

Cambridge University Press

978-0-521-76293-9 - Data Analysis and Graphics Using R – an Example-Based Approach: Third Edition

John Maindonald and W. John Braun

Frontmatter

[More information](#)

Data Analysis and Graphics Using R, Third Edition

Discover what you can do with R! Introducing the R system, covering standard regression methods, then tackling more advanced topics, this book guides users through the practical, powerful tools that the R system provides. The emphasis is on hands-on analysis, graphical display, and interpretation of data. The many worked examples, from real-world research, are accompanied by commentary on what is done and why. The companion website has code and data sets, allowing readers to reproduce all analyses, along with solutions to selected exercises and updates. Assuming basic statistical knowledge and some experience with data analysis (but not R), 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 third edition takes into account recent changes in R, including advances in graphical user interfaces (GUIs) and graphics packages. The treatments of the random forests methodology and one-way analysis have been extended. Both text and code have been revised throughout, and where possible simplified. New graphs and examples have been added.

JOHN MAINDONALD is Visiting Fellow at the Mathematical Sciences Institute at the 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.

W. JOHN BRAUN is Professor in the Department of Statistical and Actuarial Sciences at the 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

978-0-521-76293-9 - Data Analysis and Graphics Using R – an Example-Based Approach: Third Edition

John Maindonald and W. John Braun

Frontmatter

[More information](#)

Cambridge University Press

978-0-521-76293-9 - Data Analysis and Graphics Using R – an Example-Based Approach: Third Edition

John Maindonald and W. John Braun

Frontmatter

[More information](#)

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

Third Edition

CAMBRIDGE SERIES IN STATISTICAL AND PROBABILISTIC
MATHEMATICS

Editorial Board

- Z. Ghahramani (Department of Engineering, University of Cambridge)
 R. Gill (Mathematical Institute, Leiden University)
 F. P. Kelly (Department of Pure Mathematics and Mathematical Statistics,
 University of Cambridge)
 B. D. Ripley (Department of Statistics, University of Oxford)
 S. Ross (Department of Industrial and 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.

A complete list of books in the series can be found at
<http://www.cambridge.org/uk/series/sSeries.asp?code=CSPM>

Recent titles include the following:

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 John Braun
11. *Statistical Models*, by A. C. Davison
12. *Semiparametric Regression*, by David Ruppert, M. P. Wand and R. J. Carroll
13. *Exercises in Probability*, by Loïc 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
21. *Networks*, by Peter Whittle
22. *Saddlepoint Approximations with Applications*, by Ronald W. Butler
23. *Applied Asymptotics*, by A. R. Brazzale, A. C. Davison and N. Reid
24. *Random Networks for Communication*, by Massimo Franceschetti and Ronald Meester
25. *Design of Comparative Experiments*, by R. A. Bailey
26. *Symmetry Studies*, by Marlos A. G. Viana
27. *Model Selection and Model Averaging*, by Gerda Claeskens and Nils Lid Hjort
28. *Bayesian Nonparametrics*, edited by Nils Lid Hjort *et al*
29. *From Finite Sample to Asymptotic Methods in Statistics*, by Pranab K. Sen, Julio M. Singer and Antonio C. Pedrosa de Lima
30. *Brownian Motion*, by Peter Mörters and Yuval Peres

Cambridge University Press

978-0-521-76293-9 - Data Analysis and Graphics Using R – an Example-Based Approach: Third Edition

John Maindonald and W. John Braun

Frontmatter

[More information](#)

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

Third Edition

John Maindonald

Mathematical Sciences Institute, Australian National University

and

W. John Braun

Department of Statistical and Actuarial Sciences, University of Western Ontario



CAMBRIDGE
UNIVERSITY PRESS

Cambridge University Press

978-0-521-76293-9 - Data Analysis and Graphics Using R – an Example-Based Approach: Third Edition

John Maindonald and W. John Braun

Frontmatter

[More information](#)

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9780521762939

First edition © Cambridge University Press 2003

Second and third editions © John Maindonald and W. John Braun 2007, 2010

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 published 2007

Reprinted 2007, 2008, 2009

Third edition published 2010

Reprinted with corrections 2011

4th printing 2014

Printed in the United States of America by Sheridan Books, Inc.

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

Library of Congress Cataloguing in Publication data

Maindonald, J. H. (John Hilary), 1937–

Data analysis and graphics using R : an example-based approach / John Maindonald, W. John Braun. – 3rd ed.

p. cm. – (Cambridge series in statistical and probabilistic mathematics ; 10)

Includes bibliographical references and indexes.

ISBN 978-0-521-76293-9

1. Statistics – Data processing. 2. Statistics – Graphic methods – Data processing.

3. R (Computer program language) I. Braun, John, 1963– II. Title. III. Series.

QA276.4.M245 2010

519.50285 – dc22 2009054016

ISBN 978-0-521-76293-9 Hardback

Additional resources for this publication at www.maths.anu.edu.au/~johnm/r-book

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

978-0-521-76293-9 - Data Analysis and Graphics Using R – an Example-Based Approach: Third Edition

John Maindonald and W. John Braun

Frontmatter

[More information](#)

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

For Susan, Matthew and Phillip

Cambridge University Press

978-0-521-76293-9 - Data Analysis and Graphics Using R – an Example-Based Approach: Third Edition

John Maindonald and W. John Braun

Frontmatter

[More information](#)

Contents

Preface	<i>page</i> xix
Content – how the chapters fit together	xxv
1 A brief introduction to R	1
1.1 <i>An overview of R</i>	1
1.1.1 A short R session	1
1.1.2 The uses of R	6
1.1.3 Online help	7
1.1.4 Input of data from a file	8
1.1.5 R packages	9
1.1.6 Further steps in learning R	9
1.2 <i>Vectors, factors, and univariate time series</i>	10
1.2.1 Vectors	10
1.2.2 Concatenation – joining vector objects	10
1.2.3 The use of relational operators to compare vector elements	11
1.2.4 The use of square brackets to extract subsets of vectors	11
1.2.5 Patterned data	11
1.2.6 Missing values	12
1.2.7 Factors	13
1.2.8 Time series	14
1.3 <i>Data frames and matrices</i>	14
1.3.1 Accessing the columns of data frames – <code>with()</code> and <code>attach()</code>	17
1.3.2 Aggregation, stacking, and unstacking	17
1.3.3* Data frames and matrices	18
1.4 <i>Functions, operators, and loops</i>	19
1.4.1 Common useful built-in functions	19
1.4.2 Generic functions, and the class of an object	21
1.4.3 User-written functions	22
1.4.4 <code>if</code> Statements	23
1.4.5 Selection and matching	23
1.4.6 Functions for working with missing values	24
1.4.7* Looping	24

1.5	<i>Graphics in R</i>	25
1.5.1	The function <code>plot()</code> and allied functions	25
1.5.2	The use of color	27
1.5.3	The importance of aspect ratio	28
1.5.4	Dimensions and other settings for graphics devices	28
1.5.5	The plotting of expressions and mathematical symbols	29
1.5.6	Identification and location on the figure region	29
1.5.7	Plot methods for objects other than vectors	30
1.5.8	Lattice (trellis) graphics	30
1.5.9	Good and bad graphs	32
1.5.10	Further information on graphics	33
1.6	<i>Additional points on the use of R</i>	33
1.7	<i>Recap</i>	35
1.8	<i>Further reading</i>	36
1.9	<i>Exercises</i>	37
2	Styles of data analysis	43
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	47
2.1.3	Patterns in bivariate data	49
2.1.4	Patterns in grouped data – lengths of cuckoo eggs	52
2.1.5*	Multiple variables and times	53
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	59
2.2.2	Summaries of information from data frames	63
2.2.3	Standard deviation and inter-quartile range	65
2.2.4	Correlation	67
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	How will results be used?	70
2.3.3	Formal and informal assessments	71
2.3.4	Statistical analysis strategies	72
2.3.5	Planning the formal analysis	72
2.3.6	Changes to the intended plan of analysis	73
2.4	<i>Recap</i>	73
2.5	<i>Further reading</i>	74
2.6	<i>Exercises</i>	74
3	Statistical models	77
3.1	<i>Statistical models</i>	77
3.1.1	Incorporation of an error or noise component	78
3.1.2	Fitting models – the model formula	80

<i>Contents</i>		xi
3.2	<i>Distributions: models for the random component</i>	81
3.2.1	Discrete distributions – models for counts	82
3.2.2	Continuous distributions	84
3.3	<i>Simulation of random numbers and random samples</i>	86
3.3.1	Sampling from the normal and other continuous distributions	87
3.3.2	Simulation of regression data	88
3.3.3	Simulation of the sampling distribution of the mean	88
3.3.4	Sampling from finite populations	90
3.4	<i>Model assumptions</i>	91
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	96
3.5	<i>Recap</i>	97
3.6	<i>Further reading</i>	98
3.7	<i>Exercises</i>	98
4	A review of inference concepts	102
4.1	<i>Basic concepts of estimation</i>	102
4.1.1	Population parameters and sample statistics	102
4.1.2	Sampling distributions	102
4.1.3	Assessing accuracy – the standard error	103
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	105
4.2	<i>Confidence intervals and tests of hypotheses</i>	106
4.2.1	A summary of one- and two-sample calculations	109
4.2.2	Confidence intervals and tests for proportions	112
4.2.3	Confidence intervals for the correlation	113
4.2.4	Confidence intervals versus hypothesis tests	113
4.3	<i>Contingency tables</i>	114
4.3.1	Rare and endangered plant species	116
4.3.2	Additional notes	119
4.4	<i>One-way unstructured comparisons</i>	119
4.4.1	Multiple comparisons	122
4.4.2	Data with a two-way structure, i.e., two factors	123
4.4.3	Presentation issues	124
4.5	<i>Response curves</i>	125
4.6	<i>Data with a nested variation structure</i>	126
4.6.1	Degrees of freedom considerations	127
4.6.2	General multi-way analysis of variance designs	127
4.7	<i>Resampling methods for standard errors, tests, and confidence intervals</i>	128
4.7.1	The one-sample permutation test	128
4.7.2	The two-sample permutation test	129

4.7.3*	Estimating the standard error of the median: bootstrapping	130
4.7.4	Bootstrap estimates of confidence intervals	131
4.8*	<i>Theories of inference</i>	132
4.8.1	Maximum likelihood estimation	133
4.8.2	Bayesian estimation	133
4.8.3	If there is strong prior information, use it!	135
4.9	<i>Recap</i>	135
4.10	<i>Further reading</i>	136
4.11	<i>Exercises</i>	137
5	Regression with a single predictor	142
5.1	<i>Fitting a line to data</i>	142
5.1.1	Summary information – lawn roller example	143
5.1.2	Residual plots	143
5.1.3	Iron slag example: is there a pattern in the residuals?	145
5.1.4	The analysis of variance table	147
5.2	<i>Outliers, influence, and robust regression</i>	147
5.3	<i>Standard errors and confidence intervals</i>	149
5.3.1	Confidence intervals and tests for the slope	150
5.3.2	SEs and confidence intervals for predicted values	150
5.3.3*	Implications for design	151
5.4	<i>Assessing predictive accuracy</i>	152
5.4.1	Training/test sets and cross-validation	153
5.4.2	Cross-validation – an example	153
5.4.3*	Bootstrapping	155
5.5	<i>Regression versus qualitative anova comparisons – issues of power</i>	158
5.6	<i>Logarithmic and other transformations</i>	160
5.6.1*	A note on power transformations	160
5.6.2	Size and shape data – allometric growth	161
5.7	<i>There are two regression lines!</i>	162
5.8	<i>The model matrix in regression</i>	163
5.9*	<i>Bayesian regression estimation using the MCMCpack package</i>	165
5.10	<i>Recap</i>	166
5.11	<i>Methodological references</i>	167
5.12	<i>Exercises</i>	167
6	Multiple linear regression	170
6.1	<i>Basic ideas: a book weight example</i>	170
6.1.1	Omission of the intercept term	172
6.1.2	Diagnostic plots	173
6.2	<i>The interpretation of model coefficients</i>	174
6.2.1	Times for Northern Irish hill races	174
6.2.2	Plots that show the contribution of individual terms	177
6.2.3	Mouse brain weight example	179
6.2.4	Book dimensions, density, and book weight	181

<i>Contents</i>	xiii
6.3 <i>Multiple regression assumptions, diagnostics, and efficacy measures</i>	183
6.3.1 Outliers, leverage, influence, and Cook's distance	183
6.3.2 Assessment and comparison of regression models	186
6.3.3 How accurately does the equation predict?	187
6.4 <i>A strategy for fitting multiple regression models</i>	189
6.4.1 Suggested steps	190
6.4.2 Diagnostic checks	191
6.4.3 An example – Scottish hill race data	191
6.5 <i>Problems with many explanatory variables</i>	196
6.5.1 Variable selection issues	197
6.6 <i>Multicollinearity</i>	199
6.6.1 The variance inflation factor	201
6.6.2 Remedies for multicollinearity	203
6.7 <i>Errors in x</i>	203
6.8 <i>Multiple regression models – additional points</i>	208
6.8.1 Confusion between explanatory and response variables	208
6.8.2 Missing explanatory variables	208
6.8.3* The use of transformations	210
6.8.4* Non-linear methods – an alternative to transformation?	210
6.9 <i>Recap</i>	212
6.10 <i>Further reading</i>	212
6.11 <i>Exercises</i>	214
7 Exploiting the linear model framework	217
7.1 <i>Levels of a factor – using indicator variables</i>	217
7.1.1 Example – sugar weight	217
7.1.2 Different choices for the model matrix when there are factors	220
7.2 <i>Block designs and balanced incomplete block designs</i>	222
7.2.1 Analysis of the rice data, allowing for block effects	222
7.2.2 A balanced incomplete block design	223
7.3 <i>Fitting multiple lines</i>	224
7.4 <i>Polynomial regression</i>	228
7.4.1 Issues in the choice of model	229
7.5* <i>Methods for passing smooth curves through data</i>	231
7.5.1 Scatterplot smoothing – regression splines	232
7.5.2* Roughness penalty methods and generalized additive models	235
7.5.3 Distributional assumptions for automatic choice of roughness penalty	236
7.5.4 Other smoothing methods	236
7.6 <i>Smoothing with multiple explanatory variables</i>	238
7.6.1 An additive model with two smooth terms	238
7.6.2* A smooth surface	240
7.7 <i>Further reading</i>	240
7.8 <i>Exercises</i>	240

8	Generalized linear models and survival analysis	244
8.1	<i>Generalized linear models</i>	244
8.1.1	Transformation of the expected value on the left	244
8.1.2	Noise terms need not be normal	245
8.1.3	Log odds in contingency tables	245
8.1.4	Logistic regression with a continuous explanatory variable	246
8.2	<i>Logistic multiple regression</i>	249
8.2.1	Selection of model terms, and fitting the model	252
8.2.2	Fitted values	254
8.2.3	A plot of contributions of explanatory variables	255
8.2.4	Cross-validation estimates of predictive accuracy	255
8.3	<i>Logistic models for categorical data – an example</i>	256
8.4	<i>Poisson and quasi-Poisson regression</i>	258
8.4.1	Data on aberrant crypt foci	258
8.4.2	Moth habitat example	261
8.5	<i>Additional notes on generalized linear models</i>	266
8.5.1*	Residuals, and estimating the dispersion	266
8.5.2	Standard errors and z - or t -statistics for binomial models	267
8.5.3	Leverage for binomial models	268
8.6	<i>Models with an ordered categorical or categorical response</i>	268
8.6.1	Ordinal regression models	269
8.6.2*	Loglinear models	272
8.7	<i>Survival analysis</i>	272
8.7.1	Analysis of the <code>Aids2</code> data	273
8.7.2	Right-censoring prior to the termination of the study	275
8.7.3	The survival curve for male homosexuals	276
8.7.4	Hazard rates	276
8.7.5	The Cox proportional hazards model	277
8.8	<i>Transformations for count data</i>	279
8.9	<i>Further reading</i>	280
8.10	<i>Exercises</i>	281
9	Time series models	283
9.1	<i>Time series – some basic ideas</i>	283
9.1.1	Preliminary graphical explorations	283
9.1.2	The autocorrelation and partial autocorrelation function	284
9.1.3	Autoregressive models	285
9.1.4*	Autoregressive moving average models – theory	287
9.1.5	Automatic model selection?	288
9.1.6	A time series forecast	289
9.2*	<i>Regression modeling with ARIMA errors</i>	291
9.3*	<i>Non-linear time series</i>	298
9.4	<i>Further reading</i>	300
9.5	<i>Exercises</i>	301

10	Multi-level models and repeated measures	303
10.1	<i>A one-way random effects model</i>	304
10.1.1	Analysis with <code>aov()</code>	305
10.1.2	A more formal approach	308
10.1.3	Analysis using <code>lmer()</code>	310
10.2	<i>Survey data, with clustering</i>	313
10.2.1	Alternative models	313
10.2.2	Instructive, though faulty, analyses	318
10.2.3	Predictive accuracy	319
10.3	<i>A multi-level experimental design</i>	319
10.3.1	The anova table	321
10.3.2	Expected values of mean squares	322
10.3.3*	The analysis of variance sums of squares breakdown	323
10.3.4	The variance components	325
10.3.5	The mixed model analysis	326
10.3.6	Predictive accuracy	328
10.4	<i>Within- and between-subject effects</i>	329
10.4.1	Model selection	329
10.4.2	Estimates of model parameters	331
10.5	<i>A generalized linear mixed model</i>	332
10.6	<i>Repeated measures in time</i>	334
10.6.1	Example – random variation between profiles	336
10.6.2	Orthodontic measurements on children	340
10.7	<i>Further notes on multi-level and other models with correlated errors</i>	344
10.7.1	Different sources of variance – complication or focus of interest?	344
10.7.2	Predictions from models with a complex error structure	345
10.7.3	An historical perspective on multi-level models	345
10.7.4	Meta-analysis	347
10.7.5	Functional data analysis	347
10.7.6	Error structure in explanatory variables	347
10.8	<i>Recap</i>	347
10.9	<i>Further reading</i>	348
10.10	<i>Exercises</i>	349
11	Tree-based classification and regression	351
11.1	<i>The uses of tree-based methods</i>	352
11.1.1	Problems for which tree-based regression may be used	352
11.2	<i>Detecting email spam – an example</i>	353
11.2.1	Choosing the number of splits	356
11.3	<i>Terminology and methodology</i>	356
11.3.1	Choosing the split – regression trees	357
11.3.2	Within and between sums of squares	357
11.3.3	Choosing the split – classification trees	358
11.3.4	Tree-based regression versus loess regression smoothing	359

11.4	<i>Predictive accuracy and the cost–complexity trade-off</i>	361
11.4.1	Cross-validation	361
11.4.2	The cost–complexity parameter	362
11.4.3	Prediction error versus tree size	363
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	366
11.6	<i>Detecting email spam – the optimal tree</i>	366
11.7	<i>The randomForest package</i>	369
11.8	<i>Additional notes on tree-based methods</i>	372
11.9	<i>Further reading and extensions</i>	373
11.10	<i>Exercises</i>	374
12	Multivariate data exploration and discrimination	377
12.1	<i>Multivariate exploratory data analysis</i>	378
12.1.1	Scatterplot matrices	378
12.1.2	Principal components analysis	379
12.1.3	Multi-dimensional scaling	383
12.2	<i>Discriminant analysis</i>	385
12.2.1	Example – plant architecture	386
12.2.2	Logistic discriminant analysis	387
12.2.3	Linear discriminant analysis	388
12.2.4	An example with more than two groups	390
12.3*	<i>High-dimensional data, classification, and plots</i>	392
12.3.1	Classifications and associated graphs	394
12.3.2	Flawed graphs	394
12.3.3	Accuracies and scores for test data	398
12.3.4	Graphs derived from the cross-validation process	404
12.4	<i>Further reading</i>	406
12.5	<i>Exercises</i>	407
13	Regression on principal component or discriminant scores	410
13.1	<i>Principal component scores in regression</i>	410
13.2*	<i>Propensity scores in regression comparisons – labor training data</i>	414
13.2.1	Regression comparisons	417
13.2.2	A strategy that uses propensity scores	419
13.3	<i>Further reading</i>	426
13.4	<i>Exercises</i>	426
14	The R system – additional topics	427
14.1	<i>Graphical user interfaces to R</i>	427
14.1.1	The R Commander’s interface – a guide to getting started	428
14.1.2	The <i>rattle</i> GUI	429
14.1.3	The creation of simple GUIs – the <i>fgui</i> package	429
14.2	<i>Working directories, workspaces, and the search list</i>	430

Contents

xvii

14.2.1*	The search path	430
14.2.2	Workspace management	430
14.2.3	Utility functions	431
14.3	<i>R system configuration</i>	432
14.3.1	The R Windows installation directory tree	432
14.3.2	The library directories	433
14.3.3	The startup mechanism	433
14.4	<i>Data input and output</i>	433
14.4.1	Input of data	434
14.4.2	Data output	437
14.4.3	Database connections	438
14.5	<i>Functions and operators – some further details</i>	438
14.5.1	Function arguments	439
14.5.2	Character string and vector functions	440
14.5.3	Anonymous functions	441
14.5.4	Functions for working with dates (and times)	441
14.5.5	Creating groups	443
14.5.6	Logical operators	443
14.6	<i>Factors</i>	444
14.7	<i>Missing values</i>	446
14.8*	<i>Matrices and arrays</i>	448
14.8.1	Matrix arithmetic	450
14.8.2	Outer products	451
14.8.3	Arrays	451
14.9	<i>Manipulations with lists, data frames, matrices, and time series</i>	452
14.9.1	Lists – an extension of the notion of “vector”	452
14.9.2	Changing the shape of data frames (or matrices)	454
14.9.3*	Merging data frames – <code>merge()</code>	455
14.9.4	Joining data frames, matrices, and vectors – <code>cbind()</code>	455
14.9.5	The <code>apply</code> family of functions	456
14.9.6	Splitting vectors and data frames into lists – <code>split()</code>	457
14.9.7	Multivariate time series	458
14.10	<i>Classes and methods</i>	458
14.10.1	Printing and summarizing model objects	459
14.10.2	Extracting information from model objects	460
14.10.3	S4 classes and methods	460
14.11	<i>Manipulation of language constructs</i>	461
14.11.1	Model and graphics formulae	461
14.11.2	The use of a list to pass arguments	462
14.11.3	Expressions	463
14.11.4	Environments	463
14.11.5	Function environments and lazy evaluation	464
14.12*	<i>Creation of R packages</i>	465
14.13	<i>Document preparation – <code>Sweave()</code> and <code>xtable()</code></i>	467
14.14	<i>Further reading</i>	468
14.15	<i>Exercises</i>	469

Cambridge University Press

978-0-521-76293-9 - Data Analysis and Graphics Using R – an Example-Based Approach: Third Edition

John Maindonald and W. John Braun

Frontmatter

[More information](#)

xviii

Contents

15	Graphs in R	472
15.1	<i>Hardcopy graphics devices</i>	472
15.2	<i>Plotting characters, symbols, line types, and colors</i>	472
15.3	<i>Formatting and plotting of text and equations</i>	474
	15.3.1 Symbolic substitution of symbols in an expression	475
	15.3.2 Plotting expressions in parallel	475
15.4	<i>Multiple graphs on a single graphics page</i>	476
15.5	<i>Lattice graphics and the grid package</i>	477
	15.5.1 Groups within data, and/or columns in parallel	478
	15.5.2 Lattice parameter settings	480
	15.5.3 Panel functions, strip functions, strip labels, and other annotation	483
	15.5.4 Interaction with lattice (and other) plots – the <i>playwith</i> package	485
	15.5.5 Interaction with lattice plots – focus, interact, unfocus	485
	15.5.6 Overlaid plots with different scales	486
15.6	<i>An implementation of Wilkinson’s Grammar of Graphics</i>	487
15.7	<i>Dynamic graphics – the <code>rgl</code> and <code>rggobi</code> packages</i>	491
15.8	<i>Further reading</i>	492
	Epilogue	493
	References	495
	Index of R symbols and functions	507
	Index of terms	514
	Index of authors	523

The color plates will be found between pages 328 and 329.

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, careful critiques of analysis results, the use of modern data analysis tools and approaches, the use of simulation and other computer-intensive methods – where these provide insight or give results that are not otherwise available, attention to graphical and other presentation issues, the use of

Cambridge University Press

978-0-521-76293-9 - Data Analysis and Graphics Using R – an Example-Based Approach: Third Edition

John Maindonald and W. John Braun

Frontmatter

[More information](#)

xx

Preface

examples drawn from across the range of statistical applications, and the inclusion of code that reproduces analyses.

A substantial part of the book was derived, initially, from John Maindonald's 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.

The R system

We use the R system for computations. It began in the early 1990s as a project of Ross Ihaka and Robert Gentleman, who were both at the time working at the University of Auckland (New Zealand). 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, 20 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 15 or 20 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. 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 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-borne 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 give a list at the end of the reference section that cites 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. See, for example, the comments in Gigerenzer *et al.* (1989) on the manner in which differences of historical development 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 computing tools

We have drawn attention to advances in statistical computing methodology. These have made possible the development of 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.

New traditions of data analysis have developed – data mining, machine learning, and analytics. These emphasize new types of data, new data analysis demands, new data analysis tools, and data sets that may be of unprecedented size. Textual data and image data offer interesting new challenges for data analysis. The traditional concerns of professional data analysts remain as important as ever. Size of data set is not a guarantee of quality and of relevance to issues that are under investigation. It does not guarantee that the source population has been adequately sampled, or that the results will generalize as required to the target population.

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.

Third edition changes and additions

The second edition added new material on survival analysis, random coefficient models, the handling of high-dimensional data, and extended the account of regression methods. This third edition has a more adequate account of errors in predictor variables, extends the treatment and use of random forests, and adds a brief account of generalized linear mixed models. The treatment of one-way analysis of variance, and a major part of the chapter on regression, have been rewritten.

Two areas of especially rapid advance have been graphical user interfaces (GUIs), and graphics. There are now brief introductions to two popular GUIs for R – the R Commander (*Rcmdr*) and *rattle*. The sections on graphics have been substantially extended. There is a brief account of the *lattice* and associated *plywith* GUIs for interfacing with R graphics.

Code has again been extensively revised, simplifying it wherever possible. There are changes to some graphs, and new graphs have been added.

Acknowledgments

Many different people have helped with this project. Winfried Theis (University of Dortmund, Germany) and Detlef Steuer (University of the Federal Armed Forces, Hamburg, Germany) helped with technical \LaTeX issues, with a cvs archive for manuscript 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 highly useful comment on the manuscript. Susan Wilson (Australian National University) gave welcome encouragement. Duncan Murdoch (University of Western Ontario) helped with technical advice. Cath Lawrence (Australian National University) wrote a Python program that allowed us to extract the R code from our \LaTeX files; this has now at length become an R function.

For the second edition, Brian Ripley (University of Oxford) made extensive comments on the manuscript, leading to important corrections and improvements. We are most grateful to him, and to others who have offered comments. Alan Welsh (Australian National University) has helped work through points where it has seemed difficult to get the emphasis right. Once again, Duncan Murdoch has given much useful technical advice. Others who 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 Chapter 15), Graham Williams (<http://www.togaware.com>, on Chapter 1), and Yang Yang (University of Western Ontario, on Chapter 10). Comment that has contributed to this edition has come from Ray Balise (Stanford School of Medicine), Wenqing He and Lengyi Han (University of Western Ontario), Paul Murrell, Andrew Robinson (University of Melbourne, on Chapter 10), Phil Kocic (Australian National University, on Chapter 9), and Rob Hyndman (Monash University, on Chapter 9). Readers who have made relatively extensive comments include Bob Green (Queensland Health) and Zander Smith (SwissRe). Additionally, discussions on the R-help and R-devel email lists have been an important source of insight and understanding. 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. Following the references is a list of data sources (individuals and/or organizations) that we wish to thank and acknowledge. 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 the second and third editions, has been a marvellous source of advice and encouragement.

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

Cambridge University Press

978-0-521-76293-9 - Data Analysis and Graphics Using R – an Example-Based Approach: Third Edition

John Maindonald and W. John Braun

Frontmatter

[More information](#)

xxiv

Preface

`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>

Content – how the chapters fit together

Chapter 1 is a brief introduction to R. Readers who are new to R should as a minimum study Section 1.1, or an equivalent, before moving on to later chapters. In later study, refer back as needed to Chapter 1, or forward to Chapter 14.

Chapters 2–4: Exploratory data analysis and review of elementary statistical ideas

Chapters 2–4 cover, at greater depth and from a more advanced perspective, topics that are common in introductory courses. Different readers will use these chapters differently, depending on their statistical preparedness.

Chapter 2 (*Styles of data analysis*) places data analysis in the wider context of the research study, commenting on some of the types of graphs that may help answer questions that are commonly of interest and that will be used throughout the remainder of the text. Subsections 2.1.7, 2.2.3 and 2.2.4 introduce terminology that will be important in later chapters.

Chapter 3 (*Statistical models*) introduces the *signal + noise* form of regression model. The different models for the signal component are too varied to describe in one chapter! Coverage of models for the *noise* (random component) is, relative to their use in remaining chapters, more complete.

Chapter 4 (*A review of inference concepts*) describes approaches to generalizing from data. It notes the limitations of the formal hypothesis testing methodology, arguing that a less formal approach is often adequate. It notes also that there are contexts where a Bayesian approach is essential, in order to take account of strong prior information.

Chapters 5–13: Regression and related methodology

Chapters 5–13 are designed to give a sense of the variety and scope of methods that come, broadly, under the heading of *regression*. In Chapters 5 and 6, the models are linear in the explanatory variable(s) as well as in the parameters. A wide range of issues affect the practical use of these models: influence, diagnostics, robust and resistant methods, AIC and other model comparison measures, interpretation of coefficients, variable selection, multicollinearity, and errors in x . All these issues are relevant, in one way or another, throughout later chapters. Chapters 5 and 6 provide relatively straightforward contexts in which to introduce them.

The models of Chapters 5–13 give varying combinations of answers to the questions:

1. What is the *signal* term? Is it in some sense linear? Can it be described by a simple form of mathematical equation?
2. Is the *noise* term *normal*, or are there other possibilities?
3. Are the noise terms independent between observations?
4. Is the model specified in advance? Or will it be necessary to choose the model from a potentially large number of possible models?

In Chapters 5–8, the models become increasingly general, but always with a model that is linear in the coefficients as a starting point. In Chapters 5–7, the noise terms are normal and independent between observations. The *generalized linear models* of Chapter 8 allow non-normal noise terms. These are still assumed independent.¹ Chapter 9 (*Time series models*) and Chapter 10 (*Multilevel models and repeated measures*) introduce models that allow, in their different ways, for dependence between observations. In Chapter 9 the correlation is with observations at earlier points in time, while in Chapter 10 the correlation might for example be between different students in the same class, as opposed to different students in different classes. In both types of model, the noise term is constructed from normal components – there are normality assumptions.

Chapters 6–10 allowed limited opportunity for the choice of model and/or explanatory variables. Chapter 11 (*Tree-based classification and regression*) introduces models that are suited to a *statistical learning* approach, where the model is chosen from a large portfolio of possibilities. Moreover, these models do not have any simple form of equation. Note the usual implicit assumption of independence between observations – this imposes limitations that, depending on the context, may or may not be important for practical use.

Chapter 12 (*Multivariate data exploration and discrimination*) begins with methods that may be useful for multivariate data exploration – principal components, the use of distance measures, and multi-dimensional scaling. It describes dimension reduction approaches that allow low-dimensional views of the data. Subsection 12.2 moves to discriminant methods – i.e., to regression methods in which the outcome is categorical. Subsection 12.3 identifies issues that arise when the number of variables is large relative to the number of observations. Such data is increasingly common in many different application areas.

It is sometimes possible to replace a large number of explanatory variables by one, or a small number, of *scoring* variables that capture the relevant information in the data. Chapter 13 investigates two different ways to create scores that may be used as explanatory variables in regression. In the first example, the principal component scores are used. The second uses *propensity* scores to summarize information on a number of covariates that are thought to explain group differences that are, for the purposes of the investigation, nuisance variables.

¹ Note, however, the extension to allow models with a variance that, relative to the binomial or Poisson, is inflated.