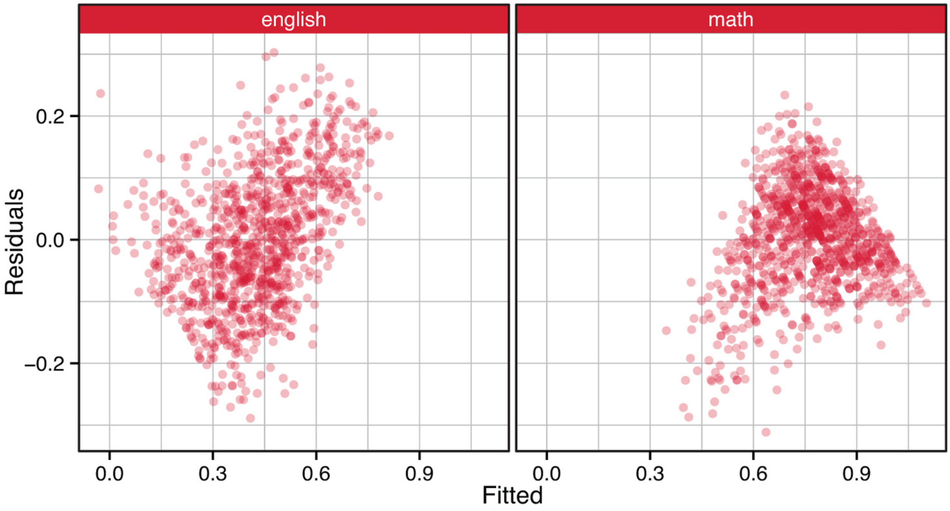
Texts in Statistical Science

Extending the Linear Model with R

Generalized Linear, Mixed Effects and Nonparametric Regression Models

SECOND EDITION



Julian J. Faraway

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Accessing the E-book edition

Using the VitalSource® ebook

Access to the VitalBook™ ebook accompanying this book is via VitalSource® Bookshelf – an ebook reader which allows you to make and share notes and highlights on your ebooks and search across all of the ebooks that you hold on your VitalSource Bookshelf. You can access the ebook online or offline on your smartphone, tablet or PC/Mac and your notes and highlights will automatically stay in sync no matter where you make them.

1. **Create a VitalSource Bookshelf account at** ***https://online.vitalsource.com/user/new*** or log into your existing account if you already have one.
2. **Redeem the code provided in the panel below to get online access to the ebook.** Log in to Bookshelf and select **Redeem** at the top right of the screen. Enter the redemption code shown on the scratch-off panel below in the **Redeem Code** pop-up and press **Redeem**. Once the code has been redeemed your ebook will download and appear in your library.

No returns if this code has been revealed.

DOWNLOAD AND READ OFFLINE

To use your ebook offline, download BookShelf to your PC, Mac, iOS device, Android device or Kindle Fire, and log in to your Bookshelf account to access your ebook:

On your PC/Mac

Go to ***https://support.vitalsource.com/hc/en-us*** and follow the instructions to download the free **VitalSource Bookshelf** app to your PC or Mac and log into your Bookshelf account.

On your iPhone/iPod Touch/iPad

Download the free **VitalSource Bookshelf** App available via the iTunes App Store and log into your Bookshelf account. You can find more information at ***https://support.vitalsource.com/hc/en-us/categories/200134217-Bookshelf-for-iOS***

On your Android™ smartphone or tablet

Download the free **VitalSource Bookshelf** App available via Google Play and log into your Bookshelf account. You can find more information at ***https://support.vitalsource.com/hc/en-us/categories/200139976-Bookshelf-for-Android-and-Kindle-Fire***

On your Kindle Fire

Download the free **VitalSource Bookshelf** App available from Amazon and log into your Bookshelf account. You can find more information at ***https://support.vitalsource.com/hc/en-us/categories/200139976-Bookshelf-for-Android-and-Kindle-Fire***

N.B. The code in the scratch-off panel can only be used once. When you have created a Bookshelf account and redeemed the code you will be able to access the ebook online or offline on your smartphone, tablet or PC/Mac.

SUPPORT

If you have any questions about downloading Bookshelf, creating your account, or accessing and using your ebook edition, please visit ***http://support.vitalsource.com/***

Extending the Linear Model with R

Generalized Linear, Mixed Effects and Nonparametric Regression Models

SECOND EDITION

CHAPMAN & HALL/CRC
Texts in Statistical Science Series

Statistical Theory: A Concise Introduction
F. Abramovich and Y. Ritov

Practical Multivariate Analysis, Fifth Edition
A. Afifi, S. May, and V.A. Clark

Practical Statistics for Medical Research
D.G. Altman

Interpreting Data: A First Course in Statistics
A.J.B. Anderson

Introduction to Probability with R
K. Baclawski

Linear Algebra and Matrix Analysis for Statistics
S. Banerjee and A. Roy

Mathematical Statistics: Basic Ideas and Selected Topics, Volume I, Second Edition
P. J. Bickel and K. A. Doksum

Mathematical Statistics: Basic Ideas and Selected Topics, Volume II
P. J. Bickel and K. A. Doksum

Analysis of Categorical Data with R
C. R. Bilder and T. M. Loughin

Statistical Methods for SPC and TQM
D. Bissell

Introduction to Probability
J. K. Blitzstein and J. Hwang

Bayesian Methods for Data Analysis, Third Edition
B.P. Carlin and T.A. Louis

Second Edition
R. Caulcutt

The Analysis of Time Series: An Introduction, Sixth Edition
C. Chatfield

Introduction to Multivariate Analysis
C. Chatfield and A.J. Collins

Problem Solving: A Statistician's Guide, Second Edition
C. Chatfield

Statistics for Technology: A Course in Applied Statistics, Third Edition
C. Chatfield

Analysis of Variance, Design, and Regression: Linear Modeling for Unbalanced Data, Second Edition
R. Christensen

Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians
R. Christensen, W. Johnson, A. Branscum, and T.E. Hanson

Modelling Binary Data, Second Edition
D. Collett

Modelling Survival Data in Medical Research, Third Edition
D. Collett

Introduction to Statistical Methods for Clinical Trials
T.D. Cook and D.L. DeMets

Applied Statistics: Principles and Examples
D.R. Cox and E.J. Snell

Multivariate Survival Analysis and Competing Risks
M. Crowder

Statistical Analysis of Reliability Data
M.J. Crowder, A.C. Kimber, T.J. Sweeting, and R.L. Smith

An Introduction to Generalized Linear Models, Third Edition
A.J. Dobson and A.G. Barnett

Nonlinear Time Series: Theory, Methods, and Applications with R Examples
R. Douc, E. Moulines, and D.S. Stoffer

Introduction to Optimization Methods and Their Applications in Statistics
B.S. Everitt

Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, Second Edition
J.J. Faraway

Linear Models with R, Second Edition
J.J. Faraway

A Course in Large Sample Theory
T.S. Ferguson

Multivariate Statistics: A Practical Approach
B. Flury and H. Riedwyl

Readings in Decision Analysis
S. French

Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data
M. Friendly and D. Meyer

Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition
D. Gamerman and H.F. Lopes

Bayesian Data Analysis, Third Edition
A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin

Multivariate Analysis of Variance and Repeated Measures: A Practical Approach for Behavioural Scientists
D.J. Hand and C.C. Taylor

Practical Longitudinal Data Analysis
D.J. Hand and M. Crowder

Logistic Regression Models
J.M. Hilbe

Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects
J.S. Hodges

Statistics for Epidemiology
N.P. Jewell

Stochastic Processes: An Introduction, Second Edition
P.W. Jones and P. Smith

The Theory of Linear Models
B. Jørgensen

Principles of Uncertainty
J.B. Kadane

Graphics for Statistics and Data Analysis with R
K.J. Keen

Mathematical Statistics
K. Knight

Introduction to Multivariate Analysis: Linear and Nonlinear Modeling
S. Konishi

Nonparametric Methods in Statistics with SAS Applications
O. Korosteleva

Modeling and Analysis of Stochastic Systems, Second Edition
V.G. Kulkarni

Exercises and Solutions in Biostatistical Theory
L.L. Kupper, B.H. Neelon, and S.M. O'Brien

Exercises and Solutions in Statistical Theory
L.L. Kupper, B.H. Neelon, and S.M. O'Brien

Design and Analysis of Experiments with R
J. Lawson

Design and Analysis of Experiments with SAS
J. Lawson

A Course in Categorical Data Analysis
T. Leonard

Statistics for Accountants
S. Letchford

Introduction to the Theory of Statistical Inference
H. Liero and S. Zwanzig

Statistical Theory, Fourth Edition
B.W. Lindgren

Stationary Stochastic Processes: Theory and Applications
G. Lindgren

Statistics for Finance
E. Lindström, H. Madsen, and J. N. Nielsen

The BUGS Book: A Practical Introduction to Bayesian Analysis
D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter

Introduction to General and Generalized Linear Models
H. Madsen and P. Thyregod

Time Series Analysis
H. Madsen

Pólya Urn Models
H. Mahmoud

Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition
B.F.J. Manly

Introduction to Randomized Controlled Clinical Trials, Second Edition
J.N.S. Matthews

Statistical Rethinking: A Bayesian Course with Examples in R and Stan
R. McElreath

Statistical Methods in Agriculture and Experimental Biology, Second Edition
R. Mead, R.N. Curnow, and A.M. Hasted

Statistics in Engineering: A Practical Approach
A.V. Metcalfe

Statistical Inference: An Integrated Approach, Second Edition
H. S. Migon, D. Gamerman, and F. Louzada

Beyond ANOVA: Basics of Applied Statistics
R.G. Miller, Jr.

A Primer on Linear Models
J.F. Monahan

Applied Stochastic Modelling, Second Edition
B.J.T. Morgan

Elements of Simulation
B.J.T. Morgan

Probability: Methods and Measurement
A. O'Hagan

Introduction to Statistical Limit Theory
A.M. Polansky

Applied Bayesian Forecasting and Time Series Analysis
A. Pole, M. West, and J. Harrison

Statistics in Research and Development, Time Series: Modeling, Computation, and Inference
R. Prado and M. West

Essentials of Probability Theory for Statisticians
M.A. Proschan and P.A. Shaw

Introduction to Statistical Process Control
P. Qiu

Sampling Methodologies with Applications
P.S.R.S. Rao

A First Course in Linear Model Theory
N. Ravishanker and D.K. Dey

Essential Statistics, Fourth Edition
D.A.G. Rees

Stochastic Modeling and Mathematical Statistics: A Text for Statisticians and Quantitative Scientists
F.J. Samaniego

Statistical Methods for Spatial Data Analysis
O. Schabenberger and C.A. Gotway

Bayesian Networks: With Examples in R
M. Scutari and J.-B. Denis

Large Sample Methods in Statistics
P.K. Sen and J. da Motta Singer

Spatio-Temporal Methods in Environmental Epidemiology
G. Shaddick and J.V. Zidek

Decision Analysis: A Bayesian Approach
J.Q. Smith

Analysis of Failure and Survival Data
P. J. Smith

Applied Statistics: Handbook of GENSTAT Analyses
E.J. Snell and H. Simpson

Applied Nonparametric Statistical Methods, Fourth Edition
P. Sprent and N.C. Smeeton

Data Driven Statistical Methods
P. Sprent

Generalized Linear Mixed Models: Modern Concepts, Methods and Applications
W. W. Stroup

Survival Analysis Using S: Analysis of Time-to-Event Data
M. Tableman and J.S. Kim

Applied Categorical and Count Data Analysis
W. Tang, H. He, and X.M. Tu

Elementary Applications of Probability Theory, Second Edition
H.C. Tuckwell

Introduction to Statistical Inference and Its Applications with R
M.W. Trosset

Understanding Advanced Statistical Methods
P.H. Westfall and K.S.S. Henning

Statistical Process Control: Theory and Practice, Third Edition
G.B. Wetherill and D.W. Brown

Generalized Additive Models: An Introduction with R
S. Wood

Epidemiology: Study Design and Data Analysis, Third Edition
M. Woodward

Practical Data Analysis for Designed Experiments
B.S. Yandell

Texts in Statistical Science

Extending the Linear Model with R

Generalized Linear, Mixed Effects and Nonparametric Regression Models

SECOND EDITION

Julian J. Faraway
University of Bath, UK

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

No claim to original U.S. Government works
Version Date: 20160301

International Standard Book Number-13: 978-1-4987-2098-4 (eBook - PDF)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface

Linear models are central to the practice of statistics. They are part of the core knowledge expected of any applied statistician. Linear models are the foundation of a broad range of statistical methodologies; this book is a survey of techniques that grow from a linear model. Our starting point is the regression model with response y and predictors $x_1, \ldots x_p$. The model takes the form:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

where ε is normally distributed. This book presents three extensions to this framework. The first generalizes the y part; the second, the ε part; and the third, the x part of the linear model.

Generalized Linear Models (GLMs)**:** The standard linear model cannot handle nonnormal responses, y, such as counts or proportions. This motivates the development of generalized linear models that can represent categorical, binary and other response types.

Mixed Effect Models: Some data has a grouped, nested or hierarchical structure. Repeated measures, longitudinal and multilevel data consist of several observations taken on the same individual or group. This induces a correlation structure in the error, ε. Mixed effect models allow the modeling of such data.

Nonparametric Regression Models: In the linear model, the predictors, x, are combined in a linear way to model the effect on the response. Sometimes this linearity is insufficient to capture the structure of the data and more flexibility is required. Methods such as additive models, trees and neural networks allow a more flexible regression modeling of the response that combines the predictors in a nonparametric manner.

This book aims to provide the reader with a well-stocked toolbox of statistical methodologies. A practicing statistician needs to be aware of and familiar with the basic use of a broad range of ideas and techniques. This book will be a success if the reader is able to recognize and get started on a wide range of problems. However, the breadth comes at the expense of some depth. Fortunately, there are book-length treatments of topics discussed in every chapter of this book, so the reader will know where to go next if needed.

R is a free software environment for statistical computing and graphics. It runs on a wide variety of platforms including the Windows, Linux and Macintosh operating systems. Although there are several excellent statistical packages, only R is both free and possesses the power to perform the analyses demonstrated in this book. While it is possible in principle to learn statistical methods from purely theoretical expositions, I believe most readers learn best from the demonstrated interplay of theory and practice. The data analysis of real examples is woven into this book and all the R commands necessary to reproduce the analyses are provided.

Prerequisites: Readers should possess some knowledge of linear models. The first chapter provides a review of these models. This book can be viewed as a sequel to *Linear Models with R*, Faraway (2014). Even so there are plenty of other good books on linear models such as Draper and Smith (1998) or Weisberg (2005), that would provide ample grounding. Some knowledge of likelihood theory is also very useful. An outline is provided in Appendix A, but this may be insufficient for those who have never seen it before. A general knowledge of statistical theory is also expected concerning such topics as hypothesis tests or confidence intervals. Even so, the emphasis in this text is on application, so readers without much statistical theory can still learn something here.

This is not a book about learning R, but the reader will inevitably pick up the language by reading through the example data analyses. Readers completely new to R will benefit from studying an introductory book such as Dalgaard (2002) or one of the many tutorials available for free at the R website. Even so, the book should be intelligible to a reader without prior knowledge of R just by reading the text and output. R skills can be further developed by modifying the examples in this book, trying the exercises and studying the help pages for each command as needed. There is a large amount of detailed help on the commands available within the software and there is no point in duplicating that here. Please refer to Appendix B for details on obtaining and installing R along with the necessary add-on packages and data necessary for running the examples in this text.

The website for this book is at `people.bath.ac.uk/jjf23/ELM` where data, updates and errata may be obtained.

Second Edition: Ten years have passed since the publication of the first edition. R has expanded enormously both in popularity and in the number of packages available. I have updated the R content to correct for changes and to take advantage of the greater functionality now available. I have revised or added several topics:

- One chapter on binary and binomial responses has been expanded to three. The analysis of strictly binary responses is sufficiently different to justify a separate treatment from the binomial response. Sections for proportion responses, quasi-binomial and beta regression have been added. Applied considerations regarding these models have been gathered into a third chapter.
- Poisson models with dispersion and zero inflated count models have new sections.
- A section on linear discriminant analysis has been added for contrast with multinomial response models.
- New sections on sandwich and robust estimation for GLMs have been added. Tweedie GLMs are now covered.
- The chapters on random effects and repeated measures have been substantially revised to reflect changes in the `lme4` package that removed many p-values from the output. We show how to do hypothesis testing for these models using other methods.

- I have added a chapter concerning the Bayesian analysis of mixed effect models. There are sufficient drawbacks to the analysis in the existing two chapters that make the Bayes approach rewarding even for non-Bayesians. We venture a little beyond the confines of R in the use of STAN (Stan Development Team (2015)). We also present the approximation method of INLA (Rue et al. (2009)).
- The chapter on generalized linear mixed models has been substantially revised to reflect the much richer choice of fitting software now available. A Bayesian approach has also been included.
- The chapter on nonparametric regression has updated coverage on splines and confidence bands. In additive models, we now use the `mgcv` package exclusively while the multivariate adaptive regression splines (MARS) section has an easier-to-use interface.
- Random forests for regression and classification have been added to the chapter on trees.
- The R code has revamped throughout. In particular, there are many plots using the `ggplot2` package.
- The exercises have been revised and expanded. They are now more point-by-point specific rather than open-ended questions. Solutions are now available.
- The text is about one third longer than the first edition.

My thanks to many past students and readers of the first edition whose comments and questions have helped me make many improvements to this edition. Thanks to the builders of R (R Core Team (2015)) who made all this possible.

This page intentionally left blank

Chapter 1

Introduction

This book is about extending the linear model methodology using R statistical software. Before setting off on this journey, it is worth reviewing both linear models and R. We shall not attempt a detailed description of linear models; the reader is advised to consult texts such as Faraway (2014) or Draper and Smith (1998). We do not intend this as a self-contained introduction to R as this may be found in books such as Dalgaard (2002) or Maindonald and Braun (2010) or from guides obtainable from the R website. Even so, a reader unfamiliar with R should be able to follow the intent of the analysis and learn a little R in the process without further preparation.

Let's consider an example. The 2000 United States Presidential election generated much controversy, particularly in the state of Florida where there were some difficulties with the voting machinery. In Meyer (2002), data on voting in the state of Georgia is presented and analyzed.

Let's take a look at this data using R. Please refer to Appendix B for details on obtaining and installing R along with the necessary add-on packages and data for running the examples in this text. In this book, we denote R commands with bold text in a grey box. You should type this in at the command prompt: >. We start by loading the data:

```
data(gavote, package="faraway")
```

The `data` command loads the particular dataset into R. The name of the dataset is `gavote` and it is being loaded from the package `faraway`. If you get an error message about a package not being found, it probably means you have not installed the `faraway` package. Please check the Appendix.

An alternative means of making the data is to load the `faraway` package:

```
library(faraway)
```

This will make all the data and functions in this package available for this R session.

In R, the object containing the data is called a *dataframe*. We can obtain definitions of the variables and more information about the dataset using the `help` command:

```
help(gavote)
```

You can use the `help` command to learn more about any of the commands we use. For example, to learn about the `quantile` command:

```
help(quantile)
```

If you do not already know or guess the name of the command you need, use:

```
help.search("quantiles")
```

to learn about all commands that refer to quantiles.

We can examine the contents of the dataframe simply by typing its name:

```
gavote
            equip   econ perAA rural    atlanta   gore   bush other  votes ballots
APPLING     LEVER   poor 0.182 rural notAtlanta   2093   3940    66   6099    6617
ATKINSON    LEVER   poor 0.230 rural notAtlanta    821   1228    22   2071    2149
....
```

The output in this text is shown in `typewriter font`. I have deleted most of the output to save space. This dataset is small enough to be comfortably examined in its entirety. Sometimes, we simply want to look at the first few cases. The `head` command is useful for this:

```
head(gavote)
          equip   econ perAA rural    atlanta gore bush other votes ballots
APPLING   LEVER   poor 0.182 rural notAtlanta 2093 3940    66  6099    6617
ATKINSON  LEVER   poor 0.230 rural notAtlanta  821 1228    22  2071    2149
BACON     LEVER   poor 0.131 rural notAtlanta  956 2010    29  2995    3347
BAKER     OS-CC   poor 0.476 rural notAtlanta  893  615    11  1519    1607
BALDWIN   LEVER middle 0.359 rural notAtlanta 5893 6041   192 12126   12785
BANKS     LEVER middle 0.024 rural notAtlanta 1220 3202   111  4533    4773
```

The cases in this dataset are the counties of Georgia and the variables are (in order) the type of voting equipment used, the economic level of the county, the percentage of African Americans, whether the county is rural or urban, whether the county is part of the Atlanta metropolitan area, the number of voters for Al Gore, the number of voters for George Bush, the number of voters for other candidates, the number of votes cast, and ballots issued.

The `str` command is another useful way to examine an R object:

```
str(gavote)
'data.frame':       159 obs. of  10 variables:
 $ equip  : Factor w/ 5 levels "LEVER","OS-CC",..: 1 1 1 2 1 1 2 3 3 2 ...
 $ econ   : Factor w/ 3 levels "middle","poor",..: 2 2 2 2 1 1 1 1 2 2 ...
 $ perAA  : num  0.182 0.23 0.131 0.476 0.359 0.024 0.079 0.079 0.282 0.107 ...
 $ rural  : Factor w/ 2 levels "rural","urban": 1 1 1 1 1 1 2 2 1 1 ...
 $ atlanta: Factor w/ 2 levels "Atlanta","notAtlanta": 2 2 2 2 2 2 2 1 2 2 ...
 $ gore   : int  2093 821 956 893 5893 1220 3657 7508 2234 1640 ...
 $ bush   : int  3940 1228 2010 615 6041 3202 7925 14720 2381 2718 ...
 $ other  : int  66 22 29 11 192 111 520 552 46 52 ...
 $ votes  : int  6099 2071 2995 1519 12126 4533 12102 22780 4661 4410 ...
 $ ballots: int  6617 2149 3347 1607 12785 4773 12522 23735 5741 4475 ...
```

We can see that some of the variables, such as the equipment type, are factors. Factor variables are categorical. Other variables are quantitative. The `perAA` variable is continuous while the others are integer valued. We also see the sample size is 159.

A potential voter goes to the polling station where it is determined whether he or she is registered to vote. If so, a ballot is issued. However, a vote is not recorded if the person fails to vote for President, votes for more than one candidate or the equipment fails to record the vote. For example, we can see that in Appling county, $6617 - 6099 = 518$ ballots did not result in votes for President. This is called the *undercount*. The purpose of our analysis will be to determine what factors affect the undercount. We will not attempt a full and conclusive analysis here because our main purpose is to illustrate the use of linear models and R. We invite the reader to fill in some of the gaps in the analysis.

Initial Data Analysis: The first stage in any data analysis should be an initial graphical and numerical look at the data. A compact numerical overview is:

```
summary(gavote)
   equip       econ         perAA         rural         atlanta
 LEVER:74   middle:69   Min.   :0.000   rural:117   Atlanta   : 15
 OS-CC:44   poor  :72   1st Qu.:0.112   urban: 42   notAtlanta:144
 OS-PC:22   rich  :18   Median :0.233
 PAPER: 2               Mean   :0.243
 PUNCH:17               3rd Qu.:0.348
                        Max.   :0.765
      gore             bush            other          votes           ballots
 Min.   :   249   Min.   :   271   Min.   :   5   Min.   :   832   Min.   :   881
 1st Qu.:  1386   1st Qu.:  1804   1st Qu.:  30   1st Qu.:  3506   1st Qu.:  3694
 Median :  2326   Median :  3597   Median :  86   Median :  6299   Median :  6712
 Mean   :  7020   Mean   :  8929   Mean   : 382   Mean   : 16331   Mean   : 16927
 3rd Qu.:  4430   3rd Qu.:  7468   3rd Qu.: 210   3rd Qu.: 11846   3rd Qu.: 12251
 Max.   :154509   Max.   :140494   Max.   :7920   Max.   :263211   Max.   :280975
```

For the categorical variables, we get a count of the number of each type that occurs. We notice, for example, that only two counties used a paper ballot. This will make it difficult to estimate the effect of this particular voting method on the undercount. For the numerical variables, we have six summary statistics that are sufficient to get a rough idea of the distributions. In particular, we notice that the number of ballots cast ranges over orders of magnitudes. This suggests that I should consider the relative, rather than the absolute, undercount. I create this new relative undercount variable, where we specify the variables using the `dataframe$variable` syntax:

```
gavote$undercount <- (gavote$ballots-gavote$votes)/gavote$ballots
summary(gavote$undercount)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0278  0.0398  0.0438  0.0565  0.1880
```

We see that the undercount ranges from zero up to as much as 19%. The mean across counties is 4.38%. Note that this is not the same thing as the overall relative undercount which is:

```
with(gavote, sum(ballots-votes)/sum(ballots))
[1] 0.03518
```

We have used `with` to save the trouble of prefacing all the subsequent variables with `gavote$`. Graphical summaries are also valuable in gaining an understanding of the data. Considering just one variable at a time, histograms are a well-known way of examining the distribution of a variable:

```
hist(gavote$undercount,main="Undercount",xlab="Percent Undercount")
```

The plot is shown in the left panel of Figure 1.1. A histogram is a fairly crude estimate of the density of the variable that is sensitive to the choice of bins. A kernel density estimate can be viewed as a smoother version of a histogram that is also a superior estimate of the density. We have added a “rug” to our display that makes it possible to discern the individual data points:

```
plot(density(gavote$undercount),main="Undercount")
rug(gavote$undercount)
```

We can see that the distribution is slightly skewed and that there are two outliers in the right tail of the distribution. Such plots are invaluable in detecting mistakes or unusual points in the data. Categorical variables can also be graphically displayed. The pie chart is a popular method. We demonstrate this on the types of voting equipment:

```
pie(table(gavote$equip),col=gray(0:4/4))
```

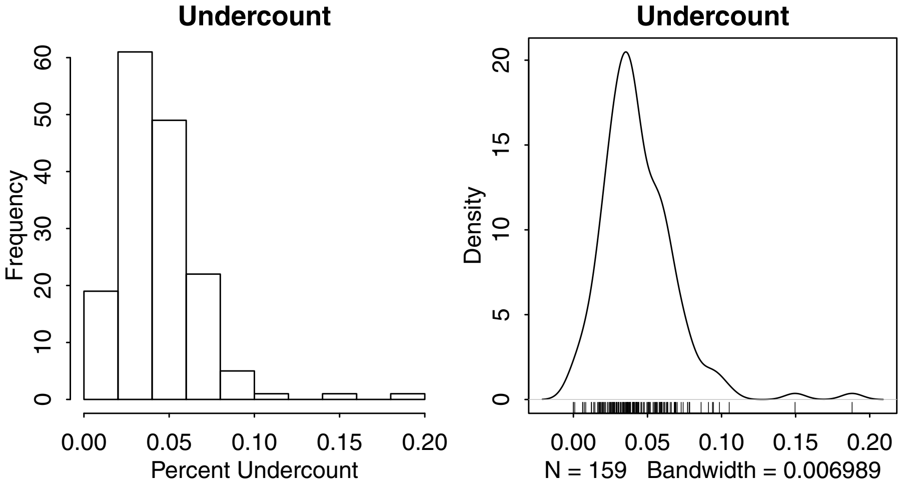


Figure 1.1 *Histogram of the undercount is shown on the left and a density estimate with a data rug is shown on the right.*

The plot is shown in the first panel of Figure 1.2. I have used shades of grey for the slices of the pie because this is a monochrome book. If you omit the `col` argument, you will see a color plot by default. Of course, a color plot is usually preferable, but bear in mind that some photocopying machines and many laser printers are black and white only, so a good greyscale plot is still needed. Alternatively, the Pareto chart is a bar plot with categories in descending order of frequency:

```
barplot(sort(table(gavote$equip),decreasing=TRUE),las=2)
```

The plot is shown in the second panel of Figure 1.2. The `las=2` argument means that the bar labels are printed vertically as opposed to horizontally, ensuring that there is enough room for them to be seen. The Pareto chart (or just a bar plot) is superior to the pie chart because lengths are easier to judge than angles.

Two-dimensional plots are also very helpful. A scatterplot is the obvious way to depict two quantitative variables. Let's see how the proportion voting for Gore relates to the proportion of African Americans:

```
gavote$pergore <- gavote$gore/gavote$votes
plot(pergore ~ perAA, gavote, xlab="Proportion African American", ylab
    ↪ ="Proportion for Gore")
```

The ↪ character just indicates that the command ran over onto a second line. Don't type ↪ in R — just type the whole command on a single line without hitting return until the end. The plot, seen in the first panel of Figure 1.3, shows a strong correlation between these variables. This is an *ecological* correlation because the data points are aggregated across counties. The plot, in and of itself, does not prove that individual African Americans were more likely to vote for Gore, although we know this to be true from other sources. We could also compute the proportion of voters for Bush, but this is, not surprisingly, strongly negatively correlated with the proportion of voters for Gore. We do not need both variables as the one explains the other. We will use the

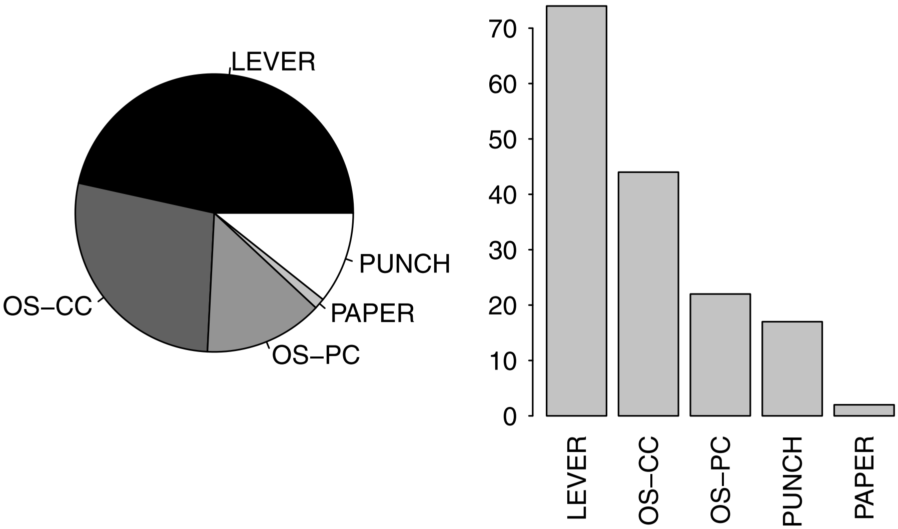


Figure 1.2 *Pie chart of the voting equipment frequencies is shown on the left and a Pareto chart on the right.*

proportion for Gore in the analysis to follow, although one could just as well replace this with the proportion for Bush. I will not consider the proportion for other voters as this has little effect on our conclusions. The reader may wish to verify this.

Side-by-side boxplots are one way of displaying the relationship between qualitative and quantitative variables:

```
plot(undercount ~ equip, gavote, xlab="", las=3)
```

The plot, shown in the second panel of Figure 1.3, shows no major differences in undercount for the different types of equipment. Two outliers are visible for the optical scan-precinct count (OS-PC) method. Plots of two qualitative variables are generally not worthwhile unless both variables have more than three or four levels. The `xtabs` function is useful for cross-tabulations:

```
xtabs(~ atlanta + rural, gavote)
            rural
atlanta       rural urban
  Atlanta         1    14
  notAtlanta    116    28
```

We see that just one county in the Atlanta area is classified as rural. We also notice that variable name `rural` is not sensible because it is the same as the name given to one of the two levels of this factor. It is best to avoid misunderstanding by using unambiguous labeling:

```
names(gavote)
 [1] "equip"      "econ"       "perAA"      "rural"      "atlanta"    "gore"
...
names(gavote)[4] <- "usage"
```

Correlations are the standard way of numerically summarizing the relationship between quantitative variables. However, not all the variables in our dataframe are

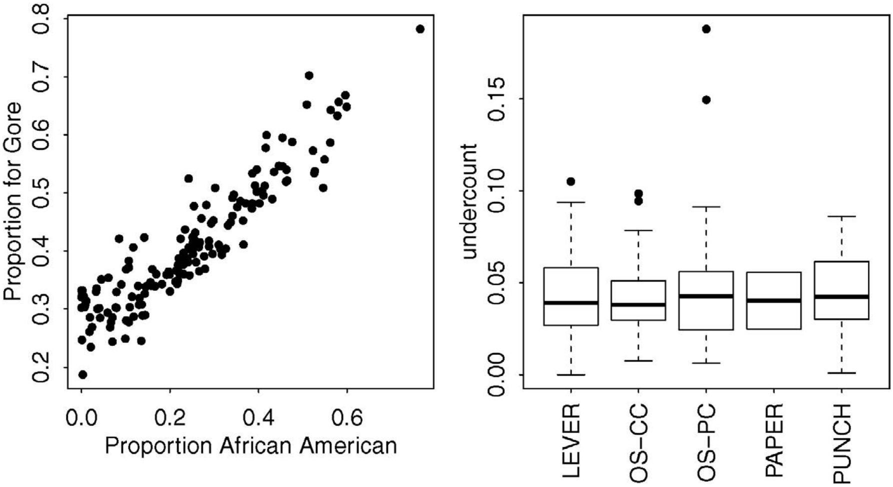


Figure 1.3 *A scatterplot plot of proportions of Gore voters and African Americans by county is shown on the left. Boxplots showing the distribution of the undercount by voting equipment are shown on the right.*

quantitative or immediately of interest. First we construct a vector using `c()` of length three which contains the indices of the variables of interest. We select these columns from the dataframe and compute the correlation. The syntax for selecting rows and/or columns is `dataframe[rows,columns]` where rows and/or columns are vectors of indices. In this case, we want all the rows, so I omit that part of the construction:

```
nix <- c(3,10,11,12)
cor(gavote[,nix])
```

```
              perAA   ballots undercount  pergore
perAA      1.000000  0.027732    0.22969 0.921652
ballots    0.027732  1.000000   -0.15517 0.095617
undercount 0.229687 -0.155172    1.00000 0.218765
pergore    0.921652  0.095617    0.21877 1.000000
```

We see some mild correlation between some of the variables except for the Gore — African Americans correlation which we know is large from the previous plot.

Defining a Linear Model: We describe this data with a linear model which takes the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon$$

where β_i, $i = 0, 1, 2, \ldots, p-1$ are unknown *parameters*. β_0 is called the *intercept* term. The *response* is Y and the *predictors* are $X_1, \ldots, X_{p-1}$. The predictors may be the original variables in the dataset or transformations or combinations of them. The error ε represents the difference between what is explained by the systematic part of the model and what is observed. ε may include measurement error although it is often due to the effect of unincluded or unmeasured variables.

The regression equation is more conveniently written as:

$$y = X\beta + \varepsilon$$

where, in terms of the n data points, $y = (y_1, \dots, y_n)^T$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$, $\beta = (\beta_0, \dots, \beta_{p-1})^T$ and:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \dots & x_{2,p-1} \\ \vdots & & & \vdots & \\ 1 & x_{n1} & x_{n2} & \dots & x_{n,p-1} \end{pmatrix}$$

The column of ones incorporates the intercept term. The *least squares* estimate of β, called $\hat{\beta}$, minimizes:

$$\sum \varepsilon_i^2 = \varepsilon^T \varepsilon = (y - X\beta)^T (y - X\beta)$$

Differentiating with respect to β and setting to zero, we find that $\hat{\beta}$ satisfies:

$$X^T X \hat{\beta} = X^T y$$

These are called the *normal equations*.

Fitting a Linear Model: Linear models in R are fit using the `lm` command. For example, suppose we model the undercount as the response and the proportions of Gore voters and African Americans as predictors:

```
lmod <- lm(undercount ~ pergore + perAA, gavote)
```

This corresponds to the linear model formula:

$$\texttt{undercount} = \beta_0 + \beta_1 \texttt{pergore} + \beta_2 \texttt{perAA} + \varepsilon$$

R uses the *Wilkinson–Rogers* notation of Wilkinson and Rogers (1973). For a straightforward linear model, such as this example, we see that it corresponds to just dropping the parameters from the mathematical form. The intercept is included by default.

We can obtain the least squares estimates of β, called the regression coefficients, $\hat{\beta}$, by:

```
coef(lmod)
(Intercept)     pergore       perAA
   0.032376    0.010979    0.028533
```

The construction of the least squares estimates does not require any assumptions about ε. If we are prepared to assume that the errors are at least independent and have equal variance, then the *Gauss–Markov* theorem tells us that the least squares estimates are the best linear unbiased estimates. Although it is not necessary, we might further assume that the errors are normally distributed, and we might compute the maximum likelihood estimate (MLE) of β (see Appendix A for more MLEs). For the linear models, these MLEs are identical with the least squares estimates. However, we shall find that, in some of the extension of linear models considered later in this book, an equivalent notion to least squares is not suitable and likelihood methods must be used. This issue does not arise with the standard linear model.

The predicted or fitted values are $\hat{y} = X\hat{\beta}$, while the residuals are $\hat{\varepsilon} = y - X\hat{\beta} = y - \hat{y}$. We can compute these as:

```
predict(lmod)
 APPLING ATKINSON    BACON    BAKER  BALDWIN    BANKS
0.041337 0.043291 0.039618 0.052412 0.047955 0.036016
...
residuals(lmod)
   APPLING   ATKINSON      BACON      BAKER    BALDWIN
 0.0369466 -0.0069949  0.0655506  0.0023484  0.0035899
...
```

where the ellipsis indicates that (much of) the output has been omitted.

It is useful to have some notion of how well the model fits the data. The residual sum of squares (RSS) is $\hat{\varepsilon}^T\hat{\varepsilon}$. This can be computed as:

```
deviance(lmod)
[1] 0.09325
```

The term *deviance* is a more general measure of fit than RSS, which we will meet again in chapters to follow. For linear models, the deviance is the RSS.

The *degrees of freedom* for a linear model is the number of cases minus the number of coefficients or:

```
df.residual(lmod)
[1] 156
nrow(gavote)-length(coef(lmod))
[1] 156
```

Let the variance of the error be σ^2, then σ is estimated by the residual standard error computed from $\sqrt{(}\mathrm{RSS}/\mathrm{df})$. For our example, this is:

```
sqrt(deviance(lmod)/df.residual(lmod))
[1] 0.024449
```

Although several useful regression quantities are stored in the `lm` model object (which we called `lmod` in this instance), we can compute several more using the `summary` command on the model object. For example:

```
lmodsum <- summary(lmod)
lmodsum$sigma
[1] 0.024449
```

R is an object-oriented language. One important feature of such a language is that *generic* functions, such as `summary`, recognize the type of object being passed to it and behave appropriately. We used `summary` for dataframes previously and now for linear models. `residuals` is another generic function and we shall see how it can be applied to many model types and return appropriately defined residuals.

The deviance measures how well the model fits in an absolute sense, but it does not tell us how well the model fits in a relative sense. The popular choice is R^2, called the *coefficient of determination* or *percentage of variance explained*:

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

where $\texttt{TSS} = \sum(y_i - \bar{y})^2$ and stands for total sum of squares. This can be most conveniently extracted as:

```
lmodsum$r.squared
[1] 0.053089
```

We see that R^2 is only about 5% which indicates that this particular model does not fit so well. An appreciation of what constitutes a good value of R^2 varies according to the application. Another way to think of R^2 is the (squared) correlation between the predicted values and the response:

```
cor(predict(lmod),gavote$undercount)^2
[1] 0.053089
```

R^2 cannot be used as a criterion for choosing models among those available because it can never decrease when you add a new predictor to the model. This means that it will favor the largest models. The adjusted R^2 makes allowance for the fact that a larger model also uses more parameters. It is defined as:

$$R_a^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}$$

Adding a predictor will only increase R_a^2 if it has some predictive value. Furthermore, minimizing $\hat{\sigma}^2$ means maximizing R_a^2 over a set of possible linear models. The value can be extracted as:

```
lmodsum$adj.r.squared
[1] 0.040949
```

One advantage of R over many statistical packages is that we can extract all these quantities individually for subsequent calculations in a convenient way. However, if we simply want to see the regression output printed in a readable way, we use the `summary`:

```
summary(lmod)
Residuals:
     Min       1Q   Median       3Q      Max
-0.04601 -0.01500 -0.00354  0.01178  0.14244

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0324     0.0128    2.54    0.012
pergore       0.0110     0.0469    0.23    0.815
perAA         0.0285     0.0307    0.93    0.355

Residual standard error: 0.0244 on 156 degrees of freedom
Multiple R-Squared: 0.0531,       Adjusted R-squared: 0.0409
F-statistic: 4.37 on 2 and 156 DF,  p-value: 0.0142
```

We have already separately computed many of the quantities given above. This summary is too verbose to my taste and I prefer a shorter `sumary` found in my R package:

```
library(faraway)
sumary(lmod)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0324     0.0128    2.54    0.012
pergore       0.0110     0.0469    0.23    0.815
perAA         0.0285     0.0307    0.93    0.355

n = 159, p = 3, Residual SE = 0.024, R-Squared = 0.05
```

You only need to load the package with `library(faraway)` once per session so this line may be skipped if you did it earlier. If you get an error message about a function not being found, it probably means you forgot to load the package that contains that function.

Qualitative Predictors: The addition of qualitative variables requires the introduction of dummy variables. Two-level variables are easy to code; consider the rural/urban indicator variable. We can code this using a dummy variable d:

$$d = \begin{cases} 0 & \text{rural} \\ 1 & \text{urban} \end{cases}$$

This is the default coding used in R. Zero is assigned to the level which is first alphabetically, unless something is done to change this (perhaps using the `relevel` command). If we add this variable to our model, it would now be:

$$\texttt{undercount} = \beta_0 + \beta_1\texttt{pergore} + \beta_2\texttt{perAA} + \beta_3\texttt{d} + \varepsilon$$

So β_3 would now represent the difference between the undercount in an urban county and a rural county. Codings other than 0-1 could be used although the interpretation of the associated parameter would not be quite as straightforward.

A more extensive use of dummy variables is needed for factors with $k > 2$ levels. We define $k-1$ dummy variables d_j for $j = 2, \ldots, k$ such that:

$$d_j = \begin{cases} 0 & \text{is not level j} \\ 1 & \text{is level j} \end{cases}$$

Interactions between variables can be added to the model by taking the columns of the model matrix X that correspond to the two variables and multiplying them together entrywise for all terms that make up the interaction.

Interpretation: Let's add some qualitative variables to the model to see how the terms can be interpreted. We have centered the `pergore` and `perAA` terms by their means for reasons that will become clear:

```
gavote$cpergore <- gavote$pergore - mean(gavote$pergore)
gavote$cperAA <- gavote$perAA - mean(gavote$perAA)
lmodi <- lm(undercount ~ cperAA+cpergore*usage+equip, gavote)
sumary(lmodi)
```

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.04330    0.00284   15.25  < 2e-16
cperAA               0.02826    0.03109    0.91   0.3648
cpergore             0.00824    0.05116    0.16   0.8723
usageurban          -0.01864    0.00465   -4.01 0.000096
equipOS-CC           0.00648    0.00468    1.39   0.1681
equipOS-PC           0.01564    0.00583    2.68   0.0081
equipPAPER          -0.00909    0.01693   -0.54   0.5920
equipPUNCH           0.01415    0.00678    2.09   0.0387
cpergore:usageurban -0.00880    0.03872   -0.23   0.8205

n = 159, p = 9, Residual SE = 0.023, R-Squared = 0.17
```

Here is the model witten in a mathematical form:

$$\texttt{undercount} = \beta_0 + \beta_1\texttt{cperAA} + \beta_2\texttt{cpergore} + \beta_2\texttt{usageurban} + \beta_4\texttt{equipOSCC} + \\ \beta_5\texttt{equipOSPC} + \beta_6\texttt{equipPAPER} + \beta_7\texttt{equipPUNCH} + \beta_8\texttt{cpergore} : \texttt{usageurban} + \varepsilon \tag{1.1}$$

The terms usageurban, equipOSCC, equipOSPC, equipPAPER and equipPUNCH are all dummy variables taking the value 1 when the county is urban or using that voting method, respectively. They take the value 0 otherwise. The term cpergore:usageurban is formed taking the product of the dummy variable for usageurban and the quantitative variable cpergore. Hence it is zero for rural counties and takes the value of cpergore for urban counties.

Consider a rural county that has an average proportion of Gore voters and an average proportion of African Americans where lever machines are used for voting. Because rural and lever are the reference levels for the two qualitative variables, there is no contribution to the predicted undercount from these terms. Furthermore, because we have centered the two quantitative variables at their mean values, these terms also do not enter into the prediction. Notice the worth of the centering because otherwise we would need to set these variables to zero to get them to drop out of the prediction equation; zero is not a typical value for these predictors. Given that all the other terms are dropped, the predicted undercount is just given by the intercept $\hat{\beta}_0$, which is 4.33%.

The interpretation of the coefficients can now be made relative to this baseline. We see that, with all other predictors unchanged, except using optical scan with precinct count (OS-PC), the predicted undercount increases by 1.56%. The other equipment methods can be similarly interpreted. Notice that we need to be cautious about the interpretation. Given two counties with the same values of the predictors, except having different voting equipment, we would *predict* the undercount to be 1.56% higher for the OS-PC county compared to the lever county. However, we cannot go so far as to say that if we went to a county with lever equipment and changed it to OS-PC that this would *cause* the undercount to increase by 1.56%.

With all other predictors held constant, we would predict the undercount to increase by 2.83% going from a county with no African Americans to all African American. Sometimes a one-unit change in a predictor is too large or too small, prompting a rescaling of the interpretation. For example, we might predict a 0.283% increase in the undercount for a 10% increase in the proportion of African Americans. Of course, this interpretation should not be taken too literally. We already know that the proportion of African Americans and Gore voters is strongly correlated so that an increase in the proportion of one would lead to an increase in the proportion of the other. This is the problem of *collinearity* that makes the interpretation of regression coefficients much more difficult. Furthermore, the proportion of African Americans is likely to be associated with other socioeconomic variables which might also be related to the undercount. This further hinders the possibility of a causal conclusion.

The interpretation of the usage and pergore cannot be done separately as there is an interaction term between these two variables. For an average number of Gore voters, we would predict a 1.86%-lower undercount in an urban county compared to a rural county. In a rural county, we predict a 0.08% increase in the undercount as the proportion of Gore voters increases by 10%. In an urban county, we predict a $(0.00824 - 0.00880) * 10 = -0.0056\%$ increase in the undercount as the proportion of Gore voters increases by 10%. Since the increase is by a negative amount, this is actually a decrease. This illustrates the potential pitfalls in interpreting the effect of a predictor in the presence of an interaction. We cannot give a simple stand-alone interpretation of the effect of the proportion of Gore voters. The effect is to increase the undercount in rural counties and to decrease it, if only very slightly, in urban counties.

Hypothesis Testing: We often wish to determine the significance of one, some or all of the predictors in a model. If we assume that the errors are independent and identically normally distributed, there is a very general testing procedure that may be used. Suppose we compare two models, a larger model Ω and a smaller model ω contained within that can be represented as a linear restriction on the parameters of the larger model. Most often, the predictors in ω are just a subset of the predictors in Ω.

Now suppose that the dimension (or number of parameters) of Ω is p and the dimension of ω is q. Then, assuming that the smaller model ω is correct, the F-statistic is:

$$F = \frac{(\mathrm{RSS}_\omega - \mathrm{RSS}_\Omega)/(p-q)}{\mathrm{RSS}_\Omega/(n-p)} \sim F_{p-q,n-p}$$

Thus we would reject the null hypothesis that the smaller model is correct if $F > F^{(\alpha)}_{p-q,n-p}$.

For example, we might compare the two linear models considered previously. The smaller model has just `pergore` and `perAA` while the larger model adds `usage` and `equip` along with an interaction. We compute the F-test as:

```
anova(lmod,lmodi)
Analysis of Variance Table

Model 1: undercount ~ pergore + perAA
Model 2: undercount ~ cperAA + cpergore * usage + equip
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1    156 0.0932
2    150 0.0818  6    0.0115 3.51 0.0028
```

It does not matter that the variables have been centered in the larger model but not in the smaller model, because the centering makes no difference to the RSS. The p-value here is small indicating the null hypothesis of preferring the smaller model should be rejected.

One common need is to test specific predictors in the model. It is possible to use the general F-testing method: fit a model with the predictor and without the predictor and compute the F-statistic. It is important to know what other predictors are also included in the models and the results may differ if these are also changed. An alternative computation is to use a t-statistic for testing the hypothesis:

$$t_i = \hat{\beta}_i / se(\hat{\beta}_i)$$

and check for significance using a t-distribution with $n-p$ degrees of freedom. This approach will produce exactly the same p-value as the F-testing method. For example, in the larger model above, the test for the significance of the proportion of African Americans gives a p-value of 0.3648. This indicates that this predictor is not statistically significant after adjusting for the effect of the other predictors on the response.

We would usually avoid using the t-tests for the levels of qualitative predictors with more than two levels. For example, if we were interested in testing the effects of the various voting equipment, we would need to fit a model without this predictor and compute the corresponding F-test. A comparison of all models with one predictor less than the larger model may be obtained conveniently as:

```
drop1(lmodi,test="F")
Single term deletions

Model:
undercount ~ cperAA + cpergore * usage + equip
               Df Sum of Sq    RSS   AIC F value Pr(>F)
<none>                      0.0818 -1186
cperAA          1   0.00045 0.0822 -1187    0.83  0.365
equip           4   0.00544 0.0872 -1184    2.50  0.045
cpergore:usage  1   0.00003 0.0818 -1188    0.05  0.821
```

We see that the equipment is barely statistically significant in that the p-value is just less than the traditional 5% significance level. You will also notice that only the interaction term `cpergore:usage` is considered and not the corresponding main effects terms, `cpergore` and `usage`. This demonstrates respect for the *hierarchy principle* which demands that all lower-order terms corresponding to an interaction be retained in the model. In this case, we see that the interaction is not significant, but a further step would now be necessary to test the main effects.

There are numerous difficulties with interpreting the results of hypothesis tests and the reader is advised to avoid taking the results too literally before understanding these problems.

Confidence Intervals: These may be constructed for β using:

$$\hat{\beta}_i \pm t_{n-p}^{(\alpha/2)} se(\hat{\beta}_i)$$

where $t_{n-p}^{(\alpha/2)}$ is the upper $\alpha/2^{th}$ quantile of a t distribution with $n-p$ degrees of freedom. A convenient way of computing the 95% confidence intervals in R is:

```
confint(lmodi)
                          2.5 %     97.5 %
(Intercept)          0.03768844  0.0489062
cperAA              -0.03317106  0.0896992
cpergore            -0.09284293  0.1093166
usageurban          -0.02782090 -0.0094523
equipOS-CC          -0.00276464  0.0157296
equipOS-PC           0.00412523  0.0271540
equipPAPER          -0.04253684  0.0243528
equipPUNCH           0.00074772  0.0275515
cpergore:usageurban -0.08529909  0.0677002
```

Confidence intervals have a duality with the corresponding t-tests in that if the p-value is greater than 5%, zero will fall in the interval and vice versa. Confidence intervals give a range of plausible values for the parameter and are more useful for judging the size of the effect of the predictor than a p-value that merely indicates statistical significance, not necessarily practical significance. These intervals are individually correct, but there is not a 95% chance that the true parameter values fall in all the intervals. This problem of *multiple comparisons* is particularly acute for the voting equipment, where five levels leads to 10 possible pairwise comparisons, more than just the four shown here.

Diagnostics: The validity of the inference depends on the assumptions concerning the linear model. One part of these assumptions is that the systematic form of the model $EY = X\beta$ is correct; we assume we have included all the right variables and transformed and combined them correctly. Another set of assumptions concerns the random part of the model: ε. We require that the errors have equal variance, be uncorrelated and have a normal distribution. We are also interested in detecting points, called *outliers*, that are unusual in that they do not fit the model that seems otherwise adequate for the rest of the data. Ideally, we would like each case to have an equal contribution to the fitted model; yet sometimes a few points have a much larger effect than others. Such points are called *influential.*

Diagnostic methods can be graphical or numerical. We prefer graphical methods because they tend to be more versatile and informative. It is virtually impossible to verify that a given model is exactly correct. The purpose of the diagnostics is more to check whether the model is not grossly wrong. Indeed, a successful data analyst should pay more attention to avoiding big mistakes than optimizing the fit.

A collection of four useful diagnostics can be simply obtained with:

```
plot(lmodi)
```

as can be seen in Figure 1.4. The plot in the upper-left panel shows the residuals plotted against the fitted values. The plot can be used to detect lack of fit. If the residuals show some curvilinear trend, this is a sign that some change to the model is required, often a transformation of one of the variables. A smoothed curve has been added to the plot to aid in this assessment. In this instance, there is no sign of such a problem. The plot is also used to check the constant variance assumption on the errors. In this case, it seems the variance is roughly constant as the fitted values vary. Assuming symmetry of the errors, we can effectively double the resolution by plotting the absolute value of the residuals against the fitted values. As it happens $|\hat{\varepsilon}|$ tends to be rather skewed and it is better to use $\sqrt{\hat{\varepsilon}}$. Such a plot is shown in the lower-left panel, confirming what we have already observed about the constancy of the variance. Notice that a few larger residuals have been labeled.

The residuals can be assessed for normality using a *QQ plot*. This compares the residuals to "ideal" normal observations. We plot the sorted residuals against $\Phi^{-1}(\frac{i}{n+1})$ for $i = 1, \ldots, n$. This can be seen in the upper-right panel of Figure 1.4. In this plot, the points follow a linear trend (except for one or two cases), indicating that normality is a reasonable assumption. If we observe a curve, this indicates skewness, suggesting a possible transformation of the response, while two tails of points diverging from linearity would indicate a long-tailed error, suggesting that we should consider robust fitting methods. Particularly for larger datasets, the normality assumption is not crucial, as the inference will be approximately correct in spite of the nonnormality. Only a clear deviation from normality should necessarily spur some action to change the model.

The fitted values can be written as $X\hat{\beta} = X(X^TX)^{-1}X^Ty = Hy$ where the *hat-matrix* $H = X(X^TX)^{-1}X^T$. $h_i = H_{ii}$ are called *leverages* and are useful diagnostics.

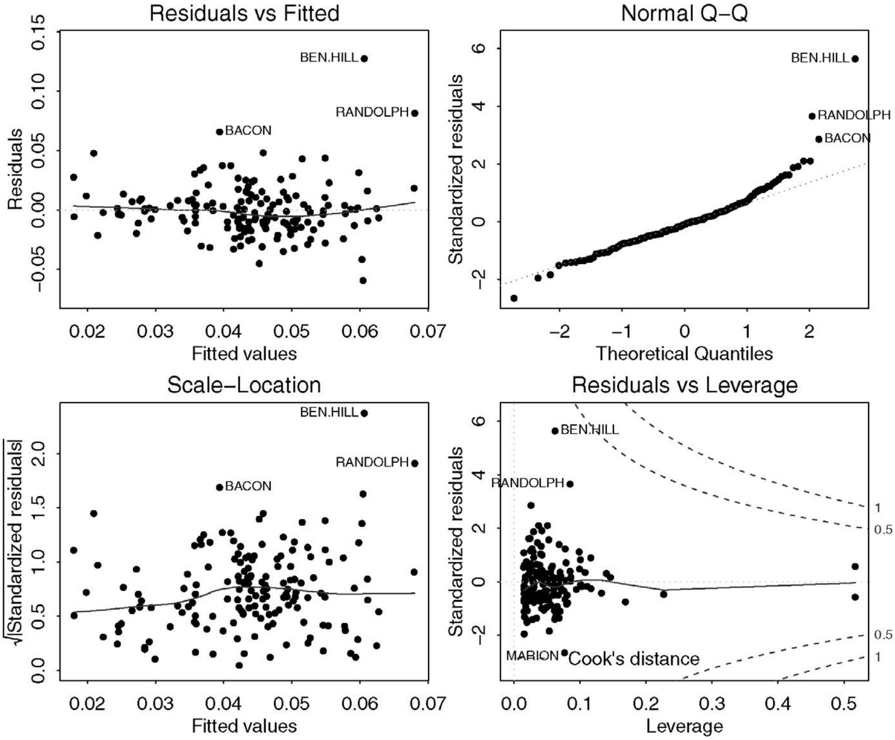


Figure 1.4 *Diagnostics obtained from plotting the model object.*

For example, since var $\hat{\varepsilon}_i = \sigma^2(1-h_i)$, a large leverage, h_i, will tend to make var $\hat{\varepsilon}_i$ small. The fit will be "forced" close to y_i. It is useful to examine the leverages to determine which cases have the power to be influential. Points on the boundary of the predictor space will have the most leverage.

The Cook statistics are a popular influence diagnostic because they reduce the information to a single value for each case. They are defined as:

$$D_i = \frac{(\hat{y}-\hat{y}_{(i)})^T(\hat{y}-\hat{y}_{(i)})}{p\hat{\sigma}^2} = \frac{1}{p}r_i^2\frac{h_i}{1-h_i}$$

where r_i are the standardized residuals. They represent a scaled measure of the change in the fit if the single case is dropped from the dataset. Information about the leverages and Cook statistics for the current model is given in the lower-right panel of Figure 1.4. A large residual combined with a large leverage will result in a larger Cook statistic. The plot shows two contour lines for the Cook statistics as these are a function of the standardized residuals and leverages.

We can see that there are a couple of cases that stick out and we should investigate more closely the influence of these points. We can pick out the top two influential cases with:

```
gavote[cooks.distance(lmodi) > 0.1,]
```

```
         equip econ perAA usage    atlanta gore bush other votes
BEN.HILL OS-PC poor 0.282 rural notAtlanta 2234 2381    46  4661
RANDOLPH OS-PC poor 0.527 rural notAtlanta 1381 1174    14  2569
         ballots undercount pergore cpergore   cperAA
BEN.HILL    5741    0.18812 0.47930 0.070975 0.039019
RANDOLPH    3021    0.14962 0.53756 0.129241 0.284019
```

Notice how we can select a subset of a dataframe using a logical expression. Here we ask for all rows in the dataframe that have Cook statistics larger than 0.1. We see that these are the same two counties that stuck out in the boxplots seen in Figure 1.3. These points are influential because they have much higher undercounts than would be expected. Their leverages are not high so they do not have unusual predictor values. The standardized residual for Ben Hill is over 5. Roughly speaking, standardized residuals exceeding 3.5 deserve closer attention so this case would attract some attention.

A useful technique for judging whether some cases in a set of positive observations are unusually extreme is the half-normal plot. Here we plot the sorted values against $\Phi^{-1}(\frac{n+i}{2n+1})$ which represent the quantiles of the upper half of a standard normal distribution. We are usually not looking for a straight line relationship since we do not necessarily expect a positive normal distribution for quantities like the leverages. We are looking for outliers, which will be apparent as points that diverge substantially from the rest of the data. Here is the half-normal plot of the leverages:

```
library(faraway)
halfnorm(hatvalues(lmodi))
```

The `halfnorm` function is part of the `faraway` package so we need to load that to access the function (if you have not already done so earlier in this session). The plot, seen in the left panel of Figure 1.5, shows two points with much higher leverage than the rest. These points are:

```
gavote[hatvalues(lmodi)>0.3,]
```

```
           equip econ perAA usage    atlanta gore bush other
MONTGOMERY PAPER poor 0.243 rural notAtlanta 1013 1465    31
TALIAFERRO PAPER poor 0.596 rural notAtlanta  556  271     5
           votes ballots undercount pergore   cpergore
MONTGOMERY  2509    2573   0.024874 0.40375 -0.0045753
TALIAFERRO   832     881   0.055619 0.66827  0.2599475
```

These are the only two counties that use a paper ballot, so they will be the only cases that determine the coefficient for paper. This is sufficient to give them high leverage as the remaining predictor values are all unremarkable. Note that these counties were not identified as influential — having high leverage alone is not necessarily enough to be influential.

Partial residual plots display $\hat{\varepsilon} + \hat{\beta}_i x_i$ against x_i. To see the motivation, look at the response with the predicted effect of the other X removed:

$$y - \sum_{j \neq i} x_j \hat{\beta}_j = \hat{y} + \hat{\varepsilon} - \sum_{j \neq i} x_j \hat{\beta}_j = x_i \hat{\beta}_i + \hat{\varepsilon}$$

The partial residual plot for `cperAA` is shown in the right panel of Figure 1.5:

```
termplot(lmodi,partial=TRUE,terms=1)
```

The line is the least squares fit to the data on this plot as well as having the same

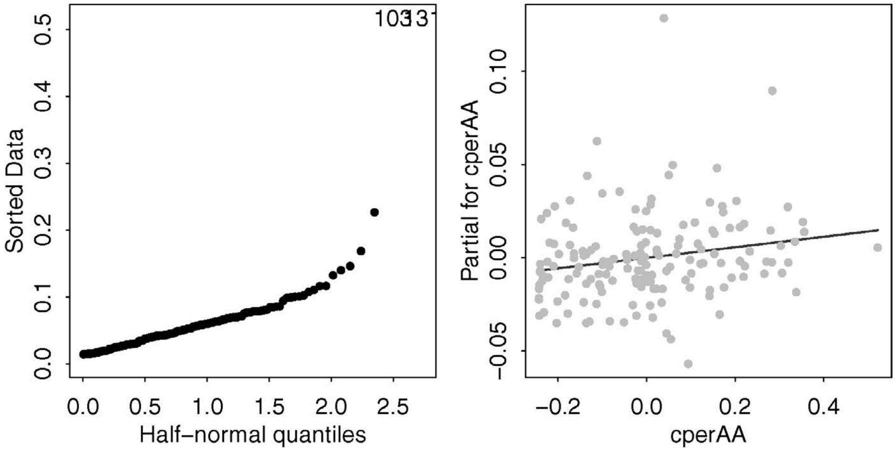


Figure 1.5 *Half-normal plot of the leverages is shown on the left and a partial residual plot for the proportion of African Americans is shown on the right.*

slope as the `cperAA` term in the current model. This plot gives us a snapshot of the marginal relationship between this predictor and the response. In this case, we see a linear relationship indicating that it is not worthwhile seeking transformations. Furthermore, there is no sign that a few points are having undue influence on the relationship.

Robust Regression: Least squares works well when there are normal errors, but performs poorly for long-tailed errors. We have identified a few potential outliers in the current model. One approach is to simply eliminate the outliers from the dataset and then proceed with least squares. This approach is satisfactory when we are convinced that the outliers represent truly incorrect observations, but even then, detecting such cases is not always easy as multiple outliers can mask each other. However, in other cases, outliers are real observations. Sometimes, removing these cases simply creates other outliers. A generally better approach is to use a robust alternative to least squares that downweights the effect of larger errors. The Huber method is the default choice of the `rlm` function and is found in the MASS package of Venables and Ripley (2002):

```
library(MASS)
rlmodi <- rlm(undercount ~ cperAA+cpergore*usage+equip, gavote)
summary(rlmodi)
```

```
Coefficients:
                    Value  Std. Error t value
(Intercept)          0.041  0.002     17.866
cperAA               0.033  0.025      1.290
cpergore            -0.008  0.042     -0.197
usageurban          -0.017  0.004     -4.406
equipOS-CC           0.007  0.004      1.802
equipOS-PC           0.008  0.005      1.695
equipPAPER          -0.006  0.014     -0.427
equipPUNCH           0.017  0.006      3.072
```

```
cpergore:usageurban  0.007  0.032      0.230

Residual standard error: 0.0172 on 150 degrees of freedom
```

Inferential methods are more difficult to apply when robust estimation methods are used, hence there is less in this output than for the corresponding `lm` output previously. The most interesting change is that the coefficient for OS-PC is now about half the size. Recall that, using the treatment coding, this represents the difference between OS-PC and the reference lever method. There is some fluctuation in the other coefficients, but not enough to change our impression of the important effects. The robust fit here has reduced the effect of the two outlying counties.

Weighted Least Squares: The sizes of the counties in this dataset vary greatly with the number of ballots cast in each county ranging from 881 to 280,975. We might expect the proportion of undercounted votes to be more variable in smaller counties than larger ones. Since the responses from the larger counties might be more precise, perhaps they should count for more in the fitting of the model. This effect can be achieved by the use of weighted least squares where we attempt to minimize $\sum w_i \varepsilon_i^2$. The appropriate choice for the weights w_i is to set them to be inversely proportional to var y_i.

Now var y for a binomial proportion is inversely proportional to the group size, in this case, the number of ballots. This suggests setting the weights proportional to the number of ballots:

```
wlmodi <- lm(undercount ~ cperAA+cpergore*usage+equip, gavote, weights
    ↪ =ballots)
```

This results in a fit that is substantially different from the unweighted fit. It is dominated by the data from a few large counties.

However, the variation in the response is likely to be caused by more than just binomial variation due to the number of ballots. There are likely to be other variables that affect the response in a way that is not proportional to ballot size. Consider the relative size of these effects. Even for the smallest county, assuming an average undercount rate, the standard deviation using the binomial is:

```
sqrt(0.035*(1-0.035)/881)
[1] 0.0061917
```

which is much smaller than the residual standard error of 0.0233. The effects will be substantially smaller for other counties. So since the other sources of variation dominate, we recommend leaving this particular model unweighted.

Transformation: Models can sometimes be improved by transforming the variables. Ideas for transformations can come from several sources. One method is to search through a family of possible transformations looking for the best fit. An example of this approach is the Box–Cox method of selecting a transformation on the response variable. Alternatively, the diagnostic plots for the current model can suggest transformations that might improve the fit or ameliorate apparent violations of the assumptions. In other situations, transformations may be motivated by theories concerning the relationship between the variables or to aid the interpretation of the model.

For this dataset, transformation of the response is problematic for both technical and interpretational reasons. The minimum undercount is exactly zero which precludes directly applying some popular transformations such as the log or inverse. An arbitrary fix for this problem is to add a small amount (say 0.005 here) to the response which would enable the use of all power transformations. The application of the Box–Cox method, using the `boxcox` function from the `MASS` package, suggests a square root transformation of the response. However, it is difficult to give an interpretation to the regression coefficients with this transformation on the response. Other than no transformation at all, a logged response does allow a simple interpretation. For an untransformed response, the coefficients represent addition to the undercount whereas for a logged response, the coefficients can be interpreted as multiplying the response. So we see that, although transformations of the response might sometimes improve the fit, they can lead to difficulties with interpretation and so should be applied with care. Another point to consider is that if the untransformed response was normally distributed, it will not be so after transformation. This suggests considering nonnormal, continuous responses as seen in Section 9.1, for example.

Transformations of the predictors are less problematic. Let's first consider the proportion of African Americans predictor in the current model. Polynomials provide a commonly used family of transformations. The use of orthogonal polynomials is recommended as these are more numerically stable and make it easier to select the correct degree:

```
plmodi <- lm(undercount ~ poly(cperAA,4)+cpergore*usage+equip, gavote)
summary(plmodi)
```

```
 Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.04346    0.00288   15.12  < 2e-16
poly(cperAA, 4)1     0.05226    0.06939    0.75   0.4526
poly(cperAA, 4)2    -0.00299    0.02613   -0.11   0.9091
poly(cperAA, 4)3    -0.00536    0.02427   -0.22   0.8254
poly(cperAA, 4)4    -0.01651    0.02420   -0.68   0.4961
cpergore             0.01315    0.05693    0.23   0.8176
usageurban          -0.01913    0.00474   -4.03 0.000088
equipOS-CC           0.00644    0.00472    1.36   0.1746
equipOS-PC           0.01559    0.00588    2.65   0.0089
equipPAPER          -0.01027    0.01720   -0.60   0.5514
equipPUNCH           0.01405    0.00687    2.05   0.0425
cpergore:usageurban -0.01054    0.04136   -0.25   0.7993

Residual standard error: 0.0235 on 147 degrees of freedom
Multiple R-Squared: 0.173,       Adjusted R-squared: 0.111
F-statistic: 2.79 on 11 and 147 DF,  p-value: 0.00254
```

The hierarchy principle requires that we avoid eliminating lower-order terms of a variable when high-order terms are still in the model. From the output, we see that the fourth-order term is not significant and can be eliminated. With standard polynomials, the elimination of one term would cause a change in the values of the remaining coefficients. The advantage of the orthogonal polynomials is that the coefficients for the lower-order terms do not change as we change the maximum degree of the model. Here we see that all the terms of `cperAA` are not significant and all can be removed. Some insight into the relationship may be gained by plotting the fit on top of the partial residuals:

```
termplot(plmodi,partial=TRUE,terms=1)
```

The plot, seen in the first panel of Figure 1.6, shows that the quartic polynomial is not so different from a constant fit, explaining the lack of significance.

Polynomial fits become less attractive with higher-order terms. The fit is not local in the sense that a point in one part of the range of the variable affects the fit across the whole range. Furthermore, polynomials tend to have rather oscillatory fits and extrapolate poorly. A more stable fit can be had using splines, which are piecewise polynomials. Various types of splines are available and they typically have the local fit and stable extrapolation properties. We demonstrate the use of cubic B-splines here:

```
library(splines)
blmodi <- lm(undercount ~ cperAA+bs(cpergore,4)+usage+equip, gavote)
```

Because the spline fit for `cperAA` was very similar to orthogonal polynomials, we consider `cpergore` here for some variety. Notice that we have eliminated the interaction with `usage` for simplicity. The complexity of the B-spline fit may be controlled by specifying the degrees of freedom. We have used four here. The nature of the fit can be seen in the second panel of Figure 1.6:

```
termplot(blmodi,partial=TRUE,terms=2)
```

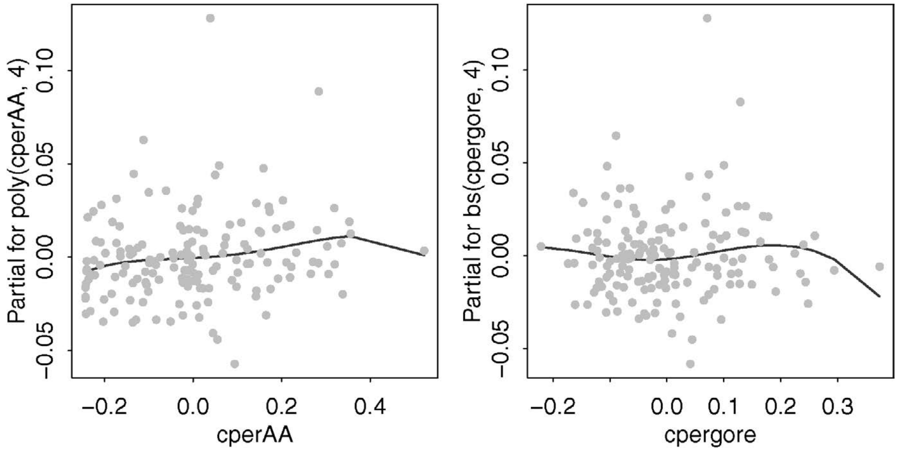


Figure 1.6 *Partial fits using orthogonal polynomials for* `cperAA` *(shown on the left) and cubic B-splines for* `cpergore` *(shown on the right).*

We see that the curved fit is not much different from a constant. More details about splines can be found in Section 14.2.

Variable Selection: One theoretical view of the problem of variable selection is that one subset of the available variables represents the correct model for the data and that any method should be judged by its success in identifying this correct model. While this may be a tempting world in which to test competing variable selection methods, it seems unlikely to match with reality. Even if we believe that a correct model even exists, it is more than likely that we will not have recorded all the relevant variables or not have chosen the correct transformations or functional form for the model amongst the set we choose to consider. We might then retreat from this ideal view and hope to identify the best model from the available set. Even then, we would need to define what is meant by best.

Linear modeling serves two broad goals. Some build linear models for the purposes of prediction — they expect to observe new X and wish to predict y, along with measures of uncertainty in the prediction. Prediction performance is improved by removing variables that contribute little or nothing to the model. We can define a criterion for prediction performance and search for the model that optimizes that criterion. One such criterion is the adjusted R^2 previously mentioned. The `regsubsets` function in the `leaps` package implements this search. For problems involving a moderate number of variables, it is possible to exhaustively search all possible models for the best. As the number of variables increases, exhaustive search becomes prohibitive and various stepwise methods must be used to search the model space. The implementation also has the disadvantage that it can only be applied to quantitative predictors.

Another popular criterion is the Akaike Information Criterion or AIC defined as:

$$\text{AIC} = -2 \text{ maximum log likelihood} + 2p$$

where p is the number of parameters. This criterion has the advantage of generality and can be applied far beyond normal linear models. The `step` command implements a stepwise search strategy through the space of possible models. It does allow qualitative variables and respects the hierarchy principle. We start by defining a rather large model:

```
biglm <- lm(undercount ~ (equip+econ+usage+atlanta)^2+(equip+econ+
    ↪ usage+atlanta)*(perAA+pergore), gavote)
```

This model includes up to all two-way interactions between the qualitative variables along with all two-way interaction between a qualitative and a quantitative variable. All main effects are included. The `step` command sequentially eliminates terms to minimize the AIC:

```
smallm <- step(biglm,trace=FALSE)
```

The resulting model includes interactions between `equip` and `econ`, `econ` and `perAA`, and `usage` and `perAA`, together with the associated main effects. The `trace=FALSE` argument blocks the large amount of intermediate model information that we would otherwise see.

Linear modeling is also used to try to understand the relationship between the variables — we want to develop an explanation for the data. For this dataset, we are much more interested in explanation than prediction. However, the two goals are not mutually exclusive and often the same methods are used for variable selection in both cases. Even so, when explanation is the goal, it may be unwise to rely on completely automated variable selection methods. For example, the proportion of voters for Gore was eliminated from the model by the AIC-based `step` method and yet we know this variable to be strongly correlated with the proportion of African Americans which is in the model. It would be rash to conclude that the latter variable is important and the former is not — the two are intertwined. Researchers interested in explaining the relationship may prefer a more manual variable selection approach that takes into account background information and is geared toward the substantive questions of interest.

The other major class of variable selection methods is based on testing. We can use F-tests to compare larger models with smaller nested models. A stepwise testing approach can then be applied to select a model. The consensus view among statisticians is that this is an inferior method to variable selection compared to the criterion-based methods. Nevertheless, testing-based methods are still useful, particularly when under manual control. They have the advantage of applicability across a wide class of models where tests have been developed. They allow the user to respect restrictions of hierarchy and situations where certain variables must be included for explanatory purposes. Let's compare the AIC-selected models above to models with one fewer term:

```
drop1(smallm,test="F")
Single term deletions

Model:
undercount ~ equip + econ + usage + perAA + equip:econ + equip:perAA +
    usage:perAA
            Df Sum of Sq    RSS   AIC F value  Pr(F)
<none>                   0.0536 -1231
equip:econ   6    0.0075 0.0612 -1222    3.25 0.0051
equip:perAA  4    0.0068 0.0605 -1220    4.43 0.0021
usage:perAA  1    0.0010 0.0546 -1230    2.65 0.1060
```

We see that the `usage:perAA` can be dropped. A subsequent test reveals that `usage` can also be removed. This gives us a final model of:

```
finalm <- lm(undercount~equip + econ  + perAA + equip:econ + equip:
    ↪ perAA, gavote)
sumary(finalm)
Coefficients: (2 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.04187    0.00503    8.33  6.5e-14
equipOS-CC          -0.01133    0.00737   -1.54  0.12670
equipOS-PC           0.00858    0.01118    0.77  0.44429
equipPAPER          -0.05843    0.03701   -1.58  0.11669
equipPUNCH          -0.01575    0.01875   -0.84  0.40218
econpoor             0.02027    0.00553    3.67  0.00035
econrich            -0.01697    0.01239   -1.37  0.17313
perAA               -0.04204    0.01659   -2.53  0.01239
equipOS-CC:econpoor -0.01096    0.00988   -1.11  0.26922
equipOS-PC:econpoor  0.04838    0.01380    3.51  0.00061
equipPUNCH:econpoor -0.00356    0.01243   -0.29  0.77492
equipOS-CC:econrich  0.00228    0.01538    0.15  0.88246
equipOS-PC:econrich -0.01332    0.01705   -0.78  0.43615
equipPUNCH:econrich  0.02003    0.02200    0.91  0.36405
equipOS-CC:perAA     0.10725    0.03286    3.26  0.00138
equipOS-PC:perAA    -0.00591    0.04341   -0.14  0.89198
equipPAPER:perAA     0.12914    0.08181    1.58  0.11668
equipPUNCH:perAA     0.08685    0.04650    1.87  0.06388

n = 159, p = 18, Residual SE = 0.020, R-Squared = 0.43
```

Because there are only two paper-using counties, there is insufficient data to estimate the interaction terms involving paper. This model output is difficult to interpret because of the interaction terms.

Conclusion: Let's attempt an interpretation of this final model. Certainly we should explore more models and check more diagnostics, so our conclusions can only be tentative. The reader is invited to investigate other possibilities.

To interpret interactions, it is often helpful to construct predictions for all the levels of the variables involved. Here I generate all combinations of `equip` and `econ` for a median proportion of `perAA`:

```
pdf <- data.frame(econ=rep(levels(gavote$econ), 5),  equip=rep(levels(
    ↪ gavote$equip), rep(3,5)), perAA=0.233)
```

We now compute the predicted undercount for all 15 combinations and display the result in a table:

```
pp <- predict(finalm,new=pdf)
xtabs(round(pp,3) ~ econ + equip, pdf)
         equip
econ      LEVER  OS-CC  OS-PC  PAPER  PUNCH
  middle  0.032  0.046  0.039  0.004  0.037
  poor    0.052  0.055  0.108  0.024  0.053
  rich    0.015  0.031  0.009 -0.013  0.040
```

We can see that the undercount is lower in richer counties and higher in poorer counties. The amount of difference depends on the voting system. Of the three most commonly used voting methods, the LEVER method seems best. It is hard to separate the two optical scan methods, but there is clearly a problem with the precinct count in poorer counties, which is partly due to the two outliers we observed earlier. We notice one impossible prediction — a negative undercount in rich paper-using counties, but given the absence of such data (there were no such counties), we are not too disturbed.

We use the same approach to investigate the relationship between the proportion of African Americans and the voting equipment. We set the proportion of African Americans at three levels — the first quartile, the median and the third quartile — and then compute the predicted undercount for all types of voting equipment. We set the `econ` variable to middle:

```
pdf <- data.frame(econ=rep("middle",15), equip=rep(levels(gavote$equip
    ↪ ),  rep(3,5)), perAA=rep(c(.11,0.23,0.35),5))
pp <- predict(finalm,new=pdf)
```

We create a three-level factor for the three levels of `perAA` to aid the construction of the table:

```
propAA <- gl(3,1,15,labels=c("low","medium","high"))
xtabs(round(pp,3) ~ propAA + equip,pdf)
        equip
propAA    LEVER  OS-CC  OS-PC  PAPER  PUNCH
  low     0.037  0.038  0.045 -0.007  0.031
  medium  0.032  0.046  0.039  0.003  0.036
  high    0.027  0.053  0.034  0.014  0.042
```

We see that the effect of the proportion of African Americans on the undercount is mixed. High proportions are associated with higher undercounts for OS-CC and PUNCH and associated with lower undercounts for LEVER and OS-PC.

In summary, we have found that the economic status of a county is the clearest factor determining the proportion of undercounted votes, with richer counties having lower undercounts. The type of voting equipment and the proportion of African Americans do have some impact on the response, but the direction of the effects is not simply stated. We would like to emphasize again that this dataset deserves further analysis before any definitive conclusions are drawn.

Exercises

Since this is a review chapter, it is best to consult the recommended background texts for specific questions on linear models. However, it is worthwhile gaining some practice using R on some real data. Your data analysis should consist of:

- An initial data analysis that explores the numerical and graphical characteristics of the data.
- Variable selection to choose the best model.
- An exploration of transformations to improve the fit of the model.
- Diagnostics to check the assumptions of your model.
- Some predictions of future observations for interesting values of the predictors.
- An interpretation of the meaning of the model with respect to the particular area of application.

There is always some freedom in deciding which methods to use, in what order to apply them, and how to interpret the results. So there may not be one clear right answer and good analysts may come up with different models.

Here are some datasets which should provide some good practice at building linear models:

1. The `swiss` data — use `Fertility` as the response.
2. The `rock` data — use `perm` as the response.
3. The `mtcars` data — use `mpg` as the response.
4. The `attitude` data — use `rating` as the response.
5. The `prostate` data — use `lpsa` as the response.
6. The `teengamb` data — use `gamble` as the response.

Bibliography

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. New York: Wiley.

Agresti, A. (2013). *Categorical Data Analysis* (3 ed.). New York: John Wiley.

Allison, T. and D. Cicchetti (1976). Sleep in mammals: Ecological and constitutional correlates. *Science 194*, 732–734.

Andrews, D. and A. Herzberg (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. New York: Springer-Verlag.

Appleton, D., J. French, and M. Vanderpump (1996). Ignoring a covariate: An example of Simpson's paradox. *American Statistician 50*, 340–341.

Bates, D. (2005). Fitting linear mixed models in R. *R News 5*(1), 27–30.

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton University Press.

Bergman, B. and A. Hynen (1997). Dispersion effects from unreplicated designs in the 2^{k-p} series. *Technometrics 39*, 191–198.

Bickel, P. and K. Doksum (2015). *Mathematical Statistics: Basic Ideas and Selected Topics* (2 ed.). Boca Raton, FL: CRC Press.

Bilder, C. R. and T. M. Loughin (2014). *Analysis of Categorical Data with R*. Boca Raton, FL: CRC Press.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

Bishop, Y., S. Fienberg, and P. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Blasius, J. and M. Greenacre (1998). *Visualization of Categorical Data*. San Diego: Academic Press.

Bliss, C. I. (1935). The calculation of the dose-mortality curve. *Annals of Applied Biology 22*, 134–167.

Bliss, C. I. (1967). *Statistics in Biology*. New York: McGraw Hill.

Bowman, A. and A. Azzalini (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford: Oxford University Press.

Box, G. and R. Meyer (1986). Dispersion effects from fractional designs. *Technometrics 28*, 19–27.

Box, G. P., S. Bisgaard, and C. Fung (1988). An explanation and critique of

Taguchi's contributions to quality engineering. *Quality and Reliability Engineering International 4*, 123–131.

Box, G. P., W. G. Hunter, and J. S. Hunter (1978). *Statistics for Experimenters*. New York: Wiley.

Breiman, L. (2001a). Random forests. *Machine Learning 45*(1), 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science 16*, 199–231.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall.

Breiman, L. and J. H. Friedman (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association 80*, 580–598.

Breslow, N. (1982). Covariance adjustment of relative-risk estimates in matched studies. *Biometrics 38*, 661–672.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*, 9–25.

Cameron, A. and P. Trivedi (1998). *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Christensen, R. (1997). *Log-Linear Models and Logistic Regression* (2 ed.). New York: Springer.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association 74*, 829–836.

Clogg, C. and E. Shihadeh (1994). *Statistical Models for Ordinal Variables*. Thousands Oaks, CA: Sage.

Cochran, W. (1954). Some methods of strengthening the common χ^2 tests. *Biometrics 10*, 417–451.

Collett, D. (2003). *Modelling Binary Data* (2 ed.). London: Chapman & Hall.

Comizzoli, R. B., J. M. Landwehr, and J. D. Sinclair (1990). Robust materials and processes: Key to reliability. *AT&T Technical Journal 69*(6), 113–128.

Cox, D. (1970). *Analysis of Binary Data*. London: Spottiswoode, Ballantyne and Co.

Cox, D. and D. Hinkley (1974). *Theoretical Statistics*. London: Chapman & Hall.

Crainiceanu, C. and D. Ruppert (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B 66*, 165–185.

Crowder, M. (1978). Beta-binomial anova for proportions. *Applied Statistics 27*, 34–37.

Crowder, M. J. and D. J. Hand (1990). *Analysis of Repeated Measures*. London: Chapman & Hall.

Dalal, S., E. Fowlkes, and B. Hoadley (1989). Risk analysis of the space shuttle:

Pre-Challenger prediction of failure. *Journal of the American Statistical Association 84*, 945–957.

Dalgaard, P. (2002). *Introductory Statistics with R*. New York: Springer.

Daubechies, I. (1991). *Ten Lectures on Wavelets*. Philadelphia: SIAM.

Davies, O. (1954). *The Design and Analysis of Industrial Experiments*. New York: Wiley.

Dey, D., S. Ghosh, and B. Mallick (2000). *Generalized Linear Models: A Bayesian Perspective*. New York: Marcel Dekker.

Diggle, P. J., P. Heagerty, K. Y. Liang, and S. L. Zeger (2013). *Analysis of Longitudinal Data* (2 ed.). Oxford: Oxford University Press.

Dobson, A. and A. Barnett (2008). *An Introduction to Generalized Linear Models* (3 ed.). London: Chapman & Hall.

Draper, N. and H. Smith (1998). *Applied Regression Analysis* (3rd ed.). New York: Wiley.

Dunn, P. K. and G. K. Smyth (2005). Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing 15*(4), 267–280.

Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer.

Faraway, J. (2014). *Linear Models with R* (2 ed.). Boca Raton, FL: Chapman & Hall/CRC.

Faraway, J. and C. Chatfield (1998). Time series forecasting with neural networks: A case study. *Applied Statistics 47*, 231–250.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*, 27–38.

Fitzmaurice, G., N. Laird, and J. Ware (2004). *Applied Longitudinal Analysis*. Hoboken, NJ: Wiley-Interscience.

Fletcher, R. (1987). *Practical Methods of Optimization* (2 ed.). Chichester, UK: John Wiley.

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software 8*(15), 1–27.

Frees, E. (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge: Cambridge University Press.

Friedman, J. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics 19*, 1–141.

Frome, E. and R. DuFrain (1986). Maximum likelihood estimation for cytogenic dose-response curves. *Biometrics 42*, 73–84.

Gelman, A. (2005). Analysis of variance — why it is more important than ever (with discussion). *Annals of Statistics 33*, 1–53.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis* (3 ed.). Boca Raton, FL: CRC Press.

Gelman, A. and J. Hill (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press.

Gelman, A. and Y.-S. Su (2013). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models.* R package version 1.6-10.

Gill, J. (2001). *Generalized Linear Models: A Unified Approach.* Thousand Oaks, CA: Sage.

Goldstein, H. (1995). *Multilevel Statistical Models* (2 ed.). London: Arnold.

Green, P. and B. Silverman (1993). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* London: Chapman & Hall.

Gu, C. (2002). *Smoothing Spline ANOVA Models.* New York: Springer-Verlag.

Haberman, S. (1977). *The Analysis of Frequency Data.* Chicago, IL: University of Chicago Press.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software 33*(2), 1–22.

Halekoh, U. and S. Højsgaard (2014). A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models—the R package pbkrtest. *Journal of Statistical Software 59*, 1–32.

Hall, S. (1994). Analysis of defectivity of semiconductor wafers by contigency table. *Proceedings of the Institute of Environmental Sciences 1*, 177–183.

Hallin, M. and J.-F. Ingenbleek (1983). The Swedish automobile portfolio in 1977. A statistical study. *Scandinavian Actuarial Journal 83*, 49–64.

Hardin, J. and J. Hilbe (2003). *Generalized Estimating Equations.* Boca Raton, FL: Chapman & Hall/CRC Press.

Härdle, W. (1991). *Smoothing Techniques with Implementation in S.* New York: Springer.

Harrell, F. (2001). *Regression Modelling Strategies.* New York: Springer-Verlag.

Hart, J. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests.* New York: Springer.

Hartigan, J. and B. Kleiner (1981). Mosaics for contingency tables. In W. Eddy (Ed.), *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, New York, pp. 268–273. Springer-Verlag.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models.* London: Chapman & Hall.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Hauck, W. and A. Donner (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association 72*, 851–853.

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation* (2 ed.). Upper

Saddle River, NJ: Prentice Hall.

Hertz, J., A. Krogh, and R. Palmer (1991). *Introduction to the Theory of Neural Computation*. Redwood City, CA: Addison–Wesley.

Hilbe, J. M. (2009). *Logistic Regression Models*. Boca Raton, FL: CRC Press.

Hill, M. S. (1992). *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, CA: Sage.

Hinde, J. and C. Demetrio (1988). Overdispersion: Models and estimation. *Computational Statistics and Data Analysis 27*, 151–170.

Højsgaard, S., U. Halekoh, and J. Yan (2005). The R package geepack for generalized estimating equations. *Journal of Statistical Software 15*(2), 1–11.

Hornik, K., M. Stinchcombe, and H. White (1989). Multilayer feedforward networks are universal approximators. *Neural Networks 2*, 359–366.

Hosmer, D. and S. Lemeshow (2013). *Applied Logistic Regression* (3 ed.). New York: Wiley.

Johnson, M. P. and P. H. Raven (1973). Species number and endemism: The Galápagos archipelago revisited. *Science 179*, 893–895.

Kenward, M. and J. Roger (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics 53*, 983–997.

Kleinbaum, D. G. and M. Klein (2002). *Logistic Regression: A Self-Learning Text*. New York: Springer.

Kosmidis, I. (2013). *brglm: Bias Reduction in Binary-Response Generalized Linear Models*. R package version 0.5-9.

Kovac, A. and B. W. Silverman (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *Journal of the American Statistical Association 95*(449), 172–183.

Kunsch, H. R., L. A. Stefanski, and R. J. Carroll (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association 84*(406), 460–466.

Lawless, J. (1987). Negative binomial and mixed poisson regression. *Canadian Journal of Statistics 15*, 209–225.

Le, C. T. (1998). *Applied Categorical Data Analysis*. Newbury Park, CA: Wiley.

Leonard, T. (2000). *A Course in Categorical Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC Press.

Liaw, A. and M. Wiener (2002). Classification and regression by randomforest. *R News 2*(3), 18–22.

Lindsey, J. K. (1997). *Applying Generalized Linear Models*. New York: Springer.

Lindsey, J. K. (1999). *Models for Repeated Measurements* (2 ed.). Oxford: Oxford University Press.

Loader, C. (1999). *Local Regression and Likelihood*. New York: Springer.

Long, J. S. (1990). The origins of sex differences in science. *Social Forces 68*(4), 1297–1316.

Lowe, C., C. Roberts, and S. Lloyd (1971). Malformations of the central nervous system and softness of local water supplies. *British Medical Journal 15*, 357–361.

Lunn, D., C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter (2012). *The BUGS Book: A Practical Introduction to Bayesian Analysis.* Boca Raton, FL: CRC Press.

Maindonald, J. and J. Braun (2010). *Data Analysis and Graphics Using R* (3 ed.). Cambridge, UK: Cambridge University Press.

Manly, B. (1978). Regression models for proportions with extraneous variance. *Biometrie-Praximetrie 18*, 1–18.

Mantel, N. and W. Haenszel (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute 22*, 719–748.

McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics 11*, 59–67.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models* (2 ed.). London: Chapman & Hall.

McCulloch, C. and S. Searle (2002). *Generalized, Linear, and Mixed Models.* New York: Wiley.

McCulloch, W. and W. Pitts (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics 5*, 115–133.

Mehta, C. and N. Patel (1995). Exact logistic regression: Theory and examples. *Statistics in Medicine 14*, 2143–2160.

Menard, S. (2002). *Applied Logistic Regression Analysis* (2 ed.). Thousands Oaks, CA: Sage.

Meyer, M. (2002). Uncounted votes: Does voting equipment matter? *Chance 15*(4), 33–38.

Milliken, G. A. and D. E. Johnson (1992). *Analysis of Messy Data*, Volume 1. New York: Van Nostrand Reinhold.

Morgan, J. and J. Sonquist (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association 58*, 415–434.

Mortimore, P., P. Sammons, L. Stoll, D. Lewis, and R. Ecob (1988). *School Matters.* Wells, UK: Open Books.

Müller, S., J. L. Scealy, and A. H. Welsh (2013). Model selection in linear mixed models. *Statistical Science 28*(2), 135–167.

Myers, R. and D. Montgomery (1997). A tutorial on generalized linear models. *Journal of Quality Technology 29*, 274–291.

Myers, R., D. Montgomery, and G. Vining (2002). *Generalized Linear Models: With Applications in Engineering and the Sciences.* New York: Wiley.

Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika 78*, 691–692.

Nason, G. (2013). *wavethresh: Wavelets Statistics and Transforms.* R package version 4.6.6.

Neal, R. (1996). *Bayesian Learning for Neural Networks.* New York: Springer–Verlag.

Nelder, J., Y. Lee, B. Bergman, A. Hynen, A. Huele, and J. Engel (1998). Letter to editor: Joint modeling of mean and dispersion. *Technometrics 40*, 168–175.

Nelder, J. and R. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A 132*, 370–384.

Payne, C. (1987). *The GLIM System Release 3.77 Manual* (2 ed.). Oxford: Numerical Algorithms Group.

Pignatiello, J. J. and J. S. Ramberg (1985). Contribution to discussion of offline quality control, parameter design and the Taguchi method. *Journal of Quality Technology 17*, 198–206.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-Effects Models in S and S-PLUS.* New York: Springer.

Powers, D. and Y. Xie (2000). *Statistical Methods for Categorical Data Analysis.* San Diego, CA: Academic Press.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics 9*, 705–724.

Purott, R. and E. Reeder (1976). The effect of changes in dose rate on the yield of chromosome aberrations in human lymphocytes exposed to gamma radiation. *Mutation Research 35*, 437–444.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufman.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rasmussen, C. and C. Williams (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: The MIT Press.

Raudenbush, S. and A. Bryk (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2 ed.). Thousand Oaks, CA: Sage.

Rice, J. (1998). *Mathematical Statistics and Data Analysis.* Monterey, CA: Brooks Cole.

Ripley, B. (1996). *Pattern Recognition and Neural Networks.* Cambridge, UK: Cambridge University Press.

Rosenman, R. H., R. J. Brand, C. D. Jenkins, M. Friedman, R. Straus, and M. Wurm (1975). Coronary heart disease in the western collaborative group study: Final follow-up experience of 8 1/2 years. *JAMA 233*(8), 872–877.

Rosenstone, S. J., D. R. Kinder, and W. E. Miller (1997). *American National Election Study*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research.

Rue, H., S. Martino, and N. Chopin (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B 71*(2), 319–392.

Santner, T. and D. Duffy (1989). *The Statistical Analysis of Discrete Data.* New York: Springer.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika 78*(4), 719–727.

Scheffé, H. (1959). *The Analysis of Variance.* New York: Wiley.

Scheipl, F., S. Greven, and H. Kuechenhoff (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics and Data Analysis 52*(7), 3283–3299.

Scrucca, L. (2012). *dispmod: Dispersion Models.* R package version 1.1.

Searle, S., G. Casella, and C. McCulloch (1992). *Variance Components.* New York: Wiley.

Seshadri, V. (1993). *The Inverse Gaussian Distribution.* Oxford: Clarendon.

Sheldon, F. (1960). Statistical techniques applied to production situations. *Industrial and Engineering Chemistry 52*, 507–509.

Simonoff, J. (1996). *Smoothing Methods in Statistics.* New York: Springer.

Simonoff, J. (2003). *Analyzing Categorical Data.* New York: Springer.

Simpson, D. P., T. G. Martins, A. Riebler, G.-A. Fuglstad, H. Rue, and S. H. Sørbye (2014). Penalising model component complexity: A principled, practical approach to constructing priors. *arXiv:1403.4630.*

Simpson, E. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B 13*, 238–241.

Smyth, G., F. Huele, and A. Verbyla (2001). Exact and approximate reml for heteroscedastic regression. *Statistical Modelling: An International Journal 1*, 161–175.

Snedecor, G. and W. Cochran (1989). *Statistical Methods* (8 ed.). Ames, IA: Iowa State University Press.

Snee, R. (1974). Graphical display of two-way contingency tables. *American Statistician 28*, 9–12.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(4), 583–639.

Stan Development Team (2015). *Stan Modeling Language Users Guide and Reference Manual, Version 2.8.0.*

Steele, R. (1998). *Effect of Surface and Vision on Balance.* Ph. D. thesis, Department of Physiotherapy, University of Queensland.

Stone, C. (1985). Additive regression and other nonparametric models. *Annals of Statistics 13*, 689–705.

Stram, D. and J. Lee (1994). Variance components testing in the longitudinal mixed-effects model. *Biometrics 50*, 1171–1179.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika 42*, 412–416.

Thall, P. F. and S. C. Vail (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics 46*, 657–671.

Tukey, J. (1977). *Exploratory Data Analysis*. New York: Addison Wesley.

Venables, W. and B. Ripley (2002). *Modern Applied Statistics with S* (4 ed.). New York: Springer.

Verbeke, G. and G. Molenberghs (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.

Wand, M. and M. Jones (1995). *Kernel Smoothing*. London: Chapman & Hall.

Wang, J., R. Zamar, A. Marazzi, V. Yohai, M. Salibian-Barrera, R. Maronna, E. Zivot, D. Rocke, D. Martin, M. Maechler, and K. Konis. (2014). *robust: Robust Library*. R package version 0.4-16.

Wedderburn, R. W. M. (1974). Quasilikelihood functions, generalized linear models and the Gauss–Newton method. *Biometrika 61*, 439–447.

Weisberg, S. (2005). *Applied Linear Regression* (3 ed.). New York: Wiley.

Wheeler, B. (2013). *SuppDists: Supplementary Distributions*. R package version 1.1-9.1.

Whitmore, G. (1986). Inverse Gaussian ratio estimation. *Applied Statistics 35*, 8–15.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Wickham, H. and R. Francois (2015). *dplyr: A Grammar of Data Manipulation*. R package version 0.4.1.

Wilkinson, G. and C. Rogers (1973). Symbolic description of factorial models for the analysis of variance. *Applied Statistics 22*, 392–399.

Williams, D. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics 31*, 144–148.

Williams, D. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics 36*, 181–191.

Wold, S., A. Ruhe, H. Wold, and W. Dunn (1984). The collinearity problem in linear regression: The partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing 5*, 735–743.

Wood, S. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistal Society, Series B 62*, 413–428.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: CRC Press.

Wood, S. (2015). *Core Statistics*. Cambridge: Cambridge University Press.

Yee, T. (2010). The vgam package for categorical data analysis. *Journal of Statistical Software 32*(10), 1–34.

Yule, G. (1903). Notes on the theory of association of attributes in statistics. *Biometrika 2*, 121–134.

Zeileis, A. (2004). Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software 11*(10), 1–17.

Zeileis, A., C. Kleiber, and S. Jackman (2008). Regression models for count data in R. *Journal of Statistical Software 27*(8), 1–25.