Journal of Statistical Software

April 2011, Volume 40, Book Review 1. http://www.jstatsoft.org/

Reviewer: Dirk Eddelbuettel
Debian Project

R in a Nutshell

Joseph Adler
O'Reilly, Sebastopol, 2009.
ISBN 978-0-596-80170-0. 640 pp. USD 49.99.
http://oreilly.com/catalog/9780596801717

Introduction

R in a Nutshell by Adler is an impressively comprehensive introduction to R. It also marks the first serious effort by O'Reilly, a publisher long known for its focus on programming titles, particularly using open source languages, to garner a corner of the rapidly expanding market for R books. This review will provide some brief comments to assess if (and if so, how well) the author and the publisher have succeeded in presenting R.

Covering a language and environment a rich as R in a single book or article is certainly challenging. To address the wide range in topics, the book runs to over 600 pages, organized into four parts and 24 chapters:

Part One discusses R basics in four chapters on the installation, the user interface, a first short tutorial, as well as packages.

Part Two introduces the R language in seven chapters starting with a language overview, followed by R syntax, R objects, symbols and the environment, functions, object-oriented programming, as well as high-performance R.

Part Three covers working with data in four chapters on data input/output, data preparation, (base) graphics, and the **lattice** graphics package.

Part Four concentrates on statistics using R in nine chapters covering material stretching from data analysis, probability distributions, statistical tests, power tests, regression, classification, machine learning, time-series analysis to bioinformatics.

This is suitable split, and the book manages to address several possible audiences at once. New users will find help for their first installation and initial steps, while more advanced users can brush up on particular topics, language aspects, programming modes or modeling topics.

This reviewer will not pretend to have read all 600 pages with the same attention to detail. However, several of the chapters that received closer scrutiny passed with ease—and just some minor issues require comments. We will cover the positive aspects first before listing some points which could be improved upon for a second edition.

Good things

Adler clearly knows R. To mention just one example, recommending environments for data lookup (p. 136) is a very good yet too rarely made point. More generally, the selection of chapters and topics is appropriate and shows reasonable balance between using R for data analysis, modeling, or for 'programming with data' (to quote Chambers).

Adler also knows modeling. The fourth part of the book which covers statistical modeling deserves a special mention for its treatment of modern approaches not typically seen in introductory texts. The regression chapter starts with the basic linear model, and uses this to discuss key aspects of modeling with R such as use of fit objects, accessor functions, summary and prediction methods, accessing goodness of fit, updating, more technical details and a brief reminder on the modeling assumptions. Adler then continues with robust and resistant regression, subset selection and shrinkage methods, ridge and least angle regression / lasso to principal component regression and partial least squares. This is followed by nonlinear models including generalized linear models, nonlinear least squares, survival models, smoothing and an entire section on machine-learning-based regression models such as regression trees, bagging and boosting, random forests and multivariate adaptive regression splines, neural nets, projection pursuit and generalized additive models and support vector machines. It is a fairly impressive and complete overview. This is followed by a chapter on classification (using linear methods, LDA, k-nearest neighbors, tree methods, neural nets, SVMs, and random forests) and a short chapter on machine learning which covers association rules and clustering.

The introduction of R as a programming language is also successful. Language constructs, syntax, object types, environments and functions are covered with the right balance between sufficient detail and required brevity; the discussion on object-oriented programming is also suitable, even if a little unusual in introducing S4 classes before S3 classes. The third part devoted to working with data is also important, and well executed. It is bound to contain helpful tips for novices as well as more experienced R users.

The provision of the corresponding CRAN (Comprehensive R Archive Network) package **nutshell** must also be applauded. It comprises 64 megabytes of data which allows readers to easily recreate many of the excellent examples provided throughout the book.

Lastly, how can one not love a book which illustrates a topic such as autocorrelation and possible errors in extrapolation with an xkcd comic (http://xkcd.com/605/) as Adler does with Figure 23.1?

Not so good things

Many of the topics touched upon in just subsections have received full book-length treatment elsewhere, so a lack of detail throughout the book—while understandable given the breadth of scope—is at times frustrating especially as further references are sometimes given (e.g., for queuing theory on p. 136) but more generally omitted.

Reviewers will always have preferences for particular topics they like to see more broadly treated. Mine skew towards high-performance and parallel computing with R. That last topic receives a short section in the chapter on high-performance R. Unfortunately, the parallel computing discussion covers only offerings by one particular vendor yet earlier work such as the packages **snow** and **Rmpi** are not discussed at all. That is a pretty odd omission, and doubly so given the recommendations in the survey paper of Schmidberger *et al.* (2009). The vendor also gets special treatment in the focus on its non-free version and its MKL extension for accelerated BLAS—while the comparable offerings (GotoBLAS) discussed in the R installation manual are not even mentioned. Lastly, that `foreach`—a useful extension—is treated in more detail than the core language function `for` itself seems inappropriate.

The omission of a discussion of how to write a simple package is surprising. Likewise, the lack of coverage of compiled extensions via `.C()` and `.Call()` (even though building R itself is discussed in a slightly odd place in the context of high-performance R) can only be regretted. On the other hand, devoting an entire chapter to **Bioconductor** is a little odd given that it is 'merely' a domain-specific extension and application of R—but given how very successful and influential **Bioconductor** is, the choice is defensible. Some users, though, may have prefered a discussion of, say, **ggplot2** or seen the pages used to discuss extending R or package writing.

A more serious concern is the lack of editorial oversight. As noted above, Adler excels in integrating useful examples provided via a companion package **nutshell** on CRAN which makes the data readily available to the R user. But for example the data set on housing was not compiled by 'Schiller' (a long-deceased poet?) but rather by the economist Robert Shiller. Similarly, on pp. 20–21, the same example is needlessly repeated twice; p. 375 lists an incomplete formula; one has to load **nnet** before using the `multinom` function and so on. Also, the split between Chapters 20 (60 pages, or roughly ten percent of the book), 21 (18 pages) and 22 (8 pages) is uneven; maybe a cut into more standard regression methods and those based on machine learning approaches would have been an alternative. Lastly, a minor but irritating point is the poor choice of defaults for charts leading a poor 'ink-to-paper' ratio which something as simple as a suitable global `par()` setting could have improved.

Overall

Adler (and O'Reilly) have succeeded with this first R title which can serve simultaneously as a reference manual kept near one's workstation, an occasional refresher on some topics or even a first introduction for some readers. The range in examples is excellent, and the support via the corresponding package is to be commended. *R in a Nutshell* is a very useful resource for new and experienced R users alike.

References

Schmidberger M, Morgan M, Eddelbuettel D, Yu H, Tierney L, Mansmann U (2009). "State of the Art in Parallel Computing with R." *Journal of Statistical Software*, **31**(1), 1–27. URL http://www.jstatsoft.org/v31/i01/.

Reviewer:

Dirk Eddelbuettel
Debian Project
Chicago, IL, United States of America
E-mail: edd@debian.org
URL: http://dirk.eddelbuettel.com/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 40, Book Review 1 *Published:* 2011-04-04
April 2011