Review of "The R Book"

Michael J. Crawley, Wiley, 2007

by Friedrich Leisch

The back cover of this physically impressive 1000-page volume advertises it as *"...the first comprehemsive reference manual for the R language..."* which *"...introduces all the statistical models covered by R..."*. Considering (a) that the R Core Team considers its own language manual (R Development Core Team, 2007b) a draft, and only the source code the ultimate reference in more cases than we like, and (b) the multitude of models implemented by R packages on CRAN or Bioconductor, I thought I would be in for an interesting read.

The book has 27 chapters. Chapters 1–8 give an introduction to R, starting where to obtain and how to install the software, describing the language, data input and data manipulation, graphics, tables, mathematical calculations, and classical statistical tests. Chapters 9–20 on statistical modelling form the main part of the book with a detailed coverage of the linear regression model and its extensions like GLMs, GAMs, mixed effects and non-linear least squares. Chapters 20–27 show "other" topics like trees, time series, multivariate and spatial statistics, survival analysis, using R for simulation models and low-level graphics commands.

The preface states that the book is *"aimed at beginners and intermediate users"* and can be used *"as a text ... as well as a reference manual"*. I find that the book in its present form is not optimal for either purpose. The first section on "getting started" has a few minor problems, like using many functions without quoted character arguments. In some cases this is a matter of style (like `library(foo)` or `help(foo)`), but some instances simply do not work (`find(foo)`, `apropos(foo)`). Packages are often called libraries (which is a different thing), input lines can be longer than 128 characters (the current limit is 8192), and recommending MS Word as a source code editor is at least debatable even for Windows users. I personally find the R code throughout the book hard to read: it is typeset in a proportional font, uses no spaces around the assignment operator `<-`, no line indentation for nested code blocks, and path names sometimes contain erroneous spaces, especially after backslashes.

The chapter on "essentials of the R language" gives an introduction to the language and many non-statistical functions like string processing and regular expressions. What I found very confusing is the lack of clear structure. The book uses only one level of numbering (chapters), and this chapter is 100 pages long. E.g., on page 47 there are two headings: "The match function" and "Writing functions in R". Both seem to have the same font size and hence are on equal level. However, as is to be expected given the two topics, the section on the match function is 2 paragraphs long, how to write functions takes the next 20 pages, with many intermezzos and headings in two different sizes. The author also jumps around a lot, many concepts are discussed or introduced as a side note for a different theme, and it is often unclear where examples end. E.g., how formal and actual arguments are matched in a funcion call is the first paragraph in the section on "Saving data to disc". All of this will be confusing for beginners and makes it hard to use the book as a reference manual.

In the chapter on mathematics a dozen pages is used on introducing the OLS estimate (typo in several equations: $\hat{\beta} = X'X - 1X'y$), including a stepwise implementation of (the correct version of) this formula. Although the next page in the book starts with solving linear equations via `solve()`, it is not even mentioned that it is numerically not the best idea to compute regression coefficients using the formula above.

The quality of the book increases considerably in the chapters on statistical modelling. A minor drawback is that it sometimes gives the impression that linear models are the only ones available, even discriminant analysis is not considered a model, because response variables cannot be multi-level categorical according to the cookbook recipe on page 324. However, there is a nice general introduction to statistical modelling and model selection, and linear modelling is covered in depth with many examples.

Once the author leaves the territory of linear models (and their extensions), quality decreases again. The chapter on trees uses package **tree**, although even the author of **tree** recommends using package **rpart** (Venables and Ripley, 2002, p. 266). The chapter on multivariate statistics basically recommends not doing multivariate statistics at all, because one is too likely to shoot oneself into the foot.

In summary, the book fails to meet the high expectations that the title and cover texts raise. In this review I could list only a selection of problems I found, and of course there are good things too, like the detailed explanation on how to enter data into a spreadsheet to form a proper data frame. The book is definitely not a reference manual for the R system or R language, but a book on applied linear modelling with (many pages of) lower-quality additional material to give the impression of universal coverage. There are better introductionary books for beginners, and Venables and Ripley (2002) is still "the R book" when it comes to a reference text for applied statistics.

A (symptomatic) end note: The author gives detailed credit to the R core team and wider R com-

munity in the acknowledgements (thanks!). On page one he recommends the `citation()` function to users to give credit to developers (yes!), however he seems not to have used the function too often, because R Development Core Team (2007a,b) and many others are missing from the references, which cover only 4 of 1000 pages.

Bibliography

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007a. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

R Development Core Team. *R Language Definition*. R Foundation for Statistical Computing, Vienna, Austria, 2007b. URL `http://www.R-project.org`. ISBN 3-900051-13-5.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S. Fourth Edition*. Springer, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4/`. ISBN 0-387-95457-0.

Friedrich Leisch
Ludwig-Maximilians-Universität München, Germany
`Friedrich.Leisch@R-project.org`

Changes in R 2.6.0

by the R Core Team

User-visible changes

- `integrate()`, `nlm()`, `nlminb()`, `optim()`, `optimize()` and `uniroot()` now have `...` much earlier in their argument list. This reduces the chances of unintentional partial matching but means that the later arguments must be named in full.
- The default type for `nchar()` is now `"chars"`. This is almost always what was intended, and differs from the previous default only for non-ASCII strings in a MBCS locale. There is a new argument `allowNA`, and the default behaviour is now to throw an error on an invalid multibyte string if `type = "chars"` or `type = "width"`.
- Connections will be closed if there is no R object referring to them. A warning is issued if this is done, either at garbage collection or if all the connection slots are in use.

New features

- `abs()`, `sign()`, `sqrt()`, `floor()`, `ceiling()`, `exp()` and the gamma, trig and hyperbolic trig functions now only accept one argument even when dispatching to a `Math` group method (which may accept more than one argument for other group members).
- `abbreviate()` gains a `method` argument with a new option `"both.sides"` which can make shorter abbreviations.
- `aggregate.data.frame()` no longer changes the group variables into factors, and leaves alone the levels of those which are factors. (Inter alia grants the wish of PR#9666.)
- The default `max.names` in `all.names()` and `all.vars()` is now `-1` which means unlimited. This fixes PR#9873.
- `as.vector()` and the default methods of `as.character()`, `as.complex()`, `as.double()`, `as.expression()`, `as.integer()`, `as.logical()` and `as.raw()` no longer duplicate in most cases where the object is unchanged. (Beware: some code has been written that invalidly assumes that they do duplicate, often when using `.C/.Fortran(DUP = FALSE)`.)
- `as.complex()`, `as.double()`, `as.integer()`, `as.logical()` and `as.raw()` are now primitive and internally generic for efficiency. They no longer dispatch on S3 methods for `as.vector()` (which was never documented). `as.real()` and `as.numeric()` remain as alternative names for `as.double()`.

  `expm1()`, `log()`, `log1p()`, `log2()`, `log10()`, `gamma()`, `lgamma()`, `digamma()` and `trigamma()` are now primitive. (Note that `logb()` is not.)

  The `Math2` and `Summary` groups (`round`, `signif`, `all`, `any`, `max`, `min`, `sum`, `prod`, `range`) are now primitive.

  See under Section "**methods** Package" below for some consequences for S4 methods.
- `apropos()` now sorts by name and not by position on the search path.