

Cambridge University Press

978-0-521-86116-8 - Data Analysis and Graphics Using R - an Example-Based Approach, Second Edition

John Maindonald and W. John Braun

Frontmatter

[More information](#)

---

## **Data Analysis and Graphics Using R, Second Edition**

Join the revolution ignited by the ground-breaking R system! Starting with an introduction to R, covering standard regression methods, then presenting more advanced topics, this book guides users through the practical and powerful tools that the R system provides. The emphasis is on hands-on analysis, graphical display and interpretation of data. The many worked examples, taken from real-world research, are accompanied by commentary on what is done and why. A website provides computer code and data sets, allowing readers to reproduce all analyses. Updates and solutions to selected exercises are also available. Assuming basic statistical knowledge and some experience of data analysis, the book is ideal for research scientists, final-year undergraduate or graduate level students of applied statistics, and practicing statisticians. It is both for learning and for reference.

This second edition reflects changes in R since 2003. There is new material on survival analysis, random coefficient models and the handling of high-dimensional data. The treatment of regression methods has been extended, including a brief discussion of errors in predictor variables. Both text and code have been revised throughout, and where possible simplified. New graphs have been added.

JOHN MAINDONALD is Visiting Fellow at the Centre for Mathematics and its Applications, Australian National University. He has collaborated extensively with scientists in a wide range of application areas, from medicine and public health to population genetics, machine learning, economic history and forensic linguistics.

JOHN BRAUN is Associate Professor of Statistical and Actuarial Sciences, University of Western Ontario. He has collaborated with biostatisticians, biologists, psychologists and most recently has become involved with a network of forestry researchers.

Cambridge University Press

978-0-521-86116-8 - Data Analysis and Graphics Using R - an Example-Based  
Approach, Second Edition

John Maindonald and W. John Braun

Frontmatter

[More information](#)

---

Data Analysis and Graphics  
Using R – an Example-Based Approach  
Second Edition

Cambridge University Press  
 978-0-521-86116-8 - Data Analysis and Graphics Using R - an Example-Based  
 Approach, Second Edition  
 John Maindonald and W. John Braun  
 Frontmatter  
[More information](#)

---

CAMBRIDGE SERIES IN STATISTICAL AND PROBABILISTIC  
 MATHEMATICS

---

*Editorial Board*

R. Gill (Department of Mathematics, Utrecht University)  
 B. D. Ripley (Department of Statistics, University of Oxford)  
 S. Ross (Department of Industrial & Systems Engineering, University of Southern  
 California)  
 B. W. Silverman (St. Peter's College, Oxford)  
 M. Stein (Department of Statistics, University of Chicago)

This series of high-quality upper-division textbooks and expository monographs covers all aspects of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization, and mathematical programming. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

Already published

1. *Bootstrap Methods and Their Application*, by A. C. Davison and D. V. Hinkley
2. *Markov Chains*, by J. Norris
3. *Asymptotic Statistics*, by A. W. van der Vaart
4. *Wavelet Methods for Time Series Analysis*, by Donald B. Percival and Andrew T. Walden
5. *Bayesian Methods*, by Thomas Leonard and John S. J. Hsu
6. *Empirical Processes in M-Estimation*, by Sara van de Geer
7. *Numerical Methods of Statistics*, by John F. Monahan
8. *A User's Guide to Measure Theoretic Probability*, by David Pollard
9. *The Estimation and Tracking of Frequency*, by B. G. Quinn and E. J. Hannan
10. *Data Analysis and Graphics using R*, by John Maindonald and W. John Braun
11. *Statistical Models*, by A. C. Davison
12. *Semiparametric Regression*, by D. Ruppert, M. P. Wand, R. J. Carroll
13. *Exercises in Probability*, by Loic Chaumont and Marc Yor
14. *Statistical Analysis of Stochastic Processes in Time*, by J. K. Lindsey
15. *Measure Theory and Filtering*, by Lakhdar Aggoun and Robert Elliott
16. *Essentials of Statistical Inference*, by G. A. Young and R. L. Smith
17. *Elements of Distribution Theory*, by Thomas A. Severini
18. *Statistical Mechanics of Disordered Systems*, by Anton Bovier
19. *The Coordinate-Free Approach to Linear Models*, by Michael J. Wichura
20. *Random Graph Dynamics*, by Rick Durrett

Cambridge University Press  
978-0-521-86116-8 - Data Analysis and Graphics Using R - an Example-Based  
Approach, Second Edition  
John Maindonald and W. John Braun  
Frontmatter  
[More information](#)

---

---

# Data Analysis and Graphics Using R – an Example-Based Approach

Second Edition

*John Maindonald*

Centre for Mathematics and its Applications, Australian National University

*and*

*W. John Braun*

Department of Statistical and Actuarial Science, University of Western Ontario



**CAMBRIDGE**  
UNIVERSITY PRESS

Cambridge University Press  
978-0-521-86116-8 - Data Analysis and Graphics Using R - an Example-Based  
Approach, Second Edition  
John Maindonald and W. John Braun  
Frontmatter  
[More information](#)

---

CAMBRIDGE UNIVERSITY PRESS  
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press  
The Edinburgh Building, Cambridge CB2 2RU, UK

Published in the United States of America by Cambridge University Press, New York

[www.cambridge.org](http://www.cambridge.org)  
Information on this title: [www.cambridge.org/9780521861168](http://www.cambridge.org/9780521861168)

© Cambridge University Press 2003  
© John Maindonald and W. John Braun 2007

This publication is in copyright. Subject to statutory exception  
and to the provisions of relevant collective licensing agreements,  
no reproduction of any part may take place without  
the written permission of Cambridge University Press.

First published 2003  
Second edition 2007

Printed in the United Kingdom at the University Press, Cambridge

*A catalogue record for this publication is available from the British Library*

ISBN-13 978-0-521-86116-8 hardback  
ISBN-10 0-521-86116-0 hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs  
for external or third-party internet websites referred to in this publication, and does not  
guarantee that any content on such websites is, or will remain, accurate or appropriate.

Cambridge University Press  
978-0-521-86116-8 - Data Analysis and Graphics Using R - an Example-Based  
Approach, Second Edition  
John Maindonald and W. John Braun  
Frontmatter  
[More information](#)

---

It is easy to lie with statistics. It is hard to tell the truth without statistics.

[Andrejs Dunkels]

... technology tends to overwhelm common sense.

[D. A. Freedman]

Cambridge University Press  
978-0-521-86116-8 - Data Analysis and Graphics Using R - an Example-Based  
Approach, Second Edition  
John Maindonald and W. John Braun  
Frontmatter  
[More information](#)

---

For Amelia and Luke  
also Shireen, Peter, Lorraine, Evan and Winifred

For Susan, Matthew and Phillip

Contents

|   |                 |
|---|-----------------|
| <b>Preface</b>  | <i>page xix</i> |
| <b>1 A brief introduction to R</b>                            | <b>1</b>        |
| 1.1 <i>An overview of R</i>                                   | 1               |
| 1.1.1 A short R session                                       | 1               |
| 1.1.2 The uses of R   | 5               |
| 1.1.3 Online help   | 6               |
| 1.1.4 Further steps in learning R                             | 8               |
| 1.2 <i>Data input, packages and the search list</i>           | 8               |
| 1.2.1 Reading data from a file                                | 8               |
| 1.2.2 R packages  | 9               |
| 1.3 <i>Vectors, factors and univariate time series</i>        | 10              |
| 1.3.1 Vectors in R  | 10              |
| 1.3.2 Concatenation – joining vector objects                  | 10              |
| 1.3.3 Subsets of vectors                                      | 11              |
| 1.3.4 Patterned data  | 11              |
| 1.3.5 Missing values  | 12              |
| 1.3.6 Factors   | 13              |
| 1.3.7 Time series   | 14              |
| 1.4 <i>Data frames and matrices</i>                           | 14              |
| 1.4.1 The attaching of data frames                            | 16              |
| 1.4.2 Aggregation, stacking and unstacking                    | 17              |
| 1.4.3* Data frames and matrices                               | 17              |
| 1.5 <i>Functions, operators and loops</i>                     | 18              |
| 1.5.1 Built-in functions                                      | 18              |
| 1.5.2 Generic functions and the class of an object            | 20              |
| 1.5.3 User-written functions                                  | 21              |
| 1.5.4 Relational and logical operators and operations         | 22              |
| 1.5.5 Selection and matching                                  | 23              |
| 1.5.6 Functions for working with missing values               | 23              |
| 1.5.7* Looping  | 24              |
| 1.6 <i>Graphics in R</i>                                      | 24              |
| 1.6.1 The function <code>plot ( )</code> and allied functions | 25              |
| 1.6.2 The use of color  | 27              |



|          |   |           |
|----------|---|-----------|
| 1.6.3    | The importance of aspect ratio  | 27        |
| 1.6.4    | Dimensions and other settings for graphics devices  | 28        |
| 1.6.5    | The plotting of expressions and mathematical symbols                                      | 28        |
| 1.6.6    | Identification and location on the figure region  | 29        |
| 1.6.7    | Plot methods for objects other than vectors   | 29        |
| 1.6.8    | Lattice graphics versus base graphics – <code>xypplot()</code> versus <code>plot()</code> | 30        |
| 1.6.9    | Further information on graphics   | 30        |
| 1.6.10   | Good and bad graphs   | 30        |
| 1.7      | <i>Lattice (trellis) graphics</i>   | 31        |
| 1.8      | <i>Additional points on the use of R</i>  | 33        |
| 1.9      | <i>Recap</i>  | 36        |
| 1.10     | <i>Further reading</i>  | 36        |
| 1.10.1   | References for further reading  | 37        |
| 1.11     | <i>Exercises</i>  | 37        |
| <b>2</b> | <b>Styles of data analysis</b>  | <b>43</b> |
| 2.1      | <i>Revealing views of the data</i>  | 43        |
| 2.1.1    | Views of a single sample  | 44        |
| 2.1.2    | Patterns in univariate time series  | 48        |
| 2.1.3    | Patterns in bivariate data  | 50        |
| 2.1.4    | Patterns in grouped data  | 52        |
| 2.1.5*   | Multiple variables and times  | 54        |
| 2.1.6    | Scatterplots, broken down by multiple factors   | 56        |
| 2.1.7    | What to look for in plots   | 58        |
| 2.2      | <i>Data summary</i>   | 59        |
| 2.2.1    | Counts  | 60        |
| 2.2.2    | Summaries of information from data frames   | 63        |
| 2.2.3    | Standard deviation and inter-quartile range   | 66        |
| 2.2.4    | Correlation   | 68        |
| 2.3      | <i>Statistical analysis questions, aims and strategies</i>                                | 69        |
| 2.3.1    | How relevant and how reliable are the data?   | 70        |
| 2.3.2    | Helpful and unhelpful questions   | 70        |
| 2.3.3    | How will results be used?   | 71        |
| 2.3.4    | Formal and informal assessments   | 72        |
| 2.3.5    | Statistical analysis strategies   | 73        |
| 2.3.6    | Planning the formal analysis  | 73        |
| 2.3.7    | Changes to the intended plan of analysis  | 74        |
| 2.4      | <i>Recap</i>  | 74        |
| 2.5      | <i>Further reading</i>  | 75        |
| 2.5.1    | References for further reading  | 75        |
| 2.6      | <i>Exercises</i>  | 75        |
| <b>3</b> | <b>Statistical models</b>   | <b>78</b> |
| 3.1      | <i>Regularities</i>   | 79        |
| 3.1.1    | Deterministic models  | 79        |

| <i>Contents</i>  | xi         |
|--|------------|
| 3.1.2 Models that include a random component               | 79         |
| 3.1.3 Fitting models – the model formula                   | 82         |
| 3.2 <i>Distributions: models for the random component</i>  | 83         |
| 3.2.1 Discrete distributions                               | 84         |
| 3.2.2 Continuous distributions                             | 86         |
| 3.3 <i>The uses of random numbers</i>                      | 88         |
| 3.3.1 Simulation   | 88         |
| 3.3.2 Sampling from populations                            | 89         |
| 3.4 <i>Model assumptions</i>                               | 90         |
| 3.4.1 Random sampling assumptions – independence           | 91         |
| 3.4.2 Checks for normality                                 | 92         |
| 3.4.3 Checking other model assumptions                     | 95         |
| 3.4.4 Are non-parametric methods the answer?               | 95         |
| 3.4.5 Why models matter – adding across contingency tables | 95         |
| 3.5 <i>Recap</i>   | 96         |
| 3.6 <i>Further reading</i>                                 | 97         |
| 3.6.1 References for further reading                       | 97         |
| 3.7 <i>Exercises</i>                                       | 97         |
| <b>4 An introduction to formal inference</b>               | <b>101</b> |
| 4.1 <i>Basic concepts of estimation</i>                    | 101        |
| 4.1.1 Population parameters and sample statistics          | 101        |
| 4.1.2 Sampling distributions                               | 102        |
| 4.1.3 Assessing accuracy – the standard error              | 102        |
| 4.1.4 The standard error for the difference of means       | 103        |
| 4.1.5* The standard error of the median                    | 104        |
| 4.1.6 The sampling distribution of the <i>t</i> -statistic | 104        |
| 4.2 <i>Confidence intervals and hypothesis tests</i>       | 107        |
| 4.2.1 One- and two-sample intervals and tests for means    | 107        |
| 4.2.2 Confidence intervals and tests for proportions       | 113        |
| 4.2.3 Confidence intervals for the correlation             | 113        |
| 4.2.4 Confidence intervals versus hypothesis tests         | 114        |
| 4.3 <i>Contingency tables</i>                              | 115        |
| 4.3.1 Rare and endangered plant species                    | 117        |
| 4.3.2 Additional notes                                     | 119        |
| 4.4 <i>One-way unstructured comparisons</i>                | 120        |
| 4.4.1 Displaying means for the one-way layout              | 123        |
| 4.4.2 Multiple comparisons                                 | 124        |
| 4.4.3 Data with a two-way structure, that is, two factors  | 125        |
| 4.4.4 Presentation issues                                  | 126        |
| 4.5 <i>Response curves</i>                                 | 126        |
| 4.6 <i>Data with a nested variation structure</i>          | 127        |
| 4.6.1 Degrees of freedom considerations                    | 128        |
| 4.6.2 General multi-way analysis of variance designs       | 129        |

|          |   |            |
|----------|---|------------|
| xii      | <i>Contents</i>   |            |
| 4.7      | <i>Resampling methods for standard errors, tests and confidence intervals</i> | 129        |
| 4.7.1    | The one-sample permutation test   | 129        |
| 4.7.2    | The two-sample permutation test   | 130        |
| 4.7.3*   | Estimating the standard error of the median: bootstrapping                    | 131        |
| 4.7.4    | Bootstrap estimates of confidence intervals                                   | 133        |
| 4.8*     | <i>Theories of inference</i>  | 134        |
| 4.8.1    | Maximum likelihood estimation   | 135        |
| 4.8.2    | Bayesian estimation   | 136        |
| 4.8.3    | If there is strong prior information, use it!                                 | 136        |
| 4.9      | <i>Recap</i>  | 137        |
| 4.10     | <i>Further reading</i>  | 138        |
| 4.10.1   | References for further reading  | 138        |
| 4.11     | <i>Exercises</i>  | 139        |
| <b>5</b> | <b>Regression with a single predictor</b>                                     | <b>144</b> |
| 5.1      | <i>Fitting a line to data</i>   | 144        |
| 5.1.1    | Lawn roller example   | 145        |
| 5.1.2    | Calculating fitted values and residuals                                       | 146        |
| 5.1.3    | Residual plots  | 147        |
| 5.1.4    | Iron slag example: is there a pattern in the residuals?                       | 148        |
| 5.1.5    | The analysis of variance table  | 150        |
| 5.2      | <i>Outliers, influence and robust regression</i>                              | 151        |
| 5.3      | <i>Standard errors and confidence intervals</i>                               | 153        |
| 5.3.1    | Confidence intervals and tests for the slope                                  | 153        |
| 5.3.2    | SEs and confidence intervals for predicted values                             | 154        |
| 5.3.3*   | Implications for design   | 155        |
| 5.4      | <i>Regression versus qualitative anova comparisons</i>                        | 157        |
| 5.4.1    | Issues of power   | 157        |
| 5.4.2    | The pattern of change   | 158        |
| 5.5      | <i>Assessing predictive accuracy</i>  | 158        |
| 5.5.1    | Training/test sets and cross-validation                                       | 158        |
| 5.5.2    | Cross-validation – an example   | 159        |
| 5.5.3*   | Bootstrapping   | 161        |
| 5.6*     | <i>A note on power transformations</i>  | 164        |
| 5.6.1*   | General power transformations   | 164        |
| 5.7      | <i>Size and shape data</i>  | 165        |
| 5.7.1    | Allometric growth   | 166        |
| 5.7.2    | There are two regression lines!   | 167        |
| 5.8      | <i>The model matrix in regression</i>   | 168        |
| 5.9      | <i>Recap</i>  | 169        |
| 5.10     | <i>Methodological references</i>  | 170        |
| 5.11     | <i>Exercises</i>  | 170        |

|          |  |            |
|----------|--|------------|
| <b>6</b> | <b>Multiple linear regression</b>                                      | <b>173</b> |
| 6.1      | <i>Basic ideas: book weight and brain weight examples</i>              | 173        |
| 6.1.1    | Omission of the intercept term   | 176        |
| 6.1.2    | Diagnostic plots   | 176        |
| 6.1.3    | Example: brain weight  | 178        |
| 6.1.4    | Plots that show the contribution of individual terms                   | 180        |
| 6.2      | <i>Multiple regression assumptions and diagnostics</i>                 | 182        |
| 6.2.1    | Influential outliers and Cook’s distance                               | 183        |
| 6.2.2    | Influence on the regression coefficients                               | 184        |
| 6.2.3*   | Additional diagnostic plots  | 185        |
| 6.2.4    | Robust and resistant methods   | 185        |
| 6.2.5    | The uses of model diagnostics  | 185        |
| 6.3      | <i>A strategy for fitting multiple regression models</i>               | 186        |
| 6.3.1    | Preliminaries  | 186        |
| 6.3.2    | Model fitting  | 187        |
| 6.3.3    | An example – the Scottish hill race data                               | 187        |
| 6.4      | <i>Measures for the assessment and comparison of regression models</i> | 193        |
| 6.4.1    | $R^2$ and adjusted $R^2$   | 193        |
| 6.4.2    | AIC and related statistics   | 194        |
| 6.4.3    | How accurately does the equation predict?                              | 194        |
| 6.5      | <i>Interpreting regression coefficients</i>                            | 196        |
| 6.5.1    | Book dimensions and book weight  | 196        |
| 6.6      | <i>Problems with many explanatory variables</i>                        | 199        |
| 6.6.1    | Variable selection issues  | 200        |
| 6.7      | <i>Multicollinearity</i>   | 202        |
| 6.7.1    | A contrived example  | 202        |
| 6.7.2    | The variance inflation factor  | 206        |
| 6.7.3    | Remedies for multicollinearity   | 206        |
| 6.8      | <i>Multiple regression models – additional points</i>                  | 207        |
| 6.8.1    | Errors in $x$  | 207        |
| 6.8.2    | Confusion between explanatory and response variables                   | 210        |
| 6.8.3    | Missing explanatory variables  | 210        |
| 6.8.4*   | The use of transformations   | 212        |
| 6.8.5*   | Non-linear methods – an alternative to transformation?                 | 212        |
| 6.9      | <i>Recap</i>   | 214        |
| 6.10     | <i>Further reading</i>   | 214        |
| 6.10.1   | References for further reading   | 215        |
| 6.11     | <i>Exercises</i>   | 216        |
| <b>7</b> | <b>Exploiting the linear model framework</b>                           | <b>219</b> |
| 7.1      | <i>Levels of a factor – using indicator variables</i>                  | 220        |
| 7.1.1    | Example – sugar weight   | 220        |
| 7.1.2    | Different choices for the model matrix when there are factors          | 223        |

|          |   |            |
|----------|---|------------|
| xiv      | <i>Contents</i>   |            |
| 7.2      | <i>Block designs and balanced incomplete block designs</i>        | 224        |
| 7.2.1    | Analysis of the rice data, allowing for block effects             | 224        |
| 7.2.2    | A balanced incomplete block design                                | 226        |
| 7.3      | <i>Fitting multiple lines</i>                                     | 227        |
| 7.4      | <i>Polynomial regression</i>                                      | 231        |
| 7.4.1    | Issues in the choice of model                                     | 233        |
| 7.5*     | <i>Methods for passing smooth curves through data</i>             | 234        |
| 7.5.1    | Scatterplot smoothing – regression splines                        | 235        |
| 7.5.2*   | Penalized splines and generalized additive models                 | 239        |
| 7.5.3    | Other smoothing methods   | 239        |
| 7.6      | <i>Smoothing terms in additive models</i>                         | 241        |
| 7.6.1*   | The fitting of penalized spline terms                             | 243        |
| 7.7      | <i>Further reading</i>  | 243        |
| 7.7.1    | References for further reading                                    | 243        |
| 7.8      | <i>Exercises</i>  | 243        |
| <b>8</b> | <b>Generalized linear models and survival analysis</b>            | <b>246</b> |
| 8.1      | <i>Generalized linear models</i>                                  | 246        |
| 8.1.1    | Transformation of the expected value on the left                  | 246        |
| 8.1.2    | Noise terms need not be normal                                    | 247        |
| 8.1.3    | Log odds in contingency tables                                    | 247        |
| 8.1.4    | Logistic regression with a continuous explanatory variable        | 248        |
| 8.2      | <i>Logistic multiple regression</i>                               | 251        |
| 8.2.1    | Selection of model terms and fitting the model                    | 253        |
| 8.2.2    | A plot of contributions of explanatory variables                  | 256        |
| 8.2.3    | Cross-validation estimates of predictive accuracy                 | 257        |
| 8.3      | <i>Logistic models for categorical data – an example</i>          | 258        |
| 8.4      | <i>Poisson and quasi-Poisson regression</i>                       | 260        |
| 8.4.1    | Data on aberrant crypt foci                                       | 260        |
| 8.4.2    | Moth habitat example  | 263        |
| 8.5      | <i>Additional notes on generalized linear models</i>              | 269        |
| 8.5.1*   | Residuals, and estimating the dispersion                          | 269        |
| 8.5.2    | Standard errors and z- or t-statistics for binomial models        | 270        |
| 8.5.3    | Leverage for binomial models                                      | 270        |
| 8.6      | <i>Models with an ordered categorical or categorical response</i> | 271        |
| 8.6.1    | Ordinal regression models   | 271        |
| 8.6.2*   | Loglinear models  | 274        |
| 8.7      | <i>Survival analysis</i>  | 275        |
| 8.7.1    | Analysis of the Aids2 data  | 276        |
| 8.7.2    | Right censoring prior to the termination of the study             | 278        |
| 8.7.3    | The survival curve for male homosexuals                           | 279        |
| 8.7.4    | Hazard rates  | 279        |
| 8.7.5    | The Cox proportional hazards model                                | 280        |
| 8.8      | <i>Transformations for count data</i>                             | 282        |
| 8.9      | <i>Further reading</i>  | 283        |

|           |  |            |
|-----------|--|------------|
| 8.9.1     | References for further reading                                     | 283        |
| 8.10      | <i>Exercises</i>   | 284        |
| <b>9</b>  | <b>Time series models</b>  | <b>286</b> |
| 9.1       | <i>Time series – some basic ideas</i>                              | 286        |
| 9.1.1     | Preliminary graphical explorations                                 | 286        |
| 9.1.2     | The autocorrelation function                                       | 287        |
| 9.1.3     | Autoregressive models  | 288        |
| 9.1.4*    | Autoregressive moving average models – theory                      | 290        |
| 9.2*      | <i>Regression modeling with moving average errors</i>              | 291        |
| 9.3*      | <i>Non-linear time series</i>                                      | 297        |
| 9.4       | <i>Other time series packages</i>                                  | 298        |
| 9.5       | <i>Further reading</i>   | 298        |
| 9.5.1     | Spatial statistics   | 299        |
| 9.5.2     | References for further reading                                     | 299        |
| 9.6       | <i>Exercises</i>   | 299        |
| <b>10</b> | <b>Multi-level models and repeated measures</b>                    | <b>301</b> |
| 10.1      | <i>A one-way random effects model</i>                              | 302        |
| 10.1.1    | Analysis with <code>aoV()</code>                                   | 303        |
| 10.1.2    | A more formal approach   | 306        |
| 10.1.3    | Analysis using <code>lmer()</code>                                 | 308        |
| 10.2      | <i>Survey data, with clustering</i>                                | 311        |
| 10.2.1    | Alternative models   | 311        |
| 10.2.2    | Instructive, though faulty, analyses                               | 316        |
| 10.2.3    | Predictive accuracy  | 317        |
| 10.3      | <i>A multi-level experimental design</i>                           | 317        |
| 10.3.1    | The anova table  | 319        |
| 10.3.2    | Expected values of mean squares                                    | 320        |
| 10.3.3*   | The sums of squares breakdown                                      | 321        |
| 10.3.4    | The variance components  | 324        |
| 10.3.5    | The mixed model analysis   | 325        |
| 10.3.6    | Predictive accuracy  | 327        |
| 10.3.7    | Different sources of variance – complication or focus of interest? | 327        |
| 10.4      | <i>Within- and between-subject effects</i>                         | 328        |
| 10.4.1    | Model selection  | 329        |
| 10.4.2    | Estimates of model parameters                                      | 330        |
| 10.5      | <i>Repeated measures in time</i>                                   | 332        |
| 10.5.1    | Example – random variation between profiles                        | 334        |
| 10.5.2    | Orthodontic measurements on children                               | 339        |
| 10.6      | <i>Error structure considerations</i>                              | 343        |
| 10.6.1    | Predictions from models with a complex error structure             | 343        |
| 10.6.2    | Error structure in explanatory variables                           | 344        |

|           |   |            |
|-----------|---|------------|
| 10.7      | <i>Further notes on multi-level and other models with correlated errors</i> | 344        |
| 10.7.1    | An historical perspective on multi-level models                             | 344        |
| 10.7.2    | Meta-analysis   | 346        |
| 10.7.3    | Functional data analysis  | 346        |
| 10.8      | <i>Recap</i>  | 346        |
| 10.9      | <i>Further reading</i>  | 347        |
| 10.9.1    | References for further reading  | 347        |
| 10.10     | <i>Exercises</i>  | 348        |
| <b>11</b> | <b>Tree-based classification and regression</b>                             | <b>350</b> |
| 11.1      | <i>The uses of tree-based methods</i>                                       | 351        |
| 11.1.1    | Problems for which tree-based regression may be used                        | 351        |
| 11.2      | <i>Detecting email spam – an example</i>                                    | 352        |
| 11.2.1    | Choosing the number of splits   | 355        |
| 11.3      | <i>Terminology and methodology</i>  | 355        |
| 11.3.1    | Choosing the split – regression trees                                       | 355        |
| 11.3.2    | Within and between sums of squares  | 356        |
| 11.3.3    | Choosing the split – classification trees                                   | 357        |
| 11.3.4    | Tree-based regression versus loess regression smoothing                     | 358        |
| 11.4      | <i>Predictive accuracy and the cost–complexity tradeoff</i>                 | 360        |
| 11.4.1    | Cross-validation  | 361        |
| 11.4.2    | The cost–complexity parameter   | 361        |
| 11.4.3    | Prediction error versus tree size   | 362        |
| 11.5      | <i>Data for female heart attack patients</i>                                | 363        |
| 11.5.1    | The one-standard-deviation rule   | 365        |
| 11.5.2    | Printed information on each split   | 365        |
| 11.6      | <i>Detecting email spam – the optimal tree</i>                              | 366        |
| 11.7      | <i>The randomForest package</i>   | 368        |
| 11.8      | <i>Additional notes on tree-based methods</i>                               | 371        |
| 11.8.1    | The combining of tree-based methods with other approaches                   | 371        |
| 11.8.2    | Models with a complex error structure                                       | 372        |
| 11.8.3    | Pruning as variable selection   | 372        |
| 11.8.4    | Other types of tree   | 372        |
| 11.8.5    | Factors as predictors   | 372        |
| 11.8.6    | Summary of pluses and minuses of tree-based methods                         | 372        |
| 11.9      | <i>Further reading</i>  | 373        |
| 11.9.1    | References for further reading  | 373        |
| 11.10     | <i>Exercises</i>  | 374        |
| <b>12</b> | <b>Multivariate data exploration and discrimination</b>                     | <b>375</b> |
| 12.1      | <i>Multivariate exploratory data analysis</i>                               | 376        |
| 12.1.1    | Scatterplot matrices  | 376        |
| 12.1.2    | Principal components analysis   | 377        |
| 12.1.3    | Multi-dimensional scaling   | 383        |

|           | <i>Contents</i>  | xvii       |
|-----------|--|------------|
| 12.2      | <i>Discriminant analysis</i>   | 384        |
| 12.2.1    | Example – plant architecture   | 384        |
| 12.2.2    | Logistic discriminant analysis   | 386        |
| 12.2.3    | Linear discriminant analysis   | 387        |
| 12.2.4    | An example with more than two groups                                     | 388        |
| 12.3*     | <i>High-dimensional data, classification and plots</i>                   | 390        |
| 12.3.1    | Classifications and associated graphs                                    | 392        |
| 12.3.2    | Flawed graphs  | 393        |
| 12.3.3    | Accuracies and scores for test data                                      | 397        |
| 12.3.4    | Graphs derived from the cross-validation process                         | 403        |
| 12.4      | <i>Further reading</i>   | 405        |
| 12.4.1    | References for further reading   | 406        |
| 12.5      | <i>Exercises</i>   | 406        |
| <b>13</b> | <b>Regression on principal component or discriminant scores</b>          | <b>408</b> |
| 13.1      | <i>Principal component scores in regression</i>                          | 408        |
| 13.2*     | <i>Propensity scores in regression comparisons – labor training data</i> | 412        |
| 13.2.1    | Regression analysis, using all covariates                                | 415        |
| 13.2.2    | The use of propensity scores   | 417        |
| 13.3      | <i>Further reading</i>   | 419        |
| 13.3.1    | References for further reading   | 419        |
| 13.4      | <i>Exercises</i>   | 420        |
| <b>14</b> | <b>The R system – additional topics</b>                                  | <b>421</b> |
| 14.1      | <i>Working directories, workspaces and the search list</i>               | 421        |
| 14.1.1*   | The search path  | 421        |
| 14.1.2    | Workspace management   | 421        |
| 14.1.3    | Utility functions  | 423        |
| 14.2      | <i>Data input and output</i>   | 423        |
| 14.2.1    | Input of data  | 424        |
| 14.2.2    | Data output  | 428        |
| 14.3      | <i>Functions and operators – some further details</i>                    | 429        |
| 14.3.1    | Function arguments   | 430        |
| 14.3.2    | Character string and vector functions                                    | 431        |
| 14.3.3    | Anonymous functions  | 431        |
| 14.3.4    | Functions for working with dates (and times)                             | 432        |
| 14.3.5    | Creating groups  | 433        |
| 14.3.6    | Logical operators  | 434        |
| 14.4      | <i>Factors</i>   | 434        |
| 14.5      | <i>Missing values</i>  | 437        |
| 14.6*     | <i>Matrices and arrays</i>   | 439        |
| 14.6.1    | Matrix arithmetic  | 440        |
| 14.6.2    | Outer products   | 441        |
| 14.6.3    | Arrays   | 442        |



|          |   |            |
|----------|---|------------|
| 14.7     | <i>Manipulations with lists, data frames and matrices</i>           | 443        |
| 14.7.1   | Lists – an extension of the notion of “vector”                      | 443        |
| 14.7.2   | Changing the shape of data frames                                   | 445        |
| 14.7.3*  | Merging data frames – <code>merge()</code>                          | 445        |
| 14.7.4   | Joining data frames, matrices and vectors – <code>cbind()</code>    | 446        |
| 14.7.5   | The <code>apply</code> family of functions                          | 446        |
| 14.7.6   | Splitting vectors and data frames into lists – <code>split()</code> | 448        |
| 14.7.7   | Multivariate time series  | 448        |
| 14.8     | <i>Classes and methods</i>  | 449        |
| 14.8.1   | Printing and summarizing model objects                              | 449        |
| 14.8.2   | Extracting information from model objects                           | 450        |
| 14.8.3   | S4 classes and methods  | 450        |
| 14.9     | <i>Manipulation of language constructs</i>                          | 451        |
| 14.9.1   | Model and graphics formulae   | 451        |
| 14.9.2   | The use of a list to pass parameter values                          | 452        |
| 14.9.3   | Expressions   | 453        |
| 14.9.4   | Environments  | 453        |
| 14.9.5   | Function environments and lazy evaluation                           | 455        |
| 14.10    | <i>Document preparation — Sweave()</i>                              | 456        |
| 14.11    | <i>Graphs in R</i>  | 457        |
| 14.11.1  | Hardcopy graphics devices   | 457        |
| 14.11.2  | Multiple graphs on a single graphics page                           | 457        |
| 14.11.3  | Plotting characters, symbols, line types and colors                 | 457        |
| 14.12    | <i>Lattice graphics and the grid package</i>                        | 462        |
| 14.12.1  | Interaction with plots  | 464        |
| 14.12.2* | Use of <code>grid.text()</code> to label points                     | 464        |
| 14.12.3* | Multiple lattice graphs on a graphics page                          | 465        |
| 14.13    | <i>Further reading</i>  | 466        |
| 14.13.1  | Vignettes   | 466        |
| 14.13.2  | References for further reading                                      | 466        |
| 14.14    | <i>Exercises</i>  | 467        |
|          | <b>Epilogue – models</b>  | <b>470</b> |
|          | <b>References</b>   | <b>474</b> |
|          | <b>Index of R Symbols and Functions</b>                             | <b>485</b> |
|          | <b>Index of Terms</b>   | <b>491</b> |
|          | <b>Index of Authors</b>   | <b>501</b> |
|          | <b>Color Plates after Page 502</b>                                  |            |

Cambridge University Press

978-0-521-86116-8 - Data Analysis and Graphics Using R - an Example-Based Approach, Second Edition

John Maindonald and W. John Braun

Frontmatter

[More information](#)

---

## Preface

This book is an exposition of statistical methodology that focuses on ideas and concepts, and makes extensive use of graphical presentation. It avoids, as much as possible, the use of mathematical symbolism. It is particularly aimed at scientists who wish to do statistical analyses on their own data, preferably with reference as necessary to professional statistical advice. It is intended to complement more mathematically oriented accounts of statistical methodology. It may be used to give students with a more specialist statistical interest exposure to practical data analysis.

While no prior knowledge of specific statistical methods or theory is assumed, there is a demand that readers bring with them, or quickly acquire, some modest level of statistical sophistication. Readers should have some prior exposure to statistical methodology, some prior experience of working with real data, and be comfortable with the typing of analysis commands into the computer console. Some prior familiarity with regression and with analysis of variance will be helpful.

We cover a range of topics that are important for many different areas of statistical application. As is inevitable in a book that has this broad focus, there will be investigators working in specific areas – perhaps epidemiology, or psychology, or sociology, or ecology – who will regret the omission of some methodologies that they find important.

We comment extensively on analysis results, noting inferences that seem well-founded, and noting limitations on inferences that can be drawn. We emphasize the use of graphs for gaining insight into data – in advance of any formal analysis, for understanding the analysis, and for presenting analysis results.

The data sets that we use as a vehicle for demonstrating statistical methodology have been generated by researchers in many different fields, and have in many cases featured in published papers. As far as possible, our account of statistical methodology comes from the coalface, where the quirks of real data must be faced and addressed. Features that may challenge the novice data analyst have been retained. The diversity of examples has benefits, even for those whose interest is in a specific application area. Ideas and applications that are useful in one area often find use elsewhere, even to the extent of stimulating new lines of investigation. We hope that our book will stimulate such cross-fertilization.

To summarize: the strengths of this book include the directness of its encounter with research data, its advice on practical data analysis issues, the inclusion of code that reproduces analyses, careful critiques of analysis results, attention to graphical and other

Cambridge University Press

978-0-521-86116-8 - Data Analysis and Graphics Using R - an Example-Based Approach, Second Edition

John Maindonald and W. John Braun

Frontmatter

[More information](#)

xx

*Preface*

presentation issues, and the use of examples drawn from across the range of statistical applications.

John Braun wrote the initial drafts of Subsections 4.7.3, 4.7.4, 5.5.3, 6.8.5, 8.4.1 and Section 9.3. Initial drafts of remaining material were, mostly, from John Maindonald's hand. A substantial part was derived, initially, from the lecture notes of courses for researchers, at the University of Newcastle (Australia) over 1996–1997 and at The Australian National University over 1998–2001. Both of us have worked extensively over the material in these chapters. John Braun has taken primary responsibility for maintenance of the *DAAG* package.

**The R system**

We use the R system for the computations. The R system implements a dialect of the influential S language, developed at AT&T Bell Laboratories by Rick Becker, John Chambers and Allan Wilks, which is the basis for the commercial S-PLUS system. It follows S in its close linkage between data analysis and graphics. Versions of R are available, at no charge, for 32-bit versions of Microsoft Windows, for Linux and other Unix systems, and for the Macintosh. It is available through the Comprehensive R Archive Network (CRAN). Go to <http://cran.r-project.org/>, and find the nearest mirror site.

The development model used for R has proved highly effective in marshalling high levels of computing expertise for continuing improvement, for identifying and fixing bugs, and for responding quickly to the evolving needs and interests of the statistical community. Oversight of “base R” is handled by the R Core Team, whose members are widely drawn internationally. Use is made of code, bug fixes and documentation from the wider R user community. Especially important are the large number of packages that supplement base R, and that anyone is free to contribute. Once installed, these attach seamlessly into the base system.

Many of the analyses offered by R's packages were not, 10 years ago, available in any of the standard statistical packages. What did data analysts do before we had such packages? Basically, they adapted more simplistic (but not necessarily simpler) analyses as best they could. Those whose skills were unequal to the task did unsatisfactory analyses. Those with more adequate skills carried out analyses that, even if not elegant and insightful by current standards, were often adequate. Tools such as are available in R have reduced the need for the adaptations that were formerly necessary. We can often do analyses that better reflect the underlying science. There have been challenging and exciting changes from the methodology that was typically encountered in statistics courses 10 or 15 years ago.

In the ongoing development of R, priorities have been: the provision of good data manipulation abilities; flexible and high-quality graphics; the provision of data analysis methods that are both insightful and adequate for the whole range of application area demands; seamless integration of the different components of R; and the provision of interfaces to other systems (editors, databases, the web, etc.) that R users may require.

Cambridge University Press

978-0-521-86116-8 - Data Analysis and Graphics Using R - an Example-Based Approach, Second Edition

John Maindonald and W. John Braun

Frontmatter

[More information](#)

Ease of use is important, but not at the expense of power, flexibility and checks against answers that are potentially misleading.

Depending on the user's level of skill with R, there will be some relatively routine tasks where another system may seem simpler to use. Note however the availability of interfaces, notably John Fox's *Rcmdr*, that give a graphical user interface (GUI) to a limited part of R. Such interfaces will develop and improve as time progresses. They may in due course, for many users, be the preferred means of access to R. Be aware that the demand for simple tools will commonly place limitations on the tasks that can, without professional assistance, be satisfactorily undertaken.

Primarily, R is designed for scientific computing and for graphics. Among the packages that have been added are many that are not obviously statistical – for drawing and coloring maps, for map projections, for plotting data collected by balloon-born weather instruments, for creating color palettes, for working with bitmap images, for solving sudoku puzzles, for creating magic squares, for reading and handling shapefiles, for solving ordinary differential equations, for processing various types of genomic data, and so on. Check through the list of R packages that can be found on any of the CRAN sites, and you may be surprised at what you find!

The citation for John Chambers' 1998 Association for Computing Machinery Software award stated that S has "forever altered how people analyze, visualize and manipulate data." The R project enlarges on the ideas and insights that generated the S language. We are grateful to the R Core Team, and to the creators of the various R packages, for bringing into being the R system – this marvellous tool for scientific and statistical computing, and for graphical presentation. We list at the end of the reference section the authors and compilers of packages that have been used in this book.

### **Influences on the modern practice of statistics**

The development of statistics has been motivated by the demands of scientists for a methodology that will extract patterns from their data. The methodology has developed in a synergy with the relevant supporting mathematical theory and, more recently, with computing. This has led to methodologies and supporting theory that are a radical departure from the methodologies of the pre-computer era.

Statistics is a young discipline. Only in the 1920s and 1930s did the modern framework of statistical theory, including ideas of hypothesis testing and estimation, begin to take shape. Different areas of statistical application have taken these ideas up in different ways, some of them starting their own separate streams of statistical tradition. Gigerenzer *et al.* (1989, "The Empire of Statistics") examine the history, commenting on the different streams of development that have influenced practice in different research areas.

Separation from the statistical mainstream, and an emphasis on "black box" approaches, have contributed to a widespread exaggerated emphasis on tests of hypotheses, to a neglect of pattern, to the policy of some journal editors of publishing only those studies that show a statistically significant effect, and to an undue focus on the individual study. Anyone

who joins the R community can expect to witness, and/or engage in, lively debate that addresses these and related issues. Such debate can help ensure that the demands of scientific rationality do in due course win out over influences from accidents of historical development.

### **New tools for effective data analysis**

We have drawn attention to advances in statistical computing methodology. These have led to new powerful tools for exploratory analysis of regression data, for choosing between alternative models, for diagnostic checks, for handling non-linearity, for assessing the predictive power of models, and for graphical presentation. In addition, we have new computing tools that make it straightforward to move data between different systems, to keep a record of calculations, to retrace or adapt earlier calculations, and to edit output and graphics into a form that can be incorporated into published documents.

The best any analysis can do is to highlight the information in the data. No amount of statistical or computing technology can be a substitute for good design of data collection, for understanding the context in which data are to be interpreted, or for skill in the use of statistical analysis methodology. Statistical software systems are one of several components of effective data analysis.

The questions that statistical analysis is designed to answer can often be stated simply. This may encourage the layperson to believe that the answers are similarly simple. Often, they are not. Be prepared for unexpected subtleties. Effective statistical analysis requires appropriate skills, beyond those gained from taking one or two undergraduate courses in statistics. There is no good substitute for professional training in modern tools for data analysis, and experience in using those tools with a wide range of data sets. No-one should be embarrassed that they have difficulty with analyses that involve ideas that professional statisticians may take 7 or 8 years of professional training and experience to master.

### **Changes in this second edition**

This new edition takes account of changes in R since 2003. There is new material on survival analysis, random coefficient models and the handling of high-dimensional data. The treatment of regression methods has been extended, including in particular a brief discussion of errors in predictor variables. Both the text and R code have been extensively revised. Code has, wherever possible, been simplified. Some examples have been reworked. There are changes to some graphs, and new graphs have been added.

### **Acknowledgments**

Many different people have helped us with this project. Winfried Theis (University of Dortmund, Germany) and Detlef Steuer (University of the Federal Armed Forces, Hamburg, Germany) helped with technical aspects of working with  $\text{\LaTeX}$ , with setting up a cvs server to manage the  $\text{\LaTeX}$  files, and with helpful comments. Lynne Billard

(University of Georgia, USA), Murray Jorgensen (University of Waikato, NZ) and Berwin Turlach (University of Western Australia) gave valuable help in the identification of errors and text that required clarification. Susan Wilson (Australian National University) gave welcome encouragement. Duncan Murdoch (University of Western Ontario) helped set up the *DAAG* package, and has supplied valuable technical advice. Thanks also to Cath Lawrence (Australian National University) for her Python program that allowed us to extract the R code, as and when required, from our  $\text{\LaTeX}$  files; this has now at length become an R function. Many of the tables in this book were generated, in first draft form, using the `xtable()` function from the *xtable* package for R.

For this second edition, Brian Ripley (University of Oxford) has gone through the manuscript and made extensive comments, leading to important corrections and improvements. We are most grateful to him, and to others who have commented on the manuscript. Alan Welsh (Australian National University) has been helpful in working through points where it has seemed difficult to get the emphasis right. Once again, Duncan Murdoch has given much useful technical advice. Others who have made helpful comments and/or pointed out errors include Jeff Wood (Australian National University), Nader Tajvidi (University of Lund), Paul Murrell (University of Auckland, on Section 14.11), Graham Williams (<http://www.togaware.com>, on Chapter 1) and Yang Yang (University of Western Ontario, on Chapter 10). The failings that remain are, naturally, our responsibility.

A strength of this book is the extent to which it has drawn on data from many different sources. We give a list, following the list of references for the data near the end of the book, of individuals and/or organizations to whom we are grateful for allowing use of data. We are grateful to those who have allowed us to use their data. At least these data will not, as often happens once data have become the basis for a published paper, gather dust in a long-forgotten folder! We are grateful, also, to the many researchers who, in their discussions with us, have helped stimulate our thinking and understanding. We apologize if there is anyone that we have inadvertently failed to acknowledge.

Diana Gillooly of Cambridge University Press, taking over from David Tranah for this new edition, has been a marvellous source of advice and encouragement throughout the revision process.

### Conventions

Text that is R code, or output from R, is printed in a verbatim text style. For example, in Chapter 1 we will enter data into an R object that we call `austpop`. We will use the `plot()` function to plot these data. The names of R packages, including our own *DAAG* package, are printed in italics.

Starred exercises and sections identify more technical items that can be skipped at a first reading.

### Solutions to exercises

Solutions to selected exercises, R scripts that have all the code from the book and other supplementary materials are available via the link given at <http://www.maths.anu.edu.au/~johnm/r-book>