Journal of Statistical Software

August 2010, Volume 36, Book Review 2. *http://www.jstatsoft.org/*

Reviewer: Chee Kian Leong
SIM University

R in a Nutshell

Joseph Adler
O'Reilly, Sebastopol, 2010.
ISBN 978-0-596-80170-0. 611 pp. USD 49.99 (P).
`http://oreilly.com/catalog/9780596801700`

From its humble beginning in 1996, R has rapidly established itself as the lingua franca of statisticians worldwide. In recent years, the open source software has gone beyond the confines of academia and rapidly gained momentum in business analytical applications. It is perhaps a sign of R's increasing popularity as a programming language and statistical computing environment of choice that prompts a mainstream publisher O'Reilly to publish "R in a Nutshell".

The book is organized into four parts. The first part, consisting of four chapters, deals with the basics of R. In Chapter 1, readers gain a good understanding of how to install R for the main operating systems: Windows, Max OS X and Linux/Unix systems. Chapter 2 discusses the nitty-gritty of the user interface and here the reader will find it interesting and very useful to learn that R can be run from inside Microsoft Excel 2007. A short tutorial of R in Chapter 3 is illustrated with many practical examples. The strength of R is in its exponentially increasing number of packages and learning how to leverage on these packages is the subject of Chapter 4.

Part II of the book devotes seven chapters to the R language. Since this part is intended for those who are enthusiastic to learn the technicalities of R as a programming language, readers who are only interested in building some statistical models with data could probably skim this part on first reading. An overview of the language is provided in Chapter 5 while Chapters 6 to 9 describe the R syntax, objects, symbols and environment and functions. Support for object-oriented programming is discussed in Chapter 10. The materials covered very similar grounds to Braun and Murdoch (2007). What differentiates the book is Chapter 11 which delves on how to deliver performance when analyzing very large data sets, which are common in business analytical applications involving data warehouses. The discussion is necessarily brief and this reflects the challenges of massive data sets facing the R development community. On the other hand, the author could have devoted space (perhaps in a future edition) for Monte Carlo simulation, matrix and numerical computation, as these are part and parcel of a good foundation in statistical programming.

GIGO (garbage in, garbage out) is an adage that data analysts should be wise enough to heed. Data preparation and visualization in the preliminary stages provide the data audit for data understanding and data quality for statistical analysis. Thus, the author has rightly devoted Part III to working with data in R. Because the data available for analysis may come in different formats and from many databases, learning how to import and export data is important and well-documented in Chapter 12. A nifty feature is the real time retrieval of data from a single URL and this is illustrated using an example on fetching the closing price of the S&P500 index from Yahoo! Finance. Data extracted and imported into R are seldom clean unless they are part of some pedagogical texts. Hence, data preparation is critical. Unfortunately, data preparation is seldom (if ever) discussed in statistical texts or books on R. Consequently, Chapter 13 on data preparation is probably worth the price of the book alone: the reader will be pleasantly surprised by the embarras de richesses of data preparation techniques as implemented in R, including data transformation, data binning and data cleaning. The next two chapters present an overview of data visualization techniques, with the basics in Chapter 14 and an extensive explanation of lattice graphics in Chapter 15. As always, the presentation is enlivened by excellent real world examples, such as the San Francisco Real Estate Prices data set.

Once the data are sufficiently explored and prepared, statistical modeling can take place. This is the subject of Part IV, composed of nine chapters. Chapters 16 to 19 contain staples that are de rigueur in most statistics textbook such as correlation and covariance, probability distributions, principal components analysis, factor analysis, experimental design, power tests, *t* tests, ANOVA test and other statistical tests for both continuous and discrete data. An exhaustive survey of regression methods is provided in Chapter 20. What distinguishes this chapter is the inclusion of semiparametric and nonparametric methods as well machine learning algorithms for regression. Most machine learning techniques (such as tree models, neural network and support vector machines) are typically employed for classification models (the subject of Chapter 21) and their applications in regression are rarely mentioned in texts on data mining or machine learning. The discussion is enriched by well-chosen examples. Furthermore, the author makes a special effort to interpret the results of the modeling instead of leaving the readers to figure out the results for themselves. Readers who are into data mining or business analytics will be pleased with Chapters 21 and 22. Chapter 21 covers data mining techniques for classification or predictive modeling while Chapter 22 techniques for association or market basket analysis and clustering. The popular algorithms are present: logistic regression, linear discriminant, classification tree, neural network and support vector machines and also included are succinct exposition on ensemble methods such as bagging, boosting and random forests as well as silhouette plot for evaluating clustering models. A glaring omission is Bayesian classifiers such as Naives Bayes and Bayesian networks or graphical models but this is probably due to the lack of a well-developed package for such analyses. The same cannot be excused of the time series coverage in Chapter 23, which stands out for its brevity (six pages) in spite of the large number of packages for time series analysis and forecasting (such as the **Rmetrics** packages (Würtz 2010). As far as time series models are concerned, the state of the art in the book stops with ARIMA models. Since time series analysis practically dominates econometric analysis and forecasting is an integral part of business and financial data analysis, the lacuna should and could be addressed in a future edition by enriching the content on GARCH, state-space and long-memory models as well as Fourier and wavelet analysis. The final chapter focuses on the **Bioconductor** project and will appeal to analysts working in bioinformatics. The chapter begins with an example based on a data set from the Gene Expression Omnibus (GEO) website and the rest of the chapter basically expands on this example. The author wisely adds a section "Where to Go Next" with references to resources outside the Bioconductor project, vignettes, courses and books, though similar sections should probably be useful for other chapters in Part II, III and IV.

The book is intended to be a concise but comprehensive desktop reference to R. To this end, the author Joseph Adler has succeeded more than admirably. In about 600 pages, the book covers a lot of ground and the presentation is succinct with many well-chosen and well-illustrated examples. Above all, the author should be commended on the excellent and comprehensible codes, which should provide a benchmark for other authors intending to offer R codes as part of their books. In conclusion, the book deserves the strongest recommendation as a comprehensive resource for learning R and more importantly, as a highly practical guide on how to conduct data analysis professionally.

References

Braun W, Murdoch D (2007). *A First Course in Statistical Programming with R*. Cambridge University Press, Cambridge.

Würtz D (2010). "**Rmetrics**: An Environment for Teaching Financial Engineering and Computational Finance with R." URL http://www.Rmetrics.org/.

Reviewer:

Chee Kian Leong
SIM University
School of Business
535A Clementi Road
Singapore 599490
E-mail: ckleong@unisim.edu.sg

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/

Volume 36, Book Review 2 *Published:* 2010-08-19
August 2010