Journal of Statistical Software

August 2012, Volume 50, Book Review 1. http://www.jstatsoft.org/

Reviewer: Dirk Eddelbuettel
Debian Project

R for Dummies

Andrie de Vries, Joris Meys
John Wiley & Sons, Chichester, 2012.
ISBN 978-1-119-96284-7. 408 pp. USD 29.99.
http://www.wiley.com/WileyCDA/WileyTitle/productCd-1119962846.html

---

R for Dummies by Andrie de Vries and Joris Meys poses some challenges for this reviewer who will not deny a dislike of the "For Dummies" series, its commercial success notwithstanding. The fluorescent yellow cover, the rather loud typesetting conventions, the forced-humour of its comic style and, last but not least, the very concept of "dumbing down" a given topic worthy of serious book length discussions has been irritating to me for the more than two decades that the series has graced bookstores. Now the series reached the topic of statistical programming as well as interactive investigation, modeling and visualization in the form of a book on R. And lo and behold, this book by De Vries and Meys is actually a success.

In their book, De Vries and Meys offer a nice logical progression that differs from other introductions to R. By delaying the more instant gratification of graphics and visualization until the end – after data manipulation and transformation has been covered in a reasonably thorough manner – the authors rightly stress the point that the majority of data work often is data preparation. The book is divided into six parts providing a total of twenty chapters. In Part 1, a first general introduction provides a high-level introduction through three chapters covering a big picture overview, initial explorations and some fundamentals. This is followed by a second part introducing basics of operators, vectors, arithmetic, initial reading/writing of data, factor types, date types and multidimensional objects including data frames. Part 3, also spanning four chapters, is unusual for an introductory book and goes straight to programming with R. It introduces functions, control flow language elements, debugging approaches and the help systems. Part 4, again over four chapters, is focused on data and covers input/output, manipulation and processing, data description and summaries, as well as an introduction to statistical tests and estimation. Part 5 contains three chapters on graphics, covering base graphics, **lattice** and **ggplot2**. The final Part 6 contains two chapters that bear the imprint of the publishing series: a first covers "things otherwise done in **Excel**" before a last chapter provides the unavoidable list of ten things that the "Dummies" series forces into each title, for better or worse. Finally, an appendix covers initial installation and configuration.

The focus on programming and scripting is welcome, and is well executed. It is clear that De Vries and Meys know their material – and that they have a high degree of enthusiasm about

their R. The range of topics is appropriate for the audience, and helpful tips are provided throughout the book. The emphasis on **RStudio** as a unifying programming environment is also a good choice, and could have been made stronger by de-emphasizing Windows-specific aspects as one of the many strengths of R is its platform independence.

A sore point, however, is the number of small inaccuracies. Page 12 states that **foreign** is part of base R which is not correct: while it is distributed with R, it is still only a recommended package. Or on page 27 it is suggested to "type `?paste`" to get help for the `save()` command. On page 213, the **RPostgresSQL** package is wrongly named without the initial letter. Page 10 calls S a "commercial programming language" which is incorrect (S-PLUS is/was a commercial implementation). On page 208, it is incorrectly stated that `stringsAsFactors` defaults to `FALSE` when the opposite is true (as shown also on the following pages). And sometimes the error is more grave: Page 11 refers to limitations of the GPL as requiring release of code in case users "change or redistribute" which is false: it is only on code change *and* redistribution that these terms bind – a crucial difference.

But large parts of the book are indeed well done, and introduce R in a good and useful way. For example, the discussion in Chapter 3 on variable naming, and the suggestion to use descriptive names, is very good. The authors rightly point out how one of the strengths of R is the strong support from the user and developer community, as well as the availability of CRAN. The chapter on dates and times – an under-appreciated strength of R – is very good. Yet it too fails to mention the fractional second resolution, an underused `POSIXct` strength, and it might also have pointed out that the `POSIXlt` range for months is from zero to eleven, a frequent source of confusion when one to twelve is assumed.

R for Dummies may well break new ground, and introduce R to new audiences. The pricing is aggressive and should help the book to find its way into the hands of a large number of students, data analysis practitioners as well as researchers. They will find a well-written and easy-to-read introduction to the language and environment – if they can overcome any initial bias against a *Dummies* title as this reviewer did.

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 50, Book Review 1 *Published:* 2012-08-06
August 2012