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.

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

Third Edition

CAMBRIDGE SERIES IN STATISTICAL AND PROBABILISTIC MATHEMATICS

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

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

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 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

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

For Susan, Matthew and Phillip

Contents

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

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 sudoko 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 *latticist* and associated *playwith* 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 Kokic (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

`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.