Journal of Statistical Software

August 2006, Volume 16, Book Review 6. http://www.jstatsoft.org/

Reviewer: Joseph M. Hilbe
Arizona State University

A Handbook of Statistical Analyses Using R

Brian S. Everitt and Torsten Hothorn
Chapman & Hall/CRC, Boca Raton, Florida, 2006.
ISBN 1-58488-539-4. 304 pp. USD 49.95.
http://CRAN.R-project.org/src/contrib/Descriptions/HSAUR.html

---

Introductory comments

This text is one of a series of five handbooks that present an overview on how to use a major statistical software package. Handbooks include S-PLUS, Stata, SPSS, SAS, and R. Although R is not strictly speaking a statistical package, it is a currently popular statistical language that is downloaded into ones computer from various mirror sites. It is similar in logic to the S language of the 1980s, which later became transformed to the S-PLUS commercial package.

Brian Everitt, an Emeritus Professor with the Department of Biostatistics and Computing at King's College, London, is the primary author of the series. He is a co-author of each book in the series, serving as the lead author in several, including the subject text of this review. Other series authors are uniquely competent in the use of the particular statistical package of the title.

As other handbooks, this text on R comes in both hardback and paperback. Libraries tend to purchase the hardback editions, all other tend to prefer the paperback. The list price of the paperback edition is USD 49.95, but it can be purchased through Amazon or Barnes & Noble for USD 44.95. This is a reasonable cost for an academic text of 304 pages.

The book has fifteen chapters, each devoted to particular aspect of the software. Each chapter ends with a list of three to five exercise questions, based on the subject of the related chapter. The bibliography contains in excess of two hundred entries, providing the reader with an excellent resource of primary readings.

R's higher-level computing language and statistical, data management, and graphical capabilities are outlined in the text. Useful examples are presented to assist understanding. In addition, examples incorporate the R commands which produce the output of interest. A package containing the data sets used for examples can be downloaded from the Comprehensive R Archive Network (CRAN). I shall outline the contents of each chapter, offering comments along the way.

Chapter 1: An introduction to R

This chapter introduces the reader to the R language. Hints are given regarding installation of R from mirror sites, and examples are provided about using the R help and documentation facility. In addition, the authors give a summary on how to understand R data objects, how to import and export data, how to engage in simple data manipulation, and how to produce both summary statistics and essential graphical plots, e.g., histograms, bar graphs, and so forth. Chapter 1 should be considered as necessary reading for those without a prior knowledge of R. Those with competency using R can quickly skim the chapter.

Chapter 2: Simple inference

After presenting several data sets that will be used later in the chapter, as well as in several subsequent chapters, the authors review the theory of t tests, simple non-parametric tests, and tests of contingency tables. Discussion is set at the introductory course level. Examples of most of the discussed tests are then given using R. together with related graphical support. Interpretation of the tests and graphical results are discussed, which greatly adds to the value of the chapter.

I suggest that the first two chapters be read by all who purchase the book, no matter what the level of statistical expertise. The authors do a fine job in guiding the reader through the use of R in analyzing and graphically representing data. If the basics are well understood, more advanced applications of R with data will be better easier to handle.

Chapter 3: Conditional inference

This chapter can be thought of as a follow-up to Chapter 2, presenting more detailed use of R in structuring tables of data. To this end the authors give an analysis of a randomized clinical trial which requires the use of conditional inference procedures. Emphasized in this chapter are various techniques of randomization and assignment to data to treatment groups. Several worked out examples are given, which can later be used by the reader as paradigms for their own work.

This chapter is essential reading for those readers involved in clinical trials and experimental designs. I must give a caveat however. The discussion is tight, without a lot of reiteration of points made. It is best read while simultaneously working out the examples on the computer. It will be easy to forget the many techniques used if one does not also incorporate other modalities in the learning process. This is especially the case for readers new to the R language.

Chapter 4: Analysis of variance

Following a brief overview of ANOVA theory, the authors use four example data sets to demonstrate the use of R in designing appropriate ANOVA tables. Included are examples for evaluating weight gain, deciding on the benefits of foster feeding in rats, testing water hardness, and classifying male Egyptian skulls. For each example, the data is read into memory, summarized, and subjected to the appropriate ANOVA or MANOVA command. The authors also show how to use R to effect Tukey's honest significant differences procedure

in evaluating factor levels. Other tests demonstrated include Hotelling-Lawley, Wilks, Pillai, and so forth. The authors also show how R can be used to develop scatterplot matrices and comparison graphs to accompany ANOVA results.

Chapter 5: Multiple linear regression

The chapter is devoted to showing readers how R can be used for multiple linear regression (OLS). A cloud seeding example is used to demonstrate the procedure. Data is read into memory, and the model is fit. Special emphasis is given to model diagnostics, including residual plots and statistical tests. As usual, an overview of the chapter subject matter is given.

The authors cover most traditional areas related to performing OLS regression, but really nothing is discussed about variable selection techniques. This would be a drawback if the book is adopted as a primary text for an introduction to statistics course. However, I would hope that the book is not used in this manner. It is best used to acquaint those already having a solid background in statistics with the R language, and how it can be used to analyze and graph data. Although some theory is presented, it is merely ancillary to the purpose of the text.

Chapter 6: Logistic regression and generalized linear models

I have argued for the last ten years or more that logistic regression be discussed in introduction to statistics courses. Previously the procedure was not available in all major software packages. But as computing power became ever more enhanced, procedures requiring iterations to fit the model were made more widely available. The fact that so many data modeling situations involve a binary response, or dependent variable, makes it vital that anyone completing an introductory level course have a basic knowledge of the procedure. Fortunately, at some institutions, traditional introductory texts are being augmented by materials outlining logistic regression.

Everitt and Hothorn do an excellent job of presenting an overview of logistic regression in the space of only eighteen pages. Although the logistic or logit model may be estimated using maximum likelihood (ML), it can also be estimated within the framework of generalized linear models (GLM). The authors choose the GLM approach. They do not tell us why they chose this method over ML, but I suspect that it is because subjects discussed later in the book depend on GLM methodology, e.g., GEE, GLMM, and so forth.

The authors give a brief review of GLM and show how both normal (or Gaussian) regression and logistic regression are members of the GLM family. Very little GLM theory is given. I would have liked to see a bit more, especially since GLMs are used later in the text as well. I also think that it would have been helpful to explain a bit about the two types of estimation, and why logistic regression is discussed from the GLM viewpoint. I don't find fault with this approach – not at all. I simply think that it needs more explanation. I would also like there to be some mention of GLM family members aside from binomial – of which logistic is a member – normal, Poisson, and quasi-Poisson. A novice reader may be left with the impression that these are the only GLM families, which is not the case (there are also the gamma, inverse Gaussian, and negative binomial, which incorporates the geometric). Moreover, traditional GLM theory has no quasi-Poisson family. The authors correctly state

that the multiplication of a value to the variance function makes a quasi-likelihood model. However, the quasi-Poisson "family" function used by R is not a quasi-likelihood model. All that was done to the original Poisson model shown on page 105 was to apply a χ^2 scaling factor to the variance. The χ^2 scale is based on the χ^2 dispersion, or the Pearson χ^2 statistic divided by the residual degrees of freedom (model observations minus predictors, including constant). The model standard errors are multiplied by the inverse square of the dispersion statistic after the model parameters have been estimated. It is not a quasi-likelihood model, rather, it is traditonally termed a model with scaled standard errors. The authors are aware of these distinctions, yet did not point them out in the chapter.

Another shortcoming of the chapter relates to logistic regression diagnostics. Most major software packages incorporate Hosmer & Lemeshow diagnostics into their logistic regression procedure. Usually these diagnostics follow when using ML methods of estimation – however, this does not need to be the case. Regardless, Hosmer & Lemeshow diagnostics are now considered somewhat as a standard when dealing with logistic regression models. Yet nothing at all is mentioned of these diagnostics, nor does it appear that R has this capability. If another R function allows it, the authors should mention it.

Chapter 7: Density estimation

Chapter 7 deals with kernal density estimation. Various methods and theory are discussed and examples are presented. Waiting times for the Yellowstone Park geyser "Old Faithful" eruptions are used as the first example. The authors show how to use R to employ a variety of different kernels to derive density estimates of eruption times. They also show the use of contour plots and bivariate plots with the data.

The authors conclude the chapter with a description of how to do bootstrapping in R, together with the production of so-called "bootplots". I found this topic to be well written and quite interesting.

Chapter 8: Recursive partitioning

Recursive partitioning is also known as "classification trees". This capability was an important feature of the S language, and its subsequent commercial version, S-PLUS. The authors explain how to use R in pruning trees and in using procedures such as the random forest method and conditional inference tree. The graphic produced on Glaucoma data using a conditional inference tree is quite spectacular. All of the commands required to produce the graphs and statistical output are included.

Chapter 9: Survival analysis

More space is given to theory in this chapter than any previous chapter in the book. A good overview of survivor and hazard functions is presented, together with a discussion of the logic of Cox regression.

A fully worked out survival analysis example is given for Glioma radioimmunotherapy data. Graphs are developed and interpretation given. The authors even tie in recursive partitioning techniques learned in the previous chapter to survivor function data, producing another very

nice graph. Diagnostics, e.g., martingale residuals, are discussed and plotted.

Chapter 10: Analyzing longitudinal data I

This is the first of two chapters dealing with longitudinal and clustered data. The first one deals with how R can be used with 1) linear mixed effects for repeated measures data, and 2) random effects models. Of particular interest are discussions of empirical Bayes estimates and the problem of dropouts. Dropouts are a major problem when dealing with longitudinal data. The book explores different tactics, and shows how R can be used to handle the problem.

The chapter is one of the most demanding of the book. A prior knowledge of the area is required in order to fully appreciate how R can be used to engage in these models. The chapter is well written, but the code is rather advanced.

Chapter 11: Analyzing longitudinal data II

Generalized estimating equations (GEE) is the sole model discussed in this chapter. There is more space devoted to this single topic than to any other in the book. Correlation structures are examined and several models are developed as examples. Code is shown on how to present graphical representations across time periods – which I found quite helpful. The only downside is the lack of good references.

Chapter 12: Meta-analysis

One of the shorter chapters, the theory of meta-analysis is discussed at some length. Examples using both fixed and random effects models are used to show how R can be used as a meta-analytic tool. The authors take a good amount of space in describing how graphics can assist in meta-analysis, and provide complete code to replicate the examples given in the text. This is a subject area not discussed in other handbook texts. It is also rarely discussed in other introductory level statistics texts. I applaud the authors for showing how R can be used in such analysis.

Chapter 13: Principal components analysis

Principal components analysis, on the other hand, is a commonly discussed theme in the handbook series. Little theory is presented, but an example using results from the 1988 Olympic Heptathlon competition provides an interesting application of the method. I found the results quite consistent with my experience in the area (as a US Olympic Decathlon Trials competitor, 2-time national Pentathlon champion, and lead competition official at the 1984 Los Angeles Olympic Games decathlon and heptathlon). The example spurred my interest to further examine the data. Complete R code is provided to allow the reader to follow and expand on the analysis.

Chapter 14: Multidimensional scaling

The authors discuss both classical and non-parametric multidimensional scaling techniques.

One of the two examples used relates to how members of the New Jersey US House of representatives voted on some 19 environmental bills. The other dealt with a comparison of 13 characteristics of British and continental water voles. Although both example data sets were used to demonstrate the use of R code, primary emphasis was given to the water vole data. In any case, theory is presented, including a good discussion of measurement distances, e.g., Euclidian distance.

Chapter 15: Cluster analysis

Cluster analysis is the final topic of the text. A brief, but satisfactory account of the many concepts involved in cluster analysis is presented. Emphasis is given to k-means clustering, model-based clustering, and classification maximum likelihood. For an example, the authors selected data from exoplanet classification study. This is timely, given the many planets now being discovered each year around stars in our galactic vicinity. Having an interest in astrostatistics, I found this example not only interesting, but extremely well presented. The use of R to produce associated graphics should be of value to anyone interested in the area.

Summary

Everitt and Hothorn have written an excellent tutorial on using R to analyze data using a wide range of standard statistical methods. They use numerous examples throughout the text, present 100 figures, and show 54 tables to augment discussion. And this is all done in a book of only 275 pages in length.

I highly recommend the text or anyone learning R, and who want to use it for the sophisticated analysis of data. No knowledge of R is presumed, but it is expected that the reader have a basic well-rounded knowledge of statistics.

Reviewer:

Joseph M. Hilbe
Emeritus Professor, University of Hawaii, and
Adjunct Professor, Sociology and Statistics, Arizona State University
Tempe, Arizona, United States of America
E-mail: hilbe@asu.edu or jhilbe@aol.com

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 16, Book Review 6 *Published:* 2006-08-12
August 2006