briefly outlined in Appendix A.8.4. Except for residual analysis described in Chapter 9, using weights is essentially a “set and forget” procedure. Output from statistical packages for OLS and WLS will appear identical, and can be interpreted identically.

Strong Interaction

The purpose of the experiment described here is to study the interactions of unstable elementary particles in collision with proton targets (Weisberg et al., 1978). These particles interact via the so-called strong interaction force that holds nuclei together. Although the electromagnetic force is well understood, the strong interaction is somewhat mysterious, and this experiment was designed to test certain theories of the nature of the strong interaction.

The experiment was carried out with beams having various values of incident momentum, or equivalently for various values of s , the square of the total energy in the center-of-mass frame of reference system. For each value of s , we observe the *scattering cross-section* y , measured in millibarns (mb). A theoretical model of the strong interaction force predicts that

$$E(y|s) = \beta_0 + \beta_1 s^{-1/2} + \text{relatively small terms} \quad (7.5)$$

The theory makes quantitative predictions about β_0 and β_1 and their dependence on particular input and output particle type.

The data given in Table 7.1 and in the file physics summarize the results of experiments when both the input and output particle was the π^- meson. A very large number of particles was counted at each setting of s , and as a result, the values of $\text{Var}(y|s = s_i) = \sigma^2/w_i$ are known almost exactly; the square roots of these values are given in the third column of Table 7.1, labeled SD_i .

Ignoring the smaller terms, mean function (7.5) is a simple linear regression mean function with regressors for an intercept and $x = s^{-1/2}$. We should use WLS because the variances are different for each value of s . Because of the very large sample sizes, we are in the unusual situation that we not only know the weights, *but we know the value of σ^2/w_i for each value of i* . There are 11

Table 7.1 The Strong Interaction Data

$x = s^{-1/2}$	y(mb)	SD _i
0.345	367	17
0.287	311	9
0.251	295	9
0.225	268	7
0.207	253	7
0.186	239	6
0.161	220	6
0.132	213	6
0.084	193	5
0.060	192	5

Table 7.2 wls Estimates for the Strong Interaction Data

	Estimate	Std. Error	t-Value	Pr(> t)
(Intercept)	148.4732	8.0786	18.38	0.0000
x	530.8354	47.5500	11.16	0.0000

$\hat{\sigma} = 1.6565$ with 8 df, $R^2 = 0.9397$

quantities w_1, \dots, w_{10} and σ^2 that describe the values of only 10 variances, so we have too many parameters, and we are free to specify one of the 11 parameters to be any nonzero value we choose. The simplest approach is to set $\sigma^2 = 1$, and then the last column of Table 7.1 gives $1/\sqrt{w_i}$, $i = 1, 2, \dots, n$, and so the weights are just the inverse squares of the last column of this table.

The fit of the simple regression model via wls is summarized in Table 7.2. The summary for the fit is the same as for ols, and interpretation is the same. The standard errors shown are the correct wls estimated standard errors. The t-tests for both coefficients are very large with corresponding p-values of effectively 0. The value of R^2 is large.

Interestingly, the estimate $\hat{\sigma} = 1.66$ is larger than the assumed value of $\sigma = 1$, which could indicate that the straight-line mean function (7.5) does not provide an adequate summary of these data. We explore this graphically in Figure 7.1. The solid line on the figure shows the wls fit of (7.5). The dashed curve matches the points more closely, and so we should not trust the usefulness of this model to describe this experiment.

A simple alternative to (7.5) is to add a quadratic regressor in $x = s^{-1/2}$ to the mean function to get

$$E(y|s) = \beta_0 + \beta_1 s^{-1/2} + \beta_2 s + \text{relatively small terms} \quad (7.6)$$

and this is the model that is fit to get the dashed line in Figure 7.1. One can show that adding the additional regressor gives $\hat{\sigma} = 0.679$, increases R^2 to 0.991, providing nearly a perfect fit. An F-test comparing for adding the

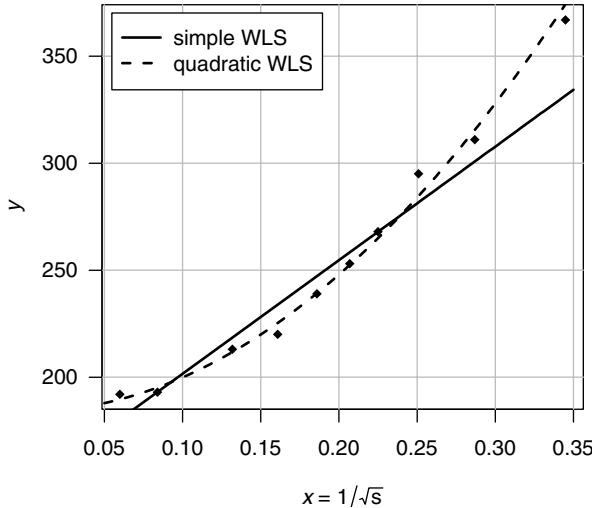


Figure 7.1 Scatterplot for the strong interaction data.

quadratic term has p -value smaller than 0.001, suggesting the alternative model with the quadratic term provides a superior fit.

One exception to the “set and forget” metaphor for wls is in prediction of future values, given in Section 3.5 for ols. A point prediction for a new observation \mathbf{x}_* given a fitted wls model is $\tilde{y}_* = \boldsymbol{\beta}'\mathbf{x}_*$. The variance of a prediction is the sum of two components. The first component is the variance of the fitted value \tilde{y}_* . As with ols, it is estimated by the square of (3.25). The second component is the variance of the unobservable error for the new observation at \mathbf{x}_* , and this depends on weights. In the physics example, if $x_* = x_j$, one of the observed values of x in the data, we would take the known variance SD_j^2 as the variance of the future value. The standard error of prediction would then be

$$\text{sepred}(\tilde{y}_*|x = x_j) = \sqrt{SD_j^2 + \text{sefit}(\tilde{y}_*|\mathbf{x}_*)^2}$$

In wls more generally, the variance of a future value will be σ^2/w_* , where w_* is the weight that is appropriate for the future value, so we would need to know w_* to compute a standard error of prediction,

$$\text{sepred}(\tilde{y}_*|x) = \sqrt{\sigma^2/w_* + \text{sefit}(\tilde{y}_*|\mathbf{x}_*)^2}$$

7.1.1 Weighting of Group Means

The data used in Problem 1.6 in the file `Rateprof` on professor ratings from the website RateMyProfessor.com provide another use of weights in fitting models. These data consist of the averages of many student ratings for

each instructor. We take as the response variable $\bar{y}_i = \text{quality}$, the average rating for quality for the i th instructor in the data. All the ratings in these data are on a 1 to 5 scale, with 5 the highest, and so average ratings can be any number between 1 and 5. We take as the regressors $\bar{x}_{1i} = \text{easiness}$, the average rating for easiness of the i th instructor's course or courses, and $\bar{x}_{2i} = \text{raterinterest}$, the average student interest in the material covered by the i th instructor. Also given in the data file is $n_i = \text{numRaters}$, the number of ratings that were averaged for the i th instructor.

Suppose we let $(y_{ij}, x_{1ij}, x_{2ij})$ be, respectively, the quality rating, easiness rating, and rater interest score of the j th student who rated the i th instructor. These values for the individual raters are not given in the data. Nevertheless, we can write for instructor i and for $j = 1, \dots, n_i$

$$\mathbb{E}(y_{ij}|X) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} \quad \text{Var}(y_{ij}|X) = \sigma^2 \quad (7.7)$$

The observed rating for instructor i is $\bar{y}_i = \sum y_{ij}/n_i$ and

$$\begin{aligned} \mathbb{E}(\bar{y}_i|X) &= \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbb{E}(y_{ij}|X) \\ &= \frac{1}{n_i} \sum_{j=1}^{n_i} [\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij}] \\ &= \beta_0 + \beta_1 \bar{x}_{1i} + \beta_2 \bar{x}_{2i} \end{aligned} \quad (7.8)$$

Assuming the ratings are independent, the variance of the rating of instructor i is

$$\begin{aligned} \text{Var}(\bar{y}_i|X) &= \frac{1}{n_i^2} \sum_{j=1}^{n_i} \text{Var}(y_{ij}|X) \\ &= \sigma^2/n_i \end{aligned} \quad (7.9)$$

Thus, to estimate the parameters in (7.7) from the rater averages, we should use wls.¹ Comparing to (7.2), the weights are $w_i = n_i = \text{numRaters}$. This weighting correctly pays more attention to instructor ratings with n_i large than with n_i small.

Inferring from a large group, here the average of many raters, to individuals, here each student rater, is called an *ecological regression* (Robinson, 2009). If (7.7) holds for all students and all instructors, then the inference from means to individuals is completely justified by the derivation that leads to (7.8). On the other hand, if different parameters β_{1i} and β_{2i} are required for each instructor, then (7.8) may not hold, even as a reasonable approximation. If the data

¹In Problem 6.10 weighting was ignored for these data. The reader is invited to repeat that problem using weights.

consist only of the instructor averages, we cannot tell if the assumption of the same parameters for all instructors is acceptable or not.

If the n_i observations on the i th instructor are correlated, the variance formula (7.9) is not correct. If all the students in this example were from the same class in a single year, then they may influence each other and induce correlation. We will touch briefly on how to model correlated data in Section 7.4.

7.1.2 Sample Surveys

Sample surveys (Cochran, 1977; Lohr, 2009; Lumley, 2010) are often used to collect data. Suppose we have a finite population of N units, and inferences of interest about these N units are to be based on a subset of n of the units. In a *simple random sample*, all possible samples of n of the N observations are equally likely to be the sample actually collected, and as a result, all units have the sample *inclusion probability*, the probability that a particular unit is included in the sample, of $\pi = n/N$.

Few large-scale surveys actually use a simple random sample, however. For example, in a *stratified random sample*, a population is divided into J subpopulations or strata, with sizes N_1, \dots, N_J . If the within-stratum sample sizes are n_1, \dots, n_J , the inclusion probability for units in stratum j is $\pi_j = n_j/N_j$. The π_j can be different in each stratum. Another alternative that may lead to unequal probability of inclusion is *cluster sampling* or *multistage sampling*. For example, to study schoolchildren, researchers could first take a simple random sample of schools in the school district of interest, and then take a simple random sample of children in a school to study. If the number of children selected is the same in each school, then the inclusion probability for a particular child will depend on the number of children in his or her school, and this is likely to be different for each school.

For illustration, we return to the UN data from Section 3.1 of estimating the regression of `lifeExpF` on `log(ppgdp)` and `fertility`. We will now treat the $N = 199$ localities/countries in the data set as if they were a population. Suppose we divide the world into three strata according to the value of the variable `group`. The number of countries/localities in each of the levels of `group` is

	oecd	other	africa
Count	31	115	53

We take a simple random sample of size $n_j = 20$ separately from each of the three strata, for a total sample size of $n = 60$. In the `oecd` stratum, the inclusion probability is $\pi_1 = 20/31 = 0.65$. The inverse of the inclusion

probability $1/\pi_1 = 31/20 = 1.55$ is called the *sampling weight*; all observations in the `oecd` statum have the same sampling weight. We can interpret the sampling weight as the number of units in a stratum that are represented by each observation in the sample from that stratum. For the `oecd` stratum, each observation represents 1.55 countries in the `oecd`. For `other`, the sampling weight is $1/\pi_2 = 5.75$ and for `africa`, it is $1/\pi_3 = 2.65$, so each observation in these latter two strata represents more countries in their strata than do the observations from the `oecd`. When fitting regression using survey data, wls is appropriate with weights given by the sampling weights. The sampling weights account for differing inclusion probabilities for the units in the sample, not for nonconstant variance.

With survey data, if we fit a regression model like

$$\begin{aligned} E(\text{lifeExpF} \mid \log(\text{ppgdp}), \text{fertility}) \\ = \beta_0 + \beta_1 \log(\text{ppgdp}) + \beta_2 \text{fertility} \end{aligned}$$

there are two distinct approaches to interpreting the β -coefficients. In the *finite population approach*, the β s computed from the OLS regression of the response on the regressors in the whole population are parameters, and these are to be estimated from the data in the sample. The *superpopulation* approach would treat the population as if it were a random sample from a theoretical superpopulation, and then interpret coefficients in the usual way as described in this book. Further discussion of weighting for survey data is given by Lumley (2010, section 5.3).

7.2 MISSPECIFIED VARIANCES

There are many WLS estimators of a regression parameter β , one for each specification of \mathbf{W} . Gilstein and Leamer (1983) described the set of all possible WLS estimates. We consider a more limited goal of describing the bias and variance of estimates based on a misspecified set of weights. The bottom line is: WLS produces unbiased estimates for any choice of \mathbf{W} with positive diagonal elements, but the variance of the estimate is not the matrix produced by regression software, and unless corrected for misspecification, confidence statements and tests can be incorrect.

Using matrix notation, suppose the true regression model is

$$E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \quad \text{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2 \mathbf{W}^{-1}$$

where \mathbf{W} has positive weights on the diagonal and zeroes elsewhere. We get the weights wrong, and fit using OLS, corresponding to assuming

$$E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \quad \text{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2 \mathbf{I}$$

The OLS estimator is

$$\hat{\boldsymbol{\beta}}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Similar to the correct WLS estimate, this estimate is unbiased, $E(\hat{\boldsymbol{\beta}}_0|\mathbf{X}) = \boldsymbol{\beta}$. The variance of this estimate is

$$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \quad (7.10)$$

If $\mathbf{W} = \mathbf{I}$ then the last two terms in (7.10) cancel to give the usual formula for the variance of the WLS estimate (A.28). When $\mathbf{W} \neq \mathbf{I}$, the variance of the estimator has an interesting “sandwich” form.

7.2.1 Accommodating Misspecified Variance

To estimate $\text{Var}(\hat{\boldsymbol{\beta}}_0|\mathbf{X})$ requires estimating $\sigma^2\mathbf{W}^{-1}$ in (7.10). Suppose we let $\hat{e}_i = y_i - \hat{\boldsymbol{\beta}}_0'\mathbf{x}_i$ be the i th residual from the misspecified model. Then \hat{e}_i^2 is an estimate of σ^2/w_i , the i th diagonal element of $\sigma^2\mathbf{W}^{-1}$. Although this seems like a rather poor estimate—after all, it is based mostly on the single observation y_i —theoretical work by Eicker (1963, 1967), Huber (1967), and White (1980), among others, have shown that replacing $\sigma^2\mathbf{W}^{-1}$ by a diagonal matrix with the \hat{e}_i^2 on the diagonal produces a consistent estimate of $\text{Var}(\hat{\boldsymbol{\beta}}_0|\mathbf{X})$.

Several variations of this estimate that are equivalent in large samples but have better small sample behavior have been proposed (Long and Ervin, 2000). The method that appears to be most commonly used is called HC3, and this method estimates the variance (7.10) by

$$\text{Var}(\hat{\boldsymbol{\beta}}_0|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1} \left[\mathbf{X}' \text{diag} \left(\frac{\hat{e}_i^2}{(1-h_{ii})^2} \right) \mathbf{X} \right] (\mathbf{X}'\mathbf{X})^{-1} \quad (7.11)$$

where $\text{diag}()$ means a diagonal matrix, and h_{ii} is the i th leverage, a number between 0 and 1 to be described in Section 9.1.2. An estimator of this type is often called a *sandwich estimator*.

In some fields, (7.11) is used routinely to get standard errors in t -tests for individual coefficients, sometimes without comment. The significance levels obtained by using sandwich estimates are generally larger than significance levels assuming weights are correctly specified, and so the tests done this way are often conservative. The sandwich estimators can be used in F -tests using (6.21) with $\hat{\mathbf{V}}$ given by the right-hand side of (7.11).

Corrections for different types of variance misspecification, such as clustering or autoregressive errors, are described by Long and Ervin (2000) and Zeileis (2004).

Sniffer Data

When gasoline is pumped into a tank, hydrocarbon vapors are forced out of the tank and into the atmosphere. To reduce this significant source of air pollution, devices are installed to capture the vapor. In testing these vapor recovery systems, a “sniffer” measures the amount recovered. To estimate the efficiency of the system, some method of estimating the total amount given off must be used. To this end, a laboratory experiment was conducted in which the amount of vapor given off was measured under controlled conditions. Four predictors are relevant for modeling:

TankTemp = initial tank temperature (degrees F)

GasTemp = temperature of the dispensed gasoline (degrees F)

TankPres = initial vapor pressure in the tank (psi)

GasPres = vapor pressure of the dispensed gasoline (psi)

The response is the hydrocarbons Y emitted in grams. The data, kindly provided by John Rice, are given in the data file `sniffer`, and are shown in Figure 7.2. The clustering of points in many of the frames of this scatterplot matrix is indicative of the attempt of the experimenters to set the predictors at a few nominal values, but the actual values of the predictors measured during the experiment were somewhat different from the nominal. We also see that the predictors are generally linearly related. Some of the predictors, notably the two pressure variables, are closely linearly related, suggesting, as we will see in Chapter 10, that using both in the mean function may not be desirable. For now, however, we will use all four predictors as regressors. Table 7.3 gives the OLS estimates, OLS standard errors, and standard errors based on HC3. The standard errors for TankPres and GasPres are about 25% larger with the HC3 method, but otherwise the standard errors are similar.

7.2.2 A Test for Constant Variance

Cook and Weisberg (1983), with similar work by Breusch and Pagan (1979), provided a diagnostic test for nonconstant variance by building a simple model for the weights w_i . We suppose that for some parameter vector λ and some vector of regressors Z

$$\text{Var}(Y|X, Z = \mathbf{z}) = \sigma^2 \exp(\lambda' \mathbf{z}) \quad (7.12)$$

In this equation the weights are given by $w = 1/\exp(\lambda' \mathbf{z}) = \exp(-\lambda' \mathbf{z})$. If $\lambda = \mathbf{0}$, then (7.12) corresponds to constant variance, so a test of NH: $\lambda = \mathbf{0}$ versus AH: $\lambda \neq \mathbf{0}$ is a test for nonconstant variance. There is great latitude in specifying Z . If $Z = Y$, then variance depends on the response. Similarly, Z may be the same as X , a subset of X , or indeed it could be completely different from X , perhaps

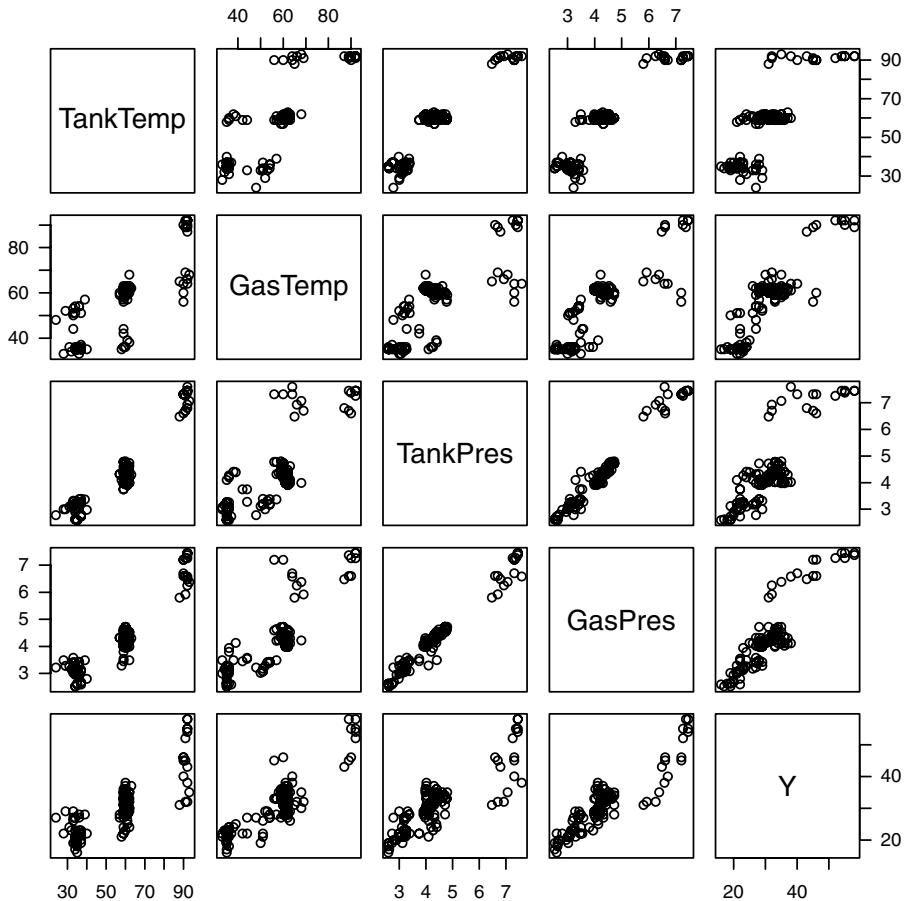


Figure 7.2 Scatterplot matrix for the sniffer data.

Table 7.3 Sniffer Data Estimates and Standard Errors

	OLS Est	OLS SE	HC3 SE
(Intercept)	0.154	1.035	1.047
TankTemp	-0.083	0.049	0.044
GasTemp	0.190	0.041	0.034
TankPres	-4.060	1.580	1.972
GasPres	9.857	1.625	2.056

indicating spatial location or time of observation. In (7.12) $\text{Var}(Y|X, Z = \mathbf{z}) > 0$ for all \mathbf{z} because the exponential function is never negative. The variance is monotonic, either increasing or decreasing, in each component of Z unless interactions are included. The results of Chen (1983) suggest that the tests described here are not very sensitive to the exact functional form used in (7.12), and so the use of the exponential function is relatively benign, and any

form that depends on the linear combination $\lambda'x$ would lead to very similar inference.

Assuming normal errors, a score test of constant variance is available in many statistics packages; Problem 7.5 outlines how to compute this test with any statistical package. The test statistic has an approximate $\chi^2(q)$ distribution, where q is the number of (linearly independent) regressors in Z .

Sniffer Data

In the last section, we found that using the HC3 method to estimate variances give somewhat different answers for two of the coefficients, but we did not provide evidence that nonconstant variance was indeed present. Residual plots can help find nonconstant variance. If the residuals appear to increase or decrease in magnitude as a function of the variable on the horizontal axis, then nonconstant variance may be present.² Figure 7.3a is the plot of residuals versus fitted values. While this plot is far from perfect, it does not suggest the need to worry much about the assumption of nonconstant variance. Figures 7.3b and c, which are plots of residuals against TankTemp and GasPres, respectively, give a somewhat different picture, as particularly in Figure 7.3c variance does appear to increase from left to right. Because none of the graphs in Figure 7.2 have clearly nonlinear mean functions, the inference that variance may not be constant can be tentatively adopted from the residual plots.

Table 7.4 gives the results of several nonconstant variance score tests, each computed using a different choice for Z . The plot shown in Figure 7.3d has estimates of $\hat{\lambda}'x$ on the horizontal axis, where $\hat{\lambda}$ was estimated using the method outlined in Problem 7.5, with Z corresponding to all four of the regressors.

From Table 7.4, we would diagnose nonconstant variance as a function of various choices of Z . We can compare *nested* choices for Z by taking the difference between the score tests and comparing the result with the χ^2 distribution with df equal to the difference in their df (Hinkley, 1985). For example, to compare the 4 df choice of Z to $Z = (\text{TankTemp}, \text{GasPres})$, we can compute $13.76 - 11.78 = 1.98$ with $4 - 2 = 2 df$, to get a p -value of about 0.37, and so the simpler Z with two regressors is adequate. Comparing $Z = (\text{TankTemp}, \text{GasPres})$ with $Z = \text{GasPres}$, the test statistic is $11.78 - 9.71 = 2.07$ with $2 - 1 = 1 df$, giving a p -value of about 0.15, so once again the simpler choice of Z seems adequate. A reasonable approach to working with these data is to assume that

$$\text{Var}(Y|X, Z) = \sigma^2 \times \text{GasPres}$$

and use $1/\text{GasPres}$ as weights in weighted least squares.

²We will see in Chapter 9 that other problems besides nonconstant variance may produce this pattern if the regressors have nonlinear relationships.

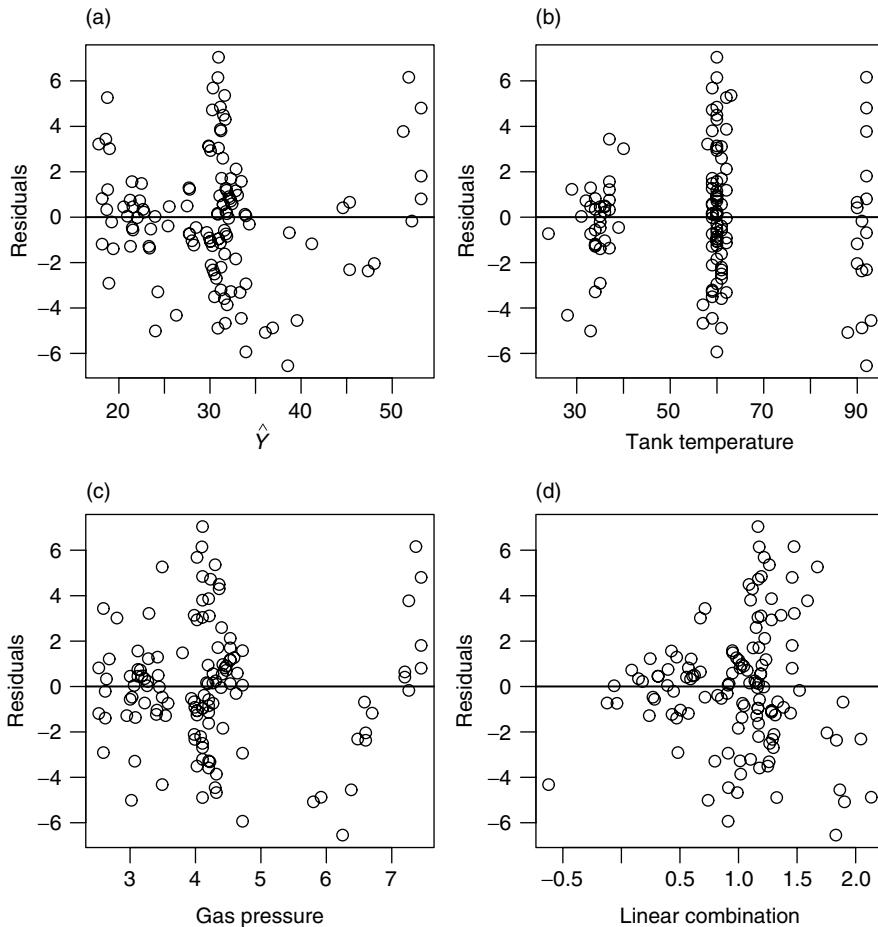


Figure 7.3 Residuals plots for the sniffer data with variance assumed to be constant.

Table 7.4 Score Tests for the Sniffer Data

Choice for Z	df	Test stat.	p -Value
GasPres	1	5.50	.019
TankTemp	1	9.71	.002
TankTemp, GasPres	2	11.78	.003
TankTemp, GasTemp, TankPres, GasPres	4	13.76	.008
Fitted values	1	4.80	.028

Pinheiro and Bates (2000, section 5.2) start with the model for nonconstant variance (7.12) and discuss methodology for both testing $\lambda = \mathbf{0}$, and for estimating λ . They also consider a number of alternative models for nonconstant variance.

7.3 GENERAL CORRELATION STRUCTURES

The *generalized least squares* or GLS model extends WLS one step further, and starts with

$$E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \quad \text{Var}(\mathbf{Y}|\mathbf{X}) = \boldsymbol{\Sigma} \quad (7.13)$$

where $\boldsymbol{\Sigma}$ is an $n \times n$ positive definite symmetric matrix. The WLS model uses $\boldsymbol{\Sigma} = \sigma^2 \mathbf{W}^{-1}$, and the OLS model uses $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$. If $\boldsymbol{\Sigma}$ is fully known, meaning that all $n(n + 1)/2$ unique variances and covariances between the elements of the response are completely known, then GLS estimation is completely analogous to WLS estimation with $\boldsymbol{\Sigma}$ substituted for $\sigma^2 \mathbf{W}^{-1}$, and $\sigma^2 = 1$.

If we have n observations and $\boldsymbol{\Sigma}$ is completely unknown, then the total number of parameters is the number of regression coefficients p' plus n variances on the diagonal of $\boldsymbol{\Sigma}$ plus $n(n - 1)/2$ covariances on the off-diagonals of $\boldsymbol{\Sigma}$, many more parameters than observations. The only hope in fitting (7.13) with $\boldsymbol{\Sigma}$ unknown is in introducing some structure in $\boldsymbol{\Sigma}$ so it depends on a small number of parameters. Three of many possible choices are:

Compound Symmetry If all the observations are equally correlated, then

$$\boldsymbol{\Sigma}_{\text{CS}} = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

has only two parameters, ρ and σ^2 . Generalized least squares software, such as the `gls` function in the `nlme` package in R (Pinheiro and Bates, 2000), can be used to do the estimation. Interpretation would be similar to WLS fitting.

Autoregressive This form is generally associated with time series (Box et al., 2008; Tsay, 2005). If data are time ordered and equally spaced, the lag-1 autoregressive covariance structure is

$$\boldsymbol{\Sigma}_{\text{AR}} = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{pmatrix}$$

In this form, the correlation between two observations t time units apart is ρ^t . Again, there are only two parameters in the covariance matrix.

Block Diagonal A block diagonal form for $\boldsymbol{\Sigma}$ can arise if observations are sampled clusters. For example, a study of school performance might

sample m children from each of k classrooms. The m children within a classroom may be correlated because they all have the same teacher, but children in different classrooms are independent. If $\Sigma_{CS}(m)$ is an $m \times m$ covariance matrix for compound symmetry, then Σ could be a matrix with $\Sigma_{CS}(m)$ repeated k times along its diagonal, and with 0 in all other locations. This matrix also has only two parameters, but hints at the general structures that are possible with just a few parameters.

7.4 MIXED MODELS

An important and popular extension of the linear model theory used in this book is to *mixed models*. This is an enormous topic, and comprehensive treatment of it is beyond the scope of this book. We present here an example to illustrate a few of the possibilities. Useful and more comprehensive references include Fitzmaurice et al. (2011), Goldstein (2010), McCulloch et al. (2008), Raudenbush and Bryk (2002), and Zuur et al. (2009).

Psychophysics is the branch of psychology that explores the relationship between physical stimuli and psychological responses (Stevens, 1966; Varshney and Sun, 2013). Theory suggests that psychological responses are proportional to the magnitude of the stimulus raised to a power, so the regression of the log-response on the log-stimulus should have a linear mean function.

In a classic experiment, S. S. Stevens (1906–1973) and his colleagues played tones of varying loudness to each of $m = 10$ subjects. Each subject heard three replications of tones at 50, 60, 70, 80, and 90 decibels (db), presented in random order. The subject was asked to draw a line whose length in cm corresponds to the loudness of the tone. The data file *Stevens* presents the data, including variables *y*, the average of the logarithms of the three lengths for each value of *loudness*, and a factor for *subject*. Since the decibel scale is logarithmic, within subject the regression of *y* on *loudness* should be have a straight-line mean function.

In Figure 7.4 a line joins the data for each of the subjects, so 10 lines are shown on the figure. The psychophysical model seems plausible here, but each subject may have a different slope and intercept.

This experiment has several features that were not apparent in earlier examples in this book. First, each subject was measured several times. While observations on different subjects are likely to be independent, observations on the same subject but at different values of *loudness* are likely to be correlated. Analysis should account for this correlation.

A second important feature of this experiment is that the levels of the factor *subject* can be thought of as a random sample of possible subjects who might have been included in the experiment. The fit of a model for a particular subject may not be of particular interest. In contrast, in the UN data example the fit of a model within each level of *group* might be of primary interest.

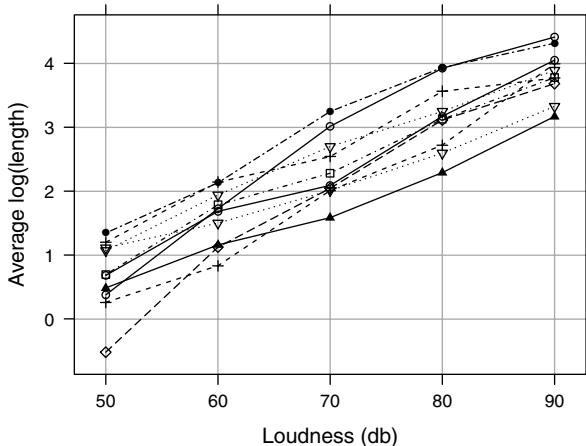


Figure 7.4 Psychophysics example.

The random coefficients model, as special case of mixed models, allows for appropriate inferences. First, we hypothesize the existence of a population regression mean function,

$$E(y| \text{loudness} = x) = \beta_0 + \beta_1 x \quad (7.14)$$

Primary interest could be in the estimates of these parameters, and on whether or not the psychophysics model that leads to (7.14) is supported by the data.

Subject effects are not included in (7.14). To add them we hypothesize that each of the subjects may have his or her own slope and intercept. Let y_{ij} , $i = 1, \dots, 10$, $j = 1, \dots, 5$ be the log-response for subject i measured at the j th level of loudness. For the i th subject,

$$E(y_{ij} | \text{loudness} = x, b_{0i}, b_{1i}) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) \text{loudness}_{ij} \quad (7.15)$$

The b_{0i} and b_{1i} are the deviations from the population intercept and slope for the i th subject. The new feature is that we treat the b_{ji} as random variables,

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{pmatrix}\right)$$

Inferences about (β_0, β_1) concern population behavior. Inferences about (τ_0^2, τ_1^2) concern the variation of the intercepts and slopes between individuals in the population. Model (7.15) cannot be fit directly because the b_{ji} are unobservable random variables. In fitting we need to average over their distribution,

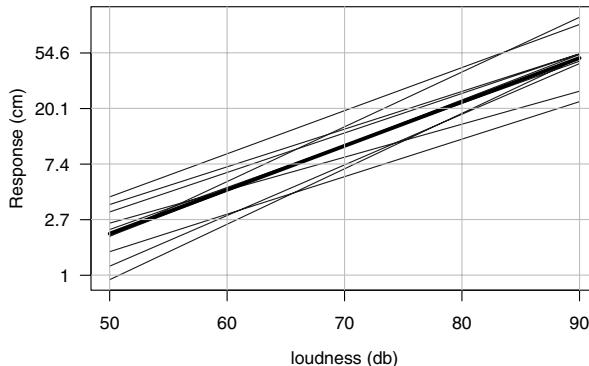


Figure 7.5 Fitted mixed model for the psychophysics example. The thick line is the population estimate, while the thinner lines are predicted lines for individuals.

and this will induce correlation between observations on the same subject, as desired.

Figure 7.5 summarizes the fitted model for these data. The thick line has the estimates of β_0 and β_1 as intercept and slope. The gray lines are “predicted” lines for each individual subject. The gray lines indicate the variation in fitted models that could be expected for other subjects from the same population.

There is plenty of standard jargon that describes this model. This is a *repeated measures* problem because each experimental unit, a subject, is measured repeatedly. It is a *random coefficients* model because each subject has his or her own slope and intercept. If there were additional regressors that describe the subjects, for example, gender, age, or others, then this is a *hierarchical* or *multilevel* problem.

7.5 VARIANCE STABILIZING TRANSFORMATIONS

Suppose that the response is strictly positive, and the variance function is

$$\text{Var}(Y|X = \mathbf{x}) = \sigma^2 g(\mathbb{E}(Y|X = \mathbf{x})) \quad (7.16)$$

where $g(\mathbb{E}(Y|X = \mathbf{x}))$ is a function that is increasing with the value of its argument. For example, if the distribution of $Y|X$ has a Poisson distribution, then $g(\mathbb{E}(Y|X = \mathbf{x})) = \mathbb{E}(Y|X = \mathbf{x})$, since for Poisson variables, the mean and variance are equal. Although the regression models for Poisson and for binomial responses introduced in Chapter 12 are commonly used, an alternative approach is to transform the response so the transformed response has an approximately constant variance function (Scheffé, 1959, section 10.7). Table 7.5 lists the common variance stabilizing transformations. Of course, transforming away nonconstant variance can introduce nonlinearity into the mean function, so this option may not always be reasonable.

Table 7.5 Common Variance Stabilizing Transformations

Y_T	Comments
\sqrt{Y}	Used when $\text{Var}(Y X) \propto E(Y X)$, as for Poisson distributed data. $Y_T = \sqrt{Y} + \sqrt{Y+1}$ can be used if many of the counts are small (Freeman and Tukey, 1950).
$\log(Y)$	Use if $\text{Var}(Y X) \propto [E(Y X)]^2$. In this case, the errors behave like a percentage of the response, $\pm 10\%$, rather than an absolute deviation, ± 10 units.
$1/Y$	The inverse transformation stabilizes variance when $\text{Var}(Y X) \propto [E(Y X)]^4$. It can be appropriate when responses are mostly close to 0, but occasional large values occur.
$\sin^{-1}(\sqrt{Y})$	The <i>arcsine square-root</i> transformation is used if Y is a proportion between 0 and 1, but it can be used more generally if y has a limited range by first transforming Y to the range $(0, 1)$, and then applying the transformation.

The square root, $\log(Y)$, and $1/Y$ are appropriate when variance increases or decreases with the response, but each is more severe than the one before it. The square-root transformation is relatively mild and is most appropriate when the response follows a Poisson distribution. The logarithm is the most commonly used transformation; the base of the logarithms is irrelevant. It is appropriate when the error standard deviation is a percentage of the response, such as $\pm 10\%$ of the response, not ± 10 units, so $\text{Var}(Y|X) \propto \sigma^2 [E(Y|X)]^2$.

The reciprocal or inverse transformation is often applied when the response is a time until an event, such as time to complete a task, or until healing. This converts times per event to a rate per unit time; often the transformed measurements may be multiplied by a constant to avoid very small numbers. Rates can provide a natural measurement scale.

7.6 THE DELTA METHOD

The *delta method* is used to obtain standard errors for nonlinear combinations of estimated coefficients. For example, we have seen at Equation (5.10) in Section 5.3 that the value of the predictor that will maximize or minimize a quadratic, depending on the signs of the β s, is $x_M = -\beta_1/(2\beta_2)$. This is a nonlinear combination of the β s, and so its estimate, $\hat{x}_M = -\hat{\beta}_1/(2\hat{\beta}_2)$, is a nonlinear combination of estimates. The delta method provides an approximate standard error of a nonlinear combination of estimates that is accurate in large samples. The derivation of the delta method, and possibly its use, requires elementary calculus.

We will use different notation for this derivation to emphasize that the results are much more general than just for ratios of coefficient estimates in

multiple linear regression. Let $\boldsymbol{\theta}$ be a $k \times 1$ parameter vector, with estimator $\hat{\boldsymbol{\theta}}$ such that

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \sigma^2 \mathbf{D}) \quad (7.17)$$

where \mathbf{D} is a known, positive definite, matrix. Equation (7.17) can be exact, as it is for the multiple linear regression model with normal errors, or asymptotically valid, as in nonlinear or generalized linear models. In some problems, σ^2 may be known, but in the multiple linear regression problem, it is usually unknown and will be estimated from data.

Suppose $g(\boldsymbol{\theta})$ is a nonlinear continuous function of $\boldsymbol{\theta}$ that we would like to estimate and that $\boldsymbol{\theta}^*$ is the true value of $\boldsymbol{\theta}$. To approximate $g(\hat{\boldsymbol{\theta}})$, we can use a Taylor series expansion, as in Section 11.1, about $g(\boldsymbol{\theta}^*)$,

$$\begin{aligned} g(\hat{\boldsymbol{\theta}}) &= g(\boldsymbol{\theta}^*) + \sum_{j=1}^k \frac{\partial g}{\partial \theta_j} (\hat{\theta}_j - \theta_j^*) + \text{small terms} \\ &\approx g(\boldsymbol{\theta}^*) + \dot{\mathbf{g}}(\boldsymbol{\theta}^*)' (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \end{aligned} \quad (7.18)$$

where we have defined

$$\dot{\mathbf{g}}(\boldsymbol{\theta}^*) = \frac{\partial g}{\partial \boldsymbol{\theta}} = \left(\frac{\partial g}{\partial \theta_1}, \dots, \frac{\partial g}{\partial \theta_k} \right)'$$

evaluated at $\boldsymbol{\theta}^*$. The vector $\dot{\mathbf{g}}$ has dimension $k \times 1$. We have expressed in (7.18) our estimate $g(\hat{\boldsymbol{\theta}})$ as approximately a constant $g(\boldsymbol{\theta}^*)$ plus a linear combination of data. The variance of a constant is 0, as is the covariance between a constant and a function of data. We can therefore approximate the variance of $g(\hat{\boldsymbol{\theta}})$ by

$$\begin{aligned} \text{Var}[g(\hat{\boldsymbol{\theta}})] &= \text{Var}[g(\boldsymbol{\theta}^*)] + \text{Var}[\dot{\mathbf{g}}(\boldsymbol{\theta}^*)' (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)] \\ &= \dot{\mathbf{g}}(\boldsymbol{\theta}^*)' \text{Var}(\hat{\boldsymbol{\theta}}) \dot{\mathbf{g}}(\boldsymbol{\theta}^*) \\ &= \sigma^2 \dot{\mathbf{g}}(\boldsymbol{\theta}^*)' \mathbf{D} \dot{\mathbf{g}}(\boldsymbol{\theta}^*) \end{aligned} \quad (7.19)$$

This equation is the heart of the delta method, so we will write it out again as a scalar equation. Let \dot{g}_i be the j th element of $\dot{\mathbf{g}}(\hat{\boldsymbol{\theta}})$, so \dot{g}_i is the partial derivative of $g(\boldsymbol{\theta})$ with respect to θ_i , and let d_{ij} be the (i,j) -element of the matrix \mathbf{D} . Then the estimated variance of $g(\hat{\boldsymbol{\theta}})$ is

$$\text{Var}[g(\hat{\boldsymbol{\theta}})] = \sigma^2 \sum_{i=1}^k \sum_{j=1}^k \dot{g}_i \dot{g}_j d_{ij} \quad (7.20)$$

In practice, all derivatives are evaluated at $\hat{\boldsymbol{\theta}}$, and σ^2 is replaced by its estimate.

In large samples and under regularity conditions, $g(\hat{\boldsymbol{\theta}})$ will be normally distributed with mean $g(\boldsymbol{\theta}^*)$ and variance (7.19). In small samples, the normal approximation may be poor, and inference based on the bootstrap might be preferable.

For quadratic regression at Equation (5.9), the minimum or maximum occurs at $g(\boldsymbol{\beta}) = -\beta_1/(2\beta_2)$, which is estimated by $g(\hat{\boldsymbol{\beta}})$. To apply the delta method, we need the partial derivative, evaluated at $\hat{\boldsymbol{\beta}}$,

$$\left(\frac{\partial g}{\partial \boldsymbol{\beta}} \right)' = \left(0, -\frac{1}{2\hat{\beta}_2}, \frac{\hat{\beta}_1}{2\hat{\beta}_2^2} \right)$$

Using (7.20), straightforward calculation gives

$$\text{Var}(g(\hat{\boldsymbol{\beta}})) = \frac{1}{4\hat{\beta}_2^2} \left(\text{Var}(\hat{\beta}_1) + \frac{\hat{\beta}_1^2}{\hat{\beta}_2^2} \text{Var}(\hat{\beta}_2) - \frac{2\hat{\beta}_1}{\hat{\beta}_2} \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \right) \quad (7.21)$$

The variances and covariances in (7.21) are elements of the matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, and so the estimated variance is obtained from $\hat{\sigma}^2 \mathbf{D} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$.³

As a modestly more complicated example, the estimated mean function for palatability for the cake data (Section 5.3.1) when the temperature is 350 degrees is given by (5.14). The estimated maximum palatability occurs when the baking time is

$$\hat{x}_M = -\frac{\hat{\beta}_1 + \hat{\beta}_{12}(350)}{2\hat{\beta}_{11}} = 36.2 \text{ min}$$

which depends on the estimate $\hat{\beta}_{12}$ for the interaction as well as on the linear and quadratic terms for X_1 . The standard error from the delta method can be computed to be 0.4 minutes. If we can believe the normal approximation, a 95% confidence interval for x_M is $36.2 \pm 1.96 \times 0.4$ or about 35.4–37.0 minutes.

Writing a function for computing the delta method is not particularly hard using a language such as Maple, Mathematica, Matlab, or R that can do symbolic differentiation to get \dot{g} . If your package will not do the differentiation for you, then you can still compute the derivatives by hand and use (7.20) to get the estimated standard error. The estimated variance matrix $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ is computed by all standard regression programs, although getting access to it may not be easy in all programs.

7.7 THE BOOTSTRAP

The *bootstrap* provides a computationally intensive alternative method used primarily for computing standard errors, confidence intervals, and tests when

³The estimated variance $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ could be replaced by a sandwich estimator, as in Section 7.2.1.

either the assumptions needed for standard methods are questionable, or where standard methods are not readily available. We start with a simple example of the latter, and then turn to applying the bootstrap to regression problems.

Suppose we have a sample y_1, \dots, y_n from a particular distribution G , for example, a standard normal distribution. What is a confidence interval for the population median?

Rather than developing the theory to compute this interval, or googling the problem to see if someone else has done the work, we can obtain an approximate answer to this question by computer simulation, set up as follows:

1. Obtain a simulated random sample y_1^*, \dots, y_n^* from the known distribution G . Most statistical computing languages and even spreadsheet programs include functions for simulating random deviates (see Thisted, 1988, for computational methods).
2. Compute and save the median of the sample in step 1.
3. Repeat steps 1 and 2 a large number of times, say B times. The larger the value of B , the more precise the ultimate answer.
4. If we take $B = 999$, a simple *percentile-based* 95% confidence interval for the median is the interval between the 25th smallest value and the 975th largest value, which are the sample 2.5 and 97.5 percentiles, respectively.

In most interesting problems, we will not actually know G and so this simulation is not possible. Efron (1979) pointed out that the observed data can be used to estimate G , and then we can sample from the estimate \hat{G} . The algorithm becomes as follows:

1. Obtain a random sample y_1^*, \dots, y_n^* from \hat{G} by sampling *with replacement* from the observed values y_1, \dots, y_n . In particular, the j th element of the sample y_i^* is equally likely to be any of the original y_1, \dots, y_n . Some of the y_i will appear several times in the random sample, while others will not appear at all.
2. Continue with steps 2–4 of the first algorithm.

A test at the 5% level concerning the population median can be rejected if the hypothesized value of the median does not fall in the confidence interval.

Efron called this method the bootstrap, and we call B the number of bootstrap samples. Excellent references for the bootstrap are the books by Efron and Tibshirani (1993) and Davison and Hinkley (1997).

7.7.1 Regression Inference without Normality

Bootstrap methods can be applied in more complex problems like regression. Inferences and accurate standard errors for parameters and mean functions

require either normality of regression errors or large sample sizes. In small samples without normality, standard inference methods can be misleading, and in these cases, a bootstrap can be used for inference.

Transactions Data

For each of n branches of a large Australian bank, we have recorded the number t_1 of type 1 transactions, the number t_2 of type 2 transactions, and the total number of minutes time of labor used by the branch. For $j = 1, 2$ suppose the time of a transaction of type j is like a draw from a distribution with mean β_j minutes, and for simplicity suppose the standard deviation is τ minutes for either type of transaction. The total number of minutes is expected to be

$$E(\text{time}|t_1, t_2) = \beta_0 + \beta_1 t_1 + \beta_2 t_2 \quad (7.22)$$

possibly with $\beta_0 = 0$ because 0 transactions should imply 0 time spent. The response time is a sum, and the variance is a sum of the individual variances,

$$\text{Var}(\text{time}|t_1, t_2) = (t_1 + t_2)\tau^2$$

meaning that the variance should be larger in branches with more transactions.

The data are displayed in Figure 7.6, and are given in the data file `Transact`. The key features of the scatterplot matrix are (1) the marginal response plots in the last row appear to have reasonably linear mean functions; (2) there appear to be a number of branches with no t_1 transactions but many t_2 transactions; and (3) in the plot of time versus t_2 , variability appears to increase from left to right, as expected by the derivation in the last paragraph.

A *case resampling* bootstrap (Davison and Hinkley, 1997, p. 264) is computed as follows:

1. Number the cases in the data set from 1 to n . Take a random sample *with replacement* of size n from these case numbers.
2. Create a data set from the original data, but repeating each row in the data set the number of times that row was selected in the random sample in step 1. Some cases will appear several times, and others will not appear at all. Compute the regression using this data set, and save the values of the coefficient estimates.
3. Repeat steps 1 and 2 a large number of times, say, B times.
4. Estimate a 95% confidence interval for each of the estimates by the 2.5 and 97.5 percentiles of the sample of B bootstrap samples. A more accurate method called the *bias corrected and accelerated* or BCa method, discussed by Efron and Tibshirani (1993, chapter 14) and Davison and

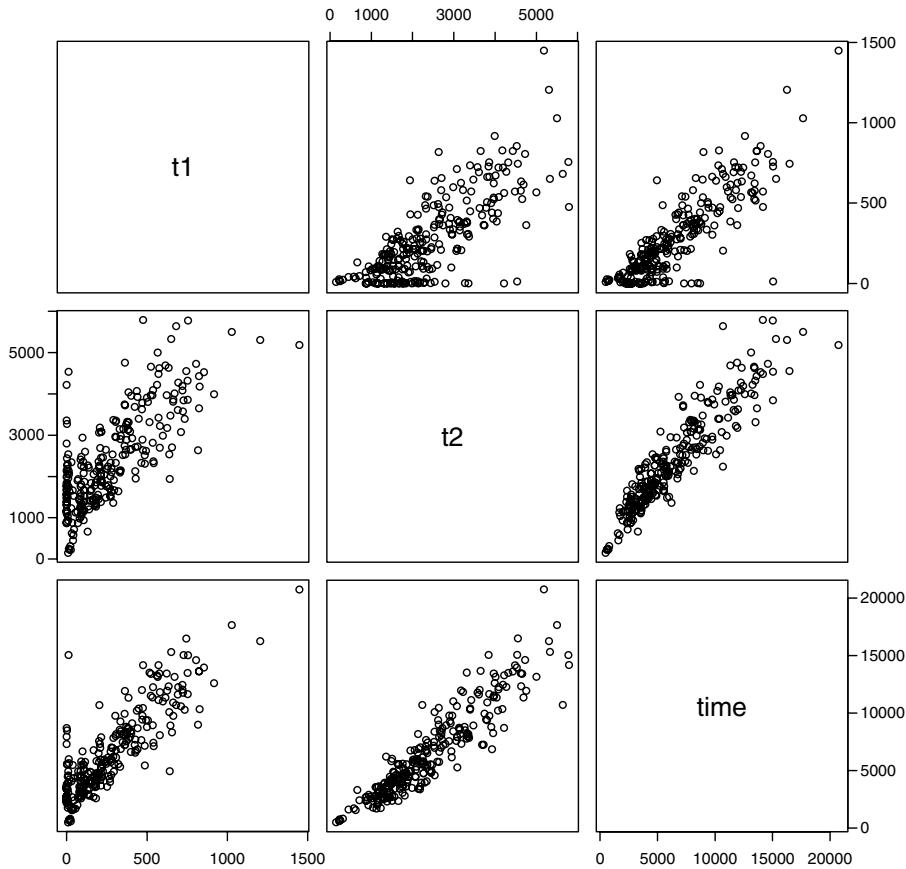


Figure 7.6 Scatterplot matrix for the transactions data.

Hinkley (1997, section 3.9), produces an interval based on different sample quantiles that depend on the data. The BCa method generally produces narrower intervals and is the usual default method used in statistical software.

Many standard computing packages, including R, SAS, Stata, and SPSS, have procedures available that implement the bootstrap for regression.

Table 7.6 summarizes some of the bootstrap results. The column marked “OLS” shows OLS estimates from the regression of time on t1 and t2. The column marked “boot” gives the average of $B = 999$ case bootstraps. In principle, the values in both columns are estimating the same quantities. The difference between these two columns is called the bootstrap bias, and it is given in the third column in the table. The estimates of t1 and t2 agree to two digits, while the bias in the intercept appears to be larger. The “OLS.SE” is the standard error from the usual regression formula (3.14), while the “boot.SE” is the

Table 7.6 Summary Statistics for Case Bootstrap in the Transactions Data

	OLS	boot	bias	OLS.SE	boot.SE
(Intercept)	144.37	159.20	-14.83	170.54	188.54
t1	5.46	5.51	-0.04	0.43	0.66
t2	2.03	2.02	0.01	0.09	0.15

Table 7.7 95% Confidence Intervals for the Transactions Data

Method	(Intercept)	t1	t2
Normal theory	(-191.47, 480.21)	(4.61, 6.32)	(1.85, 2.22)
Normal with boot SE	(-240.00, 499.08)	(4.12, 6.72)	(1.75, 2.34)
Percentile	(-204.33, 538.92)	(4.20, 6.80)	(1.73, 2.32)
BCa	(-259.44, 487.18)	(3.88, 6.64)	(1.79, 2.38)

estimated standard errors from the bootstrap, which is the standard deviation of the $B = 999$ bootstrap estimates. We see first that the biases are small relative to either set of standard errors, and the standard errors from OLS are overly optimistic relative to the bootstrap estimated standard errors, particularly for the coefficients for t1 and t2.

Table 7.7 reports confidence intervals for the three coefficients. The first row of the table gives the standard method based on OLS estimates and large sample theory, from (2.14). The second method also uses (2.14), but it substitutes the bootstrap estimated standard error for the OLS standard error. From Table 7.6, the bootstrap standard errors are uniformly larger than the OLS standard errors, and so these intervals are larger than the OLS intervals but are still symmetric about the OLS point estimate. The percentile and BCa methods are given in the last two rows. The normal theory confidence intervals are probably too short, and any of the other three methods appear to provide more reasonable intervals.

7.7.2 Nonlinear Functions of Parameters

Suppose we wanted to get a confidence interval for the ratio β_1/β_2 in the transactions data. The point estimate from the OLS fit is $\hat{\beta}_1/\hat{\beta}_2 = 2.68$. Section 7.6 shows how to compute an approximate standard error for this ratio using normal theory via the delta method, but we can use the bootstrap without the need for any additional theoretical calculations.

In the $B = 999$ bootstrap samples, the mean ratio was 2.76. The standard deviation of the B ratios is the bootstrap estimated standard error of the ratio, which turns out to be 0.52. The percentile-based confidence interval for the ratio is from the 25th smallest of the ratios to the 975th largest, or from 1.85 to 3.89.

7.7.3 Residual Bootstrap

The case resampling plan outlined in Section 7.7.1 resamples from the joint distribution of the response and the terms. An alternative uses *residual resampling* in which the residuals from the initial fit are resampled. Here is the general algorithm:

1. Given data (\mathbf{x}_i, y_i) , $i = 1, \dots, n$, fit the linear regression model $E(Y|X = \mathbf{x}) = \boldsymbol{\beta}'\mathbf{x}$ and compute the OLS estimator $\hat{\boldsymbol{\beta}}$, and the residuals, $\hat{e}_i = y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i$.
2. Randomly sample from the residuals to get a new sample (e_1^*, \dots, e_n^*) , where e_i^* is equally likely to be any of $(\hat{e}_1, \dots, \hat{e}_n)$. A modified definition of residuals can be used that may slightly improve performance by correcting for unequal variances of the residuals; see Davison and Hinkley (1997, p. 262).
3. Create a bootstrap response with elements $y_i^* = \hat{\boldsymbol{\beta}}'\mathbf{x}_i + \hat{e}_i^*$. Compute the regression of the bootstrap response on X , and get the coefficient estimates or other summary statistic of interest.
4. Repeat steps 2 and 3 B times. Confidence intervals are obtained as with the case bootstrap.

This sampling procedure assumes the linear mean function is correct, and that variance is constant. If these assumptions are correct, this resampling method can be more accurate than case resampling.

7.7.4 Bootstrap Tests

The presentation of the bootstrap here has emphasized its use to estimate variances and to compute confidence intervals. The bootstrap can also be used for testing hypotheses. When testing a single coefficient to be equal to 0, one could reject the null hypothesis at level one minus the stated confidence level if the bootstrap confidence interval for that coefficient does not include the null value of 0 (Davison and Hinkley, 1997, section 5.5). More complicated tests equivalent to F -tests or tests for several coefficients simultaneously (Chapter 6) can also be carried out with minimal assumptions using a bootstrap, but the methodology is beyond the level of presentation in this book. See Davison and Hinkley (1997, section 6.3.2) for a readable summary.

7.8 PROBLEMS

- 7.1** Sue fits a WLS regression with all weights equal to 2. Joe fits a WLS regression to the same data with all weights equal to 1. What are the differences

in estimates of coefficients, standard errors, σ^2 , F -tests between Sue's and Joe's analyses?

- 7.2** (Data file: *physics1*) The data file *physics1* gives the results of the experiment described in Section 7.1, except in this case, the input is the π^- meson as before, but the output is the π^+ meson.

Analyze these data following the analysis done in the text, and summarize your results.

- 7.3** Large public surveys such as the Youth Risk Behavior Survey conducted by the Centers for Disease Control (2013) often provide weights to be used in an analysis.

- 7.3.1** Subpopulations may be oversampled to insure that the number of participants in a particular subpopulation is large enough to get estimates of the desired precision. For example, if estimates are required for each state separately, then states with smaller population would need to be oversampled relative to states with a large population to get the same precision.

In combining data over subpopulations, should larger weight be given to observations from the oversampled subpopulation to those not in the oversampled subpopulation?

- 7.3.2** Nonresponse is a common problem in surveys. For example, if a sample size of 1000 is planned in a particular state but in that state only 600 responses are obtained, if we are prepared to believe that the responders in the subpopulation are no different from the nonresponders, then we can weight the responders to represent the nonresponders.

In combining data over subpopulations, should larger weight be given to observations that represent the nonresponders in the subpopulation or smaller weight?

- 7.4** (Data file: *salarygov*) Refer to Problems 5.9 and 6.11.

- 7.4.1** The data as given have as its unit of analysis the *job class*. In a study of the dependence of maximum salary on skill, one might prefer to have the *employee* as the unit of analysis. Explain why changing the unit of analysis to the employee rather than the job class would suggest using wls. What are the relevant weights?

- 7.4.2** Repeat Problem 6.11, but use wls. Do any conclusions change?

- 7.5** The score test for nonconstant variance, Section 7.2.2, is available as an option in many standard regression packages. If not available, it can be computed using the following prescription. Suppose X represents the regressors in the mean function and Z the regressors in the variance function (7.12).

1. Assume $\lambda = \mathbf{0}$ and use OLS to fit with the mean function $E(Y|X=\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x}$. Save the residuals \hat{e}_i and the residual sum of squares RSS.
2. Compute scaled squared residuals $u_i = n\hat{e}_i^2/\text{RSS}$. We combine the u_i into a regressor U .
3. Fit OLS with the mean function $E(U|Z=\mathbf{z}) = \lambda_0 + \boldsymbol{\lambda}'\mathbf{z}$. Obtain SSreg for this regression with $df = q$, the number of components in Z . If variance is thought to be a function of the responses, then in this regression, replace Z by the fitted values from the regression in step 1. The SSreg then will have 1 df.
4. Compute the score test, $S = \text{SSreg}/2$. The significance level for the test can be obtained by comparing S with its asymptotic distribution, which, under the hypothesis $\lambda = \mathbf{0}$, is $\chi^2(q)$. If $\lambda \neq \mathbf{0}$, then S will be too large, so large values of S provide evidence against the hypothesis of constant variance.

Reproduce the score tests given for the sniffer data in Section 7.2.1.

- 7.6** (Data file: stopping) The (hypothetical) data in the file give automobile stopping Distance in feet and Speed in mph for $n = 62$ trials of various automobiles (Ezekiel and Fox, 1959).

- 7.6.1** Draw a scatterplot of Distance versus Speed. Explain why this graph supports fitting a quadratic regression model.
- 7.6.2** Fit the quadratic model but with constant variance. Compute the score test for nonconstant variance for the alternatives that (a) variance depends on the mean; (b) variance depends on Speed; and (c) variance depends on Speed and Speed². Is adding Speed² helpful?
- 7.6.3** Refit the quadratic regression model assuming $\text{Var}(\text{Distance}|\text{Speed}) = \text{Speed } \sigma^2$. Compare the estimates and their standard errors with the unweighted case.
- 7.6.4** Based on the unweighted model, use a sandwich estimator of variance to correct for nonconstant variance. Compare with the results of the last subproblem.
- 7.6.5** Fit the unweighted quadratic model, but use a case resampling bootstrap to estimate standard errors, and compare with the previous methods.

- 7.7 Galton's sweet peas** (Data file: galtonpeas) Many of the ideas of regression first appeared in the work of Sir Francis Galton (1822–1911) on the inheritance of characteristics from one generation to the next. In Galton (1877), he discussed experiments on sweet peas. By comparing the sweet peas produced by parent plants to those produced by offspring plants, he could observe inheritance from one generation to the next. Galton categorized parent plants according to the typical diameter of the

peas they produced. For seven size classes from 0.15 to 0.21 inches, he arranged for each of nine of his friends to grow 10 plants from seed in each size class; however, two of the crops were total failures. A summary of Galton's data were later published in Pearson (1930). The data file includes Parent diameter, Progeny diameter, and SD the standard deviation of the progeny diameters. Sample sizes are unknown but are probably large.

- 7.7.1** Draw the scatterplot of Progeny versus Parent.
- 7.7.2** Assuming that the standard deviations given are population values, compute the weighted regression of Progeny on Parent. Draw the fitted mean function on your scatterplot.
- 7.7.3** Galton took the average size of all peas produced by a plant to determine the size class of the parental plant. Yet for seeds to represent that plant and produce offspring, Galton chose seeds that were as close to the overall average size as possible. Thus, for a small plant, the exceptional large seed was chosen as a representative, while larger, more robust plants were represented by relatively smaller seeds. What effects would you expect these experimental biases to have on (1) estimation of the intercept and slope and (2) estimates of error?
- 7.8 Jevons's gold coins** (Data file: *jevons*) The data in this example are deduced from a diagram in Jevons (1868) and provided by Stephen M. Stigler. In a study of coinage, Jevons weighed 274 gold sovereigns that he had collected from circulation in Manchester, England. For each coin, he recorded the weight after cleaning to the nearest 0.001 g, and the date of issue. The data file includes Age, the age of the coin in decades, n, the number of coins in the age class, Weight, the average weight of the coins in the age class, SD, the standard deviation of the weights. The minimum Min and maximum Max of the weights are also given. The standard weight of a gold sovereign was 7.9876 g; the minimum legal weight was 7.9379 g.
- 7.8.1** Draw a scatterplot of Weight versus Age, and comment on the applicability of the usual assumptions of the linear regression model. Also draw a scatterplot of SD versus Age, and summarize the information in this plot.
- 7.8.2** To fit a simple linear regression model with Weight as the response, wls should be used with variance function $\text{Var}(\text{Weight}|\text{Age}) = n\sigma^2/\text{SD}^2$. Sample sizes are large enough to assume the SD are population values. Fit the wls model.
- 7.8.3** Is the fitted regression consistent with the known standard weight for a new coin?
- 7.8.4** For previously unsampled coins of Age = 1, 2, 3, 4, 5, estimate the probability that the weight of the coin is less than the legal minimum.

(*Hints:* The standard error of prediction is the square root of the sum of two terms, the assumed known variance of an unsampled coin of known Age, which is different for each age, and the estimated variance of the fitted value for that Age; the latter is computed from the formula for the variance of a fitted value. You should use the normal distribution rather than a *t* to get the probabilities.)

- 7.8.5** Determine the Age at which the predicted weight of coins is equal to the legal minimum, and use the delta method to get a standard error for the estimated age. This problem is called *inverse regression*, and is discussed by Brown (1993).

7.9 Bootstrap for a median (Data file: UN11)

- 7.9.1** Find a 95% confidence interval for the mean of $\log(\text{fertility})$. Then, obtain an approximate 95% confidence interval for the median of fertility by exponentiating the end points of the interval for the mean of $\log(\text{fertility})$.
- 7.9.2** Use the bootstrap to obtain a 95% confidence interval for the median of fertility . Compare with the interval for the mean of fertility from Problem 7.9.1.

7.10 (Data file: fuel2001)

- 7.10.1** Use the bootstrap to estimate confidence intervals for the coefficients in the fuel data, and compare the results with the usual large sample OLS estimates.
- 7.10.2** Examine the histograms of the bootstrap replications for each of the coefficients. Are the histograms symmetric or skewed? Do they look like normally distributed data, as they would if the large sample normal theory applied to these data? Do the histograms support or refute the differences between the bootstrap and large sample confidence intervals found in Problem 7.10.1?

7.11 (Data file: cakes) Refer to Problem 5.8, which uses the model given at (5.12),

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 \quad (7.23)$$

Estimate the optimal (X_1, X_2) combination $(\tilde{X}_1, \tilde{X}_2)$ that maximizes the fitted response and find the standard errors of \tilde{X}_1 and \tilde{X}_2 . (*Hint:* You will need to differentiate (7.23) with respect to both X_1 and X_2 and then find the maximizers as functions of the β s.)

7.12 (Data file: mile) The data file gives the world record times for the one-mile run (Perkiomäki, 1997). For males, the records are for the period

from 1861 to 2003, and for females, for the period 1967–2003. The variables in the file are `Year`, year of the record, `Time`, the record time, in seconds, `Name`, the name of the runner, `Country`, the runner's home country, `Place`, the place where the record was run (missing for many of the early records), and `Gender`, either male or female.

- 7.12.1** Draw a scatterplot of `Time` versus `Year`, using a different symbol for men and women. Comment on the graph.
 - 7.12.2** Fit a regression model with intercepts and slopes for each gender. Provide an interpretation of the slopes.
 - 7.12.3** Find the year in which the female record is expected to be 240 seconds, or 4 minutes. This will require inverting the fitted regression equation. Use the delta method to estimate the standard error of this estimate.
 - 7.12.4** Using the model fit in Problem 7.12.2, estimate the year in which the female record will match the male record, and use the delta method to estimate the standard error of the year in which they will agree. Comment on whether you think using the point at which the fitted regression lines cross is a reasonable estimator of the crossing time.
- 7.13** (Data file: `Transact`) Use the delta method to get a 95% confidence interval for the ratio β_1/β_2 for the transactions data, and compare with the bootstrap interval obtained at the end of Section 7.7.1.
- 7.14 Windmill data** (Data file: `wm1`) These data were discussed in Problem 2.21. Use $B = 999$ replications of the bootstrap to estimate a 95% confidence interval for the long-term average wind speed at the candidate site and compare this with the prediction interval in Problem 2.21.5. See the comment at the end of Problem 2.21.4 to justify using a bootstrap confidence interval for the mean as a prediction interval for the long-term mean.

Transformations

There are exceptional problems for which we know the correct regressors to make $E(Y|X)$ a linear regression mean function. For example, if (Y, X) has a joint normal distribution, then as in Section 4.4, the conditional distribution of $Y|X$ has a linear mean function. Sometimes, the mean function may be determined by a theory, apart from parameter values, as in the strong interaction data in Section 7.1. Often no theory tells us the correct form for the mean function, and any parametric form we use is little more than an approximation that we hope is adequate for the problem at hand. Replacing either the predictors, the response, or both by nonlinear transformations of them is an important tool that the analyst can use to extend the number of problems for which linear regression methodology is appropriate. This brings up two important questions: How do we choose transformations? How do we decide if an approximate model is adequate for the data at hand? We address the first of these questions in this chapter, and the second in Chapter 9.

8.1 TRANSFORMATION BASICS

The most frequent purpose of transformations is to achieve a mean function that is linear in the transformed scale. In problems with only one predictor and a response, the mean function can be visualized in a scatterplot, and we can attempt to select transformations so the resulting scatterplot has an approximate straight-line mean function. With many predictors, selection of transformations can be harder, so we consider the one predictor case first. We seek a transformation so if X is the regressor obtained by transforming the predictor and Y is the transformed response, then the mean function in the transformed scale is

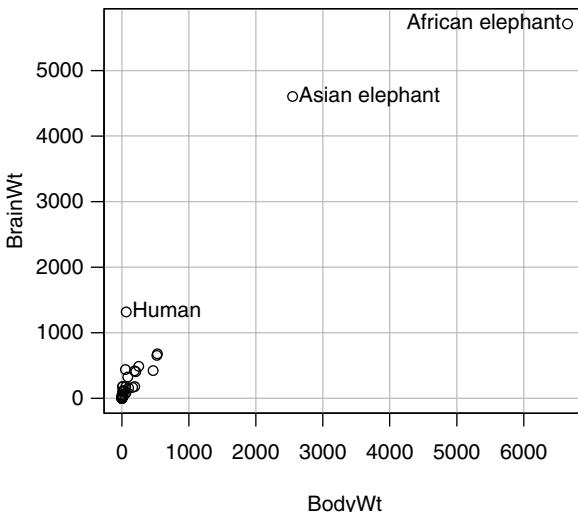


Figure 8.1 Plot of BrainWt versus BodyWt for 62 mammal species.

$$E(Y|X = x) \approx \beta_0 + \beta_1 x$$

where we have used “ \approx ” rather than “ $=$ ” to recognize that this relationship may be an approximation and not exactly true.

Figure 8.1 contains a plot of body weight BodyWt in kilograms and brain weight BrainWt in grams for 62 species of mammals (Allison and Cicchetti, 1976), using the data in the file brains. Apart from the three separated points for humans and two species of elephants, the clumping of points in the lower left of the plot hides any useful visual information about the mean of BrainWt given BodyWt. Little or no evidence for a straight-line mean function is available from this graph. Both variables range over several orders of magnitude from tiny species with body weights of just a few grams to huge animals of over 6600 kg. Transformations can help in this problem.

8.1.1 Power Transformations

A *transformation family* is a collection of transformations that are indexed by one or a few parameters that the analyst can select. The family that is used most often is called the *power family*, defined for a strictly positive variable U by

$$\psi(U, \lambda) = U^\lambda \tag{8.1}$$

This family includes the square root and cube root transformations, $\lambda = 1/2$ or $1/3$, the inverse, $\lambda = -1$, and untransformed, $\lambda = 1$. We will interpret the value of $\lambda = 0$ to be a log transformation. The values of λ used most often in practice

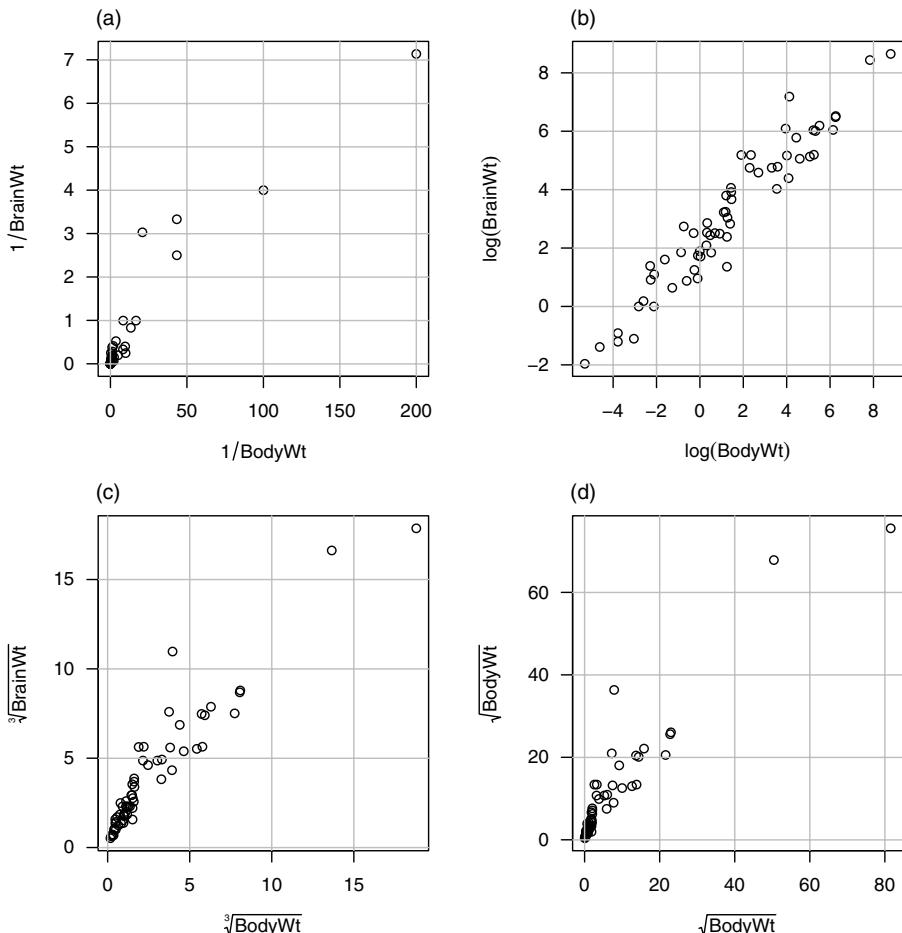


Figure 8.2 Scatterplots for the brain weight data with four possible transformations.

are in the range $[-1, 1]$, or less often in the range $[-2, 2]$. The variable U must be strictly positive for these transformations to be used, but we will have more to say later about transforming variables that may be 0 or negative. We have introduced the ψ -notation¹ because we will later consider other families of transformations, and having this notation will allow more clarity in the discussion.

Figure 8.2 shows plots of $\psi(\text{BrainWt}, \lambda)$ versus $\psi(\text{BodyWt}, \lambda)$ with *the same* λ for both variables, for $\lambda = -1, 0, 1/3, 1/2$. There is no requirement that the same transformation is used for both variables, but it is reasonable here because both are weights. If we allowed each variable to have its own transformation parameter, the visual search for a transformation is harder because

¹ ψ is the Greek letter *psi*.

more possibilities need to be considered. A negative power like $\lambda = -1$ reorders the points, so the isolated point in the upper right of Figure 8.2a is for the smallest animal in the data, but in the remaining plots the largest animals are in the upper right.

The clear choice from the four graphs in Figure 8.2 is to replace the variables by their logarithms. The mean function appears to be a straight line in this scale. As a bonus, the variance function in the log plot appears to be constant because the variation is uniform across the plot.

The use of logarithms for the brain weight data may not be particularly surprising, in light of the following two empirical rules that are often helpful in linear regression modeling:

The log rule If the values of a variable range over more than one order of magnitude and the variable is strictly positive, then replacing the variable by its logarithm is likely to be helpful.

The range rule If the range of a variable is considerably less than one order of magnitude, then any transformation of that variable is unlikely to be helpful.

The log rule is satisfied for both `BodyWt`, with range 0.005 kg to 6654 kg, and for `BrainWt`, with range 0.14 g to 5712 g, so log transformations would have been indicated as a starting point for examining these variables for transformations.

Simple linear regression seems to be appropriate with both variables in log scale. This corresponds to the *physical model*

$$\text{BrainWt} = \alpha \times \text{BodyWt}^{\beta_1} \times \delta \quad (8.2)$$

where δ is a *multiplicative error*. For example, if $\delta = 1.1$ for a particular species, then the `BrainWt` for that species is 1.1 times the expected `BrainWt` for all species with the same `BodyWt`. On taking logarithms and setting $\beta_1 = \log(\alpha)$ and $e = \log(\delta)$,

$$\log(\text{BrainWt}) = \beta_0 + \beta_1 \log(\text{BodyWt}) + e$$

which for `BodyWt` fixed is the simple linear regression model. Scientists who study the relationships between attributes of individuals or species call (8.2) an *allometric* model (Gould, 1966, 1973; Hahn, 1979), and the value of β_1 plays an important role in allometric studies. We emphasize, however, that not all useful transformations will correspond to interpretable physical models.

8.1.2 Transforming One Predictor Variable

Transformations of both the response and the predictor are required to get a linear mean function in the brain weight example. In other problems,

transformation of only one of these variables may be desirable. For selecting a transformation, it is convenient to introduce the family of *scaled power transformations*, defined for strictly positive X by

$$\psi_s(X, \lambda) = \begin{cases} (X^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log(X) & \text{if } \lambda = 0 \end{cases} \quad (8.3)$$

$\psi_s(X, \lambda)$ is continuous as a function of λ . Since $\lim_{\lambda \rightarrow 0} \psi_s(X, \lambda) = \log(X)$, the logarithmic transformation is a member of this family with $\lambda = 0$. Scaled power transformations preserve the direction of association, in the sense that if (X, Y) are positively related, then $(\psi_s(X, \lambda), Y)$ are positively related for all values of λ . With basic power transformations, the direction of association changes when $\lambda < 0$.

If we find an appropriate power to use for a scaled power transformation, we would in practice use the basic power transformation $\psi(X, \lambda)$ in regression modeling, since the two differ only by a scale, location, and possibly a sign change. The scaled transformations are generally used to select a transformation only.

If transforming only the predictor and using a choice from the power family, we begin with the mean function

$$E(Y|X) = \beta_0 + \beta_1 \psi_s(X, \lambda) \quad (8.4)$$

If we know λ , we can fit (8.4) via OLS and get the residual sum of squares, $RSS(\lambda)$. An estimate $\hat{\lambda}$ of λ is the value of λ that minimizes $RSS(\lambda)$. We do not need to know λ very precisely, and selecting λ to minimize $RSS(\lambda)$ from $\lambda \in \{-1, -1/2, 0, 1/3, 1/2, 1\}$ is usually adequate.

As an example, consider the dependence of tree Height in decimeters on Dbh, the diameter of the tree in mm at 137 cm above the ground, for a sample of western cedar trees in 1991 in the Upper Flat Creek stand of the University of Idaho Experimental Forest (courtesy of Andrew Robinson). The data are in the file `ufcwc`. Figure 8.3 is the scatterplot of the data, and on this plot we have superimposed curved lines corresponding to the fit of (8.4) for $\lambda \in \{-1, 0, 1\}$. For these values of λ we get

λ	$RSS(\lambda)$
-1	197,352
0	152,232
1	193,740

The value of $RSS(0)$ is much lower than the RSS for the other two values, in agreement with the visual fit of the line on log scale compared with the other two. By solving a nonlinear least squares problem, we can find the value of λ that minimizes the RSS ; it is given by $\hat{\lambda} = 0.05$. The corresponding fitted line,

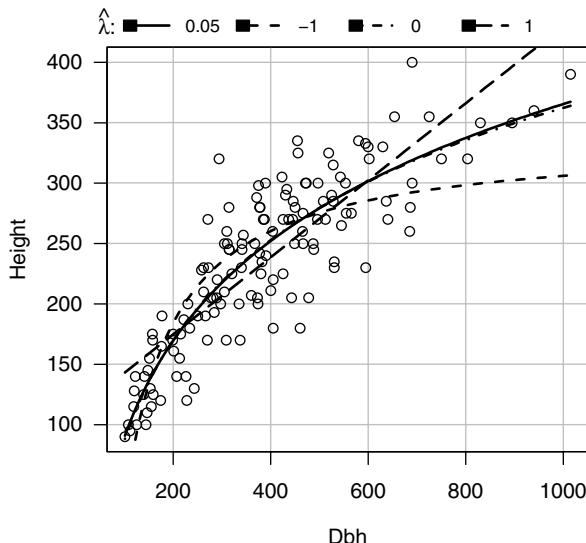


Figure 8.3 Height versus Dbh for the red cedar data from Upper Flat Creek.

shown in Figure 8.3, is essentially identical to the line for the log transformation. The plot of Height versus $\log(\text{Dbh})$ is shown in Figure 8.4.

8.1.3 The Box–Cox Method

Box and Cox (1964) provided a general method for selecting a transformation from a family indexed by a parameter λ . This method is usually applied in the important problem of choosing a response transformation as will be presented in Section 8.3. We introduce the Box–Cox method now because it can also be used to select transformations of many predictors simultaneously, as presented in Section 8.2.2.

We use a slightly more complicated version of the power family called the *modified power family*, defined by Box and Cox (1964) for strictly positive Y to be

$$\begin{aligned}\psi_M(Y, \lambda_y) &= \psi_s(Y, \lambda_y) \times \text{gm}(Y)^{1-\lambda_y} \\ &= \begin{cases} \text{gm}(Y)^{1-\lambda_y} \times (Y^{\lambda_y} - 1)/\lambda_y & \text{if } \lambda_y \neq 0 \\ \text{gm}(Y) \times \log(Y) & \text{if } \lambda_y = 0 \end{cases} \quad (8.5)\end{aligned}$$

where $\text{gm}(Y)$ is the *geometric mean* of the untransformed variable: if the values of Y are y_1, \dots, y_n , the geometric mean of Y is $\text{gm}(Y) = \exp(\sum \log(y_i)/n)$.

Suppose that the mean function

$$E(\psi_M(Y, \lambda_y)|X = \mathbf{x}) = \boldsymbol{\beta}' \mathbf{x} \quad (8.6)$$

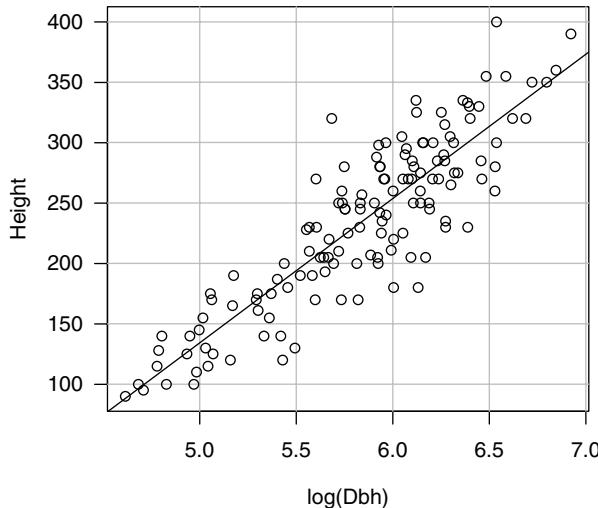


Figure 8.4 The red cedar data from Upper Flat Creek transformed.

holds for some λ_y . If λ_y were known, we could fit the mean function (8.6) using OLS because the transformed response $\psi_M(Y, \lambda_y)$ would then be completely specified. Write the residual sum of squares from this regression as $\text{RSS}(\lambda_y)$. Multiplication of the scaled power transformation by $g_m(Y)^{1-\lambda}$ guarantees that the units of $\psi_M(Y, \lambda_y)$ are the same for all values of λ_y , and so all the $\text{RSS}(\lambda_y)$ are in the same units. We estimate λ_y to be the value of the transformation parameter that minimizes $\text{RSS}(\lambda_y)$. From a practical point of view, we can again select λ_y from $\lambda_y \in \{-1, -1/2, 0, 1/3, 1/2, 1\}$.

The Box–Cox method is not transforming for linearity, but rather it is transforming for *normality*: λ is chosen to make the residuals from the regression of $\psi(Y, \lambda_y)$ on X as close to normally distributed as possible. Hernandez and Johnson (1980) point out that “as close to normal as possible” need not be very close to normal, and so graphical checks are desirable after selecting a transformation. The Box–Cox method will also produce a confidence interval for the transformation parameter; see Appendix A.12.1 for details.

8.2 A GENERAL APPROACH TO TRANSFORMATIONS

The data described in Table 8.1 and given in the data file Highway are taken from an unpublished master’s paper in civil engineering by Carl Hoffstedt. They relate the automobile accident rate in accidents per million vehicle miles to several potential predictors. The data include 39 sections of large highways in the state of Minnesota in 1973. The goal of this analysis is to understand the impact on accidents of the design variables `slim`, `sigs`, and `shld` that are under the control of the highway department. The other variables are thought

Table 8.1 The Highway Accident Data

Variable	Description
rate	1973 accident rate per million vehicle miles
len	Length of the segment in miles
adt	Estimated average daily traffic count in thousands
trucks	Truck volume as a percentage of the total volume
slim	1973 speed limit
shld	Shoulder width in feet of outer shoulder on the roadway
sigs	Number of signalized interchanges per mile in the segment

to be important determinants of accidents but are more or less beyond the control of the highway department and are included to reduce variability due to these uncontrollable factors. We have no particular reason to believe that `rate` will be a linear function of the predictors, or any theoretical reason to prefer any particular form for the mean function.

An important first step in this analysis is to examine the scatterplot matrix of all the predictors and the response, as given in Figure 8.5. Here are some observations about this scatterplot matrix that might help in selecting transformations:

1. The variable `sigs`, the number of traffic lights per mile, is 0 for freeway-type road segments but can be well over 2 for other segments. We can't use power transformations or logs because of the 0 values. A simple expedient with variables that can be 0 is to add a small constant before transforming. The variable `sigs` is a rate per mile, so we add the constant to the number of signals in the segment, and then recompute a rate,

$$\text{sigs1} = \frac{\text{sigs} \times \text{len} + 1}{\text{len}}$$

This variable is always positive and can be transformed using the power family.

2. `adt` and `len` have wide ranges, and logarithms are likely to be appropriate for them.
3. `slim` varies only from 40 mph to 70 mph, with most values in the range from 50 to 60. Transformations are unlikely to be much use here.
4. Each of the predictors seems to be at least modestly associated with `rate`, as the mean function for each of the plots in the top row of Figure 8.5 is not flat.
5. Many of the predictors are also related to each other. In some cases, the mean functions for the plots of predictor versus predictor appear to be linear; in other cases, they are not linear.

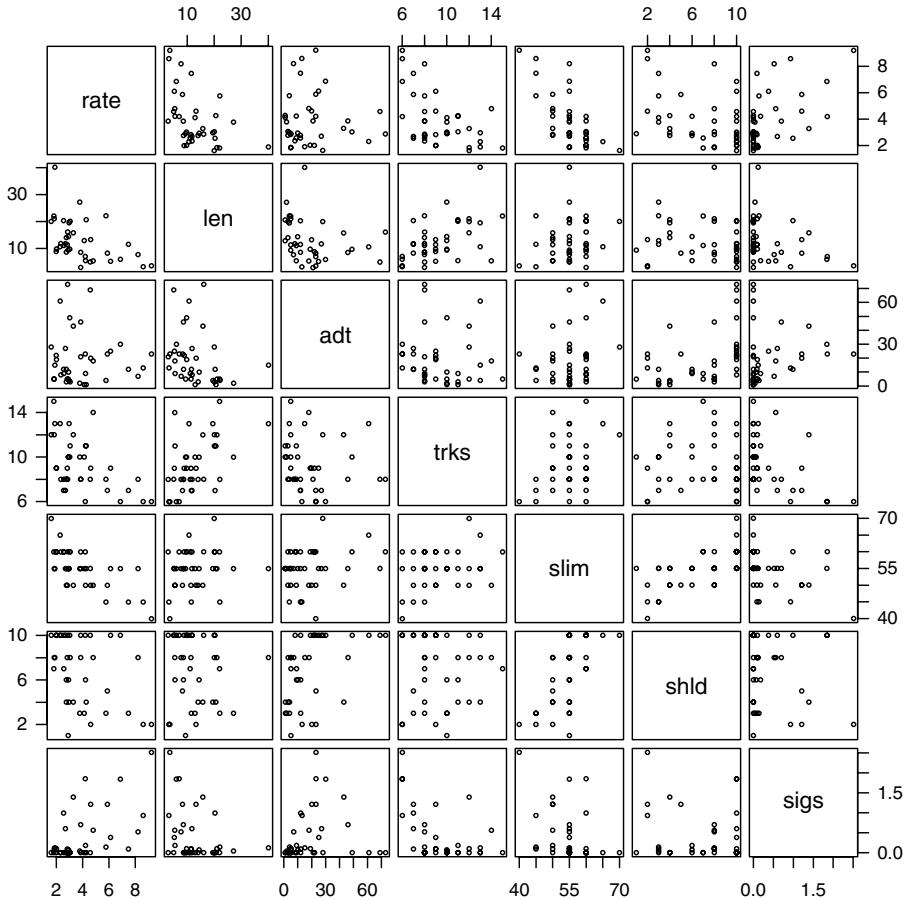


Figure 8.5 The highway accident data, no transformations.

Given these preliminary views of the scatterplot matrix, we now have the daunting task of finding good transformations to use. This raises immediate questions: What are the goals in selecting transformations? How can we decide if we have made a good choice?

The overall goal is to find transformations in which multiple linear regression matches the data to a reasonable approximation. The connection between this goal and choosing transformations that make the 2D plots of predictors have linear mean functions is not entirely obvious. Important work by Brillinger (1983) and Li and Duan (1989) provides a theoretical connection. Suppose we have a response variable Y and a set of regressors X derived from the predictors, and suppose it were true that

$$E(Y|X = \mathbf{x}) = g(\boldsymbol{\beta}'\mathbf{x}) \quad (8.7)$$

for some *completely unknown and unspecified function* g . According to this, the mean of Y depends on X only through a linear combination of the regressors in X , and if we could draw a graph of Y versus $\beta'x$, this graph would have g as its mean function. We could then either estimate g , or we could transform Y to make the mean function linear. All this depends on estimating β without specifying anything about g . Are there conditions under which the OLS regression of Y on X can help us learn about β ?

8.2.1 The 1D Estimation Result and Linearly Related Regressors

We will say that X is a set of *linearly related regressors* if the graph of any linear combination of the regressors in X versus any other linear combination of the regressions in X has a straight-line mean function. The condition that all the graphs in a scatterplot matrix of X have straight-line mean functions is weaker than the condition for linearly related regressors, but it is a reasonable condition that we can check in practice. Requiring that X is multivariate normal is much stronger than linearly related regressors. Hall and Li (1993) show that the condition for linearly related regressors holds approximately as the number of predictors grows large, so in very large problems, transformation becomes less important because the assumption of linearly related regressors will hold approximately without any transformations.

Given that linearly related regressors hold at least to a reasonable approximation, and assuming that $E(Y|X = \mathbf{x}) = g(\beta'\mathbf{x})$, then the OLS estimate $\hat{\beta}$ is a consistent estimate of $c\beta$ for some constant c that is usually nonzero (Cook, 1998; Li and Duan, 1989). Given this theorem, a useful general procedure for applying multiple linear regression analysis is

1. Transform predictors to get regressors for which the condition for linearly related regressors holds, at least approximately. The regressors in X may include dummy variables that represent factors, which should not be transformed, as well as transformations of continuous predictors.
2. We can estimate g from the 2D scatterplot of Y versus $\hat{\beta}'\mathbf{x}$, where $\hat{\beta}$ is the OLS estimator from the regression of Y on X . Almost equivalently, we can estimate a transformation of Y either from the inverse plot of $\hat{\beta}'\mathbf{x}$ versus Y or from using the Box–Cox method.

This is a general and powerful approach to building regression models that match data well, based on the assumption that (8.7) is appropriate for the data. We have already seen mean functions in Chapter 5 for which (8.7) does not hold because of the inclusion of interaction regressors, and so transformations chosen using the methods discussed here may not provide a comprehensive mean function when interactions are present.

The Li–Duan theorem is actually much more general and has been extended to problems with interactions present and to many other estimation methods

beyond OLS. See Cook and Weisberg (1999a, chapters 18–20) and, at a higher mathematical level, Cook (1998).

8.2.2 Automatic Choice of Transformation of Predictors

Using the results of Section 8.2.1, we seek to transform the predictors so that all plots of one predictor versus another have a linear mean function, or at least have mean functions that are not too curved. Velilla (1993) proposed a multivariate extension of the Box–Cox method to select transformations to linearity, and this method can often suggest a very good starting point for selecting transformations of predictors. Starting with k untransformed strictly positive predictors $X = (X_1, \dots, X_k)$, we will apply a modified power transformation to each X_j , and so there will be k transformation parameters collected into $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_k)'$. We will write $\psi_M(X, \lambda)$ to be the set of variables

$$\psi_M(X, \lambda) = (\psi_M(X_1, \lambda_1), \dots, \psi_M(X_k, \lambda_k))'$$

Let $\mathbf{V}(\lambda)$ be the sample covariance matrix of the transformed data $\psi_M(X, \lambda)$. The value $\hat{\lambda}$ is selected as the value of λ that minimizes the logarithm of the determinant of $\mathbf{V}(\lambda)$. If special purpose software is not available, this minimization can be carried using a general function minimizer included in many high-level languages, such as R, Maple, Mathematica, or even Microsoft Excel. The minimizers generally require only specification of the function to be minimized and a set of starting values for the algorithm. The starting values can be taken to be $\lambda = \mathbf{0}$, $\lambda = \mathbf{1}$, or some other appropriate vector of zeros and ones.

Returning to the highway data, we eliminate `slim` as a variable to be transformed because its range is too narrow. For the remaining predictors, we get the summary of transformations using the multivariate Box–Cox method in Table 8.2. The table gives the value of $\hat{\lambda}$ in the column marked “Est. power.” The standard errors are computed as outlined in Appendix A.12.2. The standard errors can be treated like standard errors of regression coefficients. The next two columns provide a 95% confidence interval for each λ . The estimated powers for `len`, `adt`, `trks`, and `sigs1` are close to 0, and the power for `shld` does not appear to be different from 1. Table 8.2b includes three *likelihood ratio tests*. The first of these tests is that all powers are 0; this is firmly rejected as the approximate test, with an approximate $\chi^2(5)$ distribution,² has a tiny significance level. Similarly, the test for no transformation ($\lambda = \mathbf{1}$) is firmly rejected. The test that the first three variables should be in log scale, the next untransformed, and the last in log scale, has a p -value 0.29 and suggests using these simple transformations in further analysis with these data. The predictors in transformed scale, along with the response, are shown in Figure 8.6. All these 2D plots have a linear mean function, or at least

²The df are the number estimated elements in λ .

Table 8.2 Power Transformations to Normality for the Highway Data

(a) Estimated Powers				
	Est.Power	Std.Err.	Lower	Conf. Int.
len	0.144	0.213	-0.273	0.561
adt	0.051	0.121	-0.185	0.287
trks	-0.703	0.618	-1.913	0.508
shld	1.346	0.363	0.634	2.057
sigs1	-0.241	0.150	-0.534	0.052

(b) Test Statistics			
	LRT	df	p-Value
LR test, lambda = (0 0 0 0 0)	23.32	5	0.00
LR test, lambda = (1 1 1 1 1)	132.86	5	0.00
LR test, lambda = (0 0 0 1 0)	6.09	5	0.30

are not strongly nonlinear. They provide a good place to start regression modeling.

8.3 TRANSFORMING THE RESPONSE

Once the predictors are transformed, we can turn our attention to transforming the response.

Cook and Weisberg (1994) suggested that the methodology of Section 8.1.2 for transforming a single predictor can be adapted to visualizing the need to transform a response, and to select an appropriate transformation from the graph. Start with

$$E(\hat{Y}|Y) = \alpha_0 + \alpha_1 \psi_s(Y, \lambda_y)$$

where \hat{Y} are the fitted values from the regression of the untransformed Y on the appropriately transformed regressors X . This implies examining a graph with \hat{Y} on the vertical axis and Y on the horizontal axis, called an *inverse fitted value plot*. If the regressors are approximately linearly related, the transformation with parameter λ_y can be selected either to minimize the RSS or visually.

Figure 8.7a is the inverse fitted value plot for the highway data using the transformed regressors determined in the last section. We can use the method of Section 8.1.2 to select a transformation for `rate`. The best-fitting curve with $\hat{\lambda} = 0.19$ and the log curve for $\lambda = 0$ are nearly identical, and so a log transformation of `rate` is suggested.

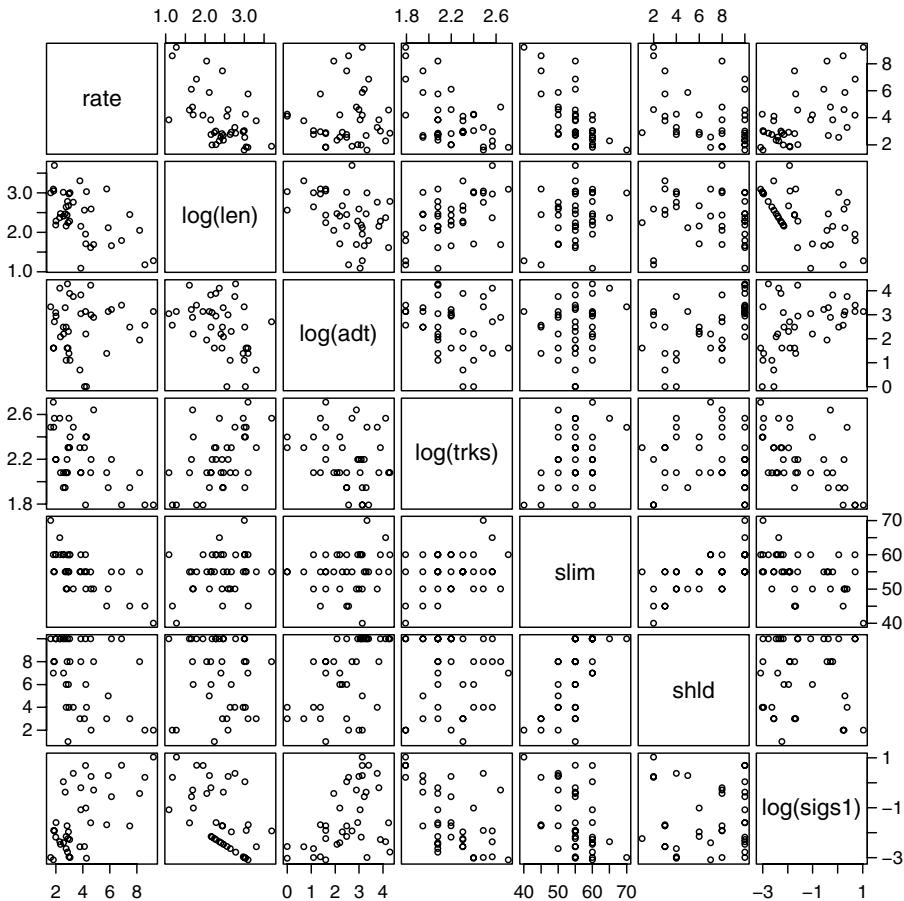


Figure 8.6 Transformed predictors for the highway data.

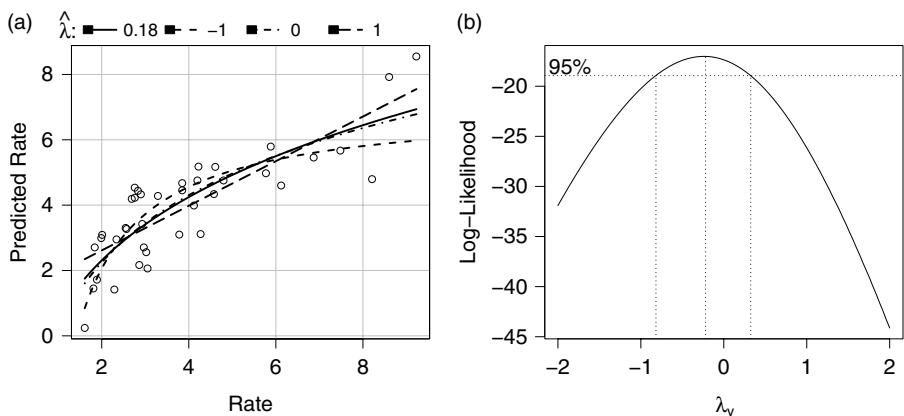


Figure 8.7 (a) Inverse fitted value plot for the highway data. (b) Profile log-likelihood for the Box-Cox method.

The Box–Cox method is the most commonly used procedure for finding a transformation of the response. This method is often summarized by a graph with λ_y on the horizontal axis and either $\text{RSS}(\lambda_y)$ or better yet $\log[L(\lambda_y)] = -(n/2) \log[\text{RSS}(\lambda_y)/n]$ on the vertical axis. With this latter choice, the estimate $\hat{\lambda}_y$ is the point that maximizes the curve, and a 95% confidence interval for the estimate is given by the set of all λ_y with $\log[L(\hat{\lambda}_y)] - \log[L(\lambda_y)] < 1.92$; see Appendix A.12.1. This graph for the highway data is shown in Figure 8.7b, with $\hat{\lambda} \approx -0.2$ and the confidence interval of about -0.8 to $+0.3$. The log transformation is in the confidence interval, agreeing with the inverse fitted value plot.

The two transformation methods for the response seem to agree for the highway data, but there is no theoretical reason why they need to give the same transformation. The following path is recommended for selecting a response transformation:

1. With approximately linearly related regressors, draw the inverse response plot of \hat{y} versus the response. If this plot shows a clear nonlinear trend, then the response should be transformed to match the nonlinear trend. There is no reason why only power transformations should be considered. For example, the transformation could be selected using a smoother. If there is no clear nonlinear trend, transformation of the response is unlikely to be helpful.
2. The Box–Cox procedure can be used to select a transformation to normality. It requires the use of a transformation family.

For the highway data, we now have a reasonable starting point for regression, with several of the predictors and the response all transformed to log scale. We will continue with this example in Chapter 10.

8.4 TRANSFORMATIONS OF NONPOSITIVE VARIABLES

Several transformation families for a variable U that includes negative values have been suggested. The central idea is to use the methods discussed in this chapter for selecting a transformation from a family but to use a family that permits U to be nonpositive. One possibility is to consider transformations of the form $(U + \gamma)^{\lambda}$, where γ is sufficiently large to ensure that $U + \gamma$ is strictly positive. We used a variant of this method with the variable `sigs` in the highway data. In principle, (γ, λ) could be estimated simultaneously, although in practice, estimates of γ are highly variable and unreliable. Alternatively, Yeo and Johnson (2000) proposed a family of transformations that can be used without restrictions on U that have many of the good properties of the Box–Cox power family. These transformations are defined by

$$\psi_{YJ}(U, \lambda) = \begin{cases} \psi_M(U + 1, \lambda) & \text{if } U \geq 0 \\ -\psi_M(-U + 1, 2 - \lambda) & \text{if } U < 0 \end{cases} \quad (8.8)$$

If U is strictly positive, then the Yeo–Johnson transformation is the same as the Box–Cox power transformation of $(U + 1)$. If U is strictly negative, then the Yeo–Johnson transformation is the Box–Cox power transformation of $(-U + 1)$, but with power $2 - \lambda$. With both negative and positive values, the transformation is a mixture of these two, so different powers are used for positive and negative values. In this latter case, interpretation of the transformation parameter is difficult, as it has a different meaning for $U \geq 0$ and for $U < 0$.

8.5 ADDITIVE MODELS

Additive models provide an alternative to the methods for selecting transformations for predictors. Suppose we have a regression problem with regressors for factors and other variables that do not need transformation given in a vector \mathbf{z} , and additional predictors that may need to be transformed in $\mathbf{x}' = (x_1, \dots, x_q)$. We consider the mean function

$$E(Y|\mathbf{z}, \mathbf{x}) = \beta' \mathbf{z} + \sum g_j(x_j) \quad (8.9)$$

where $g_j(x_j)$ is some unknown function that is essentially a transformation of x_j . Additive models proceed by estimating the functions g_j . Regression parameters for the x_j do not appear in (8.9) because they are absorbed into the estimates of the g_j (this takes some getting used to). Methodology that uses splines to estimate the g_j is discussed in a fine book by Wood (2006), with accompanying software available in R in the `mgcv` package.

8.6 PROBLEMS

8.1 (Data file: `baesk1`) These data were collected in a study of the effect of dissolved sulfur on the surface tension of liquid copper (Baes and Kellogg, 1953). The predictor `Sulfur` is the weight percent sulfur, and the response is `Tension`, the decrease in surface tension in dynes per centimeter. Two replicate observations were taken at each value of `Sulfur`. These data were previously discussed by Sclove (1968).

8.1.1 Draw the plot of `Tension` versus `Sulfur` to verify that a transformation is required to achieve a straight-line mean function.

8.1.2 Set $\lambda = -1$, and fit the mean function

$$E(\text{Tension}|\text{Sulfur}) = \beta_0 + \beta_1 \text{Sulfur}^\lambda$$

using `OLS`; that is, fit the `OLS` regression with `Tension` as the response and `1/Sulfur` as the regressor. Add a line for the fitted values from this fit to the plot you drew in Problem 8.1.2. If you do

not have a program that will do this automatically, you can let `new` be a vector of 100 equally spaced values between the minimum value of `Sulfur` and its maximum value. Compute the fitted values $\text{Fit.new} = \hat{\beta}_0 + \hat{\beta}_1 \text{new}^\lambda$, and a line joining these points to your graph. Repeat for $\lambda = 0, 1$, and so in the end you will have three lines on your plot. Which of these three choices of λ gives fitted values that match the data most closely?

- 8.1.3** Replace `Sulfur` by its logarithm, and consider transforming the response `Tension`. To do this, draw the inverse fitted value plot with the fitted values from the regression `Tension ~ log(Sulfur)` on the vertical axis and `Tension` on the horizontal axis. Repeat the methodology of Problem 8.1.2 to decide if further transformation of the response will be helpful.
- 8.2** (Data file: `stopping`) We reconsider the stopping distance data used in Problem 7.6.
- 8.2.1** Using `Speed` as the only regressor, find an appropriate transformation for `Distance` that can linearize this regression.
- 8.2.2** Using `Distance` as the response, transform the predictor `Speed` using a power transformation with each $\lambda \in \{-1, 0, 1\}$, and show that none of these transformations is adequate.
- 8.2.3** Show that using $\lambda = 2$ does match the data well. This suggests using a quadratic polynomial for regressors, including both `Speed` and `Speed2`.
- 8.2.4** Hald (1960) suggested on the basis of a theoretical argument using a quadratic mean function for `Distance` given `Speed`, with $\text{Var}(\text{Distance}|\text{Speed}) = \sigma^2 \text{Speed}^2$. Draw the plot of `Distance` versus `Speed`, and add a line on the plot of the fitted curve from Hald's model. Then obtain the fitted values from the fit of the transformed `Distance` on `Speed`, using the transformation you found in Problem 8.2.1. Transform these fitted values to the `Distance` scale (for example, if you fit the regression `sqrt(Distance) ~ Speed`, then the fitted values would be in square-root scale and you would square them to get the original `Distance` scale). Add to your plot the line corresponding to these transformed fitted values. Compare the fit of the two models.
- 8.3** (Data file: `water`) A major source of water in Southern California is the Owens Valley. This water supply is in turn replenished by spring runoff from the Sierra Nevada mountains. If runoff could be predicted, engineers, planners, and policy makers could do their jobs more efficiently. The data file contains snowfall depth measurements over 43 years taken at six sites in the mountains, in inches, and stream runoff volume at a site near Bishop, California. The three sites with names starting with "O" are fairly close to

each other, and the three sites starting with “A” are also fairly close to each other. Year is also given in the data file, but should not be used as a predictor.

- 8.3.1** Construct the scatterplot matrix of the data, and provide general comments about relationships among the variables.
- 8.3.2** Using the methodology for automatic choice of transformations outlined in Section 8.2.2, find transformations to make the transformed predictors as close to linearly related as possible. Obtain a test of the hypothesis that all $\lambda_j = 0$ against a general alternative, and summarize your results. Do the transformations you found appear to achieve linearity? How do you know?
- 8.3.3** Given log transformations of the predictors, show that a log transformation of the response is reasonable.
- 8.3.4** Consider the multiple linear regression model with mean function given by

$$\log(\text{BSAAM}) \sim \log(\text{APMAM}) + \log(\text{APSAB}) + \log(\text{APSLAKE}) + \\ \log(\text{OPBPC}) + \log(\text{OPRC}) + \log(\text{OPSLAKE})$$

with constant variance function. Estimate the regression coefficients using OLS. You will find that two of the estimates are negative; Which are they? Does a negative coefficient make any sense? Why are the coefficients negative?

- 8.3.5** Test the hypothesis that the coefficients for the three “O” log predictors are equal against the alternative that they are not all equal. Repeat for the “A” predictors. Explain why these might be interesting hypotheses. (*Hint:* The geometric mean of the regressors OPBPC, OPRC, OPSLAKE is equal to $\exp[(\log(\text{OPBPC}) + \log(\text{OPRC}) + \log(\text{OPSLAKE}))/3]$, and so the sum $[\log(\text{OPBPC}) + \log(\text{OPRC}) + \log(\text{OPSLAKE})]$ is proportional to the logarithm of the geometric mean of these regressors. If the three coefficients are equal, then we are essentially replacing the three predictors by one regressor equivalent to the logarithm of their geometric mean.)

- 8.4** (Data file: salarygov) In Problem 5.9 we modeled MaxSalary as the response and modeled the predictor Score using regressors from a B-spline basis to account for curvature.

- 8.4.1** As an alternative, show that the regression model $\log(\text{MaxSalary}) \sim \text{Score}$ has an approximately linear mean function with approximately constant variance. Most studies of salary and income are done in log scale.
- 8.4.2** As in Problem 5.9.3, define a factor with two levels for male- and female-dominated job classes, and fit appropriate models to explore differences between the two classes of jobs.

8.5 World cities (Data file: BigMac2003) The Union Bank of Switzerland publishes a report entitled “Prices and Earnings Around the Globe” (2009). The data described in Table 8.3 are taken from their 2003 version for 69 world cities.

- 8.5.1** Draw the scatterplot with BigMac on the vertical axis and FoodIndex on the horizontal axis. Provide a qualitative description of this graph. Two of the cities had very high cost for BigMac. What are they?
- 8.5.2** Use the Box–Cox method and, if available, an inverse response plot to find a transformation of BigMac so that the resulting scatterplot has a linear mean function.
- 8.5.3** An advantage of the inverse response plot is that we can find individual points that may be important in fitting curves to the plot. These influential points will often be at the extreme left or at the extreme right of the plot. Removing these points could change the fitted curve substantially. Less often, points that are separated vertically from the fitted line can be influential.
The two cities at the far right of the inverse response plot are candidates for influential points. To verify this, refit the transformation methods without these two points and summarize the changes, if any, in choice of transformation.
- 8.5.4** Draw the scatterplot matrix of the three variables (BigMac, Rice, Bread), and use the multivariate Box–Cox procedure to decide on normalizing transformations. Test the null hypothesis that $\lambda = (1, 1, 1)'$ against a general alternative. Does deleting Karachi and Nairobi change your conclusions?
- 8.5.5** Set up the regression using the four regressors, $\log(\text{Bread})$, $\log(\text{Bus})$, $\log(\text{TeachGI})$, and $\text{Apt}^{0.33}$, and with response BigMac. Draw the

Table 8.3 Global Price Comparison Data

Variable	Description
BigMac	Minutes of labor to buy a BigMac hamburger based on a typical wage averaged over 13 occupations
Bread	Minutes of labor to buy 1 kg bread
Rice	Minutes of labor to buy 1 kg of rice
Bus	Lowest cost of 10 km public transit
FoodIndex	Food price index, Zurich = 100
TeachGI	Primary teacher’s gross annual salary, thousands of U.S. dollars
TeachNI	Primary teacher’s net annual salary, thousands of U.S. dollars
TaxRate	$100 \times (\text{TeachGI} - \text{TeachNI})/\text{TeachGI}$. In some places, this is negative, suggesting a government subsidy rather than tax
TeachHours	Teacher’s hours per week of work
Apt	Monthly rent in U.S. dollars of a typical three-room apartment

inverse fitted value plot of \hat{y} versus BigMac. Estimate the best power transformation. Check on the adequacy of your estimate by refitting the regression model with the transformed response and drawing the inverse response plot again. If transformation was successful, this second inverse fitted value plot should have a linear mean function.

8.6 (Data file: `W001`) These data were introduced in Section 5.2. For this problem, we will start with `cycles`, rather than its logarithm, as the response. Remember that you may need to declare `len`, `amp`, and `load` as factors.

- 8.6.1** Draw the scatterplot matrix for these data and summarize the information in this plot. (*Warning:* The predictors are factors, not continuous variables, so the plotting program might label the levels as 1, 2, and 3, rather than the actual numeric value of the variable.)
- 8.6.2** View all three predictors as factors with three levels, and *without transforming* `cycles`, fit the second-order mean function with regressors for all main effects and all two-factor interactions. Summarize results of the `amp` by `load` interaction with an effects plot.
- 8.6.3** Fit the first-order mean function consisting only of the main effects. From Problem 8.6.2, this mean function is not adequate for these data based on using `cycles` as the response because the tests for each of the two-factor interactions indicate that these are likely to be nonzero. Use the Box–Cox method to select a transformation for `cycles` based on the first-order mean function.
- 8.6.4** In the transformed scale, fit both the first-order model and the second-order model, and compute an F -test comparing these two models. This is a nonstandard test because it is simultaneously testing all interactions to be equal to zero. Then provide an effects plot for the `len` by `amp` interaction. This will of course be three parallel lines. Then redraw this effects plot with `cycles` rather than $\log(\text{cycles})$ on the horizontal axis, and compare with the effects plot you drew in Problem 8.6.2.

8.7 (Data file: `fuel2001`) Justify transforming Miles in the fuel data.

Regression Diagnostics

Graphs so far have mostly been used to help us decide what to do before fitting a regression model. *Regression diagnostics* are used *after* fitting to check if a fitted mean function and assumptions are consistent with observed data. The basic statistics here are the residuals or possibly rescaled residuals. If the fitted model does not give a set of residuals that appear to be reasonable, then some aspect of the model, either the assumed mean function or assumptions concerning the variance function, may be called into doubt. A related issue is the importance of each case on estimation and other aspects of the analysis. In some data sets, the observed statistics may change in important ways if a few cases are deleted from the data. Such cases are called *influential*, and we shall learn to detect such cases. We will be led to study and use two relatively unfamiliar diagnostic statistics, called *distance measures* and *leverage values*. We concentrate on graphical diagnostics but include numerical quantities that can aid in interpretation of the graphs.

9.1 THE RESIDUALS

We begin by deriving the properties of residuals using the matrix notation outlined in Chapter 3. The basic multiple linear regression model is given by

$$E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \quad \text{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2 \mathbf{I} \quad (9.1)$$

where \mathbf{X} is a known matrix with n rows and p' columns, including a column of ones for the intercept if the intercept is included in the mean function. We will further assume that we have selected a parameterization for the mean function so that \mathbf{X} has full column rank, meaning that the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ exists; as we

have seen previously, this is not an important limitation on regression models because we can always delete regressors from the mean function, or equivalently delete columns from \mathbf{X} , until we have full rank. The $p' \times 1$ vector $\boldsymbol{\beta}$ is the unknown parameter vector.

In fitting model (9.1), we estimate $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, and the fitted values $\hat{\mathbf{Y}}$ corresponding to the observed values \mathbf{Y} are then given by

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{H}\mathbf{Y}\end{aligned}\tag{9.2}$$

where \mathbf{H} is the $n \times n$ matrix defined by

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tag{9.3}$$

\mathbf{H} is called the *hat matrix* because it transforms the vector of observed responses \mathbf{Y} into the vector of fitted responses $\hat{\mathbf{Y}}$. The vector of residuals $\hat{\mathbf{e}}$ is defined by

$$\begin{aligned}\hat{\mathbf{e}} &= \mathbf{Y} - \hat{\mathbf{Y}} \\ &= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{Y}\end{aligned}\tag{9.4}$$

9.1.1 Difference between $\hat{\mathbf{e}}$ and \mathbf{e}

In this book the vector of errors \mathbf{e} has been defined implicitly by

$$\begin{aligned}\mathbf{e} &= \mathbf{Y} - E(\mathbf{Y}|\mathbf{X}) \\ &= \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\end{aligned}\tag{9.5}$$

The errors \mathbf{e} are unobservable random variables, with $E(\mathbf{e}|\mathbf{X}) = \mathbf{0}$ and $\text{Var}(\mathbf{e}|\mathbf{X}) = \sigma^2\mathbf{I}$. The residuals $\hat{\mathbf{e}}$ are computed quantities that can be graphed or otherwise studied. Their mean and variance, using (9.4) and Appendix A.7, are

$$\begin{aligned}E(\hat{\mathbf{e}}|\mathbf{X}) &= \mathbf{0} \\ \text{Var}(\hat{\mathbf{e}}|\mathbf{X}) &= \sigma^2(\mathbf{I} - \mathbf{H})\end{aligned}\tag{9.6}$$

Like the errors, each of the residuals has zero mean. Unlike the errors, each residual may have a different variance, and in general, the residuals are correlated. If the errors are normally distributed, so are the residuals. From (9.4),

if the errors are not normally distributed, then the residuals may still be nearly normal because sums of nonnormal variables are approximately normal. If the intercept is included in the model, then $\sum \hat{e}_i = 0$. In scalar form, the variance of the i th residual is

$$\text{Var}(\hat{e}_i) = \sigma^2(1 - h_{ii}) \quad (9.7)$$

where h_{ii} is the i th diagonal element of \mathbf{H} . Diagnostic procedures are based on the residuals which we would like to assume behave as the unobservable errors would. The usefulness of this assumption depends on the hat matrix, since it is \mathbf{H} that relates \mathbf{e} to $\hat{\mathbf{e}}$ and also gives the variances and covariances of the residuals.

9.1.2 The Hat Matrix

\mathbf{H} is $n \times n$ and symmetric with many special properties that are easy to verify directly from (9.3). Multiplying \mathbf{X} on the left by \mathbf{H} leaves \mathbf{X} unchanged, $\mathbf{H}\mathbf{X} = \mathbf{X}$. Similarly, $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$. The property $\mathbf{HH} = \mathbf{H}^2 = \mathbf{H}$ also shows that $\mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{0}$, so the covariance between the fitted values \mathbf{HY} and residuals $(\mathbf{I} - \mathbf{H})\mathbf{Y}$ is

$$\begin{aligned} \text{Cov}(\hat{\mathbf{Y}}, \hat{\mathbf{e}}|\mathbf{X}) &= \text{Cov}(\mathbf{HY}, (\mathbf{I} - \mathbf{H})\mathbf{Y}|\mathbf{X}) \\ &= \sigma^2 \mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{0} \end{aligned}$$

Another name for \mathbf{H} is the *orthogonal projection* on the column space of \mathbf{X} . The elements of \mathbf{H} , the h_{ij} , are given by

$$h_{ij} = \mathbf{x}'_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_j = \mathbf{x}'_j (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i = h_{ji} \quad (9.8)$$

Many helpful relationships can be found between the h_{ij} . For example,

$$\sum_{i=1}^n h_{ii} = p' \quad (9.9)$$

and, if the mean function includes an intercept,

$$\sum_{i=1}^n h_{ij} = \sum_{j=1}^n h_{ij} = 1 \quad (9.10)$$

Each diagonal element h_{ii} is bounded below by $1/n$ and above by $1/r$, if r is the number of rows of \mathbf{X} that are identical to \mathbf{x}_i .

As can be seen from (9.7), cases with large values of h_{ii} will have small values for $\text{Var}(\hat{e}_i|\mathbf{X})$; as h_{ii} gets closer to 1, this variance will approach 0. For such a case, no matter what value of y_i is observed for the i th case, we are

nearly certain to get a residual near 0. Hoaglin and Welsch (1978) pointed this out using a scalar version of (9.2),

$$\hat{y}_i = \sum_{j=1}^n h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j \quad (9.11)$$

In combination with (9.10), Equation (9.11) shows that as h_{ii} approaches 1, \hat{y}_i gets closer to y_i . For this reason, they called h_{ii} the *leverage* of the i th case.

Cases with large values of h_{ii} will have unusual values for \mathbf{x}_i . Assuming that the intercept is in the mean function, and using the notation of the deviations from the average cross-products matrix discussed in Chapter 3, h_{ii} can be written as

$$h_{ii} = \frac{1}{n} + (\mathbf{x}_i^* - \bar{\mathbf{x}})' (\mathcal{X}' \mathcal{X})^{-1} (\mathbf{x}_i^* - \bar{\mathbf{x}}) \quad (9.12)$$

where $\mathbf{x}_i' = (1, \mathbf{x}_i^{**})$, and $\bar{\mathbf{x}}$ is the mean of the \mathbf{x}_i^* . The second term on the right-hand side of (9.12) is the equation of an ellipsoid centered at $\bar{\mathbf{x}}$.

For example, consider again the United Nations data, Section 3.1. The plot of $\log(\text{ppgdp})$ versus pctUrban is given in the scatterplot in Figure 9.1. The ellipses drawn on graph correspond to elliptical contours of constant h_{ii} for $h_{ii} = 0.01, 0.03, 0.05$, and 0.07 . Any point that falls exactly on the outer contour would have $h_{ii} = 0.07$, while points on the innermost contour have $h_{ii} = 0.01$. Points near the long or major axis of the ellipsoid need to be much farther

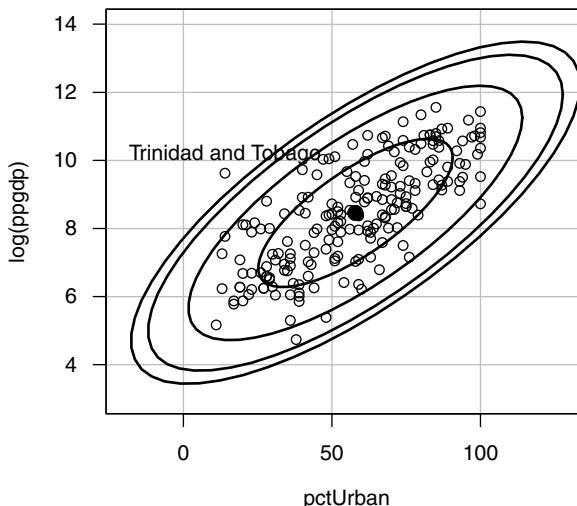


Figure 9.1 Contours of constant leverage in two dimensions.

away from $\bar{\mathbf{x}}$, in the usual Euclidean distance sense, than do points closer to the minor axis, to have the same values for h_{ii} .¹

In the example, the localities with the highest level of urbanization, which are Anguilla, Bermuda, Cayman Islands, Hong Kong, Macao, Nauru, and Singapore, all with 100% urbanization, do not have particularly high leverage, as all the points for these places are between the contour for $h_{ii} = 0.02$ and 0.04. None of the h_{ii} is very large, with the largest value for the marked point for Trinidad and Tobago, which has relatively high income for relatively low urbanization. High leverage points with values close to one can occur, and identifying these cases is very useful in understanding a regression problem.

9.1.3 Residuals and the Hat Matrix with Weights

When $\text{Var}(\mathbf{e}|\mathbf{X}) = \sigma^2 \mathbf{W}^{-1}$ with \mathbf{W} a known diagonal matrix of positive weights as in Section 7.1, all the results so far in this section require some modification. A useful version of the hat matrix is given by

$$\mathbf{H} = \mathbf{W}^{1/2} \mathbf{X} (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}^{1/2} \quad (9.13)$$

and the leverages are the diagonal elements of this matrix. The fitted values are given as usual by $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, where now $\hat{\boldsymbol{\beta}}$ is the wls estimator.

The definition of the residuals is a little trickier. The “obvious” definition of a residual is, in scalar version, $y_i - \hat{\boldsymbol{\beta}}' \mathbf{x}_i$, but this choice has important deficiencies. First, the sum of squares of these residuals will *not* equal the residual sum of squares because the weights are ignored. Second, the variance of the i th residual will depend on the weight of case i .

Both of these problems can be solved by defining residuals for weighted least squares for $i = 1, \dots, n$ by

$$\hat{e}_i = \sqrt{w_i} (y_i - \hat{\boldsymbol{\beta}}' \mathbf{x}_i) \quad (9.14)$$

The sum of squares of these residuals is the residual sum of squares. The variance of these residuals does not depend on the weight. When all the weights are equal to 1, (9.14) reduces to (9.4). In drawing graphs and other diagnostic procedures discussed in this book, (9.14) should be used to define residuals. Some computer packages use the unweighted residuals rather than (9.14) by default. The residuals defined by (9.14) are generally called *Pearson residuals*. In this book \hat{e} and $\hat{\mathbf{e}}$ always refer to the residuals defined by (9.14).

¹The regressor `pctUrban` is a percentage between 0 and 100. Contours of constant leverage corresponding to `pctUrban < 0` or `pctUrban > 100` are shown to give the shape of the contours, even though in this particular problem points could not occur in this region.

9.1.4 Residual Plots When the Model Is Correct

Residuals are generally used in scatterplots of the residuals \hat{e} against a regressor or linear combination of regressors in the mean function that we will call U . The key features of these residual plots when the correct model is fit are as follows:

1. The mean function is $E(\hat{e}|U) = \mathbf{0}$. This means that the scatterplot of residuals on the vertical axis versus *any linear combination of the regressors* should have a constant mean function equal to 0. When the model is correct, residual plots should look like null plots.
2. Since $\text{Var}(\hat{e}_i|U) = \sigma^2(1 - h_{ii})$ even if the fitted model is correct, the variance function is not quite constant. The variability will be smaller for high-leverage cases with h_{ii} close to 1.
3. The residuals are correlated, but this correlation is generally unimportant and not visible in residual plots.

9.1.5 The Residuals When the Model Is Not Correct

If the fitted model is based on incorrect assumptions there is a U for which the plot of residuals versus U is not a null plot. Figure 9.2 shows several generic residual plots for a simple linear regression problem. Figure 9.2a is a null plot that indicates no problems with the fitted model. If obtained from simple regression, Figure 9.2b–d would suggest nonconstant variance as a function of the quantity plotted on the horizontal axis. The curvature apparent in

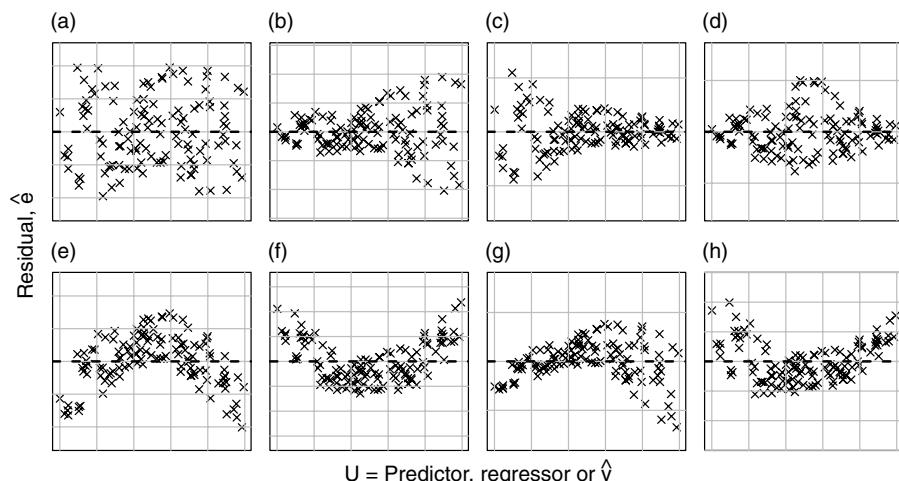


Figure 9.2 Residual plots: (a) null plot; (b) right-opening megaphone; (c) left-opening megaphone; (d) double outward bow; (e) and (f) nonlinearity; (g) and (h) combinations of nonlinearity and nonconstant variance function.

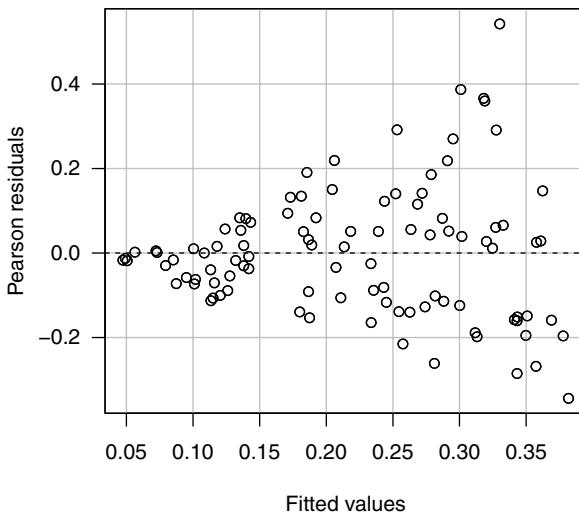


Figure 9.3 Residual plot for the `caution` data.

Figure 9.2e–h suggests an incorrectly specified mean function. Figure 9.2g,h suggest both curvature and nonconstant variance.

In models with many regressors, we cannot necessarily associate shapes in a residual plot with a particular problem with the assumptions. Figure 9.3 shows a residual plot for the fit of the mean function $E(Y|X = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ for the artificial data given in the file `caution` from Cook and Weisberg (1999b). The right-opening megaphone is clear in this graph, suggesting non-constant variance. But these data were actually generated using a mean function

$$E(Y|X = \mathbf{x}) = \frac{|x_1|}{2 + (1.5 + x_2)^2} \quad (9.15)$$

and so the real problem is that the mean function is wrong. A nonnull residual plot in multiple regression indicates that something is wrong but does not necessarily tell what is wrong.

Residual plots in multiple regression *can* be interpreted just as residual plots in simple regression if two conditions are satisfied. First, the predictors should be approximately linearly related (Section 8.2.1). The second condition is on the mean function: we must be able to write the mean function in the form $E(Y|X = \mathbf{x}) = g(\boldsymbol{\beta}'\mathbf{x})$ for some unspecified function g . If either of these conditions fails, then residual plots cannot be interpreted as in simple regression (Cook and Weisberg, 1999b). In the `caution` data, the second condition fails because (9.15) cannot be written as a function of a single linear combination of the regressors.

9.1.6 Fuel Consumption Data

According to theory, if the mean function and other assumptions are correct, then *all possible residual plots* of residuals versus any function of the regressors or predictors should resemble a null plot, so many plots of residuals should be examined. Usual choices include plots versus each of the regressors and versus fitted values, as shown in Figure 9.4 for the fuel consumption data. None of the plots versus individual regressors in Figure 9.4a–d suggests any particular problems, apart from the relatively large positive residual for Wyoming and large negative residual for Alaska. In some of the graphs, the point for the District of Columbia is separated from the others. Wyoming is large but sparsely populated with a well-developed road system. Driving long distances for the necessities of life, such as going to see a doctor, will be common in this state. While Alaska is also very large and sparsely populated, most people live in relatively small areas around cities. Much of Alaska is not accessible by road. These conditions should result in lower use of motor fuel than might otherwise be expected. The District of Columbia is a very compact urban area with good rapid transit, so use of cars will generally be less. It has a small residual but unusual values for the regressors in the mean function, so it is

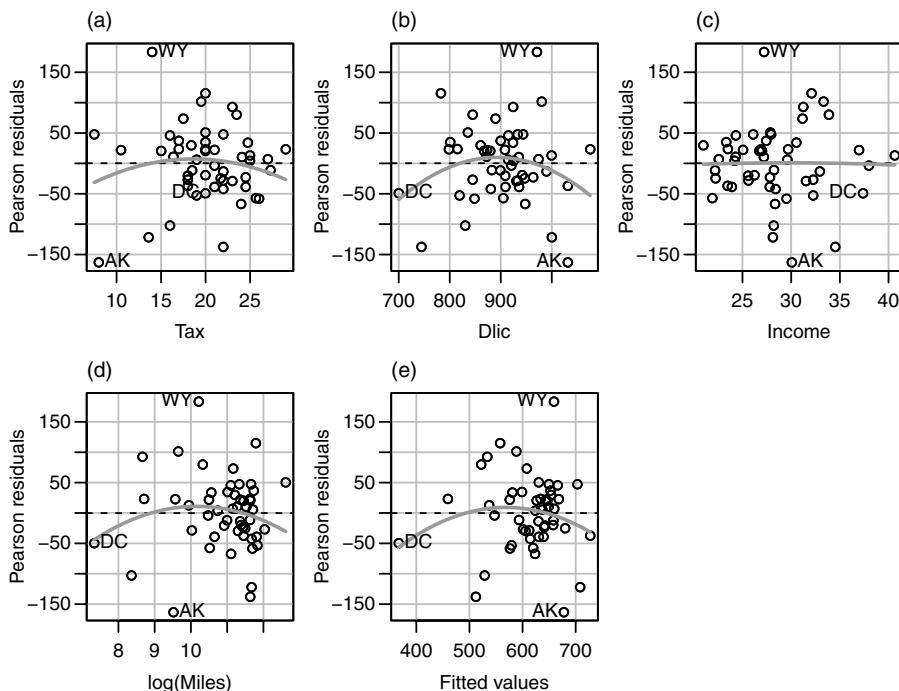


Figure 9.4 Residual plots for the fuel consumption data. The curves are quadratic fits used in lack-of-fit testing.

separated horizontally from most of the rest of the data. The District of Columbia has high leverage ($h_{9,9} = 0.415$), while the other two are candidates for outliers.

Figure 9.4e is a plot of residuals versus the fitted values, which are just a linear combination of the regressors. Some computer packages will produce this graph as the *only* plot of residuals and if only one plot were possible, this would be the plot to draw, as it contains some information from all the regressors in the mean function. There is a hint of curvature in this plot, possibly suggesting that the mean function is not adequate for the data. We will look at this more carefully in the next section.

9.2 TESTING FOR CURVATURE

Tests can be computed to help decide if residual plots such as those in Figure 9.4 are null plots or not. One helpful test looks for curvature in this plot. Suppose we have a plot of residuals \hat{e} versus a quantity U , where U could be a regressor in the mean function or a combination of regressors.² A simple test for curvature is to refit the original mean function with an additional regressor for U^2 added. The test for curvature is then based on the t -statistic for testing the coefficient for U^2 to be 0. If U does not depend on estimated coefficients, then a usual t -test of this hypothesis can be used. If U is equal to the fitted values so that it depends the estimated coefficients, then the test statistic should be compared with the standard normal distribution to get significance levels. This latter case is called *Tukey's test for nonadditivity* (Tukey, 1949).

The lack-of-fit tests for the residual plots in Figure 9.4 are the following:

	Test Stat	<i>p</i> -Value
Tax	-1.08	0.29
Dlic	-1.92	0.06
Income	-0.08	0.93
log(Miles)	-1.35	0.18
Tukeytest	-1.45	0.15

None of the tests has small significance levels, providing no evidence against the mean function.

As a second example, consider again the United Nations data with the model $\text{fertility} \sim \log(\text{ppgdp}) + \text{pctUrban}$. Plots of residuals versus the two regressors and versus fitted values are shown in Figure 9.5. Even

²This procedure is not recommended for factor, polynomial, or spline regressors.

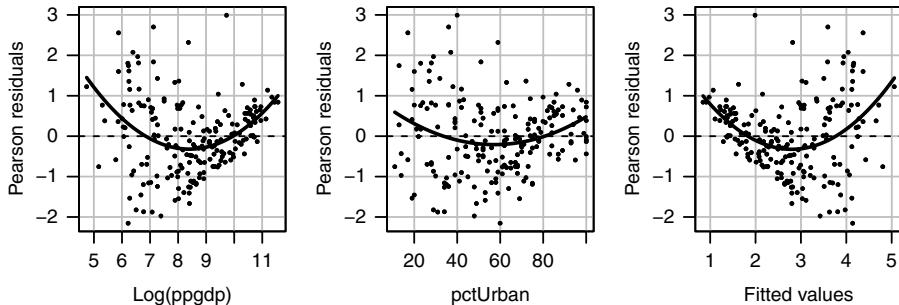


Figure 9.5 Residual plots for the UN data. The curved lines are quadratic polynomials fit to the residual plot and do not correspond exactly to the lack-of-fit tests that add a quadratic regressor to the original mean function.

without reference to the curved lines shown on the plot, the visual appearance of these plots suggests curvature, as confirmed by the lack-of-fit tests:

	Test Stat	p-Value
log (ppgdp)	5.41	0.000
pctUrban	3.29	0.001
Tukey test	5.42	0.000

All p -values are 0 to two decimal places, suggesting that the mean function is not adequate for these data.

Since the fit is inadequate, we should consider modification to get a mean function that matches the data well. One approach is to include both quadratic regressors and an interaction between $\log(\text{ppgdp})$ and pctUrban . Using the methods described elsewhere in this book, we conclude that the mean function $\text{fertility} \sim \log(\text{ppgdp}) + \text{pctUrban} + \log(\text{ppgdp}) : \text{pctUrban}$ matches adequately, with a p -value for Tukey's test of 0.09. Addition of a quadratic term in $\log(\text{ppgdp})$ would also provide a minor improvement, but we omit this because transforming a transformed predictor is unusual. The effects plot for this model is shown in Figure 9.6. For mostly rural countries, fertility is estimated to be very high with low ppgdp , and decline most rapidly as ppgdp increases. The effects are attenuated as pctUrban increases. The graph is slightly misleading, however, because there are few relatively wealthy rural countries, and no relatively poor urban countries.

9.3 NONCONSTANT VARIANCE

A nonconstant variance function in a residual plot may indicate that a constant variance assumption is false. There are at least four basic remedies for non-constant variance. The first is to use a variance stabilizing transformation,

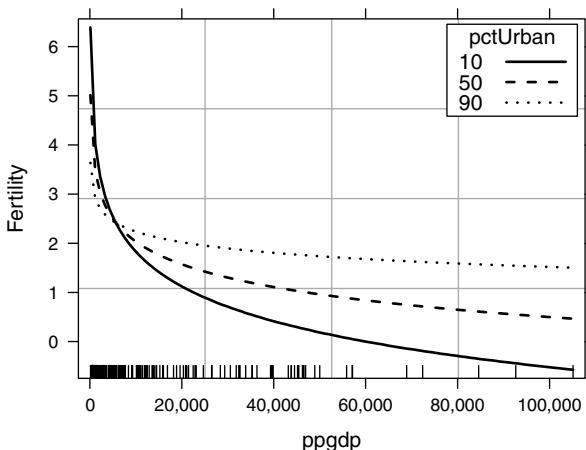


Figure 9.6 Effects plot for the UN data with an interaction.

Section 7.5, since replacing Y by Y_T may induce constant variance in the transformed scale. A second option is to find empirical weights that could be used in weighted least squares. Weights that are simple functions of single predictors, such as $\text{Var}(Y|X) = \sigma^2 X_1$, with $X_1 > 0$, can sometimes be justified theoretically. If replication is available, then within-group variances may be used to provide approximate weights. The third option is to do nothing and use the corrections for misspecified variances described in Section 7.2.1 at the cost of decreased efficiency of estimates.

The final option is to use *generalized linear models* that account for the nonconstant variance that is a function of the mean, introduced in Chapter 12.

9.4 OUTLIERS

In some problems, the observed response for a few of the cases may not seem to correspond to the model fitted to the bulk of the data. In a simple regression problem, such as displayed in Figure 1.9c, Section 1.4, this may be obvious from a plot of the response versus the predictor, where most of the cases lie near a fitted line but a few do not. Cases that do not follow the same model as the rest of the data are called *outliers*, and identifying these cases can be useful.

We use the *mean shift outlier model* to define outliers. Suppose that the i th case is a candidate for an outlier. We assume that the mean function for all other cases is $E(Y|X = \mathbf{x}_j) = \mathbf{x}_j' \boldsymbol{\beta}$, but for case i , the mean function is $E(Y|X = \mathbf{x}_i) = \mathbf{x}_i' \boldsymbol{\beta} + \delta$. The expected response for the i th case is shifted by an amount δ , and a test of $\delta = 0$ is a test for a single outlier in the i th case. In this development, we assume $\text{Var}(Y|X) = \sigma^2$.

Cases with large residuals are candidates for outliers. Not all large residual cases are outliers, since large errors e_i will occur with the frequency prescribed by the generating probability distribution. Whatever testing procedure we develop must offer protection against declaring too many cases to be outliers. This leads to the use of simultaneous testing procedures. Also, not all outliers are bad. For example, a geologist searching for oil deposits may be looking for outliers, if the oil is in the places where a fitted model does not match the data. Outlier identification is done relative to a specified model. If the form of the model is modified, the status of individual cases as outliers may change. Finally, some outliers will have greater effect on the regression estimates than will others, a point that is pursued shortly.

9.4.1 An Outlier Test

Suppose that the i th case is suspected to be an outlier. Define a new regressor U to be a dummy variable that has a 1 for its i th element and 0 for all other elements. Compute the regression of the response on both the regressors in X and U . The estimated coefficient for U is the estimate of the mean shift δ . The t -statistic for testing $\delta = 0$ against a two-sided alternative is the appropriate test statistic. Normally distributed errors are required for this test, and then the test will be distributed as Student's t with $n - p' - 1$ df.

We will now consider an alternative approach that will lead to the same test, but from a different point of view. The equivalence of the two approaches is left as an exercise.

Again suppose that the i th case is suspected to be an outlier. We can proceed as follows:

1. Delete the i th case from the data, so $n - 1$ cases remain in the reduced data set.
2. Using the reduced data set, estimate β and σ^2 . Call these estimates $\hat{\beta}_{(i)}$ and $\hat{\sigma}_{(i)}^2$ to remind us that case i was not used in estimation. The estimator $\hat{\sigma}_{(i)}^2$ has $n - p' - 1$ df.
3. For the deleted case, compute the fitted value $\hat{y}_{i(i)} = \mathbf{x}'_{(i)} \hat{\beta}_{(i)}$. Since the i th case was not used in estimation, y_i and $\hat{y}_{i(i)}$ are independent. The variance of $y_i - \hat{y}_{i(i)}$ is given by

$$\text{Var}(y_i - \hat{y}_{i(i)} | \mathbf{X}) = \sigma^2 + \sigma^2 \mathbf{x}'_{(i)} (\mathbf{X}'_{(i)} \mathbf{X}_{(i)})^{-1} \mathbf{x}_{(i)} \quad (9.16)$$

where $\mathbf{X}_{(i)}$ is the matrix \mathbf{X} with the i th row deleted. This variance is estimated by replacing σ^2 with $\hat{\sigma}_{(i)}^2$ in (9.16).

4. Now $E(y_i - \hat{y}_{i(i)} | \mathbf{X}) = \delta$, is 0 under the null hypothesis that case i is not an outlier but nonzero otherwise. Assuming normal errors, a Student's t -test of the hypothesis $\delta = 0$ is given by

$$t_i = \frac{y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)} \sqrt{1 + \mathbf{x}'_i (\mathbf{X}'_{(i)} \mathbf{X}_{(i)})^{-1} \mathbf{x}_i}} \quad (9.17)$$

This test has $n - p' - 1$ df, and is identical to the t -test suggested in the first paragraph of this section.

There is a simple computational formula for t_i in (9.17). We first define an intermediate quantity, often called a *standardized residual*, by

$$r_i = \frac{\hat{e}_i}{\hat{\sigma} \sqrt{1 - h_{ii}}} \quad (9.18)$$

where the h_{ii} is the leverage for the i th case, defined at (9.8). Like the residuals \hat{e}_i , the r_i have mean 0, but unlike the \hat{e}_i , the variances of the r_i are all equal to 1. Because the h_{ii} need not all be equal, the r_i are not just a rescaling of the \hat{e}_i . With the aid of Appendix A.13, one can show that t_i can be computed as

$$t_i = r_i \left(\frac{n - p' - 1}{n - p' - r_i^2} \right)^{1/2} = \frac{\hat{e}_i}{\hat{\sigma}_{(i)} \sqrt{1 - h_{ii}}} \quad (9.19)$$

A statistic divided by its estimated standard deviation is usually called a *studentized statistic*, in honor of W.S. Gosset, who first wrote about the t -distribution using the pseudonym Student.³ The residual t_i is called a *studentized residual*. We see that r_i and t_i carry the same information since one can be obtained from the other via a simple formula. Also, this result shows that t_i can be computed from the residuals, the leverages and $\hat{\sigma}^2$, so we don't need to delete the i th case, or to add a variable U , to get the outlier test.

Studentized or standardized residuals are sometimes used in place of the Pearson residuals in residual plots described in Section 9.1. This has the advantage of removing some of the nonconstant variance in the plotted residuals, but the disadvantage of losing the units of the residuals. Any of these sets of residuals can be used in graphical methods with little difference in interpretation.

9.4.2 Weighted Least Squares

If we initially assumed that $\text{Var}(Y|X) = \sigma^2/w$ for known positive weights w , then in Equation (9.18), we compute the residuals \hat{e}_i using the correct weighted formula (9.14) and leverages are the diagonal elements of (9.13). Otherwise, no changes are required.

³See St. Andrews University (2003) for a biography of Student.

9.4.3 Significance Levels for the Outlier Test

If the analyst suspects in advance that the i th case is an outlier, then t_i should be compared with the central t -distribution with the appropriate number of df . The analyst rarely has a prior choice for the outlier. Testing the case with the largest value of $|t_i|$ to be an outlier is like performing n significance tests, one for each of n cases. If, for example, $n = 65$, $p' = 4$, the probability that a t -statistic with 60 df exceeds 2.000 in absolute value is 0.05; however, the probability that the largest of 65 independent t -tests exceeds 2.000 is 0.964, suggesting quite clearly the need for a different critical value for a test based on the maximum of many tests. Since tests based on the t_i are correlated, this computation is only a guide. Bretz et al. (2010) discuss multiple testing problems in more generality.

For outlier testing, the usual correction for multiple testing is based on the *Bonferroni inequality*, which states that for n tests each of size α , the probability of falsely labeling at least one case as an outlier is no greater than na . This procedure is *conservative* and provides an upper bound on the probability. For example, the Bonferroni inequality specifies only that the probability of the maximum of 65 tests exceeding 2.00 is no greater than $65(0.05)$, which is larger than 1. Choosing the critical value to be the $(\alpha/n) \times 100\%$ point of t will give a significance level of no more than $n(\alpha/n) = \alpha$. We would choose a level of $0.05/65 = 0.00077$ for each test to give an overall level of no more than $65(0.00077) = 0.05$.

Standard functions for the t -distribution can be used to compute p -values for the outlier test: simply compute the p -value as usual and then multiply by the sample size. If this number is smaller than 1, then this is the p -value adjusted for multiple testing. If this number exceeds 1, then the p -value is 1.

In Forbes's data, Example 1.1, case 12 was suspected to be an outlier because of its large residual. To perform the outlier test, we first need the standardized residual, which is computed using (9.18) from $\hat{e}_i = 1.36$, $\hat{\sigma} = 0.379$, and $h_{12,12} = 0.0639$,

$$r_{12} = \frac{1.359}{0.379\sqrt{1 - 0.0639}} = 3.708$$

and the outlier test is

$$t_i = 3.708 \left(\frac{17 - 2 - 1}{17 - 2 - 3.708^2} \right)^{1/2} = 12.41$$

The nominal two-sided p -value corresponding to this test statistic when compared with the $t(14)$ distribution is 6.13×10^{-9} . If the location of the outlier was not selected in advance, the Bonferroni-adjusted p -value is $17 \times 6.13 \times 10^{-9} = 1.04 \times 10^{-7}$. This very small value supports case 12 as an outlier.

The test locates an outlier, but it does not tell us what to do about it. If we believe that the case is an outlier because of a blunder, for example, an unusually large measurement error, or a recording error, then we might delete the outlier and analyze the remaining cases without the suspected case. Sometimes, we can try to figure out why a particular case is outlying, and finding the cause may be the most important part of the analysis. All this depends on the context of the problem you are studying.

9.4.4 Additional Comments

There is a vast literature on methods for handling outliers, including Barnett and Lewis (1994), Beckman and Cook (1983), and Hawkins (1980). If a set of data has more than one outlier, a sequential approach can be recommended, but the cases may mask each other, making finding groups of outliers difficult. Cook and Weisberg (1982, p. 28) provide the generalization of the mean shift model given here to multiple cases. Hawkins et al. (1984) provide a promising method for searching all subsets of cases for outlying subsets. Bonferroni bounds for outlier tests are discussed by Cook and Prescott (1981). They find that for one-case-at-a-time methods, the bound is very accurate, but it is much less accurate for multiple-case methods.

The testing procedure helps find outliers, to make them available for further study. Alternatively, we could design robust statistical methods that can tolerate or accommodate some proportion of bad or outlying data; see, for example, Staudte and Sheather (1990).

9.5 INFLUENCE OF CASES

Single cases or small groups of cases can strongly influence the fit of a regression model. In Anscombe's example in Figure 1.9d, the fitted model depends entirely on the one point with $x = 19$. If that case were deleted, we could not estimate the slope. If it were perturbed, moved around a little, the fitted line would follow the point. In contrast, if any of the other cases were deleted or moved around, the change in the fitted mean function would be quite small.

The general idea of *influence analysis* is to study changes in a specific part of the analysis when the data are slightly perturbed. Whereas statistics such as residuals are used to find problems with a model, influence analysis is done as if the model were correct, and we study the robustness of the conclusions, given a particular model, to the perturbations. The most useful and important method of perturbing the data is deleting the cases from the data one at a time. We then study the effects or influence of each individual case by comparing the full data analysis to the analysis obtained with a case removed. Cases whose removal causes major changes in the analysis are called influential.

Using the notation from the last section, a subscript (i) means "with the i th case deleted," so, for example, $\beta_{(i)}$ is the estimate of β computed without case

i , $\mathbf{X}_{(i)}$ is the $(n - 1) \times p'$ matrix obtained from \mathbf{X} by deleting the i th row, and so on. In particular, then,

$$\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'_{(i)} \mathbf{X}_{(i)})^{-1} \mathbf{X}'_{(i)} \mathbf{Y}_{(i)} \quad (9.20)$$

A simple computing formula for $\hat{\boldsymbol{\beta}}_{(i)}$ is derived in Appendix A.13.

Figure 9.7 is a scatterplot matrix of coefficient estimates for the three parameters in the UN data, based on the model $\text{fertility} \sim \log(\text{ppgdp}) + \text{lifeExpF}$ obtained by deleting cases one at a time. Every time a case is deleted, different coefficient estimates are obtained. Apart from the points for Botswana, Lesotho, South Africa, and Swaziland, all 2D plots in Figure 9.7 are more or less elliptically shaped, which is a common characteristic of the deletion estimates. Deletion of any of the four African countries

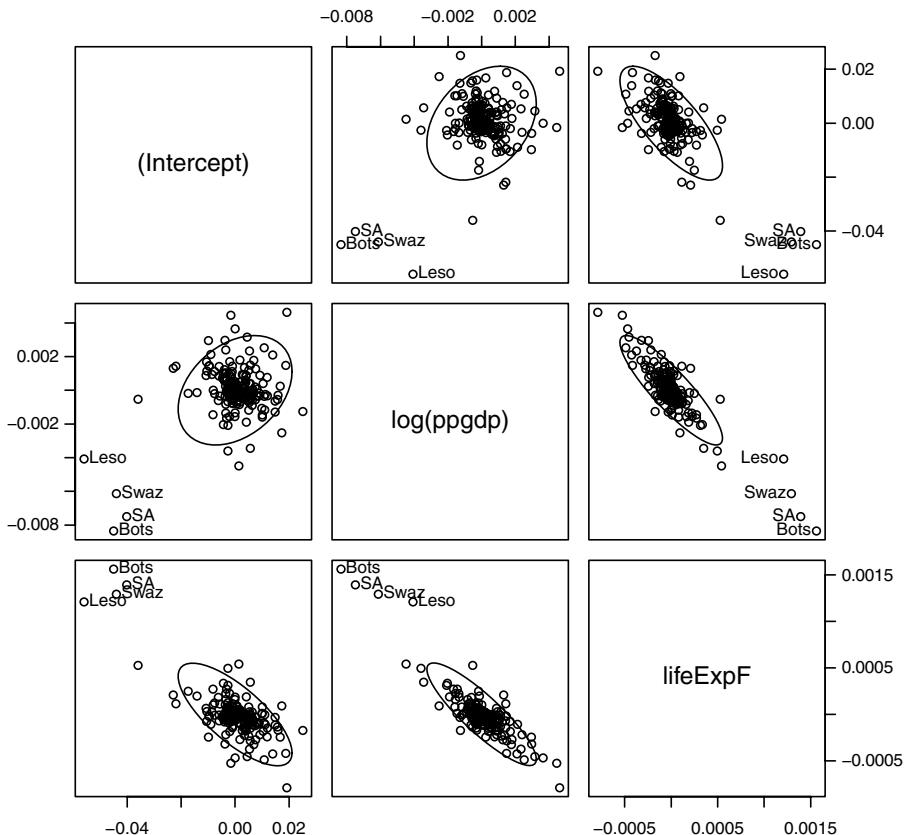


Figure 9.7 Estimates of parameters in the UN data obtained by deleting one case at a time. The ellipses shown on the plots would be 95% confidence regions for the bivariate mean in each plot if the points in the plot were a sample from a bivariate normal distribution.

Table 9.1 Coefficient Estimates and Standard Errors for Fitting with the UN Data, with All the Data and 4 Countries Removed

	Est, All	SE, All	Est, Reduced	SE, Reduced
(Intercept)	3.5074	0.1271	3.7322	0.1289
log(ppgdp)	-0.0654	0.0178	-0.0336	0.0182
lifeExpF	-0.0282	0.0027	-0.0349	0.0029

decreases the estimated intercept and coefficient for $\log(\text{ppgdp})$, and increases the coefficient for lifeExpF . If we refit the regression model after deleting all four of these countries, we get the coefficient estimates and standard errors shown in Table 9.1. Removing the four countries does not materially change the SEs of the estimates, but the slope estimate for $\log(\text{ppgdp})$ is reduced by almost 50% and the slope estimate for lifeExpF is increased in magnitude by about 25%. The analysis is unstable and changes in important ways, depending on the cases that are included in the data set. This could reflect either a few countries that are really unusual, or, as is more likely here, a difference in the relationship between these variables for African countries in general, as discussed in Section 5.1.

9.5.1 Cook's Distance

Cook (1977) suggested a method that can be used to summarize the difference between $\hat{\beta}$ and $\hat{\beta}_{(i)}$ with a single number. We define *Cook's distance* D_i to be

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(\mathbf{X}'\mathbf{X})(\hat{\beta}_{(i)} - \hat{\beta})}{p'\hat{\sigma}^2} \quad (9.21)$$

This statistic has several desirable properties. First, contours of constant D_i are ellipsoids. Second, the contours can be thought of as defining the distance from $\hat{\beta}_{(i)}$ to $\hat{\beta}$. Third, D_i does not depend on parameterization, so if the columns of \mathbf{X} are modified by linear transformation, D_i is unchanged. Finally, if we define vectors of fitted values as $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ and $\hat{\mathbf{Y}}_{(i)} = \mathbf{X}\hat{\beta}_{(i)}$, then (9.21) can be rewritten as

$$D_i = \frac{(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})'(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})}{p'\hat{\sigma}^2} \quad (9.22)$$

so D_i is the ordinary Euclidean distance between $\hat{\mathbf{Y}}$ and $\hat{\mathbf{Y}}_{(i)}$. Cases for which D_i is large have substantial influence on both the estimate of β and on fitted values, and deletion of them may result in important changes in conclusions.

9.5.2 Magnitude of D_i

Cases with large values of D_i are the ones whose deletion will result in substantial changes in the analysis. Typically, the case with the largest D_i , or in large data sets the cases with the largest few D_i , will be of interest. One method of calibrating D_i is obtained by analogy to confidence regions. If D_i were exactly equal to the $\alpha \times 100\%$ point of the F -distribution with p' and $n - p' df$, then deletion of the i th case would move the estimate of $\hat{\beta}$ to the edge of a $(1 - \alpha) \times 100\%$ confidence region based on the complete data. Since for most F -distributions the 50% point is near one, 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 one, deletion of a case will not change the estimate of $\hat{\beta}$ 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.

9.5.3 Computing D_i

From the derivation of Cook's distance, it is not clear that using these statistics is computationally convenient. However, the results sketched in Appendix A.13 can be used to write D_i using more familiar quantities. A simple form for D_i is

$$D_i = \frac{1}{p'} r_i^2 \frac{h_{ii}}{1 - h_{ii}} \quad (9.23)$$

D_i is a product of the square of the i th standardized residual r_i and a monotonic function of the leverage h_{ii} . If p' is fixed, the size of D_i will be determined by two different sources: the size of r_i , a random variable reflecting lack of fit of the model at the i th case, and h_{ii} , reflecting the location of \mathbf{x}_i relative to $\bar{\mathbf{x}}$. A large value of D_i may be due to large r_i , large h_{ii} , or both.

Rat Data

An experiment was conducted to investigate the amount of a particular drug present in the liver of a rat. Nineteen rats were randomly selected, weighed, placed under light ether anesthesia and given an oral dose of the drug. Because large livers would absorb more of a given dose than smaller livers, the actual dose an animal received was approximately determined as 40 mg of the drug per kilogram of body weight. Liver weight is known to be strongly related to body weight. After a fixed length of time, each rat was sacrificed, the liver weighed, and the percentage of the dose in the liver determined. The experimental hypothesis was of no relationship between the percentage of the dose in the liver y and the body weight `BodyWt`, liver weight `LiverWt`, and relative Dose. The data, provided by Dennis Cook and given in the file `rat`, are shown in Figure 9.8. As had been expected, the marginal summary plots for y versus

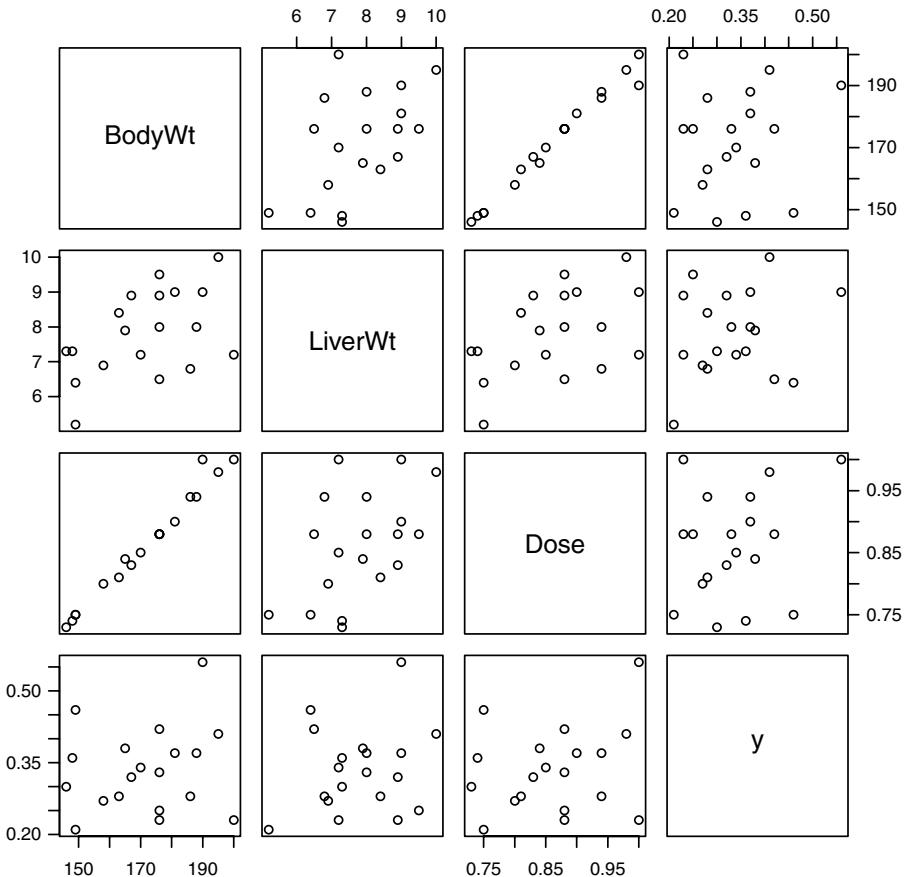


Figure 9.8 Scatterplot matrix for the rat data.

Table 9.2 Regression Summary for the Rat Data

	Estimate	Std. Error	t Value	Pr(> t)
(Intercept)	0.2659	0.1946	1.37	0.1919
BodyWt	-0.0212	0.0080	-2.66	0.0177
LiverWt	0.0143	0.0172	0.83	0.4193
Dose	4.1781	1.5226	2.74	0.0151

$\hat{\sigma} = 0.0773$ with 15 df, $R^2 = 0.3639$.

each of the predictors suggests no relationship, and none of the simple regressions is significant, all having t -values less than 1.

The fitted regression summary for the regression of y on the three predictors is shown in Table 9.2. `BodyWt` and `Dose` have significant t -tests, with $p < 0.05$ in both cases, indicating that the two measurements combined are a

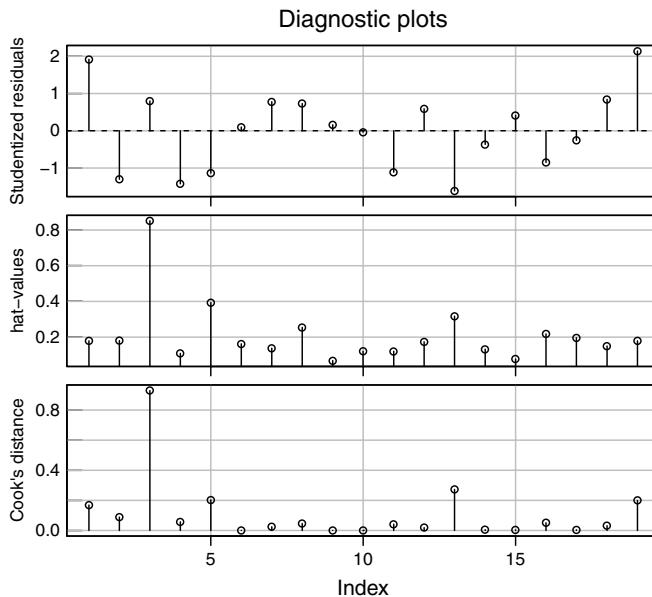


Figure 9.9 Diagnostic statistics for the rat data.

useful indicator of Y ; if `LiverWt` is dropped from the mean function, the same phenomenon appears. The analysis so far, based only on summary statistics, might lead to the conclusion that while neither `BodyWt` or `Dose` are associated with the response when the other is ignored, in combination, they are associated with the response. But, from Figure 9.8, `Dose` and `BodyWt` are almost perfectly linearly related, so they measure the same thing!

We turn to diagnostics to attempt to resolve this paradox. Figure 9.9 displays diagnostic statistics for the mean function with all the regressors included. The studentized residuals that test for outliers are not particularly large. However, Cook's distance immediately locates a possible cause: $D_3 = 0.93$ is much larger than all the other values of Cook's distance, suggesting that the third case may have large enough influence on the fit to induce the anomaly. The value of $h_{33} = 0.85$ indicates that the problem is an unusual set of predictors for case 3.

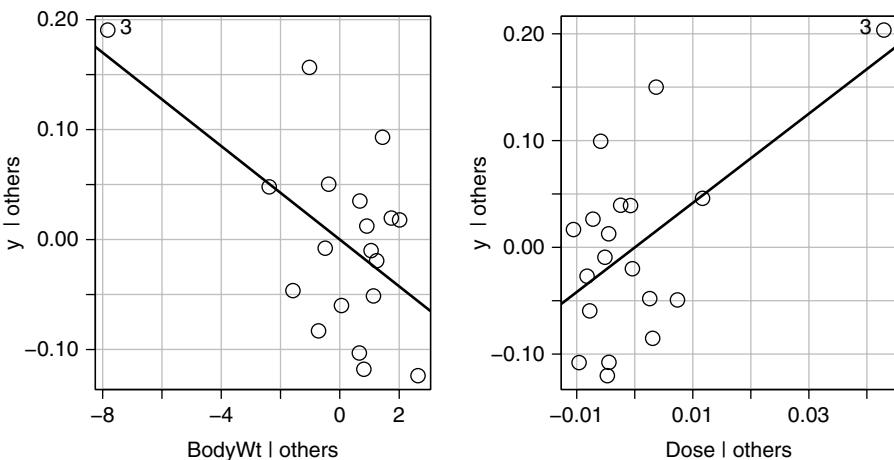
The fit without the third case is shown in Table 9.3. The paradox dissolves and the apparent relationship found in the first analysis can thus be ascribed to the third case alone.

Once again, the diagnostic analysis finds a problem, but does not tell us what to do next, and this will depend on the context of the problem. Rat number 3, with weight 190 g, was reported to have received a full dose of 1.000, which was a larger dose than it should have received, according to the rule for assigning doses; for example, rat number 8 with weight of 195 g got a lower dose of 0.98. A number of causes for the result found in the first analysis are

Table 9.3 Regression Summary for the Rat Data with Case 3 Deleted

	Estimate	Std. Error	t Value	Pr(> t)
(Intercept)	0.3114	0.2051	1.52	0.1512
BodyWt	-0.0078	0.0187	-0.42	0.6838
LiverWt	0.0090	0.0187	0.48	0.6374
Dose	1.4849	3.7131	0.40	0.6953

$\hat{\sigma} = 0.0782$ with 14 df, $R^2 = 0.0211$.

**Figure 9.10** Added-variable plots for BodyWt and Dose.

possible: (1) the dose or weight recorded for case 3 was in error, so the case should probably be deleted from the study, or (2) the regression fit in the second analysis is not appropriate except in the region defined by the 18 points excluding case 3. This has many implications concerning the experiment. It is possible that the combination of dose and rat weight chosen was fortuitous, and that the lack of relationship found would not persist for any other combinations of them, since inclusion of a data point apparently taken under different conditions leads to a different conclusion. This suggests the need for collection of additional data, with dose determined by some rule other than a constant proportion of weight.

9.5.4 Other Measures of Influence

The added-variable plots introduced in Section 3.1 provide a graphical diagnostic for influence. Cases corresponding to points at the left or right of an added-variable plot that do not match the general trend in the plot are likely to be influential for the variable that is to be added. For example, Figure 9.10 shows the added-variable plots for BodyWt and for Dose for the rat data. The

point for case 3 is clearly separated from the others, and is a likely influential point based on these graphs. The added-variable plot does not correspond exactly to Cook's distance, but to *local influence* defined by Cook (1986).

As with the outlier problem, influential groups of cases may serve to mask each other and may not be found by examination of cases one at a time. In some problems, multiple-case methods may be desirable; see Cook and Weisberg (1982, section 3.6).

9.6 NORMALITY ASSUMPTION

The assumption of normal errors plays only a minor role in linear regression analysis. It is needed primarily for inference with small samples, and even then the bootstrap outlined in Section 7.7 can be used for inference. Furthermore, nonnormality of the unobservable errors is very difficult to diagnose in small samples by examination of residuals. The relationship between the errors and the residuals is

$$\begin{aligned}\hat{\mathbf{e}} &= (\mathbf{I} - \mathbf{H})\mathbf{Y} \\ &= (\mathbf{I} - \mathbf{H})(\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}) \\ &= (\mathbf{I} - \mathbf{H})\mathbf{e}\end{aligned}$$

because $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$. In scalar form, the i th residual is

$$\hat{e}_i = e_i - \left(\sum_{j=1}^n h_{ij} e_j \right) \quad (9.24)$$

The first term on the right of (9.24) is the i th error. The second term is a linear combination of all the errors, and by the central limit theorem, this will generally be nearly normally distributed even if the e_i are not normally distributed. With a small or moderate sample size n , the second term can dominate the first, and the residuals can behave like a normal sample even if the errors are not normal. Gnanadesikan (1997) refers to this as the *supernormality* of residuals.

As n increases for fixed p' , the second term in (9.24) has small variance compared with the first term, and the distribution of the residuals will more closely resemble the distribution of the errors, and so the residuals could be used to test for normality. Should a test of normality be desirable, a *normal probability plot* can be used. A general treatment of probability plotting is given by Gnanadesikan (1997). Suppose we have a sample of n numbers z_1, z_2, \dots, z_n , and we wish to examine the hypothesis that the z 's are a sample from a normal distribution with unknown mean μ and variance σ^2 . A useful way to proceed is as follows:

1. Order the z 's to get $z_{(1)} \leq z_{(2)} \leq \dots \leq z_{(n)}$. The ordered z 's are called the *sample order statistics*.
2. Now, consider a standard normal sample of size n . Let $u_{(1)} \leq u_{(2)} \leq \dots \leq u_{(n)}$ be the mean values of the order statistics that would be obtained if we repeatedly took samples of size n from the standard normal. The $u_{(i)}$'s are called the *expected order statistics*. The $u_{(i)}$ are available in printed tables or can be well approximated using a computer program.⁴
3. If the z s are normal, then

$$E(z_{(i)}) = \mu + \sigma u_{(i)}$$

so that the regression of $z_{(i)}$ on $u_{(i)}$ will be a straight line. If it is not straight, we have evidence against normality.

Judging whether a probability plot is sufficiently straight requires experience. Daniel and Wood (1980) provided many pages of plots to help the analyst learn to use these plots; this can be easily recreated using a computer package that allows one quickly to look at many plots. Atkinson (1985) used a variation of the bootstrap to calibrate probability plots.

Many statistics have been proposed for testing a sample for normality. One of these that works extremely well is the Shapiro and Wilk (1965) W statistic, which is essentially the square of the correlation between the observed order statistics and the expected order statistics. Normality is rejected if W is too small. Royston (1982a–c) provides details and computer routines for the calculation of the test and for finding p -values.

Figure 9.11 shows normal probability plots of the residuals for the heights data (Section 1.1) and for the transactions data (Section 7.7.1). Both have large enough samples for normal probability plots to be useful. For the heights data, the plot is very nearly straight, indicating no evidence against normality. For the transactions data, normality is in doubt because the plot is not straight. In particular, there are very large positive residuals well away from a fitted line. This supports the earlier claim that the errors for this problem are likely to be skewed with too many large values.

9.7 PROBLEMS

- 9.1** (Data file: `Rpdata`) The data in this file has a response y and six regressors x_1, \dots, x_6 . The data are artificial, to make a few points.

⁴Suppose $\Phi(x)$ is a function that returns the area p to the left of x under a standard normal distribution, and $\Phi^{-1}(p)$ computes the inverse of the normal, so for a given value of p , it returns the associated value of x . Then the i th expected normal order statistic is approximately $\Phi^{-1}[(i - (3/8))/(n + (1/4))]$ (Blom, 1958).

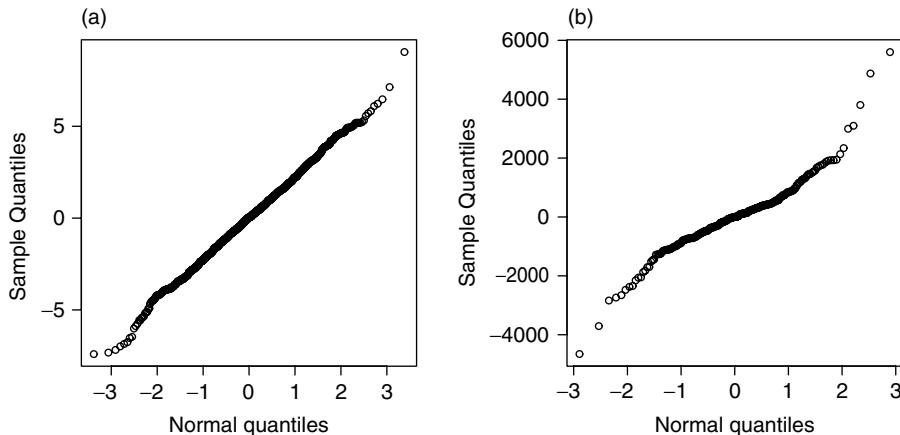


Figure 9.11 Normal probability plots of residuals for (a) the heights data and (b) the transactions data.

- 9.1.1 First draw a scatterplot matrix of all data and comment. Is there anything strange?
- 9.1.2 Fit the OLS regression $y \sim x_1 + x_2 + x_3 + x_4 + x_5 + x_6$. Is there anything strange?
- 9.1.3 Draw a plot of residuals versus fitted values. Is there anything strange? See Stefanski (2007) if you want to find out how this data set came about.

9.2 Working with the hat matrix

- 9.2.1 Prove the results given by (9.9) and (9.10).
- 9.2.2 Prove that $1/n \leq h_{ii} \leq 1/r$, where h_{ii} is a diagonal entry in \mathbf{H} , and r is the number of rows in \mathbf{X} that are exactly the same as \mathbf{x}_i .

9.3 Alaska pipeline faults (Data file: `pipeline`) This example compares in-field ultrasonic measurements of the depths of defects, `Field`, in the Alaska oil pipeline with measurements of the same defects in a laboratory, `Lab`. The lab measurements were done in six different batches, in the variable `Batch`. The goal is to decide if the field measurement can be used to predict the more accurate lab measurement. The lab measurement is the response variable and the field measurement is the predictor variable. The data are from the National Institute of Science and Technology (2012, section 6).

- 9.3.1 Draw the scatterplot of `Lab` versus `Field`, and comment on the applicability of the simple linear regression model.
- 9.3.2 Fit the simple regression model, get the residual plot, and summarize. Explain why the plot suggests nonconstant variance and provide a test for nonconstant variance.

9.3.3 Having diagnosed nonconstant variance, consider four options for summarizing these data: (1) do nothing; use the OLS fit computed previously; (2) use OLS for fitting but the bootstrap to estimate standard errors; (3) use WLS with the variance function $\text{Var}(\text{LablField}) = \sigma^2 \times \text{Field}$; and (4) use OLS for fitting by the correction for nonconstant variance described in Section 7.2.1. Compare the solutions for the slope and its standard error.

9.4 Simple regression Consider the simple regression model, $E(Y|X = x) = \beta_0 + \beta_1 x$, $\text{Var}(Y|X = x) = \sigma^2$.

9.4.1 Find a formula for the h_{ij} and for the leverages h_{ii} .

9.4.2 In a 2D plot of the response versus the predictor in a simple regression problem, explain how high-leverage points can be identified.

9.4.3 Make up a predictor X so that the value of the leverage in simple regression for one of the cases is equal to 1.

9.5 QR factorization and the hat matrix Using the QR factorization defined in Appendix A.13, show that $\mathbf{H} = \mathbf{Q}\mathbf{Q}'$. Hence, if \mathbf{q}_i is the i th row of \mathbf{Q} ,

$$h_{ii} = \mathbf{q}_i' \mathbf{q}_i \quad h_{ij} = \mathbf{q}_i' \mathbf{q}_j$$

This means that if the QR factorization of \mathbf{X} has been computed, h_{ii} is the sum of squares of the elements of \mathbf{q}_i , and the less-frequently used off-diagonal elements h_{ij} are the sums of products of the elements of \mathbf{q}_i and \mathbf{q}_j .

9.6 Let \mathbf{U} be an $n \times 1$ vector with 1 as its first element and 0s elsewhere. Consider computing the regression of \mathbf{U} on an $n \times p'$ full rank matrix \mathbf{X} . As usual, let $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ be the hat matrix with elements h_{ij} .

9.6.1 Show that the elements of the vector of fitted values from the regression of \mathbf{U} on \mathbf{X} are the h_{1j} , $j = 1, 2, \dots, n$.

9.6.2 Show that the first element of the vector of residuals is $1 - h_{11}$, and the other elements are $-h_{1j}$, $j > 1$.

9.7 Two $n \times n$ matrices \mathbf{A} and \mathbf{B} are *orthogonal* if $\mathbf{AB} = \mathbf{BA} = \mathbf{0}$. Show that $\mathbf{I} - \mathbf{H}$ and \mathbf{H} are orthogonal. Use this result to show that as long as the intercept is in the mean function, the slope of the regression of $\hat{\mathbf{e}}$ on $\hat{\mathbf{Y}}$ is 0. What is the slope of the regression of $\hat{\mathbf{e}}$ on \mathbf{Y} ?

9.8 California water (Data file: `water`) Draw residual plots for the mean function described in Problem 8.3.4 for the California water data, and comment on your results. Test for curvature as a function of fitted values.

Table 9.4 Crustacean Zooplankton Species Data (Dodson, 1992)

Variable	Description
Species	Number of zooplankton species
MaxDepth	Maximum lake depth, m
MeanDepth	Mean lake depth, m
Cond	Specific conductance, micro Siemans
Elev	Elevation, m
Lat	N latitude, degrees
Long	W longitude, degrees
Dist	Distance to nearest lake, km
NLakes	Number of lakes within 20 km
Photo	Rate of photosynthesis, mostly by the ^{14}C method
Area	Surface area of the lake, in hectares
Lake	Name of lake

9.9 Lake diversity (Data file: `lakes`) The number of crustacean zooplankton species present in a lake can be different, even for two nearby lakes. The data from Dodson (1992) give the number of known crustacean zooplankton species for 69 world lakes. Also included are a number of characteristics of each lake. There are some missing values; most computer programs will delete all rows of data that are missing any of the predictors and the response, so your analysis will likely be based on the 42 fully observed lakes. The goal of the analysis is to understand how the number of species present depends on the other measured variables that are characteristics of the lake. The variables are described in Table 9.4.

Decide on appropriate transformations of the data to be used in this problem. Then, fit appropriate linear regression models, and summarize your results. Include residual analysis to support your conclusions.

9.10 In an unweighted regression problem with $n = 54$, $p' = 5$, the results included $\hat{\sigma} = 4.0$ and the following statistics for four of the cases:

\hat{e}_i	h_{ii}
1.000	0.9000
1.732	0.7500
9.000	0.2500
10.295	0.1850

For each of these four cases, compute r_i , D_i , and t_i . Test each of the four cases to be an outlier. Make a qualitative statement about the influence of each case on the analysis.

9.11 (Data file: `fuel2001`) In the fuel consumption data, consider fitting the mean function

$$E(\text{Fuel}|X) = \beta_0 + \beta_1 \text{Tax} + \beta_2 \text{Dlic} + \beta_3 \text{Income} + \beta_4 \log(\text{Miles})$$

For this regression, we find $\hat{\sigma} = 64.891$ with $46 df$, and the diagnostic statistics for four states and the District of Columbia were the following:

	Fuel	\hat{e}_i	h_{ii}
Alaska	514.279	-163.145	0.256
New York	374.164	-137.599	0.162
Hawaii	426.349	-102.409	0.206
Wyoming	842.792	183.499	0.084
District of Columbia	317.492	-49.452	0.415

Compute D_i and t_i for each of these cases, and test for one outlier. Which is most influential?

- 9.12** The matrix $(\mathbf{X}'_{(i)} \mathbf{X}_{(i)})$ can be written as $(\mathbf{X}'_{(i)} \mathbf{X}_{(i)}) = \mathbf{X}' \mathbf{X} - \mathbf{x}_i \mathbf{x}'_i$, where \mathbf{x}'_i is the i th row of \mathbf{X} . By direct multiplication, use this definition to verify that (A.44) holds.

- 9.13** The quantity $y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{(i)}$ is the residual for the i th case when $\boldsymbol{\beta}$ is estimated without the i th case. Use (A.44) to show that

$$y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{(i)} = \frac{\hat{e}_i}{1 - h_{ii}}$$

This quantity is called the *predicted residual*, or the *PRESS residual*.

- 9.14** Use Appendix A.13 to verify (9.23).

- 9.15** (Data file: lathe) Refer to the lathe data in Problem 5.12.

- 9.15.1** Starting with the full second-order model, use the Box–Cox method to show that an appropriate scale for the response is the logarithmic scale.

- 9.15.2** Find the two cases that are most influential in the fit of the quadratic mean function for $\log(\text{Life})$, and explain why they are influential. Delete these points from the data, refit the quadratic mean function, and compare with the fit with all the data.

- 9.16 Florida election 2000** (Data file: florida) In the 2000 election for U.S. president, the counting of votes in Florida was controversial. In Palm Beach County in south Florida, for example, voters used a so-called butterfly ballot. Some believe that the layout of the ballot caused some voters to cast votes for Buchanan when their intended choice was Gore.

The data from Smith (undated) has four variables, County, the county name, and Gore, Bush, and Buchanan, the number of votes for each of these three candidates. Draw the scatterplot of Buchanan versus Bush, and test the hypothesis that Palm Beach County is an outlier relative to the simple linear regression mean function for $E(\text{Buchanan}|\text{Bush})$. Identify another county with an unusual value of the Buchanan vote, given its Bush vote, and test that county to be an outlier. State your conclusions from the test, and its relevance, if any, to the issue of the butterfly ballot.

Next, repeat the analysis, but first consider transforming the variables in the plot to better satisfy the assumptions of the simple linear regression model. Again test to see if Palm Beach County is an outlier, and summarize.

- 9.17** (Data file: landrent) These data were collected by Douglas Tiffany to study the variation in rent paid in 1977 for agricultural land planted to alfalfa. The variables are average rent per acre Y planted to alfalfa, average rent paid X_1 for all tillable land, density of dairy cows X_2 (number per square mile), proportion X_3 of farmland used as pasture, and $X_4 = 1$ if liming is required to grow alfalfa and 0 otherwise.

The unit of analysis is a county in Minnesota; the 67 counties with appreciable rented farmland are included. Alfalfa is a high protein crop that is suitable feed for dairy cows. 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. Use all the techniques learned so far to explore these data with regard to understanding rent structure. Summarize your results.

- 9.18** (Data file: cloud) The data summarize the results of the first Florida Area Cumulus Experiment, or FACE-1, designed to study the effectiveness of cloud seeding to increase rainfall in a target area (Woodley et al., 1977). A fixed target area of approximately 3000 square miles was established to the north and east of Coral Gables, Florida. During the summer of 1975, each day was judged on its suitability for seeding. The decision to use a particular day in the experiment was based primarily on a suitability criterion S depending on a mathematical model for rainfall. Days with $S > 1.5$ were chosen as experimental days; there were 24 days chosen in 1975. On each day, the decision to seed was made by flipping a coin; as it turned out, 12 days were seeded, 12 unseeded. On seeded days, silver iodide was injected into the clouds from small aircraft. The predictors and the response are defined in Table 9.5.

The goal of the analysis is to decide if there is evidence that cloud seeding is effective in increasing rainfall. Begin your analysis by drawing appropriate graphs. Obtain appropriate transformations of predictors. Fit

Table 9.5 The Florida Area Cumulus Experiment on Cloud Seeding

Variable	Description
A	Action: 1 = seed, 0 = do not seed
D	Days after the first day of the experiment (June 16, 1975 = 0)
S	Suitability for seeding
C	Percentage cloud cover in the experimental area, measured using radar in Coral Gables, Florida
P	Prewetness, amount of rainfall in the hour preceding seeding in 10^7 cubic meters
E	Echo motion category, either 1 or 2, a measure of the type of cloud
Rain	Rainfall following the action of seeding or not seeding in 10^7 cubic meters

Table 9.6 The Drug Cost Data

Variable	Description
COST	Average cost to plan for one prescription for one day, dollars
RXPM	Average number of prescriptions per member per year
GS	Percentage generic substitution used by the plan
RI	Restrictiveness index (0 = none, 100 = total)
COPAY	Average member copayment for prescriptions
AGE	Average member age
F	Percentage female members
MM	Member months, a measure of the size of the plan
ID	An identifier for the name of the plan

appropriate mean functions and summarize your results. (*Hint:* Be sure to check for influential observations and outliers.)

9.19 (Data file: drugcost) Health plans use many tools to try to control the cost of prescription medicines. For older drugs, generic substitutes that are equivalent to name-brand drugs are sometimes available at a lower cost. Another tool that may lower costs is restricting the drugs that physicians may prescribe. For example, if several similar drugs are available for treating the same symptoms, a health plan may require physicians to prescribe only a few of them. Since the usage of the chosen drug will be higher, the health plan may be able to negotiate a lower price for that drug.

The data described in Table 9.6, provided by Mark Siracuse, can be used to explore the effectiveness of these two strategies in controlling drug costs. The response variable is COST, the average cost of drugs per prescription per day. The data are from the mid-1990s, and are for 29

plans throughout the United States with pharmacies administered by a national insurance company.

Provide a complete analysis of these data, paying particular regard to possible outliers and influential cases. Summarize your results with regard to the importance of GS and RI. In particular, can we infer that more use of GS and RI will reduce drug costs?

Variable Selection

The methods suggested in the last few chapters can go a long way toward helping an analyst to build a useful regression model. Main-effects and interactions, Chapters 4 and 5, illustrate how to include information about qualitative predictors in a model. Transformations, Chapter 8, can help select useful scales for quantitative predictors. Regressors derived from basis functions, such as polynomials, Section 5.3, and splines, Section 5.4, further enhance our ability to model the effects of predictors on a response. The diagnostic methods of Chapters 9 can confirm that a model appears to match data.

Some problems have many potential predictors and responses. For example, a manufacturer studying the factors that impact the quality of its product may have many measures of quality, and possibly hundreds or even thousands of potential predictors of quality, including characteristics of the manufacturing process, training of employees, suppliers of raw materials, and many others. In a medical setting, to model the size of tumor, we might have predictors that describe the status of the patient, treatments given, and environmental factors thought to be relevant. In both of these settings, and in many others, we can have too many predictors.

The purpose of this chapter is to outline methods to select predictors, and the regressors derived from them, to use in a regression problem of interest. The methodology to be used depends on the goal of the regression analysis, and for this we distinguish three general cases:

- Many regression problems have as their primary goal assessing the effect of one, or at most a few, predictors on a response. In this case, including additional predictors beyond the ones of primary interest could be desirable for either interpretability of the results, or for increasing precision of tests and estimates. Including too many additional predictors could decrease precision.

- Interest could center on discovering the predictors that are associated with the response. The goal is to divide potential predictors into two sets, the *active* predictors and the *inactive* ones. This can be surprisingly difficult if predictors are related to each other.
- The goal of regression could be prediction of future values of a response given predictors. Including too many predictors can lead to relatively inaccurate predictions because a fitted model could account for the quirks of the observed data that are not present in future observations, while using a model that is too small can also lead to relatively inaccurate predictions if important predictors are missed.

The literature on the problem of model selection is enormous, concentrating mostly on the second and third goals. The fields of *machine learning* and to some extent *data mining* provide techniques for these problems. An introduction to these areas is given by Hastie et al. (2009). In this chapter we emphasize the first problem, briefly discuss the second problem, and mostly summarize general approaches to the prediction problem.

10.1 VARIABLE SELECTION AND PARAMETER ASSESSMENT

Suppose the primary goal of the analysis is to test the “effect” of a *focal* predictor on the response of interest, and the problem faced is selecting additional predictors to include when performing the test. As a reminder from Chapter 4, regression coefficients are defined as a characteristic of the conditional distribution of the response given the predictors. As long as assumptions of a linear regression model are satisfied for the set of predictors used, the test for the focal predictor will give an appropriate inference for that particular conditional distribution.

For simplicity in this discussion, we suppose that the predictor is represented by the regressor X_1 , and we consider only two regression models. The first is the simple regression model using only the focal regressor,

$$\begin{aligned} E(Y|X_1 = x_1) &= \beta_0 + \beta_1 x_1 \\ \text{Var}(Y|X_1) &= \sigma_1^2 \end{aligned} \tag{10.1}$$

The second model adds q additional regressors,

$$\begin{aligned} E(Y|X_1 = x_1, X_2 = \mathbf{x}_2) &= \gamma_0 + \gamma_1 x_1 + \boldsymbol{\gamma}'_2 \mathbf{x}_2 \\ \text{Var}(Y|X_1, X_2) &= \sigma_{12}^2 \end{aligned} \tag{10.2}$$

Adopting the assumption of linearly related regressors, Section 8.2.1, if the linear regression model is appropriate for (10.2), then it is also appropriate for (10.1). The usual notation in this book is to use the same name for the

parameters regardless of the mean function, but for this section we use β s in (10.1) and γ s in (10.2) to remind us that the meaning of a regression coefficient depends on all the regressors in a mean function. Similarly, we use different subscripts for σ^2 in the two models to remind us that the variance also depends on the regressors. β_1 measures the expected change in the response when X_1 is changed, while γ_1 is the expected change in the response when X_1 is changed and X_2 is held fixed at its current value. Thus, the tests of $\beta_1 = 0$ and of $\gamma_1 = 0$ test different hypotheses concerning the effect of X_1 . The further assumption is that the analyst would be willing to summarize the effect of X_1 relative to either of the candidate models.

Let $R_{Y,(X_1,X_2)}^2$ be the value of R^2 for the linear regression with response Y and regressors given by (X_1, X_2) and the intercept. Similarly, define $R_{Y,X_1|X_2}^2 = R_{Y,(X_1,X_2)}^2 - R_{Y,X_2}^2$ to be the increase in R^2 when X_1 is added to a regression model already using X_2 as regressors. This could be called a partial R^2 . If t_γ is the value of the t -test for testing the coefficient of $\gamma_1 = 0$, and t_β is the test for $\beta_1 = 0$, then if $t_\beta^2 \neq 0$, one can show that

$$t_\gamma^2 = t_\beta^2 \left(\frac{n-q-2}{n-2} \right) \left\{ \frac{R_{Y,X_1|X_2}^2}{R_{Y,X_1}^2 (1 - R_{Y,X_2|X_1}^2)} \right\} \quad (10.3)$$

The ratio involving sample size is a correction for df , so all else being equal, adding regressors will decrease the size of the test statistic. Recall from Section 4.1.5 that the quantity R_{X_1,X_2}^2 measures the collinearity between X_1 and the remaining regressors. Although this quantity does not appear in (10.3), all the other terms can be understood as a function of this one quantity.

If $R_{X_1,X_2}^2 \approx 0$, then the sum of squares explained by X_1 should be about the same if X_2 is included or excluded from the model. This means that $(R_{Y,X_1|X_2}^2 / R_{Y,X_1}^2) \approx 1$. We will also have in this case that $R_{Y,X_2|X_1}^2 \approx R_{Y,X_2}^2$, and so the term in curly brackets in (10.3) will be large if X_2 contains useful regressors. Unless q is so large that the first term dominates the result, including X_2 should increase the size of the test statistic.

The primary and possibly the only case with R_{X_1,X_2}^2 exactly equal to 0 occurs in designed experiments in which levels of factors are assigned to units at random using an orthogonal design. Additional covariates, if included, would then be at least approximately uncorrelated with the design predictors and the regressors derived from them, implying low collinearity. The results here confirm the usual practice in this case of fitting a model with factors, interactions, and covariates, and then using the residual mean square from a large model to test for the effect of focal predictors.

When R_{X_1,X_2}^2 is small, inference about a focal predictor from the larger model will be appropriate as long as the number of regressors q is not too large. Again, this seems to be standard practice in many areas, particularly in social sciences, in which selection of variables before testing appears to be the exception rather than the rule. If q is too large, then the benefit of adding variables could be outweighed by the correction for df in (10.3).

In the collinear case of $R_{X_1, X_2}^2 \approx 1$, the numerator of the term in curly brackets in (10.3) will be close to 0. If $R_{Y, X_2|X_1}^2$ is large as well, then (10.3) will behave like the ratio of two small numbers, and the test t_γ will be of little value because it will be very unstable. If $R_{Y, X_2|X_1}^2$ is not large, then $t_\gamma \approx 0$. Exactly how to proceed depends on the context of the problem. In the Minnesota water use example described in Section 4.1.5, the focal predictor is `year`, and X_2 consists of the remaining regressors including `log(muniPop)`. Water use was seen to increase with `year` if `log(muniPop)` is ignored, but after adjusting for `log(muniPop)`, the coefficient for `year` is negative with a small t -value. Because of the high collinearity between `year` and `log(muniPop)`, we expect the estimate for `year` adjusted for X_2 to be of little use. The importance of the focal predictor is therefore ambiguous and depends on whether or not adjustment is made for X_2 . Ambiguity is perhaps the correct inference for this problem.

This leaves the case of R_{X_1, X_2}^2 not close to 0 or to 1. Including additional regressors that are correlated with the response will generally increase t_γ^2 , while including regressors not correlated with the response will generally decrease t_γ^2 . Selection methods, such as the stepwise procedure illustrated in Section 10.2.2, can be useful in this circumstance if the analyst is uncertain about the appropriate conditional distribution to study to learn about the effect of the focal predictor.

If the focal predictor were a factor, interaction or other term that requires r regressors, the appropriate test will be an F -test rather than a t -test. The result corresponding to (10.3) is

$$F_\gamma = F_\beta \left(\frac{n-q-r-1}{n-r-1} \right) \left\{ \frac{R_{Y, X_1|X_2}^2}{R_{Y, X_1}^2 (1 - R_{Y, X_2|X_1}^2)} \right\} \quad (10.4)$$

which differs from (10.3) by a change in notation for the test statistic and a modification in the ratio of degrees of freedom.

Comparing t_β and t_γ should be based on their power (Section 6.4), not on the values of the statistics, but that would require additional notation and assumptions concerning the data. The discussion here can provide general guidelines on how to select predictors for study of a focal predictor.

10.2 VARIABLE SELECTION FOR DISCOVERY

The second use of variable selection is to discover which of a many predictors in a problem are active. Section 6.6.2 outlined an extreme example of this from Ioannidis (2005) where the goal is to find the few active genes that are associated with a particular trait from a pool of thousands of possible genes. The idea is that we have a pool of predictors in X , and seek to find a partition $X = (X_A, X_I)$, where X_A is the set of active regressors, and X_I is the set of inactive regressors, such that X_A is the smallest subset of the regressors,

subject to the marginality principle, such that $E(Y|X) = E(Y|X_A)$. In Ioannidis's example, X_A would consist of the regressors for active genes.

Unless the regressors are all uncorrelated, deciding on the regressors that should be considered active is not easy. For example, in a problem to find the active predictors that relate to a child's school achievement, the predictors $X_1 = \text{mother's years of education}$ and $X_2 = \text{father's years of education}$ are likely to be very highly correlated, and a method that determines one of these to be active and the other inactive based solely on a numeric criterion seems arbitrary. For this particular case, replacing X_1 and X_2 by $X_3 = (X_1 + X_2)/2$ and $X_4 = X_1 - X_2$ could solve the problem because both these predictors are on the same scale and their average and difference are meaningful. This can be generalized: before applying selection methods, combine predictors in sensible ways suggested by subject-matter considerations.

The approach to finding the active predictors we pursue here is to consider all possible choices for X_A , and then select the one that optimizes some selection criterion. To implement this method, we need to select a criterion function, and also to face the possibly daunting task of fitting hundreds or even thousands of models. Stepwise fitting using an information criterion like AIC, both to be defined later, is probably the most common computational compromise.

10.2.1 Information Criteria

Let X_c be a candidate subset of p_c of the regressors in X . We want to assess the candidate model

$$\begin{aligned} E(Y|X_c = x_{pc}) &= \beta_0 + \boldsymbol{\beta}'_{pc} \mathbf{x}_{pc} \\ \text{Var}(Y|X_c) &= \sigma^2 \end{aligned} \tag{10.5}$$

for different candidate sets X_c . Using (10.5) will be a reasonable mean function for the regression problem if the methods of earlier chapters in this book suggest that $E(Y|X)$ has a linear mean function with constant variance, and the assumption of linearly related predictors described in Section 8.2.1 is sensible.

Criteria for comparing various candidate subsets are based on the lack of fit of a model and its *complexity*. Lack of fit for a candidate subset X_c is measured by its residual sum of squares RSS_{pc} . Complexity for multiple linear regression models is measured by the number of regressors p_c in X_c , including the intercept.¹ The most common criterion that is useful in multiple linear regression and many other problems where model selection is at issue is the *Akaike Information Criterion*, or AIC. Ignoring constants that are the same for

¹The complexity may also be defined as the number of parameters estimated in the regression as a whole, which is equal to the number of regressors plus 1 for estimating σ^2 .

every candidate subset, AIC is given for linear regression by Sakamoto et al. (1986),

$$\text{AIC} = n \log(\text{RSS}_{pc}/n) + 2p_c \quad (10.6)$$

Small values of AIC are preferred, so better candidate sets will have smaller RSS and a smaller number of terms p_c . An alternative to AIC is the *Bayes Information Criterion*, or BIC, given by Schwarz (1978),

$$\text{BIC} = n \log(\text{RSS}_{pc}/n) + \log(n)p_c \quad (10.7)$$

which provides a different balance between lack of fit and complexity. Once again, smaller values are preferred.

As the sample size n increases for fixed p_c , the lack-of-fit term in AIC increases with n , while the complexity term stays constant. The BIC criterion, however, pays more attention to sample size as the complexity term increases with n , although at a slower rate than the lack-of-fit term. If there really exists a partition of the regressors into active and inactive regressors, then as n increases, BIC will select X_A with probability approaching 1 (Nishii, 1984). In many problems, the linear model is only an approximation to the real data-generating process, and there may be no X_A . In this case, for a large enough sample, AIC will perform better (Yang, 2005). Although these large sample results do not guarantee much in a finite sample, both AIC and BIC, or modifications of them with other measures of complexity, are often used in practice to select regressors.

10.2.2 Stepwise Regression

There are potentially 2^p possible choices of X_A obtained from all possible subsets of the regressors.² If $p = 5$, there are only $2^5 = 32$ choices for X_A , and all 32 possible can be fit and compared. If $p = 10$, there are 1024 choices, and fitting such a large number of models is possible but still an unpleasant prospect.

For $p \leq 30$, the *leaps and bounds* algorithm (Furnival et al., 1974) can be used to find the few candidate models that minimize AIC or BIC without actually computing all possible models. The algorithm has been implemented in statistical packages and in subroutine libraries (Orestes Cerdeira et al., 2012; Rogue Wave Software, 2013). This algorithm doesn't work well with predictors that are represented with several regressors like factors, interactions, and polynomials, and so it is not frequently used in practice. Stepwise methods are not

²There are fewer possible models if the regressors represent factors and interactions that are treated as a group. In addition, with interactions, permissible models should obey the marginality principle, again decreasing the number of possible models.

guaranteed to find the candidate subset that is optimal according to any criterion function, but they often give useful results.

Stepwise methods have three basic variations. For this section, we will define a *term* to all the regressors that represent a factor or an interaction, or a single regressor that represents a predictor or its transformation. Suppose AIC were the criterion function of interest. *Forward selection* starts with a current subset consisting of only the intercept and any regressors to be included in all models.

[FS] Consider all candidate subsets consisting of one additional term beyond the current subset, such that the models considered do not violate the marginality principle from Section 6.2, so an interaction is never added unless all the lower order effects in the interaction are already included. Compute AIC for each of these models. If the AIC for all the candidate models exceeds the AIC for the current model, stop and accept the current model. Otherwise, accept the subset model that minimizes AIC as the current model. If more regressors are available for fitting, repeat this step; otherwise, stop.

If the number of terms beyond the intercept is k , this algorithm will consider at most $k + (k - 1) + \dots + 1 = k + (k + 1)/2$ of the 2^k possible subsets. For $k = 10$, the number of subsets considered is 45 of the 1024 possible subsets.

Backward elimination works in the opposite direction. Set the model with all terms to be the current model.

[BE] Consider candidate models that differ from the current model by the deletion of one term, subject to the marginality principle, and compute AIC for each. Accept the current model if the AIC for all the candidate models exceeds the AIC for the current model; otherwise, set the current model to the candidate with the minimum AIC. If no regressors remain in the current model, stop; otherwise, repeat this step.

As with the forward selection method, only $k(k + 1)/2$ subsets are considered. The subsets considered by forward selection and by backward elimination may not be the same.

The forward and backward algorithms can be combined into a *stepwise* method, where at step subsets are considered that either add a term or delete a term. Start with a candidate model as with either the forward or backward algorithm.

[SW] The candidate models consist of all subsets obtained from the current subset by either adding or deleting a term, subject to the marginality principle. Accept as the new candidate model the subset with the smallest AIC. If the new candidate was the same as at the last step, stop; otherwise, repeat this step.

Highway Accidents

We will use the highway accident data described in Section 8.2 concerning frequency of accident on segments of highways in Minnesota, modeled by characteristics of the segment. The response is $\log(\text{rate})$, the log of the

accidents per million vehicle miles of travel on the segment. The regressors include the four log-transformed predictors found in Section 8.2.2 as well as `lane`, the number of lanes; `slim`, the speed limit; `shld`, the width of the roadway shoulder in feet; `lwid`, the width of a driving lane in feet; `acpt`, the number of access points per mile; and `itg`, the number of freeway-type interchanges per mile. Finally, a factor `htype` with 4 levels is included, indicating the type of funding that supports the highway. The various funding types have different design requirements, and so the levels of `htype` could be associated with the response. The number of segments is $n = 39$.

Suppose `a` were the number of accidents in a segment, `v` were the number of vehicles that used the segment in the period, in millions, and `len` is the length of the segment in miles. Then the accident rate is $\text{rate} = \text{a}/(\text{v} \times \text{len})$. Accidents generally occur at a few relatively rare “bad spots” on a highway, and increasing the length of a segment by a small amount is unlikely to add a “bad spot.” Thus, increasing `len` will decrease `rate` because the other two components that go into computing `rate` should be nearly constant. Consequently, the response and $\log(\text{len})$ should be negatively correlated, and we should consider only models that include $\log(\text{len})$.³

Suppose the goal of the analysis were to discover the set of active regressors. The regressors in the highway data exhibit moderate collinearity, with the values of $R^2_{X_1, X_2}$ varying from 0.49 for $X_1 = \log(\text{trucks})$ to 0.97 for $X_1 = \log(\text{adt})$, and this suggests that the regressors contain redundant information and that selecting a subset of regressors may be helpful.

We illustrate using forward selection based on AIC. The first step considers all choices for the active regressors X_A that include $\log(\text{len})$ and one additional term. There are 9 such choices, and each of these is summarized in Table 10.1. The first column of the table gives the name of the term added. The column marked *df* is the number of regressors in the added term, which is 1 whenever a single regressor is added and 3 for adding the factor `htype`. The row labeled [none] corresponds to using only $\log(\text{len})$ and no additional regressors. The `RSS` is the residual sum of squares for the model including $\log(\text{len})$ and the additional term. The remaining columns give the values of AIC and BIC. The rows are ordered by the value of AIC, and all the rows with AIC less than the AIC for [none] would be an improvement over the current model using AIC as the criterion. The model listed first, adding `slim`, gives the smallest value of AIC so this is the model that would be accepted at this step. If [none] had the smallest AIC, selection would have stopped before adding any regressors.

The next step starts with the model `log(rate) ~ log(len) + slim`. Again consider all models obtained by adding a term. This is continued

³A possible alternative here would be to use as a response the logarithm of the number of accidents per million vehicles $\log(y) = \log(a/v)$, but this is likely to change the negative relationship between the response and $\log(\text{len})$ to a positive relationship. Can you see why this is likely to be true?

Table 10.1 First Step in Forward Stepwise Regression

Add ...	<i>df</i>	RSS	AIC	BIC
+slim	1	2.94	-94.87	-89.88
+acpt	1	3.38	-89.36	-84.36
+shld	1	3.78	-85.05	-80.06
+log(sig1)	1	4.52	-78.03	-73.03
+htype	3	4.14	-77.44	-69.12
+log(trks)	1	4.76	-76.06	-71.07
+log(adt)	1	5.06	-73.68	-68.69
[none]		5.48	-72.51	-69.18
+lane	1	5.22	-72.42	-67.43
+itg	1	5.27	-72.08	-67.09
+lwid	1	5.30	-71.85	-66.86

Table 10.2 Forward Stepwise Fit for the Highway Data

	Estimate	Std. Error	<i>t</i> Value	Pr(> <i>t</i>)
(Intercept)	4.1665	0.7411	5.62	0.0000
log(len)	-0.2357	0.0849	-2.78	0.0089
slim	-0.0319	0.0103	-3.10	0.0038
acpt	0.0110	0.0067	1.65	0.1081
log(trks)	-0.3290	0.2135	-1.54	0.1325

$\hat{\sigma} = 0.2698$ with 34 *df*, $R^2 = 0.6961$.

until adding a term would increase AIC. For these data, the following models are fit:

Step	Model	AIC
2	log(rate) ~ log(len)	-72.51
3	log(rate) ~ log(len) + slim	-94.87
4	log(rate) ~ log(len) + slim + acpt	-96.89
5	log(rate) ~ log(len) + slim + acpt + log(trucks)	-97.53

At this point, selection stops because adding more regressors increases AIC. Although coefficient estimation and tests are not the primary concerns in variable discovery, the fitted model from the forward stepwise procedure is summarized in Table 10.2.

A curious feature of this model is that the coefficient for `slim` is negative. This nominally implies that higher speed limits are associated with *fewer* accidents. While inferring causation here might please those who like to drive fast, it could well be that highway officials lower speed limits on roads with high accident rates, so `slim` could be caused by the response. The significance levels reported for `slim` and for the other regressors in Table 10.2 are not to be

trusted, as the t -values may not follow t -distributions even approximately after subset selection.

These data were originally collected to study the effects of design variables on accident rates. We now change our emphasis and assume that the goal is to learn about the effect of the focal predictor `shld` on accident rate, as in Section 10.1. Four selection methods were used, requiring that `log(len)` and `shld` are included in all models, with the following results:

Method	$\hat{\beta}_{\text{shld}}$	t_{shld}	AIC
None	-0.070	-4.034	-85.05
Forward	-0.045	-2.687	-96.72
Backward	0.007	0.284	-101.41
All	0.003	0.087	-94.20

The row marked “None” includes only `log(len)` and `shld` as regressors, and the row marked “All” includes all the regressors. The rows “Forward” and “Backward” are the models selected by forward selection and backward elimination, respectively, considering only models including `log(len)` and `shld` using AIC as the stopping criterion. Given in the table are the coefficient estimate and t -value for `shld`, along with the value of AIC for the different models. Both the model “None” and the model from forward selection suggest that `shld` has a large negative effect; since the response is in log scale, increasing `shld` by one foot is associated with a decrease in `rate` of about 7% ignoring all other terms except for `log(length)` or about 4% conditioning on the additional variables in the model selected by forward selection, $\text{log}(\text{rate}) \sim \text{log}(\text{len}) + \text{shld} + \text{acpt} + \text{log}(\text{sigs1})$. The model selected by backward elimination, $\text{log}(\text{rate}) \sim \text{log}(\text{len}) + \text{shld} + \text{log}(\text{adt}) + \text{slim} + \text{log}(\text{sigs1}) + \text{htype}$, has more regressors and has a much smaller value of AIC. Conditioning on these regressors, the effect of `shld` is now slightly positive, although the corresponding t -value is very small. The “size” of the `shld` effect depends on the other regressors in the mean function, just as its interpretation depends on the other regressors. The inclusion, or not, of the factor `htype`, the classification of the highway, changes the inference about the focal predictor `shld`. The interstate highways, `htype = fai`, all have `shld = 10 ft`, while the other types generally have lower values of `shld`. If adjustment is made for `htype`, `shld` has little remaining variability, and so it is not related to the response.

The model selected by forward selection forcing `shld` into all subsets has somewhat higher AIC than the model selected without forcing `shld` into all models, as should be expected. Using backward elimination in this example found a model including `shld` with a smaller AIC than the forward model that did not force `shld` into all models. There is no requirement that forward selection and backward elimination finish with the same model, and neither is guaranteed to find the model with the smallest AIC. If the focus is on `shld`,

then finding an optimal model is of little consequence in any case, since the main finding is that either wider shoulders, or wider shoulders with the extra design and maintenance features that define the levels of `htype`, are associated with lower accident rates in these data.

10.2.3 Regularized Methods

A different approach to discovering relevant variables starts from an assumption of *sparsity*, that only a small number of predictors are required to model a response. The justification for this belief is interesting: “*Use a procedure that does well in sparse problems, since no procedure does well in dense problems*” [italics in the original] (Hastie et al., 2009, p. 611). The problem cited by Ioannidis (2005) in Section 6.6.2 is an example of an application of the sparcity principle, in which it is thought that only a few of the many thousands of available genes are active in determining a mutation of interest. This assumption could be exactly true, it could be true enough, meaning that while many genes are active, knowledge of only a few are required for building therapeutic methods, or it could simply reflect reliance on a simplified understanding of the mechanism that causes the mutation that may be useful but inaccurate.

Using an information criterion like `AIC` or `BIC` does not incorporate sparcity directly into the procedure. The *lasso* (Tibshirani, 1996) is typical of methods that use sparsity directly. Start with the usual linear regression model assuming $E(Y|X = \mathbf{x}) = \beta_0 + \boldsymbol{\beta}'\mathbf{x}$ and $\text{Var}(Y|X) = \sigma^2$. The goal is to obtain an estimate of $\boldsymbol{\beta}$ that has most of its elements equal to 0. Regressors with nonzero estimates are selected for the active regressors. The lasso estimate minimizes

$$\hat{\boldsymbol{\beta}}_{\lambda}^{lasso} = \arg \min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \sum (y_i - \beta_0 - \boldsymbol{\beta}'\mathbf{x}_i)^2 + \lambda \sum |\beta_j| \right\} \quad (10.8)$$

This is just the usual least squares criterion plus a penalty given by a tuning parameter λ times the sum of the absolute values of the coefficients. If $\lambda = 0$ the lasso is the same as OLS; as λ increases, the second term can swamp the first, and for large enough λ , all the estimates will equal 0. Modifications to the lasso, such as the *elastic net* (Zou and Hastie, 2005) and *scad* (Fan and Li, 2001), use different penalties to get estimates with better properties. As a class, these are regularized estimation methods.

Study of methods that use sparseness is an active area of research, and there is no consensus on applied methodology. There are, however, a few important conceptual limitations. The methods generally depend on the scaling of the regressors. If one of the regressors is divided by a constant c , then its corresponding regression coefficient is multiplied by c (see Problem 2.9), and so a penalty that depends on the size of the coefficients will change. We have also seen in Chapters 4–5 that the meaning and values of coefficients change depending on the other regressors included in `mean` function. The penalized methods will give different results if, for example, the baseline level of a factor

Table 10.3 Results of a Simulated Example

Method	Number of Regressors	R^2	p -Value of Overall F	Coefs. with p -Value <.05
No selection	50	0.497	0.5442	2
Forward selection	19	0.427	0.0002	8
Backward elimination	12	0.352	0.0002	9

is changed. If the data are a random sample from a population, and no factors and interactions are included, then prescaling to correlation scale can produce sensible results, but with either nonrandom sampling or factors and interactions models fit to a particular data set may not apply when used for prediction of future values.

10.2.4 Subset Selection Overstates Significance

All selection methods can overstate significance. Consider a simulated example. A data set of $n = 100$ cases with a response Y and $k = 50$ regressors $X = (X_1, \dots, X_{50})$ was generated using standard normal random deviates, so there are no active regressors, and the unobservable population multiple correlation between Y and X is 0. The sample multiple correlation R^2 for the regression of Y on X in this single simulation was 0.497. This may seem surprisingly large, considering that all the data are independent random numbers. The overall F -test, which is in a scale more easily calibrated, gives a p -value of 0.544 for the data; Rencher and Pun (1980) and Freedman (1983) report similar simulations with the overall p -value varying from near 0 to near 1, as it should since the null hypothesis of $\beta = \mathbf{0}$ is true.

Table 10.3 reports the summary of this regression and the final model using both forward selection and backward elimination. As expected, the value of R^2 for the subset models is less than the R^2 for the model with all the regressors. Perhaps unexpectedly, the significance level for the overall F is now tiny for both selection algorithms. In addition, many of the coefficients for the selected regressors have significance levels less than 0.05. Tests from subset models cannot be trusted. See Hurvich and Tsai (1990) for more discussion.

10.3 MODEL SELECTION FOR PREDICTION

The basic idea in prediction is to use observable values of predictors for a new subject or case to predict the value of an interesting response. Here are a few examples.

Epworth Sleepiness Scale

A primary result of a sleep disorder in humans is daytime sleepiness. To diagnose the existence of a disorder, patients fill out a questionnaire called the Epworth Sleepiness Scale or ESS. The ESS consists of 8 standard questions.

One of the questions is: “How likely are you to doze while sitting and reading?” The answer x_j to the j th question is an integer between 0 for “no chance of dozing” and 3 for “high chance of dozing.” The computed scale is the sum of the scores, $\sum \beta_j x_j$, with all the $\beta_j = 1$, so this is in the form of a linear regression mean function. Johns (1991) reported that the ESS can distinguish between normal subjects and subjects with sleep disorders, with high values of the scale corresponding to the sleep disorder group.

Prediction functions that are simply a sum of scores are very common. Other examples include most tests given for certification or licensing for some particular skill, where a minimum number correct is required to predict that a candidate is proficient. Simple prediction functions like the ESS are important because they are easily used and explained. The ESS has been validated with data, but it was not based on fitting a model because an objective measure of sleepiness is not available for a construction set of subjects. Were such data available, then estimating a prediction function could give better predictions. Replacing an easily administered method like the ESS by a more complicated but possibly more precise prediction method may or may not improve clinical outcomes.

Credit Scoring

A credit score is a measure of a person’s credit worthiness, with higher numbers indicating that the person is more likely to repay a loan than is a person with a lower credit score. Credit scores are generally predictions from models fit to data on other people for whom outcomes of loans and predictors such as the person’s characteristics are known. In the United States, many private companies compute credit scores, with the FICO score, sold by the Fair Isaac Corporation, the most prominent of these.

Because credit scoring affects so many people, both the predictors and the regression coefficients or weights for the predictors can be very contentious. The FICO score, for example, is explained on a website (Fair Isaac Corporation, 2013) as a linear combination of components, much like a linear regression equation. Many details are hidden, however, and exactly which regressors are used to represent “payment history,” for example, are not included on the website. The FICO score is based on fitting models to data and applying the model to future individuals.

The Epworth Sleepiness Scale is a straightforward equation used for prediction. The FICO score is explained on its website as if it were also based on a simple equation, although the actual proprietary prediction method is probably more complicated.

Weather Forecasting

Weather forecasting can provide an example in which the prediction method is not simply explained. For example, the University of Washington Probability Forecast (University of Washington Applied Physics Laboratory, 2013) provides real-time weather forecasts by averaging predictions from a variety of

sources. Each source may use different methodology, different predictors, and have different accuracy for a particular location. The predictions are combined using a weighted average to give overall predictions and the uncertainty in the predictions. Averaging many models will often give better predictions than will using any one model, at the cost of greatly increased complexity. See Fraley et al. (2011) for application to weather forecasting, Hoeting et al. (1999) for a tutorial on Bayesian model averaging in general, or Yang (2001) for other approaches.

The general problem of formulating predictions from training data has spawned a new field of *machine learning* with a huge literature of its own. Dozens of methods, with wonderful names like *neural networks* and *random forests*, and features for these methods like *boosting* and *bagging*, make this an exciting area of study. Hastie et al. (2009) provides a readable introduction. The linear regression methodology in this book will generally produce prediction methods that are worse than the newer machine learning methods, but the improvements obtained by the more complex methods are often small.

10.3.1 Cross-Validation

We conclude with an example using *cross-validation*, a general method that can be used to judge how well a procedure will predict with future data sampled from the same population or data-generating mechanism that produced the current data. The idea is to divide the available data into two parts at random, a *construction set* and a *validation set*. The construction set is used to obtain a model for prediction. The fitted model is then applied to the validation set, and prediction errors, observed minus fit, are computed. These are then summarized, typically by the SD of the prediction errors. It is usual to divide to use between 50% and 75% of the data for the construction set and the remainder for the validation set.

10.3.2 Professor Ratings

Suppose we were interested in modeling professor's quality rating as a function of the characteristics of the professor and some characteristics of the student raters, using the regressors used in Problem 6.10. Because collinearity is low in this problem, the methods in this book should be adequate for the purposes of either selecting a set of active predictors or for obtaining a prediction equation. The numeric predictors, `raterInterest`, `easiness`, and `numYears`, are used without transformation.⁴

The data were divided at random into a set of 250 observations for a construction set and the remaining 116 for a validation set. Several methods listed in Table 10.4 were fit to the data in the construction set. As suggested in Section

⁴The response is bounded between 1 and 5, and so models we fit could give predictions outside that range. This could suggest rescaling the response, an option not pursued here.

Table 10.4 SD of Prediction Errors for Several Methods of Fitting the Professor Ratings Data

Method	Construction	Validation
No regressors	0.824	0.863
First-order	0.565	0.640
Second-order	0.541	0.661
Stepwise	0.554	0.635
Random forest	0.341	0.631
Lasso	0.588	0.645
Elastic net	0.593	0.651
Average	0.514	0.623

7.1.1, we used `numRaters`, the number of raters averaged to get `quality`, as weights. Predictions were obtained for both the construction and validation set with the resulting SDs reported in the table. The row in the table marked “No regressors” estimates `quality` by the weighted mean `quality` in the construction set. The weighted mean in the validation set is somewhat different and the SD is a little larger in the validation set. The first-order model uses all the regressors, and the second-order model uses all the main effects and two-factor interactions. The second-order model has somewhat smaller prediction SD on the construction set, as it must because bigger models must result in smaller residual error, but it has higher prediction SD than the first-order model. Similarly, the value of $AIC = 511.8$ for the second-order model is considerably larger than the value $AIC = 491.1$ for the first-order model.

The stepwise method started with the second-order model using AIC to remove/add terms. The selected model has $AIC = 486.5$, but it also has prediction SD on the validation set that is larger than the prediction SD for the first-order model. The next three lines of the table refer to methods that are not described in this book. The first uses a random forest (Hastie et al., 2009, chapter 15) to get predictions, and the remaining two lines summarize the lasso and a version of the elastic net. The random forest method fits the observed data exceedingly well and does about as well as the first-order model on the validation set. Because of low collinearity, the lasso and the elastic net perform about the same in this example. The last line of the table used predictions obtained by averaging the predictions from the other 6 methods excluding the no regressors case. The average does the best, although in this example, none of the methods differ by much.

10.4 PROBLEMS

- 10.1** Suppose the regressors in a problem are divided into the focal predictor X_1 and the remaining regressors collected into X_2 . For the purpose of estimating the effect of X_1 , collinearity is a problem if $R^2_{X_1, X_2}$ is large.

Many computer programs allow computing a quantity called the *variance inflation factor* given by $1/(1 - R_{X_1, X_2}^2)$, from which R_{X_1, X_2}^2 can be easily computed.

10.1.1 (Data file: MinnWater) In the Minnesota water use data, suppose the response is `log(irrUse)`, the logarithm of the amount of water used in irrigation of crops, and the regressors are `agPrecipitation`, `Year`, and `log(statePop)`. Compute R_{X_1, X_2}^2 , selecting each of the three variables in turn as the focal regressor, and summarize your findings.

10.1.2 (Data file: UN11) With the United Nations data, use `lifeExpF` as the response, and `log(ppgdp)`, `fertility`, and `pctUrban` as regressors, compute the collinearity measure assuming each of these three is the focal predictor in turn, and summarize results.

10.2 (Data file: Highway)

10.2.1 For the highway accident data, use your software to verify the forward selection and backward elimination subsets that are given in Section 10.2.2.

10.2.2 Use as response `log(rate * len)` and treat `lwid` as the focal regressor. Use both forward selection and backward elimination to assess the importance of `lwid`. Summarize your results.

10.2.3 Using the identity $\log(\text{rate} \times \text{len}) = \log(\text{rate}) + \log(\text{len})$, we can write

$$E(\log(\text{rate} \times \text{len}) | X = \mathbf{x}) = \beta_0 + \boldsymbol{\beta}' \mathbf{x}$$

$$E(\log(\text{rate}) + \log(\text{len}) | X = \mathbf{x}) = \beta_0 + \boldsymbol{\beta}' \mathbf{x}$$

$$E(\log(\text{rate}) | X = \mathbf{x}) = \beta_0 + \boldsymbol{\beta}' \mathbf{x} - \log(\text{len})$$

In this last equation, the variable $\log(\text{len})$ is on the right side of the equation with an implied known regression coefficient equal to -1 . A regressor with a known regression coefficient is called an *offset*, and most modern regression software allows you to include offsets in fitting a model. The difference between an offset and a regressor is that no coefficient will be estimated for the offset.

Repeat Problem 10.2.2, but use `log(rate)` as the response and $-\log(\text{len})$ as an offset. Is the analysis the same or different? Explain.

10.3 (Data file: mantel) Using these “data” with a response Y and three regressors X_1 , X_2 , and X_3 from Mantel (1970), apply the forward selection and backward elimination algorithms, using AIC as a criterion function. Also, find AIC and BIC for all possible models and compare results. Which appear to be the active regressors?

Table 10.5 Oxygen Update Experiment

Variable	Description
Day	Day number
BOD	Biological oxygen demand
TKN	Total Kjeldahl nitrogen
TS	Total solids
TVS	Total volatile solids
COD	Chemical oxygen demand
O2UP	Oxygen uptake

10.4 (Data file: `BGSboys`) For the boys in the Berkeley Guidance Study in Problem 3.3, find a model for `HT18` as a function of the other variables for ages 9 and earlier. Perform a complete analysis, including selection of transformations and diagnostic analysis, and summarize your results.

10.5 (Data file: `dwaste`) An experiment was conducted to study `O2UP`, oxygen uptake in milligrams of oxygen per minute, given five chemical measurements shown in Table 10.5 (Moore, 1975). The data were collected on samples of dairy wastes kept in suspension in water in a laboratory for 220 days. All observations were on the same sample over time. We desire an equation relating $\log(O2UP)$ to the other variables. The goal is to find variables that should be further studied with the eventual goal of developing a prediction equation; `Day` cannot be used as a predictor.

Complete the analysis of these data, including a complete diagnostic analysis. What diagnostic indicates the need for transforming `O2UP` to a logarithmic scale?

10.6 Galápagos Islands (Data file: `galapagos`) The Galápagos Islands off the coast of Ecuador provide an excellent laboratory for studying the factors that influence the development and survival of different species. Johnson and Raven (1973) have presented data in the file `galapagos`, giving the number of species and related variables for 29 different islands (Table 10.6). Counts are given for both the total number of species and

Table 10.6 Galápagos Island Data

Variable	Description
Island	Island name
NS	Number of species
ES	Number of endemic species (occurs only on that island)
Area	Surface area of island, hectares
Anear	Area of closest island, hectares
Dist	Distance to closest island, kilometers
DistSC	Distance from Santa Cruz Island, kilometers
Elevation	Elevation in m, missing values given as zero
EM	1 if elevation is observed, 0 if missing

the number of species that occur only on that one island (the endemic species).

Use these data to find factors that influence diversity, as measured by some function of the number of species and the number of endemic species, and summarize your results. One complicating factor is that elevation is not recorded for six very small islands, so some provision must be made for this. Four possibilities are (1) find the elevations; (2) delete these six islands from the data; (3) ignore elevation as a predictor of diversity, or (4) substitute a plausible value for the missing data. Examination of large-scale maps suggests that none of these elevations exceed 200 m.

Nonlinear Regression

A regression mean function cannot always be written as a linear combination of the regressors. For example, the mean function

$$E(Y|X = x) = \theta_1 + \theta_2 [1 - \exp(-\theta_3 x)] \quad (11.1)$$

was suggested for the turkey diet supplement experiment described in Section 1.1, where Y was three-week weight gain from baseline and X the amount of supplement added to the turkey diet. This mean function has three parameters, θ_1 , θ_2 , and θ_3 , but only one regressor, X . The mean function is a nonlinear mean function because it is not a linear combination of the parameters. In (11.1), θ_2 multiplies $[1 - \exp(-\theta_3 x)]$, and θ_3 enters through the exponent.

Another nonlinear mean function we have already seen was used in estimating transformations of predictors to achieve linearity, given by

$$E(Y|X = x) = \beta_0 + \beta_1 \psi_s(x, \lambda) \quad (11.2)$$

where $\psi_s(x, \lambda)$ is the scaled power transformation defined by (8.3). This is a nonlinear model because the slope parameter β_1 multiplies $\psi_s(x, \lambda)$, which depends on the parameter λ . The transformation parameter λ was estimated visually in Chapter 8, and then the β s are estimated from the linear model assuming λ is fixed at its estimated value. If we estimate all three parameters simultaneously, then the mean function is nonlinear.

The parameters of the nonlinear mean function often have a useful interpretation. In the turkey growth example, when $X = 0$, $E(Y|X = 0) = \theta_1$, so θ_1 is the expected weight gain with no supplementation. Assuming $\theta_3 > 0$, as X increases, $E(Y|X = x)$ will approach $\theta_1 + \theta_2$, so the sum of the first two parameters is the maximum growth possible for any dose called an *asymptote*, and θ_2 is the maximum additional growth due to supplementation. The final

parameter θ_3 is a rate parameter; for larger values of θ_3 , the expected growth approaches its maximum more quickly than it would if θ_3 were smaller.

11.1 ESTIMATION FOR NONLINEAR MEAN FUNCTIONS

Here is the general setup for nonlinear regression. We have a set of p regressors X , and a vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$ of parameters such that the mean function relating the response Y to X is given by

$$E(Y|X = \mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}) \quad (11.3)$$

We call the function m a *kernel mean function*. The two examples of m we have seen so far in this chapter are in (11.1) and (11.2), but there are of course many other choices, both simpler and more complex. The linear kernel mean function, $m(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}'\boldsymbol{\theta}$ is a special case of the nonlinear kernel mean function. Many nonlinear mean functions impose restrictions on the parameters, like $\theta_3 > 0$ in (11.1).

As with linear models, we also need to specify the variance function, and for this we will use the same structure as for the linear model and assume

$$\text{Var}(Y|X = \mathbf{x}_i) = \sigma^2 / w_i \quad (11.4)$$

where, as before, the w_i are known, positive weights, and σ^2 is an unknown positive number. Equations (11.3) and (11.4) together with the assumption that observations are independent of each other define the nonlinear regression model. The only difference between the nonlinear regression model and the linear regression model is the form of the mean function, and so we should expect that there will be many parallels that can be exploited.

The data consists of observations (\mathbf{x}_i, y_i) , $i = 1, \dots, n$. Because we have retained the assumption that observations are independent and that the variance function (11.4) is known apart from the scale factor σ^2 , we can use least squares to estimate the unknown parameters, so we need to minimize over all permitted values of $\boldsymbol{\theta}$ the residual sum of squares function,

$$\text{RSS}(\boldsymbol{\theta}) = \sum_{i=1}^n w_i (y_i - m(\mathbf{x}_i, \boldsymbol{\theta}))^2 \quad (11.5)$$

We have OLS if all the weights are equal and WLS if they are not all equal.

The solution $\hat{\boldsymbol{\theta}}$ that minimizes (11.5) for linear models is available at (A.21) in Appendix A.7. For nonlinear regression, there generally is no formula, and minimization of (11.5) is a numerical problem. We present some theory now that will approximate (11.5) at each iteration of a computing algorithm by a nearby linear regression problem. Not only will this give one of the standard

computing algorithms used for nonlinear regression but will also provide expressions for approximate standard errors and point out how to do approximate tests. The derivation uses some calculus.

We begin with a brief refresher on approximating a function using a Taylor series expansion.¹ In the scalar version, suppose we have a function $g(\beta)$, where β is a scalar. We want to approximate $g(\beta)$ for values of β close to some fixed value β^* . The Taylor series approximation is

$$g(\beta) = g(\beta^*) + (\beta - \beta^*) \frac{dg(\beta)}{d\beta} + \frac{1}{2}(\beta - \beta^*)^2 \frac{d^2g(\beta)}{d\beta^2} + \text{Remainder} \quad (11.6)$$

All the derivatives in Equation (11.6) are evaluated at β^* , and so the Taylor series approximates $g(\beta)$, the function on the left side of (11.6) using the polynomial in β on the right side of (11.6). We have only shown a two-term Taylor expansion and have collected all the higher-order terms into the remainder. By taking enough terms in the Taylor expansion, any smooth function g can be approximated as closely as wanted. In most statistical applications, only one or two terms of the Taylor series are needed to get an adequate approximation. Indeed, in the application of the Taylor expansion here, we will mostly use a one-term expansion that includes the quadratic term in the remainder.

When $g(\theta)$ is a function of a vector valued parameter θ , the two-term Taylor series is very similar,

$$g(\theta) = g(\theta^*) + (\theta - \theta^*)' \mathbf{u}(\theta^*) + \frac{1}{2}(\theta - \theta^*)' \mathbf{H}(\theta^*)(\theta - \theta^*) + \text{Remainder} \quad (11.7)$$

where we have defined two new quantities in (11.7), the *score vector* $\mathbf{u}(\theta^*)$, and the *Hessian matrix* $\mathbf{H}(\theta^*)$. If θ^* has k elements, then $\mathbf{u}(\theta^*)$ also has k elements, and its j th element is given by $\partial g(\mathbf{x}, \theta)/\partial \theta_j$, evaluated at $\theta = \theta^*$. The Hessian is a $k \times k$ symmetric matrix whose (ℓ, j) element is the partial second derivative $\partial^2 g(\mathbf{x}, \theta)/(\partial \theta_\ell \partial \theta_j)$, evaluated at $\theta = \theta^*$.

We return to the problem of minimizing (11.5). Suppose we have a current guess θ^* of the value of θ that will minimize (11.5). The general idea is to approximate $m(\theta, \mathbf{x}_i)$ using a Taylor approximation around θ^* . Using a one-term Taylor series, ignoring the term with the Hessian in (11.7), we get

$$m(\theta, \mathbf{x}_i) \approx m(\theta^*, \mathbf{x}_i) + \mathbf{u}_i(\theta^*)'(\theta - \theta^*) \quad (11.8)$$

We have put the subscript i on the \mathbf{u} because the value of the derivatives can be different for every value of \mathbf{x}_i . The $\mathbf{u}_i(\theta^*)$ play the same role as the regressors in the multiple linear regression model. There are as many elements

¹Jerzy Neyman (1894–1981), one of the major figures in the development of statistics in the twentieth century, often said that arithmetic had five basic operations: addition, subtraction, multiplication, division, and Taylor series.

of $\mathbf{u}_i(\boldsymbol{\theta})$ as parameters in the mean function. The difference between nonlinear and linear models is that the $\mathbf{u}_i(\boldsymbol{\theta}^*)$ may depend on unknown parameters, while in multiple linear regression, the regressors depend only on the predictors.

Substitute the approximation (11.8) into (11.5) and simplify to get

$$\begin{aligned}\text{RSS}(\boldsymbol{\theta}) &= \sum_{i=1}^n w_i [y_i - m(\boldsymbol{\theta}, \mathbf{x}_i)]^2 \\ &\approx \sum_{i=1}^n w_i [y_i - m(\boldsymbol{\theta}^*, \mathbf{x}_i) - \mathbf{u}_i(\boldsymbol{\theta}^*)'(\boldsymbol{\theta} - \boldsymbol{\theta}^*)]^2 \\ &= \sum_{i=1}^n w_i [\hat{e}_i^* - \mathbf{u}_i(\boldsymbol{\theta}^*)'(\boldsymbol{\theta} - \boldsymbol{\theta}^*)]^2\end{aligned}\quad (11.9)$$

where $\hat{e}_i^* = y_i - m(\boldsymbol{\theta}^*, \mathbf{x}_i)$ is the *i*th *working residual* that depends on the current guess $\boldsymbol{\theta}^*$. The approximate $\text{RSS}(\boldsymbol{\theta})$ is now in the same form as the residual sum of squares function for multiple linear regression (7.3), with response given by the working residuals, regressors given by $\mathbf{u}_i(\boldsymbol{\theta}^*)$, parameter given by $\boldsymbol{\theta} - \boldsymbol{\theta}^*$, and weights w_i . We switch to matrix notation and let $\mathbf{U}(\boldsymbol{\theta}^*)$ be an $n \times k$ matrix with *i*th row $\mathbf{u}_i(\boldsymbol{\theta}^*)'$, \mathbf{W} is an $n \times n$ diagonal matrix of weights, and $\hat{\mathbf{e}}^* = (\hat{e}_1^*, \dots, \hat{e}_n^*)'$. The least squares solution is then

$$\widehat{\boldsymbol{\theta} - \boldsymbol{\theta}^*} = [\mathbf{U}(\boldsymbol{\theta}^*)' \mathbf{W} \mathbf{U}(\boldsymbol{\theta}^*)]^{-1} \mathbf{U}(\boldsymbol{\theta}^*)' \mathbf{W} \hat{\mathbf{e}}^* \quad (11.10)$$

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + [\mathbf{U}(\boldsymbol{\theta}^*)' \mathbf{W} \mathbf{U}(\boldsymbol{\theta}^*)]^{-1} \mathbf{U}(\boldsymbol{\theta}^*)' \mathbf{W} \hat{\mathbf{e}}^* \quad (11.11)$$

We will use (11.10) in two ways, first to get a computing algorithm for estimating $\boldsymbol{\theta}$ in the rest of this section and then as a basis for inference in the next section.

Here is the *Gauss–Newton algorithm* that is suggested by (11.10) and (11.11):

1. Select an initial guess $\boldsymbol{\theta}^{(0)}$ for $\boldsymbol{\theta}$, and compute $\text{RSS}(\boldsymbol{\theta}^{(0)})$.
2. Set the iteration counter at $j = 0$.
3. Compute $\mathbf{U}(\boldsymbol{\theta}^{(j)})$ and $\hat{\mathbf{e}}^{(j)}$ with *i*th element $y_i - m(\mathbf{x}_i, \boldsymbol{\theta}^{(j)})$. Evaluating (11.11) obtains the solution of a weighted linear least squares problem, with response $\hat{\mathbf{e}}^{(j)}$, predictors $\mathbf{U}(\boldsymbol{\theta}^{(j)})$, and weights given by the w_i . The new estimator is $\boldsymbol{\theta}^{(j+1)}$. Also, compute the residuals sum of squares $\text{RSS}(\boldsymbol{\theta}^{(j+1)})$.
4. Stop if $\text{RSS}(\boldsymbol{\theta}^{(j)}) - \text{RSS}(\boldsymbol{\theta}^{(j+1)})$ is sufficiently small, in which case there is convergence. Otherwise, set $j = j + 1$. If j is too large, stop, and declare that the algorithm has failed to converge. If j is not too large, go to step 3.

The Gauss–Newton algorithm estimates the parameters of a nonlinear regression problem by a sequence of approximating linear WLS calculations.

Most statistical software for nonlinear regression uses the Gauss–Newton algorithm, or a modification of it, for estimating parameters. Some programs

allow using a general function minimizer based on some other algorithm to minimize (11.5). We provide some references at the end of the chapter.

There appear to be two impediments to the use of the Gauss–Newton algorithm. First, the score vectors, which are the derivatives of m with respect to the parameters, are needed. Some software may require the user to provide expressions for the derivatives, but many packages compute derivatives using either symbolic or numeric differentiation. Also, the user must provide starting values $\boldsymbol{\theta}^{(0)}$; there appears to be no general way to avoid specifying starting values. The optimization routine may also converge to a local minimum of the residuals sum of squares function rather than a global minimum, and so finding good starting values can be very important in some problems. With poor starting values, an algorithm may fail to converge to any estimate. We will shortly discuss starting values in the context of an example.

11.2 INFERENCE ASSUMING LARGE SAMPLES

We repeat (11.11), but now we reinterpret $\boldsymbol{\theta}^*$ as the *true, unknown value of $\boldsymbol{\theta}$* . In this case, the working residuals $\hat{\mathbf{e}}^*$ are now the actual errors \mathbf{e} , the differences between the response and the true means. We write

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + [\mathbf{U}(\boldsymbol{\theta}^*)' \mathbf{W} \mathbf{U}(\boldsymbol{\theta}^*)]^{-1} \mathbf{U}(\boldsymbol{\theta}^*)' \mathbf{W} \mathbf{e} \quad (11.12)$$

This equation is based on the assumption that the nonlinear kernel mean function m can be accurately approximated close to $\boldsymbol{\theta}^*$ by the linear approximation (11.8), and this can be guaranteed only if the sample size n is large enough. We then see that $\hat{\boldsymbol{\theta}}$ is equal to the true value plus a linear combination of the elements of \mathbf{e} , and by the central limit theorem $\hat{\boldsymbol{\theta}}$ under regularity conditions will be approximately normally distributed,

$$\hat{\boldsymbol{\theta}} | X \sim N(\boldsymbol{\theta}^*, \sigma^2 [\mathbf{U}(\boldsymbol{\theta}^*)' \mathbf{W} \mathbf{U}(\boldsymbol{\theta}^*)]^{-1}) \quad (11.13)$$

An estimate of the large sample variance is obtained by replacing the unknown $\boldsymbol{\theta}^*$ by $\hat{\boldsymbol{\theta}}$ on the right side of (11.13),

$$\widehat{\text{Var}}(\hat{\boldsymbol{\theta}} | X) = \hat{\sigma}^2 [\mathbf{U}(\hat{\boldsymbol{\theta}})' \mathbf{W} \mathbf{U}(\hat{\boldsymbol{\theta}})]^{-1} \quad (11.14)$$

where the estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\text{RSS}(\hat{\boldsymbol{\theta}})}{n - k} \quad (11.15)$$

and k is the number of parameters estimated in the mean function.

These results closely parallel the results for the linear model, and consequently the inferential methods such as F - and t -tests and the analysis of

variance for comparing nested mean functions, can be used for nonlinear models. One change that is recommended is to use the normal distribution rather than the t for inferences where the t would be relevant, but since (11.13) is really expected to be valid only in large samples, this is hardly important. *We emphasize that in small samples, large sample inferences may be inaccurate.*

11.3 STARTING VALUES

We can illustrate using these results with the turkey growth experiment. The experiment was conducted to study the effects on turkey growth of different amounts A of methionine, ranging from a control with no supplementation to 0.44% of the total diet. The experimental unit was a pen of young turkeys, and treatments were assigned to pens at random so that 10 pens get the control (no supplementation) and five pens received each of the other five amounts used in the experiment, for a total of 35 pens. Pen weights, the average weight of the turkeys in the pen, were obtained at the beginning and the end of the experiment 3 weeks later. The response variable is Gain , the average weight gain in grams per turkey in a pen. The weight gains are given in the file `turk0` (Cook and Witmer, 1985). The primary goal of this experiment is to understand how expected weight gain $E(\text{Gain}|A)$ changes as A is varied. The data are shown in Figure 11.1.

In Figure 11.1, $E(\text{Gain}|A)$ appears to increase with A , at least over the range of values of A in the data. In addition, there is considerable pen-to-pen variation, reflected by the variability between repeated observations at the same value of A . The mean function is certainly not a straight line since the difference in the means when $A > 0.3$ is much smaller than the difference in means when $A < 0.2$. While a polynomial of degree two or three might well match the

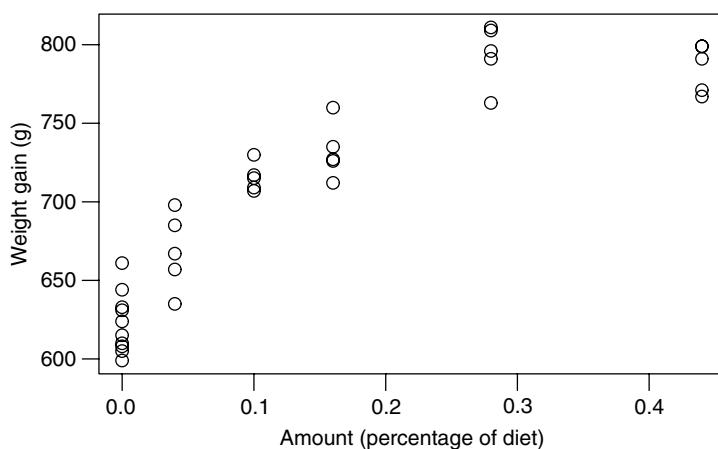


Figure 11.1 Turkey data.

mean at the six values of A in the experiment, it will surely not match the data outside the range of A , and the parameters would have no practical interpretation.

For turkey growth as a function of an amount of an amino acid, the mean function

$$E(\text{Gain}|A) = \theta_1 + \theta_2(1 - \exp(-\theta_3 A)) \quad (11.16)$$

was suggested by Parks (1982). To estimate the parameters in (11.16), we need starting values for θ . While there is no absolute rule for selecting starting values, the following approaches are often useful:

Guessing Sometimes, starting values can be obtained by guessing values for the parameters. In the turkey data, from Figure 11.1, the intercept is about 620 and the asymptote is around 800. This leads to starting values $\theta_1^{(0)} = 620$ and $\theta_2^{(0)} = 800 - 620 = 180$. Guessing a value for the rate parameter θ_3 is harder.

Solving equations for a subset of the data Select as many distinct data points as parameters, and solve the equations for the unknown parameters. The hope is that the equations will be easy to solve. Selecting data points that are diverse often works well. In the turkey data, given $\theta_1^{(0)} = 620$ and $\theta_2^{(0)} = 180$ from the graph, we can get an initial estimate for θ_3 by solving only one equation in one unknown. For example, when $A = 0.16$, a plausible value of Gain is Gain = 750, so

$$750 = 620 + 180(1 - \exp(-\theta_3^{(0)}(0.16)))$$

which is easily solved to give $\theta_3^{(0)} \approx 8$. Thus, we now have starting values for all three parameters.

Linearization If possible, transform to a multiple linear regression mean function, and fit it to get starting values. In the turkey data, we can move the parameters θ_1 and θ_2 to the left side of the mean function to get

$$\frac{(\theta_1 + \theta_2) - y_i}{\theta_2} = \exp(-\theta_3 A)$$

Taking logarithms of both sides,

$$\log\left(\frac{(\theta_1 + \theta_2) - y_i}{\theta_2}\right) = -\theta_3 D$$

Substituting initial guesses $\theta_1^{(0)} = 620$ and $\theta_2^{(0)} = 180$ on the left side of this equation, we can compute an initial guess for θ_3 by the linear regression of $\log[(y_i - 800)/180]$ on $-D$, through the origin. The OLS estimate in this approximate problem is $\theta_3^{(0)} \approx 12$.

Table 11.1 Nonlinear Least Squares Fit of (11.16)

	Estimate	Std. Error	z-Value	Pr(> z)
$\hat{\theta}_1$	622.958	5.901	105.57	<2e-16
$\hat{\theta}_2$	178.252	11.636	15.32	2.74e-16
$\hat{\theta}_3$	7.122	1.205	5.91	1.41e-06

$\hat{\sigma} = 19.66$, $df = 32$.

Many computer packages for nonlinear regression require specification of the function m using an expression such as

$$y \sim th1 + th2 * (1 - \exp(-th3 * A))$$

As with the Wilkinson and Rogers (1973) notation for linear models, the \sim is read “is modeled as,” and the left side is the name of the response. The right side uses syntax similar to a computer language like C, Basic or Fortran to specify the model, including both variable names and parameter names. In contrast, the Wilkinson–Rogers notation for linear models omits parameter names because of the implied relationship between regressors and parameters in linear models.

If the starting values are adequate and the nonlinear optimizer converges, output including the quantities in Table 11.1 will be produced. This table is very similar to the usual output for linear regression. The column marked “Estimate” gives $\hat{\theta}$. Since there is no necessary connection between regressors and parameters, the lines of the table are labeled with the names of the parameters, not the names of the regressors. The next column labeled “Std. Error” gives the square root of the diagonal entries of the matrix given at (11.14), so the standard errors are based on large sample approximation. The column labeled “z-value” is the ratio of the estimate to its large sample standard error and can be used for a test of the null hypothesis that a particular parameter is equal to zero against either a general or one-sided alternative. The column marked “P(>|z|)” is the significance level for this test, using a normal reference distribution rather than a t -distribution. Given at the foot of the table is the estimate $\hat{\sigma}$ and its df , which is the number of cases minus the number of elements in θ that were estimated, $df = 35 - 3 = 32$.

Since this example has only one predictor, Figure 11.1 is a summary graph for this problem. Figure 11.2 repeats this figure, but with the fitted mean function $\hat{E}(\text{Gain}|A = a) = 622.958 + 178.252(1 - \exp(-7.122a))$ added to the graph. The fitted mean function does not reproduce the possible decline of response for the largest value of A because it is constrained to increase toward an asymptote. For $A = 0.28$, the fitted function is somewhat less than the mean of the observed values, while at $A = 0.44$, it is somewhat larger than the mean of

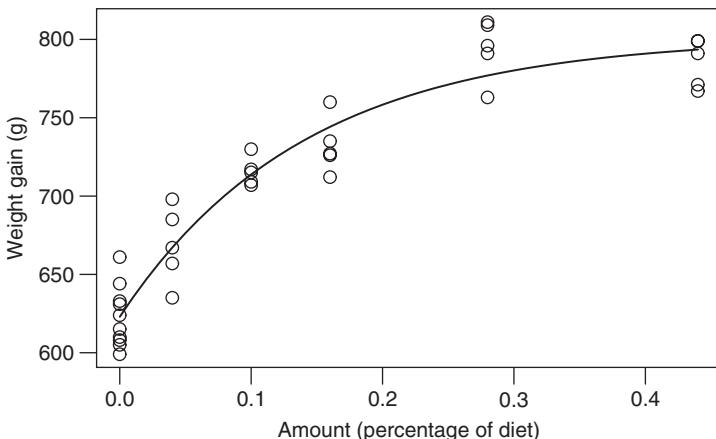


Figure 11.2 Fitted mean function.

the observed values. If we believe that an asymptotic form is really appropriate for these data, then the fit of this mean function seems to be very good.

Factors with Nonlinear Models

The primary purpose of the turkey growth experiment was to compare weight gain curves for methionine derived from three different sources, indicated in the data file by a factor S with three levels. The most general model is, for $j = 1, 2, 3$,

$$\begin{aligned} E(\text{Gain}|A = a, S = j) &= \theta_{1j} + \theta_{2j}[1 - \exp(-\theta_{3j}a)] \\ \text{Var}(\text{Gain}|A = a, S = j) &= \sigma_j^2 \end{aligned} \quad (11.17)$$

This model allows each level of S to have its own curve and its own variance. Fitting (11.17) may be straightforward in many statistical packages.² If the variance is thought to be the same for the various levels of S , then $\hat{\sigma}^2 = \sum_{j=1}^3 \text{RSS}_j / \sum_{j=1}^3 df_j$ is the pooled estimate of σ^2 , with RSS_j and df_j the RSS and residual df for the j th level, and it can be used in testing and confidence statements.

For the weight gain example model (11.17) is inappropriate because a dose of $A = 0$ from source 1 is the same as $A = 0$ with any of the sources, so the expected response at $A = 0$ must be the same for all three sources. This requires that the intercept parameters are all equal, $\theta_{11} = \theta_{12} = \theta_{13}$. To fit this model may require using dummy variables, as most packages cannot interpret factors in nonlinear fitting.

²In R the function `nlsList` in the `nlme` package can be used.

For $j = 1, 2, 3$, let S_j be a dummy variable for level j of S , with value 1 when $S = j$ and 0 otherwise. A model with a common intercept but separate rate and asymptote parameters is

$$\begin{aligned} E(\text{Gain}|A = a, S_1, S_2, S_3) &= \theta_1 + S_1[\theta_{21}(1 - \exp(-\theta_{31}a))] \\ &\quad + S_2[\theta_{22}(1 - \exp(-\theta_{32}a))] \\ &\quad + S_3[\theta_{23}(1 - \exp(-\theta_{33}a))] \end{aligned} \quad (11.18)$$

$$\text{Var}(\text{Gain}|A = a, S = j) = \sigma^2$$

The common variance assumption could be relaxed in some circumstances as discussed in Chapter 7. Another reasonable mean function assumes common asymptote but different rate parameters,

$$E(\text{Gain}|A = a, S_1, S_2, S_3) = \theta_1 + \theta_2 \left[1 - \exp \left(-\sum \theta_{3i} S_i a \right) \right] \quad (11.19)$$

The model of no group differences is given by (11.16).

The data from this experiment are given in the data file `turkey`. For each combination of A and S the file contains m , the number of pens of turkeys with that combination of settings, the average weight gain Gain , and the SD of those weight gains. Assuming variance between pens treated alike is the same in all combinations of A and S ,

$$\hat{\sigma}_{pe}^2 = \frac{\sum(m-1)\text{SD}^2}{\sum(m-1)} = \frac{19916}{57} = 349.40 \quad (11.20)$$

provides a model-free estimate of the variance σ^2 with 57 df that can be used in testing. This estimate is called the mean square for pure error.

The data are shown in Figure 11.3, with the fitted lines from (11.18). A separate symbol was used for each of the three groups. Each point shown is an average over m pens, where $m = 5$ for every point except at $A = 0$, where $m = 10$. The point at $A = 0$ is common to all three groups.

The four mean functions (11.16)–(11.19) are fit using nonlinear weighted least squares, with weights equal to m . Starting values for the estimates can be obtained in the same way as for fitting for one group. For testing, all we need are the RSS and df for each fitted model: and df for each.

	Model	df	RSS
Common mean function	(11.16)	10	4326.1
Different rates	(11.19)	8	2568.4
Common intercept	(11.18)	6	2040.0
Separate regressions	(11.17)	4	1151.2

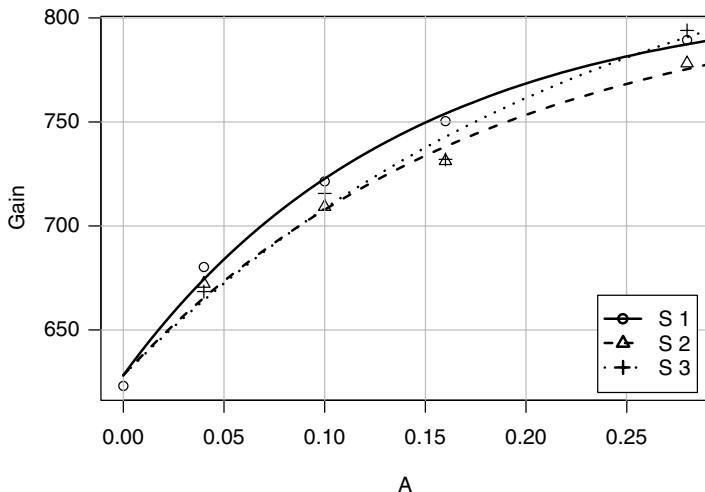


Figure 11.3 Turkey growth as a function of methionine added for three sources of methionine. The lines shown on the figure are for the fit of (11.18), the most general reasonable mean function for these data.

Testing can proceed as in Section 6.1, using $\hat{\sigma}_{pe}^2$ as the denominator of the F -tests. For example, to test the NH given by the common mean function (11.16) versus the AH given by (11.19), the test, using $\hat{\sigma}_{pe}^2$ in the denominator, is

$$F = \frac{(4326.1 - 2040.0)/(10 - 6)}{349.4} = \frac{571.5}{349.4} = 1.63 \sim F(2, 57)$$

for which $p = 0.27$, suggesting no evidence against the simpler mean function. Further testing is not needed in this problem because the separate regression model is not relevant to these data, and the different rates model is intermediate between the models for which no difference can be found. If we did not have a pure error estimate of variance, the estimate of variance from the most general mean function (11.18) would be used in the F -tests.

11.4 BOOTSTRAP INFERENCE

The inference methods based on large samples introduced in the last section *may be inaccurate and misleading in small samples*. We cannot tell in advance if the large sample inference will be accurate or not, as it depends not only on the mean function but also on the way we parameterize it, since there are many ways to write the same nonlinear mean function, and on the actual values of the predictors and the response. Because of this possible inaccuracy,

computing inferences in some other way, at least as a check on the large sample inferences, is a good idea.

One generally useful approach is to use the bootstrap introduced in Section 7.7. We illustrate with data in the file `segreg`, which consists of measurements of electricity consumption in kilowatt-hours and mean temperature in degrees Fahrenheit for one building on the University of Minnesota's Twin Cities campus for 39 months in 1988–1992, courtesy of Charles Ng. The goal is to model consumption as a function of temperature. Higher temperature causes the use of air conditioning, so high temperatures should mean high consumption. This building is steam heated, so electricity is not used for heating. Figure 11.4 is a plot of C = consumption in KWH/day versus Temp , the mean temperature in degrees F.

The mean function for these data is

$$E(C|\text{Temp}) = \begin{cases} \theta_0 & \text{Temp} \leq \gamma \\ \theta_0 + \theta_1(\text{Temp} - \gamma) & \text{Temp} > \gamma \end{cases}$$

This mean function has three parameters, the *level* θ_0 of the first phase; the *slope* θ_1 of the second phase; and the *knot*, γ , and assumes that energy consumption is unaffected by temperature when the temperature is below the knot, but the mean increases linearly with temperature beyond the knot. The goal is to estimate the parameters.

The mean function can be combined into a single equation by writing

$$E(C|\text{Temp}) = \theta_0 + \theta_1(\max(0, \text{Temp} - \gamma))$$

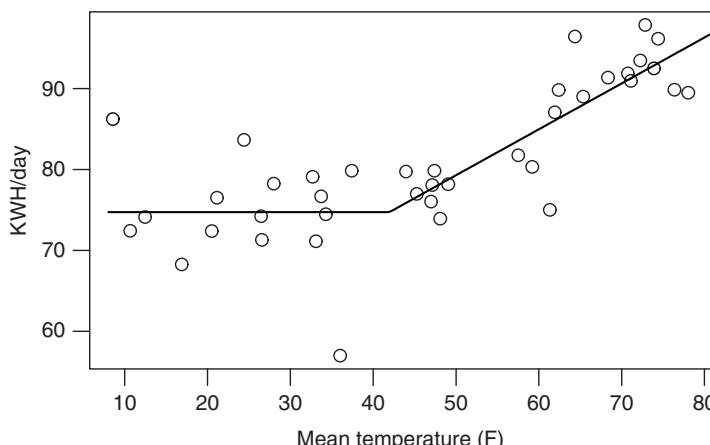


Figure 11.4 Electrical energy consumption per day as a function of mean temperature for one building. The line shown on the graph is the least squares fit.

Table 11.2 Regression Summary Segmented Regression Example

	Estimate	Std. Error	t-Value	Pr(> t)
$\hat{\theta}_0$	74.6953	1.3433	55.607	<2e-16
$\hat{\theta}_1$	0.5674	0.1006	5.641	2.10e-06
$\hat{\gamma}$	41.9512	4.6583	9.006	9.43e-11

$\hat{\sigma} = 5.373$, $df = 36$.

Starting values can be easily obtained from the graph, with $\theta_0^{(0)} = 70$, $\theta_1^{(0)} = 0.5$, and $\gamma^{(0)} = 40$. The fitted model is summarized in Table 11.2. The baseline electrical consumption is estimated to be about $\hat{\theta}_0 \approx 75$ KWH/day. The knot is estimated to be at $\hat{\gamma} \approx 42^\circ\text{F}$, and the increment in consumption beyond that temperature is about $\hat{\theta}_2 \approx 0.6$ KWH per degree increase.

From Figure 11.4, one might get the impression that information about the knot is asymmetric: γ could be larger than 42 but is unlikely to be substantially less than 42. We might expect that in this case, confidence or test procedures based on asymptotic normality will be quite poor. We can confirm this using the bootstrap.

Figure 11.5 is a scatterplot matrix of $B = 999$ case resampling bootstrap replications. All three parameters are estimated on each replication. The diagonals contain histograms of the 999 estimates of each of the parameters. If the normal approximation were adequate, we would expect that each of these histograms would look like a normal density function. While this may be so for θ_1 , this is not the case for θ_2 or γ . As expected, the histogram for γ is skewed to the right, meaning that estimates of γ much larger than about 40 occasionally occur, but smaller values almost never occur. The univariate normal approximations are therefore poor.

The other graphs in the scatterplot matrix tell us about the distributions of the estimated parameters taken two at a time. If the normal approximation were to hold, these graphs should have approximately straight-line mean functions. The smoothers on Figure 11.5 are generally far from straight, and so the large sample inferences are likely to be badly in error.

In contrast, Figure 11.6 is the bootstrap summary for the first source in the turkey growth data. Normality is apparent in histograms on the diagonal, and a linear mean function seems plausible for most of the scatterplots, and so the large sample inference is adequate here.

Table 11.3 compares the estimates and 95% confidence intervals produced by the asymptotic z -approximation and by the percentile bootstrap. Although the bootstrap SDs match the large sample standard errors reasonably well, the confidence intervals for both θ_1 and for γ are shifted toward smaller values than the more accurate bootstrap estimates.

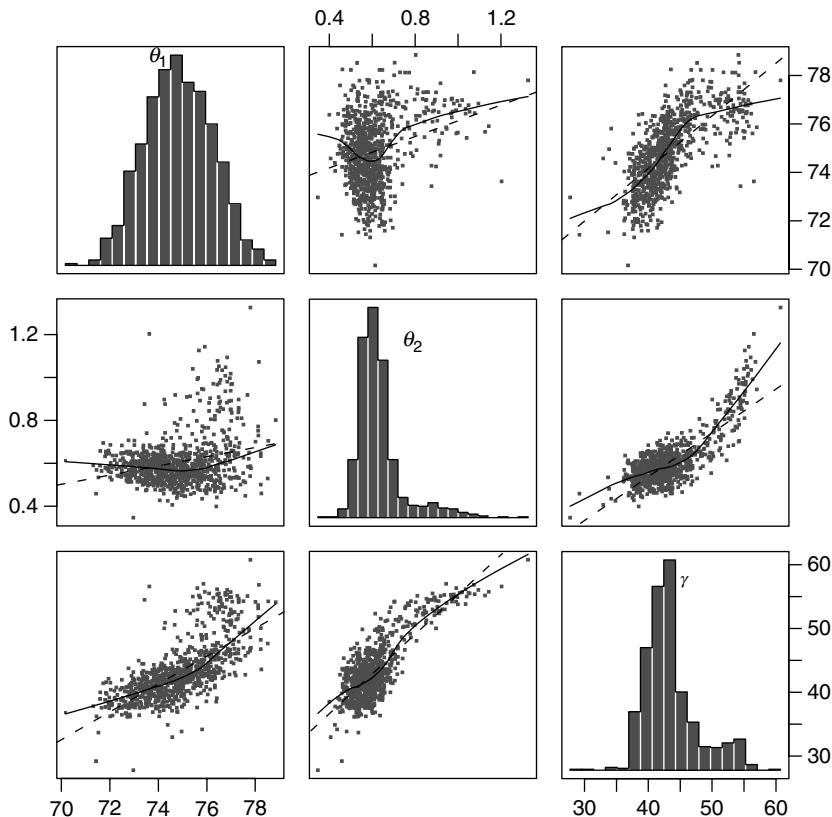


Figure 11.5 Scatterplot matrix of estimates of the parameters in the segmented regression example, computed from $B = 999$ case bootstraps.

11.5 FURTHER READING

Seber and Wild (1989) and Bates and Watts (1988) provide textbook-length treatments of nonlinear regression problems. Computational issues are also discussed in these references and in Thisted (1988, chapter 4). Ratkowsky (1990) provides an extensive listing of nonlinear mean functions that are commonly used in various fields of application.

11.6 PROBLEMS

- 11.1** (Data file: `sleep1`) Suppose we have a response Y , a predictor X , and a factor G with g levels. A generalization of the concurrent regression mean function given by Model 3 of Section 5.1.3, is, for $j = 1, \dots, g$,

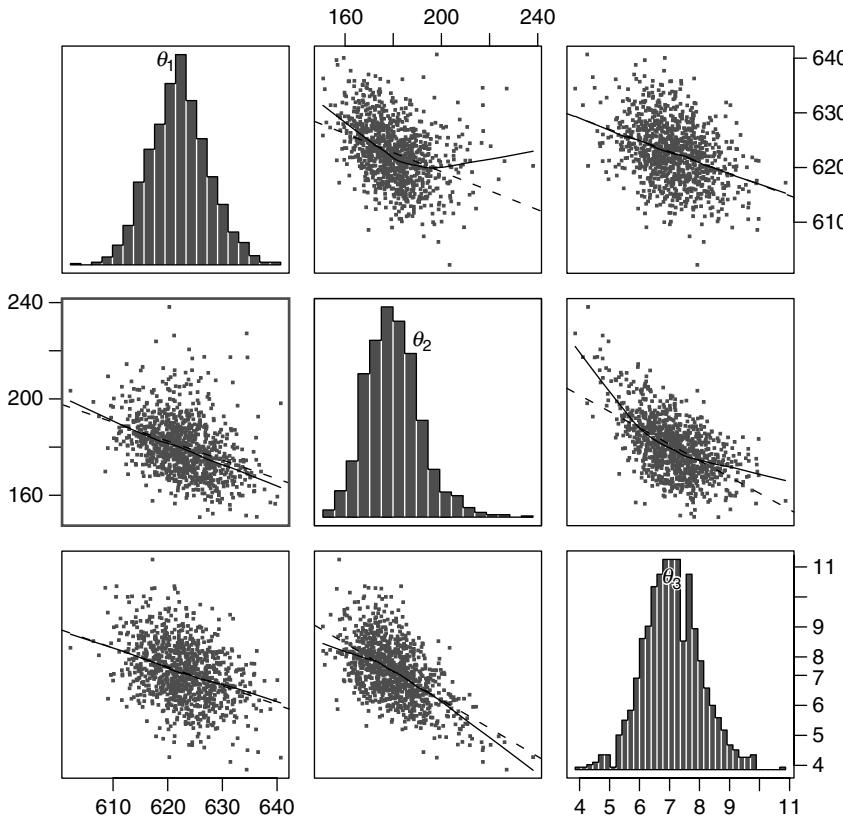


Figure 11.6 Scatterplot matrix of bootstrap estimates for the turkey growth data. Two of the replicates were very different from the others and were deleted before graphing.

Table 11.3 Comparison of Large-Sample and Bootstrap Inference for the Segmented Regression Data

	Large Sample			Bootstrap			
	θ_0	θ_1	γ	Mean	θ_0	θ_1	γ
Estimate	74.70	0.57	41.95	Mean	74.92	0.62	43.60
SE	1.34	0.10	4.66	SD	1.47	0.13	4.81
2.5%	72.06	0.37	32.82	2.5%	71.96	0.47	37.16
97.5%	77.33	0.76	51.08	97.5%	77.60	0.99	55.59

$$E(Y|X = x, G = j) = \beta_0 + \beta_{1j}(x - \gamma) \quad (11.21)$$

for some point of concurrence γ .

- 11.1.1** Explain why (11.21) is a nonlinear mean function. Describe in words what this mean function specifies.

- 11.1.2** Allison and Cicchetti (1976) provided data on the typical sleeping habits of mammal species. Of interest here is $T\$$, the total hours of sleep per day; $BodyWt$, the typical body weight of the species; and D , a discrete variable with five values, from $D = 1$, animals with low danger to $D = 5$ for animals with very high danger. Fit the mean function

$$E(T\$|\log(BodyWt) = x, D = j) = \beta_0 + \beta_{1j}(x - \gamma)$$

To get starting values, fit the concurrent regression model with $\gamma = 0$. The estimate of γ will be very highly variable, as is often the case with centering parameters like γ in this mean function.

- 11.2** (Date file: `lakemary`) In fisheries studies, the most commonly used mean function for expected length of a fish at a given age is the *von Bertalanffy* function (Bertalanffy, 1938; Haddon and Haddon, 2010), given by

$$E(Length|Age = t) = L_\infty(1 - \exp(-K(t - t_0))) \quad (11.22)$$

The parameter L_∞ is the expected value of `Length` for extremely large ages, and so it is the asymptotic or upper limit to growth, and K is a growth rate parameter that determines how quickly the upper limit to growth is reached. When `Age` = t_0 , the expected length of the fish is 0, which allows fish to have nonzero length at birth if $t_0 < 0$.

- 11.2.1** The data in the file gives the `Age` in years and `Length` in millimeters for a sample of 78 bluegill fish from Lake Mary, Minnesota, in 1981 (courtesy of Richard Frie). `Age` is determined by counting the number of rings on a scale of the fish. This is a cross-sectional data set, meaning that all the fish were measured once. Draw a scatterplot of the data.

- 11.2.2** Use nonlinear regression to fit the von Bertalanffy function to these data. To get starting values, first guess at L_∞ from the scatterplot to be a value larger than any of the observed values in the data. Next, divide both sides of (11.22) by the initial estimate of L_∞ , and rearrange terms to get just $\exp(-K(t - t_0))$ on the right of the equation. Take logarithms, to get a linear mean function, and then use OLS for the linear mean function to get the remaining starting values. After getting the fitted model, draw the fitted mean function on your scatterplot.

- 11.2.3** Obtain a 95% confidence interval for L_∞ using the large sample approximation, and using the bootstrap.
- 11.3** (Data file: walleye) The data in the file walleye give the length in mm and the age in years of a sample of over 3,000 male walleye, a popular game fish, captured in Butternut Lake in Northern Wisconsin (LeBeau, 2004). The fish are also classified according to the time period in which they were captured, with `period` = 1 for pre-1990, `period` = 2 for 1990–1996, and `period` = 3 for 1997–2000. Management practices on the lake were different in each of the periods, so it is of interest to compare the length at age for the three time periods.
Using the von Bertalanffy length at age function (11.22), compare the three time periods. If different, are all the parameters different, or just some of them? Which ones? Summarize your results.
- 11.4 A quadratic polynomial as a nonlinear model** (Data file: swan96) The data were collected by the Minnesota Department of Natural Resources to study the abundance of black crappies, a species of fish, on Swan Lake, Minnesota in 1996. The response variable is `LCPUE`, the logarithm of the catch of 200 mm or longer black crappies per unit of fishing effort. It is believed that `LCPUE` is proportional to abundance. The single predictor is `Day`, the day on which the sample was taken, measured as the number of days after June 19, 1996. Some of the measurements were taken the following spring on the same population of fish before the young of the year are born in late June. No samples are taken during the winter months when the lake surface was frozen.
- 11.4.1** For these data, fit the quadratic polynomial

$$E(LCPUE|Day = x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

assuming $\text{Var}(LCPUE|Day = x) = \sigma^2$. Draw a scatterplot of `LCPUE` versus `Day`, and add the fitted curve to this plot.

- 11.4.2** Using the delta method described in Section 7.6, obtain the estimate and variance for the value of `Day` that maximizes $E(LCPUE|Day)$.
- 11.4.3** Another parameterization of the quadratic polynomial is

$$E(Y|X) = \theta_1 - 2\theta_2\theta_3 x + \theta_3 x^2$$

where the θ s can be related to the β s by

$$\theta_1 = \beta_0, \quad \theta_2 = -\beta_1/2\beta_2, \quad \theta_3 = \beta_2$$

In this parameterization, θ_1 is the intercept, θ_2 is the value of the predictor that gives the maximum value of the response, and θ_3 is

a measure of curvature. This is a nonlinear model because the mean function is a nonlinear function of the parameters. Its advantage is that at least two of the parameters, the intercept θ_1 and the value of x that maximizes the response θ_2 , are directly interpretable. Use nonlinear least squares to fit this mean function. Compare your results with the first two parts of this problem.

- 11.5** (Data file: Highway) Nonlinear regression can be used to select transformations for a linear regression mean function. As an example, consider the highway accident data, described in Table 8.1, with response $\log(\text{rate})$ and two predictors $X_1 = \text{len}$ and $X_2 = \text{adt}$. Fit the nonlinear mean function

$$E(\log(\text{Rate})|X_1 = x_1, X_2 = x_2, X_3 = x_3) = \beta_0 + \beta_1 \psi_S(X_1, \lambda_1) + \beta_2 \psi_S(X_2, \lambda_2)$$

where the scaled power transformations $\psi_S(X_j, \lambda_j)$ are defined at (8.3). Compare the results you get to results obtained using the transformation methodology in Chapter 8.

Binomial and Poisson Regression

In this chapter we show how nearly all the methods described in this book can be extended to problems in which the response variable is a count rather than a measured variable. We will consider binomial regression, which includes a binary categorical response, and also the closely related Poisson regression. We first review a bit about the binomial and Poisson distributions, and then describe the regression models with counted responses with either binomial or Poisson distributions, emphasizing the connections to the rest of this book.

Books dedicated to binomial regression include Collett (2003) and Hosmer et al. (2013). Counted data more generally is covered in Agresti (2007, 2013). Counted data models can also be studied in the framework of *generalized linear models*, in which the linear model, binomial model, and Poisson model are all special cases. McCullagh and Nelder (1989) provided the basis for this approach.

12.1 DISTRIBUTIONS FOR COUNTED DATA

12.1.1 Bernoulli Distribution

Suppose the random variable Y has two possible values, perhaps called “success” or “failure,” with probability of success equal to θ where $0 \leq \theta \leq 1$. We label the possible outcomes of Y as $y = 1$ if success occurs, and $y = 0$ if success does not occur. We will say that Y with these characteristics has a *Bernoulli distribution*¹ with probability of success θ . Using Appendix A.2,

$$\mathrm{E}(Y) = \theta \quad \mathrm{Var}(Y) = \theta(1 - \theta) \tag{12.1}$$

¹Named for Jacob Bernoulli, 1654–1705.

An important feature of the Bernoulli distribution is that the variance depends on the mean, $\text{Var}(Y) = E(Y)(1 - E(Y))$. The variance is largest for $\theta = 1/2$, and smallest for θ close to 0 or 1. The Bernoulli is the only distribution for a random variable with sample space $\{0, 1\}$.

12.1.2 Binomial Distribution

The binomial distribution generalizes the Bernoulli. Suppose we have m random variables B_1, B_2, \dots, B_m , such that (1) each B_j has a Bernoulli distribution with the same probability θ of success; and (2) all the B_j are independent. Then if Y is the number of successes in the m trials, $Y = \sum B_j$, we say that Y has a binomial distribution with m trials and probability of success θ . We write this as $Y \sim \text{Bin}(m, \theta)$. Each of the Bernoulli variables is $B_j \sim \text{Bin}(1, \theta)$.

The *probability mass function* for the binomial is

$$\Pr(Y = j) = \binom{m}{j} \theta^j (1 - \theta)^{(m-j)} \quad (12.2)$$

for $j \in \{0, 1, \dots, m\}$. The mean and variance of a binomial are

$$E(Y) = m\theta \quad \text{Var}(Y) = m\theta(1 - \theta) \quad (12.3)$$

Since m is known, both the mean and variance are determined by θ only. Both assumptions of constant θ and of independence are required for the binomial distribution to apply to the number of successes in m trials. For example, in a survey of family members about their view on a particular political issue, the number in favor of the issue will likely not be binomially distributed because the views of members of the same family are unlikely to be independent.

12.1.3 Poisson Distribution

Whereas the binomial distribution concerns the distribution of the number of successes in a fixed number m of trials, the Poisson distribution² is the number of events of a specific type that occur in a fixed time or space. A Poisson variable Y can take the value of any nonnegative integer $\{0, 1, 2, \dots\}$. For example, if customers to an ice cream store arrive independently, at random but at a constant rate, then the number of customers arriving in any fixed period of time will follow a Poisson distribution.

We will say Y has a Poisson distribution with rate λ , $Y \sim \text{Po}(\lambda)$ if

$$\Pr(Y = y) = \exp(-\lambda) \lambda^y / y! \quad y = 0, 1, \dots \quad (12.4)$$

²Named for Siméon Denis Poisson, 1781–1840.

Using Appendix A.2, it is not hard to show that

$$\mathrm{E}(Y) = \lambda \quad \mathrm{Var}(Y) = \lambda \quad (12.5)$$

so the mean and variance of a Poisson variable are equal.

In the ice cream store arrival example, the count Y depends on both the assumption of independence of customers and constant arrival rate. The rate could vary with time of day or outdoor temperature, and so a Poisson assumption may be appropriate for short time intervals but not for longer intervals. Arrivals could be correlated if customers arrive in groups, for example, at the end of a high school sports event, again suggesting that the number of arrivals may not follow a Poisson distribution.

There are many interesting and useful relationships between the Poisson and the binomial that suggest that regression models for both types of responses should be studied together. In particular, suppose that $Y_1 \sim \mathrm{Po}(\lambda_1)$ and $Y_2 \sim \mathrm{Po}(\lambda_2)$, and suppose Y_1 and Y_2 are independent. Y_1 could be the number of ice cream customers who arrive and buy an ice cream cone, and Y_2 could be the number who arrive and buy a yogurt cone. Then the sum $(Y_1 + Y_2) \sim \mathrm{Po}(\lambda_1 + \lambda_2)$ is the number of customers who arrive in the time period and buy a cone of either type. The conditional distribution of Y_1 given the total number of customers arriving is binomial, $Y_1 | (Y_1 + Y_2) \sim \mathrm{Bin}(Y_1 + Y_2, \lambda_1 / (\lambda_1 + \lambda_2))$.

12.2 REGRESSION MODELS FOR COUNTS

The big idea is that the parameter for the counted distribution, θ for the binomial or λ for the Poisson, can depend on the values of predictors.

12.2.1 Binomial Regression

We consider the binomial case first. We assume that $\theta(\mathbf{x})$ depends on the values \mathbf{x} of the regressors only through a linear combination $\boldsymbol{\beta}'\mathbf{x}$ for some unknown $\boldsymbol{\beta}$. We can write $\theta(\mathbf{x})$ as a function of $\boldsymbol{\beta}'\mathbf{x}$,

$$\theta(\mathbf{x}) = m(\boldsymbol{\beta}'\mathbf{x}) \quad (12.6)$$

The quantity $\boldsymbol{\beta}'\mathbf{x}$ is called the *linear predictor*. As in nonlinear models, the function m is called a kernel mean function. Because the left side of (12.6) is a probability, $m(\boldsymbol{\beta}'\mathbf{x})$ must map $\boldsymbol{\beta}'\mathbf{x}$, which can take any possible value in $(-\infty, \infty)$, into the range $(0, 1)$. The most frequently used kernel mean function for binomial regression, and the only one we discuss in this book, is the *logistic function*,

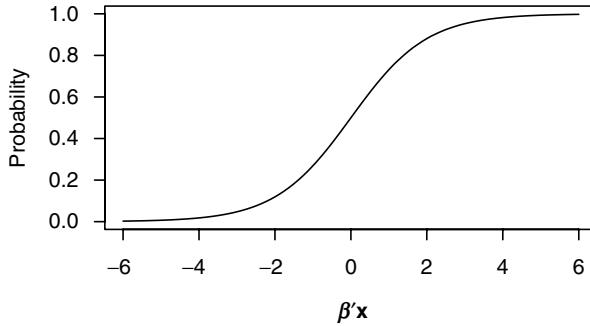


Figure 12.1 The logistic kernel mean function.

$$\theta(x) = \text{m}(\beta'x) = \frac{\exp(\beta'x)}{1 + \exp(\beta'x)} = \frac{1}{1 + \exp(-\beta'x)} \quad (12.7)$$

The last two forms are equivalent representations for the same function. A graph of the logistic function is shown in Figure 12.1.

Most presentations of logistic regression work with the *link function*, which is the inverse of the kernel mean function. Solving (12.7) for $\beta'x$, we find

$$\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta'x \quad (12.8)$$

The left side of (12.8) is called a *logit* or *log-odds* and the right side is the linear predictor $\beta'x$. If we were to draw a graph of $\log\{\theta(x)/[1 - \theta(x)]\}$ versus $\beta'x$, we would get a straight line.

The ratio $\theta(x)/[1 - \theta(x)]$ is the *odds of success*. For example, if the probability of success is 0.25, the odds of success are $0.25/(1 - 0.25) = 1/3$, one success to each three failures. If the probability of success is 0.8, then the odds of success are $0.8/0.2 = 4$, or four successes to one failure. Whereas probabilities are bounded between 0 and 1, odds can be any nonnegative number. The logit is the logarithm of the odds. According to Equation (12.8), the logit is equal to a linear combination of the regressors.

The data for logistic regression for the i th observation consist of the observed number of successes y_i , the observed number of trials m_i , and a vector \mathbf{x}_i of regressors computed from the predictors in the problem, as in Chapters 4 and 5. The three components of the logistic regression model are the following:

Distribution The distribution of $(Y_i|X_i = \mathbf{x}_i) \sim \text{Bin}(m_i, \theta(\mathbf{x}_i))$. Both the mean and the variance of Y_i depend only on the known m_i and on $\theta(\mathbf{x}_i)$.

Linear predictor The parameter $\theta(\mathbf{x}_i)$ depends on \mathbf{x}_i only through the linear predictor $\beta'x_i$ for some unknown parameter vector β .

Link function There is a link function, or equivalently its inverse the kernel mean function, that specifies the connection between $\theta(\mathbf{x}_i)$ and the linear related regressors $\boldsymbol{\beta}'\mathbf{x}_i$, such as the logit link at (12.8).

Logistic regression models are not fit with OLS. Rather, maximum likelihood estimation is used, based on the binomial distribution; see Appendix A.11.2. Most statistics packages will make fitting logistic models easy, and the results will look just like the results for fitting OLS.

Blowdown

On July 4, 1999, a storm with winds exceeding 90 miles per hour hit the Boundary Waters Canoe Area Wilderness in northeastern Minnesota, causing serious damage to the forest. Rich et al. (2007) studied the effects of this storm using a very extensive ground survey of the area, determining status, either alive or dead, of more than 3600 trees. Suppose $\theta(\mathbf{x})$ is the probability of death by blowdown for a tree with characteristics given by the regressors \mathbf{x} .

We start with one predictor, and consider the dependence of the probability of blowdown on the diameter d of the tree, measured to the nearest 0.5 cm, for black spruce trees only. We will use $\log(d)$ as the single regressor beyond the intercept. The data file BlowBS contains d , the number of trees m of that diameter that were measured, and died , the number of trees that died in the blowdown. We view m as fixed, and model died given m as a binomial response, with regressors for the intercept and $\log(d)$. The data file has $n = 35$ rows, corresponding to the 35 unique values of d . It represents a total of $\sum m = 659$ trees, with m ranging from 1 tree for some of the larger diameters to 91 for $d = 6$ cm.

As usual, we begin by graphing the data in Figure 12.2, with the blowdown fraction died/m on the vertical axis and the regressor $\log(d)$ on the horizontal axis. Since the samples sizes m are highly variable in this example, points in the graph are drawn with area proportional to m . Larger points with more trees are more important in fitting. Concentrating on the the larger points, the probability of blowdown increases with $\log(d)$. The points based on few trees are further from a trend, as should be expected. For example, if $m = 3$, the value of died/m could only equal 0, 1/2, or 1, and all these values are likely to be far from any fitted regression line.

The results fitting the logistic model are summarized in Table 12.1. The coefficient summary is similar to the output for simple regression, giving the

Table 12.1 Logistic Regression Summary for the Black Spruce Blowdown Data

	Estimate	Std. Error	<i>z</i> -Value	$\text{Pr}(z)$
(Intercept)	-7.8925	0.6325	-12.48	0.0000
$\log(d)$	3.2643	0.2761	11.82	0.0000

Residual deviance = 49.891 (33 df). Null deviance = 250.856 (34 df).

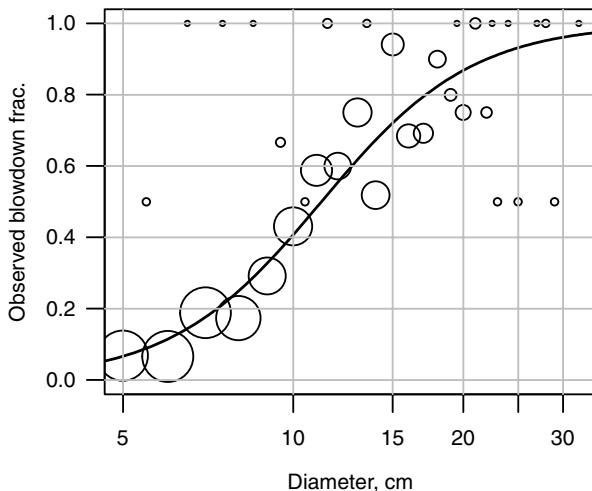


Figure 12.2 Plot of the the blowdown fraction versus d , with the horizontal axis in log scale. The plotted points have area proportional to the sample size for that point. The solid line is the fitted logistic regression.

name of the regressor, its coefficient estimate, and standard error. The ratio of the estimate to its standard error is called a z -value rather than a t -value. Since logistic regression does not have a variance parameter, the t -distribution is not appropriate, and the large sample normal approximation is used to get the significance levels in the last column. Also included in the output is the deviance, which is analogous to the residual sum of squares in linear regression and is described in Section 12.2.2.

The very small p -values suggest that both the intercept and the coefficient for $\log(d)$ are unlikely to be equal to 0. Since the logistic model is in the scale of the logarithm of the odds, we can use Section 4.1.6 on responses in log scale to interpret coefficients. From (4.5), if the diameter d is increased by 10%, then the odds of death by blowdown are multiplied by $\exp[\log(1.1) \times 3.264] \approx 1.34$.

The fitted curve, with equation $1/[1 + \exp[-(-7.8925 + 3.2643 \times \log(d))]]$ is shown on Figure 12.2. The agreement of the line to the points with larger m is encouraging.

To account for the place-to-place variation in storm intensity that is likely to change the probability of death by blowdown, a measure s of local severity of the storm was computed as the fraction of the total basal area of trees of four major species that died near the measured tree. The data file `Blowdown` includes d , s , and a factor `spp` for species with 9 levels for all $n = 3666$ trees, with one row in the file for each tree that was measured. The variable y equals 1 if a particular tree died as a result of the blowdown, and 0 if it survived. The data on black spruce trees used previously are included and require 659 rows, one for each of the trees.

The Wilkinson–Rogers notation for specifying a model was designed for linear models, but it often used for logistic, Poisson, and other generalized linear models. The model we consider is $y \sim \log(d) + s + \log(d):s$, allowing for an interaction between s and $\log(d)$. The “ \sim ” should be read as “is modeled using,” since the response depends on the regressors only through the dependence of the log-odds on the regressors.

The regression summary using the data in Blowdown with spp equal to black spruce, and therefore ignoring trees of all other species, is given in Table 12.2. Since an interaction is included, the z -tests for main effects are relevant only if the test for the interaction suggests the interaction is not needed. The tiny p -value for the interaction suggests all the terms should be maintained, even though the z -value for the main effect of $\log(d)$ is small.

The effects plot is shown in Figure 12.3. The model returns fitted values in the logit scale, and these were transformed first to fitted odds by

Table 12.2 Black Spruce Blowdown Data

	Estimate	Std. Error	z -Value	$\text{Pr}(> z)$
(Intercept)	-3.678	1.425	-2.58	0.010
$\log(d)$	0.578	0.633	0.91	0.361
s	-11.205	3.638	-3.08	0.002
$\log(d):s$	7.085	1.645	4.31	0.000

Residual deviance = 541.7 (655 df). Null deviance = 856.2 (658 df).

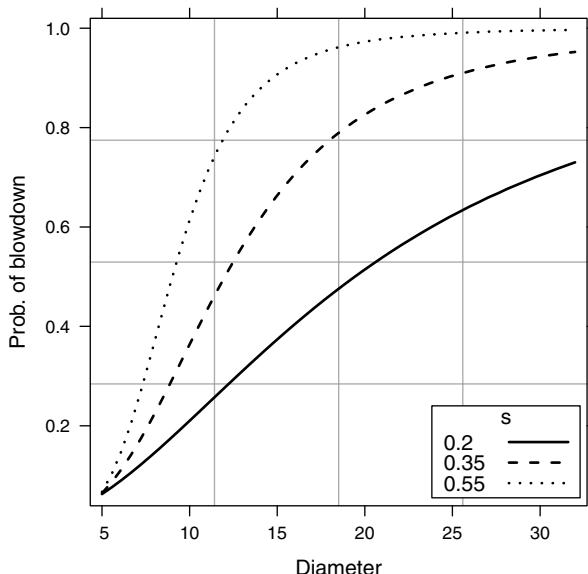


Figure 12.3 Effects plot for black spruce trees in the blowdown data with both d and s as predictors.

exponentiating, and then to the fitted probabilities shown on the plot. The horizontal axis is the diameter d . Because of the interaction, the dependence of the probability of blowdown on d will be different for each value of s . Three curves are shown in Figure 12.3, corresponding roughly to the 25%, 50%, and 75% points of the distribution of s for the black spruce trees. The effect of d increases fastest for the highest quartile and slowest for the lowest quartile. This plot could be drawn with many variations, including using $\log(d)$ on the horizontal axis, using odds or log-odds on the vertical axis, and reversing the roles of d and s in the plot, but the presentation here seems to catch the essence of the solution.

12.2.2 Deviance

In multiple linear regression (Chapter 6), the residual sum of squares provides the basis for tests for comparing mean functions. In logistic and Poisson regression, the residual sum of squares is replaced by the *deviance*, which is often called G^2 . The deviance is defined for logistic regression to be

$$G^2 = 2 \sum_{i=1}^n \left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (m_i - y_i) \log\left(\frac{m_i - y_i}{m_i - \hat{y}_i}\right) \right] \quad (12.9)$$

where $\hat{y}_i = m_i \hat{\theta}(\mathbf{x}_i)$ are the fitted number of successes in m_i trials. The df associated with the deviance is equal to the number of cases n used in the calculation minus the number of elements of β that were estimated; for the black spruce data fit in Table 12.2, $df = 659 - 4 = 655$.

Methodology for comparing models parallels the results in Section 6.1. Write $\beta' \mathbf{x} = \beta'_1 \mathbf{x}_1 + \beta'_2 \mathbf{x}_2$, and consider testing

$$\begin{aligned} \text{NH: } \theta(\mathbf{x}) &= m(\beta'_1 \mathbf{x}_1) \\ \text{AH: } \theta(\mathbf{x}) &= m(\beta'_1 \mathbf{x}_1 + \beta'_2 \mathbf{x}_2) \end{aligned}$$

Obtain the deviance G_{NH}^2 and degrees of freedom df_{NH} under the null hypothesis, and then obtain G_{AH}^2 and df_{AH} under the alternative hypothesis. As with linear models, we will have evidence against the null hypothesis if $G_{NH}^2 - G_{AH}^2$ is too large. To get a p -value, we compare the difference $G_{NH}^2 - G_{AH}^2$ with the χ^2 distribution with $df = df_{NH} - df_{AH}$, not with an F -distribution as was done for linear models.

The NH that the probability of blowdown is constant versus AH that the probability depends on a set of regressors is equivalent to the overall test in linear models, and it is based on the difference between the null deviance and the residual deviance. For the model summarized in Table 12.1, this is $G^2 = 250.86 - 49.89 = 200.97$, which is compared with the χ^2 distribution with $34 - 33 = 1$ df . The significance level is very tiny, suggesting the hypothesis of probability of blowdown independent of d is firmly rejected. Similarly, the

overall test for the model summarized in Table 12.2 is $G^2 = 856.2 - 541.7 = 314.5$, which is compared with the χ^2 distribution with $658 - 655 = 3$ df. Once again the significance level is very small, providing evidence that the probability of blowdown is not constant.

For the next example, we consider testing NH: $y \sim \log(d)$ versus AH: $y \sim \log(d) + s + \log(d) : s$. The NH model was fit in Table 12.1 using the group of trees with the same diameter as the unit of analysis, while the AH model was fit using each tree as the unit of analysis. We need to refit the NH model using the tree as the unit of analysis, with the data file Blowdown. Estimates and standard errors are the same fitting using the grouped binomial data or the individual Bernoulli data, but the deviance is different. Fitting to the individual trees, the residual deviance is 655.24 with 657 df, and the test is $G^2 = 113.50$ with 2 df. Once again, the significance level is very small, and the NH firmly rejected.

Tests with logistic models, and with the Poisson models to be introduced later, are often summarized in an *Analysis of Deviance* table that is directly analogous to the Analysis of Variance table used to study linear models. As an example, we consider a third model for blowdown probability that uses all the data, adding a factor with nine levels for species, and allowing each species to have its own $\log(d) : s$ interaction. This corresponds to fitting with main effects, all two-factor interactions, and a three-factor interaction. The results may be summarized in the Analysis of Deviance table in Table 12.3. This is a Type II table, as in Section 6.2, and is interpreted and used in the same way. Starting at the bottom of the table, the three-factor interaction test is considered first. Testing stops because it has a very small significance level, as lower-order effects are not tested when higher-order effects in the same predictors are nonzero.

Figure 12.4 is a summary effects plot for the blowdown data. The model fit suggests that effects of s and d are needed for most tree species. We see there are interesting differences between species. For red pine, the probability of blowdown appears to decrease with d , while for jack pine, the probability of blowdown may be independent of d . Cedar trees were relatively immune to blowdown except in areas of very high severity. Further analysis of these

Table 12.3 Analysis of Deviance for Blowdown

	df	G^2	$\text{Pr}(>\chi^2)$
$\log(d)$	1	227.8	0.0000
S	1	594.0	0.0000
Spp	8	509.5	0.0000
$\log(d) : s$	1	41.6	0.0000
$\log(d) : spp$	8	71.9	0.0000
$s : spp$	8	36.5	0.0000
$\log(d) : s : spp$	8	20.4	0.0088

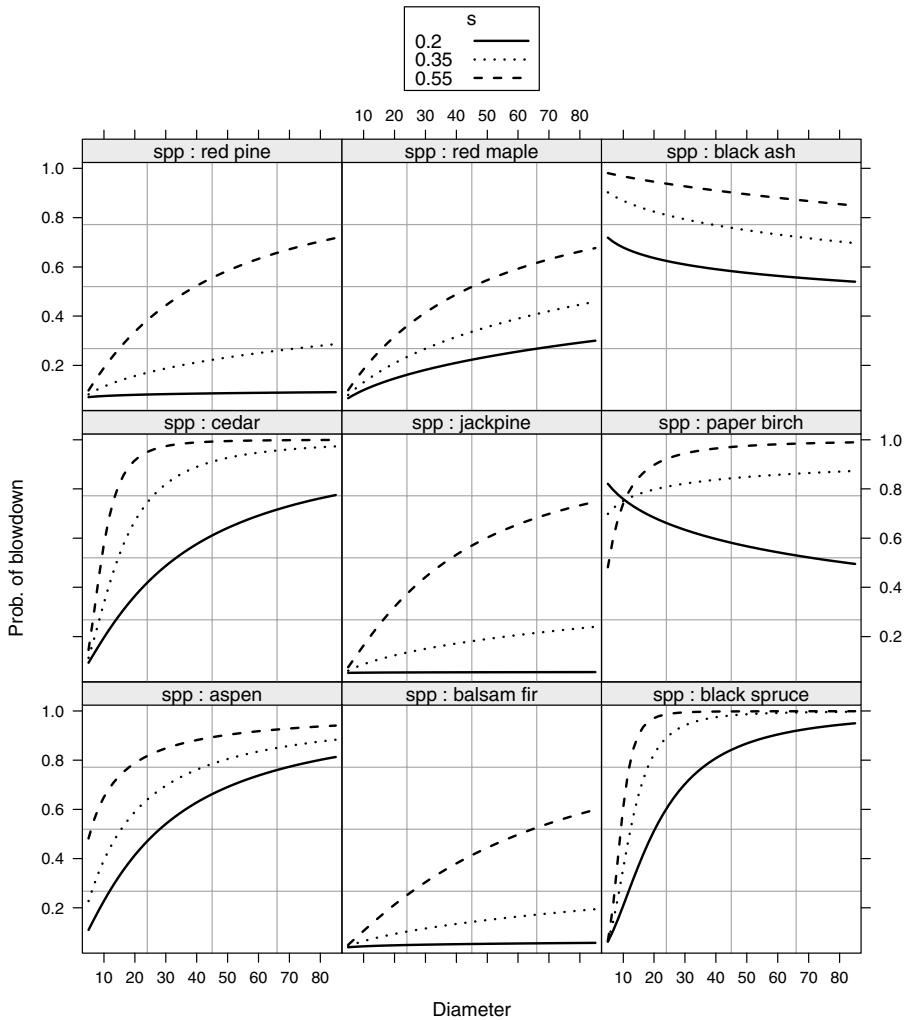


Figure 12.4 Effects plots in the blowdown data with both d and s as predictors.

data would require more work, and quite likely a separate analysis for each species separately could be enlightening.

12.3 POISSON REGRESSION

When the data are to be modeled as if they are Poisson counts, the rate parameter is assumed to depend on the regressors with linear predictor $\beta'x$ through the link function

$$\log[\lambda(\beta'x)] = \beta'x \quad (12.10)$$

Poisson regression models are often called *log-linear models*.

The data for Poisson regression for the i th observation consist of the observed number of events y_i , and the values of the regressors. The three components of the Poisson regression model are as follows:

Distribution The distribution of $(Y_i | X_i = \mathbf{x}_i) \sim \text{Po}[\lambda(\mathbf{x})]$.

Linear predictor The parameter $\lambda(\mathbf{x}_i)$ depends on \mathbf{x}_i only through the linear predictor $\boldsymbol{\beta}'\mathbf{x}_i$ for some unknown parameter vector $\boldsymbol{\beta}$.

Link function The link function is the log-link (12.10).

Maximum likelihood estimation is the usual method used to fit Poisson regression models. The deviance for Poisson regression is given by

$$G^2 = 2 \sum_{i=1}^n [\log(y_i/\hat{y}_i) - (y_i - \hat{y}_i)] \quad (12.11)$$

where \hat{y}_i is the fitted value $\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_i)$.

Mathematical Sciences PhDs

The data in Table 12.4 gives the number of PhD degrees awarded in the mathematical sciences in the United States in 2008–2009 (Phipps et al., 2009). The rows of the table correspond to the factor `type`, with six levels. The first four rows correspond to mathematics departments grouped into Type I public and private for the largest universities, and Types II and III for smaller universities. Type IV corresponds to programs in biostatistics or statistics in any university. Type Va is for applied mathematics in any university. Columns subdivide the counts further by sex and citizenship of the PhD recipient. We can view each of the cell counts as a Poisson random variable with possibly different rates. The data are in the file `AMSSurvey` in a format that is suitable for fitting with Poisson regression. The data file has one row for each cell in the table, so there are $n = 24$ rows. Columns are given for `type`, `sex`, and `citizen`. An additional column called `count` gives the cell counts shown in the table.

Table 12.4 American Mathematics Society PhD Survey, 2008–2009

Level	Non-U.S.		U.S.	
	Female	Male	Female	Male
I(Pu)	29	130	35	132
I(Pr)	25	79	20	87
II	50	89	47	96
III	39	53	32	47
IV	105	122	54	71
Va	12	28	14	34

Table 12.5 Analysis of Deviance for Mathematical Sciences PhDs

	<i>df</i>	<i>G</i> ²	Pr(> χ^2)
Type	5	233.3	0.0000
Sex	1	183.0	0.0000
Citizen	1	5.9	0.0149
type:sex	5	69.1	0.0000
type:citizen	5	24.0	0.0002
sex:citizen	1	0.5	0.4635
Type:sex:citizen	5	1.4	0.9222

Table 12.6 Poisson Regression Summary for the Mathematical Sciences PhDs

	Estimate	Std. Error	<i>z</i> -Value	Pr(> <i>z</i>)
(Intercept)	3.0992	0.1646	18.83	0.0000
typeI (Pu)	0.3417	0.2143	1.59	0.1109
typeII	0.7681	0.2026	3.79	0.0002
typeIII	0.5436	0.2150	2.53	0.0114
typeIV	1.5310	0.1870	8.19	0.0000
typeVa	-0.6296	0.2814	-2.24	0.0253
sexMale	1.3053	0.1681	7.77	0.0000
citizenUS	0.0284	0.1377	0.21	0.8364
typeI (Pu) : sexMale	0.1041	0.2184	0.48	0.6335
typeII:sexMale	-0.6597	0.2097	-3.15	0.0017
typeIII:sexMale	-0.9628	0.2288	-4.21	0.0000
typeIV:sexMale	-1.1115	0.1993	-5.58	0.0000
typeVa:sexMale	-0.4363	0.2878	-1.52	0.1296
typeI (Pu) : citizenUS	0.0207	0.1767	0.12	0.9070
typeII:citizenUS	-0.0001	0.1821	-0.00	0.9997
typeIII:citizenUS	-0.1808	0.2061	-0.88	0.3805
typeIV:citizenUS	-0.6251	0.1771	-3.53	0.0004
typeVa:citizenUS	0.1539	0.2545	0.60	0.5455

Residual deviance = 1.957 (6 *df*). Null deviance = 521.444 (23 *df*).

Table 12.5 gives the Type II Analysis of Deviance table for the fit of log-linear Poisson model with all main effects, two-factor interactions, and the three-factor interaction. Starting as usual at the bottom, both the type:sex:citizen and the sex:citizen interactions have large *p*-values and appear to be negligible. The remaining two-factor interactions have small *p*-values and are not negligible. Before further summarization, we fit the Poisson model including only the important two-factor interactions.

The regression summary is provided in Table 12.6. As is true with any regression model with interactions present, interpretation of coefficient estimates is challenging because the parameters depend on the choice of regressors used to represent the factors. The interaction parameters are the most easily interpretable. For example, the coefficient for U.S. citizens at type IV institutions is -0.6251, and this describes the difference between citizens and

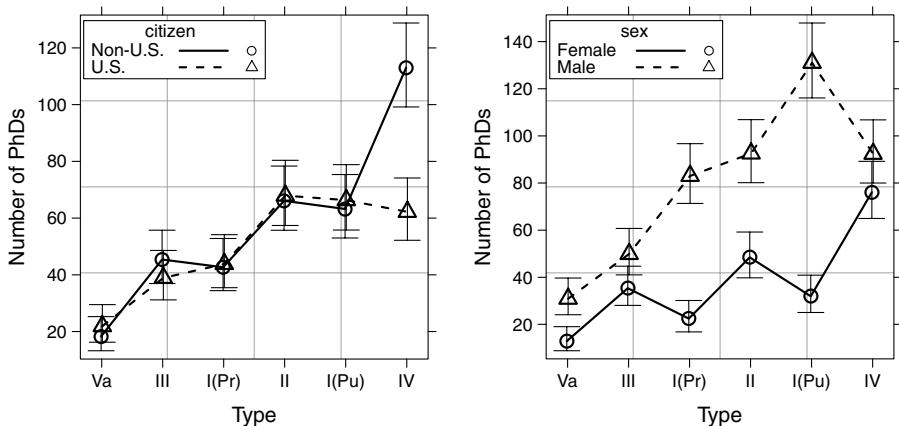


Figure 12.5 Effects plots for Mathematical Sciences PhDs.

noncitizens. Since the Poisson model uses a log-link, the expected number of U.S. citizen PhDs is $\exp(-0.6251) \approx 0.5$ times the expected non-U.S. citizens. The coefficients for the main effects are not easily interpretable. The difference between Male and Female is not reflected by the coefficient for Sex because this difference depends on the value of type.

A better way to understand the fitted model is to get estimated cell counts for each of the 24 cells based on the model, and then view them in the effects plots shown in Figure 12.5, one for each of the two-factor interactions. In both plots, the horizontal axis is levels of type. The levels have been ordered according to the total number PhD awards granted, as this makes the graphs easier to read. The vertical axis is the fitted number of PhDs. The Poisson model used a log-link, so an alternative of plotting log-fitted values on the vertical axis could have been used. The lines joining the points in the plot are just visual aids, as fitted values are available only at the points shown. The error bars are 95% confidence intervals, without adjusting for multiple inferences, for the estimated number of PhDs awarded.

Figure 12.5a is for the type:citizen interaction. The number of PhDs for citizens and noncitizens are essentially the same for all types of institutions except for Type IV, statistics and biostatistics programs, which have many more noncitizen PhDs awarded, although this difference is exaggerated because the vertical axis is not in log scale. The picture for the type:sex interaction is a little more complicated. Males outnumber females at all levels of type, except perhaps for Type IV. The sex differences vary by type, and are largest in the Type I public and private universities.

12.3.1 Goodness of Fit Tests

If a Poisson mean function is correctly specified, the residual deviance G^2 will be distributed as a $\chi^2(n - p')$ random variable, where n is the number of cells

and p' is the number of regressors fit. If the mean function is not correctly specified, or if the Poisson assumption is wrong, then G^2 will generally be too large, and so a lack of fit test can be obtained by comparing the value of G^2 to the relevant χ^2 distribution. For the model summarized in Table 12.6, the deviance is $G^2 = 1.96$, and when compared with the $\chi^2(6)$ distribution, we get a significance level of 0.92, suggesting no lack of fit of the model used.

The same idea can be used for binomial regression when the sample sizes m_i are larger than 1. For Table 12.1, we have $G^2 = 49.89$ with 33 df , corresponding to a p -value of 0.03, providing modest evidence of lack of fit. Since we found that adding s to that model improved the fit, finding that the initial model is inadequate is not surprising.

An alternative to using G^2 for lack of fit testing is to use Pearson's X^2 for testing, given by the familiar formula

$$X^2 = \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i} \quad (12.12)$$

Like G^2 , X^2 is compared with $\chi^2(n - p')$ to get significance levels. In large samples, the two tests will give the same inference, but in smaller samples χ^2 is generally more powerful.

In binomial regression with all or nearly all the $m_i = 1$, neither G^2 nor X^2 provides a lack of fit test.

12.4 TRANSFERRING WHAT YOU KNOW ABOUT LINEAR MODELS

Most of the methodology developed in this book transfers to problems with binomial or Poisson responses. In this section, important connections are briefly summarized.

12.4.1 Scatterplots and Regression

Graphing data, Chapter 1, is just as important in binomial and Poisson regression as it is in linear regression. In problems with a binary response, plots of the response versus predictors or regressors are generally not very helpful because the response only has two values. Smoothers, however, can help look at these plots as well. Plots of predictors with color used to indicate the level of the response can also be helpful.

12.4.2 Simple and Multiple Regression

The general ideas in Chapters 2 and 3 apply to binomial and Poisson models, even if the details differ. With the counted data models, estimates $\hat{\beta}$ and $\text{Var}(\hat{\beta}|X)$ are computed using the appropriate maximum likelihood methods,

not with the formulas in these chapters. Once these are found, they can be used in the formulas and methods given the text. For example, a point estimate and standard error for a linear combination of the elements of β is given by (3.26), but with $\hat{\sigma}^2$ set equal to 1, and $(\mathbf{X}'\mathbf{X})^{-1}$ replaced by the covariance matrix of $\hat{\beta}$ from the binomial or Poisson fit. Confidence intervals and tests use the standard normal rather than a t -distribution.

12.4.3 Model Building

Chapters 4 and 5 apply with modest modification in binomial and Poisson regression. Since both binomial and Poisson models use logarithms in their link functions, the results of Section 4.1.7 can be useful.

12.4.4 Testing and Analysis of Deviance

The t -tests discussed in Chapters 2, 3, and 6 are replaced by z -tests for binomial and Poisson models. The F -tests in Chapter 6 are replaced by χ^2 tests based on changes in deviance. The marginality principle, Section 6.2, is the guiding principle for testing with counted responses.

In linear models, the t -tests and F -tests for the same hypothesis have the same value, and so they are identical. With binomial and Poisson responses, the tests are identical only for very large samples, and in small samples they can give conflicting summaries. The G^2 tests are generally preferred.

12.4.5 Variances

Failure of the assumptions needed for binomial or Poisson fitting may be reflected in overdispersion, meaning that the variation between observations given the predictors is larger than the value required by the model. One general approach to overdispersion is to fit models that allow for it, such as the binomial or Poisson mixed models similar to those in Section 7.4. Other models, for example, using negative binomial distributions rather than binomial (Hilbe, 2011), can account for overdispersion. Alternatively, variance corrections like those in Section 7.2.1 are also available, and some software packages including Stata offer them as “robust” standard errors.

12.4.6 Transformations

Transformation of the response is not relevant with binomial and Poisson models. Transformation of predictors is relevant, however, and all the methodology in Chapter 8 can be used.

12.4.7 Regression Diagnostics

Many diagnostic methods depend on residuals. In binomial and Poisson models, the variance depends on the mean, and any useful residuals must be

scaled to account for variance. A generalization of the Pearson residuals defined in Section 9.1.3, is appropriate for most purposes. Fox and Weisberg (2011, chapter 6) provide examples of applying diagnostic methods for binomial and Poisson models.

12.4.8 Variable Selection

All the ideas discussed in Chapter 10 carry over to binomial and Poisson models.

12.5 GENERALIZED LINEAR MODELS

The multiple linear regression, logistic, and Poisson log-linear models are particular instances of *generalized linear models*. They share three basic characteristics:

1. The conditional distribution of the response $Y|X$ is distributed according to an *exponential family distribution*. The important members of this class include the normal, binomial, Poisson, and gamma distributions.
2. The response Y depends on the regressors only through the linear combination of terms $\beta'x$.
3. The mean $E(Y|X = \mathbf{x}) = m(\beta'x)$ for some kernel mean function m . For the multiple linear regression model, m is the identity function, and for logistic regression it is the logistic function. The Poisson was specified using the log link, so its m is the inverse of the log, or the exponential function. Other choices of the kernel mean function are possible but are used less often in practice.

These three components are enough to specify completely a regression problem along with methods for computing estimates and making inferences. The methodology for these models generally builds on the methods in this book, usually with only minor modification. Generalized linear models were first suggested by Nelder and Wedderburn (1972) and are discussed at length by McCullagh and Nelder (1989). Some statistical packages use common software to fit all generalized linear models, including the multiple linear regression model.

12.6 PROBLEMS

12.1 (Data file: Blowdown)

- 12.1.1 Create a table that gives the number of trees that survived and the number that died of each of the nine species.

- 12.1.2** Select the rows from the data file with `spp` equal to black spruce to get the data on the balsam fir trees only. Draw the graph of status `y` versus $\log(d)$, and add a smoother. Does the graph support fitting a logistic model?
- 12.1.3** Fit the same model as is used in Table 12.1, but fit the Bernoulli regression model by fitting to the individual trees. Show that the estimates and standard errors are identical to those in Table 12.1, but the deviance and df are different.
- 12.1.4** Add $(\log(d))^2$ to the mean function to allow for a possible decline in the probability of blowdown for the largest trees. Obtain the z -test that the coefficient for the quadratic term is 0, and also obtain the G^2 test for the same hypothesis. Show that these two tests are not identical, that is $G^2 \neq z^2$, and state the conclusions from the tests. Draw the effects plot for `d`; does the quadratic model allow for declining probabilities?

12.2 Professor ratings (Data file: `Ratprof`) Problem 6.10 concerned learning about quality rating as a function of several other predictors. In this problem, take as the response variable `pepper` with values yes and no, where yes means that the consensus of the raters is that the instructor is physically attractive. The predictors are `gender`, `discipline`, `quality`, `easiness`, and `raterInterst`. Find a set of regressors that appears to model the probability that `pepper = yes`, and summarize your results. (*Hint:* In some computer programs you may need to convert the values no and yes to 0 and 1. R will do this automatically, ordering the levels alphabetically.)

12.3 Downer data (Data file: `Downer`) For unknown reasons, dairy cows sometimes become recumbent—they lay down. Called *downers*, these cows may have a serious illness that may lead to their death. These data are from a study of blood samples of over 400 downer cows studied at the Ruakura New Zealand Animal Health Laboratory during 1983–1984. A variety of blood tests were performed, and for many of the animals, the outcome (survived, died) was determined. The goal is to see if survival can be predicted from the blood measurements. The variables in the data file are described in Table 12.7. These data were collected from veterinary records, and not all variables were recorded for all cows.

- 12.3.1** Consider first predicting `outcome` from `myopathy`. Find the fraction of surviving cows of `myopathy = 0` and for `myopathy = 1`.
- 12.3.2** Fit the logistic regression `outcome ~ myopathy`. Write a sentence that explains the meaning of each of the coefficient estimates, and provide 95% confidence intervals. Obtain the estimated probability of survival when `myopathy = 0` and when

Table 12.7 The Downer Data

Variable	<i>n</i>	Description
ast	429	Serum asparate amino transferase (U/L at 30°C)
calving	431	Factor with levels before and after calving
ck	413	Serum creatine phosphokinase (U/L at 30°C)
daysrec	432	Days recumbent when measurements were done
inflamat	136	Is inflammation present? no or yes
myopathy	222	Is muscle disorder present? a factor with levels absent and present
pcv	175	Packed cell volume (hematocrit), percentage
urea	266	Serum urea (mmol/l)
outcome	435	survived or died

Source: Clark et al. (1987).

myopathy = 1, and compare with the observed survival fractions in Problem 12.3.1.

- 12.3.3** Next, consider the regression problem with only *ck* as a predictor. Since *ck* is observed more often than is *myopathy*, this regression will be based on more cases than were used in the first two parts of this problem. Fit the logistic regression mean function with $\log(ck)$ as the only regressor beyond the intercept. Summarize results.
- 12.3.4** Fit the logistic mean function $y \sim \text{myopathy} + \log(ck) + \text{myopathy} : \log(ck)$. Obtain a Type II Analysis of Deviance table and summarize the results. Draw and interpret an effects plot, assuming the interaction is significant.

12.4 Starting with (12.7), derive (12.8).

12.5 Donner party (Data file: *Donner*) In the winter of 1846–1847, about 90 wagon train emigrants in the Donner party were unable to cross the Sierra Nevada Mountains of California before winter, and almost half of them starved to death. The data in file *Donner* from Johnson (1996) include some information about each of the members of the party. The variables include *age*, the age of the person; *sex*, whether male or female; *status*, whether the person was a member of a family group, a hired worker for one of the family groups, or a single individual who did not appear to be a hired worker or a member of any of the larger family groups; and *y*, a factor with levels *died* and *survived*.

- 12.5.1** How many men and women were in the Donner Party? What was the survival rate for each sex? Obtain a test that the survival rates were the same against the alternative that they were different. What do you conclude?

- 12.5.2** Fit the logistic regression model $y \sim \text{age}$, and provide an interpretation for the fitted coefficient for age.
- 12.5.3** Use your computer package to draw a scatterplot of the Pearson residuals from the model fit in Section 12.5 versus age. The residual plot will consist of two curves, with the curve of all positive values corresponding to survivors and the negative curve for deaths. As in Chapter 9, residual plots can be used to diagnose curvature, but this is quite hard without the aid of a smoother added to the plot.
- 12.5.4** Fit the logistic regression model $y \sim \text{age} + \text{age}^2 + \text{sex} + \text{status}$ and summarize results.

12.6 Challenger (Data file: `Challeng`) These data from Dalal et al. (1989) records performance of O-rings for the 23 U.S. space shuttle missions prior to the Challenger disaster of January 20, 1986. For each of the previous missions, the temperature at takeoff and the pressure of a prelaunch test were recorded, along with the number of O-rings that failed out of 6.

Use these data to try to understand the probability of failure as a function of temperature, and of temperature and pressure. Use your fitted model to estimate the probability of failure of an O-ring when the temperature was 31°F, the launch temperature on January 20, 1986.

12.7 Titanic (Data file: `Whiteststar`) The Titanic was a British luxury passenger liner that sank when it struck an iceberg about 640 km south of Newfoundland on April 14–15, 1912, on its maiden voyage to New York City from Southampton, England. Of 2201 known passengers and crew, only 711 are reported to have survived. These data from Dawson (1995) classify the people on board the ship according to their `sex` as male or female; `age`, either child or adult; and `class`, either first, second, third, or crew. Not all combinations of the three-factors occur in the data, since no children were members of the crew. For each `age`/`sex`/`class` combination, the number of people `m` and the number surviving `Surv` are also reported. The data are shown in Table 12.8.

- 12.7.1** Fit a logistic regression model with terms for factors `sex`, `age`, and `class`. On the basis of examination of the data in Table 12.8, explain why you expect that this mean function will be inadequate to explain these data.
- 12.7.2** Fit a logistic regression model that includes all the terms of the last part, plus all the two-factor interactions. Use appropriate testing procedures to decide if any of the two-factor interactions can be eliminated. Assuming that the mean function you have obtained matches the data well, summarize the results you have obtained by interpreting the parameters to describe different

Table 12.8 Data from the Titanic Disaster of 1912. Each Cell Gives `Surv/m`, the Number of Survivors, and the Number Number of People in the Cell

Class	Female		Male	
	Adult	Child	Adult	Child
Crew	20/23		192/862	
First	140/144	1/1	57/175	5/5
Second	80/93	13/13	14/168	11/11
Third	76/165	14/31	75/462	13/48

survival rates for various factor combinations. (*Hint:* How does the survival of the crew differ from the passengers? First class from third class? Males from females? Children versus adults? Did children in first class survive more often than children in third class?)

12.8 More Blowdown (Data file: `Blowdown`)

12.8.1 For the blowdown example, fit the model $y \sim \log(d) + s + \log(d) : s$ for `spp = paper birch` and summarize results.

12.8.2 Repeat for `spp = aspen`.

12.9 (Data file: `AMSSurvey`)

12.9.1 The example discussed in Section 12.3 concerns 2008–2009 PhDs in the mathematical sciences. Also included in the data file is an additional variable `count11` that gives the number of mathematical sciences PhDs in 2011–2012. Analyze these data to parallel the analysis in the text and summarize your results.

12.9.2 View these data as a four-dimensional table, with the dimensions `type`, `sex`, `citizen`, and `year`, where `year` is either 2008–2009 or 2011–2012. Fit models to this four-factor problem, and summarize results. (*Hint:* This will require that you reform the data file to have 48 rows. For example, in R, the user-unfriendly `reshape` function can do this.

```
> AMS1 <- reshape(AMSSurvey,
+ varying=c("count", "count11"),
+ v.names="y", times=c("2008-09", "2011-12"),
+ timevar="year", direction="long")
> AMS1$year <- factor(AMS1$year)
The variable year is now a factor and y is the count of PhDs.)
```

Appendix

A.1 WEBSITE

The web address for this book is <http://z.umn.edu/alr4ed>.

The website includes information about using R with this book, a description of an R package called `alr4` that includes all the data files described, and solutions to odd-numbered problems.

A.2 MEANS, VARIANCES, COVARIANCES, AND CORRELATIONS

Suppose we let u_1, u_2, \dots, u_n be n random variables.¹ Also let a_0, a_1, \dots, a_n be $n + 1$ known constants.

A.2.1 The Population Mean and E Notation

The symbol $E(u_i)$ is read as the expected value of the random variable u_i . The phrase “expected value” is the same as the phrase “mean value.” Informally, the expected value of u_i is the average value of a very large sample drawn from the distribution of u_i . If $E(u_i) = 0$, then the average value we would get for u_i if we sampled its distribution repeatedly is 0. Since u_i is a random variable, any particular realization of u_i is likely to be nonzero.

The expected value is a linear operator, which means

$$\begin{aligned} E(a_0 + a_1 u_1) &= a_0 + a_1 E(u_1) \\ E\left(a_0 + \sum a_i u_i\right) &= a_0 + \sum a_i E(u_i) \end{aligned} \tag{A.1}$$

¹Formally, we have random variables U_1, \dots, U_n and u_1, \dots, u_n that are realizations of the U_i , but we ignore here the distinction between a random variable and its realization.

For example, suppose all the u_i have the same expected value and we write $E(u_i) = \mu$, $i = 1, \dots, n$. The sample mean of the u_i is $\bar{u} = \sum u_i/n = \sum(1/n)u_i$, and the expected value of the sample mean is

$$E(\bar{u}) = E\left(\sum \frac{1}{n}u_i\right) = \frac{1}{n}\sum E(u_i) = \frac{1}{n}(n\mu) = \mu$$

We say that \bar{u} is an *unbiased* estimate of the population mean μ , since its expected value is μ .

A.2.2 Variance and Var Notation

The symbol $\text{Var}(u_i)$ is for the variance of u_i . The variance is defined by the equation $\text{Var}(u_i) = E[u_i - E(u_i)]^2$, the expected squared difference between an observed value for u_i and its mean value. The larger $\text{Var}(u_i)$, the more variable observed values for u_i are likely to be. The symbol σ^2 is often used for a variance, or σ_u^2 might be used for the variance of the identically distributed u_i if several variances are being discussed. The square root of a variance, often σ or σ_u , is the standard deviation, and is in the same units as the units of the random variable u_i . For example, if the u_i are heights in centimeters, then units of σ_u are also centimeters. The units of σ_u^2 are cm^2 , which can be much harder to interpret.

The general rule for the variance of a sum of *uncorrelated* random variables is

$$\text{Var}\left(a_0 + \sum a_i u_i\right) = \sum a_i^2 \text{Var}(u_i) \quad (\text{A.2})$$

The a_0 term vanishes because the variance of a constant is 0. Assuming that $\text{Var}(u_i) = \sigma^2$, we can find the variance of the sample mean of independently, identically distributed u_i :

$$\text{Var}(\bar{u}) = \text{Var}\left(\sum \frac{1}{n}u_i\right) = \frac{1}{n^2}\sum \text{Var}(u_i) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

The standard deviation of a sum is found by computing the variance of the sum and then taking a square root.

A.2.3 Covariance and Correlation

The symbol $\text{Cov}(u_i, u_j)$ is read as the covariance between the random variables u_i and u_j and is also an expected value defined by the equation

$$\text{Cov}(u_i, u_j) = E\{[u_i - E(u_i)][u_j - E(u_j)]\} = \text{Cov}(u_j, u_i)$$

The covariance describes the way two random variables vary jointly. If the two variables are independent, then $\text{Cov}(u_i, u_j) = 0$, but zero correlation does not imply independence. The variance is a special case of covariance, since $\text{Cov}(u_i, u_i) = \text{Var}(u_i)$.

When covariance is nonzero, common language is to say that two variables are *correlated*. Formally, the *correlation coefficient* is defined by

$$\rho(u_i, u_j) = \frac{\text{Cov}(u_i, u_j)}{\sqrt{\text{Var}(u_i)\text{Var}(u_j)}}$$

The correlation does not depend on units of measurement and has a value between -1 and 1 , with $\rho(u_i, u_j) = 0$ only if $\text{Cov}(u_i, u_j) = 0$.

The rule for covariances is

$$\text{Cov}(a_0 + a_1 u_1, a_3 + a_2 u_2) = a_1 a_2 \text{Cov}(u_1, u_2)$$

It is left as an exercise to show that

$$\rho(a_0 + a_1 u_1, a_3 + a_2 u_2) = \rho(u_1, u_2)$$

so the unit-free correlation coefficient does not change if the random variables are rescaled or centered.

The general form for the variance of a linear combination of random variables is

$$\text{Var}\left(a_0 + \sum_{i=1}^n a_i u_i\right) = \sum_{i=1}^n a_i^2 \text{Var}(u_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n a_i a_j \text{Cov}(u_i, u_j) \quad (\text{A.3})$$

A.2.4 Conditional Moments

Throughout the book, we use notation like $E(Y|X)$ or $E(Y|X = x)$ to denote the mean of the random variable Y in the population for which the value of X is fixed. Similarly, $\text{Var}(Y|X)$ or $\text{Var}(Y|X = x)$ is the variance of the random variable Y in the population for which X is fixed.

There are simple relationships between the *conditional* mean and variance of Y given X and the *unconditional* mean and variances (Casella and Berger, 2001):

$$E(Y) = E[E(Y|X)] \quad (\text{A.4})$$

$$\text{Var}(Y) = E[\text{Var}(Y|X)] + \text{Var}[E(Y|X)] \quad (\text{A.5})$$

For example, suppose that when we condition on the predictor X we have a simple linear regression mean function with constant variance, $E(Y|X=x) = \beta_0 + \beta_1 x$, $\text{Var}(Y|X=x) = \sigma^2$. In addition, suppose the unconditional

moments of the predictor are $E(X) = \mu_x$ and $\text{Var}(X) = \tau_x^2$. Then for the unconditional random variable Y ,

$$\begin{aligned} E(Y) &= E[E(Y|X=x)] \\ &= E[\beta_0 + \beta_1 x] \\ &= \beta_0 + \beta_1 \mu_x \end{aligned}$$

$$\begin{aligned} \text{Var}(Y) &= E[\text{Var}(Y|X=x)] + \text{Var}[E(Y|X=x)] \\ &= E[\sigma^2] + \text{Var}[\beta_0 + \beta_1 x] \\ &= \sigma^2 + \beta_1^2 \tau_x^2 \end{aligned}$$

The expected value of the unconditional variable Y is obtained by substituting the expected value of the unconditional variable X into the conditional expected value formula, and the unconditional variance of Y equals the conditional variance plus an additional quantity that depends on both β_1^2 and on τ_x^2 .

A.3 LEAST SQUARES FOR SIMPLE REGRESSION

The OLS estimates of β_0 and β_1 in simple regression are the values that minimize the residual sum of squares function,

$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (\text{A.6})$$

One method of finding the minimizer is to differentiate with respect to β_0 and β_1 , set the derivatives equal to 0, and solve

$$\begin{aligned} \frac{\partial \text{RSS}(\beta_0, \beta_1)}{\beta_0} &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial \text{RSS}(\beta_0, \beta_1)}{\beta_1} &= -2 \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{aligned}$$

Upon rearranging terms, we get

$$\begin{aligned} \beta_0 n + \beta_1 \sum x_i &= \sum y_i \\ \beta_0 \sum x_i + \beta_1 \sum x_i^2 &= \sum x_i y_i \end{aligned} \quad (\text{A.7})$$

Equations (A.7) are called the *normal equations* for the simple linear regression model (2.1). The normal equations depend on the data only through the sufficient statistics $\sum x_i$, $\sum y_i$, $\sum x_i^2$, and $\sum x_i y_i$. Using the formulas

$$\begin{aligned} S_{XX} &= \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 \\ S_{XY} &= \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y} \end{aligned} \quad (\text{A.8})$$

equivalent and numerically more stable sufficient statistics are given by \bar{x} , \bar{y} , S_{XX} , and S_{XY} . Solving (A.7), we get

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \quad (\text{A.9})$$

A.4 MEANS AND VARIANCES OF LEAST SQUARES ESTIMATES

The least squares estimates are linear combinations of the observed values y_1, \dots, y_n of the response, so we can apply the results of Appendix A.2 to the estimates found in Appendix A.3 to get the means, variances, and covariances of the estimates. Assume the simple regression model (2.1) is correct. The estimator $\hat{\beta}_1$ given at (A.9) can be written as $\hat{\beta}_1 = \sum c_i y_i$, where for each i , $c_i = (x_i - \bar{x})/S_{XX}$. Since we are conditioning on the values of X , the c_i are fixed numbers. By (A.1),

$$\begin{aligned} E(\hat{\beta}_1|X) &= E\left(\sum c_i y_i | X = x_i\right) = \sum c_i E(y_i | X = x_i) \\ &= \sum c_i (\beta_0 + \beta_1 x_i) \\ &= \beta_0 \sum c_i + \beta_1 \sum c_i x_i \end{aligned}$$

By direct summation, $\sum c_i = 0$ and $\sum c_i x_i = 1$, giving

$$E(\hat{\beta}_1|X) = \beta_1$$

which shows that $\hat{\beta}_1$ is unbiased for β_1 . A similar computation will show that $\hat{\beta}_0$ is an unbiased estimate of β_0 .

Since the y_i are assumed independent, the variance of $\hat{\beta}_1$ is found by an application of (A.2),

$$\begin{aligned} \text{Var}(\hat{\beta}_1|X) &= \text{Var}\left(\sum c_i y_i | X = x_i\right) \\ &= \sum c_i^2 \text{Var}(y_i | X = x_i) \\ &= \sigma^2 \sum c_i^2 \\ &= \sigma^2 / S_{XX} \end{aligned}$$

This computation also used $\sum c_i^2 = \sum (x_i - \bar{x})^2 / S_{XX}^2 = 1 / S_{XX}$. Computing the variance of $\hat{\beta}_0$ requires an application of (A.3). We write

$$\begin{aligned}\text{Var}(\hat{\beta}_0|X) &= \text{Var}(\bar{y} - \hat{\beta}_1 \bar{x}|X) \\ &= \text{Var}(\bar{y}|X) + \bar{x}^2 \text{Var}(\hat{\beta}_1|X) - 2\bar{x}\text{Cov}(\bar{y}, \hat{\beta}_1|X)\end{aligned}\quad (\text{A.10})$$

To complete this computation, we need to compute the covariance,

$$\begin{aligned}\text{Cov}(\bar{y}, \hat{\beta}_1|X) &= \text{Cov}\left(\frac{1}{n} \sum y_i, \sum c_i y_i|X\right) \\ &= \frac{1}{n} \sum c_i \text{Cov}(y_i, y_i|X) \\ &= \frac{\sigma^2}{n} \sum c_i \\ &= 0\end{aligned}$$

because the y_i are independent and $\sum c_i = 0$. Substituting into (A.10) and simplifying,

$$\text{Var}(\hat{\beta}_0|X) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right)$$

Finally,

$$\begin{aligned}\text{Cov}(\hat{\beta}_0, \hat{\beta}_1|X) &= \text{Cov}(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1|X) \\ &= \text{Cov}(\bar{y}, \hat{\beta}_1|X) - \bar{x} \text{Cov}(\hat{\beta}_1, \hat{\beta}_1|X) \\ &= 0 - \sigma^2 \frac{\bar{x}}{S_{XX}} \\ &= -\sigma^2 \frac{\bar{x}}{S_{XX}}\end{aligned}$$

Further application of these results gives the variance of a fitted value, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$:

$$\begin{aligned}\text{Var}(\hat{y}|X = x) &= \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x|X = x) \\ &= \text{Var}(\hat{\beta}_0|X = x) + x^2 \text{Var}(\hat{\beta}_1|X = x) + 2x\text{Cov}(\hat{\beta}_0, \hat{\beta}_1|X = x) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right) + \sigma^2 x^2 \frac{1}{S_{XX}} - 2\sigma^2 x \frac{\bar{x}}{S_{XX}} \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}} \right)\end{aligned}\quad (\text{A.11})$$

A prediction \tilde{y}_* at the future value x_* is just $\hat{\beta}_0 + \hat{\beta}_1 x_*$. The variance of a prediction consists of the variance of the fitted value at x_* given by (A.11) plus σ^2 , the variance of the error that will be attached to the future value,

$$\text{Var}(\tilde{y}_*|X = x_*) = \sigma^2 \left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\text{SXX}} \right) + \sigma^2$$

as given by (2.16).

A.5 ESTIMATING $E(Y|X)$ USING A SMOOTHER

For a 2D scatterplot of Y versus X , a scatterplot *smoother* provides an estimate of the mean function $E(Y|X = x)$ as x varies, without making parametric assumptions about the mean function. Many smoothing methods are used, and the smoother we use most often in this book is the simplest case of the `loess` smoother, Cleveland (1979); see also the first step in Algorithm 6.1.1 in Härdle (1990, p. 192). This smoother estimates $E(Y|X = x_g)$ by \tilde{y}_g via a weighted least squares (WLS) simple regression, giving more weight to points close to x_g than to points distant from x_g . Here is the method:

1. Select a value for a *smoothing parameter* f , a number between 0 and 1. Values of f close to 1 will give curves that are too smooth and will be close to a straight line, while small values of f give curves that are too rough and match all the wiggles in the data. The value of f must be chosen to balance the bias of oversmoothing with the variability of undersmoothing. Remarkably, for many problems $f \approx 2/3$ is a good choice. There is a substantial literature on the appropriate ways to estimate a smoothing parameter for `loess` and for other smoothing methods, but for the purposes of using a smoother to help us look at a graph, optimal choice of a smoothing parameter is not critical.
2. Find the fn closest points to x_g . For example, if $n = 100$, and $f = 0.6$, then find the $fn = 60$ closest points to x_g . Every time the value of x_g is changed, the points selected may change.
3. Among these fn nearest neighbors to x_g , compute the WLS estimates for the simple regression of $Y \sim X$, with weights determined so that points close to x_g have the highest weight, and the weights decline toward 0 for points farther from x_g . We use a triangular weight function that gives maximum weight to data at x_g , and weights that decrease linearly to 0 at the edge of the neighborhood. If a different weight function is used, answers are somewhat different.
4. The value of \tilde{y}_g is the fitted value at x_g from the WLS regression using the nearest neighbors found at step 2 as the data, and the weights from step 3.

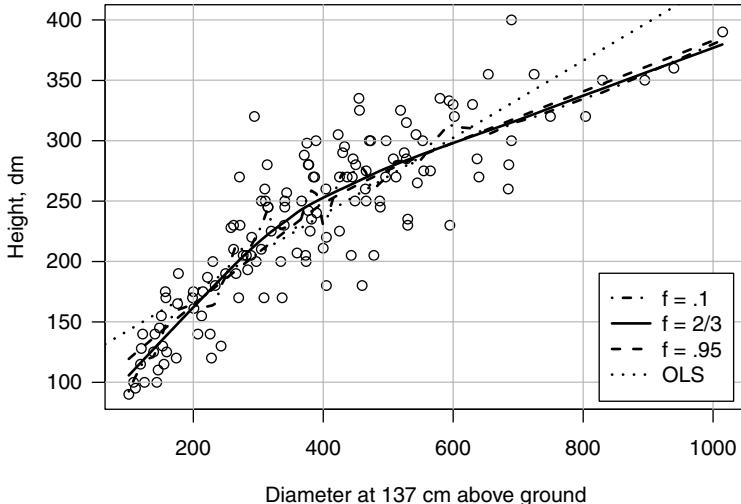


Figure A.1 Three choices of the smoothing parameter for a `loess` smooth. The data used in this plot are discussed in Section 8.1.2.

5. Repeat steps 1–4 for many values of x_g that form a grid of points that cover the interval on the x -axis of interest. Join the points.

Figure A.1 shows a plot of Height versus Diameter for western cedar trees in the Upper Flat Creek data, along with four smoothers. The first smoother is the `ols` simple regression line, which does not match the data well because the mean function for the data in this figure is probably curved, not straight. The `loess` smooth with $f = 0.1$ is as expected very wiggly, matching the local variation rather than the mean. The line for $f = 2/3$ seems to match the data very well, while the `loess` fit for $f = .95$ is nearly the same as for $f = 2/3$, but it tends toward oversmoothing and attempts to match the `ols` line. We would conclude from this graph that a straight-line mean function is likely to be inadequate because it does not match the data very well. Loader (2004) presents a bootstrap based lack-of-fit test based on comparing parametric and nonparametric estimates of the mean function.

The `loess` smoother is an example of a *nearest neighbor* smoother. Local polynomial regression smoothers and kernel smoothers are similar to `loess`, except they give positive weight to all cases within a fixed distance of the point of interest rather than a fixed number of points. There is a large literature on nonparametric regression, for which scatterplot smoothing is a primary tool. Recent reference on this subject include Simonoff (1996), Bowman and Azzalini (1997), and Loader (1999).

A.6 A BRIEF INTRODUCTION TO MATRICES AND VECTORS

We provide only a brief introduction to matrices and vectors. More complete references include Seber (2008), Schott (2005), or any good linear algebra book.

Boldface type is used to indicate matrices and vectors. We will say that \mathbf{X} is an $r \times c$ matrix if it is an array of numbers with r rows and c columns. A specific 4×3 matrix \mathbf{X} is

$$\mathbf{X} = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 1 & 5 \\ 1 & 3 & 4 \\ 1 & 8 & 6 \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \end{pmatrix} = (x_{ij}) \quad (\text{A.12})$$

The element x_{ij} of \mathbf{X} is the number in the i th row and the j th column. For example, in the preceding matrix, $x_{32} = 3$.

A *vector* is a matrix with just one column. A specific 4×1 matrix \mathbf{y} , which is a vector of length 4, is given by

$$\mathbf{y} = \begin{pmatrix} 2 \\ 3 \\ -2 \\ 0 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}$$

The elements of a vector are generally singly subscripted; thus, $y_3 = -2$. A *row vector* is a matrix with one row. We do not use row vectors in this book. If a vector is needed to represent a row, a transpose of a column vector will be used, Appendix A.6.4.

A *square matrix* has the same number of rows and columns, so $r = c$. A square matrix \mathbf{Z} is symmetric if $z_{ij} = z_{ji}$ for all i and j . A square matrix is *diagonal* if all elements off the main diagonal are 0, $z_{ij} = 0$, unless $i = j$. The matrices \mathbf{C} and \mathbf{D} below are symmetric and diagonal, respectively:

$$\mathbf{C} = \begin{pmatrix} 7 & 3 & 2 & 1 \\ 3 & 4 & 1 & -1 \\ 2 & 1 & 6 & 3 \\ 1 & -1 & 3 & 8 \end{pmatrix} \quad \mathbf{D} = \begin{pmatrix} 7 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 8 \end{pmatrix}$$

The diagonal matrix with all elements on the diagonal equal to 1 is called the identity matrix, for which the symbol \mathbf{I} is used. The 4×4 identity matrix is

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

A *scalar* is a 1×1 matrix, an ordinary number.

A.6.1 Addition and Subtraction

Two matrices can be added or subtracted only if they have the same number of rows and columns. The sum $\mathbf{C} = \mathbf{A} + \mathbf{B}$ of $r \times c$ matrices is also $r \times c$. Addition is done elementwise:

$$\mathbf{C} = \mathbf{A} + \mathbf{B} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix} = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \\ a_{31} + b_{31} & a_{32} + b_{32} \end{pmatrix}$$

Subtraction works the same way, with the “+” signs changed to “−” signs. The usual rules for addition of numbers apply to addition of matrices, namely commutativity, $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$, and associativity, $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$.

A.6.2 Multiplication by a Scalar

If k is a number and \mathbf{A} is an $r \times c$ matrix with elements (a_{ij}) , then $k\mathbf{A}$ is an $r \times c$ matrix with elements (ka_{ij}) . For example, the matrix $\sigma^2 \mathbf{I}$ has all diagonal elements equal to σ^2 and all off-diagonal elements equal to 0.

A.6.3 Matrix Multiplication

Multiplication of matrices follows rules that are more complicated than are the rules for addition and subtraction. For two matrices to be multiplied together in the order \mathbf{AB} , the number of columns of \mathbf{A} must equal the number of rows of \mathbf{B} . For example, if \mathbf{A} is $r \times c$, and \mathbf{B} is $c \times q$, then $\mathbf{C} = \mathbf{AB}$ is $r \times q$. If the elements of \mathbf{A} are (a_{ij}) and the elements of \mathbf{B} are (b_{ij}) , then the elements of $\mathbf{C} = (c_{ij})$ are given by the formula

$$c_{ij} = \sum_{k=1}^c a_{ik} b_{kj}$$

This formula says that c_{ij} is formed by taking the i th row of \mathbf{A} and the j th column of \mathbf{B} , multiplying the first element of the specified row in \mathbf{A} by the first element in the specified column in \mathbf{B} , multiplying second elements, and so on, and then adding the products together.

If **A** is $1 \times c$ and **B** is $c \times 1$, then the product **AB** is 1×1 , an ordinary number. For example, if **A** and **B** are

$$\mathbf{A} = (1 \ 3 \ 2 \ -1) \quad \mathbf{B} = \begin{pmatrix} 2 \\ 1 \\ -2 \\ 4 \end{pmatrix}$$

then the product **AB** is

$$\mathbf{AB} = (1 \times 2) + (3 \times 1) + (2 \times -2) + (-1 \times 4) = -3$$

AB is not the same as **BA**. For the preceding matrices, the product **BA** will be a 4×4 matrix:

$$\mathbf{BA} = \begin{pmatrix} 2 & 6 & 4 & -2 \\ 1 & 3 & 2 & -1 \\ -2 & -6 & -4 & 2 \\ 4 & 12 & 8 & -4 \end{pmatrix}$$

The following small example illustrates what happens when all the dimensions are bigger than 1. A 3×2 matrix **A** times a 2×2 matrix **B** is given as

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \\ a_{31}b_{11} + a_{32}b_{21} & a_{31}b_{12} + a_{32}b_{22} \end{pmatrix}$$

Using numbers, an example of multiplication of two matrices is

$$\begin{pmatrix} 3 & 1 \\ -1 & 0 \\ 2 & 2 \end{pmatrix} \begin{pmatrix} 5 & 1 \\ 0 & 4 \end{pmatrix} = \begin{pmatrix} 15+0 & 3+4 \\ -5+0 & -1+0 \\ 10+0 & 2+8 \end{pmatrix} = \begin{pmatrix} 15 & 4 \\ -5 & -1 \\ 10 & 10 \end{pmatrix}$$

In this example, **BA** is not defined because the number of columns of **B** is not equal to the number of rows of **A**. However, the associative law holds: If **A** is $r \times c$, **B** is $c \times q$, and **C** is $q \times p$, then $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$, and the result is an $r \times p$ matrix.

A.6.4 Transpose of a Matrix

The transpose of an $r \times c$ matrix **X** is a $c \times r$ matrix called **X'** such that if the elements of **X** are (x_{ij}) , then the elements of **X'** are (x_{ji}) . For the matrix **X** given at (A.12),

$$\mathbf{X}' = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & 3 & 8 \\ 1 & 5 & 4 & 6 \end{pmatrix}$$

The transpose of a column vector is a row vector. The transpose of a product $(\mathbf{AB})'$ is the product of the transposes, in opposite order, so $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.

Suppose that \mathbf{a} is an $r \times 1$ vector with elements a_1, \dots, a_r . Then the product $\mathbf{a}'\mathbf{a}$ will be a 1×1 matrix or scalar, given by

$$\mathbf{a}'\mathbf{a} = a_1^2 + a_2^2 + \dots + a_r^2 = \sum_{i=1}^r a_i^2 \quad (\text{A.13})$$

Thus, $\mathbf{a}'\mathbf{a}$ provides a compact notation for the sum of the squares of the elements of a vector \mathbf{a} . The square root of this quantity $(\mathbf{a}'\mathbf{a})^{1/2}$ is called the *norm* or *length* of the vector \mathbf{a} . Similarly, if \mathbf{a} and \mathbf{b} are both $r \times 1$ vectors, then we obtain

$$\mathbf{a}'\mathbf{b} = a_1 b_1 + a_2 b_2 + \dots + a_n b_n = \sum_{i=1}^r a_i b_i = \sum_{i=1}^r b_i a_i = \mathbf{b}'\mathbf{a}$$

The fact that $\mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a}$ is often quite useful in manipulating the vectors used in regression calculations.

Another useful formula in regression calculations is obtained by applying the distributive law

$$(\mathbf{a} - \mathbf{b})'(\mathbf{a} - \mathbf{b}) = \mathbf{a}'\mathbf{a} + \mathbf{b}'\mathbf{b} - 2\mathbf{a}'\mathbf{b} \quad (\text{A.14})$$

A.6.5 Inverse of a Matrix

For any real number $c \neq 0$, there is another number called the *inverse* of c , say d , such that the product $cd = 1$. For example, if $c = 3$, then $d = 1/c = 1/3$, and the inverse of 3 is 1/3. Similarly, the inverse of 1/3 is 3. The number 0 does not have an inverse because there is no other number d such that $0 \times d = 1$.

Square matrices can also have an inverse. We will say that the inverse of a matrix \mathbf{C} is another matrix \mathbf{D} , such that $\mathbf{CD} = \mathbf{I}$, and we write $\mathbf{D} = \mathbf{C}^{-1}$. Not all square matrices have an inverse. The collection of matrices that have an inverse are called full rank, invertible, or nonsingular. A square matrix that is not invertible is of less than full rank, or singular. If a matrix has an inverse, it has a unique inverse.

The inverse is easy to compute only in special cases, and its computation in general can require a very tedious calculation that is best done on a computer. High-level matrix and statistical languages such as Matlab, Maple, Mathematica and R include functions for inverting matrices, or returning an appropriate message if the inverse does not exist.

The identity matrix \mathbf{I} is its own inverse. If \mathbf{C} is a diagonal matrix, say

$$\mathbf{C} = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

then \mathbf{C}^{-1} is the diagonal matrix

$$\mathbf{C}^{-1} = \begin{pmatrix} \frac{1}{3} & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

as can be verified by direct multiplication. For any diagonal matrix with nonzero diagonal elements, the inverse is obtained by inverting the diagonal elements. If any of the diagonal elements are 0, then no inverse exists.

A.6.6 Orthogonality

Two vectors \mathbf{a} and \mathbf{b} of the same length are *orthogonal* if $\mathbf{a}'\mathbf{b} = 0$. An $r \times c$ matrix \mathbf{Q} has *orthonormal columns* if its columns, viewed as a set of $c \leq r$ different $r \times 1$ vectors, are orthogonal and in addition have length 1. This is equivalent to requiring that $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$, the $r \times r$ identity matrix. A square matrix \mathbf{A} is *orthogonal* if $\mathbf{A}'\mathbf{A} = \mathbf{A}\mathbf{A}' = \mathbf{I}$, and so $\mathbf{A}^{-1} = \mathbf{A}'$. For example, the matrix

$$\mathbf{A} = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \end{pmatrix}$$

can be shown to be orthogonal by showing that $\mathbf{A}'\mathbf{A} = \mathbf{I}$, and therefore

$$\mathbf{A}^{-1} = \mathbf{A}' = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{pmatrix}$$

A.6.7 Linear Dependence and Rank of a Matrix

Suppose we have a $n \times p$ matrix \mathbf{X} with columns given by the vectors $\mathbf{x}_1, \dots, \mathbf{x}_p$; we consider only the case $p \leq n$. We will say that $\mathbf{x}_1, \dots, \mathbf{x}_p$ are *linearly dependent* if we can find multipliers a_1, \dots, a_p , not all of which are 0, such that

$$\sum_{i=1}^p a_i \mathbf{x}_i = \mathbf{0} \quad (\text{A.15})$$

If no such multipliers exist, then we say that the vectors are *linearly independent*, and the matrix is *full rank*. In general, the *rank* of a matrix is the maximum number of \mathbf{x}_i that form a linearly independent set.

For example, the matrix \mathbf{X} given at (A.12) can be shown to have linearly independent columns because no a_i not all equal to zero can be found that satisfy (A.15). On the other hand, the matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 2 & 5 \\ 1 & 1 & 4 \\ 1 & 3 & 6 \\ 1 & 8 & 11 \end{pmatrix} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \quad (\text{A.16})$$

has linearly dependent columns and is singular because $\mathbf{x}_3 = 3\mathbf{x}_1 + \mathbf{x}_2$. The matrix has rank 2, because the linearly independent subset of the columns with the most elements has two elements.

The matrix $\mathbf{X}'\mathbf{X}$ is a $p \times p$ matrix. If \mathbf{X} has rank p , so does $\mathbf{X}'\mathbf{X}$. Full-rank square matrices always have an inverse. Square matrices of less than full rank never have an inverse.

A.7 RANDOM VECTORS

An $n \times 1$ vector \mathbf{Y} is a *random vector* if each of its elements is a random variable. The mean of an $n \times 1$ random vector \mathbf{Y} is also an $n \times 1$ vector whose elements are the means of the elements of \mathbf{Y} . The variance of an $n \times 1$ vector \mathbf{Y} is an $n \times n$ square symmetric matrix, often called a *covariance matrix*, written $\text{Var}(\mathbf{Y})$ with $\text{Var}(y_i)$ as its (i, i) element and $\text{Cov}(y_i, y_j) = \text{Cov}(y_j, y_i)$ as both the (i, j) and (j, i) element.

The rules for means and variances of random vectors are matrix equivalents of the scalar versions in Appendix A.2. If \mathbf{a}_0 is a vector of constants, and \mathbf{A} is a matrix of constants,

$$\mathbf{E}(\mathbf{a}_0 + \mathbf{A}\mathbf{Y}) = \mathbf{a}_0 + \mathbf{A}\mathbf{E}(\mathbf{Y}) \quad (\text{A.17})$$

$$\text{Var}(\mathbf{a}_0 + \mathbf{A}\mathbf{Y}) = \mathbf{A}\text{Var}(\mathbf{Y})\mathbf{A}' \quad (\text{A.18})$$

A.8 LEAST SQUARES USING MATRICES

The multiple linear regression model can be written as

$$E(Y|X = \mathbf{x}) = \boldsymbol{\beta}'\mathbf{x} \quad \text{Var}(Y|X = \mathbf{x}) = \sigma^2$$

The matrix version is

$$E(\mathbf{Y}|X) = \mathbf{X}\boldsymbol{\beta} \quad \text{Var}(\mathbf{Y}|X) = \sigma^2 \mathbf{I}$$

where \mathbf{Y} is the $n \times 1$ vector of response values and \mathbf{X} is a $n \times p'$ matrix. If the mean function includes an intercept, then the first column of \mathbf{X} is a vector of ones, and $p' = p + 1$. If the mean function does not include an intercept, then the column of one is not included in \mathbf{X} and $p' = p$. The i th row of the $n \times p'$ matrix \mathbf{X} is \mathbf{x}_i' , $\boldsymbol{\beta}$ is a $p' \times 1$ vector of parameters for the mean function.

The OLS estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is given by the arguments that minimize the residual sum of squares function,

$$\text{RSS}(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Using (A.14)

$$\text{RSS}(\boldsymbol{\beta}) = \mathbf{Y}'\mathbf{Y} + \boldsymbol{\beta}'(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} \quad (\text{A.19})$$

$\text{RSS}(\boldsymbol{\beta})$ depends on only three functions of the data: $\mathbf{Y}'\mathbf{Y}$, $\mathbf{X}'\mathbf{X}$, and $\mathbf{Y}'\mathbf{X}$. Any two data sets that have the same values of these three quantities will have the same least squares estimates. Using (A.8), the information in these quantities is equivalent to the information contained in the sample means of the regressors plus the sample covariances of the regressors and the response.

To minimize (A.19), differentiate with respect to $\boldsymbol{\beta}$ and set the result equal to 0. This leads to the matrix version of the normal equations,

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y} \quad (\text{A.20})$$

The OLS estimates are any solution to these equations. If the inverse of $(\mathbf{X}'\mathbf{X})$ exists, as it will if the columns of \mathbf{X} are linearly independent, the OLS estimates are unique and are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (\text{A.21})$$

If the inverse does not exist, then the matrix $(\mathbf{X}'\mathbf{X})$ is of less than full rank, and the OLS estimate is not unique. In this case, most computer programs will use a linearly independent subset of the columns of \mathbf{X} in fitting the model, so that the reduced model matrix does have full rank. This is discussed in Section 4.1.4.

A.8.1 Properties of Estimates

Using the rules for means and variances of random vectors, (A.17) and (A.18), we find

$$\begin{aligned} E(\hat{\beta}|\mathbf{X}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}|\mathbf{X}] \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']E(\mathbf{Y}|\mathbf{X}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta \\ &= \beta \end{aligned} \quad (\text{A.22})$$

so $\hat{\beta}$ is unbiased for β , as long as the mean function that was fit is the true mean function. The variance of $\hat{\beta}$ is

$$\begin{aligned} \text{Var}(\hat{\beta}|\mathbf{X}) &= \text{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}|\mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\text{Var}(\mathbf{Y}|\mathbf{X})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\sigma^2\mathbf{I}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned} \quad (\text{A.23})$$

The variances and covariances are compactly determined as σ^2 times a matrix whose elements are determined only by \mathbf{X} and not by \mathbf{Y} .

A.8.2 The Residual Sum of Squares

Let $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ be the $n \times 1$ vector of fitted values corresponding to the n cases in the data, and $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}}$ is the vector of residuals. One representation of the residual sum of squares, which is the residual sum of squares function evaluated at $\hat{\beta}$, is

$$\text{RSS} = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = \hat{\mathbf{e}}'\hat{\mathbf{e}} = \sum_{i=1}^n \hat{e}_i^2$$

which suggests that the residual sum of squares can be computed by squaring the residuals and adding them up. In multiple linear regression, it can also be computed more efficiently on the basis of summary statistics. Using (A.19) and the summary statistics $\mathbf{X}'\mathbf{X}$, $\mathbf{X}'\mathbf{Y}$, and $\mathbf{Y}'\mathbf{Y}$, we write

$$\text{RSS} = \text{RSS}(\hat{\beta}) = \mathbf{Y}'\mathbf{Y} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} - 2\mathbf{Y}'\mathbf{X}\hat{\beta}$$

We will first show that $\hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{Y}'\mathbf{X}\hat{\beta}$. Substituting for one of the $\hat{\beta}$ s, we get

$$\hat{\beta}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \hat{\beta}'\mathbf{X}'\mathbf{Y} = \mathbf{Y}'\mathbf{X}\hat{\beta}$$

The last result follows because taking the transpose of a 1×1 matrix does not change its value. The residual sum of squares function can now be rewritten as

$$\begin{aligned} \text{RSS} &= \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{Y}}'\hat{\mathbf{Y}} \end{aligned}$$

where $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ are the fitted values. The residual sum of squares is the difference in the squares of the lengths of the two vectors \mathbf{Y} and $\hat{\mathbf{Y}}$. Another useful form for the residual sum of squares is

$$\text{RSS} = \mathbf{S}\mathbf{Y}\mathbf{Y}(1 - R^2)$$

where R^2 is the square of the sample correlation between $\hat{\mathbf{Y}}$ and \mathbf{Y} .

A.8.3 Estimate of Variance

Under the assumption of constant variance, the estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\text{RSS}}{d} \quad (\text{A.24})$$

with d df, where d is equal to the number of cases n minus the number of regressors with estimated coefficients in the model. If the matrix \mathbf{X} is of full rank, then $d = n - p'$, where $p' = p$ for mean functions without an intercept, and $p' = p + 1$ for mean functions with an intercept. The number of estimated coefficients will be less than p' if \mathbf{X} is not of full rank.

A.8.4 Weighted Least Squares

From Section 7.1, the WLS model can be written in matrix notation as

$$\mathbf{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \quad \text{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2 \mathbf{W}^{-1} \quad (\text{A.25})$$

To distinguish OLS and WLS results, we will use a subscript W on several quantities. In practice, there is no need to distinguish between OLS and WLS, and this subscript is dropped elsewhere in the book.

- The WLS estimator $\hat{\boldsymbol{\beta}}_W$ of $\boldsymbol{\beta}$ is given by the arguments that minimize the residual sum of squares function,

$$\begin{aligned} \text{RSS}_W(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}'\mathbf{W}\mathbf{Y} + \boldsymbol{\beta}'(\mathbf{X}'\mathbf{W}\mathbf{X})\boldsymbol{\beta} - 2\mathbf{Y}'\mathbf{W}\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

- The wls estimator solves the weighted normal equations

$$\mathbf{X}'\mathbf{WX}\boldsymbol{\beta} = \mathbf{X}'\mathbf{WY}$$

- The wls estimate is

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{X}'\mathbf{WX})^{-1}\mathbf{X}'\mathbf{WY} \quad (\text{A.26})$$

- $\hat{\boldsymbol{\beta}}_w$ is unbiased:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_w | \mathbf{X}) &= E[(\mathbf{X}'\mathbf{WX})^{-1}\mathbf{X}'\mathbf{WY} | \mathbf{X}] \\ &= (\mathbf{X}'\mathbf{WX})^{-1}\mathbf{X}'\mathbf{W}E(\mathbf{Y} | \mathbf{X}) \\ &= (\mathbf{X}'\mathbf{WX})^{-1}\mathbf{X}'\mathbf{WX}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned} \quad (\text{A.27})$$

- The variance of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}_w | \mathbf{X}) &= \text{Var}((\mathbf{X}'\mathbf{WX})^{-1}\mathbf{X}'\mathbf{WY} | \mathbf{X}) \\ &= (\mathbf{X}'\mathbf{WX})^{-1}\mathbf{X}'\mathbf{W}[\text{Var}(\mathbf{Y} | \mathbf{X})]\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{WX})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}[\sigma^2\mathbf{W}^{-1}]\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{WX})^{-1} \end{aligned} \quad (\text{A.28})$$

- The RSS_w can be computed from

$$\text{RSS}_w = \mathbf{Y}'\mathbf{WY} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{WX}\hat{\boldsymbol{\beta}}$$

- The estimated variance is

$$\hat{\sigma}^2 = \frac{\text{RSS}_w}{d} \quad (\text{A.29})$$

with d df , where d is equal to the number of cases n minus the number of regressors with estimated coefficients in the model.

- Confidence intervals are the same for both ols and wls as long as (A.28) and (A.29) are used. Testing procedures in Chapter 6 are the same with ols and wls subject to the changes described here. In particular, standard computer programs produce output that will look the same with ols and wls and the output can be interpreted similarly.

A.9 THE QR FACTORIZATION

Most of the formulas given in this book are convenient for derivations but can be inaccurate when used on a computer because inverting a matrix such as

$(\mathbf{X}'\mathbf{X})$ leaves open the possibility of introducing significant rounding errors into calculations. Most statistical packages will use better methods of computing, and understanding how they work is useful.

We start with the basic $n \times p'$ matrix \mathbf{X} of regressors. Suppose we could find an $n \times p'$ matrix \mathbf{Q} and a $p' \times p'$ matrix \mathbf{R} such that (1) $\mathbf{X} = \mathbf{QR}$; (2) \mathbf{Q} has orthonormal columns, meaning that $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_{p'}$; and (3) \mathbf{R} is an upper triangular matrix, meaning that all the entries in \mathbf{R} below the diagonal are equal to 0, but those on or above the diagonal can be nonzero.

Using the basic properties of matrices, we can write

$$\mathbf{X} = \mathbf{QR}$$

$$\mathbf{X}'\mathbf{X} = (\mathbf{QR})'(\mathbf{QR}) = \mathbf{R}'\mathbf{R}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{R}'\mathbf{R})^{-1} = \mathbf{R}^{-1}(\mathbf{R}')^{-1} \quad (\text{A.30})$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{R}^{-1}(\mathbf{Q}'\mathbf{Y}) \quad (\text{A.31})$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{QQ}' \quad (\text{A.32})$$

Equation (A.30) follows because \mathbf{R} is a square matrix, and the inverse of the product of square matrices is the product of the inverses in opposite order. From (A.31), to compute $\hat{\boldsymbol{\beta}}$, first compute $\mathbf{Q}'\mathbf{Y}$, which is a $p' \times 1$ vector, and multiply on the left by \mathbf{R} to get

$$\mathbf{R}\hat{\boldsymbol{\beta}} = \mathbf{Q}'\mathbf{Y} \quad (\text{A.33})$$

This last equation is very easy to solve because \mathbf{R} is a triangular matrix and so we can use *backsolving*. For example, to solve the equations

$$\begin{pmatrix} 7 & 4 & 2 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix} \hat{\boldsymbol{\beta}} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$$

first solve the last equation, so $\hat{\beta}_3 = 1$, substitute into the equation above it, so $2\hat{\beta}_2 + 1 = 2$, so $\hat{\beta}_2 = 1/2$. Finally, the first equation is $7\hat{\beta}_1 + 2 + 2 = 3$, so $\hat{\beta}_1 = -1/7$.

Equation (A.32) shows how the elements of the $n \times n$ hat matrix \mathbf{H} can be computed without inverting a matrix, and without using all the storage needed to save \mathbf{H} in full. If \mathbf{q}_i is the i th column of \mathbf{Q} , then an element h_{ij} of the \mathbf{H} matrix is simply computed as $h_{ij} = \mathbf{q}'_i \mathbf{q}_j$.

Golub and Van Loan (1996) provide a complete treatment on computing and using the **QR** factorization. Very high-quality computer code for computing this and related quantities for statistics is provided in the publicly available Lapack package, described on the internet at <http://www.netlib.org/lapack/lug/>. This code is also used in many standard statistical packages.

A.10 SPECTRAL DECOMPOSITION

The spectral decomposition provides a very useful representation of a square symmetric matrix (Schott, 2005; Christensen, 2011; Golub and Van Loan, 1996). Suppose \mathbf{S} is a $p \times p$ symmetric matrix. Then the spectral theorem says that there exists a matrix \mathbf{U} that is $p \times p$ and orthogonal, so $\mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathbf{I}$, and a diagonal matrix \mathbf{D} with diagonal elements $d_1 \geq d_2 \geq \dots \geq d_p \geq 0$ such that

$$\mathbf{S} = \mathbf{UDU}' \quad (\text{A.34})$$

The d_j are called the *eigenvalues* of \mathbf{S} , and the columns $(\mathbf{u}_1', \dots, \mathbf{u}_p')$ of \mathbf{U} are called the corresponding eigenvectors. The eigenvectors are unique if all the eigenvalues are unequal. The number of nonzero eigenvalues of \mathbf{S} is equal to the rank of \mathbf{S} . If all the eigenvalues are positive, then

$$\mathbf{S}^{-1} = \mathbf{U}(\mathbf{D})^{-1}\mathbf{U}'$$

This is particularly useful in computations because inverting \mathbf{S} requires only inverting a diagonal matrix \mathbf{D} .

Equation (A.34) can be rewritten in scalar form as

$$\mathbf{S} = \sum_{j=1}^p d_j \mathbf{u}_j \mathbf{u}_j'$$

For any vector \mathbf{a} with $\mathbf{a}'\mathbf{a} = 1$,

$$\mathbf{a}'\mathbf{S}\mathbf{a} = \sum_{j=1}^p d_j (\mathbf{a}'\mathbf{u}_j)(\mathbf{u}_j'\mathbf{a}) = \sum_{j=1}^p d_j (\mathbf{a}'\mathbf{u}_j)^2$$

Now for each j , $(\mathbf{a}'\mathbf{u}_j)^2$ is bounded between 0 and 1. If we set $\mathbf{a} = \mathbf{u}_1$, then $(\mathbf{a}'\mathbf{u}_1) = (\mathbf{u}_1'\mathbf{u}_1) = 1$, and $(\mathbf{a}'\mathbf{u}_j) = (\mathbf{u}_1'\mathbf{u}_j) = 0$ for all $j > 1$ because \mathbf{U} is an orthogonal matrix. For this case the sum in the last equation is equal to d_1 , and this is the largest possible value of $\mathbf{a}'\mathbf{S}\mathbf{a}$.

A.11 MAXIMUM LIKELIHOOD ESTIMATES

A.11.1 Linear Models

Maximum likelihood estimation is probably the most frequently used method of deriving estimates in statistics. A general treatment is given by Casella and Berger (2001, section 7.2.2); here we derive the maximum likelihood estimates for the linear regression model assuming normality, without proof or much explanation. Our goal is to establish notation and define quantities that will

be used in the discussion of Box–Cox transformations, and estimation for generalized linear models in Chapter 12.

The normal multiple linear regression model specifies for the i th observation that

$$(y_i | \mathbf{x}_i) \sim N(\boldsymbol{\beta}' \mathbf{x}_i, \sigma^2)$$

Given this model, the density for the i th observation y_i is the normal density function,

$$f_{y_i}(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \boldsymbol{\beta}' \mathbf{x}_i)^2}{2\sigma^2}\right)$$

Assuming the observations are independent, the likelihood function is just the product of the densities for each of the n observations, viewed as a function of the parameters with the data fixed rather than a function of the data with the parameters fixed:

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2 | Y) &= \prod_{i=1}^n f_{y_i}(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \boldsymbol{\beta}' \mathbf{x}_i)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \boldsymbol{\beta}' \mathbf{x}_i)^2\right) \end{aligned}$$

The maximum likelihood estimates are simply the values of $\boldsymbol{\beta}$ and σ^2 that maximize the likelihood function.

The values that maximize the likelihood will also maximize the logarithm of the likelihood

$$\log[L(\boldsymbol{\beta}, \sigma^2 | Y)] = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \boldsymbol{\beta}' \mathbf{x}_i)^2 \quad (\text{A.35})$$

The log-likelihood function (A.35) is a sum of three terms. Since $\boldsymbol{\beta}$ is included only in the third term and this term has a negative sign in front of it, we recognize that maximizing the log-likelihood over $\boldsymbol{\beta}$ is the same as minimizing the third term, which, apart from constants, is the same as the residual sum of squares function (see Section 3.4.3). We have just shown that the maximum likelihood estimate of $\boldsymbol{\beta}$ for the normal linear regression problem is the same as the OLS estimator. Fixing $\boldsymbol{\beta}$ at the OLS estimator $\hat{\boldsymbol{\beta}}$, (A.35) becomes

$$\log[L(\hat{\boldsymbol{\beta}}, \sigma^2 | Y)] = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \text{RSS} \quad (\text{A.36})$$

and differentiating (A.36) with respect to σ^2 and setting the result to 0 gives the maximum likelihood estimator for σ^2 as RSS/n , the same estimate we have been using, apart from division by n rather than $n - p'$.

Maximum likelihood estimation has many important properties that make them useful. These estimates are approximately normally distributed in large samples, and the large sample variance achieves the lower bound for the variance of all unbiased estimates.

A.11.2 Logistic Regression

In logistic regression we have (y_1, \dots, y_n) independent with $y_i \sim \text{Bin}(m_i, \theta(\mathbf{x}))$. The likelihood based on (y_1, \dots, y_n) is obtained by multiplying the likelihood for each observation,

$$\begin{aligned} L &= \prod_{i=1}^n \binom{m_i}{y_i} (\theta(\mathbf{x}_i))^{y_i} (1 - \theta(\mathbf{x}_i))^{m_i - y_i} \\ &\propto \prod_{i=1}^n (\theta(\mathbf{x}_i))^{y_i} (1 - \theta(\mathbf{x}_i))^{m_i - y_i} \end{aligned}$$

In the last expression, we have dropped the binomial coefficients $\binom{m_i}{y_i}$ because they do not depend on parameters. After minor rearranging, the log-likelihood is

$$\log(L) \propto \sum_{i=1}^n \left[y_i \log\left(\frac{\theta(\mathbf{x}_i)}{1 - \theta(\mathbf{x}_i)}\right) + m_i \log(1 - \theta(\mathbf{x}_i)) \right]$$

Next, we substitute for $\theta(\mathbf{x}_i)$ using Equation (12.8) to get

$$\log(L(\boldsymbol{\beta})) = \sum_{i=1}^n [(\boldsymbol{\beta}' \mathbf{x}_i) y_i - m_i \log(1 + \exp(\boldsymbol{\beta}' \mathbf{x}_i))] \quad (\text{A.37})$$

The log-likelihood depends on the regression parameters $\boldsymbol{\beta}$ explicitly, and we can maximize (A.37) to get estimates. An iterative procedure is required. Most computer packages use the *Fisher scoring* algorithm for the computing, which amounts to a sequence of weighted least squares computations with the weights depending on the estimates (Fox and Weisberg, 2011, section 5.12). The more general *Newton–Raphson* algorithm can also be used. Details of the computational method are provided by McCullagh and Nelder (1989, section 2.5), Collett (2003, section 3.12), Hosmer et al. (2013), and Agresti (2013), among others.

The estimated covariance matrix of the estimates is given by

$$\text{Var}(\hat{\beta}) = (\mathbf{X}' \hat{\mathbf{W}} \mathbf{X})^{-1}$$

where $\hat{\mathbf{W}}$ is a diagonal matrix with entries $m_i \hat{\theta}(\mathbf{x}_i)(1 - \hat{\theta}(\mathbf{x}_i))$, and \mathbf{X} is a matrix with i th row \mathbf{x}' .

A.12 THE BOX-COX METHOD FOR TRANSFORMATIONS

A.12.1 Univariate Case

Box and Cox (1964) derived the Box–Cox method for selecting a transformation using a likelihood-like method. They supposed that, for some value of λ , $\psi_M(Y, \lambda)$ given by (8.5) in Section 8.1.3, is normally distributed. With n independent observations, therefore, the log-likelihood function for $(\beta, \sigma^2, \lambda)$ is given by (A.35), but with y_i replaced by $\psi_M(y_i, \lambda)$,²

$$\log(L(\hat{\beta}, \sigma^2, \lambda | Y)) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (\psi_M(y_i, \lambda) - \beta' \mathbf{x}_i)^2 \quad (\text{A.38})$$

For a fixed value of λ , (A.38) is the same as (A.35), and so the maximum likelihood estimates for β and σ^2 are obtained from the regression of $\psi_M(Y, \lambda)$ on X , and the value of the log-likelihood evaluated at these estimates is

$$\log(L(\beta(\lambda), \sigma^2(\lambda), \lambda | Y)) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\text{RSS}(\lambda)/n) - \frac{n}{2} \quad (\text{A.39})$$

where $\text{RSS}(\lambda)$ is the residual sum of squares in the regression of $\psi_M(Y, \lambda)$ on X , as defined in Section 8.1.3. Only the second term in (A.39) involves data, and so the global maximum likelihood estimate of λ minimizes $\text{RSS}(\lambda)$.

Standard likelihood theory can be applied to get a $(1 - \alpha) \times 100\%$ confidence interval for λ to be the set

$$\left\{ \lambda | 2 \left[\log(L(\beta(\hat{\lambda}), \sigma^2(\hat{\lambda}), \hat{\lambda} | Y)) - \log(L(\beta(\lambda), \sigma^2(\lambda), \lambda | Y)) \right] < \chi^2(1, 1 - \alpha) \right\}$$

Or, setting $\alpha = .05$ so $\chi^2(1, .95) = 3.84$, and using (A.39)

$$\left\{ \lambda | (n/2)(\log(\text{RSS}(\lambda)) - \log(\text{RSS}(\hat{\lambda}))) < 1.92 \right\} \quad (\text{A.40})$$

²As λ is varied, the *units* of $\psi_M(Y, \lambda)$ can change, and so the joint density of the transformed data should require a Jacobian term; see Casella and Berger (2001, section 4.3). The modified power transformations are defined so the Jacobian of the transformation is always equal to 1, and it can therefore be ignored.

Many statistical packages will have routines that will provide a graph of $\text{RSS}(\lambda)$ versus λ , or of $(n/2) \log(\text{RSS}(\lambda))$ versus λ as shown in Figure 8.7, for the highway accident data. Equation (A.40) shows that the confidence interval for λ includes all values of λ for which the log-likelihood is within 1.92 units of the maximum value of the log-likelihood, or between the two vertical lines in the figure.

A.12.2 Multivariate Case

Although the material in this section uses more mathematical statistics than most of this book, it is included because the details of computing the multivariate extension of Box–Cox transformations are not published elsewhere. The basic idea was proposed by Velilla (1993).

Suppose X is a set of p variables we wish to transform and define

$$\psi_M(X, \boldsymbol{\lambda}) = (\psi_M(X_1, \lambda_1), \dots, \psi_M(X_k, \lambda_k))$$

We have used the modified power transformations (8.5) for each element of X , but the same general idea can be applied using other transformations such as the Yeo–Johnson family introduced in Section 8.4. In analogy to the univariate case, we assume that for some $\boldsymbol{\lambda}$, we will have

$$\psi_M(X, \boldsymbol{\lambda}) \sim N(\boldsymbol{\mu}, \mathbf{V})$$

where \mathbf{V} is an unknown positive definite symmetric matrix that needs to be estimated. If \mathbf{x}_i is the observed value of X for the i th observation, then the likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\mu}, \mathbf{V}, \boldsymbol{\lambda}|X) &= \prod_{i=1}^n \frac{1}{(2\pi|\mathbf{V}|)^{1/2}} \\ &\times \exp\left(-\frac{1}{2}(\psi_M(\mathbf{x}_i, \boldsymbol{\lambda}) - \boldsymbol{\mu})' \mathbf{V}^{-1} (\psi_M(\mathbf{x}_i, \boldsymbol{\lambda}) - \boldsymbol{\mu})\right) \end{aligned} \quad (\text{A.41})$$

where $|\mathbf{V}|$ is the determinant.³ After rearranging terms, the log-likelihood is given by

$$\begin{aligned} \log(L(\boldsymbol{\mu}, \mathbf{V}, \boldsymbol{\lambda}|X)) &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(|\mathbf{V}|) \\ &- \frac{1}{2} \sum_{i=1}^n \mathbf{V}^{-1} (\psi_M(\mathbf{x}_i, \boldsymbol{\lambda}) - \boldsymbol{\mu}) (\psi_M(\mathbf{x}_i, \boldsymbol{\lambda}) - \boldsymbol{\mu})' \end{aligned} \quad (\text{A.42})$$

³The determinant is defined in any linear algebra textbook.

If we fix λ , then (A.42) is the standard log-likelihood for the multivariate normal distribution. The values of \mathbf{V} and μ that maximize (A.42) are the sample mean and sample covariance matrix, the latter with divisor n rather than $n - 1$,

$$\begin{aligned}\mu(\lambda) &= \frac{1}{n} \sum_{i=1}^n \psi_M(\mathbf{x}_i, \lambda) \\ \mathbf{V}(\lambda) &= \frac{1}{n} \sum_{i=1}^n (\psi_M(\mathbf{x}_i, \lambda) - \mu(\lambda))(\psi_M(\mathbf{x}_i, \lambda) - \mu(\lambda))'\end{aligned}$$

Substituting these estimates into (A.42) gives the profile log-likelihood for λ ,

$$\log(L(\mu(\lambda), \mathbf{V}(\lambda), \lambda | X)) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(|\mathbf{V}(\lambda)|) - \frac{n}{2} \quad (\text{A.43})$$

This equation will be maximized by minimizing the determinant of $\mathbf{V}(\lambda)$ over values of λ . This is a numerical problem for which there is no closed-form solution, but it can be solved using a general-purpose function minimizer.

Standard theory for maximum likelihood estimates can provide tests concerning λ and standard errors for the elements of λ . To test the hypothesis that $\lambda = \lambda_0$ against a general alternative, compute

$$G^2 = 2 \left[\log(L(\mu(\hat{\lambda}), \mathbf{V}(\hat{\lambda}), \hat{\lambda})) - \log(L(\mu(\lambda_0), \mathbf{V}(\lambda_0), \lambda_0)) \right]$$

and compare G^2 with a chi-squared distribution with p df. The standard error of $\hat{\lambda}$ is obtained from the inverse of the expected information matrix evaluated at $\hat{\lambda}$. The expected information for λ is just the matrix of second derivatives of (A.43) with respect to λ evaluated at $\hat{\lambda}$. Many optimization routines, such as `optim` in R, will return the matrix of estimated second derivatives if requested; all that is required is inverting this matrix, and then the square roots of the diagonal elements are the estimated standard errors.

A.13 CASE DELETION IN LINEAR REGRESSION

Suppose \mathbf{X} is the $n \times p'$ matrix of regressors with linearly independent columns. We use the subscript “ (i) ” to mean “without case i ,” so that $\mathbf{X}_{(i)}$ is an $(n - 1) \times p'$ matrix. We can compute $(\mathbf{X}'_{(i)} \mathbf{X}_{(i)})^{-1}$ from the remarkable formula

$$(\mathbf{X}'_{(i)} \mathbf{X}_{(i)})^{-1} = (\mathbf{X}' \mathbf{X})^{-1} + \frac{(\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}'_i (\mathbf{X}' \mathbf{X})^{-1}}{1 - h_{ii}} \quad (\text{A.44})$$

where $h_{ii} = \mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ is the i th leverage value, a diagonal value from the hat matrix. This formula was used by Gauss (1821); a history of it and many variations is given by Henderson and Searle (1981). It can be applied to give all the results that one would want relating multiple linear regression with and without the i th case. For example,

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \hat{e}_i}{1-h_{ii}} \quad (\text{A.45})$$

Writing $r_i = \hat{e}_i / \hat{\sigma} \sqrt{1-h_{ii}}$, the estimate of variance is

$$\hat{\sigma}_{(i)}^2 = \hat{\sigma}^2 \left(\frac{n-p'-1}{n-p'-r_i^2} \right)^{-1} \quad (\text{A.46})$$

and the studentized residual t_i is

$$t_i = r_i \left(\frac{n-p'-1}{n-p'-r_i^2} \right)^{1/2} \quad (\text{A.47})$$

The diagnostic statistics examined in this book were first thought to be practical because of simple formulas used to obtain various statistics when cases are deleted that avoided recomputing estimates. Advances in computing in the last 30 years have made the computational burden of recomputing without a case much less onerous, and so diagnostic methods equivalent to those discussed here can be applied to problems other than linear regression where the updating formulas are not available.

References

In a few instances, the URL given for an article refers to the website <http://dx.doi.org/>, used to resolve a digital object identifier (DOI) and send you to the correct website. This may lead you to a page requesting payment before viewing an article. Many research libraries subscribe to journals and may use a different method to resolve a DOI so you can get to articles for free. Ask your librarian, or see <http://doi.org>.

An on-line version of this bibliography with clickable links is available on the website for this book, <http://z.umn.edu/alr4ed>.

- AGRESTI, A. (2007). *An Introduction to Categorical Data Analysis*. 2nd ed. Wiley, Hoboken, NJ.
- AGRESTI, A. (2013). *Categorical Data Analysis*. 3rd ed. Wiley, Hoboken, NJ.
- ALLISON, P. D. (2001). *Missing Data*. Quantitative Applications in the Social Sciences. Sage, Thousand Oaks, CA.
- ALLISON, T. and CICCHETTI, D. V. (1976). Sleep in mammals: Ecological and constitutional correlates. *Science*, **194**, 732–734. URL: <http://www.jstor.org/stable/1743947>.
- ANScombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, **27**, 17–21. URL: <http://www.jstor.org/stable/2682899>.
- ATKINSON, A. C. (1985). *Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Clarendon Press, Oxford.
- BAES, C. and KELLOGG, H. (1953). Effects of dissolved sulphur on the surface tension of liquid copper. *Journal of Metals*, **5**, 643–648.
- BARNETT, V. and LEWIS, T. (1994). *Outliers in Statistical Data*. 3rd ed. Wiley, Hoboken, NJ.
- BATES, D. and WATTS, D. (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley, Hoboken, NJ.
- BECKMAN, R. J. and COOK, R. D. (1983). Outliers. *Technometrics*, **25**, 119–149. URL: <http://www.jstor.org/stable/1268541>.
- BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289–300. URL: <http://www.jstor.org/stable/2346101>.

- BERTALANFFY, L. (1938). A quantitative theory of organic growth (inquiries on growth laws II). *Human Biology*, **10**, 181–213. URL: <http://www.jstor.org/stable/41447359>.
- BERZUINI, C., DAWID, P., and BERNARDINELLI, L. (eds.) (2012). *Causality: Statistical Perspectives and Applications*. Wiley, Hoboken, NJ.
- BLESKE-RECHEK, A. and FRITSCH, A. (2011). Student consensus on ratemyprofessors.com. *Practical Assessment, Research & Evaluation*, **16**. (Online; last accessed—August 1, 2013), URL: <http://pareonline.net/getvn.asp?v=16&n=18>.
- BLOM, G. (1958). *Statistical Estimates and Transformed Beta Variables*. Wiley, New York.
- BOWMAN, A. W. and AZZALINI, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford University Press, Oxford.
- BOX, G., JENKINS, G., and REINSEL, G. (2008). *Time Series Analysis: Forecasting and Control*. 4th ed. Wiley, Hoboken, NJ.
- BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, **26**, 211–252. URL: <http://www.jstor.org/stable/2984418>.
- BRETZ, F., HOTHORN, T., WESTFALL, P., and WESTFALL, P. (2010). *Multiple Comparisons Using R*. Chapman & Hall/CRC, Boca Raton, FL.
- BREUSCH, T. S. and PAGAN, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, **47**, 1287–1294. URL: <http://www.jstor.org/stable/1911963>.
- BRILLINGER, D. (1983). A generalized linear model with “Gaussian” regressor variables. *A Festschrift for Erich L. Lehmann in Honor of His Sixty-Fifth Birthday*, 97–114.
- BROWN, P. (1993). *Measurement, Regression, and Calibration*. Oxford Scientific Publications, Oxford.
- BURNSIDE, O. C., WILSON, R. G., WEISBERG, S., and HUBBARD, K. G. (1996). Seed longevity of 41 weed species buried 17 years in Eastern and Western Nebraska. *Weed Science*, **44**, 74–86. URL: <http://www.jstor.org/stable/4045786>.
- BURT, C. (1966). The genetic determination of differences in intelligence: A study of monozygotic twins reared together and apart. *British Journal of Psychology*, **57**, 137–153. URL: <http://dx.doi.org/10.1111/j.2044-8295.1966.tb01014.x>.
- CARPENTER, J. and KENWARD, M. (2012). *Multiple Imputation and Its Application*. Wiley, Hoboken, NJ. (Online; last accessed August 1, 2013), URL: <http://missingdata.lshtm.ac.uk>.
- CASELLA, G. and BERGER, R. (2001). *Statistical Inference*. Duxbury Press, Pacific Grove, CA.
- CENTERS FOR DISEASE CONTROL (2013). Youth risk behavior surveillance system. (Online; last accessed August 1, 2013), URL: <http://www.cdc.gov/HealthyYouth/yrbs/index.htm>.
- CHEN, C.-F. (1983). Score tests for regression models. *Journal of the American Statistical Association*, **78**, 158–161. URL: <http://www.jstor.org/stable/2287123>.
- CHRISTENSEN, R. (2011). *Plane Answers to Complex Questions: The Theory of Linear Models*. 4th ed. Springer, New York.

- CLAPHAM, A. (1934). *English Romanesque Architecture after the Conquest*. Clarendon Press, Oxford.
- CLARK, R., HENDERSON, H., HOGGARD, G., ELLISON, R., and YOUNG, B. (1987). The ability of biochemical and haematological tests to predict recovery in periparturient recumbent cows. *New Zealand Veterinary Journal*, **35**, 126–133. URL: <http://dx.doi.org/10.1080/00480169.1987.35410>.
- CLAUSIUS, R. (1850). Über die bewegende Kraft der Wärme und die Gezette welche sich daraus für Wärmelehre selbst ableiten lassen. *Annalen der Physik*, **79**, 500–524.
- CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatter-plots. *Journal of the American Statistical Association*, **74**, 829–836. URL: <http://www.jstor.org/stable/2286407>.
- COCHRAN, W. (1977). *Sampling Techniques*. 3rd ed. Wiley, Hoboken, NJ.
- COLLETT, D. (2003). *Modelling Binary Data*. 2nd ed. Chapman & Hall, Boca Raton, FL.
- COLORADO CLIMATE CENTER (2012). Colorado climate center monthly data access. (Online; last accessed August 1, 2013), URL: <http://ccc.atmos.colostate.edu/cgi-bin/monthlydata.pl>.
- COOK, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, **19**, 15–18. URL: <http://www.jstor.org/stable/1268249>.
- COOK, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological)*, **48**, 133–169. URL: <http://www.jstor.org/stable/2345711>.
- COOK, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, Hoboken, NJ.
- COOK, R. D. and PRESCOTT, P. (1981). On the accuracy of Bonferroni significance levels for detecting outliers in linear models. *Technometrics*, **23**, 59–63. URL: <http://www.jstor.org/stable/1267976>.
- COOK, R. D. and WEISBERG, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall/CRC, Boca Raton, FL. (Online; last accessed August 1, 2013), URL: <http://conservancy.umn.edu/handle/37076>.
- COOK, R. D. and WEISBERG, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika*, **70**, 1–10. URL: <http://www.jstor.org/stable/2335938>.
- COOK, R. D. and WEISBERG, S. (1994). Transforming a response variable for linearity. *Biometrika*, **81**, 731–737. URL: <http://www.jstor.org/stable/2337076>.
- COOK, R. D. and WEISBERG, S. (1999a). *Applied Regression Including Computing and Graphics*. Wiley, New York.
- COOK, R. D. and WEISBERG, S. (1999b). Graphs in statistical analysis: Is the medium the message? *The American Statistician*, **53**, 29–37. URL: <http://www.jstor.org/stable/2685649>.
- COOK, R. D. and WITMER, J. A. (1985). A note on parameter-effects curvature. *Journal of the American Statistical Association*, **80**, 872–878. URL: <http://www.jstor.org/stable/2288546>.
- COX, D. (1958). *Planning of Experiments*. Wiley, Hoboken, NJ.
- CUNNINGHAM, R. and HEATHCOTE, C. (1989). Estimating a non-Gaussian regression model with multicollinearity. *Australian & New Zealand Journal of Statistics*, **31**, 12–17.

- 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. URL: <http://www.jstor.org/stable/2290069>.
- DANIEL, C. and WOOD, F. (1980). *Fitting Equations to Data: Computer Analysis of Multifactor Data*. Wiley, Hoboken, NJ.
- DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge.
- DAWSON, R. (1995). The “unusual episode” data revisited. *Journal of Statistics Education*, **3**. URL: <http://www.amstat.org/publications/JSE/v3n3/datasets.dawson.html>.
- DE BOOR, C. (1978). *A Practical Guide to Splines*. Springer, New York.
- DERRICK, A. (1992). Development of the measure-correlate-predict strategy for site assessment. In *Proceedings of the 14th BWEA Conference*. 259–265.
- DODSON, S. (1992). Predicting crustacean zooplankton species richness. *Limnology and Oceanography*, **37**, 848–856. URL: <http://www.jstor.org/stable/2837943>.
- EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**, 1–26. URL: <http://www.jstor.org/stable/2958830>.
- EFRON, B. and TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton, FL.
- EICKER, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *The Annals of Mathematical Statistics*, **34**, 447–456. URL: <http://www.jstor.org/stable/2238390>.
- EICKER, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. University of California Press, Berkeley, 59–82.
- EZEKIEL, M. and FOX, K. (1959). *Methods of Correlation and Regression Analysis: Linear and Curvilinear*. Wiley, Hoboken, NJ.
- FAIR ISAAC CORPORATION (2013). myfico. (Online; last accessed August 1, 2013), URL: <http://www.myfico.com/CreditEducation/WhatsInYourScore.aspx>.
- FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360. URL: <http://www.jstor.org/stable/3085904>.
- FEDERAL HIGHWAY ADMINISTRATION (2001). Highway statistics 2001. (Online; last accessed August 1, 2012), URL: <http://www.fhwa.dot.gov/ohim/hs01/index.htm>.
- FINKELSTEIN, M. O. (1980). The judicial reception of multiple regression studies in race and sex discrimination cases. *Columbia Law Review*, **80**, 737–754. URL: <http://www.jstor.org/stable/1122138>.
- FISHER, R. and MACKENZIE, W. (1923). Studies in crop variation. II. The manurial response of different potato varieties. *The Journal of Agricultural Science*, **13**, 311–320. URL: <http://digital.library.adelaide.edu.au/dspace/bitstream/2440/15179/1/32.pdf>.
- FITZMAURICE, G., LAIRD, N., and WARE, J. (2011). *Applied Longitudinal Analysis*. 2nd ed. Wiley, Hoboken, NJ.
- FORBES, J. D. (1857). XIV.—Further experiments and remarks on the measurement of heights by the boiling point of water. *Transactions of the Royal Society of Edinburgh*, **21**, 235–243. URL: http://journals.cambridge.org/article_S0080456800032075.

- Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software*, **8**, 1–27. URL: <http://www.jstatsoft.org/v08/i15>.
- Fox, J. and WEISBERG, S. (2011). *An R Companion to Applied Regression*. 2nd ed. Sage, Thousand Oaks, CA. URL: <http://z.umn.edu/carbook>.
- FRALEY, C., RAFTERY, A. E., GNEITING, T., SLOUGHTER, J., and BERROCAL, V. J. (2011). Probabilistic weather forecasting in R. *R Journal*, **3**, 55–63. (Online; last accessed August 1, 2013), URL: http://journal.r-project.org/archive/2011-1/RJournal_2011-1_Fraley~et~al.pdf.
- FREEDMAN, D. and LANE, D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, **1**, 292–298. URL: <http://www.jstor.org/stable/1391660>.
- FREEDMAN, D. A. (1983). A note on screening regression equations. *The American Statistician*, **37**, 152–155. URL: <http://www.jstor.org/stable/2685877>.
- FREEMAN, M. and TUKEY, J. (1950). Transformations related to the angular and the square root. *The Annals of Mathematical Statistics*, **21**, 607–611.
- FURNIVAL, G. M., WILSON, J., and ROBERT W. (1974). Regressions by leaps and bounds. *Technometrics*, **16**, 499–511. URL: <http://www.jstor.org/stable/1267601>.
- GALTON, F. (1877). Typical laws of heredity. *Proceedings of the Royal Institution*, **8**, 282–301. (Online; last accessed August 1, 2013), URL: <http://galton.org/essays/1870-1879/galton-1877-roy-soc-typical-laws-heredity.pdf>.
- GALTON, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, **15**, 246–263. URL: <http://www.jstor.org/stable/2841583>.
- GAUSS, C. (1821). Anzeige: Theoria combinationis observationum erroribus minimis obnoxiae: Pars prior (theory of the combination of observations which leads to the smallest errors). *Göttingische gelehrte Anzeigen*, **33**, 321–327. Reprinted by the Society for Industrial and Applied Mathematics, 1987, URL: <http://pubs.siam.org/doi/pdf/10.1137/1.9781611971248.fm>.
- GILSTEIN, C. Z. and LEAMER, E. E. (1983). The set of weighted regression estimates. *Journal of the American Statistical Association*, **78**, 942–948. URL: <http://www.jstor.org/stable/2288208>.
- GNANADESIKAN, R. (1997). *Methods for Statistical Data Analysis of Multivariate Observations*. 2nd ed. Wiley, Hoboken, NJ.
- GOLDSTEIN, H. (2010). *Multilevel Statistical Models*. 4th ed. Wiley, Hoboken, NJ.
- GOLUB, G. and VAN LOAN, C. (1996). *Matrix Computations*, vol. 3. Johns Hopkins University Press, Baltimore, MD.
- GOULD, S. (1966). Allometry and size in ontogeny and phylogeny. *Biological Reviews*, **41**, 587–638. URL: <http://dx.doi.org/10.1111/j.1469-185X.1966.tb01624.x>.
- GOULD, S. J. (1973). The shape of things to come. *Systematic Zoology*, **22**, 401–404. URL: <http://www.jstor.org/stable/2412947>.
- GREEN, P. and SILVERMAN, B. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, vol. 58. Chapman & Hall/CRC, Boca Raton, FL.
- GREENE, W. (2003). *Econometric Analysis*. 5th ed. Prentice Hall, Upper Saddle River, NJ.

- HADDON, M. and HADDON, M. (2010). *Modelling and Quantitative Methods in Fisheries*. Chapman & Hall/CRC, Boca Raton, FL.
- HAHN, A. (ed.) (1979). *Development and Evolution of Brain Size*. Academic Press, New York.
- HALD, A. (1960). *Statistical Theory with Engineering Applications*. Wiley, Hoboken, NJ.
- HALL, P. and LI, K.-C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, **21**, 867–889. URL: <http://www.jstor.org/stable/2242265>.
- HÄRDLE, W. (1990). *Applied Nonparametric Regression*, vol. 26. Cambridge University Press, Cambridge, MA.
- HART, C. W. M. (1943). The Hawthorne experiments. *The Canadian Journal of Economics and Political Science/Revue Canadienne d'Economique et de Science politique*, **9**, 150–163. URL: <http://www.jstor.org/stable/137416>.
- HASTIE, T., TIBSHIRANI, R., and FRIEDMAN, J. H. (2009). *The Elements of Statistical Learning*. 2nd ed. Springer, New York.
- HAWKINS, D. M. (1980). *Identification of Outliers*. Chapman & Hall/CRC, Boca Raton, FL.
- HAWKINS, D. M., BRADU, D., and KASS, G. V. (1984). Location of several outliers in multiple-regression data using elemental sets. *Technometrics*, **26**, 197–208. URL: <http://www.jstor.org/stable/1267545>.
- HENDERSON, H. V. and SEARLE, S. R. (1981). On deriving the inverse of a sum of matrices. *SIAM Review*, **23**, 53–60. URL: <http://www.jstor.org/stable/2029838>.
- HERNANDEZ, F. and JOHNSON, R. A. (1980). The large-sample behavior of transformations to normality. *Journal of the American Statistical Association*, **75**, 855–861. URL: <http://www.jstor.org/stable/2287172>.
- HILBE, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press, Cambridge.
- HINKLEY, D. (1985). Transformation diagnostics for linear models. *Biometrika*, **72**, 487–496. URL: <http://www.jstor.org/stable/2336721>.
- HOAGLIN, D. C. and WELSCH, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician*, **32**, 17–22. URL: <http://www.jstor.org/stable/2683469>.
- HOCKING, R. (1985). *The Analysis of Linear Models*. Brooks Cole, Monterey, CA.
- HOCKING, R. (2003). *Methods and Applications of Linear Models: Regression and the Analysis of Variance*. Wiley, Hoboken, NJ.
- HOETING, J. A., MADIGAN, D., RAFTERY, A. E., and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, **14**, 382–401. URL: <http://www.jstor.org/stable/2676803>.
- HOSMER, D. W., LEMESHOW, S., and MAY, S. (2008). *Applied Survival Analysis*. 2nd ed. Wiley, Hoboken, NJ.
- HOSMER, D. W., LEMESHOW, S., and STURDIVANT, R. (2013). *Applied Logistic Regression*. 3rd ed. Wiley, Hoboken, NJ.
- HUBER, P. (1967). The behavior of maximum likelihood estimates under non-standard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 221–33.

- HUBER, P. and RONCHETTI, E. M. (2009). *Robust Statistics*. 2nd ed. Wiley, Hoboken, NJ.
- HURVICH, C. M. and TSAI, C.-L. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, **44**, 214–217. URL: <http://www.jstor.org/stable/2685338>.
- IOANNIDIS, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, **2**, e124. URL: <http://dx.doi.org/10.1371%2Fjournal.pmed.0020124>.
- JEVONS, W. S. (1868). On the condition of the metallic currency of the United Kingdom, with reference to the question of international coinage. *Journal of the Statistical Society of London*, **31**, 426–464. URL: <http://www.jstor.org/stable/2338797>.
- JOHNS, M. W. (1991). A new method for measuring daytime sleepiness: The Epworth sleepiness scale. *Sleep*, **16**, 540–545. URL: <http://www.ncbi.nlm.nih.gov/pubmed/1798888>.
- JOHNSON, K. (1996). *Unfortunate Emigrants: Narratives of the Donner Party*. Utah State University Press, Logan, UT.
- JOHNSON, M. P. and RAVEN, P. H. (1973). Species number and endemism: The Galápagos archipelago revisited. *Science*, **179**, 893–895. URL: <http://www.jstor.org/stable/1735348>.
- JOINER, B. L. (1981). Lurking variables: Some examples. *The American Statistician*, **35**, 227–233. URL: <http://www.jstor.org/stable/2683295>.
- KENNEDY, W. and GENTLE, J. (1980). *Statistical Computing*, vol. 33. CRC, Boca Raton, FL.
- LEBEAU, M. (2004). *Evaluation of the Intraspecific Effects of a 15-Inch Minimum Size Limit on Walleye Populations in Northern Wisconsin*. PhD thesis, University of Minnesota.
- LEHRER, J. (2010). The truth wears off. *The New Yorker*, **13**. (Online; last accessed August 1, 2012), URL: http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer?currentPage=all.
- LENTH, R. V. (2006–2009). Java applets for power and sample size [computer software]. (Online; last accessed August 1, 2013), URL: <http://www.stat.uiowa.edu/~rlelenth/Power>.
- LENTH, R. V. (2013). *lsmeans: Least-squares means*. R package version 1.06-05, URL: <http://CRAN.R-project.org/package=lsmeans>.
- LI, K.-C. and DUAN, N. (1989). Regression analysis under link violation. *The Annals of Statistics*, **17**, 1009–1052. URL: <http://www.jstor.org/stable/2241708>.
- LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*. 2nd ed. Wiley, Hoboken, NJ.
- LOADER, C. (1999). *Local Regression and Likelihood*. Springer, New York.
- LOADER, C. (2004). Smoothing: Local regression techniques. In *Handbook of Computational Statistics: Concepts and Methods*. Springer, New York, 539–563.
- LOHR, S. (2009). *Sampling: Design and Analysis*. 2nd ed. Duxbury Press, Pacific Grove, CA.
- LONG, J. S. and ERVIN, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, **54**, 217–224. URL: <http://www.jstor.org/stable/2685594>.

- LUMLEY, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. Wiley, Hoboken, NJ. (Supplement online; last accessed August 1, 2013), URL: <http://faculty.washington.edu/tlumley/svybook>.
- MANTEL, N. (1970). Why stepdown procedures in variable selection. *Technometrics*, **12**, 621–625. URL: <http://www.jstor.org/stable/1267207>.
- MCCULLAGH, P. (2002). What is a statistical model? *The Annals of Statistics*, **30**, 1225–1267. URL: <http://www.jstor.org/stable/1558705>.
- MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*. 2nd ed. Chapman & Hall/CRC, Boca Raton, FL.
- MCCULLOCH, C., SEARLE, S., and NEUHAUS, J. (2008). *Generalized Linear Mixed Models*. 2nd ed. Wiley, Hoboken, NJ.
- MILLER, R. (1981). *Simultaneous Statistical Inference*. Springer, New York.
- MONTGOMERY, D. (2012). *Design and Analysis of Experiments*. 8th ed. Wiley, Hoboken, NJ.
- MOORE, J. A. (1975). *Total Biomedical Oxygen Demand of Animal Wastes*. PhD thesis, University of Minnesota.
- MOSTELLER, F. and TUKEY, J. W. (1977). *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley, Reading, MA.
- MOSTELLER, F. and WALLACE, D. (1964). *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, MA.
- NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY (2012). Engineering statistics. (Online; last accessed August 1, 2013), URL: <http://www.itl.nist.gov/div898/handbook/>.
- NCAR (2013). CISL research data archive. (Online; last accessed August 1, 2013), URL: <http://rda.ucar.edu/datasets/ds090.0/>.
- NELDER, J. A. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society. Series A (General)*, **140**, 48–77. URL: <http://www.jstor.org/stable/2344517>.
- NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, **135**, 370–384. URL: <http://www.jstor.org/stable/2344614>.
- NISHII, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, **12**, 758–765. URL: <http://www.jstor.org/stable/2241410>.
- NOLL, S., WAIBEL, P., COOK, R., and WITMER, J. (1984). Biopotency of methionine sources for young turkeys. *Poultry Science*, **63**, 2458–2470. URL: <http://dx.doi.org/10.3382/ps.0632458>.
- OEHLMER, G. (2000). *A First Course in Design and Analysis of Experiments*. Freeman, New York. This is out-of-print, but now available under a Creative Commons license. (Online; last accessed August 1, 2013), URL: <http://users.stat.umn.edu/~gary/Book.html>.
- ORESTES CERDEIRA, J., DUARTE SILVA, P., CADIMA, J., and MINHOTO, M. (2012). Subselect: Selecting variable subsets. R package version 0.12-2, URL: <http://CRAN.R-project.org/package=subselect>.
- PARKS, J. (1982). *A Theory of Feeding and Growth of Animals*, vol. 11. Springer, Berlin.
- PEARSON, K. (1930). *Life and Letters and Labours of Francis Galton*. Cambridge University Press, Cambridge.

- PEARSON, K. and LEE, A. (1903). On the laws of inheritance in man: I. Inheritance of physical characters. *Biometrika*, **2**, 357–462. URL: <http://www.jstor.org/stable/2331507>.
- PERKIÖMÄKI, M. (1997). World-wide track & field statistics. (Online; last accessed August 1, 2013), URL: <http://www.saunalahti.fi/~sut/eng/>.
- PHIPPS, P., MAXWELL, J. W., and ROSE, C. (2009). 2009 Annual Survey of the Mathematical Sciences. *Notices of the American Mathematical Society*, **57**, 250–259. (Online; last accessed August 1, 2013), URL: <http://www.ams.org/profession/data/annual-survey/docsgrt>.
- PINHEIRO, J. C. and BATES, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.
- R CORE TEAM (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: <http://www.R-project.org/>.
- RATKOWSKY, D. A. (1990). *Handbook of Nonlinear Regression Models*. Marcel Dekker, New York.
- RAUDENBUSH, S. and BRYK, A. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, Thousand Oaks, CA.
- RENCHER, A. C. and PUN, F. C. (1980). Inflation of R^2 in best subset regression. *Technometrics*, **22**, 49–53. URL: <http://www.jstor.org/stable/1268382>.
- RICH, R. L., FRELICH, L. E., and REICH, P. B. (2007). Wind-throw mortality in the Southern Boreal forest: Effects of species, diameter and stand age. *Journal of Ecology*, **95**, 1261–1273. URL: <http://www.jstor.org/stable/4496078>.
- ROBINSON, W. (2009). Ecological correlations and the behavior of individuals. *International Journal of Epidemiology*, **38**, 337–341. URL <http://ije.oxfordjournals.org/content/38/2/337.short>.
- ROGUE WAVE SOFTWARE (2013). IMSL numerical libraries. (Online; last accessed August 1, 2013), URL: <http://www.roguewave.com/products/imsl-numerical-libraries.aspx>.
- ROYSTON, J. P. (1982a). Algorithm AS 177: Expected normal order statistics (exact and approximate). *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **31**, 161–165. URL: <http://www.jstor.org/stable/2347982>.
- ROYSTON, J. P. (1982b). Algorithm AS 181: The W test for normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **31**, 176–180. URL: <http://www.jstor.org/stable/2347986>.
- ROYSTON, J. P. (1982c). An extension of Shapiro and Wilk's W test for normality to large samples. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **31**, 115–124. URL: <http://www.jstor.org/stable/2347973>.
- RUBIN, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592. URL: <http://www.jstor.org/stable/2335739>.
- SAKAMOTO, Y., ISHIGURO, M., and KITAGAWA, G. (1986). *Akaike Information Criterion Statistics*. D. Reidel Publishing, Boston.
- SAS INSTITUTE, INC. (2013). Lsmeans statement. (Online; last accessed August 1, 2013), URL: <http://z.umn.edu/saslsmeans>.
- SAW, J. (1966). A conservative test for the concurrence of several regression lines and related problems. *Biometrika*, **53**, 272–275.

- SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, Boca Raton, FL.
- SCHAFER, J. L. and GRAHAM, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, **7**, 147. URL: <http://dx.doi.org/10.1037/1082-989X.7.2.147>.
- SCHEFFÉ, H. (1959). *The Analysis of Variance*. Wiley, New York.
- SCHOTT, J. (2005). *Matrix Analysis for Statistics*. Wiley, Hoboken, NJ.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464. URL: <http://www.jstor.org/stable/2958889>.
- SCLOVE, S. L. (1968). Improved estimators for coefficients in linear regression. *Journal of the American Statistical Association*, **63**, 596–606. URL: <http://www.jstor.org/stable/2284030>.
- SEBER, G. (2008). *A Matrix Handbook for Statisticians*, vol. 746. Wiley, Hoboken, NJ.
- SEBER, G. A. F. and WILD, C. J. (1989). *Nonlinear Regression*. Wiley, Hoboken, NJ.
- SHAPIRO, S. S. and WILK, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591–611. URL: <http://www.jstor.org/stable/2333709>.
- SIMONOFF, J. (1996). *Smoothing Methods in Statistics*. Springer, New York.
- SMITH, R. (no date). A statistical assessment of Buchanan's vote in Palm Beach County. (Online; last accessed August 1, 2013), URL: <http://www.stat.unc.edu/faculty/rs/palmbeach.html>.
- ST. ANDREWS UNIVERSITY (2003). William Sealy Gosset. (Online; last accessed August 1, 2013), URL: <http://www-gap.dcs.st-and.ac.uk/history/Mathematicians/Gosset.html>.
- STAUDTE, R. and SHEATHER, S. (1990). *Robust Estimation and Testing*. Wiley Online Library.
- STEFANSKI, L. A. (2007). Residual (sur)realism. *The American Statistician*, **61**, 163–177. (Online; last accessed August 1, 2013), URL: <https://www.amstat.org/about/pdfs/NCSUStatsProfSurpriseHomework.pdf>.
- STEVENS, S. S. (1966). A metric for the social consensus. *Science*, **151**, 530–541. URL: <http://www.jstor.org/stable/1717034>.
- TAFF, S. J. and WEISBERG, S. (2007). Compensated short-term conservation restrictions may reduce sale prices. *Appraisal Journal*, **75**, 45–55.
- TAFF, S. J., TIFFANY, D. G., and WEISBERG, S. (1996). Measured effects of feedlots on residential property values in Minnesota: A report to the legislature. Staff Papers 14121, University of Minnesota, Department of Applied Economics. URL: <http://ideas.repec.org/p/ags/umaesp/14121.html>.
- THISTED, R. (1988). *Elements of Statistical Computing: Numerical Computation*, vol. 1. Chapman & Hall/CRC, Boca Raton, FL.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 267–288. URL: <http://www.jstor.org/stable/2346178>.
- TSAY, R. (2005). *Analysis of Financial Time Series*. Wiley, Hoboken, NJ.
- TUDDENHAM, R. and SNYDER, M. (1954). Physical growth of California boys and girls from birth to eighteen years. *Publications in Child Development. University of California, Berkeley*, **1**, 183.

- TUKEY, J. W. (1949). One degree of freedom for non-additivity. *Biometrics*, **5**, 232–242. URL: <http://www.jstor.org/stable/3001938>.
- UBS (2009). Prices and earnings: A comparison of purchasing power around the globe. (Online; last accessed August 1, 2013), URL: http://www.ubs.com/global/en/wealth_management/wealth_management_research/prices_earnings.html.
- UNITED NATIONS (2011). Social indicators. (Online; last accessed August 1, 2013), URL: <http://unstats.un.org/unsd/demographic/products/socind/>.
- UNIVERSITY OF WASHINGTON APPLIED PHYSICS LABORATORY (2013). University of Washington probability forecast. (Online; last accessed August 1, 2013), URL: <http://probcast.washington.edu/>.
- U.S. CENSUS (undated). Imputation of unreported data items. (Online; last accessed August 1, 2013), URL: <http://www.census.gov/cps/methodology/unreported.html>.
- VARSHNEY, L. R. and SUN, J. Z. (2013). Why do we perceive logarithmically? *Significance*, **10**, 28–31. (Online; last accessed August 1, 2013), URL: <http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2013.00636.x/pdf>.
- VELILLA, S. (1993). A note on the multivariate Box-Cox transformation to normality. *Statistics and Probability Letters*, **17**, 259–263. URL: [http://dx.doi.org/10.1016/0167-7152\(93\)90200-3](http://dx.doi.org/10.1016/0167-7152(93)90200-3).
- WEISBERG, H., BEIER, E., BRODY, H., PATTON, R., RAYCHAUDHURI, K., TAKEDA, H., THERN, R., and VAN BERG, R. (1978). s-dependence of proton fragmentation by hadrons. ii. Incident laboratory momenta 30–250 gev/c. *Physical Review D*, **17**, 2875. URL: <http://dx.doi.org/10.1103/PhysRevD.17.2864>.
- WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**, 817–838. URL: <http://www.jstor.org/stable/1912934>.
- WILKINSON, G. N. and ROGERS, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **22**, 392–399. URL: <http://www.jstor.org/stable/2346786>.
- WILM, H. (1950). Statistical control in hydrologic forecasting. Tech. Rep. 71, Pacific Northwest Forest Range Experiment Station, Oregon.
- WOOD, S. (2006). *Generalized Additive Models: An Introduction with R*, vol. 66. Chapman & Hall/CRC, Boca Raton, FL.
- WOODLEY, W. L., SIMPSON, J., BIONDINI, R., and BERKELEY, J. (1977). Rainfall results, 1970–1975: Florida area cumulus experiment. *Science*, **195**, 735–742. URL: <http://www.jstor.org/stable/1743962>.
- YANG, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association*, **96**, 574–588. URL: <http://amstat.tandfonline.com/doi/abs/10.1198/016214501753168262>.
- YANG, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, **92**, 937–950. URL: <http://www.jstor.org/stable/20441246>.
- YEO, I.-K. and JOHNSON, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, **87**, 954–959. URL: <http://www.jstor.org/stable/2673623>.
- ZEILEIS, A. (2004). Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, **11**, 1–17. URL: <http://www.jstatsoft.org/v11/i10>.

- ZIPF, G. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Reading, MA.
- ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **67**, 301–320. URL: <http://www.jstor.org/stable/3647580>.
- ZUUR, A., IENO, E., WALKER, N., SAVELIEV, A., and SMITH, G. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer, New York.

Author Index

- Agresti, A., 270
Ahlstrom, M., 50
Allison, P. D., 119
Allison, T., 186, 267
Anscombe, F. J., 12, 27
Azzalini, A., 14

Baes, C., 199
Barnett, V., 218
Bates, D. M., 167–168, 265
Beckman, R. J., 218
Berger, R., 29, 90, 138, 143
Bertalanffy, L., 267–268
Berzuini, C., 87
Bowman, A. W., 14
Box, G. E. P., 108, 190–191
Bretz, F., 103, 217
Breusch, T. S., 164
Brillinger, D., 193
Bryk, A., 169

Carpenter, J., 121
Casella, J., 29, 90, 138, 143
Cerdeira, O., 239
Chen, C.-F., 164
Christensen, R., 79, 135, 144
Cicchetti, D. V., 186
Clapham, A., 129
Clausius, R., 42
Cleveland, W. S., 14
Cochran, W., 161
Collett, D., 270
Cook, R. D., 164, 194–196, 210, 218, 220–225,
 257
Cox, D., 87, 108, 190–191
Cunningham, R., 94

Dalal, S. R., 288
Daniel, C., 225
Davison, A. C., 175–177, 179
Dawson, R., 288
Derrick, A., 154
Duan, N., 193–194

Efron, B., 175–176
Eicker, F., 163
Ervin, L. H., 163

Fisher, R. A., 135
Fitzmaurice, G., 169
Forbes, J. D., 6, 24–25, 32, 37–38
Fox, J., 74, 79, 285
Freedman, D., 149
Frie, R., 267
Furnival, G. M., 239

Galton, F., 181–182
Gentle, J., 94
Gilstein, C. Z., 162
Gnanadesikan, R., 225
Goldstein, H., 169
Gosset, W. S., 212
Gould, S. J., 129, 188
Graham, J. W., 119
Green, P., 14

Haddon, M., 267
Hahn, A., 188

- Hald, A., 200
 Hall, P., 194
 Hardle, W., 14
 Hart, C. W. M., 149
 Hastie, T., 235, 244
 Hawkins, D. M., 218
 Heathcote, C., 94
 Hinkley, D. V., 175–177, 179
 Hoaglin, D. C., 207
 Hocking, R., 141
 Hoeting, J. A., 247
 Hoffstedt, C., 191–196
 Hosmer, D. W., 121, 270
 Huber, P., 163
 Ioannidis, J. P. A., 147–148, 235, 244
 Jevons, W. S., 182
 Johnson, M. P., 250
 Johnson, R. A., 198–199
 Joiner, B. L., 88
 Kellogg, H., 199
 Kennedy, W., 94
 Kenward, M., 121
 Lane, D., 149
 Leamer, E. E., 162
 Lee, A., 2–3
 Lenth, R. V., 103
 Lewis, T., 218
 Li, K.-C., 193–194
 Little, R. J. A., 119
 Lohr, S., 161
 Long, J. S., 163
 Lumley, T., 161
 Mackenzie, W., 135
 McCullagh, P., 141, 270, 285
 McCulloch, S., 169
 Miller, Rolf, 50
 Mosteller, F., 48, 99
 Nelder, J. A., 139, 270, 285
 Neyman, J., 254
 Ng, C., 263
 Nishii, R., 239
 Noll, S., 9
 Oehlert, G., 87, 94, 103, 111
 Pagan, A. R., 164
 Pearson, K., 2–3
 Pinheiro, J. C., 167–168
 Ratkowsky, D. A., 265
 Raudenbush, S., 169
 Raven, P. H., 250
 Rice, John, 164
 Robinson, A., 189
 Robinson, W., 160
 Rogers, C. E., 64, 101, 106–109, 139, 151, 259
 Royston, J. P., 226
 Rubin, D. B., 119, 121
 Sakamoto, Y., 239
 Schafer, J. L., 119
 Scheffé, 171
 Schwarz, G., 239
 Seber, G. A. F., 265
 Shapiro, S. S., 225
 Silverman, B., 14
 Simonoff, J., 14
 Syracuse, M., 232
 Snyder, M., 70
 Stevens, S. S., 169
 Stigler, Steven M., 182
 Sun, J. Z., 81, 169
 Taff, S. J., 87, 119
 Tibshirani, R., 175–176, 244
 Tuddenham, R., 70
 Tukey, J. W., 99, 212
 Varshney, L. R., 81, 169
 Vellila, S., 195
 Wald, A., 129
 Wallace, D., 48
 Watts, D., 265
 Wedderburn, R. W. M., 285
 Weisberg, S., 79, 119, 157, 164, 195–196, 210,
 218, 225, 285
 Welsch, R. E., 207
 White, H., 163
 Wild, C. J., 265
 Wilk, M. B., 225
 Wilkinson, G. N., 64, 101, 106–109, 139, 151, 259
 Wilm, 48
 Witmer, J. A., 257
 Wood, F., 225
 Wood, S., 115, 199
 Yang, Y., 239, 247
 Yeo, L.-K., 198–199
 Zeileis, A., 163
 Zipf, G., 48–49
 Zuur, A., 169

Subject Index

- Active predictors, 235, 238–245
Added-variable plots, 51–55, 67–68, 70, 71, 78
properties, 69
as a regression diagnostic, 224–225
relation to t -test, 68
Additive models, 199
Akaike Information Criterion (AIC), 238–245, 248
Allometric model, 187–188
Analysis of covariance, 107–108
Analysis of deviance, 284
Poisson regression, 278–279, 281–283
Analysis of variance (ANOVA), 133, 138–150
types, 140–141
unbalanced, 140
Autoregressive errors, 163, 168
- Backsolving, 308
Backward elimination, 240–243
Basis functions, 113–116
Bayes Information Criterion (BIC), 239, 244–245
Bernoulli distribution, counted data, 270–271
Bertalanffy function, 267–268
Best linear unbiased estimates, 29
Binomial distribution, 271
Binomial regression. *See* Logistic regression
Block diagonal, generalized least squares model, 168–169
Bonferroni inequality, regression diagnostics, outliers, 217–218
- Bootstrap analysis, 149, 174–179
bias, 177
bias corrected and accelerated (BCA), 176–178
case resampling, 176–178
for a median, 175
nonlinear regression, 225, 226
residual resampling, 179
tests, 179
- Box–Cox transformation method, 190–191, 312–314
automatic choice of predictor transformations, 195–196
linearly related regressors, 194–195
nonpositive variables, 198–199
response transformations, 196–198
- Boxplots
one-way factor model, 99–102
transformations, 186
- Causation, 87
Censoring, 121–122
Central composite design, 128
Central limit theorem, nonlinear regression, 256–257
Chi-squared random variable, 27
noncentral, 143
Clausius–Clapeyron formula, 42
Cluster sampling, 161
Coefficient of determination (R^2)
adjusted, 36
multiple linear regression, 66–68, 92–93
regression through origin, 93

- Coefficient of determination (R^2) (continued)
 sampling conditions, 91–93
 simple linear regression, 35–36, 91–92
 variable selection, 235–237
- Coefficient estimates
 logistic regression, 311–312
 multiple regression, 55–65
 simple regression, 22–28
 spline basis, 116
- Collinearity, 79–81
- Complex regressors
 factors, 56, 98–109
 interactions, 56, 104–108
 polynomials, 56, 109–113
 principal components, 57, 116–119
 splines, 57, 113–116
- Complexity, predictor variable discovery, 238–239
- Confidence intervals, 30
 bootstrap analysis, 175–179
 Box–Cox transformations, 191
 Poisson and binomial regression, 284
 simple linear regression, 30–34
- Constant variance, test for, 164–167
- Construction set, variable selection, 247
- Cook’s distance, 220–225
- Correlation, 23, 284
 matrix, 58, 119,
 partial, 54, 68
 relation to *t*-test, 46
 sample, 174
- Counted response models, 270
 distributions, 270–272
 regression models, 272–279
 simple and multiple regression, 283–284
- Covariance
 matrix, 295, 302
 multiple linear regression, 66
 sample, 59
 simple linear regression, 29
- Credit scoring, 246
- Cross-sectional data, 8
- Cross-validation, variable selection, 247
- Curvature testing, 212–213
- Data mining, 235
- Data sets
AMSSurvey, 280, 289
anscombe, 12
baesk1, 199
BGSall, 70, 129
BGSboys, 70, 250
BGSgirls, 70, 93
BigMac2003, 41, 202
- BlowBS*, 274
Blowdown, 275, 276, 278, 285, 289
brains, 186
cakes, 111, 126, 152, 183
cathedral, 45, 129
caution, 210
Challeng, 288
cloud, 231
domedata, 131
domedata1, 132
Donner, 287
Downer, 286
drugcost, 232
dwaste, 250
florida, 230
Forbes, 5, 24, 42, 129
ftcollinssnow, 8, 41
ftcollinstemp, 41
fuel2001, 15, 57, 151, 183, 203, 229
galapagos, 250
galtonpeas, 181
Heights, 2, 46
Highway, 191, 249, 269
Hooker, 42, 129
Htwt, 38
jevons, 182
lakes, 239
landrent, 231
lathe, 230
lathe1, 128
mantel, 249
mile, 183
MinnLand, 119, 123, 127, 128, 153, 154
MinnWater, 79, 97, 249
Mitchell, 18
MWwords, 48
oldfaith, 18, 49
physics, 157
physics1, 180
pipeline, 237
prodscor, 155
rat, 221
Rateprof, 19, 117, 152, 159, 286
Rpdata, 226
salary, 130, 136
salarygov, 126, 153, 180, 201
segreg, 263
sleep1, 265
snake, 48
sniffer, 164
Stevens, 169
stopping, 181, 200
swan96, 268

- Transact, 94, 184
turk0, 257
turkey, 9, 261
twins, 153
UBSprices, 39, 40, 41
ufcwc, 189
UN11, 17, 47, 51, 69, 95, 123, 150, 151, 183,
 249
walleye, 268
water, 19, 71, 200, 228
wblake, 7, 18, 47
Whitestar, 288
wm1, 49, 184
wm2, 154
Wool, 108, 131, 153, 203
Degrees of freedom, 26
Delta method, 172–174
Dependent variable. *See* Response variable
Deviance, logistic, and Poisson regression,
 277–279
Discovery, variable selection, 237–245
 information criteria, 238–239
 regularized methods, 244–245
 stepwise regression, 239–244
 subset selection, 245
Discovery probability, hypothesis testing,
 147–148
Dummy variables, 56–57, 100–102
 nonlinear regression, 260–262
 outlier testing, 215–216
Ecological regression, 160–162
Effects plots, 74–75, 105–108, 111–113,
 141–142, 278–279
Eigenvectors and eigenvalues, 118–119, 309
Elastic net, 244–245
Epworth sleepiness scale, 245–246
Errors e
 assumptions, 21
 multiple linear regression, 61
 regression diagnostic residuals, 205–206
Estimated variances, simple linear
 regression, 29–30
Examples
 Alaska pipeline, 227
 Anscombe, 12–13
 Berkeley Guidance study, 75–78, 119, 149,
 250
 blowdown, 274–279
 brain weight, 187–188
 cakes, 111–113, 138
 California water, 19, 200, 228
 cathedrals, 129–130
 caution, 210
Challenger, 288
cloud seeding, 281
credit scoring, 246
Donner party, 287
downer, 286–287
drug costs, 232
electrical energy consumption, 263–265
Epworth sleepiness scale, 245–246
feedlots, 87–89
Florida election in 2000, 230
Forbes's data, 5–7, 11–12, 24–27, 30–38,
 135–138
Ft. Collins snowfall, 8, 11, 31, 41, 136
Ft. Collins temperature, 41
fuel consumption, 15–17, 57–59, 63–66,
 73–79, 211–212
Galápagos Island data, 250–251
Galton's peas, 181
Government salary, 126–127, 153, 180, 201
height inheritance, 2–5, 10–12, 14–15, 36–37,
 91–93
Hooker, 44, 129
Jevon's gold coins, 182
lake diversity, 229
Lake Mary, 267
land rent, 231
land valuation, 155
lathe, 128, 230
mammal species, 186
Mantel, 249
mathematical sciences PhDs, 280–283
metrodome, 131
mile run, 183
Minnesota agricultural land sales, 119–122,
 144–145
Minnesota highway accidents, 191–196,
 240–245
Minnesota water use, 79–81
Old Faithful Geyser, 18–19, 49, 113–116
Oxygen uptake, 250
professor ratings, 117–119, 159–162,
 247–248
rat data, 221–225
segmented regression, 263
sex discrimination, 130
sleep in mammals, 267
smallmouth bass, 7–8, 10–12, 14–15
Snake River, 48
sniffer data, 164–167
Stefanski, 226–227
stopping distances, 181
strong interaction, 157–159
surface tension, 199
Swan Lake black crappies, 268

- Examples (*continued*)
 Titanic, 288
 transactions, 94, 175–178, 184, 226, 227
 turkey growth, 9–10, 252–259
 twin study, 153
 UBS prices, 39
 United Nations, 51–55, 98–108, 137–138,
 140, 149, 207–208, 212–214,
 219–225
 Upper Flat creek, 189–191
 walleye growth, 268
 weather prediction, 8–9, 246–247
 weight gain, 260–262
 Whitestar, 288
 windmills, 49, 154–155, 184
 wool, 108–109, 136–138, 141–142
 Zipf's law, 48
- Expected information matrix, 314
 Experimentation vs. observation, 86–89
 Exponential family distribution, 285
- Factors, 56, 98–109
 nonlinear regression, 260–262
- False discovery, 147–148, 150
 Family-wise error rate, 150
 FICO score, 246
 File drawer effects, 150
 Finite population approach to sample
 surveys, 162
 Fisher scoring, 311
 Fitted mean function
 multiple regression model, 55
 nonlinear regression, 259–262
- Fitted values
 inverse fitted value plot, 196–198
 multiple linear regression, 68–69
 simple linear regression
 confidence intervals, 33–34
 estimation of, 32–34
 ordinary least squares, 22–24
- Fixed-significance level, *p*-value
 interpretation, 147
- Focal predictors, variable selection, 235–237
 Forward selection, predictor variable
 discovery, 240–242
- F*-tests, 133–138
 analysis of variance, 139–142
 interpretation, 146–150
 overall test, 135–138
 power and non-null distributions, 144–145
 Wald tests, 145–146
- Gauss–Markov theorem, 28–29
 Gauss–Newton algorithm, 255–256
- General correlation structure, 168–169
 General likelihood ratio tests, 138
 General linear hypothesis, Wald tests, 146
 Generalized least squares (GLS),
 autoregressive, 168
 block diagonal, 168–168
 compound symmetry, 168
- Generalized linear models, 285
 Geometric mean, 190–191
 Goodness of fit tests, Poisson regression,
 282–283
- Hat matrix, 205–208
 Hawthorne effect, 150
 HC3 estimates, misspecified variances,
 163–167
- Hessian matrix, nonlinear regression,
 254–256
- Hierarchical regression, mixed models, 171
 Hot deck, missing data, 122
 Hyperplane, multiple regression model, 55
 Hypothesis testing
 analysis of variance and, 133–150
 counted response, 284
 false results, 147–148
 file drawer effects, 150
 general linear hypothesis, 146
 goodness of fit, 282
 Hawthorne effect, 150
 interpreting *p*-values, 146–150
 likelihood ratio tests, 138, 146, 195
 logistic regression, 277–279
 marginality principle, 139
 multiple testing, 150
 nonadditivity, 212
 one coefficient, 67–68
 Poisson regression, 279–281
 population vs. sampling, 149
 power, 143–145
 reported significance levels, 149
 t-tests, 30–34, 67–68
 types, analysis of variance, 135–136
 unbalanced, 135
- Imputation, missing data, 122
 Inclusion probability, sample surveys,
 161–162
- Independent variable. *See* Predictor
 variables
- Influence, 204, 218–225
 Cook's distance, 220–224
- Information criteria, 238–239
 Interactions, 56, 104–106, 139–142,
 211–213

- Intercept, 21–22, 56, 100–102
confidence interval, 30–34
- Interquartile range (IQR), 99–102
- Invariance, 43
- Inverse fitted value plot, 196–198
- Inverse regression, 183
- Inverse response plot, 198, 202, 203
- Jittering scatterplots, 3–5
- Kernel mean function, 253, 272–277
- Lack-of-fit testing, 211–212
- Lasso, 244–245
- Leaps and bounds algorithm, 239–245
- Least squares estimates. *See* Ordinary least squares
- Level means comparison, factor models, 102–103
- Leverage values, 204
residuals, 207–209
scatterplots, 4–5
- Li–Duan theorem, 194–195
- Likelihood ratio tests, 134
transformations, automatic predictor selection, 195–196
Wald tests comparison, 146
- Linear dependence, 78–79
- Linear independence, 78–79
- Linear predictor
binomial regression, 272–277
Poisson regression, 280–283
- Linear regression
basic properties, 1–2
coefficients, 133
F-tests, 134–138
mean functions, 10–12
multiple linear regression, 51–69
scatterplots, 2–10
simple linear regression, 21–38
summary graph, 12–13
variable selection, 235–237
- Linearly related regressors, 194–195
- Link function
binomial regression, 273–277
Poisson regression, 279–283
- loess smoother, 14–15
- Log rule, power transformations, 188
- Logarithms
base, 24
power transformations, 187–188
regressors in, 81–82
- response in, 82–83
variance stabilization, 172
- Logistic regression, 272–277
deviance, 277–279
goodness of fit tests, 282–283
- Logit function, 273–277
- Log-likelihood profile, Box–Cox method, 196–198
- Log-odds, 273–277
- Longitudinal studies, 8
- Lsmeans, 103, 108, 153
- Lurking variables, 88–89
- Machine learning, 235, 247–248
- Main effects interpretation, 73–93
analysis of variance, 139–142
experimentation vs. observation, 86–89
factor models
continuous predictors, 104–106
one-factor model, 106–108
multiple factors, 109
normal population sampling, 89–91
parameter estimates, 73–83
regressor omission, 84–86
- Marginal plot, 52–55
- Marginality principle, analysis of variance, 139–142
- Matrices, 290–309
inverse, 301
multiple linear regression, 60–61
partitioned matrix, 71–72
QR factorization, 307–308
rank, 76–81, 301
scatterplot matrices, 15–17
simple regression, 63–66
spectral decomposition, 309
- Maximum likelihood estimates, 309–313
Poisson regression, 280–283
regression parameters, 90–91
- Mean functions
additive model transformation, 199
Box–Cox transformation, 190–191
F-tests, 135–138
main effects 109
multiple linear regression, 58–59
nonlinear regression, 252–256
one-factor models, 100–102
outlier models, 214–218
parameter estimation, 75–78
parameter regressors, omission, 84–86
polynomial regression, 109–113
quadratic regression, 109–113
rank deficient and overparameterized mean functions, 78–79

- Mean functions (*continued*)
 regression, 10–12
 scaled power transformations, 189–190
 simple linear regression, 21–22
 least squares estimates, 29
 regressor addition, 51–55
 smoothers, 14–15
- Mean shift outlier model, regression
 diagnostics, 214–218
- Mean square, 26, 134–138
- Means comparison
 analysis of variance, 142
 level means, 102–103
- Measure, correlate, predict method, 154–155
- Missing data, 119–122
 missing at random (MAR), 121–122
 multiple imputation, 122
- Misspecified variance, 162–167
 accommodation, 163–164
 constant variance test, 164–167
- Mixed models, 169–171
- Model averaging, 247
- Multilevel and hierarchical models, 171
- Multiple comparisons, 102, 108
- Multiple correlation coefficient. *See*
 Coefficient of determination
- Multiple linear regression, 51–69
 coefficient of determination (R^2), 66–67,
 92–93
 collinearity, 79–81
 delta method, 173–174
 factors, 98–108
 model, 55
 ordinary least squares, 58–68
 overall F -test, 136
 predictions, fitted values, and linear
 combinations, 68–69
 regressors, 51–58
 residual plots, 210
 transformations, 193–196
- Multiple testing, 150
- Multiplicative error, 187–188
- Multistage sample surveys, 161–162
- Multivariate normality, 89–91
- Natural logarithms. *See Logarithms*
- Neural networks, 247
- Newton–Raphson algorithm, 311
- Noncentrality parameter, power
 and non-null distributions, 143–145
- Nonconstant variance
 regression diagnostics, 213–214
 tests for, 164–167
- Nonlinear regression, 252–269
 bootstrap inference, 262–265
 large sample inference, 256–257
 literature sources, 265
 mean function estimation, 253–256
 starting values, 257–262
- Non-null distributions, analysis of variance,
 143–145
- Nonparametric estimation, mean functions,
 10–12
- Nonpositive variables, transformation,
 198–199
- Normal distribution
 multivariate, 89–91
 sampling from, 89–91
- Normal equations, 293
- Normal probability plot, 225–226
- Normality
 Box–Cox transformation to, 191
 power transformations to, 195–196
- Normality assumption, regression
 diagnostics, 225–226
- Notation
 AIC, 238
 ANOVA, 139
 BIC, 239
 case, 2
 correlation ρ , 292
 covariance, Cov, 291
 df, 26
 expectation E, 290
 GLS, 168
 hats, 22
 h_{ii} , 207
 NID, 29
 ols, 22
 p' , 64
 predictor, 16
 $R^2_{Y,X}$, 236
 regressor, 16
 RSS, 24
 r_{xy} , 23
 SD, 23
 se, 28
 SSreg, 35
 SXX, 23
 s_{xy} , 23
 SXY, 23
 SYy, 23
 typewriter font, 2
 variance VAR, 291
 WLS, 156
 \bar{x} , 23

- Null plot
characteristics, 14
simple linear regression, 36–38
- Observational data, 75
- Odds of success, binomial regression, 273–277
- Offset, 249
- One-dimensional estimation,
linearly related regressors, 194–195
- One-factor model, one-way ANOVA, 99–102
- Ordinary least squares (OLS) estimation, 22, 24–26, 58–68
computing formulas, 61
matrix version, 304
misspecified variances, 163–167
nonlinear regression, 258–259
properties, 27–29, 305–307
- Orthogonal factors, 141–142
- Orthogonal polynomials, 112–113
- Orthogonal projection, 206–208
- Outliers, 214–218
scatterplots, 4–5, 13
- Overall *F*-test
multiple regression, 136
simple regression, 135–136
- Overparameterized mean function
one-factor models, 100–102
parameter estimates, 78–79
- Pairwise comparisons, 102–103
- Parameters, 73–93, 95–114
aliased, 78
collinearity, 79–81
F-tests, 138
intercept, 10, 21
multiple regression model, 55
not the same as estimates, 24
partial slope, 73
rank deficient or overparameterized mean
functions, 78–79
signs of estimates, 75
simple linear regression, 21–22
slope, 10, 21
variable selection and assessment of,
235–237
- Partial R^2 , 236
- Partial slope, 73
- Pearson residuals, 208
Poisson and binomial regression, 284–285
- Pearson's χ^2 , 283
- Per-test error rate, 150
- Poisson distribution, 271–272
generalized linear models, 283–285
variance stabilizing transformations, 171–172
- Poisson regression, 270–289
deviance, 277–279
goodness of fit tests, 282–283
- Polynomial regressors, 109–113
multiple predictors, 111–112
multiple regression model, 56
numerical issues, 112–113
- Power calculators, 144
- Power family
modified power family, 190–191
scaled power transformations, 188–190
transformations, 186–188
- Power of the test, analysis of variance, 143–145
- Predicted residual (PRESS residual), 230
- Prediction, 32–34
weighted least squares, 159
- Predictor variables. *See also* Regressors
active vs. inactive, 235
complex regressors, 98–122
principal components, 117–119
discovery, 238–245
experimentation vs. observation, 86–89
multiple linear regression, 55–58, 68–69
one-factor models, 100–102
polynomial regression, 109–113
scatterplots, 2–5
matrix, 16–17
selection methods, 234–251
single variable transformation, 188–190
transformations, 193–196
automatic selection, 195–196
- Principal component analysis
complex regressors, 116–119
multiple regression model, predictors and
regressors, 57
- Probability plot, 225–226
- p*-value
hypothesis testing, 133
interpretation, 146–147
means comparison, 103
outlier tests, 217–218
power and non-null distributions, 144–145
Wald tests, 145–146

- QR factorization, 228, 307–308
 Quadratic regression, 109–113
 curvature testing with, 212–213
 delta method for a maximum or minimum, 174
- R packages
alr4, ii, 290
car, 140
effects, 108, 153
lsmeans, 153
nlme, 168
- R*². *See* Coefficient of determination
- Random coefficients model, 170–171
- Random forests, 247
- Random vectors, 303
- Range rule, power transformations, 188
- Rank deficient mean function, 78–79
- Regression coefficients
 complex models, 98–113
 interpretation, 73–91
- Regression diagnostics, 204–233
 hat matrix, 205
 weighted hat matrix, 208
 influential cases, 218–225
 added-variable plots, 224–225
 Cook's distance, 220–221
 nonconstant variance, 213–214
 normality assumption, 225–226
 outliers, 214–218
 level significance, 217–218
 methodology, 218
 test, 215–216
 weighted least squares, 216
 Poisson and binomial regression, 284–285
 residuals, 204–212
 curvature testing, 212–213
 error vectors, 205–206
 hat matrix, 206–208
 plots of, 209–210
 weighted hat matrix, 208
- Regression through the origin, 93
- Regressors, 16, 51, 55–58
 class variable, 101
 collinear, 79
 dropping, 84
 dummy variables, 56, 100
 effects coding, 125
 factors, 98–109
 intercept, 56
 linearly dependent, 78
 linearly related, 194–195
 polynomial, 56, 109–113
 principal component, 116–119
- splines, 113–116
 transformed predictors, 56
- Regularized methods, 244–245
- Reliability of hypothesis testing, 148
- Repeated measures, 171
- Reported significance levels, 149
- Research findings, test interpretation, 147–148
- Residual mean square, 26–27
- Residual plots, 166, 209–226
- Residual sampling, bootstrap analysis, 179
- Residual variance, 90–91
- Residuals, 23, 25, 35–38, 204–218
 Pearson, 208
 predicted, 230
 standardized, 216
 studentized, 216
 supernormality, 225–226
 weighted, 156
- Response variable
 logarithmic scale, 82–83
 scatterplots, 2–5
 transformations, 196–198
- Sample surveys, 161–162
- Sampling weight, 162
- Sandwich estimator, 163–167
- Scad, 244
- Scaled power transformations, 189–190
 Box–Cox method, 191
- Scatterplot, 2
- Scatterplot matrix, 15–17
- Score test, nonconstant variance, 166–167
- Score vector, 254–256
- Second-order mean function
 analysis of variance, 141–142
 polynomial regressors, 111–113
- Segmented regression, 263–265
- Separated points, scatterplots, 4–5
- Sequential analysis of variance (Type I), 140–141
- Signs of parameter estimates, 75
- Single coefficient hypotheses, 133
 multiple linear regression, 68–69
 Wald tests, 145–146
- Single linear combination, Wald tests, 146
- Size, scatterplots, 14
- Slices, scatterplots, 4–5
- Slope parameter
 estimates, 73–83
 simple linear regression, 21–22
- Smoothers
 loess, 14, 296–298
 splines, 113–116

- Sparcity principle, 244–245
 Spectral decomposition, 309
 Splines, 113–116
 Square-root transformation, variance stabilization, 172
 Stacking the deck, hypothesis testing, 149–150
 Standard deviation, simple linear regression, 29–30
 Standard error of prediction, 33, 68, 159
 bootstrap analysis, 176–179
 delta method, 172–174
 Standard error of regression, 29–30, 61
 Starting values, nonlinear regression, 257–262
 Statistical error, 21–22
 Stepwise regression, 238, 239–245
 Stratified random sample, sample surveys, 161–162
 Summary graph, 12–14
 Sums of squares
 regression, 35, 63, 134
 residual, 22, 24, 63
 total, 35
 Superpopulation, sample surveys, 162
 Symbols, definitions table, 23
- Taylor series approximation, 254–256
 Test interpretation, 146–150
 bootstrap analysis, 179
 Poisson and binomial regression, 284
 regression diagnostics, outliers, 215–218
 Term. *See* Regressors
 Test statistics, power transformations,
 automatic predictor selection, 195–196
 Third-order mean function, 109
 Transformation family, 186–188
 Transformations, 56, 185–203
 additive models, 199
 automatic predictor selection, 195–196
 basic power transformation, 186
 basic principles, 185–186
 Box–Cox method, 190–191, 194–199,
 312–314
 linearly related regressors, 194–195
 log rule, 188
 modified power, 190
 methodology and examples, 191–196
 multivariate, 195
 nonpositive variables, 198–199
 power transformations, 186–188
 range rule, 188
 response, 196–198
 scaled power, 189, 252
 scatterplots, 14
 single predictor variable, 188–190
 variance stabilization, 171–172
 Yeo–Johnson, 198–199
 True discovery, hypothesis testing, 147–148
t-Tests
 misspecified variances, 163–167
 multiple linear regression, 68
 one-factor models, 102
 main effects model, 107–108
 Poisson and binomial regression, 284
 regression diagnostics, outliers, 217–218
 simple linear regression, 30–34
 two sample, 44
 Tukey’s test for nonadditivity, 212–213
 Type II analysis of variance, 140–141
- Uncorrected sums of squares, 61–62
 Uncorrelated data, scatterplots, 8–9
 Unexplained variation
 multiple linear regression, coefficient of determination (R^2), 67–68
 simple linear regression, coefficient of determination (R^2), 35–36
 Univariate summary statistics
 multiple regression, 57–58
 simple linear regression, 23–24
- Validation set, variables selection, 247
 Variable selection, 234–251
 discovery, 237–245
 information criteria, 238–239
 regularized methods, 244–245
 stepwise regression, 239–244
 subset selection, 245
 parameter assessment, 235–237
 Poisson and binomial regression, 285
 prediction, model selection for, 245–248
 cross-validation, 247
 professor ratings, 247–248
 Variance estimation
 bootstrap method, 174–179
 nonlinear parameter functions, 178
 regression inference, no normality, 175–178
 residual bootstrap, 179
 delta method, 174
 multiple linear regression, 66
 nonlinear regression, 253–256
 simple linear regression, 26–27
 tests, 179

- Variance inflation factor, 249
Variances
general correlation structures, 168–169
misspecified variance, 162–167
accommodation, 163–164
constant variance test, 164–167
mixed models, 169–171
multiple linear regression, 58–59
overview, 156–179
Poisson and binomial regression, 284
scatterplots, 12–14
simple linear regression, 21–22
stabilizing transformations, 171–172
weighted least squares, 156–162
- Wald tests, 133, 145–146
likelihood ratio test comparison, 146
single coefficient hypotheses, 145–146
- Weighted least squares (WLS)
constant variance test, 166–167
regression diagnostics
outliers, 216
weighted hat matrix, residuals, 208
variances, 156–162
group means weighting, 159–161
sample surveys, 161–162
- Wilkinson–Rogers notation, 101, 106–109,
139, 151, 259
binomial regression, 276–277
- Working residual, nonlinear mean function
estimation, 255
- W* statistic, regression diagnostics, 226
- Yeo–Johnson transformation, nonpositive
variables, 198–199
- Zipf's law, 48

WILEY SERIES IN PROBABILITY AND STATISTICS
ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Harvey Goldstein, Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*

Editors Emeriti: *Vic Barnett, J. Stuart Hunter, Joseph B. Kadane, Jozef L. Teugels*

The **Wiley Series in Probability and Statistics** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

- † ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data, *Second Edition*
AGRESTI · An Introduction to Categorical Data Analysis, *Second Edition*
AGRESTI · Categorical Data Analysis, *Third Edition*
ALTMAN, GILL, and McDONALD · Numerical Issues in Statistical Computing for the Social Scientist
AMARATUNGA and CABRERA · Exploration and Analysis of DNA Microarray and Protein Array Data
AMARATUNGA, CABRERA, and SHKEDY . Exploration and Analysis of DNA Microarray and Other High-Dimensional Data, *Second Edition*
ANDĚL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, *Third Edition*
* ANDERSON · The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) · Advances in Biometry
ARNOLD, BALAKRISHNAN, and NAGARAJA · Records
* ARTHANARI and DODGE · Mathematical Programming in Statistics
* BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BAJORSKI · Statistics for Imaging, Optics, and Photonics
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications
BALAKRISHNAN and NG · Precedence-Type Tests and Applications
BARNETT · Comparative Statistical Inference, *Third Edition*
BARNETT · Environmental Statistics
BARNETT and LEWIS · Outliers in Statistical Data, *Third Edition*
BARTHOLOMEW, KNOTT, and MOUSTAKI · Latent Variable Models and Factor Analysis: A Unified Approach, *Third Edition*
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference, *Second Edition*
BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- BATES and WATTS · Nonlinear Regression Analysis and Its Applications
- BECHHOFER, SANTNER, and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons
- BEIRLANT, GOEGEBEUR, SEGERS, TEUGELS, and DE WAAL · Statistics of Extremes: Theory and Applications
- † BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression
- † BELSLEY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
- BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, *Fourth Edition*
- BERNARDO and SMITH · Bayesian Theory
- BHAT and MILLER · Elements of Applied Stochastic Processes, *Third Edition*
- BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications
- BIEMER, GROVES, LYBERG, MATHIOWETZ, and SUDMAN · Measurement Errors in Surveys
- BILLINGSLEY · Convergence of Probability Measures, *Second Edition*
- BILLINGSLEY · Probability and Measure, *Anniversary Edition*
- BIRKES and DODGE · Alternative Methods of Regression
- BISGAARD and KULAHCI · Time Series Analysis and Forecasting by Example
- BISWAS, DATTA, FINE, and SEGAL · Statistical Advances in the Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics
- BLISCHKE and MURTHY (editors) · Case Studies in Reliability and Maintenance
- BLISCHKE and MURTHY · Reliability: Modeling, Prediction, and Optimization
- BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, *Second Edition*
- BOLLEN · Structural Equations with Latent Variables
- BOLLEN and CURRAN · Latent Curve Models: A Structural Equation Perspective
- BOROVKOV · Ergodicity and Stability of Stochastic Processes
- BOSQ and BLANKE · Inference and Prediction in Large Dimensions
- BOULEAU · Numerical Methods for Stochastic Processes
- * BOX and TIAO · Bayesian Inference in Statistical Analysis
- BOX · Improving Almost Anything, *Revised Edition*
- * BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement
- BOX and DRAPER · Response Surfaces, Mixtures, and Ridge Analyses, *Second Edition*
- BOX, HUNTER, and HUNTER · Statistics for Experimenters: Design, Innovation, and Discovery, *Second Edition*
- BOX, JENKINS, and REINSEL · Time Series Analysis: Forecasting and Control, *Fourth Edition*
- BOX, LUCEÑO, and PANIAGUA-QUÍONES · Statistical Control by Monitoring and Adjustment, *Second Edition*
- * BROWN and HOLLANDER · Statistics: A Biomedical Introduction
- CAIROLI and DALANG · Sequential Stochastic Optimization
- CASTILLO, HADI, BALAKRISHNAN, and SARABIA · Extreme Value and Related Models with Applications in Engineering and Science
- CHAN · Time Series: Applications to Finance with R and S-Plus®, *Second Edition*
- CHARALAMBIDES · Combinatorial Methods in Discrete Distributions
- CHATTERJEE and HADI · Regression Analysis by Example, *Fourth Edition*
- CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression
- CHERNICK · Bootstrap Methods: A Guide for Practitioners and Researchers, *Second Edition*
- CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences
- CHILÈS and DELFINER · Geostatistics: Modeling Spatial Uncertainty, *Second Edition*
- CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies, *Third Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- CLARKE · Linear Models: The Theory and Application of Analysis of Variance
- CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, *Second Edition*
- * COCHRAN and COX · Experimental Designs, *Second Edition*
- COLLINS and LANZA · Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences
- CONGDON · Applied Bayesian Modelling
- CONGDON · Bayesian Models for Categorical Data
- CONGDON · Bayesian Statistical Modelling, *Second Edition*
- CONOVER · Practical Nonparametric Statistics, *Third Edition*
- COOK · Regression Graphics
- COOK and WEISBERG · An Introduction to Regression Graphics
- COOK and WEISBERG · Applied Regression Including Computing and Graphics
- CORNELL · A Primer on Experiments with Mixtures
- CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Third Edition*
- COX · A Handbook of Introductory Statistical Methods
- CRESSIE · Statistics for Spatial Data, *Revised Edition*
- CRESSIE and WIKLE · Statistics for Spatio-Temporal Data
- CSÖRGÖ and HORVÁTH · Limit Theorems in Change Point Analysis
- DAGPUNAR · Simulation and Monte Carlo: With Applications in Finance and MCMC
- DANIEL · Applications of Statistics to Industrial Experimentation
- DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Eighth Edition*
- * DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*
- DASU and JOHNSON · Exploratory Data Mining and Data Cleaning
- DAVID and NAGARAJA · Order Statistics, *Third Edition*
- * DEGROOT, FIENBERG, and KADANE · Statistics and the Law
- DEL CASTILLO · Statistical Process Adjustment for Quality Control
- DEMARIS · Regression with Social Data: Modeling Continuous and Limited Response Variables
- DEMIDENKO · Mixed Models: Theory and Applications with R, *Second Edition*
- DENISON, HOLMES, MALLICK and SMITH · Bayesian Methods for Nonlinear Classification and Regression
- DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis
- DEY and MUKERJEE · Fractional Factorial Plans
- DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
- * DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*
- * DOOB · Stochastic Processes
- DOWDY, WEARDEN, and CHILKO · Statistics for Research, *Third Edition*
- DRAPER and SMITH · Applied Regression Analysis, *Third Edition*
- DRYDEN and MARDIA · Statistical Shape Analysis
- DUDEWICZ and MISHRA · Modern Mathematical Statistics
- DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, *Fourth Edition*
- DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations
- EDLER and KITSOS · Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment
- * ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
- ENDERS · Applied Econometric Time Series, *Third Edition*
- † ETHIER and KURTZ · Markov Processes: Characterization and Convergence

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- EVANS, HASTINGS, and PEACOCK · Statistical Distributions, *Third Edition*
 EVERITT, LANDAU, LEESE, and STAHL · Cluster Analysis, *Fifth Edition*
 FEDERER and KING · Variations on Split Plot and Split Block Experiment Designs
 FELLER · An Introduction to Probability Theory and Its Applications, Volume I, *Third Edition*, Revised; Volume II, *Second Edition*
 FITZMAURICE, LAIRD, and WARE · Applied Longitudinal Analysis, *Second Edition*
 * FLEISS · The Design and Analysis of Clinical Experiments
 FLEISS · Statistical Methods for Rates and Proportions, *Third Edition*
 † FLEMING and HARRINGTON · Counting Processes and Survival Analysis
 FUJIKOSHI, ULYANOV, and SHIMIZU · Multivariate Statistics: High-Dimensional and Large-Sample Approximations
 FULLER · Introduction to Statistical Time Series, *Second Edition*
 † FULLER · Measurement Error Models
 GALLANT · Nonlinear Statistical Models
 GEISSER · Modes of Parametric Statistical Inference
 GELMAN and MENG · Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives
 GEWEKE · Contemporary Bayesian Econometrics and Statistics
 GHOSH, MUKHOPADHYAY, and SEN · Sequential Estimation
 GIESBRECHT and GUMPERTZ · Planning, Construction, and Statistical Analysis of Comparative Experiments
 GIFI · Nonlinear Multivariate Analysis
 GIVENS and HOETING · Computational Statistics
 GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems
 GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*
 GOLDSTEIN · Multilevel Statistical Models, *Fourth Edition*
 GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues
 GOLDSTEIN and WOOFF · Bayes Linear Statistics
 GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing
 GROSS, SHORTLE, THOMPSON, and HARRIS · Fundamentals of Queueing Theory, *Fourth Edition*
 GROSS, SHORTLE, THOMPSON, and HARRIS · Solutions Manual to Accompany Fundamentals of Queueing Theory, *Fourth Edition*
 * HAHN and SHAPIRO · Statistical Models in Engineering
 HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners
 HALD · A History of Probability and Statistics and their Applications Before 1750
 † HAMPEL · Robust Statistics: The Approach Based on Influence Functions
 HARTUNG, KNAPP, and SINHA · Statistical Meta-Analysis with Applications
 HEIBERGER · Computation for the Analysis of Designed Experiments
 HEDAYAT and SINHA · Design and Inference in Finite Population Sampling
 HEDEKER and GIBBONS · Longitudinal Data Analysis
 HELLER · MACSYMA for Statisticians
 HERITIER, CANTONI, COPT, and VICTORIA-FESER · Robust Methods in Biostatistics
 HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design, *Second Edition*
 HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 2: Advanced Experimental Design
 HINKELMANN (editor) · Design and Analysis of Experiments, Volume 3: Special Designs and Applications
 HOAGLIN, MOSTELLER, and TUKEY · Fundamentals of Exploratory Analysis of Variance

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- * HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes
- * HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory Data Analysis
- HOCHBERG and TAMHANE · Multiple Comparison Procedures
- HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variance, *Third Edition*
- HOEL · Introduction to Mathematical Statistics, *Fifth Edition*
- HOGG and KLUGMAN · Loss Distributions
- HOLLANDER, WOLFE, and CHICKEN · Nonparametric Statistical Methods, *Third Edition*
- HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*
- HOSMER, LEMESHOW, and MAY · Applied Survival Analysis: Regression Modeling of Time-to-Event Data, *Second Edition*
- HUBER · Data Analysis: What Can Be Learned From the Past 50 Years
- HUBER · Robust Statistics
- † HUBER and RONCHETTI · Robust Statistics, *Second Edition*
- HUBERTY · Applied Discriminant Analysis, *Second Edition*
- HUBERTY and OLEJNIK · Applied MANOVA and Discriminant Analysis, *Second Edition*
- HUITEMA · The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, *Second Edition*
- HUNT and KENNEDY · Financial Derivatives in Theory and Practice, *Revised Edition*
- HURD and MIAMEE · Periodically Correlated Random Sequences: Spectral Theory and Practice
- HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek—with Commentary
- HUZURBAZAR · Flowgraph Models for Multistate Time-to-Event Data
- JACKMAN · Bayesian Analysis for the Social Sciences
- † JACKSON · A User's Guide to Principle Components
- JOHN · Statistical Methods in Engineering and Quality Assurance
- JOHNSON · Multivariate Statistical Simulation
- JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz
- JOHNSON, KEMP, and KOTZ · Univariate Discrete Distributions, *Third Edition*
- JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present
- JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 1, *Second Edition*
- JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 2, *Second Edition*
- JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions
- JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE · The Theory and Practice of Econometrics, *Second Edition*
- JUREK and MASON · Operator-Limit Distributions in Probability Theory
- KADANE · Bayesian Methods and Ethics in a Clinical Trial Design
- KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence
- KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, *Second Edition*
- KARIYA and KURATA · Generalized Least Squares
- KASS and VOS · Geometrical Foundations of Asymptotic Inference
- † KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis
- KEDEM and FOKIANOS · Regression Models for Time Series Analysis
- KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- KHURI · Advanced Calculus with Applications in Statistics, *Second Edition*
 KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models
- * KISH · Statistical Design for Research
 KLEIBER and KOTZ · Statistical Size Distributions in Economics and Actuarial Sciences
 KLEMELÄ · Smoothing of Multivariate Data: Density Estimation and Visualization
 KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions, *Third Edition*
 KLUGMAN, PANJER, and WILLMOT · Loss Models: Further Topics
 KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions, *Third Edition*
- KOSKI and NOBLE · Bayesian Networks: An Introduction
 KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions, Volume 1, *Second Edition*
 KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index
 KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Supplement Volume
 KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 1
 KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 2
- KOWALSKI and TU · Modern Applied U-Statistics
 KRISHNAMOORTHY and MATHEW · Statistical Tolerance Regions: Theory, Applications, and Computation
 KROESE, TAIMRE, and BOTEV · Handbook of Monte Carlo Methods
 KROONENBERG · Applied Multiway Data Analysis
 KULINSKAYA, MORGENTHALER, and STAUDTE · Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence
 KULKARNI and HARMAN · An Elementary Introduction to Statistical Learning Theory
 KUROWICKA and COOKE · Uncertainty Analysis with High Dimensional Dependence Modelling
- KVAM and VIDAKOVIC · Nonparametric Statistics with Applications to Science and Engineering
 LACHIN · Biostatistical Methods: The Assessment of Relative Risks, *Second Edition*
 LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction
 LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*
 LAWLESS · Statistical Models and Methods for Lifetime Data, *Second Edition*
 LAWSON · Statistical Methods in Spatial Epidemiology, *Second Edition*
 LE · Applied Categorical Data Analysis, *Second Edition*
 LE · Applied Survival Analysis
 LEE · Structural Equation Modeling: A Bayesian Approach
 LEE and WANG · Statistical Methods for Survival Data Analysis, *Fourth Edition*
 LEPAGE and BILLARD · Exploring the Limits of Bootstrap
 LESSLER and KALSBEK · Nonsampling Errors in Surveys
 LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics
 LIAO · Statistical Group Comparison
 LIN · Introductory Stochastic Analysis for Finance and Insurance
 LINDLEY · Understanding Uncertainty, *Revised Edition*
 LITTLE and RUBIN · Statistical Analysis with Missing Data, *Second Edition*
 LLOYD · The Statistical Analysis of Categorical Data
 LOWEN and TEICH · Fractal-Based Point Processes
 MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

- MALLER and ZHOU · Survival Analysis with Long Term Survivors
- MARCHETTE · Random Graphs for Statistical Pattern Recognition
- MARDIA and JUPP · Directional Statistics
- MARKOVICH · Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice
- MARONNA, MARTIN and YOHAI · Robust Statistics: Theory and Methods
- MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, *Second Edition*
- McCULLOCH, SEARLE, and NEUHAUS · Generalized, Linear, and Mixed Models, *Second Edition*
- McFADDEN · Management of Data in Clinical Trials, *Second Edition*
- * McLACHLAN · Discriminant Analysis and Statistical Pattern Recognition
- McLACHLAN, DO, and AMBROISE · Analyzing Microarray Gene Expression Data
- McLACHLAN and KRISHNAN · The EM Algorithm and Extensions, *Second Edition*
- McLACHLAN and PEEL · Finite Mixture Models
- MCNEIL · Epidemiological Research Methods
- MEEKER and ESCOBAR · Statistical Methods for Reliability Data
- MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice
- MENGERSEN, ROBERT, and TITTERINGTON · Mixtures: Estimation and Applications
- MICKEY, DUNN, and CLARK · Applied Statistics: Analysis of Variance and Regression, *Third Edition*
- * MILLER · Survival Analysis, *Second Edition*
- MONTGOMERY, JENNINGS, and KULAHCI · Introduction to Time Series Analysis and Forecasting
- MONTGOMERY, PECK, and VINING · Introduction to Linear Regression Analysis, *Fifth Edition*
- MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness
- MUIRHEAD · Aspects of Multivariate Statistical Theory
- MULLER and STOYAN · Comparison Methods for Stochastic Models and Risks
- MURTHY, XIE, and JIANG · Weibull Models
- MYERS, MONTGOMERY, and ANDERSON-COOK · Response Surface Methodology: Process and Product Optimization Using Designed Experiments, *Third Edition*
- MYERS, MONTGOMERY, VINING, and ROBINSON · Generalized Linear Models. With Applications in Engineering and the Sciences, *Second Edition*
- NATVIG · Multistate Systems Reliability Theory With Applications
- † NELSON · Accelerated Testing, Statistical Models, Test Plans, and Data Analyses
- † NELSON · Applied Life Data Analysis
- NEWMAN · Biostatistical Methods in Epidemiology
- NG, TAIN, and TANG · Dirichlet Theory: Theory, Methods and Applications
- OKABE, BOOTS, SUGIHARA, and CHIU · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, *Second Edition*
- OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis
- PALTA · Quantitative Methods in Population Health: Extensions of Ordinary Regressions
- PANJER · Operational Risk: Modeling and Analytics
- PANKRATZ · Forecasting with Dynamic Regression Models
- PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
- PARDOUX · Markov Processes and Applications: Algorithms, Networks, Genome and Finance
- PARMIGIANI and INOUE · Decision Theory: Principles and Approaches
- * PARZEN · Modern Probability Theory and Its Applications
- PEÑA, TIAO, and TSAY · A Course in Time Series Analysis

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- PESARIN and SALMASO · Permutation Tests for Complex Data: Applications and Software
- PIANTADOSI · Clinical Trials: A Methodologic Perspective, *Second Edition*
- POURAHMADI · Foundations of Time Series Analysis and Prediction Theory
- POURAHMADI · High-Dimensional Covariance Estimation
- POWELL · Approximate Dynamic Programming: Solving the Curses of Dimensionality, *Second Edition*
- POWELL and RYZHOV · Optimal Learning
- PRESS · Subjective and Objective Bayesian Statistics, *Second Edition*
- PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach
- PURI, VILAPLANA, and WERTZ · New Perspectives in Theoretical and Applied Statistics
- † PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming
- QUI · Image Processing and Jump Regression Analysis
- * RAO · Linear Statistical Inference and Its Applications, *Second Edition*
- RAO · Statistical Inference for Fractional Diffusion Processes
- RAUSAND and HØYLAND · System Reliability Theory: Models, Statistical Methods, and Applications, *Second Edition*
- RAYNER, THAS, and BEST · Smooth Tests of Goodness of Fit: Using R, *Second Edition*
- RENCHER and SCHAALJE · Linear Models in Statistics, *Second Edition*
- RENCHER and CHRISTENSEN · Methods of Multivariate Analysis, *Third Edition*
- RENCHER · Multivariate Statistical Inference with Applications
- RIGDON and BASU · Statistical Methods for the Reliability of Repairable Systems
- * RIPLEY · Spatial Statistics
- * RIPLEY · Stochastic Simulation
- ROHATGI and SALEH · An Introduction to Probability and Statistics, *Second Edition*
- ROLSKI, SCHMIDL, SCHMIDT, and TEUGELS · Stochastic Processes for Insurance and Finance
- ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice
- ROSSI, ALLENBY, and MCCULLOCH · Bayesian Statistics and Marketing
- † ROUSSEEUW and LEROY · Robust Regression and Outlier Detection
- ROYSTON and SAUERBREI · Multivariate Model Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modeling Continuous Variables
- * RUBIN · Multiple Imputation for Nonresponse in Surveys
- RUBINSTEIN and KROESE · Simulation and the Monte Carlo Method, *Second Edition*
- RUBINSTEIN and MELAMED · Modern Simulation and Modeling
- RUBINSTEIN, RIDDER, and VAISMAN · Fast Sequential Monte Carlo Methods for Counting and Optimization
- RYAN · Modern Engineering Statistics
- RYAN · Modern Experimental Design
- RYAN · Modern Regression Methods, *Second Edition*
- RYAN · Sample Size Determination and Power
- RYAN · Statistical Methods for Quality Improvement, *Third Edition*
- SALEH · Theory of Preliminary Test and Stein-Type Estimation with Applications
- SALTELLI, CHAN, and SCOTT (editors) · Sensitivity Analysis
- SCHERER · Batch Effects and Noise in Microarray Experiments: Sources and Solutions
- * SCHEFFE · The Analysis of Variance
- SCHIMEK · Smoothing and Regression: Approaches, Computation, and Application
- SCHOTT · Matrix Analysis for Statistics, *Second Edition*
- SCHOUTENS · Levy Processes in Finance: Pricing Financial Derivatives
- SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization
- * SEARLE · Linear Models
- † SEARLE · Linear Models for Unbalanced Data

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- [†] SEARLE · Matrix Algebra Useful for Statistics
[†] SEARLE, CASELLA, and McCULLOCH · Variance Components
 SEARLE and WILLETT · Matrix Algebra for Applied Economics
 SEBER · A Matrix Handbook For Statisticians
[†] SEBER · Multivariate Observations
 SEBER and LEE · Linear Regression Analysis, *Second Edition*
[†] SEBER and WILD · Nonlinear Regression
 SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems
* SERFLING · Approximation Theorems of Mathematical Statistics
 SHAFER and VOVK · Probability and Finance: It's Only a Game!
 SHERMAN · Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties
 SILVAPULLE and SEN · Constrained Statistical Inference: Inequality, Order, and Shape Restrictions
 SINGPURWALLA · Reliability and Risk: A Bayesian Perspective
 SMALL and MCLEISH · Hilbert Space Methods in Probability and Statistical Inference
 SRIVASTAVA · Methods of Multivariate Statistics
 STAPLETON · Linear Statistical Models, *Second Edition*
 STAPLETON · Models for Probability and Statistical Inference: Theory and Applications
 STAUDTE and SHEATHER · Robust Estimation and Testing
 STOYAN · Counterexamples in Probability, *Second Edition*
 STOYAN, KENDALL, and MECKE · Stochastic Geometry and Its Applications, *Second Edition*
 STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics
 STREET and BURGESS · The Construction of Optimal Stated Choice Experiments: Theory and Methods
 STYAN · The Collected Papers of T. W. Anderson: 1943–1985
 SUTTON, ABRAMS, JONES, SHELDON, and SONG · Methods for Meta-Analysis in Medical Research
 TAKEZAWA · Introduction to Nonparametric Regression
 TAMHANE · Statistical Analysis of Designed Experiments: Theory and Applications
 TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
 THOMPSON · Empirical Model Building: Data, Models, and Reality, *Second Edition*
 THOMPSON · Sampling, *Third Edition*
 THOMPSON · Simulation: A Modeler's Approach
 THOMPSON and SEBER · Adaptive Sampling
 THOMPSON, WILLIAMS, and FINDLAY · Models for Investors in Real World Markets
 TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
 TSAY · Analysis of Financial Time Series, *Third Edition*
 TSAY · An Introduction to Analysis of Financial Data with R
 TSAY · Multivariate Time Series Analysis: With R and Financial Applications
 UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
[†] VAN BELLE · Statistical Rules of Thumb, *Second Edition*
 VAN BELLE, FISHER, HEAGERTY, and LUMLEY · Biostatistics: A Methodology for the Health Sciences, *Second Edition*
 VESTRUP · The Theory of Measures and Integration
 VIDAKOVIC · Statistical Modeling by Wavelets
 VIERTL · Statistical Methods for Fuzzy Data
 VINOD and REAGLE · Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments

*Now available in a lower priced paperback edition in the Wiley Classics Library.

[†]Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- WALLER and GOTWAY · Applied Spatial Statistics for Public Health Data
WEISBERG · Applied Linear Regression, *Fourth Edition*
WEISBERG · Bias and Causation: Models and Judgment for Valid Comparisons
WELSH · Aspects of Statistical Inference
WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment
* WHITTAKER · Graphical Models in Applied Multivariate Statistics
WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting
WOODWORTH · Biostatistics: A Bayesian Introduction
WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data,
Second Edition
WU and HAMADA · Experiments: Planning, Analysis, and Parameter Design
Optimization, *Second Edition*
WU and ZHANG · Nonparametric Regression Methods for Longitudinal Data Analysis
YIN · Clinical Trial Design: Bayesian and Frequentist Adaptive Methods
YOUNG, VALERO-MORA, and FRIENDLY · Visual Statistics: Seeing Data with
Dynamic Interactive Graphics
ZACKS · Examples and Problems in Mathematical Statistics
ZACKS · Stage-Wise Adaptive Designs
* ZELLNER · An Introduction to Bayesian Inference in Econometrics
ZELTERMAN · Discrete Distributions—Applications in the Health Sciences
ZHOU, OBUCHOWSKI, and McCLISH · Statistical Methods in Diagnostic Medicine,
Second Edition

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.