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

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.

Already published
1. *Bootstrap Methods and Their Application*, A.C. Davison and D.V. Hinkley
2. *Markov Chains*, J. Norris
3. *Asymptotic Statistics*, A.W. van der Vaart
4. *Wavelet Methods for Time Series Analysis*, D.B. Percival and A.T. Walden
5. *Bayesian Methods*, T. Leonard and J.S.J. Mu
6. *Empirical Processes in M-Estimation*, S. van de Geer
7. *Numerical Methods of Statistics*, J. Monahan
8. *A User's Guide to Measure-Theoretic Probability*, D. Pollard
9. *The Estimation and Tracking of Frequency*, B.G. Quinn and E.J. Hannan

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

John Maindonald
Centre for Bioinformation Science, John Curtin School of Medical Research
and Mathematical Sciences Institute, Australian National University

and

John Braun
Department of Statistical and Actuarial Science, University of Western Ontario

CAMBRIDGE
UNIVERSITY PRESS

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

Cambridge University Press
40 West 20th Street, New York, NY 10011-4211, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521813365

First published 2003

Reprinted 2004, 2005

Printed in the United States of America

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 and John Braun.
p. cm. – (Cambridge series in statistical and probabilistic mathematics)
Includes bibliographical references and index.
ISBN 0 521 81336 0
1. Statistical – Data processing. 2. Statistics – Graphic methods – Data processing. 3. R (Computer program language) I. Braun, John, 1963– II. Title. III. Cambridge series on statistical and probabilistic mathematics.
QA276.4.M245 2003
519.5′0285–dc21 2002031560

ISBN-13 978-0-521-81336-5 hardback
ISBN-10 0-521-81336-0 hardback

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

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]

Contents

Cambridge University Press
0521813360 - Data Analysis and Graphics Using R: An Example-based Approach
John Maindonald and 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.

The authors can claim, between them, 40 years of experience in working with researchers from many different backgrounds. Initial drafts of the monograph were constructed from notes that the first author prepared for courses for researchers, first of all at the University of Newcastle (Australia) over 1996–1997, and greatly developed and extended in the course of work in the Statistical Consulting Unit at The Australian National University over 1998–2001. We are grateful to those who have discussed their research with us, brought us their data for analysis, and allowed us to use it in the examples that appear in the present monograph. 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 have covered a range of topics that we consider important for many different areas of statistical application. This diversity of sources of examples has benefits, even for those whose interests are in one 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. As is inevitable in a book that has this broad focus, there will be specific areas – perhaps epidemiology, or psychology, or sociology, or ecology – that will regret the omission of some methodologies that they find important.

We use the R system for the computations. The R system implements a dialect of the influential S language that is the basis for the commercial S-PLUS system. It follows S in its close linkage between data analysis and graphics. Its development is the result of a co-operative international effort, bringing together an impressive array of statistical computing expertise. It has quickly gained a wide following, among professionals and non-professionals alike. At the time of writing, R users are restricted, for the most part, to a command line interface. Various forms of graphical user interface will become available in due course.

The R system has an extensive library of packages that offer state-of-the-art-abilities. Many of the analyses that they offer were not, 10 years ago, available in any of the standard 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.

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.

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) 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 Statistical Computing

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.

One can think of an effective statistical analysis package as a workshop (this analogy appears in a simpler form in the JMP Start Statistics Manual (SAS Institute Inc. 1996, p. xiii).). The tools are the statistical and computing abilities that the package provides. The layout of the workshop, the arrangement both of the tools and of the working area, is important. It should be easy to find each tool as it is needed. Tools should float back of their own accord into the right place after use! In other words, we want a workshop where mending the rocking chair is a pleasure!

The workshop analogy is worth pursuing further. Different users have different requirements. A hobbyist workshop will differ from a professional workshop. The hobbyist may have less sophisticated tools, and tools that are easy to use without extensive training or experience. That limits what the hobbyist can do. The professional needs powerful and highly flexible tools, and must be willing to invest time in learning the skills needed to use them.

Good graphical abilities, and good data manipulation abilities, should be a high priority for the hobbyist statistical workshop. Other operations should be reasonably easy to implement when carried out under the instructions of a professional. Professionals also require top rate graphical abilities. The focus is more on flexibility and power, both for graphics and for computation. Ease of use is important, but not at the expense of power and flexibility.

A Note on the R System

The R system implements a dialect of the S language that was developed at AT&T Bell Laboratories by Rick Becker, John Chambers and Allan Wilks. Versions of R are available, at no cost, 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 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 Development 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.

Acknowledgements

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 LATEX, with setting up a cvs server to manage the 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. 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 LATEX files. The failings that remain are, naturally, our responsibility.

There are a large number of people who have helped with the providing of data sets. We give a list, following the list of references for the data near the end of the book. We apologize if there is anyone that we have inadvertently failed to acknowledge. Finally, thanks to David Tranah of Cambridge University Press, for his encouragement and help in bringing the writing of this monograph to fruition.

References

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. & Krüger, L. 1989. *The Empire of Chance*. Cambridge University Press.

SAS Institute Inc. 1996. *JMP Start Statistics*. Duxbury Press, Belmont, CA.

These (and all other) references also appear in the consolidated list of references near the end of the book.

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.

Web sites for supplementary information

The DAAG package, the R scripts that we present, and other supplementary information, are available from

http://cbis.anu.edu/DAAG
http://www.stats.uwo.ca/DAAG

Solutions to exercises

Solutions to selected exercises are available from the website

http://www.maths.anu.edu.au/∼johnm/r-book.html
See also www.cambridge.org/0521813360

A Chapter by Chapter Summary

Chapter 1: A Brief Introduction to R

This chapter aims to give enough information on the use of R to get readers started.

Note R's extensive online help facilities. Users who have a basic minimum knowledge of R can often get needed additional information from the help pages as the demand arises. A facility in using the help pages is an important basic skill for R users.

Chapter 2: Style of Data Analysis

Knowing how to explore a set of data upon encountering it for the first time is an important skill. What graphs should one draw?

Different types of graph give different views of the data. Which views are likely to be helpful?

Transformations, especially the logarithmic transformation, may be a necessary preliminary to data analysis.

There is a contrast between exploratory data analysis, where the aim is to allow the data to speak for themselves, and confirmatory analysis (which includes formal estimation and testing), where the form of the analysis should have been largely decided before the data were collected.

Statistical analysis is a form of data summary. It is important to check, as far as this is possible that summarization has captured crucial features of the data. Summary statistics, such as the mean or correlation, should always be accompanied by examination of a relevant graph. For example, the correlation is a useful summary, if at all, only if the relationship between two variables is linear. A scatterplot allows a visual check on linearity.

Chapter 3: Statistical Models

Formal data analyses assume an underlying statistical model, whether or not it is explicitly written down.

Many statistical models have two components: a *signal* (or deterministic) component; and a *noise* (or error) component.

Data from a sample (commonly assumed to be randomly selected) are used to fit the model by estimating the signal component.

The fitted model determines *fitted* or *predicted* values of the signal. The *residuals* (which estimate the noise component) are what remain after subtracting the fitted values from the observed values of the signal.

The normal distribution is widely used as a model for the noise component.

Haphazardly chosen samples should be distinguished from random samples. Inference from haphazardly chosen samples is inevitably hazardous. Self-selected samples are particularly unsatisfactory.

Chapter 4: An Introduction to Formal Inference

Formal analysis of data leads to inferences about the population(s) from which the data were sampled. Statistics that can be computed from given data are used to convey information about otherwise unknown population parameters.

The inferences that are described in this chapter require randomly selected samples from the relevant populations.

A *sampling distribution* describes the theoretical distribution of sample values of a statistic, based on multiple *independent* random samples from the population.

The standard deviation of a sampling distribution has the name *standard error*.

For sufficiently large samples, the normal distribution provides a good approximation to the true sampling distribution of the mean or a difference of means.

A *confidence interval* for a parameter, such as the mean or a difference of means, has the form

$$\text{statistic} \pm t\text{-critical-value} \;\times\; \text{standard error}.$$

Such intervals give an assessment of the level of uncertainty when using a sample statistic to estimate a population parameter.

Another viewpoint is that of *hypothesis testing*. Is there sufficient evidence to believe that there is a difference between the means of two different populations?

Checks are essential to determine whether it is plausible that confidence intervals and hypothesis tests are valid. Note however that plausibility is not proof!

Standard chi-squared tests for two-way tables assume that items enter independently into the cells of the table. Even where such a test is not valid, the standardized residuals from the "no association" model can give useful insights.

In the one-way layout, in which there are several independent sets of sample values, one for each of several groups, data structure (e.g. compare treatments with control, or focus on a small number of "interesting" contrasts) helps determine the inferences that are appropriate. In general, it is inappropriate to examine all possible comparisons.

In the one-way layout with quantitative levels, a regression approach is usually appropriate.

Chapter 5: Regression with a Single Predictor

Correlation can be a crude and unduly simplistic summary measure of dependence between two variables. Wherever possible, one should use the richer regression framework to gain deeper insights into relationships between variables.

The line or curve for the regression of a response variable y on a predictor x is different from the line or curve for the regression of x on y. Be aware that the inferred relationship is conditional on the values of the predictor variable.

The model matrix, together with estimated coefficients, allows for calculation of predicted or fitted values and residuals.

Following the calculations, it is good practice to assess the fitted model using standard forms of graphical diagnostics.

Simple alternatives to straight line regression using the data in their raw form are

- transforming x and/or y,
- using polynomial regression,
- fitting a smooth curve.

For size and shape data the allometric model is a good starting point. This model assumes that regression relationships among the logarithms of the size variables are linear.

Chapter 6: Multiple Linear Regression

Scatterplot matrices may provide useful insight, prior to fitting a regression model.

Following the fitting of a regression, one should examine relevant diagnostic plots.

Each regression coefficient estimates the effect of changes in the corresponding explanatory variable when other explanatory variables are held constant.

The use of a different set of explanatory variables may lead to large changes in the coefficients for those variables that are in both models.

Selective influences in the data collection can have a large effect on the fitted regression relationship.

For comparing alternative models, the AIC or equivalent statistic (including Mallows C_p) can be useful. The R^2 statistic has limited usefulness.

If the effect of variable selection is ignored, the estimate of predictive power can be grossly inflated.

When regression models are fitted to observational data, and especially if there are a number of explanatory variables, estimated regression coefficients can give misleading indications of the effects of those individual variables.

The most useful test of predictive power comes from determining the predictive accuracy that can be expected from a new data set.

Cross-validation is a powerful and widely applicable method that can be used for assessing the expected predictive accuracy in a new sample.

Chapter 7: Exploiting the Linear Model Framework

In the study of regression relationships, there are many more possibilities than regression lines! If a line is adequate, use that. But one is not limited to lines!

A common way to handle qualitative factors in linear models is to make the initial level the baseline, with estimates for other levels estimated as offsets from this baseline.

Polynomials of degree n can be handled by introducing into the model matrix, in addition to a column of values of x, columns corresponding to $x^2, x^3, \ldots, x^n$. Typically, $n = 2, 3$ or 4.

Multiple lines are fitted as an interaction between the variable and a factor with as many levels as there are different lines.

Scatterplot smoothing, and smoothing terms in multiple linear models, can also be handled within the linear model framework.

Chapter 8: Logistic Regression and Other Generalized Linear Models

Generalized linear models (GLMs) are an extension of linear models, in which a function of the expectation of the response variable y is expressed as a linear model. A further generalization is that y may have a binomial or Poisson or other non-normal distribution.

Common important GLMs are the logistic model and the Poisson regression model.

Survival analysis may be seen as a further specific extension of the GLM framework.

Chapter 9: Multi-level Models, Time Series and Repeated Measures

In a multi-level model, the random component possesses structure; it is a sum of distinct error terms.

Multi-level models that exhibit suitable balance have traditionally been analyzed within an analysis of variance framework. Unbalanced multi-level designs require the more general multi-level modeling methodology.

Observations taken over time often exhibit time-based dependence. Observations that are close together in time may be more highly correlated than those that are widely separated. The autocorrelation function can be used to assess levels of serial correlation in time series.

Repeated measures models have measurements on the same individuals at multiple points in time and/or space. They typically require the modeling of a correlation structure similar to that employed in analyzing time series.

Chapter 10: Tree-based Classification and Regression

Tree-based models make very weak assumptions about the form of the classification or regression model. They make limited use of the ordering properties of continuous or ordinal explanatory variables. They are unsuitable for use with small data sets.

Tree-based models can be an effective tool for analyzing data that are non-linear and/or involve complex interactions.

The decision trees that tree-based analyses generate may be complex, giving limited insight into model predictions.

Cross-validation, and the use of training and test sets, are essential tools both for choosing the size of the tree and for assessing expected accuracy on a new data set.

Chapter 11: Multivariate Data Exploration and Discrimination

Principal components analysis is an important multivariate exploratory data analysis tool.

Examples are presented of the use of two alternative discrimination methods – logistic regression including multivariate logistic regression, and linear discriminant analysis.

Both principal components analysis, and discriminant analysis, allow the calculation of scores, which are values of the principal components or discriminant functions, calculated observation by observation. The scores may themselves be used as variables in, e.g., a regression analysis.

Chapter 12: The R System – Additional Topics

This final chapter gives pointers to some of the further capabilities of R. It hints at the marvellous power and flexibility that are available to those who extend their skills in the use of R beyond the basic topics that we have treated. The information in this chapter is intended, also, for use as a reference in connection with the computations of earlier chapters.