Journal of Statistical Software

September 2010, Volume 36, Book Review 6. http://www.jstatsoft.org/

Reviewer: Juana Sanchez
University of California, Los Angeles

Data Analysis and Graphics Using R – An Example-Based Approach (3rd Edition)

John Maindonald and W. John Braun
Cambridge University Press, New York, NY, 2010.
ISBN 978-0-521-76293-9. 525 pp. USD 80.00.
http://www.maths.anu.edu.au/~johnm/r-book/daagur3.html

Maindonald and Braun's book is a hands-on guide to data analysis with R unlike other statistics books that rely on R for computing. The book is not as specialized as those in the UseR! series of Springer. It also is more general and more focused on the ideas and concepts of statistics than Verzani (2005) or Dalgaard (2008). The contents are comprehensive enough for a statistical methodology course taught to juniors, seniors and graduate students previously exposed to introductory statistics and regression analysis. The authors' experience in statistical consulting is revealed throughout the book in the emphasis they put on communication, interaction with the science behind the data and the attention paid to detail and reporting of results. In terms of scope and objectives, this book is more far reaching and mentoring in data analysis than Venables and Ripley (2002) for S users in the 1990s. Keeping mathematical formulas to a minimum, illustrations of good data practices are tied to data from many different sources. Limitations and advantages of the statistical methods, critical perspectives on what can go wrong when assumptions are not met, and alternative paths of analyses with simulation, all are presented rigorously with an engaging narrative that flatters the readers' intelligence and imagination.

For researchers, this book is a data analyst's dream come true. In addition to providing the most modern base R commands to do standard statistical analyses, the authors take a very active role in suggesting how to start and how to proceed and end the analysis; they explain what the results mean under different circumstances and they discuss how to present the results to others. No matter where you open the book, you will find an excellent explanation of a data analysis problem from beginning to end. The R code blends very well with the text; although inserted within the narrative, code appears in a much smaller font than the regular text to diminish its intrusion with the explanations. The multidisciplinary datasets come from the **datasets** package in base R or the **DAAG** package of this book. A complete reference list of data sources is presented at the end of the book. A variety of exercises from basic to advanced help assess the understanding and learning of the material discussed in the main body of each chapter. The exercises are followed by a list of references suggesting

further reading on the statistical methodology, the issues arising in the data analysis that are not method related, and R. There is also a website with overheads for a course taught with this book, lab sessions, and other supplementary material. Overall, the whole package (book and website), is a great resource for applying statistical methods with R and for motivating them with a narrative that mixes the historical and the contemporary perspectives of data analysis.

The chapters in the book are very well tied together. After an introduction to R in Chapter 1, the reader could jump to Chapter 11 and learn more R, including package creation, document preparation with `Sweave` and **xtable**, and other housekeeping and data management features. Chapter 2 is very rich in suggestions to use statistical graphics to reveal different aspects of the data and on how to wisely transform the graphs for presentation. The book uses **lattice** plots, but not **ggplot2**. Chapter 3 explains why models play such a big role in statistics. Common probability models are motivated by their role as potential error component models. Random sampling and the central limit theorem are introduced in this chapter. Chapter 4 is a review of inference concepts: estimation and hypothesis testing for discrete or categorical data, including maximum likelihood estimation. Chapters 5–7 are about simple and multiple regression analysis, with and without standard assumptions. The remaining chapters contain more advanced examples of data analysis: generalized linear models (Chapter 8), time series models (Chapter 9), multi-level and repeated measures models (Chapter 10), tree-based classification and regression (Chapter 11), multivariate data exploration and discrimination (Chapter 12). As the rest of the book, Chapter 13 assumes that the reader knows about the methods presented, namely principal components and propensity scores. But nonetheless the chapter reviews them using a labor earnings case study and pointing out aspects of the data that may make the analysis invalid or complicated. As in other chapters, issues that arise in this type of modeling are listed and discussed. In all the chapters, the authors bring up how the data sets were obtained, design and data management issues. There are no chapters dedicated to Bayesian estimation or modern Markov chain Monte Carlo methods (except for a mention in Section 4.8.2).

Overall, the third edition of Maindonald and Braun's book should be in the library of every applied researcher in the physical, life and social sciences. The practice of data analysis is necessary to value the material learned in theory courses and to be able to read the applied literature. Often such practice is jeopardized by lack of understanding of the problems that one can run into when faced with a data set. Intimidation is also due to lack of good explanations of what software can do for you. This book succeeds in making the practice of statistics easier by avoiding those two problems. It shows how you can find the intricacies of a data analysis and the many different angles that a study presents using R. No questions are left unanswered after running the R code. The book is very well written and thinks for the reader, and the supplements are good for self-learning or instruction.

References

Dalgaard P (2008). *Introductory Statistics with R*. 2nd edition. Springer-Verlag, New York.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York.

Verzani J (2005). *Using R for Introductory Statistics*. Chapman & Hall/CRC, Boca Raton.

Reviewer:

Juana Sanchez
University of California, Los Angeles
Department of Statistics
8125 Math Sciences Building, Box 951554
Los Angeles, CA 90095-1554, United States of America
E-mail: jsanchez@stat.ucla.edu
URL: http://www.stat.ucla.edu/~jsanchez/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 36, Book Review 6 *Published:* 2010-09-21
September 2010