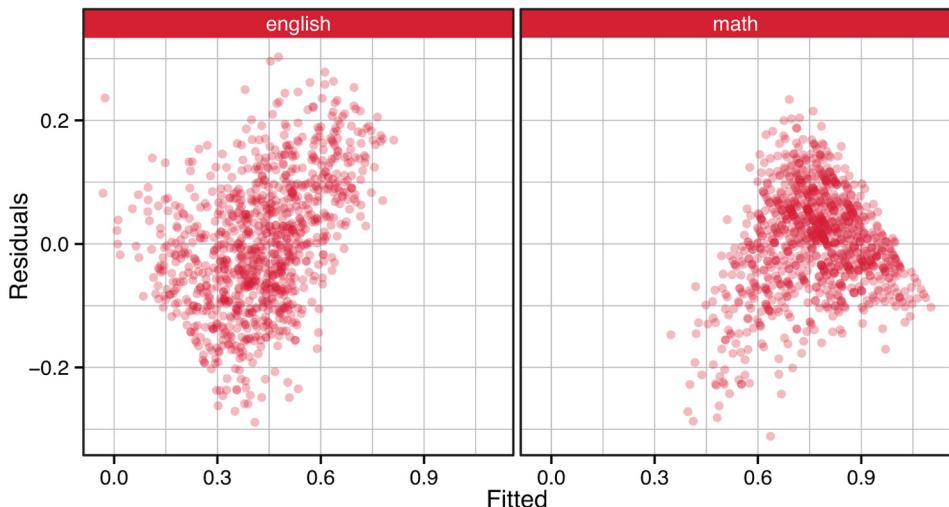


Texts in Statistical Science

# 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**

### **Series Editors**

Francesca Dominici, *Harvard School of Public Health, USA*

Julian J. Faraway, *University of Bath, UK*

Martin Tanner, *Northwestern University, USA*

Jim Zidek, *University of British Columbia, Canada*

### **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. Riedwy1
- 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

© 2016 by Taylor & Francis Group, LLC  
CRC Press is an imprint of Taylor & Francis Group, an Informa business

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

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

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access [www.copyright.com](http://www.copyright.com) (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

**Trademark Notice:** Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

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

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

---

# Contents

---

<b>Preface</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Binary Response</b>	<b>25</b>
2.1 Heart Disease Example	25
2.2 Logistic Regression	28
2.3 Inference	32
2.4 Diagnostics	34
2.5 Model Selection	38
2.6 Goodness of Fit	40
2.7 Estimation Problems	44
<b>3 Binomial and Proportion Responses</b>	<b>51</b>
3.1 Binomial Regression Model	51
3.2 Inference	53
3.3 Pearson's $\chi^2$ Statistic	54
3.4 Overdispersion	55
3.5 Quasi-Binomial	60
3.6 Beta Regression	64
<b>4 Variations on Logistic Regression</b>	<b>67</b>
4.1 Latent Variables	67
4.2 Link Functions	68
4.3 Prospective and Retrospective Sampling	71
4.4 Prediction and Effective Doses	74
4.5 Matched Case-Control Studies	75
<b>5 Count Regression</b>	<b>83</b>
5.1 Poisson Regression	83
5.2 Dispersed Poisson Model	87
5.3 Rate Models	89
5.4 Negative Binomial	92
5.5 Zero Inflated Count Models	94

<b>6 Contingency Tables</b>	<b>103</b>
6.1 Two-by-Two Tables	103
6.2 Larger Two-Way Tables	108
6.3 Correspondence Analysis	110
6.4 Matched Pairs	112
6.5 Three-Way Contingency Tables	114
6.6 Ordinal Variables	121
<b>7 Multinomial Data</b>	<b>129</b>
7.1 Multinomial Logit Model	129
7.2 Linear Discriminant Analysis	134
7.3 Hierarchical or Nested Responses	136
7.4 Ordinal Multinomial Responses	139
<b>8 Generalized Linear Models</b>	<b>151</b>
8.1 GLM Definition	151
8.2 Fitting a GLM	154
8.3 Hypothesis Tests	156
8.4 GLM Diagnostics	159
8.5 Sandwich Estimation	168
8.6 Robust Estimation	169
<b>9 Other GLMs</b>	<b>175</b>
9.1 Gamma GLM	175
9.2 Inverse Gaussian GLM	181
9.3 Joint Modeling of the Mean and Dispersion	184
9.4 Quasi-Likelihood GLM	186
9.5 Tweedie GLM	188
<b>10 Random Effects</b>	<b>195</b>
10.1 Estimation	196
10.2 Inference	201
10.3 Estimating Random Effects	207
10.4 Prediction	208
10.5 Diagnostics	210
10.6 Blocks as Random Effects	211
10.7 Split Plots	215
10.8 Nested Effects	219
10.9 Crossed Effects	222
10.10 Multilevel Models	224
<b>11 Repeated Measures and Longitudinal Data</b>	<b>237</b>
11.1 Longitudinal Data	238
11.2 Repeated Measures	243
11.3 Multiple Response Multilevel Models	248

<b>12 Bayesian Mixed Effect Models</b>	<b>255</b>
12.1 STAN	256
12.2 INLA	266
12.3 Discussion	271
<b>13 Mixed Effect Models for Nonnormal Responses</b>	<b>275</b>
13.1 Generalized Linear Mixed Models	275
13.2 Inference	275
13.3 Binary Response	277
13.4 Count Response	284
13.5 Generalized Estimating Equations	291
<b>14 Nonparametric Regression</b>	<b>297</b>
14.1 Kernel Estimators	299
14.2 Splines	302
14.3 Local Polynomials	306
14.4 Confidence Bands	307
14.5 Wavelets	308
14.6 Discussion of Methods	313
14.7 Multivariate Predictors	315
<b>15 Additive Models</b>	<b>321</b>
15.1 Modeling Ozone Concentration	323
15.2 Additive Models Using <code>mgcv</code>	324
15.3 Generalized Additive Models	330
15.4 Alternating Conditional Expectations	331
15.5 Additivity and Variance Stabilization	334
15.6 Generalized Additive Mixed Models	336
15.7 Multivariate Adaptive Regression Splines	337
<b>16 Trees</b>	<b>343</b>
16.1 Regression Trees	343
16.2 Tree Pruning	346
16.3 Random Forests	351
16.4 Classification Trees	354
16.5 Classification Using Forests	360
<b>17 Neural Networks</b>	<b>365</b>
17.1 Statistical Models as NNs	366
17.2 Feed-Forward Neural Network with One Hidden Layer	367
17.3 NN Application	368
17.4 Conclusion	371

<b>A Likelihood Theory</b>	<b>375</b>
A.1 Maximum Likelihood	375
A.2 Hypothesis Testing	378
A.3 Model Selection	381
<b>B About R</b>	<b>383</b>
<b>Bibliography</b>	<b>385</b>
<b>Index</b>	<b>395</b>

---

# 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  $x_1, \dots, x_p$ . The model takes the form:

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

where  $\varepsilon$  is normally distributed. This book presents three extensions to this framework. The first generalizes the  $y$  part; the second, the  $\varepsilon$  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,  $\varepsilon$ . 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](http://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 Mairdonald 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 2 2 ...
 $ perAA : num  0.182 0.23 0.131 0.476 0.359 0.024 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 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))
```

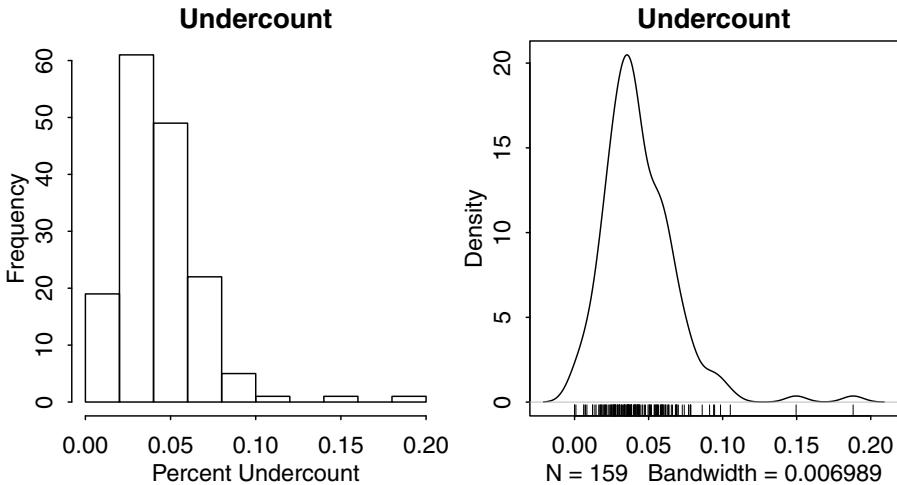


Figure 1.1 *Histogram of the undercount is shown on the left and a density estimate with a data rug is shown on the right.*

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$peregore <- gavote$gore/gavote$votes
plot(peregore ~ 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

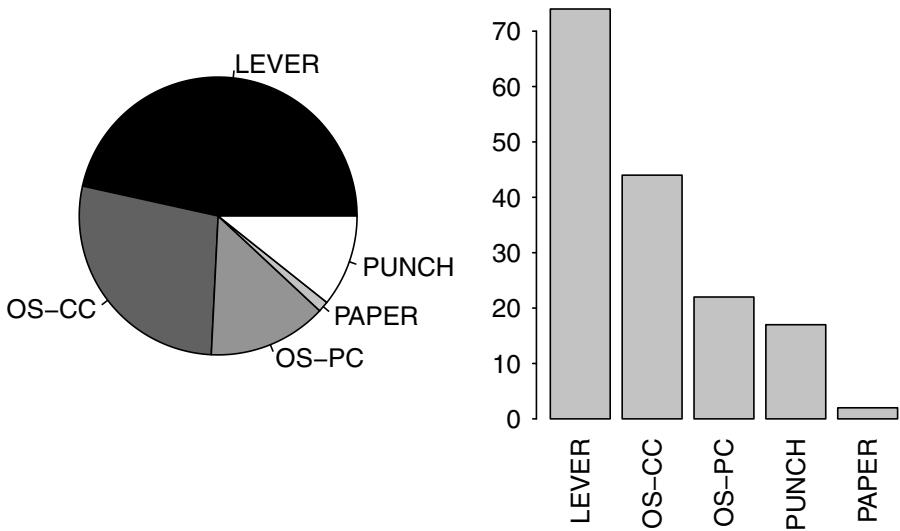


Figure 1.2 *Pie chart of the voting equipment frequencies is shown on the left and a Pareto chart on the right.*

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

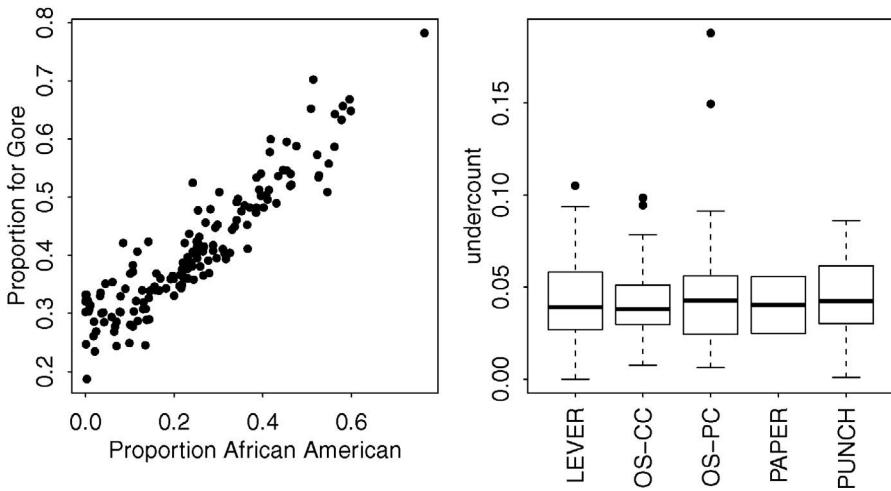


Figure 1.3 A scatterplot of proportions of Gore voters and African Americans by county is shown on the left. Boxplots showing the distribution of the undercount by voting equipment are shown on the right.

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:

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

where  $\beta_i$ ,  $i = 0, 1, 2, \dots, p - 1$  are unknown *parameters*.  $\beta_0$  is called the *intercept* term. The *response* is  $Y$  and the *predictors* are  $X_1, \dots, X_{p-1}$ . The predictors may be the original variables in the dataset or transformations or combinations of them. The error  $\epsilon$  represents the difference between what is explained by the systematic part of the model and what is observed.  $\epsilon$  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:

$$y = X\beta + \epsilon$$

where, in terms of the  $n$  data points,  $y = (y_1, \dots, y_n)^T$ ,  $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$ ,  $\beta = (\beta_0, \dots, \beta_{p-1})^T$  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  $\beta$ , called  $\hat{\beta}$ , minimizes:

$$\sum \epsilon_i^2 = \epsilon^T \epsilon = (y - X\beta)^T (y - X\beta)$$

Differentiating with respect to  $\beta$  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:

$$\text{undercount} = \beta_0 + \beta_1 \text{pergore} + \beta_2 \text{perAA} + \epsilon$$

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  $\beta$ , 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  $\epsilon$ . 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  $\beta$  (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  $\hat{\epsilon} = y - X\hat{\beta} = y - \hat{y}$ . 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  $\hat{\epsilon}^T \hat{\epsilon}$ . 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  $\sigma^2$ , then  $\sigma$  is estimated by the residual standard error computed from  $\sqrt{(\text{RSS}/\text{df})}$ . 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*:

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

where  $\text{TSS} = \sum(y_i - \bar{y})^2$  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)
```

```
summary(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:

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

So  $\beta_3$  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  $j = 2, \dots, k$  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)
summary(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 written in a mathematical form:

$$\text{undercount} = \beta_0 + \beta_1 \text{cperAA} + \beta_2 \text{cpergore} + \beta_3 \text{usageurban} + \beta_4 \text{equipOSCC} + \beta_5 \text{equipOSPC} + \beta_6 \text{equipPAPER} + \beta_7 \text{equipPUNCH} + \beta_8 \text{cpergore : usageurban} + \varepsilon \quad (1.1)$$

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  $\Omega$  and a smaller model  $\omega$  contained within that can be represented as a linear restriction on the parameters of the larger model. Most often, the predictors in  $\omega$  are just a subset of the predictors in  $\Omega$ .

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

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

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

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:

	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  $\beta$  using:

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

where  $t_{n-p}^{(\alpha/2)}$  is the upper  $\alpha/2^{\text{th}}$  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:  $\epsilon$ . 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(lmobj)
```

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  $|\hat{\epsilon}|$  tends to be rather skewed and it is better to use  $\sqrt{|\hat{\epsilon}|}$ . 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  $i = 1, \dots, n$ . 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  $X\hat{\beta} = X(X^T X)^{-1}X^T y = Hy$  where the *hat-matrix*  $H = X(X^T X)^{-1}X^T$ .  $h_i = H_{ii}$  are called *leverages* and are useful diagnostics.

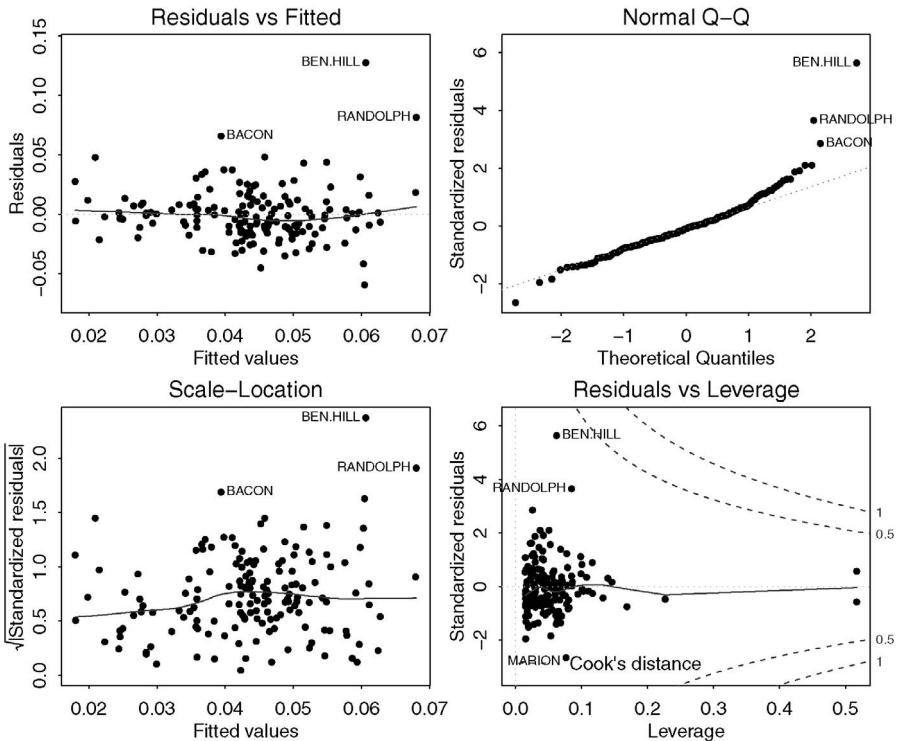


Figure 1.4 *Diagnostics obtained from plotting the model object.*

For example, since  $\text{var } \hat{\epsilon}_i = \sigma^2(1 - h_i)$ , a large leverage,  $h_i$ , will tend to make  $\text{var } \hat{\epsilon}_i$  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  $\Phi^{-1}\left(\frac{n+i}{2n+1}\right)$  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  $\hat{y} + \hat{\beta}_i x_i$  against  $x_i$ . To see the motivation, look at the response with the predicted effect of the other  $X$  removed:

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

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

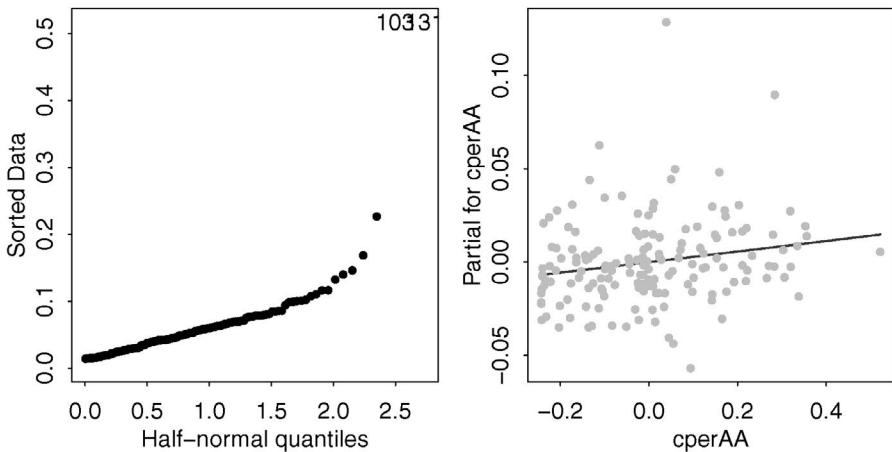


Figure 1.5 *Half-normal plot of the leverages is shown on the left and a partial residual plot for the proportion of African Americans is shown on the right.*

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
equipPOS-CC	0.007	0.004	1.802
equipPOS-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  $\text{var } y_i$ .

Now  $\text{var } y$  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 pre-

cludes 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)
```

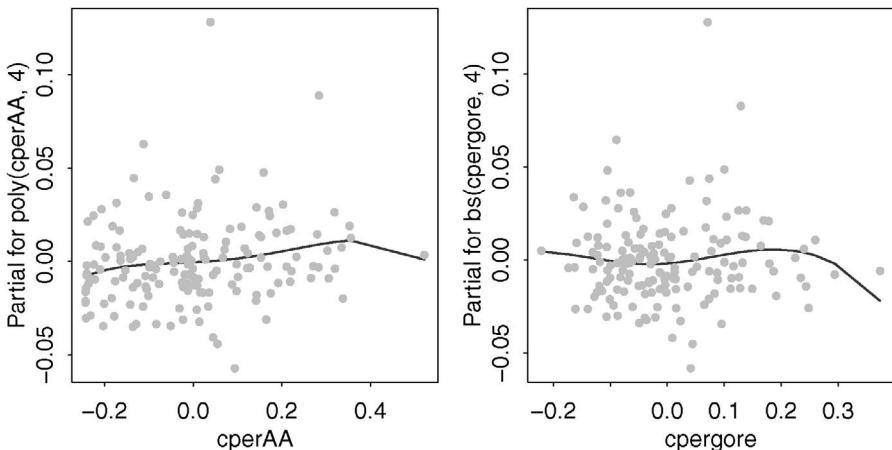


Figure 1.6 *Partial fits using orthogonal polynomials for `cperAA` (shown on the left) and cubic B-splines for `cpergore` (shown on the right).*

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(bigm, 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)
summary(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 esti-

mate 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.

## Binary Response

---

### 2.1 Heart Disease Example

What might affect the chance of getting heart disease? One of the earliest studies addressing this issue started in 1960 and used 3154 healthy men, aged from 39 to 59, from the San Francisco area. At the start of the study, all were free of heart disease. Eight and a half years later, the study recorded whether these men now suffered from heart disease along with many other variables that might be related to the chance of developing this disease. We load a subset of this data from the Western Collaborative Group Study described in Rosenman et al. (1975):

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

We start by focusing on just three of the variables in the dataset:

```
summary(wcgs[, c("chd", "height", "cigs")])
```

	chd	height	cigs
no :2897	Min. :60.0	Min. : 0.0	
yes: 257	1st Qu.:68.0	1st Qu.: 0.0	
	Median :70.0	Median : 0.0	
	Mean :69.8	Mean :11.6	
	3rd Qu.:72.0	3rd Qu.:20.0	
	Max. :78.0	Max. :99.0	

We see that only 257 men developed heart disease as given by the factor variable chd. The men vary in height (in inches) and the number of cigarettes (cigs) smoked per day. We can plot these data using R base graphics:

```
plot(height ~ chd, wcgs)
wcgs$y <- ifelse(wcgs$chd == "no", 0, 1)
plot(jitter(y, 0.1) ~ jitter(height), wcgs, xlab="Height", ylab="Heart
    ↪ Disease", pch=".")
```

The first panel in Figure 2.1 shows a boxplot. This shows the similarity in the distribution of heights of the two groups of men with and without heart disease. But the heart disease is the response variable so we might prefer a plot which treats it as such. We convert the absence/presence of disease into a numerical 0/1 variable and plot this in the second panel of Figure 2.1. Because heights are reported as round numbers of inches and the response can only take two values, it is sensible to add a small amount of noise to each point, called *jittering*, so that we can distinguish them. Again we can see the similarity in the distributions. We might think about fitting a line to this plot.

More informative plots may be obtained using the `ggplot2` package of Wickham (2009). In the first panel of Figure 2.2, we see two histograms showing the distribution of heights for both those with and without heart disease. The *dodge* option ensures that the two histograms are interleaved. We see that the two distributions are

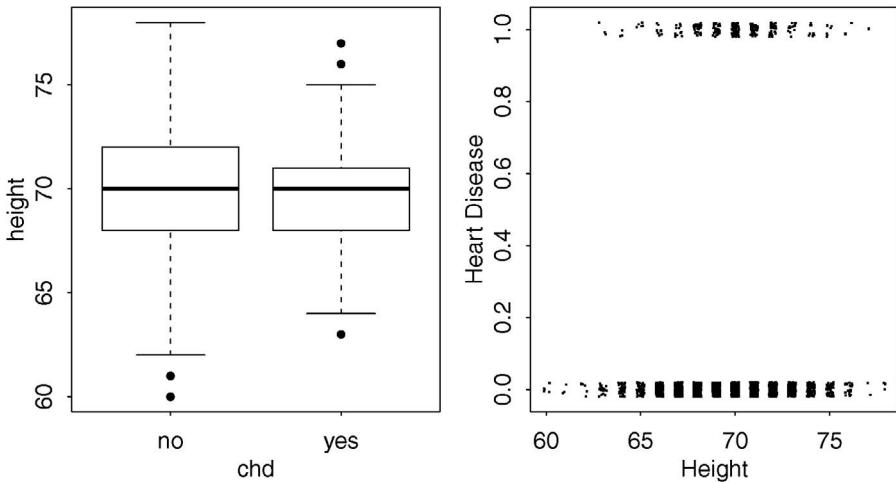


Figure 2.1 Plots of the presence/absence of heart disease according to height in inches.

similar. We also had to set the bin width of the histogram. It was natural to use one inch as all the height measurements are rounded to the nearest inch. In the second panel of Figure 2.2, we see the corresponding histograms for smoking. In this case, we have shown the frequency rather than the count version of the histogram. We see that smokers are more likely to get heart disease.

```
library(ggplot2)
ggplot(wcgs, aes(x=height, color=chd)) + geom_histogram(position="dodge",
  binwidth=1)
ggplot(wcgs, aes(x=cigs, color=chd)) + geom_histogram(position="dodge",
  binwidth=5, aes(y=..density..))
```

It is also helpful to plot both predictors and the response on the same display as seen in Figure 2.3. Some special effort is necessary in constructing this plot because of the large number of men that have the same height and cigarette usage. We use both jittering and partial transparency, controlled by the parameter alpha to provide a clearer sense of the numbers of such repeated points. With a smaller number of points, it is possible to combine both levels of the response in a single plot by using two colors or plotting symbols. But this does not work well with a large number of points as it becomes difficult to distinguish the two. We have chosen to display the two levels on separate panels called *facets* in the ggplot2 package. It is worth the effort of plotting data well to gain insight into the best approaches to modeling.

```
ggplot(wcgs, aes(x=height,y=cigs))+geom_point(alpha=0.2, position=
  position_jitter())+facet_grid(~ chd)
```

We would like to predict the heart disease outcome for a given individual and also to explain the relationship between height, cigarette usage and heart disease. We observe that, for the same height and cigarette consumption, both outcomes occur. This occurs quite regularly. Hence it makes better sense to model the probability of getting heart disease rather than the outcome itself.

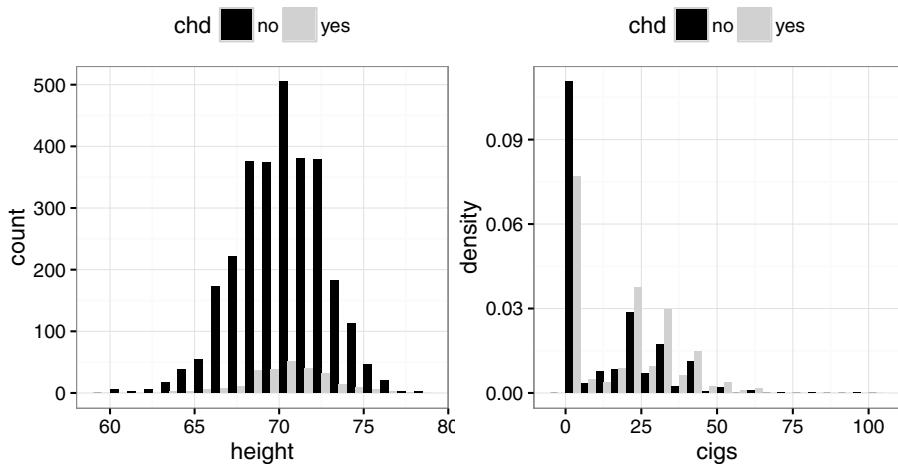


Figure 2.2 *Interleaved histograms of the distribution of heights and cigarette usage for men with and without heart disease.*

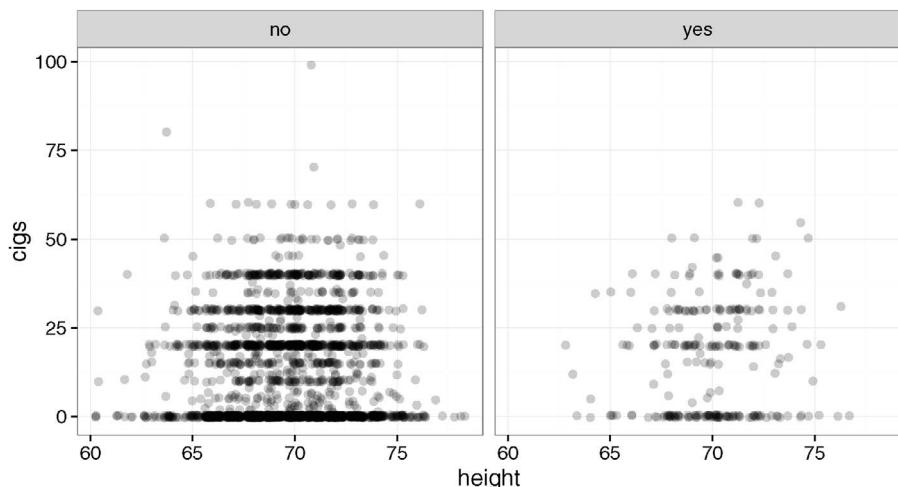


Figure 2.3 *Height and cigarette consumption for men without heart disease on the left and with heart disease on the right. Some jittering and transparency have been used to reduce overplotting problems.*

We might envisage fitting a line to the data in the second panel of Figure 2.1. But there are several problems with this approach. A line would eventually extend above one or below zero and these are not valid probabilities. We could truncate the predictions in these regions to  $[0, 1]$  but this would result in predictor values where the outcome was considered certain. That does not seem reasonable in this dataset, nor would it in many others. So using a linear model for binary response data is usually not a sensible idea.

## 2.2 Logistic Regression

Suppose we have a response variable  $Y_i$  for  $i = 1, \dots, n$  which takes the values zero or one with  $P(Y_i = 1) = p_i$ . This response may be related to a set of  $q$  predictors  $(x_{i1}, \dots, x_{iq})$ . We need a model that describes the relationship of  $x_1, \dots, x_q$  to the probability  $p$ . Following the linear model approach, we construct a *linear predictor*:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_q x_{iq}$$

Since the linear predictor can accommodate quantitative and qualitative predictors with the use of dummy variables and also allows for transformations and combinations of the original predictors, it is very flexible and yet retains interpretability. The idea that we can express the effect of the predictors on the response solely through the linear predictor is important. The idea can be extended to models for other types of response and is one of the defining features of the wider class of generalized linear models (GLMs) discussed later in Chapter 8.

We have seen previously that the linear relation  $\eta_i = p_i$  is not workable because we require  $0 \leq p_i \leq 1$ . Instead we shall use a *link function*  $g$  such that  $\eta_i = g(p_i)$ . We need  $g$  to be monotone and be such that  $0 \leq g^{-1}(\eta) \leq 1$  for any  $\eta$ . The most popular choice of link function in this situation is the *logit*. It is defined so that:

$$\eta = \log(p/(1-p))$$

or equivalently:

$$p = \frac{e^\eta}{1 + e^\eta}$$

Combining the use of the logit link with a linear predictor gives us the term *logistic regression*. Other choices of link function are possible but we will defer discussion of these until later. The logit and its inverse are defined as `logit` and `ilogit` in the `faraway` package. The relationship between  $p$  and the linear predictor  $\eta$  is shown in Figure 2.4.

```
library(faraway)
curve(ilogit(x), -6, 6, xlab=expression(eta), ylab="p")
```

The logistic curve is almost linear in its midrange. This means that for modeling responses that are all of moderate probability, logistic and linear regression will not behave very differently. The curve approaches one at its upper end and zero at its lower end but never reaches these bounds. This means that logistic regression will never predict that anything is inevitable or impossible. Notice that changes in the linear predictor will result in different changes in the probability depending on where

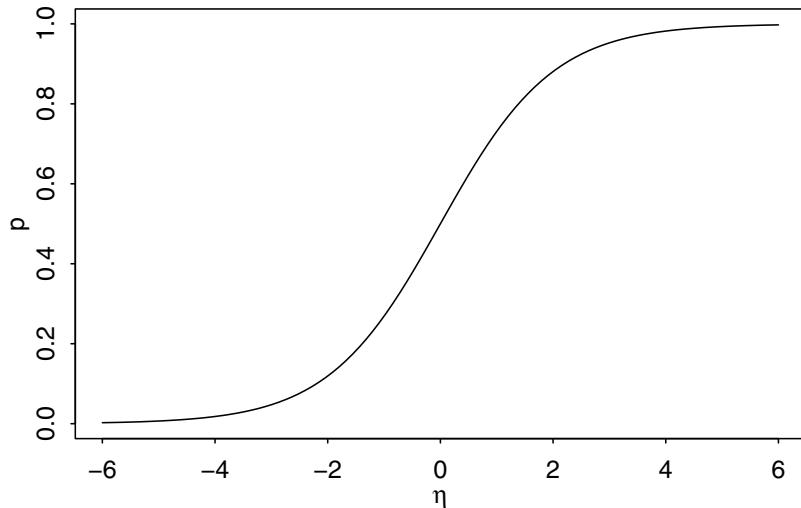


Figure 2.4 A logistic relationship between the probability of the response,  $p$ , and the linear predictor,  $\eta$ .

the change is made. This is in contrast to a linear relationship and makes the interpretation of logistic regression coefficients more difficult.

Now we estimate the parameters of the model. We use the method of maximum likelihood; see Appendix A for a brief introduction to this method. The log-likelihood using the logit link is given by:

$$l(\beta) = \sum_{i=1}^n [y_i \eta_i - \log(1 + e^{\eta_i})]$$

We can maximize this to obtain the maximum likelihood estimates  $\hat{\beta}$  and use the standard theory to obtain approximate standard errors. An algorithm to perform the maximization will be discussed in Chapter 8. We fit the model in R like this:

```
lmod <- glm(chd ~ height + cigs, family = binomial, wcgs)
```

This model is a special case of wider class called a generalized linear model (GLM) so we use the `glm` fitting function and specify the distribution of the response, in this case binomial, which identifies the member of this wider family of models. The model formula is specified in the same way as for linear models. A binary response is a special case of the binomial and requires that the response have only two levels. By default, the first level alphabetically will be associated with  $y = 0$  and the second with  $y = 1$ . If you want to change this, use the `relevel` command.

We can examine the standard summary output:

```
summary(lmod)
```

Call:

```
glm(formula = chd ~ height + cigs, family = binomial, data = wcgs)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-----	----	--------	----	-----

```
-1.004 -0.443 -0.363 -0.350 2.436
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-4.50161	1.84186	-2.44	0.015
height	0.02521	0.02633	0.96	0.338
cigs	0.02313	0.00404	5.72	1e-08

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 1781.2 on 3153 degrees of freedom
Residual deviance: 1749.0 on 3151 degrees of freedom
AIC: 1755
```

Number of Fisher Scoring iterations: 5

A shorter version is available in the `faraway` package as `sumary` (a shortened `summary`). Incidentally, we don't need to keep loading the `faraway` package within a single R session. We have included it here just as a reminder of where the `sumary` function comes from. If you get a "not found" message, it means you forgot to load this package.

```
library(faraway)
sumary(lmod)

Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.50161 1.84186 -2.44 0.015
height 0.02521 0.02633 0.96 0.338
cigs 0.02313 0.00404 5.72 1e-08
```

n = 3154 p = 3

Deviance = 1749.049 Null Deviance = 1781.244 (Difference = 32.195)

You can always use the `summary` version if you prefer but henceforth we will use this shorter `sumary`. Let's start with the interpretation of the regression coefficients:  $\hat{\beta}_0 = -4.50$ ,  $\hat{\beta}_1 = 0.025$  and  $\hat{\beta}_2 = 0.023$ . We can compute the probability of heart disease given the values of the predictors. We can vary the height for fixed levels of cigarette consumption — nonsmoker and 20 a day. For this we need to extract the coefficients:

```
(beta <- coef(lmod))

(Intercept) height cigs
-4.501614 0.025208 0.023127
```

We then construct the plot using a point plotting symbol and add the fitted curves:

```
plot(jitter(y,0.1) ~ jitter(height), wccgs, xlab="Height", ylab="Heart
→ Disease",pch=".")
curve(ilogit(beta[1] + beta[2]*x + beta[3]*0),add=TRUE)
curve(ilogit(beta[1] + beta[2]*x + beta[3]*20),add=TRUE,lty=2)
```

In the first panel of Figure 2.5, we see that the probability of heart disease increases slightly with height. We also see a somewhat higher risk for smokers. We produce a similar plot varying the cigarette consumption rather than the height:

```
plot(jitter(y,0.1) ~ jitter(cigs), wccgs, xlab="Cigarette Use", ylab="

→ Heart Disease",pch=".")
curve(ilogit(beta[1] + beta[2]*60 + beta[3]*x),add=TRUE)
curve(ilogit(beta[1] + beta[2]*78 + beta[3]*x),add=TRUE,lty=2)
```

In the second panel of Figure 2.5, we see how the probability of heart disease increases with the amount of smoking.

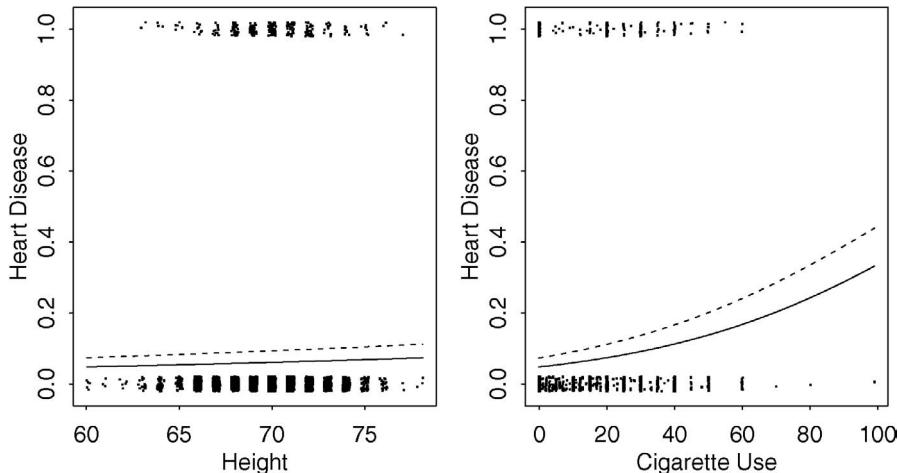


Figure 2.5 *Predicted probability of heart disease as height and cigarette consumption vary. In the first panel, the solid line represents a nonsmoker, while the dashed line is a pack-a-day smoker. In the second panel, the solid line represents a very short man (60 in. tall) while the dashed line represents a very tall man (78 in. tall.)*

**Interpreting Odds:** Odds are an alternative scale to probability for representing chance. They arose as a way to express the payoffs for bets. An *evens* bet means that the winner gets paid an equal amount to that staked. A 3–1 *against* bet would pay \$3 for every \$1 bet, while a 3–1 *on* bet would pay only \$1 for every \$3 bet. If these bets are *fair* in the sense that a bettor would break even in the long-run average, then we can make a correspondence to probability. Let  $p$  be the probability and  $o$  be the odds, where we represent 3–1 against as  $1/3$  and 3–1 on as 3, then the following relationship holds:

$$\frac{p}{1-p} = o \quad \text{or} \quad p = \frac{o}{1+o}$$

One mathematical advantage of odds is that they are unbounded above, which makes them more convenient for some modeling purposes.

Odds also form the basis of a subjective assessment of probability. Some probabilities are determined from considerations of symmetry or long-term frequencies, but such information is often unavailable. Individuals may determine their subjective probability for events by considering what odds they would be prepared to offer on the outcome. Under this theory, other potential persons would be allowed to place bets for or against the event occurring. Thus the individual would be forced to make an honest assessment of probability to avoid financial loss.

If we have two covariates  $x_1$  and  $x_2$ , then the logistic regression model is:

$$\log(\text{odds}) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

or

$$\text{odds} = e^{\beta_0} \cdot e^{\beta_1 x_1} \cdot e^{\beta_2 x_2}$$

Now  $\beta_1$  can be interpreted as follows: a unit increase in  $x_1$  with  $x_2$  held fixed increases the log-odds of success by  $\beta_1$  or increases the odds of success by a factor of  $e^{\beta_1}$ . So the exponentiated coefficients are more useful:

```
exp(beta)
(Intercept)      height       cigs
0.011091     1.025528    1.023397
```

We can say that the odds of heart disease increase by 2.6% with each additional inch in height and by 2.3% with each additional cigarette smoked per day. Note that  $\exp x \approx 1 + x$  for small values of  $x$ . We observe that  $\hat{\beta}_2 = 0.023$  so the 2.3% increase in odds due to smoking a cigarette a day could have been quickly estimated from the original model output without further computation. This approximation is useful for quick intuitions.

It is more natural to compute the effect of a pack a day (20 cigarettes):

```
exp(beta[3]*20)
cigs
1.5881
```

So we have 59% increase in the odds of heart disease due to smoking. Of course, the usual interpretational difficulties regarding causation apply as in standard regression. We cannot conclude, just from this study, that smoking causes an increase in the risk of heart disease.

Most people are more comfortable with a probability scale so it is a good idea to compute the difference in the predicted probabilities as a predictor is changed over a sensible range as we have done above. Odds are more convenient from a mathematical perspective and do allow us to express the effects of smoking in this example in a compact way. One disadvantage of odds ratios is that they only express the relative difference. Although smoking is associated with a strongly increased risk of heart disease, the absolute probabilities are still not large.

An alternative notion to odds is relative risk. Suppose the probability of “success” in the presence of some condition is  $p_1$  and  $p_2$  in its absence. The relative risk is  $p_1/p_2$ . For example, the predicted probabilities of getting heart disease for a 68in.-tall man who does not smoke and who smokes 20 a day are, respectively:

```
c(ilogit(sum(beta*c(1, 68, 20))), ilogit(sum(beta*c(1, 68, 0))))
[1] 0.089079 0.058004
```

The relative risk is then:

```
ilogit(sum(beta*c(1, 68, 20)))/ilogit(sum(beta*c(1, 68, 0)))
[1] 1.5357
```

The relative risk of 1.54 is not far different from the odds ratio of 1.59. For low probability outcomes, the relative risk and the odds ratio will be very similar, but for larger probabilities, there may be substantial differences.

## 2.3 Inference

Consider two models, a larger model with  $l$  parameters and likelihood  $L_L$  and a smaller model with  $s$  parameters and likelihood  $L_S$  where the smaller model represents a subset (or more generally a linear subspace) of the larger model. Likelihood

methods suggest the likelihood ratio statistic:

$$2 \log \frac{L_L}{L_S} \quad (2.1)$$

as an appropriate test statistic for comparing the two models. Now suppose we choose a saturated larger model — such a model typically has as many parameters as cases and has fitted values  $\hat{p}_i = y_i$ . The test statistic becomes:

$$D = -2 \sum_{i=1}^n \hat{p}_i \text{logit}(\hat{p}_i) + \log(1 - \hat{p}_i)$$

where  $\hat{p}_i$  are the fitted values from the smaller model.  $D$  is called the *deviance* and is useful in making hypothesis tests to compare models.

In other examples of GLMs, the deviance is a measure of how well the model fit the data but in this case,  $D$  is just a function of the fitted values  $\hat{p}$  so it cannot be used for that purpose. Other methods must be used to judge goodness of fit for binary data — for example, the Hosmer-Lemeshow test described in Section 2.6.

In the summary output previously, we had:

```
Deviance = 1749.049 Null Deviance = 1781.244 (Difference = 32.195)
```

The Deviance is the deviance for the current model while the Null Deviance is the deviance for a model with no predictors and just an intercept term.

We can use the deviance to compare two nested models. The test statistic in (2.1) becomes  $D_S - D_L$ . This test statistic is asymptotically distributed  $\chi^2_{l-s}$ , assuming that the smaller model is correct and the distributional assumptions hold. For example, we can compare the fitted model to the null model (which has no predictors) by considering the difference between the residual and null deviances. For the heart disease example, this difference is 32.2 on two degrees of freedom (one for each predictor). Hence, the  $p$ -value for the test of the hypothesis that at least one of the predictors is related to the response is:

```
1-pchisq(32.2, 2)
```

```
[1] 1.0183e-07
```

Since this value is so small, we are confident that there is some relationship between the predictors and the response. Note that the expected value of a  $\chi^2$ -variate with  $d$  degrees of freedom is simply  $d$  so we knew the  $p$ -value would be small before even calculating it.

We can test the individual predictors by fitting models that drop these predictors and computing the difference in the deviance observed. We test the significance of height in the model as follows:

```
lmodc <- glm(chd ~ cigs, family = binomial, wcls)
```

```
anova(lmodc, lmod, test="Chi")
```

```
Analysis of Deviance Table
```

	Model 1: chd ~ cigs	Model 2: chd ~ height + cigs			
	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	3152	1750			
2	3151	1749	1	0.92	0.34

The analysis of deviance table displays all the information. We see that height is not significant in a model that already includes cigarette consumption. We can readily test all the predictors in the model using the `drop1` function:

```
drop1(lmod,test="Chi")
```

Single term deletions

```
Model:
chd ~ height + cigs
      Df Deviance AIC    LRT Pr(>Chi)
<none>     1749 1755
height   1     1750 1754  0.92     0.34
cigs     1     1780 1784 31.07  2.5e-08
```

An alternative to this test is the  $z$ -value, which is  $\hat{\beta}/se(\hat{\beta})$ , which is approximately normally distributed. Recalling the output from before:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-4.50161	1.84186	-2.44	0.015
height	0.02521	0.02633	0.96	0.338
cigs	0.02313	0.00404	5.72	1e-08

In this case, the outcomes are similar to previous tests but not identical. This is in contrast to the normal (Gaussian) linear model where the two would be identical. In this particular example, there is no practical difference, but in some cases, especially with sparse data, the standard errors can be overestimated and so the  $z$ -value is too small and the significance of an effect could be missed. This is known as the Hauck–Donner effect — see Hauck and Donner (1977). So the deviance-based test is preferred.

Confidence intervals for the regression parameters can be constructed using normal approximations for the parameter estimates. A  $100(1 - \alpha)\%$  confidence interval for  $\beta_i$  would be:

$$\hat{\beta}_i \pm z^{\alpha/2} se(\hat{\beta}_i)$$

where  $z^{\alpha/2}$  is a quantile from the normal distribution. Thus a 95% confidence interval for  $\beta_1$  in our model would be:

```
0.02521 + c(-1,1) * 1.96 * 0.02633
[1] -0.026397  0.076817
```

A better way is to construct a profile likelihood-based confidence interval:

```
confint(lmod)
```

```
Waiting for profiling to be done...
      2.5 %    97.5 %
(Intercept) -8.134755 -0.912970
height       -0.026199  0.077028
cigs         0.015149  0.031005
```

The profile likelihood method is generally preferable for the same Hauck–Donner reasons discussed above although it is more work to compute.

## 2.4 Diagnostics

Regression diagnostics are useful in checking the assumptions of the model and in identifying any unusual points. As with linear models, residuals are the most important means of determining how well the data fits the model and where any changes or

improvements might be advisable. We can compute residuals as a difference between observed and fitted values. There are two kinds of fitted (or predicted) values:

```
linpred <- predict(lmod)
predprob <- predict(lmod, type="response")
```

The former is the predicted value in the linear predictor scale,  $\eta$ , while the latter is the predicted probability  $p = \text{logit}^{-1}(\eta)$ . Here are the first few values and a confirmation of the relationship between them:

```
head(linpred)
```

2001	2002	2003	2004	2005	2006
-2.0833	-2.2745	-2.7623	-2.3249	-2.2745	-2.6867

```
head(predprob)
```

2001	2002	2003	2004	2005	2006
0.110734	0.093255	0.059397	0.089079	0.093255	0.063766

```
head(ilogit(linpred))
```

2001	2002	2003	2004	2005	2006
0.110734	0.093255	0.059397	0.089079	0.093255	0.063766

We compute the *raw residuals* as  $y - \hat{p}$ :

```
rawres <- wcgs$y - predprob
```

These can also be obtained as `residuals(lmod, type="response")`. Following the standard practice for diagnostics in linear models, we plot the residuals against the fitted values:

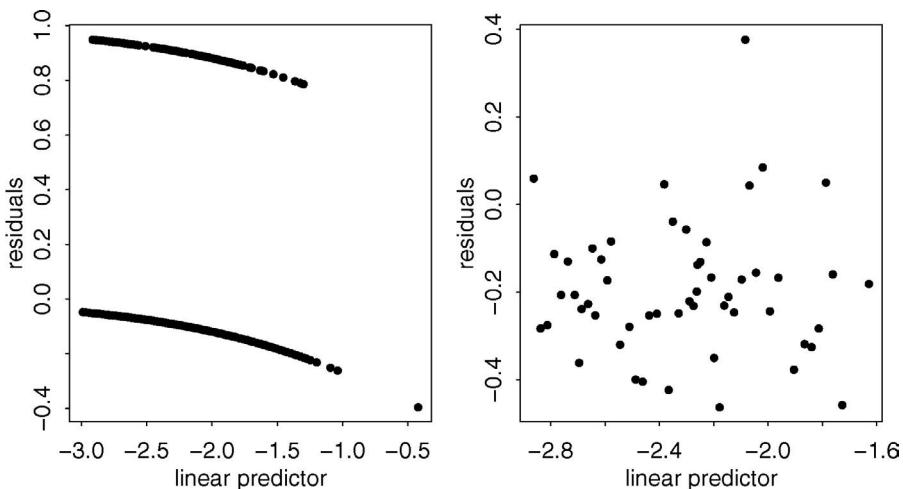


Figure 2.6 The panel on the left shows the raw residuals and linear predictor. The two lines are due to the binary response. The panel on the right shows the binned version of the plot.

```
plot(rawres ~ linpred, xlab="linear predictor", ylab="residuals")
```

The plot, as seen in the first panel of Figure 2.6, is not very helpful. Because  $y = 0$  or  $1$ , the residual can take only two values given a fixed linear predictor. The upper line in the plot corresponds to  $y = 1$  and the lower line to  $y = 0$ . We gain no insight into the fit of the model. We have chosen to plot the linear predictor rather than the predicted probability on the horizontal axis because the former provides a better spacing of the points in this direction.

Another problem with the raw residuals is that we do not expect them to have equal variance. Binary variance is given by  $p(1 - p)$  so there will be more variation from probabilities in the midrange. Some form of standardization is desirable. In the standard linear model, the residual sum of squares is given by  $\sum_i \hat{\epsilon}_i^2$ . For a logistic regression model, the equivalent quantity is the deviance. The *deviance residuals*,  $r_i$ , are defined by analogy using  $\sum r_i^2 = \text{Deviance}$ . We ensure these residuals have the right sign by defining:

$$r_i = \text{sign}(y_i - \hat{p}_i) \sqrt{r_i^2}$$

The deviance residuals are the default choice and are given by `residuals(lmod)`.

We now construct a more useful residuals plot by grouping the residuals into bins where the bins are based on similar predictor values. The choice of the number of bins depends on the size of the dataset. We choose 100 bins so that we have roughly 30 observations per bin. Some effort is required to construct this plot. The `dplyr` package of Wickham and Francois (2015) is useful for this task. First, we add the residuals and linear predictor into the data frame.

```
library(dplyr)
wcgs <- mutate(wcgs, residuals=residuals(lmod), linpred=predict(lmod))
The next step is to create the bins:
gdf <- group_by(wcgs, cut(linpred, breaks=unique(quantile(linpred,
  ↪ (1:100)/101))))
```

The `cut` function divides a variable into a factor with levels defined by breakpoints given by `breaks`. We would like to have a roughly equal number of points per bin so we divide the range of the linear predictor based on quantiles. As it happens, there are many tied values in this variable so we insist on only the `unique` breakpoints.

Now within each of these bins, we ask for the means of the residuals and linear predictors:

```
diagdf <- summarise(gdf, residuals=mean(residuals), linpred=mean(
  ↪ linpred))
```

The residuals plot can be seen in the second panel of Figure 2.6 as:

```
plot(residuals ~ linpred, diagdf, xlab="linear predictor")
```

The deviance residuals are not constrained to have mean zero so the mean level in the plot is not of interest. We see an even variation as the linear predictor varies so this plot reveals no inadequacy in the model. A similar plot can be constructed using the `binnedplot` function of the `arm` package of Gelman and Hill (2006).

We should also plot the binned residuals against the predictors. We make a similar calculation:

```
gdf <- group_by(wcgs, height)
diagdf <- summarise(gdf, residuals=mean(residuals))
ggplot(diagdf, aes(x=height,y=residuals)) + geom_point()
```

The height predictor only takes a limited number of integer values so we just group by height directly. The plot is shown in the first panel of Figure 2.7 where we see nothing remarkable except perhaps for a large residual for a height of 77 in. Let's look at the data for this height:

```
filter(wcgs, height==77) %>% select(height, cigs, chd, residuals)
  height cigs chd residuals
1      77    0  no   -0.38579
2      77    0  yes   2.29566
3      77    5  no   -0.40785
```

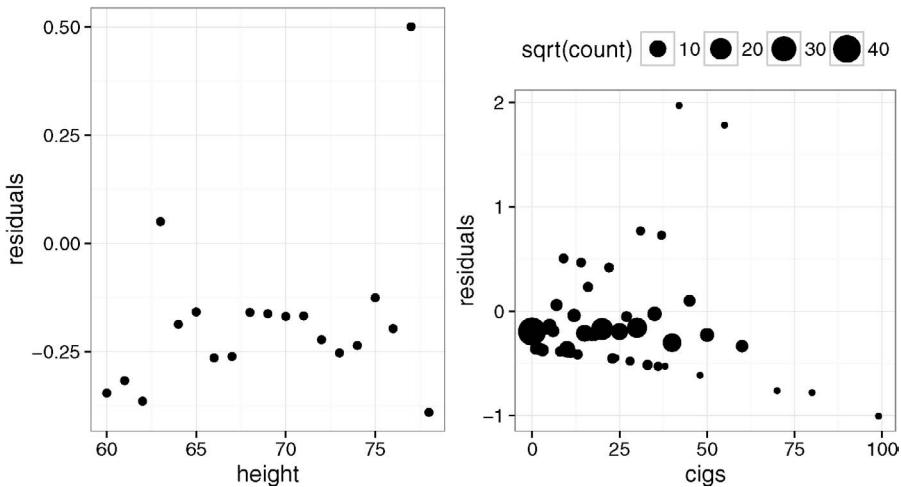


Figure 2.7 *Binned residuals plots for the predictors.*

We use some features from the `dplyr` package. The `filter` command picks out the subset of the rows of the data with height of 77 in. The output is piped to the next command using the operator `%>%`. The `select` command picks out the variables we want to see. From this we see that this bin consists of only three cases, of which one is a man who has heart disease. Hence, this point on the residual plot is not exceptional.

We now construct the corresponding plot for cigarette consumption as seen in the second panel of Figure 2.7.

```
group_by(wcgs, cigs) %>% summarise(residuals=mean(residuals), count=n()
  ↪ () %>% ggplot(aes(x=cigs, y=residuals, size=sqrt(count))) +
  ↪ geom_point())
```

We use the `%>%` operator to chain together the intermediate steps. This time we keep a count of the number of residuals in each bin using the `n()` function within `summarise`. We use this count to control the size of the plotted point. We take square roots because the SD is proportional to the square root of the sample size so this gives the appropriate visual impression. Many men in the sample do not smoke at all so there is a large bin for zero consumption. Other round numbers such as 20 a day are similarly highlighted. We see the more unusual points on the display corresponding to small bin sizes and so they can be discounted in interpreting the plot. Focusing on the larger bins, we see no evidence of lack of fit or a need to transform the predictor.

We now turn attention to the detection of unusual points. A QQ plot of the residuals is standard practice in checking linear models. Such a plot is shown in the first panel of Figure 2.8.

```
qqnorm(residuals(lmod))
```

We see that the plot is very far from the desired linear relationship. We see two clusters of points corresponding to  $y = 0$  and  $y = 1$ . But there is no reason to expect these residuals to be normally distributed so this does not raise any concern. The

largest residuals will be observed in cases where  $y_i = 1$  when  $\hat{p}_i$  is small and where  $y_i = 0$  when  $\hat{p}_i$  is close to 1. These represent occasions when something unexpected happened. But even in such cases, the residual cannot become particularly large.

We can detect cases which are unusual in the predictor space by examining the leverages just as in standard linear models. These are best examined using a half-normal plot as seen in the second panel of Figure 2.8.

```
halfnorm(hatvalues(lmod))
```

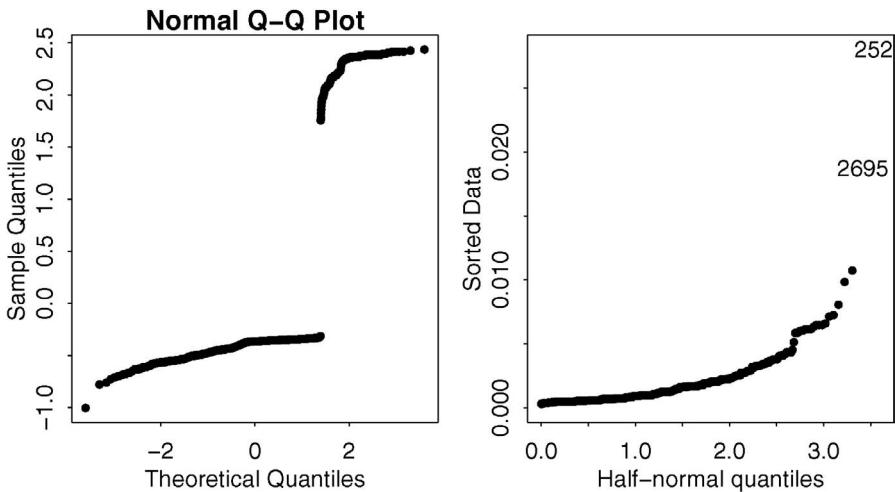


Figure 2.8 A *QQ* plot of the deviance residuals is shown on the left and a half-normal plot of the leverages is shown on the right.

The two outlying points can be identified:

```
filter(wcgs, hatvalues(lmod) > 0.015) %>% select(height, cigs, chd)
height cigs chd
1      71   99  no
2      64   80  no
```

These are the two men with the very highest cigarette consumption which can be seen near the top of Figure 2.3. Given the relatively large size of the dataset and the fact that these two points are not particularly extreme, we are not concerned.

## 2.5 Model Selection

The analysis thus far has used only two of the predictors available but we might construct a better model for the response if we used some of the other predictors. We might find that not all these predictors are helpful in explaining the response. We would like to identify a subset of the predictors that model the response well without including any superfluous predictors.

We could use the inferential methods to construct hypothesis tests to compare various candidate models and use this as a mechanism for choosing a model. Back-

ward elimination is one such method which is relatively easy to implement. The method proceeds sequentially:

1. Start with the full model including all the available predictors. We can add derived predictors formed from transformations or interactions between two or more predictors.
2. Compare this model with all the models consisting of one less predictor. Compute the  $p$ -value corresponding to each dropped predictor. The `drop1` function in R can be used for this purpose.
3. Eliminate the term with largest  $p$ -value that is greater than some preset critical value, say 0.05. Return to the previous step. If no such term meets this criterion, stop and use the current model.

Thus predictors are sequentially eliminated until a final model is settled upon. Unfortunately, this is an inferior procedure. Although the algorithm is simple to use, it is hard to identify the problem to which it provides a solution. It does not identify the best set of predictors for predicting future responses. It is not a reliable indication of which predictors are the best explanation for the response. Even if one believes the fiction that there is a true model, this procedure would not be best for identifying such a model.

The Akaike information criterion (AIC) is a popular way of choosing a model — see Section A.3 for more. The criterion for a model with likelihood  $L$  and number of parameters  $q$  is defined by

$$AIC = -2 \log L + 2q$$

We select the model with the smallest value of AIC among those under consideration. Any constant terms in the definition of log-likelihood can be ignored when comparing different models that will have the same constants. For this reason we can use  $AIC = \text{deviance} + 2q$ .

We would like to consider models consisting of all possible subsets of the available predictors. For large values of  $q$ , there may be a very large number of such subsets so a sequential search through this set may be necessary to reduce the computation. Such a sequential search is provided by the `step` command:

```
wcgs$bmi <- with(wcgs, 703*wcgs$weight/(wcgs$height^2))
lmod <- glm(chd ~ age + height + weight + bmi + sdp + dbp + chol +
             → dibep + cigs + arcus, family=binomial, wcgs)
lmodr <- step(lmod, trace=0)
summary(lmodr)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-15.95760	2.28608	-6.98	2.9e-12
age	0.06159	0.01240	4.97	6.8e-07
height	0.05016	0.02782	1.80	0.071
bmi	0.06038	0.02660	2.27	0.023
sdp	0.01773	0.00415	4.27	2.0e-05
chol	0.01071	0.00153	7.01	2.4e-12
dibepB	0.65762	0.14590	4.51	6.6e-06
cigs	0.02104	0.00426	4.94	8.0e-07
arcuspresent	0.21100	0.14372	1.47	0.142

n = 3140 p = 9

Deviance = 1569.325 Null Deviance = 1769.171 (Difference = 199.846)

We have derived a body mass index (BMI) variable from height and weight. The factor of 703 is required to convert to the metric units used by BMI. The `trace=0` argument prevents the printing of the intermediate stages of the search. Omit this if you would like to see these. In this case, only two predictors are dropped from the full set: weight and diastolic blood pressure.

The urge to houseclean may tempt us to use model selection methods such as AIC. If the purpose of the model is to predict future observations, then AIC-based model selection may provide a decent choice for this purpose. There are other criteria, similar to AIC, that can also be used. On the other hand, we may be more interested in explaining the response. For example, we may wish to know what the risk factors are for coronary heart disease. We might be tempted to think that since diastolic blood pressure has been dropped from the model that it has no relation to the chance of heart disease. Even so, considered alone:

```
drop1(glm(chd ~ dbp, family=binomial, wcds), test="Chi")
```

Single term deletions

```
Model:
chd ~ dbp
      Df Deviance AIC LRT Pr(>Chi)
<none>      1752 1756
dbp     1      1781 1783 29.6  5.5e-08
```

We see that diastolic blood pressure is related to the chance of heart disease. For this and other reasons, model selection is not a magical sword that can cut through the knotty problem of determining the risk factors for heart disease let alone what might be causing it.

## 2.6 Goodness of Fit

As mentioned earlier, we cannot use the deviance for a binary response GLM as a measure of fit. We can use diagnostic plots of the binned residuals to help us identify inadequacies in the model but these cannot tell us whether the model fits or not. Even so the process of binning can help us develop a test for this purpose. We divide the observations up into  $J$  bins based on the linear predictor. Let the mean response in the  $j^{th}$  bin be  $y_j$  and the mean predicted probability be  $\hat{p}_j$  with  $m_j$  observations within the bin. We compute these values:

```
wcds <- na.omit(wcds)
wcds <- mutate(wcds, predprob=predict(lmod, type="response"))
gdf <- group_by(wcds, cut(linpred, breaks=unique(quantile(linpred,
  ~ (1:100)/101))))
hldf <- summarise(gdf, y=sum(y), ppred=mean(predprob), count=n())
```

There are a few missing values in the data. The default method is to ignore these cases. The `na.omit` command drops these cases from the data frame for the purposes of this calculation. We use the same method of binning the data as for the residuals but now we need to compute the number of observed cases of heart disease and total observations within each bin. We also need the mean predicted probability within each bin.

When we make a prediction with probability  $p$ , we would hope that the event occurs in practice with that proportion. We can check that by plotting the observed

proportions against the predicted probabilities as seen in Figure 2.9. For a well-calibrated prediction model, the observed proportions and predicted probabilities should be close.

```
hldf <- mutate(hldf, se.fit=sqrt(ppred*(1-ppred)/count))
ggplot(hldf,aes(x=ppred,y=y/count,ymin=y/count-2*se.fit,ymax=y/count+2
  ↪ *se.fit))+geom_point() + geom_linerange(color=grey(0.75)) + geom_
  ↪ abline(intercept=0,slope=1)+xlab("Predicted Probability") + ylab(
  ↪ "Observed Proportion")
```

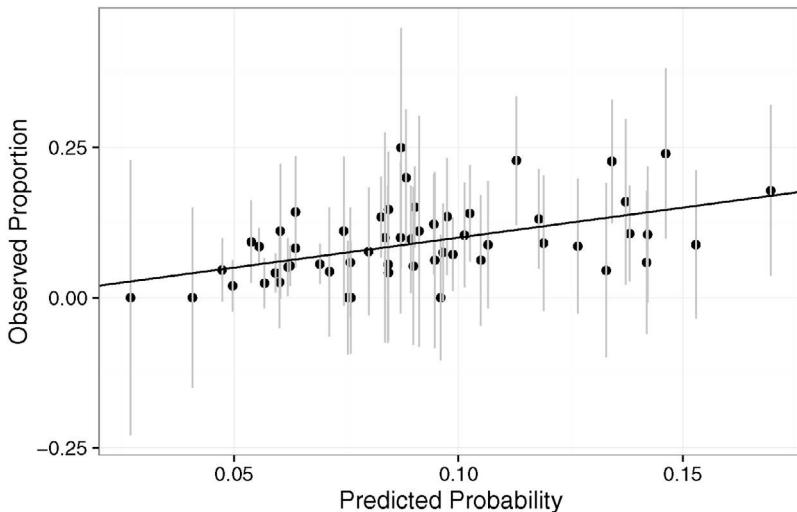


Figure 2.9 Binned predicted probabilities and observed proportions for the heart disease model.

Although we can see there is some variation, there is no consistent deviation from what is expected. We have computed approximate 95% confidence intervals using the binomial variation. The line passes through most of these intervals confirming that the variation from the expected is not excessive.

The Hosmer-Lemeshow statistic formalizes this assessment and is defined as:

$$X_{HL}^2 = \sum_{j=1}^J \frac{(y_j - m_j \hat{p}_j)^2}{m_j \hat{p}_j (1 - \hat{p}_j)}$$

This statistic has an approximate  $\chi^2$  distribution with  $J - 1$  degrees of freedom. We have some freedom to decide on the binning. We need sufficient observations per bin to ensure the accuracy of the  $\chi^2$  approximation yet not so few bins that the fit is barely tested.

```
hlstat <- with(hldf, sum( (y-count*ppred)^2/(count*ppred*(1-ppred))))
```

```
[1] 63.212 56.000
```

The  $p$ -value is then given by

```
1-pchisq(63.212, 56-1)
```

```
[1] 0.20898
```

Since the  $p$ -value is moderate, we detect no lack of fit. In marginal cases, it is worth experimenting with differing numbers of bins to check the robustness of the conclusion.

We might want a method for assessing the quality of predictions that does not depend on binning methods, as these involve some arbitrary choices about numbers and spacing of bins. *Scoring* methods were developed, initially by weather forecasters, to assess predictions. The most popular method is logarithmic scoring which is given as  $y * \log(\hat{p}) + (1 - y) * \log(1 - \hat{p})$ . So if the event  $y = 1$  occurs, we score  $\log \hat{p}$  and score  $\log(1 - \hat{p})$  if it does not. Better prediction methods have higher scores. Notice if we predict something is certain to happen ( $\hat{p} = 1$ ) but it does not occur then we score  $-\infty$ , so this method of scoring does not encourage certainty.

The model can be used to predict the outcome for each man in the dataset. Sometimes it is natural to think of the model as classifying observations. When  $\hat{p}_i < 0.5$  we classify the case into no heart disease but when  $\hat{p}_i \geq 0.5$  we put the case into the yes category. Of course, for the observed cases we already know the outcome but for future unknown cases, this classification may be helpful.

```
wcgsm <- mutate(wcgsm, predout=ifelse(predprob < 0.5, "no", "yes"))
xtabs(~ chd + predout, wcgsm)
predout
chd      no   yes
no    2882     3
yes    253     2
```

We see how the classifications match with the observed outcomes. The correct classification rate is:

```
(2882+2) / (2882+3+253+2)
[1] 0.91847
```

So the error or misclassification rate is about 8%. This can be a helpful way of judging the fit of the model although it can conceal some important variation. Suppose we were to use this model prospectively to predict the outcomes of men with given characteristics. Perhaps we might recommend certain treatments based on these predictions. For the men who will not develop heart disease, what fraction are predicted to not develop the disease? This is known as the *specificity* of the test. We calculate this as  $2882/(2882 + 3) = 0.999$ , which is very high. In contrast, the proportion of those who will develop heart disease that are correctly identified by the model is called the *sensitivity*. Here this is  $2/(253 + 2) = 0.00784$ , which is extremely low. Our prediction process is very unlikely to detect which men will develop heart disease. We see that the overall error rate of 8% conceals important information if this procedure were to be used as a diagnostic test.

We can improve the sensitivity of the test by dropping the 0.5 threshold we used for the classification. We compute how the sensitivity and specificity change as the threshold changes for a range of values up to 0.5:

```
thresh <- seq(0.01, 0.5, 0.01)
Sensitivity <- numeric(length(thresh))
Specificity <- numeric(length(thresh))
for(j in seq(along=thresh)){
  pp <- ifelse(wcgsm$predprob < thresh[j], "no", "yes")
  xx <- xtabs(~ chd + pp, wcgsm)
  Specificity[j] <- xx[1,1]/(xx[1,1]+xx[1,2])
```

```
Sensitivity[j] <- xx[2,2]/(xx[2,1]+xx[2,2])
}
```

We can now plot how these change as a function of the threshold:

```
matplot(thresh, cbind(Sensitivity, Specificity), type="l", xlab="Threshold
    ↪ ", ylab="Proportion", lty=1:2)
```

The plot, shown in the first panel of Figure 2.10, demonstrates how the sensitivity falls but the specificity rises as the threshold is increased. The costs of the two possible kinds of errors in such diagnostic tests are rarely equal. We would need to balance the costs and risks of treating a man who will not get heart disease against the costs of failing to treat a man who will. This is a difficult calculation but at least we now have an understanding of the properties of the test. An alternative and popular way to display the same information is the *receiver operating characteristic* (ROC) curve which plots the sensitivity (true positive rate) against one minus the specificity (false positive rate) as seen in the second panel of Figure 2.10.

```
plot(1-Specificity, Sensitivity, type="l")
abline(0,1, lty=2)
```

A useless test (one that decides at random) would fall on the  $y = x$  line whereas a very good test has a curve that is pulled out into the upper left corner. The area under the ROC curve can be used as a measure of the performance of the test and might be used to compare different tests.

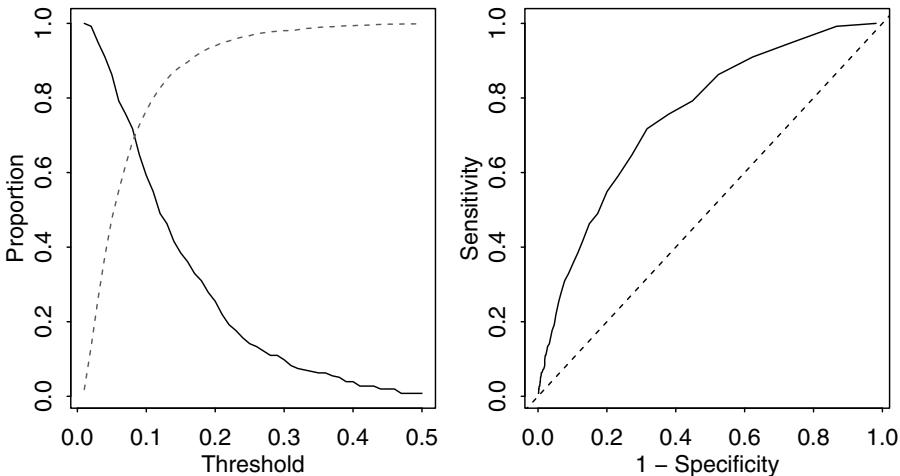


Figure 2.10 *Sensitivity and specificity for the heart disease model plotted as a function of the probability threshold on the left and as the receiver operating characteristic curve on the right.*

The proportion of variance explained or  $R^2$  is a popular measure of fit for normal linear models. We might consider applying the same concept to binomial regression models by using the proportion of deviance explained. However, a better statistic is

due to Nagelkerke (1991):

$$R^2 = \frac{1 - (\hat{L}_0/\hat{L})^{2/n}}{1 - \hat{L}_o^{2/n}} = \frac{1 - \exp((D - D_{null})/n)}{1 - \exp(-D_{null}/n)}$$

where  $n$  is the number of binary observations and  $\hat{L}_0$  is the maximized likelihood under the null. The numerator can be seen as a ratio of the relative likelihood with the  $1/n$  power having the effect of a geometric mean on the observations. The denominator simply normalizes so that  $0 \leq R^2 \leq 1$ . For example, for the current model, the  $R^2$  is:

```
lmodr <- glm(chd ~ age + height + bmi + sdp + chol + dibep + cigs +
  ↪ arcus, family=binomial, wcols)
(1-exp((lmodr$dev-lmodr$null)/3140)) / (1-exp(-lmodr$null/3140))
[1] 0.14315
```

This gives the impression of a fairly poor fit when judged from the experience of linear models. However, this is misleading. In a standard linear model, it is possible for the observed and fitted values to be very close showing a strong fit and an  $R^2$  close to one. This simply isn't possible for binary response models given the natural variation. It is quite common to see low values of Naglekerke's and other  $R^2$  substitutes even when the model is good. For this reason, it may be best to avoid this statistic except perhaps for the purpose of comparing compatible models.

## 2.7 Estimation Problems

Estimation of the logistic regression model using the Fisher scoring algorithm, described in Section 8.2, is usually fast. However, difficulties can sometimes arise. When convergence fails, it is sometimes due to a problem exhibited by the following dataset. We take a subset of the famous Fisher Iris data to consider only two of the three species of Iris and use only two of the potential predictors:

```
library(dplyr)
irisr <- filter(iris, Species != "virginica") %>% select(Sepal.Width,
  ↪ Sepal.Length, Species)
```

We plot the data using a different shape of plotting symbol for the two species:

```
(p <- ggplot(irisr, aes(x=Sepal.Width, y=Sepal.Length, shape=Species))
  ↪ +geom_point())
```

We now fit a logistic regression model to see if the species can be predicted from the two sepal dimensions.

```
lmod <- glm(Species ~ Sepal.Width + Sepal.Length, family=binomial,
  ↪ irisr)
```

Warning messages:

```
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

We see that there were problems with the convergence. A look at the summary reveals further evidence:

```
summary(lmod)

Estimate Std. Error z value Pr(>|z|)
(Intercept) -361     195973      0      1
Sepal.Width   -110      55362      0      1
Sepal.Length    132      64577      0      1
```

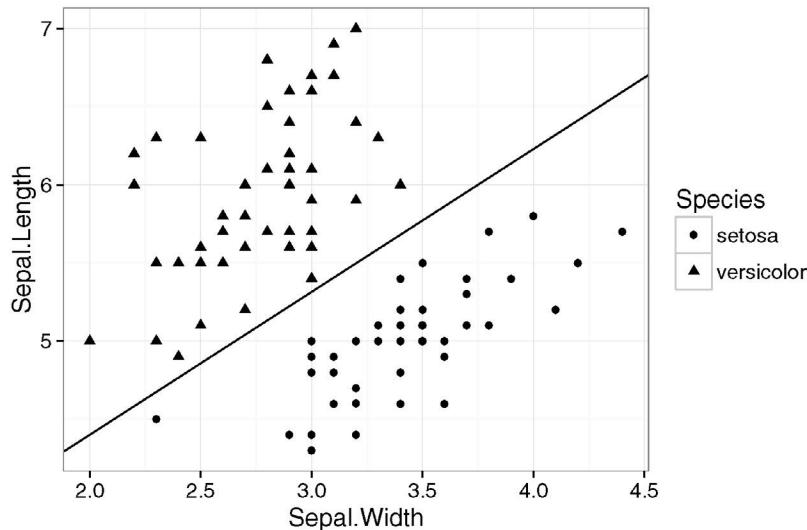


Figure 2.11 Two species of iris by sepal length and width. The line is computed from the bias-reduced GLM fit.

$n = 100 \ p = 3$

Deviance = 0.000 Null Deviance = 138.629 (Difference = 138.629)

Notice that the residual deviance is zero indicating a perfect fit and yet none of the predictors are significant due to the high standard errors. A look at the data reveals the reason for this. We see that the two groups are *linearly separable* so that a perfect fit is possible. We suffer from an embarrassment of riches in this example — we can fit the data perfectly. Unfortunately, this results in unstable estimates of the parameters and their standard errors and would (probably falsely) suggest that perfect predictions can be made. An alternative fitting approach might be considered in such cases called *exact logistic regression*. See Cox (1970) or Mehta and Patel (1995). Implementations can be found in the *elrm* and *logistix* packages in R.

An alternative to exact methods is the bias reduction method of Firth (1993) and implemented in the *brglm* package of Kosmidis (2013). For the maximum likelihood estimate (MLE),  $E\hat{\beta} \neq \beta$  and indeed a sensible unbiased estimator would be difficult to obtain. Firth's method removes the  $O(1/n)$  term from the asymptotic bias of estimated coefficients. These estimates have the advantage of always being finite:

```
library(brglm)
bmod <- brglm(Species ~ Sepal.Width + Sepal.Length, family=binomial,
  ↪ irisr)
summary(bmod)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-24.51	12.49	-1.96	0.0498
Sepal.Width	-8.90	2.75	-3.24	0.0012
Sepal.Length	9.73	3.33	2.92	0.0035

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 130.638 on 99 degrees of freedom
Residual deviance: 3.323 on 97 degrees of freedom
```

We can see that this results in significant predictors which we expect given Figure 2.11. We compute and display the line corresponding to a predicted probability of 1/2.

```
p + geom_abline(intercept=(0.5+24.51)/9.73, slope=8.9/9.73)
```

Instability in parameter estimation will also occur in datasets that approach linear separability. Care will also be needed in such cases.

**Further Reading:** See books by Collett (2003), Hosmer and Lemeshow (2013), Cox (1970), Harrell (2001), Menard (2002), Christensen (1997), Kleinbaum and Klein (2002) and Hilbe (2009).

## Exercises

1. The dataset `wbca` comes from a study of breast cancer in Wisconsin. There are 681 cases of potentially cancerous tumors of which 238 are actually malignant. Determining whether a tumor is really malignant is traditionally determined by an invasive surgical procedure. The purpose of this study was to determine whether a new procedure called fine needle aspiration, which draws only a small sample of tissue, could be effective in determining tumor status.
  - (a) Plot the relationship between the classification and `BNucl`.
    - i. Explain why  
`plot(Class ~ BNucl, wbca)`  
 does not work well.
    - ii. Create a factor version of the response and produce a version of the first panel of Figure 2.1. Comment on the shape of the boxplots.
    - iii. Produce a version of the second panel of Figure 2.1. What does this plot say about the distribution?
    - iv. Produce a version of the interleaved histogram shown in Figure 2.2 and comment on the distribution.
  - (b) Produce a version of Figure 2.3 for the predictors `BNucl` and `Thick`. Produce an alternative version with only one panel but where the two types are plotted differently. Compare the two plots and describe what they say about the ability to distinguish the two types using these two predictors.
  - (c) Fit a binary regression with `Class` as the response and the other nine variables as predictors. Report the residual deviance and associated degrees of freedom. Can this information be used to determine if this model fits the data? Explain.
  - (d) Use AIC as the criterion to determine the best subset of variables. (Use the `step` function.)
  - (e) Suppose that a cancer is classified as benign if  $p > 0.5$  and malignant if  $p < 0.5$ . Compute the number of errors of both types that will be made if this method is applied to the current data with the reduced model.

- (f) Suppose we change the cutoff to 0.9 so that  $p < 0.9$  is classified as malignant and  $p > 0.9$  as benign. Compute the number of errors in this case.
- (g) Produce an ROC plot and comment on effectiveness of the new diagnostic test.
- (h) It is usually misleading to use the same data to fit a model and test its predictive ability. To investigate this, split the data into two parts — assign every third observation to a test set and the remaining two thirds of the data to a training set. Use the training set to determine the model and the test set to assess its predictive performance. Compare the outcome to the previously obtained results.
2. The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix. The purpose of the study was to investigate factors related to diabetes. The data may be found in the dataset `pima`.
- Create a factor version of the test results and use this to produce an interleaved histogram to show how the distribution of insulin differs between those testing positive and negative. Do you notice anything unbelievable about the plot?
  - Replace the zero values of insulin with the missing value code `NA`. Recreate the interleaved histogram plot and comment on the distribution.
  - Replace the incredible zeroes in other variables with the missing value code. Fit a model with the result of the diabetes test as the response and all the other variables as predictors. How many observations were used in the model fitting? Why is this less than the number of observations in the data frame.
  - Refit the model but now without the `insulin` and `triceps` predictors. How many observations were used in fitting this model? Devise a test to compare this model with that in the previous question.
  - Use AIC to select a model. You will need to take account of the missing values. Which predictors are selected? How many cases are used in your selected model?
  - Create a variable that indicates whether the case contains a missing value. Use this variable as a predictor of the test result. Is missingness associated with the test result? Refit the selected model, but now using as much of the data as reasonable. Explain why it is appropriate to do this.
  - Using the last fitted model of the previous question, what is the difference in the odds of testing positive for diabetes for a woman with a BMI at the first quartile compared with a woman at the third quartile, assuming that all other factors are held constant? Give a confidence interval for this difference.
  - Do women who test positive have higher diastolic blood pressures? Is the diastolic blood pressure significant in the regression model? Explain the distinction between the two questions and discuss why the answers are only apparently contradictory.
3. A study was conducted on children who had corrective spinal surgery. We are interested in factors that might result in kyphosis (a kind of deformation) after surgery. The data can be loaded by

```
data(kyphosis, package="rpart")
```

Consult the help page on the data for further details.

- (a) Make plots of the response as it relates to each of the three predictors. You may find a jittered scatterplot more effective than the interleaved histogram for a dataset of this size. Comment on how the predictors appear to be related to the response.
  - (b) Fit a GLM with the kyphosis indicator as the response and the other three variables as predictors. Plot the deviance residuals against the fitted values. What can be concluded from this plot?
  - (c) Produce a binned residual plot as described in the text. You will need to select an appropriate amount of binning. Comment on the plot.
  - (d) Plot the residuals against the Start predictor, using binning as appropriate. Comment on the plot.
  - (e) Produce a normal QQ plot for the residuals. Interpret the plot.
  - (f) Make a plot of the leverages. Interpret the plot.
  - (g) Check the goodness of fit for this model. Create a plot like Figure 2.9. Compute the Hosmer-Lemeshow statistic and associated  $p$ -value. What do you conclude?
  - (h) Use the model to classify the subjects into predicted outcomes using a 0.5 cutoff. Produce cross-tabulation of these predicted outcomes with the actual outcomes. When kyphosis is actually present, what is the probability that this model would predict a present outcome? What is the name for this characteristic of the test?
4. Treatment of prostate cancer depends on whether the cancer has spread to the surrounding lymph nodes. This can be determined using a surgical procedure but it would be better if noninvasive methods could be used. Load in the data and learn about the variables by:

```
data(nodal, package="boot")
help(nodal, package="boot")
```

- (a) A plot consisting of a binary image of the data can be constructed as:

```
nodal$m <- NULL
image(as.matrix(nodal))
```

Improve this plot by ordering the cases on the response and labeling the axes informatively using the `axis` command.

- (b) Fit an appropriate model with nodal outcome as the response and the other five variables as predictors. Is there evidence that at least some of the five predictors are related to the response?
- (c) Fit a smaller model that removes `aged` and `grade` from the model. Can this smaller model be used in preference to the larger model?
- (d) How much does having a serious x-ray result increase the odds of nodal involvement compared to a nonserious result? (Use the smaller model.) Give a 95% confidence interval for the odds.
- (e) Fit a model with all five predictors and all their two-way interactions. Explain why the standard errors of the coefficients are so large.

- (f) Use the bias-reduced model fitting method described in the chapter to fit the model of the previous question. Which interaction is largest?
- (g) If the predicted response probability exceeds 0.5, the case is classified positively and, if not, negatively. Use the bias-reduced model to classify the cases in the dataset. Compare these to the actual classifications. How many were wrongly classified? Repeat this comparison for the model in (b). Do you think these misclassification rates are a reasonable estimate of how these models will perform in the future?
5. A study was conducted to determine the effectiveness of a new teaching method in economics. The data may be found in the dataset `spector`. Write a report on how well the new method works. You should include suitable graphical depictions of the data, diagnostics on your chosen model and an interpretation of the effects of the predictors on the response.
6. Additional datasets with binary responses are readily accessible for more practice:
- lizards in the `brglm` package
  - shuttle, crabs, birthwt, biopsy, bacteria in the `MASS` package
  - urine in the `boot` package

**This page intentionally left blank**

## Binomial and Proportion Responses

---

Sometimes we observe more than one binary outcome for each combination of the predictors so that the response becomes more than simply binary but some number of one type and some number of the other type. In such situations, we have a binomial rather than just a Bernoulli distributed response.

### 3.1 Binomial Regression Model

Suppose the response variable  $Y_i$  for  $i = 1, \dots, n$  is binomially distributed  $B(m_i, p_i)$  so that:

$$P(Y_i = y_i) = \binom{m_i}{y_i} p_i^{y_i} (1 - p_i)^{m_i - y_i}$$

We further assume that the  $Y_i$  are independent. The individual outcomes or *trials* that compose the response  $Y_i$  are all subject to the same  $q$  predictors  $(x_{i1}, \dots, x_{iq})$ . The group of trials is known as a *covariate class*. For example, we might record whether customers of a particular type make a purchase or not. Conventionally, one outcome is labeled a *success* (say, making purchase in this example) and the other outcome is labeled as a *failure*. No emotional meaning should be attached to success and failure in this context. For example, success might be the label given to a patient death with survival being called a failure. Because we need to have multiple trials for each covariate class, data for binomial regression models is more likely to result from designed experiments with a few predictors at chosen values rather than observational data which is likely to be more sparse.

As in the binary case, we construct a *linear predictor*:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_q x_{iq}$$

We can use a logistic link function  $\eta_i = \log(p_i / (1 - p_i))$ . The log-likelihood is then given by:

$$l(\beta) = \sum_{i=1}^n \left[ y_i \eta_i - m_i \log(1 + e^{\eta_i}) + \log \binom{m_i}{y_i} \right]$$

Let's work through an example to see how the analysis differs from the binary response case.

In January 1986, the space shuttle Challenger exploded shortly after launch. An investigation was launched into the cause of the crash and attention focused on the rubber O-ring seals in the rocket boosters. At lower temperatures, rubber becomes more brittle and is a less effective sealant. At the time of the launch, the temperature

was 31°F. Could the failure of the O-rings have been predicted? In the 23 previous shuttle missions for which data exists, some evidence of damage due to blow by and erosion was recorded on some O-rings. Each shuttle had two boosters, each with three O-rings. For each mission, we know the number of O-rings out of six showing some damage and the launch temperature. This is a simplification of the problem — see Dalal et al. (1989) for more details.

We plot the proportion of damaged O-rings against temperature in Figure 3.1:

```
data(orings, package="faraway")
plot(damage/6 ~ temp, orings, xlim=c(25, 85), ylim = c(0,1), xlab=
  "Temperature", ylab="Prob of damage")
```

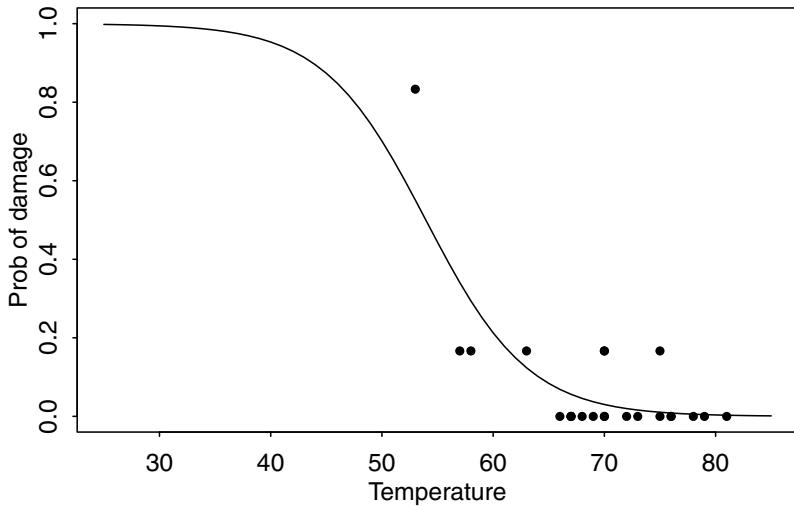


Figure 3.1 *Damage to O-rings in 23 space shuttle missions as a function of launch temperature. Logistic fit is shown.*

We are interested in how the probability of failure in a given O-ring is related to the launch temperature and predicting that probability when the temperature is 31°F. We estimate the regression parameters for the Challenger data:

```
lmod <- glm(cbind(damage, 6-damage) ~ temp, family=binomial, orings)
summary(lmod)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	11.6630	3.2963	3.54	0.0004
temp	-0.2162	0.0532	-4.07	0.000048

n = 23 p = 2

Deviance = 16.912 Null Deviance = 38.898 (Difference = 21.985)

For binomial response data, we need two pieces of information about the response values —  $y_i$  and  $m_i$ . We express this as a two-column matrix with the first column representing the number of successes  $y$  and the second column the number of failures  $m - y$ . In this case, an O-ring damage incident is a “success.”

We show the logit fit to the data as seen in Figure 3.1:

```
x <- seq(25, 85, 1)
```

```
lines(x, ilogit(11.6630 - 0.2162*x))
```

We can predict the response at 31°F:

```
ilogit(11.6630 - 0.2162*31)
```

```
[1] 0.99304
```

We see a very high probability of damage although we still need to develop some inferential techniques before we leap to any conclusion.

### 3.2 Inference

We use the same likelihood-based methods as in Section 2.3 to derive the binomial deviance:

$$D = 2 \sum_{i=1}^n \{y_i \log y_i / \hat{y}_i + (m_i - y_i) \log(m_i - y_i) / (m_i - \hat{y}_i)\}$$

where  $\hat{y}_i$  are the fitted values from the model.

Provided that  $Y$  is truly binomial and that the  $m_i$  are relatively large, the deviance is approximately  $\chi^2$  distributed with  $n - q - 1$  degrees of freedom if the model is correct. Thus we can use the deviance to test whether the model is an adequate fit. For the logit model of the Challenger data, we may compute:

```
pchisq(deviance(lmod), df.residual(lmod), lower=FALSE)
```

```
[1] 0.71641
```

Since this  $p$ -value is well in excess of 0.05, we conclude that this model fits sufficiently well. Of course, this does not mean that this model is correct or that a simpler model might not also fit adequately. Even so, for the null model:

```
pchisq(38.9, 22, lower=FALSE)
```

```
[1] 0.014489
```

We see that the fit is inadequate, so we cannot ascribe the response to simple variation not dependent on any predictor. Note that a  $\chi_d^2$  variable has mean  $d$  and standard deviation  $\sqrt{2d}$  so that it is often possible to quickly judge whether a deviance is large or small without explicitly computing the  $p$ -value. If the deviance is far in excess of the degrees of freedom, the null hypothesis can be rejected.

The  $\chi^2$  distribution is only an approximation that becomes more accurate as the  $m_i$  increase. The approximation is very poor for small  $m_i$  and fails entirely in binary cases where  $m_i = 1$ . Although it is not possible to say exactly how large  $m_i$  should be for an adequate approximation,  $m_i \geq 5 \forall i$  has often been suggested. Permutation or bootstrap methods might be considered as an alternative.

We can also use the deviance to compare two models, with smaller model  $S$  representing a subspace (usually a subset) of a larger model  $L$ . The likelihood ratio test statistic becomes  $D_S - D_L$ . This test statistic is asymptotically distributed  $\chi_{L-S}^2$ , assuming that the smaller model is correct and the distributional assumptions hold. We can use this to test the significance of temperature by computing the difference in the deviances between the model with and without temperature. The model without temperature is just the null model and the difference in degrees of freedom or parameters is one:

```
pchisq(38.9 - 16.9, 1, lower=FALSE)
```

```
[1] 2.7265e-06
```

Since the  $p$ -value is so small, we conclude that the effect of launch temperature is statistically significant. An alternative to this test is the  $z$ -value, which is  $\hat{\beta}/se(\hat{\beta})$ , here equal to  $-4.07$  with a  $p$ -value of  $4.8e-05$ . As in the binary case, the deviance-based test is preferred. Again, there are concerns with the accuracy of the approximation, but the test involving differences of deviances is generally more accurate than the goodness of fit test involving a single deviance.

What if all the cases that form a covariate class have not been grouped together? Let's see what happens when each individual O-ring is listed separately:

```
erings <- with(orings, data.frame(temp=rep(temp,each=6), damage=as.
  ↪ vector(sapply(orings$damage, function(x) rep(c(0,1), times=c(6-
  ↪ x,x))))))
```

The first launch was at 53 degrees and resulted in five damaged O-rings out of six. This has been divided into six individual trials:

```
head(erings)
```

	temp	damage
1	53	0
2	53	1
3	53	1
4	53	1
5	53	1
6	53	1

Now we fit the model and examine the output:

```
emod <- glm(damage ~ temp, family=binomial, erings)
summary(emod)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	11.6630	3.2962	3.54	0.0004
temp	-0.2162	0.0532	-4.07	0.000048

$n = 138$   $p = 2$

Deviance = 54.759 Null Deviance = 76.745 (Difference = 21.985)

We see that the parameter estimates, standard errors and the difference in the deviances are the same as before so we will make the same conclusions. The only thing we have lost in this version of the output is the ability to make the goodness of fit test from the residual deviance.

Confidence intervals for the regression parameters may be constructed as with the binary regression model:

```
confint(lmod)
```

	Waiting for profiling to be done...
(Intercept)	2.5 % 97.5 %
temp	5.57543 18.73812
temp	-0.33267 -0.12018

We prefer this profile likelihood method to the  $\hat{\beta} \pm 2se(\hat{\beta})$  approach.

### 3.3 Pearson's $\chi^2$ Statistic

The deviance is one measure of how well the model fits the data, but there are alternatives. The Pearson's  $X^2$  statistic takes the general form:

$$X^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

where  $O_i$  is the observed count and  $E_i$  is the expected count for case  $i$ . For a binomial response, we count the number of successes for which  $O_i = y_i$  while  $E_i = n_i \hat{p}_i$  and failures for which  $O_i = n_i - y_i$  and  $E_i = n_i(1 - \hat{p}_i)$ , which results in:

$$X^2 = \sum_{i=1}^n \frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i (1 - \hat{p}_i)}$$

If we define *Pearson residuals* as:

$$r_i^P = (y_i - n_i \hat{p}_i) / \sqrt{\text{var } \hat{y}_i}$$

which can be viewed as a type of standardized residual, then  $X^2 = \sum_{i=1}^n (r_i^P)^2$ . So the Pearson's  $X^2$  is analogous to the residual sum of squares used in normal linear models.

The Pearson  $X^2$  will typically be close in size to the deviance and can be used in the same manner. Alternative versions of the hypothesis tests described above might use the  $X^2$  in place of the deviance with the same approximate null distributions. However, some care is necessary because the model is fit to minimize the deviance and not the Pearson's  $X^2$ . This means that it is possible, although unlikely, that the  $X^2$  could increase as a predictor is added to the model.  $X^2$  can be computed like this:

```
[1] 28.067
```

Compare this to:

```
deviance(1mod)
```

```
[1] 16.912
```

In this case there is more than the typical small difference between  $X^2$  and the deviance. However, a test for model fit:

```
1-pchisq(28.067, 21)
```

```
[1] 0.13826
```

results in a moderate sized  $p$ -value which would not reject this model which agrees with decision based on the deviance statistic.

### 3.4 Overdispersion

If the binomial model specification is correct, we expect that the residual deviance will be approximately distributed  $\chi^2$  with the appropriate degrees of freedom. Sometimes, we observe a deviance that is much larger than would be expected if the model were correct. We must then determine which aspect of the model specification is incorrect.

The most common explanation is that we have the wrong structural form for the model. We have not included the right predictors or we have not transformed or combined them in the correct way. We have a number of ways of determining the importance of potential additional predictors and diagnostics for determining better transformations — see Section 8.4. Suppose, however, that we are able to exclude this explanation. This is difficult to achieve, but when we have only one or two predictors, it is feasible to explore the model space quite thoroughly and be sure that there is not a plausible superior model formula.

Another common explanation for a large deviance is the presence of a small

number of outliers. Fortunately, these are easily checked using diagnostic methods. When larger numbers of points are identified as outliers, they become unexceptional, and we might more reasonably conclude that there is something amiss with the error distribution.

Sparse data can also lead to large deviances. In the extreme case of a binary response, the deviance is not even approximately  $\chi^2$ . In situations where the group sizes are simply small, the approximation is poor. Because we cannot judge the fit using the deviance, we shall exclude this case from further consideration in this section.

Having excluded these other possibilities, we might explain a large deviance by deficiencies in the random part of the model. A binomial distribution for  $Y$  arises when the probability of success  $p$  is independent and identical for each trial within the group. If the group size is  $m$ , then  $\text{var } Y = mp(1 - p)$  if the binomial assumptions are correct. However, if the assumptions are broken, the variance may be greater. This is *overdispersion*. In rarer cases, the variance is less and *underdispersion* results.

There are two main ways that overdispersion can arise — the independent or identical assumptions can be violated. We look at the constant  $p$  assumption first. It is easy to see how there may be some unexplained heterogeneity within a group that might lead to some variation in  $p$ . For example, in the shuttle disaster case study of Section 3.1, the position of the O-ring on the booster rocket may have some effect on the failure probability. Yet this variable was not recorded and so we cannot include it as a predictor. Heterogeneity can also result from clustering. Suppose a population is divided into clusters, so that when you take a sample, you actually get a sample of clusters. This would be common in epidemiological applications.

Let the sample size be  $m$ , the cluster size be  $k$  and the number of clusters be  $l = m/k$ . Let the number of successes in cluster  $i$  be  $Z_i \sim B(k, p_i)$ . Now suppose that  $p_i$  is a random variable such that  $E p_i = p$  and  $\text{var } p_i = \tau^2 p(1 - p)$ . Let the total number of successes be  $Y = Z_1 + \dots + Z_l$ . Then:

$$EY = \sum EZ_i = \sum_{i=1}^l kp = mp$$

as in the standard case, but:

$$\text{var } Y = \sum \text{var } Z_i = \sum \{E(\text{var } (Z_i|p_i)) + \text{var } (E(Z_i|p_i))\} = (1 + (k-1)\tau^2)mp(1-p)$$

So  $Y$  is overdispersed since  $1 + (k-1)\tau^2 \geq 1$ . Notice that in the sparse case,  $m = 1$ , and this problem cannot arise.

Overdispersion can also result from dependence between trials. If the response has a common cause, say a disease is influenced by genes, the responses will tend to be positively correlated. For example, subjects in human or animal trials may be influenced in their responses by other subjects. If the food supply is limited, the probability of survival of an animal may be increased by the death of others. This circumstance would result in underdispersion.

The simplest approach for modeling overdispersion is to introduce an additional dispersion parameter, so that  $\text{var } Y = \sigma^2 mp(1 - p)$ . In the standard binomial case

$\sigma^2 = 1$ . We now let  $\sigma^2$  vary and estimate using the data. Notice the similarity to linear regression. The dispersion parameter may be estimated using:

$$\hat{\sigma}^2 = \frac{X^2}{n - p}$$

Using the deviance in place of the Pearson's  $X^2$  is not recommended as it may not be consistent. The estimation of  $\beta$  is unaffected since  $\sigma^2$  does not change the mean response but:

$$\text{var } \hat{\beta} = \hat{\sigma}^2 (X^T \hat{W} X)^{-1}$$

So we need to scale up the standard errors by a factor of  $\hat{\sigma}$ .

We cannot use the difference in deviances when comparing models, because the test statistic will be distributed  $\sigma^2 \chi^2$ . Since  $\sigma^2$  is not known and must be estimated in the overdispersion situation, an F-statistic must be used:

$$F = \frac{(D_{\text{small}} - D_{\text{large}}) / (df_{\text{small}} - df_{\text{large}})}{\hat{\sigma}^2}$$

This statistic is only approximately F distributed, in contrast to the Gaussian case.

In Manly (1978), an experiment is reported where boxes of trout eggs were buried at five different stream locations and retrieved at four different times, specified by the number of weeks after the original placement. The number of surviving eggs was recorded. The box was not returned to the stream. The data is also analyzed by Hinde and Demetrio (1988). We can construct a tabulation of the data by:

```
data(troutegg, package="faraway")
ftable(xtabs(cbind(survive,total) ~ location+period, troutegg))
```

location	period		
		survive	total
1	4	89	94
	7	94	98
	8	77	86
	11	141	155
2	4	106	108
	7	91	106
	8	87	96
	11	104	122
3	4	119	123
	7	100	130
	8	88	119
	11	91	125
4	4	104	104
	7	80	97
	8	67	99
	11	111	132
5	4	49	93
	7	11	113
	8	18	88
	11	0	138

Notice that in one case, all the eggs survive, while in another, none of the eggs survive. We now fit a binomial GLM for the two main effects:

```
bmod <- glm(cbind(survive,total-surveive) ~ location+period, family=
  ↪ binomial,troutegg)
summary(bmod)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	4.636	0.281	16.48	< 2e-16
location2	-0.417	0.246	-1.69	0.09
location3	-1.242	0.219	-5.66	1.5e-08
location4	-0.951	0.229	-4.16	3.2e-05
locations5	-4.614	0.250	-18.44	< 2e-16
period7	-2.170	0.238	-9.10	< 2e-16
period8	-2.326	0.243	-9.57	< 2e-16
period11	-2.450	0.234	-10.47	< 2e-16

n = 20 p = 8

Deviance = 64.495 Null Deviance = 1021.469 (Difference = 956.974)

The deviance of 64.5 on 12 degrees of freedom seems to show that this model does not fit. Before we conclude that there is overdispersion, we need to eliminate other potential explanations. With about 100 eggs in each box, we have no problem with sparseness, but we do need to check for outliers and look at the model formula. A half-normal plot of the residuals is a good way to check for outliers:

```
halfnorm(residuals(bmod))
```

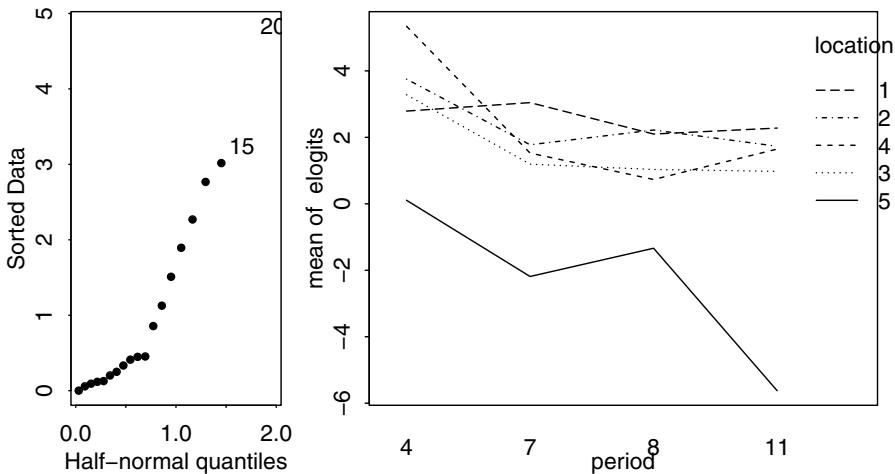


Figure 3.2 *Diagnostic plots for the trout egg model. A half-normal plot of the residuals is shown on the left and an interaction plot of the empirical logits is shown on the right.*

The half-normal plot is shown in the left panel of Figure 3.2. No single outlier is apparent. Perhaps one can discern a larger number of residuals which seem to follow a more dispersed distribution than the rest.

We can also check whether the predictors are correctly expressed by plotting the *empirical logits*. These are defined as:

$$\log \left( \frac{y + 1/2}{m - y + 1/2} \right)$$

The halves are added to prevent infinite values for groups consisting of all successes or failures. We now construct an interaction plot of the empirical logits:

```
elogits <- with(troutegg, log((survive+0.5)/(total- survive+0.5)))
with(troutegg, interaction.plot(period, location, elogits))
```

Interaction plots are difficult to interpret conclusively, but there is no obvious sign of large interactions. So there is no evidence that the linear model is inadequate. We do not have any outliers and the functional form of the model appears to be suitable, but the deviance is still larger than should be expected. Having eliminated these more obvious causes as the source of the problem, we may now put the blame on overdispersion. Possible reasons for the overdispersion include inhomogeneous trout eggs, variation in the experimental procedures or unknown variables affecting survival.

We can estimate the dispersion parameter as:

```
(sigma2 <- sum(residuals(bmod, type="pearson"))^2) / 12)
[1] 5.3303
```

We see that this is substantially larger than one as it would be in the standard binomial GLM. We can now make *F*-tests on the predictors using:

```
drop1(bmod, scale=sigma2, test="F")
```

```
Single term deletions
scale: 5.3303
```

	Df	Deviance	AIC	F value	Pr(F)
<none>		64	157		
location	4	914	308	39.5	8.1e-07
period	3	229	182	10.2	0.0013

Warning message:  
F test assumes quasibinomial family in:  
drop1.glm(bmod, scale = sigma2, test = "F")

We see that both terms are clearly significant. It is necessary to specify the `scale` argument using the estimated value of  $\sigma^2$ . If this argument is omitted, the deviance will be used in the estimation of the dispersion parameter. For this particular dataset, it makes very little difference, but in some cases, using the deviance to estimate the dispersion gives inconsistent results. The warning message reminds us that the use of free dispersion parameter results in a model that is no longer a true binomial GLM, but rather what is known as a *quasi-binomial* GLM. More on such models may be found in Section 3.5.

No goodness of fit test is possible because we have a free dispersion parameter. We can use the dispersion parameter to scale up the estimates of the standard error as in:

```
summary(bmod, dispersion=sigma2)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	4.636	0.649	7.14	9.5e-13
location2	-0.417	0.568	-0.73	0.463
location3	-1.242	0.507	-2.45	0.014
location4	-0.951	0.528	-1.80	0.072
location5	-4.614	0.578	-7.99	1.4e-15
period7	-2.170	0.550	-3.94	8.1e-05
period8	-2.326	0.561	-4.15	3.4e-05
period11	-2.450	0.540	-4.53	5.8e-06

```
overdispersion parameter = 5.330
```

```
n = 20 p = 8
Deviance = 64.495 Null Deviance = 1021.469 (Difference = 956.974)
```

We see that the differences in the location become less pronounced with only the fifth location being clearly different.

This dispersion parameter method is only appropriate when the covariate classes are roughly equal in size. If not, more sophisticated methods should be used. One such approach uses the beta-binomial distribution where we assume that  $p$  follows a beta distribution. This approach is discussed in Williams (1982) and Crowder (1978) and can be implemented using the `dispmod` package of Scrucia (2012) in R:

```
library(dispmod)
dmod <- glm.binomial.disp(bmod)
summary(dmod)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	4.518	0.621	7.28	3.3e-13
location2	-0.377	0.560	-0.67	0.501
location3	-1.210	0.507	-2.39	0.017
location4	-0.956	0.520	-1.84	0.066
location5	-4.468	0.559	-8.00	1.3e-15
period7	-2.086	0.520	-4.01	6.1e-05
period8	-2.227	0.523	-4.26	2.0e-05
period11	-2.362	0.519	-4.56	5.2e-06

```
n = 20 p = 8
Deviance = 12.400 Null Deviance = 190.186 (Difference = 177.786)
```

As can be seen, the results are quite similar to the previous approach because the covariate class sizes do not vary much.

### 3.5 Quasi-Binomial

In the previous section, we have demonstrated ways to model data where the supposedly binomial response is more variable than should be expected. A *quasi-binomial* model is another way to allow for extra-binomial variation. We will explain the method in greater generality than immediately necessary because the idea can be used across a wider range of response types.

The idea is to specify only how the mean and variance of the response are connected to the linear predictor. The method of weighted least squares, as used for standard linear models, would be a simple example of this. An examination of the fitting of the binomial model reveals that this only requires the mean and variance information and does not use any additional information about the binomial distribution. Hence, we can obtain the parameter estimates  $\hat{\beta}$  and standard errors without making the full binomial assumption.

The problem arises when we attempt to do inference. To construct a confidence interval or perform an hypothesis test, we need some distributional assumptions. Previously we have used the deviance, but for this we need a likelihood and to compute a likelihood we need a distribution. Now we need a suitable substitute for a likelihood that can be computed without assuming a distribution.

Let  $Y_i$  have mean  $\mu_i$  and variance  $\phi V(\mu_i)$ . We assume that  $Y_i$  are independent. We

define a score,  $U_i$ :

$$U_i = \frac{Y_i - \mu_i}{\phi V(\mu_i)}$$

Now:

$$EU_i = 0$$

$$\text{var } U_i = \frac{1}{\phi V(\mu_i)}$$

$$-E \frac{\partial U_i}{\partial \mu_i} = -E \frac{-\phi V(\mu_i) - (Y_i - \mu_i)\phi V'(\mu_i)}{[\phi V(\mu_i)]^2} = \frac{1}{\phi V(\mu_i)}$$

These properties are shared by the derivative of the log-likelihood,  $l'$ . This suggests that we can use  $U$  in place of  $l'$ . So we define:

$$Q_i = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} dt$$

The intent is that  $Q$  should behave like the log-likelihood. We then define the log *quasi-likelihood* for all  $n$  observations as:

$$Q = \sum_{i=1}^n Q_i$$

The usual asymptotic properties expected of maximum likelihood estimators also hold for quasi-likelihood-based estimators as may be seen in McCullagh (1983).

Notice that the quasi-likelihood depends directly only on the variance function and that the choice of distribution also determines only the variance function. So the choice of variance function is associated with the random structure of the model while the link function determines the relationship with the systematic part of the model.

For the standard linear model, the quasi-likelihood corresponds exactly to the log-likelihood. Here the *dispersion parameter*  $\phi$  is  $\sigma^2$  so nothing is gained by this approach. However, for the binomial model, the introduction of  $\phi$  provides an additional dimension of flexibility to the model, which is useful in modeling overdispersion. One curious possibility is that some choices of  $V(\mu)$  may not correspond to a known, or even any, distribution.

$\hat{\beta}$  is obtained by maximizing  $Q$ . Everything proceeds as before except for the estimation of  $\phi$  since the likelihood approach is not reliable here. We recommend:

$$\hat{\phi} = \frac{X^2}{n - p}$$

Although quasi-likelihood estimators are attractive because they require fewer assumptions, they are generally less efficient than the corresponding regular likelihood-based estimator. So if you have information about the distribution, you are advised to use it.

The inferential procedures are similar to those used before. Recall that the regular

deviance for a model is formed from the difference in log-likelihoods for the model and the saturated model:

$$D(y, \hat{\mu}) = -2\phi \sum_i (l(\hat{\mu}_i | y_i) - l(y_i | y_i))$$

so by analogy the quasi-deviance is  $-2\phi Q$  because the contribution from the saturated model is zero. The  $\phi$  cancels, so the quasi-deviance is just:

$$Q = -2 \sum_i \int_{y_i}^{\mu_i} \frac{y_i - t}{V(t)} dt$$

In Allison and Cicchetti (1976), data on the sleep behavior of 62 mammals is presented. Suppose we are interested in modeling the proportion of sleep spent dreaming as a function of the other predictors: the weight of the body and the brain, the gestation period, the lifespan and the three constructed indices measuring vulnerability to predation, exposure while sleeping and overall danger:

```
data(mammalsleep, package="faraway")
mammalsleep$pdr <- with(mammalsleep, dream/sleep)
summary(mammalsleep$pdr)
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
	0.000	0.118	0.176	0.186	0.243	0.462	14.000

We notice that the proportion of time spent dreaming varies from zero up to almost half the time. For some datasets with a proportion response, the range of proportions might never come close to zero or one. For such datasets, using a normal Gaussian model (with perhaps some weights) might be acceptable. But for this dataset, we have some very small proportions as response values and so a Gaussian model will not work. We attempt to model the proportion response directly. A logit link seems sensible since the response is restricted between zero and one. Furthermore, we might expect the variance to be greater for moderate values of the proportion  $\mu$  and less as  $\mu$  approaches zero or one because of the nature of the measurements. This suggests a variance function of the approximate form  $\mu(1 - \mu)$ . This corresponds to the binomial GLM with the canonical logit link and yet the response is not binomial. We can use a quasi-binomial:

```
mod1 <- glm(pdr ~ log(body) + log(brain) + log(lifespan) + log(
  ↪ gestation) + predation + exposure + danger, family=
  ↪ quasibinomial, mammalsleep)
```

where we have logged many of the predictors because of skewness. Since we now have a free dispersion parameter, we must use  $F$ -tests to compare models:

```
drop1(mod1, test="F")
```

	Df	Deviance	F value	Pr(F)
<none>		1.57		
log(body)	1	1.78	4.51	0.041
log(brain)	1	1.59	0.33	0.568
log(lifespan)	1	1.65	1.79	0.189
log(gestation)	1	1.62	1.15	0.292
predation	1	1.57	0.10	0.749
exposure	1	1.58	0.32	0.575
danger	1	1.58	0.31	0.579

We might eliminate predation as the least significant variable. Further sequential backward elimination results in:

```
mdl <- glm(pdr ~ log(body) + log(lifespan) + danger, family=
  ↪ quasibinomial, mammalsleep )
summary(mdl)

Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.4932    0.2913   -1.69  0.09796
log(body)    0.1463    0.0384    3.81  0.00046
log(lifespan) -0.2866   0.1080   -2.65  0.01126
danger      -0.1732   0.0600   -2.89  0.00615
```

```
overdispersion parameter = 0.041
n = 45 p = 4
Deviance = 1.732 Null Deviance = 2.509 (Difference = 0.777)
```

Notice that the dispersion parameter is far less than the default value of one that we would see for a binomial. We see the proportion of time spent dreaming increases for heavier mammals that live less time and live in less danger. Notice that the relatively large residual deviance compared to the null deviance indicates that this is not a particularly well-fitting model.

The usual diagnostics should be performed. Here are two selected plots that have some interest:

```
ll <- row.names(na.omit(mammalsleep[,c(1,6,10,11)]))
halfnorm(cooks.distance(mdl), labs=ll)
plot(predict(mdl), residuals(mdl,type="pearson"), xlab="Linear
  ↪ Predictor", ylab="Pearson Residuals")
```

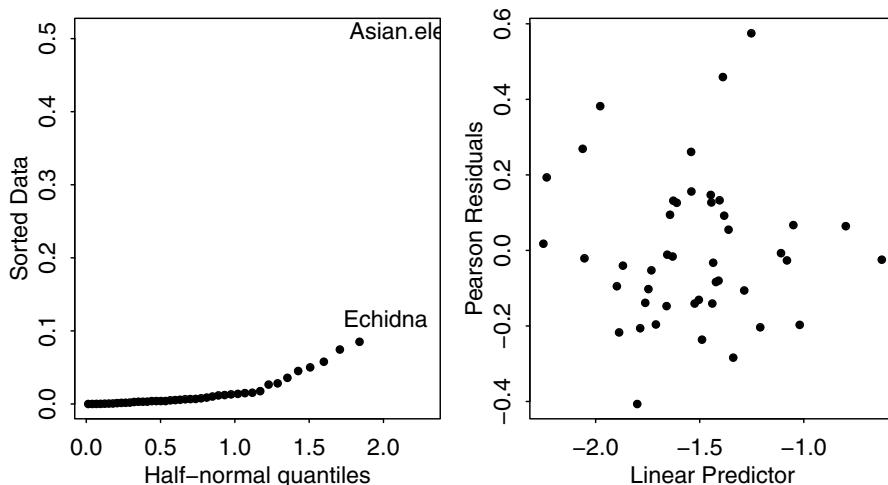


Figure 3.3 A half-normal plot of the Cook statistics is shown on the left and a plot of the Pearson residuals against the fitted linear predictors is shown on the right.

In the first panel of Figure 3.3, we see that the Asian elephant is quite influential and a fit without this case should be considered. In the second panel, we see that a pattern of constant variation indicating that our choice of variance function was reasonable. We used the Pearson residuals because these explicitly normalize the raw residuals

using the variance function making the check more transparent. Even so, the deviance residuals would have served the same purpose.

### 3.6 Beta Regression

Beta regression is useful for responses that are bounded in  $(0, 1)$  such as proportions. It could also be used for variables that are bounded in some other finite interval simply by rescaling to  $(0, 1)$ . A Beta-distributed random variable  $Y$  has density:

$$f(y|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1} (1-y)^{b-1}$$

for parameters  $a, b$  and Gamma function  $\Gamma()$ . It is more convenient to transform the parameters so  $\mu = a/(a+b)$  and  $\phi = a+b$  so that  $EY = \mu$  and  $\text{var } Y = \mu(1-\mu)/(1+\phi)$ . We can then link the linear predictor  $\eta$  using  $\eta = g(\mu)$  using a link function  $g$  where any of the choices used for the binomial model would be suitable.

An implementation of the Beta regression model can be found in the `mgcv` package of Wood (2006). We can apply this to the `mammalsleep` also used in the previous section.

```
data(mammalsleep, package="faraway")
mammalsleep$pdr <- with(mammalsleep, dream/sleep)
library(mgcv)
modb <- gam(pdr ~ log(body)+log(lifespan), family=betar(), mammalsleep
             )
summary(modb)
Family: Beta regression(8.927)
Link function: logit

Formula:
pdr ~ log(body) + log(lifespan)

Parametric coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.3779    0.3732   1.01    0.31
log(body)   0.2680    0.0551   4.86  1.2e-06
log(lifespan) -0.9227   0.1658  -5.56  2.6e-08

R-sq.(adj) = -0.178 Deviance explained = 73.5%
-REML = -47.801 Scale est. = 1 n = 45
```

The default choice of link is the logit function. The estimated value of  $\phi$  is 8.927. A comparison of the fitted values of this model and the quasi-binomial model fitted earlier reveals no substantial difference. The advantage of the Beta-based model is the full distributional model which would allow the construction of full predictive distributions rather than just a point estimate and standard error.

### Exercises

1. The question concerns data from a case-control study of esophageal cancer in Ille-et-Vilaine, France. The data is distributed with R and may be obtained along with a description of the variables by:

```
data(esoph)
help(esoph)
```

- (a) Plot the proportion of cases against each predictor using the size of the point to indicate the number of subject as seen in Figure 2.7. Comment on the relationships seen in the plots.
  - (b) Fit a binomial GLM with interactions between all three predictors. Use AIC as a criterion to select a model using the `step` function. Which model is selected?
  - (c) All three factors are ordered and so special contrasts have been used appropriate for ordered factors involving linear, quadratic and cubic terms. Further simplification of the model may be possible by eliminating some of these terms. Use the `unclass` function to convert the factors to a numerical representation and check whether the model may be simplified.
  - (d) Use the summary output of the factor model to suggest a model that is slightly more complex than the linear model proposed in the previous question.
  - (e) Does your final model fit the data? Is the test you make accurate for this data?
  - (f) Check for outliers in your final model.
  - (g) What is the predicted effect of moving one category higher in alcohol consumption?
  - (h) Compute a 95% confidence interval for this predicted effect.
2. Incubation temperature can affect the sex of turtles. An experiment was conducted with three independent replicates for each temperature and the number of male and female turtles born was recorded and can be found in the `turtle` dataset.
    - (a) Plot the proportion of males against the temperature. Comment on the nature of the relationship.
    - (b) Fit a binomial response model with a linear term in temperature. Does this model fit the data?
    - (c) Is this data sparse?
    - (d) Check for outliers.
    - (e) Compute the empirical logits and plot these against temperature. Does this indicate a lack of fit?
    - (f) Add a quadratic term in temperature. Is this additional term a significant predictor of the response. Does the quadratic model fit the data?
    - (g) There are three replicates for each value of temperature. Assuming independent binomial variation, how much variation would be expected in the three proportions observed? Compare this to the observed variation in these proportions. Do they approximately agree or is there evidence of greater variation?
    - (h) If the three replicates are homogenous, they could be combined so that the dataset would have only five cases in total. Create this dataset and fit a model linear in temperature. Compare the fit seen for this model with that found in (b).

3. A biologist analyzed an experiment to determine the effect of moisture content on seed germination. Eight boxes of 100 seeds each were treated with the same moisture level. Four boxes were covered and four left uncovered. The process was repeated at six different moisture levels.
- (a) Plot the germination percentage against the moisture level on two side-by-side plots according to the coverage of the box. What relationship do you see?
  - (b) Create a new factor describing the box (the data are ordered in blocks of 6 observations per box). Add lines to your previous plot that connect observations from the same box. Is there an indication of a box effect?
  - (c) Fit a binomial response model including the coverage, box and moisture predictors. Use the plots to determine an appropriate choice of model.
  - (d) Test for the significance of a box effect in your model. Repeat the same test but using the Pearson's Chi-squared statistic instead of the deviance.
  - (e) At what value of moisture does the predicted maximum germination occur for noncovered boxes? For covered boxes?
  - (f) Produce a plot of the residuals against the fitted values and interpret.
  - (g) Plot the residuals against moisture while distinguishing the covering. Interpret.
  - (h) Plot the residuals against the leverages. Are there any influential points?
4. This problem concerns the modeling of the quantitative structure-activity relationships (QSAR) of the inhibition of dihydrofolate reductase (DHFR) by pyrimidines. We want to relate the physicochemical and/or structural properties as exhibited by the 26 predictors in pyrimidines with an activity level. We have structural information on 74 2,4-diamino- 5-(substituted benzyl) pyrimidines used as inhibitors of DHFR in *E. coli*. All the variables lie in [0,1].
- (a) Plot the activity (response) against the first three predictors. Are any outliers in the response apparent? Remove any such cases.
  - (b) Fit a Gaussian linear model for the response with all 26 predictors. How well does this model fit the data in terms of  $R^2$ ? Plot the residuals against the fitted values. Is there any evidence of a violation of the standard assumptions?
  - (c) Fit a quasi-binomial model for the activity response. Compare the predicted values for this model to those for the Gaussian linear model. Take care to compute the predicted values in the appropriate scale. Compare the fitted coefficients between the two models. Are there any substantial differences?
  - (d) Fit a Gaussian linear model with the logit transformation applied to the response. Compare the coefficients of this model with the quasi-binomial model.
  - (e) Fit a Beta regression model. Compare the coefficients of this model with that of logit response regression model.
  - (f) What property of the response leads to the similarity of the models considered thus far in this question?

## Variations on Logistic Regression

---

### 4.1 Latent Variables

Suppose that students answer questions on a test and that a specific student has an aptitude  $T$ . A particular question might have difficulty  $d$  and the student will get the answer correct only if  $T > d$ . Now if we consider  $d$  fixed and  $T$  as a random variable with density  $f$  and distribution function  $F$ , then the probability that the student will get the answer wrong is:

$$p = P(T \leq d) = F(d)$$

$T$  is called a *latent variable*. Suppose that the distribution of  $T$  is *logistic*:

$$F(y) = \frac{\exp(y - \mu)/\sigma}{1 + \exp(y - \mu)/\sigma}$$

So

$$\text{logit}(p) = -\mu/\sigma + d/\sigma$$

If we set  $\beta_0 = -\mu/\sigma$  and  $\beta_1 = 1/\sigma$ , we now have a logistic regression model. We can illustrate this in the following example where we set  $d = 1$  and let  $T$  have mean  $-1$  and  $\sigma = 1$ :

```
x <- seq(-6, 4, 0.1)
y <- dlogis(x, location=-1)
plot(x,y,type="l",ylab="density",xlab="t")
ii <- (x <= 1)
polygon(c(x[ii],1,-6),c(y[ii],0,0),col='gray')
```

The plot in Figure 4.1 shows a logistically distributed latent variable. We can see that this distribution is apparently very similar to the normal distribution. The shaded area represents the probability of getting an answer wrong. As the mean aptitude of this student is somewhat less than the difficulty of the question, this probability is substantially greater than one half.

This idea also arises in a *bioassay* where we might treat an animal, plant or person with some concentration of a treatment and observe the outcome. For example, suppose we are interested in the concentration of insecticide to be used in exterminating a pest. Insects will have varying tolerances for the toxin and will survive if their tolerance is greater than the dose. In this context, the term *tolerance distribution* for  $T$  is used. Applications in several other areas exist where we observe only a binary outcome but believe this to be generated by some continuous but unobserved variable.

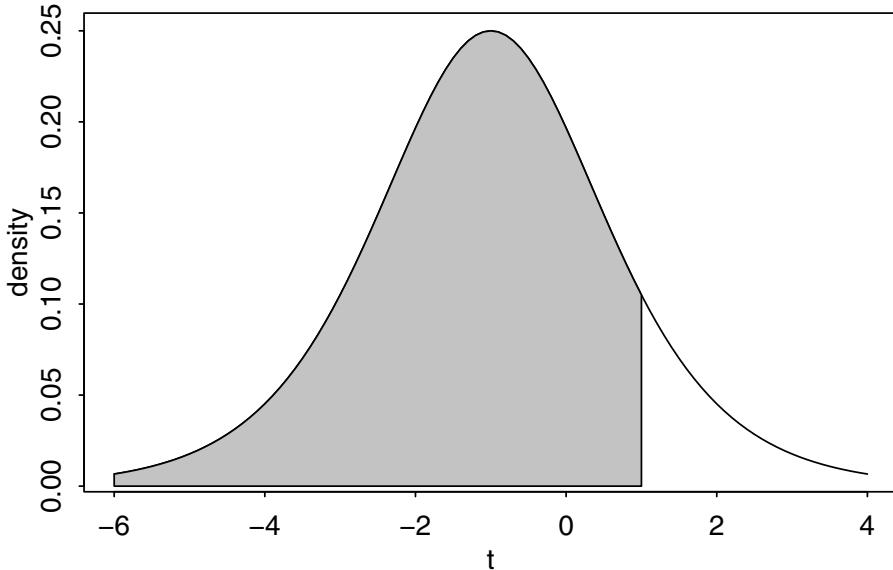


Figure 4.1 *Probability of getting the answer wrong for logistic latent variable.*

## 4.2 Link Functions

Until now we have used logit link function to connect the probability and the linear predictor. But other choices of link function are reasonable. We need a function that bounds the probability between zero and one. We also expect the link function to be monotone. It is conceivable that the success probability may go up and down as the linear predictor increases but this circumstance is best modeled by adding nonlinear components to the linear predictors such as quadratic terms rather than modifying the link function. The latent variable formulation suggests some other possibilities for the link function. Here are some choices which are implemented in the `glm()` function:

1. Probit:  $\eta = \Phi^{-1}(p)$  where  $\Phi$  is the normal cumulative distribution function. This arises from a normally distributed latent variable.
2. Complementary log-log:  $\eta = \log(-\log(1-p))$ . A Gumbel-distributed latent variable will lead to this.
3. Cauchit:  $\eta = \tan^{-1}(\pi(p-1/2))$  which is motivated by a Cauchy-distributed latent variable.

We can illustrate the choices using some data from Bliss (1935) on the numbers of insects dying at different levels of insecticide concentration. We fit all four link functions:

```
data(bliss, package="faraway")
bliss
  dead alive conc
1     2    28    0
```

```

2     8     22     1
3    15     15     2
4    23      7     3
5    27      3     4

```

```

mlogit <- glm(cbind(dead,alive) ~ conc, family=binomial, data=bliss)
mprobit <- glm(cbind(dead,alive) ~ conc, family=binomial(link=probit),
  ↪ data=bliss)
mcloglog <- glm(cbind(dead,alive) ~ conc, family=binomial(link=cloglog
  ↪ ), data=bliss)
mcauchit <- glm(cbind(dead,alive) ~ conc, family=binomial(link=cauchit
  ↪ ), data=bliss)

```

We start by considering the fitted values:

```

fitted(mlogit)
  1       2       3       4       5
0.089172 0.238323 0.500000 0.761677 0.910828

```

or from `predict(mlogit, type="response")`. These are constructed using linear predictor,  $\eta$ :

```

coef(mlogit)[1]+coef(mlogit)[2]*bliss$conc
[1] -2.3238 -1.1619  0.0000  1.1619  2.3238

```

Alternatively, these values may be obtained from `modl$linear.predictors` or `predict(modl)`. The fitted values are then:

```

library(faraway)
ilogit(mlogit$lin)
  1       2       3       4       5
0.089172 0.238323 0.500000 0.761677 0.910828

```

Notice the need to distinguish between predictions in the scale of the response and the link. Now compare the logit, probit and complementary log-log fits:

```

predval <- sapply(list(mlogit,mprobit,mcloglog,mcauchit),fitted)
dimnames(predval) <- list(0:4,c("logit","probit","cloglog","cauchit"))
round(predval,3)
  logit probit cloglog cauchit
0 0.089  0.084  0.127  0.119
1 0.238  0.245  0.250  0.213
2 0.500  0.498  0.455  0.506
3 0.762  0.752  0.722  0.791
4 0.911  0.914  0.933  0.882

```

These are not very different, but now look at a wider range from  $[-4, 8]$ . We apply the `predict` function to each of the four models forming a matrix of predicted values. We label the columns and add the information about the dose. The `tidyR` package is useful for reformatting data from a wide format of multiple measured values per row to a long format where there is only one response value per row. This is accomplished using the `gather()` function. This format, where each row is an observation and each column is a variable, is the most convenient form for many R analyses. Finally, the `ggplot2` package is useful for completing and well-labeled plot.

```

dose <- seq(-4,8,0.2)
predval <- sapply(list(mlogit,mprobit,mcloglog,mcauchit),function(m)
  ↪ predict(m,data.frame(conc=dose),type="response"))
colnames(predval) <- c("logit","probit","cloglog","cauchit")
predval <- data.frame(dose,predval)
library(tidyR)
mpv <- gather(predval, link, probability, -dose)
library(ggplot2)
ggplot(mpv, aes(x=dose,y=probability,linetype=link))+geom_line()

```

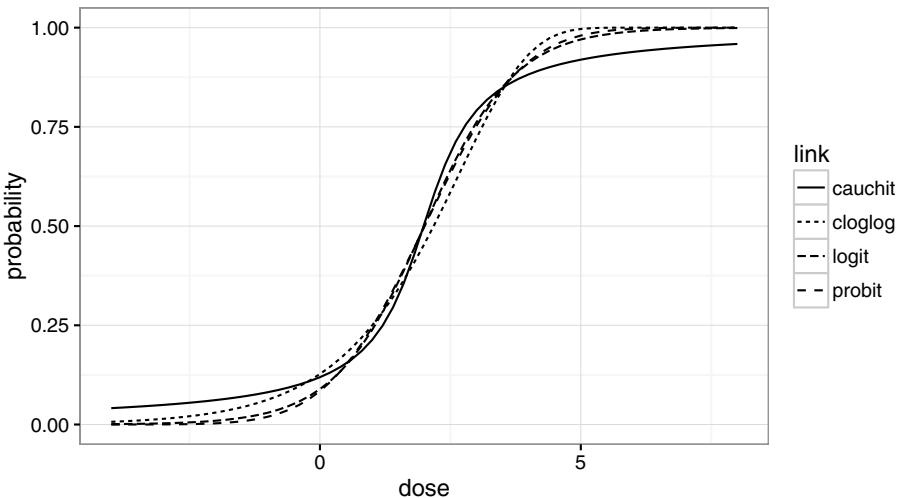


Figure 4.2 Probit, logit, complementary log-log and cauchit compared. The data range from 0 to 4. We see that the links are similar in this range and only begin to diverge as we extrapolate.

The predicted probabilities as shown in Figure 4.2 do not differ much in the range of the data but the differences become more apparent as we extrapolate. The cauchit link is most different as the corresponding latent variable is most variable and would be slowest to converge to zero on the left and one on the right. The complementary log-log can also be clearly distinguished from the logit and probit. But the logit and probit are harder to separate.

The difference between the probit and the logit becomes apparent when we examine the ratio of the probabilities in the lower and upper tails:

```
ggplot(predval, aes(x=dose, y=probit/logit))+geom_line() + xlim(c(-4, 0))
ggplot(predval, aes(x=dose, y=(1-probit)/(1-logit)))+geom_line() + xlim(c
  ↪ (-4, 8))
```

In Figure 4.3, we see that the probit and logit differ substantially in the tails. The same phenomenon is observed for the complementary log-log. This is problematic since the former plot indicates it would be difficult to distinguish between the two using the data we have. This is an issue in trials of potential carcinogens and other substances that must be tested for possible harmful effects on humans. Some substances are highly poisonous in that their effects become immediately obvious at doses that might normally be experienced in the environment. It is not difficult to detect such substances. However, there are other substances whose harmful effects only become apparent at large dosages where the observed probabilities are sufficiently larger than zero to become estimable without immense sample sizes. In order to estimate the probability of a harmful effect at a low dose, it would be necessary to select an appropriate link function and yet the data for high dosages will be of little help in doing this. As Paracelsus (1493–1541) said, “All substances are poisons; there is none which is not a poison. The right dose differentiates a poison.”

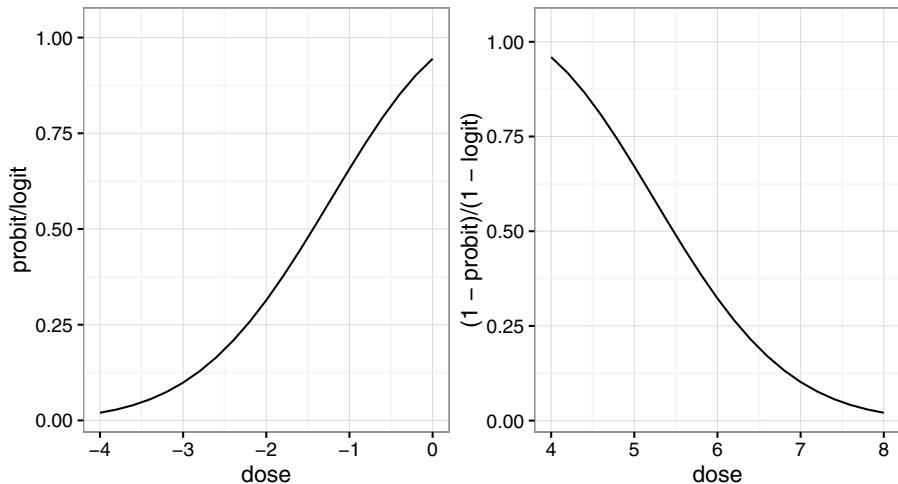


Figure 4.3 *Ratio of predicted probit to logit probabilities in the lower tail on the left and in the upper tail to the right.*

A good example of this problem is asbestos. Information regarding the harmful effects of asbestos derives from historical studies of workers in industries exposed to very high levels of asbestos dust. However, we would like to know the risk to individuals exposed to low levels of asbestos dust such as those found in old buildings. It is very difficult to accurately determine this risk. We cannot accurately measure exposure or outcome. This is not to argue that nothing should be done, but that decisions should be made in recognition of the uncertainties.

We must choose a link function to specify a binomial regression model. It is usually not possible to make this choice based on the data alone. For regions of moderate  $p$ , that is not close to zero or one, the link functions we have proposed are quite similar and so a very large amount of data would be necessary to distinguish between them. Larger differences are apparent in the tails, but for very small  $p$ , one needs a very large amount of data to obtain just a few successes, making it expensive to distinguish between link functions in this region. So usually, the choice of link function is made based on assumptions derived from physical knowledge or simple convenience.

The default choice is the logit link. There are three advantages: it leads to simpler mathematics than the probit due the intractability of  $\Phi$ , it is easier to interpret using odds and it allows easier analysis of retrospectively sampled data as we shall see in the next section.

### 4.3 Prospective and Retrospective Sampling

Consider the data shown in Table 4.1 from a study on infant respiratory disease which shows the proportions of children developing bronchitis or pneumonia in their first year of life by type of feeding and sex, which may be found in Payne (1987):

	Bottle Only	Some Breast with Supplement	Breast Only
Boys	77/458	19/147	47/494
Girls	48/384	16/127	31/464

Table 4.1 *Incidence of respiratory disease in infants to the age of 1 year.*

We can recover the layout above with the proportions as follows:

```
data(babyfood, package="faraway")
xtabs(disease/(disease + nondisease) ~ sex + food, babyfood)
  food
sex     Bottle   Breast   Suppl
Boy  0.16812 0.095142 0.12925
Girl 0.12500 0.066810 0.12598
```

In *prospective* sampling, the predictors are fixed and then the outcome is observed. This is also called a *cohort study*. In the infant respiratory disease example shown in Table 4.1, we would select a sample of newborn girls and boys whose parents had chosen a particular method of feeding and then monitor them for their first year.

In *retrospective* sampling, the outcome is fixed and then the predictors are observed. This is also called a *case-control study*. Typically, we would find infants coming to a doctor with a respiratory disease in the first year and then record their sex and method of feeding. We would also obtain a sample of respiratory disease-free infants and record their information. The method for obtaining the samples is important — we require that the probability of inclusion in the study is independent of the predictor values.

Since the question of interest is how the predictors affect the response, prospective sampling seems to be required. Let's focus on just boys who are breast or bottle fed. The data we need is:

```
babyfood[c(1,3),1]
  disease nondisease sex   food
1      77          381 Boy  Bottle
3      47          447 Boy Breast
```

As we have seen in Section 2.2, the log-odds is a sensible way to measure the strength of the association of a predictor with the outcome.

- Given the infant is *breast* fed, the log-odds of having a respiratory disease are  $\log 47/447 = -2.25$ .
- Given the infant is *bottle* fed, the log-odds of having a respiratory disease are  $\log 77/381 = -1.60$ .

The difference between these two log-odds,  $\Delta = -1.60 - -2.25 = 0.65$ , represents the increased risk of respiratory disease incurred by bottle feeding relative to breast feeding. This is the log-odds ratio. In this case, the log-odds ratio is positive indicating a greater risk of respiratory disease for bottle-fed compared to breast-fed babies.

Now suppose that this had been a retrospective study — we could compute the log-odds of feeding type given respiratory disease status and then find the difference.

Notice that this would give the same result because:

$$\Delta = \log 77/47 - \log 381/447 = \log 77/381 - \log 47/447 = 0.65$$

This shows that a retrospective design is as effective as a prospective design for estimating  $\Delta$ . This manipulation is not possible for other links such as the probit so we have to use the logit.

Retrospective designs have several advantages compared to prospective designs. In this example, we would need to wait a year before the outcome for a particular baby is known when using the cohort approach. For many other responses such as cancer or heart disease and even nonmedical outcomes related to career achievements, we may need to wait a long time before the results are known. In contrast, the case-control study can be done quickly. Also, we can investigate many possible predictors using the retrospective approach whereas in the prospective study, we must specify these in advance and we may not be gifted with foresight. Retrospective studies are particularly valuable for rare outcomes, otherwise a very large cohort will be necessary to guarantee any “positive” responses.

Prospective designs have other advantages. They are less susceptible to bias in the selection of the sample. Retrospective studies rely on historical records which may be of unknown accuracy and completeness. They may also rely on the memory of the subject which may be unreliable. Prospective studies also allow the study of more than one outcome. For example, we might also measure babies’ success in learning to walk. This is problematic in the case-control approach because more than one outcome would lead to different samples. We shall also see that only prospective studies can generate models that predict the probability of an outcome.

In most practical situations, we will also need to account for the effects of covariates  $X$ . Let  $\pi_0$  be the probability that an individual is included in the study if they do *not* have the disease, while let  $\pi_1$  be the probability of inclusion if they do have the disease. For a prospective study,  $\pi_0 = \pi_1$  because we have no knowledge of the outcome, while for a retrospective study typically  $\pi_1$  is much greater than  $\pi_0$ . Suppose that for given  $x$ ,  $p^*(x)$  is the conditional probability that an individual has the disease given that he or she was included in the study, while  $p(x)$  is the unconditional probability that he or she has the disease as we would obtain from a prospective study. Now by Bayes theorem:

$$p^*(x) = \frac{\pi_1 p(x)}{\pi_1 p(x) + \pi_0(1 - p(x))}$$

which can be rearranged to show that:

$$\text{logit}(p^*(x)) = \log \frac{\pi_1}{\pi_0} + \text{logit}(p(x))$$

So the only difference between the retrospective and the prospective study would be the difference in the intercept:  $\log(\pi_1/\pi_0)$ . Generally  $\pi_1/\pi_0$  would not be known, so we would not be able to estimate  $\beta_0$ , but knowledge of the other  $\beta$  would be most important since this can be used to assess the *relative* effect of the covariates.

We could not, however, estimate the absolute effect. Hence, a retrospective study might tell us that a particular lifestyle choice would double the risk of an unpleasant outcome but it cannot tell us the probability of that outcome. If the probability is not high, we may not care.

#### 4.4 Prediction and Effective Doses

Sometimes we wish to predict the outcome for given values of the covariates. For binomial data this will mean estimating the probability of success. Given covariates  $x_0$ , the predicted response on the link scale is  $\hat{\eta} = x_0\hat{\beta}$  with variance given by  $x_0^T(X^TWX)^{-1}x_0$ . Approximate confidence intervals may be obtained using a normal approximation. To get an answer in the probability scale, it will be necessary to transform back using the inverse of the link function. We predict the response for the insect data:

```
data(bliss, package="faraway")
lmod <- glm(cbind(dead,alive) ~ conc, family=binomial, data=bliss)
lmodsum <- summary(lmod)
```

We show how to predict the response at a dose of 2.5:

```
x0 <- c(1,2.5)
eta0 <- sum(x0*coef(lmod))
ilogit(eta0)
[1] 0.64129
```

A 64% predicted chance of death at this dose — now compute a 95% confidence interval (CI) for this probability. First, extract the variance matrix of the coefficients:

```
(cm <- lmodsum$cov.unscaled)
            (Intercept)      conc
(Intercept)    0.174630 -0.065823
conc          -0.065823  0.032912
```

The standard error on the logit scale is then:

```
se <- sqrt( t(x0) %*% cm %*% x0)
```

so the CI on the probability scale is:

```
ilogit(c(eta0-1.96*se, eta0+1.96*se))
[1] 0.53430 0.73585
```

A more direct way of obtaining the same result is:

```
predict(lmod, newdata=data.frame(conc=2.5), se=T)
$fit
[1] 0.58095
```

```
$se.fit
[1] 0.2263
ilogit(c(0.58095-1.96*0.2263, 0.58095+1.96*0.2263))
[1] 0.53430 0.73585
```

Note that in contrast to the linear regression situation, there is no distinction possible between confidence intervals for a future observation and those for the mean response. Now we try predicting the response probability at the low dose of  $-5$ :

```
x0 <- c(1,-5)
se <- sqrt( t(x0) %*% cm %*% x0)
eta0 <- sum(x0*lmod$coeff)
ilogit(c(eta0-1.96*se, eta0+1.96*se))
[1] 2.3577e-05 3.6429e-03
```

This is not a wide interval in absolute terms, but in relative terms, it certainly is. The upper limit is about 100 times larger than the lower limit.

When there is a single (continuous) covariate or when other covariates are held fixed, we sometimes wish to estimate the value of  $x$  corresponding to a chosen  $p$ . For example, we may wish to determine which dose,  $x$ , will lead to a probability of success  $p$ . ED50 stands for the *effective dose* for which there will be a 50% chance of success. When the objective is to kill the subjects or determine toxicity, as when using insecticides, the term LD50 would be used. LD stands for *lethal dose*. Other percentiles are also of interest. For a logit link, we can set  $p = 1/2$  and then solve for  $x$  to find:

$$\widehat{\text{ED50}} = -\hat{\beta}_0/\hat{\beta}_1$$

Using the Bliss data, the LD50 is:

```
(ld50 <- -lmod$coef[1]/lmod$coef[2])
(Intercept)
2
```

To determine the standard error, we can use the delta method. The general expression for the variance of  $g(\hat{\theta})$  for multivariate  $\theta$  is given by

$$\text{var } g(\hat{\theta}) \approx g'(\hat{\theta})^T \text{var } \hat{\theta} g'(\hat{\theta})$$

which, in this example, works out as:

```
dr <- c(-1/lmod$coef[2], lmod$coef[1]/lmod$coef[2]^2)
sqrt(dr %*% lmodsum$cov.un %*% dr)[,]
[1] 0.17844
```

So the 95% CI is given by:

```
c(2-1.96*0.178, 2+1.96*0.178)
[1] 1.6511 2.3489
```

Other levels may be considered — the effective dose  $x_p$  for probability of success  $p$  is:

$$x_p = \frac{\text{logit}(p) - \beta_0}{\beta_1}$$

So, for example:

```
(ed90 <- (logit(0.9)-lmod$coef[1])/lmod$coef[2])
(Intercept)
3.8911
```

More conveniently, we may use the `dose.p` function in the MASS package:

```
library(MASS)
dose.p(lmod, p=c(0.5, 0.9))
      Dose          SE
p = 0.5: 2.0000 0.17844
p = 0.9: 3.8911 0.34499
```

## 4.5 Matched Case-Control Studies

In a case-control study, we try to determine the effect of certain risk factors on the outcome. We understand that there are other confounding variables that may affect the outcome. One approach to dealing with these is to measure or record them, include them in the logistic regression model as appropriate and thereby control for

their effect. But this method requires that we model these confounding variables with the correct functional form. This may be difficult. Also, making an appropriate adjustment is problematic when the distribution of the confounding variables is quite different in the cases and controls. So we might consider an alternative where the confounding variables are explicitly adjusted for in the design.

In a *matched case-control study*, we match each case (diseased person, defective object, success, etc.) with one or more controls that have the same or similar values of some set of potential confounding variables. For example, if we have a 56-year-old, Hispanic male case, we try to match him with some number of controls who are also 56-year-old Hispanic males. This group would be called a *matched set*. Obviously, the more confounding variables one specifies, the more difficult it will be to make the matches. Loosening the matching requirements, for example, accepting controls who are 50–60 years old, might be necessary. Matching also gives us the possibility of adjusting for confounders that are difficult to measure. For example, suppose we suspect an environmental effect on the outcome. However, it is difficult to measure exposure, particularly when we may not know which substances are relevant. We could match subjects based on their place of residence or work. This would go some way to adjusting for the environmental effects.

Matched case-control studies also have some disadvantages apart from the difficulties of forming the matched sets. One loses the possibility of discovering the effects of the variables used to determine the matches. For example, if we match on sex, we will not be able to investigate a sex effect. Furthermore, the data will likely be far from a random sample of the population of interest. So although relative effects may be found, it may be difficult to generalize to the population.

Sometimes, cases are rare but controls are readily available. A  $1 : M$  design has  $M$  controls for each case.  $M$  is typically small and can even vary in size from matched set to matched set due to difficulties in finding matching controls and missing values. Each additional control yields a diminished return in terms of increased efficiency in estimating risk factors — it is usually not worth exceeding  $M = 5$ .

For individual  $i$  in the  $j^{th}$  matched set, we also observe a covariate vector  $x_{ij}$  which will include the risk factors of interest plus any other variables that we may wish to adjust for, but were unable for various reasons to include among the criteria used to match the sets. It is important that the decision to include a subject in the study be independent of the risk factors as in the unmatched case-control studies. Suppose we have  $n$  matched sets and that we take  $i = 0$  to represent the case and  $i = 1, \dots, M$  to represent the controls. We propose a logistic regression model of the following form:

$$\text{logit}(p_j(x_{ij})) = \alpha_j + \beta^T x_{ij}$$

The  $\alpha_j$  models the effect of the confounding variables in the  $j^{th}$  matched set. Given a matched set  $j$  of  $M + 1$  subjects known to have one case and  $M$  controls, the conditional probability of the observed outcome, or, in other words, that subject  $i = 0$  is the case and the rest are controls, is:

$$\frac{\exp \beta^T x_{0j}}{\sum_{i=0}^M \exp \beta^T x_{ij}}$$

Notice that  $\alpha_j$  cancels out in this expression. We may then form the conditional likelihood for the model by taking the product over all the matched sets:

$$L(\beta) = \prod_{j=1}^n \left\{ 1 + \sum_{i=1}^M \exp[\beta^T (x_{ij} - x_{0j})] \right\}^{-1}$$

We may now employ standard likelihood methods to make inference — see Breslow (1982) for details. The likelihood takes the same form as that used for the proportional hazards model used in survival analysis. This is convenient because we may use software developed for those models as we demonstrate below. Since the  $\alpha$ s are not estimated, we cannot make predictions about individuals, but only make statements about the relative risks as measured by the  $\beta$ s. This same restriction also applies to the unmatched model, so this is nothing new.

In Le (1998), a matched case-control study is presented concerning the association between x-rays and childhood acute myeloid leukemia. The sets are matched on age, race and county of residence. For the most part, there is only one control for each case, but there are a few instances of two controls. We start with a look at the data:

```
data(amlxray, package="faraway")
head(amlxray)
```

	ID	disease	Sex	downs	age	Mray	MupRay	MlowRay	Fray	Cray	CnRay	
1	7004		F	no	0	no		no	no	no	no	1
2	7004		F	no	0	no		no	no	no	no	1
3	7006		M	no	6	no		no	no	yes		3
4	7006		M	no	6	no		no	no	yes		2
5	7009		F	no	8	no		no	no	no	no	1
6	7009		F	no	8	no		no	no	no	no	1

Only the age is presented here as one of the matching variables. In the three sets shown here, we see that both subjects have the same age and the first is the case and the second is the control. The other variables are risk factors of interest.

Downs syndrome is known to be a risk factor. There are only seven such subjects in the dataset:

```
amlxray[amlxray$downs=="yes", 1:4]
```

	ID	disease	Sex	downs
7	7010		M	yes
17	7018		F	yes
78	7066		F	yes
88	7077		M	yes
173	7146		F	yes
196	7176		F	yes
210	7189		F	yes

We see that all seven subjects are cases. If we include this variable in the regression, its coefficient is infinite. Given this and the prior knowledge, it is simplest to exclude all these subjects and their associated matched subjects:

```
(ii <- which(amlxray$downs=="yes"))
[1] 7 17 78 88 173 196 210
amlxray <- amlxray[-c(ii, ii+1), ]
```

The variables Mray, MupRay and MlowRay record whether the mother has ever had an x-ray, ever had an upper body x-ray and ever had a lower body x-ray, respectively.

These variables are closely associated, so we will pick just Mray for now and investigate the others more closely if indicated. We will also use CnRay, a four-level ordered factor grouping the number of x-rays that the child has received in preference to Cray which merely indicates whether the child has ever had an x-ray.

The clogit function fits a conditional logit model. Since the likelihood is identical with that from a proportional hazards model, it may be found in the survival package. The matched sets must be designated by the strata function:

```
library(survival)
cmod <- clogit(disease ~ Sex+Mray+Fray+CnRay+strata(ID), ramlxray)
summary(cmod)
```

	coef	exp(coef)	se(coef)	z	p
SexM	0.156	1.17	0.386	0.405	0.6900
Mrayyes	0.228	1.26	0.582	0.391	0.7000
Frayyes	0.693	2.00	0.351	1.974	0.0480
CnRay.L	1.941	6.96	0.621	3.127	0.0018
CnRay.Q	-0.248	0.78	0.582	-0.426	0.6700
CnRay.C	-0.580	0.56	0.591	-0.982	0.3300

	exp(coef)	exp(-coef)	lower .95	upper .95
SexM	1.17	0.855	0.549	2.49
Mrayyes	1.26	0.796	0.401	3.93
Frayyes	2.00	0.500	1.005	3.98
CnRay.L	6.96	0.144	2.063	23.51
CnRay.Q	0.78	1.281	0.249	2.44
CnRay.C	0.56	1.786	0.176	1.78

Rsquare= 0.089 (max possible= 0.499 )  
 Likelihood ratio test= 20.9 on 6 df, p=0.00192  
 Wald test = 14.5 on 6 df, p=0.0246  
 Score (logrank) test = 18.6 on 6 df, p=0.0049

The overall tests for significance of the predictors indicate that at least some of the variables are significant. We see that Sex and whether the mother had an x-ray are not significant. There seems little point in investigating the other x-ray variables associated with the mother. An x-ray on the father is marginally significant. However, the x-ray on the child has the clearest effect. Because this is an ordered factor, we have used linear, quadratic and cubic contrasts. Only the linear effect is significant.

The second table of coefficients gives us information helpful for interpreting the size of the effects. We see that the father having had an x-ray doubles the odds of the disease. The interpretation of the number of x-rays of the child is more difficult to interpret because of the coding. Since we have found only a linear effect, we convert CnRay to the numerical values 1–4 using unclass. We also drop the insignificant predictors:

```
cmodr <- clogit(disease ~ Fray+unclass(CnRay)+strata(ID), ramlxray)
summary(cmodr)
```

	coef	exp(coef)	se(coef)	z	p
Frayyes	0.670	1.96	0.344	1.95	0.05100
unclass(CnRay)	0.814	2.26	0.237	3.44	0.00058

	exp(coef)	exp(-coef)	lower .95	upper .95
Frayyes	1.96	0.512	0.996	3.84
unclass(CnRay)	2.26	0.443	1.419	3.59

The codes for Cnray are 1 = none, 2 = 1 or 2 x-rays, 3 = 3 or 4 x-rays and 4 = 5 or

more x-rays. We see that the odds of the disease increase by a factor of 2.26 as we move between adjacent categories. Notice that the father's x-ray variable is now just insignificant in this regression underlining its borderline status.

An incorrect analysis of this data ignores the matching structure and simply uses a binomial GLM:

```
gmod <- glm(disease ~ Fray+unclass(CnRay), family=binomial, ramlxray)
```

```
summary(gmod)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.162	0.301	-3.86	0.00011
Frayyes	0.500	0.308	1.63	0.10405
unclass(CnRay)	0.601	0.177	3.39	0.00071

The results are somewhat different.

Although we have found an effect due to x-rays of the child, we cannot conclude the effect is causal. After all, subjects only have x-rays when something is wrong, so it is quite possible that the x-rays are linked to some unknown causal factor.

Other examples of matched data may be found in Section 6.4.

## Exercises

1. The Chicago insurance dataset found in `chredlin` concerns the problem of redlining in insurance. Read the `help` page for background. Use `involact` as the response and ignore `volact`.
  - (a) Plot a histogram of the distribution of `involact` taking care to choose the bin width to illustrate the issue with zero values. What fraction of the responses is zero?
  - (b) Fit a Gaussian linear model with `involact` as the response with the other five variables as predictors. Use a log transformation for income. Describe the relationship between these predictors and the response.
  - (c) Plot the residuals against the fitted values. How are the zero response values manifested on the plot? What impact do these cases have on the interpretation of the plot?
  - (d) Create a binary response variable which distinguishes zero values of `involact`. Fit a logistic regression model with this response but with the same five predictors. What problem occurred during this fit? Explain why this happened.
  - (e) Fit a smaller model using only `race` and `age`. Interpret the z-statistics. Test for the significance of the two predictors using the difference-in-deviances test. Which test for the significance of the predictors should be preferred?
  - (f) Make plot of `race` against `age` which also distinguishes the two levels of the response variable. Interpret the plot and connect it to the previous model output.
  - (g) Refit the logit model but use a probit link. Compare the model output between the logit and probit models. Which parts are similar and which parts differ substantively? Plot the predicted values on the probability scale against each other and comment on what you see.

- (h) Which binary response model is most comparable to the Gaussian linear model? Contrast the drawbacks between the Gaussian and binary response models.
2. Aflatoxin B1 was fed to lab animals at various doses and the number responding with liver cancer recorded. The data may be found in the dataset `aflatoxin`.
- Plot the proportion of animals with liver cancer against the dose.
  - Fit a logistic regression for the number of animals with liver cancer as a function of the dose. Assess the statistical significance of the dose with two different tests. Which test is preferable?
  - For the fitted model, compute the predicted probability of liver cancer over the range of the data. Show the predicted curve on top of the data shown in (a).
  - Change to a probit link. Compute the predicted probability under this model and add the curve to the existing plot.
  - Compute the predicted probability of liver cancer for a dose of 25 ppb on the link scale. Compute a 95% confidence interval. Transform onto the probability scale.
  - Compute the predicted probability directly on the probability scale and report the standard error. Use this standard error to construct a 95% confidence interval for the probability. Which of the two methods of computing the CI as seen in this and the previous question is preferable?
  - Compute the effective dose at 1%, 10% and at 50% levels with associated standard error. Remark on any incongruity in your calculated effective doses.
3. The `infert` dataset presents data from a study of secondary infertility (failure to conceive after at least one previous conception). The factors of interest are induced abortions and spontaneous abortions (e.g., miscarriages). The study matched each case of infertility with two controls who were not infertile, matching on age, education and parity (number of prior pregnancies).
- Construct cross-classified tables by number of spontaneous and induced abortions separately for cases and controls. Comment on the differences between the two tables.
  - Fit a binary response model with only spontaneous and induced as predictors. Determine the statistical significance of these predictors. Express the effects of the predictors in terms of odds.
  - Fit a binary response model with only education, age and parity as predictors. Explain how the significance (or lack thereof) of these predictors should be interpreted.
  - Now put all five predictors in a binary response model. Interpret the results in terms of odds.
  - Fit a matched case control model appropriate to the data. Interpret the output and compare the odds to those found in the previous model.
  - The spontaneous and induced predictors could be viewed as ordinal due to the grouping in the highest level. Refit the model using ordinal factors rather than numerical variables for these two predictors. Is there evidence that the ordinal representation is necessary?

4. The data in `downs.bc` found in the `boot` package describes the proportion of births with Down's syndrome by the ages of the mothers collected in British Columbia.
- Plot the proportion of Down's births against age and comment on the relationship.
  - Fit a binomial response model for the Down's births with age as a linear predictor. Comment on the significance of the age effect.
  - Display the predicted probability on the plot from (a). Comment on the quality of the fit.
  - Add a quadratic term in age to the model and display the predicted probability on the same plot. Is the fit improved?
  - Compute the predicted probability of a Down's birth at age 30 and at age 40 using the previous model. What is the odds ratio of these two probabilities? Now compute the predicted response on the link scale at age 30 and age 40. Exponentiate the difference between these two predictions. Explain why the answer is identical to that computed in the first part of this question.
  - Fit a binomial response model with just a linear predictor in age but now using the complementary log-log link. Display the fitted probability curve on top of the observed proportions. Comment on the fit.
  - Exponentiate the coefficient of age in the previous model. Now compute the predicted probability of a Down's birth at age 30 and at age 31. Compute the ratio of these two probabilities. Explain why this ratio is the same as the exponentiated coefficient. Give an interpretation in the context of this data.

**This page intentionally left blank**

## Count Regression

---

When the response is an unbounded count ( $0, 1, 2, 3, \dots$ ), we can use a count regression model to explain this in terms of the given predictors. Sometimes other models may be appropriate. When the count is bounded, a binomial-type response regression as discussed in the previous chapters is sensible. In some cases, the counts might be sufficiently large that a normal approximation is justified so that a normal linear model may be used. We shall consider two distributions for counts: the Poisson and, less commonly, the negative binomial.

### 5.1 Poisson Regression

If  $Y$  is Poisson with mean  $\mu > 0$ , then:

$$P(Y = y) = \frac{e^{-\mu}\mu^y}{y!}, \quad y = 0, 1, 2, \dots$$

Three examples of the Poisson density are depicted in Figure 5.1. In the left panel, we see a distribution that gives highest probability to  $y = 0$  and falls rapidly as  $y$  increases. In the center panel, we see a skew distribution with longer tail on the right. Even for a not so large  $\mu = 5$ , we see the distribution become more normally shaped. This becomes more pronounced as  $\mu$  increases.

```
barplot(dpois(0:5, 0.5), xlab="y", ylab="Probability", names=0:5, main="  
    ↪ mean = 0.5")  
barplot(dpois(0:10, 2), xlab="y", ylab="Probability", names=0:10, main="  
    ↪ mean = 2")  
barplot(dpois(0:15, 5), xlab="y", ylab="Probability", names=0:15, main="  
    ↪ mean = 5")
```

The expectation and variance of a Poisson are the same:  $EY = \text{var } Y = \mu$ . The Poisson distribution arises naturally in several ways:

1. If the count is some number out of some possible total, then the response would be more appropriately modeled as a binomial. However, for small success probabilities and large totals, the Poisson is a good approximation and can be used. For example, in modeling the incidence of rare forms of cancer, the number of people affected is a small proportion of the population in a given geographical area. Specifically, if  $\mu = np$  while  $n \rightarrow \infty$ , then  $B(n, p)$  is well approximated by  $Pois(\mu)$ . Also, for small  $p$ , note that  $\text{logit}(p) \approx \log p$ , so that the use of the Poisson with a log link is comparable to the binomial with a logit link. Where  $n$  varies between cases, a *rate model* can be used as described in Section 5.3.

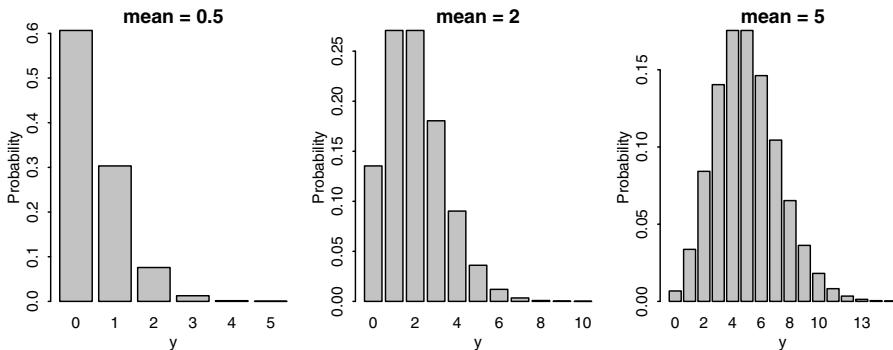


Figure 5.1 *Poisson probabilities for  $\mu = 0.5, 2$  and  $5$ , respectively.*

2. Suppose the probability of occurrence of an event in a given time interval is proportional to the length of that time interval and independent of the occurrence of other events. In this case the number of events in any specified time interval will be Poisson distributed. Examples include modeling the number of incoming telephone calls to a service center or the number of earthquakes in a fixed period of time. However, in any real application, the assumptions are likely to be violated. For example, the rate of incoming telephone calls is likely to vary with the time of day while the timing of earthquakes are unlikely to be completely independent. Nevertheless, the Poisson may be a good approximation.
3. Poisson distributions also arise naturally when the time between events is independent and identically exponentially distributed. We count the number of events in a given time period. This is effectively equivalent to the previous case, since the exponential distribution between events will result from the assumption of constant and independent probability of occurrence of an event in an interval.

If the count is the number falling into some level of a given category, then a multinomial response model or categorical data analysis should be used. For example, if we have counts of how many people have type O, A, B or AB blood and are interested in how that relates to race and gender, then a straight Poisson regression model will not be appropriate. We will see later that the Poisson distribution still comes into play in Chapter 7.

An important result concerning Poisson random variables is that their sum is also Poisson. Specifically, suppose that  $Y_i \sim Pois(\mu_i)$  for  $i = 1, 2, \dots$  and are independent, then  $\sum_i Y_i \sim Pois(\sum_i \mu_i)$ . This is useful because sometimes we have access only to the aggregated data. If we assume the individual-level data is Poisson, then so is the summed data and Poisson regression can still be applied.

For 30 Galápagos Islands, we have a count of the number of plant species found on each island and the number that are endemic to that island. We also have five geographic variables for each island. The data was presented by Johnson and Raven (1973) and also appear in Weisberg (2005). We have filled in a few missing values

that appeared in the original dataset for simplicity. We model the number of species using normal linear regression:

```
data(gala, package="faraway")
gala <- gala[,-2]
```

We throw out the Endemics variable (which falls in the second column of the dataframe) since we won't be using it in this analysis. We fit a linear regression and look at the residual vs. fitted plot:

```
mod1 <- lm(Species ~ . , gala)
plot(mod1, 1)
```

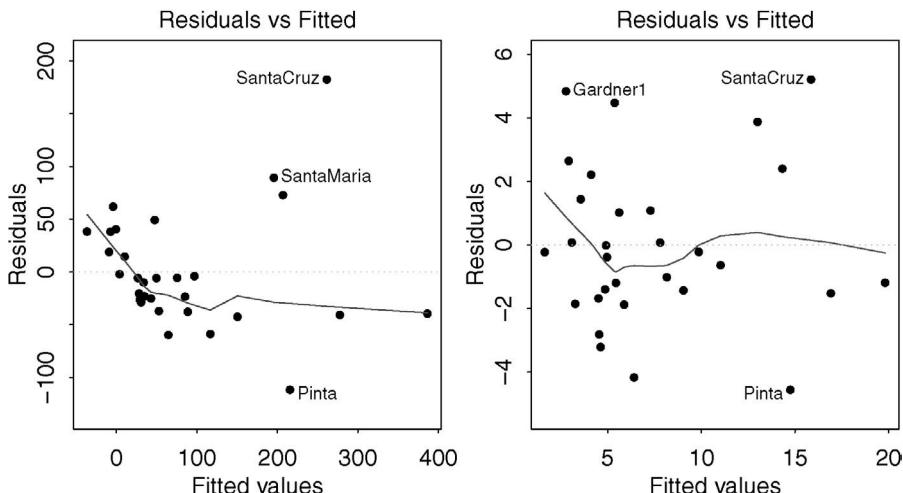


Figure 5.2 *Residual-fitted plots for the Galápagos dataset. The plot on the left is for a model with the original response while that on the right is for the square-root transformed response.*

We see clear evidence of nonconstant variance in the left panel of Figure 5.2. Some experimentation (or the use of the Box–Cox method) reveals that a square-root transformation is best:

```
modt <- lm(sqrt(Species) ~ . , gala)
plot(modt, 1)
```

We now see in the right panel of Figure 5.2 that the nonconstant variance problem has been cleared up. Let's take a look at the fit:

```
library(faraway)
summary(modt)
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.391924	0.871268	3.89	0.00069
Area	-0.001972	0.001020	-1.93	0.06508
Elevation	0.016478	0.002441	6.75	5.5e-07
Nearest	0.024933	0.047950	0.52	0.60784
Scruz	-0.013483	0.009798	-1.38	0.18151
Adjacent	-0.003367	0.000805	-4.18	0.00033

$n = 30$ ,  $p = 6$ , Residual SE = 2.774, R-Squared = 0.78

We see a fairly good fit ( $R^2 = 0.78$ ) considering the nature of the variables. However, we achieved this fit at the cost of transforming the response. This makes interpretation more difficult. Furthermore, some of the response values are quite small (single digits) which makes us question the validity of the normal approximation. This model may be adequate, but perhaps we can do better. We develop a Poisson regression model.

Suppose we have count responses  $Y_i$  that we wish to model in terms of a vector of predictors  $x_i$ . Now if  $Y_i \sim \text{Pois}(\mu_i)$ , we need some way to link the  $\mu_i$  to the  $x_i$ . We use a linear combination of the  $x_i$  to form the linear predictor  $\eta_i = x_i^T \beta$ . Since we require that  $\mu_i \geq 0$ , we can ensure this by using a log link function, that is:

$$\log \mu_i = \eta_i = x_i^T \beta$$

So, as with the binomial regression models of the previous chapter, this model also has a linear predictor and a link function. The log-likelihood is:

$$l(\beta) = \sum_{i=1}^n (y_i x_i^T \beta - \exp(x_i^T \beta) - \log(y_i!))$$

Differentiating with respect to  $\beta_j$  gives the MLE as the solution to:

$$\sum_{i=1}^n (y_i - \exp(x_i^T \hat{\beta})) x_{ij} = 0 \quad \forall j$$

which can be more compactly written as:

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

The normal equations for the least squares estimate of  $\beta$  in Gaussian linear models take the same form when we set  $\hat{\mu} = X \hat{\beta}$ . The equations for  $\beta$  for a binomial regression with a logit link also take the same form. This would not be true for other link functions. The link function having this property is known as the *canonical link*.

However, there is no explicit formula for  $\hat{\beta}$  for the Poisson (or binomial) regression and we must resort to numerical methods to find a solution. We fit the Poisson regression model to the Galápagos data:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	3.1548079	0.0517495	60.96	< 2e-16
Area	-0.0005799	0.0000263	-22.07	< 2e-16
Elevation	0.0035406	0.0000874	40.51	< 2e-16
Nearest	0.0088256	0.0018213	4.85	0.0000013
Scruz	-0.0057094	0.0006256	-9.13	< 2e-16
Adjacent	-0.0006630	0.0000293	-22.61	< 2e-16

n = 30 p = 6

Deviance = 716.846 Null Deviance = 3510.729 (Difference = 2793.883)

Using the same arguments as for binary response regression as seen in Section 2.3, we develop a deviance for the Poisson regression:

$$D = 2 \sum_{i=1}^n (y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i))$$

This Poisson deviance is also known as the *G-statistic*.

The same asymptotic inference may be employed as for the binomial model. We can judge the goodness of fit of a proposed model by checking the deviance of the model against a  $\chi^2$  distribution with degrees of freedom equal to that of the model. We can compare nested models by taking the difference of the deviances and comparing to a  $\chi^2$  distribution with degrees of freedom equal to the difference in the number of parameters for the two models. We can test the significance of individual predictors and construct confidence intervals for  $\beta$  using the standard errors,  $se(\hat{\beta})$ , although, as before, it is better to use profile likelihood methods.

An alternative and perhaps better-known goodness of fit measure is the Pearson's  $X^2$  statistic:

$$X^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

In this example, we see that the residual deviance is 717 on 24 degrees of freedom, which indicates an ill-fitting model if indeed the Poisson is the correct model for the response. A common drawback in using the standard Poisson regression is that responses are often more variable than the Poisson model would imply. The Poisson distribution has only one parameter  $\mu$  which represents both the mean and the variance. In contrast, the standard linear model has a variance  $\sigma^2$  that is independent of the mean  $\mu$  and hence is more flexible.

## 5.2 Dispersed Poisson Model

We can modify the standard Poisson model to allow for more variation in the response. But before we do that, we must check whether the large size of deviance might be related to some other cause.

In the Galápagos example, we check the residuals to see if the large deviance can be explained by an outlier:

```
halfnorm(residuals(modp))
```

The half-normal plot of the (absolute value of the) residuals shown in Figure 5.3 shows no outliers. It could be that the structural form of the model needs some improvement, but some experimentation with different forms for the predictors will reveal that there is little scope for improvement. Furthermore, the proportion of deviance explained by this model,  $1 - 717/3510 = 0.796$ , is about the same as in the linear model above.

For a Poisson distribution, the mean is equal to the variance. Let's investigate this relationship for this model. It is difficult to estimate the variance for a given value of the mean, but  $(y - \hat{\mu})^2$  does serve as a crude approximation. We plot this estimated variance against the mean, as seen in the second panel of Figure 5.3:

```
plot(log(fitted(modp)), log((gala$Species-fitted(modp))^2), xlab=
  ↪ expression(hat(mu)), ylab=expression((y-hat(mu))^2))
abline(0, 1)
```

We see that the variance is proportional to, but larger than, the mean. When the variance assumption of the Poisson regression model is broken but the link function and choice of predictors are correct, the estimates of  $\beta$  are consistent, but the standard er-

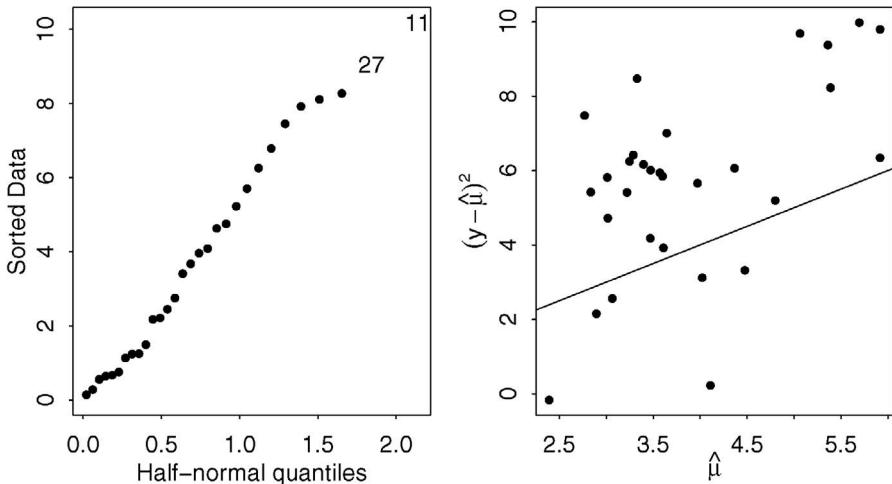


Figure 5.3 *Half-normal plot of the residuals of the Poisson model is shown on the left and the relationship between the mean and variance is shown on the right. A line representing mean equal to variance is also shown.*

rors will be wrong. We cannot determine which predictors are statistically significant in the above model using the output we have.

The Poisson distribution has only one parameter and so is not very flexible for empirical fitting purposes. We can generalize by allowing ourselves a dispersion parameter. Over- or underdispersion can occur in various ways in Poisson models. For example, suppose the Poisson response  $Y$  has rate  $\lambda$  which is itself a random variable. The tendency to fail for a machine may vary from unit to unit even though they are the same model. We can model this by letting  $\lambda$  be gamma distributed with  $E\lambda = \mu$  and  $\text{var } \lambda = \mu/\phi$ . Now  $Y$  is negative binomial with mean  $EY = \mu$ . The mean is the same as the Poisson, but the variance  $\text{var } Y = \mu(1 + \phi)/\phi$  which is not equal to  $\mu$ . In this case, overdispersion would occur and could be modeled using a negative binomial model as demonstrated in Section 5.4.

If we know the specific mechanism, as in the above example, we could model the response as a negative binomial or other more flexible distribution. If the mechanism is not known, we can introduce a dispersion parameter  $\phi$  such that  $\text{var } Y = \phi EY = \phi\mu$ .  $\phi = 1$  is the regular Poisson regression case, while  $\phi > 1$  is overdispersion and  $\phi < 1$  is underdispersion.

The dispersion parameter may be estimated using:

$$\hat{\phi} = \frac{X^2}{n - p} = \frac{\sum_i (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i}{n - p}$$

We estimate the dispersion parameter in our example by:

```
(dp <- sum(residuals(modp, type="pearson")^2) / modp$df.res)
[1] 31.749
```

We can then adjust the standard errors and so forth in the summary as follows:

```
summary(modp, dispersion=dp)
    Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.154808  0.291590 10.82 < 2e-16
Area        -0.000580  0.000148 -3.92 8.9e-05
Elevation   0.003541  0.000493  7.19 6.5e-13
Nearest      0.008826  0.010262  0.86  0.39
Scruz       -0.005709  0.003525 -1.62  0.11
Adjacent    -0.000663  0.000165 -4.01 6.0e-05

overdispersion parameter = 31.749
n = 30 p = 6
Deviance = 716.846 Null Deviance = 3510.729 (Difference = 2793.883)
```

Notice that the estimation of the dispersion and the regression parameters is independent, so choosing a dispersion other than one has no effect on the regression parameter estimates. Notice also that there is some similarity in which variables are picked out as significant when compared with the linear regression model.

The same result can be achieved by using the *quasi-Poisson* model which directly incorporates the dispersion parameter:

```
modd <- glm(Species ~ ., family=quasipoisson, gal)
```

When comparing Poisson models with overdispersion, an *F*-test rather than a  $\chi^2$  test should be used. As in normal linear models, the variance, or dispersion parameter in this case, needs to be estimated. This requires the use of the *F*-test. We can test the significance of each of the predictors relative to the full model:

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

```
Model:
Species ~ Area + Elevation + Nearest + Scruz + Adjacent
          Df Deviance AIC F value Pr(>F)
<none>      717  890
Area        1    1204 1375  16.32 0.00048
Elevation   1    2390 2560  56.00 1e-07
Nearest      1     739  910   0.76 0.39336
Scruz        1     814  984   3.24 0.08444
Adjacent     1    1341 1512  20.91 0.00012
```

The *z*-statistics from the `summary()` are less reliable and so the *F*-test is preferred. In this example, there is little practical difference between the two.

### 5.3 Rate Models

The number of events observed may depend on a size variable that determines the number of opportunities for the events to occur. For example, if we record the number of burglaries reported in different cities, the observed number will depend on the number of households in these cities. In other cases, the size variable may be time. For example, if we record the number of customers served by a sales worker, we must take account of the differing amounts of time worked.

Sometimes, it is possible to analyze such data using a binomial response model. For the burglary example above, we might model the number of burglaries out of the number of households. However, if the proportion is small, the Poisson approxima-

tion to the binomial is effective. Furthermore, in some examples, the total number of potential cases may not be known exactly. The modeling of rare diseases illustrates this issue as we may know the number of cases but not have precise population data. Sometimes, the binomial model simply cannot be used. In the burglary example, some households may be robbed more than once. In the customer service example, the size variable is not a count. An alternative approach is to model the ratio. However, there are often difficulties with normality and unequal variance when taking this approach, particularly if the counts are small.

In Purott and Reeder (1976), some data is presented from an experiment conducted to determine the effect of gamma radiation on the numbers of chromosomal abnormalities (ca) observed. The number (cells), in hundreds of cells exposed in each run, differs. The dose amount (doseamt) and the rate (doserate) at which the dose is applied are the predictors of interest. We may format the data for observation like this:

```
data(dicentric, package="faraway")
round(xtabs(ca/cells ~ doseamt+doserate, dicentric), 2)
```

doserate	0.1	0.25	0.5	1	1.5	2	2.5	3	4
doseamt	1	0.05	0.05	0.07	0.07	0.06	0.07	0.07	0.07
	2.5	0.16	0.28	0.29	0.32	0.38	0.41	0.41	0.37
	5	0.48	0.82	0.90	0.88	1.23	1.32	1.34	1.24

We can plot the data as seen in the first panel of Figure 5.4:

```
with(dicentric, interaction.plot(doseamt, doserate, ca/cells, legend=
  FALSE))
```

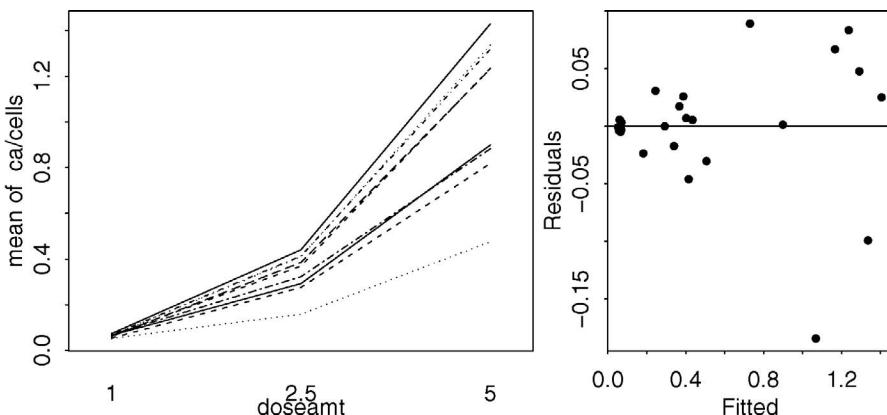


Figure 5.4 Chromosomal abnormalities rate response is shown on the left and a residuals vs. fitted plot of a linear model fit to these data is shown on the right.

We might try modeling the rate directly. We see that the effect of the dose rate may be multiplicative, so we log this variable in the following model:

```
lmod <- lm(ca/cells ~ log(doserate)*factor(doseamt), dicentric)
summary(lmod)$adj
[1] 0.98444
```

As can be seen from the adjusted  $R^2$ , this model fits well. However, a look at the diagnostics reveals a problem, as seen in the second panel of Figure 5.4:

```
plot(residuals(lmod) ~ fitted(lmod), xlab="Fitted", ylab="Residuals")
abline(h=0)
```

We might prefer an approach that directly models the count response. We need to use the log of the number of cells because we expect this to have a multiplicative effect on the response:

```
dicentric$dosef <- factor(dicentric$doseamt)
pmod <- glm(ca ~ log(cells)+log(doserate)*dosef, family=poisson,
             ↪ dicentric)
summary(pmod)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.7653	0.3812	-7.25	4e-13
log(cells)	1.0025	0.0514	19.52	< 2e-16
log(doserate)	0.0720	0.0355	2.03	0.04240
dosef2.5	1.6298	0.1027	15.87	< 2e-16
dosef5	2.7667	0.1229	22.52	< 2e-16
log(doserate):dosef2.5	0.1611	0.0484	3.33	0.00087
log(doserate):dosef5	0.1932	0.0430	4.49	7e-06

n = 27 p = 7

Deviance = 21.748 Null Deviance = 916.127 (Difference = 894.379)

We can relate this Poisson model with a log link back to a linear model for the ratio response:

$$\log(\text{ca}/\text{cells}) = X\beta$$

This can be rearranged as

$$\log \text{ca} = \log \text{cells} + X\beta$$

We are using `log cells` as a predictor. Checking above, we can see that the coefficient of 1.0025 is very close to one. This suggests fitting a model with the coefficient fixed as one. In this manner, we are modeling the rate of chromosomal abnormalities while still maintaining the count response for the Poisson model. This is known as a *rate model*. We fix the coefficient as one by using an *offset*. Such a term on the predictor side of the model equation has no parameter attached:

```
rmod <- glm(ca ~ offset(log(cells))+log(doserate)*dosef, family=
             ↪ poisson, dicentric)
summary(rmod)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.7467	0.0343	-80.16	< 2e-16
log(doserate)	0.0718	0.0352	2.04	0.04130
dosef2.5	1.6254	0.0495	32.86	< 2e-16
dosef5	2.7611	0.0435	63.49	< 2e-16
log(doserate):dosef2.5	0.1612	0.0483	3.34	0.00084
log(doserate):dosef5	0.1935	0.0424	4.56	0.0000051

n = 27 p = 6

Deviance = 21.750 Null Deviance = 4753.004 (Difference = 4731.254)

Not surprisingly, the coefficients are only slightly different from the previous model. We see from the residual deviance that the model fits well. Previous analyses have posited a quadratic effect in dose; indeed, the observed coefficients speak against

a purely linear effect. However, given that we have only three dose levels, we can hardly check whether quadratic is appropriate. Given the significant interaction effect, we can see that the effect of the dose rate is different depending on the overall dose. We can see that the combination of a high dose, delivered quickly, has a greater combined effect than the main effect estimates would suggest. More on the analysis of this data may be found in Frome and DuFrain (1986).

## 5.4 Negative Binomial

Given a series of independent trials, each with probability of success  $p$ , let  $Z$  be the number of trials until the  $k^{\text{th}}$  success. Then:

$$P(Z = z) = \binom{z-1}{k-1} p^k (1-p)^{z-k} \quad z = k, k+1, \dots$$

The negative binomial can arise naturally in several ways. Imagine a system that can withstand  $k$  hits before failing. The probability of a hit in a given time period is  $p$  and we count the number of time periods until failure. The negative binomial also arises from a generalization of the Poisson where the parameter  $\lambda$  is gamma distributed. The negative binomial also comes up as a limiting distribution for urn schemes that can be used to model contagion.

We get a more convenient parameterization if we let  $Y = Z - k$  and  $p = (1 + \alpha)^{-1}$  so that:

$$P(Y = y) = \binom{y+k-1}{k-1} \frac{\alpha^y}{(1 + \alpha)^{y+k}}, \quad y = 0, 1, 2, \dots$$

then  $EY = \mu = k\alpha$  and  $\text{var } Y = k\alpha + k\alpha^2 = \mu + \mu^2/k$ .

The log-likelihood is then:

$$\sum_{i=1}^n \left( y_i \log \frac{\alpha}{1 + \alpha} - k \log(1 + \alpha) + \sum_{j=0}^{y_i-1} \log(j+k) - \log(y_i!) \right)$$

The most convenient way to link the mean response  $\mu$  to a linear combination of the predictors  $X$  is:

$$\eta = x^T \beta = \log \frac{\alpha}{1 + \alpha} = \log \frac{\mu}{\mu + k}$$

We can regard  $k$  as fixed and determined by the application or as an additional parameter to be estimated. More on regression models for negative binomial responses may be found in Cameron and Trivedi (1998) and Lawless (1987).

Consider this example. ATT ran an experiment varying five factors relevant to a wave-soldering procedure for mounting components on printed circuit boards. The response variable, `skips`, is a count of how many solder skips appeared to a visual inspection. The data comes from Comizzoli et al. (1990). We start with a Poisson regression:

```
data(solder, package="faraway")
modp <- glm(skips ~ ., family=poisson, data=solder)
c(deviance(modp), df.residual(modp))
```

```
[1] 1829 882
```

We see that the full model has a residual deviance of 1829 on 882 degrees of freedom. This is not a good fit. Perhaps including interaction terms will improve the fit:

```
modp2 <- glm(skips ~ (Opening + Solder + Mask + PadType + Panel)^2 ,
               family=poisson, data=solder)
deviance(modp2)
[1] 1068.8
pchisq(deviance(modp2), df.residual(modp2), lower=FALSE)
[1] 1.1307e-10
```

The fit is improved but not enough to conclude that the model fits. We could try adding more interactions but that would make interpretation increasingly difficult. A check for outliers reveals no problem.

An alternative model for counts is the negative binomial. The functions for fitting come from the MASS package — see Venables and Ripley (2002) for more details. We can specify the link parameter  $k$ . Here we choose  $k = 1$  to demonstrate the method, although there is no substantive motivation from this application to use this value. Note that the  $k = 1$  case corresponds to an assumption of a *geometric* distribution for the response.

```
library(MASS)
modn <- glm(skips ~ ., negative.binomial(1), solder)
modn
```

Coefficients:

(Intercept)	OpeningM	OpeningS	SolderThin	MaskA3
-1.6993	0.5085	1.9997	1.0489	0.6571
MaskA6	MaskB3	MaskB6	PadTypeD6	PadTypeD7
2.5265	1.2726	2.0803	-0.4612	0.0161
PadTypeL4	PadTypeL6	PadTypeL7	PadTypeL8	PadTypeL9
0.4688	-0.4711	-0.2949	-0.0849	-0.5213
PadTypeW4	PadTypeW9	Panel		
-0.1425	-1.4836	0.1693		

Degrees of Freedom: 899 Total (i.e. Null); 882 Residual

Null Deviance: 1740

Residual Deviance: 559 AIC: 3880

We could experiment with different values of  $k$ , but there is a more direct way of achieving this by allowing the parameter  $k$  to vary and be estimated using maximum likelihood in:

```
modn <- glm.nb(skips ~ ., solder)
summary(modn)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.4225	0.1427	-9.97	< 2e-16
OpeningM	0.5029	0.0798	6.31	2.9e-10
OpeningS	1.9132	0.0715	26.75	< 2e-16
SolderThin	0.9393	0.0536	17.52	< 2e-16
MaskA3	0.5898	0.0965	6.11	9.9e-10
MaskA6	2.2673	0.1018	22.27	< 2e-16
MaskB3	1.2110	0.0964	12.57	< 2e-16
MaskB6	1.9904	0.0922	21.58	< 2e-16
PadTypeD6	-0.4659	0.1124	-4.15	3.4e-05
PadTypeD7	-0.0331	0.1067	-0.31	0.75611
PadTypeL4	0.3827	0.1026	3.73	0.00019
PadTypeL6	-0.5784	0.1141	-5.07	4.0e-07

PadTypeL7	-0.3666	0.1109	-3.30	0.00095
PadTypeL8	-0.1589	0.1082	-1.47	0.14199
PadTypeL9	-0.5660	0.1139	-4.97	6.8e-07
PadTypeW4	-0.2004	0.1087	-1.84	0.06526
PadTypeW9	-1.5646	0.1362	-11.49	< 2e-16
Panel	0.1637	0.0314	5.21	1.8e-07

(Dispersion parameter for Negative Binomial(4.3972) family taken to be 1)

Null deviance: 4043.3 on 899 degrees of freedom  
 Residual deviance: 1008.3 on 882 degrees of freedom  
 AIC: 3683

Theta: 4.397  
 Std. Err.: 0.495  
 2 x log-likelihood: -3645.309

We see that  $\hat{k} = 4.397$  with a standard error of 0.495. We can compare negative binomial models using the usual inferential techniques.

## 5.5 Zero Inflated Count Models

Sometimes we see count response data where the number of zeroes appearing is significantly greater than the Poisson or negative binomial models would predict. Consider the number of arrests for criminal offenses incurred by individuals. A large number of people have never been arrested by the police while a smaller number have been detained on multiple occasions. Modifying the Poisson by adding a dispersion parameter does not adequately model this divergence from the standard count distributions.

We consider a sample of 915 biochemistry graduate students as analyzed by Long (1990). The response is the number of articles produced during the last three years of the PhD. We are interested in how this is related to the gender, marital status, number of children, prestige of the department and productivity of the advisor of the student. The dataset may be found in the `pscl` package of Zeileis et al. (2008) which also provides the new model fitting functions needed in this section. We start by fitting a Poisson regression model:

```
library(pscl)
modp <- glm(art ~ ., data=bioChemists, family=poisson)
summary(modp)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.30462	0.10298	2.96	0.0031
femWomen	-0.22459	0.05461	-4.11	0.0000392
marMarried	0.15524	0.06137	2.53	0.0114
kid5	-0.18488	0.04013	-4.61	0.0000041
phd	0.01282	0.02640	0.49	0.6271
ment	0.02554	0.00201	12.73	< 2e-16

n = 915 p = 6

Deviance = 1634.371 Null Deviance = 1817.405 (Difference = 183.034)

We can see that deviance is significantly larger than the degrees of freedom. Some experimentation reveals that this cannot be solved by using a richer linear predictor or by eliminating some outliers. We might consider a dispersed Poisson model or negative binomial but some thought suggests that there are good reasons why a

student might produce no articles at all. We count and predict how many students produce between zero and seven articles. Very few students produce more than seven articles so we ignore these. The `predprob` function produces the predicted probabilities for each case. By summing these, we get the expected number for each article count.

```
ocount <- table(bioChemists$art)[1:8]
pcount <- colSums(predprob(modp)[,1:8])
plot(pcount, ocount, type="n", xlab="Predicted", ylab="Observed")
text(pcount, ocount, 0:7)
```

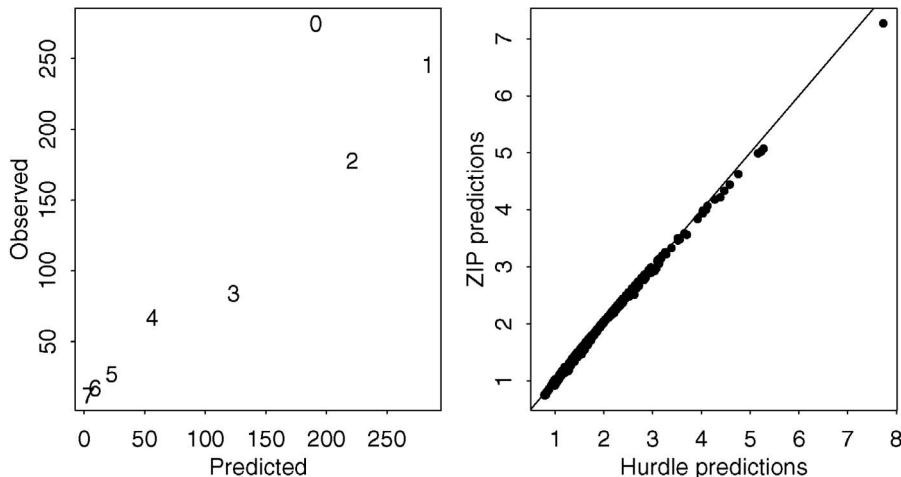


Figure 5.5. On the left, the predicted and observed counts for number of articles 0-7 is shown. On the right, the predictions from the hurdle and zero-inflated Poisson model are compared.

In the left panel of Figure 5.5, we see that there are many more unproductive students than would be predicted by the Poisson model. In contrast, the relationship between observed and predicted is linear for the students who produce at least one article.

There are two main ways of modeling an excess of zero counts. First we consider the *hurdle* model. This can be understood in terms of a latent variable. In our example, this latent variable might measure the propensity of the student to publish. We can imagine a hurdle which this latent variable must exceed for a publication to be produced. If the latent variable is less than the hurdle, nothing appears, but as it exceeds the hurdle, more articles may be produced. The latent variable is related to the predictors in some manner we may wish to uncover. A general specification of the model is:

$$\begin{aligned} P(Y = 0) &= f_1(0) \\ P(Y = j) &= \frac{1 - f_1(0)}{1 - f_2(0)} f_2(j), \quad j > 0 \end{aligned}$$

The first part of the model describes the probability of zero being observed (or not). We shall use a binary response model that can link this probability to the predictors.

The second part of the model specifies the probability of outcomes greater than zero. We shall use a Poisson for  $f_2$  but this now describes a *truncated* Poisson, as zero is not an allowable outcome and we must rescale the distribution accordingly. Zero is the hurdle here but we could specify higher hurdles if the application warranted it. We can fit such a model as:

```
modh <- hurdle(art ~ ., data=bioChemists)
summary(modh)

Count model coefficients (truncated poisson with log link):
Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.67114 0.12246 5.48 4.2e-08
femWomen -0.22858 0.06522 -3.51 0.00046
marMarried 0.09648 0.07283 1.32 0.18521
kid5 -0.14219 0.04845 -2.93 0.00334
phd -0.01273 0.03130 -0.41 0.68434
ment 0.01875 0.00228 8.22 < 2e-16

Zero hurdle model coefficients (binomial with logit link):
Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.2368 0.2955 0.80 0.423
femWomen -0.2512 0.1591 -1.58 0.114
marMarried 0.3262 0.1808 1.80 0.071
kid5 -0.2852 0.1111 -2.57 0.010
phd 0.0222 0.0796 0.28 0.780
ment 0.0801 0.0130 6.15 7.5e-10
```

Number of iterations in BFGS optimization: 12

Log-likelihood: -1.61e+03 on 12 Df

Notice that the model produces two sets of coefficients — one for the truncated Poisson part of the model and one for the binary response part of the model. As can be seen, these can be quite different. By default a binomial model with a logit link is used for the zero component and a Poisson with a log link is used for the count component. Sensible alternatives can be specified. Before interpreting the model, we present a second form of model for zero inflated count data.

Suppose we ask people how many games of chess they have played in the last month. Some will say zero because they do not play chess but some zero responses will be from chess players who have not managed a game in the last month. Circumstances such as these require a *mixture* model. A general specification of this model takes the form:

$$\begin{aligned} P(Y = 0) &= \phi + (1 - \phi)f(0) \\ P(Y = j) &= (1 - \phi)f(j), \quad j > 0 \end{aligned}$$

The parameter  $\phi$  represents the proportion who will always respond zero. We can model this proportion using binary response model. The distribution  $f$  models the counts of those individuals that can have a positive response. In some cases, these individuals will have a zero response which combines with the always-zero subjects. We can use a Poisson for  $f$  in which case this is called *zero-inflated Poisson* or ZIP model. We can fit this:

```
modz <- zeroinfl(art ~ ., data=bioChemists)
summary(modz)

Count model coefficients (poisson with log link):
Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) 0.64084 0.12131 5.28 1.3e-07
femWomen -0.20914 0.06340 -3.30 0.00097
marMarried 0.10375 0.07111 1.46 0.14456
kid5 -0.14332 0.04743 -3.02 0.00251
phd -0.00617 0.03101 -0.20 0.84238
ment 0.01810 0.00229 7.89 3.1e-15
```

Zero-inflation model coefficients (binomial with logit link):

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.57706	0.50939	-1.13	0.257
femWomen	0.10975	0.28008	0.39	0.695
marMarried	-0.35401	0.31761	-1.11	0.265
kid5	0.21710	0.19648	1.10	0.269
phd	0.00127	0.14526	0.01	0.993
ment	-0.13411	0.04524	-2.96	0.003

Number of iterations in BFGS optimization: 21

Log-likelihood: -1.6e+03 on 12 Df

The results take a similar form to the hurdle model. Focusing on the zero part of the model, we notice that the ment variable which counts the number of articles produced by the mentor is the most significant predictor in both cases but takes opposing signs. ~~This is because the hurdle approach models the probability of a positive count whereas the zero-inflated approach models the probability of a zero count.~~ Hence there is no contradiction.

It is difficult to choose between the two approaches. We can compare the fitted values:

```
plot(fitted(modh), fitted(modz), xlab="Hurdle predictions", ylab="ZIP
    ↪ predictions")
abline(0,1)
```

As can be seen in the second panel of Figure 5.5, there is not much difference between these fits. We might use our background knowledge of the application to make an appropriate choice.

We can use the standard likelihood testing theory to compare nested models. For example, suppose we consider a simplified version of the ZIP model where we now have different predictors for the two components of the model. The count part of the model is specified before the | and the zero model after.

```
modz2 <- zeroinfl(art ~ fem+kid5+ment | ment, data=bioChemists)
```

```
summary(modz2)
```

Count model coefficients (poisson with log link):

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.69452	0.05303	13.10	< 2e-16
femWomen	-0.23386	0.05840	-4.00	6.2e-05
kid5	-0.12652	0.03967	-3.19	0.0014
ment	0.01800	0.00222	8.10	5.7e-16

Zero-inflation model coefficients (binomial with logit link):

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.6849	0.2053	-3.34	0.00085
ment	-0.1268	0.0398	-3.18	0.00145

Number of iterations in BFGS optimization: 13

Log-likelihood: -1.61e+03 on 6 Df

Twice the difference in the log-likelihood will be approximately  $\chi^2$ -distributed with degrees of freedom equal to the difference in the number of parameters.

```
(1rt <- 2*(modz$loglik-modz2$loglik))
[1] 6.1728
1-pchisq(6.1728,6)
[1] 0.40411
```

Given the large  $p$ -value of 0.4, we conclude that our simplification of the model is justifiable. For interpretation, the exponentiated coefficients are more useful:

```
exp(coef(modz2))
count_(Intercept)    count_femWomen      count_kid5
                2.00274          0.79147        0.88116
count_ment   zero_(Intercept)      zero_ment
                1.01817          0.50415        0.88091
```

We see that among students who are inclined to write articles, women produce at a rate 0.79 times that for men and each child under five reduces production by about 12% while each additional article produced by the mentor increases production by about 1.8%. We see that each extra article from the mentor reduces the odds of a nonproductive student by a factor of 0.88. Of course, none of these interpretations can be taken as causal.

We can also use the model to make predictions. Consider a single male with no children whose mentor produced six articles:

```
newman <- data.frame(fem="Men",mar="Single",kid5=0,ment=6)
predict(modz2, newdata=newman, type="prob")
0           1           2           3           4           5           6
1 0.27759 0.19394 0.21636 0.16091 0.089758 0.040054 0.014895
...
1
```

We see that most likely outcome for this student is that no articles will be produced with a probability of 0.278. We can query the probability of no production from the zero part of the model:

```
predict(modz2, newdata=newman, type="zero")
1
0.19067
```

So the additional probability to make this up to 0.278 comes from the Poisson count part of the model. This difference might be attributed to the student who had the potential to write an article but just didn't do it.

## Exercises

1. The dataset `discoveries` lists the numbers of “great” inventions and scientific discoveries in each year from 1860 to 1959.
  - (a) Plot the discoveries over time and comment on the trend, if any.
  - (b) Fit a Poisson response model with a constant term. Now compute the mean number of discoveries per year. What is the relationship between this mean and the coefficient seen in the model?
  - (c) Use the deviance from the model to check whether the model fits the data. What does this say about whether the rate of discoveries is constant over time?
  - (d) Make a table of how many years had zero, one, two, three, etc. discoveries. Collapse eight or more into a single category. Under an appropriate Poisson

distribution, calculate the expected number of years with each number of discoveries. Plot the observed against the expected using a different plotting character to denote the number of discoveries. How well do they agree?

- (e) Use the Pearson's Chi-squared test to check whether the observed numbers are consistent with the expected numbers. Interpret the result.
  - (f) Fit a Poisson response model that is quadratic in the year. Test for the significance of the quadratic term. What does this say about the presence of a trend in discovery?
  - (g) Compute the predicted number of discoveries each year and show these predictions as a line drawn over the data. Comment on what you see.
2. The `salmonella` data was collected in a salmonella reverse mutagenicity assay. The predictor is the dose level of quinoline and the response is the numbers of revertant colonies of TA98 salmonella observed on each of three replicate plates.
- (a) Plot the data and comment on the relationship between dose and colonies.
  - (b) Compute the mean and variance within each set of observations with the same dose. Plot the variance against the mean and comment on what this says about overdispersion.
  - (c) Fit a model with dose treated as a six-level factor. Check the deviance to determine whether this model fits the data. Do you think it is possible to find a transformation of the dose predictor that results in a Poisson model that does fit the data?
  - (d) Make a QQ plot of the residuals from the previous model. Interpret the plot.
  - (e) Fit a Poisson model what includes an overdispersion parameter and is quadratic in the dose. Can we determine from the deviance of this model whether the fit is adequate?
  - (f) Plot the residuals against the fitted values for the previous model. Interpret the plot.
  - (g) Plot the fitted mean response of this model on top of the data.
  - (h) Give the predicted mean response for a dose of 500. Compute a 95% confidence interval.
    - (i) At what dose does the maximum predicted response occur?
3. The `ships` dataset found in the `MASS` package gives the number of wave damage incidents and aggregate months of service for different types of ships broken down by year of construction and period of operation.
- (a) Examine the data for the period of operation 1960–1974 for ships constructed in the years 1975–1979. Why are there no damage incidents?
  - (b) Make a two-way table that shows the rate of damage incidents per 1000 months of aggregate service classified by type and year of construction. Comment on the table.
  - (c) Compute the rate of damage incidents per year for all cases where some service was recorded.

- (d) Fit linear models with the observed rate of damage incidents as the response and the following three combinations of predictors: (i) All two-way interactions, (ii) main effects terms only, (iii) null (no predictors). Make sure year is treated as a factor rather than numerical variable. Which of these three models is preferred?
- (e) Fit a Poisson response model for the number of incidents with the predictors: log of service, type, year and period. Test whether the parameter associated with the service term can be one. Explain why we are interested in such a test.
- (f) Fit the Poisson rate model with all two-way interactions of the three predictors. Does this model fit the data?
- (g) Check the residuals. Are there any outliers? Plot residuals against the fitted values. Investigate any unusual features of this plot.
- (h) Now fit the rate model with just the main effects and compare it to the interaction model. Which model is preferred?
- (i) Fit quasi-Poisson versions of the two previous models and repeat the comparison.
- (j) Interpret the coefficients of the main effects of the quasi-Poisson model. What factors are associated with higher and lower rates of damage incidents?
4. The dataset `africa` gives information about the number of military coups in sub-Saharan Africa and various political and geographical information.
- Plot the response, the number of military coups against each of the other variables.
  - Use a stepwise AIC-based method to select a model that uses a smaller number of the available predictors.
  - Does the deviance of your selected model indicate a good fit to the data?
  - Make a QQ plot of the residuals and comment. Plot the residuals against the fitted values and interpret the result. What is the source of the lines of points observed on this plot?
  - Give an interpretation of the coefficients of this plot.
  - Count the number of countries with each number of military coups. Compare this with the numbers predicted by the previous model. Is there any evidence of excess of countries with zero coups? Use a Chi-squared test as implemented in `chisq.test()`.
5. The `dvisits` data comes from the Australian Health Survey of 1977–1978 and consist of 5190 single adults where young and old have been oversampled.
- Make plots which show the relationship between the response variable, `doctorco`, and the potential predictors, `age` and `illness`.
  - Combine the predictors `chcond1` and `chcond2` into a single three-level factor. Make an appropriate plot showing the relationship between this factor and the response. Comment.

- (c) Build a Poisson regression model with `doctorco` as the response and `sex`, `age`, `agesq`, `income`, `levyplus`, `freepoor`, `freerepa`, `illness`, `actdays`, `hscore` and the three-level condition factor as possible predictor variables. Considering the deviance of this model, does this model fit the data?
- (d) Plot the residuals and the fitted values — why are there lines of observations on the plot? Make a QQ plot of the residuals and comment.
- (e) Use a stepwise AIC-based model selection method. What sort of person would be predicted to visit the doctor the most under your selected model?
- (f) For the last person in the dataset, compute the predicted probability distribution for their visits to the doctor, i.e., give the probability they visit 0, 1, 2, etc. times.
- (g) Tabulate the frequencies of the number of doctor visits. Compute the expected frequencies of doctor visits under your most recent model. Compare the observed with the expected frequencies and comment on whether it is worth fitting a zero-inflated count model.
- (h) Fit a comparable (Gaussian) linear model and graphically compare the fits. Describe how they differ.
6. Components are attached to an electronic circuit card assembly by a wave-soldering process. The soldering process involves baking and preheating the circuit card and then passing it through a solder wave by conveyor. Defects arise during the process. The design is  $2^{7-3}$  with three replicates. The data is presented in the dataset `wavesolder`. You can assume that the replicates are independent.
- (a) Make plots of the number of defects against each of the predictors. Comment on the relationships you see. Check graphically that there is no trend in the replicates.
- (b) Compute the mean and variance within each group of three replicates. Plot the variance against the mean. Comment on the relationship and the viability of a Poisson model for the response. Repeat the plot, but use a log scale on both axes. Does this plot reveal anything new?
- (c) Fit a Poisson model for the number of defects with all predictors included as main effects. What does the deviance of this model say about its fit?
- (d) Make a plot of the residuals against the fitted values and comment on what is seen. Make a QQ plot of the residuals. Are there any outliers?
- (e) Refit the Poisson model but excluding the case with the largest residual. Compute the deviance. Does this model now fit the data?
- (f) Fit a quasi-poisson model with same model formula and excluded case. Estimate the value of the dispersion parameter. Check the model summary. Now use an  $F$ -test to the significance of each of the predictors. Compare the two sets of tests—one from the model summary and one from the  $F$ -test. Are they similar? Report on what predictors are significant and which level of the significant factors will lead to higher defects.
- (g) Check the diagnostics again as in (d).

7. The dataset `esdcomp` was recorded on 44 doctors working in an emergency service at a hospital to study the factors affecting the number of complaints received.
- (a) Consider the rate of complaints in terms of the number received in relation to the number of patient visits made. Plot this against each of the four potential predictors and comment on the relationships.
  - (b) Fit a binomial GLM for the number of complaints out of the number of visits with the other four variables as predictors. Does this model fit the data? Perform this check numerically and graphically.
  - (c) Check the significance of the four predictors in the binomial model. Give a numerical interpretation to the effect of any significant predictors.
  - (d) Fit an appropriate Poisson rate model for number of complaints that takes a comparable form to the binomial GLM fitted earlier. Does this model fit the data? Again check this numerically and graphically.
  - (e) Again check the significance of the predictors and provide a numerical interpretation. Compare the conclusions of the two models.
  - (f) Exchange the role of hours and visits in the Poisson rate model. Again check the significance of the predictors and interpret the outcome.
  - (g) Compare the two proposed rate models.

## Contingency Tables

---

A contingency table is used to show cross-classified categorical data on two or more variables. The variables can be *nominal* or *ordinal*. A nominal variable has categories with no natural ordering; for example, consider the automotive companies Ford, General Motors and Toyota. An ordering could be imposed using some criterion like sales, but there is nothing inherent in the categories that makes any particular ordering obvious. An ordinal variable has a natural default ordering. For example, a disease might be recorded as absent, mild or severe. The five-point Likert scale ranging through strongly disagree, disagree, neutral, agree and strongly agree is another example.

An *interval scale* is an ordinal variable that has categories with a distance measure. This is often the result of continuous data that has been discretized into intervals. For example, age groups 0–18, 18–34, 34–55 and 55+ might be used to record age information. If the intervals are relatively wide, then methods for ordinal data can be used where the additional information about the intervals may be useful in the modeling. If the intervals are quite narrow, then we could replace interval response with the midpoint of the interval and then use continuous data methods. One could argue that all so-called continuous data is of this form, because such data cannot be measured with arbitrary precision. Height might be given to the nearest centimeter, for example.

### 6.1 Two-by-Two Tables

The data shown in Table 6.1 were collected as part of a quality improvement study at a semiconductor factory. A sample of wafers was drawn and cross-classified according to whether a particle was found on the die that produced the wafer and whether the wafer was good or bad. More details on the study may be found in Hall (1994). The data might have arisen under several possible sampling schemes:

Quality	No Particles	Particles	Total
Good	320	14	334
Bad	80	36	116
Total	400	50	450

Table 6.1 *Study of the relationship between wafer quality and the presence of particles on the wafer.*

1. We observed the manufacturing process for a certain period of time and observed 450 wafers. The data were then cross-classified. We could use a Poisson model.
2. We decided to sample 450 wafers. The data were then cross-classified. We could use a multinomial model.
3. We selected 400 wafers without particles and 50 wafers with particles and then recorded the good or bad outcome. We could use a binomial model.
4. We selected 400 wafers without particles and 50 wafers with particles that also included, by design, 334 good wafers and 116 bad ones. We could use a hypergeometric model.

The first three sampling schemes are all plausible. The fourth scheme seems less likely in this example, but we include it for completeness. Such a scheme is more attractive when one level of each variable is relatively rare and we choose to oversample both levels to ensure some representation.

The main question of interest concerning these data is whether the presence of particles on the wafer affects the quality outcome. We shall see that all four sampling schemes lead to exactly the same conclusion. First, let's set up the data in a convenient form for analysis:

```
y <- c(320, 14, 80, 36)
particle <- gl(2, 1, 4, labels=c("no", "yes"))
quality <- gl(2, 2, labels=c("good", "bad"))
(wafer <- data.frame(y, particle, quality))

y   particle   quality
1  320       no      good
2   14      yes      good
3   80       no      bad
4   36      yes      bad
```

We will need the data in this form with one observation per line for our model fitting, but usually we prefer to observe it in table form:

```
(ov <- xtabs(y ~ quality+particle))
           particle
quality no yes
  good 320 14
  bad   80 36
```

**Poisson Model:** Suppose we assume that the process is observed for some period of time and we count the number of occurrences of the possible outcomes. It would be natural to view these outcomes occurring at different rates and that we could form a Poisson model for these rates. Suppose we fit an additive model:

```
mod1 <- glm(y ~ particle+quality, poisson)
library(faraway)
summary(mod1)

Estimate Std. Error z value Pr(>|z|)
(Intercept) 5.6934    0.0572  99.54  <2e-16
particleyes -2.0794    0.1500 -13.86  <2e-16
qualitybad   -1.0576    0.1078  -9.81  <2e-16
```

n = 4 p = 3

Deviance = 54.030 Null Deviance = 474.099 (Difference = 420.068)

The null model, which suggests all four outcomes occur at the same rate, does not fit because the deviance of 474.1 is very large for three degrees of freedom. The additive

model, with a deviance of 54.03, is clearly an improvement over this. We might also want to test the significance of the individual predictors. We could use the  $z$ -values, but it is better to use the likelihood ratio test based on the differences in the deviance (not that it matters much for this particular dataset):

```
drop1(mod1, test="Chi")
```

Single term deletions

Model:

	Df	Deviance	AIC	LRT	Pr(Chi)
<none>		54	84		
particle	1	364	392	310	<2e-16
quality	1	164	192	110	<2e-16

We see that both predictors are significant relative to the full model. By examining the coefficients, we see that wafers without particles occur at a significantly higher rate than wafers with particles. Similarly, we see that good-quality wafers occur at a significantly higher rate than bad-quality wafers.

The model coefficients are closely related to the marginal totals in the table. The maximum likelihood estimates satisfy:

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

where the  $X^T y$  is, in this example:

```
(t(model.matrix(mod1)) %*% y) [,1]
(Intercept) particleyes qualitybad
450           50            116
```

So we see that the fitted values,  $\hat{\mu}$ , are a function of marginal totals. This fact is exploited in an alternative fitting method known as *iterative proportional fitting*. The `glm` function in R, however, uses Fisher scoring, described in Section 8.2. In any case, the log-likelihood (ignoring any terms not involving  $\mu$ ) is:

$$\log L = \sum_i y_i \log \mu_i$$

which is maximized to obtain the fit. In fact the log-likelihood has a second term,  $\sum_i (y_i - \mu_i)$ , but this will be zero at the MLE provided the model has an intercept term, so this term can be ignored.

The analysis so far has told us nothing about the relationship between the presence of particles and the quality of the wafer. The additive model posits:

$$\log \mu = \gamma + \alpha_i + \beta_j$$

where  $\alpha$  represents the particle effect and  $\beta$  represents the quality outcome and  $i, j = 1, 2$ .  $\gamma$  is the intercept term. Due to the log link, the predicted rate for the response in any cell in the table is formed from the product of the rates for the corresponding levels of the two predictors. There is no interaction term and so good- or bad-quality outcomes occur independently of whether a particle was found on the wafer. This model has a deviance of 54.03 on one degree of freedom and so does not fit the data.

The addition of an interaction term would saturate the model and so would have zero deviance and degrees of freedom. So an hypothesis comparing the models with

and without interaction would give a test statistic of 54.03 on one degree of freedom. The hypothesis of no interaction would be rejected. We conclude that the presence of particles is related to the quality outcome.

**Multinomial Model:** Suppose we assume that the total sample size was fixed at 450 and that the frequency of the four possible outcomes was recorded. In these circumstances, it is natural to use a multinomial distribution to model the response. Let  $y_{ij}$  be the observed response in cell  $(i, j)$  and let  $p_{ij}$  be the probability that an observation falls in that cell and let  $n$  be the sample size. The probability of the observed response under the multinomial is then:

$$\frac{n!}{\prod_i \prod_j y_{ij}} \prod_i \prod_j p_{ij}^{y_{ij}}$$

Now the  $p_{ij}$  will be linked to the predictor information according to the model we choose. To estimate the parameters, we would maximize the log-likelihood:

$$\log L = \sum_i \sum_j y_{ij} \log p_{ij}$$

where terms not involving  $p_{ij}$  are ignored. Notice that this takes essentially the same form as for the Poisson model above.

The main hypothesis of interest is whether the quality and presence of a particle on the wafer are independent. Let  $p_i$  for  $i = 1, 2$  be the probabilities of the two quality outcomes and  $p_j$  for  $j = 1, 2$  be the probability of the two particle categories. Let  $p_{ij}$  be the probability of a particular joint outcome. Under independence,  $p_{ij} = p_i p_j$ . Using the fact that probabilities must sum to one, the maximum likelihood estimates are:

$$\hat{p}_i = \sum_j y_{ij}/n \quad \text{and} \quad \hat{p}_j = \sum_i y_{ij}/n$$

We can compute these for the wafer data as, respectively:

```
(pp <- prop.table( xtabs(y ~ particle)))
particle
  no      yes
0.88889 0.11111
(gp <- prop.table( xtabs(y ~ quality)))
quality
  good     bad
0.74222 0.25778
```

The fitted values are then  $\hat{\mu}_{ij} = np_i p_j = \sum_i y_{ij} \sum_j y_{ij}/n$  or:

```
(fv <- outer(gp, pp) * 450)
particle
quality   no      yes
  good 296.89 37.111
  bad 103.11 12.889
```

To test the fit, we compare this model against the saturated model, for which  $\hat{\mu}_{ij} = y_{ij}$ . So the deviance is:

$$2 \sum_i \sum_j y_{ij} \log(y_{ij}/\hat{\mu}_{ij})$$

which computes to:

```
2*sum.ov*log.ov/fv)
[1] 54.03
```

which is the same deviance we observed in the Poisson model. So we see that the test for independence in the multinomial model coincides with the test for no interaction in the Poisson model. The Poisson-based test is easier to execute in R, so we shall usually take that approach.

This connection between the Poisson and multinomial is no surprise due to the following result. Let  $Y_1, \dots, Y_k$  be independent Poisson random variables with means  $\lambda_1, \dots, \lambda_k$ , then the joint distribution of  $Y_1, \dots, Y_k | \sum_i Y_i = n$  is multinomial with probabilities  $p_j = \lambda_j / \sum_i \lambda_i$ .

One alternative to the deviance is the Pearson  $X^2$  statistic:

$$X^2 = \sum_{i,j} \frac{(y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

which takes the value:

```
sum((ov-fv)^2/fv)
[1] 62.812
```

Yates' continuity correction subtracts 0.5 from  $y_{ij} - \hat{\mu}_{ij}$  when this value is positive and adds 0.5 when it is negative. This gives superior results for small samples. This correction is implemented in:

```
prop.test.ov
```

```
2-sample test for equality of proportions with
continuity correction
```

```
data: ov
X-squared = 60.124, df = 1, p-value = 8.907e-15
```

The deviance-based test is preferred to the Pearson's  $X^2$ .

**Binomial:** It would also be natural to view the presence of the particle as affecting the quality of wafer. We would view the quality as the response and the particle status as a predictor. We might fix the number of wafers with no particles at 400 and the number with particles as 50 and then observe the outcome. We could then use a binomial model for the response for both groups. Let's see what happens:

```
(m <- matrix(y, nrow=2))
[,1] [,2]
[1,] 320   80
[2,] 14    36
modb <- glm(m ~ 1, family=binomial)
deviance(modb)
[1] 54.03
```

We fit the null model which suggests that the probability of the response is the same in both the particle and no-particle group. This hypothesis of *homogeneity* corresponds exactly to the test of independence and the deviance is exactly the same.

For larger contingency tables, where there are more than two rows (or columns), we can use a multinomial model for each row. This model is more accurately called a *product* multinomial model to distinguish it from the unrestricted multinomial model introduced above.

**Hypergeometric:** The remaining case is where both sets of marginal totals are fixed. This situation is rather less common in practice, but does suggest a more accurate test for independence. This sampling scheme can arise when classifying objects into one of two types when the true proportions of each type are known in advance. For example, suppose you are given 10 true or false statements and told that 5 are true and 5 are false. You are asked to sort the statements into true and false. We can generate a two-by-two table of the correct classification against the observed classification generated. Under the hypergeometric distribution and the assumption of independence, the probability of the observed table is:

$$\frac{(y_{11} + y_{12})!(y_{11} + y_{21})!(y_{12} + y_{22})!(y_{21} + y_{22})!}{y_{11}!y_{12}!y_{21}!y_{22}!n!}$$

If we fix any number in the table, say  $y_{11}$ , the remaining three numbers are completely determined because the row and column totals are known. There is a limited number of values which  $y_{11}$  can possibly take and we can compute the probability of all these outcomes. Specifically, we can compute the total probability of all outcomes more extreme than the one observed. This method is called *Fisher's exact test*. We may execute it as follows:

```
fisher.test(ov)
```

Fisher's Exact Test for Count Data

```
data: ov
p-value = 2.955e-13
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 5.0906 21.5441
sample estimates:
odds ratio
 10.213
```

The odds ratio, which is  $(y_{11}y_{22})/(y_{12}y_{21})$ , takes the value:

```
(320*36) / (14*80)
```

```
[1] 10.286
```

and is a measure of the association for which an exact confidence interval may be calculated as we see in the output.

Fisher's test is attractive because the null distribution for the deviance and Pearson's  $\chi^2$  test statistics is only approximately  $\chi^2$  distributed. This approximation is particularly suspect for tables with small counts making an exact method valuable. The Fisher test becomes more difficult to compute for larger tables and some approximations may be necessary. However, for larger tables, the  $\chi^2$  approximation will tend to be very accurate.

## 6.2 Larger Two-Way Tables

Snee (1974) presents data on 592 students cross-classified by hair and eye color:

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

	y	eye	hair
1	5	green	BLACK
2	29	green	BROWN

```
..etc..
16 7 brown BLOND
```

The data is more conveniently displayed using:

```
(ct <- xtabs(y ~ hair + eye, haireye))
  eye
```

hair	green	hazel	blue	brown
BLACK	5	15	20	68
BROWN	29	54	84	119
RED	14	14	17	26
BLOND	16	10	94	7

We can execute the usual Pearson's  $\chi^2$  test for independence as:

```
summary(ct)
```

```
Call: xtabs(formula = y ~ hair + eye, data = haireye)
```

```
Number of cases in table: 592
```

```
Number of factors: 2
```

```
Test for independence of all factors:
```

```
Chisq = 138, df = 9, p-value = 2.3e-25
```

where we see that hair and eye color are clearly not independent.

One option for displaying contingency table data is the *dotchart*:

```
dotchart(ct)
```

which may be seen in the first panel of Figure 6.1. The mosaic plot, described in

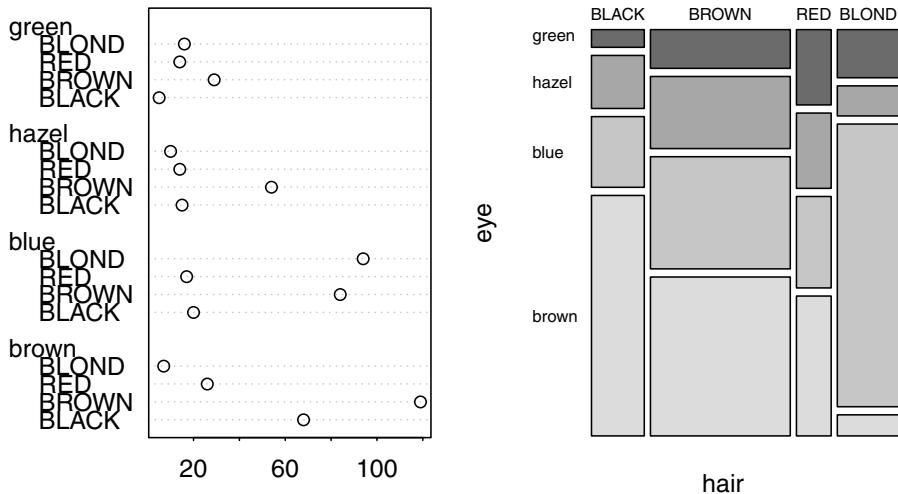


Figure 6.1 *Dotchart and mosaic plot*.

Hartigan and Kleiner (1981), divides the plot region according to the frequency of each level in a recursive manner:

```
mosaicplot(ct, color=TRUE, main=NULL, las=1)
```

In the plot shown in the second panel of Figure 6.1, the area is first divided according to the frequency of hair color. Within each hair color, the area is then divided according to the frequency of eye color. A different plot could be constructed by reversing the order of hair and eye in the *xtabs* command above. We can now readily see the frequency of various outcomes. We see, for example, that brown hair and brown

eyes is the most common combination while green eyes and black hair is the least common. When the two categories are independent, the rectangles in the plot will align close to a grid. In this example, the horizontal divisions do not line up at all so we can tell there is dependence.

Now we fit the Poisson GLM:

```
modc <- glm(y ~ hair+eye, family=poisson, haireye)
summary(modc)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	2.458	0.152	16.14	<2e-16
hairBROWN	0.974	0.113	8.62	<2e-16
hairRED	-0.419	0.153	-2.75	0.006
hairBLOND	0.162	0.131	1.24	0.216
eyehazel	0.374	0.162	2.30	0.021
eyeblue	1.212	0.142	8.51	<2e-16
eyebrown	1.235	0.142	8.69	<2e-16

n = 16 p = 7

Deviance = 146.444 Null Deviance = 453.308 (Difference = 306.864)

We see that most of the levels of hair and eye color show up as significantly different from the reference levels of black hair and green eyes. But this merely indicates that there are higher numbers of people with some hair colors than others and some eye colors than others. We already know this. We are more interested in the relationship between hair and eye color. The deviance of 146.44 on just nine degrees freedom shows that they are clearly dependent.

### 6.3 Correspondence Analysis

The analysis of the hair-eye color data in the previous section revealed how hair and eye color are dependent. But this does not tell us how they are dependent. To study this, we can use a kind of residual analysis for contingency tables called *correspondence analysis*.

Compute the Pearson residuals  $r_P$  and write them in the matrix form  $R_{ij}$ , where  $i = 1, \dots, r$  and  $j = 1, \dots, c$ , according to the structure of the data. Perform the singular value decomposition:

$$R_{r \times c} = U_{r \times w} D_{w \times w} V_{w \times c}^T$$

where  $r$  is the number of rows,  $c$  is the number of columns and  $w = \min(r, c)$ .  $U$  and  $V$  are called the right and left singular vectors, respectively.  $D$  is a diagonal matrix with sorted elements  $d_i$ , called *singular values*. Another way of writing this is:

$$R_{ij} = \sum_{k=1}^w U_{ik} d_k V_{jk}$$

As with eigendecompositions, it is not uncommon for the first few singular values to be much larger than the rest. Suppose that the first two dominate so that:

$$R_{ij} \approx U_{i1} d_1 V_{j1} + U_{i2} d_2 V_{j2}$$

We usually absorb the  $d$ s into  $U$  and  $V$  for plotting purposes so that we can assess the relative contribution of the components. Thus:

$$\begin{aligned} R_{ij} &\approx (U_{i1}\sqrt{d_1}) \times (V_{j1}\sqrt{d_1}) + (U_{i2}\sqrt{d_2}) \times (V_{j2}\sqrt{d_2}) \\ &\equiv U_{i1}V_{j1} + U_{i2}V_{j2} \end{aligned}$$

where in the latter expression we have redefined the  $U$ s and  $V$ s to include the  $\sqrt{d}$ .

The two-dimensional correspondence plot displays  $U_{i2}$  against  $U_{i1}$  and  $V_{j2}$  against  $V_{j1}$  on the same graph. So the points on the plot will either represent a row level ( $U$ ) or a column level ( $V$ ). We compute the plot for the hair and eye color data:

```
z <- xtabs(residuals(modc, type="pearson") ~ hair+eye, haireye)
svdz <- svd(z, 2, 2)
leftsv <- svdz$u %*% diag(sqrt(svdz$d[1:2]))
rightsv <- svdz$v %*% diag(sqrt(svdz$d[1:2]))
ll <- 1.1*max(abs(rightsv), abs(leftsv))
plot(rbind(leftsv, rightsv), asp=1, xlim=c(-11,11), ylim=c(-11,11), xlab=
  "SV1", ylab="SV2", type="n")
abline(h=0, v=0)
text(leftsv, dimnames(z)[[1]])
text(rightsv, dimnames(z)[[2]])
```

The plot is shown in Figure 6.2. The correspondence analysis plot can be interpreted in light of the following observations:

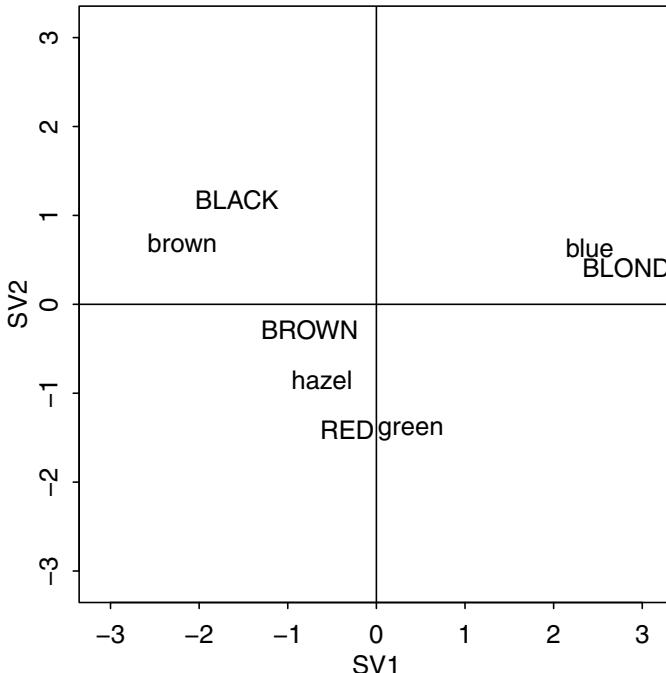


Figure 6.2 *Correspondence analysis for hair-eye combinations. Hair colors are given in uppercase letters and eye colors are given in lowercase letters.*

- $\sum d_i^2 = \text{Pearson's } X^2$  is called the *inertia*. When  $r = c$ ,  $d_i^2$  are the eigenvalues of  $R$ .
- Look for large values of  $|U_i|$  indicating that the row  $i$  profile is different. For example, the point for blonds in Figure 6.2 is far from the origin indicating that the distribution of eye colors within this group of people is not typical. In contrast, we see that the point for people with brown hair is close to the origin, indicating an eye color distribution that is close to the overall average. The same type of observation is true for the columns,  $|V_j|$ . Points distant from the origin mean that the level associated with the column  $j$  profile is different in some way.
- If row and column levels appear close together on the plot and far from the origin, we can see that there will be a large positive residual associated with this particular combination indicating a strong positive association. For example, we see that blue eyes and blond hair occur close together on the plot and far from the origin indicating a strong association. On the other hand, if the two points are situated diametrically apart on either side of the origin, we may expect a large negative residual indicating a strong negative association. For example, there are relatively fewer people with blond hair and brown eyes than would be expected under independence.
- If points representing two rows or two column levels are close together, this indicates that the two levels will have a similar pattern of association. In some cases, one might consider combining the two levels. For example, people with hazel or green eyes have similar hair color distributions and we might choose to combine these two categories.
- Because the distance between points is of interest, it is important that the plot is scaled so that the visual distance is proportionately correct. This does not happen automatically, because the default behavior of plots is to fill the plot region out to the specified aspect ratio.

There are several competing ways to construct contingency tables. See Venables and Ripley (2002) who provide the function `corresp` in the MASS package. See also Blasius and Greenacre (1998) for a survey of methods for visualizing categorical data.

## 6.4 Matched Pairs

In the typical two-way contingency tables, we display accumulated information about two categorical measures on the same object. In matched pairs, we observe one measure on two matched objects.

In Stuart (1955), data on the vision of a sample of women is presented. The left and right eye performance is graded into four categories:

```
data(eyegrade)
(ct <- xtabs(y ~ right+left, eyegrade))
  left
right    best second third worst
  best     1520   266    124    66
  second    234   1512    432    78
  third     117   362   1772   205
  worst      36     82    179   492
```

If we check for independence:

```
summary(ct)
Call: xtabs(formula = y ~ right + left, data = eyegrade)
Number of cases in table: 7477
Number of factors: 2
Test for independence of all factors:
Chisq = 8097, df = 9, p-value = 0
```

We are not surprised to find strong evidence of dependence. Most people's eyes are similar. A more interesting hypothesis for such matched pair data is symmetry. Is  $p_{ij} = p_{ji}$ ? We can fit such a model by defining a factor where the levels represent the symmetric pairs for the off-diagonal elements. There is only one observation for each level down the diagonal:

```
(symfac <- factor(apply(eyegrade[,2:3],1, function(x) paste(sort(x),
  &collapse="-"))))

[1] best-best    best-second   best-third   best-worst
[5] best-second   second-second  second-third  second-worst
[9] best-third    second-third   third-third   third-worst
[13] best-worst   second-worst  third-worst   worst-worst
10 Levels: best-best best-second best-third ... worst-worst
```

We now fit this model:

```
mods <- glm(y ~ symfac, eyegrade, family=poisson)
c(deviance(mods),df.residual(mods))
[1] 19.249 6.000
pchisq(deviance(mods),df.residual(mods),lower=F)
[1] 0.0037629
```

Here, we see evidence of a lack of symmetry. It is worth checking the residuals:

```
round(xtabs(residuals(mods) ~ right+left, eyegrade),3)

left
right   best   second   third   worst
best     0.000  1.001  0.317  2.008
second  -1.023  0.000  1.732 -0.225
third   -0.320 -1.783  0.000  0.928
worst   -2.219  0.223 -0.949  0.000
```

We see that the residuals above the diagonal are mostly positive, while they are mostly negative below the diagonal. So there are generally more poor left, good right eye combinations than the reverse. Furthermore, we can compute the marginals:

```
margin.table(ct,1)
right
  best second third worst
  1976  2256  2456  789
margin.table(ct,2)
left
  best second third worst
  1907  2222  2507  841
```

We see that there are somewhat more poor left eyes and good right eyes, so perhaps marginal homogeneity does not hold here. The assumption of symmetry implies marginal homogeneity (the reverse is not necessarily true). We may observe data where there is a difference in the frequencies of the levels of the rows and columns, but still be interested in symmetry. Suppose we set:

$$p_{ij} = \alpha_i \beta_j \gamma_{ij}$$

where  $\gamma_{ij} = \gamma_{ji}$ . This will allow for some symmetry while allowing for different

marginals. This is the *quasi-symmetry* model. Now:

$$\log EY_{ij} = \log np_{ij} = \log n + \log \alpha_i + \log \beta_j + \log \gamma_{ij}$$

So we can fit this model using:

```
modq <- glm(y ~ right+left+symfac, eyegrade, family=poisson)
pchisq(deviance(modq), df.residual(modq), lower=F)
[1] 0.06375
```

We see that this model does fit. It can be shown that marginal homogeneity together with quasi-symmetry implies symmetry. One can test for marginal homogeneity by comparing the symmetry and quasi-symmetry models:

```
anova(mods, modq, test="Chi")
```

Analysis of Deviance Table

	Model 1: y ~ symfac	Model 2: y ~ right + left + symfac			
	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi )
1	6	19.25			
2	3	7.27	3	11.98	0.01

So we find evidence of a lack of marginal homogeneity. This test is only appropriate if quasi-symmetry already holds.

When we examine the data here, we do see that many people do have symmetric vision. These entries lie down the diagonal. We might ask whether there is independence between left and right eyes among those people whose vision is not symmetric. This is the *quasi-independence* hypothesis and we can test it by omitting the data from the diagonal:

```
modqi <- glm(y ~ right + left, eyegrade, family=poisson, subset= -c
               ↪ (1, 6, 11, 16))
pchisq(deviance(modqi), df.residual(modqi), lower=F)
[1] 4.4118e-41
```

This model does not fit. This is not surprising since we can see that the entries adjacent to the diagonal are larger than those farther away. The difference in vision between the two eyes is likely to be smaller than expected under independence.

## 6.5 Three-Way Contingency Tables

In Appleton et al. (1996), a 20-year follow-up study on the effects of smoking is presented. In the period 1972–1974, a larger study, which also considered other issues, categorized women into smokers and nonsmokers and according to their age group. In the follow-up, the researchers recorded whether the subjects were dead or still alive. Only smokers or women who had never smoked are presented here. Relatively few smokers quit and these women have been excluded from the data. The cause of death is not reported here. Here is the data:

```
data(femsmoke)
femsmoke
      y smoker dead   age
1     2     yes  yes 18-24
2     1     no   yes 18-24
3     3     yes  yes 25-34
.....
28    0     no   no   75+
```

We can combine the data over age groups to produce:

```
(ct <- xtabs(y ~ smoker+dead, femsmoke))
  dead
smoker yes no
  yes 139 443
  no 230 502
```

We can compute the proportions of dead and alive for smokers and nonsmokers:

```
prop.table(ct,1)
  dead
smoker yes     no
  yes 0.23883 0.76117
  no  0.31421 0.68579
```

We see that 76% of smokers have survived for 20 years while only 69% of nonsmokers have survived. Thus smoking appears to have a beneficial effect on longevity. We can check the significance of this difference:

```
summary(ct)
Call: xtabs(formula = y ~ smoker + dead, data = femsmoke)
Number of cases in table: 1314
Number of factors: 2
Test for independence of all factors:
  Chisq = 9.1, df = 1, p-value = 0.0025
```

So the difference cannot be reasonably ascribed to chance variation. However, if we consider the relationship within a given age group, say 55–64:

```
(cta <- xtabs(y ~ smoker+dead, femsmoke, subset=(age=="55-64")))
  dead
smoker yes no
  yes 51 64
  no 40 81
prop.table(cta,1)
  dead
smoker yes     no
  yes 0.44348 0.55652
  no  0.33058 0.66942
```

We see that 56% of the smokers have survived compared to 67% of the nonsmokers. This advantage to nonsmokers holds throughout all the age groups. Thus the *marginal* association where we add over the age groups is different from the *conditional* association observed within age groups. Data where this effect is observed are an example of *Simpson's paradox*. The paradox is named after Simpson (1951), but dates back to Yule (1903).

Let's see why the effect occurs here:

```
prop.table(xtabs(y ~ smoker+age, femsmoke),2)
  age
smoker 18-24 25-34 35-44 45-54 55-64 65-74 75+
  yes 0.47009 0.44128 0.47391 0.62500 0.48729 0.21818 0.16883
  no  0.52991 0.55872 0.52609 0.37500 0.51271 0.78182 0.83117
```

We see that smokers are more concentrated in the younger age groups and younger people are more likely to live for another 20 years. This explains why the marginal table gave an apparent advantage to smokers which is, in fact, illusory because once we control for age, we see that smoking has a negative effect on longevity.

It is interesting to note that the dependence in the 55–64 age group is not statistically significant:

```
fisher.test(cta)
```

Fisher's Exact Test for Count Data

```
data: cta
p-value = 0.08304
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.92031 2.83340
sample estimates:
odds ratio
1.6103
```

However, this is just a subset of the data. Suppose we compute the odds ratios in all the age groups:

```
ct3 <- xtabs(y ~ smoker+dead+age, femsmoke)
apply(ct3, 3, function(x) (x[1,1]*x[2,2])/(x[1,2]*x[2,1]))
 18-24   25-34   35-44   45-54   55-64   65-74   75+
2.30189 0.75372 2.40000 1.44175 1.61367 1.14851   NaN
```

We see that there is some variation in the odds ratio, but they are all greater than one with the exception of the 25–34 age group. We could test for independence in each  $2 \times 2$  table, but it is better to use a combined test. The Mantel–Haenszel test is designed to test independence in  $2 \times 2$  tables across  $K$  strata. It only makes sense to use this test if the relationship is similar in each stratum. For this data, the observed odds ratios do not vary greatly, so the use of the test is justified.

Let the entries in the  $2 \times 2 \times K$  table be  $y_{ijk}$ . If we assume a hypergeometric distribution in each  $2 \times 2$  table, then  $y_{11k}$  is sufficient for each table given that we assume that the marginal totals for each table carry no information. The Mantel–Haenszel statistic is:

$$\frac{(|\sum_k y_{11k} - \sum_k E y_{11k}| - 1/2)^2}{\sum_k \text{var } y_{11k}}$$

where the expectation and variance are computed under the null hypothesis of independence in each stratum. The statistic is approximately  $\chi_1^2$  distributed under the null, although it is possible to make an exact calculation for smaller datasets. The statistic as stated above is due to Mantel and Haenszel (1959), but a version without the half-continuity correction was published by Cochran (1954). For this reason, it is sometimes known as the Cochran–Mantel–Haenszel statistic.

We compute the statistic for the data here:

```
mantelhaen.test(ct3, exact=TRUE)
Exact conditional test of independence in 2 x 2 x k tables
```

```
data: ct3
S = 139, p-value = 0.01591
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
1.0689 2.2034
sample estimates:
common odds ratio
1.5303
```

We used the exact method in preference to the approximation. We see that a statistically significant association is revealed once we combine the information across strata.

Now let's consider a linear model approach to investigating how the three factors interact. Let  $p_{ijk}$  be the probability that an observation falls into the  $(i, j, k)$  cell. Let

$p_i$  be the marginal probability that the observation falls into the  $i^{th}$  cell of the first variable,  $p_j$  be the marginal probability that the observation falls into the  $j^{th}$  cell of the second variable and  $p_k$  be the marginal probability that the observation falls into the  $k^{th}$  cell of the third variable.

**Mutual Independence:** If all three variables are independent, then:

$$p_{ijk} = p_i p_j p_k$$

Now  $EY_{ijk} = np_{ijk}$  so:

$$\log EY_{ijk} = \log n + \log p_i + \log p_j + \log p_k$$

So the main effects-only model corresponds to mutual independence. The coding we use will determine exactly how the parameters relate to the margin totals of the table although typically we will not be especially interested in these. Since independence is the simplest possibility, this model is the null model in an investigation of this type. The model  $\log EY_{ijk} = \mu$  would suggest that all the cells have equal probability. It is rare that such a model would have any interest so the model above makes for a more appropriate null.

We can test for independence using the Pearson's  $\chi^2$  test:

```
summary(ct3)
```

Call: xtabs(formula = y ~ smoker + dead + age, data = femsmoke)

Number of cases in table: 1314

Number of factors: 3

Test for independence of all factors:

Chisq = 791, df = 19, p-value = 2.1e-155

We can also fit the appropriate linear model:

```
modi <- glm(y ~ smoker + dead + age, femsmoke, family=poisson)
c(deviance(modi), df.residual(modi))
```

[1] 735 19

Although the statistics for the two tests are somewhat different, in either case, we see a very large value given the degrees of freedom. We conclude that this model does not fit the data.

We can show that the coefficients of this model correspond to the marginal proportions. For example, consider the smoker factor:

```
(coefs烟民 <- exp(c(0, coef(modi)[2])))
```

smokerno

1.0000 1.2577

```
coefs烟民/sum(coefs烟民)
```

smokerno

0.44292 0.55708

We see that these are just the marginal proportions for the smokers and nonsmokers in the data:

```
prop.table(xtabs(y ~ smoker, femsmoke))
```

smoker

yes no

0.44292 0.55708

This just serves to emphasize the point that the main effects of the model just convey information that we already know and is not the main interest of the study.

**Joint Independence:** Let  $p_{ij}$  be the (marginal) probability that the observation falls into a  $(i, j, \cdot)$  cell where any value of the third variable is acceptable. Now suppose that the first and second variable are dependent, but jointly independent of the third. Then:

$$p_{ijk} = p_{ij}p_k$$

We can represent this as:

$$\log EY_{ijk} = \log n + \log p_{ij} + \log p_k$$

Using the hierarchy principle, we would also include the main effects corresponding to the interaction term  $\log p_{ij}$ . So the log-linear model with just one interaction term corresponds to joint independence. The specific interaction term tells us which pair of variables is dependent. For example, we fit a model that says age is jointly independent of smoking and life status:

```
modj <- glm(y ~ smoker*dead + age, femsmoke, family=poisson)
c(deviance(modj), df.residual(modj))
```

[1] 725.8 18.0

Although this represents a small improvement over the mutual independence model, the deviance is still very high for the degrees of freedom and it is clear that this model does not fit the data. There are two other joint independence models that have the two other interaction terms. These models also fit badly.

**Conditional Independence:** Let  $p_{ijk|k}$  be the probability that an observation falls in cell  $(i, j, \cdot)$  given that we know the third variable takes the value  $k$ . Now suppose we assert that the first and second variables are independent given the value of the third variable, then:

$$p_{ijk|k} = p_{i|k}p_{j|k}$$

which leads to:

$$p_{ijk} = p_{ik}p_{jk}/p_k$$

This results in the model:

$$\log EY_{ijk} = \log n + \log p_{ik} + \log p_{jk} - \log p_k$$

Again, using the hierarchy principle, we would also include the main effects corresponding to the interaction terms and we would have model with main effects and two interaction terms. The minus for the  $\log p_k$  term is irrelevant because we do not care about the main effects parameters in these models. The nature of the conditional independence can be determined by observing which of one of the three possible two-way interactions does not appear in the model.

The most plausible conditional independence model for our data is:

```
modc <- glm(y ~ smoker*age + age*dead, femsmoke, family=poisson)
c(deviance(modc), df.residual(modc))
```

[1] 8.327 7.000

We see that the deviance is only slightly larger than the degrees of freedom indicating a fairly good fit. This indicates that smoking is independent of life status given age. However, bear in mind that we do have some zeroes and other small numbers in the

table and so there is some doubt as to the accuracy of the  $\chi^2$  approximation here. It is generally better to compare models rather than assess the goodness of fit.

**Uniform Association:** We might consider a model with all two-way interactions:

$$\log EY_{ijk} = \log n + \log p_i + \log p_j + \log p_k + \log p_{ij} + \log p_{ik} + \log p_{jk}$$

The model has no three-way interaction and so it is not saturated. There is no simple interpretation in terms of independence. Consider our example:

```
modu <- glm(y ~ (smoker+age+dead)^2, femsmoke, family=poisson)
```

Now we compute the fitted values and determine the odds ratios for each age group based on these fitted values:

```
ctf <- xtabs(fitted(modu) ~ smoker+dead+age, femsmoke)
apply(ctf, 3, function(x) (x[1,1]*x[2,2])/(x[1,2]*x[2,1]))
18-24 25-34 35-44 45-54 55-64 65-74 75+
1.5333 1.5333 1.5333 1.5333 1.5333 1.5333 1.5333
```

We see that the odds ratio is the same for every age group. Thus the uniform association model asserts that for every level of one variable, we have the same association for the other two variables.

The information may also be extracted from the coefficients of the fit. Consider the odds ratio for smoking and life status for a given age group:

$$(EY_{11k}EY_{22k})/(EY_{12k}EY_{21k})$$

This will be precisely the coefficient for the smoking and life-status term. We extract this:

```
exp(coef(modu)[ 'smokerno:deadno' ])
smokerno:deadno
1.5333
```

We see that this is exactly the odds ratio we found above. The other interaction terms may be interpreted similarly.

**Model Selection:** Log-linear models are hierarchical, so it makes sense to start with the most complex model and see how far it can be reduced. We can use analysis of deviance to compare models. We start with the saturated model:

```
modsat <- glm(y ~ smoker*age*dead, femsmoke, family=poisson)
drop1(modsat,test="Chi")
```

Single term deletions

Model:  
y ~ smoker \* age \* dead  
Df Deviance AIC LRT Pr(Chi)  
<none> 3.0e-10 190.2  
smoker:age:dead 6 2.4 180.6 2.4 0.88

We see that the three-way interaction term may be dropped. Now we consider dropping the two-way terms:

```
drop1(modu,test="Chi")
```

Single term deletions

Model:  
y ~ (smoker + age + dead)^2  
Df Deviance AIC LRT Pr(Chi)  
<none> 2 181

```
smoker:age   6      93 259  90  <2e-16
smoker:dead  1      8 185   6   0.015
age:dead     6      632 798 630  <2e-16
```

Two of the interaction terms are strongly significant, but the `smoker:dead` term is only just statistically significant. This term corresponds to the test for conditional independence of smoking and life status given age group. We see that the conditional independence does not hold. This tests the same hypothesis as the Mantel-Haenszel test above. In this case the *p*-values for the two tests are very similar.

**Binomial Model:** For some three-way tables, it may be reasonable to regard one variable as the response and the other two as predictors. In this example, we could view life status as the response. Since this variable has only two levels, we can model it using a binomial GLM. For more than two levels, a multinomial model would be required.

We construct a binomial response model:

```
ybin <- matrix(femsmoke$y, ncol=2)
modbin <- glm(ybin ~ smoker*age, femsmoke[1:14,], family=binomial)
```

This model is saturated, so we investigate a simplification:

```
drop1(modbin, test="Chi")
```

Single term deletions

```
Model:
ybin ~ smoker * age
          Df Deviance AIC LRT Pr(Chi)
<none>      5.3e-10 75.0
smoker:age  6      2.4 65.4  2.4    0.88
```

We see that the interaction term may be dropped, but now we check if we may drop further terms:

```
modbinr <- glm(ybin ~ smoker+age, femsmoke[1:14,], family=binomial)
drop1(modbinr, test="Chi")
```

Single term deletions

```
Model:
ybin ~ smoker + age
          Df Deviance AIC LRT Pr(Chi)
<none>      2   65
smoker  1      8   69   6   0.015
age     6      632 683 630  <2e-16
```

We see that both main effect terms are significant, so no further simplification is possible. This model is effectively equivalent to the uniform association model above. Check the deviances:

```
deviance(modu)
```

```
[1] 2.3809
```

```
deviance(modbinr)
```

```
[1] 2.3809
```

We see that they are identical. We can extract the same odds ratio from the parameter estimates as above:

```
exp(-coef(modbinr) [2])
```

```
smokerno
```

```
1.5333
```

The change in sign is simply due to which outcome is considered a success in the binomial GLM. So we can identify the binomial GLM with a corresponding Pois-

son GLM and the numbers we will obtain will be identical. We would likely prefer the binomial analysis where one factor can clearly be identified as the response and we would prefer the Poisson GLM approach when the relationship between the variables is more symmetric. However, there is one important difference between the two approaches. The null model for the binomial GLM:

```
modbinull <- glm(ybin ~ 1, femsmoke[1:14,], family=binomial)
deviance(modbinull)
```

[1] 641.5

is associated with this two-way interaction model for the Poisson GLM:

```
modj <- glm(y ~ smoker*age + dead, femsmoke, family=poisson)
deviance(modj)
```

[1] 641.5

So the binomial model implicitly assumes an association between `smoker` and `age`. In this particular dataset, there are more younger smokers than older ones, so the association is present. However, what if there was no association? One could argue that the Poisson GLM approach would be superior because it would allow us to drop this term and achieve a simpler model. On the other hand, one could argue that if the relationship between the response and the two predictors is the main subject of interest, then we lose little by conditioning out the marginal combined effect of age and smoking status, whether it is significant or not.

Our conclusion is that smoking is associated with higher mortality after we adjust for age. We have seen this in three different, but related, ways. The Mantel-Haenszel test, the preference for the uniform association model and the binomial response model all point in this direction. The test for the fit of the conditional independence model suggested no association between smoking and mortality adjusting for age. However, this goodness of fit test is of inferior quality to the other tests so we prefer the association conclusion.

**Correspondence Analysis:** We cannot directly apply the correspondence analysis method described above for two-way tables. However, we could combine two of the factors into a single factor by considering all possible combinations of the two levels. To make the choice of which two levels to combine, we would pick the pair whose association is least interesting to us. We could apply this to the smoking dataset here, but because there are only two levels of smoking and life status, the plot is not very interesting.

## 6.6 Ordinal Variables

Some variables have a natural order. We can use the methods for nominal variables described earlier in this chapter, but more information can be extracted by taking advantage of the structure of the data. Sometimes we might identify a particular ordinal variable as the response. In such cases, the methods of Section 7.4 can be used. However, sometimes we are interested in modeling the association between ordinal variables. Here the use of *scores* can be helpful.

Consider a two-way table where both variables are ordinal. We may assign scores  $u_i$  and  $v_j$  to the rows and columns such that  $u_1 \leq u_2 \leq \dots \leq u_I$  and  $v_1 \leq v_2 \leq \dots \leq v_J$ . The assignment of scores requires some judgment. If you have no particular prefer-

ence, even spacing allows for the simplest interpretation. If you have an interval scale, for example, 0–10 years old, 10–20 years old, 20–40 years old and so on, midpoints are often used. It is a good idea to check that the inference is robust to the assignment of scores by trying some reasonable alternative choices. If your qualitative conclusions are changed, this is an indication that you cannot make any strong finding.

Now fit the *linear-by-linear association* model:

$$\log EY_{ij} = \log \mu_{ij} = \log np_{ij} = \log n + \alpha_i + \beta_j + \gamma u_i v_j$$

So  $\gamma = 0$  means independence while  $\gamma$  represents the amount of association and can be positive or negative.  $\gamma$  is rather like an (unscaled) correlation coefficient. Consider underlying (latent) continuous variables which are discretized by the cutpoints  $u_i$  and  $v_j$ . We can then identify  $\gamma$  with the correlation coefficient of the latent variables.

Consider an example drawn from a subset of the 1996 American National Election Study (Rosenstone et al. (1997)). Using just the data on party affiliation and level of education, we can construct a two-way table:

```
data(nes96)
xtabs(~ PID + educ, nes96)
```

	educ						
PID	MS	HSdrop	HS	Coll	CCdeg	BAdeg	MAdeg
strDem	5	19	59	38	17	40	22
weakDem	4	10	49	36	17	41	23
indDem	1	4	28	15	13	27	20
indInd	0	3	12	9	3	6	4
indRep	2	7	23	16	8	22	16
weakRep	0	5	35	40	15	38	17
strRep	1	4	42	33	17	53	25

Both variables are ordinal in this example. We need to convert this to a dataframe with one count per line to enable model fitting.

```
(partyed <- as.data.frame.table(xtabs(~ PID + educ, nes96)))
```

	PID	educ	Freq
1	strDem	MS	5
2	weakDem	MS	4
3	indDem	MS	1
...etc....			

If we fit a nominal-by-nominal model, we find no evidence against independence:

```
nomod <- glm(Freq ~ PID + educ, partyed, family= poisson)
pchisq(deviance(nomod), df.residual(nomod), lower=F)
[1] 0.26961
```

However, we can take advantage of the ordinal structure of both variables and define some scores. As there seems to be no strong reason to the contrary, we assign evenly spaced scores—one to seven for both PID and educ:

```
partyed$oPID <- unclass(partyed$PID)
partyed$oeduc <- unclass(partyed$educ)
```

The `unclass` function converts a factor to an integer, numbered by level. It achieves exactly the assignment of scores we desire here. Now fit the linear-by-linear association model and compare to the independence model:

```
ormod <- glm(Freq ~ PID + educ + I(oPID*oeduc), partyed, family=
  ↪ poisson)
anova(nomod, ormod, test="Chi")
```

## Analysis of Deviance Table

```

Model 1: Freq ~ PID + educ
Model 2: Freq ~ PID + educ + I(oPID * oeduc)
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1          36      40.7
2          35      30.6  1      10.2    0.0014

```

We see that there is evidence of an association. We find that using the ordinal information gives us more power to detect an association. We can examine  $\hat{\gamma}$ :

```
summary(ormod)$coef['I(oPID * oeduc)',]
```

Estimate	Std. Error	z value	Pr(> z )
0.0287446	0.0090617	3.1720850	0.0015135

We see that  $\hat{\gamma}$  is 0.0287. The *p*-value here can also be used to test the significance of the association although, as a Wald test, it is less reliable than the likelihood ratio test we used first. We see that  $\hat{\gamma}$  is positive, which, given the way that we have assigned the scores, means that a higher level of education is associated with a greater probability of tending to the Republican end of the spectrum.

Just to check the robustness of the assignment of the scores, it is worth trying some different choices. For example, suppose we choose scores so that there is more of a distinction between Democrats and Independents as well as Independents and Republicans. Our assignment of scores for apid below achieves this. Another idea might be that people who complete high school or less are not different; that those who go to college, but do not get a BA degree are not different and that those who get a BA or higher are not different. My assignment of scores in aedu achieves this:

```

apid <- c(1,2,5,6,7,10,11)
aedu <- c(1,1,1,2,2,3,3)
ormoda <- glm(Freq ~ PID + educ + I(apid[oPID]*aedu[oeduc]), partyed,
  ↪ family= poisson)
anova(nomod,ormoda,test="Chi")

```

## Analysis of Deviance Table

```

Model 1: Freq ~ PID + educ
Model 2: Freq ~ PID + educ + I(apid[oPID] * aedu[oeduc])
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1          36      40.7
2          35      30.9  1      9.8    0.0017

```

The numerical outcome is slightly different, but the result is still significant. Some experimentation with other plausible choices indicates that we can be fairly confident about the association here.

The association parameter may be interpreted in terms of log-odds. For example, consider the log-odds ratio for adjacent entries in both rows and columns:

$$\log \frac{\mu_{ij}\mu_{i+1,j+1}}{\mu_{i,j+1}\mu_{i+1,j}} = \gamma(u_{i+1} - u_i)(v_{j+1} - v_j)$$

For evenly spaced scores, these log-odds ratios will all be equal. For our example, where the scores are spaced one apart, the log-odds ratio is  $\gamma$ . To illustrate this point, consider the fitted values under the linear-by-linear association model:

```
round(xtabs(predict(ormod,type="response")) ~ PID + educ, partyed),2)
```

educ	PID	MS	HSdrop	HS	Coll	CCdeg	BAdeg	MAdeg
------	-----	----	--------	----	------	-------	-------	-------

strDem	3.58	13.36	59.22	41.34	18.34	42.46	21.71
weakDem	2.92	11.22	51.20	36.78	16.80	40.02	21.06
indDem	1.59	6.27	29.45	21.78	10.23	25.09	13.59
indind	0.49	2.00	9.65	7.34	3.55	8.96	5.00
indRep	1.12	4.71	23.41	18.33	9.13	23.70	13.60
weakRep	1.61	6.95	35.59	28.68	14.69	39.28	23.19
strRep	1.69	7.49	39.48	32.74	17.26	47.49	28.85

Now compute log-odds ratio for, say, the lower two-by-two table:

```
log(39.28*28.85/(47.49*23.19))
```

```
[1] 0.028585
```

We see this is, but for rounding, equal to  $\hat{\gamma}$ .

It is always worth examining the residuals to check if there is more structure than the model suggests. We use the raw response residuals (the unscaled difference between observed and expected) because we would like to see effects that are large in an absolute sense.

```
round(xtabs(residuals(ormod, type="response")) ~ PID + educ, partyed), 2)
```

	educ	MS	HSDrop	HS	Coll	CCdeg	BAdeg	MAdeg
PID								
strDem	1.42	5.64	-0.22	-3.34	-1.34	-2.46	0.29	
weakDem	1.08	-1.22	-2.20	-0.78	0.20	0.98	1.94	
indDem	-0.59	-2.27	-1.45	-6.78	2.77	1.91	6.41	
indind	-0.49	1.00	2.35	1.66	-0.55	-2.96	-1.00	
indRep	0.88	2.29	-0.41	-2.33	-1.13	-1.70	2.40	
weakRep	-1.61	-1.95	-0.59	11.32	0.31	-1.28	-6.19	
strRep	-0.69	-3.49	2.52	0.26	-0.26	5.51	-3.85	

We do see some indications of remaining structure. For example, we see many more weak Republicans with some college than expected while fewer Republicans with master's degrees or higher. There may not be a monotone relationship between party affiliation and educational level.

To investigate this effect, we might consider an ordinal-by-nominal model where we now treat education as a nominal variable. This is called a *column effects* model because the columns (which are the education levels here) are not assigned scores and we will estimate their effect instead. A *row effects* model is effectively the same model except with the roles of the variables reversed. The model takes the form:

$$\log EY_{ij} = \log \mu_{ij} = \log np_{ij} = \log n + \alpha_i + \beta_j + u_i \gamma_j$$

where the  $\gamma_j$  are called the column effects. Equality of the  $\gamma_j$ 's corresponds to the hypothesis of independence. We fit this model for our data:

```
cmod <- glm(Freq ~ PID + educ + educ:oPID, partyed, family= poisson)
```

We can compare this to the independence model:

```
anova(nomod, cmod, test="Chi")
```

Analysis of Deviance Table

```
Model 1: Freq ~ PID + educ
Model 2: Freq ~ PID + educ + educ:oPID
Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      36      40.7
2      30      22.8  6     18.0   0.0063
```

We find that the column-effects model is preferred. Now examine the fitted coefficients, looking at just the interaction terms as the main effects have no particular interest:

```
summary(cmod)$coef[14:19,]
```

	Estimate	Std. Error	z value	Pr(> z )
educMS:oPID	-0.3122169	0.154051	-2.026710	0.042692
educHSdrop:oPID	-0.1944513	0.077228	-2.517891	0.011806
educHS:oPID	-0.0553470	0.048196	-1.148384	0.250810
educColl:oPID	0.0044605	0.050603	0.088147	0.929760
educCCdeg:oPID	-0.0086994	0.060667	-0.143395	0.885978
educBAdeg:oPID	0.0345539	0.048782	0.708330	0.478740

The last coefficient, educMAdeg:oPID, is not identifiable and so this may be taken as zero. If there was really a monotone trend in the effect of educational level on party affiliation, we would expect these coefficients to be monotone. However, we can see that they are not. However, if we compare this to the linear-by-linear association model:

```
anova(ormod, cmod, test="Chi")
```

Analysis of Deviance Table

Model	Freq ~ PID + educ + I(oPID * oeduc)	Model 2: Freq ~ PID + educ + educ:oPID	
Resid.	Df	Resid. Dev Df Deviance P(> Chi )	
1	35	30.57	
2	30	22.76	5 7.81 0.17

We see that the simpler linear-by-linear association is preferred to the more complex column-effects model. Nevertheless, if the linear-by-linear association were a good fit, we would expect the observed column-effect coefficients to be roughly evenly spaced. Looking at these coefficients, we observe that for high school and above, the coefficients are not significantly different from zero while for the lowest two categories, there is some difference. This suggests an alternate assignment of scores for education:

```
aedu <- c(1,1,2,2,2,2,2)
ormodb <- glm(Freq ~ PID + educ + I(oPID*aedu[oeduc]),
  partyed, family= poisson)
deviance(ormodb)
[1] 28.451
deviance(ormod)
[1] 30.568
```

We see that the deviance of this model is even lower than our original model. This gives credence to the view that whether a person finishes high school or not is the determining factor in party affiliation. However, since we used the data itself to assign the scores and come up with this hypothesis, we would be tempting fate to then use the data again to test this hypothesis.

The use of scores can be helpful in reducing the complexity of models for categorical data with ordinal variables. It is especially useful in higher dimensional tables where a reduction in the number of parameters is particularly welcome. The use of scores can also sharpen our ability to detect associations.

**Further Reading:** See books by Agresti (2013), Bishop et al. (1975), Haberman (1977), Le (1998), Leonard (2000), Powers and Xie (2000), Santner and Duffy (1989), Simonoff (2003) and Bilder and Loughin (2014).

### Exercises

1. The dataset `parstum` contains cross-classified data on marijuana usage by college students as it relates to the alcohol and drug usage of the parents.
  - (a) Display the data in a two-way table. Make a graphical display of the data and comment on the evidence of dependence between the two factors.
  - (b) Fit a nominal-by-nominal model for the count response. Does this model fit the data? What does this say about dependence?
  - (c) Make a Chi-squared test for independence. Compare to the previous test.
  - (d) Assign evenly spaced scores in the appropriate order to each of the two factors. Fit an ordinal-by-ordinal model. Report and interpret the coefficient of the interaction term in your model. What does the deviance of this model say about the quality of the fit?
  - (e) Test the significance of the interaction term by comparing this model to the nominal by nominal model.
  - (f) A researcher proposes that it does not matter whether one or both parents use drugs or alcohol. Make an assignment of scores consistent with this belief and compare the fit to the previous model.
2. The dataset `melanoma` gives data on a sample of patients suffering from melanoma (skin cancer) cross-classified by the type of cancer and the location on the body.
  - (a) Display the data in a two-way table. Make a mosaic plot and comment on the evidence for independence.
  - (b) Check for independence between site and tumor type using a Chi-squared test.
  - (c) Fit a Poisson GLM model and use it to check for independence.
  - (d) Make a two-way table of the deviance residuals from the last model. Comment on the larger residuals.
  - (e) Construct the correspondence plot. Interpret the plot.
  - (f) Omit all the head location data and repeat the test for independence. What does this indicate?
3. Data on social mobility of men in the United Kingdom may be found in `cmob`. A sample of men aged 45–64 was drawn from the 1971 census and 1981 census and the social class of man was recorded at each time point. The classes are I = professional, II = semiprofessional, IIIN = skilled nonmanual, IIIM = skilled manual, IV = semiskilled, V = unskilled.
  - (a) Display the data as a two-way table and comment on the arrangement around the diagonal.
  - (b) Fit a Poisson model and use it to check for independence.
  - (c) Create a symmetry factor and use it to check for symmetry between class over the two time points.
  - (d) Compute the difference between the matrix of data as seen in (a) and its transpose. Comment on the distribution of numbers above and below the diagonal. What does this say about the direction of social mobility?

- (e) Check for quasi-symmetry.
  - (f) Divide the data by ten to simulate the effect of a smaller sample and repeat the test for quasi-symmetry. What can be concluded from this?
  - (g) If possible, check for marginal homogeneity.
  - (h) Check for quasi-independence.
- (i) Make an assignment of scores to the levels of each factor and fit the ordinal by ordinal model. Is the association parameter significant?
4. The dataset `death` contains data on murder cases in Florida in 1977. The data is cross-classified by the race (black or white) of the victim, of the defendant and whether the death penalty was given.
- (a) Consider the frequency with which the death penalty is applied to black and white defendants, both marginally and conditionally, with respect to the race of the victim. Is this an example of Simpson's paradox? Are the observed differences in the frequency of application of the death penalty statistically significant?
  - (b) Determine the most appropriate dependence model between the variables.
  - (c) Fit a binomial regression with death penalty as the response and show the relationship to your model in the previous question.
5. The dataset `sexfun` comes from a questionnaire from 91 couples in the Tucson, Arizona, area. Subjects answered the question "Sex is fun for me and my partner." The possible answers were "never or occasionally," "fairly often," "very often" and "almost always."
- (a) Check for symmetry, quasi-symmetry, marginal homogeneity and quasi-independence.
  - (b) Develop a score-based model. Find some good-fitting scores.
6. The dataset `suicide` contains one year of suicide data from the United Kingdom cross-classified by sex, age and method.
- (a) Determine the most appropriate dependence model between the variables.
  - (b) Collapse the sex and age of the subject into a single six-level factor containing all combinations of sex and age. Conduct a correspondence analysis and give an interpretation of the plot.
  - (c) Repeat the correspondence analysis separately for males and females. Does this analysis reveal anything new compared to the combined analysis in the previous question?
7. A student newspaper conducted a survey of student opinions about the Vietnam War in May 1967. Responses were classified by sex, year in the program and one of four opinions. The survey was voluntary. The data may be found in the dataset `uncviet`.
- (a) Conduct an analysis of the patterns of dependence in the data assuming that all variables are nominal.

- (b) Assign scores to the year and opinion and fit an appropriate model. Interpret the trends in opinion over the years. Check the sensitivity of your conclusions to the assignment of the scores.
8. The dataset `HairEyeColor` contains the same data analyzed in this chapter as `haireye`. Repeat the analysis in the text for each sex and make a comparison of the conclusions.
9. A sample of psychiatry patients was cross-classified by their diagnoses and whether a drug treatment was prescribed. The data may be found in `drugpsy`. Is the chance that drugs will be prescribed constant across diagnoses?
10. The `UCBAdmissions` dataset presents data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex.
- (a) Show that this provides an example of Simpson's paradox.
  - (b) Determine the most appropriate dependence model between the variables.
  - (c) Fit a binomial regression with admissions status as the response and show the relationship to your model in the previous question.

## Multinomial Data

---

The multinomial distribution is an extension of the binomial where the response can take more than two values. Let  $Y_i$  be a random variable that falls into one of a finite number of categories, labeled  $1, 2, \dots, J$ . Let  $p_{ij} = P(Y_i = j)$  so  $\sum_{j=1}^J p_{ij} = 1$ . As with binary data (the case where  $J = 2$ ), we may encounter both grouped and ungrouped data. Let  $Y_{ij}$  be the number of observations falling into category  $j$  for group or individual  $i$  and let  $n_i = \sum_j Y_{ij}$ . For ungrouped data,  $n_i = 1$  and one and only one of  $Y_{i1}, \dots, Y_{iJ}$  is equal to one and the rest are zero. The  $Y_{ij}$ , conditional on the total  $n_i$ , follow a *multinomial* distribution:

$$P(Y_{i1} = y_{i1}, \dots, Y_{iJ} = y_{iJ}) = \frac{n_i}{y_{i1}! \cdots y_{iJ}!} p_{i1}^{y_{i1}} \cdots p_{iJ}^{y_{iJ}}$$

We must also distinguish between *nominal* multinomial data where there is no natural order to the categories and *ordinal* multinomial data where there is an order. The *multinomial logit* model is intended for nominal data. It can be used for ordinal data, but the information about order will not be used.

### 7.1 Multinomial Logit Model

As with the binary response model, we must find a way to link the probabilities  $p_{ij}$  to the predictors  $x_i$ , while ensuring that the probabilities are restricted between zero and one. We can use a similar idea:

$$\eta_{ij} = x_i^T \beta_j = \log \frac{p_{ij}}{p_{i1}}, \quad j = 2, \dots, J$$

We must obey the constraint that  $\sum_{j=1}^J p_{ij} = 1$ , so it is convenient to declare one of the categories as the *baseline*, say,  $j = 1$ . So we get  $p_{i1} = 1 - \sum_{j=2}^J p_{ij}$  and have:

$$p_{ij} = \frac{\exp(\eta_{ij})}{1 + \sum_{j=2}^J \exp(\eta_{ij})}$$

Note that  $\eta_{i1} = 0$ . It does not matter which category is declared as the baseline although some choices may be more convenient for interpretation. We may estimate the parameters of this model using maximum likelihood and then use the standard methods of inference.

Consider an example drawn from a subset of the 1996 American National Election Study (Rosenstone et al. (1997)). For simplicity, we consider only the age,

education level and income group of the respondents. Our response will be the party identification of the respondent: Democrat, Independent or Republican. The original data involved more than three categories; we collapse this to three, again for simplicity of the presentation.

```
data(nes96, package="faraway")
party <- nes96$PID
levels(party) <- c("Democrat", "Democrat", "Independent", "Independent",
  ↪ "Independent", "Republican", "Republican")
inca <- c(1.5, 4, 6, 8, 9.5, 10.5, 11.5, 12.5, 13.5, 14.5, 16, 18.5, 21, 23.5,
  ↪ 27.5, 32.5, 37.5, 42.5, 47.5, 55, 67.5, 82.5, 97.5, 115)
income <- inca[unclass(nes96$income)]
rnes96 <- data.frame(party, income, education=nes96$educ, age=nes96$age)
summary(rnes96)
```

	party	income	education	age
Democrat	:380	Min. : 1.5	MS : 13	Min. :19
Independent	:239	1st Qu.: 23.5	HSdrop: 52	1st Qu.:34
Republican	:325	Median : 37.5	HS :248	Median :44
		Mean : 46.6	Coll :187	Mean :47
		3rd Qu.: 67.5	CCdeg : 90	3rd Qu.:58
		Max. :115.0	BAdeg :227	Max. :91
			MAdeg :127	

The income variable in the original data was an ordered factor with income ranges. We have converted this to a numeric variable by taking the midpoint of each range. This corresponds to a score-based approach of handling interval-valued variables.

Let's start with a graphical look at the relationship between the predictors and the response. The response is at the individual level and so we need to group the data just to get a sense of how the party identification is associated with the predictors. Some manipulation of the data is facilitated with the `dplyr` package. We group the data by education and party affiliation, count the number in each category and then count the number in each education category, using this to compute the proportion supporting each party for each party affiliation. The resulting plot is shown in the first panel of Figure 7.1. We see that proportion of Democrats falls with educational status, reaching a plateau for the college educated. We see the proportion of Republicans rising with educational level and reaching a similar plateau.

```
library(dplyr)
egp <- group_by(rnes96, education, party) %>% summarise(count=n()) %>%
  ↪ group_by(education) %>% mutate(etotal=sum(count), proportion=
  ↪ count/etotal)
ggplot(egp, aes(x=education, y=proportion, group=party, linetype=party
  ↪ ))+geom_line()
```

A similar calculation is made for income but unfortunately there are relatively low counts in some income categories making the computation of a stable proportion difficult. To overcome this, we first group the income into seven groups of roughly equal size and then follow much the same calculation as for education.

```
igp <- mutate(rnes96, incomegp=cut_number(income,7)) %>% group_by(
  ↪ incomegp, party) %>% summarise(count=n()) %>% group_by(incomegp
  ↪ ) %>% mutate(etotal=sum(count), proportion=count/etotal)
ggplot(igp, aes(x=incomegp, y=proportion, group=party, linetype=party
  ↪ ))+geom_line()
```

The plot is shown in the second panel of Figure 7.1. As income increases, we observe

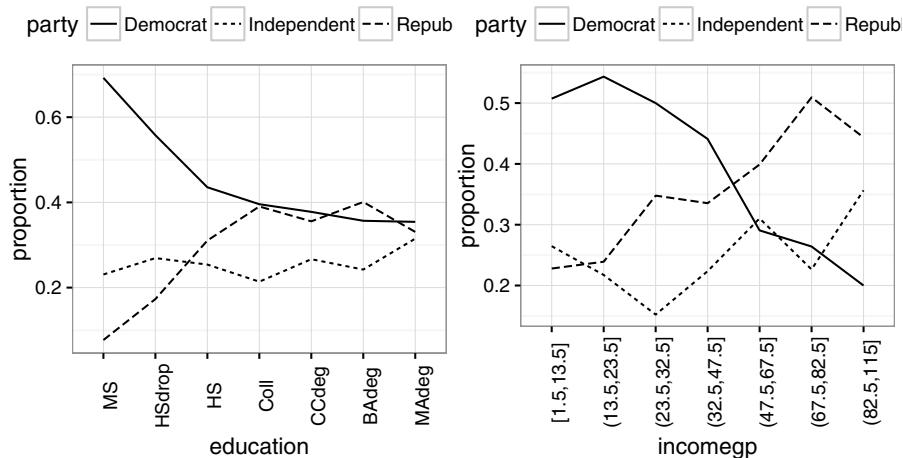


Figure 7.1 *Relationship between party affiliation and education, age and income. Democrats are shown with solid line, Republicans with a dashed line and Independents with a dotted line. Education is categorized into seven levels described in the text. Income is in thousands of dollars.*

an increase in the proportion of Republicans and Independents and a decrease in the proportion of Democrats. A similar plot can be made for age where the relationship of party to age is not clear. This is cross-sectional rather than longitudinal data, so we cannot say anything about what might happen to an individual with, for example, increasing income. We can only expect to make conclusions about the relative probability of party affiliations for different individuals with different incomes.

We might ask whether the trends we see in the observed proportions are statistically significant. We need to model the data to answer this question. We fit a multinomial logit model. The `multinom` function is part of the `nnet` package described in Venables and Ripley (2002):

```
library(nnet)
mmod <- multinom(party ~ age + education + income, rnes96)
# weights: 30 (18 variable)
initial value 1037.090001
iter 10 value 990.568608
iter 20 value 984.319052
final value 984.166272
converged
```

The program uses the optimization method from the neural net trainer in `nnet` as described in Chapter 17 to compute the maximum likelihood. There is no deeper connection to neural networks — we just need the optimization found in that package.

We can select which variables to include in the model based the AIC criterion using a stepwise search method (output edited to show only the decision information):

```
mmodi <- step(mmod)
Start: AIC=2004.3
party ~ age + education + income
```

```
Df      AIC
- education 6 1996.5
- age       16 2003.6
<none>     18 2004.3
- income    16 2045.9
```

```
Step: AIC=1996.5
party ~ age + income
Df      AIC
- age     4 1993.4
<none>   6 1996.5
- income  4 2048.9
```

```
Step: AIC=1993.4
party ~ income
Df      AIC
<none>   4 1993.4
- income  2 2045.3
```

At the first stage of the search, we see that omitting education would be the best option to reduce the AIC criterion. At the next step, age is removed resulting in a model with only income.

We can also use the standard likelihood methods to derive a test to compare nested models. For example, we can fit a model without education and then compare the deviances:

```
mmod <- multinom(party ~ age + income, rnes96)
deviance(mmod) - deviance(mmod)
[1] 16.206
pchisq(16.206, mmod$edf-mmod$edf, lower=F)
[1] 0.18198
```

We see that education is not significant relative to the full model. This may seem somewhat surprising given the plot in Figure 7.1, but the large differences between proportions of Democrats and Republicans occur for groups with low education which represent only a small number of people.

We can obtain predicted values for specified values of income. We can compute the probability of party affiliation for a range of incomes in [0, 110].

```
inclevels <- 0:110
preds <- data.frame(income=inclevels, predict(mmodi, data.frame(income=
  ↪ inclevels), type="probs"))
library(tidyr)
lpred <- gather(preds, party, probability, -income)
ggplot(lpred, aes(x=income, y=probability, group=party, linetype=party)) +
  ↪ geom_line()
```

We see in Figure 7.2 how the probability of being Republican or Independent increases with income while the probability of being Democrat drops. The default form just gives the most probable category:

```
predict(mmodi, data.frame(income=inclevels))
[1] Democrat   Democrat   Democrat   Democrat   Democrat
.....
[111] Republican
```

The multinomial logit model can also be used in a predictive sense to classify new observations. We can see how well the selected model does in predicting the party of the known individuals in our dataset:

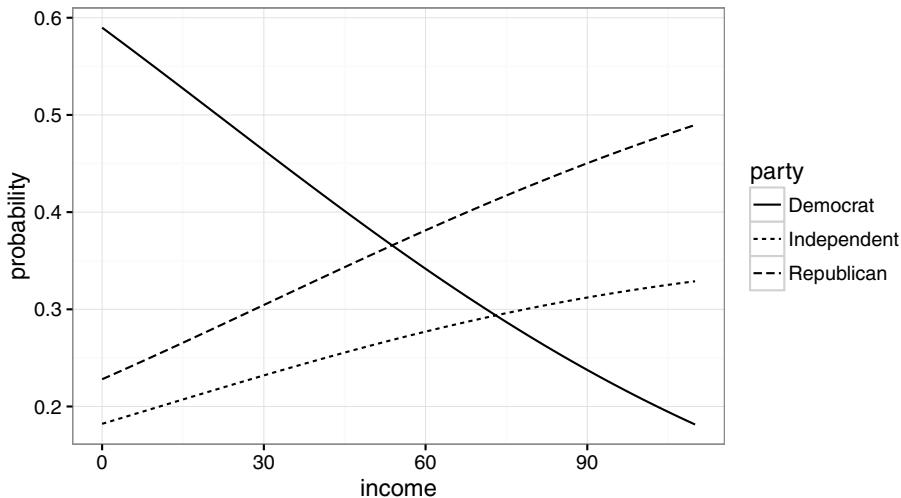


Figure 7.2 *Predicted probabilities of party affiliation as income varies (thousands).*

```
xtabs(~ predict(mmodi) + rnes96$party)
      rnes96$party
predict(mmodi) Democrat Independent Republican
Democrat        284         123        166
Independent        0          0          0
Republican        96         116        159
```

We can compute the proportion correctly classified as:

```
(284+0+159)/nrow(rnes96)
```

```
[1] 0.46928
```

We see that only 47% of the current data are correctly classified and we could expect the performance on new individuals to be slightly worse than this. No cases are classified as independents because the probability of the other two outcomes always dominates this outcome. We see that the majority of actual Republicans are classified as Democrats. So the performance is not impressive but given the large overlap in the predictor space between the three levels (or classes), it is not surprising.

The multinomial logit model is not usually the best choice for classification performance. Better performance can be obtained by methods specialized for this purpose such as random forests or support vector machines. Even so, the multinomial logit model often provides better insight into how and which predictors affect the classification. This understanding of how a classification method works can be invaluable in predicting how it will behave in new circumstances or how predictors might be added or modified. The more sophisticated methods are often opaque as to meaning so while they may perform well, they may add little to our understanding.

We can examine the coefficients to gain an understanding of the relationship between the predictor and the response:

```
summary(mmodi)
```

Coefficients:

```
(Intercept) nincome
```

```
Independent -1.17493 0.016087
Republican -0.95036 0.017665
```

Std. Errors:

	(Intercept)	nincome
Independent	0.15361	0.0028497
Republican	0.14169	0.0026525

Residual Deviance: 1985.4

AIC: 1993.4

The intercept terms model the probabilities of the party identification for an income of zero. We can see the relationship from this calculation:

```
cc <- c(0, -1.17493, -0.95036)
exp(cc) / sum(exp(cc))
[1] 0.58982 0.18216 0.22802
predict(mmodi, data.frame(income=0), type="probs")
```

Democrat	Independent	Republican
0.58982	0.18216	0.22802

The slope terms represent the log-odds of moving from the baseline category of Democrat to Independent and Republican, respectively, for a unit change of \$1000 in income. We can see more explicitly what this means by predicting probabilities for incomes \$1000 apart and then computing the log-odds:

```
(pp <- predict(mmodi, data.frame(income=c(0, 1)), type="probs"))
Democrat Independent Republican
1 0.58982      0.18216     0.22802
2 0.58571      0.18382     0.23047
log(pp[1,1]*pp[2,2]/(pp[1,2]*pp[2,1]))
[1] 0.016087
log(pp[1,1]*pp[2,3]/(pp[1,3]*pp[2,1]))
[1] 0.017665
```

Log-odds can be difficult to interpret particularly with many predictors and interactions. Sometimes, computing predicted probabilities for a selected range of predictors can provide better intuition.

## 7.2 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a method based on linear combinations of the predictors that classifies cases into one of a fixed number of possibilities. This prediction is called *classification* and LDA is usually considered as a classification method. A multinomial logit model explains the relationship between a response that can take one of a fixed number of levels and linear combinations of the predictors. Although LDA focuses more on prediction and the multinomial logit more on explanation, we can learn from comparing the two. We give only a brief introduction to LDA here and the reader is encouraged to consult other sources regarding this method. We focus on the explanatory insights offered by LDA in this section.

Consider the vector of predictors  $\mathbf{x}_i$  for  $i = 1, \dots, n$ . We can compute a measure of total variation in the data using the total sum of squares  $S$ :

$$S = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

Suppose we have  $G$  groups and we add a group subscript  $g$  to denote which group a case belongs to. We define the within-group covariance  $W$  as:

$$W = \sum_{g=1}^G \sum_{i=1}^{n_g} (\mathbf{x}_{gi} - \bar{\mathbf{x}}_g)(\mathbf{x}_{gi} - \bar{\mathbf{x}}_g)^T$$

We define the between-groups covariance  $B$  as:

$$B = \sum_{g=1}^G (\bar{\mathbf{x}}_g - \bar{\mathbf{x}})(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})^T$$

where  $n_g$  is the number of cases in group  $g$ . We have  $S = W + B$ . We want to form linear combinations  $\mathbf{a}$  of the predictors to maximize the separation between the groups. To achieve this, we choose  $\mathbf{a}$  to maximize:

$$\frac{\mathbf{a}^T B \mathbf{a}}{\mathbf{a}^T W \mathbf{a}}$$

The solution can be found using eigendecomposition of  $W^{-1}B$ . The first eigenvector represents the combination which maximizes the ratio with subsequent eigenvectors representing the next best solutions subject to the requirement that they are orthogonal to previous eigenvectors.

The calculation can be performed using the MASS library.

```
library(MASS)
mlda <- lda(party ~ age + income, rnes96)
mlda
```

Prior probabilities of groups:  
Democrat    Independent    Republican  
0.40254    0.25318    0.34428

Group means:  
                    age income  
Democrat    47.066 37.632  
Independent 46.506 51.655  
Republican 47.412 53.298

Coefficients of linear discriminants:  
                    LD1       LD2  
age    0.0053257 0.060821  
income 0.0332155 -0.000114

Proportion of trace:  
LD1    LD2  
0.9926 0.0074

The prior probabilities are chosen by default as the observed proportion in the data. One can specify the prior if, for example, we have a biased sample in which the class sizes do not correspond to the population. The means in the groups,  $\bar{\mathbf{x}}_g$ , reveal that there is not much difference in the ages of the three groups but there are noticeable differences in income. The proportion of the trace is computed using the eigenvalues of the decomposition. In this case, we see that the first component is strongly dominant and so the classification will depend mostly on this. The coefficients give us

the **a**. We see that the first combination is dominated by the income while the second by the age. However, the second combination counts for little so the classification is based mainly on the income.

We can use the LDA to classify the current data:

```
preds <- predict(mlda)
head(preds$posterior)
```

	Democrat	Independent	Republican
1	0.58666	0.18945	0.22389
2	0.59537	0.19135	0.21328
3	0.59322	0.19089	0.21590
4	0.59105	0.19042	0.21854
5	0.56846	0.18533	0.24620
6	0.59483	0.19124	0.21393

The posterior probabilities computed from the model can be used to classify the cases. The first six cases are all classified as Democrats as this is highest probability in each case.

We can get the most likely outcome from each case and compare it against the observed class:

```
xtabs(~ predict(mlda)$class + rnes96$party)
```

	rnes96\$party			
	predict(mlda)\$class	Democrat	Independent	Republican
Democrat	298	141	184	
Independent	0	0	0	
Republican	82	98	141	

The result is quite similar to the multinomial logit in that no cases are classified as Independent and many true Republicans are classified as Democrats.

### 7.3 Hierarchical or Nested Responses

Consider the following data collected by Lowe et al. (1971) by way of McCullagh and Nelder (1989) concerning live births with deformations of the central nervous system in south Wales:

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

	Area	NoCNS	An	Sp	Other	Water	Work
1	Cardiff	4091	5	9	5	110	NonManual
2	Newport	1515	1	7	0	100	NonManual
3	Swansea	2394	9	5	0	95	NonManual
4	GlamorganE	3163	9	14	3	42	NonManual
5	GlamorganW	1979	5	10	1	39	NonManual
6	GlamorganC	4838	11	12	2	161	NonManual
7	MonmouthV	2362	6	8	4	83	NonManual
8	MonmouthOther	1604	3	6	0	122	NonManual
9	Cardiff	9424	31	33	14	110	Manual
10	Newport	4610	3	15	6	100	Manual
11	Swansea	5526	19	30	4	95	Manual
12	GlamorganE	13217	55	71	19	42	Manual
13	GlamorganW	8195	30	44	10	39	Manual
14	GlamorganC	7803	25	28	12	161	Manual
15	MonmouthV	9962	36	37	13	83	Manual
16	MonmouthOther	3172	8	13	3	122	Manual

NoCNS indicates no central nervous system (CNS) malformation. An denotes anencephalus while Sp denotes spina bifida and Other represents other malformations. Water is water hardness and the subjects are categorized by the type of work performed by the parents. We might consider a multinomial response with four categories. However, we can see that most births suffer no malformation and so this category dominates the other three. It is better to consider this as a hierarchical response as depicted in Figure 7.3. Now consider the multinomial likelihood for the  $i^{th}$

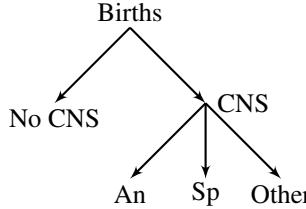


Figure 7.3 *Hierarchical response for birth types.*

observation which is proportional to:

$$p_{i1}^{y_{i1}} p_{i2}^{y_{i2}} p_{i3}^{y_{i3}} p_{i4}^{y_{i4}}$$

Define  $p_{ic} = p_{i2} + p_{i3} + p_{i4}$  which is probability of a birth with some kind of CNS malformation. We can then write the likelihood as:

$$p_{i1}^{y_{i1}} p_{ic}^{y_{i2}+y_{i3}+y_{i4}} \times \left( \frac{p_{i2}}{p_{ic}} \right)^{y_{i2}} \left( \frac{p_{i3}}{p_{ic}} \right)^{y_{i3}} \left( \frac{p_{i4}}{p_{ic}} \right)^{y_{i4}}$$

The first part of the product is now a binomial likelihood for a CNS vs. NoCNS response. The second part of the product is now a multinomial likelihood for the three CNS categories conditional of the presence of CNS. For example,  $p_{i2}/p_{ic}$  is the conditional probability of an anencephalus birth given that a malformation has occurred for the  $i^{th}$  observation. We can now separately develop a binomial model for whether malformation occurs and a multinomial model for the type of malformation.

We start with the binomial model. First we accumulate the number of CNS births and plot the data with the response on the logit scale as shown in the first panel of Figure 7.4:

```
cns$CNS <- cns$An+cns$Sp+cns$Other
plot(log(CNS/NoCNS) ~ Water, cns, pch=as.character(Work))
```

We observe that the proportion of CNS births falls with increasing water hardness and is higher for manual workers. We observe one observation (manual, Newport) that may be an outlier. Notice that the Area is confounded with the Water hardness, so we cannot put both these predictors in our model. But Water does represent a subspace of Area so it is legitimate to compare:

```
binmodw <- glm(cbind(CNS, NoCNS) ~ Water + Work, cns, family=binomial)
binmoda <- glm(cbind(CNS, NoCNS) ~ Area + Work, cns, family=binomial)
anova(binmodw, binmoda, test="Chi")
```

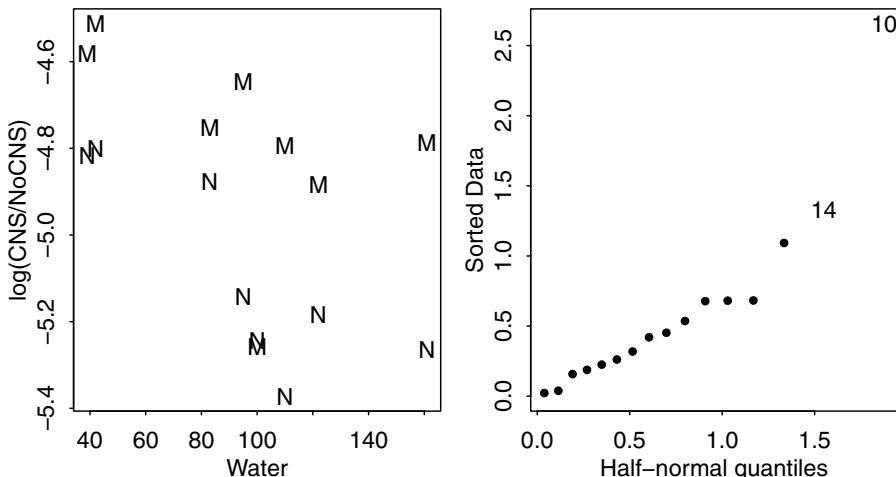


Figure 7.4 The first plot shows the empirical logits for the proportion of CNS births related to water hardness and profession (M=Manual, N=Nonmanual). The second is a half-normal plot of the residuals of the chosen model.

#### Analysis of Deviance Table

```
Model 1: cbind(CNS, NoCNS) ~ Water + Work
Model 2: cbind(CNS, NoCNS) ~ Area + Work
Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      13     12.36
2       7      3.08  6      9.29    0.16
```

One can view this test as a check for linear trend in the effect of water hardness. We find that the simpler model using Water is acceptable. A check for an interaction effect revealed nothing significant although a look at the residuals is worthwhile:

```
library(faraway)
halfnorm(residuals(binmodw))
```

In the second plot of Figure 7.4, we see an outlier corresponding to Newport manual workers. This case deserves closer examination. Finally, a look at the chosen model:

```
summary(binmodw)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-4.432580	0.089789	-49.37	< 2e-16
Water	-0.003264	0.000968	-3.37	0.00075
WorkNonManual	-0.339058	0.097094	-3.49	0.00048

```
n = 16 p = 3
Deviance = 12.363 Null Deviance = 41.047 (Difference = 28.685)
```

The residual deviance is close to the degrees of freedom indicating a reasonable fit to the data. We see that since:

```
exp(-0.339058)
[1] 0.71244
```

births to nonmanual workers have a 29% lower chance of CNS malformation. Water hardness ranges from about 40 to 160. So a difference of 120 would decrease the odds of CNS malformation by about 32%.

Now consider a multinomial model for the three malformation types conditional on a malformation having occurred. As this data is grouped, in contrast to the `nes96` example, it is most convenient to present the response as a matrix:

```
cmmod <- multinom(cbind(An, Sp, Other) ~ Water + Work, cns)
```

We find that neither predictor has much effect:

```
nmod <- step(cmmod)
```

	Df	AIC
- Water	4	1381.1
- Work	4	1381.2
<none>	6	1383.5

	Df	AIC
- Work	2	1378.5
<none>	4	1381.1

which leaves us with a null final model:

```
nmod
```

Coefficients:	
(Intercept)	
Sp	0.28963
Other	-0.98083

Residual Deviance: 1374.5

The fitted proportions are:

```
cc <- c(0, 0.28963, -0.98083)
names(cc) <- c("An", "Sp", "Other")
exp(cc)/sum(exp(cc))
```

An	Sp	Other
0.36888	0.49279	0.13833

So we find that water hardness and parents' professions are related to the probability of a malformed birth, but that they have no effect on the type of malformation.

Observe that if we fit a multinomial logit model to all four categories:

```
multinom(cbind(NoCNS, An, Sp, Other) ~ Water + Work, cns)
```

Coefficients:		Water	WorkNonManual
(Intercept)			
An	-5.4551	-0.00290884	-0.36388
Sp	-5.0710	-0.00432305	-0.24359
Other	-6.5947	-0.00051358	-0.64219

Residual Deviance: 9391

AIC: 9409

We find that both `Water` and `Work` are significant, but that the fact that they do not distinguish the type of malformation is not easily discovered from this model.

## 7.4 Ordinal Multinomial Responses

Suppose we have  $J$  ordered categories and that for individual  $i$ , with ordinal response  $Y_i$ ,  $p_{ij} = P(Y_i = j)$  for  $j = 1, \dots, J$ . With an ordered response, it is often easier to work with the cumulative probabilities,  $\gamma_{ij} = P(Y_i \leq j)$ . The cumulative probabilities are increasing and invariant to combining adjacent categories. Furthermore,  $\gamma_{iJ} = 1$ , so we need only model  $J - 1$  probabilities.

As usual, we must link the  $\gamma$ s to the covariates  $x$ . We will consider three possibilities which all take the form:

$$g(\gamma_{ij}) = \theta_j - x_i^T \beta$$

Possible link functions,  $g$ , are the logit, the probit and the complementary log-log. We have explicitly specified the intercepts,  $\theta_j$ , so that the vector  $x_i$  does not include an intercept. Furthermore,  $\beta$  does not depend on  $j$  so that we assume that the predictors have a uniform effect on the response categories in a sense that we will shortly make clear.

Suppose that  $Z_i$  is some unobserved continuous variable that might be thought of as the real underlying latent response. We only observe a discretized version of  $Z_i$  in the form of  $Y_i$  where  $Y_i = j$  is observed if  $\theta_{j-1} < Z_i \leq \theta_j$ . Further suppose that  $Z_i - \beta^T x_i$  has distribution  $F$ , then:

$$P(Y_i \leq j) = P(Z_i \leq \theta_j) = P(Z_i - \beta^T x_i \leq \theta_j - \beta^T x_i) = F(\theta_j - \beta^T x_i)$$

Now if, for example,  $F$  follows the logistic distribution, where  $F(x) = e^x / (1 + e^x)$ , then:

$$\gamma_{ij} = \frac{\exp(\theta_j - \beta^T x_i)}{1 + \exp(\theta_j - \beta^T x_i)}$$

and so we would have a logit model for the cumulative probabilities  $\gamma_{ij}$ . Choosing the normal distribution for the latent variable leads to a probit model, while the choice of an extreme value distribution leads to the complementary log-log. This *latent* variable explanation for the model is displayed in Figure 7.5. We include the R code for reference.

```
x <- seq(-5, 5, by=0.1)
xa <- c(21, 41, 71)
plot(x, dlogis(x), type="l", axes=FALSE, ylab="", xlab="")
segments(x[xa], rep(0, 3), x[xa], dlogis(x[xa]))
axis(1, at=x[xa], c(expression(theta[1]), expression(theta[2]), expression
  → (theta[3])))
```

Notice that if  $\beta > 0$ , as  $x_i$  increases,  $P(Y_i = J)$  will also increase. This explains the use of the minus sign in the definition of the model because it allows for the more intuitive interpretation of the sign of  $\beta$ .

**Proportional Odds Model:** Let  $\gamma_j(x_i) = P(Y_i \leq j | x_i)$ , then the *proportional odds* model, which uses the logit link, is:

$$\log \frac{\gamma_j(x_i)}{1 - \gamma_j(x_i)} = \theta_j - \beta^T x_i, \quad j = 1, \dots, J - 1$$

It is called this because the relative odds for  $Y \leq j$  comparing  $x_1$  and  $x_2$  are:

$$\left( \frac{\gamma_j(x_1)}{1 - \gamma_j(x_1)} \right) / \left( \frac{\gamma_j(x_2)}{1 - \gamma_j(x_2)} \right) = \exp(-\beta^T (x_1 - x_2))$$

This does not depend on  $j$ . Of course, the assumption of proportional odds does need to be checked for a given dataset.

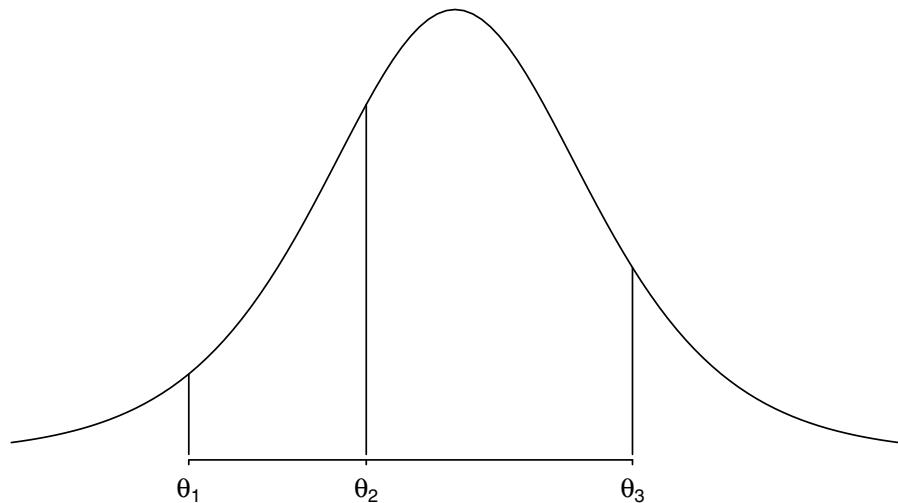


Figure 7.5 *Latent variable view of an ordered multinomial response. Here, four discrete responses can occur, depending on the position of Z relative to the cutpoints  $\theta_j$ . As x changes, the cutpoints will move together to change the relative probabilities of the four responses.*

Returning to the `nes96` dataset, suppose we assume that Independents fall somewhere between Democrats and Republicans. We would then have an ordered multinomial response. We can then fit this using the `polr` function from the MASS library described in Venables and Ripley (2002):

```
library(MASS)
pomod <- polr(party ~ age + education + income, rnes96)
```

The deviance and number of parameters for this model are:

```
c(deviance(pomod), pomod$edf)
```

```
[1] 1984.2 10.0
```

which can be compared to the corresponding multinomial logit model:

```
c(deviance(mmod), mmod$edf)
```

```
[1] 1968.3 18.0
```

The proportional odds model uses fewer parameters, but does not fit quite as well. Typically, the output from the proportional odds model is easier to interpret. We may use an AIC-based variable selection method:

```
pomodi <- step(pomod)
```

Start: AIC=2004.2

party ~ age + education + income

	Df	AIC
- education	6	2003
<none>		2004
- age	1	2004
- income	1	2039

Step: AIC=2002.8

party ~ age + income

```
Df AIC
- age      1 2001
<none>    2003
- income   1 2047
```

Step: AIC=2001.4  
party ~ income

```
Df AIC
<none> 2001
- income 1 2045
```

Thus we finish with a model including just income as we did with the earlier multinomial model. We could also use a likelihood ratio test to compare the models:

```
deviance(pomodi)-deviance(pomod)
[1] 11.151
pchisq(11.151,pomod$edf-pomodi$edf,lower=F)
[1] 0.13217
```

We see that the simplification to just income is justifiable. We can check the proportional odds assumption by computing the observed odds proportions with respect to, in this case, income levels. We have computed the log-odds difference between  $\gamma_1$  and  $\gamma_2$ :

```
pim <- with(rnes96,prop.table(table(income,party),1))
logit(pim[,1])-logit(pim[,1]+pim[,2])
 1.5      4       6      8     9.5     10.5     11.5
-0.90079 -2.06142 -0.75769 -1.00330 -2.30259 -0.30830 -0.79851
 12.5     13.5    14.5     16     18.5     21     23.5
-1.89712 -1.25276 -1.17865 -0.41285 -0.35424 -1.51413 -1.65345
 27.5     32.5    37.5     42.5    47.5     55     67.5
-0.74678 -0.52252 -0.92326 -1.02962 -0.82198 -1.42760 -1.18261
 82.5     97.5    115
-0.98676 -1.48292 -1.70660
```

It is questionable whether these can be considered sufficiently constant, but at least there is no trend. Now consider the interpretation of the fitted coefficients:

```
summary(pomodi)
Coefficients:
            Value Std. Error t value
income 0.013120 0.0019708 6.6572

Intercepts:
            Value Std. Error t value
Democrat|Independent 0.209 0.112 1.863
Independent|Republican 1.292 0.120 10.753
```

Residual Deviance: 1995.36

AIC: 2001.36

We can say that the odds of moving from Democrat to Independent/Republican category (or from Democrat/Independent to Republican) increase by a factor of  $\exp(0.013120) = 1.0132$  as income increases by one unit (\$1000). Notice that the log-odds are similar to those obtained in the multinomial logit model. The intercepts correspond to the  $\theta_j$ . So for an income of \$0, the predicted probability of being a Democrat is:

```
ilogit(0.209)
```

```
[1] 0.55206
```

while that of being an Independent is:

```
ilogit(1.292)-ilogit(0.209)
```

```
[1] 0.23242
```

with the remainder being Republicans. We can compute predicted values:

```
inclevels <- seq(0,100,by=20)
predict(pomodi,data.frame(income=inclevels, row.names=inclevels), type=
  ↪ "probs")
```

	Democrat	Independent	Republican
0	0.55209	0.23232	0.21559
20	0.48668	0.25007	0.26325
40	0.42173	0.26109	0.31718
60	0.35937	0.26411	0.37652
80	0.30143	0.25877	0.43980
100	0.24920	0.24569	0.50511

Notice how the probability of being a Democrat uniformly decreases with income while that for being a Republican uniformly increases as income increases, but that the middle category of Independent increases then decreases. This type of behavior can be expected from the latent variable representation of the model.

We can illustrate the latent variable interpretation of proportional odds by computing the cutpoints for incomes of \$0, \$50,000 and \$100,000:

```
x <- seq(-4,4,by=0.05)
plot(x,dlogis(x),type="l")
abline(v=c(0.209,1.292))
abline(v=c(0.209,1.292)-50*0.013120,lty=2)
abline(v=c(0.209,1.292)-100*0.013120,lty=5)
```

The plot is shown in Figure 7.6.

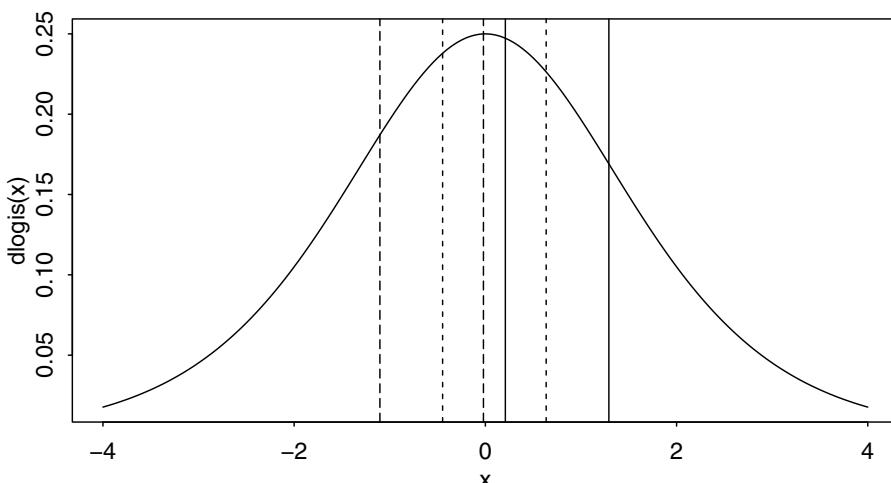


Figure 7.6 Solid lines represent an income of \$0, dotted lines are for \$50,000 and dashed lines are for \$100,000. The probability of being a Democrat is given by the area lying to the left of the leftmost of each pair of lines, while the probability of being a Republican is given by the area to the right of the rightmost of the pair. Independents are represented by the area in-between.

**Generalization:** The proportional odds models can be generalized by allowing  $\beta$  to vary by

$$\log \frac{\gamma_j(x)}{1 - \gamma_j(x)} = \theta_j - \beta_j^T x \quad j = 1, \dots, k - 1$$

but this loses the proportionality property. Such models cannot be fitted using `polr()` but are provided for in the `VGAM` package of Yee (2010). Here's how we fit such a model to the current data:

```
library(VGAM)
nmod <- vglm(party ~ income, family=cumulative(parallel=FALSE), rnes96)
summary(nmod)

Coefficients:
Estimate Std. Error z value
(Intercept):1 0.3289 0.12156 2.71
(Intercept):2 1.1483 0.12872 8.92
income:1 -0.0162 0.00236 -6.86
income:2 -0.0105 0.00220 -4.76
```

Residual deviance: 1987.5 on 1884 degrees of freedom

The parameterization is different for this package in that  $\theta_j + \beta_j^T x$  is used for the linear predictor replacing the  $-$  with a  $+$  in the `polr` form. The argument `parallel = FALSE` indicates we have chosen the nonproportional (or parallel) option.

We compare this to the proportional odds model and use the standard likelihood theory to compare the models:

```
pmod <- vglm(party ~ income, family=cumulative(parallel=TRUE), rnes96)
deviance(pmod) - deviance(nmod)
[1] 7.8241
1-pchisq(7.82, 1)
[1] 0.0051671
```

There is a difference of only one parameter between the two models. We see that the simplification to the proportional odds model is not justified here because the *p*-value is small. However, we should bear in mind that we have a fairly large sample size of almost 1000 cases so even relatively small differences are prone to be significant. So our previous use of the proportional odds model was not completely unreasonable.

Predictions can be made in a straightforward manner. Interpreting the model is somewhat more difficult due to the varying slopes. Nonparallel lines will cross somewhere. For this model that point is:

```
(1.148-0.329)/(0.0105-0.0162)
[1] -143.68
```

So at an income below  $-\$143,680$ , the predicted probability of being Democrat would exceed the probability of being Democrat or Independent. Clearly this is impossible and the model predictions are ridiculous. This is not a practical problem in this example as  $-\$143,680$  is far outside the range of incomes we would consider. But in other situations, particularly where more predictors are used, this could become a serious issue. This is one good reason to prefer the proportional model if it is viable.

**Ordered Probit Model:** If the latent variable  $Z_i$  has a standard normal distribution, then:

$$\Phi^{-1}(\gamma_j(x_i)) = \theta_j - \beta_j^T x_i \quad j = 1, \dots, J - 1$$

Applying this model to the nes96 data, we find:

```
opmod <- polr(party ~ income, method="probit", rnes96)
summary(opmod)
```

Coefficients:

	Value	Std. Error	t value
nincome	0.008182	0.0012078	6.7745

Intercepts:

	Value	Std. Error	t value
Democrat Independent	0.128	0.069	1.851
Independent Republican	0.798	0.072	11.040

Residual Deviance: 1994.89

AIC: 2000.89

The deviance is similar to the logit version of this model, but the coefficients appear to be different. However, if we compute the same predictions:

```
dems <- pnorm(0.128-inclevels*0.008182)
demind <- pnorm(0.798-inclevels*0.008182)
data.frame(Democrat=dems, Independent=demind-dems, Republican=1-demind,
           ↪ row.names=inclevels)
Democrat Independent Republican
0      0.55093     0.23664    0.21244
20     0.48578     0.25129    0.26292
40     0.42102     0.26006    0.31892
60     0.35833     0.26228    0.37939
80     0.29925     0.25778    0.44297
100    0.24503     0.24691    0.50806
```

We see that the predicted values are very similar to those seen for the logit. If the coefficients are appropriately rescaled, they are also very similar.

**Proportional Hazards Model:** A concept of *hazard* was developed in insurance applications. When issuing a life insurance policy, the insurer is interested in the probability that the person will die during the term of the policy given that they are alive now. This is not the same as the unconditional probability of death during the same time period. In other words, for example, we want to know the chance that a 55-year-old man will die in the next year, given that he is alive and aged 55. The unconditional probability that a man will die aged 55 is not particularly useful for the purposes of insurance.

Suppose we use the complementary log-log in place of the logit above, that is:

$$\log(-\log(1 - \gamma_j(x_i))) = \theta_j + \beta^T x_i$$

Then the *hazard* of category  $j$  is the probability of falling in category  $j$  given that your category is greater than  $j$ :

$$\text{Hazard}(j) = P(Y_i = j | Y_i \geq j) = \frac{P(Y_i = j)}{P(Y_i \geq j)} = \frac{p_j}{1 - \gamma_{i,j-1}} = \frac{\gamma_{ij} - \gamma_{i,j-1}}{1 - \gamma_{i,j-1}}$$

The corresponding latent variable distribution is the extreme value:

$$F(x) = 1 - \exp(-\exp(x))$$

The extreme value distribution is not symmetric like the logistic and normal and so there seems little justification for applying it to the nes96 data, but the command is:

```
polr(party ~ income, method="cloglog")
```

**Assigning Scores:** When the ordinal response has a larger number of categories, it may be sensible to assign scores to each level and then model these scores as the response in a standard linear model. Suppose we reconsider the nes96 data but use the income as the response variable. In the original form of the data this is an ordinal variable:

```
levels(nes96$income)
```

```
[1] "$3Kminus"   "$3K-$5K"    "$5K-$7K"    "$7K-$9K"    "$9K-$10K"  
[6] "$10K-$11K"  "$11K-$12K"  "$12K-$13K"  "$13K-$14K"  "$14K-$15K"  
[11] "$15K-$17K"  "$17K-$20K"  "$20K-$22K"  "$22K-$25K"  "$25K-$30K"  
[16] "$30K-$35K"  "$35K-$40K"  "$40K-$45K"  "$45K-$50K"  "$50K-$60K"  
[21] "$60K-$75K"  "$75K-$90K"  "$90K-$105K" "$105Kplus"
```

This is a special case of ordinal variable called an *interval* variable because a continuous variable has been discretized into intervals. For such variables, it is natural to assign scores equal to the midpoints of the intervals:

```
inca <- c(1.5, 4, 6, 8, 9.5, 10.5, 11.5, 12.5, 13.5, 14.5, 16, 18.5, 21, 23.5,  
        ↪ 27.5, 32.5, 37.5, 42.5, 47.5, 55, 67.5, 82.5, 97.5, 115)  
nes96$sincome <- inca[unclass(nes96$income)]
```

The unclass function converts a categorical variable into an integer variable with the levels numbered in the order in which they appear. We have to choose a score for the first and last levels. It is natural to suppose that zero is the lower bound giving a choice of 1.5 for the first interval. We have chosen 115 for the last interval but other choices may be reasonable. For noninterval valued ordinal variables the choice of scores is much more subjective and should reflect an understanding of the relative difference between adjacent levels.

Now we can fit a linear model using the same variables considered previously (but using the original form of the party identification variable):

```
lmod <- lm(sincome ~ age + educ + PID, nes96)  
summary(lmod)
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	45.8380	3.3533	13.67	< 2e-16
age	-0.0496	0.0582	-0.85	0.394
educ.L	39.2854	5.0850	7.73	2.9e-14
educ.Q	2.8554	4.7508	0.60	0.548
educ.C	3.4486	4.0083	0.86	0.390
educ^4	2.5299	3.2825	0.77	0.441
educ^5	-4.4705	2.7992	-1.60	0.111
educ^6	-0.5773	2.3419	-0.25	0.805
PID.L	12.7103	2.1634	5.88	5.9e-09
PID.Q	-6.3140	2.8613	-2.21	0.028
PID.C	1.0306	2.3740	0.43	0.664
PID^4	6.2413	2.9296	2.13	0.033
PID^5	-2.5174	2.5585	-0.98	0.325
PID^6	-0.7923	3.6532	-0.22	0.828

n = 944, p = 14, Residual SE = 27.899, R-Squared = 0.2

Both the education and party identification variable are ordinal and so are represented using polynomial contrasts. These can be hard to interpret when we have several levels. One option is to recode these predictors as nominal factors which is more straightforward to understand but loses some information. In this case, we notice that the linear term for both education and party identification is the only strongly

significant term. This provides the hint that we might simplify this model by also assigning scores to these two predictors. We also notice that age appears not to be needed. We can then compare this simplified model to the original model:

```
lmod2 <- lm(sincome ~ unclass(educ) + unclass(PID), nes96)
anova(lmod2, lmod)
```

Analysis of Variance Table

Model 1: sincome ~ unclass(educ) + unclass(PID)

Model 2: sincome ~ age + educ + PID

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	941	737108				
2	930	723881	11	13227	1.54	0.11

We see that the simplification is justifiable. We need to know the ordering of the scores before interpreting the model:

```
levels(nes96$educ)
[1] "MS"      "HSdrop"  "HS"       "Coll"    "CCdeg"   "BAdeg"   "MAdeg"
levels(nes96$PID)
[1] "strDem"  "weakDem" "indDem"   "indind"  "indRep"  "weakRep"
[7] "strRep"
```

Now we consider the model summary:

```
summary(lmod2)
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.510     3.038    1.16    0.25
unclass(educ) 7.359     0.573   12.84 < 2e-16
unclass(PID)  2.463     0.403    6.11   1.4e-09
```

$n = 944$ ,  $p = 3$ , Residual SE = 27.988, R-Squared = 0.19

We see that each one-step increase in the level of education is associated with a \$7359 increase in income. We also see that each step along our political scale from strongly Democrat to strongly Republican is associated with \$2463 increase income. This association should only be viewed predictively and not causally. It would certainly be surprising if changing your political opinions caused you to earn more or less. Furthermore, more intelligent and/or privileged people would tend to earn more regardless of their level of formal education.

**Further Reading:** For more on the analysis of ordered categorical data see the books by Agresti (1984), Clogg and Shihadeh (1994), Powers and Xie (2000) and Simonoff (2003).

## Exercises

- The hsb data was collected as a subset of the High School and Beyond study conducted by the National Education Longitudinal Studies program of the National Center for Education Statistics. The variables are gender; race; socioeconomic status (SES); school type; chosen high school program type; scores on reading, writing, math, science, and social studies. We want to determine which factors are related to the choice of the type of program — academic, vocational or general — that the students pursue in high school. The response is multinomial with three levels.

- Make a table showing the proportion of males and females choosing the three

- different programs. Comment on the difference. Repeat this comparison but for SES rather than gender.
- (b) Construct a plot like the right panel of Figure 7.1 that shows the relationship between program choice and reading score. Comment on the plot. Repeat for math in place of reading.
  - (c) Compute the correlation matrix for the five subject scores.
  - (d) Fit a multinomial response model for the program choice and examine the fitted coefficients. Of the five subjects, one gives unexpected coefficients. Identify this subject and suggest an explanation for this behavior.
  - (e) Construct a derived variable that is the sum of the five subject scores. Fit a multinomial model as before except with this one sum variable in place of the five subjects separately. Compare the two models to decide which should be preferred.
  - (f) Use a stepwise method to reduce the model. Which variables are in your selected model?
  - (g) Construct a plot of predicted probabilities from your selected model where the math score varies over the observed range. Other predictors should be set at the most common level or mean value as appropriate. Your plot should be similar to Figure 7.2. Comment on the relationship.
  - (h) Compute a table of predicted probabilities cross-classified by SES and school type. Fix the other predictors at their mean values. Comment on how SES and school type are related to the response.
  - (i) Compute the predicted outcome for the student with ID 99. What did this student actually choose?
  - (j) Construct a table of the most likely predicted outcomes and observed outcomes. In what proportion of cases is the predicted and observed outcome the same?
2. Data were collected from 39 students in a University of Chicago MBA class and may be found in the dataset happy.
- (a) Make plots of the relationship between the response, happiness and each of the four predictors. Transform sex and love to appropriate factors.
  - (b) Fit a proportional odds model with the main effects of each of the four predictors. Extract the estimated values of  $\theta$  using the `zeta` component of the fit. Make an index plot of these and comment on any deviation from linearity and what this may signify.
  - (c) Use stepwise AIC to select a reduced model. Which variables are eliminated. Make a test to compare this reduced model to the original model.
  - (d) Compare a model where love is treated as an ordered factor with one where it is treated as a numerical predictor. Which model is preferred?
  - (e) Examine the regression coefficients of your selected model and provide an interpretation of their meaning.

- (f) Varying money across its range and holding love and work fixed at their median values, predict the most likely happiness level. Repeat the process varying only love and then work. What do we learn about the effect of the three predictors on the response?
- (g) Predict the happiness distribution for subject whose parents earn \$30,000 a year, who is lonely, not sexually active and has no job.
3. A student newspaper conducted a survey of student opinions about the Vietnam War in May 1967. Responses were classified by sex, year in the program and one of four opinions. The survey was voluntary. The data may be found in the dataset `uncviet`.
- Compute the proportion favoring each policy within each year by sex combination. Plot these proportions as year varies with a different line type for each policy. Plot men and women on separate panels. Your plot should look similar to Figure 7.1.
  - Fit a proportional odds model with policy as the response with sex and year as predictors (include their interaction). Use the `weight` argument to set the number of respondents for each case. Why is it sensible to include an interaction term?
  - Now fit a model with main effects only, excluding the interaction. Compare this model to the previous one. Which is preferred?
  - Compute the predicted proportion for each case and plot in the same format as (a). Use the `type="probs"` argument to predict. You will need to select only one of the four probabilities for each case. Comment on the plot and compare it to (a).
  - Compute the raw residuals as the difference between the predicted and observed proportions. Use the same format to plot these residuals as the predicted proportions. Comment on the plot and suggest how the definition of the residual might be improved.
  - Examine the regression summary output to find the coefficients that are relevant to the opinions of women. What does the significance (or lack thereof) say about how the opinions of women vary across the year groups?
4. The `pneumo` data gives the number of coal miners classified by radiological examination into one of three categories of pneumoconiosis and by the number of years spent working at the coal face divided into eight categories.
- Make a plot showing how the proportion on miners in the three categories at each year point varies over time. Comment on the relationship.
  - Treating the pneumoconiosis status as response variable as nominal, build a model for predicting the frequency of the three outcomes in terms of length of service. What does the model say about the similarity of the proportions falling into the mild and severe categories?
  - Would it be better to use `log(year)` as the predictor?
  - Produce a plot of the predicted probabilities in the same format as (a).

- (e) Fit a proportional odds model to the data. Take care to order response correctly. You will need to specify the number falling in each case using the `weight` argument. What is the estimated value of  $\theta_1$  and how should it be interpreted?
  - (f) Repeat the analysis with the pneumoconiosis status being treated as ordinal.
  - (g) Extract the predicted probabilities from the model and plot in the same format as (a). Compare to the predictions from the nominal model.
  - (h) Fit a hierarchical model to the status response. First fit a binomial response—normal and not normal. Interpret the effect of year on the odds of getting the lung disease.
  - (i) Now fit a binomial model for mild vs. severe lung disease. Is the year effect significant. What is the probability of a mild disease within the diseased subgroup?
  - (j) Compute the predicted probabilities of the three categories by combining the two binomial model predictions. Plot in the same format as (a) and comment.
5. The debt data arise from a large postal survey on the psychology of debt. The frequency of credit card use is a three-level factor ranging from never, through occasionally to regularly.
- (a) Declare the response as an ordered factor and make a plot showing the relationship to `prodebt`. Comment on the plot. Use a table or plot to display the relationship between the response and the income group.
  - (b) Fit a proportional odds model for credit card use with all the other variables as predictors. What are the two most significant predictors (largest  $t$ -values) and what is their qualitative effect on the response? What is the least significant predictor?
  - (c) Fit a proportional odds model using only the least significant predictor from the previous model. What is the significance of this predictor in this small model? Are the conclusions regarding this predictor contradictory for the two models?
  - (d) Use stepwise AIC to select a smaller model than the full set of predictors. You will need to handle the missing values carefully. Report on the qualitative effect of the predictors in your chosen model. Can we conclude that the predictors that were dropped from the model have no relation to the response?
  - (e) Compute the median values of the predictors in your selected model. At these median values, contrast the predicted outcome probabilities for both smokers and nonsmokers.
  - (f) Fit a proportional hazards model to the same set of predictors and recompute the two sets of probabilities from the previous question. Does it make a difference to use this type of model?

# Generalized Linear Models

---

In previous chapters, we have seen how to model a binomial or Poisson response. Multinomial response models can often be recast as Poisson responses and the standard linear model with a normal (Gaussian) response is already familiar. Although these models each have their distinctive characteristics, we see some common features in all of them. We can abstract these to form the *generalized linear model* (GLM). By developing a theory and constructing general methods for GLMs, we are able to tackle a wider range of data with different types of response variables. GLMs were introduced by Nelder and Wedderburn (1972), while McCullagh and Nelder (1989) provide a book-length treatment.

## 8.1 GLM Definition

A GLM is defined by specifying two components. The response should be a member of the exponential family distribution and the link function describes how the mean of the response and a linear combination of the predictors are related.

**Exponential Family:** In a GLM the distribution of  $Y$  is from the exponential family of distributions which take the general form:

$$f(y|\theta, \phi) = \exp \left[ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right]$$

The  $\theta$  is called the *canonical parameter* and represents the location, while  $\phi$  is called the *dispersion parameter* and represents the scale. We may define various members of the family by specifying the functions  $a$ ,  $b$  and  $c$ . The most commonly used examples are:

1. Normal or Gaussian:

$$\begin{aligned} f(y|\theta, \phi) &= \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{(y-\mu)^2}{2\sigma^2} \right] \\ &= \exp \left[ \frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2} \left( \frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right) \right] \end{aligned}$$

So we can write  $\theta = \mu$ ,  $\phi = \sigma^2$ ,  $a(\phi) = \phi$ ,  $b(\theta) = \theta^2/2$  and  $c(y, \phi) = -(y^2/\phi + \log(2\pi\phi))/2$ .

2. Poisson:

$$f(y|\theta, \phi) = e^{-\mu} \mu^y / y!$$

$$= \exp(y \log \mu - \mu - \log y!)$$

So we can write  $\theta = \log(\mu)$ ,  $\phi \equiv 1$ ,  $a(\phi) = 1$ ,  $b(\theta) = \exp(\theta)$  and  $c(y, \phi) = -\log y!$ .

3. Binomial:

$$\begin{aligned} f(y|\theta, \phi) &= \binom{n}{y} \mu^y (1-\mu)^{n-y} \\ &= \exp \left( y \log \mu + (n-y) \log(1-\mu) + \log \binom{n}{y} \right) \\ &= \exp \left( y \log \frac{\mu}{1-\mu} + n \log(1-\mu) + \log \binom{n}{y} \right) \end{aligned}$$

So we see that  $\theta = \log \frac{\mu}{1-\mu}$ ,  $b(\theta) = -n \log(1-\mu) = n \log(1 + \exp \theta)$  and  $c(y, \phi) = \log \binom{n}{y}$ .

The gamma and inverse Gaussian are other less used members of the exponential family that are covered in Chapter 9. Notice that in the normal density, the  $\phi$  parameter is free (as it is also for the gamma density), while for the Poisson and binomial it is fixed at one. This is because the Poisson and binomial are one-parameter families, while the normal and gamma have two parameters. In fact, some authors reserve the term *exponential family* distribution for cases where  $\phi$  is not used, while using the term *exponential dispersion family* for cases where it is. This has important consequences for the analysis.

Some other densities, such as the negative binomial and the Weibull distribution, are not members of the exponential family, but they are sufficiently close that the GLM can be fit with some modifications. It is also possible to fit distributions that are not in the exponential family using the GLM-style approach, but there are some additional complications.

We now derive the mean and variance of the exponential family distributions. The log-likelihood for a single  $y$  is given by:

$$l(\theta) = (y\theta - b(\theta))/a(\phi) + c(y, \phi)$$

Taking derivatives with respect to  $\theta$  gives:

$$l'(\theta) = (y - b'(\theta))/a(\phi)$$

Taking the expectation over  $y$  gives:

$$El'(\theta) = (EY - b'(\theta))/a(\phi)$$

From general likelihood theory, we know that  $El'(\theta) = 0$  at the true value of  $\theta$  which implies that:

$$EY = \mu = b'(\theta)$$

Now taking second derivatives:

$$l''(\theta) = -b''(\theta)/a(\phi)$$

General likelihood theory tells us that  $E l''(\theta) = -E[(l'(\theta))^2]$ . We evaluate at the true value of  $\theta$  to obtain

$$b''(\theta)/a(\phi) = E[(Y - b'(\theta))^2]/a^2(\phi)$$

which gives us

$$\text{var } Y = b''(\theta)a(\phi)$$

The mean is a function of  $\theta$  only, while the variance is a product of functions of the location and the scale.  $V(\mu) = b''(\theta)/w$  is called the *variance function* and describes how the variance relates to the mean using the known relationship between  $\theta$  and  $\mu$ .

In the Gaussian case,  $b''(\theta) = 1$  and so the variance is independent of the mean. For other distributions, this is not true, making the Gaussian case exceptional. We can introduce weights by setting:

$$a(\phi) = \phi/w$$

where  $w$  is a known weight that varies between observations. We could allow for general choices of  $a$  but the simplified form, including weights if needed, covers all the situations we currently need.

**Link Function:** Suppose we may express the effect of the predictors on the response through a *linear predictor*:

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = x^T \beta$$

The link function,  $g$ , describes how the mean response,  $EY = \mu$ , is linked to the covariates through the linear predictor:

$$\eta = g(\mu)$$

In principle, any monotone continuous and differentiable function will do, but there are some convenient and common choices for the standard GLMs.

In the Gaussian linear model, the identity link,  $\eta = \mu$ , is the obvious selection, but another choice would give  $y = g^{-1}(x^T \beta) + \varepsilon$ . This does not correspond directly to a transform on the response:  $g(y) = x^T \beta + \varepsilon$  as, for example, in a Box–Cox type transformation. In a GLM, the link function is assumed known whereas in a *single index model*,  $g$  is estimated.

For the Poisson GLM, the mean  $\mu$  must be positive so  $\eta = \mu$  will not work conveniently since  $\eta$  can be negative. The standard choice is  $\mu = e^\eta$ , so that  $\eta = \log \mu$  which ensures  $\mu > 0$ . This log link means that additive effects of  $x$  lead to multiplicative effects on  $\mu$ .

For the binomial GLM, let  $p$  be the probability of success and let this be our  $\mu$  if we define the response as the proportion rather than the count. This requires that  $0 \leq p \leq 1$ . There are several commonly used ways to ensure this: the logistic, probit and complementary log-log links. These are discussed in detail in Chapter 2.

The *canonical link* has  $g$  such that  $\eta = g(\mu) = \theta$ , the canonical parameter of the exponential family distribution. This means that  $g(b'(\theta)) = \theta$ . The canonical links

Family	Link	Variance Function
Normal	$\eta = \mu$	1
Poisson	$\eta = \log \mu$	$\mu$
Binomial	$\eta = \log(\mu/(1-\mu))$	$\mu(1-\mu)$
Gamma	$\eta = \mu^{-1}$	$\mu^2$
Inverse Gaussian	$\eta = \mu^{-2}$	$\mu^3$

Table 8.1 *Canonical links for GLMs.*

for the common GLMs are shown in Table 8.1. If a canonical link is used,  $X^T Y$  is *sufficient* for  $\beta$ . The canonical link is mathematically and computationally convenient and is often the natural choice of link. However, one is not required to use the canonical link and sometimes context may compel another choice.

## 8.2 Fitting a GLM

The parameters,  $\beta$ , of a GLM can be estimated using maximum likelihood. The log-likelihood for a single observation, where  $a_i(\phi) = \phi/w_i$ , is:

$$l(\beta; y_i) = w_i \left[ \frac{y_i \theta_i - b(\theta_i)}{\phi} \right] + c(y_i, \phi)$$

So for independent observations, the log-likelihood will be  $\sum_i l(\beta; y_i)$ . We want to maximize this over  $\beta$  and start by taking partial derivatives with respect to the components  $\beta_j$ :

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_i w_i \left( y_i \frac{\partial \theta_i}{\partial \beta_j} - b'(\theta_i) \frac{\partial \theta_i}{\partial \beta_j} \right)$$

The chain rule gives us:

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j}$$

Using the fact that  $\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i)$ , we have

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_i \frac{(y_i - b'(\theta_i))}{b''(\theta_i)/w_i} \frac{\partial \mu_i}{\partial \beta_j}$$

We now substitute in the known relations for the mean and variance functions. We set the partial derivatives to zero to obtain the maximum likelihood estimates as the solution to:

$$\sum_i \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0 \quad \forall j$$

Now suppose we knew the variance function  $V(\mu)$ , then we would obtain this same set of equations if we had started out to minimize the weighted least squares criterion:

$$\sum_i \frac{(y_i - \mu_i)^2}{V(\mu_i)}$$

This suggests how we might develop an algorithm for deriving the estimates using iteratively reweighted least squares (IRWLS). Full details can be found in McCullagh and Nelder (1989) or Wood (2006).

The procedure requires an initial guess of  $\hat{\mu}^0$  from which  $\hat{\eta}^0$  where the superscript indicates the iteration number.

1. Form the “adjusted dependent variable”  $z^i = \hat{\eta}^i + (y - \hat{\mu}^i) \frac{d\eta}{d\mu}|_{\hat{\eta}^i}$ .
2. Form the weights  $1/w^0 = \left(\frac{d\eta}{d\mu}\right)^2|_{\hat{\eta}^0} V(\hat{\mu}^0)$ .
3. Reestimate to get  $\hat{\beta}^{i+1}$  and hence  $\hat{\eta}^{i+1}$ .

Repeat these steps until convergence.

Notice that the fitting procedure uses only  $\eta = g(\mu)$  and  $V(\mu)$ , but requires no further knowledge of the distribution of  $y$ . This point will be important later when considering the quasi-likelihood method in Section 9.4. Estimates of variance may be obtained using standard likelihood theory from:

$$\text{var}(\hat{\beta}) = (X^T W X)^{-1} \hat{\phi}$$

where  $W$  is a diagonal matrix formed from the weights  $w$ . This is comparable to the form used in weighted least squares with the exception that the weights are now a function of the response for a GLM.

Let’s implement the procedure explicitly to understand how the fitting algorithm works. We use the Bliss data from Section 4.2 to illustrate this. Here is the fit we are trying to match:

```
data(bliss, package="faraway")
mod1 <- glm(cbind(dead, alive) ~ conc, family=binomial, bliss)
summary(mod1)$coef
Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.3238 0.41789 -5.5608 2.6854e-08
conc 1.1619 0.18142 6.4046 1.5077e-10
```

For a binomial response, we have:

$$\eta = \log \frac{\mu}{1-\mu} \quad \frac{d\eta}{d\mu} = \frac{1}{\mu(1-\mu)} \quad V(\mu) = \mu(1-\mu)/n \quad w = n\mu(1-\mu)$$

where the variance is computed with the understanding that  $y$  is the proportion, not the count. We use  $y$  for our initial guess for  $\hat{\mu}$  which works here because none of the observed proportions are zero or one:

```
y <- bliss$dead/30; mu <- y
library(faraway)
eta <- logit(mu)
z <- eta + (y-mu) / (mu*(1-mu))
w <- 30*mu*(1-mu)
lmod <- lm(z ~ conc, weights=w, bliss)
coef(lmod)
(Intercept) conc
-2.3025 1.1536
```

It is interesting how close these initial estimates are to the converged values given above. This is not uncommon. Even so, to get a more precise result, iteration is necessary. We do five iterations here:

```

for(i in 1:5){
  eta <- lmod$fit
  mu <- ilogit(eta)
  z <- eta + (y-mu)/(mu*(1-mu))
  w <- 30*mu*(1-mu)
  lmod <- lm(z ~ bliss$conc, weights=w)
  cat(i,coef(lmod),"\n")
}
1 -2.3237 1.1618
2 -2.3238 1.1619
3 -2.3238 1.1619
4 -2.3238 1.1619
5 -2.3238 1.1619

```

We can see that convergence is fast in this case. In most cases, the convergence is rapid. If there is a failure to converge, this is often a sign of some problem with the model specification or an unusual feature of the data. An example of such a problem with the estimation may be seen in Section 2.7. A look at the final (weighted) linear model reveals that:

```

summary(lmod)
Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.3238     0.1462   -15.9  0.00054
bliss$conc    1.1619     0.0635    18.3  0.00036

```

$n = 5, p = 2, \text{Residual SE} = 0.350, R\text{-Squared} = 0.99$

The standard errors are not correct and can be computed (rather inefficiently) as follows:

```

xm <- model.matrix(lmod)
wm <- diag(w)
sqrt(diag(solve(t(xm) %*% wm %*% xm)))
[1] 0.41787 0.18141

```

Now  $\hat{\text{var}}(\hat{\beta}) = (X^T W X)^{-1}$  because  $\phi = 1$  for the binomial model but in the Gaussian linear model  $\hat{\text{var}}(\hat{\beta}) = (X^T W X)^{-1} \hat{\sigma}^2$ . To get the correct standard errors from the `lm` fit, we need to scale out the  $\hat{\sigma}$  as follows:

```

summary(lmod)$coef[,2]/summary(lmod)$sigma
(Intercept)      conc
0.41789       0.18142

```

These calculations are shown for illustration purposes only and are done more efficiently and reliably by the `glm` function.

### 8.3 Hypothesis Tests

When considering the choice of model for some data, we must define the range of possibilities. The *null* model is the smallest model we will entertain while the *full* or *saturated* model is the most complex.

The null model represents the situation where there is no relation between the predictors and the response. Usually this means we fit a common mean  $\mu$  for all  $y$ , that is, one parameter only. For the Gaussian GLM, this is the model  $y = \mu + \varepsilon$ . For some contingency table models, there will be additional parameters that represent row or column totals or other such constraints. In these cases, the null model will have more than one parameter.

In the saturated model, the response is explained in the best achievable way. There are usually several ways in which this can be done by adding sufficiently many parameters. One way that will always be available is to define a factor with one level for each unique combination of the predictors appearing in the data. In many datasets, there will be only one observation for each unique combination. In such cases there will be  $n$  parameters for  $n$  data points and  $\hat{\mu} = y$ . In other examples, where replication occurs, we will need fewer than  $n$  parameters.

A statistical model describes how we partition the data into systematic structure and random variation. The null model represents one extreme where the data is represented entirely as random variation, while the saturated or full model represents the data as being entirely systematic, excepting any variation that no model could explain.

We would like a measure of how well our model fits. We can do this by comparing our model to the full model. Hence we consider the difference between the log-likelihood for the full model,  $l(y, \phi|y)$ , and that for the model under consideration,  $l(\hat{\mu}, \phi|y)$ , expressed as a log likelihood ratio statistic:

$$2(l(y, \phi|y) - l(\hat{\mu}, \phi|y))$$

The factor of two is used for convenience in the subsequent distributional results. Provided that the observations are independent and for an exponential family distribution, when  $a_i(\phi) = \phi/w_i$ , this simplifies to:

$$\sum_i 2w_i(y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i))/\phi$$

where  $\tilde{\theta}$  are the estimates under the full (saturated) model and  $\hat{\theta}$  are the estimates under the model of interest. We write this quantity as  $D(y, \hat{\mu})/\phi$  where  $D(y, \hat{\mu})$  is called the *deviance* and  $D(y, \hat{\mu})/\phi$  is called the *scaled deviance*. Deviances for the common GLMs are shown in Table 8.2.

GLM	Deviance
Gaussian	$\sum_i (y_i - \hat{\mu}_i)^2$
Poisson	$2\sum_i [y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)]$
Binomial	$2\sum_i [y_i \log(y_i/\hat{\mu}_i) + (m - y_i) \log((m - y_i)/(m - \hat{\mu}_i))]$
Gamma	$2\sum_i [-\log(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i]$
Inverse Gaussian	$\sum_i (y_i - \hat{\mu}_i)^2 / (\hat{\mu}_i^2 y_i)$

Table 8.2 For the binomial  $y_i \sim B(m, p_i)$  and  $\mu_i = mp_i$ , that is,  $\mu$  is the count and not proportion in this formula. For the Poisson, the deviance is known as the G-statistic. The second term  $\sum_i (y_i - \hat{\mu}_i)$  is usually zero if an intercept term is used in the model.

An alternative measure of fit that is sometimes used in place of the deviance is the Pearson's  $X^2$  statistic:

$$X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

where  $V(\hat{\mu}) = \text{var } (\hat{\mu})$ .

There are two main types of hypothesis test. The goodness of fit test asks whether the current model fits the data. The other type of test compares two nested models where the smaller model represents a linear restriction on the parameters of the larger model. The goodness of fit test can be viewed as a model comparison test if we identify the smaller model with the model of interest and the larger model with the full or saturated model.

For the goodness of fit test, we use the fact that, under certain conditions, provided the model is correct, the scaled Deviance and the Pearson's  $X^2$  statistic are both asymptotically  $\chi^2$  with degrees of freedom equal to the number of observations minus the number of identifiable parameters. For GLMs with a dispersion parameter, such as the Gaussian, we usually do not know the value of the  $\phi$ , and so this test usually cannot be used. For the binomial and the Poisson,  $\phi = 1$ , and so the test is practical. However, the accuracy of the asymptotic approximation is dubious for smaller datasets. For a binary response, the approximation is worthless as explained in Section 2.3.

For comparing a larger model,  $\Omega$ , to a smaller nested model,  $\omega$ , the difference in the scaled deviances,  $D_\omega - D_\Omega$ , is asymptotically  $\chi^2$  with degrees of freedom equal to the difference in the number of identifiable parameters in the two models. For the Gaussian model and other models where the dispersion  $\phi$  is usually not known, this test cannot be directly used. However, if we insert an estimate of  $\phi$  we may compute an  $F$ -statistic of the form:

$$\frac{(D_\omega - D_\Omega) / (df_\omega - df_\Omega)}{\hat{\phi}}$$

where  $\hat{\phi} = X^2 / (n - p)$  is a good estimate of the dispersion. The degrees of freedom  $df$  are typically the number of observations minus the number of parameters. For the Gaussian model, a sensible estimate is  $\hat{\phi} = RSS_\Omega / df_\Omega$ , and the resulting  $F$ -statistic has an exact  $F$  distribution for the null. For other GLMs with free dispersion parameters, the statistic is only approximately  $F$  distributed.

For every GLM except the Gaussian, an approximate null distribution must be used whose accuracy may be in doubt particularly for smaller samples. However, the approximation is better when comparing models than for the goodness of fit statistic.

Let's consider the possible tests on the Bliss insect data:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.324	0.418	-5.56	2.7e-08
conc	1.162	0.181	6.40	1.5e-10

```
n = 5 p = 2
Deviance = 0.379 Null Deviance = 64.763 (Difference = 64.385)
```

We are able to make a goodness of fit test by examining the size of the residual deviance compared to its degrees of freedom:

```
1-pchisq(deviance(mod1), df.residual(mod1))
[1] 0.9446
```

where we see the  $p$ -value is large indicating no evidence of a lack of fit. As with lack of fit tests for Gaussian linear models, this outcome does not mean that this model is correct or that no better models exist. We can also see that the null model would be

inadequate for the data since the null deviance of 64.7 is very large for four degrees of freedom.

We can test for the significance of the linear concentration term by comparing the current model to the null model:

```
anova(mod1, test="Chi")
Analysis of Deviance Table
Model: binomial, link: logit

Terms added sequentially (first to last)
  Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL             4      64.8
conc  1       64.4      3      0.4     1e-15
```

We see that the concentration term is clearly significant. We can also fit and test a more complex model:

```
mod12 <- glm(cbind(dead,alive) ~ conc+I(conc^2), family=binomial, bliss)
anova(mod1, mod12, test="Chi")

  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1          3      0.379
2          2      0.195  1      0.183     0.669
```

We can see that there is no need for a quadratic term in the model. The same information could be extracted with:

```
anova(mod12, test="Chi")
```

We may also take a Wald test approach by taking the standard error of the parameter estimates to construct a  $z$ -statistic of the form  $\hat{\beta}/se(\hat{\beta})$ . This has an asymptotically normal null distribution. For the Bliss data, for the concentration term, we have  $z = 1.162/0.181 = 6.40$ . Thus the (approximate)  $p$ -value for the Wald test of the concentration parameter being equal to zero is  $1.5e^{-10}$  and thus we clearly reject the null here. Remember that this is again only an approximate test except in the special case of the Gaussian GLM where the  $z$ -statistic is the  $t$ -statistic and has an exact  $t$ -distribution. The difference of deviances test is preferred to the Wald test, due, in part, to the problem noted by Hauck and Donner (1977).

## 8.4 GLM Diagnostics

As with standard linear models, it is important to check the adequacy of the assumptions that support the GLM. The diagnostic methods for GLMs mirror those used for Gaussian linear models. However, some adaptations are necessary and, depending on the type of GLM, not all diagnostic methods will be applicable.

**Residuals:** Residuals represent the difference between the data and the model and are essential to explore the adequacy of the model. In the Gaussian case, the residuals are  $\hat{\epsilon} = y - \hat{\mu}$ . These are called response residuals for GLMs, but since the variance of the response is not constant for most GLMs, some modification is necessary. We would like residuals for GLMs to be defined such that they can be used in a similar way as in the Gaussian linear model.

The *Pearson residual* is comparable to the standardized residuals used for linear models and is defined as:

$$r_P = \frac{y - \hat{\mu}}{\sqrt{V(\hat{\mu})}}$$

where  $V(\mu) \equiv b''(\theta)$ . These are just a rescaling of  $y - \hat{\mu}$ . Notice that  $\sum r_P^2 = X^2$  and hence the name. Pearson residuals can be skewed for nonnormal responses.

The *deviance residuals* are defined by analogy to Pearson residuals. The Pearson residual was  $r_P$  such that  $\sum r_P^2 = X^2$ , so we set the deviance residual as  $r_D$  such that  $\sum r_D^2 = \text{Deviance} = \sum d_i$ . Thus:

$$r_D = \text{sign}(y - \hat{\mu}) \sqrt{d_i}$$

For example, in the Poisson:

$$r_D = \text{sign}(y - \hat{\mu}) [2(y \log y / \hat{\mu} - y + \hat{\mu})]^{1/2}$$

Let's examine the types of residuals available to us using the Bliss data. We can obtain the deviance residuals as:

```
residuals(modl)
[1] -0.451015 0.359696 0.000000 0.064302 -0.204493
```

These are the default choice of residuals. The Pearson residuals are:

```
residuals(modl, "pearson")
 1      2      3      4      5 
-0.432523 0.364373 0.000000 0.064147 -0.208107
```

which are just slightly different from the deviance residuals. The response residuals are:

```
residuals(modl, "response")
 1      2      3      4      5 
-0.0225051 0.0283435 0.0000000 0.0049898 -0.0108282
```

which is just the response minus the fitted value:

```
bliss$dead/30 - fitted(modl)
 1      2      3      4      5 
-0.0225051 0.0283435 0.0000000 0.0049898 -0.0108282
```

Finally, the so-called working residuals are:

```
residuals(modl, "working")
 1      2      3      4      5 
-0.277088 0.156141 0.000000 0.027488 -0.133320
```

```
modl$residuals
 1      2      3      4      5 
-0.277088 0.156141 0.000000 0.027488 -0.133320
```

Note that it is important to use the `residuals()` function to get the deviance residuals which are most likely what is needed for diagnostic purposes. Using `$residuals` gives the working residuals which is not usually needed for diagnostics. We can now identify the working residuals as a by-product of the IRWLS fitting procedure from Section 8.2.

```
residuals(lmod)
 1      2      3      4      5 
-2.7709e-01 1.5614e-01 -3.8463e-16 2.7488e-02 -1.3332e-01
```

**Leverage and Influence:** For a linear model,  $\hat{y} = Hy$ , where  $H = X(X^T X)^{-1}X^T$  is the *hat matrix* that projects the data onto the fitted values. The leverages  $h_i$  are given by the diagonal of  $H$  and represent the potential of the point to influence the fit. They are solely a function of  $X$  and whether they are in fact influential will also depend on  $y$ . Leverages are somewhat different for GLMs. The IRWLS algorithm used to fit the GLM uses weights,  $w$ . These weights are just part of the IRWLS

algorithm and are not user assigned. However, these do affect the leverage. We form a matrix  $W = \text{diag}(w)$  and the hat matrix is:

$$H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$$

We extract the diagonal elements of  $H$  to get the leverages  $h_i$ . A large value of  $h_i$  indicates that the fit may be sensitive to the response at case  $i$ . Large leverages typically mean that the predictor values are unusual in some way. One important difference from the linear model case is that the leverages are no longer just a function of  $X$  and now depend on the response through the weights  $W$ . The leverages may be calculated as:

```
influence(mod1)$hat
```

1	2	3	4	5
0.42550	0.41331	0.32238	0.41331	0.42550

As in the linear model case, we might choose to studentize the residuals as follows:

$$r_{SD} = \frac{r_D}{\sqrt{\hat{\phi}(1 - h_i)}}$$

or compute jackknife residuals representing the difference between the observed response for case  $i$  and that predicted from the data with case  $i$  excluded, scaled appropriately. These are expensive to compute exactly and so an approximation due to Williams (1987) can be used:

$$\text{sign}(y - \hat{y}) \sqrt{(1 - h_i)r_{SD}^2 + h_i r_{SP}^2}$$

where  $r_{SP} = r_P / \sqrt{1 - h_i}$ . These may be computed as:

```
rstudent(mod1)
```

1	2	3	4	5
-0.584786	0.472135	0.000000	0.083866	-0.271835

Outliers may be detected by observing particularly large jackknife residuals.

Leverage only measures the potential to affect the fit whereas measures of influence more directly assess the effect of each case on the fit. We can examine the change in the fit from omitting a case by looking at the changes in the coefficients:

```
influence(mod1)$coef
```

	(Intercept)	conc
1	-0.2140015	0.0806635
2	0.1556719	-0.0470873
3	0.0000000	0.0000000
4	-0.0058417	0.0084177
5	0.0492639	-0.0365734

Alternatively, we can examine the Cook statistics:

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T (X^TWX)(\hat{\beta}_{(i)} - \hat{\beta})}{p\hat{\phi}}$$

which may be calculated as:

```
cooks.distance(mod1)
```

1	2	3	4	5
0.1205927	0.0797100	0.0000000	0.0024704	0.0279174

We can see that the biggest change would occur by omitting the first observation. However, since this is a very small dataset with just five observations, we would not contemplate dropping cases. In any event, we see that the change in the coefficients would not qualitatively change the conclusion.

**Model Diagnostics:** We may divide diagnostic methods into two types. Some methods are designed to detect single cases or small groups of cases that do not fit the pattern of the rest of the data. Outlier detection is an example of this. Other methods are designed to check the assumptions of the model. These methods can be subdivided into those that check the structural form of the model, such as the choice and transformation of the predictors, and those that check the stochastic part of the model, such as the nature of the variance about the mean response. Here, we focus on methods for checking the assumptions of the model.

For linear models, the plot of residuals against fitted values is probably the single most valuable graphic. For GLMs, we must decide on the appropriate scale for the fitted values. Usually, it is better to plot the linear predictors  $\hat{\eta}$  rather than the predicted responses  $\hat{\mu}$ . We revisit the model for Galápagos data first presented in Section 5.1. Consider first a plot using  $\hat{\mu}$  presented in the first panel of Figure 8.1:

```
data(gala, package="faraway")
gala <- gala[,-2]
modp <- glm(Species ~ ., family=poisson, gala)
plot(residuals(modp) ~ predict(modp, type="response"),
     xlab=expression(hat(mu)), ylab="Deviance residuals")
```

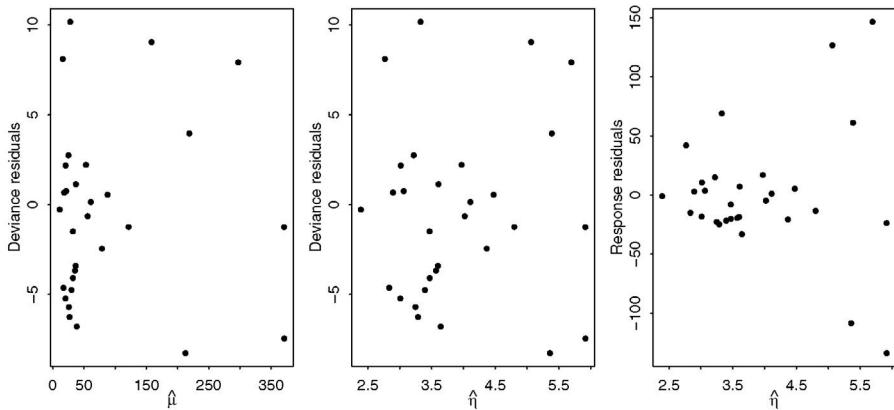


Figure 8.1 *Residual vs. fitted plots for the Galápagos model. The first uses fitted values in the scale of the response while the second uses fitted values in the scale of the linear predictor. The third plot uses response residuals while the first two use deviance residuals.*

There are just a few islands with a large predicted number of species while most predicted response values are small. This makes it difficult to see the relationship between the residuals and the fitted values because most of the points are compressed on the left of the display. Now we try plotting  $\hat{\eta}$ :

```
plot(residuals(modp) ~ predict(modp, type="link"),
     xlab=expression(hat(eta)), ylab="Deviance residuals")
```

Now the points, shown in the second panel of Figure 8.1, are more evenly spaced in the horizontal direction. We are looking for two main features in such a plot. Is there any nonlinear relationship between the predicted values and the residuals? If so, this would be an indication of a lack of fit that might be rectified by a change in the model. For a linear model, we might consider a transformation of the response, but this is usually impractical for a GLM since it would change the assumed distribution of the response. We might also consider a change to the link function, but often this is undesirable since there are a few choices of link function that lead to easily interpretable models. It is best if a change in the choice of predictors or transformations on these predictors can be made since this involves the least disruption to the GLM. For this particular plot, there is no evidence of nonlinearity.

The variance of the residuals with respect to the fitted values should also be inspected. The assumptions of the GLM would require constant variance in the plot and, in this case, this appears to be the case. A violation of this assumption would prompt a change in the model. We might consider a change in the variance function  $V(\mu)$ , but this would involve abandoning the Poisson GLM since this specifies a particular form for the variance function. We would need to use a quasi-likelihood GLM described in Section 9.4. Alternatively, we could employ a different GLM for a count response such as the negative binomial. Finally, we might use weights if we could identify some feature of the data that would suggest a suitable choice.

For all GLMs but the Gaussian, we have a nonconstant variance function. However, by using deviance residuals, we have already scaled out the variance function and so, provided the variance function is correct, we do expect to see constant variance in the plot. If we use response residuals, that is  $y - \hat{\mu}$ , as seen in the third panel of Figure 8.1:

```
plot(residuals(modp, type="response") ~ predict(modp, type="link"),
     xlab=expression(hat(eta)), ylab="Response residuals")
```

We see a pattern of increasing variation consistent with the Poisson.

In some cases, plots of the residuals are not particularly helpful. For a binary response, the residual can only take two possible values for given predicted response. This is the most extreme situation, but similar discreteness can occur for binomial responses with small group sizes and Poisson responses that are small. Plots of residuals in these cases tend to show curved lines of points corresponding to the limited number of observed responses. Such artifacts can obscure the main purpose of the plot. Difficulties arise for binomial data where the covariate classes have very different sizes. Points on plots may represent just a few or a large number of individuals.

Investigating the nature of the relationship between the predictors and the response is another primary objective of diagnostic plots. Even before a model is fit to the data, we might simply plot the response against the predictors. For the Galápagos data, consider a plot of the number of species against the area of the island shown in the first panel of Figure 8.2:

```
plot(Species ~ Area, gala)
```

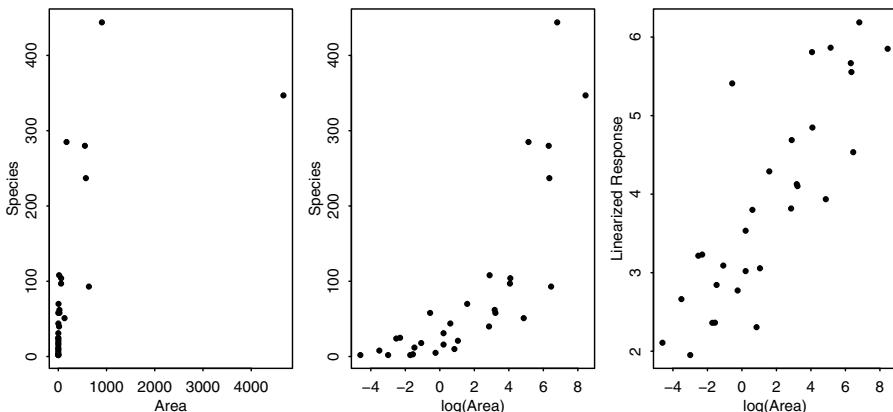


Figure 8.2 *Plots of the number of species against area for the Galápagos data. The first plot clearly shows a need for transformation, the second shows the advantage of using logged area, while the third shows the value of using the linearized response.*

We see that both variables have skewed distributions. We start with a log transformation on the predictor as seen in the second panel of Figure 8.2:

```
plot(Species ~ log(Area), gala)
```

We see a curvilinear relationship between the predictor and the response. However, the default Poisson GLM uses a log link which we need to take into account. To allow for the choice of link function, we can plot the linearized response:

$$z = \eta + (y - \mu) \frac{d\eta}{d\mu}$$

as we see in the third panel of Figure 8.2:

```
mu <- predict(modp, type="response")
z <- predict(modp) + (gala$Species-mu) / mu
plot(z ~ log(Area), gala, ylab="Linearized Response")
```

We now see a linear relationship suggesting that no further transformation of area is necessary. Notice that we used the current model in the computation of  $z$ . Some might prefer to use an initial guess here to avoid presuming the choice of model. For this dataset, we find that a log transformation of all the predictors is helpful:

```
modp1 <- glm(Species ~ log(Area) + log(Elevation) + log(Nearest) +
  log(Scruz+0.1) + log(Adjacent), family=poisson, gala)
c(deviance(modp), deviance(modp1))
[1] 716.85 359.12
```

We see that this results in a substantial reduction in the deviance.

The disadvantage of simply examining the raw relationship between the response and the other predictors is that it fails to take into account the effect of the other predictors. Partial residual plots are used for linear models to make allowance for the effect of the other predictors while focusing on the relationship of interest. These can be adapted for use in GLMs by plotting  $z - \hat{\eta} + \beta_j x_j$  vs.  $x_j$ . The interpretation is the same as in the linear model case. We compute the partial residual plot for the (now logged) area, as shown in the first panel of Figure 8.3:

```
mu <- predict(modpl, type="response")
u <- (gala$Species-mu)/mu + coef(modpl)[2]*log(gala$Area)
plot(u ~ log(Area), gala, ylab="Partial Residual")
abline(0, coef(modpl)[2])
```

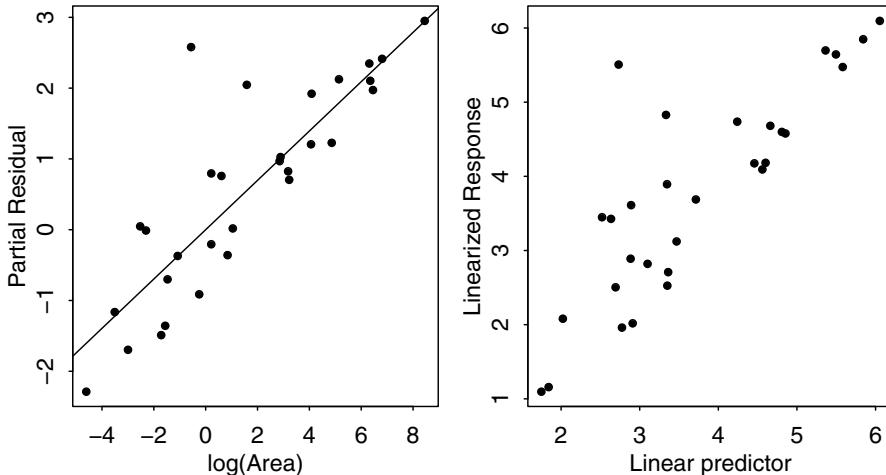


Figure 8.3 A partial residual plot for  $\log(\text{Area})$  is shown on the left while a diagnostic for the link function is shown on the right.

In this plot, we see no reason for concern. There is no nonlinearity indicating a need to transform nor are there any obvious outliers or influential points. Partial residuals can also be obtained from `residuals(., type="partial")` although an offset will be necessary if you want the regression line displayed correctly on the plot.

One can search for good transformations of the predictors in nongraphical ways. Polynomial terms or spline functions of the predictors can be experimented with, but generalized additive models, described in Chapter 15, offer a more direct way to discover some good transformations.

The link function is a fundamental assumption of the GLM. Quite often the choice of link function is set by the characteristics of the response, such as positivity, or by ease of interpretation, as with logit link for binomial GLMs. It is often difficult to contemplate alternatives. Nevertheless, it is worth checking to see whether the link assumption is not grossly wrong. Before doing this, it is important to eliminate other simpler violations of the assumptions that are more easily rectified such as outliers or transformations of the predictors. After these concerns have been eliminated, one can check the link assumption by making a plot of the linearized response  $z$  against linear predictor  $\hat{\eta}$ . An example of this is shown in the second panel of Figure 8.3:

```
z <- predict(modpl)+(gala$Species-mu)/mu
plot(z ~ predict(modpl), xlab="Linear predictor",
     ylab="Linearized Response")
```

In this case, we see no indication of a problem.

An alternative approach to checking the link function is to propose a family of link functions of which the current choice is a member. A range of links can then be

fit and compared to the current choice. The approach is analogous to the Box–Cox method used for linear models. Alternative choices are easier to explore within the quasi-likelihood framework described in Section 9.4.

**Unusual Points:** We have already described the raw material of residuals, leverage and influence measures that can be used to check for points that do not fit the model or influence the fit unduly. Let's now see how to use graphical methods to examine these quantities.

The QQ plot of the residuals is the standard way to check the normality assumption on the errors typically made for a linear model. For a GLM, we do not expect the residuals to be normally distributed, but we are still interested in detecting outliers. For this purpose, it is better to use a half-normal plot that compares the sorted absolute residuals and the quantiles of the half-normal distribution:

$$\Phi^{-1} \left( \frac{n+i}{2n+1} \right) \quad i = 1, \dots, n$$

The residuals are not expected to be normally distributed, so we are not looking for an approximate straight line. We only seek outliers which may be identified as points off the trend. A half-normal plot is better for this purpose because in a sense the resolution of the plot is doubled by having all the points in one tail.

Since we are more specifically interested in outliers, we should plot the jackknife residuals. An example for the Galápagos model is shown in the first panel of Figure 8.4:

```
halfnorm(rstudent(modp1))
```

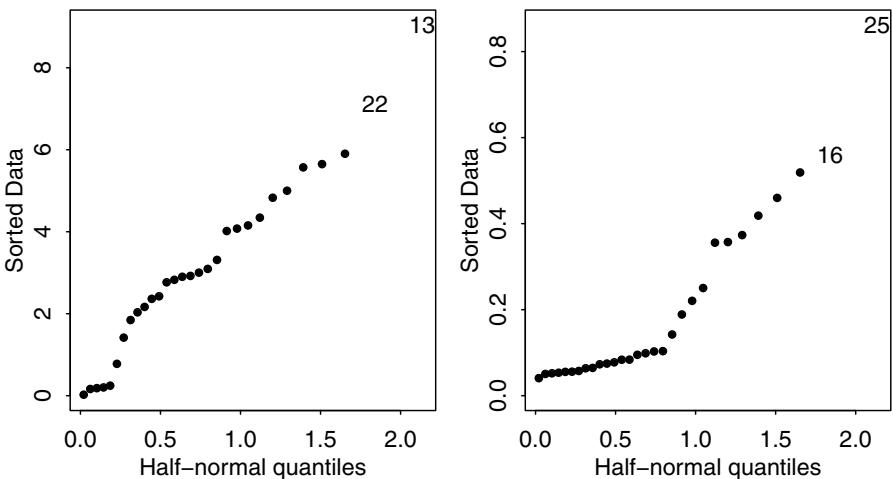


Figure 8.4 Half-normal plots of the jackknife residuals on the left and the leverages on the right.

We see no sign of outliers in the plot. The half-normal plot is also useful for positive-valued diagnostics such as the leverages and the Cook statistics. A look at the leverages is shown in the second panel of Figure 8.4:

```
gali <- influence(modpl)
halfnorm(gali$hat)
```

There is some indication that case 25, Santa Cruz island, may have some leverage. The predictor `Scruz` is the distance from Santa Cruz island which is zero for this case. This posed a problem for making the log transformation and explains why we added 0.1 to this variable. However, there is some indication that this inelegant fix may be causing some difficulty.

Moving on to influence, a half-normal plot of the Cook statistics is shown in the first panel of Figure 8.5:

```
halfnorm(cooks.distance(modpl))
```

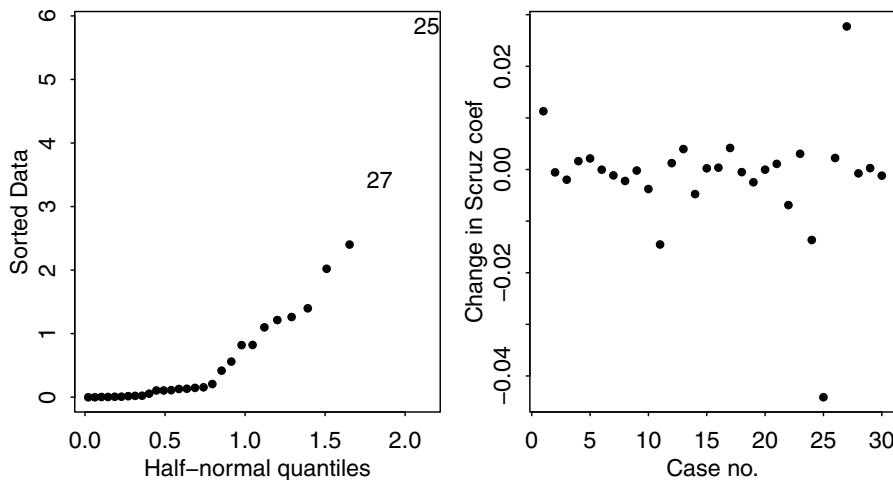


Figure 8.5 Half-normal plot of the Cook statistics is shown on the left and an index plot of the change in the `Scruz` coefficient is shown on the right.

Again we have some indication that Santa Cruz island is influential. We can examine the change in the fitted coefficients. For example, consider the change in the `Scruz` coefficient as shown in the second panel of Figure 8.5:

```
plot(gali$coef[,5], ylab="Change in Scruz coef", xlab="Case no.")
```

We see a substantial change for case 25. If we compare the full fit to a model without this case, we find:

```
modplr <- glm(Species ~ log(Area) + log(Elevation) + log(Nearest)
+ log(Scruz+0.1) + log(Adjacent), family=poisson, gala, subset=-25)
cbind(coef(modpl), coef(modplr))
```

	[,1]	[,2]
(Intercept)	3.287941	3.050699
log(Area)	0.348445	0.334530
log(Elevation)	0.036421	0.059603
log(Nearest)	-0.040644	-0.052548
log(Scruz + 0.1)	-0.030045	0.015919
log(Adjacent)	-0.089014	-0.088516

We see a sign change for the `Scruz` coefficient. This is interesting since in the full

model, the coefficient is more than twice the standard error way from zero indicating some significance. A simple solution is to add a larger amount, say 0.5, to `Scruz`.

Other than this user-introduced anomaly, we find no difficulty. Using our earlier discovery of the log transformation, some variable selection and allowing for remaining overdispersion, our final model is:

```
modpla <- glm(Species ~ log(Area)+log(Adjacent), family=poisson, gala)
dp <- sum(residuals(modpla,type="pearson")^2)/modpla$df.res
summary(modpla,dispersion=dp)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	3.2767	0.1794	18.26	< 2e-16
log(Area)	0.3750	0.0326	11.50	< 2e-16
log(Adjacent)	-0.0957	0.0249	-3.85	0.00012

overdispersion parameter = 16.527

n = 30 p = 3

Deviance = 395.543 Null Deviance = 3510.729 (Difference = 3115.186)

Notice that the deviance is much lower and the elevation variable is not used when compared with our model choice in Section 5.1.

This example concerned a Poisson GLM. Diagnostics for binomial GLMs are similar, but see Pregibon (1981) and Collett (2003) for more details.

## 8.5 Sandwich Estimation

Inference from GLMs depends on the accuracy of the assumptions of the model. Diagnostics are one way in which we can detect violations for these assumptions. Sometimes we can modify the model so that these violations no longer occur or are sufficiently minor to ignore. But sometimes we may not be able or wish to change the model. Instead we would like to modify the inferential procedures so that they are insensitive to these violations. Two common forms of violation are heteroscedascity (unequal variance) and autocorrelation in the response. In GLMs, excepting the Gaussian case, the variance of the response varies as a function of the mean response. By heteroscedascity, we mean the problem where this variation goes beyond what is expected from the mean.

In the standard linear model, ordinary least squares estimation is not optimal in the presence of heteroscedascity and/or correlation in the errors. Nevertheless, it still retains some desirable properties (unbiasedness) and we might retain these estimates especially if we were unsure of a modification that would surely improve them. For GLMs, we also wish to retain the standard MLEs of  $\beta$  because these will still be reasonable. For example, we have seen that these estimators are unchanged in the presence of overdispersion. However, the estimate of  $\text{var } \hat{\beta}$  is much more sensitive to these violations of the assumptions. *Sandwich estimation* is a method for obtaining better estimates of  $\text{var } \hat{\beta}$  in these circumstances.

We describe how these are applied for standard linear models (LM) to avoid complicating the exposition but the extension to GLMs is straightforward. We have

$$y = X\beta + \varepsilon \quad \text{where} \quad \text{var } \varepsilon = \Omega$$

Using the standard ordinary least squares estimate  $\hat{\beta}$  gives

$$\text{var } \hat{\beta} = (X^T X)^{-1} X^T \Omega X (X^T X)^{-1}$$

The usual assumption of independent and identical errors has  $\Omega = \sigma^2 I$  which means we reduce to  $\text{var } \hat{\beta} = \sigma^2 (X^T X)^{-1}$ . But suppose we believe that this assumption has been violated. Let's focus on the heteroscedastic case with no correlation. We might estimate  $\Omega = \text{diag}(\omega_1, \dots, \omega_n)$  in order to obtain a better estimate of the variance. Various estimates have been proposed but one that has been shown to behave well in the small sample case is:

$$\omega_i = \frac{\hat{\varepsilon}_i^2}{(1 - h_i)^2}$$

for leverages  $h_i$ . This allows the computation of a sandwich estimate of the variance of  $\hat{\beta}$ . The estimator gets its name from the bread,  $(X^T X)^{-1}$  and the meat,  $X^T \Omega X$ .

The estimator can be extended to handle autocorrelation in the response which might be a concern for time-ordered data. We have described the LM solution but the extension to GLMs is straightforward. For more details, see Zeileis (2004).

Sandwich estimators are implemented in the sandwich package. We illustrate their application to the Galápagos data using the final model of the previous section:

```
data(gala, package="faraway")
library(sandwich)
modpla <- glm(Species ~ log(Area) + log(Adjacent), family=poisson, gala)
(sebeta <- sqrt(diag(vcovHC(modpla))))
```

(Intercept)	log(Area)	log(Adjacent)
0.181162	0.034877	0.026938

Notice that these standard errors for the regression coefficients are quite similar to the previous model where we used a dispersion parameter to account for the known violation of the mean equal to variance assumption of the strictly Poisson GLM. Hence the sandwich estimator also takes care of the overdispersion problem and it is not necessary to introduce a dispersion parameter into the sandwich-based solution. Wald tests of the parameters can be obtained by simply dividing the parameter estimates by their standard errors as before with much the same outcome.

## 8.6 Robust Estimation

When we compute the MLE for a GLM we are required to solve a set of equations as seen in Section 8.2:

$$\sum_i \gamma_i \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0 \quad \forall j$$

We have added weights  $\gamma_i$  to give us some additional flexibility. These weights are distinct from previously defined weights  $w_i$  and are used for the specific purpose of downweighting extreme observations where  $y_i$  is far from  $\mu_i$ . One problem with the standard MLE is that as a particular  $y_i$  becomes more extreme, the influence on  $\hat{\beta}$  is unbounded. We want to define  $\gamma_i$  in such a way that each observation has *bounded influence*. Now consider

$$E \gamma \frac{(y - \mu)}{V(\mu)} \frac{\partial \mu}{\partial \beta_j}$$

When  $\gamma = 1$  we know from standard likelihood theory that this expectation is equal to zero. We will need  $\gamma$  to be defined as a function of  $y$ ,  $x$  and  $\beta$  to ensure we achieve bounded influence but we still want the expectation to be zero as in the standard case. In other words, we want these to be *conditionally unbiased* given  $x$ .

These considerations lead to the definition of the CUBIF (conditionally unbiased influence function) method of robust estimation. Our description has been intuitive but for the full details see Kunsch et al. (1989). This method is implemented in the robust package of Wang et al. (2014) which we apply to the same Galápagos model of the previous sections:

```
data(gala, package="faraway")
library(robust)
rmodpla <- glmrob(Species ~ log(Area)+log(Adjacent), family=poisson,
                     ~ data=gala)
summary(rmodpla)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.98321   0.05198  57.4   <2e-16
log(Area)    0.42115   0.00925  45.5   <2e-16
log(Adjacent) -0.11214   0.00660 -17.0   <2e-16

Robustness weights w.r * w.x:
 9 weights are ~= 1. The remaining 21 ones are summarized as
   Min. 1st Qu. Median Mean 3rd Qu. Max.
 0.0879  0.2540  0.3820  0.4340  0.6300  0.8330
```

This implementation uses a Huber-style influence function. There are various fitting parameters which can be tweaked but we have shown the default choices here.

There are two strategies for using robust models. One approach is to use them as the primary method of estimation. We can use them to make inferences about relationships between the variables and to predict future observations. The drawback is that the statistical methodology to achieve this is more complicated and less developed than for standard GLMs. Progress can be made but it is more difficult. In some applications, a very large number of models will be fitted or future data will be used so that it will be impractical for a human analyst to judge the results. In such circumstances, we may prefer the robust modeling approach as it will be less sensitive to unexpected or aberrant observations.

The other approach is to view the robust fit as an ancillary to the standard GLM fit. We can compare the results of these two fits. In this particular example, we find that the coefficients are not substantially different. We might proceed to make conclusions using the standard fit with the added confidence that these results were not unduly sensitive to some extreme observations. If we had seen a substantially different fit, we would investigate to find the cause of the difference.

It is worth checking which observations received the lowest weighting in the robust model:

```
wts <- rmodpla$w.r
names(wts) <- row.names(gala)
head(sort(wts))

Gardner     Rabida     Bartolome      Seymour Daphne.Minor
  0.087863   0.114472   0.214854   0.216445   0.228807
```

Our previous diagnostic analyses also detected unusual points but these methods struggle with the problem that such points often influence the fit itself. At most, they

can check the effect of excluding one case whereas robust methods are able to handle multiple unusual cases. In this example, we might examine these cases more closely to see what makes them unusual.

**Further Reading:** The canonical book on GLMs is McCullagh and Nelder (1989). Other books include Dobson and Barnett (2008), Lindsey (1997), Myers et al. (2002), Gill (2001) and Fahrmeir and Tutz (2001). For a Bayesian perspective, see Dey et al. (2000).

### Exercises

1. Data is generated from the exponential distribution with density  $f(y) = \lambda \exp(-\lambda y)$  where  $\lambda, y > 0$ .
  - (a) Identify the specific form of  $\theta, \phi, a(), b()$  and  $c()$  for the exponential distribution.
  - (b) What is the canonical link and variance function for a GLM with a response following the exponential distribution?
  - (c) Identify a practical difficulty that may arise when using the canonical link in this instance.
  - (d) When comparing nested models in this case, should an  $F$  or  $\chi^2$  test be used? Explain.
  - (e) Express the deviance in this case in terms of the responses  $y_i$  and the fitted values  $\hat{\mu}_i$ .
2. The Conway–Maxwell–Poisson distribution has probability function:

$$P(Y = y) = \frac{\lambda^y}{(y!)^\nu} \frac{1}{Z(\lambda, \nu)} \quad y = 0, 1, 2, \dots$$

where

$$Z(\lambda, \nu) = \sum_{i=0}^{\infty} \frac{\lambda^i}{(i!)^\nu}$$

Place this in exponential family form, identifying all the relevant components necessary for use in a GLM.

3. Consider the following distributions and determine whether they can be put in exponential family form. If it is not possible, consider whether there is a special case which follows the form.

- The Uniform distribution:

$$f(y|\alpha, \beta) = \frac{1}{\beta - \alpha} I(y \in [\alpha, \beta])$$

- The Weibull distribution: ( $y > 0$ )

$$f(y|\alpha, \beta) = \alpha \beta^{-\alpha} y^{\alpha-1} e^{-(y/\beta)^\alpha}$$

- The Gumbel distribution:  $(y > 0)$

$$f(y|\alpha, \beta) = \frac{1}{\beta} \exp((y - \alpha)/\beta - \exp((y - \alpha)/\beta))$$

4. Consider the Galápagos data and model analyzed in this chapter. The purpose of this question is to reproduce the details of the GLM fitting of this data.
- Fit a Poisson model to the species response with the five geographic variables as predictors. Do not use the endemics variable. Report the values of the coefficients and the deviance.
  - For a Poisson GLM, derive  $\eta$ ,  $d\eta/d\mu$ ,  $V(\mu)$  and the weights to be used in an iteratively fit GLM. What is the form of the adjusted dependent variable here?
  - Using the observed response as initial values, compute the first stage of the iteration, stopping after the first linear model fit. Compare the coefficients of this linear model to those found in the GLM fit. How close are they?
  - Continue the iteration to get the next  $\eta$  and  $\mu$ . Use this to compute the current value of the deviance. How close is this to the deviance from the GLM?
  - Compute one more iteration of the GLM fit, reporting the next calculation of the coefficients and deviance. How close are these to target now?
  - Repeat these iterations a few more times, computing the deviance in each time. Stop when the deviance does not change much. Compare your final estimated coefficients to that produced by the GLM fit.
  - Use your final iterated linear model fit to produce standard errors for the coefficients. How close are these to that produced by the direct GLM fit?
5. Again using the Galápagos data, fit a Poisson model to the species response with the five geographic variables as predictors. Do not use the endemics variable. The purpose of this question is to compare six different ways of testing the significance of the elevation predictor, i.e.,  $H_0 : \beta_{Elev} = 0$ . In each case, report the  $p$ -value.
- Use the  $z$ -statistic from the model summary.
  - Fit a model without elevation and use the difference in deviances to make the test.
  - Use the Pearson Chi-squared statistic in place of the deviance in the previous test.
  - Fit the Poisson model with a free dispersion parameter as described in Section 5.2. Make the test using the model summary.
  - Use the sandwich estimation method for the standard errors in the original model. Use these to compute  $z$ -statistics.
  - Use the robust GLM estimation method and report the test result from the summary.
  - Compare all six results. Pick the best one and justify your choice.

6. The `worldcup` data were collected at the 2010 World Cup. We are interested in modeling the number of shots taken by each player. As goalkeepers do not normally shoot, you should remove them from the dataset. Due to substitution and varying success in the tournament, the number of minutes played by each player is quite variable. For this reason, compute new variables that represent the number of tackles and passes made per 90-minute game.
- (a) Fit a Poisson model with the number of shots as the response and team, position, tackles and passes per game as predictor. Note that time played is a rate variable and should be accounted for as described in Section 5.3. Interpret the effect of tackles and passes on shots.
  - (b) Calculate the leverages for the current model. Report which player has the highest leverage and suggest why this might be so. Make an appropriate plot of the leverages and comment on whether any leverage is exceptional.
  - (c) Compute the change in the regression coefficients when each case is dropped. In particular, examine the change in the tackle coefficient identifying the player which causes the greatest absolute change in this value. What is unusual about this player? Plot the change in the tackle coefficient and determine if any of the values is particularly large.
  - (d) Calculate the Cook Statistics. Which player has the largest such statistic and what is unusual about him?
  - (e) Find the jackknife residuals. Find the player with the largest absolute residual of this kind. How did he come to be the largest?
  - (f) Plot the residuals against the appropriate fitted values. Explain the source of the lines of points appearing on the plot. What does this plot indicate?
  - (g) Make the following three plots:
    - i. Plot raw shots against tackles.
    - ii. Plot shots per game against tackles per game.
    - iii. Plot the linearized response against tackles per game.Make an interpretation of each plot and choose the best one for discovering the relationship between this predictor and the response. Justify your choice.
  - (h) Construct the partial residual plot for tackles and interpret. Is the point on the far right really influential?
  - (i) Make a diagnostic plot to check the choice of the (default) link function.

**This page intentionally left blank**

# Other GLMs

---

The binomial, Gaussian and Poisson GLMs are by far the most commonly used, but there are a number of less popular GLMs that are useful for particular types of data. The gamma and inverse Gaussian are intended for continuous, skewed responses. In some cases, we are interested in modeling both the mean and the dispersion of the response and so we present dual GLMs for this purpose. The quasi-GLM is a model that is useful for nonstandard responses where we are unwilling to specify the distribution but can state the link and variance functions.

## 9.1 Gamma GLM

The density of the gamma distribution is usually given by:

$$f(y) = \frac{1}{\Gamma(v)} \lambda^v y^{v-1} e^{-\lambda y} \quad y > 0$$

where  $v$  describes the shape and  $\lambda$  describes the scale of the distribution. However, for the purposes of a GLM, it is convenient to reparameterize by putting  $\lambda = v/\mu$  to get:

$$f(y) = \frac{1}{\Gamma(v)} \left(\frac{v}{\mu}\right)^v y^{v-1} e^{-\left(\frac{y}{\mu}\right)} \quad y > 0$$

Now  $EY = \mu$  and  $\text{var } Y = \mu^2/v = (EY)^2/v$ . The dispersion parameter is  $\phi = v^{-1}$ . Here we plot a gamma density with three different values of the shape parameter  $v$  (the scale parameter would just have a multiplicative effect) as seen in Figure 9.1:

```
x <- seq(0, 8, by=0.1)
plot(x, dgamma(x, 0.75), type="l", ylab="", xlab="", ylim=c(0,1.25),
      ↪ xaxs="i", yaxs="i")
plot(x, dgamma(x, 1.0), type="l", ylab="", xlab="", ylim=c(0,1.25), xaxs
      ↪ ="i", yaxs="i")
plot(x, dgamma(x, 2.0), type="l", ylab="", xlab="", ylim=c(0,1.25), xaxs
      ↪ ="i", yaxs="i")
```

The gamma distribution can arise in various ways. The sum of  $v$  independent and identically distributed exponential random variables with rate  $\lambda$  has a gamma distribution. The  $\chi^2$  distribution is a special case of the gamma where  $\lambda = 1/2$  and  $v = df/2$ .

The canonical parameter is  $-1/\mu$ , so the canonical link is  $\eta = -1/\mu$ . However, we typically remove the minus (which is fine provided we take account of this in any derivations) and just use the inverse link. We also have  $b(\theta) = \log(1/\mu) = -\log(-\theta)$

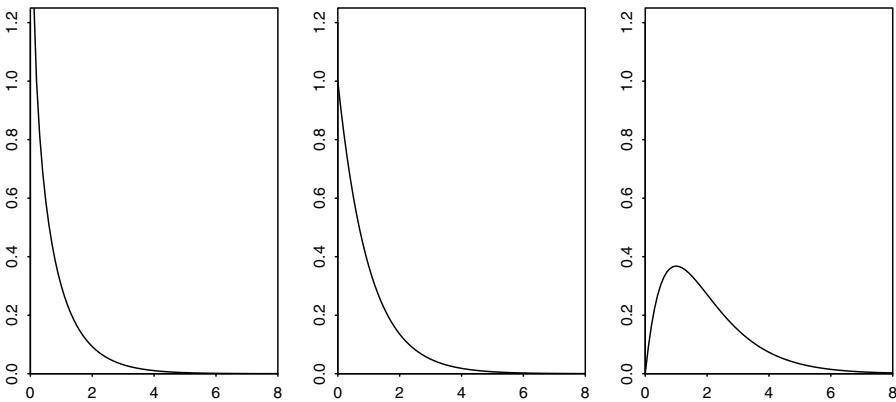


Figure 9.1 *The gamma density explored. In the first panel  $v = 0.75$  and we see that the density is unbounded at zero. In the second panel,  $v = 1$  which is the exponential density. In the third panel,  $v = 2$  and we see a skewed distribution.*

and so  $b''(\theta) = \mu^2$  is the variance function. The (unscaled) deviance is:

$$D(y, \hat{\mu}) = -2 \sum \{ \log y_i / \hat{\mu}_i - (y_i - \hat{\mu}_i) / \hat{\mu}_i \}$$

The utility of the gamma GLM arises in two different ways. Certainly, if we believe the response to have a gamma distribution, the model is clearly applicable. However, the model can also be useful in other situations where we may be willing to speculate on the relationship between the mean and the variance of the response but are not sure about the distribution. Indeed, it is possible to grasp the mean-to-variance relationship from graphical displays with relatively small datasets, while assertions about the response distribution would require a lot more data.

In the Gaussian linear model,  $\text{var } Y$  is constant as a function of the mean response. This is a fundamental assumption necessary for the optimality of least squares. However, sometimes contextual knowledge of the data or diagnostics shows that  $\text{var } Y$  is nonconstant. When the form of nonconstancy is known exactly, then weighted least squares can be used. In practice, however, the form is often not known exactly. Alternatively, the transformation of  $Y$  may lead to a constant variance model. The difficulty here is that while the original scale  $Y$  may be meaningful,  $\log Y$  or  $\sqrt{Y}$ , for example, may not. An example is where a sum of the  $Y$ s may be of interest. In such a case, transformation would be a hindrance.

If  $\text{var } Y \propto EY$ , then  $\sqrt{Y}$  is the variance stabilizing transform. If one wants to avoid a transformation, a GLM approach can be used. When  $Y$  has a Poisson distribution, then  $\text{var } Y \propto EY$ , suggesting the use of a Poisson GLM. Now one might object that the Poisson is a distribution for discrete data, which would seem to disallow its use for continuous responses. However, fitting a GLM only depends on the mean and variance of a distribution; the other moments are not used. This is important because it indicates that we need specify nothing more than the mean and variance. The distribution could be discrete or continuous and it would make no difference.

For some data, we might expect the standard deviation to increase linearly with the response. If so, the coefficient of variation,  $SD Y/EY$ , would be constant and  $\text{var } Y \propto (EY)^2$ . For example, measurements of larger objects do tend to have more error than smaller ones.

If we wanted to apply a Gaussian linear model, the log transform is indicated. This would imply a lognormal distribution for the original response. Alternatively, if  $Y \sim \text{gamma}$ , then  $\text{var } Y \propto (EY)^2$ , so a gamma GLM is also appropriate in this situation. In a few cases, one may have some knowledge concerning the true distribution of the response which would drive the choice. However, in many cases, it would be difficult to distinguish between these two options on the basis of the data alone and the choice would be driven by the purpose and desired interpretation of the model.

There are three common choices of link function:

1. The canonical link is  $\eta = \mu^{-1}$ . Since  $-\infty < \eta < \infty$ , the link does not guarantee  $\mu > 0$  which could cause problems and might require restrictions on  $\beta$  or on the range of possible predictor values. On the other hand, the reciprocal link has some advantages. The Michaelis–Menten model, used in biochemistry, has:

$$Ey = \mu = \frac{\alpha_0 x}{1 + \alpha_1 x}$$

which can be represented after some reexpression as:

$$\eta = \alpha_1/\alpha_0 + 1/(\alpha_0 x) = \mu^{-1}$$

As  $x$  increases,  $\eta \rightarrow \alpha_1/\alpha_0$ , which means that the mean  $\mu$  will be bounded. The inverse link can be useful in such situations where we know the mean response to be bounded.

2. The log link,  $\eta = \log \mu$ , should be used when the effect of the predictors is suspected to be multiplicative on the mean. When the variance is small, this approach is similar to a Gaussian model with a logged response.
3. The linear link,  $\eta = \mu$ , is useful for modeling sums of squares or variance components which are  $\chi^2$ . This is a special case of the gamma.

The general GLM procedures apply to the analysis and fitting. To estimate the dispersion, McCullagh and Nelder (1989) recommend the use of:

$$\hat{\phi} = \frac{1}{\hat{v}} = \frac{X^2}{n - p}$$

The maximum likelihood estimator and the usual estimator,  $D/(n - p)$ , are both sensitive to unusually small values of the response and are not consistent estimates of the coefficient of variation when the gamma distribution assumption does not hold.

Myers and Montgomery (1997) present data from a step in the manufacturing process for semiconductors. Four factors are believed to influence the resistivity of the wafer and so a full factorial experiment with two levels of each factor was run. Previous experience led to the expectation that resistivity would have a skewed distribution and so the need for transformation was anticipated. We start with a look at the data:

```
data(wafer, package="faraway")
summary(wafer)

x1      x2      x3      x4      resist
-:8     -:8     -:8     -:8    Min.   :166
+:8     +:8     +:8     +:8   1st Qu.:201
                           Median :214
                           Mean   :229
                           3rd Qu.:259
                           Max.   :340
```

The application of the Box–Cox method or past experience suggests the use of a log transformation on the response. We fit the full model and then reduce it using AIC-based model selection:

```
l1mdl <- lm(log(resist) ~ .^2, wafer)
rlmdl <- step(l1mdl)
library(faraway)
summary(rlmdl)

Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.3111  0.0476 111.53 4.7e-14
x1+         0.2009  0.0476   4.22  0.00292
x2+        -0.2107  0.0476  -4.43  0.00221
x3+         0.4372  0.0673   6.49  0.00019
x4+         0.0354  0.0476   0.74  0.47892
x1+:x3+   -0.1562  0.0673  -2.32  0.04896
x2+:x3+   -0.1782  0.0673  -2.65  0.02941
x3+:x4+   -0.1830  0.0673  -2.72  0.02635
```

$n = 16$ ,  $p = 8$ , Residual SE = 0.067, R-Squared = 0.95

We find a model with three two-way interactions, all with  $x3$ .

Now we fit the corresponding gamma GLM and again select the model using the AIC criterion. Note that the family must be specified as Gamma rather than gamma to avoid confusion with the  $\Gamma$  function. We use the log link to be consistent with the linear model. This must be specified as the default is the inverse link:

```
gmdl <- glm(resist ~ .^2, family=Gamma(link=log), wafer)
rgmdl <- step(gmdl)
summary(rgmdl)

Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.3120  0.0476 111.68 4.6e-14
x1+         0.2003  0.0476   4.21  0.00295
x2+        -0.2110  0.0476  -4.44  0.00218
x3+         0.4367  0.0673   6.49  0.00019
x4+         0.0354  0.0476   0.74  0.47836
x1+:x3+   -0.1555  0.0673  -2.31  0.04957
x2+:x3+   -0.1763  0.0673  -2.62  0.03064
x3+:x4+   -0.1819  0.0673  -2.70  0.02687
```

Dispersion parameter = 0.005

$n = 16$   $p = 8$

Deviance = 0.036 Null Deviance = 0.698 (Difference = 0.662)

In this case, we see that the coefficients are remarkably similar to the linear model with the logged response. Even the standard errors are almost identical and the square root of the dispersion corresponds to the residual standard error of the linear model:

```
sqrt(summary(rgmdl)$dispersion)
```

[1] 0.067267

The maximum likelihood estimate of  $\phi$  may be computed using the MASS package:

```
library(MASS)
gamma.dispersion(rgmdl)
[1] 0.0022657
```

We see that this gives a substantially smaller estimate, which would suggest smaller standard errors. However, it is not consistent with our experience with the Gaussian linear model in this example.

In this example, because the value of  $v = 1/\phi$  is large (221), the gamma distribution is well approximated by a normal. Similarly, for the logged response linear model, a lognormal distribution with a small variance ( $\sigma = 0.0673$ ) is also very well approximated by a normal. For this reason, there is not much to distinguish these two models. The gamma GLM has the advantage of modeling the response directly while the lognormal has the added convenience of working with a standard linear model.

Let us examine another example where there is more distinction between the two approaches. In Hallin and Ingenbleek (1983) data on payments for insurance claims for various areas of Sweden in 1977 are presented. The data is further subdivided by mileage driven, the bonus from not having made previous claims and the type of car. We have information on the number of insured, measured in policy-years, within each of these groups. Since we expect that the total amount of the claims for a group will be proportionate to the number of insured, it makes sense to treat the log of the number insured as an offset for similar reasons to those in Section 5.3. Attention has been restricted to data from Zone 1. After some model selection, a gamma GLM of the following form was found:

```
data(motorins, package="faraway")
motori <- motorins[motorins$Zone == 1,]
g1 <- glm(Payment ~ offset(log(Insured)) + as.numeric(Kilometres) +
  ~ Make+Bonus, family=Gamma(link=log), motori)
summary(g1)
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	6.5273	0.1777	36.72	< 2e-16
as.numeric(Kilometres)	0.1201	0.0311	3.85	0.00014
Make2	0.4070	0.1782	2.28	0.02313
Make3	0.1553	0.1796	0.87	0.38767
Make4	-0.3439	0.1915	-1.80	0.07355
Make5	0.1447	0.1810	0.80	0.42473
Make6	-0.3456	0.1782	-1.94	0.05352
Make7	0.0614	0.1824	0.34	0.73689
Make8	0.7504	0.1873	4.01	0.000079
Make9	0.0320	0.1782	0.18	0.85778
Bonus	-0.2007	0.0215	-9.33	< 2e-16

```
Dispersion parameter = 0.555
n = 295 p = 11
Deviance = 155.056 Null Deviance = 238.974 (Difference = 83.918)
```

In comparison, the lognormal model, where we have used the `glm` function for compatibility, looks like this:

```
llg <- glm(log(Payment) ~ offset(log(Insured))+as.numeric(Kilometres) +
  ~ Make+Bonus,family=gaussian , motori)
summary(llg)
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	6.51403	0.18634	34.96	< 2e-16

```

as.numeric(Kilometres) 0.05713 0.03265 1.75 0.0813
Make2                  0.36387 0.18686 1.95 0.0525
Make3                  0.00692 0.18824 0.04 0.9707
Make4                  -0.54786 0.20076 -2.73 0.0067
Make5                  -0.02179 0.18972 -0.11 0.9087
Make6                  -0.45881 0.18686 -2.46 0.0147
Make7                  -0.32118 0.19126 -1.68 0.0942
Make8                  0.20958 0.19631 1.07 0.2866
Make9                  0.12545 0.18686 0.67 0.5025
Bonus                 -0.17806 0.02254 -7.90 6.2e-14

```

```

Dispersion parameter = 0.611
n = 295 p = 11
Deviance = 173.529 Null Deviance = 238.565 (Difference = 65.035)

```

Notice that there are now important differences between the two models. We see that mileage class given by Kilometers is statistically significant in the gamma GLM, but not in the lognormal model. Some of the coefficients are quite different. For example, we see that for make 8, relative to the reference level of make 1, there are  $\exp(0.7504) = 2.1178$  times as much payment when using the gamma GLM, while the comparable figure for the lognormal model is  $\exp(0.20958) = 1.2332$ .

These two models are not nested and have different distributions for the response, which makes direct comparison problematic. The AIC criterion, which is minus twice the maximized likelihood plus twice the number of parameters, has often been used as a way to choose between models. Smaller values are preferred. However, when computing a likelihood, it is common practice to discard parts that are not functions of the parameters. This has no consequence when models with same distribution for the response are compared since the parts discarded will be equal. For responses with different distributions, it is essential that all parts of the likelihood be retained. The large difference in AIC for these two models indicates that this precaution was not taken. Nevertheless, we note that the null deviance for both models is almost the same while the residual deviance is smaller for the gamma GLM. This improvement relative to the null indicates that the gamma GLM should be preferred here. Note that purely numerical comparisons such as this are risky and that some attention to residual diagnostics, scientific context and interpretation is necessary.

We compare the shapes of the distributions for the response using the dispersion estimates from the two models, as seen in Figure 9.2:

```

x <- seq(0,5,by=0.05)
plot(x,dgamma(x,1/0.55597,scale=0.55597),type="l",ylab="", xlab="",
      → yaxs="i",ylim=c(0,1))
plot(x,dlnorm(x,meanlog=-0.30551,sdlog=sqrt(0.61102)),type="l", ylab=
      → "",xlab="",yaxs="i",ylim=c(0,1))

```

We see the greater peakedness of the lognormal indicating more small payments which are balanced by more large payments. The lognormal thus has higher kurtosis.

We may also make predictions from both models. Here is a plausible value of the predictors:

```
x0 <- data.frame(Make="1",Kilometres=1,Bonus=1,Insured=100)
```

and here is predicted response for the gamma GLM:

```
predict(gl,new=x0,se=T,type="response")
```

```
Sfit
```

```
[1] 63061
```

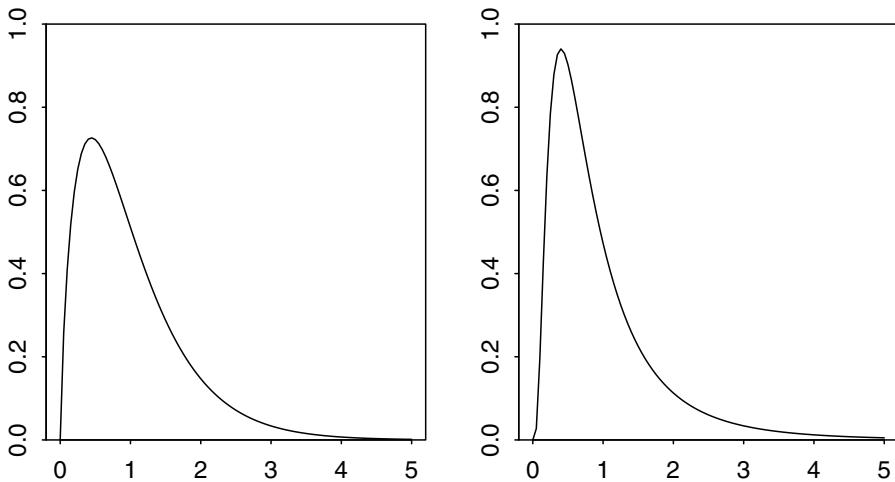


Figure 9.2 *Gamma density for observed shape of  $1/0.55597$  is shown on the left and lognormal density for an observed SD on the log scale of  $\sqrt{0.61102}$ . The means have been set to one in both cases.*

```
$se.fit
[1] 9711.5
```

For the lognormal, we have:

```
predict(l1g,new=x0,se=T,type="response")
$fit
[1] 10.998
```

```
$se.fit
[1] 0.16145
```

so that the corresponding values on the original scale would be:

```
c(exp(10.998),exp(10.998)*0.16145)
[1] 59754.5 9647.4
```

where we have used the delta method to estimate the standard error on the original scale.

## 9.2 Inverse Gaussian GLM

The density of an inverse Gaussian random variable,  $Y \sim IG(\mu, \lambda)$ , is:

$$f(y|\mu, \lambda) = (\lambda/2\pi)^{1/2} \exp[-\lambda(y-\mu)^2/2\mu^2y] \quad y, \mu, \lambda > 0$$

The mean is  $\mu$  and the variance is  $\mu^3/\lambda$ . The canonical link is  $\eta = 1/\mu^2$  and the variance function is  $V(\mu) = \mu^3$ . The deviance is given by:

$$D = \sum_i (y_i - \hat{\mu}_i)^2 / (\hat{\mu}_i^2 y_i)$$

Plots of the inverse Gaussian density for a range of values of the shape parameter,  $\lambda$ , are shown in Figure 9.3. We require the `SuppDists` package of Wheeler (2013).

```
library(SuppDists)
x <- seq(0,8,by=0.1)
plot(x,dinvGauss(x,1,0.5),type="l",ylab="",xlab="",ylim=c(0,1.5),
     xaxs="i",yaxs="i")
plot(x,dinvGauss(x,1,1),type="l",ylab="",xlab="",ylim=c(0,1.5),
     xaxs="i",yaxs="i")
plot(x,dinvGauss(x,1,5),type="l",ylab="",xlab="",ylim=c(0,1.5),
     xaxs="i",yaxs="i")
```

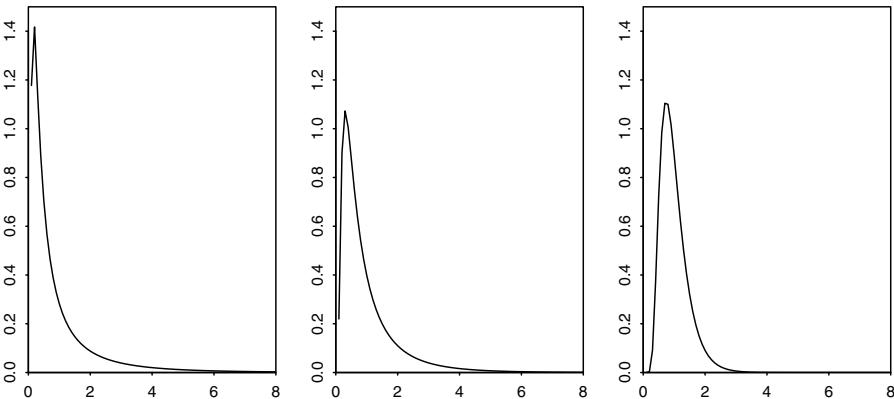


Figure 9.3 Inverse Gaussian densities for  $\lambda = 0.5$  on the left,  $\lambda = 1$  in the middle and  $\lambda = 5$  on the right.  $\mu = 1$  in all three cases.

The case of  $\mu = 1$  is known as the *Wald* distribution. The inverse Gaussian has found application in the modeling of lifetime distributions with nonmonotone failure rates and in the first passage times of Brownian motions with drift. See Seshadri (1993) for a book-length treatment.

Notice that the variance function for the inverse Gaussian GLM increases more rapidly with the mean than the gamma GLM, making it suitable for data where this occurs.

In Whitmore (1986), some sales data on a range of products is presented for the projected,  $x_i$ , and actual,  $y_i$ , sales for  $i = 1, \dots, 20$ . We consider a model,  $y_i = \beta x_i$  where  $\beta$  would represent the relative bias in the projected sales. Since the sales vary over a wide range from small to large, a normal error would be unreasonable because  $Y$  is positive and violations of this constraint could easily occur. We start with a look at the normal model:

```
data(cpd, package="faraway")
lmod <- lm(actual ~ projected-1, cpd)
summary(lmod)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
projected 0.9940    0.0172   57.9    <2e-16
plot(actual ~ projected, cpd)
abline(lmod)
```

Now consider the inverse Gaussian GLM where we must specify an identity link because we have  $y_i = \beta x_i$ :

```
igmod <- glm(actual ~ projected-1, family=inverse.gaussian(link=
  ↪ identity"), cpd)
summary(igmod)

Estimate Std. Error t value Pr(>|t|)
projected   1.1036     0.0614     18  2.2e-13

Dispersion parameter = 0.000
n = 20 p = 1
Deviance = 0.003 Null Deviance = Inf (Difference = Inf)
abline(igmod, lty=2)
```

We see that there is a clear difference in the estimates of the slope. The fits are shown in the first panel of Figure 9.4. We should check the diagnostics on the inverse

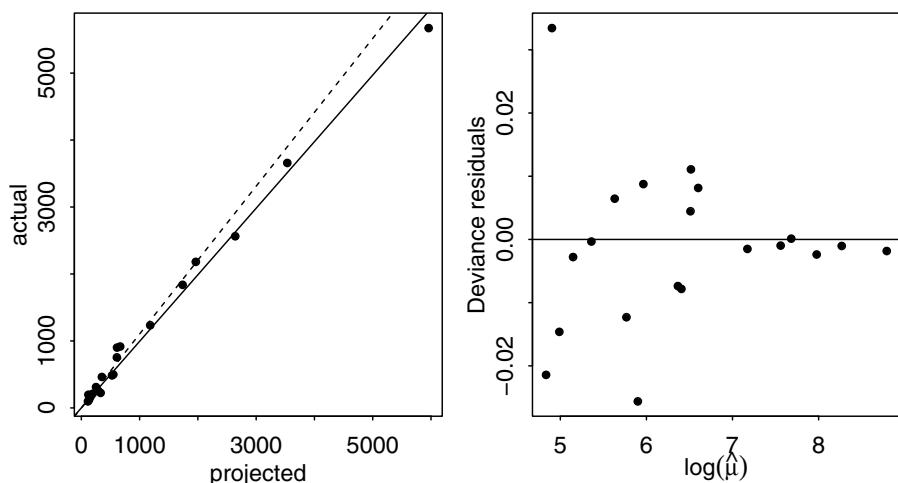


Figure 9.4 Projected and actual sales are shown for 20 products on the left. The linear model fit is shown as a solid line and the inverse Gaussian GLM fit is shown with a dotted line. A residual-fitted plot for the inverse Gaussian GLM is shown on the right.

Gaussian GLM:

```
plot(residuals(igmod) ~ log(fitted(igmod)), ylab="Deviance residuals",
  ↪ xlab=expression(log(hat(mu))))
```

We see in the second panel of Figure 9.4 that the variance of the residuals is decreasing with error, indicating that the inverse Gaussian variance function is too strong for this data. We have used  $\log(\hat{\mu})$  so that the points are more evenly spread horizontally making it easier, in this case, to see the variance relationship. A gamma GLM is a better choice here. In Whitmore (1986), a different variance function is used, but we do not pursue this here as this would not be a GLM.

### 9.3 Joint Modeling of the Mean and Dispersion

All the models we have considered so far have modeled the mean response  $\mu = EY$  where the variance takes a known form:  $\text{var } Y_i = \phi V(\mu_i)$  where the dispersion parameter  $\phi$  is the variance in the Gaussian model, the squared coefficient of variation in the gamma model and one in the binomial and Poisson models. We can generalize a little to allowing weights by letting  $\phi \equiv \phi_i = \phi w_i$  when the weights are known.

In this section, we are interested in examples where  $\phi_i$  varies with the covariates  $X$ . This is a particular issue that arises in industrial experiments. We wish to manufacture an item with a target mean or optimized response. We set the predictors to produce items as close as possible to the target mean or to optimize the mean. This requires a model for the mean. We would also prefer that the variance of the response be small at the chosen value of the predictors for production. So we need to model the variance as a function of the predictors.

We take, as an example, an experiment to determine which recipe will most reliably produce the best cake. The data comes from Box et al. (1988) and is shown in Table 9.1. The objective is to bake a cake reliably no matter how incompetent the

Design Vars			Environmental Vars					
F	S	E	T	0	-	+	-	+
t	0	-	-	+	+			
0	0	0		6.7	3.4	5.4	4.1	3.8
-	-	-		3.1	1.1	5.7	6.4	1.3
+	-	-		3.2	3.8	4.9	4.3	2.1
-	+	-		5.3	3.7	5.1	6.7	2.9
+	+	-		4.1	4.5	6.4	5.8	5.2
-	-	+		6.3	4.2	6.8	6.5	3.5
+	-	+		6.1	5.2	6.0	5.9	5.7
-	+	+		3.0	3.1	6.3	6.4	3.0
+	+	+		4.5	3.9	5.5	5.0	5.4

Table 9.1  $F$ =Flour,  $S$ =Shortening,  $E$ =Eggs,  $T$ =Oven temperature and  $t$ =Baking time. “+” indicates a higher-than-normal setting while “-” indicates a lower-than-normal setting. “0” indicates the standard setting.

cook. For this data, we can see, by examination, that a response of 6.8 is possible for lower flour and shortening content and higher egg content if the temperature is on the high side and the cooking time on the short side. However, we cannot be sure that the consumer will be able to set the temperature and cooking time correctly. Perhaps their oven is not correctly calibrated or they are just incompetent. If they happen to bake the cake for longer than the set time, they will produce a cake with a 3.5 rating. They will blame the product and not themselves and not buy that mix again. If on the other hand we produce the mix with high flour and eggs and low shortening, the worst the customer can do is a 5.2 and will do better than that for other combinations of time and temperature.

Here we need a combination of a high mean with respect to the design factors,

flour, eggs and shortening, and a low variance with respect to the environmental factors, temperature and time. In this example, the answer is easily seen by inspection, but usually more formal model fitting methods will be needed.

**Joint Model Specification:** We use the standard GLM approach for the mean:

$$EY_i = \mu_i \quad \eta_i = g(\mu_i) = \sum_j x_{ij}\beta_j \quad \text{var } Y_i = \phi_i V(\mu_i) \quad w_i = 1/\phi_i$$

Now the dispersion,  $\phi_i$ , is no longer considered fixed. Suppose we find an estimate,  $d_i$ , of the dispersion and model it using a gamma GLM:

$$Ed_i = \phi_i \quad \zeta_i = \log(\phi_i) = \sum_j z_{ij}\gamma_j \quad \text{var } d_i = \tau\phi_i^2$$

Notice the connection between the two models. The model for the mean produces the response for the model for the dispersion, which in turn produces the weights for the mean model. In principle, something other than a gamma GLM could be used for the dispersion, although since we wish to model a strictly positive, continuous and typically skewed dispersion, the gamma is the obvious choice. The dispersion predictors  $Z$  are usually a subset of the mean model predictors  $X$ .

For unreplicated experiments,  $r_P^2$  and  $r_D^2$  are two possible choices for  $d_i$ . If replications are available, then a more direct estimate of dispersion would be possible. For more details on the formulation, estimation and inference for these kinds of model, see McCullagh and Nelder (1989), Box and Meyer (1986), Bergman and Hynen (1997) and Nelder et al. (1998).

In the last three citations, data from a welding-strength experiment was analyzed. There were nine two-level factors and 16 unreplicated runs. Previous analyses have differed on which factors are significant for the mean. We found that two factors, Drying and Material, were apparently strongly significant, while the significance of others, including Preheating, was less clear. We fit a linear model for the mean using these three predictors:

```
data(weldstrength, package="faraway")
lmod <- lm(Strength ~ Drying + Material + Preheating, weldstrength)
summary(lmod)
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	43.625	0.262	166.25	< 2e-16
Drying	2.150	0.262	8.19	2.9e-06
Material	-3.100	0.262	-11.81	5.8e-08
Preheating	-0.375	0.262	-1.43	0.18

$n = 16$ ,  $p = 4$ , Residual SE = 0.525, R-Squared = 0.95

Following a suggestion of Smyth et al. (2001), we use the squared studentized residuals,  $(y_i - \hat{y}_i)^2/(1 - h_i)$ , as the response in the dispersion with a gamma GLM using a log-link and weights of  $1 - h_i$ . Again, we follow the suggestion of some previous authors as to which predictors are important for modeling the dispersion:

```
h <- influence(lmod)$hat
d <- residuals(lmod)^2/(1-h)
gmod <- glm(d ~ Material+Preheating, family=Gamma(link=log),
             ← weldstrength, weights=1-h)
```

Now feed back the estimated weights to the linear model:

```
w <- 1/fitted(gmod)
lmod <- lm(Strength ~ Drying + Material + Preheating, weldstrength,
           weights=w)
```

We now iterate until convergence, where we find that:

```
summary(lmod)
Estimate Std. Error t value Pr(>|t|)
(Intercept) 43.825     0.108   406.83 < 2e-16
Drying       1.869     0.045    41.53  2.5e-14
Material     -3.234     0.108   -30.03  1.2e-12
Preheating   -0.239     0.101    -2.35   0.036
```

$n = 16$ ,  $p = 4$ , Residual SE = 1.000, R-Squared = 1

We note that Preheating is now significant in contrast to the initial mean model fit.

The output for the dispersion model is:

```
summary(gmod)
Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.064     0.356   -8.60 0.0000010
Material     -3.037     0.413   -7.35 0.0000056
Preheating    2.904     0.413    7.03 0.0000089
```

overdispersion parameter = 0.500

$n = 16$   $p = 3$

Deviance = 20.943 Null Deviance = 57.919 (Difference = 36.976)

The standard errors are not correct in this output and further calculation, described in Smyth et al. (2001), would be necessary. This would result in somewhat larger standard errors (about twice the size), but the two factors would still be significant.

## 9.4 Quasi-Likelihood GLM

An examination of the fitting procedure for GLMs as described in Section 8.2 reveals that it is necessary only to know the link and variance functions to fit the model. This might be convenient since we may wish to minimize the assumptions by avoiding the specification of a distribution for the response. Even so, the distribution is necessary to compute the likelihood and hence the deviance. We need this for the inference. Fortunately, we can compute a substitute for the real likelihood, called the quasi-likelihood. This has already been demonstrated in Sections 3.5 and 5.2 where the quasi-likelihood method is introduced. We return to the idea again later in Section 13.2 in the context of dependent data.

We show here how the quasi method can help clarify the choice of the link and power function. Let's return to the wafer example of Section 9.1. We refit the gamma GLM model with just the main effects for simplicity:

```
g1 <- glm(resist ~ x1 + x2 + x3 + x4, family=Gamma(link=log), wafer)
```

The equivalent quasi version of this model is:

```
g1q <- glm(resist ~ x1 + x2 + x3 + x4, family=quasi(link=log, variance=
           ~ "mu^2"), wafer)
```

Notice that we need to specify both the link and the variance functions. We compare the deviance, degrees of freedom and log-likelihood for these two models. First the gamma GLM:

```
c(deviance(g1), df.residual(g1), logLik(g1))
[1] 0.12418 11.00000 -70.45392
```

and now the quasi GLM:

```
c(deviance(g1q), df.residual(g1q), logLik(g1q))
[1] 0.12418 11.00000 NA
```

The fitted coefficients, the deviance and the degrees of freedom are the same for both models. There is no log-likelihood for the quasi GLM because it is not a true likelihood fit. There's not much to recommend the quasi GLM in this instance as the additional assumption of the gamma distribution of the response makes the analysis easier. We can check this assumption so we are not being presumptuous.

The quasi GLM becomes more interesting if we vary the link function. We can consider link functions of the form  $\eta = \mu^p$  using the `power` function. We consider values of  $p$  in  $[0, 1]$ , where  $p = 0$  is considered as  $\log \mu$ . We fit a model for a range of powers and save the deviance obtained.

```
linkpow <- seq(0, 1, by=0.1)
devpow <- numeric(length(linkpow))
for(i in seq(along=linkpow)){
  glqq <- glm(resist ~ x1 + x2 + x3 + x4, family=quasi(link=power(
    ↪ linkpow[i]), variance="mu^2"), wafer)
  devpow[i] <- deviance(glqq)
}
```

A plot of the deviance against the power is shown in the first panel of Figure 9.5. We see that log is indeed the best choice. We would like to investigate  $p < 0$  but that facility is not available in the `power` function.

```
plot(linkpow, devpow, type="l", xlab="Link power", ylab="Deviance")
```

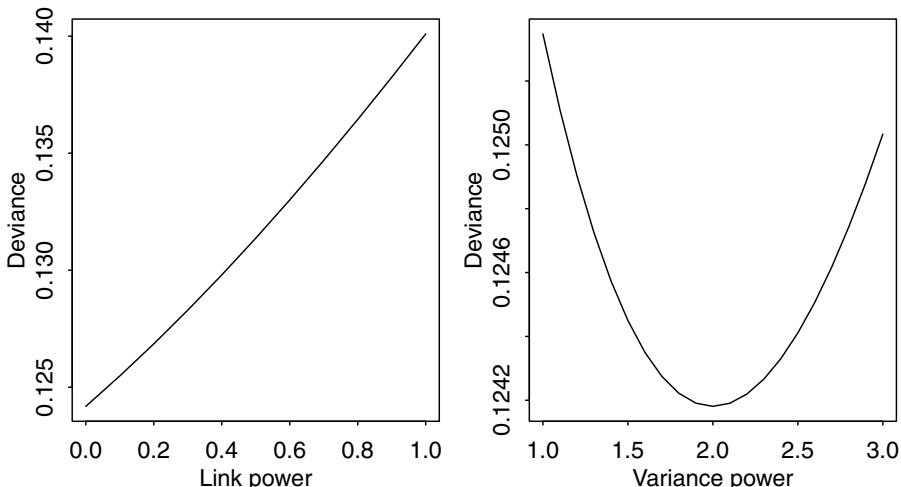


Figure 9.5 Deviance as the power of the link function is varied on the left and the variance function on the right.

We can also change the variance function. We might consider variance functions of the form  $V(\mu) = \mu^p$ . Again we can investigate how the fit of the model changes over a range of values of  $p$ , in this case in the range  $[1, 3]$ :

```
powfam <- quasi(link="log", variance="mu^2")
varpow <- seq(1, 3, by=0.1)
```

```
devpow <- numeric(length(varpow))
for(i in seq(along=varpow)){
  powfam[["variance"]] <- function(mu) mu^varpow[i]
  glqq <- glm(resist ~ x1 + x2 + x3 + x4, family=powfam, wafer)
  devpow[i] <- deviance(glqq)
}
```

The relationship is shown in the second panel of Figure 9.5. We see that  $p = 2$  is the best choice. Fortunately, that choice is implicit in the use of a gamma GLM so we are vindicated.

```
plot(varpow, devpow, type="l", xlab="Variance power", ylab="Deviance")
```

## 9.5 Tweedie GLM

In Section 8.1, we derived the mean and variance for an exponential (dispersion) family as

$$EY = \mu = b'(\theta) \quad \text{var } Y = b''(\theta)\phi$$

The variance function  $V(\mu) = b''(\theta)$  defines how the variance is connected to the mean. We can essentially define the distribution by specifying the variance function since  $\mu$  can be derived from this. Suppose we consider variance functions of the form  $V(\mu) = \mu^p$ . We have already seen four examples: the Gaussian ( $p = 0$ ), the Poisson ( $p = 1$ ), the gamma ( $p = 2$ ) and the inverse Gaussian ( $p = 3$ ).

We could take the quasi-likelihood approach and simply choose the link and the variance functions as demonstrated in the previous section. We can specify our chosen link and variance and use the `quasi` family in fitting the `glm()`. Unfortunately, only the four values of  $p = 0, 1, 2, 3$  are pre-programmed in base R. For these four choices, it is usually more convenient to use the one of the four distributional GLMs rather than the `quasi` model since we have the advantage of a full rather than quasi-likelihood. We could use other, perhaps noninteger values of  $p$  as demonstrated earlier but it is difficult to interpret such choices meaningfully.

Fortunately, any choice of  $p$ , except for the open interval  $(0, 1)$ , defines a so-called Tweedie distribution. Values of  $p$  in the interval  $(1, 2)$  are particularly interesting since these result in a compound Poisson-Gamma distribution. Observations from this distribution are generated as the sum of a Poisson number of gamma distributed variables in the following manner:

$$Y = \sum_{i=1}^N X_i, \quad N \sim \text{Poisson} \text{ and, independently, } X_i \sim \text{Gamma}$$

The mean of the Poisson is given by  $\mu^{2-p}/((2-p)\phi)$ . The gamma variables have shape parameter  $(2-p)/(p-1)$  and scale parameter  $\phi(p-1)\mu^{p-1}$ .

The interesting feature of this distribution is the possibility that  $N = 0$  and so  $Y = 0$ . This is useful for modeling responses where there is a nonzero probability that  $Y = 0$  but is otherwise a positive value. One example of this is insurance claims by customers that are often zero but can be positive. Rainfall amounts in a short period of time can often be zero but are otherwise some positive quantity. The zero inflated Poisson discussed in Section 5.5 also models excess zeroes but otherwise

the distribution is integer so may not be suitable. Further discussion of the Tweedie distribution may be found in Dunn and Smyth (2005).

The chredlin dataset concerns insurance redlining in Chicago in the 1970s. The response, involact, is the rate of usage of a default insurance plan available to those denied regular insurance in a given zip code. We want to relate this to five demographic predictors. A worked analysis of this data as a standard linear model can be found in Faraway (2014). The response has several zero values out of 47 total observations.

```
data(chredlin, package="faraway")
sum(chredlin$involact == 0)
[1] 15
```

This makes the assumption of normal distribution for the response somewhat problematic. We could try a Tweedie GLM but we have no particular preference for the value of  $p$ . We consider this an additional parameter that can be estimated by likelihood methods. An implementation is found in the mgcv package of Wood (2006):

```
library(mgcv)
twmod <- gam(involact ~ race+fire+theft+age+log(income), family=tw(
  ↪ link="log"), data=chicago)
summary(twmod)

Family: Tweedie(p=1.108)
Link function: log

Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.69565 9.10630 -1.50 0.14025
race 0.02159 0.00554 3.90 0.00035
fire 0.04850 0.01519 3.19 0.00270
theft -0.01022 0.00504 -2.03 0.04895
age 0.03458 0.00911 3.80 0.00048
log(income) 1.03099 0.93234 1.11 0.27526

R-sq. (adj) = 0.753 Deviance explained = 69.7%
-REML = 42.572 Scale est. = 0.30847 n = 47
```

The gam function is used instead of glm but otherwise the syntax is familiar. We are free to specify a link function and we make the default choice of log. It seems reasonable that the predictors would have a multiplicative rather than additive effect on the mean response. In the standard linear model, this would be achieved by logging the response, but that is not possible when some response values are zero. We avoid that obstacle by using the log in the link. We see that the estimated value of  $p$  is 1.108. Otherwise, the general significance and direction of the effect of the predictors and the quality of the fit are quite similar to that achieved by the linear model.

Consider the first zip code in the dataset. What is the predicted distribution for this case? We set up a grid to compute the density and extract the  $p$  and  $\phi$  values:

```
xgrid <- seq(1e-10, 1.25, len=100)
p <- 1.108
phi <- 0.30847
```

The expected mean response for the first case is:

```
(mu <- twmod$fit[1])
[1] 0.17818
```

From this we can compute the mean of the Poisson as

```
(poismean <- mu^(2-p) / (phi * (2-p)))
[1] 0.78018
```

which means the probability of a zero response is:

```
(p0 <- exp(-poismean))
[1] 0.45832
```

Now we compute the density for the nonzero part of the response, scale it down for the probability of a nonzero response and plot. We add a line segment to denote the point mass at zero and show the result in Figure 9.6.

```
twden <- exp(1dTweedie(xgrid, mu, p=p, phi=phi) [,1])
plot(xgrid, twden*(1-p0), type="l", xlab="x", ylab="Density")
dmax <- max(twden*(1-p0))
segments(0, 0, 0, dmax, lwd=2)
text(0.05, dmax, paste0("p=", signif(p0, 3)))
```

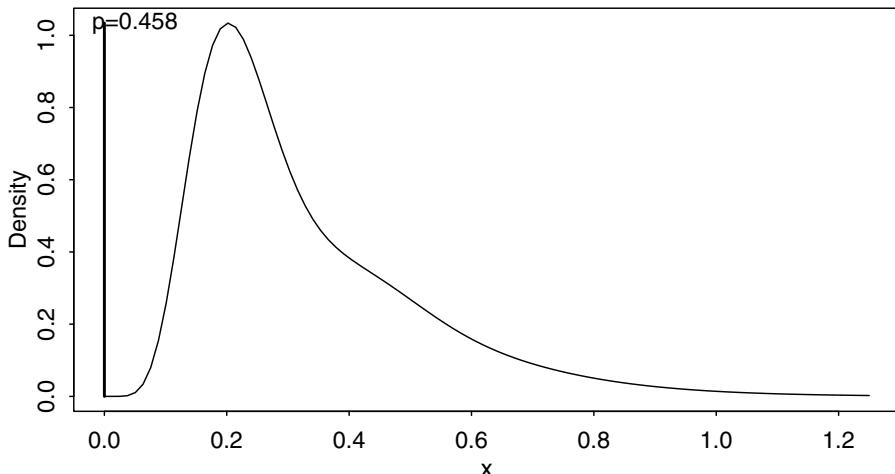


Figure 9.6 Predictive density for the first zip code based on the fitted Tweedie GLM. Response is zero with probability 0.458 but otherwise takes a continuous positive value.

## Exercises

1. The relationship between corn yield (bushels per acre) and nitrogen (pounds per acre) fertilizer application was studied in Wisconsin. The data may be found in `cornnit`.
  - (a) Make a plot of the data and comment on the relationship. Do you think the effect of nitrogen on yield is likely to be additive or multiplicative? Do you think yield will always increase with more nitrogen or will there be an optimum application of nitrogen?
  - (b) Using a (Gaussian) linear model, represent the relationship between the yield response and the nitrogen predictor taking into account your answers to the previous questions. Check the residual vs. fitted plot and the QQ plot and comment on any anomalies.
  - (c) Remove any outliers and repeat the model diagnostics.

- (d) At what level of nitrogen does the maximum yield occur according to your model? Find a 95% confidence interval for the expected yield at this dose of nitrogen.
- (e) Now develop a GLM for the data that does not (explicitly) transform the response. Pick a form for this model which is analogous to the Gaussian linear model just fitted. Check the same diagnostics as before.
- (f) Remove any outliers and repeat the model diagnostics.
- (g) Find the optimum level of nitrogen and form a 95% confidence interval for the expected yield.
- (h) Make a plot of the data that shows the fitted curves from both models.
2. An experiment was conducted as part of an investigation to combat the effects of certain toxic agents. The survival time of rats depended on the type of poison used and the treatment applied. The data is found in `rats`.
- (a) Make plots of the data and comment on differences between the treatments and poisons.
  - (b) Fit a linear model with an interaction between the two predictors. Use the Box-Cox method to determine an optimal transformation on the response. Can this optimal transformation be rounded to a more interpretable function?
  - (c) Refit the model using your chosen interpretable transformation. Check the standard diagnostics and report any problems.
  - (d) Is the interaction term significant? Simplify the model if justified and indicate which poison and treatment will result in the shortest survival time.
  - (e) Build an inverse Gaussian GLM for this data. Select an appropriate link function consistent with the previous modeling. Perform diagnostics to verify your choice. Which poison and treatment will result in the shortest survival time?
  - (f) Compare the predicted values on the original scale of the response for the two models. Do the two models achieve a similar fit?
3. Components are attached to an electronic circuit card assembly by a wave-soldering process. The soldering process involves baking and preheating the circuit card and then passing it through a solder wave by conveyor. Defects arise during the process. The data is found in `wavesolder`.
- (a) Make boxplots of the number of defects for each of the predictors and comment on the relationships seen.
  - (b) Compute the mean and variance within each group of three replicates. Plot the variance against the mean. What is the relation?
  - (c) Fit a quasi-poisson model for the mean number of defects using the main effects of all the predictors. Which predictors seem to be significant and which not? Make a test to check whether the insignificant predictors can be removed from the model.
  - (d) Fit a gamma GLM with a log link for the variance response using all the main effects of the predictors. What problem arises?

- (e) Now fit a standard linear model to the log of the variance. Check whether the same set of predictors can be removed from this model as previously.
- (f) Try fitting the gamma GLM with a log link for the variances, now using a reduced set of predictors as determined from the previous question. Compare the fitted values from both models of the variance. Are they close?
- (g) Examine the final models for the mean and variance. State the qualitative effect of the chosen predictors on the mean and variance.
- (h) Compare the fitted values from the model for the mean with those for the variance. What is the relationship? Are the fitted variances consistent with what might be expected given a quasi-Poisson model?
4. Data were collected from 39 students in a University of Chicago MBA class and presented in `happy`. Happiness was measured on a 10 point scale.
- (a) Plot the happiness response on each of the four predictors. Comment on the relationships.
  - (b) Rescale the response and fit an appropriate quasi-likelihood model. Which predictors are not statistically significant in your model?
  - (c) Consider dropping each of the predictors from the full model using an *F*-test to check the significance of each predictor. Remove the least significant predictor and repeat the consideration of single predictors relative to the current model. Iterate as necessary.
  - (d) Make a plot of the residuals and predicted values from your selected model. Interpret it. Which subject is the best example of someone happier than s/he deserves to be?
  - (e) Using your model, indicate how much familial income would compensate for a two level drop in `love`.
5. The `leafblotch` data shows the percentage leaf area affected by leaf blotch on 10 varieties of barley at nine different sites. The data comes from Wedderburn (1974).
- (a) Make plots of the blotch response against the two predictors. Describe the relationships seen.
  - (b) Fit a binomial GLM with blotch as the response and site and variety as predictors. Find the deviance and degrees of freedom and comment on what this says about the fit of this model.
  - (c) Fit a quasi-binomial model and report the value of the dispersion parameter. Could this dispersion estimate be derived from the binomial GLM? Demonstrate how if possible.
  - (d) Plot the residuals against the predicted values. Interpret the plot.
  - (e) A better variance function is  $\mu^2(1 - \mu)^2$  and yet this is not one of the available choices in R. However, the effect may be obtained by the appropriate use of weighting. Define weights as a function of  $\mu$  that, when used in conjunction with a variance function of  $\mu(1 - \mu)$ , achieve the effect of a  $\mu^2(1 - \mu)^2$  variance function. Repeat the plot of the residuals and comment.

- (f) Test the significance of each of the predictors relative to the full model using an  $F$ -test.
- (g) There may be an interaction between sites and varieties although this is difficult to check. Which site and variety combination shows the greatest indication of an interaction?
6. Consider the `orings` data regarding space shuttle flights prior to the Challenger disaster from Chapter 3.
- Fit a quasi-likelihood model with an identity link to the data. You will need to rescale the response to  $[0, 1]$ . You should use the usual binomial variation. Why does the model fitting fail?
  - Now try fitting the same model but now setting the starting values for the regression coefficients using the `start` argument. You will need to make a valid choice for the coefficients. You may also find it helpful to set the argument `trace=TRUE` to check the convergence. What happens?
  - Examine the summary output of your fitted model and determine the temperature above which the predicted probability of failure would be invalid.
  - Fit a quasi-likelihood model but with the usual logit link. Compare this model to the identity link model. Determine which model is superior.
  - Display curves showing the predicted probabilities as temperature varies for both models. Show these curves on a single plot, superimposed on the data.
  - Compute the predicted probability of failure at  $31^{\circ}\text{F}$  for both models. Compare the results.
7. Fit the `orings` data with a binomial response and a logit link as in Chapter 3.
- Construct the appropriate test statistic for testing the effect of the temperature. State the appropriate null distribution and give the  $p$ -value.
  - Generate data under the null distribution for the previous test. Use the `rbinom` function with the average proportion of damaged O-rings. Recompute the test statistic and compute the  $p$ -value.
  - Repeat the process of the previous question 1000 times, saving the test statistic each time. Compare the empirical distribution of these simulated test statistics with the nominal null distribution stated in the first part of this question.
  - Compare the critical values for a 5% level test computed using these two methods.
8. One hundred twenty-five fruitflies were divided randomly into five groups of 25 each. The response was the longevity of the fruitfly in days. One group was kept solitary, while another was kept individually with a virgin female each day. Another group was given eight virgin females per day. As an additional control the fourth and fifth groups were kept with one or eight pregnant females per day. Pregnant fruitflies will not mate. The thorax length of each male was measured as this was known to affect longevity. The data is presented in `fruitfly`.
- Make a plot of the data.

- (b) Fit a standard linear model with longevity as the response and thorax length and activity as predictors, including an interaction term. Remove the interaction term if it is not significant.
  - (c) Plot the residuals and fitted values. Interpret the plot.
  - (d) Transform the response in an appropriate way and recheck the residual plot.
  - (e) Fit a gamma GLM with an appropriate link. Plot the residuals against the predicted values. Interpret the plot.
  - (f) Consider a fruit fly with mean thorax length in the isolated activity group. Use the fit from the model to plot the gamma density for the lifetime of such a fly.
9. The truck data concerns an experiment to optimize the production of leaf springs for trucks. A heat treatment is designed so that the free height of the spring should come as close to eight inches as possible. We can vary five factors at two levels each. A  $2^{5-1}$  fractional factorial experiment with three replicates was carried out. The data comes from Pignatiello and Ramberg (1985).
- (a) Plot the data and comment on the effect of the factors on the response.
  - (b) Fit a model with all two-way interactions for the height response. Why is it that not all two-way terms can be estimated?
  - (c) Use `step()` to choose a smaller model. What is the largest interaction term?
  - (d) Compute the predicted height for each of the 16 possible combinations of the factors. Order them in closeness to the target of 8 inches.
  - (e) Compute the variance in each group of three replicates. Use a gamma GLM with a log link to model the variances using all two-way interactions. Use `step()` to select a model. Comment on the factors which have an effect on the variance.
  - (f) Compute the predicted variance for each of the 16 combinations and display them in increasing order. Which combination produces the lowest variation?
  - (g) Discuss the problem of selecting a combination that comes close to the target of 8 inches while minimizing the variation.

---

## Chapter 10

# Random Effects

---

Grouped data arise in almost all areas of statistical application. Sometimes the grouping structure is simple, where each case belongs to a single group and there is only one grouping factor. More complex datasets have a hierarchical or nested structure or include longitudinal or spatial elements. Sometimes the grouping arises because the same individual is measured repeatedly or sometimes each individual is measured once only but these individuals have some form of group structure. We defer examination of the repeated measurement of individuals to the next chapter, although the statistical methodology used is the same.

All such data share the common feature of correlation of observations within the same group and so analyses that assume independence of the observations will be inappropriate. The use of random effects is one common and convenient way to model such grouping structure.

A *fixed effect* is an unknown constant that we try to estimate from the data. Almost all the parameters used in the linear and generalized linear models we have presented earlier in this book are fixed effects. In contrast, a *random effect* is a random variable. It does not make sense to estimate a random effect; instead, we try to estimate the parameters that describe the distribution of this random effect.

Consider an experiment to investigate the effect of several drug treatments on a sample of patients. Typically, we are interested in specific drug treatments and so we would treat the drug effects as fixed. However, it makes most sense to treat the patient effects as random. It is often reasonable to treat the patients as being randomly selected from a larger collection of patients whose characteristics we would like to estimate. Furthermore, we are not particularly interested in these specific patients, but in the whole population of patients. A random effects approach to modeling effects is more ambitious in the sense that it attempts to say something about the wider population beyond the particular sample. Blocking factors can often be viewed as random effects, because these often arise as a random sample of those blocks potentially available.

There is some judgment required in deciding when to use fixed and when to use random effects. Sometimes the choice is clear, but in other cases, reasonable statisticians may differ. In some analyses, random effects are used simply to induce a certain correlation structure in the data and there is sense in which the chosen levels represent a sample from a population. Gelman (2005) remarks on the variety of definitions for random effects and proposes a particular straightforward solution to the dilemma of whether to use fixed or random effects — he recommends always using random effects.

A *mixed effects* model has both fixed and random effects. A simple example of such a model would be a two-way analysis of variance (ANOVA):

$$y_{ijk} = \mu + \tau_i + v_j + \varepsilon_{ijk}$$

where the  $\mu$  and  $\tau_i$  are fixed effects and the error  $\varepsilon_{ijk}$  and the random effects  $v_j$  are independent and identically distributed  $N(0, \sigma^2)$  and  $N(0, \sigma_v^2)$ , respectively.

We would want to estimate the  $\tau_i$  and test the hypothesis  $H_0 : \tau_i = 0 \forall i$  while we would estimate  $\sigma_v^2$  and might test  $H_0 : \sigma_v^2 = 0$ . Notice the difference: we need to estimate and test several fixed effect parameters while we need only estimate and test a single random effect parameter.

In the following sections, we consider estimation and inference for mixed effects models and then illustrate the application to several common designs.

## 10.1 Estimation

This is not as simple as it was for fixed effects models, where least squares is an easily applied method with many good properties. Let's start with the simplest possible random effects model: a one-way ANOVA design with a factor at  $a$  levels:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, \dots, a \quad j = 1, \dots, n$$

where the  $\alpha$ s and  $\varepsilon$ s have mean zero, but variances  $\sigma_\alpha^2$  and  $\sigma_\varepsilon^2$ , respectively. These variances are known as the variance components. Notice that this induces a correlation between observations at the same level equal to:

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}$$

This is known as the *intraclass correlation coefficient* (ICC). In the limiting case when there is no variation between the levels,  $\sigma_\alpha = 0$  so then  $\rho = 0$ . Alternatively, when the variation between the levels is much larger than that within the levels, the value of  $\rho$  will approach 1. This illustrates how random effects generate correlations between observations.

For simplicity, let there be an equal number of observations  $n$  per level. We can decompose the variation as follows (where dot in the subscript indicates the average over that index):

$$\sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^a \sum_{j=1}^n (\bar{y}_i - \bar{y}_{..})^2$$

or  $SST = SSE + SSA$ , respectively. SSE is the residual sum of squares, SST is the total sum of squares (corrected for the mean) and SSA is the sum of squares due to  $\alpha$ . These quantities are often displayed in an ANOVA table along with the degrees of freedom associated with each sum of squares. Dividing through by the respective degrees of freedom, we obtain the mean squares, MSE and MSA. Now we find that:

$$E(SSE) = a(n-1)\sigma_\varepsilon^2, \quad E(SSA) = (a-1)(n\sigma_\alpha^2 + \sigma_\varepsilon^2)$$

which suggests using the estimators:

$$\hat{\sigma}_\epsilon^2 = SSE/(a(n-1)) = MSE, \quad \hat{\sigma}_\alpha^2 = \frac{SSA/(a-1) - \hat{\sigma}_\epsilon^2}{n} = \frac{MSA - MSE}{n}$$

This method of estimating variance components can be used for more complex designs. The ANOVA table is constructed, the expected mean squares calculated and the variance components obtained by solving the resulting equations. These estimators are known as *ANOVA estimators*. These were the first estimators developed for variance components. They have the advantage of taking explicit forms suitable for hand calculation which was important in precomputing days, but they have a number of disadvantages:

1. The estimates can take negative values. For example, in our situation above, if  $MSA < MSE$  then  $\hat{\sigma}_\alpha^2 < 0$ . This is rather embarrassing since variances cannot be negative. Various fixes have been proposed, but these all take away from the original simplicity of the estimation method.
2. A balanced design has an equal number of observations per cell, where cell is defined as the finest subdivision of the data according to the factors. In such circumstances, the ANOVA decomposition into sums of squares is unique. For unbalanced data, this is not true and we must choose which ANOVA decomposition to use, which will in turn affect the estimation of the variance components. Various rules have been suggested about how the decomposition should be done, but none of these have universal appeal.
3. The need for complicated algebraic calculations. Formulae for the simpler models are easy to find and coded in software. More complex models will require difficult and opaque constructions.

We would like a method that would avoid negative variances, work unambiguously for unbalanced data and that can be applied in a transparent and straightforward manner. Maximum likelihood (ML) estimation satisfies these requirements. This does require that we assume some distribution for the errors and the random effects. The usual assumption is normality; ML would work for other distributions, but these are rarely considered in this context.

For a fixed effect model with normal errors, we can write:

$$y = X\beta + \epsilon \quad \text{or} \quad y \sim N(X\beta, \sigma^2 I)$$

where  $X$  is an  $n \times p$  model matrix and  $\beta$  is a vector of length  $p$ . We can generalize to a mixed effect model with a vector  $\gamma$  of  $q$  random effects with associated model matrix  $Z$  which has dimension  $n \times q$ . Then we can model the response  $y$ , given the value of the random effects as:

$$y = X\beta + Z\gamma + \epsilon \quad \text{or} \quad y|\gamma \sim N(X\beta + Z\gamma, \sigma^2 I)$$

If we further assume that the random effects  $\gamma \sim N(0, \sigma^2 D)$  then  $\text{var } y = \text{var } Z\gamma + \text{var } \epsilon = \sigma^2 ZDZ^T + \sigma^2 I$  and we can write the unconditional distribution of  $y$  as:

$$y \sim N(X\beta, \sigma^2(I + ZDZ^T))$$

If we knew  $D$ , then we could estimate  $\beta$  using generalized least squares; see, for example, Chapter 8 in Faraway (2014). However, the estimation of the variance components,  $D$ , is often one purpose of the analysis. Standard maximum likelihood is one method of estimation that can be used here. If we let  $V = I + ZDZ^T$ , then the joint density for the response is:

$$\frac{1}{2\pi^{n/2}|\sigma^2 V|^{1/2}} e^{-\frac{1}{2\sigma^2}(y-X\beta)^T V^{-1}(y-X\beta)}$$

so that the log-likelihood for the data is:

$$l(\beta, \sigma, D | y) = -\frac{n}{2} \log 2\pi - \frac{1}{2} \log |\sigma^2 V| - \frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta)$$

This can be optimized to find maximum likelihood estimates of  $\beta$ ,  $\sigma^2$  and  $D$ . This is straightforward in principle, but there may be difficulties in practice. More complex models involving larger numbers of random effects parameters can be difficult to estimate. Sometimes the MLE of a variance parameter may be zero which occurs on the boundary of its domain. The derivative of the likelihood may not be zero in this boundary state which causes problems for many optimization methods.

Standard errors can be obtained using the usual large sample theory for maximum likelihood estimates. The variance can be estimated using the inverse of the negative second derivative of the log-likelihood evaluated at the MLE.

MLEs have some drawbacks. One particular problem is that they are biased. For example, consider an i.i.d. sample of normal data  $x_1, \dots, x_n$ , then the MLE is:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

A denominator of  $n - 1$  is needed for an unbiased estimator. Similar problems occur with the estimation of variance components. Given that the number of levels of a factor may not be large, the bias of the MLE of the variance component associated with that factor may be quite large. *Restricted maximum likelihood* (REML) estimators are an attempt to get round this problem. The idea is to find all independent linear combinations of the response,  $k$ , such that  $k^T X = 0$ . Form matrix  $K$  with columns  $k$ , so that:

$$K^T y \sim N(0, K^T V K)$$

We can then proceed to maximize the likelihood based on  $K^T y$  which does not involve any of the fixed effect parameters. Once the random effect parameters have been estimated, it is simple enough to obtain the fixed effect parameter estimates. REML generally produces less biased estimates. For balanced data, the REML estimates are usually the same as the ANOVA estimates.

We illustrate the fitting methods using some data from an experiment to test the paper brightness depending on a shift operator described in Sheldon (1960). We start with a fixed effects one-way ANOVA:

```
data(pulp, package="faraway")
op <- options(contrasts=c("contr.sum", "contr.poly"))
lmod <- aov(bright ~ operator, pulp)
summary(lmod)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
operator	3	1.340	0.447	4.2	0.023
Residuals	16	1.700	0.106		

We can plot the data as seen in Figure 10.1.

```
library(ggplot2)
ggplot(pulp, aes(x=operator, y=bright))+geom_point(position = position_jitter(width=0.1, height=0.0))
```

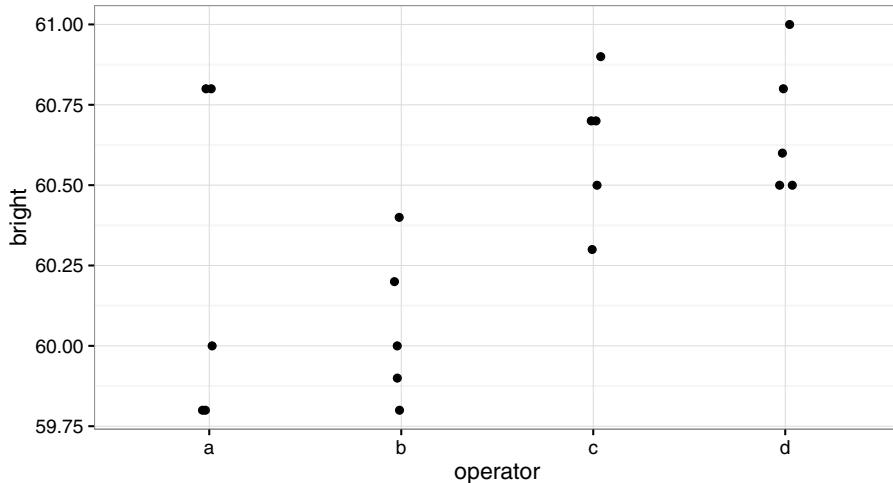


Figure 10.1 Paper brightness varying by operator. Some jittering has been used to make co-incident points apparent.

```
coef(lmod)
(Intercept) operator1 operator2 operator3
  60.40      -0.16     -0.34      0.22
options(op)
```

We have specified sum contrasts here instead of the default treatment contrasts to make the later connection to the corresponding random effects clearer. The `aov` function is just a wrapper for the standard `lm` function that produces results more appropriate for ANOVA models. We see that the operator effect is significant with a *p*-value of 0.023. The estimate of  $\sigma^2$  is 0.106 and the estimated overall mean is 60.4. For sum contrasts,  $\sum \alpha_i = 0$ , so we can calculate the effect for the fourth operator as  $0.16 + 0.34 - 0.22 = 0.28$ .

Turning to the random effects model, we can compute the variance of the operator effects,  $\sigma_{\alpha}^2$ , using the formula above as:

```
(0.447-0.106)/5
[1] 0.0682
```

Now we demonstrate the maximum likelihood estimators. The original R package for fitting mixed effects models was `nlme` as described in Pinheiro and Bates (2000). Subsequently Bates (2005) introduced the package `lme4`. The syntax for these two packages is somewhat different. The `lme4` package is generally more capable especially for larger datasets. The estimates these two packages produce for smaller, simpler datasets as considered in this chapter will generally be the same. However,

there are some crucial differences in the approach to inference that we will discuss later. We use the `lme4` packages in preference to `nlme`:

```
library(lme4)
mmod <- lmer(bright ~ 1+(1|operator), pulp)
summary(mmod)
```

Linear mixed model fit by REML ['lmerMod']  
 Formula: bright ~ 1 + (1 | operator)  
 Data: pulp

REML criterion at convergence: 18.6

Scaled residuals:

Min	1Q	Median	3Q	Max
-1.467	-0.759	-0.124	0.628	1.601

Random effects:

Groups	Name	Variance	Std.Dev.
operator	(Intercept)	0.0681	0.261
Residual		0.1062	0.326

Number of obs: 20, groups: operator, 4

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	60.400	0.149	404

The model has fixed and random effects components. The fixed effect here is just the intercept represented by the first 1 in the model formula. The random effect is represented by `(1|operator)` indicating that the data is grouped by `operator` and the 1 indicating that the **random effect is constant within each group**. The parentheses are necessary to ensure that expression is parsed in the correct order.

The default fitting method is REML. We see that this gives identical estimates to the ANOVA method above —  $\hat{\sigma}^2 = 0.106$ ,  $\hat{\sigma}_{\alpha}^2 = 0.068$  and  $\hat{\mu} = 60.4$ . For unbalanced designs, the REML and ANOVA estimators are not necessarily identical. The standard deviations are simply the square roots of the variances and not estimates of the uncertainty in the variances.

As with the GLM summary output, we find it rather verbose and prefer our own abbreviated version (which is adapted from the `display()` function in the `arm` package of Gelman and Su (2013)):

```
sumary(mmod)
```

Fixed Effects:

coef.est	coef.se
60.40	0.15

Random Effects:

Groups	Name	Std.Dev.
operator	(Intercept)	0.26
Residual		0.33

---

number of obs: 20, groups: operator, 4  
 AIC = 24.6, DIC = 14.4  
 deviance = 16.5

This output contains just the information we need. It is better to use standard devi-

ations rather than variances as the former are measured in the units of the response and so much easier to interpret.

The maximum likelihood estimates may also be computed:

```
smod <- lmer(bright ~ 1+(1|operator), pulp, REML=FALSE)
summary(smod)

Fixed Effects:
coef.est  coef.se
 60.40     0.13

Random Effects:
 Groups   Name        Std.Dev.
 operator (Intercept) 0.21
 Residual            0.33
---
number of obs: 20, groups: operator, 4
AIC = 22.5, DIC = 16.5
deviance = 16.5
```

The between-subjects SD, 0.21, is smaller than with the REML method as the ML method biases the estimates towards zero. The fixed effects are unchanged.

## 10.2 Inference

**Test Statistic:** We follow a general procedure. Decide which component(s) of the model you wish to test. These can be fixed and/or random effects. Specify two models: a null  $H_0$  which does not contain your specified component(s) and an alternative  $H_1$  which does include your component(s). The other terms in the models must be the same. These other terms (usually) make a difference to the result and must be chosen with care.

Using standard likelihood theory, we may derive a test to compare two nested hypotheses,  $H_0$  and  $H_1$ , by computing the likelihood ratio test statistic:

$$2(l(\hat{\beta}_1, \hat{\sigma}_1, \hat{D}_1 | y) - l(\hat{\beta}_0, \hat{\sigma}_0, \hat{D}_0 | y))$$

where  $\hat{\beta}_0, \hat{\sigma}_0, \hat{D}_0$  are the MLEs of the parameters under the null hypothesis and  $\hat{\beta}_1, \hat{\sigma}_1, \hat{D}_1$  are the MLEs of the parameters under the alternative hypothesis.

If you plan to use the likelihood ratio test to compare two nested models that differ only in their fixed effects, you cannot use the REML estimation method. The reason is that REML estimates the random effects by considering linear combinations of the data that remove the fixed effects. If these fixed effects are changed, the likelihoods of the two models will not be directly comparable. Use ordinary maximum likelihood in this situation if you also wish to use the likelihood ratio test.

**Approximate Null Distribution:** This test statistic is approximately chi-squared with degrees of freedom equal to the difference in the dimensions of the two parameters spaces (the difference in the number of parameters when the models are identifiable). Unfortunately, this test is not exact and also requires several assumptions — see a text such as Cox and Hinkley (1974) for more details. Serious problems can arise with this approximation.

One crucial assumption is that the parameters under the null are not on the boundary of the parameter space. Since we are often interested in testing hypotheses about

the random effects that take the form  $H_0 : \hat{\sigma}^2 = 0$ , this is a common problem which makes the asymptotic inference invalid. If you do use the  $\chi^2$  distribution with the usual degrees of freedom, then the test will tend to be conservative — the  $p$ -values will tend to be larger than they should be. This means that if you observe a significant effect using the  $\chi^2$  approximation, you can be fairly confident that it is actually significant. The  $p$ -values generated by the likelihood ratio test for fixed effects are also approximate and unfortunately tend to be too small, thereby sometimes overstating the importance of some effects.

Regrettably the  $p$ -value based on the  $\chi^2$  approximation can either be entirely or just somewhat wrong. Perhaps with sufficient data and favorable models, the approximation may be satisfactory but it is difficult to say exactly when such propitious conditions may arise. Hence the safest advice is to not use this approximation.

**Expected mean squares:** Another method of hypothesis testing is based on the sums of squares found in the ANOVA decompositions. These tests are sometimes more powerful than their likelihood ratio test equivalents. However, the correct derivation of these tests usually requires extensive tedious algebra that must be recalculated for each type of model. Furthermore, the tests cannot be used (at least without complex and unsatisfactory adjustments) when the experiment is unbalanced. This method only works for simple models and balanced data.

**$F$ -tests for fixed effects:** We might try to use the  $F$ -test used in standard linear models to perform hypothesis tests regarding the fixed effects. The  $F$ -statistic is based on residual sums of squares and degrees of freedom as described in Chapter 3 of Faraway (2014). This is the method used in the `nlme` package. In the standard linear model setting, provided the normality assumption is correct, the null distribution has an exact  $F$ -distribution. Unfortunately, problems arise in transferring this method to mixed effect models. Firstly, the definition of degrees of freedom becomes murky in the presence of random effect parameters. Secondly, the test statistic is not necessarily  $F$ -distributed.

For some simple models with balanced data, the  $F$ -test is correct but in other cases with more complex models or unbalanced data, the  $p$ -values can be substantially incorrect. It is difficult to specify exactly when this test may be relied upon. For this reason, the `lme4` now declines to state  $p$ -values. Furthermore, the  $t$ -statistics that one might generate to test or form a confidence interval for a single fixed effect parameter also rely on the same problematic approximations.

**Strategies for inference:** We have good test statistics in the likelihood ratio test (LRT) or  $F$ -statistic but as yet no universally reliable way to obtain a null distribution. One solution would be to ignore the possible problem and use either the `nlme` package or the `lmerTest` package (which restores the questionable  $p$ -values to `lme4`). In certain known simple models with balanced data, this will produce accurate results but it would be speculative to report such results in other situations without at least verifying the results using other methods. A number of alternatives exist.

The standard degrees of freedom for the  $F$ -statistic in mixed models are not always reliable. Various researchers have developed methods for adjusting these degrees of freedom. One popular method is due to Kenward and Roger (1997). We will illustrate the use of this method later in this chapter. Even if the adjustment is opti-

mal, there remains the problem that the null distribution may not be  $F$ . Furthermore, the method is relevant only for the testing of fixed effects.

We can use bootstrap methods to find more accurate  $p$ -values for the likelihood ratio test. The usual bootstrap approach is nonparametric in that no distribution is assumed. Since we are willing to assume normality for the errors and the random effects, we can use a technique called the *parametric bootstrap*. We generate data under the null model using the fitted parameter estimates. We compute the likelihood ratio statistic for this generated data. We repeat this many times and use this to judge the significance of the observed test statistic. This approach will be demonstrated below. The problem may also be addressed by using Bayesian methods to fit the models. We discuss these in Chapter 12.

**Model Selection:** For comparing larger numbers of models, it is unwise to take a testing-based approach to selection. The problems are similar to those encountered in model selection for standard linear models. When the number of models considered becomes more than a handful, the issue of multiple testing arises and  $p$ -values lose their normal meaning. Instead it is better to take a criterion-based approach to model selection. Although we can develop the ideas of model selection of linear models and extend them to linear mixed models, there are some important additional difficulties which means that this extension is not straightforward. Firstly, the dependent response means that effective sample size is less than the total number of cases. Secondly, we have two kinds of parameters, some for the fixed effects and some for the random effects. It is not clear how these two types of parameters should be counted together. Thirdly, most criteria are based on the likelihood which does not behave well at the boundary of the parameter space as can occur with variance parameters.

The Akaike Information Criterion (AIC) and its variations are the most popular model selection criterion. In the `lme4` package, AIC is defined as:

$$-2(\max \log \text{likelihood}) + 2p$$

where  $p$  is the total number of parameters. We can confidently use this criterion to compare models which differ only in their fixed effects, as the number of random effect parameters will be the same for all models considered. If we compare models where the random effects are also varied, then we must think more carefully about how to count the random effect parameters. This is problematic due to the aforementioned boundary problems.

Other criteria can be considered. The Bayes Information Criterion (BIC) replaces the  $2p$  in the AIC with  $p \log n$  and tends to prefer smaller models to the AIC. Another popular criterion used with mixed effect models is the Deviance Information Criterion (DIC) of Spiegelhalter et al. (2002). This criterion is more suited to the Bayesian models discussed in Chapter 12. For a discussion of model selection criteria, see Section A.3. For the specific application to linear mixed models, see Müller et al. (2013). For most of the examples considered in this chapter, there are only a few variables so we are able to rely on testing methods to choose between just a few models. We defer an example of using these methods to Section 10.10.

**Example:** Now let's demonstrate these inferential methods on the `pulp` data. The fixed effect analysis shows that the operator effects are statistically significant

with a  $p$ -value of 0.023. A random effects analysis using the expected mean squares approach yields exactly the same  $F$ -statistic for the one-way ANOVA. This method works exactly for such a simple model.

We can also employ the likelihood ratio approach to test the null hypothesis that the variance between the operators is zero. In the fixed effects model, we tested the hypothesis that the four operators had the same effect. In the mixed effect model where the operators are treated as random, the hypothesis that this variance is zero claims that there is no differences between operators in the population. This is a stronger claim than the fixed effect model hypothesis about just the four chosen operators.

We first fit the null model:

```
nullmod <- lm(bright ~ 1, pulp)
```

As there are no random effects in this model, we must use `lm`. For models of the same class, we could use `anova` to compute the LRT and its  $p$ -value. Here, we need to compute this directly:

```
lrtstat <- as.numeric(2*(logLik(smod)-logLik(nullmod)))
pvalue <- pchisq(lrtstat,1,lower=FALSE)
data.frame(lrtstat, pvalue)
```

```
      lrtstat    pvalue
1 2.5684 0.10902
```

The  $p$ -value is now well above the 5% significance level. We cannot say that this result is necessarily wrong, but the use of the  $\chi^2$  approximation does cause us to doubt the result.

We can use the parametric bootstrap approach to obtain a more accurate  $p$ -value. We need to estimate the probability, given that the null hypothesis is true, of observing an LRT of 2.5684 or greater. Under the null hypothesis,  $y \sim N(\mu, \sigma^2)$ . A simulation approach generates data under this model, fits the null and alternative models and computes the LRT statistic. The process is repeated a large number of times and the proportion of LRT statistics exceeding the observed value of 2.5684 is used to estimate the  $p$ -value. In practice, we do not know the true values of  $\mu$  and  $\sigma$ , but we can use the estimated values; this distinguishes the parametric bootstrap from the purely simulation approach. The `simulate` function makes it simple to generate a sample from a model:

```
y <- simulate(nullmod)
```

Now taking the data we generate, we fit both the null and alternative models and then compute the LRT. We repeat the process 1000 times:

```
lrstat <- numeric(1000)
set.seed(123)
for(i in 1:1000) {
  y <- unlist(simulate(nullmod))
  bnull <- lm(y ~ 1)
  balt <- lmer(y ~ 1 + (1|operator), pulp, REML=FALSE)
  lrstat[i] <- as.numeric(2*(logLik(balt)-logLik(bnull)))
}
```

We have set the random number seed here so that the results will reproduce exactly if you run the same code. You do not need to set a seed for your own data unless you need to achieve the same reproducibility. Be aware that simulation naturally contains

some variation. If this variation might make a difference to your conclusions, you need to use a larger number of bootstrap samples.

We may examine the distribution of the bootstrapped LRTs. We compute the proportion that are close to zero:

```
mean(lrstat < 0.00001)
[1] 0.703
```

We see there is a 70% chance that the likelihoods for the null and alternatives are virtually identical giving an LRT statistic of practically zero. The LRT clearly does not have a  $\chi^2$  distribution. There is some discussion of this matter in Stram and Lee (1994), who propose a 50:50 mixture of a  $\chi^2$  and a mass at zero. Unfortunately, as we can see, the relative proportions of these two components vary from case to case. Crainiceanu and Ruppert (2004) give a more complete solution to the one-way ANOVA problem, but there is no general and exact result for this and more complex problems. The parametric bootstrap may be the simplest approach. The method we have used above is transparent and could be computed much more efficiently if speed is an issue.

Our estimated *p*-value is:

```
mean(lrstat > 2.5684)
[1] 0.019
```

We can compute the standard error for this estimate by:

```
sqrt(0.019*0.981/1000)
[1] 0.0043173
```

So we can be fairly sure it is under 5%. If in doubt, do some more replications to make sure; this only costs computer time. As it happens, this *p*-value is close to the fixed effects *p*-value.

The RLsim package of Scheipl et al. (2008) can be used to test random effect terms:

```
library(RLsim)
exactLRT(smod, nullmod)
```

No restrictions on fixed effects. REML-based inference preferable.

simulated finite sample distribution of LRT. (*p*-value based  
on 10000 simulated values)

```
data:  
LRT = 2.5684, p-value = 0.0213
```

The result is obtained with less computing time than our explicitly worked example. The difference in the outcomes is within the sampling error. As the output points out, it is slightly better to use REML when testing the random effects (although remember that REML would be invalid for testing fixed effects). We can make this computation:

```
exactRLRT(mmod)  
simulated finite sample distribution of RLRT.
```

(*p*-value based on 10000 simulated values)

```
data:  
RLRT = 3.4701, p-value = 0.021
```

Notice that the testing function is now exactRLRT and that only the alternative model needs to be specified as there is only one random effect component. The outcome is very similar to those obtained previously.

The parametric bootstrap can also be used to construct confidence intervals for the parameters. We simulate data from the chosen model and estimate the parameters. We repeat this process many times, storing the results each time. Quantiles of the bootstrapped estimates are then used to compute the intervals. We need to be able to extract the parameter estimates from the model. We can view the estimates of variance parameters using:

```
VarCorr(mmod)
```

Groups	Name	Std.Dev.
operator	(Intercept)	0.261
Residual		0.326

A more convenient form for extracting the values can be obtained as:

```
as.data.frame(VarCorr(mmod))
```

grp	var1	var2	vcov	sdcor
operator	(Intercept)	<NA>	0.068083	0.26093
Residual		<NA>	0.106250	0.32596

Now we are ready to bootstrap:

```
bsd <- numeric(1000)
for(i in 1:1000 {
    y <- unlist(simulate(mmod))
    bmod <- refit(mmod, y)
    bsd[i] <- as.data.frame(VarCorr(bmod))$sdcor[1]
}
```

The `refit` function changes only the response in a model we have already fit. This is significantly faster than fitting the model from scratch as the overhead in setting up the model is avoided. The 95% bootstrap confidence interval for  $\sigma_\alpha$  is:

```
quantile(bsd, c(0.025, 0.975))
  2.5%   97.5%
0.00000 0.51335
```

Essentially the same result can be obtained more directly using the `confint` function:

```
confint(mmod, method="boot")
```

Computing bootstrap confidence intervals ...		
	2.5 %	97.5 %
sd_(Intercept) operator	0.00000	0.51539
sigma	0.21347	0.45522
(Intercept)	60.09417	60.69724

Nevertheless, it is worth understanding the detailed method of construction to know how it works and to allow one to modify the method if circumstances require it.

In this case, the lower bound is zero. This is not surprising given our earlier uncertainty over whether there really is a difference between the operators. In simpler circumstances, there is a duality between confidence intervals and hypothesis tests in that the outcome of a test can be determined by whether the point null hypothesis lies within the confidence interval. Unfortunately, this duality does not apply in all circumstances, this being a case in point. If you want to do a hypothesis test, use the method described earlier and not the confidence interval.

In this example, the random and fixed effect tests gave similar outcomes. However, the hypotheses in random and fixed effects are intrinsically different. To generalize somewhat, it is easier to conclude there is an effect in a fixed effects model since the conclusion applies only to the levels of the factor used in the experiment, while for random effects, the conclusion extends to levels of the factor not considered.

Since the range of the random effect conclusions is greater, the evidence necessarily has to be stronger.

### 10.3 Estimating Random Effects

In a fixed effects model, the effects are represented by parameters and it makes sense to estimate them. For example, in the one-way ANOVA model:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

We can calculate  $\hat{\alpha}_i$ . We do need to resolve the identifiability problem with the  $\alpha$ s and the  $\mu$ , but once we decide on this, the meaning of the  $\hat{\alpha}$ s is clear enough. We can then proceed to make further inference such as multiple comparisons of these levels.

In a model with random effects, the  $\alpha$ s are no longer parameters, but random variables. Using the standard normal assumption:

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

It does not make sense to estimate the  $\alpha$ s because they are random variables. So instead, we might think about the expected values. However:

$$E\alpha_i = 0 \quad \forall i$$

which is clearly not very interesting. If one looks at this from a Bayesian point of view, as described in, for example, Gelman et al. (2013), we have a prior density on the  $\alpha$ s. The prior mean is  $E\alpha_i = 0$ . Let  $f$  represent density, then the posterior density for  $\alpha$  is given by:

$$f(\alpha_i|y) \propto f(y|\alpha_i)f(\alpha_i)$$

We can then find the posterior mean, denoted by  $\hat{\alpha}$  as:

$$E(\alpha_i|y) = \int \alpha_i f(\alpha_i|y) d\alpha_i$$

For the general case, this works out to be:

$$\hat{\alpha} = DZ^T V^{-1} (y - X\beta)$$

Now a purely Bayesian approach would specify the parameters of the prior (or specify priors for these) and compute a posterior distribution for  $\alpha$ . Here we take an empirical Bayes point of view and substitute the MLEs into  $D$ ,  $V$  and  $\beta$  to obtain the predicted random effects. These may be computed as:

```
raneff(mmod) $operator
```

(Intercept)	
a	-0.12194
b	-0.25912
c	0.16767
d	0.21340

The predicted random effects are related to the fixed effects. These fixed effects are:

```
(cc <- model.tables(lmod) )
```

Tables of effects

```
operator
operator
  a      b      c      d
-0.16 -0.34  0.22  0.28
```

Let's compute the ratio to the random effects as:

```
cc[[1]]$operator/ranef(mmod)$operator
X.Intercept.
a      1.3121
b      1.3121
c      1.3121
d      1.3121
```

We see that the predicted random effects are exactly in proportion to the fixed effects. Typically, the predicted random effects are smaller and could be viewed as a type of *shrinkage* estimate.

The 95% confidence intervals for the random effects can be calculated and displayed as seen in Figure 10.2.

```
library(lattice)
dotplot(ranef(mmod, condVar=TRUE))
```

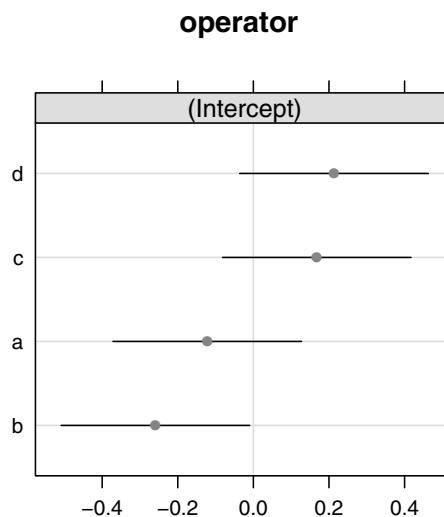


Figure 10.2 *Confidence intervals for the random effects in the pulp data.*

## 10.4 Prediction

Suppose we wish to predict a new value. If the prediction is to be made for a new operator or unknown operator, the best we can do is give  $\hat{\mu} = 60.4$ . If we know the operator, then we can combine this with our fixed effects to produce what are known as the *best linear unbiased predictors* (BLUPs) as follows:

```
fixef(mmod) + ranef(mmod)$operator
(Intercept)
a      60.278
b      60.141
c      60.568
d      60.613
```

We can also use the predict function to construct these predictions. First consider the prediction for a new or unknown operator. We specify the random effects part of the prediction as  $\sim 0$  meaning that this term is not present. In more complex models with more than one random effect, more choices are available. By default this would produce a fitted value for each case in the data but since these are identical we take only the first value:

```
predict(mmod, re.form=~0) [1]
1
60.4
```

Now we specify that the operator is ‘a’:

```
predict(mmod, newdata=data.frame(operator="a"))
1
60.278
```

The predict function for mixed model objects does not compute standard errors or prediction intervals. For this simple model, it would be possible to compute these explicitly but for more general models, it becomes much more difficult. For this reason, we present a parametric bootstrap method for computing these as it is clearer how the bands are computed. We start with the unknown operator case:

```
group.sd <- as.data.frame(VarCorr(mmod))$sdcor[1]
resid.sd <- as.data.frame(VarCorr(mmod))$sdcor[2]
pv <- numeric(1000)
for(i in 1:1000){
  y <- unlist(simulate(mmod))
  bmod <- refit(mmod, y)
  pv[i] <- predict(bmod, re.form=~0)[1] + rnorm(n=1, sd=group.sd) +
    rnorm(n=1, sd=resid.sd)
}
quantile(pv, c(0.025, 0.975))
2.5% 97.5%
59.535 61.286
```

As in previous bootstraps, the first step is to simulate from the fitted model. We refit the model with the simulated response and generate a predicted value. But there are two additional sources of variation. We have variation due to the new operator and also due to a new observation from that operator. For this reason, we add normal sample values with standard deviations equal to those estimated earlier. If you really want a confidence interval for the mean prediction, you should not add these extra error terms. We repeat this 1000 times and take the appropriate quantiles to get a 95% interval.

Some modification is necessary if we know the operator we are making the prediction interval for. We use the option `use.u=TRUE` in the `simulate` function indicating that we should simulate new values conditional on the estimated random effects. We need to do this because otherwise we would simulate an entirely new ‘a’ effect in each replication. Instead, we want to preserve the originally generated ‘a’ effect.

```

for(i in 1:1000){
  y <- unlist(simulate(mmod, use.u=TRUE))
  bmod <- refit(mmod, y)
  pv[i] <- predict(bmod, newdata=data.frame(operator="a")) + rnorm(n=1,
    sd=resid.sd)
}
quantile(pv, c(0.025, 0.975))
  2.5% 97.5%
59.606 61.023

```

In a simple model such as this, we could mathematically calculate the standard error formulas and use this to compute these intervals more efficiently. However, the bootstrap is more general and is easier to apply in more complex situations. More bootstrapping functionality can be found in the `lme4::bootMer()` function and also in the `merTools` package. Bootstrapping is fast enough for simple models but greater efficiency is needed in more complex cases.

## 10.5 Diagnostics

It is important to check the assumptions made in fitting the model. Diagnostic methods available for checking linear mixed models largely mirror those used for linear models but there are some variations. Residuals are commonly defined as the difference between the observed and fitted values. In mixed models, there is more than one kind of fitted (or predicted) value resulting in more than one kind of residual. The default predicted values and residuals use the estimated random effects. This means these residuals can be regarded as estimates of  $\epsilon$  which is usually what we want.

As with linear models, this pair of diagnostics plots is most valuable:

```

qqnorm(residuals(mmod), main="")
plot(fitted(mmod), residuals(mmod), xlab="Fitted", ylab="Residuals")
abline(h=0)

```

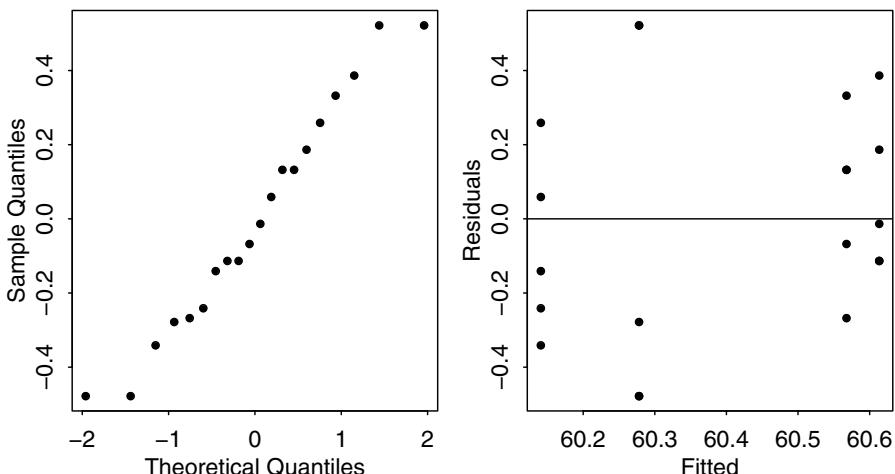


Figure 10.3 *Diagnostic plots for the one-way random effects model.*

The plots are shown in Figure 10.3 and indicate no particular problems. Random effects models are particularly sensitive to outliers, because they depend on variance components that can be substantially inflated by unusual points. The QQ plot is one way to pick out outliers. We also need the normality for the testing. The residual-fitted plot is also important because we made the assumption that the error variance was constant.

If we had more than four groups, we could also look at the normality of the group level effects and check for constant variance also. With so few groups, it is not sensible to do this. Also note that there is no particular reason to think about multiple comparisons. These are for comparing selected levels of a factor. For a random effect, the levels were randomly selected, so such comparisons have less motivation.

## 10.6 Blocks as Random Effects

Blocks are properties of the experimental units. The blocks are either clearly defined by the conditions of the experiment or they are formed with the judgment of the experimenter. Sometimes, blocks represent groups of runs completed in the same period of time. Typically, we are not interested in the block effects specifically, but must account for their effect. It is therefore natural to treat blocks as random effects.

We illustrate with an experiment to compare four processes, A, B, C and D, for the production of penicillin. These are the treatments. The raw material, corn steep liquor, is quite variable and can only be made in blends sufficient for four runs. Thus a randomized complete block design is suggested by the nature of the experimental units. The data comes from Box et al. (1978). We start with the fixed effects analysis:

```
data(penicillin, package="faraway")
summary(penicillin)
```

treat	blend	yield
A:5	Blend1:4	Min. :77
B:5	Blend2:4	1st Qu.:81
C:5	Blend3:4	Median :87
D:5	Blend4:4	Mean :86
	Blend5:4	3rd Qu.:89
		Max. :97

We plot the data as seen in Figure 10.4. We create a version of the blend variable to get neater labeling.

```
penicillin$Blend <- gl(5, 4)
ggplot(penicillin, aes(y=yield, x=treat, shape=Blend)) +geom_point() +
  ← xlab("Treatment")
ggplot(penicillin, aes(y=yield, x=Blend, shape=treat)) +geom_point()
```

It is convenient to use sum contrasts rather than the default treatment contrasts for the purpose of comparison to the mixed effect modeling to come.

```
op <- options(contrasts=c("contr.sum", "contr.poly"))
lmod <- aov(yield ~ blend + treat, penicillin)
summary(lmod)
```

	Df	Sum Sq	Mean Sq	F	value	Pr(>F)
blend	4	264.0	66.0	3.50	0.041	
treat	3	70.0	23.3	1.24	0.339	
Residuals	12	226.0	18.8			

```
coef(lmod)
```

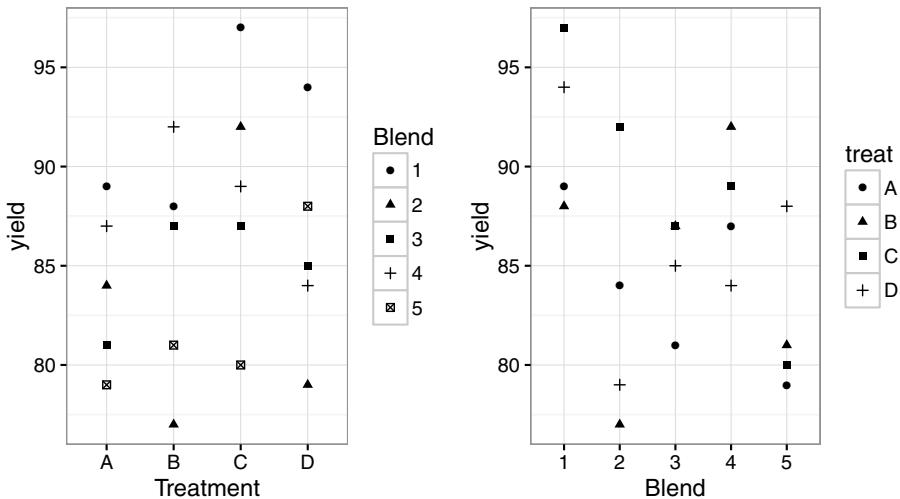


Figure 10.4 *Yield from penicillin blends varying by treatment.*

	blend1	blend2	blend3	blend4
treat1	86	6	-3	-1
treat2		treat3		2
treat3	-2	-1	3	

From this we see that there is no significant difference between the treatments, but there is between the blends. Now let's fit the data with a mixed model, where we have fixed treatment effects, but random blend effects. This seems natural since the blends we use can be viewed as having been selected from some notional population of blends.

```
mmod <- lmer(yield ~ treat + (1|blend), penicillin)
summary(mmod)
```

Fixed Effects:

	coef.est	coef.se
(Intercept)	86.00	1.82
treat1	-2.00	1.68
treat2	-1.00	1.68
treat3	3.00	1.68

Random Effects:

Groups	Name	Std.Dev.
blend	(Intercept)	3.43
	Residual	4.34

---

number of obs: 20, groups: blend, 5

AIC = 118.6, DIC = 128

deviance = 117.3

```
options(op)
```

We notice a few connections. The residual variance is the same in both cases:  $18.8 = 4.34^2$ . This is because we have a balanced design and so REML is equivalent to the

ANOVA estimator. The treatment effects are also the same as is the overall mean. The BLUPs for the random effects are:

```
ranef(mmod)$blend
```

	(Intercept)
Blend1	4.28788
Blend2	-2.14394
Blend3	-0.71465
Blend4	1.42929
Blend5	-2.85859

which, as with the one-way ANOVA, are a shrunken version of the corresponding fixed effects. The usual diagnostics show nothing amiss.

We have a number of options in testing the fixed effects in this example. For this simple balanced model, the `aov` function can be used:

```
amod <- aov(yield ~ treat + Error(blend), penicillin)
summary(amod)
```

Error: blend

Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	4	264	66	

Error: Within

Df	Sum Sq	Mean Sq	F value	Pr(>F)
treat	3	70	23.3	1.24 0.34
Residuals	12	226	18.8	

Notice how the random effects term for `blend` is specified. The *p*-value for testing the treatment effects is 0.34 indicating no significant effect. This test is exact and works well here but only works for simple balanced models. For example, a single missing value would invalidate this test so we have good reason to explore more general methods.

We might try to base a test on the *F*-statistic which can be obtained like this:

```
anova(mmod)
```

Analysis of Variance Table

Df	Sum Sq	Mean Sq	F value
treat	3	70	23.3
			1.24
Residuals	12	226	18.8

For this simple balanced case, it can be shown that this *F*-statistic has a true *F* null distribution with usual degrees of freedom from which we could derive a *p*-value. Unfortunately, as with the `aov` method, this result does not generalize well.

More reliable *F*-tests can be achieved by using adjusted degrees of freedom. The `pbkrtest` package (Halekoh and Højsgaard (2014)) implements the Kenward-Roger (Kenward and Roger (1997)) method:

```
library(pbkrtest)
```

```
amod <- lmer(yield ~ treat + (1|blend), penicillin, REML=FALSE)
nmod <- lmer(yield ~ 1 + (1|blend), penicillin, REML=FALSE)
KRmodcomp(amod, nmod)
```

F-test with Kenward-Roger approximation; computing time: 0.14 sec.

large	yield ~ treat + (1   blend)
small	yield ~ 1 + (1   blend)
stat	ndf ddf F.scaling p.value
Ftest	1.24 3.00 12.00 1 0.34

It is essential to use the ML method of estimation when testing fixed effects. Since we wish to test the treatment effect, we fit the model with this term and the same model but without this term. As can be seen, it produces an identical result to the `aov`

output with the same degrees of freedom and  $p$ -value. The advantage of this method is that it can be generalized to a much wider class of problems.

We can also use the parametric bootstrap. First we compute the LRT statistic:

```
as.numeric(2*(logLik(amod)-logLik(nmod)))
[1] 4.0474
```

Just for reference, we could use the  $\chi^2$  approximation to quickly compute a  $p$ -value:

```
1-pchisq(4.0474,3)
[1] 0.25639
```

This is just an approximation of unknown quality. We aim to do better than this.

We can improve the accuracy with the parametric bootstrap approach. We can generate a response from the null model and use this to compute the LRT. We repeat this 1000 times, saving the LRT each time:

```
lrstat <- numeric(1000)
for(i in 1:1000){
  ryield <- unlist(simulate(nmod))
  nmodr <- refit(nmod, ryield)
  amodr <- refit(amod, ryield)
  lrstat[i] <- 2*(logLik(amodr)-logLik(nmodr))
}
```

Notice how we have used `refit` to speed up the computation. Under the standard likelihood theory, the LRT statistic here should have a  $\chi^2_3$  distribution. A QQ plot of these simulated LRT values indicates that this is a poor approximation. We can compute our estimated  $p$ -value as:

```
mean(lrstat > 4.0474)
[1] 0.353
```

which is much closer to the  $F$ -test result than the  $\chi^2_3$ -based approximation.

The `pbkrtest` package offers a convenient way to perform the parametric bootstrap for fixed effect terms:

```
pmod <- PBmodcomp(amod, nmod)
summary(pmod)

Parametric bootstrap test; time: 32.22 sec; samples: 1000 extremes: 333;
large : yield ~ treat + (1 | blend)
small : yield ~ 1 + (1 | blend)
      stat   df ddf p.value
PBtest    4.05        0.33
Gamma     4.05        0.33
Bartlett 3.42 3.00   0.33
F         1.35 3.00 12.9   0.30
LRT       4.05 3.00   0.26
```

The parametric bootstrap  $p$ -value is 0.33, which is similar to our previous results. Remember that bootstrap is based on random sampling so if you repeat this, you will get slightly different results. Since this  $p$ -value is not close to significance, we have no worries about this. Notice that the output also produces the  $\chi^2$ -based LRT result along with three other versions that are explained in the documentation for the `pbkrtest` package. The package also offers the possibility of using the multiple cores available now on most computers. This parallel computing can be helpful as the parametric bootstrap is computationally expensive.

We can also test the significance of the blends. As with a fixed effects analysis, we are typically not directly interested in size of the blocking effects. Once having decided to design the experiment with blocks, we must retain them in the model.

However, we may wish to examine the blocking effects for information useful for the design of future experiments. We can fit the model with and without random effects and compute the LRT:

```
rmod <- lmer(yield ~ treat + (1|blend), penicillin)
nlmod <- lm(yield ~ treat, penicillin)
as.numeric(2*(logLik(rmod)-logLik(nlmod, REML=TRUE)))
[1] 2.7629
```

We need to specify the nondefault REML option for null model to ensure that the LRT is computed correctly. Now we perform the parametric bootstrap much as before:

```
lrstatf <- numeric(1000)
for(i in 1:1000){
  ryield <- unlist(simulate(nlmod))
  nlmodr <- lm(ryield ~ treat, penicillin)
  rmodr <- refit(rmod, ryield)
  lrstatf[i] <- 2*(logLik(rmodr)-logLik(nlmodr, REML=TRUE))
}
```

Again, the distribution is far from  $\chi^2_1$  which is clear when we examine the proportion of generated LRTs which are close to zero:

```
mean(lrstatf < 0.00001)
[1] 0.551
```

We can see from this that the LRT is clearly not  $\chi^2_1$  distributed. Even the nonzero values seem to have some other distribution. This makes it clear that asymptotic approximations cannot be relied on these circumstances.

We can compute the estimated *p*-value as:

```
mean(lrstatf > 2.7629)
[1] 0.043
```

So we find a significant blend effect. The *p*-value is close to that observed for the fixed effects analysis. Given that the *p*-value is close to 5%, we might wish to increase the number of bootstrap samples to increase our confidence in the result.

We can also use RLRsim to obtain a *p*-value.

```
library(RLRsim)
exactRLRT(rmod)
simulated finite sample distribution of RLRT.

(p-value based on 10000 simulated values)

data:
RLRT = 2.7629, p-value = 0.0406
```

In this example, we saw no major advantage in modeling the blocks as random effects, so we might prefer to use the fixed effects analysis as it is simpler to execute. However, in subsequent analyses, we shall see that the use of random effects will be mandatory as equivalent results may not be obtained from a purely fixed effects analysis.

## 10.7 Split Plots

Split plot designs originated in agriculture, but occur frequently in other settings. As the name implies, main plots are split into several subplots. The main plot is treated with a level of one factor while the levels of some other factor are allowed to vary

with the subplots. The design arises as a result of restrictions on a full randomization. For example, a field may be divided into four subplots. It may be possible to plant different varieties in the subplots, but only one type of irrigation may be used for the whole field. Note the distinction between split plots and blocks. Blocks are features of the experimental units which we have the option to take advantage of in the experimental design. Split plots impose restrictions on what assignments of factors are possible. They impose requirements on the design that prevent a complete randomization. Split plots often arise in nonagricultural settings when one factor is easy to change while another factor takes much more time to change. If the experimenter must do all runs for each level of the hard-to-change factor consecutively, a split-plot design results with the hard-to-change factor representing the whole plot factor.

Consider the following example. In an agricultural field trial, the objective was to determine the effects of two crop varieties and four different irrigation methods. Eight fields were available, but only one type of irrigation may be applied to each field. The fields may be divided into two parts with a different variety planted in each half. The whole plot factor is the method of irrigation, which should be randomly assigned to the fields. Within each field, the variety is randomly assigned. Here is a summary of the data:

```
data(irrigation, package="faraway")
summary(irrigation)
```

field	irrigation	variety	yield	
f1	:2	i1:4	v1:8	Min. :34.8
f2	:2	i2:4	v2:8	1st Qu.:37.6
f3	:2	i3:4		Median :40.1
f4	:2	i4:4		Mean :40.2
f5	:2			3rd Qu.:42.7
f6	:2			Max. :47.6
(Other):4				

We can plot the data as seen in Figure 10.5.

```
ggplot(irrigation, aes(y=yield, x=field, shape=irrigation, color=
  ↪ variety)) + geom_point()
```

The irrigation and variety are fixed effects, but the field is clearly a random effect. We must also consider the interaction between field and variety, which is necessarily also a random effect because one of the two components is random. The fullest model that we might consider is:

$$y_{ijk} = \mu + i_i + v_j + (iv)_{ij} + f_k + (vf)_{jk} + \varepsilon_{ijk}$$

$\mu, i_i, v_j, (iv)_{ij}$  are fixed effects; the rest are random having variances  $\sigma_f^2$ ,  $\sigma_{vf}^2$  and  $\sigma_\varepsilon^2$ . Note that we have no  $(if)_{ik}$  term in this model. It would not be possible to estimate such an effect since only one type of irrigation is used on a given field; the factors are not crossed. We would fit such a model using the expression

```
lmer(yield ~ irrigation * variety + (1|field) + (1|field:variety),
  ↪ irrigation)
```

However, if you try to fit such a model, it will fail because it is not possible to distinguish the variety within the field variation. We would need more than one observation per variety within each field for us to separate the two variabilities. We resort to a simpler model that omits the variety by field interaction random effect:

$$y_{ijk} = \mu + i_i + v_j + (iv)_{ij} + f_k + \varepsilon_{ijk}$$

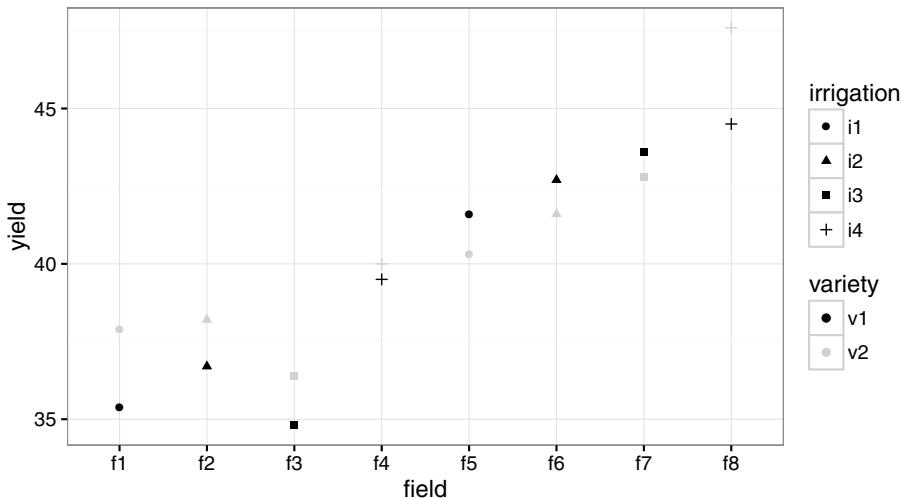


Figure 10.5 Yield on fields with different irrigation methods.

```
lmod <- lmer(yield ~ irrigation * variety + (1|field), irrigation)
summary(lmod)
```

Fixed Effects:

	coef.est	coef.se
(Intercept)	38.50	3.03
irrigationi2	1.20	4.28
irrigationi3	0.70	4.28
irrigationi4	3.50	4.28
varietyv2	0.60	1.45
irrigationi2:varietyv2	-0.40	2.05
irrigationi3:varietyv2	-0.20	2.05
irrigationi4:varietyv2	1.20	2.05

Random Effects:

Groups	Name	Std.Dev.
field	(Intercept)	4.02
Residual		1.45

---

number of obs: 16, groups: field, 8

AIC = 65.4, DIC = 91.8

deviance = 68.6

We can see that the largest variance component is that due to the field effect:  $\hat{\sigma}_f = 4.02$  with  $\hat{\sigma}_\epsilon = 1.45$ .

The relatively large standard errors compared to the fixed effect estimates suggest that there may be no significant fixed effects. We can check this sequentially using *F*-tests with adjusted degrees of freedom:

```
library(pbkrtest)
lmoda <- lmer(yield ~ irrigation + variety + (1|field), data=irrigation
               ↪ )
KRmodcomp(lmod, lmoda)
```

F-test with Kenward-Roger approximation; computing time: 0.07 sec.

```

large : yield ~ irrigation * variety + (1 | field)
small : yield ~ irrigation + variety + (1 | field)
      stat  ndf  ddf F.scaling p.value
Ftest 0.25 3.00 4.00      1     0.86

```

We find there is no significant interaction term. We can now test each of the main effects starting with the variety:

```
lmodi <- lmer(yield ~ irrigation + (1|field), irrigation)
```

```
KRmodcomp(lmoda, lmodi)
```

F-test with Kenward-Roger approximation; computing time: 0.06 sec.

```

large : yield ~ irrigation + variety + (1 | field)
small : yield ~ irrigation + (1 | field)
      stat  ndf  ddf F.scaling p.value
Ftest 1.58 1.00 7.00      1     0.25

```

Dropping variety from the model seems reasonable since the  $p$ -value of 0.25 is large.

We can test irrigation in a similar manner:

```
lmodv <- lmer(yield ~ variety + (1|field), irrigation)
```

```
KRmodcomp(lmoda, lmodv)
```

F-test with Kenward-Roger approximation; computing time: 0.06 sec.

```

large : yield ~ irrigation + variety + (1 | field)
small : yield ~ variety + (1 | field)
      stat  ndf  ddf F.scaling p.value
Ftest 0.39 3.00 4.00      1     0.77

```

Irrigation also fails to be significant.

We should check the diagnostic plots to make sure there is nothing amiss:

```
plot(fitted(lmod), residuals(lmod), xlab="Fitted", ylab="Residuals")
qqnorm(residuals(lmod), main="")
```

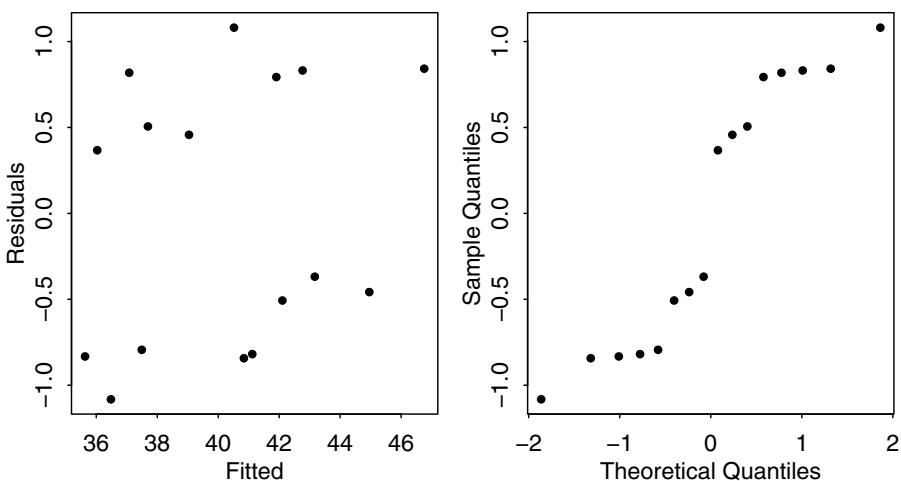


Figure 10.6 *Diagnostic plots for the split plot example.*

We can see in Figure 10.6 that there is no problem with the nonconstant variance, but that the residuals indicate a bimodal distribution caused by the pairs of observations in each field. This type of divergence from normality is unlikely to cause any major problems with the estimation and inference.

We can test the random effects term like this:

```
library(RLRLsim)
exactRLRT(lmod)
```

simulated finite sample distribution of RLRT.

(p-value based on 10000 simulated values)

data:

RLRT = 6.1118, p-value = 0.0098

We see that the fields do seem to vary as the result is clearly significant.

Sometimes analysts ignore the split-plot variable as in:

```
mod <- lm(yield ~ irrigation * variety, data=irrigation)
anova(mod)
```

Analysis of Variance Table

Response: yield

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
irrigation	3	40.2	13.4	0.73	0.56
variety	1	2.2	2.2	0.12	0.73
irrigation:variety	3	1.6	0.5	0.03	0.99
Residuals	8	146.5	18.3		

The results will not be the same. This last model is incorrect because it fails to take into account the restrictions on the randomization introduced by the fields and the additional variability thereby induced.

## 10.8 Nested Effects

When the levels of one factor vary only within the levels of another factor, that factor is said to be *nested*. For example, when measuring the performance of workers at several different job locations, if the workers only work at one location, the workers are nested within the locations. If the workers work at more than one location, then the workers are *crossed* with locations.

Here is an example to illustrate nesting. Consistency between laboratory tests is important and yet the results may depend on who did the test and where the test was performed. In an experiment to test levels of consistency, a large jar of dried egg powder was divided up into a number of samples. Because the powder was homogenized, the fat content of the samples is the same, but this fact is withheld from the laboratories. Four samples were sent to each of six laboratories. Two of the samples were labeled as G and two as H, although in fact they were identical. The laboratories were instructed to give two samples to two different technicians. The technicians were then instructed to divide their samples into two parts and measure the fat content of each. So each laboratory reported eight measures, each technician four measures, that is, two replicated measures on each of two samples. The data comes from Bliss (1967):

```
data(eggs, package="faraway")
summary(eggs)
```

Fat	Lab	Technician	Sample
Min. :0.060	I :8	one:24	G:24
1st Qu.:0.307	II :8	two:24	H:24
Median :0.370	III:8		

```
Mean   :0.388   IV :8
3rd Qu.:0.430   V  :8
Max.   :0.800   VI :8
```

We can plot the data as seen in Figure 10.7.

```
library(ggplot2)
ggplot(eggs, aes(y=Fat, x=Lab, color=Technician, shape=Sample)) + geom_point(position = position_jitter(width=0.1, height=0.0))+scale_color_grey()
```

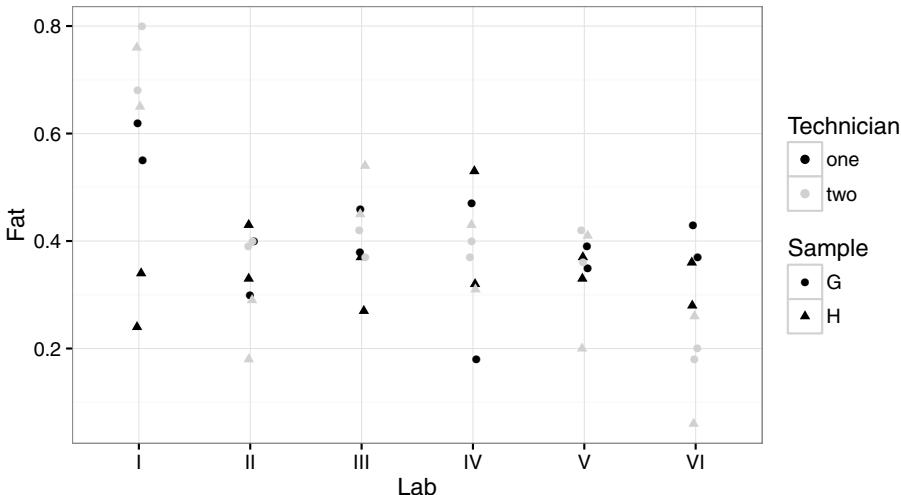


Figure 10.7 *Fat content of homogenous powdered egg as tested by different laboratories, technicians and samples.*

Although the technicians have been labeled “one” and “two,” they are two different people in each lab. Thus the technician factor is nested within laboratories. Furthermore, even though the samples are labeled “H” and “G,” these are not the same samples across the technicians and the laboratories. Hence we have samples nested within technicians. Technicians and samples should be treated as random effects since we may consider these as randomly sampled. If the labs were specifically selected, then they should be taken as fixed effects. If, however, they were randomly selected from those available, then they should be treated as random effects. If the purpose of the study is to come to some conclusion about consistency across laboratories, the latter approach is advisable.

For the purposes of this analysis, we will treat labs as random. So all our effects (except the grand mean) are random. The model is:

$$y_{ijkl} = \mu + L_i + T_{ij} + S_{ijk} + \varepsilon_{ijkl}$$

This can be fit using:

```
cmod <- lmer(Fat ~ 1 + (1|Lab) + (1|Lab:Technician) + (1|Lab:
  ↪ Technician:Sample), data=eggs)
summary(cmod)
```

```

Fixed Effects:
coef.est  coef.se
  0.39      0.04

Random Effects:
Groups           Name        Std.Dev.
Lab:Technician:Sample (Intercept) 0.06
Lab:Technician       (Intercept) 0.08
Lab                 (Intercept) 0.08
Residual            0.08
---
number of obs: 48, groups: Lab:Technician:Sample, 24; Lab:Technician, 12; Lab, 6
AIC = -54.2, DIC = -73.3
deviance = -68.8

```

So we have  $\hat{\sigma}_L = 0.08$ ,  $\hat{\sigma}_T = 0.08$ ,  $\hat{\sigma}_S = 0.06$  and  $\hat{\sigma}_e = 0.08$ . So all four variance components are of a similar magnitude. The lack of consistency in measures of fat content can be ascribed to variance between labs, variance between technicians, variance in measurement due to different labeling and just plain measurement error. We can see if the model can be simplified by removing the lowest level of the variance components. Again the parametric bootstrap can be used:

```

cmodr <- lmer(Fat ~ 1 + (1|Lab) + (1|Lab:Technician), data=eggs)
lrstat <- numeric(1000)
for(i in 1:1000){
  rFat <- unlist(simulate(cmodr))
  nmod <- lmer(rFat ~ 1 + (1|Lab) + (1|Lab:Technician), data=eggs)
  amod <- lmer(rFat ~ 1 + (1|Lab) + (1|Lab:Technician) +
    (1|Lab:Technician:Sample), data=eggs)
  lrstat[i] <- 2*(logLik(amod)-logLik(nmod))
}
mean(lrstat > 2*(logLik(cmod)-logLik(cmodr)))
[1] 0.092

```

We do not reject  $H_0 : \sigma_S^2 = 0$ . A similar computation may be made using the RLRsim package. This requires us to specify another model where only the tested random effect is included:

```

library(RLRsim)
cmods <- lmer(Fat ~ 1 + (1|Lab:Technician:Sample), data=eggs)
exactRLRT(cmods, cmod, cmodr)
simulated finite sample distribution of RLRT.

(p-value based on 10000 simulated values)

```

```

data:
RLRT = 1.6034, p-value = 0.1056

```

An examination of the reduced model is interesting:

```

VarCorr(cmodr)
Groups           Name        Std.Dev.
Lab:Technician (Intercept) 0.0895
Lab             (Intercept) 0.0769
Residual        0.0961

```

The variation due to samples has been absorbed into the other components.

So we can reasonably say that the variation due to samples can be ignored. We may now test the significance of the variation between technicians. Using the same method above, this is found to be significant.

Although the data has a natural hierarchical structure which suggests a particular order of testing, we might reasonably wonder which of the components contribute substantially to the overall variation. Why test the sample effect first? A look at the confidence intervals reveals the problem:

```
confint(cmod, method="boot")
2.5 % 97.5 %
sd_(Intercept) | Lab:Technician:Sample 0.000000 0.097527
sd_(Intercept) | Lab:Technician      0.000000 0.136021
sd_(Intercept) | Lab              0.000000 0.152663
sigma            0.058872 0.107040
(Intercept)       0.299666 0.473920
```

We might drop any of the three random effect terms but it is not possible to be sure which is best to go. It is safest to conclude there is some variation in the fat measurement coming from all three sources.

## 10.9 Crossed Effects

Effects are said to be crossed when they are not nested. In full factorial designs, effects are completely crossed because every level of one factor occurs with every level of another factor. However, in some other designs, crossing is not complete. An example of less than complete crossing is a latin square design, where there is one treatment factor and two blocking factors. Although not all combinations of factors occur, the blocking factors are not nested. When at least some crossing occurs, methods for nested designs cannot be used. We consider a latin square example.

In an experiment reported by Davies (1954), four materials, A, B, C and D, were fed into a wear-testing machine. The response is the loss of weight in 0.1 mm over the testing period. The machine could process four samples at a time and past experience indicated that there were some differences due to the position of these four samples. Also some differences were suspected from run to run. A fixed effects analysis of this dataset may be found in Faraway (2014). Four runs were made. The latin square structure of the design may be observed:

```
data(abrasion, package="faraway")
matrix(abrasion$material, 4, 4)
 [,1] [,2] [,3] [,4]
[1,] "C"  "A"  "D"  "B"
[2,] "D"  "B"  "C"  "A"
[3,] "B"  "D"  "A"  "C"
[4,] "A"  "C"  "B"  "D"
```

We can plot the data as seen in Figure 10.8.

```
library(ggplot2)
ggplot(abrasion, aes(x=material, y=wear, shape=run, color=position))+
  → geom_point(position = position_jitter(width=0.1, height=0.0))+
  → scale_color_grey()
```

A fixed effects analysis of the data reveals:

```
lmod <- aov(wear ~ material + run + position, abrasion)
summary(lmod)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
material	3	4622	1540	25.15	0.00085
run	3	986	329	5.37	0.03901

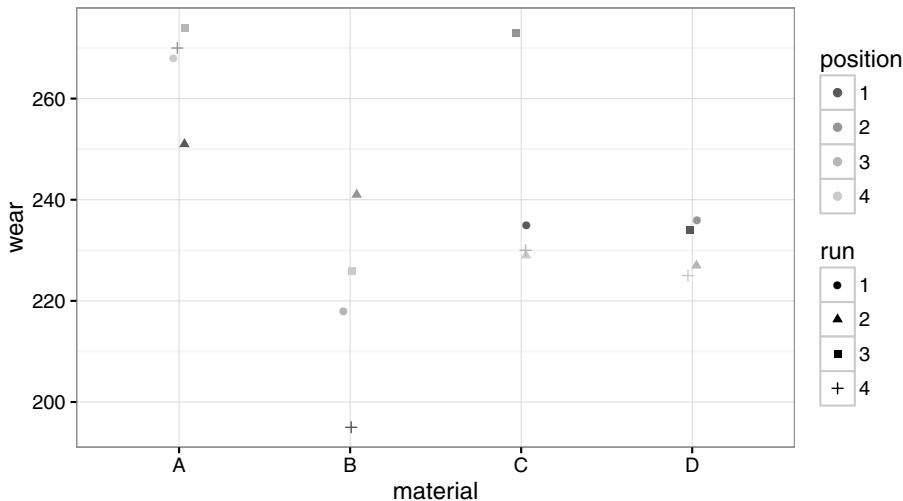


Figure 10.8 Abrasion wear on material according to run and position.

```
position      3    1468      489     7.99 0.01617
Residuals     6    367       61
```

All the effects are significant. However, we might regard the run and position as random effects. The appropriate model is then:

```
mmod <- lmer(wear ~ material + (1|run) + (1|position), abrasion)
summary(mmod)
```

Fixed Effects:

	coef.est	coef.se
(Intercept)	265.75	7.67
materialB	-45.75	5.53
materialC	-24.00	5.53
materialD	-35.25	5.53

Random Effects:

Groups	Name	Std.Dev.
run	(Intercept)	8.18
position	(Intercept)	10.35
Residual		7.83

---

```
number of obs: 16, groups: run, 4; position, 4
AIC = 114.3, DIC = 140.4
deviance = 120.3
```

The `lmer` function is able to recognize that the run and position effects are crossed and fit the model appropriately. We can test the random effects using the `RLRsim` package. We need to fit both models that use just one random effect:

```
library(RLRsim)
mmoddp <- lmer(wear ~ material + (1|position), abrasion)
mmodr <- lmer(wear ~ material + (1|run), abrasion)
exactRLRT(mmoddp, mmod, mmodr)
simulated finite sample distribution of RLRT.
```

```
(p-value based on 10000 simulated values)
```

```
data:  
RLRT = 4.5931, p-value = 0.0139
```

This first comparison tests the significance of the position term. The first model in the `exactRLRT` specifies the model with only that random effect term being tested. The second and third terms specify the alternative and null models under the hypothesis being tested. We see that the position variance is statistically significant. We can also test the run term:

```
exactRLRT (mmodr, mmmod, mmodp)  
simulated finite sample distribution of RLRT.
```

```
(p-value based on 10000 simulated values)
```

```
data:  
RLRT = 3.0459, p-value = 0.0345
```

We see that the run variation is also statistically significant. Since the design of this experiment has already restricted the randomization to allow for these effects, we would keep these terms in the model even if they were found not to be significant. This information would only be valuable for future experiments.

The fixed effect term can be tested using the `pbkrtest` package. Given the small balanced nature of the experiment, we can feel confident in using the Kenward-Roger adjustment. Note that we need to use ML estimation for the fixed effect comparison.

```
library(pbkrtest)  
mmod <- lmer(wear ~ material + (1|run) + (1|position), abrasion, REML=  
    ↪ FALSE)  
nmod <- lmer(wear ~ 1 + (1|run) + (1|position), abrasion, REML=FALSE)  
KRmodcomp(mmod, nmod)  
F-test with Kenward-Roger approximation; computing time: 0.15 sec.  
large : wear ~ material + (1 | run) + (1 | position)  
small : wear ~ 1 + (1 | run) + (1 | position)  
      stat  ndf ddf F.scaling p.value  
Ftest 25.1  3.0  6.0          1 0.00085
```

We find that there is a clearly significant difference in the materials.

The fixed effects analysis was somewhat easier to execute, but the random effects analysis has the advantage of producing estimates of the variation in the blocking factors which will be more useful in future studies. Fixed effects estimates of the run effect for this experiment are only useful for the current study.

## 10.10 Multilevel Models

*Multilevel models* is a term used for models for data with hierarchical structure. The term is most commonly used in the social sciences. We can use the methodology we have already developed to fit some of these models.

We take as our example some data from the Junior School Project collected from primary (U.S. term is elementary) schools in inner London. The data is described in detail in Mortimore et al. (1988) and a subset is analyzed extensively in Goldstein (1995).

The variables in the data are the `school`, the `class` within the school (up to

four), gender, social class of the father (I=1; II=2; III nonmanual=3; III manual=4; IV=5; V=6; Long-term unemployed=7; Not currently employed=8; Father absent=9), raven's test in year 1, student id number, english test score, mathematics test score and school year (coded 0, 1 and 2 for years one, two and three). So there are up to three measures per student. The data was obtained from the *Multilevel Models project*.

We shall take as our response the math test score result from the final year and try to model this as a function of gender, social class and the Raven's test score from the first year which might be taken as a measure of ability when entering the school. We subset the data to ignore the math scores from the first two years:

```
data(jsp, package="faraway")
jspr <- jsp[jsp$year==2, ]
```

We start with two plots of the data. Due to the discreteness of the score results, it is helpful to *jitter* (add small random perturbations) the scores to avoid overprinting. The use of transparency, specified using the alpha parameter, also helps with dense data.

```
ggplot(jspr, aes(x=raven, y=math)) + xlab("Raven Score") + ylab("Math
  ↪ Score") + geom_point(position = position_jitter(), alpha=0.3)
ggplot(jspr, aes(x=social, y=math)) + xlab("Social Class") + ylab("Math
  ↪ Score") + geom_boxplot()
```

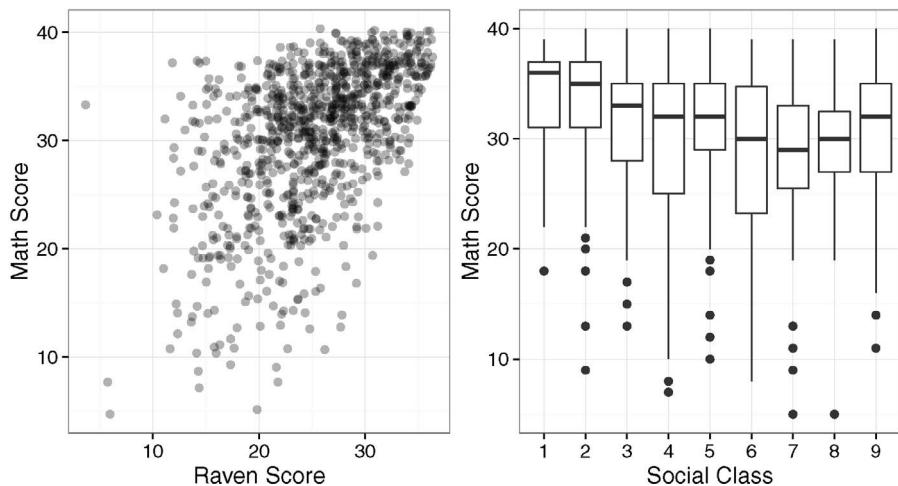


Figure 10.9 Plots of the Junior School Project data.

In Figure 10.9, we can see the positive correlation between the Raven's test score and the final math score. The maximum math score was 40, which reduces the variability at the upper end of the scale. We also see how the math scores tend to decline with social class.

One possible approach to analyzing these data is multiple regression. For example, we could fit:

```
glin <- lm(math ~ raven*gender*social, jspr)
anova(glin)
```

Analysis of Variance Table

Response: math	Df	Sum Sq	Mean Sq	F value	Pr(>F)
raven	1	11481	11481	368.06	<2e-16
gender	1	44	44	1.41	0.2347
social	8	779	97	3.12	0.0017
raven:gender	1	0.01145	0.01145	0.00037	0.9847
raven:social	8	583	73	2.33	0.0175
gender:social	8	450	56	1.80	0.0727
raven:gender:social	8	235	29	0.94	0.4824
Residuals	917	28603	31		

It would seem that gender effects can be removed entirely, giving us:

```
glin <- lm(math ~ raven+social, jspr)
anova(glin)
```

Analysis of Variance Table

Response: math	Df	Sum Sq	Mean Sq	F value	Pr(>F)
raven	1	11481	11481	365.72	<2e-16
social	8	778	97	3.10	0.0019
raven:social	8	564	71	2.25	0.0222
Residuals	935	29351	31		

This is a fairly large dataset, so even small effects can be significant. Even though the raven:social term is significant at the 5% level, we remove it to simplify interpretation:

```
glin <- lm(math ~ raven+social, jspr)
summary(glin)
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	17.0248	1.3745	12.39	<2e-16
raven	0.5804	0.0326	17.83	<2e-16
social2	0.0495	1.1294	0.04	0.965
social3	-0.4289	1.1957	-0.36	0.720
social4	-1.7745	1.0599	-1.67	0.094
social5	-0.7823	1.1892	-0.66	0.511
social6	-2.4937	1.2609	-1.98	0.048
social7	-3.0485	1.2907	-2.36	0.018
social8	-3.1175	1.7749	-1.76	0.079
social9	-0.6328	1.1273	-0.56	0.575

n = 953, p = 10, Residual SE = 5.632, R-Squared = 0.29

We see that the final math score is strongly related to the entering Raven score and that the math scores of the lower social classes are lower, even after adjustment for the entering score. Of course, any regression analysis requires more investigation than this; there are diagnostics and transformations to be considered and more. However, even if we were to do this, there would still be a problem with this analysis. We are assuming that the 953 students in the dataset are independent observations. This is not a tenable assumption as the students come from 50 different schools. The number coming from each school varies:

```
table(jspr$school)
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
26	11	14	24	26	18	11	27	21	0	11	23	22	13	7	16	6	18	14	13	28
22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42

```
14 18 21 14 20 22 15 13 27 35 23 44 27 16 28 17 12 14 10 10 41
44 45 46 47 48 49 50
5 11 15 33 63 22 14
```

It is highly likely that students in the same school (and perhaps class) will show some dependence. So we have somewhat less than 953 independent cases worth of information. Any analysis that pretends these are independent is likely to overstate the significance of the results. Furthermore, the analysis above tells us nothing about the variation between and within schools. People will certainly be interested in this. We could aggregate the results across schools but this would lose information and expose us to the dangers of an ecological regression.

We need an analysis that uses the individual-level information, but also reflects the grouping in the data. Our first model has fixed effects representing all interactions between raven, social and gender with random effects for the school and the class nested within the school:

```
mmod <- lmer(math ~ raven*social*gender + (1|school) + (1|school:class),
  ↪ data=jspr)
```

A look at the summary output from this model suggests that gender may not be significant. We can test this using the Kenward-Roger adjusted  $F$ -test from the pbkrtest package:

```
mmodr <- lmer(math ~ raven*social + (1|school) + (1|school:class),   data=
  ↪ jspr)
KRmodcomp(mmod, mmodr)

F-test with Kenward-Roger approximation; computing time: 0.39 sec.
large : math ~ raven * social * gender + (1 | school) + (1 | school:class)
small : math ~ raven * social + (1 | school) + (1 | school:class)
      stat    ndf    ddf F.scaling p.value
Ftest  1.01  18.00 892.94        1  0.44
```

This can be verified using the parametric bootstrap although with a dataset of this size, it does take some time to run. The size of the dataset means that we can be quite confident about the adjusted  $F$ -test in any case.

In this example, we have more than a handful of potential models we might consider even if we vary only the fixed effect part of the model. In such circumstances, we might prefer to take a criterion-based approach to model selection. One approach is to specify all the models we wish to consider:

```
all13 <- lmer(math ~ raven*social*gender + (1|school) + (1|school:class),
  ↪ data=jspr, REML=FALSE)
all12 <- update(all13, . ~ . - raven:social:gender)
notrs <- update(all12, . ~ . -raven:social)
notrg <- update(all12, . ~ . -raven:gender)
notsg <- update(all12, . ~ . -social:gender)
onlyrs <- update(all12, . ~ . -social:gender - raven:gender)
all11 <- update(all12, . ~ . -social:gender - raven:gender - social:
  ↪ raven)
nogen <- update(all11, . ~ . -gender)
```

It is important to use the ML method for constructing the AICs. As explained previously, it is not sensible to use the REML method when comparing models with different fixed effects. We have specified models with a three-way interaction, all two-way interactions, models leaving out each two-way interaction, a model excluding any interaction involving gender, a model with just main effects and finally a

model without gender entirely. Now we can create a table showing the AIC and BIC values:

```
anova(all3, all2, notrs, notrg, notsg, onlyrs, all1, nogen) [,1:4]
   Df AIC BIC logLik
all1 14 5956 6024 -2964
nogen 21 5949 6051 -2954
onlyrs 22 5950 6057 -2953
notrs 23 5962 6073 -2958
notsg 23 5952 6064 -2953
notrg 30 5956 6102 -2948
all2 31 5958 6108 -2948
all3 39 5967 6156 -2944
```

The anova output produces chi-squared tests for comparing the models. This is not correct here as the sequence of models is not nested and furthermore, these tests are inaccurate for reasons previously explained. We exclude this part of the output using [,1:4]. We can see that the AIC is minimized by the model that removes gender entirely. This confirms our hypothesis-testing based approach to selecting the model but rather more thoroughly by also considering the intermediate models.

The BIC criterion commonly prefers models that are smaller than the AIC. We see that illustrated in this example as BIC picks the model with only the main effects. We might reasonably add other models to the comparison. It becomes tedious to list all the possibilities when there are more variables but it requires some more complex R code to generate these automatically.

Given that we have decided that gender is not important, we simplify to:

```
jspr$craven <- jspr$raven-mean(jspr$raven)
mmod <- lmer(math ~ craven*social+(1|school)+(1|school:class), jspr)
summary(mmod)
```

Fixed Effects:

	coef.est	coef.se
(Intercept)	31.91	1.20
craven	0.61	0.19
social2	0.02	1.27
social3	-0.63	1.31
social4	-1.97	1.20
social5	-1.36	1.30
social6	-2.27	1.37
social7	-2.55	1.41
social8	-3.39	1.80
social9	-0.83	1.25
craven:social2	-0.13	0.21
craven:social3	-0.22	0.22
craven:social4	0.04	0.19
craven:social5	-0.15	0.21
craven:social6	-0.04	0.23
craven:social7	0.40	0.23
craven:social8	0.26	0.26
craven:social9	-0.08	0.21

Random Effects:

Groups	Name	Std.Dev.
school:class	(Intercept)	1.08
school	(Intercept)	1.77
Residual		5.21

```
---
number of obs: 953, groups: school:class, 90; school, 48
AIC = 5963.2, DIC = 5893.6
deviance = 5907.4
```

We centered the Raven score about its overall mean. This means that we can interpret the social effects as the predicted differences from social class one at the mean Raven score. If we did not do this, these parameter estimates would represent differences for `raven=0` which is not very useful. We can see the math score is strongly related to the entering Raven score. We see that for the same entering score, the final math score tends to be lower as social class goes down. Note that class 9 here is when the father is absent and class 8 is not necessarily worse than 7, so this factor is not entirely ordinal. We also see the most substantial variation at the individual level with smaller amounts of variation at the school and class level.

We check the standard diagnostics first:

```
diagd <- fortify(mmod)
ggplot(diagd, aes(sample=.resid)) + stat_qq()
ggplot(diagd, aes(x=.fitted, y=.resid)) + geom_point(alpha=0.3) + geom_
  ↪ hline(yintercept=0) + xlab("Fitted") + ylab("Residuals")
```

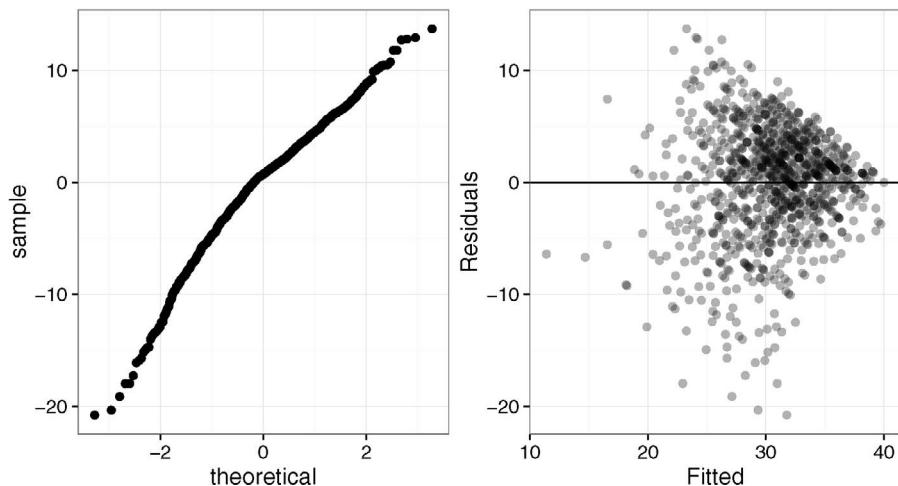


Figure 10.10 *Diagnostic plots for the Junior Schools Project model.*

In Figure 10.10, we see that the residuals are close to normal, but there is a clear decrease in the variance with an increase in the fitted values. This is due to the reduced variation in higher scores already observed. We might consider a transformation of the response to remove this effect.

We can also check the assumption of normally distributed random effects. We can do this at the school and class level:

```
qqnorm(ranef(mmod)$school[[1]], main="School effects")
qqnorm(ranef(mmod)$"school:class"[[1]], main="Class effects")
```

We see in Figure 10.11 that there is approximate normality in both cases with some

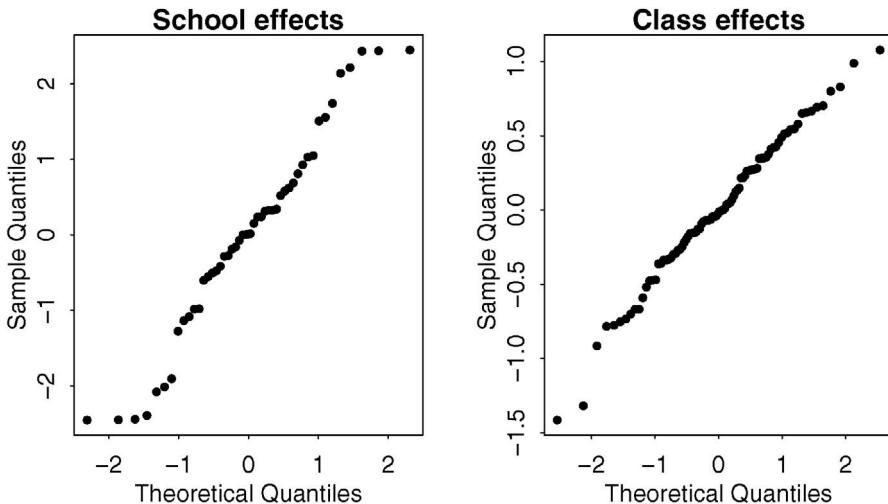


Figure 10.11 *QQ plots of the random effects at the school and class levels.*

evidence of short tails for the school effects. It is interesting to look at the sorted school effects:

```
adjscores <- ranef(mmod)$school[[1]]
```

These represent a ranking of the schools adjusted for the quality of the intake and the social class of the students. The difference between the best and the worst is about five points on the math test. Of course, we must recognize that there is variability in these estimated effects before making any decisions about the relative strengths of these schools. Compare this with an unadjusted ranking that simply takes the average score achieved by the school, centered by the overall average:

```
rawscores <- coef(lm(math ~ school-1, jspr))
rawscores <- rawscores-mean(rawscores)
```

We compare these two measures of school quality in Figure 10.12:

```
plot(rawscores, adjscores)
sint <- c(9, 14, 29)
text(rawscores[sint], adjscores[sint]+0.2, c("9", "15", "30"))
```

School 10 is listed but has no students, hence the need to adjust the labeling. There are some interesting differences. School 15 looks best on the raw scores but after adjustment, it drops to 15th place. This is a school that apparently performs well, but when the quality of the incoming students is considered, its performance is not so impressive. School 30 illustrates the other side of the coin. This school looks average on the raw scores, but is doing quite well given the ability of the incoming students. School 9 is actually doing a poor job despite raw scores that look quite good.

It is also worth plotting the residuals and the random effects against the predictors. We would be interested in finding any inhomogeneity or signs of structure that might lead to an improved model.

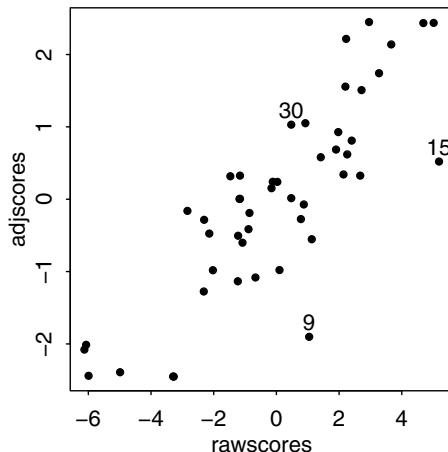


Figure 10.12 Raw and adjusted school-quality measures. Three selected schools are marked.

We may also be interested to know whether there really is much variation between schools or classes within schools. We can investigate this by testing the random effect terms using the `RLRsim` package. We need to fit models without each of the random effect terms.

```
library(RLRsim)
mmodc <- lmer(math ~ craven*social+(1|school:class), jspr)
mmodm <- lmer(math ~ craven*social+(1|school), jspr)
```

We can test the class effect:

```
exactRLRT(mmodc, mmodm, mmodc)
simulated finite sample distribution of RLRT.

(p-value based on 10000 simulated values)
```

```
data:
RLRT = 2.3903, p-value = 0.0549
```

The evidence for a class effect is quite marginal. We would certainly choose to include it for testing fixed effect terms as we would rather be sure that it had been taken account of. Even so we can see that the class effect may be quite small. In contrast, we can test for a school effect:

```
exactRLRT(mmodm, mmodm, mmodc)
simulated finite sample distribution of RLRT.

(p-value based on 10000 simulated values)
```

```
data:
RLRT = 7.1403, p-value = 0.0033
```

The school effect comes through strongly. It seems schools matter more than specific teachers.

**Compositional Effects:** Fixed effect predictors in this example so far have been at the lowest level, the student, but it is not improbable that factors at the school or

class level might be important predictors of success in the math test. We can construct some such predictors from the individual-level information; such factors are called *compositional effects*. For example, the average entering score for a school might be an important predictor. The ability of one's fellow students may have an impact on future achievement. We construct this variable:

```
schraven <- lm(raven ~ school, jspr)$fit
```

and insert it into our model:

```
mmodc <- lmer(math ~ craven*social+schraven*social+(1|school)+ (1 |
  ↪ school:class), jspr)
KRmodcomp(mmod, mmodc)
```

F-test with Kenward-Roger approximation; computing time: 0.16 sec.

```
large : math ~ craven * social + schraven * social + (1 | school) + (1 |
  school:class)
small : math ~ craven * social + (1 | school) + (1 | school:class)
      stat    ndf    ddf F.scaling p.value
Ftest  0.68   9.00 640.14     0.997    0.73
```

We see that this new effect is not significant. We are not constrained to taking means. We might consider various quantiles or measures of spread as potential compositional variables.

Much remains to be investigated with this dataset. We have only used the simplest of error structures and we should investigate whether the random effects may also depend on some of the other covariates.

**Further Reading:** The classical approach to random effects can be found in many older books such as Snedecor and Cochran (1989) or Scheffé (1959). More recent books such as Searle et al. (1992) also focus on the ANOVA approach. A wide range of models are explicitly considered in Milliken and Johnson (1992). Multilevel models are covered in Goldstein (1995), Raudenbush and Bryk (2002) and Gelman and Hill (2006). The predecessor to the `lme4` package was `nlme` which is described in Pinheiro and Bates (2000), but the book still contains much general material of interest.

## Exercises

1. The `denim` dataset concerns the amount of waste in material cutting for a jeans manufacturer due to five suppliers.
  - (a) Plot the data and comment.
  - (b) Fit the linear fixed effects model. Is the operator significant?
  - (c) Make a useful diagnostic plot for this model and comment.
  - (d) Analyze the data with supplier as a random effect. What are the estimated standard deviations of the effects?
  - (e) Test the significance of the supplier term.
  - (f) Compute confidence intervals for the random effect SDs.
  - (g) Locate two outliers and remove them from the data. Repeat the fitting, testing and computation of the confidence intervals, commenting on the differences you see from the complete data.

- (h) Estimate the effect of each supplier. If only one supplier will be used, choose the best.
2. The coagulation dataset comes from a study of blood coagulation times. Twenty-four animals were randomly assigned to four different diets and the samples were taken in a random order.
- Plot the data and comment.
  - Fit a fixed effects model and construct a prediction together with a 95% prediction interval for the response of a new animal assigned to diet D.
  - Now fit a random effects model using REML. A new animal is assigned to diet D. Predict the blood coagulation time for this animal along with a 95% prediction interval.
  - A new diet is given to a new animal. Predict the blood coagulation time for this animal along with a 95% prediction interval
  - A new diet is given to the first animal in the dataset. Predict the blood coagulation time for this animal with a prediction interval. You may assume that the effects of the initial diet for this animal have washed out.
3. The eggprod dataset concerns an experiment where six pullets were placed into each of 12 pens. Four blocks were formed from groups of three pens based on location. Three treatments were applied. The number of eggs produced was recorded.
- Make suitable plots of the data and comment.
  - Fit a fixed effects model for the number of eggs produced with the treatments and blocks as predictors. Determine the significance of the two predictors and perform a basic diagnostic check.
  - Fit a model for the number of eggs produced with the treatments as fixed effects and the blocks as random effects. Which treatment is best in terms of maximizing production according to the model? Are you sure it is better than other two treatments?
  - Use the Kenward-Roger approximation for an *F*-test to check for differences between the treatments. How does the result compare to the fixed effects result?
  - Perform the same test but using a bootstrap method. How do the results compare?
  - Test for the significance of the blocks. Does the outcome agree with the fixed effects result?
4. Data on the cutoff times of lawnmowers may be found in the dataset lawn. Three machines were randomly selected from those produced by manufacturers A and B. Each machine was tested twice at low speed and high speed.
- Make plots of the data and comment.
  - Fit a fixed effects model for the cutoff time response using just the main effects of the three predictors. Explain why not all effects can be estimated.

- (c) Fit a mixed effects model with manufacturer and speed as main effects along with their interaction and machine as a random effect. If the same machine were tested at the same speed, what would be the SD of the times observed? If different machines were sampled from the same manufacturer and tested at the same speed once only, what would be the SD of the times observed?
- (d) Test whether the interaction term of the model can be removed. If so, go on to test the two main fixed effects terms.
- (e) Check whether there is any variation between machines.
- (f) Fit a model with speed as the only fixed effect and manufacturer as a random effect with machines also as a random effect nested within manufacturer. Compare the variability between machines with the variability between manufacturers.
- (g) Construct bootstrap confidence intervals for the terms of the previous model. Discuss whether the variability can be ascribed solely to manufacturers or to machines.
5. A number of growers supply broccoli to a food processing plant. The plant instructs the growers to pack the broccoli into standard-size boxes. There should be 18 clusters of broccoli per box. Because the growers use different varieties and methods of cultivation, there is some variation in the cluster weights. The plant manager selected three growers at random and then four boxes at random supplied by these growers. Three clusters were selected from each box. The data may be found in the `broccoli` dataset. The weight in grams of the cluster is given.
- (a) Plot the data and comment on the nature of the variation seen.
- (b) Compute the mean weights within growers. Compute the mean weights within boxes.
- (c) Fit an appropriate mixed effects model. Comment on how the variation is assigned to the possible sources.
- (d) Test whether there may be no variation attributable to growers.
- (e) Test whether there may be no variation attributable to boxes.
- (f) Compute confidence intervals for the SD components in your full model.
6. An experiment was conducted to select the supplier of raw materials for production of a component. The breaking strength of the component was the objective of interest. Four suppliers were considered. The four operators can only produce one component each per day. A latin square design is used and the data is presented in `breaking`.
- (a) Plot the data and interpret.
- (b) Fit a fixed effects model for the main effects. Determine which factors are significant.
- (c) Fit a mixed effects model with operators and days as random effects but the suppliers as fixed effects. Why is this a natural choice of fixed and random effects? Which supplier results in the highest breaking point? What is the nature of the variation between operators and days?

- (d) Test the operator and days effects.
  - (e) Test the significance of the supplier effect.
  - (f) For the best choice of supplier, predict the proportion of components produced in the future that will have a breaking strength less than 1000.
7. An experiment was conducted to optimize the manufacture of semiconductors. The semicond data has the resistance recorded on the wafer as the response. The experiment was conducted during four different time periods denoted by ET and three different wafers during each period. The position on the wafer is a factor with levels 1 to 4. The Grp variable is a combination of ET and wafer. Analyze the data as a split plot experiment where ET and position are considered as fixed effects. Since the wafers are different in experimental time periods, the Grp variable should be regarded as the block or group variable.
- (a) Plot the data appropriately and comment.
  - (b) Fit a fixed effects model with an interaction between ET and position (no other predictors). What terms are significant? What is wrong with using this model to make inference about these predictors?
  - (c) Fit a model appropriate to the split plot design used here. Comment on the relative variation between and within the groups (Grp).
  - (d) Test for the effect of position.
  - (e) Which level of ET results in the highest resistance? Can we be sure that this is really better than the second highest level?
  - (f) Make a plot of the residuals and fitted values and interpret. Make a QQ plot and comment.
8. Redo the Junior Schools Project data analysis in the text with the final year English score as the response. Highlight any differences from the analysis of the final year Math scores.
9. An experiment was conducted to determine the effect of recipe and baking temperature on chocolate cake quality. Fifteen batches of cake mix for each recipe were prepared. Each batch was sufficient for six cakes. Each of the six cakes was baked at a different temperature which was randomly assigned. Several measures of cake quality were recorded of which breaking angle was just one. The dataset is presented as choccake.
- (a) Plot the data and comment.
  - (b) Fit linear model with an interaction between recipe and temperature as fixed effects and no random effects. Which terms are significant? Why is this analysis unreliable?
  - (c) Fit a mixed effects model that takes account of the batch structure, identifying the design type. Compare the temperature effect (minimum to maximum) with the likely difference between batches. How do they compare?
  - (d) Test for a recipe effect.
  - (e) Check the following diagnostic plots and comment.
    - i. The residuals against fitted values.

- ii. A QQ plot of the residuals.
- iii. A QQ plot of the batch random effects.

## Repeated Measures and Longitudinal Data

---

In repeated measures designs, there are several individuals (or units) and measurements are taken repeatedly on each individual. When these repeated measurements are taken over time, it is called a *longitudinal* study or, in some applications, a *panel* study. Typically various covariates concerning the individual are recorded and the interest centers on how the response depends on the covariates over time. Often it is reasonable to believe that the response of each individual has several components: a fixed effect, which is a function of the covariates; a random effect, which expresses the variation between individuals; and an error, which is due to measurement or unrecorded variables.

Suppose each individual has response  $y_i$ , a vector of length  $n_i$  which is modeled conditionally on the random effects  $\gamma_i$  as:

$$y_i | \gamma_i \sim N(X_i \beta + Z_i \gamma_i, \sigma^2 \Lambda_i)$$

Notice this is very similar to the model used in the previous chapter with the exception of allowing the errors to have a more general covariance  $\Lambda_i$ . As before, we assume that the random effects  $\gamma_i \sim N(0, \sigma^2 D)$  so that:

$$y_i \sim N(X_i \beta, \Sigma_i)$$

where  $\Sigma_i = \sigma^2(\Lambda_i + Z_i D Z_i^T)$ . Now suppose we have  $M$  individuals and we can assume the errors and random effects between individuals are uncorrelated, then we can combine the data as:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_M \end{bmatrix} \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_M \end{bmatrix} \quad \gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \dots \\ \gamma_M \end{bmatrix}$$

and  $\tilde{D} = diag(D, D, \dots, D)$ ,  $Z = diag(Z_1, Z_2, \dots, Z_M)$ ,  $\Sigma = diag(\Sigma_1, \Sigma_2, \dots, \Sigma_M)$  and  $\Lambda = diag(\Lambda_1, \Lambda_2, \dots, \Lambda_M)$ . Now we can write the model as

$$y \sim N(X\beta, \Sigma) \quad \Sigma = \sigma^2(\Lambda + Z\tilde{D}Z^T)$$

The log-likelihood for the data is then computed as previously and estimation, testing, standard errors and confidence intervals all follow using standard likelihood theory as before. There is no strong distinction between the methodology used in this and the previous chapter.

This general structure encompasses a wide range of possible models for different types of data. We explore some of these in the following three examples:

## 11.1 Longitudinal Data

The Panel Study of Income Dynamics (PSID), begun in 1968, is a longitudinal study of a representative sample of U.S. individuals described in Hill (1992). The study is conducted at the Survey Research Center, Institute for Social Research, University of Michigan, and is still continuing. There are currently 8700 households in the study and many variables are measured. We chose to analyze a random subset of this data, consisting of 85 heads of household who were aged 25–39 in 1968 and had complete data for at least 11 of the years between 1968 and 1990. The variables included were annual income, gender, years of education and age in 1968:

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

```
head(psid)
```

	age	educ	sex	income	year	person
1	31	12	M	6000	68	1
2	31	12	M	5300	69	1
3	31	12	M	5200	70	1
4	31	12	M	6900	71	1
5	31	12	M	7500	72	1

Now plot the data:

```
library(dplyr)
psid20 <- filter(psid, person <= 20)
library(ggplot2)
ggplot(psid20, aes(x=year, y=income)) + geom_line() + facet_wrap(~ person)
```

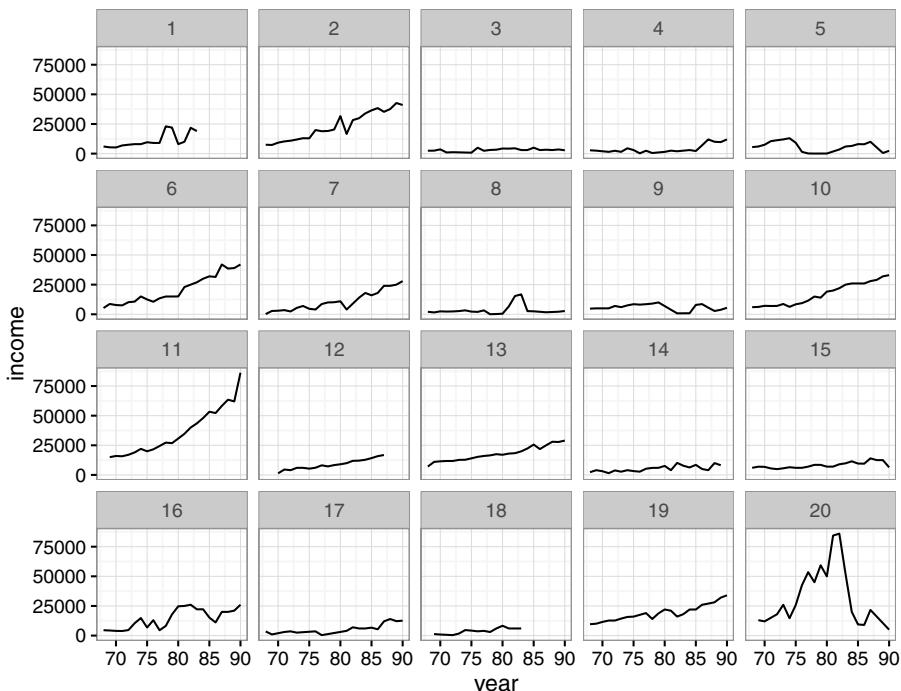


Figure 11.1 The first 20 subjects in the PSID data. Income is shown over time.

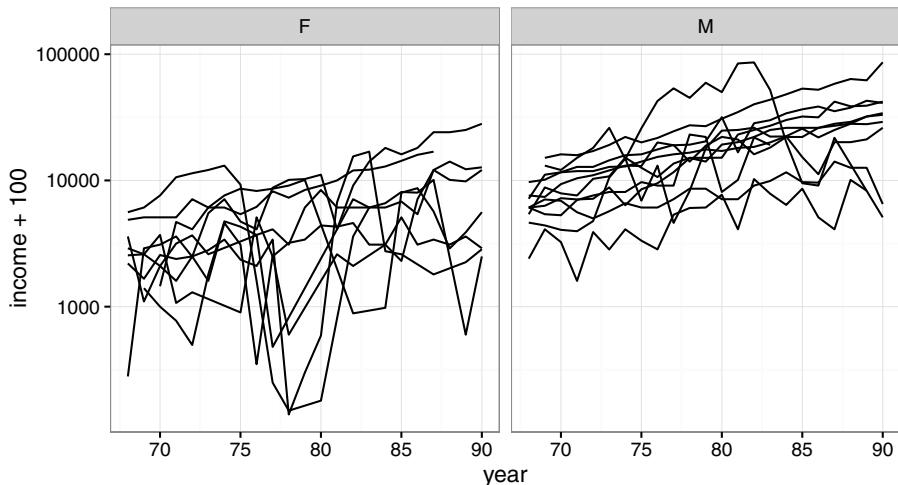


Figure 11.2 *Income change in the PSID data grouped by sex.*

The first 20 subjects are shown in Figure 11.1. We see that some individuals have a slowly increasing income, typical of someone in steady employment in the same job. Other individuals have more erratic incomes. We can also show how the incomes vary by sex. Income is more naturally considered on a log-scale:

```
ggplot(psid20, aes(x=year, y=income+100, group=person)) +geom_line() +
  facet_wrap(~ sex) + scale_y_log10()
```

See Figure 11.2. We added \$100 to the income of each subject to remove the effect of some subjects having very low incomes for short periods of time. These cases distorted the plots without the adjustment. We see that men's incomes are generally higher and less variable while women's incomes are more variable, but are perhaps increasing more quickly. We could fit a line to each subject starting with the first:

```
lmod <- lm(log(income) ~ I(year-78), subset=(person==1), psid)
coef(lmod)
```

(Intercept)	I(year - 78)
9.399957	0.084267

We have centered the predictor at the median value so that the intercept will represent the predicted log income in 1978 and not the year 1900 which would be nonsense. We now fit a line for all the subjects and plot the results:

```
library(lme4)
ml <- lmList(log(income) ~ I(year-78) | person, psid)
intercepts <- sapply(ml, coef)[1,]
slopes <- sapply(ml, coef)[2,]
```

The `lmList` command fits a linear model to each group within the data, here specified by `person`. A list of linear models, one for each group, is returned from which we extract the intercepts and slopes.

```
plot(intercepts, slopes, xlab="Intercept", ylab="Slope")
psex <- psid$sex[match(1:85, psid$person)]
boxplot(split(slopes, psex))
```

In the first panel of Figure 11.3, we see how the slopes relate to the intercepts — there is little correlation. This means we can test incomes and income growths separately. In the second panel, we compare the income growth rates where we see these as higher and more variable for women compared to men. We can test the difference in

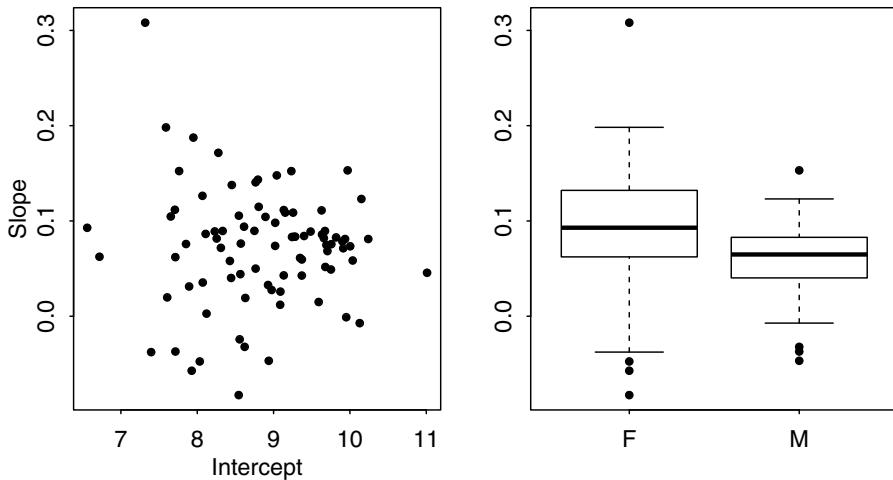


Figure 11.3 *Slopes and intercepts for the individual income growth relationships are shown on the left. A comparison of income growth rates by sex is shown on the right.*

income growth rates for men and women:

```
t.test(slopes[psex=="M"], slopes[psex=="F"])
```

Welch Two Sample t-test

```
data: slopes[psex == "M"] and slopes[psex == "F"]
t = -2.3786, df = 56.736, p-value = 0.02077
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.0591687 -0.0050773
sample estimates:
mean of x mean of y
0.056910 0.089033
```

We see that women have a significantly higher growth rate than men. We can also compare the incomes at the intercept (which is 1978):

```
t.test(intercepts[psex=="M"], intercepts[psex=="F"])
```

Welch Two Sample t-test

```
data: intercepts[psex == "M"] and intercepts[psex == "F"]
t = 8.2199, df = 79.719, p-value = 3.065e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.87388 1.43222
sample estimates:
mean of x mean of y
9.3823 8.2293
```

We see that men have significantly higher incomes.

This is an example of a *response feature* analysis. It requires choosing an important characteristic. We have chosen two here: the slope and the intercept. For many datasets, this is not an easy choice and at least some information is lost by doing this.

Response feature analysis is attractive because of its simplicity. By extracting a univariate response for each individual, we are able to use a wide array of well-known statistical techniques. However, it is not the most efficient use of the data as all the additional information besides the chosen response feature is discarded. Notice that having additional data on each subject would be of limited value.

Suppose that the income change over time can be partly predicted by the subject's age, sex and educational level. We do not expect a perfect fit. The variation may be partitioned into two components. Clearly there are other factors that will affect a subject's income. These factors may cause the income to be generally higher or lower or they may cause the income to grow at a faster or slower rate. We can model this variation with a random intercept and slope, respectively, for each subject. We also expect that there will be some year-to-year variation within each subject. For simplicity, let us initially assume that this error is homogeneous and uncorrelated, that is,  $\Lambda_i = I$ . We also center the year to aid interpretation as before. We may express these notions in the model:

```
library(lme4)
psid$cyear <- psid$year - 78
mmod <- lmer(log(income) ~ cyear*sex + age+educ+(cyear|person), psid)
```

This model can be written as:

$$\begin{aligned} \log(\text{income})_{ij} = & \mu + \beta_y \text{year}_i + \beta_s \text{sex}_j + \beta_{ys} \text{sex}_j \times \text{year}_i + \beta_e \text{educ}_j + \beta_a \text{age}_j \\ & + \gamma_j^0 + \gamma_j^1 \text{year}_i + \varepsilon_{ij} \end{aligned}$$

where  $i$  indexes the year and  $j$  indexes the individual. We have:

$$\begin{pmatrix} \gamma_k^0 \\ \gamma_k^1 \end{pmatrix} \sim N(0, \sigma^2 D)$$

The model summary is:

```
summary(mmod, digits=3)
Fixed Effects:
  coef.est  coef.se
(Intercept) 6.674    0.543
cyear       0.085    0.009
sexM        1.150    0.121
age         0.011    0.014
educ        0.104    0.021
cyear:sexM -0.026    0.012

Random Effects:
 Groups   Name       Std.Dev. Corr
 person   (Intercept) 0.531
          cyear      0.049    0.187
 Residual            0.684
---
number of obs: 1661, groups: person, 85
```

```
AIC = 3839.8, DIC = 3751.2
deviance = 3785.5
```

Let's start with the fixed effects. We see that income increases about 10% for each additional year of education. We see that age does not appear to be significant. For females, the reference level in this example, income increases about 8.5% a year, while for men, it increases about  $8.5 - 2.6 = 5.9\%$  a year. We see that, for this data, the incomes of men are  $\exp(1.15) = 3.16$  times higher.

We know the mean for males and females, but individuals will vary about this. The standard deviation for the intercept and slope are 0.531 and 0.049 ( $\sigma\sqrt{D_{11}}$  and  $\sigma\sqrt{D_{22}}$ ), respectively. These have a correlation of 0.189 ( $\text{cor}(\gamma^0, \gamma^1)$ ). Finally, there is some additional variation in the measurement not so far accounted for having standard deviation of 0.684 ( $sd(\epsilon_{ijk})$ ). We see that the variation in increase in income is relatively small while the variation in overall income between individuals is quite large. Furthermore, given the large residual variation, there is a large year-to-year variation in incomes.

We can test the fixed effect terms for significance. We use the Kenward-Roger adjusted *F*-test:

```
library(pbkrttest)
mmod <- lmer(log(income) ~ cyear*sex + age+educ+(cyear|person),psid,
  ↪ REML=FALSE)
mmodr <- lmer(log(income) ~ cyear + sex + age+educ+(cyear|person),psid,
  ↪ REML=FALSE)
KRmodcomp(mmod, mmodr)

F-test with Kenward-Roger approximation; computing time: 0.30 sec.
large : log(income) ~ cyear + sex + age + educ + (cyear | person) + cyear:sex
small : log(income) ~ cyear + sex + age + educ + (cyear | person)
      stat   ndf   ddf F.scaling p.value
Ftest 4.61 1.00 81.33          1  0.035
```

We have tested the interaction term between year and sex as this is the most complex term in the model. We see that this term is marginally significant so there is no justification to simplify the model by removing this term. Female incomes are increasing faster than male incomes.

We could test the random effect terms using perhaps the parametric bootstrap method. It is less trouble to create confidence intervals for all the parameters:

```
confint(mmod, method="boot")
              2.5 %    97.5 %
sd_(Intercept)|person 0.440965 0.6095268
cor_cyear.(Intercept)|person -0.044677 0.4486294
sd_cyear|person        0.039271 0.0582838
sigma                 0.658930 0.7087268
(Intercept)           5.571034 7.7676101
cyear                 0.067160 0.1027455
sexM                  0.899570 1.3772171
age                   -0.017808 0.0365997
educ                  0.064944 0.1530431
cyear:sexM            -0.051526 -0.0028899
```

We see that all the standard deviations are clearly well above zero. There might be a case for removing the correlation between the intercept and slope but this term is difficult to interpret and little would be gained from removing it. It is simpler just to leave it in.

There is a wider range of possible diagnostic plots that can be made with longitudinal data than with a standard linear model. In addition to the usual residuals, there are random effects to be examined. We may wish to break the residuals down by sex as seen in the QQ plots in Figure 11.4:

```
diagd <- fortify(mmod)
ggplot(diagd, aes(sample=.resid)) + stat_qq() + facet_grid(~sex)
```

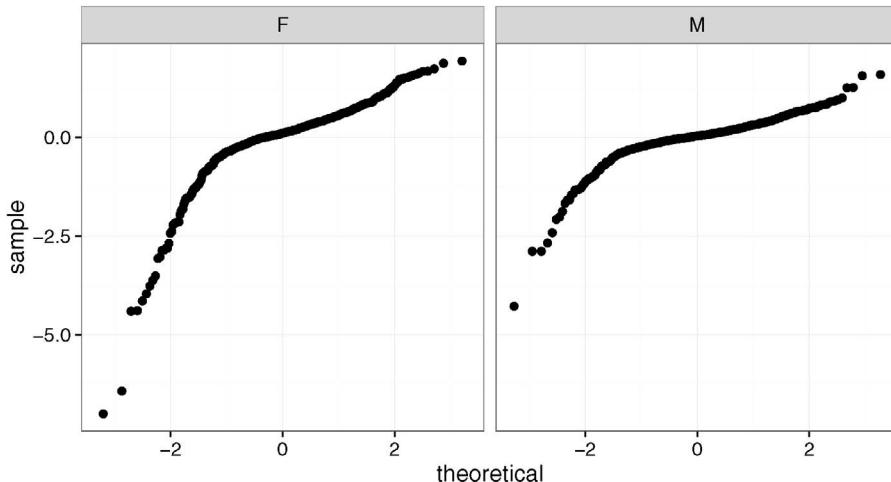


Figure 11.4 *QQ plots by sex.*

We see that the residuals are not normally distributed, but have a long tail for the lower incomes. We should consider changing the log transformation on the response. Furthermore, we see that there is greater variance in the female incomes. This suggests a modification to the model. We can make the same plot broken down by subject although there will be rather too many plots to be useful.

Plots of residuals and fitted values are also valuable. We have broken education into three levels: less than high school, high school or more than high school:

```
diagd$edulevel <- cut(psid$educ,c(0,8.5,12.5,20), labels=c("lessHS",
  ↪ "HS", "moreHS"))
ggplot(diagd, aes(x=.fitted,y=.resid)) + geom_point(alpha=0.3) + geom_
  ↪ hline(yintercept=0) + facet_grid(~ edulevel) + xlab("Fitted") +
  ↪ ylab("Residuals")
```

See Figure 11.5. Again, we can see evidence that a different response transformation should be considered. Plots of the random effects would also be useful here.

## 11.2 Repeated Measures

The acuity of vision for seven subjects was tested. The response is the lag in milliseconds between a light flash and a response in the cortex of the eye. Each eye is tested at four different powers of lens. An object at the distance of the second number appears to be at distance of the first number. The data is given in Table 11.1. The data comes from Crowder and Hand (1990) and was also analyzed by Lindsey (1999).

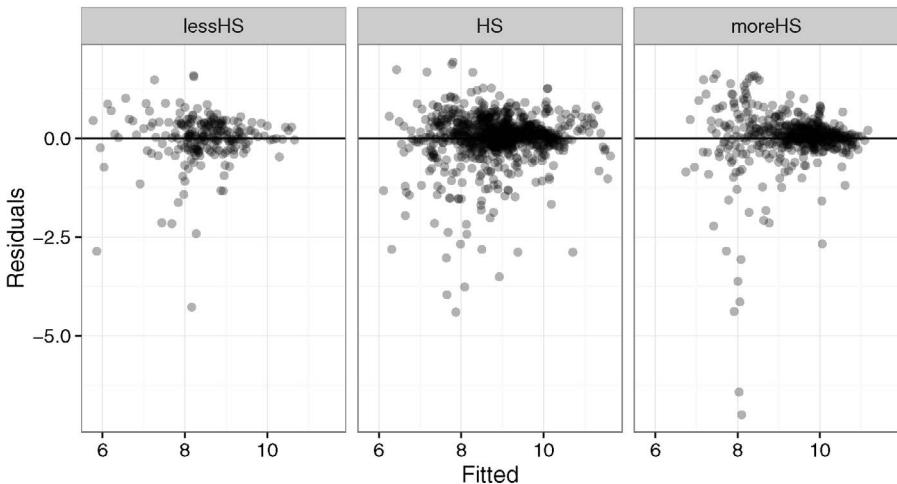


Figure 11.5 *Residuals vs. fitted plots for three levels of education: less than high school on the left, high school in the middle and more than high school on the right.*

		Power							
		6/6	6/18	6/36	6/60	6/6	6/18	6/36	6/60
		Left				Right			
		116	119	116	124	120	117	114	122
		110	110	114	115	106	112	110	110
		117	118	120	120	120	120	120	124
		112	116	115	113	115	116	116	119
		113	114	114	118	114	117	116	112
		119	115	94	116	100	99	94	97
		110	110	105	118	105	105	115	115

Table 11.1 *Visual acuity of seven subjects measured in milliseconds of lag in responding to a light flash. The power of the lens causes an object six feet in distance to appear at a distance of 6, 18, 36 or 60 feet.*

We start by making some plots of the data. We create a numerical variable representing the power to complement the existing factor so that we can see how the acuity changes with increasing power:

```
data(vision, package="faraway")
vision$npower <- rep(1:4, 14)
ggplot(vision, aes(y=acuity, x=npower, linetype=eye)) + geom_line() +
  facet_wrap(~ subject, ncol=4) + scale_x_continuous("Power",
  breaks=1:4, labels=c("6/6", "6/18", "6/36", "6/60"))
```

See Figure 11.6. There is no apparent trend or difference between right and left eyes.

However, individual #6 appears anomalous with a large difference between the eyes. It also seems likely that the third measurement on the left eye is in error for this individual.

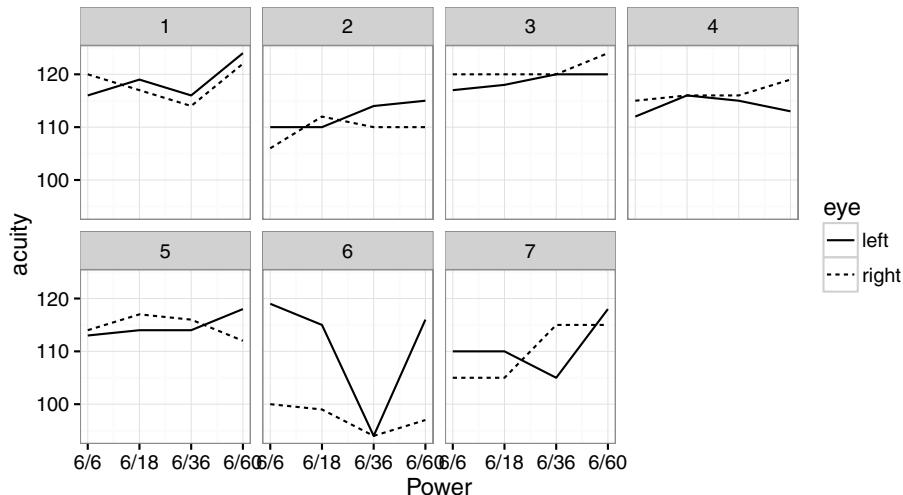


Figure 11.6 *Visual acuity profiles. The left eye is shown as a solid line and the right as a dashed line. The four powers of lens displayed are 6/6, 6/18, 6/36 and 6/60.*

We must now decide how to model the data. The power is a fixed effect. In the model below, we have treated it as a nominal factor, but we could try fitting it in a quantitative manner. The subjects should be treated as random effects. Since we do not believe there is any consistent right-left eye difference between individuals, we should treat the eye factor as nested within subjects. We start with this model:

```
mmod <- lmer(acuity~power + (1|subject) + (1|subject:eye), vision)
```

Note that if we did believe there was a consistent left vs. right eye effect, we would have used a fixed effect, putting eye in place of  $(1|\text{subject}:\text{eye})$ .

We can write this (nested) model as:

$$y_{ijk} = \mu + p_j + s_i + e_{ik} + \varepsilon_{ijk}$$

where  $i = 1, \dots, 7$  runs over individuals,  $j = 1, \dots, 4$  runs over power and  $k = 1, 2$  runs over eyes. The  $p_j$  term is a fixed effect, but the remaining terms are random. Let  $s_i \sim N(0, \sigma_s^2)$ ,  $e_{ik} \sim N(0, \sigma_e^2)$  and  $\varepsilon_{ijk} \sim N(0, \sigma^2 \Sigma)$  where we take  $\Sigma = I$ . The summary output is:

```
summary(mmod)
```

Fixed Effects:

	coef.est	coef.se
(Intercept)	112.64	2.23
power6/18	0.79	1.54
power6/36	-1.00	1.54
power6/60	3.29	1.54

```

Random Effects:
Groups      Name      Std.Dev.
subject:eye (Intercept) 3.21
subject     (Intercept) 4.64
Residual            4.07
---
number of obs: 56, groups: subject:eye, 14; subject, 7
AIC = 342.7, DIC = 349.6
deviance = 339.2

```

We see that the estimated standard deviation for subjects is 4.64 and that for eyes for a given subject is 3.21. The residual standard deviation is 4.07. The random effects structure we have used here induces a correlation between measurements on the same subject and another between measurements on the same eye. We can compute these two correlations, respectively, as:

```

4.64^2/(4.64^2+3.21^2+4.07^2)
[1] 0.44484
(4.64^2+3.21^2)/(4.64^2+3.21^2+4.07^2)
[1] 0.65774

```

As we might expect, there is a stronger correlation between observations on the same eye than between the left and right eyes of the same individual.

We can check for a power effect using a Kenward-Roger adjusted  $F$ -test:

```

library(pbkrttest)
mmodr <- lmer(acuity~power+(1|subject)+(1|subject:eye), vision, REML=
  ↪ FALSE)
nmodr <- lmer(acuity~1+(1|subject)+(1|subject:eye), vision, REML=FALSE)
KRmodcomp(mmodr, nmodr)
F-test with Kenward-Roger approximation; computing time: 0.16 sec.
large : acuity ~ power + (1 | subject) + (1 | subject:eye)
small : acuity ~ 1 + (1 | subject) + (1 | subject:eye)
      stat   ndf   ddf F.scaling p.value
Ftest 2.83  3.00 39.00      1  0.051

```

We see the result is just above the 5% level. We might expect some trend in acuity with power, but the estimated effects do not fit with this trend. While acuity is greatest at the highest power, 6/60, it is lowest for the second highest power, 6/36. A look at the data makes one suspect the measurement made on the left eye of the sixth subject at this power. If we omit this observation and refit the model, we find:

```

mmodr <- lmer(acuity~power+(1|subject)+(1|subject:eye), vision, REML=
  ↪ FALSE, subset=-43)
nmodr <- lmer(acuity~1+(1|subject)+(1|subject:eye), vision, REML=FALSE,
  ↪ subset=-43)
KRmodcomp(mmodr, nmodr)
F-test with Kenward-Roger approximation; computing time: 0.15 sec.
large : acuity ~ power + (1 | subject) + (1 | subject:eye)
small : acuity ~ 1 + (1 | subject) + (1 | subject:eye)
      stat   ndf   ddf F.scaling p.value
Ftest 3.6  3.0 38.0      1  0.022

```

Now the power effect is significant, but it appears this is due to an effect at the highest power only. We can check that the highest power has a higher acuity than the average of the first three levels by using Helmert contrasts:

```

op <- options(contrasts=c("contr.helmert", "contr.poly"))
mmodr <- lmer(acuity~power+(1|subject)+(1|subject:eye), vision, subset
  ↪ ==-43)

```

```
summary(mmodr)
```

Fixed Effects:

	coef.est	coef.se
(Intercept)	113.79	1.76
power1	0.39	0.54
power2	0.04	0.32
power3	0.71	0.22

By looking at the standard errors relative to the effect sizes, we can see that only the third contrast is of significance. We remember to reset the contrasts back to the default or subsequent output will be surprising:

```
options(op)
```

The Helmert contrast matrix is

```
contr.helmert(4)
```

	[,1]	[,2]	[,3]
1	-1	-1	-1
2	1	-1	-1
3	0	2	-1
4	0	0	3

We can see that the third contrast (column) represents the difference between the average of the first three levels and the fourth level, scaled by a factor of three. In the output, we can see that this is significant while the other two contrasts are not.

We finish with some diagnostic plots. The residuals and fitted values and the QQ plot of random effects for the eyes are shown in Figure 11.7:

```
plot(resid(mmodr) ~ fitted(mmodr), xlab="Fitted", ylab="Residuals")
abline(h=0)
qqnorm(ranef(mmodr)$"subject:eye"[[1]], main="")
```

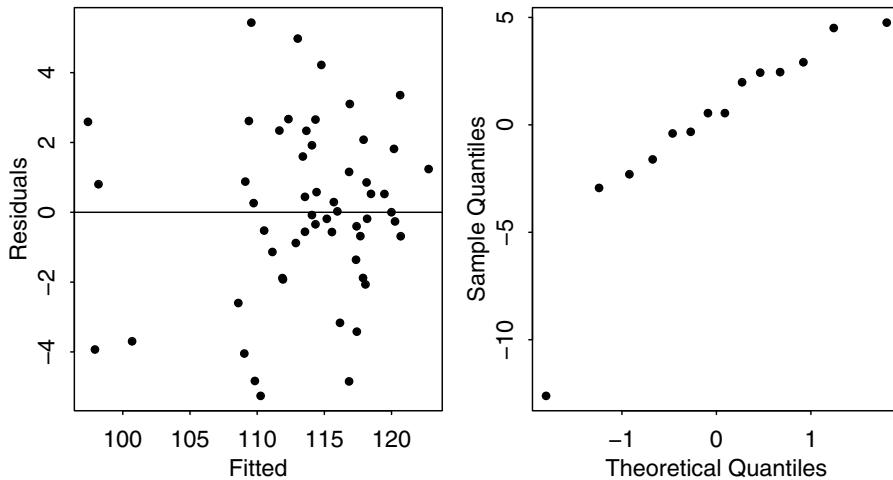


Figure 11.7 *Residuals vs. fitted plot* is shown on the left and a *QQ plot* of the random effects for the eyes is shown on the right.

The outlier corresponds to the right eye of subject #6. For further analysis, we should consider dropping subject #6. There are only seven subjects altogether, so we would

certainly regret losing any data, but this may be unavoidable. Ultimately, we may need more data to make definite conclusions.

### 11.3 Multiple Response Multilevel Models

In Section 10.10, we analyzed some data from the Junior Schools Project. In addition to a math test, students also took a test in English. Although it would be possible to analyze the English test results in the same way that we analyzed the math scores, additional information may be obtained from analyzing them simultaneously. Hence we view the data as having a bivariate response with English and math scores for each student. The student is a nested factor within the class which is in turn nested within the school. We express the multivariate response for each individual by introducing an additional level of nesting at the individual level. So we might view this as just another nested model except that there is a fixed subject effect associated with this lowest level of nesting.

We set up the data in a format with one test score per line with an indicator subject identifying which type of test was taken. We scale the English and math test scores by their maximum possible values, 40 and 100, respectively, to aid comparison:

```
data(jsp, package="faraway")
jspr <- jsp[jsp$year==2,]
mjspr <- data.frame(rbind(jspr[,1:6], jspr[,1:6]), subject=factor(rep(
  c("english", "math"), c(953, 953))), score=c(jspr$english/100,
  jspr$math/40))
```

We can examine the relationship between subject, gender and scores, as seen in Figure 11.8:

```
ggplot(mjspr, aes(x=raven, y=score)) + geom_jitter(alpha=0.25) + facet_
  ~ grid(gender ~ subject)
```

We now fit a model for the data that includes all the variables of interest that incorporates some of the interactions that we suspect might be present:

```
mjspr$craven <- mjspr$raven - mean(mjspr$raven)
mmod <- lmer(score ~ subject*gender + craven*subject + social + (1|
  school) + (1|school:class) + (1|school:class:id), mjspr)
```

The model being fit for school  $i$ , class  $j$ , student  $k$  in subject  $l$  is:

$$\begin{aligned} score_{ijkl} = & \text{subject}_l + \text{gender}_k + \text{raven}_k + \text{social}_k + (\text{subject} \times \text{gender})_{lk} + \\ & (\text{raven} \times \text{subject})_{lk} + \text{school}_i + \text{class}_j + \text{student}_k + \varepsilon_{ijkl} \end{aligned}$$

where the Raven score has been mean centered and school, class and student are random effects with the other terms, apart from  $\varepsilon$ , being fixed effects. The summary output:

```
summary(mmod)
```

Fixed Effects:

	coef.est	coef.se
(Intercept)	0.442	0.026
subjectmath	0.367	0.008
gendergirl	0.063	0.010
craven	0.017	0.001
social2	0.014	0.027

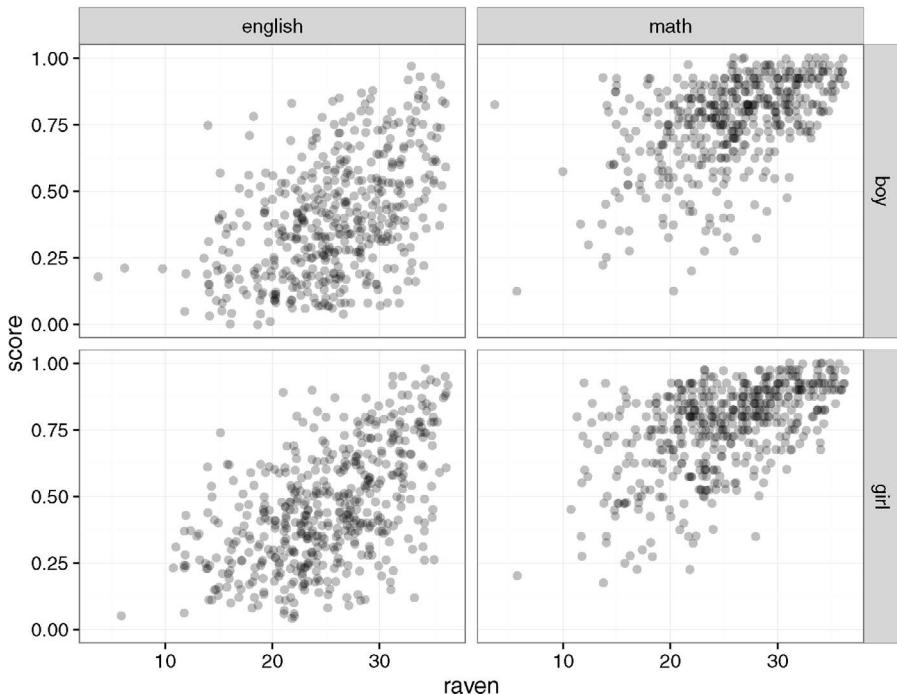


Figure 11.8 Scores on test compared to Raven score for subjects and genders.

social3	-0.021	0.029
social4	-0.071	0.026
social5	-0.050	0.029
social6	-0.088	0.031
social7	-0.099	0.032
social8	-0.082	0.042
social9	-0.047	0.027
subjectmath:gendergirl	-0.059	0.011
subjectmath:craven	-0.004	0.001

#### Random Effects:

Groups	Name	Std.Dev.
school:class:id	(Intercept)	0.101
school:class	(Intercept)	0.024
school	(Intercept)	0.047
Residual		0.117

---

number of obs: 1906, groups: school:class:id, 953; school:class, 90; school, 48  
AIC = -1705.6, DIC = -1951.1  
deviance = -1846.4

Starting with the fixed effects, we see that the math subject scores were about 37% higher than the English scores. This may just reflect the grading scale and difficulty of the test and so perhaps nothing in particular should be concluded from this except, of course, that it is necessary to have this term in the model to control for this difference.

Since gender has a significant interaction with subject, we must interpret these terms together. We see that on the English test, which is the reference level, girls score 6.3% higher than boys. On the math test, the difference is  $6.3 - 5.9 = 0.4\%$  which is negligible. We see that the scores are strongly related to the entering Raven score although the relation is slightly less strong for math than English (slope is 0.017 for English but  $0.017 - 0.004 = 0.013$  for math). We also see the declining performance as we move down the social class scale as we found in the previous analysis.

We can test the fixed effects using an  $F$ -test incorporating the Kenward-Roger  $F$ -test degrees of freedom adjustment:

```
library(pbkrtest)
mmod <- lmer(score ~ subject*gender+craven+subject+social+ (1|school)
  ↪ +(1|school:class)+(1|school:class:id), mjspr, REML=FALSE)
mmodr <- lmer(score ~ subject*gender+craven+subject+social+ (1|school)
  ↪ +(1|school:class)+(1|school:class:id), mjspr, REML=FALSE)
KRmodcomp(mmod, mmodr)
```

$F$ -test with Kenward-Roger approximation; computing time: 0.63 sec.

```
large : score ~ subject + gender + craven + social + (1 | school) + (1 |
  school:class) + (1 | school:class:id) + subject:gender +
  subject:craven
small : score ~ subject * gender + craven + subject + social + (1 | school) +
  (1 | school:class) + (1 | school:class:id)
  stat ndf ddf F.scaling p.value
Ftest 16 1 950 1 6.9e-05
```

Here we test for a subject by gender interaction. We can see that this effect is strongly statistically significant.

Moving to the random effects, we can see from Figure 11.9 that the standard deviation of the residual error in the math scores is smaller than that seen in the English scores. Perhaps this can be ascribed to the greater ease of consistent grading of math assignments or perhaps just greater variation is to be expected in English performance. The correlation between the English and math scores after adjusting for the other effects is also of interest. The last two terms in the model,  $student_k + \epsilon_{ijkl}$ , represent a  $2 \times 2$  covariance matrix for the residual scores for the two tests. We can compute the correlation as:

```
0.101^2/(0.101^2+0.117^2)
[1] 0.427
```

giving a moderate positive correlation between the scores. Various diagnostic plots can be made. An interesting one is:

```
diagd <- fortify(mmod)
ggplot(diagd, aes(x=.fitted,y=.resid)) + geom_point(alpha=0.3) + geom_
  ↪ hline(yintercept=0) + facet_grid(~ subject) + xlab("Fitted") +
  ↪ ylab("Residuals")
```

as seen in Figure 11.9. There is somewhat greater variance in the verbal scores. The truncation effect of the maximum score is particularly visible for the math scores.

**Further Reading:** Longitudinal data analysis is explicitly covered in books by Verbeke and Molenberghs (2000), Fitzmaurice et al. (2004), Gelman and Hill (2006), Diggle et al. (2013) and Frees (2004). Books stating repeated measures in the title, such as Lindsey (1999), cover much the same material.

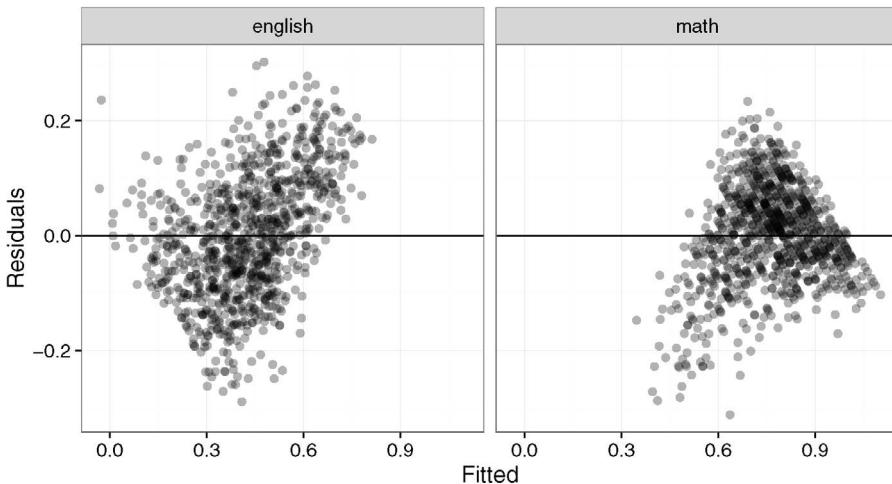


Figure 11.9 *Residuals vs. fitted plot broken down by subject.*

### Exercises

1. The `ratdrink` data consist of five weekly measurements of body weight for 27 rats. The first 10 rats are on a control treatment while 7 rats have thyroxine added to their drinking water. Ten rats have thiouracil added to their water.
  - (a) Plot the data showing how weight increases with age on a single panel, taking care to distinguish the three treatment groups. Now create a three-panel plot, one for each group. Discuss what can be seen.
  - (b) Fit a linear longitudinal model that allows for a random slope and intercept for each rat. Each group should have a different mean line. Give interpretation for the following estimates:
    - i. The fixed effect intercept term.
    - ii. The interaction between thiouracil and week.
    - iii. The intercept random effect SD.
  - (c) Check whether there is a significant treatment effect.
  - (d) Construct diagnostic plots showing the residuals against the fitted values and a QQ plot of the residuals. Interpret.
  - (e) Construct confidence intervals for the parameters of the model. Which random effect terms may not be significant? Is the thyroxine group significantly different from the control group?
2. Data on housing prices in 36 metropolitan statistical areas (MSAs) over nine years from 1986–1994 were collected and can be found in the dataset `hprice`.
  - (a) Make a plot of the data on a single panel to show how housing prices increase by year. Describe what can be seen in the plot.

- (b) Fit a linear model with the (log) house price as the response and all other variables (except `msa`) as fixed effect predictors. Which terms are statistically significant? Discuss the coefficient for time.
- (c) Make a plot that shows how per-capita income changes over time. What is the nature of the increase? Make a similar plot to show how income growth changes over time. Comment on the plot.
- (d) Create a new variable that is the per-capita income for the first time period for each MSA. Refit the same linear model but now using the initial income and not the income as it changes over time. Compare the two models.
- (e) Fit a mixed effects model that has a random intercept for each MSA. Why might this be reasonable? The rest of the model should have the same structure as in the previous question. Make a numerical interpretation of the coefficient of time in your model.
- (f) Make the following diagnostic plots and interpret: (i) Residuals vs. Fitted plot, (ii) QQ plot of the residuals, (iii) QQ plot of the random effects.
- (g) Fit a model that omits the adjacent to water and rent control predictors. Test whether this reduction in the model can be supported.
- (h) It is possible that the increase in prices may not be linear in year. Fit a model where year is treated as a factor rather than a linear term. Is this a better model than the previous choice? Make a plot of the coefficients of the time factor that shows how prices have increased over time.
- (i) Interpret the coefficients in the previous model for the initial annual income, growth and regulation predictors.
3. The `nepali` data is a subset from public health study on Nepalese children. In this question we develop a model for the weight of the child as he or she ages. You may use `mage`, `lit`, `died`, `gender` and `alive` (but not `ht`) as predictors.
- (a) Remove first the height variable and then the missing values from the dataset. You may find it cleaner to recode the sex variable to have better labels. Plot the data using two panels, one for each sex, showing how weight increases with age. Comment on the plot.
- (b) Fit a fixed effects model with weight as the response and age, sex, mother's age, literacy and other deaths in the family as predictors. Which terms are significant in this model?
- (c) Fit a mixed effects model with weight as the response. Include an interaction between age and sex and main effects in the other two predictors. Use a random intercept term for the child. What is the predicted difference in child weight between a 15- and a 25-year-old mother? What difference in weights would be expected for identical twins according to the model? Do you think this is reasonable?
- (d) Make the following diagnostic plots and interpret: (i) Residuals vs. Fitted plot, (ii) QQ plot of the residuals, (iii) QQ plot of the random effects.
- (e) Fit a model with age and mother's age as the only fixed effects and compare it to the previous model.

- (f) Now elaborate the previous model to include a random slope in age. Use AIC to choose between this model and the previous one. For your chosen model, describe how children are expected to increase in weight as they age.
- (g) Extract information about panchayat, ward, household and birth order from the `id` variable. You may find the `substring` command useful. Now fit a random intercept mixed effects model which allows for the nested random effects structure of child within household within ward within panchayat. Construct bootstrap confidence intervals to get a sense of which random effects are important. Compare the relative sizes of the random effects to the fixed effects and interpret.
4. The `attenu` data gives peak accelerations measured at various observation stations for 23 earthquakes in California. The data has been used by various workers to estimate the attenuating effect of distance on ground acceleration.
- Plot lines showing how the acceleration increases with distance for each quake. Make transformations of both axes so that the relationship is easier to see and replot.
  - Fit a mixed effects model with the transformed variables which takes account of both events and stations as random effects. Express the effect of magnitude on the acceleration.
  - Does adding a quadratic term in distance improve the model?
  - Can we remove the station variation term?
  - For a new magnitude 6 quake, predict the acceleration for up to a distance of 200 miles. Make a plot of the data and show your predicted curve on top of the data in a different color.
  - Predict how the acceleration varied for the first event where only one observation was available. Show the predicted acceleration up to 200 miles in a plot. Add the actual observation to the plot.
5. The `sleepstudy` data found in the `lme4` package describes the reaction times of subjects who are progressively sleep deprived.
- Plot the data taking care to distinguish the trajectories of the different subjects. Comment on the pattern of variation.
  - Fit a mixed effects model that describes how the reaction time varies linearly with days and allows for random variation in both the slope and intercepts of the subject lines. Under this model, would it be unusual for an individual to have a reaction time that does not increase over time?
  - Allow for quadratic effects in the previous model. Does the data support the inclusion of quadratic effects?
  - Make the following diagnostic plots and interpret: (i) Residuals vs. Fitted plot, (ii) QQ plot of the residuals, (iii) QQ plot of both random effects, (iv) a scatterplot of the random effects.
  - Identify any outlying cases and mark these on top of your initial plot. Try refitting the model without these cases and identify the largest change in the model fit.

- (f) Simulate the response under your first model and plot it. Does the simulated data look like the actual data?

---

## Chapter 12

---

# Bayesian Mixed Effect Models

---

There are a number of drawbacks to likelihood-based estimation of mixed effect models. We have seen in the previous two chapters that inference is often difficult. Furthermore, we may encounter problems even in the maximization of the likelihood leading to untrustworthy results. This leads us to explore alternative approaches to inference for these models.

Bayesian methods for inference are well established. In the past, fitting Bayesian methods had been difficult in comparison to the often one-line R command used to fit likelihood-based models. The `glm` and `lmer` commands in R are powerful in their scope. In contrast, fitting Bayesian models required more effort in setting up and implementation. In recent years, Bayesian modeling has become more accessible. It still requires more effort but one does not have to be an expert to make some progress.

It is important to understand that Bayesian inference is not simply a replacement for classical (Frequentist) inference. We can equally well address the substantive issues of the specific application but the questions and answers will be somewhat different. In particular, the classical theory of hypothesis testing does not fit well with the Bayesian approach and so it would be a mistake to expect the exact equivalence of  $p$ -values and the like. The answers will come in a different form. Historically, there have been purists who insist on an entirely Frequentist or Bayesian solution, but now many would recognize that both methods have strengths and weaknesses and one can learn much by using both.

A full explanation of Bayesian methods would take a whole book — I recommend Gelman et al. (2013) as a good starting point. Here we shall simply give a sufficient overview so that we can get started on analyses of the types of data we have seen in the previous two chapters. Likelihood-based methods as used by the `lme4` package depend on the likelihood  $l(\theta|y)$  for parameters  $\theta$  and data  $y$ . We discussed how this likelihood can be constructed in Chapter 10. We can maximize this likelihood to produce maximum likelihood estimates and use the general theory of likelihood to construct hypothesis tests and confidence intervals. We considered the parameters  $\theta$  to be fixed (but unknown) and the data  $y$  to be random. The Bayes approach considers the  $\theta$  to be random but the data  $y$ , once it has been observed, to be fixed. We have, from Bayes theorem:

$$p(\theta|y) \propto l(\theta|y)p(\theta)$$

The term  $p(\theta|y)$  is called the posterior probability (or density for continuous  $y$ ). This expresses our knowledge about the parameter values after seeing the data. We can use this probability to make decisions or predictions. The price of getting a prob-

ability on the parameters is that we must first specify a prior probability,  $p(\theta)$ , on the parameters. This should be done before looking at the data. Sometimes we know very little about these parameters. In such cases, we want an *uninformative* prior. In other cases, we may know rather more, in which case, the prior should reflect that knowledge.

The specification of the prior is the most important task that distinguishes Bayesian modeling. Many researchers feel uncomfortable about this choice because of the subjective choice involved. However, one should recognize that the Frequentist analysis is not entirely objective and deterministic either. The analyst often has to make strong assumptions to specify the model. Diagnostics can check some of these assumptions but not entirely or sometimes at all. One has to accept that there will be a subjective element to the analysis whichever method is used. It is best simply to be honest about this. In some cases, the prior enables us to use subject matter knowledge in the analysis more effectively. Even when we wish to assume as little as possible, there are some good ways to choose the prior to match this assumed lack of knowledge.

The formula linking the posterior to the likelihood and prior is rarely a means by which the posterior can be explicitly calculated. We demonstrate two distinct solutions in the chapter. The STAN language, described in Stan Development Team (2015), implements a simulation-based solution. In contrast, the INLA method, described in Rue et al. (2009), uses an approximation.

## 12.1 STAN

STAN is an example of a simulation-based approach to computing posterior densities. The idea is to set up a Markov chain of realizations of  $\theta$  such that the equilibrium distribution is the posterior distribution. These methods aim to give you a sample from the posterior distribution that you can use to answer your questions. This class of methods is known as MCMC (Monte Carlo Markov chain). We do not explain how MCMC works. The reader is referred to Gelman et al. (2013) or other texts. Here we simply present the implementation by means of examples.

The first major general software for Bayesian computing was BUGS (Bayesian inference Using Gibbs Sampling) introduced in 1989. The book by Lunn et al. (2012) provides a general introduction. BUGS developed into WinBUGS and more recently OpenBUGS. JAGS is another offshoot of the original BUGS. Around 2012, STAN became available and is described in Stan Development Team (2015). The languages used in all these software packages are quite similar although underneath there are substantial differences in the implementation. I have chosen STAN here because it was originally motivated by the fitting of just the types of models we are interested in. It is much more general but it does work particularly well for our applications. Another reason is that STAN is faster for our particular problems. Even so, this is not a strong preference and one could very reasonably choose one of the other options.

STAN will require us to venture beyond R a little, but we will be able to prepare the data, call the fitting method and process the results in R by means of the RStan package. There are R packages which provide an entirely R-based MCMC solution. For example, the `MCMCglmm` package of Hadfield (2010) offers some flexibility

in fitting Bayesian models to grouped, hierarchical and longitudinal data. However, it is inevitably limited in the functionality it provides. Of course, R is a very programmable language so it is entirely possible to implement methods to fit virtually any model. However, there are reasons beyond simply the programming effort required why this would be less than effective. Fitting Bayesian models using MCMC or related techniques is inherently computationally intensive. R has many virtues but speed is not its strongest point. This leads one to go beyond R in search of a language for describing Bayesian models that will be both more efficient and also convenient for the specification of Bayesian models.

Before proceeding, you will need to install STAN as described on the STAN website at [mc-stan.org](http://mc-stan.org). This is more work than simply installing the R package, and the details will vary according to the operating system you are running. STAN derives its speed from translating its code into C++, then compiling and running this. This means you will need a C++ compiler (but no knowledge of C++).

**One-Way ANOVA:** Let's see how we can use STAN to fit a Bayesian model to the pulp data that we have already explored with `lme4` in Chapter 10. See Figure 10.1 for a plot of the data. The model has a single factor at  $a$  levels:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, \dots, a \quad j = 1, \dots, n_i$$

where the  $\alpha$ s and  $\varepsilon$ s are independently distributed  $N(0, \sigma_\alpha^2)$  and  $N(0, \sigma_\varepsilon^2)$ , respectively. There is a single fixed effect parameter  $\mu$  and two random effect parameters  $\sigma_\alpha^2$  and  $\sigma_\varepsilon^2$ . We must specify prior distributions for these three parameters.

Here is the STAN code we have used to fit our model. It could be used for any one-way ANOVA data. Let's suppose we have  $N$  observations in total with  $J$  groups with a response and predictor named just so. You don't need to understand the STAN code to follow the output so you can skip to the discussion of the output unless you are interested in writing your own STAN programs. You should copy this code into a file called `oneway.stan` and save it into the working directory for your R session.

```
data {
  int<lower=0> N;
  int<lower=0> J;
  int<lower=1,upper=J> predictor[N];
  vector[N] response;
}
parameters {
  vector[J] eta;
  real mu;
  real<lower=0> sigmaalpha;
  real<lower=0> sigmaepsilon;
}
transformed parameters {
  vector[J] a;
  vector[N] yhat;
}
a <- mu + sigmaalpha * eta;
for (i in 1:N)
  yhat[i] <- a[predictor[i]];
```

```

}
model {
  eta ~ normal(0, 1);

  response ~ normal(yhat, sigmaepsilon);
}

```

The first block of code describes the format of the data. C++ is a much fussier language than R regarding the declaration of variables and the types of variables that are allowed. This means we often have to do a little work to prepare the data in a format that the STAN code will allow. In this case, we need numerical values for the factors rather than alphanumeric strings. We specify  $N, J$ , the response and the predictor. We need to tell it the length of the two vectors (which is  $N$ ). We can also specify bounds on the values between the angle brackets. Here we have required that  $N$  and  $J$  be strictly positive and that the predictor values lie in the set  $\{1, \dots, J\}$ . These bounds are not essential but it is a good idea to put them in as it will catch some forms of invalid data that might be supplied.

The next block contains the parameters. The `mu`, `sigmaalpha` and `sigmaepsilon` correspond to the  $\mu$ ,  $\sigma_\alpha$  and  $\sigma_\epsilon$  parameters used in previous modeling. The `eta` parameter (actually a vector of parameters) will be needed in the construction of the model. We have specified lower bounds of zero on `sigmaalpha` and `sigmaepsilon`. This is essential to the proper treatment of these parameters.

The transformed parameters block combines the previously declared parameters in preparation for declaring the model in the next block. The `a` vector corresponds to the random effects declared previously although this includes the mean term  $\mu$ . The mean values of the response are given by `yhat` and are just the appropriate value from the `a` vector.

The final `model` block defines the model. Here we declare the `eta` variables to be standard normal. We then specify the response to be normally distributed with mean and standard deviation `sigmaepsilon` (not the variance!).

You may have noticed that we have not explicitly specified the priors for the three parameters: `mu`, `sigmaalpha` and `sigmaepsilon`. By default, improper uniform priors will be declared. For `mu`, this is the whole real line. Because both limits have not been specified, this is not a proper uniform distribution—that is why it is called *improper*. For the two standard deviation parameters, the prior is uniform on the positive real line. This explains the necessity of specifying bounds on the parameters earlier so that appropriate priors will be used. Our intent in choosing these particular priors is to express our lack of knowledge about these parameters.

We need to arrange the data in R into a list format consistent with the `data` statement of the STAN program:

```

data(pulp, package="faraway")
pulpdat <- list(N=20, J=4, response=pulp$bright, predictor=as.numeric(
  → pulp$operator))

```

We load the library that interfaces STAN to R:

```

library(rstan)

```

When loading the library, you may receive a prompt about how to use the multiple cores in your CPU. It is worth taking this advice as it will speed up the computations significantly. Now we can fit the STAN model as follows:

```
fit <- stan(file = "oneway.stan", data = pulpdat)
```

This assumes that the `oneway.stan` file containing the STAN code is in the working directory used by R. This command involves three major steps. The first is to translate the STAN code into C++. If you have made a syntax error in your STAN code, it will be revealed here. The second step is to compile the C++ code. We might change the data or simply want to run the MCMC again with different settings but this would not require repeating these two steps. There is a way to split the process into three separate steps if this would save some time. The third step actually runs the MCMC. A large amount of output will be produced, most of which can be ignored. Compiler warnings are usually not important. There are typically some cryptic warnings regarding the progress of the Markov chains. These can be ignored provided there are not too many. The entire process might take a few minutes depending on the quality of your computer.

The first step is to examine the model fitting diagnostics as seen in Figure 12.1. These diagnostics are entirely distinct from usual residual-based diagnostics which check the adequacy of the model. We are checking whether the MCMC process has produced a satisfactory result. The Markov chain needs to start from some initial values which should be feasible values of the parameters but may well be unlikely values. This means it may take some time before the chain settles into anything resembling the equilibrium distribution. This initial period is called the *burn-in* or *warm-up* time. We can determine an appropriate length for this period by plotting the chain(s). STAN recognizes this potential sensitivity to initial values and, by default, computes four chains from different randomly chosen initial values. We start with the `mu` parameter:

```
traceplot(fit, pars="mu", inc_warmup = TRUE)
```

The plot, seen in the first panel of Figure 12.1, shows the four complete chains starting from the initial values. The greyed area shows the warm-up period. We can see that all four chains have stabilized well within this period. If this stabilization fails to occur within the period, we can increase the length of the warm-up period or perhaps give the program some hints about suitable initial values.

For the subsequent computations, STAN discards this warm-up half of the chain. The resolution of the first plot means that it is difficult to see the behavior of the chains once they become more stationary. This requires that we redo the plot without the warm-up period as seen in the second panel of Figure 12.1:

```
traceplot(fit, pars="mu", inc_warmup = FALSE)
```

The Markov chain, by its nature, is dependent. We can still use it to estimate the posterior density or functions of it, such as the mean. However, a positively dependent chain contains less information than an independent sample of the same size. The less dependent the chain, the better. A chain that is not strongly dependent is said to *mix* well. In less favorable situations, a chain may become “stuck” in some sub-region of the parameter space for long sequences. Such chains are said to mix poorly. The four chains shown in the second panel of Figure 12.1 exhibit strong mixing as they all vary randomly around a constant level with no obvious dependence.

If you find a poorly mixing chain, you can run the chain for longer. The default chain length is 2000 iterations and this can be substantially increased. For simpler

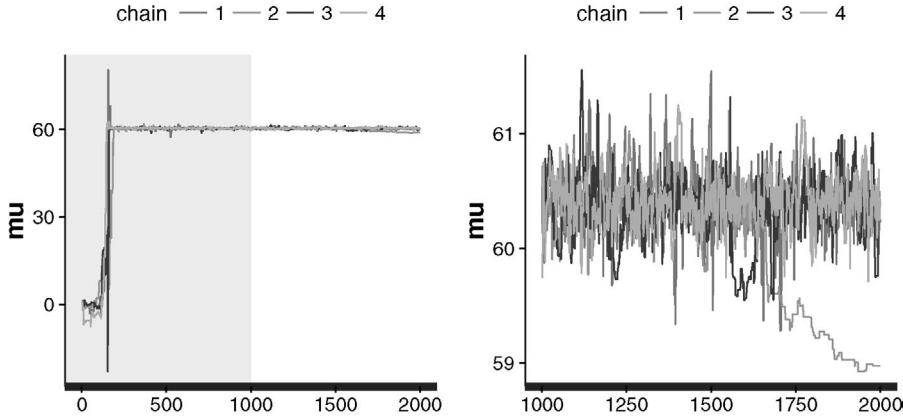


Figure 12.1 *Diagnostic plots for  $\mu$  from the STAN model for the pulp data. Version with a warm-up period is shown on the left. Four chains are shown in each case.*

models, this additional computational time may not matter much. However, in more extreme cases, one may be concerned that equilibrium has not been reached despite the extra iterations and that the chain is unreliable. In such cases, one may need to tinker with the construction of the chain or even change the prior.

We can make similar plots for `sigmaalpha` and `sigmaepsilon`, which both show a satisfactory outcome. Now we can examine the printed output:

**fit**

```
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.
```

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
eta[1]	-0.33	0.02	0.65	-1.62	-0.74	-0.32	0.10	0.89	910	1.01
eta[2]	-0.70	0.02	0.70	-2.13	-1.15	-0.65	-0.23	0.55	999	1.00
eta[3]	0.43	0.03	0.69	-0.92	-0.03	0.42	0.89	1.82	562	1.02
eta[4]	0.57	0.03	0.69	-0.73	0.11	0.54	1.02	1.96	420	1.02
mu	60.42	0.03	0.32	59.77	60.26	60.40	60.54	61.39	124	1.05
sigmaalpha	0.52	0.04	0.49	0.06	0.23	0.37	0.61	2.14	129	1.03
sigmaepsilon	0.36	0.00	0.07	0.25	0.31	0.35	0.40	0.53	1085	1.00
a[1]	60.28	0.00	0.15	59.98	60.19	60.28	60.38	60.59	2526	1.00
a[2]	60.14	0.00	0.17	59.82	60.03	60.14	60.25	60.48	1413	1.00
a[3]	60.57	0.00	0.16	60.26	60.46	60.56	60.67	60.88	1485	1.00
a[4]	60.61	0.00	0.16	60.29	60.50	60.61	60.72	60.93	2042	1.00
yhat[1]	60.28	0.00	0.15	59.98	60.19	60.28	60.38	60.59	2526	1.00
yhat[2]	60.28	0.00	0.15	59.98	60.19	60.28	60.38	60.59	2526	1.00
...										
lp_	7.80	0.12	2.81	1.35	6.11	8.27	9.91	12.03	582	1.00

For each parameter, `n_eff` is a crude measure of effective sample size, and `Rhat` is the potential scale reduction factor on split chains (at convergence, `Rhat=1`).

We have 2000 iterations for each of four chains but 1000 iterations are discarded for the first half of each chain, making 4000 iterations in all. If you use a much larger number of iterations, you may generate more posterior samples than convenient. You

can reduce this by thinning. STAN does not thin by default hence the `thin=1`. Setting `thin=10` would take every tenth observation and discard the rest.

For each posterior distribution, we have a dependent sample of size 4000. Various summary statistics are calculated as seen in the output. The samples are dependent but this is not a problem unless this is too strong. We can use the methods of time series analysis to compute various measures of dependence. The `Rhat` statistic is a measure of dependence. Values close to one indicate less dependence. The effective sample size tells us the equivalent independent sample size. These would be equal to the iteration total of 4000 for independent samples. We see that all these are well into the hundreds which is sufficient for most purposes.

The `eta` values represent scaled versions of the random effects. These are not particularly useful to us. The `mu` posterior mean is 60.4 just as in previous analyses. The `sigmaalpha` and `sigmaepsilon` posterior means are 0.52 and 0.36, respectively. We can also see how the posterior means for the four groups vary in the `a` parameters. The `yhat` posterior expected means for the response for each case just take the corresponding value of `a` depending on the group. For this reason, we have not printed all these out. The final `lp` term is the “log probability” and can be used to diagnose some problems with the chain.

The 2.5 and 97.5 percentiles form 95% credible intervals for the parameters. We claim a 0.95 probability that the parameter lies within the interval. This interpretation is different from a confidence interval where the 0.95 probability refers to the chance that the interval contains the parameter.

We can now examine the posterior distributions of  $\sigma_\alpha$  and  $\sigma_\epsilon$ :

```
library(reshape2)
postsig <- extract(fit, pars=c("sigmaalpha", "sigmaepsilon"))
ref <- melt(postsig, value.name="bright")
ggplot(data=ref, aes(x=bright, linetype=L1)) + geom_density() + xlim(0, 2)
  ↪ +scale_linetype(name="SD", labels=c("operator", "error"))
```

These can be seen in the first panel of Figure 12.2. We see that the posterior for the error SD is much more concentrated than the operator SD. The operator SD could be negligibly small but there is also a small chance that it could be substantially larger than the error SD.

In previous analysis of this data in Chapter 10, we tested a null hypothesis that there was no difference between the operators by setting  $H_0 : \sigma_\alpha = 0$ . The hypothesis testing formulation does not make sense within the Bayesian approach. Even so, we may still be interested whether there is a difference between the operators. We might formulate this question in terms of  $P(\sigma_\alpha = 0)$ . We have set a continuous prior and so we get a continuous posterior. This means the answer to our question is inevitably zero. For there to be a chance of a nonzero answer, we would need to specify a prior that gave strictly positive probability to the event that  $\sigma_\alpha = 0$ .

There are three reasons why we do not want to set such a prior. Firstly, for various technical reasons, the computation of the posterior resulting from such priors is difficult. Secondly, we would need to specify the prior probability that  $P(\sigma_\alpha = 0)$ . We may accept that the specification of a model requires some subjectivity but setting a probability like this feels uncomfortably close to fixing the conclusion. Thirdly, we might not view this as a sensible question. If the operators are four different people,

why would we expect them to behave identically? Instead, we might ask whether the difference between the operators is less than some specified small value. For example, we could compute  $P(\sigma_\alpha < 0.1)$ .

```
mean(postsig$sigmaalpha < 0.1)
```

```
[1] 0.05725
```

This is not a  $p$ -value. We have had to specify what we meant by “small,” but this is a judgement that we might reasonably ask of anyone with knowledge of the application. In this case, the response is recorded with only one figure after the decimal point so 0.1 is a reasonable choice, but choices will differ according to the situation. In this example, Figure 12.2 tells us that the difference between operators is likely to be considerable.

We may also be interested in the differences between specific operators. We can plot these posterior distributions:

```
opre <- rstan::extract(fit, pars="a")
ref <- melt(opre, value.name="bright")
ggplot(data=ref, aes(x=bright, linetype=factor(Var2)))+geom_density()+
  → scale_linetype(name="operator", labels=LETTERS[1:4])
```

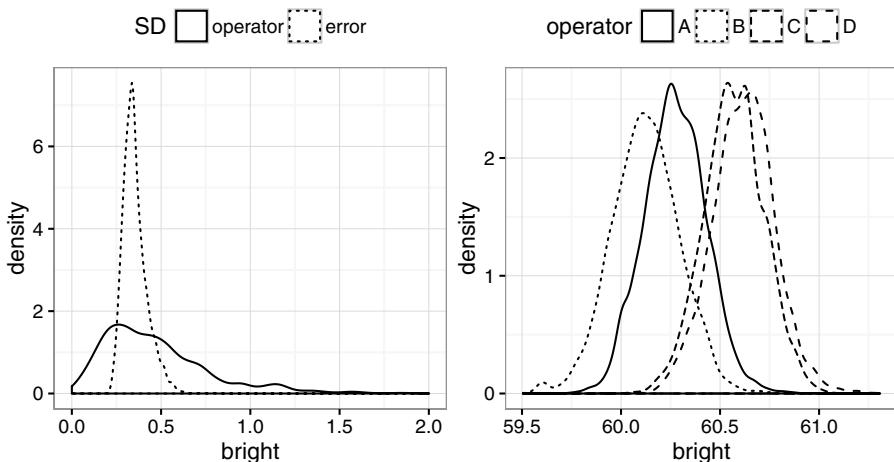


Figure 12.2 Posterior distributions for the standard deviations on the left and for the operator effects on the right.

We can see the posterior distributions for these effects in the second panel of Figure 12.2. We see that A and B tend to have lower responses than C and D but there is a substantial overlap such that we cannot be sure of any difference between the operators. We can compute a probability like  $P(\alpha_A > \alpha_D)$  from the posterior samples:

```
mean(ref[, 1] > ref[, 4])
```

```
[1] 0.0665
```

We have used uninformative priors for the parameters but we should think more carefully about this choice. In particular, we would have good reason to expect that there would not be extreme differences between and within operators. For this reason we might specify a prior that puts most mass on small to moderate values but does not completely exclude the possibility of much larger values. A good distribution

for this is the Cauchy as the Gaussian (or similar) distributions have light tails and do not allow any reasonable possibility of extremes. In contrast, the Cauchy does have heavy tails but will give most weight to the small/moderate values that we most anticipate. For strictly positive parameters like standard deviations, we can use the half-Cauchy.

We modify the `oneway.stan` file to add two new lines to the model block:

```
sigmaalpha ~ cauchy(0, 1);
sigmaepsilon ~ cauchy(0, 1);
```

We call the modified file `onewaycauchy.stan` and rerun it:

```
fit <- stan(file = "onewaycauchy.stan", data = pulpdat)
```

We can now examine the output:

```
print(fit, pars=c("mu", "sigmaalpha", "sigmaepsilon", "a"))
      mean se_mean sd 2.5% 25% 50% 75% 97.5% n_eff Rhat
mu       60.38    0.04 0.27 59.65 60.26 60.40 60.52 60.86    47 1.10
sigmaalpha  0.43    0.03 0.28  0.06  0.23  0.35  0.55  1.11    71 1.05
sigmaepsilon 0.35    0.00 0.07  0.25  0.30  0.34  0.39  0.52   381 1.01
a[1]      60.28    0.00 0.15 59.99 60.19 60.28 60.38 60.57  1765 1.00
a[2]      60.14    0.00 0.17 59.83 60.04 60.14 60.25 60.46  1099 1.00
a[3]      60.56    0.00 0.15 60.27 60.47 60.56 60.66 60.86  1588 1.00
a[4]      60.61    0.00 0.16 60.32 60.51 60.61 60.71 60.92  1301 1.00
```

We see that the posterior means for the two SDs are  $\hat{\sigma}_\alpha = 0.43$  and  $\hat{\sigma}_\epsilon = 0.35$ , which is a little smaller than the previous result although not radically different. We can also plot the posteriors in the same way as previously to find a slightly more concentrated distribution for  $\sigma_\alpha$ . We are reassured that the results are not particularly sensitive to the choice of priors. We prefer the latter result since this is more consistent with reasonable assumptions about the data-generating mechanism.

**Randomized Block Design:** In Section 10.6, we analyzed some data on penicillin production that had treatments as fixed effects and blends (of the raw material) being random effects. We repeat this analysis here using STAN. We create STAN code in a file called `rbd.stan` in the R working directory:

```
data {
  int<lower=0> N;
  int<lower=0> Nt;
  int<lower=0> Nb;
  int<lower=1,upper=Nt> treat[N];
  int<lower=1,upper=Nb> blk[N];
  vector[N] y;
}
parameters {
  vector[Nb] eta;
  vector[Nt] trt;
  real<lower=0> sigmablk;
  real<lower=0> sigmaepsilon;
}
transformed parameters {
  vector[Nb] bld;
  vector[N] yhat;

  bld <- sigmablk * eta;

  for (i in 1:N)
    yhat[i] <- trt[treat[i]]+bld[blk[i]];
}
```

```

}
model {
  eta ~ normal(0, 1);

  y ~ normal(yhat, sigmaepsilon);
}

```

The STAN code is a generalization of the one-way ANOVA example. We load and put the data into a list consistent with the `data` declaration in the STAN program:

```

data(penicillin, package="faraway")
penidat <- list(N=20, Nt=4, Nb=5, y=penicillin$yield, treat=as.numeric
  ↪ (penicillin$treat), blk=as.numeric(penicillin$blend))

```

We break up the fitting process into the three component parts:

```

rt <- stanc(file="rbd.stan")
sm <- stan_model(stanc_ret = rt, verbose=FALSE)
fit <- sampling(sm, data=penidat)

```

The first step converts the STAN code to C++ and the second step compiles that code. We only need to do this once. We could save this and load it back in a future session like this:

```

save(sm, file="rbd.RData")
load("rbd.RData")

```

Since we often want to refit a model or try it with different data, this saves us some time by avoiding the repetition of the same steps. We must check the convergence of the MCMC. Trouble is most likely to appear in the block SD parameter,  $\sigma_b$ , so we check that:

```
traceplot(fit, pars="sigmablk", inc_warmup = FALSE)
```

The plot, shown in the first panel of Figure 12.3, reveals no problem. Checks of the other parameters are similarly uneventful. Now we study the summary output, selecting only the parts of immediate interest:

```

print(fit, pars=c("trt","sigmablk","sigmaepsilon","bld"), probs=c
  ↪ (0.025,0.5,0.975))
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

```

	mean	se_mean	sd	2.5%	50%	97.5%	n_eff	Rhat
trt[1]	83.97	0.35	3.74	76.87	83.81	92.06	114	1.03
trt[2]	85.03	0.30	3.58	77.46	85.01	92.39	139	1.03
trt[3]	89.07	0.37	3.75	81.81	88.99	98.79	105	1.04
trt[4]	86.01	0.33	3.54	78.98	85.88	95.16	118	1.03
sigmablk	5.35	0.41	3.99	0.68	4.27	17.70	95	1.03
sigmaepsilon	4.95	0.04	1.11	3.29	4.78	7.57	674	1.01
bld[1]	4.12	0.31	3.63	-2.11	3.93	11.75	138	1.03
bld[2]	-2.09	0.32	3.48	-11.73	-1.64	4.05	119	1.03
bld[3]	-0.59	0.31	3.34	-7.76	-0.50	6.10	115	1.04
bld[4]	1.30	0.33	3.42	-6.73	1.20	8.17	106	1.04
bld[5]	-2.77	0.37	3.67	-12.68	-2.40	3.69	101	1.04

The effective sample sizes are large enough so we are satisfied with the fit. The posterior means are comparable to the REML estimates obtained from the `lme4` fit.

We can plot the posterior densities for the two random SD parameters:

```

postsig <- rstan::extract(fit, pars=c("sigmablk", "sigmaepsilon"))
ref <- melt(postsig,value.name="yield")

```

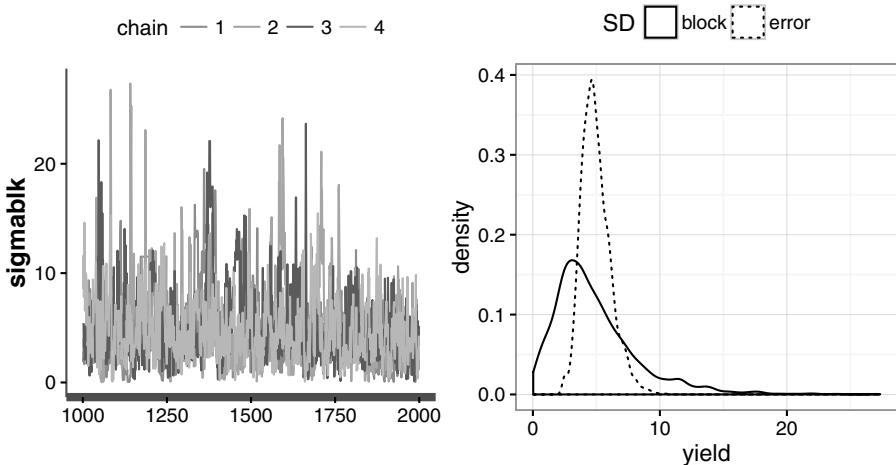


Figure 12.3 *Diagnostic plot for the MCMC for  $\sigma_b$  is shown on the left. Posterior densities for the random SDs are shown on the right.*

```
ggplot(data=ref, aes(x=yield, linetype=L1)) + geom_density() + scale_
  ↪ linetype(name="SD", labels=c("block", "error"))
```

We see in the second panel of Figure 12.3 that the error SD distribution is quite concentrated while the block SD is more diffuse. As with the previous example, it makes little sense to ask whether the blend SD is zero. The plot makes it clear that there probably is substantial variation between blends. We could ask whether the blend SD is less than some small value—one seems a reasonable limit in this example:

```
mean(postsig$sigmablk < 1)
[1] 0.0435
```

This confirms that there is a small chance that the blend SD is negligibly small but it seems much more likely that it is considerable. Given the relatively small effective sample size of 95 for this parameter, we would want more MCMC iterations to get a more accurate estimate if this was needed. We can plot the blend posterior distributions:

```
bldeff <- rstan::extract(fit, pars="bld")
rdf <- data.frame(yield=unlist(bldeff), blend=g1(5,4000))
ggplot(data=rdf, aes(x=yield, linetype=blend)) + geom_density()
```

We see in the first panel of Figure 12.4 that although there is some separation in the distributions, it would be difficult to conclude that one was clearly better than another. The bumps in the tails of the densities are caused by the small sample size and could be removed with a larger number of MCMC iterations if desired.

We can construct the posteriors for the treatment effects also:

```
trteff <- rstan::extract(fit, pars="trt")
rdf <- data.frame(yield=unlist(trteff), treat=g1(4,4000, labels=
  ↪ LETTERS[1:4]))
ggplot(data=rdf, aes(x=yield, linetype=treat)) + geom_density()
```

In the second panel of Figure 12.4 we see that there is considerable overlap between

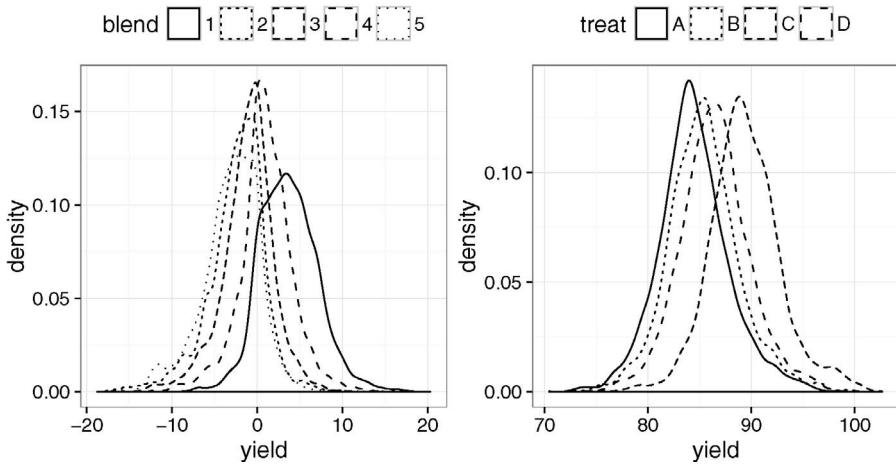


Figure 12.4 *Posterior densities for the blend effects are shown on the left, while the treatment effects are shown on the right.*

the distributions with A appearing worst and C best. We can estimate the probability that A is actually better than C as:

```
mean(trteff$trt[,1] > trteff$trt[,3])
[1] 0.05725
```

We conclude there is no clear difference between the treatments. If forced to choose, perhaps with information about the relative costs of the treatments, we could use the posterior distributions to make a rational decision.

It would also be reasonable to consider the treatments as a random effect. We could modify the STAN program to accommodate this and analyze accordingly.

STAN analyses of the other mixed effect models from previous chapters may be found in the online materials.

## 12.2 INLA

INLA stands for integrated nested Laplace approximation and was introduced by Rue et al. (2009). A Laplace approximation is a method for computing integrals of the form  $\int \exp f(x)dx$ . It requires only the maximum of  $f$  and the second derivative of  $f(x)$  at that point. It can be surprisingly accurate given how little information is required. The INLA method builds on this idea and can be used for a wide class known as Gaussian Markov random field models. This class includes all the models considered in the previous two chapters. Because the method requires no simulation, it is much faster than MCMC-based approaches.

INLA is an R package that provides an interface to the INLA method. Visit <http://www.r-inla.org> for installation instructions as the package is not available from CRAN.

**One-Way ANOVA:** Let's use INLA to fit a model to the pulp data that we have

already explored with `lme4` in Chapter 10 and with STAN in Section 12.1. See Figure 10.1 for a plot of the data and Section 12.1 for a specification of the model.

We load the library and data before fitting the model.

```
library(INLA)
data(pulp, package="faraway")
formula <- bright ~ f(operator, model="iid")
result <- inla(formula, family="gaussian", data=pulp)
```

There is only the fixed effect intercept term which is included implicitly by default in the `formula` for the model. We specify how the random operator enters the model using the `f()` term. We expect no correlation in the observations for each operator so we set the `iid` option. We use the Gaussian family for the response when fitting the model with `inla()`. We examine the result and look at a somewhat-edited output:

```
summary(result)
```

Fixed effects:

	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
(Intercept)	60.4	0.0876	60.226	60.4	60.573	60.4	0

Random effects:

Name	Model
operator	IID model

Model hyperparameters:

	mean	sd	0.025quant
Precision for the Gaussian observations	6.895	2.131	3.494
Precision for operator	18377.553	18267.005	1205.289
	0.5quant	0.975quant	mode
Precision for the Gaussian observations	6.647	11.78	6.183
Precision for operator	12974.392	66416.77	3248.557

Expected number of effective parameters(std dev): 1.013(0.0132)

Number of equivalent replicates : 19.74

We see that the intercept of 60.4 is the same as in previous model fits. INLA uses the precision, which is the inverse of the variance, for its internal computations. We need to compute SDs for interpretative purposes. We have a posterior mean for  $\hat{\sigma}_\epsilon$  of  $1/\sqrt{6.895} = 0.38$ , but the precision for the operator term is very large and so the posterior mean for  $\sigma_\alpha$  is close to zero. Even without our previous modeling experience with this data, we know from the plots of the data that we could not possibly be this confident about a conclusion of no variation between operators. Something is wrong.

The problem is the default priors chosen by INLA for the random effects precision terms. A gamma distribution with a very large variance is used. Past experience with LMMs reveals that this can sometimes lead to anomalous results. Indeed, this illustrates a general problem with so-called noninformative priors. We would like to have some default choice that will not have much effect on the outcome, but sometimes it will fail. One solution is to use a more informative prior that uses the weak information that we are likely to possess.

Let a random effect have SD  $\sigma$  and choose an exponential prior on  $\sigma$  such that a tail probability  $P(\sigma > U) = \alpha$ . We can pick a small value for  $\alpha$  such as 0.01. Let's allow ourselves to look at the SD of the response and set  $U = 3SD(Y)$ . We are not constraining  $\sigma$  very much but we still allow a small probability that it could be even

three times larger than the response SD. This is an example of a penalized complexity prior as explained in Simpson et al. (2014).

```
sdres <- sd(pulp$bright)
pcprior <- list(prec = list(prior="pc.prec", param = c(3*sdres, 0.01)))
formula <- bright ~ f(operator, model="iid", hyper = pcprior)
result <- inla(formula, family="gaussian", data=pulp)
result <- inla.hyperpar(result)
summary(result)
```

We specify  $U = 3SD(Y)$  and  $\alpha = 0.01$  in the penalized complexity prior. We use this prior for  $\sigma_\alpha$ , but retain the default prior for  $\sigma_\epsilon$ . The estimates for the posterior densities are improved by `inla.hyperpar`. The summary output shows precisions that are believable so we move onto a consideration of the posteriors.

Unlike MCMC programs, INLA does not just give you a sample from the posterior. It gives you the posterior density itself. Since we cannot easily plot the full multivariate posterior, we resort to looking at the marginals. The precisions are not very interpretable so we naturally want to convert it to the SD scale. Since we are transforming a density, the calculation is exact and can be achieved as follows:

```
sigmaalpha <- inla.tmarginal(function(x) 1/sqrt(exp(x)), result$ 
  ↪ internal.marginals.hyperpar[[2]])
sigmaepsilon <- inla.tmarginal(function(x) 1/sqrt(exp(x)), result$ 
  ↪ internal.marginals.hyperpar[[1]])
```

We can plot these as seen in the first panel of Figure 12.5:

```
ddf <- data.frame(rbind(sigmaalpha,sigmaepsilon),errterm=gl(2,1024,
  ↪ labels = c("alpha","epsilon")))
ggplot(ddf, aes(x,y, linetype=errterm))+geom_line() +xlab("bright")+
  ↪ ylab("density") +xlim(0,2)
```

We see that the posteriors for the error and operator SDs are similar to those seen in Figure 12.2.

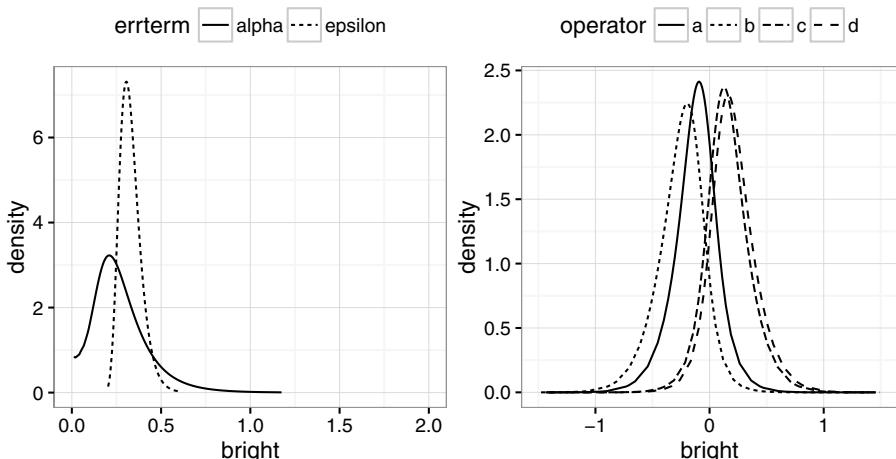


Figure 12.5 Posterior densities for the SDs are shown on the right. On the left, we see the posterior densities for the operator effects.

In the STAN analysis, we asked whether the operator SD was small. Specifically we computed  $P(\sigma_\alpha < 0.1)$ . We can compute this here by:

```
inla.pmarginal(0.1, sigmaalpha)
```

```
[1] 0.095447
```

The probability is still small but larger than that seen in the STAN-based analysis. Our prior for INLA put more weight on  $\sigma = 0$ , so this is not surprising.

We can plot the posteriors for the operator effects as seen in the second panel of Figure 12.5:

```
rdf <- do.call(rbind.data.frame, result$marginals.random$operator)
rdf <- cbind(operator=gl(4,nrow(rdf)/4,labels=letters[1:4]),rdf)
ggplot(rdf, aes(x=x,y=y,linetype=operator))+geom_line() + xlim(-1.5,
  ↪ 1.5) + xlab("bright") + ylab("density")
```

There is considerable overlap between the densities, meaning that it will be difficult to distinguish between specific operators.

Finally, we can obtain a numerical summary of the parameters of interest by:

```
restab <- sapply(result$marginals.fixed, function(x) inla.zmarginal(x,
  ↪ silent=TRUE))
restab <- cbind(restab, inla.zmarginal(sigmaalpha,silent=TRUE))
restab <- cbind(restab, inla.zmarginal(sigmaepsilon,silent=TRUE))
restab <- cbind(restab, sapply(result$marginals.random$operator,
  ↪ function(x) inla.zmarginal(x, silent=TRUE)))
colnames(restab) = c("mu", "alpha", "epsilon", levels(pulp$operator))
data.frame(restab)

      mu     alpha    epsilon      a      b      c      d
mean 60.4  0.24684  0.32115 -0.11008 -0.23533  0.15157  0.19329
sd    0.13238  0.11499  0.047074  0.14841  0.15769  0.1509   0.15398
quant0.025 60.03  0.041186  0.23559 -0.54174 -0.70209 -0.21423 -0.16563
quant0.25  60.297  0.16693  0.28606 -0.22994 -0.36867  0.037022  0.076544
quant0.5   60.399  0.24632  0.32084 -0.10922 -0.23292   0.1482   0.1894
quant0.75  60.501  0.34822  0.36309  0.0011632 -0.11752  0.27338  0.31963
quant0.975 60.768  0.66073  0.4694   0.2622  0.11642  0.59304  0.64648
```

Compare these posterior means to those obtained from the lme4 analysis in Chapter 10 and in the earlier STAN analysis. The results are comparable but not the same.

**Randomized Block Design:** In Section 10.6, we analyzed some data on penicillin production that had treatments as fixed effects and blends (of the raw material) being random effects. Refer back to this section for a description of the model and plots of the data. We repeat this analysis here using INLA:

```
data(penicillin, package="faraway")
lmod <- lm(yield ~ treat, data=penicillin)
sdres <- sd(residuals(lmod))
pcprior <- list(prec = list(prior="pc.prec", param = c(3*sdres,0.01)))
formula <- yield ~ treat + f(blend, model="iid", hyper = pcprior)
result <- inla(formula, family="gaussian", data=penicillin)
result <- inla.hyperpar(result)
summary(result)
```

Use of the default prior for the blend precision gives rise to the same problem seen in the pulp analysis. We supply a more informative prior making use of the residual SD from the fixed effects only model. This SD will be bigger than the blend or error SD. The prior allows the possibility that the SD could be bigger still so we are doing little to constrain the outcome but ensuring a reasonable result. We transform the precisions to the SD scale:

```
sigmaalpha <- inla.tmarginal(function(x) 1/sqrt(exp(x)), result$  
  ↪ internal.marginals.hyperpar[[2]])  
sigmaepsilon <- inla.tmarginal(function(x) 1/sqrt(exp(x)), result$  
  ↪ internal.marginals.hyperpar[[1]])
```

We produce a summary table of the posterior densities:

```
restab <- sapply(result$marginals.fixed, function(x) inla.zmarginal(x,  
  ↪ silent <- TRUE))  
restab <- cbind(restab, inla.zmarginal(sigmaalpha, silent=TRUE))  
restab <- cbind(restab, inla.zmarginal(sigmaepsilon, silent=TRUE))  
colnames(restab) <- c("mu", "B-A", "C-A", "D-A", "alpha", "epsilon")  
data.frame(restab)
```

	mu	B.A	C.A	D.A	alpha	epsilon
mean	84.028	0.96989	4.9565	1.9666	3.1441	4.2801
sd	2.0898	2.1297	2.1297	2.1297	1.4291	0.72534
quant0.025	78.76	-4.3684	-0.38568	-3.3725	0.48106	3.003
quant0.25	82.332	-0.7651	3.2211	0.2315	2.1311	3.7446
quant0.5	84.017	0.95897	4.9457	1.9557	3.1579	4.275
quant0.75	85.702	2.6811	6.6674	3.6777	4.3837	4.9327
quant0.975	89.28	6.2686	10.251	7.2646	7.908	6.6071

The posterior means are quite similar to the previous analyses. Particularly for asymmetric densities, the posterior mode is more analogous to a maximum likelihood estimate. These can be calculated for the two SDs as:

```
c(inla.mmarginal(sigmaalpha), inla.mmarginal(sigmaepsilon))  
[1] 2.8331 4.0070
```

We examine the posterior densities of the SDs as seen in the first panel of Figure 12.6.

```
ddf <- data.frame(rbind(sigmaalpha, sigmaepsilon), errterm=g1(2,1024,  
  ↪ labels = c("alpha", "epsilon")))  
ggplot(ddf, aes(x,y, linetype=errterm))+geom_line() +xlab("yield") +ylab  
  ↪ ("density") +xlim(0,15)
```

We see that the error SD is more precisely specified than the block SD as is expected for such experiments. As before, we might ask whether there is much difference between the blends. We might express this by  $P(\sigma_\alpha < 1)$ . We can compute this as:

```
inla.pmarginal(1, sigmaalpha)  
[1] 0.089164
```

We see that there is a small but significant chance that the blend SD is negligible.

We can also obtain the posterior densities for the treatment effects as seen in the second panel of Figure 12.6.

```
x <- seq(-15,15,length.out = 100)  
rden <- sapply(result$marginals.fixed,function(y) inla.dmmarginal(x, y)  
  ↪ )[, -1]  
ddf <- data.frame(yield=rep(x,3), density=as.vector(rden), treat=g1  
  ↪ (3,100, labels=c("B-A", "C-A", "D-A")))  
ggplot(ddf, aes(x=yield, y=density, linetype=treat))+geom_line()
```

The standard parameterization of the fixed effects means that these represent differences from the reference level, A. We can estimate the probability that A is actually better than C as:

```
inla.pmarginal(0, result$marginals.fixed$treatC)  
[1] 0.033363
```

If we use the usual  $p$ -value reasoning that doubles this for a two-sided test, we obtain a value of about 6.6%. Given that the other differences are smaller, we find there is little evidence of a significant difference between the treatments. Given this is a Bayesian analysis, the usual justification for  $p$ -values does not apply, but we can

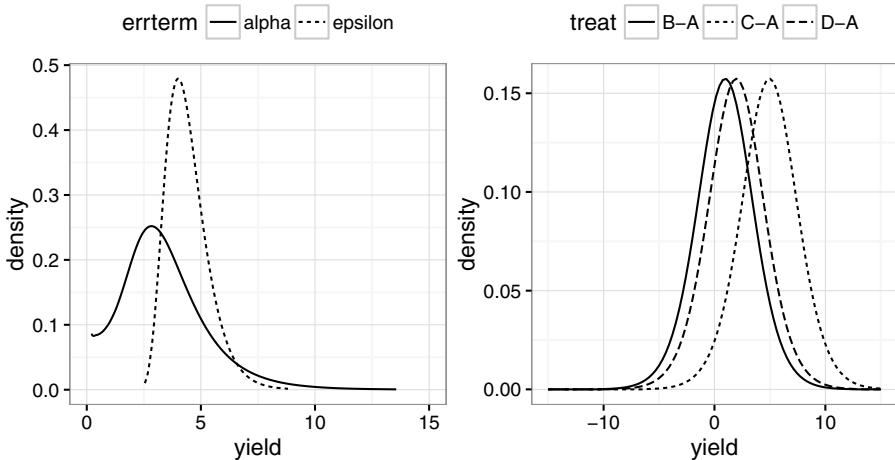


Figure 12.6 *Posterior densities for SDs on the left and for the treatment effects on the right.*

reasonably compute such quantities as summaries of the posterior densities if it is helpful.

### 12.3 Discussion

The maximum likelihood analysis of linear mixed models, demonstrated in Chapters 10 and 11, has several advantages. The models can be specified and fit with a single R command. The statistical hypothesis testing paradigm is widely accepted and may be required for the communication of some scientific research. The calculation of the  $p$ -values can be difficult, but is possible, even if simulation methods, such as the bootstrap, are required. Even so, problems may arise in fitting these models, particularly to larger datasets. Some types of valid questions cannot be answered in this mode of analysis.

The Bayesian approach offers a quite different way of analyzing this class of models. It offers several advantages in that we can use prior information to improve the inference and we can answer various relevant questions about the application in natural ways. There are some drawbacks. The models are more difficult to specify and require more programming knowledge, particularly when using STAN. The fitting process may fail in ways which are difficult to diagnose and rectify. The specification of reliable, so-called noninformative priors does not seem possible as failures producing unreasonable results are not uncommon. This requires us to think carefully about the specification of these priors. To the Bayesian, this is expected, but to others, this introduces an additional element of subjectivity which makes reaching convincing conclusions more difficult.

INLA beats STAN for speed, especially since it produces posterior densities not samples. For small datasets, the savings of a few minutes is not so important, but for larger datasets, the difference can be critical. STAN is more versatile as we are

able to program for a wider range of models than considered here. There is a general difficulty in validating a Bayesian analysis. Sometimes the outcome will be clearly flawed and we can make modifications. More worrisome are results which seem plausible, but are wrong. INLA produces approximations, but it is difficult to know when these are adequate. In principle, the MCMC analysis is essentially exact given a large enough run, but the simulation-based procedures used by STAN can fail in practice for various reasons that may be difficult to spot. For these reasons, a careful analyst is advised to use both INLA and STAN as a significant difference between the analyses will be revealing. Comparison with the `lme4`-based results provides a further check.

## Exercises

1. The `denim` dataset concerns the amount of waste in material cutting for a jeans manufacturer due to five suppliers. See another question on this dataset in Chapter 10.
  - (a) Plot the data and comment.
  - (b) Fit the one-way ANOVA model using `INLA` using the default prior. Comment on the fit.
  - (c) Refit the model but with more informative priors. Make a density plot of the error and supplier SD posterior densities.
  - (d) Calculate summaries of the posteriors from the model fit.
  - (e) Report 95% credible intervals for the SDs using the summary output. Compute the posterior modes for the error and supplier SDs and compare these to the posterior means.
  - (f) Remove the two outliers from the data and repeat the analysis. Comment on any interesting differences.
2. Use the `denim` dataset again for this question but conduct the analysis using STAN.
  - (a) Fit the one-way ANOVA model using `STAN` with the default prior. Produce diagnostic plots for the three parameters: the mean and standard deviations of the supplier and error effects.
  - (b) Report the posterior mean, 95% credible intervals and effective sample size for the three parameters.
  - (c) Make a plot of the posterior densities of the supplier and error effects. Estimate the probability that the supplier SD is bigger than the error SD.
  - (d) Plot the posterior distributions of the five suppliers. Which supplier tends to produce the least waste and which the most? What is the probability that the best supplier is better than the worst supplier?
  - (e) A plot of the data reveals two obvious outliers. Repeat the analysis without these two points and report on any interesting differences with the full data.

- (f) Instead of removing the two outliers, change the error distribution from normal to  $t_3$ . Repeat the analysis and indicate how this changes the conclusions. Discuss which approach to handling the outliers is best.
3. The `oatvar` dataset concerns a randomized block design comparing the yields of eight varieties of oats. The growing area was divided into five blocks, each planted with a single plot of each variety.
- Plot the data and comment.
  - Use `INLA` to fit a linear model with variety as a fixed effect and block as a random effect using informative priors. Make plots of the posterior SD densities of the error and block.
  - Compute a numerical summary of the posteriors. Which varieties show a significant difference from the first variety?
  - Plot the densities of the variety effects.
  - Consider a model with both variety and block as random effects. Use `INLA` to fit the model and construct a density plot of the three posterior SDs. How do these SDs compare?
4. Repeat question three using STAN in place of INLA.

**This page intentionally left blank**

# Mixed Effect Models for Nonnormal Responses

---

## 13.1 Generalized Linear Mixed Models

Generalized linear mixed models (GLMM) combine the ideas of generalized linear models with the random effects modeling ideas of the previous two chapters. The response is a random variable,  $Y_i$ , taking observed values,  $y_i$ , for  $i = 1, \dots, n$ , and follows an exponential family distribution as defined in Chapter 8:

$$f(y_i|\theta_i, \phi) = \exp \left[ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y, \phi) \right]$$

Let  $EY_i = \mu_i$  and let this be connected to the linear predictor  $\eta_i$  using the link function  $g$  by  $\eta_i = g(\mu_i)$ . Suppose for simplicity that we use the canonical link for  $g$  so that we may make the direct connection that  $\theta_i = \mu_i$ .

Now let the random effects,  $\gamma$ , have distribution  $h(\gamma|V)$  for parameters  $V$ . The fixed effects are  $\beta$ . Conditional on the random effects,  $\gamma$ ,

$$\theta_i = x_i^T \beta + z_i^T \gamma$$

where  $x_i$  and  $z_i$  are the corresponding rows from the design matrices,  $X$  and  $Z$ , for the respective fixed and random effects. Now the likelihood may be written as:

$$L(\beta, \phi, V|y) = \prod_{i=1}^n \int f(y_i|\beta, \phi, \gamma) h(\gamma|V) d\gamma$$

Typically the random effects are assumed normal:  $\gamma \sim N(0, D)$ . However, unless  $f$  is also normal, the integral remains in the likelihood, which becomes difficult to compute, particularly if the random effects structure is complicated.

## 13.2 Inference

A variety of approaches are available for estimating and performing inference for these models. All have strengths and weaknesses so it is not possible to recommend a single method to use in all circumstances. We present an overview of the theory behind these approaches before demonstrating the implementation on two examples. Later in the chapter, we discuss a related method called generalized estimating equations (GEE).

**Penalized Quasi-Likelihood (PQL):** In Section 8.2, we described a method by

which GLMs can be fit using only LMs with weights. The idea is to produce a linearized version of the response which we called the adjusted dependent variable (sometimes called the pseudo or working response) defined as

$$\tilde{y}^i = \hat{\eta}^i + (y - \hat{\mu}^i) \frac{d\eta}{d\mu}|_{\hat{\eta}^i}$$

The superscripted *i*s indicate the iteration in the optimization algorithm. We have  $E(\tilde{y}_i|\gamma) = x_i^T \beta + z_i^T \gamma$  and we may derive an expression for  $\text{var}(\tilde{y}_i|\gamma)$ . We are now able to use LMM methods with appropriate weighting. Iteration is necessary as  $\tilde{y}$  must be updated after each linear mixed model (LMM) fit. This and similar methods are described in Schall (1991) and Breslow and Clayton (1993). The name quasi-likelihood is not entirely appropriate for PQL as we still use the distributional assumptions. The GEE method described later in this chapter fits better with the “quasi” paradigm.

The PQL method has the advantage of relatively easy implementation given that existing LMM methods can be adapted. However, the inference is only asymptotically correct. Biased estimates are mostly likely to arise for binomial responses with small groups (covariate classes) and will be worst for Bernoulli responses. Similar problems will be observed for Poisson response data where the counts tend to be low. Further difficulties will arise with hypothesis testing and confidence intervals because the problems already present in LMMs are added to the approximations introduced by the linearization. We can compute *p*-values using likelihood theory-based methods but we will have limited trust in their veracity. Even so, PQL will tend to be faster and work on more complex models than some of the competitors.

**Numerical Integration:** Provided the dimension of the random effects  $\gamma$  is not too large, it is possible to use numerical methods to approximate the likelihood. The Laplace approximation is one of the least demanding methods for computing integrals of the form  $\int \exp h(x) dx$ . We need only find the maximum of  $h$  and the second derivative of  $h(x)$  at that point. For the integral in the GLMM likelihood, this can provide a surprisingly good approximation despite the integrand being evaluated at just one point. The maximization step is already familiar from simpler problems.

We can do better with more function evaluations. For these kinds of integrals, Gauss-Hermite quadrature is appropriate. The method approximates integrals of the form  $\int h(x) \exp(-x^2) dx$  by  $\sum_k w_k f(x_k)$  where the best choice weights  $w_k$  and knot-points  $x_k$  have been determined. This method is more accurate than the Laplace approach but the computational cost can become prohibitive, particularly for higher dimensional random effects. The Laplace method can be viewed as equivalent to the Gauss-Hermite method with just a single knotpoint.

Experience suggests that numerical integration methods are superior to PQL. The drawback is that they may be time-consuming or impossible to compute for more complex models. The advantage is that we have an approximation to the true likelihood rather than a quasi-likelihood. This means we have more scope and greater confidence in the inference derived from these approaches. They are not perfect since similar issues as with LMMs remain but better than PQL. See McCulloch and Searle (2002) for more discussion.

**Bayes:** As with LMMs, there is good reason to consider Bayesian methods as

an alternative to the standard likelihood-based methods. There are several advantages. Complex models can be fit with a high degree of accuracy. We can incorporate useful prior information and we have the flexibility to modify the models to allow for nonstandard features. The disadvantages are that these models may require more programming to implement and may take substantial computing resources. Furthermore, one must address technical concerns about the quality of the fit. Finally, the inferential conclusions are of a different form. This is either an advantage or disadvantage depending on your point of view. See Chapter 12 for an introduction to Bayes methods for LMMs. Extending these ideas to GLMMs is not difficult.

We now apply these methods to two examples. The first has a Bernoulli response and the second a Poisson response.

### 13.3 Binary Response

An experiment was conducted to study the effects of surface and vision on balance. The balance of subjects was observed for two different surfaces and for restricted and unrestricted vision. Balance was assessed qualitatively on an ordinal four-point scale based on observation by the experimenter. Forty subjects were studied, 20 males and 20 females ranging in age from 18 to 38, with heights given in cm and weights in kg. The subjects were tested while standing on foam or a normal surface and with their eyes closed or open or with a dome placed over their head. Each subject was tested twice in each of the surface and eye combinations for a total of 12 measures per subject. The data comes from Steele (1998) via the Australasian Data and Story Library (OzDASL).

For the purposes of this analysis, we will reduce the response to a two-point scale: whether the subject was judged completely stable (=1) or not (=0). We start by defining this response:

```
data(ctsib, package="faraway")
ctsib$stable <- ifelse(ctsib$CTSIB==1, 1, 0)
```

We can investigate the effects of the treatment variables on stability descriptively. Here is the mean response for the combined conditions:

```
xtabs(stable ~ Surface + Vision, ctsib)/80
      Vision
Surface closed   dome   open
  foam 0.0000 0.0000 0.1250
  norm 0.2125 0.2750 0.8125
```

We have divided by 80 because `xtabs` sums the values for each combination and there are 40 subjects with each combination replicated twice. We see that the normal surface with open vision leads to the highest stability. We can group the data by subject and average over the 12 observations (6 conditions, replicated twice). The plots are seen in Figure 13.1.

```
library(dplyr)
subsum <- ctsib %>% group_by(Subject) %>% summarise(Height=Height[1],
  ↪ Weight=Weight[1], stable=mean(stable), Age=Age[1], Sex=Sex[1])
library(ggplot2)
ggplot(subsum, aes(x=Height,y=stable))+geom_point()
ggplot(subsum, aes(x=Weight,y=stable))+geom_point()
ggplot(subsum, aes(x=Age,y=stable))+geom_point()
```

```
ggplot(subsum, aes(x=Sex, y=stable)) + geom_boxplot()
```

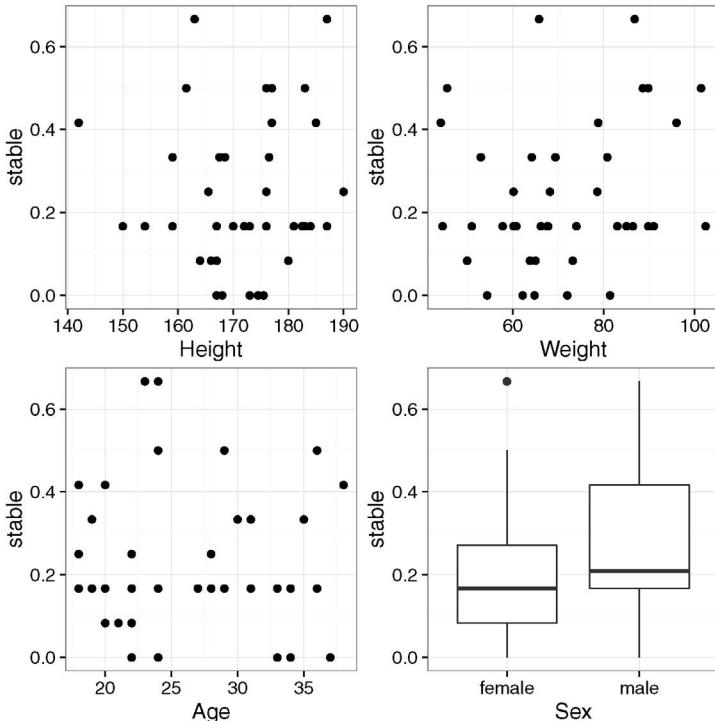


Figure 13.1 *Subject effects for the stability experiment. Response is proportion of stable over treatment conditions.*

We could fit a binomial GLM ignoring the subject information entirely:

```
gf <- glm(stable ~ Sex+Age+Height+Weight+Surface+Vision,binomial,data=
  ↪ ctsib)
summary(gf)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	7.27745	3.80399	1.91	0.05573
Sexmale	1.40158	0.51623	2.72	0.00663
Age	0.00252	0.02431	0.10	0.91739
Height	-0.09641	0.02684	-3.59	0.00033
Weight	0.04350	0.01800	2.42	0.01567
Surfacenorm	3.96752	0.44718	8.87	< 2e-16
Visiondome	0.36375	0.38322	0.95	0.34252
Visionopen	3.18750	0.41600	7.66	1.8e-14

n = 480 p = 8

Deviance = 295.203 Null Deviance = 526.254 (Difference = 231.051)

This assumes we have 480 independent observations but, in reality, we have only 40 subjects whose responses will be correlated. This analysis is likely to underestimate the standard errors and so exaggerate the significance of the experimental effects. We could also try including a fixed subject factor:

```
gfs <- glm(stable ~ Sex + Age + Height + Weight + Surface + Vision +
    ↪ factor(Subject), binomial,data=ctsib)
```

However, when we examine the summary for this model, we see problems with identifiability. This is because the subject factors cannot be completely distinguished from the four subject-specific measures: sex, age, height and weight. Even if we could get around this problem, it would hardly be appropriate to treat the subject factor as a fixed effect. We do not care about the individual subjects but we are interested in how the four subject measures might affect stability. The experimental subjects are intended as a random sample from the target population. We'd like to know something about the inherent variability in that population that is not explained by measurable variables but we don't care about the specific individuals.

There are a variety of ways of fitting GLMMs in R. First we demonstrate the PQL method implemented in the MASS package:

```
library(MASS)
modpql <- glmmPQL(stable ~ Sex + Age + Height + Weight + Surface +
    ↪ Vision, random=~1|Subject, family=binomial,data=ctsib)
summary(modpql)

Random effects:
Formula: ~1 | Subject
  (Intercept) Residual
StdDev:      3.0607  0.59062

Variance function:
Structure: fixed weights
Formula: ~invwt

Fixed effects: stable ~ Sex + Age + Height + Weight + Surface + Vision
      Value Std.Error DF t-value p-value
(Intercept) 15.5715   13.4983 437  1.1536  0.2493
Sexmale     3.3553    1.7526  35  1.9145  0.0638
Age        -0.0066   0.0820  35 -0.0810  0.9359
Height     -0.1908   0.0920  35 -2.0736  0.0455
Weight      0.0695   0.0629  35  1.1052  0.2766
Surfacenorm 7.7241   0.5736 437 13.4665  0.0000
Visiondome  0.7265   0.3259 437  2.2289  0.0263
Visionopen  6.4853   0.5440 437 11.9219  0.0000
```

The SD for the subject effect is 3.06. We can use the same ideas from logistic regression to interpret this value. We have  $\exp(3.06) = 21.3$  so the odds of stability are multiplied by this factor. Hence we can see that there is substantial variation in the inherent stability of individuals. Indeed, this variation is of comparable magnitude to the treatment effects. The residual SD is an artefact of the fitting process and does not exist in the statement of the model.

We see strongly significant surface and vision effects while some other effects have marginally significant  $p$ -values. However, this inference is based on the linearized model and rather dubious assumptions as explained in Section 10.2, so these results cannot be relied upon. Furthermore, the Bernoulli response may lead to biased estimates of regression coefficients. Hence, it would be unwise to rely entirely on this analysis without investigating alternative methods of estimation.

The numerical integration-based methods are implemented in the lme4 package. The default choice of method is the Laplace approximation.

```
library(lme4)
```

```
modlap <- glmer(stable ~ Sex + Age + Height + Weight + Surface +
    ↪ Vision + (1|Subject), family=binomial, data=ctsib)
```

Since the Laplace method is a special case of the Gauss-Hermite approximation which can only be more accurate, it is best to attempt this approach. Here we can use the maximum allowable number of quadrature points which is 25:

```
modgh <- glmer(stable ~ Sex + Age + Height + Weight + Surface +
    ↪ Vision + (1|Subject), nAGQ=25, family=binomial, data=ctsib)
```

We have a particularly simple random effects structure so we can easily afford to be profligate in the number of quadrature points (which is certainly more than we need). In more complex examples, we may need to specify much smaller numbers to allow computation in a reasonable time. Start small and increase until the estimates stop changing very much or the computation becomes infeasibly long. Now look at the output:

```
summary(modgh)
```

AIC	BIC	logLik	deviance	df.resid
247.9	285.5	-115.0	229.9	471

Scaled residuals:

Min	1Q	Median	3Q	Max
-4.884	-0.139	-0.020	-0.001	4.902

Random effects:

Groups	Name	Variance	Std.Dev.
Subject	(Intercept)	7.19	2.68

Number of obs: 480, groups: Subject, 40

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	16.17166	12.72107	1.27	0.204
Sexmale	3.09679	1.69612	1.83	0.068
Age	-0.00668	0.07646	-0.09	0.930
Height	-0.19226	0.08895	-2.16	0.031
Weight	0.07515	0.05910	1.27	0.204
Surfacenorm	7.28541	1.05516	6.90	5.0e-12
Visiondome	0.67591	0.52737	1.28	0.200
Visionopen	6.08896	0.97241	6.26	3.8e-10

Notice that we have AIC/BIC values for model comparison purposes. These are not available from PQL because it is not a true likelihood method. As it happens, the parameter estimates are quite similar to PQL which provides some reassurance.

We might ask whether any of the subject-specific variables have an effect. We can test this by fitting a model without these terms and comparing the two:

```
modgh2 <- glmer(stable ~ Surface + Vision + (1|Subject), nAGQ=25,
    ↪ family=binomial, data=ctsib)
```

```
anova(modgh, modgh2)
```

Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
modgh2	5	247	268	-119	237		
modgh	9	248	286	-115	230	7.37	4 0.12

This uses the standard likelihood-based methods to construct a chi-squared test. We have the same reasons as with LMMs to view these results with some scepticism. Even so, this is a balanced experiment of a reasonable size so this provides some confidence in the result. We see that a simplification to just the treatment variables as fixed effects seems reasonable. If we feel uncomfortable with this conclusion, we

may further point to the minimization of AIC (or BIC) as a justification for choosing the smaller model.

As with all such models, it is wise to check some diagnostics. These can be extracted using `residuals()` and `fitted()` functions. An alternative convenience is:

```
dd <- fortify(modgh2)
```

which extracts the residuals and fitted values and places them in a common data frame with the other variables. This makes the construction of some plots more convenient. For example, we might look at the QQ plots subsetted by the treatment variables:

```
ggplot(dd, aes(sample=.resid))+stat_qq() + facet_grid(Surface~Vision)
```

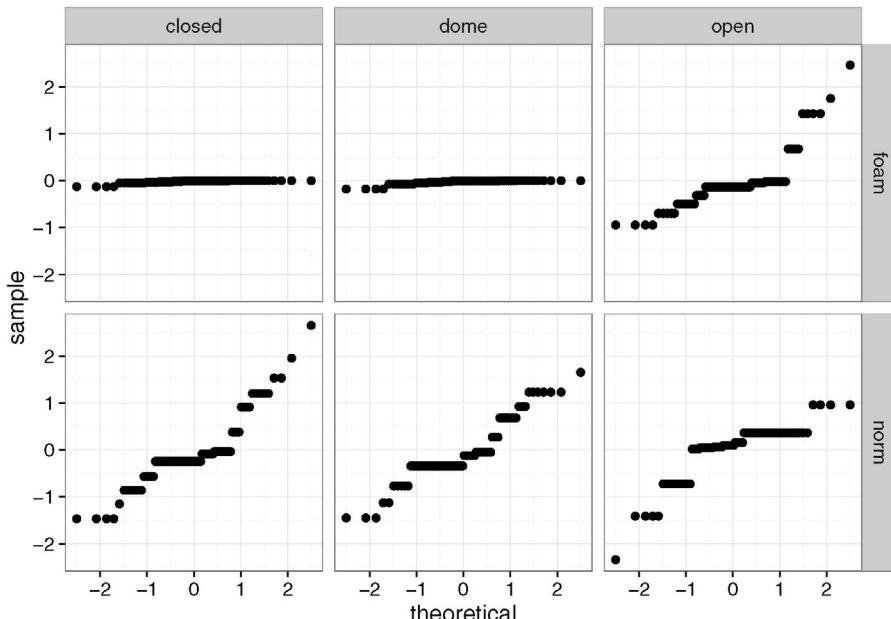


Figure 13.2 *QQ plots subsetted by treatment variables.*

In Figure 13.2, we see that the residuals are close to zero for two of the six combinations. This is because these were universally unstable conditions and have been predicted as such by the model. In the most stable, normal and open condition, larger positive residuals are not seen because there is no headroom for such cases. It would be a mistake to view this plot as indicating heteroscedascity as we have seen there are more convincing explanations for the differences in spread.

We can use INLA for a Bayesian approach to fitting these models. See Section 12.2 for an introduction. For ease of exposition, we use only the surface and vision as fixed effect predictors. The default, noninformative priors, are satisfactory:

```
library(INLA)
formula <- stable ~ Surface + Vision + f(Subject, model="iid")
result <- inla(formula, family="binomial", data=ctsib)
```

We compute the SD for the subject random effect:

```
sigmaalpha <- inla.tmarginal(function(x) 1/sqrt(x), result$marginals.
  ↪ hyperpar$"Precision for Subject")
```

The posterior density for this SD is shown in the first panel of Figure 13.3:

```
x <- seq(0,7,length.out = 100)
sdf <- data.frame(yield = x, density=inla.dmarginal(x, sigmaalpha))
ggplot(sdf,aes(x=yield,y=density))+geom_line()
```

We see that the subject effect is clear since the distribution is well away from zero but there is some uncertainty regarding the size of the effect.

We can produce a numerical summary of the posteriors:

```
restab <- sapply(result$marginals.fixed, function(x) inla.zmarginal(x,
  ↪ silent=TRUE))
restab <- cbind(restab, inla.zmarginal(sigmaalpha,silent=TRUE))
colnames(restab) = c("mu", "norm", "dome", "open", "alpha")
data.frame(restab)
```

	mu	norm	dome	open	alpha
mean	-10.298	7.3641	0.66618	6.1279	3.0248
sd	1.3507	0.92526	0.49873	0.8498	0.62416
quant0.025	-13.172	5.7182	-0.29877	4.6184	1.9838
quant0.25	-11.181	6.6971	0.32503	5.517	2.5756
quant0.5	-10.21	7.3038	0.65785	6.0704	2.9585
quant0.75	-9.334	7.9657	0.99581	6.6771	3.4007
quant0.975	-7.9392	9.3029	1.6579	7.9167	4.429

We could compute similar statistics on the subject random effects but there are too many to display them all. We see that the posterior means are quite similar to the last glmer-based fit. We can plot the posterior densities of the fixed effects as seen in the second panel of Figure 13.3:

```
x <- seq(-2,11,length.out = 100)
rden <- sapply(result$marginals.fixed,function(y) inla.dmarginal(x, y
  ↪ )[, -1]
ddf <- data.frame(yield=rep(x,3), density=as.vector(rden), treat=gl
  ↪ (3,100, labels=c("norm", "dome", "open")))
ggplot(ddf, aes(x=yield, y=density, linetype=treat))+geom_line()
```

The norm level of surface and the open level of vision are clearly different from the respective reference levels since the densities are well separated from zero. In contrast, we see there may not be much difference between the dome and closed levels of vision as this density overlaps zero. We can compute a “Bayesian  $p$ -value” as:

```
2*inla.pmarginal(0,result$marginals.fixed$Visiondome)
[1] 0.17982
```

We have multiplied by two to account for the usual two-sided testing argument. In this context,  $p$ -values do not have the same meaning. Nonetheless, it does serve as a measure of how the posterior density relates to zero. This confirms our impression that there is not much difference between the levels.

We can also use STAN for a Bayesian analysis as introduced in Section 12.1.

Here is the STAN program we need:

```
data {
  int<lower=0> Nobs;
  int<lower=0> Nsubs;
  int<lower=0> Npreds;
  int<lower=0,upper=1> y[Nobs];
  int<lower=1,upper=Nsubs> subject[Nobs];
```

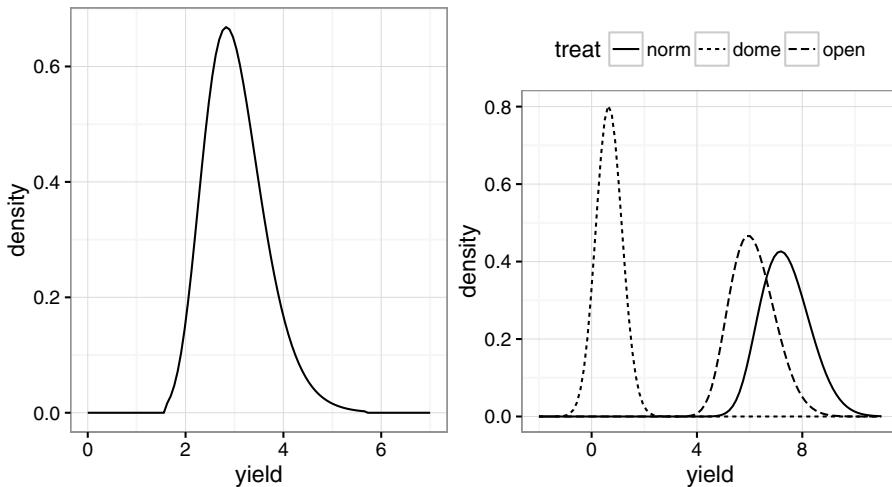


Figure 13.3 *Posterior density for the subject SD on the left and posterior densities for the treatment effects on the right.*

```

    matrix[Nobs,Npreds] x;
}
parameters {
  vector[Nsubs] subeff;
  real<lower=0> sigmasubj;
  vector[Npreds] beta;
}
model {
  subeff ~ normal(0,sigmasubj);
  sigmasubj ~ cauchy(0, 1);
  for(n in 1:Nobs) {
    y[n] ~ bernoulli_logit(x[n]*beta + subeff[subject[n]] );
  }
}

```

We have written this in a generic form so that you could use this for any grouped-by-subject data with a binary response. We use a half-Cauchy prior for the subject SD. This is somewhat more informative but seems justifiable in the context of this data. It also has the advantage of being more transparent.

We need to prepare the data in a format compatible with the data block in the code above. We form the model matrix of fixed effects,  $X$ , in advance:

```

xm <- model.matrix(~ Sex + Age + Height + Weight + Surface + Vision,
  ↪ ctsib)
stabledat <- with(ctsib, list(Nobs=nrow(ctsib),
  Nsubs=length(unique(ctsib$Subject)), Npreds=ncol(xm),
  y=stable, subject=Subject, x=xm))

```

We can now run the STAN model. We have broken the process into three steps. The first step translates the STAN code into C++, the second compiles that C++ code and the third runs the MCMC sampler. The advantage of doing it in three stages is that

one is likely only to do the first two once but the third might need to be repeated if the model or data is changed.

```
library(rstan)
rt <- stanc("glmmbin.stan")
sm <- stan_model(stanc_ret = rt, verbose=FALSE)
fit <- sampling(sm, data=stabledat)
```

This will take several minutes to run depending on the quality of your computer.

First we need to check the diagnostics of the MCMC sampling. We plot the chain for the subject SD as this is the parameter most likely to cause problems:

```
traceplot(fit, pars="sigmasubj", inc_warmup=FALSE)
```

The plot (not shown) is entirely satisfactory. We can display a summary for the parameters of interest:

```
print(fit, pars=c("sigmasubj", "beta"))
    mean se_mean    sd 2.5% 25% 50% 75% 97.5% n_eff Rhat
sigmasubj 3.60    0.04  0.81  2.28 3.03 3.48 4.09 5.36 515 1.01
beta[1]   19.25   0.73 17.56 -13.12 7.66 18.52 30.11 55.08 585 1.01
beta[2]    3.83   0.10  2.24 -0.45 2.35 3.76 5.27 8.33 531 1.01
beta[3]   -0.01   0.00  0.11 -0.23 -0.08 -0.01 0.06 0.20 836 1.00
beta[4]   -0.23   0.00  0.12 -0.48 -0.30 -0.22 -0.15 -0.01 586 1.01
beta[5]    0.09   0.00  0.08 -0.06 0.03 0.08 0.14 0.24 860 1.00
beta[6]    8.56   0.06  1.33  6.28 7.58 8.43 9.39 11.33 560 1.01
beta[7]    0.75   0.01  0.56 -0.34 0.38 0.74 1.14 1.87 4000 1.00
beta[8]    7.24   0.05  1.24  5.17 6.34 7.13 8.01 9.92 563 1.01
```

The effective sample sizes are more than satisfactory.

Now we examine the posterior distributions. We extract the parameters of interest and restore the variable names for convenience. The `reshape2` package helps us arrange the data in a format for convenient plotting. We show the estimated densities in Figure 13.4 along with a vertical line at zero.

```
ipars <- data.frame(extract(fit, pars=c("sigmasubj", "beta")))
colnames(ipars)[-1] <- colnames(xm)
library(reshape2)
rdf <- melt(ipars)
ggplot(rdf, aes(x=value)) + geom_density() + facet_wrap(~ variable,
  scales="free") + geom_vline(xintercept=0)
```

We might also be interested in how the subjects in the experiment compare. We extract the subject random effects and sort the posterior means:

```
ppars <- data.frame(extract(fit, pars="subeff"))
sort(colMeans(ppars))
subeff.3 subeff.38 subeff.37 subeff.14
-6.704126 -4.926872 -4.563769 -4.036449 ...
...edited...
subeff.17 subeff.29 subeff.25 subeff.27
3.488328 5.906336 6.735636 6.924570
```

We see that subject 3 is the least stable and subject 27 is the most stable. Since we have access to the posterior distributions, we can readily investigate which difference might be notable.

### 13.4 Count Response

In this example, we have data from a clinical trial of 59 epileptics. For a baseline, patients were observed for 8 weeks and the number of seizures recorded. The patients

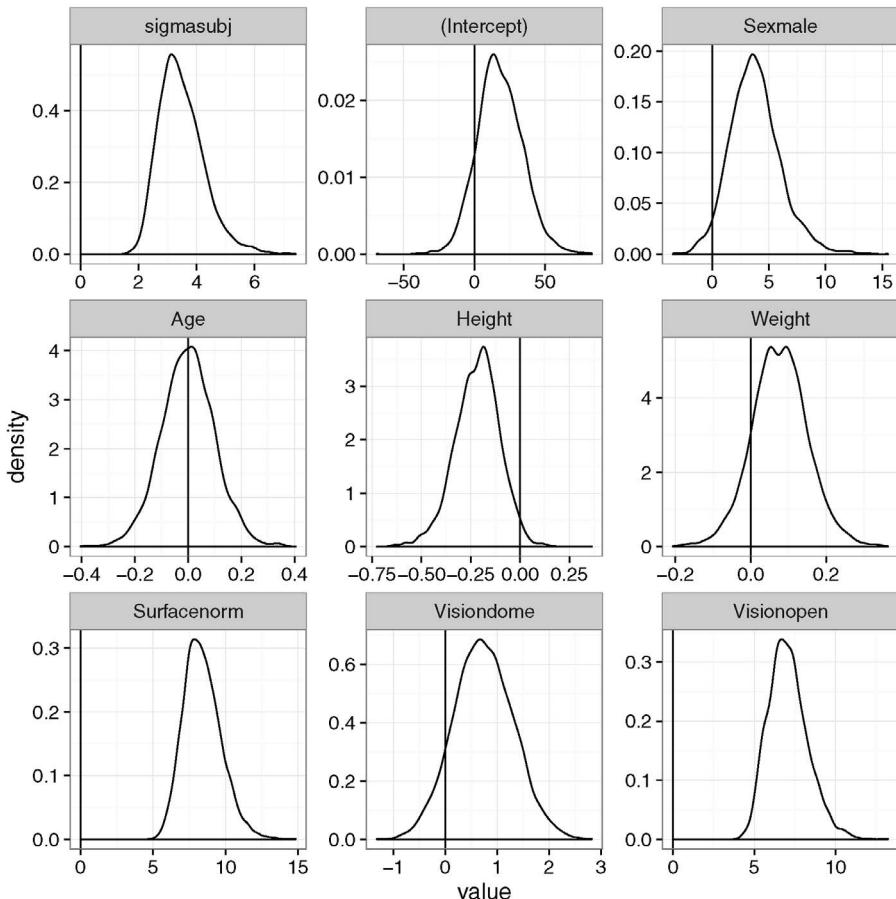


Figure 13.4 Posterior distributions as produced by the STAN fit to the epilepsy data.

were then randomized to treatment by the drug Progabide (31 patients) or to the placebo group (28 patients). They were observed for four 2-week periods and the number of seizures recorded. The data have been analyzed by many authors including Thall and Vail (1990), Breslow and Clayton (1993) and Diggle et al. (2013). Does Progabide reduce the rate of seizures?

First we create some derived variables and then look at the first two patients:

```
data(epilepsy, package="faraway")
epilepsy$period <- rep(0:4, 59)
epilepsy$drug <- factor(c("placebo", "treatment") [epilepsy$treat+1])
epilepsy$phase <- factor(c("baseline", "experiment") [epilepsy$expind
  ↪ +1])
epilepsy[epilepsy$id < 2.5, ]
```

	seizures	id	treat	expind	timeadj	age	period	drug	phase
1	11	1	0	0	8	31	0	placebo	baseline
2	5	1	0	1	2	31	1	placebo	experiment

```

3      3 1      0      1      2 31      2 placebo experiment
4      3 1      0      1      2 31      3 placebo experiment
5      3 1      0      1      2 31      4 placebo experiment
6     11 2      0      0      8 30      0 placebo baseline
7      3 2      0      1      2 30      1 placebo experiment
8      5 2      0      1      2 30      2 placebo experiment
9      3 2      0      1      2 30      3 placebo experiment
10     3 2      0      1      2 30      4 placebo experiment

```

Both were not treated (`treat=0`). The `expind` indicates the baseline phase by 0 and the treatment phase by 1. The length of these time phases is recorded in `timeadj`. We have created three new convenience variables: `period`, denoting the 2- or 8-week periods, `drug` recording the type of treatment in nonnumeric form and `phase` indicating the phase of the experiment.

We now compute the mean number of seizures *per week* broken down by the treatment and baseline vs. experimental period. The `dplyr` package is useful for these types of group summaries:

```

library(dplyr)
epilepsy %>%
  group_by(drug, phase) %>%
  summarise(rate=mean(seizures/timeadj)) %>%
  xtabs(formula=rate ~ phase + drug)

```

drug	placebo	treatment
baseline	3.8482	3.9556
experiment	4.3036	3.9839

We see that the rate of seizures in the treatment group actually increases during the period in which the drug was taken. The rate of seizures also increases even more in the placebo group. Perhaps some other factor is causing the rate of seizures to increase during the treatment period and the drug is actually having a beneficial effect. Now we make some plots to show the difference between the treatment and the control. The first plot shows the difference between the two groups during the experimental period only:

```

ggplot(epilepsy, aes(x=period, y=seizures, linetype=drug, group=id))
  + geom_line() + xlim(1,4) + scale_y_sqrt(breaks=(0:10)^2) +
  + theme(legend.position = "top", legend.direction = "horizontal")

```

We compare the two groups in the left panel of Figure 13.5 and find little to choose between them. The square-root transform is used to stabilize the variance; this is often used with count data. Now we compare the average seizure rate to the baseline for the two groups:

```

ratesum <- epilepsy %>%
  group_by(id, phase, drug) %>%
  summarise(rate=mean(seizures/timeadj))
library(tidyr)
comsum <- spread(ratesum, phase, rate)
ggplot(comsum, aes(x=baseline, y=experiment, shape=drug)) + geom_point
  + scale_x_sqrt() + scale_y_sqrt() + geom_abline(intercept=0,
  + slope=1)+ theme(legend.position = "top", legend.direction = "
  + horizontal")

```

A treatment effect, if one exists, is not readily apparent. Now we fit GLMM models. Patient #49 is unusual because of the high rate of seizures observed. We exclude it:

```

epilo <- filter(epilepsy, id != 49)
```

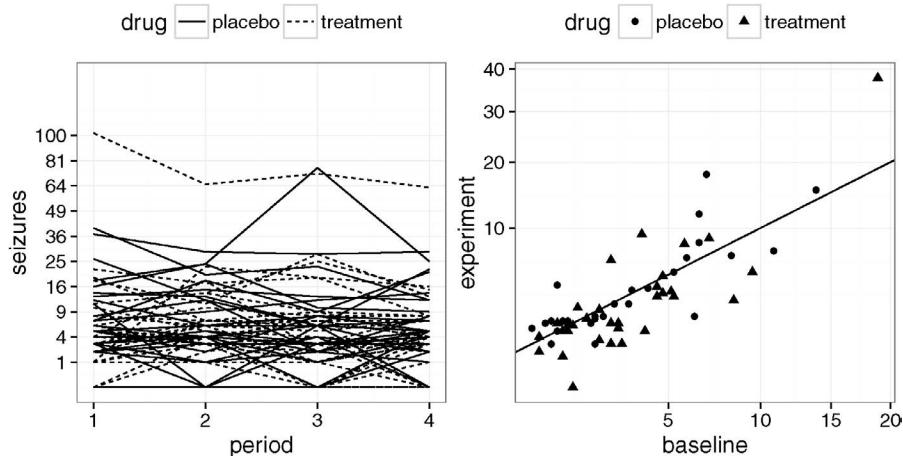


Figure 13.5 *Seizures per 2-week period on a square-root scale with treatment group shown as solid lines and the placebo group shown as dotted lines in the plot on the left. Mean seizures per week is shown on the right. We compare the baseline period with the experimental period, distinguishing those who receive treatment or control.*

Excluding a case should not be taken lightly. It is worth repeating the analysis with and without this subject. For projects where the analyst works with producers of the data, it will be possible to discuss substantive reasons for excluding cases. Exclusion of cases should always be reported and not concealed.

It is worth starting with a GLM even though the model is not correct due to the grouping of the observations. We must use an *offset* as explained in Section 5.3 to allow for the difference in lengths in the baseline and treatment periods.

```
modglm <- glm(seizures ~ offset(log(timeadj)) + expind + treat + I(
  ← expind*treat), family=poisson, data=epilo)
summary(modglm)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.3476	0.0341	39.57	< 2e-16
expind	0.1118	0.0469	2.39	0.017
treat	-0.1068	0.0486	-2.20	0.028
I(expind * treat)	-0.3024	0.0697	-4.34	0.000014

n = 290 p = 4

Deviance = 2411.550 Null Deviance = 2485.110 (Difference = 73.560)

The interaction term is the primary parameter of interest. All the subjects were untreated in the baseline, even the ones who were subsequently treated. This means that the main effect for treatment does properly measure the response to treatment because it includes the baseline period. As we have observed already, we suspect the response may have been different during the baseline time and the active period of the experiment. The interaction term represents the effect of the treatment during the baseline period after adjustment. In the output above we see that this interaction seems highly significant and negative (which is good since we want to reduce

seizures). But this inference is suspect because we have made no allowance for the correlated responses within individuals. The *p*-value is far smaller than it should be.

We might also consider allowing for overdispersion in the response by using a quasi-Poisson model as discussed in Section 9.4. However, this is a different consideration to the correlated response.

We move through the estimation options in the same order as with the binary response example earlier, starting with PQL:

```
library(MASS)
modpql <- glmmPQL(seizures ~offset(log(timeadj)) + expind + treat + I(
  ↪ expind*treat), random = ~1|id, family=poisson, data=epilo)
summary(modpql)

Formula: ~1 | id
          (Intercept) Residual
StdDev:      0.68197   1.6054

Variance function:
Structure: fixed weights
Formula: ~invwt

Fixed effects: seizures ~ offset(log(timeadj)) + expind + treat + I(expind*treat)
                Value Std.Error DF t-value p-value
(Intercept)     1.08079  0.143701 230  7.5211  0.0000
expind         0.11184  0.075767 230  1.4761  0.1413
treat        -0.00894  0.200244  56 -0.0446  0.9646
I(expind * treat) -0.30238  0.112689 230 -2.6834  0.0078
```

The parameter estimates are comparable to the GLM but the standard errors are larger as might be expected given that the correlated response has been allowed for. As with the binary response example, we still have some doubts about the accuracy of the inference. This is a particular concern when some count responses are small. A further concern is the problematic meaning of the residual SD as such a term does not appear in the statement of the model. Also we lack an AIC due to the quasi-ness of the likelihood. Even so, we do see a significant negative interaction effect indicating that the drug is effective.

Numerical quadrature can also be used. We use Gauss-Hermite in preference to Laplace as the model random effect structure is simple and so the computation is fast even though we have used the most expensive nAGQ=25 setting.

```
library(lme4)
modgh <- glmer(seizures ~offset(log(timeadj)) + expind + treat + I(
  ↪ expind*treat)+ (1|id), nAGQ=25, family=poisson, data=epilo)
summary(modgh)

Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 0.515    0.718
 Number of obs: 290, groups: id, 58
```

```
Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.03600  0.14126   7.33  2.2e-13
expind         0.11184  0.04688   2.39   0.017
treat        -0.00815  0.19652  -0.04    0.967
I(expind * treat) -0.30239  0.06971  -4.34  1.4e-05
```

We see that the interaction effect is significant. Notice that the estimate of this effect

has been quite consistent over all the estimation methods so we draw some confidence from this. We have

```
exp(-0.302)
```

```
[1] 0.73934
```

So the drug is estimated to reduce the rate of seizures by about 26%. However, the subject SD is more than twice the drug effect of  $-0.3$  at  $0.718$ . This indicates that the expected improvement in the drug is substantially less than the variation between individuals. Interpretation of the main effect terms is problematic in the presence of an interaction. For example, the treatment effect reported here represents the predicted difference in the response during the baseline period (i.e.,  $\text{expind}=0$ ). Since none of the subjects are treated during the baseline period, we are reassured to see that this effect is not significant. However, this does illustrate the danger in naively presuming that this is the treatment effect.

We can also take a Bayesian approach, starting with STAN. We can use almost the same code as the binary response example except we need to add

```
vector[Nobs] offset;
```

in the data block and replace the model line with

```
y[n] ~ poisson_log(log(offset[n])+x[n]*beta + subeff[subject[n]] );
```

We prepare the data into the required format using:

```
epilo$id[epilo$id == 59] <- 49
xm <- model.matrix(~ expind + treat + I(expind*treat), epilo)
epildat <- with(epilo, list(Nobs=nrow(epilo), Nsubs=length(unique(id)),
                           Npreds=ncol(xm),
                           y=seizures,
                           subject=id,
                           x=xm, offset=timeadj))
```

We've renumbered case 59 into the previously deleted case 49 slot. This is ugly but we need the subjects to be consecutively numbered.

Assuming that the code is placed in a file called `glmmpois.stan`, we translate, compile and run the sampler:

```
library(rstan)
rt <- stanc("glmmpois.stan")
sm <- stan_model(stanc_ret = rt, verbose=FALSE)
fit <- sampling(sm, data=epildat)
```

We can check the sampling properties of the chain by

```
traceplot(fit, pars="sigmasubj", inc_warmup=FALSE)
```

We've made the plot only for subject SD since this is the one most likely to cause problems. In this case, the plot (not shown) is satisfactory. We can review the posterior distributions:

```
ipars <- data.frame(rstan::extract(fit, pars=c("sigmasubj", "beta")))
colnames(ipars) <- c("subject", "intercept", "expind", "treat",
                     ↪ interaction")
```

We plot the two most interesting posterior distributions as seen in Figure 13.6:

```
ggplot(ipars, aes(x=subject))+geom_density()
ggplot(ipars, aes(x=interaction))+geom_density() +geom_vline(
  ↪ xintercept=0)
```

We can see the subject SD is very clearly different from zero and that it is centered on about 0.7. The interaction effect (or drug effect) is negative centered on about  $-0.3$ . This looks clearly less than zero.

We can construct a convenient summary of the results including a sort of  $p$ -value.

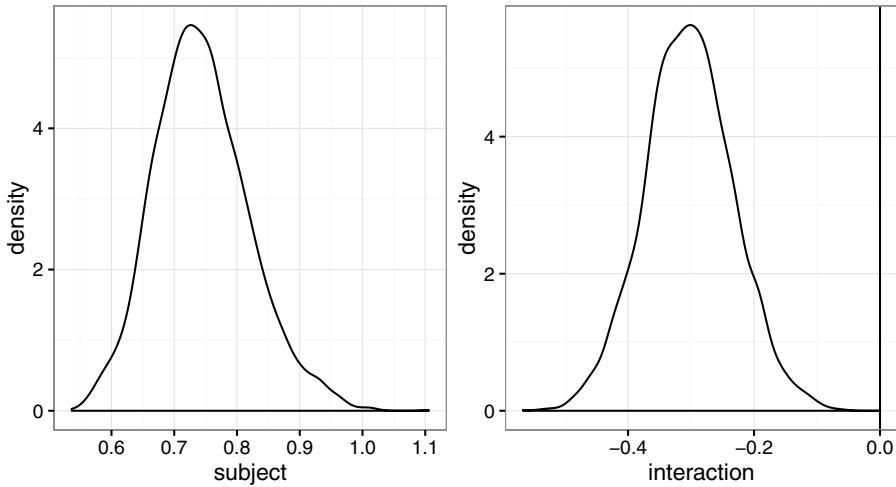


Figure 13.6 Posterior distribution of the subject SD and drug effect for the epilepsy data.

```

bayespval <- function(x) {p <- mean(x > 0); 2*min(p,1-p)}
smat <- apply(ipars, 2, function(x) c(mean(x), quantile(x,c(0.025,
  ↪ 0.975)), bayespval(x)))
row.names(smat) <- c("mean", "LCB", "UCB", "pvalue")
t(smat)
      mean      LCB      UCB pvalue
subject 0.743504 0.613355 0.90935 0.0000
intercept 1.021433 0.736583 1.30687 0.0000
expind   0.110912 0.017704 0.20310 0.0210
treat     0.005876 -0.386557 0.39178 0.9695
interaction -0.301699 -0.434090 -0.17111 0.0000

```

We see that the posterior means reported here are quite similar to the estimates computed earlier using likelihood methods. The Bayesian analysis seems rather more confident that there is a drug effect since this posterior distribution is well separated from zero.

The same model can also be fit by INLA in a straightforward way:

```

formula <- seizures ~offset(log(timeadj)) + expind + treat + I(expind*
  ↪ treat) + f(id,model="iid")
result <- inla(formula, family="poisson", data = epilo)

```

We obtain a summary of the posteriors as:

```

sigmaalpha <- inla.tmarginal(function(x) 1/sqrt(x), result$marginals.
  ↪ hyperpar$"Precision for id")
restab <- sapply(result$marginals.fixed, function(x) inla.zmarginal(x,
  ↪ silent=TRUE))
restab <- cbind(restab, inla.zmarginal(sigmaalpha,silent=TRUE))
colnames(restab) = c("mu", "expind", "treat", "interaction", "alpha")
data.frame(restab)

```

	mu	expind	treat	interaction	alpha
mean	1.036	0.11178	-0.0081489	-0.30246	0.72583
sd	0.14196	0.046891	0.19753	0.06973	0.070948

quant0.025	0.75556	0.019681	-0.39732	-0.4396	0.59944
quant0.25	0.94023	0.079891	-0.14164	-0.34985	0.67548
quant0.5	1.0354	0.11151	-0.0093067	-0.3028	0.72079
quant0.75	1.1304	0.14314	0.12302	-0.25579	0.77093
quant0.975	1.3133	0.20349	0.3787	-0.16626	0.87796

We see that the results are similar to those obtained previously. We observe that the 95% credible interval for the interaction is  $(-0.44, -0.17)$  so we are sure that this parameter differs from zero as in the STAN analysis. We can compute further numerical and graphical summaries as in previous examples obtaining very similar results.

### 13.5 Generalized Estimating Equations

The advantage of the quasi-likelihood approach as described in Section 9.4 compared to GLMs was that we did not need to specify the distribution of the response. We only needed to give the link function and the variance. We can adapt this approach for repeated measures and/or longitudinal studies. Let  $Y_i$  be a vector of random variables representing the responses on a given individual or cluster and let  $EY_i = \mu_i$  which is then linked to the linear predictor using  $g(\mu_i) = x_i^T \beta$ , where  $g$  is a link function appropriate to the response type and  $x_i$  is the predictor vector.

As with the quasi-likelihood, we also need to specify a variance function  $a()$ :

$$\text{var } Y_i = \phi a(\mu_i)$$

Certain choices of  $a()$  will be sensible depending on the type of response. The  $\phi$  is a scale parameter which may be set to one if not needed.

In addition, we must also specify how the responses within an individual or cluster are correlated with each other. We set a *working correlation matrix*  $R_i(\alpha)$  depending on a parameter  $\alpha$  which we will estimate. This results in a *working covariance matrix* for  $Y_i$ :

$$V_i = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2}$$

where  $A_i$  is a diagonal matrix formed from  $a(\mu_i)$ .

Given estimates of  $\phi$  and  $\alpha$ , we can estimate  $\beta$  by setting the (multivariate) score function to zero and solving:

$$\sum_i \left( \frac{\partial \mu_i}{\partial \beta} \right)^T V_i^{-1} (Y_i - \mu_i) = 0$$

These equations can be regarded as the multivariate analogue of those used for the quasi-likelihood models described in Section 9.4. Since  $\text{var } Y$  also depends on  $\alpha$ , we substitute any consistent estimate of  $\alpha$  in this equation and still obtain an estimate as asymptotically efficient as if  $\alpha$  were known. A similar set of equations can be derived representing the score with respect to  $\alpha$ , which may be similarly solved. We iterate between estimating  $\alpha$  and  $\beta$  until we converge at a solution.

These are called *generalized estimating equations* (GEE). Note that no specification of the distribution has been necessary which makes the fitting and specification much simpler. The estimates of  $\beta$  are consistent even if the variance is misspecified.

We use the `geepack` package as described in Højsgaard et al. (2005). The `gee` package can also fit these models with somewhat different features.

We reanalyze the stability dataset:

```
data(ctsib, package="faraway")
ctsib$stable <- ifelse(ctsib$CTSIB==1, 1, 0)
library(geepack)
modgeep <- geeglm(stable ~ Sex + Age + Height + Weight + Surface +
  → Vision, id=Subject, corstr="exchangeable", scale.fix=TRUE,
  → data=ctsib, family=binomial)
```

We have specified the same fixed effects as in the corresponding GLMM earlier. The grouping variable is specified by the `id` argument. Only simple groups are allowed while nested grouping variables cannot be accommodated easily in this function. We must choose the correlation structure within each group. If we choose no correlation, then the problem reduces to a standard GLM. Several choices are available. For this data, it seems reasonable to assume that any pair of observations from the same subject have the same correlation. This is known as an *exchangeable* correlation or, equivalently, *compound symmetry*. We have chosen to fix the scale parameter at the default value of 1 to ensure maximum compatibility with the GLMM fit. Otherwise, there would not be a strong reason to fix this. Let us now examine the output:

```
summary(modgeep)
```

Coefficients:

	Estimate	Std.err	Wald	Pr(> W )
(Intercept)	8.6233	5.9199	2.12	0.145
Sexmale	1.6449	0.9035	3.31	0.069
Age	-0.0121	0.0480	0.06	0.802
Height	-0.1021	0.0424	5.80	0.016
Weight	0.0437	0.0340	1.65	0.199
Surfacenorm	3.9163	0.5668	47.74	4.9e-12
Visiondome	0.3589	0.4040	0.79	0.374
Visionopen	3.1799	0.4606	47.66	5.1e-12

Scale is fixed.

Correlation: Structure = exchangeable Link = identity

Estimated Correlation Parameters:

Estimate	Std.err
alpha	0.218 0.0447
Number of clusters:	40 Maximum cluster size: 12

We can see that the estimated correlation between observations on the same subject is 0.22. The standard error of 0.04 indicates that we can be quite sure there is a correlation in the responses within individuals.

The standard errors are constructed using a *sandwich estimator* as described in Section 8.5. These are typically, but not always, larger than the naive standard errors from the likelihood calculations. These standard errors can be used to construct Wald statistics. We see that the treatment factors, surface and vision, are significant. Height and possibly gender are marginally significant. This part of the conclusion is similar to our GLMM results.

There is one clear difference with the GLMM output: the estimates for the GEE are about half the size of the GLMM  $\beta$ s. This is to be expected. GLMMs model the

data at the subject or individual level. The correlation between the measurements on the individual is generated by the random effect. Thus the  $\beta$ s for the GLMM represent the effect on an individual. A GEE models the data at the population level. The  $\beta$ s for a GEE represent the effect of the predictors averaged across all individuals with the same predictor values. GEEs do not use random effects but model the correlation at the marginal or correlation level.

The testing for vision is not entirely satisfactory since it has three levels meaning two tests—one being highly significant and the other not at all. If we want a single test for the significance of vision, we need to refit the model without vision and make the standard anova-type comparison:

```
modgeep2 <- geeglm(stable ~ Sex + Age + Height + Weight + Surface, id
  ↪ =Subject, corstr="exchangeable", scale.fix=TRUE, data=ctsib,
  ↪ family=binomial)
anova(modgeep2, modgeep)
```

Analysis of 'Wald statistic' Table

```
Model 1 stable ~ Sex + Age + Height + Weight + Surface + Vision
Model 2 stable ~ Sex + Age + Height + Weight + Surface
  Df   X2 P(>|Chi|)
1  2 58.4  2.1e-13
```

As expected, we see that vision is strongly significant.

The geepack package also offers the possibility of modeling an ordinal response with clusters using the `ordgee()` function. This would be appropriate for the original form of this data where the response is measured on a four-point scale.

We can also model the epilepsy data:

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

We exclude the 49th case as before. An autoregressive AR(1) model for the correlation structure is most natural since consecutive measurements will be more correlated than measurements separated in time. Note that this does require that the clusters be sorted in time order — they are in this case.

```
modgeep <- geeglm(seizures ~offset(log(timeadj)) + expind + treat + I(
  ↪ expind*treat), id=id, family=poisson, corstr="ar1", data=
  ↪ epilepsy, subset=(id!=49))
summary(modgeep)
```

Coefficients:

	Estimate	Std.err	Wald	Pr(> W )
(Intercept)	1.3138	0.1616	66.10	4.4e-16
expind	0.1509	0.1108	1.86	0.173
treat	-0.0797	0.1983	0.16	0.688
I(expind * treat)	-0.3987	0.1745	5.22	0.022

Estimated Scale Parameters:

	Estimate	Std.err
(Intercept)	10.6	2.35

Correlation: Structure = ar1 Link = identity

Estimated Correlation Parameters:

	Estimate	Std.err
--	----------	---------

alpha	0.783	0.0519
-------	-------	--------

Number of clusters: 58 Maximum cluster size: 5

The drug effects, as measured by the interaction term, has a just significant effect.

The dispersion parameter is estimated as 10.6. This means that if we did not account for the overdispersion, the standard errors would be much larger. The AR(1) correlation structure can be seen in the working correlation where adjacent measurements have 0.78 correlation.

Further analysis would involve an investigation of alternative correlation structures, the age covariate and any trend during the experimental period. The analysis of this dataset is discussed in Diggle et al. (2013).

**Further Reading:** McCulloch and Searle (2002) have some coverage of GLMMs as well as more material on GLMs. Hardin and Hilbe (2003) give a book-length treatment of GEEs. Diggle et al. (2013) discuss both topics.

## Exercises

1. The `ohio` data concern 536 children from Steubenville, Ohio and were taken as part of a study on the effects of air pollution. Children were in the study for 4 years from ages 7 to 10. The response was whether they wheezed or not. The variables are:

**resp** an indicator of wheeze status (1 = yes, 0 = no)

**id** an identifier for the child

**age** 7 yrs = -2, 8 yrs = -1, 9 yrs = 0, 10 yrs = 1

**smoke** an indicator of maternal smoking at the first year of the study (1 = smoker, 0 = nonsmoker)

- (a) Do any of the mothers in the study change their smoking status during the period of observation?
- (b) Construct a table that shows proportion of children who wheeze for 0, 1, 2, 3 or 4 years broken down by maternal smoking status.
- (c) Make plot which shows how the proportion of children wheezing changes by age with a separate line for smoking and nonsmoking mothers.
- (d) Group the data by child to count the total (out of four) years of wheezing. Fit a binomial GLM to this response to check for a maternal smoking effect. Does this prove there is a smoking effect or could there be another plausible explanation?
- (e) Fit a model for each individual response using a GLMM fit using penalized quasi-likelihood. Describe the effects of age and maternal smoking. How do the odds of wheezing change numerically over time?
- (f) Now fit the same model but using adaptive Gaussian-Hermit quadrature. Compare to the previous model fit.
- (g) Use `INLA` to fit the same model. What does this model say about the effect of age and maternal smoking?
- (h) Use `STAN` to fit the same model. Check the MCMC diagnostics and again discuss the age and maternal smoking effects.

- (i) Fit the model using GEE. Use an autoregressive rather than exchangeable error structure. Compare the results to the previous model fits. In your model, what indicates that a child who already wheezes is likely to continue to wheeze?
  - (j) What is your overall conclusion regarding the effect of age and maternal smoking? Can we trust the GLM result or are the GLMM models preferable?
2. The National Youth Survey collected a sample of 11–17 year olds, 117 boys and 120 girls, asking questions about marijuana usage. The data is presented in potuse.
- (a) Plot the total number of people falling into each usage category as it varies over time separately for each sex.
  - (b) Condense the levels of the response into whether the person did or did not use marijuana that year. Turn the year into a numerical variable. Fit a GLMM for the now binary response with an interaction between sex and year as a predictor using Gauss-Hermite quadrature. Comment on the effect of sex.
  - (c) Fit a reduced model without sex and use it to test for the significance of sex in the larger model.
  - (d) Fit a model with year as a factor. Should this model be preferred to the model with year as just a linear term? Interpret the estimated effects in the year as a factor version of the model.
  - (e) Fit GEE version of the model and compare it to the analogous GLMM fit.
3. Components are attached to an electronic circuit card assembly by a wave-soldering process. The soldering process involves baking and preheating the circuit card and then passing it through a solder wave by conveyor. Defects arise during the process. The design is  $2^{7-3}$  with three replicates and the data is found in wavesolder.
- (a) Plot the data to show how the number of defects varies with the predictors.
  - (b) Fit a Poisson GLM to the individual runs with the number of defects as the response and main effects for all the predictors. How can you tell that this model is inadequate? Fit a comparable quasi-Poisson GLM. What difference does this make to the significance of the predictors?
  - (c) Sum the defects within each replicate group of three and fit a quasi-Poisson GLM to these sums. Compare the fitted model to the previous one.
  - (d) Fit a GEE model to the individual defect responses with a fixed scale that allows for an autoregressive correlation structure within the groups (assuming that the replicates are in time order). Is it reasonable to fix the scale?
  - (e) Now refit without a fixed scale. Is there any evidence of a correlation between successive replicates?
  - (f) Finally fit a GEE model with an independent correlation structure within the replicates. Compare this model to the quasi-Poisson GLM fit.
4. The nitrofen data in boot package come from an experiment to measure the reproductive toxicity of the pesticide nitrofen on a species of zooplankton called *Ceriodaphnia dubia*. Each animal produced three broods in which the number of

live offspring was recorded. Fifty animals in total were used and divided into five batches. Each batch was treated in a solution with a different concentration of the pesticide.

- (a) Plot the total number of live offspring as they vary with concentration and comment. Now plot the numbers for each brood, taking care to distinguish the different broods. Is the trend within each brood the same?
  - (b) Fit a GLMM for the number of live offspring within each brood that varies with concentration and brood number (including an interaction). The model should take account of the dependence between observations from the same animal. Describe what the model says about how the number of live offspring change with concentration for the different broods.
  - (c) Fit an equivalent GEE model and compare it to the GLMM result.
5. The toenail data comes from a multicenter study comparing two oral treatments for toenail infection. Patients were evaluated for the degree of separation of the nail. Patients were randomized into two treatments and were followed over seven visits: four in the first year and yearly thereafter. The patients have not been treated prior to the first visit so this should be regarded as the baseline.
- (a) Calculate the proportion of patients with a normal or severe condition broken down by treatment and visit number. Plot these proportions and comment.
  - (b) Fit a GLMM for the outcome as a function of an interaction between the visit and the treatment. Since the two groups are selected at random, there should be no difference at the first visit. Does this model show a significant difference at this baseline (first visit)?
  - (c) Test for a significant treatment effect by fitting a model without treatment and comparing to the previous model.

## Nonparametric Regression

---

The generalized linear model was an extension of the linear model  $y = X\beta + \varepsilon$  to allow the responses  $y$  from the exponential family. The mixed effect models allowed for a much more general treatment of  $\varepsilon$ . We now switch our attention to the linear predictor  $\eta = X\beta$ . We want to make this more flexible. There are a wide variety of available methods, but it is best to start with simple regression. The methods developed here form part of the solution to the multiple predictor problem.

Given fixed  $x_1, \dots, x_n$ , we observe  $y_1, \dots, y_n$  where:

$$y_i = f(x_i) + \varepsilon_i$$

where the  $\varepsilon_i$  are i.i.d. and have mean zero and unknown variance  $\sigma^2$ . The problem is to estimate the function  $f$ .

A parametric approach is to assume that  $f(x)$  belongs to a parametric family of functions:  $f(x|\beta)$ . So  $f$  is known up to a finite number of parameters. Some examples are:

$$\begin{aligned}f(x|\beta) &= \beta_0 + \beta_1 x \\f(x|\beta) &= \beta_0 + \beta_1 x + \beta_2 x^2 \\f(x|\beta) &= \beta_0 + \beta_1 x^{\beta_2}\end{aligned}$$

The parametric approach is quite flexible because we are not constrained to just linear predictors as in the first model of the three above. We can add many different types of terms such as polynomials and other functions of the variable to achieve flexible fits. Nonlinear models, such as the third case above, are also parametric in nature. Nevertheless, no matter what finite parametric family we specify, it will always exclude many plausible functions.

The nonparametric approach is to choose  $f$  from some smooth family of functions. Thus the range of potential fits to the data is much larger than the parametric approach. We do need to make some assumptions about  $f$  — that it has some degree of smoothness and continuity, for example, but these restrictions are far less limiting than the parametric way.

The parametric approach has the advantage that it is more efficient if the model is correct. If you have good information about the appropriate model family, you should prefer a parametric model. Parameters may also have intuitive interpretations. Nonparametric models do not have a formulaic way of describing the relationship between the predictors and the response; this often needs to be done graphically. This relates to another advantage of parametric models in that they reduce information

necessary for prediction; you can write down the model formula, typically in a compact form. Nonparametric models are less easily communicated on paper. Parametric models also enable easy utilization of past experience.

The nonparametric approach is more flexible. In modeling new data, one often has very little idea of an appropriate form for the model. We do have a number of heuristic tools using diagnostic plots to help search for this form, but it would be easier to let the modeling approach take care of this search. Another disadvantage of the parametric approach is that one can easily choose the wrong form for the model and this results in bias. The nonparametric approach assumes far less and so is less liable to make bad mistakes. The nonparametric approach is particularly useful when little past experience is available.

For our examples we will use three datasets, one real (data on the Old Faithful geyser in Yellowstone National Park, Wyoming, USA) and two simulated, called `exa` and `exb`. The data comes from Härdle (1991). The reason we use simulated data is to see how well the estimates match the true function (which cannot usually be known for real data). We plot the data in the first three panels of Figure 14.1, using a line to mark the true function where known. For `exa`, the true function is  $f(x) = \sin^3(2\pi x^3)$ . For `exb`, it is constant zero, that is,  $f(x) = 0$ :

```
data(exa, package="faraway")
plot(y ~ x, exa, main="Example A")
lines(m ~ x, exa, lwd=2)
data(exb, package="faraway")
plot(y ~ x, exb, main="Example B")
lines(m ~ x, exb, lwd=2)
plot(waiting ~ eruptions, faithful, main="Old Faithful")
```

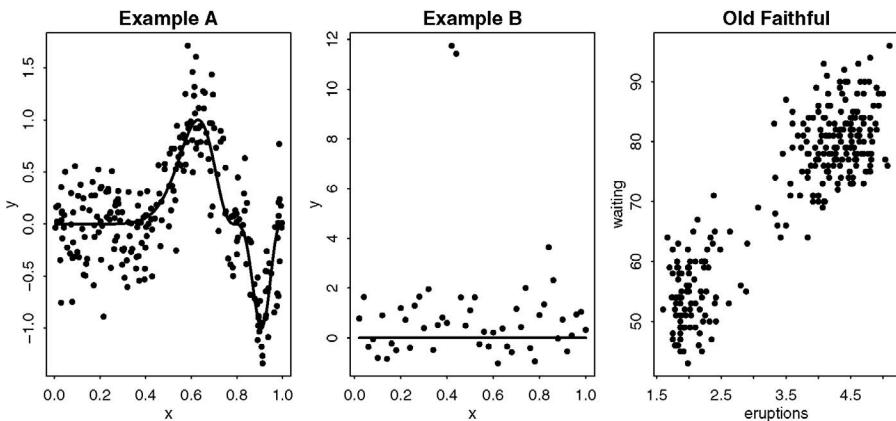


Figure 14.1 *Data examples. Example A has varying amounts of curvature, two optima and a point of inflection. Example B has two outliers. The Old Faithful provides the challenges of real data.*

We now examine several widely used nonparametric regression estimators, also known as *smoothers*.

### 14.1 Kernel Estimators

In its simplest form, this is just a moving average estimator. More generally, our estimate of  $f$ , called  $\hat{f}_\lambda(x)$ , is:

$$\hat{f}_\lambda(x) = \frac{1}{n\lambda} \sum_{j=1}^n K\left(\frac{x-x_j}{\lambda}\right) Y_j = \frac{1}{n} \sum_{j=1}^n w_j Y_j \quad \text{where} \quad w_j = K\left(\frac{x-x_j}{\lambda}\right)/\lambda$$

$K$  is a kernel where  $\int K = 1$ . The moving average kernel is rectangular, but smoother kernels can give better results.  $\lambda$  is called the bandwidth, window width or smoothing parameter. It controls the smoothness of the fitted curve.

If the  $x$ s are spaced very unevenly, then this estimator can give poor results. This problem is somewhat ameliorated by the *Nadaraya–Watson* estimator:

$$f_\lambda(x) = \frac{\sum_{j=1}^n w_j Y_j}{\sum_{j=1}^n w_j}$$

We see that this estimator simply modifies the moving average estimator so that it is a true weighted average where the weights for each  $y$  will sum to one.

It is worth understanding the basic asymptotics of kernel estimators. The optimal choice of  $\lambda$  gives:

$$\text{MSE}(x) = E(f(x) - \hat{f}_\lambda(x))^2 = O(n^{-4/5})$$

MSE stands for mean squared error and we see that this decreases at a rate proportional to  $n^{-4/5}$  with the sample size. Compare this to the typical parametric estimator where  $\text{MSE}(x) = O(n^{-1})$ , provided that the parametric model is correct. So the kernel estimator is less efficient. Indeed, the relative difference between the MSEs becomes substantial as the sample size increases. However, if the parametric model is incorrect, the MSE will be  $O(1)$  and the fit will not improve past a certain point even with unlimited data. The advantage of the nonparametric approach is the protection against model specification error. Without assuming much stronger restrictions on  $f$ , nonparametric estimators cannot do better than  $O(n^{-4/5})$ .

The implementation of a kernel estimator requires two choices: the kernel and the smoothing parameter. For the choice of kernel, smoothness and compactness are desirable. We prefer smoothness to ensure that the resulting estimator is smooth, so for example, the uniform kernel will give stepped-looking fit that we may wish to avoid. We also prefer a compact kernel because this ensures that only data, local to the point at which  $f$  is estimated, is used in the fit. This means that the Gaussian kernel is less desirable, because although it is light in the tails, it is not zero, meaning that the contribution of every point to the fit must be computed. The optimal choice under some standard assumptions is the Epanechnikov kernel:

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2) & |x| < 1 \\ 0 & \text{otherwise} \end{cases}$$

This kernel has the advantage of some smoothness, compactness and rapid computation. This latter feature is important for larger datasets, particularly when resampling

techniques like bootstrap are being used. Even so, any sensible choice of kernel will produce acceptable results, so the choice is not crucially important.

The choice of smoothing parameter  $\lambda$  is critical to the performance of the estimator and far more important than the choice of kernel. If the smoothing parameter is too small, the estimator will be too rough; but if it is too large, important features will be smoothed out.

We demonstrate the Nadaraya–Watson estimator next for a variety of choices of bandwidth on the Old Faithful data shown in Figure 14.2. We use the `ksmooth` function which is part of the R base package. This function lacks many useful features that can be found in some other packages, but it is adequate for simple use. The default uses a uniform kernel, which is somewhat rough. We have changed this to the normal kernel:

```
for(bw in c(0.1,0.5,2)) {
  with(faithful, {
    plot(waiting ~ eruptions, col=gray(0.75))
    lines(ksmooth(eruptions,waiting,"normal",bw))
  })
}
```

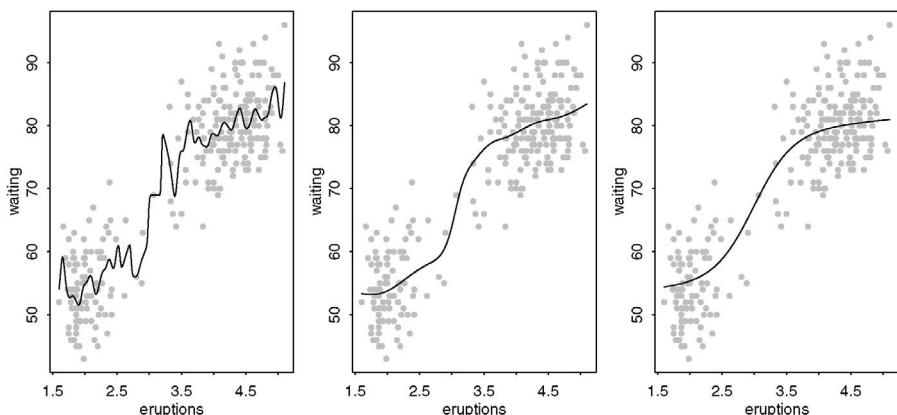


Figure 14.2 Nadaraya–Watson kernel smoother with a normal kernel for three different bandwidths on the Old Faithful data.

The central plot in Figure 14.2 is the best choice of the three. Since we do not know the true function relating waiting time and eruption duration, we can only speculate, but it does seem reasonable to expect that this function is quite smooth. The fit on the left does not seem plausible since we would not expect the mean waiting time to vary so much as a function of eruptions. On the other hand, the plot on the right is even smoother than the plot in the middle. It is not so easy to choose between these. Another consideration is that the eye can always visualize additional smoothing, but it is not so easy to imagine what a less smooth fit might look like. For this reason, we recommend picking the least smooth fit that does not show any implausible fluctuations. Of the three plots shown, the middle plot seems best. Smoothers are often used as a graphical aid in interpreting the relationship between variables. In such cases,

visual selection of the amount of smoothing is effective because the user can employ background knowledge to make an appropriate choice and avoid serious mistakes.

You can choose  $\lambda$  interactively using this subjective method. Plot  $\hat{f}_\lambda(x)$  for a range of different  $\lambda$  and pick the one that looks best as we have done above. You may need to iterate the choice of  $\lambda$  to focus your decision. Knowledge about what the true relationship might look like can be readily employed.

In cases where the fitted curve will be used to make numerical predictions of future values, the choice of the amount of smoothing has an immediate effect on the outcome. Even here subjective methods may be used. If this method of selecting the amount of smoothing seems disturbingly subjective, we should also understand that the selection of a family of parametric models for the same data would also involve a great deal of subjective choice although this is often not explicitly recognized. Statistical modeling requires us to use our knowledge of what general forms of relationship might be reasonable. It is not possible to determine these forms from the data in an entirely objective manner. Whichever methodology you use, some subjective decisions will be necessary. It is best to accept this and be honest about what these decisions are.

Even so, automatic methods for selecting the amount of smoothing are also useful. Selecting the amount of smoothing using subjective methods requires time and effort. When a large number of smooths are necessary, some automation is desirable. In other cases, the statistician will want to avoid the explicit appearance of subjectivity in the choice. Cross-validation (CV) is a popular general-purpose method. The criterion is:

$$CV(\lambda) = \frac{1}{n} \sum_{j=1}^n (y_j - \hat{f}_{\lambda(j)}(x_j))^2$$

where  $(j)$  indicates that point  $j$  is left out of the fit. We pick the  $\lambda$  that minimizes this criterion. True cross-validation is computationally expensive, so an approximation to it, known as *generalized cross-validation* or GCV, is sometimes used. There are also many other methods of automatically selecting the  $\lambda$ .

Our practical experience has been that automatic methods, such as CV, often work well, but sometimes produce estimates that are clearly at odds with the amount of smoothing that contextual knowledge would suggest. For this reason, we are unwilling to trust automatic methods completely. We recommend using them as a starting point for a possible interactive exploration of the appropriate amount of smoothing if time permits. They are also useful when very large numbers of smooths are needed such as in the additive modeling approach described in Chapter 15.

When smoothing is used to determine whether  $f$  has certain features such as multiple maximums (called *bump hunting*) or monotonicity, special methods are necessary to choose the amount of smoothing since this choice will determine the outcome of the investigation.

The `sm` library, described in Bowman and Azzalini (1997), allows the computation of the cross-validated choice of smoothing parameter. For example, we find the CV choice of smoothing parameter for the Old Faithful and plot the result:

```
library(sm)
```

```
with(faithful, sm.regression(eruptions, waiting, h=h.select(eruptions,
  ↪ waiting)))
```

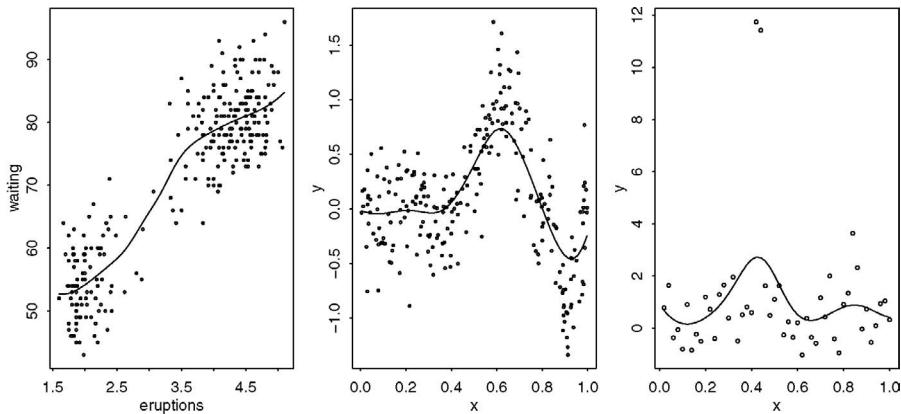


Figure 14.3 The first panel shows the kernel estimated smooth of the Old Faithful data for a cross-validated choice of smoothing parameter. The second and third panels show the resulting fits for Examples A and B, respectively.

We see the resulting fit plotted in the first panel of Figure 14.3. The `sm` package uses a Gaussian kernel where the smoothing parameter is the standard deviation of the kernel.

We repeat the exercise for Example A; the plots are shown in the second panel of Figure 14.3. The resulting fit is somewhat oversmoothed.

```
with(exa, sm.regression(x, y, h=h.select(x,y)))
```

Finally, we compute the fit for Example B as seen in the third panel of Figure 14.3.

```
with(exb, sm.regression(x, y, h=h.select(x,y)))
```

We see that the fitted curve notices the two outliers but does not reach out to them. A much smaller choice of smoothing parameter would allow the fitted curve to pass near these two points but only at the price of a much rougher fit elsewhere.

## 14.2 Splines

**Smoothing Splines:** The model is  $y_i = f(x_i) + \varepsilon_i$ , so in the spirit of least squares, we might choose  $\hat{f}$  to minimize the MSE:  $\frac{1}{n} \sum (y_i - f(x_i))^2$ . The solution is  $\hat{f}(x_i) = y_i$ . This is a “join the dots” regression that is almost certainly too rough. Instead, suppose we choose  $\hat{f}$  to minimize a modified least squares criterion:

$$\frac{1}{n} \sum (Y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx$$

where  $\lambda > 0$  is the smoothing parameter and  $\int [f''(x)]^2 dx$  is a *roughness penalty*. When  $f$  is rough, the penalty is large, but when  $f$  is smooth, the penalty is small. Thus the two parts of the criterion balance fit against smoothness. This is the *smoothing spline* fit.

For this choice of roughness penalty, the solution is of a particular form:  $\hat{f}$  is a cubic spline. This means that  $\hat{f}$  is a piecewise cubic polynomial in each interval  $(x_i, x_{i+1})$  (assuming that the  $x_i$ s are unique and sorted). It has the property that  $\hat{f}$ ,  $\hat{f}'$  and  $\hat{f}''$  are continuous. Given that we know the form of the solution, the estimation is reduced to the parametric problem of estimating the coefficients of the polynomials. This can be done in a numerically efficient way.

Several variations on the basic theme are possible. Other choices of roughness penalty can be considered, where penalties on higher-order derivatives lead to fits with more continuous derivatives. We can also use weights by inserting them in the sum of squares part of the criterion. This feature is useful when smoothing splines are means to an end for some larger procedure that requires weighting. A robust version can be developed by modifying the sum of squares criterion to:

$$\sum \rho(y_i - f(x_i)) + \lambda \int [f''(x)]^2 dx$$

where  $\rho(x) = |x|$  is one possible choice.

In R, cross-validation is used to select the smoothing parameter by default. We show this default choice of smoothing for our three test cases:

```
with(faithful, {
  plot(waiting ~ eruptions, col=gray(0.75))
  lines(smooth.spline(eruptions, waiting), lty=2)
})
with(exa, {
  plot(y ~ x, col=gray(0.75))
  lines(x, m)
  lines(smooth.spline(x, y), lty=2)
})
with(exb, {
  plot(y ~ x, col=gray(0.75))
  lines(x, m)
  lines(smooth.spline(x, y), lty=2)
})
```

The fits may be seen in Figure 14.4. The fit for the Old Faithful data looks reasonable. The fit for Example A does a good job of tracking the hills and valleys but overfits in the smoother region. The default choice of smoothing parameter given by CV is a disaster for Example B as the data is just interpolated. This illustrates the danger of blindly relying on automatic smoothing parameter selection methods.

**Regression Splines:** Regression splines differ from smoothing splines in the following way: for regression splines, the number of knots of the B-splines used for the basis are typically much smaller than the sample size. The number of knots chosen controls the amount of smoothing. For smoothing splines, the observed unique  $x$  values are the knots and  $\lambda$  is used to control the smoothing. It is arguable whether the regression spline method is parametric or nonparametric, because once the knots are chosen, a parametric family has been specified with a finite number of parameters. It is the freedom to choose the number of knots that makes the method nonparametric. One of the desirable characteristics of a nonparametric regression estimator is that it should be consistent for smooth functions. This can be achieved for regression

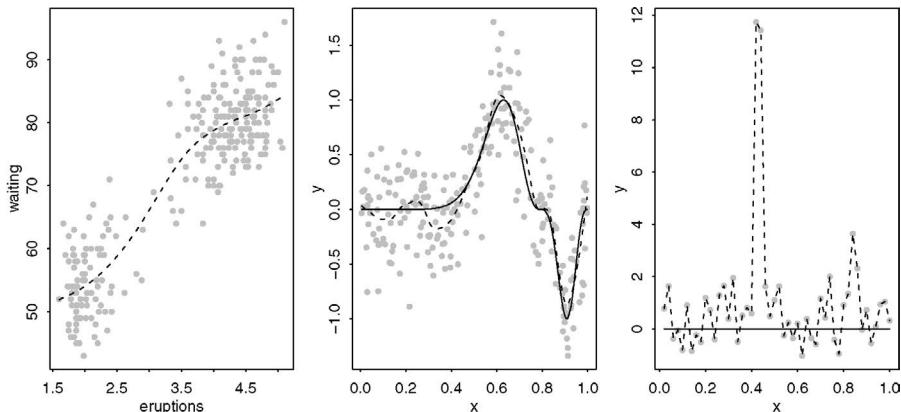


Figure 14.4 Smoothing spline fits. For Examples A and B, the true function is shown as solid and the spline fit as dashed.

splines if the number of knots is allowed to increase at an appropriate rate with the sample size.

We demonstrate some regression splines here. We use piecewise linear splines in this example, which are constructed and plotted as follows: knots

```
rhs <- function(x,c) ifelse(x>c,x-c,0)
curve(rhs(x,0.5),0,1)
```

where the spline is shown in the first panel of Figure 14.5. Now we define some knots for Example A:

```
(knots <- 0:9/10)
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
```

and compute a design matrix of splines with knots at these points for each  $x$ :

```
dm <- outer(exa$x,knots rhs)
matplot(exa$x,dm,type="l",col=1, xlab="x", ylab="")
```

where the basis functions are shown in the second panel of Figure 14.5. Now we compute and display the regression fit:

```
lmod <- lm(exa$y ~ dm)
plot(y ~ x, exa, col=gray(0.75))
lines(exa$x, predict(lmod))
```

where the plot is shown in the first panel of Figure 14.6. Because the basis functions are piecewise linear, the fit is also piecewise linear. A better fit may be obtained by adjusting the knots so that they are denser in regions of greater curvature:

```
newknots <- c(0,0.5,0.6,0.65,0.7,0.75,0.8,0.85,0.9,0.95)
dmn <- outer(exa$x,newknots,rhs)
lmod <- lm(exa$y ~ dmn)
plot(y ~ x, exa, col=gray(0.75))
lines(exa$x, predict(lmod))
```

where the plot is shown in the second panel of Figure 14.6. We obtain a better fit but only by using our knowledge of the true curvature. This knowledge would not be available for real data, so more practical methods place the knots adaptively according to the *estimated* curvature.

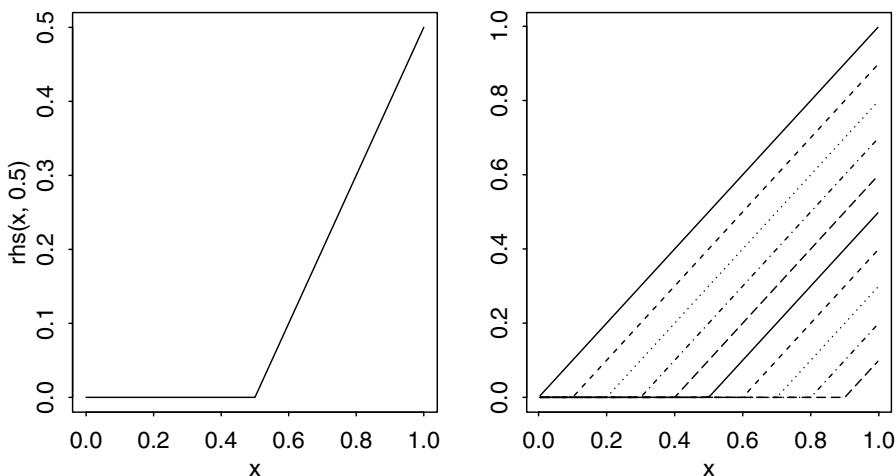


Figure 14.5 One basis function for linear regression splines shown on the left and the complete set shown on the right.

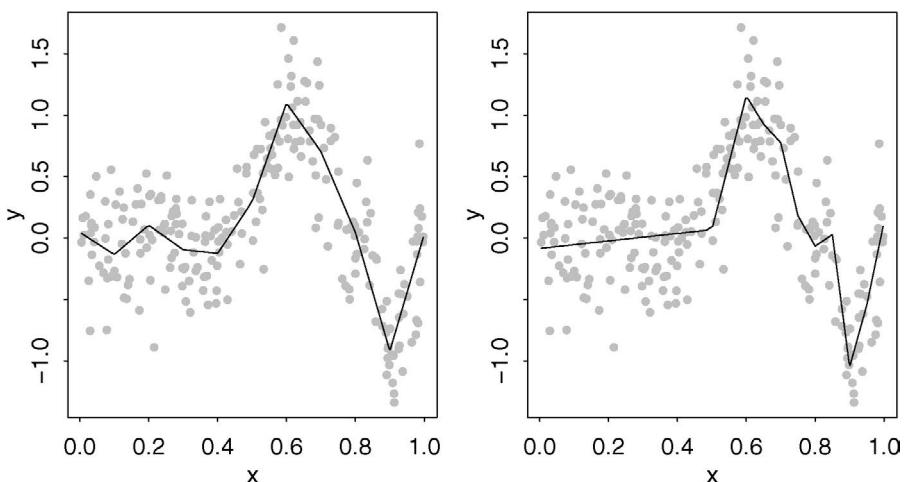


Figure 14.6 Evenly spaced knots fit shown on the left and knots spread relative to the curvature on the right.

One can achieve a smoother fit by using higher-order splines. The `bs()` function can be used to generate the appropriate spline basis. The default is cubic B-splines. We display 12 cubic B-splines evenly spaced on the  $[0,1]$  interval. The splines close to the boundary take a different form as seen in the first panel of Figure 14.7:

```
library(splines)
matplot(bs(seq(0,1,length=1000), df=12), type="l", ylab="", col=1)
```

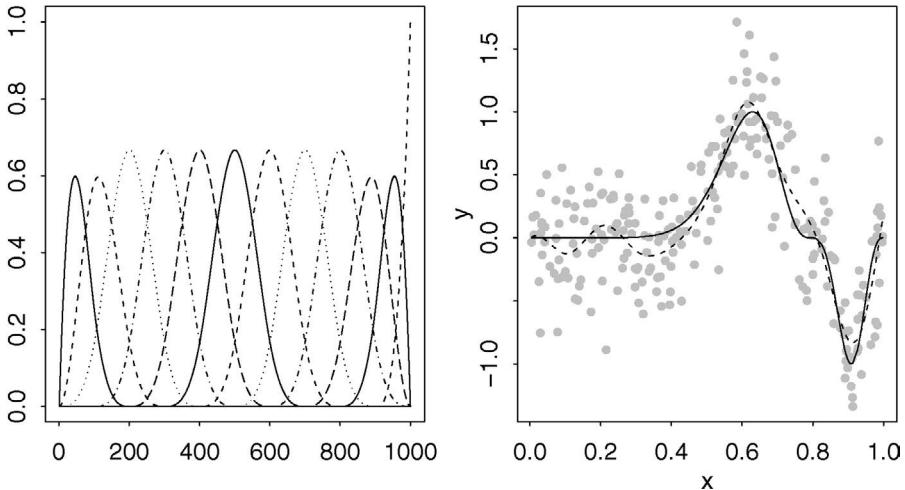


Figure 14.7 A cubic B-spline basis is shown in the left panel and the resulting fit to the Example A data is shown in the right panel.

We can now use least squares to determine the coefficients. We then display the fit as seen in the second panel of Figure 14.7:

```
lmod <- lm(y ~ bs(x, 12), exa)
plot(y ~ x, exa, col=gray(0.75))
lines(m ~ x, exa)
lines(predict(lmod) ~ x, exa, lty=2)
```

We see a smooth fit, but again we could do better by placing more knots at the points of high curvature and fewer in the flatter regions.

### 14.3 Local Polynomials

Both the kernel and spline methods have been relatively vulnerable to outliers as seen by their performance on Example B. The fits can be improved with some manual intervention, either to remove the outliers or to increase the smoothing parameters. However, smoothing is frequently just a small part of an analysis and so we might wish to avoid giving each smooth individual attention. Furthermore, habitual removal of outliers is an ad hoc strategy that is better replaced with a method that deals with long-tailed errors gracefully. The local polynomial method combines robustness ideas from linear regression and local fitting ideas from kernel methods.

First we select a window. We then fit a polynomial to the data in that window using robust methods. The predicted response at the middle of the window is the

fitted value. We then slide the window over the range of the data, repeating the fitting process as the window moves. The most well-known implementation of this type of smoothing is called *lowess* or *loess* and is due to Cleveland (1979).

As with any smoothing method, there are choices to be made. We need to choose the order of the polynomial fit. A quadratic allows us to capture peaks and valleys in the function. However, a linear term also performs well and is the default choice in the *loess* function. As with most smoothers, it is important to pick the window width well. The default choice takes three quarters of the data and may not be a good choice as we shall see below.

For the Old Faithful data, the default choice is satisfactory, as seen in the first panel of Figure 14.8:

```
with(faithful, {
  plot(waiting ~ eruptions, col=gray(0.75))
  f <- loess(waiting ~ eruptions)
  i <- order(eruptions)
  lines(f$x[i], f$fitted[i])
})
```

For Example A, the default choice is too large. The choice that minimizes the integrated squared error between the estimated and true function requires a span (proportion of the range) of 0.22. Both fits are seen in the middle panel of Figure 14.8:

```
with(exa, {
  plot(y ~ x, col=gray(0.75))
  lines(m ~ x)
  f <- loess(y ~ x)
  lines(f$x, f$fitted, lty=2)
  f <- loess(y ~ x, span=0.22)
  lines(f$x, f$fitted, lty=5)
})
```

In practice, the true function is, of course, unknown and we would need to select the span ourselves, but this optimal choice does at least show how well loess can do in the best of circumstances. The fit is similar to that for smoothing splines.

For Example B, the optimal choice of span is one (that is all the data). This is not surprising since the true function is a constant and so maximal smoothing is desired. We can see that the robust qualities of loess prevent the fit from becoming too distorted by the two outliers even with the default choice of smoothing span:

```
with(exb, {
  plot(y ~ x, col=gray(0.75))
  lines(m ~ x)
  f <- loess(y ~ x)
  lines(f$x, f$fitted, lty=2)
  f <- loess(y ~ x, span=1)
  lines(f$x, f$fitted, lty=5)
})
```

## 14.4 Confidence Bands

It is helpful to have some expression of uncertainty in the curve estimates. Both the regression spline and loess methods use (local) linear fitting using parametric methods. These same methods naturally generate a standard error which can be used to construct a confidence interval at any point in  $x$ . We may connect these intervals to

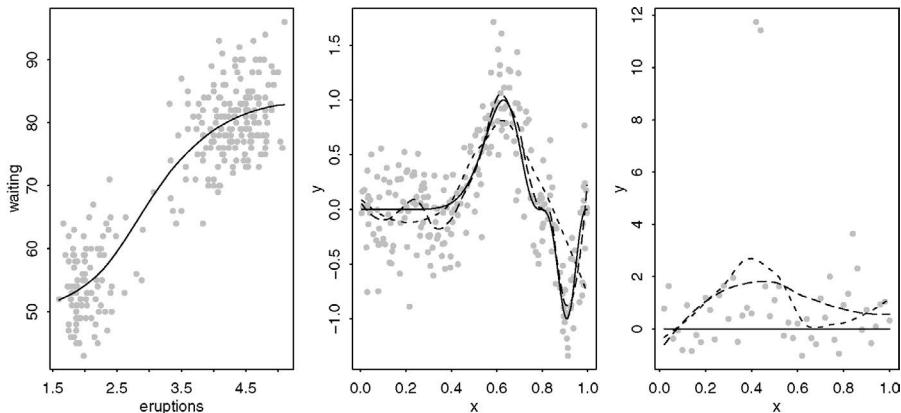


Figure 14.8 *Loess smoothing*: *Old Faithful* data is shown in the left panel with the default amount of smoothing. Example A data is shown in the middle and B in the right panel. The true function is shown as a solid line along with the default choice (dotted) and respective optimal amounts of smoothing (dashed) are also shown.

gether to form a confidence band. These are conveniently constructed and displayed using the `ggplot2` package.

We construct the 95% confidence band for Example A data using `loess`:

```
ggplot(exa, aes(x=x, y=y)) + geom_point(alpha=0.25) + geom_smooth(
  → method="loess", span=0.22) + geom_line(aes(x=x, y=m), linetype=2)
```

The plot is seen in the first panel of Figure 14.9. We have added the true function as a dashed line. We observe that the true function falls just outside the band in a few areas. However, the band we have constructed is a *pointwise* confidence band. The 95% confidence applies at each point but since we have a wide range of points, the 95% probability of the interval containing the true value cannot hold across the range. For this we would need a *simultaneous* confidence band.

We can also construct a band using splines. We need the `mgcv` package which includes a spline smoother.

```
library(mgcv)
ggplot(exa, aes(x=x, y=y)) + geom_point(alpha=0.25) + geom_smooth(
  → method="gam", formula=y ~ s(x, k=20)) + geom_line(aes(x=x, y=m),
  → linetype=2)
```

We see the resulting plot in the second panel of Figure 14.9. In this case, we have manually chosen the smoothing parameter,  $k=20$ , representing the degrees of freedom in the fit. It is larger to accommodate the variation in this function, producing a better although not perfect fit to that seen in Figure 14.6.

## 14.5 Wavelets

Regression splines are an example of a basis function approach to fitting. We approximate the curve by a family of basis functions,  $\phi_i(x)$ , so that  $\hat{f}(x) = \sum_i c_i \phi_i(x)$ . Thus the fit requires estimating the coefficients,  $c_i$ . The choice of basis functions will

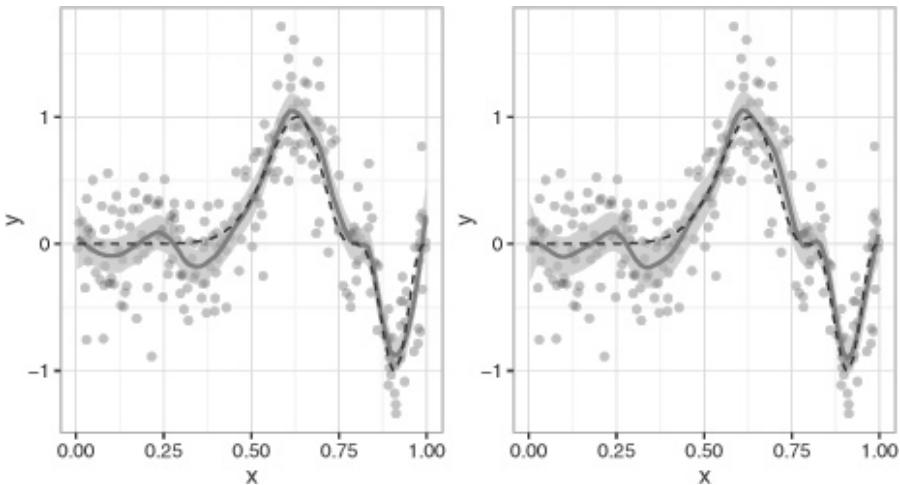


Figure 14.9 95% confidence bands for loess (left) and spline (right) fits to Example A.

determine the properties of the fitted curve. The estimation of  $c_i$  is particularly easy if the basis functions are orthogonal.

Examples of orthogonal bases are orthogonal polynomials and the Fourier basis. The disadvantage of both these families is that the basis functions are not compactly supported so that the fit of each basis function depends on the whole data. This means that these fits lack the desirable local fit properties that we have seen in previously discussed smoothing methods. Although Fourier methods are popular for some applications, particularly those involving periodic data, they are not typically used for general-purpose smoothing.

Cubic B-splines are compactly supported, but they are not orthogonal. *Wavelets* have the advantage that they are compactly supported and can be defined so as to possess the orthogonality property. They also possess the *multiresolution* property which allows them to fit the grosser features of the curve while focusing on the finer detail where necessary.

We begin with the simplest type of wavelet: the *Haar* basis. The *mother wavelet* for the Haar family is defined on the interval  $[0, 1]$  as:

$$w(x) = \begin{cases} 1 & x \leq 1/2 \\ -1 & x > 1/2 \end{cases}$$

We generate the members of the family by dilating and translating this function. The next two members of the family are defined on  $[0, 1/2)$  and  $[1/2, 1)$  by rescaling the mother wavelet to these two intervals. The next four members are defined on the quarter intervals in the same way. We can index the family members by level  $j$  and within the level by  $k$  so that each function will be defined on the interval  $[k/2^j, (k+1)/2^j)$  and takes the form:

$$h_n(x) = 2^{j/2} w(2^j x - k)$$

where  $n = 2^j + k$  and  $0 \leq k \leq 2^j$ . Plotting these functions reveals that they are orthogonal. They are also orthonormal, because they integrate to 1. Furthermore, they have a local basis where the support becomes narrower as the level is increased. Computing the coefficients is particularly quick because of these properties.

Wavelet fitting can be implemented using the `wavethresh` package of Nason (2013). The first step is to make the wavelet decomposition. We will illustrate this with Example A:

```
library(wavethresh)
wds <- wd(exa$y, filter.number=1, family="DaubExPhase")
```

The filter number specifies the complexity of the family. The Haar basis is the simplest available but is not the default choice. We can now show the mother wavelet and wavelet coefficients:

```
draw(wds)
plot(wds)
```

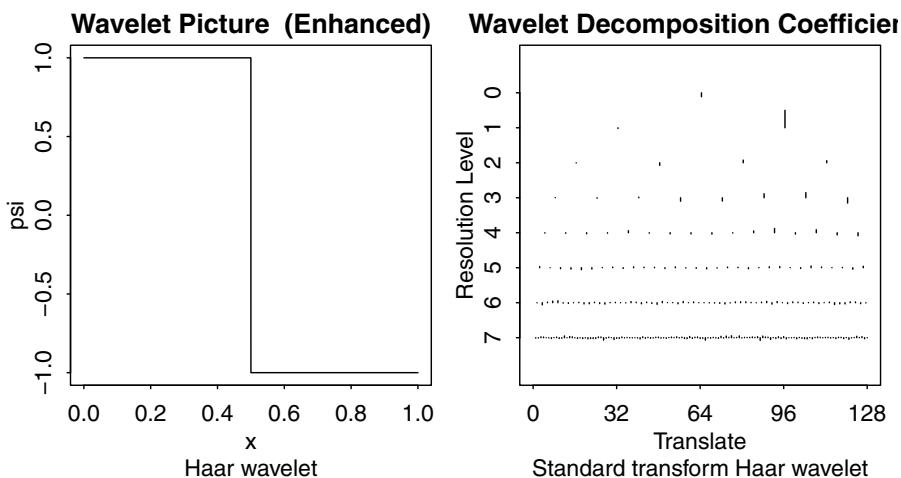


Figure 14.10 *Haar mother wavelet and wavelet coefficients from decomposition for Example A.*

We can see the Haar mother wavelet in the left panel of Figure 14.10. We see the wavelet decomposition in the right panel.

Suppose we wanted to compress the data into a more compact format. Smoothing can be viewed as a form of compression because it retains the important features of the data while discarding some of the finer detail. The smooth is described in terms of the fitted coefficients which are fewer than the number of data points. The method would be called *lossy* since some information about the original data is lost.

For example, suppose we want to smooth the data extensively. We could throw away all the coefficients of level four or higher and then reconstruct the function as follows:

```
wtd <- threshold(wds, policy="manual", value=9999)
fd <- wr(wtd)
```

Only level-three and higher coefficients are retained. There are only  $2^3 = 8$  of these.

The thresholding here applies to level four and higher only by default. Any coefficient less than 9999 in absolute value is set to zero — that is, all of them in this case. The `wr()` inverts the wavelet transform. We now plot the result as seen in the first panel of Figure 14.11:

```
plot(y ~ x, exa, col=gray(0.75))
lines(m ~ x, exa)
lines(fd ~ x, exa, lty=5, lwd=2)
```

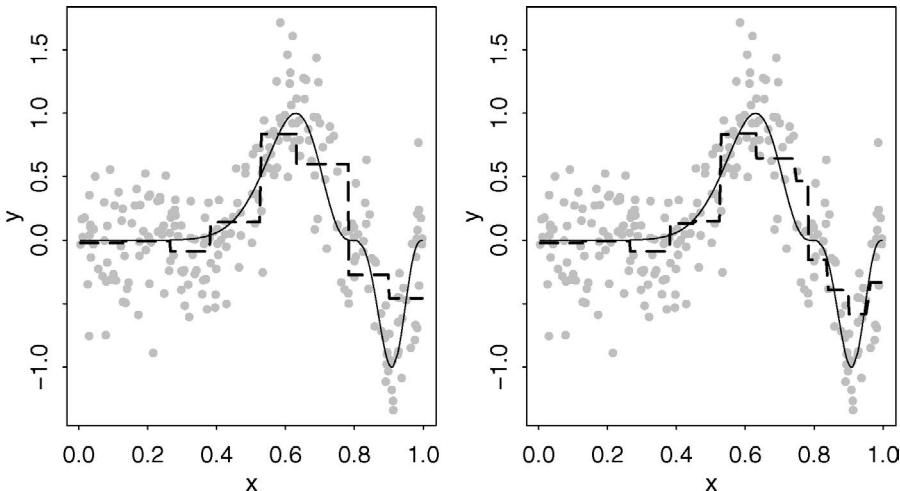


Figure 14.11 *Thresholding and inverting the transform. In the left panel all level-four and above coefficients are zeroed. In the right, the coefficients are thresholded using the default method. The true function is shown as a solid line and the estimate as a dashed line.*

We see that the fit consists of eight constant fits; we expect this since Haar basis is piecewise constant and we have thrown away the higher-order parts leaving just eight coefficients.

Instead of simply throwing away higher-order coefficients, we could zero out only the small coefficients. We choose the threshold using the default method:

```
wtd2 <- threshold(wds)
fd2 <- wr(wtd2)
```

Now we plot the result as seen in the second panel of Figure 14.11.

```
plot(y ~ x, exa, col=gray(0.75))
lines(m ~ x, exa)
lines(fd2 ~ x, exa, lty=5, lwd=2)
```

Again, we see a piecewise constant fit, but now the segments are of varying lengths. Where the function is relatively flat, we do not need the detail from the higher-order terms. Where the function is more variable, the finer detail is helpful.

We could view the thresholded coefficients as a compressed version of the original data (or signal). Some information has been lost in the compression, but the thresholding algorithm ensures that we tend to keep the detail we need, while throwing away noisier elements.

Even so, the fit is not particularly good because the fit is piecewise constant. We

would like to use continuous basis functions while retaining the orthogonality and multiresolution properties. Families of such functions were discovered and described in Daubechies (1991). We illustrate such a form in Figure 14.12 for our data:

```
wds <- wd(exa$y, filter.number=2, bc="interval")
draw(filter.number=2, family="DaubExPhase")
plot(wds)
```

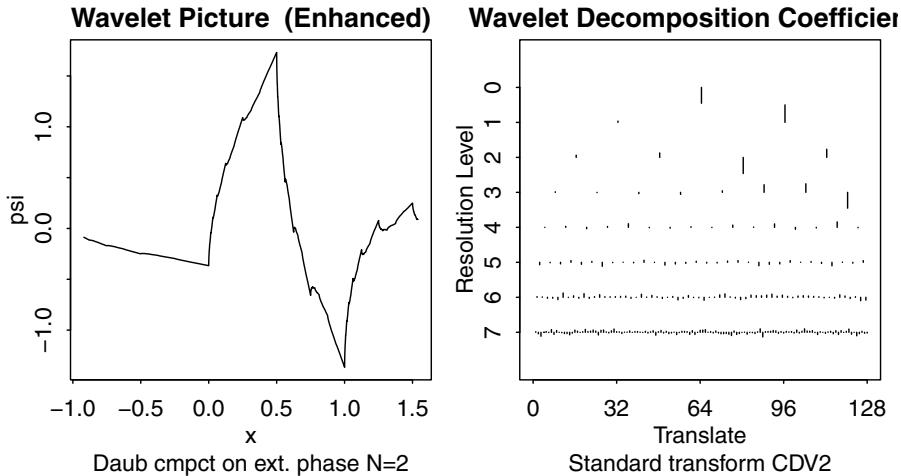


Figure 14.12 *Mother wavelet is shown in the left panel — the Daubechies orthonormal compactly supported wavelet N=2 from the extremal phase family. The right panel shows the wavelet coefficients.*

The mother wavelet takes an unusual form. The function is not explicitly defined, but is implicitly computed from the method for making the wavelet decomposition. Now we try the default thresholding and reconstruct the fit:

```
wtd <- threshold(wds)
fd <- wr(wtd)
plot(y ~ x, exa, col=gray(0.75))
lines(m ~ x, exa)
lines(fd ~ x, exa, lty=2)
```

We can see the fit in Figure 14.13. Although the fit follows the true function quite well, there is still some roughness.

The standard wavelet fitting function is designed only for evenly spaced  $x$  with a number of observations in some power of two. Fortunately, there are 256 observations in Example A, but the Old Faithful set is not so conveniently arranged. The wavethresh package implements a method described in Kovac and Silverman (2000) that handles irregularly spaced data. We illustrate how this can be applied here. The method works by interpolating an evenly spaced grid on  $[0, 1]$  using the next highest power of two beyond the size of the data — in this case 512. We need to rescale the predictor to  $[0, 1]$  first for this to work. We have taken the default options throughout except to say that the boundary condition,  $bc$ , does not assume a periodic function.

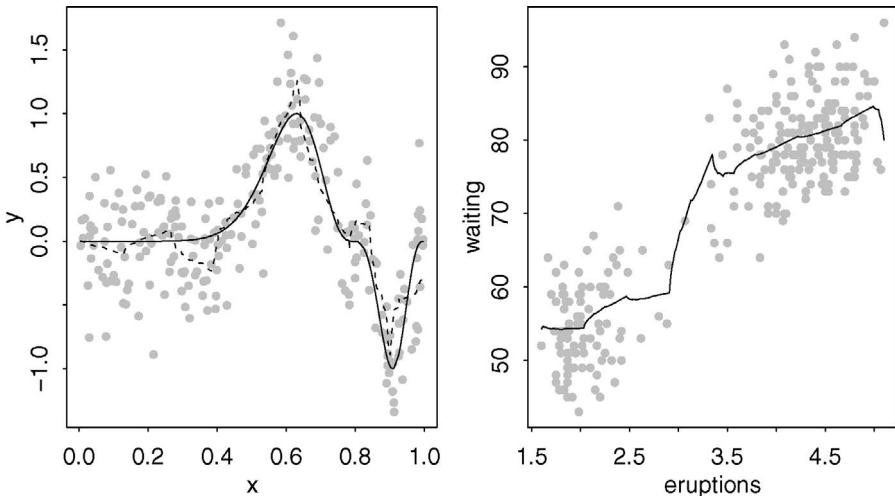


Figure 14.13 Daubechies wavelet  $N=2$  thresholded fit to the Example A data shown on the left. Irregular wavelet fit to the Old Faithful data is shown on the right.

```

x <- with(faithful, (eruptions-min(eruptions)) / (max(eruptions)-min(
  ↪ eruptions)))
gridof <- makegrid(x, faithful$waiting)
wdoft <- irregwd(gridof, bc="symmetric")
wtoft <- threshold(wdoft)
wroft <- wr(wtoft)
plot(waiting ~ eruptions, faithful, col=grey(0.75))
with(faithful, lines(seq(min(eruptions), max(eruptions), len=512), wroft))

```

The resulting plot, shown in the second panel of Figure 14.13, reveals a discontinuity in the fit not seen in previous plots. This demonstrates one of the main advantages of the wavelet method in handling and revealing discontinuities. Wavelet methods are particularly useful in processing very large datasets such as those found in image and sound files because the filtering method of thresholding coefficients can drastically reduce file size without losing much information.

## 14.6 Discussion of Methods

We have presented only a selection of the wide variety of methods available. For example, *nearest neighbor* methods adjust for varying density in the predictor space by adjusting window widths to be wider in sparser regions and narrower in denser regions. Window widths are also nonconstant in *variable bandwidth* methods. Such methods are particularly appropriate for functions like Example A where the smoothness of the function varies. We would like to use a smaller bandwidth in regions where the function changes rapidly but a wider one where it is more constant.

Bayesian methods of smoothing are evident in the *Gaussian Process* method as described in Rasmussen and Williams (2006). This method is particularly appropriate if you have prior knowledge and also works well on quite small datasets.

The construction of alternate smoothing methods has long been a popular topic of interest for statisticians and researchers in other fields. Because no definitive solution is possible, this has encouraged the development of a wide range of methods. But it is important not to allow the enthusiasm to solve an interesting technical problem to overshadow the purpose of a data analysis. We propose four possible objectives:

**Description** Sometimes we simply want to draw a line on a scatterplot to aid in the visual interpretation of the relationship displayed. We do not want to spend a lot of time or effort in constructing this smooth. We want a method that is simple and reliable.

**Auxiliary** In some applications, the smoothed fit is needed as part of some larger analysis. The smooth is not the principal aim. For example, the smooth might be used to impute some missing values. In such examples, we need to choose the smoothing method to optimize the wider objective. Sometimes, this means choosing rougher or smoother results than we would pick in a standalone problem.

**Prediction** Interpolating values in noisy data is one example where these methods could be useful. Extrapolation is more problematic as this requires some assumptions about how the function will behave outside the range of the data. Parametric methods do better here as, although extrapolation is inherently risky, it is more transparent how they will behave. We may also want to construct confidence statements which is easier to do with the basis function methods, such as splines, because we can mirror the parametric methods.

**Explanation** In linear modeling, the most common regression question concerns whether there is a relationship between  $x$  and  $y$ . No relationship corresponds to an assertion that the function is constant. We can construct confidence bands for some of the methods. We can then observe whether a constant function fits between the bands to decide the question. Even so, it would be better to directly construct a hypothesis test for which parametric methods are more amenable. Nonparametric methods do at least give us more information about situations where the assertion of no relationship only holds for part of the range of the predictor.

So when should we use nonparametric regression and which particular method should we choose? In the univariate case, we can describe three situations. When there is very little noise, interpolation (or at most, very mild smoothing) is the best way to recover the relation between  $x$  and  $y$ . Splines are good for this purpose. When there is a moderate amount of noise, nonparametric methods are most effective. There is enough noise to make smoothing worthwhile but also enough signal to justify a flexible fit. When the amount of noise becomes larger, parametric methods become relatively more attractive. There is insufficient signal to justify anything more than a simple model.

It is not reasonable to claim that any one smoother is better than the rest. The best choice of smoother will depend on the characteristics of the data and knowledge about the true underlying relationship. The choice will also depend on whether the fit is to be made automatically or with human intervention. When only a single dataset is being considered, it's simple enough to craft the fit and intervene if a particular method produces unreasonable results. If a large number of datasets are to be fit

automatically, then human intervention in each case may not be feasible. In such cases, a reliable and robust smoother may be needed.

We think the loess smoother makes a good all-purpose smoother. It is robust to outliers and yet can produce smooth fits. When you are confident that no outliers are present, smoothing splines is more efficient than local polynomials.

## 14.7 Multivariate Predictors

Given  $\mathbf{x}_1, \dots, \mathbf{x}_n$  where  $\mathbf{x} \in \mathbb{R}^p$ , we observe:

$$y_i = f(\mathbf{x}) + \varepsilon_i \quad i = 1, \dots, n$$

Many of the methods discussed previously extend naturally to higher dimensions, for example, the Nadaraya–Watson estimator becomes:

$$f_\lambda(\mathbf{x}) = \frac{\sum_{j=1}^n K\left(\frac{\mathbf{x}-\mathbf{x}_j}{\lambda}\right)Y_j}{\sum_{j=1}^n K\left(\frac{\mathbf{x}-\mathbf{x}_j}{\lambda}\right)}$$

where the kernel  $K$  is typically spherically symmetric, provided the dimensions of  $\mathbf{x}$  are scaled the same way. The spline idea can be used with the introduction of thin plate splines and local polynomials can also be naturally extended.

We can illustrate kernel smoothing in two dimensions. We see in Figure 14.14 that too little smoothing has been used in the first example while the choice for the second example appears about right.

```
data(savings, package="faraway")
y <- savings$sr
x <- cbind(savings$pop15, savings$ddpi)
sm.regression(x,y,h=c(1,1),xlab="pop15",ylab="growth",zlab="savings
  ↪ rate")
sm.regression(x,y,h=c(5,5),xlab="pop15",ylab="growth",zlab="savings
  ↪ rate")
```

We can produce a spline surface fit with the help of the `mgcv` package which is explained in detail in the next chapter:

```
library(mgcv)
amod <- gam(sr ~ s(pop15,ddpi), data=savings)
vis.gam(amod, col="gray", ticktype="detailed", theta=-35)
```

We have used the default amount of smoothing and rotated the view of the plot, as seen in the first panel of Figure 14.15, to make it similar to the kernel version. We achieve a smoother but similar fit to the kernel smooth. Two-dimensional smoothing is also possible using `loess`. We need to do more work to construct a 2D grid of predictor values on which we compute the prediction and make the perspective plot:

```
lomod <- loess(sr ~ pop15 + ddpi, data=savings)
xg <- seq(21,48,len=20)
yg <- seq(0,17,len=20)
zg <- expand.grid(pop15=xg,ddpi=yg)
persp(xg, yg, predict(lomod, zg), theta=-35, ticktype="detailed", xlab
  ↪ ="pop15", ylab="growth", zlab="savings rate", col="gray")
```

The fit, as seen in the second panel of Figure 14.15, looks similar to the previous smooths but notice the high predicted response for high growth, low `pop15` countries.

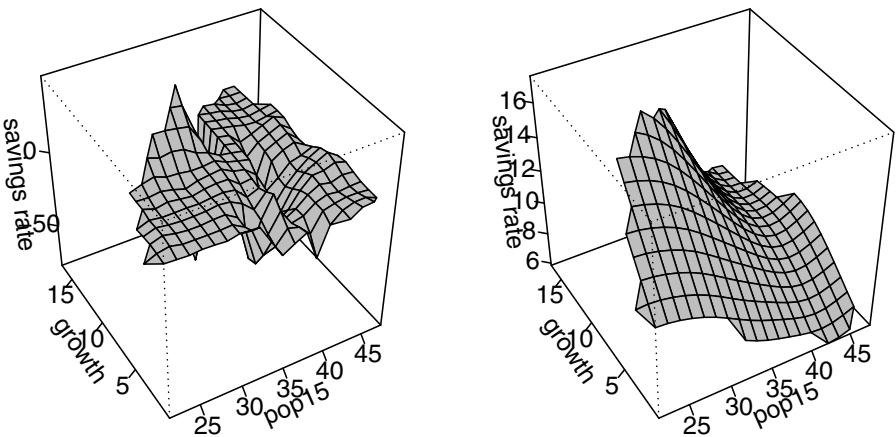


Figure 14.14 Smoothing savings rate as a function growth and population under 15. Plot on the left is too rough while that on the right seems about right.

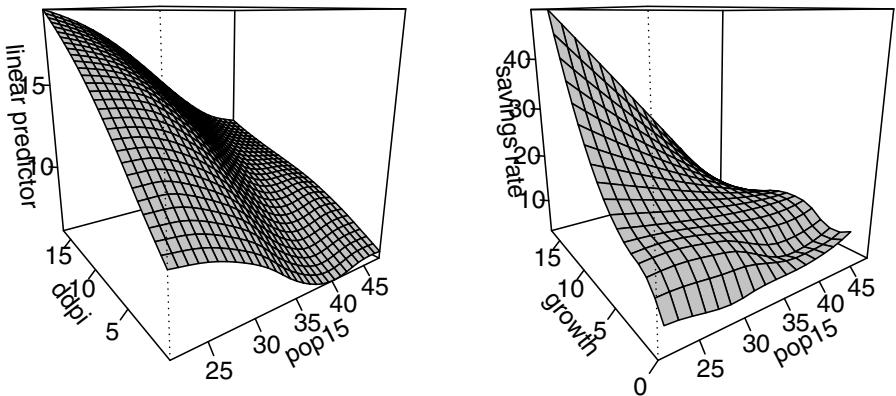


Figure 14.15 Smoothing spline fit to the savings data shown on the left. Loess smooth is shown on the right.

Checking the data, we find there are no countries in this region of the predictor space, meaning this represents an extrapolation away from the observed data. The `loess` method uses linear extrapolation, producing the observed result. The kernel-based method does not even attempt to predict outside the range whereas the spline method produced a more restrained prediction. It is difficult to say which is best as we must appeal to subject-matter knowledge to guide our choice.

Developing multivariate estimators is not so difficult but there are problems. Because nonparametric fits are quite complex, we need to visualize them to make sense of them and yet this cannot be done easily for more than two predictors. Most nonparametric regression methods rely on local smoothing; local averaging is the crudest example of this. However, to maintain a stable average we need sufficient points in

the window. For data in high dimensions, the window will need to be wide to capture sufficient points to average. You need an extremely large number of points to cover a high-dimensional space to high density. This is known as the “curse of dimensionality,” a term coined by Bellman (1961). In truth, it should not be called a curse, but rather a blessing, since information on additional variables should have some value, even if it is inconvenient. Our challenge is to make use of this information. Nonparametric regression fits are hard to interpret in higher dimensions where visualization is difficult. Simply extending the one-dimensional method is not effective.

The methods we describe in the following chapters impose additional restrictions on the fitted function to make the problem feasible and the results easier to interpret.

**Further Reading:** For a general review of smoothing methods, see Simonoff (1996). For books on specific methods of smoothing, see Loader (1999), Wahba (1990), Bowman and Azzalini (1997), Wand and Jones (1995) and Eubank (1988). The application of nonparametric regression to the goodness of fit problem may be found in Hart (1997).

## Exercises

1. The dataset `teengamb` concerns a study of teenage gambling in Britain. Take the variables `gamble` as the response and `income` as the predictor.
  - (a) Make a plot of the data.
  - (b) Fit a curve to the data using kernel smoothing with a cross-validated choice of `bandwidth`. Display the fit on the data. Does the fit look linear?
  - (c) Fit a curve using smoothing splines with the automatically chosen amount of smoothing. Display the fit and report the effective degrees of freedom. Display a fit with somewhat larger degrees of freedom. Was the automatic choice satisfactory?
  - (d) Use `loess` to fit a curve to the data with the default amount of smoothing. Display the fit.
  - (e) Produce and show a plot with a 95% confidence band for the fit. Is a linear fit plausible?
2. The dataset `uswages` is drawn as a sample from the Current Population Survey in 1988. Predict the `wage` from the years of education.
  - (a) Make a plot of the two variables of interest that makes some effort to avoid the problems of overplotting. Repeat the plot but use a log scale for the response.
  - (b) Compute the default smoothing spline fit and display on top of the data. Comment on the quality of the fit.
  - (c) Compute the default `lowess` fit and display on the fit. Does this method work better than smoothing splines in this instance?
  - (d) For each number of years of education, compute both the mean and the median wage. Construct a plot showing how these means and medians change with education. Which summary works better?

- (e) Instead of means and medians, compute the two quartiles and the median and display on top of the data. (This is a form of quantile regression).
  - (f) Display the lowess fit on the log-transformed data. Do you think it is better to work on the log scale for this data?
3. The dataset `prostate` is from a study of 97 men with prostate cancer who were due to receive a radical prostatectomy. Predict the `lweight` using the `age`.
- (a) Plot the data and comment on the relationship.
  - (b) Fit a curve using kernel methods, plotting the fit on top of the data. What is the effect of the outlier?
  - (c) Compute the smoothing spline fit with the default amount of smoothing. What type of curve has been fit to the data?
  - (d) Fit a loess curve with a 95% confidence band. Do you think a linear fit is plausible for this data?
  - (e) Display all three previous fits on top of the same display and compare.
  - (f) Introduce `lpsa` as a second predictor and show the bivariate fit to the data using smoothing splines.
4. The dataset `divusa` contains data on divorces in the United States from 1920 to 1996. We focus on the `military` variable as it changes by `year`. There really were more military personnel during World War II, so these points are not outliers.
- (a) Plot the number of military personnel and identify the three peaks in the dataset.
  - (b) Fit the data using smoothing splines with the default amount of smoothing, commenting on the quality of the fit. Repeat with lowess, again using the default amount of smoothing.
  - (c) Discuss the utility of smoothing in understanding this data.
5. The `aatemp` data comes from the U.S. Historical Climatology network. They are the annual mean temperatures (in degrees Fahrenheit) in Ann Arbor, Michigan, going back about 150 years.
- (a) Plot the temperature as a function of time and comment on the underlying trend.
  - (b) Fit a least squares line to the data and test whether the slope of the line is different from zero. What is the main drawback of this modeling approach?
  - (c) Fit a Lowess curve to the data using the default amount of smoothing. Display the fit along with a 95% confidence band. What does this say about the underlying trend in the relationship?
  - (d) Fit a regression spline basis to the data with 12 knots. Display the fit on the data.
  - (e) Compare this model to the linear fit using an  $F$ -test. Which model is preferred? What more needs to be explored with spline fit before drawing conclusions?

6. Generate simulated data according to the following model. Use 256 evenly spaced points on  $[0, 1]$  for  $x$ . Let  $y = f(x) + \varepsilon$  with

$$f(x) = \begin{cases} x & x \leq 1/2 \\ x - 1 & x > 1/2 \end{cases}$$

and  $\varepsilon \sim N(0, (0.1)^2)$ .

- (a) Plot the data along with the true function  $f$ .
- (b) Display the Lowess fit on top of the data using the default choice of smoothing.  
Comment on the quality of the fit. Repeat using a smaller span in the smooth.
- (c) Use the default method of wavelet smoothing to compute a fit to the data.  
Comment on the quality of the fit.
- (d) In the wavelet thresholding, replace the policy with universal. Does this improve the fit?

**This page intentionally left blank**

---

## Chapter 15

---

# Additive Models

---

Suppose we have a response  $y$  and predictors  $x_1, \dots, x_p$ . A linear model takes the form:

$$y = \beta_0 + \sum_{j=1}^p \beta_j X_j + \varepsilon$$

We can include transformations and combinations of the predictors among the  $xs$ , so this model can be very flexible. However, it can often be difficult to find a good model, given the wide choice of transformations available. We can try a systematic approach of fitting a family of transformations. For example, we can try polynomials of the predictors, but particularly if we include interactions, the number of terms becomes very large, perhaps greater than the sample size. Alternatively, we can use more interactive and graphical approaches that reward intuition over brute force. However, this requires some skill and effort on the part of the analyst. It is easy to miss important structure; a particular difficulty is that the methods only consider one variable at a time, when the secret to finding good transformations may require that variables be considered simultaneously.

We might try a nonparametric approach by fitting:

$$y = f(x_1, \dots, x_p) + \varepsilon$$

This avoids the necessity of parametric assumptions about the form of the function  $f$ , but for  $p$  bigger than two or three, it is simply impractical to fit such models due to large sample size requirements, as discussed at the end of the previous chapter.

A good compromise between these extremes is the *additive model*:

$$y = \beta_0 + \sum_{j=1}^p f_j(X_j) + \varepsilon$$

where the  $f_j$  are smooth arbitrary functions. Additive models were introduced by Stone (1985).

Additive models are more flexible than the linear model, but still interpretable since the functions  $f_j$  can be plotted to give a sense of the marginal relationship between the predictor and the response. Of course, many linear models discovered during a data analysis take an additive form where the transformations are determined in an ad hoc manner by the analyst. The advantage of the additive model approach is that the best transformations are determined simultaneously and without parametric assumptions regarding their form.

In its basic form, the additive model will do poorly when strong interactions exist. In this case we might consider adding terms like  $f_{ij}(x_i x_j)$  or even  $f_{ij}(x_i, x_j)$  if there is sufficient data. Categorical variables can be easily accommodated within the model using the usual regression approach. For example:

$$y = \beta_0 + \sum_{j=1}^p f_j(X_j) + Z\gamma + \varepsilon$$

where  $Z$  is the design matrix for the variables that will not be modeled additively, where some may be quantitative and others qualitative. The  $\gamma$  are the associated regression parameters. We can also have an interaction between a factor and a continuous predictor by fitting a different function for each level of that factor. For example, we might have  $f_{male}$  and  $f_{female}$ .

There are several different ways of fitting additive models in R. The `gam` package originates from the work of Hastie and Tibshirani (1990). The `mgcv` package is part of the recommended suite that comes with the default installation of R and is based on methods described in Wood (2000). The `gam` package allows more choice in the smoothers used while the `mgcv` package has an automatic choice in the amount of smoothing as well as wider functionality. The `gss` package of Gu (2002) takes a spline-based approach.

The fitting algorithm depends on the package used. The *backfitting* algorithm is used in the `gam` package. It works as follows:

1. We initialize by setting  $\beta_0 = \bar{y}$  and  $f_j(x) = \hat{\beta}_j x$  where  $\hat{\beta}$  is some initial estimate, such as the least squares, for  $j = 1, \dots, p$ .
2. We cycle  $j = 1, \dots, p, 1, \dots, p, 1, \dots$

$$f_j = S(x_j, y - \beta_0 - \sum_{i \neq j} f_i(X_i))$$

where  $S(x, y)$  means the smooth on the data  $(x, y)$ . The choice of  $S$  is left open to the user. It could be a nonparametric smoother like splines or loess, or it could be a parametric fit, say linear or polynomial. We can even use different smoothers on different predictors with differing amounts of smoothing.

The algorithm is iterated until convergence. Hastie and Tibshirani (1990) show that convergence is assured under some rather loose conditions. The term  $y - \beta_0 - \sum_{i \neq j} f_i(x_i)$  is a partial residual — the result of fitting everything except  $x_j$ , making the connection to linear model diagnostics.

The `mgcv` package employs a penalized smoothing spline approach. Suppose we represent  $f_j(x) = \sum_i \beta_i \phi_i(x)$  for a family of spline basis functions,  $\phi_i$ . We impose a penalty  $\int [f_j''(x)]^2 dx$  which can be written in the form  $\beta_j^T S_j \beta_j$ , for a suitable matrix  $S_j$  that depends on the choice of basis. We then maximize:

$$\log L(\beta) - \sum_j \lambda_j \beta_j^T S_j \beta_j$$

where  $L(\beta)$  is likelihood with respect to  $\beta$  and the  $\lambda_j$ s control the amount of smoothing for each variable. Generalized cross-validation (GCV) is used to select the  $\lambda_j$ s.

### 15.1 Modeling Ozone Concentration

We illustrate the methodology in this chapter and some of the following chapters using some data on air pollution from the Los Angeles area in 1976. The response is  $O_3$ , the atmospheric ozone concentration on a particular day. To simplify matters, we will initially reduce the predictors to just three: temperature measured at El Monte, `temp`, inversion base height at LAX, `ibh`, and inversion top temperature at LAX, `ibt`. A number of cases with missing variables have been removed for simplicity. The data were first presented by Breiman and Friedman (1985). We might start with a look at how the response is related to each of the predictors as seen in Figure 15.1. The loess fitted line and confidence band are added using the `geom_smooth()` command. We can see a somewhat nonlinear relationship in all three cases.

```
data(ozone, package="faraway")
ggplot(ozone, aes(x=temp, y=O3)) + geom_point(size=1) + geom_smooth()
ggplot(ozone, aes(x=ibh, y=O3)) + geom_point(size=1) + geom_smooth() +
  ← theme(axis.text.x = element_text(angle = 90))
ggplot(ozone, aes(x=ibt, y=O3)) + geom_point(size=1) + geom_smooth()
```

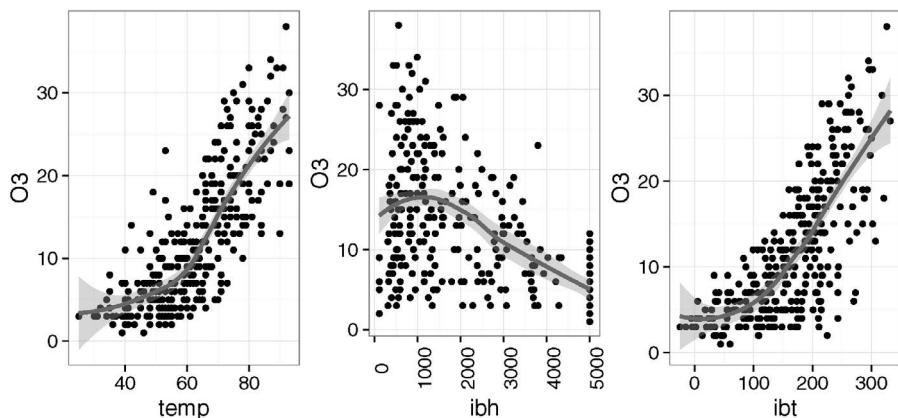


Figure 15.1 *Ozone concentration and three predictors. Loess fitted line and confidence band are shown.*

These plots show only the raw relationship between the predictors and response without allowance for the effect of other predictors. We can see that, for example,  $O_3$  increases with `ibt` but would this remain true if we adjusted for the effect of `temp` and `ibt`? A first attempt at doing this suggests a linear model, useful at least for reference purposes:

```
olm <- lm(O3 ~ temp + ibh + ibt, ozone)
summary(olm)
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-7.727982	1.621662	-4.77	2.8e-06
temp	0.380441	0.040158	9.47	< 2e-16
ibh	-0.001186	0.000257	-4.62	5.5e-06
ibt	-0.005821	0.010179	-0.57	0.57

`n = 330, p = 4, Residual SE = 4.748, R-Squared = 0.65`

Note that `ibt` is not significant in this model. Although we can determine the individual relationships between the predictors and the response by examining the regression coefficients, it can be helpful to view these effects graphically. One can use the `termplot` function but the `effects` package due to Fox (2003) has more comprehensive capabilities for achieving this functionality. We fix the other predictors at some typical value (the mean by default) and allow only the chosen predictor to vary. We plot the resulting prediction. Plots of these predictions are shown in Figure 15.2 for each of the predictors. We have added the partial residuals to the display. The dashed line represents a smooth fit to these residuals and gives a suggestion of the extent of nonlinearity.

```
library(effects)
plot(Effect("temp", olm, partial.residuals=TRUE))
plot(Effect("ibh", olm, partial.residuals=TRUE))
plot(Effect("ibt", olm, partial.residuals=TRUE))
```

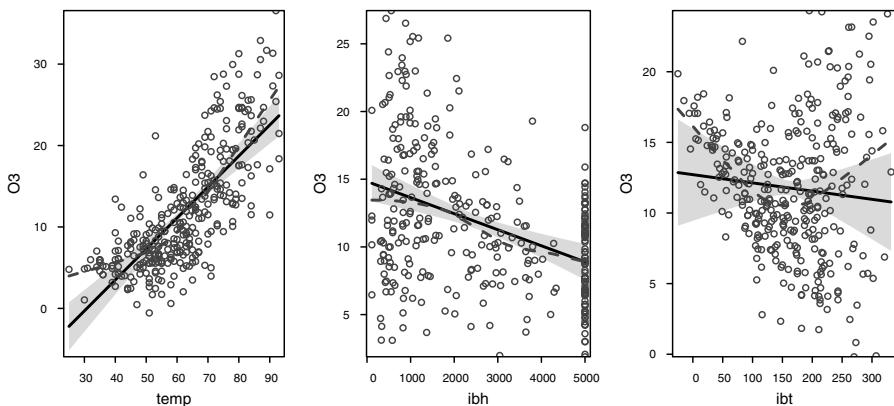


Figure 15.2 Effects of predictors in a linear model for Ozone concentration.

These plots show us how much the predicted response will vary over the observed range of the predictor and so are helpful in assessing the magnitude and direction of the effect. We can see from these plots that `ibt` does not have much effect after adjusting for the other two predictors, although there is some suggestion of a quadratic effect.

## 15.2 Additive Models Using mgcv

A method of fitting additive models is provided by the `mgcv` package of Wood (2006). Splines are the default choice of smoother with the appropriate amount of smoothing chosen automatically, unless we intervene.

```
library(mgcv)
ammgcv <- gam(O3 ~ s(temp)+s(ibh)+s(ibt), data=ozone)
summary(ammgcv)
```

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	11.776	0.238	49.4	<2e-16

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value
s(temp)	3.39	4.26	20.55	7.7e-16
s(ibh)	4.17	5.08	7.34	1.4e-06
s(ibt)	2.11	2.73	1.61	0.19

R-sq. (adj) = 0.708 Deviance explained = 71.7%  
 GCV = 19.346 Scale est. = 18.72 n = 330

The intercept is the only parametric coefficient in this model because all the predictor terms have smooths. We can compute the equivalent degrees of freedom by an analogy to linear models. For linear smoothers, the relationship between the observed and fitted values may be written as  $\hat{y} = Py$ . The trace of  $P$  then estimates the effective number of parameters. For example, in linear regression, the projection matrix is  $X(X^T X)^{-1}X^T$  whose trace is equal to the rank of  $X$  or the number of identifiable parameters. This notion can be used to obtain the degrees of freedom for additive models. The column marked Ref.df is a modified computation of the degrees of freedom which is more appropriate for use in test statistics.

Since we have sums of squares and degrees of freedom, we can compute  $F$ -statistics in the same way as linear models. However, the  $F$ -statistics quoted in the summary output have been modified to produce somewhat better statistical properties. The  $p$ -values are computed from these  $F$ -statistics and degrees of freedom although we cannot claim the null distributions are exactly  $F$ -distributed. Usually, they are good approximations. We see that the  $R^2$ , which in this case is called the “Deviance explained,” is somewhat higher than in the lm fit.

We can also examine the transformations used:

```
plot(ammgcv, residuals=TRUE, select=1)
plot(ammgcv, residuals=TRUE, select=2)
plot(ammgcv, residuals=TRUE, select=3)
```

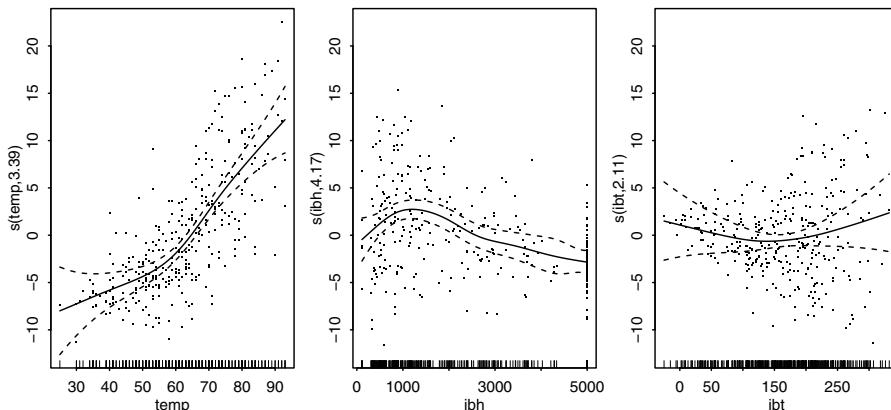


Figure 15.3 Transformation functions for the model fit by mgcv. Note how the same scale has been deliberately used on all three plots. This allows us to easily compare the relative contribution of each variable.

We see that the transformations, as seen in Figure 15.3, are nonlinear. The chosen transformation for temperature is roughly piecewise linear with a change in the slope around 60. We see a similar piecewise linear relationship with a change at about 1000 for ibh. A deeper understanding of the application might lead to some interpretation of these changepoints. We have asked for the residuals to be shown on the plot, in keeping with our previous plots, as this adds to our ability to judge the choice of smoothing. We can see the degree of smoothing appears appropriate as there is no obvious finer structure. Note that the label on the vertical axis gives the degrees of freedom for the smooth which will correspond to the shape of the fit. Something close to linear will have degrees of freedom close to one.

The confidence bands are helpful in judging the significance of features seen in the plots. We might judge whether a straight line could be drawn entirely between the two bands. This might suggest that a linear term would be sufficient. In the case of ibt, we see that a horizontal line at zero would fit between the bands. This suggests this predictor has no marginal effect on the response.

We might also be interested in whether there really is a change in the trend for temperature. We test this supposition with a linear term in temperature and then make the *F*-test:

```
am1 <- gam(O3 ~ s(temp)+s(ibh), data=ozone)
am2 <- gam(O3 ~ temp+s(ibh), data=ozone)
anova(am2, am1, test="F")
```

Analysis of Deviance Table

	Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
1	324	6950				
2	321	6054	2.7	896	17.6	7.9e-10

The *p*-value is only approximate, but we can be confident there is a change in the trend for temperature.

You can include functions of two variables with mgcv. This remains within the framework of the additive model. For example, suppose we suspect that there is an interaction between temperature and IBH. We must decide on the appropriate bivariate smoothing method. In some cases, the two variables are measured in the same units. For example, one might measure location on the north-south axis and another the location on the east-west axis. In this *isotropic* case, it is appropriate to use the same degree of smoothing in both variables. But in our example, temperature and inverse base height (IBH) are measured on different scales and are thus *anisotropic*. In this case, we will likely need a different amount of smoothing the two directions.

This can be achieved using *tensor product* smooth:

```
amint <- gam(O3 ~ te(temp,ibh)+s(ibt), data=ozone)
summary(amint)
```

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	11.776	0.239	49.4	<2e-16

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value
te(temp,ibh)	10.49	13.08	12.96	<2e-16

```
s(ibt)      1.04  1.08  0.47  0.51
```

```
R-sq.(adj) = 0.707  Deviance explained = 71.8%
GCV = 19.526  Scale est. = 18.784  n = 330
```

We see that the combined temperature and IBH term are clearly significant with the ibt term remaining redundant. This output does not tell us whether the interaction between the two terms adds anything to the model although the very small increase in the deviance explained suggests that it does not.

We can test the significance of the interaction:

```
anova(ammgcv, amint, test="F")
```

Analysis of Deviance Table

	Model 1: 03 ~ s(temp) + s(ibh) + s(ibt)	Model 2: 03 ~ te(temp, ibh) + s(ibt)				
	Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
1	319	5978				
2	317	5963	1.86	14.6	0.42	0.64

The large *p*-value confirms the insignificance of the interaction term. We can be satisfied with just the two univariate additive terms. In more interesting cases, we would be eager to visualize the interaction as seen in two different ways in Figure 15.4:

```
plot(amint, select=1)
vis.gam(amint, theta=-45, color="gray")
```

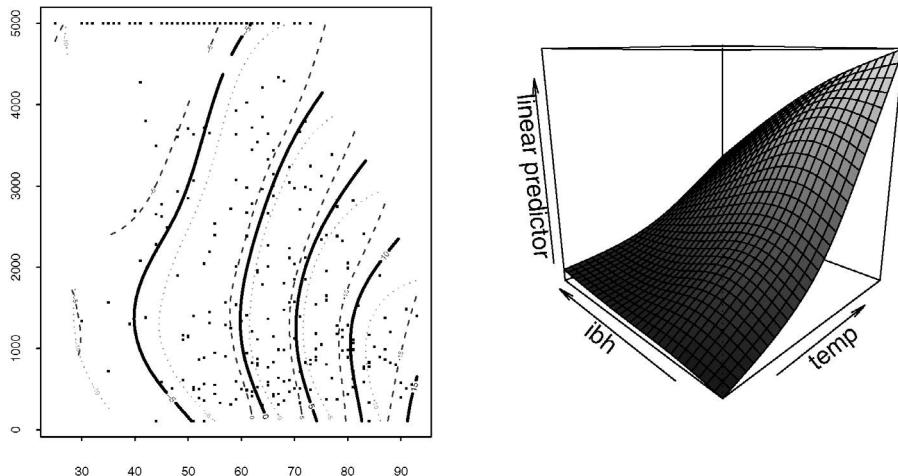


Figure 15.4 The bivariate contour plot for temperature and *ibh* is shown in the left panel. The right panel shows a perspective view of the information on the left panel.

The contours appear almost parallel. The perspective view also suggests no interaction.

One use for additive models is as an exploratory tool for standard parametric regression modeling. We can use the fitted functions to help us find suitable simple transformations of the predictors. One idea here is to model the *temp* and *ibh* ef-

fects using piecewise linear regression (also known as “broken stick” or segmented regression). We define the right and left “hockey-stick” functions:

```
rhs <- function(x,c) ifelse(x > c, x-c, 0)
lhs <- function(x,c) ifelse(x < c, c-x, 0)
```

and now fit a parametric model using cutpoints of 60 and 1000 for `temp` and `ibh`, respectively. We pick the cutpoints using the plots:

```
olm2 <- lm(O3 ~ rhs(temp, 60)+lhs(temp, 60)+rhs(ibh, 1000)+lhs(ibh, 1000),
            ↪ ozone)
summary(olm2)

Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.603832  0.622651 18.64 < 2e-16
rhs(temp, 60) 0.536441  0.033185 16.17 < 2e-16
lhs(temp, 60) -0.116173  0.037866 -3.07 0.0023
rhs(ibh, 1000) -0.0001486 0.000198 -7.49 6.7e-13
lhs(ibh, 1000) -0.0003554 0.001314 -2.71 0.0072
```

`n = 330, p = 5, Residual SE = 4.342, R-Squared = 0.71`

Compare this model to the first linear model we fit to this data. The fit is better and about as good as the additive model fit. It is unlikely we could have discovered these transformations without the help of the intermediate additive models. Furthermore, the linear model has the advantage that we can write the prediction formula in a compact form.

We can use additive models for building a linear model as above, but they can be used for inference in their own right. For example, we can predict new values with standard error:

```
predict(ammgcv, data.frame(temp=60, ibh=2000, ibt=100), se=T)
$fit
[1] 11.013
```

```
$se.fit
[1] 0.97278
```

If we try to make predictions for predictor values outside the original range of the data, we will need to linearly extrapolate the spline fits. This is dangerous for all the usual reasons:

```
predict(ammgcv, data.frame(temp=120, ibh=2000, ibt=100), se=T)
$fit
[1] 35.511
```

```
$se.fit
[1] 5.7261
```

We see that the standard error is much larger although this likely does not fully reflect the uncertainty.

We should also check the usual diagnostics:

```
plot(residuals(ammgcv) ~ predict(ammgcv), xlab="Predicted", ylab=
    ↪ "Residuals")
abline(h=0)
qqnorm(residuals(ammgcv), main="")
qqline(residuals(ammgcv))
```

We can see in Figure 15.5 there is some nonconstant variance. There are also somewhat long tails for the residuals.

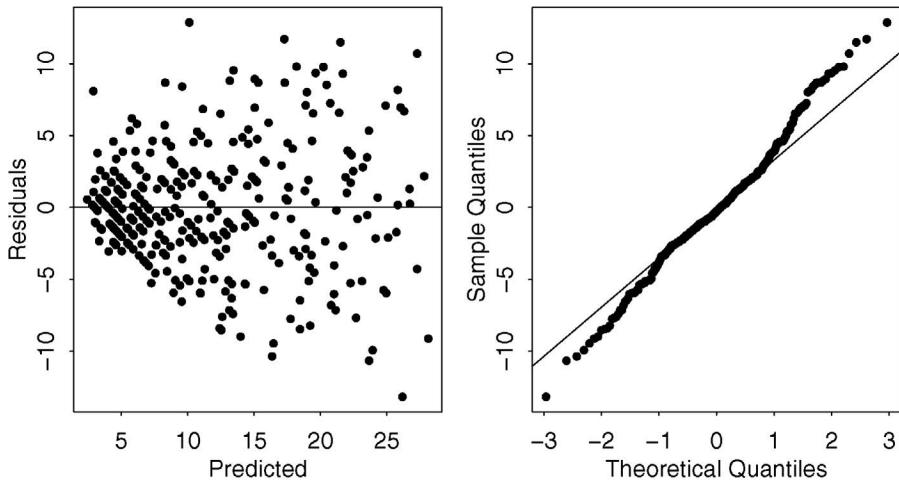


Figure 15.5 Residual plots for the additive model.

Now let's see the model for the full dataset. We found that the `ibh` and `ibt` terms were insignificant and so we removed them:

```
amred <- gam(O3 ~ s(vh)+s(wind)+s(humidity)+s(temp)+s(dpg)+ s(vis)+s(
  ↪ doy), data=ozone)
summary(amred)
```

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	11.776	0.201	58.7	<2e-16

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value
s(vh)	1.00	1.00	20.50	8.4e-06
s(wind)	1.00	1.00	6.56	0.01091
s(humidity)	1.00	1.00	14.61	0.00016
s(temp)	5.77	6.92	12.66	2.1e-14
s(dpg)	3.31	4.20	14.11	6.4e-11
s(vis)	2.22	2.77	7.20	0.00021
s(doy)	4.07	5.13	20.80	< 2e-16

R-sq.(adj) = 0.793 Deviance explained = 80.5%  
GCV = 14.113 Scale est. = 13.285 n = 330

We will compare this to the results of different modeling approaches that we will present later. We can see that we achieve a good fit with an  $R^2$  of 80.5%, but at the cost of using effectively 19.4 (sum the df) parameters including the intercept.

Also for future reference, here is the linear model with all insignificant terms removed:

```
alm <- lm(O3 ~ vis+doy+ibt+humidity+temp,data=ozone)
summary(alm)
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-10.01786	1.65306	-6.06	3.8e-09
vis	-0.00820	0.00369	-2.22	0.027

doy	-0.01020	0.00245	-4.17	3.9e-05
ibt	0.03491	0.00671	5.21	3.4e-07
humidity	0.08510	0.01435	5.93	7.7e-09
temp	0.23281	0.03607	6.45	4.0e-10

$n = 330$ ,  $p = 6$ , Residual SE = 4.430, R-Squared = 0.7

We can see that the fit is substantially worse, but uses only six parameters. Of course, we may be able to improve this fit with some manual data analysis. We could look for good transformations and check for outliers and influential points. However, since we want to compare different modeling techniques, we want to avoid making subjective interventions for the sake of a fair comparison.

### 15.3 Generalized Additive Models

In generalized linear models:

$$\eta = X\beta \quad EY = \mu \quad g(\mu) = \eta \quad Var(Y) \propto V(\mu)$$

The approach is readily extended to additive models to form generalized additive models (GAM). We replace the linear predictor with

$$\eta = \beta_0 + \sum_{j=1}^p f_j(X_j)$$

In the `mgcv` package, the  $f_j$  are represented by splines. These splines have coefficients that are just more parameters that can be estimated using the likelihood approach.

The ozone data has a response with relatively small integer values. Furthermore, the diagnostic plot in Figure 15.5 shows nonconstant variance. This suggests that a Poisson response might be suitable. We fit this using:

```
gammagcv <- gam(O3 ~ s(temp)+s(ibh)+s(ibt), family=poisson, scale=-1,
                   data=ozone)
```

```
summary(gammagcv)
```

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.293	0.023	99.5	<2e-16

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value
s(temp)	3.82	4.74	16.82	4.2e-14
s(ibh)	3.74	4.57	10.54	1.2e-08
s(ibt)	1.35	1.62	0.57	0.53

R-sq.(adj) = 0.712 Deviance explained = 72.9%  
 GCV = 1.5062 Scale est. = 1.4573 n = 330

We have set `scale=-1` because negative values for this parameter indicate that the dispersion should be estimated rather than fixed at one. Since we do not truly believe the response is Poisson, it seems wise to allow for overdispersion. The default of not specifying `scale` would fix the dispersion at one. We see that the estimated dispersion is indeed somewhat bigger than one. We see that IBT is not significant. We can check the transformations on the predictors as seen in Figure 15.6:

```
plot(gammgcv, residuals=TRUE, select=1)
plot(gammgcv, residuals=TRUE, select=2)
plot(gammgcv, residuals=TRUE, select=3)
```

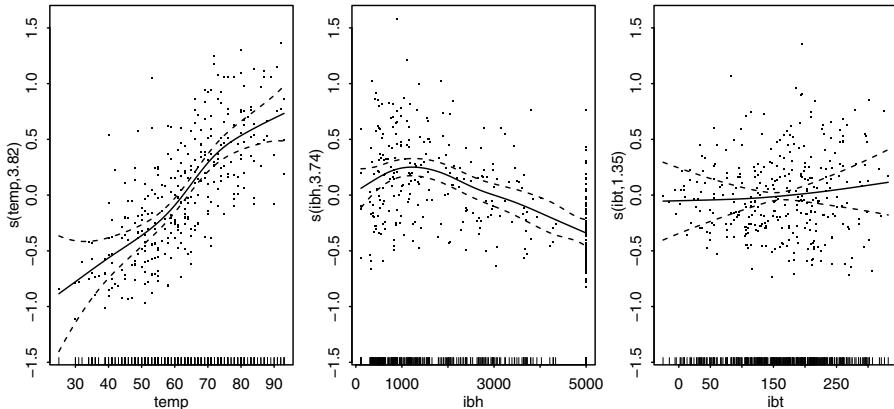


Figure 15.6 Transformation on the predictors for the Poisson GAM.

We see that the selected transformations are quite similar to those observed previously.

There are natural extensions to additive modeling for all the response types we have considered earlier in this book.

## 15.4 Alternating Conditional Expectations

In the additive model:

$$y = \alpha + \sum_{j=1}^p f_j(X_j) + \varepsilon$$

but in the transform-both-sides (TBS) model:

$$\theta(y) = \alpha + \sum_{j=1}^p f_j(X_j) + \varepsilon$$

For example,  $y = e^{x_1 + \sqrt{x_2}}$  cannot be modeled well by additive models, but can if we transform both sides:  $\log y = x_1 + \sqrt{x_2}$ . This fits within the TBS model framework. A more complicated alternative approach would be nonlinear regression. One particular way of fitting TBS models is *alternating conditional expectation* (ACE) which is designed to minimize  $\sum_i (\theta(y_i) - \sum f_j(x_{ij}))^2$ . Distractingly, this can be trivially minimized by setting  $\theta = f_j = 0$  for all  $j$ . To avoid this solution, we impose the restriction that the variance of  $\theta(y)$  be one. The fitting proceeds using the following algorithm:

1. Initialize:

$$\theta(y) = \frac{y - \bar{y}}{SD(y)} \quad f_j = \hat{\beta}_j x_j \quad j = 1, \dots, p$$

## 2. Cycle:

$$\begin{aligned} f_j &= S(x_j, \theta(y) - \sum_{i \neq j} f_i(x_i)) \\ \theta &= S(y, \sum_j f_j(x_j)) \end{aligned}$$

where  $S(x, y)$  represents a smoother returning a function on  $(x, y)$  data. Renormalize at the end of each cycle:

$$\theta(y) \leftarrow \frac{\theta(y) - \overline{\theta(y)}}{SD(\overline{\theta(y)})}$$

We repeat until convergence. ACE is comparable to the additive model, except now we allow transformation of the response as well. In principle, you can use any reasonable smoother  $S$ , but the original smoother used was the supersmooth. This cannot be easily changed in the R software implementation.

For our example, we start with the same three predictors in the ozone data:

```
x <- ozone[, c("temp", "ibh", "ibt")]
library(acepack)
acefit <- ace(x, ozone$O3)
```

Note that the `ace` function interface is quite rudimentary as we must give it the  $X$  matrix explicitly. The function returns the component `ty` which contains  $\theta(y)$  and `tx` which is a matrix whose columns contain the  $f_j(x_j)$ . We can get a sense of how well these transformations work by fitting a linear model that uses the transformed variables. We know that the intercept is zero so we exclude it to get cleaner output:

```
summary(lm(ty ~ tx-1, acefit))
Estimate Std. Error t value Pr(>|t|)
txtemp   0.9676    0.0508   19.04   <2e-16
txibh    1.1801    0.1358    8.69   <2e-16
txibt    1.3712    0.5116    2.68    0.0077
```

$n = 330$ ,  $p = 3$ , Residual SE = 0.526, R-Squared = 0.73

All three transformed predictors are strongly significant and the fit is superior to the original model. The  $R^2$  for the comparable additive model was 0.703. So the additional transformation of the response did improve the fit. Now we examine the transforms on the response and the three predictors:

```
plot(ozone$O3, acefit$ty, xlab="O3", ylab=expression(theta(O3)))
plot(x[,1], acefit$tx[,1], xlab="temp", ylab="f(temp)")
plot(x[,2], acefit$tx[,2], xlab="ibh", ylab="f(ibh)")
plot(x[,3], acefit$tx[,3], xlab="ibt", ylab="f(ibt)")
```

See Figure 15.7. The transform on the response is close to, but not quite, linear. The transformations on `temp` and `ibh` are similar to those found by the additive model. The transformation for `ibt` looks implausibly rough in some parts.

Now let's see how we do on the full data:

```
x <- ozone[,-1]
acefit <- ace(x, ozone$O3)
summary(lm(acefit$ty ~ acefit$tx-1))
Estimate Std. Error t value Pr(>|t|)
acefit$txvh      1.172     0.385    3.05   0.0025
```

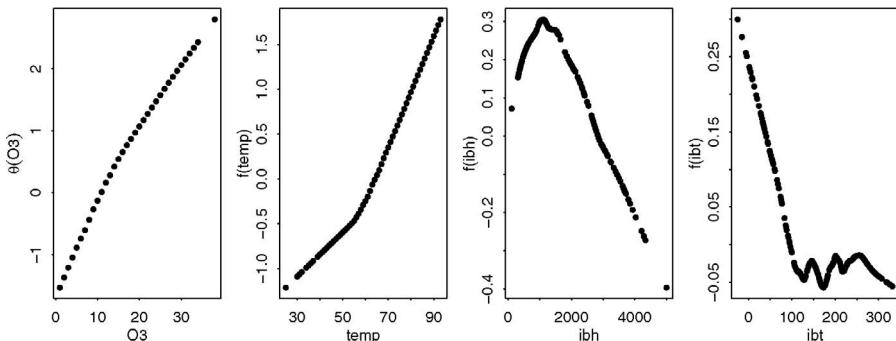


Figure 15.7 ACE transformations: the first panel shows the transformation on the response while the remaining three show the transformations on the predictors.

acefit\$txwind	1.074	0.404	2.66	0.0083
acefit\$txhumidity	0.651	0.245	2.66	0.0083
acefit\$txtemp	0.916	0.123	7.43	1.0e-12
acefit\$txibh	1.351	0.436	3.10	0.0021
acefit\$txdpq	1.322	0.167	7.92	4.0e-14
acefit\$txibt	0.926	0.196	4.71	3.7e-06
acefit\$txvis	1.386	0.230	6.03	4.5e-09
acefit\$txdoy	1.284	0.110	11.72	< 2e-16

n = 330, p = 9, Residual SE = 0.408, R-Squared = 0.84

A very good fit, but we must be cautious. Notice that all the predictors are strongly significant. This might be a reflection of reality or it could just be that the ACE model is overfitting the data by using implausible transformations as seen on the ibt variable above. Another problem is in constructing new predictions from fresh inputs. The nature of the smoothing used does not make it obvious how this can be achieved.

ACE can be useful in searching for good transformations while building a linear model. We might examine the fitted transformations as seen in Figure 15.7 to suggest appropriate parametric forms. More caution is necessary if the model is to be used in its own right, because of the tendency to overfit.

An alternative view of ACE is to consider the problem of choosing  $\theta$  and  $f_j$ 's such that  $\theta(Y)$  and  $\sum_j f_j(X_j)$  are maximally correlated. ACE solves this problem. For this reason, ACE can be viewed as a correlation method rather than a regression method.

The canonical correlation method is an ancestor to ACE. Given two sets of random variables  $X_1, \dots, X_m$  and  $Y_1, \dots, Y_n$ , we find unit vectors  $a$  and  $b$  such that:

$$\text{corr}(a^T X, b^T Y)$$

is maximized. One generalization of canonical correlation is to allow some of the  $X$ 's and  $Y$ 's to be power transforms of the original variables; this results in a parametric form of ACE. For example:

```

y <- cbind(ozone$O3, ozone$O3^2, sqrt(ozone$O3))
x <- ozone[, c("temp", "ibh", "ibt")]
cancor(x, y)
$cor
[1] 0.832346 0.217517 0.016908

$xcoef
      [,1]          [,2]          [,3]
temp -3.4951e-03 3.6335e-03 -6.7913e-03
ibh   1.3667e-05 -5.2054e-05 -5.2243e-06
ibt   1.6744e-04 -1.7384e-03 1.2436e-03

$ycoef
      [,1]          [,2]          [,3]
[1,] -0.00390830 -0.00539076 -0.1802230
[2,]  0.00009253 -0.00044172  0.0022167
[3,] -0.03928664  0.12068982  0.7948130

```

We see that it is possible to obtain a correlation of 0.832 by taking particular linear combinations of  $O3$ ,  $O3^2$  and  $\sqrt{O3}$  with the three predictors. The other two orthogonal combinations are not of interest to us here. Remember that  $R^2$  is the correlation squared in a simple linear model and  $0.832^2 = 0.692$ , so this is not a particularly competitive fit.

There are some oddities about ACE. For a single predictor, ACE is symmetric in  $X$  and  $Y$ , which is not the usual situation in regression. Furthermore, ACE does not necessarily reproduce the true model. Consider the population form of the problem and take  $Y = X + \varepsilon$  and  $\varepsilon \sim N(0, 1)$  and  $X \sim U(0, 1)$ , then  $E(Y|X) = X$  but  $E(X|Y) \neq Y$  which is not what one might expect, because  $f$  and  $\theta$  will not both be identity transformations as the model might suggest.

## 15.5 Additivity and Variance Stabilization

Additivity and variance stabilization (AVAS) is another TBS model and is quite similar to ACE. We choose the  $f_j$  to optimize the fit, but we also want constant variance for the response:

$$\text{var}[\theta(Y) | \sum_{j=1}^p f_j(X_j)] = \text{constant}$$

So we choose the  $f_j$ 's to get a good additive fit and choose the  $\theta$  to get constant variance.

Here is how the method of fitting  $\theta$  works: suppose  $\text{Var}(Y) \equiv V(Y)$  is not constant. We transform to constancy by:

$$\theta(t) = \int_0^t \frac{d\mu}{\sqrt{V(\mu)}}$$

We use data to estimate  $V(y)$ , then get  $\theta$ . The purpose of the AVAS method is to obtain additivity and variance stabilization and not necessarily to produce the best possible fit. We demonstrate its application on the ozone data:

```
avasfit <- avas(x, ozone$O3)
```

Plot the transformations selected:

```
plot(ozone$O3, avasfit$ty, xlab="O3", ylab=expression(theta(O3)))
plot(x[,1], avasfit$tx[,1], xlab="temp", ylab="f(temp)")
plot(x[,2], avasfit$tx[,2], xlab="ibh", ylab="f(ibh)")
plot(x[,3], avasfit$tx[,3], xlab="ibt", ylab="f(ibt)")
```

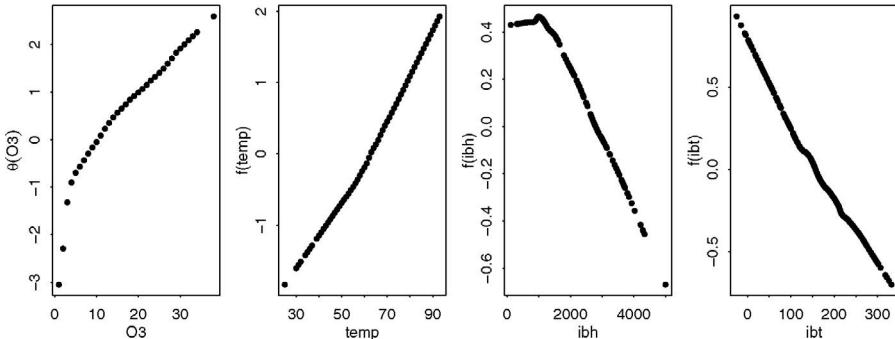


Figure 15.8 AVAS transformations — the first panel shows the transformation on the response while the remaining three show the transformations on the predictors.

See Figure 15.8. It would be convenient if the transformation on the response matched a simple functional form. We see if this is possible. We need to sort the response to get the line plots to work:

```
i <- order(ozone$O3)
plot(ozone$O3[i], avasfit$ty[i], type="l", xlab="O3", ylab=expression(
  ↪ theta(O3)))
gs <- lm(avasfit$ty[i] ~ sqrt(ozone$O3[i]))
lines(ozone$O3[i], gs$fit, lty=2)
gl <- lm(avasfit$ty[i] ~ log(ozone$O3[i]))
lines(ozone$O3[i], gl$fit, lty=5)
```

See the left panel of Figure 15.9. We have shown the square-root fit as a dotted line and log fit as a dashed line. Neither one fits well across the whole range. Now look at the overall fit:

```
lmod <- lm(avasfit$ty ~ avasfit$tx - 1)
summary(lmod)
```

	Estimate	Std. Error	t value	Pr(> t )
avasfit\$txtemp	0.9006	0.0746	12.08	< 2e-16
avasfit\$txibh	0.7950	0.1091	7.29	2.4e-12
avasfit\$txibt	0.5637	0.2389	2.36	0.019

$n = 330$ ,  $p = 3$ , Residual SE = 0.562, R-Squared = 0.69

The fit is not so good, but check the diagnostics:

```
plot(predict(lmod), residuals(lmod), xlab="Fitted", ylab="Residuals")
```

The plot is shown in the right panel of Figure 15.9.

AVAS does not optimize the fit; it trades some of the optimality in order to obtain constant variance. Whether this is a good trade depends on how much relative value you put on the accuracy of point predictions and accurate estimation of the standard error of prediction. In other words, is it more important to try to be right or to know how much you are wrong? The choice will depend on the application. An alternative

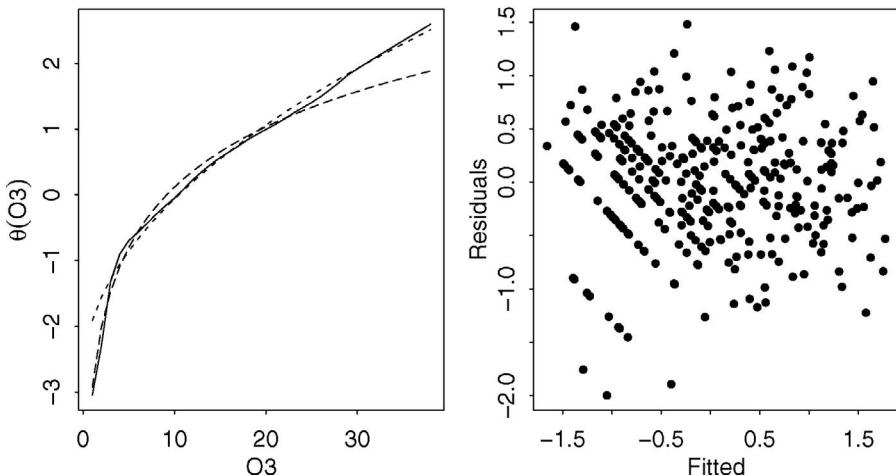


Figure 15.9 The left panel checks for simple fits to the AVAS transformation on the response given by the solid line. The log fit is given by the dashed line while the square-root fit is given by the dotted line. The right panel shows the residuals vs. fitted values plot for the AVAS model.

approach is to accept the nonconstant variance but try to model it and modify the uncertainty expressed in predictions accordingly. This sophistication in modeling is not always possible to achieve within some of the nonparametric regression frameworks.

## 15.6 Generalized Additive Mixed Models

The generalized additive mixed model (GAMM) manages to combine the three major themes of this book. The response can be nonnormal from the exponential family of distributions. The error structure can allow for grouping and hierarchical arrangements in the data. Finally we can allow for smooth transformations of the response. We demonstrate this method on the epilepsy data from Section 13.5:

```
data(epilepsy, package="faraway")
egamm <- gamm(seizures ~ offset(timeadj) + treat*expind+s(age), family
  ↪ =poisson, random=list(id=~1), data=epilepsy, subset=(id!=49))
summary(egamm$gam)
```

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-4.8392	0.1434	-33.74	<2e-16
treat	-0.0104	0.1999	-0.05	0.9584
expind	4.7255	0.0758	62.38	<2e-16
treat:expind	-0.3024	0.1127	-2.68	0.0077

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value
s(age)	1	1	0.31	0.58

```
R-sq. (adj) = 0.328
Scale est. = 2.5765    n = 290
```

We see that the age effect is not significant. Again the interaction effect is significant which shows, in this case, a beneficial effect for the drug.

## 15.7 Multivariate Adaptive Regression Splines

Multivariate adaptive regression splines (MARS) were introduced by Friedman (1991). We wish to find a model of the form:

$$\hat{f}(x) = \sum_{j=1}^k c_j B_j(x)$$

where the basis functions,  $B_j(x)$ , are formed from products of terms of the form  $[\pm(x_i - t)]_+^q$ . The  $[\cdot]_+$  denotes taking the positive part. For  $q = 1$ , this is sometimes called a hockey-stick function and can be seen in the right panel of Figure 14.5. The  $q = 1$  case is the most common choice and one might think this results in a piecewise linear fit, as in Section 14.2. However, the fit is more flexible than this. Consider the product of two hockey-stick functions in one dimension; this forms a quadratic shape. Furthermore, if we form the product of terms in two or more variables, we have an interaction term.

The model building proceeds iteratively. We start with no basis functions. We then search over all variables and possible knotpoints  $t$  to find the one basis function that produces the best fit to the data. We now repeat this process to find the next best basis function addition given that the first basis function is already included in the model. We might impose rules on what new basis functions are included. For example, we might disallow interactions or only allow up to two-way interactions. This will enhance interpretability, possibly at the cost of fit. The number of basis functions added determines the overall smoothness of the fit. We can use cross-validation to determine how many basis functions are enough.

When interactions are disallowed, the MARS approach will be a type of additive model. The MARS model building will be iterative, in contrast to the fit using the `gam` function of the `mgcv` package that fits and determines overall smoothness in one step. If a strictly additive model is all that is needed, the `mgcv` approach will typically be more effective. The MARS approach will have a relative advantage when interactions are considered, particularly when there are a larger number of variables. Here there will be a large number of potential interactions that cannot be simultaneously entertained. The iterative approach of MARS will be more appropriate here.

We apply the MARS method to the `ozone` dataset. We use the `earth` package which is built on the functionality in the original `mada` package.

```
library(earth)
mmod <- earth(O3 ~ ., ozone)
summary(mmod)

coefficients
(Intercept)      11.01166
h(5890-vh)      -0.01349
h(9-wind)        0.27685
```

```

h(humidity-41)      0.27438
h(humidity-54)      -0.26598
h(temp-58)          0.38519
h(1049-ibh)         -0.00269
h(ibh-1049)         -0.00056
h(dpg-10)           -0.10564
h(150-vis)          0.02336
h(96-doy)           -0.11977
h(doy-96)           0.03885
h(doy-158)          -0.08187

```

Selected 13 of 19 terms, and 8 of 9 predictors

Termination condition: Reached nk 21

Importance: temp, doy, humidity, dpg, ibh, vh, vis, wind, ibt-unused

Number of terms at each degree of interaction: 1 12 (additive model)

GCV 14.573    RSS 4108.1    GRSq 0.77362    RSq 0.80545

The default choice allows only additive (first-order) predictors and chooses the model size using a GCV criterion. The basis functions are expressed in terms of the breakpoints.

The fit is good in terms of  $R^2$ , but the model size is also larger. It is also an additive model, so we can reasonably compare it to the additive model presented at the end of Section 15.2. That model had an adjusted  $R^2$  of 79.3% using 19.4 parameters.

Let's reduce the model size to that used for previous models. The parameter nk controls the maximum number of model terms:

```

mmod <- earth(O3 ~ ., ozone, nk=7)
summary(mmod)

coefficients
(Intercept)    13.07286
h(58-temp)     -0.10623
h(temp-58)      0.49066
h(1049-ibh)    -0.00253
h(ibh-1049)    -0.00142
h(96-doy)       -0.09609
h(doy-96)       -0.01279

```

Selected 7 of 7 terms, and 3 of 9 predictors

Termination condition: Reached nk 7

Importance: temp, ibh, doy, vh-unused, wind-unused, humidity-unused,  
dpg-unused, ibt-unused, vis-unused

Number of terms at each degree of interaction: 1 6 (additive model)

GCV 18.284    RSS 5567.8    GRSq 0.71597    RSq 0.73631

This fit is worse, but remember we are disallowing any interaction terms. Now let's allow second-order (two-way) interaction terms. nk was chosen to get the same model size as before:

```

mmod <- earth(O3 ~ ., ozone, nk=7, degree=2)
summary(mmod)

```

```

coefficients
(Intercept)          10.03251
h(58-temp)           -0.10668
h(temp-58)            0.45593
h(ibh-1049)          -0.00115
h(55-humidity) * h(temp-58) -0.01311
h(humidity-55) * h(temp-58)  0.00599

```

```
Selected 6 of 7 terms, and 3 of 9 predictors
Termination condition: Reached nk 7
Importance: temp, ibh, humidity, vh-unused, wind-unused, dpg-unused,
ibt-unused, vis-unused, doy-unused
Number of terms at each degree of interaction: 1 3 2
GCV 18.383    RSS 5580.2    GRSq 0.71444    RSq 0.73573
```

This is a good fit. Compare this with an additive model approach. Since there are nine predictors, this would mean 36 possible two-way interaction terms. Such a model would be complex to estimate and interpret. In contrast, the MARS approach reduces the complexity by carefully selecting the interaction terms.

Now let's see how the terms enter into the model, as seen in Figure 15.10:

```
plotmo(mmod)
```

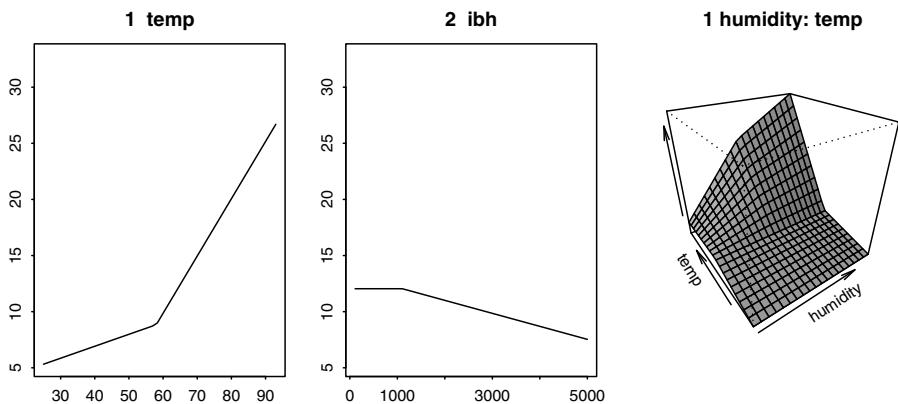


Figure 15.10 *Contribution of predictors in the MARS model.*

We see similar transformations to those used previously. Now check the diagnostics:

```
plot(mmod, 3)
plot(mmod, 4)
```

These plots, seen in Figure 15.11, show no problem with normality, but some indication of nonconstant variance.

It is interesting to compare the MARS approach to the univariate version as demonstrated in Figure 14.6. There we used a moderate number of knots in just one dimension while MARS gets by with just a few knots in higher dimensions. The key is to choose the right knots. MARS can be favorably compared to linear regression: it has additional flexibility to find nonlinearity in the predictors in higher dimensions. MARS can also be favorably compared to the tree method discussed in the next chapter: it allows for continuous fits but still maintains good interpretability.

**Further Reading:** Hastie and Tibshirani (1990) provide the original overview of additive modeling, while Wood (2006) gives a more recent introduction. Gu (2002) presents another approach to the problem. Green and Silverman (1993) show the link

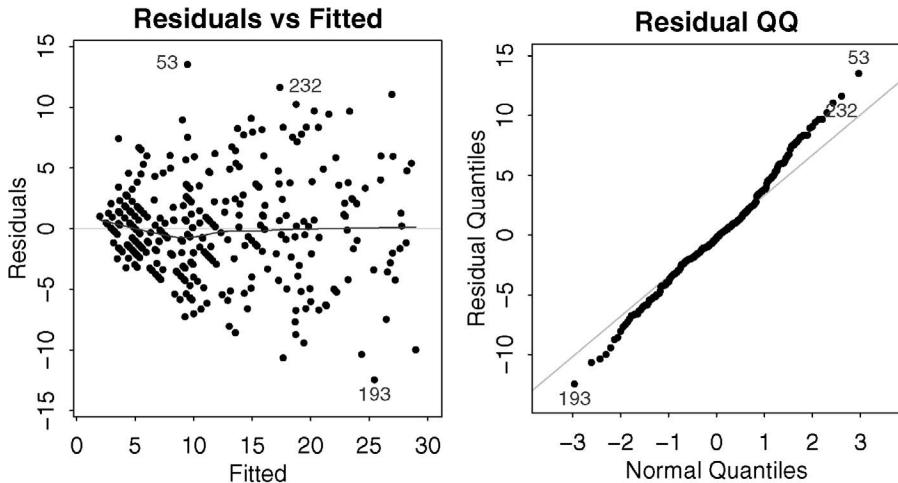


Figure 15.11 *Diagnostics for the MARS model.*

to GLMs. Hastie et al. (2001) discuss additive models as part of a larger review and compare them to competitive methods.

### Exercises

1. The `fat` data gives percentage of body fat, age, weight and height, and 10 body circumference measurements, such as the abdomen, are recorded for 252 men. Body fat is estimated through an underwater weighing technique, but this is inconvenient to use widely. In this question, we develop an additive model that allows the estimation of body fat for men using only a scale and a measuring tape. Your model should predict % body fat according to Siri. You may not use Brozek's % body fat, density or fat-free weight as predictors.
  - (a) Plot the Siri measure against each of the potential predictors and comment on the general pattern of relationships seen in these plots.
  - (b) Fit a linear model with the body fat as the response and use all the remaining variables as predictors. Use Cook Statistics to identify any extremely influential points. Eliminate these points and refit. (You may need to iterate this process.) List the influential points found and identify what particular values contribute to their status. What predictor is most significant in your final model? What value of  $R^2$  is achieved?
  - (c) Fit an additive model for the body fat with smooths on all the available predictors. Identify any influential points.
  - (d) Plot the transformations on the predictors identified by the additive model. Which predictor has the strongest relation to the response? What happens at the extreme values of  $x$  observed for some predictors?

- (e) What is the value of  $R^2$  for this additive model and how does it compare to the linear model? Does this mean the additive model is better? Which predictor makes the most nonlinear contribution to the prediction of the response? For this predictor, test whether it can be replaced with a linear term.
- (f) For the predictor identified in the previous question, plot the transformation used. Describe the nature of the relationship and attempt an interpretation. How can we tell from just the plot that this function is significantly nonlinear?
2. Find a good model for `volume` in terms of `girth` and `height` using the `trees` data. We might expect that  $\text{Volume} = c * \text{Height} * \text{Girth}^2$  suggesting a logarithmic transformation on all variables to achieve a linear model.
- Plot the relationship between the response and each of the predictors.
  - Fit a linear model with all three variables log-transformed. Check the diagnostics of this model. Do the coefficients of the model accord with the theoretically expected relationship?
  - Fit an additive model to the log-transformed variables. Plot the transformations chosen by the default additive model fit. How do these compare to the linear model?
  - Fit an additive model with a log-transformed response but with smooths on the untransformed predictors. Does this additive model point towards the expected log transformations?
  - Fit an ACE model. Plot the transformations found. Do these point towards the logged relationship that theory predicts?
3. The `pima` dataset consists of 768 female Pima Indians. We want to predict the diabetes test result from the other predictors.
- Plot the test results against each of the predictors. Identify impossible values of the predictors and replace them with the missing value code. Plot the data again and comment on the relationships.
  - Find the subset of cases that are complete (i.e., contain no missing values). From this subset, take a random sample of size 100. This is the test set. All the remaining cases, including those with missing values, are the training set.
  - Fit a GLM to the diabetes test outcome using all the predictors on the training set. Predict the response, negative if  $\hat{p} < 0.5$ , positive otherwise for the cases in the test set. Make a table showing how the test set predictions compare with the actual outcomes.
  - Use stepwise AIC variable selection on the GLM. Use this reduced model to predict the outcomes, evaluating its performance as in the last question.
  - Fit a GAM using all the predictors. Evaluate the performance on the test set.
  - Fit a GAM using only those predictors selected by the reduced GLM model. Evaluate performance.
  - Compare the results from the four different models. Is one of them clearly better than the others? Would this classification method be suitable for use in practice?

4. The `dvisits` data comes from the Australian Health Survey of 1977–1978 and consist of 5190 single adults where young and old have been oversampled.
  - (a) Build a generalized additive model with `doctorco` as the response and `sex`, `age`, `agesq`, `income`, `levyplus`, `freepoor`, `freerepa`, `illness`, `actdays`, `hscore`, `chcond1` and `chcond2` as possible predictor variables. You will need to decide which variables can reasonably be smoothed. Which variables are not statistically significant in your model.
  - (b) Fit a GAM that does not include the insignificant variables identified in the previous question. Test whether this model can be preferred to the larger model.
  - (c) Fit a GLM with the same predictor set as the previous question. Is this model superior to the GAM?
  - (d) What sort of person would be predicted to visit the doctor the most under the smaller GLM model.
  - (e) If you did the exercise on this data in Chapter 5, compare the interpretations of that selected GLM with the GAM used here.
  - (f) For the last person in the dataset, compute the predicted probability distribution for their visits to the doctor, i.e., give the probability they visit 0, 1, 2 etc. times.
5. The `ethanol` dataset in the `lattice` package presents data from ethanol fuel burned in a single-cylinder engine. The emissions of nitrogen oxides should be considered as the response and engine compression and equivalence ratio as the predictors.
  - (a) Reproduce the example plots given on the help page for `ethanol` to reveal the relationship between the variables.
  - (b) Fit a GAM with a bivariate `s()` on the predictors. Plot the relationship and comment.
  - (c) Refit the GAM but with an appropriate kind of smoothing. Plot the surface and compare.
  - (d) Fit a GAM with univariate smooths on the predictors. Can this model be used in preference to the bivariate version used in the previous question?
  - (e) Fit a MARS model with univariate smooths on the predictors. Compare it to the equivalent GAM fit.
  - (f) Now fit a bivariate MARS model and compare to the GAM version.
  - (g) Plot the relationships found by the MARS model and compare to those seen previously.

---

## Chapter 16

# Trees

---

### 16.1 Regression Trees

Regression trees are similar to additive models in that they represent a compromise between the linear model and the completely nonparametric approach. Tree methodology has roots in both the statistics and computer science literature. A precursor to current methodology was CHAID developed by Morgan and Sonquist (1963) although the book by Breiman et al. (1984) introduced the main ideas to statistics. Concurrently, tree methodology was developed in machine learning starting in the 1970s — see Quinlan (1993) for an overview.

Most statistical work starts from the specification of a model. The model says how we believe the data is generated and contains both a systematic and a random component. The model is not completely specified and so we use the data to select a particular model by either estimating parameters or perhaps by fitting functions, as in our recent nonparametric approaches. This strategy has been effective in a wide range of situations. However, the insistence on specifying a model, right from the start, does limit statistics. It is often difficult to specify a model, particularly for larger and more complex datasets. Furthermore, it is often impractical to develop inferential methods for more complex statistical models.

Tukey (1977) advocated exploratory data analysis (EDA) in his book. Graphical and descriptive statistics can sometimes make the message of the data very clear or at least suggest a suitable form for the model. However, EDA is not a complete solution and sometimes we need definite predictions or conclusions.

Regression trees are an example of a statistical method that is best described by the algorithm used in their construction. One can uncover the implicit model underlying regression trees, but the algorithm is the true starting point. Any method of analysis should ultimately be judged on whether it successfully predicts or explains something. Statistical models may achieve this, but algorithmically based methods are also competitive. The distinction between algorithm-based and model-based methods is discussed in Breiman (2001b). In the computer science literature, tree methodology has been applied to decision tree problems where there is no stochastic structure and we simply want to build a rule for making the correct decision.

To grow a tree, we use the recursive partitioning regression algorithm:

1. Consider all partitions of the region of the predictors into two regions where the division is parallel to one of the axes. In other words, we partition a single predictor by choosing a point along the range of that predictor to make the split. It does

not matter exactly where we make the split between two adjacent points so there will be at most  $(n - 1)p$  partitions to consider.

2. For each partition, we take the mean of the response in that partition. We then compute:

$$RSS(partition) = RSS(part_1) + RSS(part_2)$$

We then choose the partition that minimizes the residual sum of squares (RSS). We do need to consider many partitions, but the computations on each partition are simple, so that fit can be accomplished without excessive effort.

3. We now subpartition the partitions in a recursive manner. We only allow partitions within existing partitions and not across them. This means that the partitioning can be represented using a tree. There is no restriction preventing us from splitting the same variables consecutively.

For categorical predictors, it is possible to split on the levels of the factor. For an ordered factor with  $L$  levels, there are only  $L - 1$  possible splits. For an unordered factor, there are  $2^{L-1} - 1$  possible splits. Although this is a large number of possibilities as  $L$  grows, there is a way to limit the number that needs to be considered. Notice that there is no point in monotonely transforming a quantitative predictor as this will have no effect on the partitioning algorithm. Transforming the response will make a difference because it will change the computation of the RSS.

Missing values can be handled quite easily by tree methods. When we construct the tree, we may encounter missing values for a predictor when we are considering a split on that variable. We may simply exclude such points from the computation provided we weight appropriately. This approach is suitable for data where the observations are missing in a noninformative manner. If we believe the fact of being missing expresses some information, we might choose to treat missingness as an additional level of a factor. For continuous predictors, we could discretize the data into ranges so that it becomes a factor and then add missingness as an additional level. When we wish to predict the response for a new value with missing values, we can drop the prediction down through the tree until the missing values prevent us from going further. We can then use the mean value for that internal node.

Tree models are well suited to finding interactions. If we split on one variable and then split on another variable within the partitions of the first variable, we are finding an interaction between these two variables. As we construct further splits within splits, we are finding higher and higher order interactions. This may be a disadvantage as true high-order interactions are not common in reality. The MARS method discussed in Section 15.7 counteracts this by limiting the amount of interaction.

Trees are quite popular because the structure is easier for nontechnical people to understand. The term CART stands for Classification and Regression Trees and is also the name of a commercial software product.

We applied the regression tree methodology to study the relationship between atmospheric ozone concentration and meteorology in the Los Angeles Basin in 1976, as introduced in Section 15.1. A number of cases with missing variables have been removed for simplicity. The data were first presented by Breiman and Friedman (1985). We wish to predict the ozone level from the other predictors. We read in the data:

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

An examination of the data reveals several nonlinear relationships indicating that a linear regression might not be appropriate without the addition of some transformations. Now fit a tree:

```
library(rpart)
(tmod <- rpart(O3 ~ ., ozone))
n= 330
```

```
node), split, n, deviance, yval
 * denotes terminal node

1) root 330 21115.00 11.7760
  2) temp< 67.5 214 4114.30 7.4252 *
    4) ibh>=3573.5 108 689.63 5.1481 *
      5) ibh< 3573.5 106 2294.10 9.7453
        10) dpg< -9.5 35 362.69 6.4571 *
          11) dpg>=-9.5 71 1366.50 11.3660
            22) ibt< 159 40 287.90 9.0500 *
              23) ibt>=159 31 587.10 14.3550 *
    3) temp>=67.5 116 5478.40 19.8020
      6) ibt< 226.5 55 1276.80 15.9450
        12) humidity< 59.5 10 167.60 10.8000 *
        13) humidity>=59.5 45 785.64 17.0890 *
      7) ibt>=226.5 61 2646.30 23.2790
        14) doy>=306.5 8 398.00 16.0000 *
        15) doy< 306.5 53 1760.50 24.3770
          30) vis>=55 36 1149.90 22.9440 *
            31) vis< 55 17 380.12 27.4120 *
```

We see that the first split (nodes 2 and 3) is on temperature: 214 observations have temperatures less than 67.5 with a mean response value of 7.4, whereas 116 observations have temperatures greater than 67.5 with a mean response value of 20. The total RSS has been reduced from 21,115 to  $4114 + 5478 = 9592$ . The indentation describes the level of nesting within the tree. The children of node  $x$  are labeled as  $2x$  and  $2x + 1$ . So for example, node 8 does not appear because node 4 was not split, even though there are more than 8 nodes in this tree. Much more substantial output can be obtained from `summary(tmod)`.

Although the relevant information can be gleaned from the text-based output, a graphical display is nicer as in Figure 16.1. In the first version of the plot, the depth of the branches is proportional to the reduction in error due to the split. The disadvantage is that the labels can be hard to read in lower parts of the tree where the reduction in error is much smaller. The second version of the plot uses a uniform spacing to allow more room for labeling:

```
plot(tmod)
text(tmod)
plot(tmod, compress=T, uniform=T, branch=0.4)
text(tmod)
```

We see that the first split on temperature produces a large reduction in the RSS. Some of the subsequent splits do not do much. The immediate message is that high

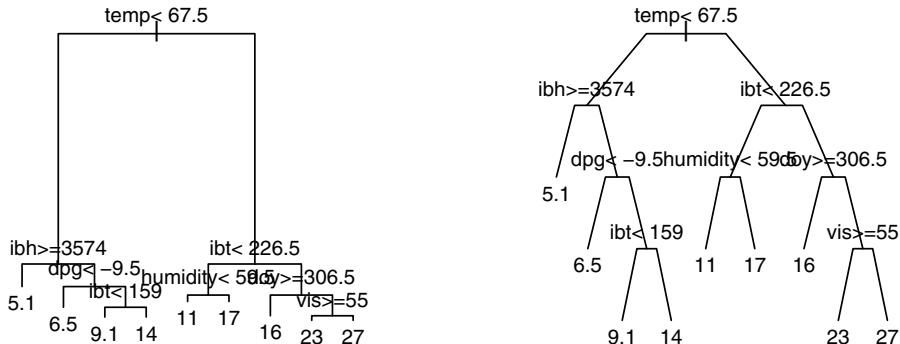


Figure 16.1 Tree model for the *ozone* data. On the left, the depth of the branches is proportional to the improvement in fit. On the right, the depth is held constant to improve readability. If the logical condition at a node is true, follow the branch to the left.

temperatures are associated with high ozone levels. A regression tree is a regression model, so diagnostics are worthwhile:

```
plot(jitter(predict(tmod)), residuals(tmod), xlab="Fitted", ylab="Residuals")
abline(h=0)
qqnorm(residuals(tmod))
qqline(residuals(tmod))
```

See Figure 16.2. There are no visible problems here. If nonconstant variance is observed, one might consider transforming the response. Trees are also somewhat sensitive to outliers as they are based on local means. Outliers may be observed in the QQ plot, but, as with linear models, they may conceal themselves and be influential on the fit. Suppose we wanted to predict the response for a new value — for example the median value in the dataset:

```
(x0 <- apply(ozone[,-1], 2, median))
vh      wind   humidity      temp      ibh      dpg      ibt      vis      doy
5760.0     5.0     64.0    62.0  2112.5    24.0    167.5   120.0   205.5
predict(tmod, data.frame(t(x0)))
```

1  
14.355

You should be able to verify this prediction by following the splits down through the tree shown in Figure 16.1.

## 16.2 Tree Pruning

The recursive partitioning algorithm describes how to grow the tree, but what is the optimal size for the tree? The default form of `rpart` does restrict the size of the tree, but some intervention is probably necessary to select the best tree size.

One possibility, called a *greedy strategy*, is to keep partitioning until the reduction in overall cost (RSS for this type of tree) is not reduced by more than  $\epsilon$ . However, it is difficult to set  $\epsilon$  in a sensible way. Furthermore, a greedy strategy may stop too soon. For example, consider data laid out in Table 16.1: neither the horizontal nor

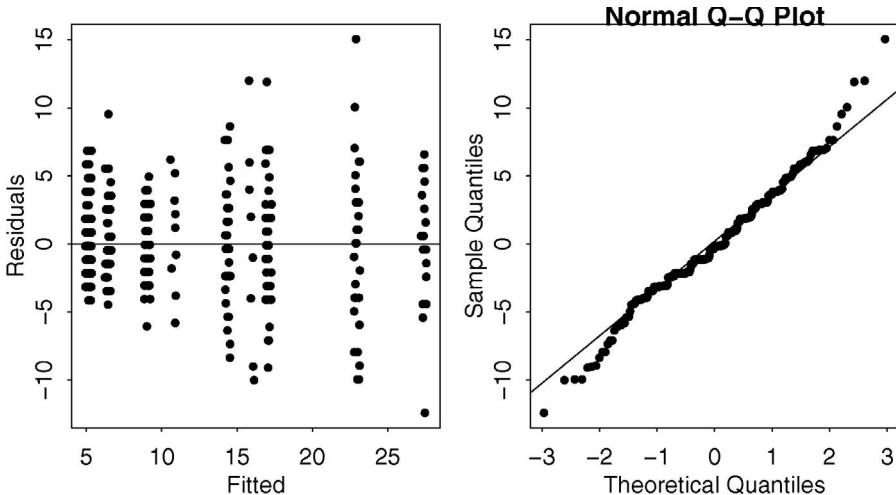


Figure 16.2 Residuals and fitted values for the tree model of the Ozone data are shown in the left panel. A QQ plot of the residuals is shown in the right panel.

$x_2$	1	2
	2	1
<hr/>		
$x_1$		

Table 16.1 There are four data points arranged in a square. The number shows the value of  $y$  at that point.

the vertical split will improve the fit at all. Both splits are required to get a better fit. We might need to look further ahead which will increase the computational cost. Furthermore, it's not clear what the best choice of  $\epsilon$  is for this strategy.

One general problem with model selection is that measures of fit such as the RSS (or deviance) usually improve as the complexity of the model increases. The measures tend to give a misleadingly optimistic impression of how well the model will predict future observations. A generic method of obtaining a better estimate of predictive ability is cross-validation (CV). For a given tree, leave out one observation, recalculate the tree and use that tree to predict the left-out observation. Repeat for all observations. For regression, this criterion would be:

$$\sum_{j=1}^n (y_j - \hat{f}_{(j)}(x_j))^2$$

where  $\hat{f}_{(j)}(x_j)$  denotes the predicted value of the tree given the input  $x_j$  when case  $j$  is not used in the construction of the tree. For other types of trees, a different criterion would be used. For classification problems, it might be the deviance.

CV is a more realistic estimate of how the tree will perform in practice.

Leave-out-one cross-validation is computationally expensive so often  $k$ -fold cross-validation is used. The data is randomly divided into  $k$  roughly equal parts. We use  $k - 1$  parts to predict the cases in the remaining part. We repeat  $k$  times, leaving out a different part each time.  $k = 10$  is a typical choice. As well as being much less expensive computationally than the full leave-out-one method, it may even work better. One drawback is that the partition is random so that repeating the method will give different numerical results.

However, there may be very many possible trees if we consider all subsets of a large tree; cross-validation would just be too expensive. We need a method to reduce the set of trees to be considered to just those that are worth considering. This is where cost-complexity pruning is useful. We define a cost-complexity function for trees:

$$CC(Tree) = \sum_{\text{terminal nodes: } i} \text{RSS}_i + \lambda(\text{number of terminal nodes})$$

If  $\lambda$  is large, then the tree that minimizes this cost will be small and vice versa. Notice the similarity to AIC. We can determine the best tree of any given size by growing a large tree and then pruning it back. Given a tree of size  $n$ , we can determine the best tree of size  $n - 1$  by considering all the possible ways of combining adjacent nodes. We pick the one that increases the fit criterion by the least amount. The strategy is akin to backward elimination in linear regression variable selection except that it can be shown that it generates the optimal sequence of trees of a given size.

We now use cross-validation to select from this sequence of trees. By default, `rpart` selects a tree size that may not be large enough to include all those trees we might want to consider. We force it to consider a larger tree and then examine the cross-validation criterion for all the subtrees. The parameter `cp` plays a similar role to the smoothing parameter in nonparametric regression and is defined as the ratio of  $\lambda$  to the RSS of the root tree (a tree with no branches). When we call `rpart` initially, it computes the whole sequence of trees and we merely need to use functions like `printcp` to examine the intermediate possibilities:

```
set.seed(123)
tmode <- rpart(O3 ~ ., ozone, cp=0.001)
printcp(tmode)
```

	CP	nsplit	rel error	xerror	xstd
1	0.54570	0	1.000	1.005	0.0767
2	0.07366	1	0.454	0.482	0.0427
3	0.05354	2	0.381	0.427	0.0393
4	0.02676	3	0.327	0.415	0.0392
5	0.02328	4	0.300	0.403	0.0389
6	0.02310	5	0.277	0.398	0.0387
7	0.01532	6	0.254	0.396	0.0385
8	0.01091	7	0.239	0.375	0.0343
9	0.00707	8	0.228	0.353	0.0323
10	0.00599	9	0.221	0.350	0.0346
11	0.00593	10	0.215	0.354	0.0350
12	0.00497	12	0.203	0.354	0.0350
13	0.00480	15	0.188	0.353	0.0355
14	0.00447	16	0.183	0.357	0.0359
15	0.00319	17	0.179	0.358	0.0357
16	0.00222	19	0.172	0.363	0.0365
17	0.00207	20	0.170	0.362	0.0361

18	0.00203	22	0.166	0.364	0.0361
19	0.00144	23	0.164	0.361	0.0361
20	0.00113	24	0.162	0.362	0.0361
21	0.00110	25	0.161	0.362	0.0361
22	0.00100	26	0.160	0.362	0.0361

In this table, we see the value of the `cp` parameter, the number of splits in the tree, the RSS of the tree divided by the RSS of the null tree. `xerror` denotes the cross-validated error which is also scaled by the RSS of the null tree. Since the partition of the data into 10 parts is random, this CV error is also random, which makes the given standard error useful. The random division also means that if you repeat this command, you will not get the same answer unless you use `set.seed`. We can select the size of the tree by minimizing the value of `xerror` and selecting the corresponding value of `CP`.

Because the selection method is quite dependent on the random division, another strategy for selecting the tree size can be considered. We can select the smallest tree with a CV error within one standard error of the minimum — in this case,  $0.350 + 0.035 = 0.385$ . So we would take the seven-split tree. We can illustrate this by plotting the CV error and a line showing one standard deviation above this value as shown in the first panel of Figure 16.3:

```
plotcp(tmod)
```

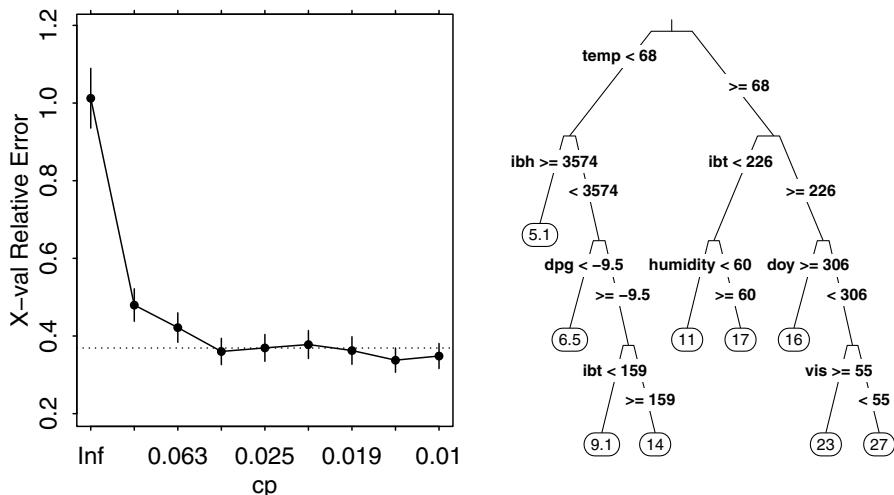


Figure 16.3 *Cross-validation plot for ozone tree model shown in the left panel and chosen tree model shown in the right panel.*

Notice that even the four-split tree comes close in terms of cross-validated error, so if we put a premium on simplicity, we might pick this tree. You can get some fancier output by using the `rpart.plot` package.

```
library(rpart.plot)
rpart.plot(tmod, type=3)
```

There are many options for the display—our chosen version is seen in the second panel of Figure 16.3. Let's compare the result to the earlier linear regression. We

achieved an  $R^2$  of about 70% using only six parameters in the previous chapter. We can select a tree with five splits and hence effectively six parameters and compare them:

```
tmodr <- prune.rpart(tmod, 0.0154)
1-sum(residuals(tmodr)^2)/sum((ozone$O3-mean(ozone$O3))^2)
[1] 0.74603
```

We see that the tree model achieved a better fit than the equivalent linear model. Of course, it would be a mistake to generalize from this, but it is a good demonstration of the value of trees. A tree fit is piecewise constant over the regions defined by the partitions, so one might not expect a particularly good fit. However, we can see that it is not necessarily worse than linear regression.

Regression methods are used for two main purposes — prediction and explanation. Tree-based methods share the transparency of linear models in that it is relatively easy to see how they work and understand how they will predict for new inputs. However, they have two disadvantages relative to linear models. They lack the inferential apparatus of prediction intervals. They have discontinuities in the prediction at the partition boundaries but are constant elsewhere. For optimal prediction performance, more flexible methods tend to do better.

If the goal of the analysis is in explaining how the predictors are related to the response, the tree model has the apparent advantage of selecting which variables have a strong relationship to the response. However, these relationships are not always stable. For example, suppose we randomly divide our data in two and see what tree is formed. We choose the same size tree as our final choice above.

```
set.seed(123)
tmod <- rpart(O3 ~ ., ozone[sample(330, 165),])
(tmods <- prune.rpart(tmod, 0.0154))
```

n= 165

```
node), split, n, deviance, yval
      * denotes terminal node

1) root 165 9743.60 10.9520
  2) temp< 72.5 127 2836.00  7.9134
    4) ibt< 118 52  241.77  4.6538 *
    5) ibt>=118 75 1658.70 10.1730
      10) doy>=357 10   18.00  4.0000 *
      11) doy< 357 65 1201.00 11.1230
        22) ibh>=2327.5 24   170.50  8.7500 *
        23) ibh< 2327.5 41   816.24 12.5120
          46) temp< 65.5 29   317.86 11.0690 *
          47) temp>=65.5 12   292.00 16.0000 *
  3) temp>=72.5 38 1817.60 21.1050
    6) ibt< 226.5 15   227.60 15.6000 *
    7) ibt>=226.5 23   838.87 24.6960 *
```

Although the initial split is about the same, the variables chosen for subsequent splits are somewhat different. We have set the random seed so we will get the same results every time, but if you repeat this, you will get a different tree every time. This instability in variable selection shows that one should avoid overinterpreting the tree model on the full data. Of course, similar problems can be seen when using linear models but the results do tend to be more stable.

Tree models are not optimal for prediction or explanation purposes. Nevertheless, they do provide a contrasting approach to linear models that may provide additional insight into the structure of your data.

### 16.3 Random Forests

The experience with fitting tree models to random partitions of the data provides the inspiration for a method that builds on trees to form a forest. The *random forest* (RF) method, introduced by Breiman (2001a), uses bootstrap aggregating, known as *bagging*. For  $b = 1, \dots, B$ ,

1. We draw a sample with replacement from  $(X, Y)$  to generate  $(X_b, Y_b)$ .
2. We fit a regression tree to  $(X_b, Y_b)$ .
3. For the set of cases not drawn in bootstrap sample (this will be about one third), we compute the mean squared error of prediction by inputting these predictor cases and comparing the predicted value to the observed value.

The latter step means that we have a measure of prediction performance that avoids the overfitting problem by not using data that was used in the construction of the given tree.

The  $B$  trees form the forest. Larger values of  $B$  are better although incremental improvement in performance levels off at some point. We will show later how we can be confident we have a sufficiently large  $B$ . We grow the trees as far as we can without going below a minimum of five cases per node. The trees in the forest will typically be larger than the one we would select as a single tree. New predictions can be made feeding the new predictor value into each of the trees in the forest and averaging the predictions made.

It has been observed that, for some datasets, certain predictors are chosen very frequently, meaning that there are strong correlations among the trees in the forest. To reduce this effect, at each node, a subsample of predictors is selected from which to choose a split. This ensures that every predictor has an opportunity to contribute to the prediction. The default choice of the subsample size is  $\sqrt{p}$  where  $p$  is the number of predictors.

Let's fit and examine the default forest. We use the `randomForest` package of Liaw and Wiener (2002).

```
library(randomForest)
fmod <- randomForest(O3 ~ ., ozone)
plot(fmod, main="")
```

The plot of the returned model objects, as seen in the first panel of Figure 16.4, shows the mean squared error (MSE) of prediction as the number of trees  $B$  used increases. There is a rapid decrease in this MSE as we progress through smaller values of  $B$  but we can see that the improvement levels off after 100. We can see that the default choice of 500 trees is more than enough in this example. Since the forest was computed quite rapidly for this size of dataset, there is no incentive to economize. For much larger datasets, one might wish to reduce the number of trees to save time. This plot suggests how much economy may be reasonable.

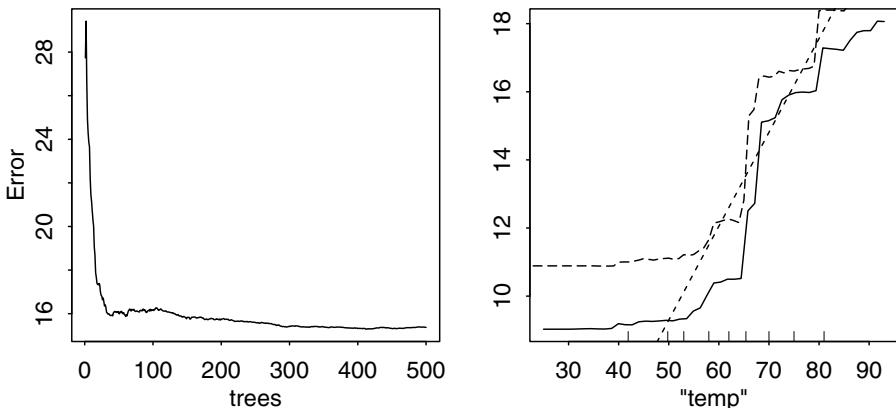


Figure 16.4 *MSE as a function of bootstrap sample size is shown on the left. Effect of temperature on ozone is shown on the right. The solid line is computed using the partial dependence method. The dashed line using the predicted RF response as temperature is varied and other predictors held at their means. The dotted line derives from a linear model.*

We must also choose the size of the subsample of predictors selected at each node (called `mtry` in the R command). This tuning parameter can have some impact on the performance of the forest. We can use cross-validation to choose from a range of values of `mtry`:

```
cvr <- rfcv(ozone[,-1], ozone[,1], step=0.9)
cbind(nvars=cvr$n.var, MSE=cvr$error.cv)
```

nvars	MSE
9	15.51
8	15.50
7	16.10
6	16.43
5	17.95
4	19.26
3	21.24
2	22.17
1	30.51

We have chosen the `step` fraction to be sufficiently large so that every subsample size is evaluated. We start with the full sample of  $p = 9$  and then move on to  $\text{step} \times p$  recursively, rounding at each stage. For larger datasets with more predictors, compute times may require us to be more economical by choosing a smaller `step` so that not too many subsample sizes are evaluated. We see that, in this example, a subsample of size 8 is suggested. However, since the performance of the full sample is hardly any different, we choose `mtry=9` for simplicity. Let's compute the  $R^2$  for comparison to previous models:

```
fmod <- randomForest(O3 ~ ., ozone, mtry=9)
1 - sum((fmod$predict - ozone$O3)^2) / sum((ozone$O3 - mean(ozone$O3))^2)
[1] 0.7455
```

The resulting fit is similar to those seen previously.

Although we can readily make a prediction using the forest model, we lack the same interpretability provided by the tree. One idea is to fix the value of all but one

predictor at some typical value, say the median, and vary the chosen predictor over its range and observe how the predictor varies. Let's perform this calculation for the temperature predictor. We form a data frame where only the temperature predictor varies:

```
tgrid <- 20:100
meds <- apply(ozone, 2, median)[-match(c("O3", "temp"), names(ozone))]
tdf <- cbind(temp=tgrid, data.frame(t(meds)))
head(tdf)
```

	temp	vh	wind	humidity	ibh	dpg	ibt	vis	doy	
1	20	5760	5		64	2112.5	24	167.5	120	205.5
2	21	5760	5		64	2112.5	24	167.5	120	205.5
3	22	5760	5		64	2112.5	24	167.5	120	205.5
4	23	5760	5		64	2112.5	24	167.5	120	205.5
5	24	5760	5		64	2112.5	24	167.5	120	205.5
6	25	5760	5		64	2112.5	24	167.5	120	205.5

Now we make predictions for all the cases in this data frame:

```
medfor <- predict(fmod, tdf)
```

We will plot these predictions but first we consider an alternative but similar approach called a *partial dependence plot* introduced by Friedman (1991). For each value of temperature on a grid, we compute the predicted value for *every* case in the dataset using the observed values of the predictors not including temperature. These predictions are then averaged. Thus this method requires  $n$  times as many calculations as our median-based method but forms an aggregate over the whole dataset. Such plots can be directly produced using:

```
partialPlot(fmod, ozone, "temp", main="")
```

We can add the median-based estimate of the effect:

```
lines(tgrid, medfor, lty=5)
```

Just for comparison purposes, we can make the same calculation using a linear model:

```
lmod <- lm(O3 ~ ., ozone)
lmpred <- predict(lmod, tdf)
lines(tgrid, lmpred, lty=2)
```

If you create all the partial dependence plots, you will find that temperature is the predictor exhibiting the strongest relationship with the response. Predictors ibh and ibt also demonstrate a lesser, but noticeable change in the predicted response over their range. In contrast, vh and wind show almost no relationship. This is an indication that, although random forests lack transparently interpretable regression coefficients, there is some indication of predictor importance. Such a measure of importance can be obtained as:

```
importance(fmod)
```

	IncNodePurity
vh	357.2
wind	177.8
humidity	891.3
temp	12820.6
ibh	1377.6
dpg	1034.4
ibt	2623.2
vis	714.0
doy	780.2

For regression-type forests, the importance is computed in the following way. For each bootstrap sample, the cases not chosen form the *out-of-bag* sample. We can use these to compute an MSE of prediction as before. Now for each predictor, we permute the values of that predictor in the out-of-bag sample while leaving the other predictors unchanged. We then recompute the MSE of prediction and compute the difference from unpermuted MSE. The permutation should destroy the relationship between that predictor and the response so if the predictor is important, the difference in the MSE will be large. We repeat this calculation over all the bootstrap samples and average the differences. This value is reported in the output of the `importance` above.

In this case, we see that temperature is by far the most important predictor. We also see which predictors we might safely discard, such as `vh` and `wind`. This method has no test for significance or criterion-based variable selection, but it does indicate the order in which we might consider discarding variables.

Random forests have demonstrated strong prediction performance in practice across a wide range of datasets. If the goal of your analysis is prediction, then the random forest is a good choice. In contrast, if your goal is to explain the relationship between the predictors and the response, random forests may provide some insight but they lack the full power of linear or additive models.

## 16.4 Classification Trees

Trees can be used for several different types of response data. For the regression tree, we computed the mean within each partition. This is just the null model for a regression. We can extend the tree method to other types of response by fitting an appropriate null model on each partition. For example, we can extend the idea to binomial, multinomial, Poisson and survival data by using a deviance, instead of the RSS, as a criterion.

Classification trees work similarly to regression trees except the residual sum of squares is no longer a suitable criterion for splitting the nodes. The splits should divide the observations within a node so that the class types within a split are mostly of one kind (or failing that, just a few kinds). We can measure the *purity* of the node with several possible measures. Let  $n_{ik}$  be the number of observations of type  $k$  within terminal node  $i$  and  $p_{ik}$  be the observed proportion of type  $k$  within node  $i$ . Let  $D_i$  be the measure for node  $i$  so that the total measure is  $\sum D_i$ . There are several choices for  $D_i$ :

1. Deviance:

$$D_i = -2 \sum_k n_{ik} \log p_{ik}$$

2. Entropy:

$$D_i = - \sum_k p_{ik} \log p_{ik}$$

3. Gini index:

$$D_i = 1 - \sum_k p_{ik}^2$$

All these measures share the characteristic that they are minimized when all members of the node are of the same type. The `rpart` function uses the Gini index by default.

We illustrate the classification tree method in a problem involving the identification of the sex and species of an historical specimen of kangaroo. We have some training data consisting of 148 cases with the following variables: there are three possible species, *Giganteus*, *Melanops* and *Fuliginosus*, the sex of the animal and 18 skull measurements. The data were published in Andrews and Herzberg (1985). The historical specimen is from the Rijksmuseum van Natuurlijke Historie in Leiden which had the following skull measurements in the same order as in the data:

```
1115 NA 748 182 NA NA 178 311 756 226 NA NA NA 48 1009 NA 204 593
```

We have a choice in how we model the response. One possibility is to form a six-level response representing all possible combinations of sex and species. Another approach is to form separate trees for identifying the sex and the species. We take the latter approach below, focusing on the species. This choice is motivated by the belief that different features are likely to discriminate the sex and the species so that attempting to model them both in the same tree might result in a larger, more complex tree that might be less powerful than two smaller trees. Even so, it would be worth trying the first approach although we shall not do so here. We start by reading in and specifying the museum case:

```
data(kanga, package="faraway")
x0 <- c(1115,NA,748,182,NA,NA,178,311,756,226,NA,NA,NA,48,1009,NA
      ↵ ,204,593)
```

We have missing values for the case to be classified. We have two options. We can build a tree model that will classify if there are missing values in the input or we can build a tree model that uses only variables that are observed. If we believe that the missing values were in some way informative, the first choice would be sensible. In this particular case, that does not seem plausible, so the latter approach is preferred. However, if we want to build a model that could be used for future unspecified cases, then we would have to deal directly with the missing values. For this special purpose situation, where we want to classify one particular kangaroo, this is not a concern.

We exclude all variables that are missing in the test case. We drop sex since we will not be modeling it yet. We form a convenient data frame:

```
kanga <- kanga[,c(T,F,!is.na(x0))]
kanga[1:2,]

  species basilar.length palate.length palate.width squamosal.depth
1 giganteus          1312           882            NA          180
2 giganteus          1439           985           230          150
  lacrymal.width zygomatic.width orbital.width foramina.length
1             394            782            249            88
2             416            824            233           100
  mandible.length mandible.depth ramus.height
1            10861           179            591
2            11582           181            643
```

We still have missing values in the training set. We have a number of options:

1. Build a tree model that discretizes the predictors into factors and then treats missing values as another level of the factors. This might be appropriate if we think missing values are informative in some way. Information would be lost in the discretization. For this data, we have no reason to believe that the data is not missing

at random and furthermore we have already decided to ignore the missing values in the test case.

2. Fill in or estimate the missing values and then build a tree. We could use missing data fill-in methods as used in other regression problems. This is not easy to implement and there are concerns about the bias caused by such methods.
3. The tree-fitting algorithm can handle missing values naturally. If a value for some case is not available, then it is simply excluded from the criterion. When we want to classify a new case with missing values, we follow the tree down until we reach a split that involves a missing value in our new case and take the majority verdict in that node. A more complicated approach is to allow a second-choice variable for splitting at a node called a *surrogate split*. Information on the surrogate splits may be obtained by using the `summary` command on the tree object.
4. Leave out the missing cases entirely.

We first check where the missing values occur:

```
apply(kanga, 2, function(x) sum(is.na(x)))
  species basilar.length palate.length palate.width
  0           1           1          24
squamosal.depth lacrymal.width zygomatic.width orbital.width
  1           0           1          0
foramina.length mandible.length mandible.depth ramus.height
  0          12           0          0
```

We observe that the majority of missing values occur in just two variables: mandible length and palate width. Suppose we throw out those variables and then remove the remaining missing cases. We compute the pairwise correlation of these variables with the other variables.

```
round(cor(kanga[,-1], use="pairwise.complete.obs")) [,c(3,9)], 2)
      palate.width mandible.length
basilar.length     0.77      0.98
palate.length      0.81      0.98
palate.width       1.00      0.81
squamosal.depth    0.69      0.80
lacrymal.width     0.77      0.92
zygomatic.width    0.78      0.92
orbital.width      0.12      0.25
foramina.length     0.19      0.23
mandible.length     0.81      1.00
mandible.depth      0.62      0.85
ramus.height        0.73      0.94
```

We see that these two variables are highly correlated with other variables in the data. We claim that there is not much additional information in these two variables and we can reasonably discard them. We do this and then remove the remaining missing cases:

```
newko <- na.omit(kanga[,-c(4,10)])
dim(newko)
[1] 144 10
```

After excluding these two variables, we only lose four cases more by throwing out all the missing value cases. Alternatively, suppose we just throw out the missing value cases on the original data:

```
dim(na.omit(kanga))
```

```
[1] 112 12
```

We would lose 36 cases by simply throwing out all the missing values. Removing a combination of variables and cases seems a better choice for this data.

We should also plot the data to see how the classes separate. An example of such a plot is:

```
ggplot(newko, aes(x=zygomatic.width, y=foramina.length, shape=species)) +
  geom_point() + theme(legend.position = "top", legend.direction =
  "horizontal", legend.title=element_blank())
```

We see in the left panel of Figure 16.5 that the classes do not separate well, at least for these two variables. We now fit a classification tree as follows: because the response

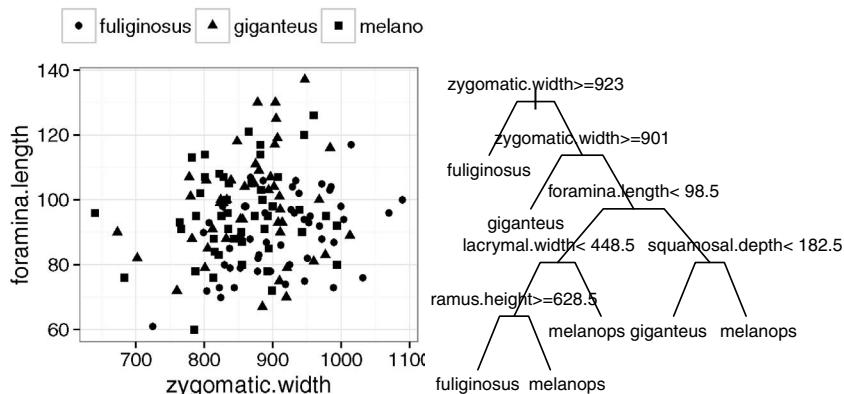


Figure 16.5 *Historical kangaroo tree model*. The left panel shows the three species as they vary with two of the measurements. The right panel shows the chosen tree.

is a factor, classification rather than regression is automatically used. Gini's index is the default choice of criterion. Here we specify a smaller value of the complexity parameter `cp` than the default, so that larger trees are also considered:

```
set.seed(123)
kt <- rpart(species ~ ., data=newko, cp=0.001)
printcp(kt)
```

Root node error: 95/144 = 0.66

n= 144

	CP	nsplit	rel error	xerror	xstd
1	0.179	0	1.00	1.20	0.051
2	0.105	1	0.82	0.97	0.061
3	0.050	2	0.72	0.86	0.063
4	0.021	6	0.52	0.82	0.063
5	0.011	7	0.49	0.84	0.063
6	0.001	8	0.48	0.87	0.062

The cross-validated error (expressed in relative terms in the `rel error` column) reaches a minimum for the six-split tree. We select this tree:

```
(ktp <- prune(kt, cp=0.0211))
```

n= 144

```

node), split, n, loss, yval, (yprob)
  * denotes terminal node

1) root 144 95 fuliginosus (0.34028 0.33333 0.32639)
  2) zygomatic.width>=923 37 13 fuliginosus (0.64865 0.16216 0.18919) *
  3) zygomatic.width< 923 107 65 giganteus (0.23364 0.39252 0.37383)
    6) zygomatic.width>=901 16  3 giganteus (0.12500 0.81250 0.06250) *
    7) zygomatic.width< 901 91 52 melanops (0.25275 0.31868 0.42857)
    14) foramina.length< 98.5 58 33 melanops (0.36207 0.20690 0.43103)
      28) lacrymal.width< 448.5 50 29 fuliginosus (0.42000 0.24000 0.34000)
        56) ramus.height>=628.5 33 14 fuliginosus (0.57576 0.18182 0.24242) *
        57) ramus.height< 628.5 17  8 melanops (0.11765 0.35294 0.52941) *
      29) lacrymal.width>=448.5 8  0 melanops (0.00000 0.00000 1.00000) *
    15) foramina.length>=98.5 33 16 giganteus (0.06061 0.51515 0.42424)
      30) squamosal.depth>=182.5 26 10 giganteus (0.07692 0.61538 0.30769) *
      31) squamosal.depth>=182.5 7  1 melanops (0.00000 0.14286 0.85714) *

plot(ktp, compress=T, uniform=T, branch=0.4)
text(ktp)

```

This tree is not particularly successful as the relative error is estimated as 80% of just guessing the species. Some of the terminal nodes are quite pure, for example, #29 and #31, while others retain much uncertainty, for example, #56 and #57. We now compute the misclassification error:

```

(tt <- table(actual=newko$species, predicted=predict(ktp, type="class"
  ↵ )))
  predicted
actual      fuliginosus giganteus melanops
  fuliginosus       43       4      2
  giganteus         12      29      7
  melanops          15       9     23
1 - sum(diag(tt))/sum(tt)
[1] 0.34028

```

We see that the error rate is 34%. We might hope to do better. We see that we can generally correctly identify *fuliginosus*, but we are more likely to be in error in distinguishing *melanops* and *giganteus*.

A look at the left panel of Figure 16.5 explains why we may have difficulty in classification. Any single measure will reflect mostly the overall size of the skull. For example, suppose we wanted to distinguish male human skulls from female human skulls. Most interesting measures will correlate strongly with size. If we just use one measure, then the rule will likely be: if the measure is small, then pick female; if it is large, pick male. This cannot be expected to work particularly well. There is something about the *relative* dimensions of the skulls that ought to be more informative.

One possibility is to allow splits on linear combinations of variables. This is allowed in some classification tree software implementations. An alternative idea is to apply the method to the principal component scores rather than the raw data. Principal components (PC) seek out the main directions of variation in the data and might generate more effective predictors for classification in this example:

```

pck <- princomp(newko[,-1])
pcdf <- data.frame(species=newko$species,pck$scores)
kt <- rpart(species ~ ., pcdf, cp=0.001)
printcp(kt)

```

```
Root node error: 95/144 = 0.66
```

```
n= 144
```

	CP	nsplit	rel error	xerror	xstd
1	0.400	0	1.00	1.18	0.053
2	0.179	1	0.60	0.76	0.063
3	0.042	2	0.42	0.52	0.060
4	0.011	3	0.38	0.54	0.060
5	0.001	5	0.36	0.61	0.062

We find a significantly smaller relative CV error (0.52). Our chosen tree is:

```
(ktp <- prune.rpart(kt, 0.0421))
```

```
n= 144
```

```
node), split, n, loss, yval, (yprob)
  * denotes terminal node
```

```
1) root 144 95 fuliginosus (0.34028 0.33333 0.32639)
  2) Comp.2< -15.13 49  8 fuliginosus (0.83673 0.04082 0.12245) *
  3) Comp.2>=-15.13 95 49 giganteus (0.08421 0.48421 0.43158)
    6) Comp.4>=-9.513 63 24 giganteus (0.11111 0.61905 0.26984)
      12) Comp.3>=-19 55 17 giganteus (0.09091 0.69091 0.21818) *
      13) Comp.3< -19 8  3 melanops (0.25000 0.12500 0.62500) *
    7) Comp.4< -9.513 32  8 melanops (0.03125 0.21875 0.75000) *
```

It is interesting that the first PC is not used. Typically, the first PC represents an overall average or total size. Other PCs represent contrasts between variables which would describe shape features in this case. We can also compute the error rate as before:

```
(tt <- table(newko$species,predict(ktp,type="class")))
  fuliginosus giganteus melanops
fuliginosus 41      5      3
giganteus   2     38      8
melanops    6     12     29
```

```
1-sum(diag(tt))/sum(tt)
```

```
[1] 0.25
```

We see that the error rate has been reduced to 25%. It would be worth considering other combinations of predictors in an attempt to reduce the error rate further.

Before we can predict the test case, we need to do some work to remove the missing values and unused variables and apply the principal component transformation:

```
nx0 <- x0[!is.na(x0)]
nx0 <- nx0[-c(3,9)]
nx0 <- (nx0-pck$center)/pck$scale
ndf <- data.frame(nx0 %*% pck$loadings)
predict(ktp, ndf)
  fuliginosus giganteus melanops
1  0.83673  0.040816  0.12245
```

We see that the test case is classified as *fuliginosus*, which agrees with the experts.

Given the structure of this dataset, with strongly correlated predictors, we might expect linear discriminant analysis, as explained in Section 7.2, to perform well. It is simple to implement:

```
library(MASS)
ldamod <- lda(species ~ ., newko)
(tt <- table(newko$species,predict(ldamod)$class))
```

```

fuliginosus giganteus melanops
fuliginosus      42       6      1
giganteus        5      32     11
melanops         3      14     30
1-sum(diag(tt))/sum(tt)
[1] 0.27778

```

We see that the default classification performance is reasonably good compared to trees.

It is a mistake to generalize from a single dataset. Different classification methods are likely to do well for certain kinds of data but perform poorly in others. It's always a good idea to try more than one.

## 16.5 Classification Using Forests

Random forests are also a natural extension to classification trees. These can be applied in much the same way as in the regression case but now we use a measure of node purity such as the Gini index in place of MSE as a measure of success. We demonstrate the random forest method on the kangaroo skull data. We start by selecting the subsample size:

```

cvr <- rfcv(newko[,-1], newko[,1], step=0.9)
cbind(nvars=cvr$n.var,error.rate=cvr$error.cv)
  nvars error.rate
9      9  0.45139
8      8  0.43750
7      7  0.43750
6      6  0.43750
5      5  0.47222
4      4  0.55556
3      3  0.55556
2      2  0.59028
1      1  0.59028

```

We see that we can go as low as six for the sub-sample size. All things be equal, a small subsample size will be faster. Now given this choice:

```

fmod <- randomForest(species ~ ., newko, mtry=6)
(tt <- table(actual=newko$species,predicted=predict(fmod)))
  predicted
actual      fuliginosus giganteus melanops
  fuliginosus      38       5      6
  giganteus        6      21     21
  melanops        10      21     16
1-sum(diag(tt))/sum(tt)
[1] 0.47917

```

The performance is quite poor. As before, we might attempt to improve matters using principal components on the predictors:

```

pck <- princomp(newko[,-1])
pcdf <- data.frame(species=newko$species,pck$scores)
fmod <- randomForest(species ~ ., pcdf, mtry=6)
tt <- table(actual=newko$species,predicted=predict(fmod))
1-sum(diag(tt))/sum(tt)
[1] 0.33333

```

The result is not impressive but random forests have a strong track record in classification across a wide range of problems so do not let this single poor outcome discourage you from using this method.

### Exercises

1. Four hundred three African Americans were interviewed in a study to understand the prevalence of obesity, diabetes and other cardiovascular risk factors in central Virginia. Data is presented in `diabetes`. In this question we build a regression tree-based model for predicting glycosolated hemoglobin (`glyhb`) in terms of the other relevant variables.
  - (a) Plot the response against each of the predictors and comment on the apparent strength of the relationships observed.
  - (b) Investigate the pattern of missing values in the data. By eliminating a combination of rows and columns, produce a reduced dataset that contains no missing values. Try to make the reduction as small as possible.
  - (c) Fit a linear model with main effects for all the predictors. Which predictors appear to be most important? Make a plot of the residuals and fitted values from this model and comment.
  - (d) Fit the default tree. From the output, answer the following questions: How many observations had `stab.glu < 158`? What is the mean response for these observations? What characterizes the largest terminal node in the tree? What is the mean response in this node?
  - (e) Make a plot of the tree. What feature of the plot reveals the most important predictor?
  - (f) Plot the residuals against the fitted values for this tree. Comment.
  - (g) Select the optimal tree using cross-validation. What would be the smallest tree that could reasonably be used?
  - (h) Predict the response for an individual with these predictor values:

```

id chol stab.glu hdl ratio   location age gender height
1004 213        72  58   3.3 Buckingham 56 female    64
weight frame bp.1s bp.1d bp.2s bp.2d waist hip time.ppn
     131 medium    108     55     NA     NA     30    40     720
  
```

- (i) Glycosolated hemoglobin greater than 7.0 is usually taken as a positive diagnosis of diabetes. Build the default classification tree for the diagnosis of diabetes and compare this model to the corresponding regression tree. Which predictors are most useful?
2. Refer to the `pima` dataset described in Question 3 of Chapter 15. If you have not done so already, answer parts (a) and (b) of that question before proceeding with the rest of this question.
  - (a) Fit the default tree model with the result of the diabetes test as the response and all the other variables as predictors using the training set. Make a nice plot of the tree and comment on the shape of the tree. Is it broad or deep?

- (b) What fraction of cases in the test set are correctly classified according to this model?
- (c) Use cross-validation to select the optimal tree size. For the selected model, check the performance on the test set.
- (d) Fit the default random forest model to the training set. Use this to classify the cases in the test set. How many are correctly classified?
- (e) Use the subsample size selection method for the random forest. For the selected model, evaluate the performance on the test set.
3. The dataset `wbca` is described in Question 2 of Chapter 2. You may find it helpful to repeat the plotting part of that question first.
- (a) Without modifying the data, fit a tree model with `Class` as the response and the other nine variables as predictors. Now change `Class` to a factor and refit the tree. What is the difference between these two trees and which one should we prefer?
- (b) Use the default (10-way) cross-validation to select a tree size. Now repeat the cross-validation. Explain why the outcome is different. Now use leave-one-out-one cross-validation (set the argument `xval` equal to the sample size). Does this calculation change if it is repeated? What model is chosen by this latter method?
- (c) Using the last model chosen in the previous question, determine how many cases in the data are correctly classified using the model. (Suppose a cancer is classified as benign if  $p > 0.5$  and malignant if  $p < 0.5$ .)
- (d) Compute the receiver operating characteristic (ROC) plot by varying the 0.5 threshold used in the previous question.
- (e) Pick out every third observation as a test set. Use the remainder as a training set with which to estimate the model. Evaluate the performance on the test set. Compare the results to (c). Which classification rate is a more reliable estimate of future performance?
4. The dataset `uswages` is drawn as a sample from the Current Population Survey in 1988.
- (a) Plot the `wage` response with each of the remaining variables as potential predictors. Try a log-scale for the response. Does this reveal more about the relationships in the data?
- (b) Fit the default tree regression model to predict `wage` using untransformed versions of all the predictors. Now transform years of experience using the function  $\log(x + 3)$ . (The  $+3$  is to cancel out some negative values.) Does this change the fit of the model? Explain.
- (c) Now fit a model with a logged response. Compare this to the previous model and explain why it is different. Compare the RSS from both models — you will need to back-transform the fitted values from the logged model for comparability.

- (d) Predict the response from the untransformed tree model for each year of education ranging from 0 up to 18 with the other predictors held at their median values. Plot the observed data on wages and years of education and display your predictions on top of this plot. Now fit a standard linear mode to the same variables, compute the predictions in the same manner and also display on the same plot. Compare the two sets of predictions.
5. The dvisits data comes from the Australian Health Survey of 1977–1978 and consist of 5190 single adults where young and old have been oversampled.
- Build a Poisson tree model with doctorco as the response and sex, age, agesq, income, levyplus, freepoor, freerepa, illness, actdays, hscore, chcond1 and chcond2 as possible predictor variables. Consult the rpart documentation for how to specify a Poisson response. What size tree might be the minimum acceptable?
  - Make a plot of the selected tree and interpret. How do the predictors explain the response.
  - For the last person in the dataset, compute the predicted probability distribution for their visits to the doctor, i.e., give the probability they visit 0, 1, 2 etc. times.
  - Fit a random forest model to the same variables. You will need to use the regression method. Construct the same predicted probabilities as in the previous question.
  - Assess the importance of each variable in the forest model. Make partial dependence plots of the most and the least important predictors. Compare the two.

**This page intentionally left blank**

## Neural Networks

---

Neural networks (NN) were originally developed as an attempt to emulate the human brain. The brain has about  $1.5 \times 10^{10}$  neurons each with 10 to  $10^4$  connections called synapses. The speed of messages between neurons is about 100 m/sec which is much slower than CPU speed. Given that our fastest reaction time is around 100 ms and neuron computation time is 1–10 ms, then the number of steps must be less than 100. This is inconceivably small for a sequential computation, even in machine code; therefore, the brain must be computing in parallel. The original idea behind neural networks was to use a computer-based model of the human brain to perform complex tasks. We can recognize people in fractions of a second, but this task is difficult for computers. So why not make software more like the human brain? Despite the promise, there are some drawbacks. The brain model of connected neurons, first suggested by McCulloch and Pitts (1943), is too simplistic given more recent research. For these and other reasons, the methodology is more properly called *artificial* neural nets. As with artificial intelligence, the promise of NNs is not matched by the reality of their performance. There was a large amount of hype concerning NNs but that is now past. Nevertheless, viewed simply as an algorithmic procedure, NNs are competitive with other less ambitiously named methods. NNs are used for various purposes. They can be used as biological models, which was the original motivation. They can also be used as a hardware implementation for adaptive control. But the area of application we are interested in is data analysis. There are NN models that rival the regression, classification and clustering methods normally used by statisticians. A perceptron is a model of a neuron and is the basic building block of a neural network as depicted in Figure 17.1. The output  $x_o$  is determined from inputs  $x_i$ :

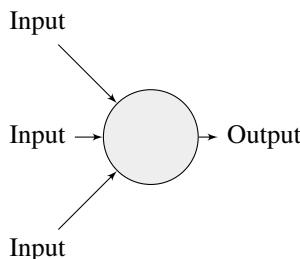


Figure 17.1 A perceptron.

$$x_o = f_o(\sum_{\text{inputs}:i} w_i x_i)$$

where  $f_o$  is called the *activation function*. Standard choices include the identity, logistic and indicator functions. The  $w_i$  are *weights*. The NN *learns* the weights from the data. A statistician would prefer to say that the NN estimates the parameters from the data. Thus NN terminology differs from statistical usage in ways that can be confusing.

## 17.1 Statistical Models as NNs

Three common statistical models are analogous to the single perceptron NN. For multiple linear regression:

$$y = \sum_i w_i x_i$$

So here  $f_o$  is the identity function. We can define  $x_1 \equiv 1$  to get an intercept term. The NN alternative is to attach a weight, called a *bias*, to each neuron:

$$f(x) = x + \theta$$

A statistician would call the bias  $\theta$  an intercept.

Logistic regression also fits easily within this framework if we define  $f_o$  as the logistic function. Of course, such an NN is not exactly equivalent to the corresponding statistical model unless that NN is fit in a very particular way.

Linear discriminant analysis is used to classify a binary response. Suppose there are two groups encoded by  $y = 0$  or  $y = 1$ . In this case  $f_o$  is the indicator function. Again the NN and statistical approach are not exactly equivalent unless the same fitting procedures are used.

Other common statistical models can be approximated by adding more neurons. Multivariate multiple linear regression with a bivariate response is  $Y = X\beta + \varepsilon$  where  $Y, X, \beta, \varepsilon$  are all matrices and is depicted as an NN in Figure 17.2: polynomial regression can be mimicked by using a different activation function and more than one layer of neurons as seen in the second panel of Figure 17.2.

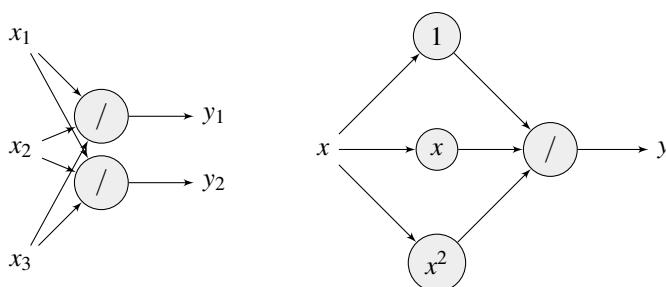


Figure 17.2 *NN equivalent of multivariate linear regression is shown on the left. It uses linear activation functions. Polynomial regression is shown on the right and uses powers for activation functions.*

## 17.2 Feed-Forward Neural Network with One Hidden Layer

The feed-forward neural network with one hidden layer is the most common choice for regression-like modeling applications. It takes the form:

$$y_o = \phi_o \left( \sum_h w_{ho} \phi_h \left( \sum_i w_{ih} x_i \right) \right)$$

The activation functions for the hidden layer,  $\phi_h$ , are almost always logistic. If identity functions are used for the hidden layer and for the output, the resulting NN is quite similar to the partial least squares approach of Wold et al. (1984). We will set one of our inputs to be constant at one so as to allow for an intercept/bias term. The choice of output activation function depends on the nature of the response. For continuous unrestricted output, an identity function is appropriate, while for response bounded between zero and one, such as a binomial proportion, a logistic function should be used. We show the feed-forward NN in Figure 17.3. Sometimes a direct

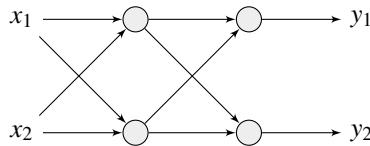


Figure 17.3 *Feed-forward neural network with one hidden layer.*

connection between the inputs and outputs is added called a *skip-layer* connection. More complexity can be added using more layers or feedbacks although this not always beneficial to the practical performance of the NN.

NNs can be elaborated to perform as universal approximators. This has been shown by authors such as Hornik et al. (1989). However, these results are of little practical value when confronted with a finite amount of data subject to noise. We must use the data to estimate the parameters of the model or, in NN-speak, use the data to train the network. The weights  $w$  are chosen to minimize a criterion, such as  $MSE = \sum(y - \hat{y})^2$  where  $y$  is the observed output and  $\hat{y}$  is the predicted output. A different criterion would be more suitable for categorical responses.

NN researchers have developed different methods of estimation motivated by brain models of learning. These methods have generally not compared well with the numerical analysis-based approaches which are generally faster and more reliable. Nevertheless, in all but the most simple NNs, the criterion is a complicated function of the parameters. The function often has many local minima making it difficult to find the true minimum. A statistician, with no pretensions to mimicking brain functions, would be inclined to use standard methods of numerical analysis such as quasi-Newton methods, conjugate gradients or simulated annealing. The `nnet` function in R uses the BFGS method, as described in Fletcher (1987).

### 17.3 NN Application

We apply the NN method to the ozone data analyzed in previous chapters. The *nnet* package, due to Venables and Ripley (2002), must be loaded first:

```
library(nnet)
data(ozone, package="faraway")
```

We start with just three variables for simplicity of exposition as in previous analyses. We fit a feed-forward NN with one hidden layer containing two units with a linear output unit:

```
set.seed(123)
nnmdl <- nnet(O3 ~ temp + ibh + ibt, ozone, size=2, linout=T)
# weights:  11
initial value 65447.874069
final value 21115.406061
converged
```

If you repeat this, your result may differ slightly because of the random starting point of the algorithm, but you will likely get a similar result. We set the random number seed for reproducibility.

The RSS of 21,115 is equal to  $\sum_i (y_i - \bar{y})^2$ , so the fit is not any better than the null model. The problem lies with the initial selection of weights. It is hard to do this well when the variables have very different scales. The solution is to rescale the data to have zero mean and unit variance:

```
sx <- scale(ozone)
```

Because a random starting point is used, the algorithm will not necessarily converge to the same solution if the fitting is repeated. Now we try refitting the model. We repeat this 100 times because of the random starting point. Here we find the best fit of the 100 attempts:

```
bestrss <- 10000
for(i in 1:100) {
  nnmdl <- nnet(O3 ~ temp + ibh + ibt, sx, size=2, linout=T, trace=F)
  cat(i, nnmdl$value, "\n")
  if(nnmdl$value < bestrss) {
    bestnn <- nnmdl
    bestrss <- nnmdl$value
  }
}
bestnn$value
```

[1] 88.031

The criterion function has 11 parameters or weights and has multiple minima. The problem is that we can never really know whether we have found the true minimum. All we can do is keep trying and stop if we do not find anything better after some number of attempts. The best strategy is not clear, although one can do better than the simple approach we have used above. Examine the estimated weights:

```
summary(bestnn)
```

```
a 3-2-1 network with 11 weights
options were - linear output units
b->h1 i1->h1 i2->h1 i3->h1
  1.12   -0.98    0.84    0.29
b->h2 i1->h2 i2->h2 i3->h2
137.89  -74.74  240.66  137.89
b->o h1->o h2->o
  2.59   -4.41    0.67
```

The notation `i2->h1`, for example, refers to the link between the second input variable and the first hidden neuron. `b` refers to the bias, which takes a constant value of one. We see that there is one skip-layer connection, `b->o`, from the bias to the output.

NNs have some drawbacks relative to competing statistical models. The parameters of an NN are uninterpretable whereas they often have some meaning in statistical models. Furthermore, NNs are not based on a probability model that expresses the structure and variation. As a consequence, there are no standard errors. It is possible to graft some statistical inference onto this NN model, but it is not easy. The bootstrap is a possible way of implementing this. The  $R^2$  for the fit is:

```
1-88.03/sum((sx[,1]-mean(sx[,1]))^2)
```

```
[1] 0.73243
```

which is very similar to the additive model fit for these predictors.

Although the NN weights may be difficult to interpret, we can get some sense of the effect of the predictors by observing the marginal effect of changes in one or more predictor as other predictors are held fixed. Here, we vary each predictor individually while keeping the other predictors fixed at their mean values. Because the data has been centered and scaled for the NN fitting, we need to restore the original scales. The fits are shown in Figure 17.4:

```
ozmeans <- colMeans(ozone)
ozscales <- apply(ozone, 2, sd)
xx <- expand.grid(temp=seq(-3, 3, 0.1), ibh=0, ibt=0)
plot(xx$temp*ozscales['temp']+ozmeans['temp'], predict(bestnn, new=xx) *
    ↪ ozscales['O3']+ozmeans['O3'], xlab="Temp", ylab="O3", type="l")
xx <- expand.grid(temp=0, ibh=seq(-3, 3, 0.1), ibt=0)
plot(xx$ibh*ozscales['ibh']+ozmeans['ibh'], predict(bestnn, new=xx) *
    ↪ ozscales['O3']+ozmeans['O3'], xlab="IBH", ylab="O3", type="l")
xx <- expand.grid(temp=0, ibh=0, ibt=seq(-3, 3, 0.1))
plot(xx$ibt*ozscales['ibt']+ozmeans['ibt'], predict(bestnn, new=xx) *
    ↪ ozscales['O3']+ozmeans['O3'], xlab="IBT", ylab="O3", type="l")
```

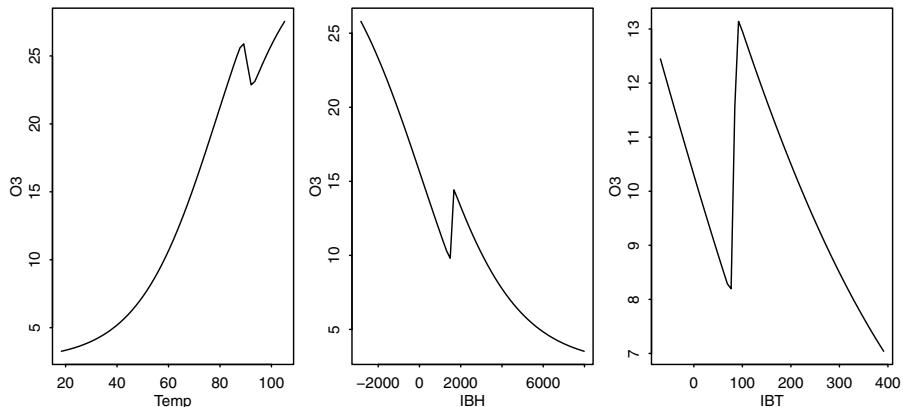


Figure 17.4 *Marginal effects of predictors for the NN fit. Other predictors are held fixed at their mean values.*

We see some surprising discontinuities in the plots which do not seem consistent with what we might expect for the effect of these predictors. If we examine the weights

for this NN above, we see several large values. Consider that all the variables have been scaled to mean zero and variance one. Products formed using the large weights will vary substantially. The situation is analogous to the collinearity problem in linear regression where unreasonably large regression coefficients are often seen. The NN is choosing extreme weights in order to optimize the fit, but the predictions will be unstable, especially for extrapolations.

We can use a penalty function, as with smoothing splines, to obtain a more stable fit. Instead of minimizing  $MSE$ , we minimize:

$$MSE + \lambda \sum_i w_i^2$$

In NN terms, this is known as *weight decay*. The idea is similar to ridge regression. Let's try  $\lambda = 0.001$  for 100 NN model fits:

```
bestrss <- 10000
for(i in 1:100) {
  nnmdl <- nnet(O3 ~ temp + ibh + ibt, sx, size=2, linout=T, decay
    ↪ = 0.001, trace=F)
  cat(i, nnmdl$value, "\n")
  if(nnmdl$value < bestrss) {
    bestnn <- nnmdl
    bestrss <- nnmdl$value
  }
}
bestnn$value
[1] 92.055
```

The value of the best RSS is somewhat larger than before. We expect this since weight decay sacrifices some fit to the current data to obtain a more stable result. We repeat the assessment of the marginal effects as before and display the results in Figure 17.5:

```
xx <- expand.grid(temp=seq(-3, 3, 0.1), ibh=0, ibt=0)
plot(xx$temp*ozscales['temp']+ozmeans['temp'], predict(bestnn, new=xx) *
  ↪ ozscales['O3']+ozmeans['O3'], xlab="Temp", ylab="O3", type="l")
xx <- expand.grid(temp=0, ibh=seq(-3, 3, 0.1), ibt=0)
plot(xx$ibh*ozscales['ibh']+ozmeans['ibh'], predict(bestnn, new=xx) *
  ↪ ozscales['O3']+ozmeans['O3'], xlab="IBH", ylab="O3", type="l")
xx <- expand.grid(temp=0, ibh=0, ibt=seq(-3, 3, 0.1))
plot(xx$ibt*ozscales['ibt']+ozmeans['ibt'], predict(bestnn, new=xx) *
  ↪ ozscales['O3']+ozmeans['O3'], xlab="IBT", ylab="O3", type="l")
```

We see that the fits are now plausibly smooth. Note that `ibh` is strictly positive in practice so the strange behavior for negative values is irrelevant. Compare these plots to Figure 15.3. The shapes are similar for temperature and `ibh`. The `ibt` plot looks quite different although we have no way to assess the significance of any of the terms in the NN fit.

NNs have interactions built in so one should also look at these. We could produce analogous plots to those in Figure 17.5 by varying two predictors at a time.

Now let's look at the full dataset. We use four hidden units because there are now more inputs.

```
bestrss <- 10000
for(i in 1:100) {
  nnmdl <- nnet(O3 ~ ., sx, size=4, linout=T, trace=F)
  cat(i, nnmdl$value, "\n")
```

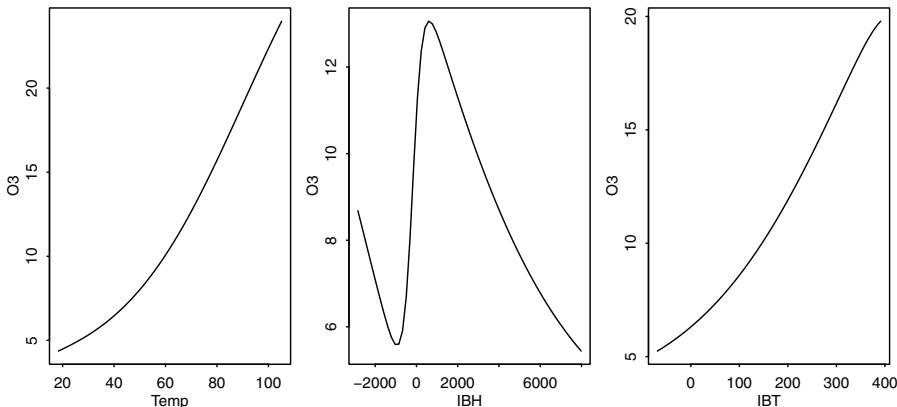


Figure 17.5 Marginal effects of predictors for the NN fit with weight decay. Other predictors are held fixed at their mean values.

```

if(nnmdl$value < bestrss) {
  bestnn <- nnmdl
  bestrss <- nnmdl$value
}
1-bestnn$value/sum((sx[,1]-mean(sx[,1]))^2)
[1] 0.85063

```

The fit is good and there may be better minimum than we have found and increasing the number of hidden units would always improve the fit. The fit can be compared to those in previous chapters. The  $R^2$ s for the linear, tree and MARS model fits were not as impressive but these approaches tell us more about the relationship between the predictors and the response, so it would not be sensible to judge by  $R^2$  alone. It would be rash to draw firm conclusions from just one dataset. Furthermore, the value of the modeling approaches needs to be judged within the context of the particular problem. If explanation is the main goal of the data analysis, NNs are not a good choice. If prediction is the objective, we cannot judge just by the fit to the data we have now. It is more important how the model performs on future observations. We do not have fresh data here as we have used it all to fit the data. It is a good idea to withhold data for use in testing the prediction performance of the models considered. NNs have been generally competitive in comparison studies but by no means dominant.

## 17.4 Conclusion

NNs, as presented here, are a controlled flexible class of nonlinear regression models. By adding more hidden units we can control the complexity of the model in a measured way from relatively simple models up to models suitable for large datasets with complex structure. NNs are also attractive because they require less expertise to use successfully compared to statistical models. Nevertheless the user must still pay attention to basic statistical issues involving transformation and scaling of the data

and outliers and influential points. See Faraway and Chatfield (1998) for an example of the application of neural networks and how they compare with statistical methods.

NNs are generally good for prediction but bad for understanding. The NN weights are almost uninterpretable. Although one can gain some insight from plotting the marginal effect of predictors, the NN inevitably introduces complex interactions that often do not reflect reality. Furthermore, without careful control, the NN can easily overfit the data resulting in overoptimistic predictions.

NNs are quite effective for large complex datasets compared to statistical methods where the burden of developing an appropriate sampling model can sometimes slow or even block progress. NNs do lack good statistical theory for inference, diagnostics and model selection. Of course, they were not developed with these statistical considerations in mind, but experience shows that such issues are often important.

NNs can outperform their statistical competitors for some problems provided they are carefully used. However, one should not be fooled by the evocative name, as NNs are just another tool in the box.

**Further Reading:** See Ripley (1996), Bishop (1995), Haykin (1998), Neal (1996) and Hertz et al. (1991) for more on NNs.

## Exercises

1. Consider the `diabetes` dataset analyzed in the exercises for Chapter 16. Before starting, plot the data and construct a reduced dataset without missing values as described in that question.
  - (a) Compute the total sum of squares for the response, `glyhb`. This is the sum of squares about the mean. Now fit a neural network model with this response and the remaining variables as predictors. Use two hidden nodes. Do not rescale and fit the model only once. Compute the residual sum of squares. How has the model performed?
  - (b) Rescale all the numerical variables in the data frame and recompute the total sum of squares. Refit the same neural network model but to the rescaled data frame. Has the model done any better?
  - (c) Fit the model of the previous question 100 times, saving the RSS on each iteration. Plot the RSS against the iteration index. Comment. Do you think more iterations would help?
  - (d) The number of steps within each model fit is limited to 100 and is controlled by the argument `maxit` in the `nnet` call. Set this argument to 200 and repeat the model fitting 100 times as in the previous question. Does increasing this argument improve the fit of the models?
  - (e) Fit the model 100 times as before, but this time save the model producing the best fit. Plot the weights of the best fitting model against the weights from the 100th model fitted. Comment on the relationship.
  - (f) Fit the model 100 times but now use four hidden units. Does this produce a better fit?

2. Refer to the `pima` dataset described in Question 3 of Chapter 15.
- (a) Find all the zero predictor values which are clearly impossible and replace them with missing values. Now fit a neural network model with the diabetes test outcome as the response and the other variables as predictors. Use two hidden units. Do not rescale the data yet. Examine the fitted values from this fit. How do they compare to the proportion of negative and positive test results in the data? How many fitted values are calculated? How does this compare to the number of rows in the data frame?
  - (b) Rescale all the predictors to standard units and refit the model. Are the fitted values more informative now?
  - (c) Fit the model 100 times and save the best fitting result. What proportion of response values are correctly classified by this best model?
  - (d) Construct a test set by taking a random sample of 100 cases from the subset of rows which contain no missing values. Use all the remaining cases as the training set. Fit the model 100 times to the training set and choose the best fit. Evaluate this best model by its performance on the training set. Compare the classification rate achieved with that of the previous question. Which rate is a more realistic representation of how the model might perform on fresh data?
  - (e) Increase the number of hidden units to eight. Evaluate the model performance on the full data (fitting 100 times and picking the best model). Now evaluate performance by fitting the models on the training set and classifying the test set. Compare the two performances and comment. Does adding more hidden units help the classification?
3. In this question we model the `Volume` in terms of `Girth` and `Height` using the `trees` data.
- (a) Plot the response on each of the predictors and comment on the relationship. Do they look linear?
  - (b) Fit a linear model with `Volume` as the response and the other two variables as predictors. Produce effect displays like those seen in Figure 15.2 for both predictors. Comment on the relationships seen.
  - (c) Rescale all the variables to standard units. Fit a neural network model with two hidden units. Repeat 100 times and pick the best fit. Compute the  $R^2$  for this model and compare to the linear model fit.
  - (d) Construct a data frame where the (standardized) height varies in the range  $[-2, 2]$  and the girth is fixed at the standardized mean of zero. Predict the response for the cases in this data frame. Plot the volume against the height and show your predicted responses as a line on top of this plot. Comment on the shape of the curve. Repeat the exercise, reversing the roles of height and girth.
  - (e) Construct a perspective plot showing the bivariate view of the fit. Look at Figure 14.14 for hints on the construction.

**This page intentionally left blank**

---

## Appendix A

---

# Likelihood Theory

---

This appendix is just an overview of the likelihood theory used in this book. For greater detail or a more gentle introduction, the reader is advised to consult a book on theoretical statistics such as Cox and Hinkley (1974), Bickel and Doksum (2015), Wood (2015) or Rice (1998).

### A.1 Maximum Likelihood

Consider  $n$  independent discrete random variables,  $Y_1, \dots, Y_n$ , with probability distribution function  $f(y|\theta)$  where  $\theta$  is the, possibly vector-valued, parameter. Suppose we observe  $\mathbf{y} = (y_1, \dots, y_n)^T$ , then we define the likelihood as:

$$P(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^n f(y_i|\theta) = L(\theta|y)$$

So the likelihood is a function of the parameter(s) given the data and is the probability of the observed data given a specified value of the parameter(s).

For continuous random variables,  $Y_1, \dots, Y_n$  with probability density function  $f(y|\theta)$ , we recognize that, in practice, we can only measure or observe data with limited precision. We may record  $y_i$ , but this effectively indicates an observation in the range  $[y_i^l, y_i^u]$  so that:

$$P(Y_i = y_i) = P(y_i^l \leq y_i \leq y_i^u) = \int_{y_i^l}^{y_i^u} f(u|\theta) du \approx f(y_i|\theta)\delta_i$$

where  $\delta_i = y_i^u - y_i^l$ . We can now write the likelihood as:

$$L(\theta|y) \approx \prod_{i=1}^n f(y_i|\theta) \prod_{i=1}^n \delta_i$$

Now provided that  $\delta_i$  is relatively small and does not depend on  $\theta$ , we may ignore it and the likelihood is the same as in the discrete case.

As an example, suppose that  $Y$  is binomially distributed  $B(n, p)$ . The likelihood is:

$$L(p|y) = \binom{n}{y} p^y (1-p)^{n-y}$$

The *maximum likelihood* estimate (MLE) is the value of the parameter(s) that gives the largest probability to the observed data or, in other words, maximizes the

likelihood function. The value at which the maximum occurs,  $\hat{\theta}$ , is the maximum likelihood estimate. In most cases, it is easier to maximize the log of likelihood function,  $l(\theta|y) = \log L(\theta|y)$ . Since log is a monotone increasing function, the maximum occurs at the same  $\hat{\theta}$ .

In a few cases, we can find an exact analytical solution for  $\hat{\theta}$ . For the binomial, we have the log-likelihood:

$$l(p|y) = \log \binom{n}{y} + y \log p + (n-y) \log(1-p)$$

The *score function*,  $u(\theta)$ , is the derivative of the log-likelihood with respect to the parameters. For this example, we have:

$$u(p) = \frac{dl(p|y)}{dp} = \frac{y}{p} - \frac{n-y}{1-p}$$

We can find the maximum likelihood estimate  $\hat{p}$  by solving  $u(p) = 0$ . We get  $\hat{p} = y/n$ . We should also verify that this stationary point actually represents a maximum.

Usually we want more than an estimate; some measure of the uncertainty in the estimate is valuable. This can be obtained via the Fisher information which is:

$$I(\theta) = \text{var } u(\theta) = -E \frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^T}$$

If there is more than one parameter,  $I(\theta)$  will be a matrix. The information at  $\hat{\theta}$  is the second derivative at the maximum. Large values indicate high curvature so that the maximum is well defined and even close alternatives will have much lower likelihood. This would indicate a high level of confidence in the estimate. One can show that the variance of  $\hat{\theta}$  can be estimated by:

$$\text{vár}(\hat{\theta}) = I^{-1}(\hat{\theta})$$

under mild conditions. Sometimes it is difficult to compute the expected value of the matrix of second derivatives. As an alternative, the observed, rather than expected, value at  $\hat{\theta}$  may be used instead. For the binomial example this gives:

$$\text{vár} \hat{p} = \hat{p}(1-\hat{p})/n$$

We illustrate these concepts by plotting the log-likelihood for two binomial datasets: one where  $n = 25, y = 10$  and another where  $n = 50, y = 20$ . We construct the log-likelihood function:

```
loglik <- function(p,y,n) lchoose(n,y) + y*log(p) + (n-y)*log(1-p)
```

For ease of presentation, we normalize by subtracting the log-likelihood at the maximum likelihood estimate:

```
nloglik <- function(p,y,n) loglik(p,y,n) - loglik(y/n,y,n)
```

Now plot the two log-likelihoods, as seen in Figure A.1:

```
pr <- seq(0.05, 0.95, by=0.01)
matplot(pr, cbind(nloglik(pr,10,25),nloglik(pr,20,50)), type="l", xlab="p", ylab="log-likelihood")
```

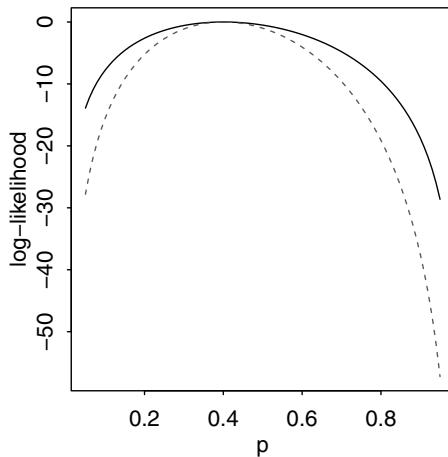


Figure A.1 Normalized binomial log-likelihood for  $n = 25, y = 10$  shown with a solid line and  $n = 50, y = 20$  shown with a dotted line.

We see that the maximum occurs at  $p = 0.4$  in each case at a value of zero because of the normalization. For the larger sample, we see greater curvature and hence more information.

Examples where likelihood can be maximized explicitly are confined to simple cases. Typically, numerical optimization is necessary. The *Newton–Raphson* method is the most well-known technique. Let  $\theta_0$  be an initial guess at  $\theta$ , then we update using:

$$\theta_1 = \theta_0 - H^{-1}(\theta_0)u(\theta_0)$$

where  $H$  is the Hessian matrix of second derivatives:

$$H(\theta) = \frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^T}$$

We iterate this method, putting  $\theta_1$  in place of  $\theta_0$  and so on, until the procedure (hopefully) converges. This method works well provided the log-likelihood is smooth and convex around the maximum and that the initial value is reasonably close. In less well-behaved cases, several things can go wrong:

- The likelihood has multiple maxima. The maximum that Newton–Raphson finds will depend on the choice of initial estimate. If you are aware that multiple maxima may exist, it is advisable to try multiple starting values to search for the overall maximum. The number and choice of these starting values are problematic. Such problems are common in fitting neural networks, but rare for generalized linear models.
- The maximum likelihood may occur at the boundary of the parameter space. This means that perhaps  $u(\hat{\theta}) \neq 0$ , which will confuse the Newton–Raphson method. Mixed effect models have several variance parameters. In some cases, these are maximized at zero, which causes difficulties in the numerical optimization.

- The likelihood has a large number of parameters and is quite flat in the neighborhood of the maximum. The Newton–Raphson method may take a long time to converge.

The Fisher scoring method replaces  $H$  with  $-I$  and sometimes gives superior results. This method is used in fitting GLMs and is equivalent to iteratively reweighted least squares.

A minimization function that uses a Newton-type method is available in R. We demonstrate its use for likelihood maximization. Note that we need to minimize  $-l$  because `nlm` minimizes, not maximizes:

```
f <- function(x) -loglik(x, 10, 25)
mm <- nlm(f, 0.5, hessian=T)
```

We use a starting value of 0.5 and find the optimum at:

```
mm$estimate
[1] 0.4
```

The inverse of the Hessian at the optimum is equal to the standard estimate of the variance:

```
c(1/mm$hessian, 0.4*(1-0.4)/25)
[1] 0.0096016 0.0096000
```

Of course, this calculation is not necessary for the binomial, but it is useful for cases where exact calculation is not possible.

## A.2 Hypothesis Testing

Consider two nested models, a larger model  $\Omega$  and a smaller model  $\omega$ . Let  $\hat{\theta}_\Omega$  be the maximum likelihood estimate under the larger model, while  $\hat{\theta}_\omega$  is the corresponding value when  $\theta$  is restricted to the range proscribed by the smaller model. The *likelihood ratio test statistic* (LRT) is:

$$2 \log(L(\hat{\theta}_\omega)/L(\hat{\theta}_\Omega)) = -2(l(\hat{\theta}_\omega) - l(\hat{\theta}_\Omega))$$

Under some regularity conditions, this statistic is asymptotically distributed  $\chi^2$  with degrees of freedom equal to the difference in the number of identifiable parameters in the two models. The approximation may not be good for small samples and may fail entirely if the regularity conditions are broken. For example, if the smaller model places some parameters on the boundary of the parameter space, the  $\chi^2$  may not be valid. This can happen in mixed effects models when testing whether a particular variance component is zero.

The *Wald test* may be used to test hypotheses of the form  $H_0 : \theta = \theta_0$  and the test statistic takes the form:

$$(\hat{\theta} - \theta_0)^T I(\hat{\theta})(\hat{\theta} - \theta_0)$$

Under the null, the test statistic has approximately a  $\chi^2$  distribution with degrees of freedom equal to the number of parameters being tested. Quite often, one does not wish to test all the parameters and the Wald test is confined to a subset. In particular, if we test only one parameter,  $H_0 : \theta_i = \theta_{i0}$ , the square root of the Wald test statistic is simply:

$$z = \frac{\hat{\theta}_i - \theta_{i0}}{se(\hat{\theta}_i)}$$

This is asymptotically normal. For a Gaussian linear model, these are the  $t$ -statistics and have an exact  $t$ -distribution, but for generalized linear and other models, the normal approximation must suffice.

The *score test* of the hypothesis  $H_0 : \theta = \theta_0$  uses the statistic:

$$u(\theta_0)^T I^{-1}(\theta_0) u(\theta_0)$$

and is asymptotically  $\chi^2$  distributed with degrees of freedom equal to the number of parameters being tested.

There is no uniform advantage to any of these three tests. The score test does not require finding the maximum likelihood estimate, while the likelihood ratio test needs this computation to be done for both models. The Wald test needs just one maximum likelihood estimate. However, although the likelihood ratio test requires more information, the extra work is often rewarded. Although the likelihood ratio test is not always the best, it has been shown to be superior in a wide range of situations. Unless one has indications to the contrary or the computation is too burdensome, the likelihood ratio test is the recommended choice.

These test methods can be inverted to produce confidence intervals. To compute a  $100(1 - \alpha)\%$  confidence interval for  $\theta$ , we calculate the range of hypothesized  $\theta_0$  such that  $H_0 : \theta_0 = 0$  would not be rejected at the  $\alpha$  level. The computation is simple for the single-parameter Wald test where the confidence interval for  $\theta_i$  is:

$$\hat{\theta}_i \pm z^{1-\alpha/2} se(\hat{\theta}_i)$$

where  $z$  is the appropriate quantile of the normal distribution. The computation is trickier for the likelihood ratio test. If we are interested in a confidence interval for a single parameter  $\theta_i$ , we will need to compute the log-likelihood for a range of  $\theta_i$  with the other  $\theta$  set to the maximizing values. This is known as the *profile likelihood* for  $\theta_i$ . Once this is computed as  $l_i(\theta_i|y)$ , the confidence interval is:

$$\{\theta_i : 2(l(\hat{\theta}_i|y) - l(\theta_i|y)) < \chi_1^{1-\alpha}\}$$

As an example, this type of calculation is used in the computation of the confidence interval for the transformation parameter used in the Box–Cox method.

We can illustrate this by considering a binomial dataset where  $n = 100$  and  $y = 40$ . We plot the normalized log-likelihood in Figure A.2 where we have drawn a horizontal line at half the distance of the 0.95 quantile of  $\chi_1^2$  below the maximum:

```
pr <- seq(0.25, 0.55, by=0.01)
plot(pr, nloglik(pr, 40, 100), type="l", xlab="p", ylab="log-likelihood")
abline(h=-qchisq(0.95, 1)/2)
```

All  $p$  that have a likelihood above the line are contained within a 95% confidence interval for  $p$ . We can compute the range by solving for the points of intersection:

```
g <- function(x) nloglik(x, 40, 100)+qchisq(0.95, 1)/2
uniroot(g, c(0.45, 0.55))$root
[1] 0.49765
uniroot(g, c(0.25, 0.35))$root
[1] 0.30743
abline(v=c(0.49765, 0.30743))
```

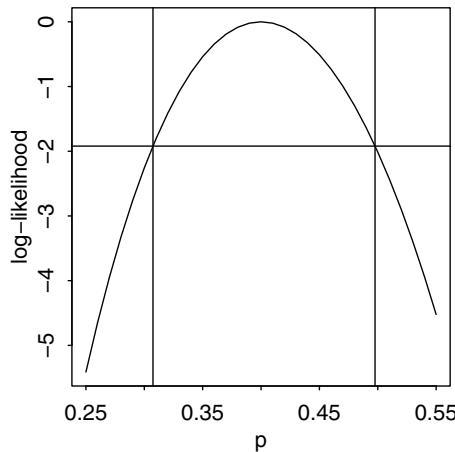


Figure A.2 Likelihood ratio test-based confidence intervals for binomial  $p$ .

The confidence interval is  $(0.307, 0.498)$  as is indicated by the vertical lines on the plot. We can compute the Wald test-based interval as:

```
se <- sqrt(0.4*(1-0.4)/100)
cv <- qnorm(0.975)
c(0.4-cv*se, 0.4+cv*se)
[1] 0.30398 0.49602
```

which is very similar, but not identical, to the LRT-based intervals.

Suppose we are interested in the hypothesis,  $H_0 : p = 0.5$ . The LRT and  $p$ -value are:

```
(lrstat <- 2*(loglik(0.4, 40, 100)-loglik(0.5, 40, 100)))
[1] 4.0271
pchisq(lrstat, 1, lower=F)
[1] 0.044775
```

So the null is barely rejected at the 5% level. The Wald test gives:

```
(z <- (0.5-0.4)/se)
[1] 2.0412
2*pnorm(z, lower=F)
[1] 0.041227
```

Again, not very different from the LRT. The score test takes more effort to compute. The observed information is:

$$\frac{-d^2l(p|y)}{dp^2} = \frac{y}{p^2} + \frac{n-y}{(1-p)^2}$$

We compute the score and information at  $p = 0.5$  and then form the test and get the  $p$ -value:

```
(sc <- 40/0.5-(100-40)/(1-0.5))
[1] -40
(obsinf <- 40/0.5^2+(100-40)/(1-0.5)^2)
[1] 400
(score.test <- 40*40/400)
[1] 4
```

```
pchisq(4, 1, lower=F)
[1] 0.0455
```

The outcome is again slightly different from the previous two tests. Asymptotically, the three tests agree. We have a moderate size sample in the example, so there is little difference. More substantial differences could be expected for smaller sample sizes.

### A.3 Model Selection

It is possible to use hypothesis testing methods to choose between models but this is not always practical or desirable. Sometimes the models we want to compare are not nested and sometimes we want to compare larger numbers of models. Comparing multiple models introduces problems with multiple testing which degrades the meaning of  $p$ -values. Furthermore, we might wish to select a model for superior prediction — hypothesis testing is not designed for this purpose.

We envisage comparing several models and require a criterion that will measure the fitness and allow us to choose the best model. Many choices have been proposed but some are only applicable to particular classes of model. We present three popular criteria:

**AIC** or Akaike Information Criterion: The Kullback-Leibler distance is a measure of the closeness of our proposed model to the true model, having density  $g(y)$ . It is defined by

$$K(f, g) = \int (\log g(y) - \log(f(y|\hat{\theta})))g(y)d(y)$$

This is the expected log-likelihood ratio of the true model and the proposed model, evaluated at the true model. This would be an ideal criterion except that we need to know the true model. However, we can estimate the expected value  $EK(f, g)$  using

$$-l(\hat{\theta}) + p + \int \log(g(y))g(y)dy$$

where  $p$  is the number of parameters. Since the last term only depends on the true model and not our proposed model, we can discard it when comparing proposed models. The AIC criterion is then:

$$AIC = -2l(\hat{\theta}) + 2p$$

The multiplication by two makes the criterion compatible with the definition of the likelihood ratio statistic which also multiplies by two (so that the null distribution is  $\chi^2$ ). AIC is not perfect — one problem is its lack of consistency. It is not guaranteed to choose the true model asymptotically. Nevertheless, it is the most widely used model selection criterion.

**BIC** or Bayes Information Criterion: This criterion has a Bayesian motivation. The *marginal likelihood* is defined as

$$\int f(y|\theta)h(\theta)d\theta$$

where  $h(\theta)$  is a prior density on  $\theta$ . The *Bayes factor* compares two models by considering the ratio of the marginal likelihoods. All things being equal, we would favor the model with the larger marginal likelihood. As we can see, this will depend on the choice of prior  $h(\theta)$  which is undesirable. However, a crude approximation to minus twice the log marginal likelihood is

$$BIC = -2l(\hat{\theta}) + p \log n$$

which does not depend on the prior. We can now use BIC to compare models without concern for the prior, making it usable for non-Bayesian model selection. Notice the similarity to AIC despite the different motivation. BIC penalizes larger models more heavily than AIC and so it prefers smaller models.

**DIC** or Deviance Information Criterion: In more complex models, particularly involving hierarchical data structures, it is not always obvious how to count  $p$  or  $n$ . DIC uses an *effective degrees of freedom*:

$$p_D = \overline{D(\theta)} - D(\bar{\theta})$$

where  $D$  is a deviance  $-2l(\theta)$ . The bar in  $\bar{x}$  denotes mean over  $x$ . The DIC is then defined as

$$DIC = D(\bar{\theta}) + 2p_D$$

DIC is more naturally computed for Bayesian models but can also be used for likelihood-based models where we assume a noninformative prior.

---

## Appendix B

---

# About R

---

**Installation:** R may be obtained free of charge from the R project Website at [www.r-project.org](http://www.r-project.org).

**How to Learn R:** The R website provides extensive resources for learning R. Several free introductory guides can be found along with a list of the many books about R which are now available. I have assumed a basic knowledge of R for readers of this text but it is not necessary to learn it in detail before starting on this book. Many readers are able to pick up the language from the examples in the text with reference to the help pages. For example, for more information on the `lm` function, type within R:

```
help(lm)
```

More advanced users may find the R mailing lists and online forums such as Stack-Exchange helpful for more difficult questions.

**Packages:** This book uses some functions and data that are not part of base R. These are collected in *packages* which you may need to install. You can find out which packages are already installed in your *library* by:

```
library()
```

Some packages, such as MASS, come with the base installation so you will not need to install them. If you have not already done so, you will need to select a location from which to download additional packages by setting your CRAN (Comprehensive R Archive Network) mirror by:

```
chooseCRANmirror()
```

You can then install packages. You will surely need to install the faraway package which contains data and functions specific to this text:

```
install.packages("faraway")
```

You may also wish to install the other packages used in this text which are:

```
acepack, brglm, dispmmod, dplyr, earth, effects, geepack, ggplot2,  
lme4, pbkrtest, pscl, randomForest, reshape2, robust, RLRsim,  
rpart.plot, rstan, sandwich, sm, SuppDists, tidyR, VGAM, wavethresh
```

All these packages are used infrequently (except for ggplot2), so you might delay installation until you need them. Note that you only need to install packages once on a given computer unless you upgrade the version of R. The standard GUIs for Windows and Mac provide menu-based methods for package installation and maintenance which you may find more convenient. The packages MASS, rpart, mgcv, nnet, splines and survival are part of the “recommended” R installation; you will have these already.

Any time you want to use these particular data or functions, you will need to load the package. For example,

```
library(faraway)
```

makes the data from this text available. If you get an error message about some data or functions not being found, it's quite possible that you have forgotten to load the necessary package. The `faraway` package is mostly data but does provide the `halfnorm`, `ilogit`, `logit` and `sumary` functions used in this text.

**Customization:** There are several ways to customize the look of R. I set the following options to achieve the output seen in this book:

```
options(width=70,digits=5,show.signif.stars=FALSE,scipen=2)
```

The `width=70` controls the width of the printed output. You may wish to increase this, depending on your screen or the width of your page. The `digits=5` reduces the number of digits shown when printing numbers from the default of seven. Note that this does not reduce the precision with which these numbers are internally stored. One might take this further — anything more than two or three significant digits in a displayed table is usually unnecessary and more importantly, distracting. `show.signif.stars=FALSE` prevents the printing of ugly significance stars in model summaries. `scipen=2` penalizes the use of scientific notation in reporting numbers. This notation has its uses but can make it harder to appreciate the size of a number. I have also sometimes edited the R output in the text to remove extraneous output or to improve the formatting.

**Updates and Errata:** The code and output shown in this book were generated under R version 3.2. R is regularly updated and improved so more recent versions may show some differences in the output. Sometimes these changes can break the code in the text. This is more likely to occur in the additional packages. Please refer to the book website at <http://people.bath.ac.uk/jjf23/ELM/> for information about such changes or errors found in the text.

**Interfaces:** The R website provides a GUI for both Mac and Windows installations. More functionality can be found elsewhere in ESS (which runs within the Emacs environment) and RStudio. Advanced users of R often find it more convenient to prepare R commands in a separate script before running these in R. Customizations to write R for many popular text editing programs are available.

**Reproducible Research:** It is important that you be able to reproduce your analysis in the future if needed. Sometimes, other people will need to verify your findings or, more likely, you yourself will need to modify or update your analysis. At a minimum you should maintain a commented set of R commands that will reproduce your analysis.

Sometimes you will want produce a document that includes the R commands, text describing the analysis and the output including graphs. This might look something like the text of this book. The `knitr` and `rmarkdown` packages provide this along with ability to use formats compatible with Microsoft Word and HTML for web display.

---

## 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 scatter-plots. *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  $\chi^2$  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 cytogenetic 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 Multi-level/Hierarchical Models*. Cambridge: Cambridge University Press.
- Gelman, A. and Y.-S. Su (2013). *arm: Data Analysis Using Regression and Multi-level/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 contingency 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). *dispmmod: 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 Statistical 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.

Since the publication of the bestselling, highly recommended first edition, R has considerably expanded both in popularity and in the number of packages available. **Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, Second Edition** takes advantage of the greater functionality now available in R and substantially revises and adds several topics.

### New to the Second Edition

- Expanded coverage of binary and binomial responses, including proportion responses, quasibinomial and beta regression and applied considerations regarding these models
- New sections on Poisson models with dispersion, zero inflated count models, linear discriminant analysis and sandwich and robust estimation for generalized linear models (GLMs)
- Revised chapters on random effects and repeated measures that reflect changes in the lme4 package and show how to perform hypothesis testing for the models using other methods
- New chapter on the Bayesian analysis of mixed effect models that illustrates the use of STAN and presents the approximation method of INLA
- Revised chapter on generalized linear mixed models to reflect the much richer choice of fitting software now available
- Updated coverage of splines and confidence bands in the chapter on nonparametric regression
- New material on random forests for regression and classification
- Revamped R code throughout, particularly the many plots using the ggplot2 package
- Revised and expanded exercises with solutions now included

This textbook continues to cover a range of techniques that grow from the linear regression model. It presents three extensions to the linear framework: GLMs, mixed effect models and nonparametric regression models. The book explains data analysis using real examples and includes all the R commands necessary to reproduce the analyses.