Journal of Statistical Software

December 2009, Volume 32, Book Review 2. http://www.jstatsoft.org/

Reviewer: Dirk Eddelbuettel
Debian Project

Computational Statistics: An Introduction to R

Günther Sawitzki
Chapman & Hall/CRC, Boca Raton, FL, 2009.
ISBN 978-1-4200-8678-2. 251 pp. USD 79.95.
http://sintro.R-Forge.R-project.org/

Computational Statistics: An Introduction to R by Günther Sawitzki offers a fresh perspective on teaching statistics. By not concerning himself with the somewhat standard 'first course' covering basic statistics, Sawitzki is free to concentrate on material suitable for a short and compact course covering *computational statistics*—an increasingly important topic for anyone dealing with statistics.

The books introduces itself with a nice and suitably detailed yet still short overview. Several pages cover the origins and core aspects of the R systems. It situates the language nicely: *"the S language using the R implementation"* and *"developed for practical work in statistics"* as well as with *"usability given priority over abstract design principles"*. This is well done, besides the minor surprise of mentioning a commercial variant called S-PLUS sold by a company that had essentially ceased to exist by the time the book went to press.

Following this introduction, the book is divided into four core chapters. They cover, respectively, (i) basic data analysis: one-sample regression and distributions, (ii) regression, (iii) comparisons: two-sample problems and comparison of distributions, and (iv) multivariate analysis: dimensions $1, 2, \ldots, \infty$ which is followed by a lengthy appendix with more detail on the R language and system. This is described as suitable for a four-day course, or a week if used with additional exercises.

The book introduces its topics and the corresponding methodologies well. Section 1.1 has a nice overview of basic R conventions. The discussion of pseudo-versus-real random-number generators in Section 1.2 is both nice and appropriate this early in the text. This leads to a nice first case study entitled 'distribution diagnostics' of histograms and distribution testing. Section 1.5 is useful, but unfortunately repeats the old sermon of 'replacing loops with more efficient constructs' as if the performance of `apply` were still that much different from loops. However, this does not take away from the nice achievement of introducing many language constructs along one lengthy case study that even includes Monte Carlo simulations.

One purely editorial feature struck me as inappropriate. In several instances, Sawitzki has chosen to include an entire help page from the R system. On-line help in an interactive

programming environment and a learned book such as this serve different audiences, and have different strengths. Mingling them as was done here does both a disservice. And it does not help when the editing is done sloppily: in three instances on pages 74, 123 and 158, the bold markup formatting has lost its closing brace leaving much longer segments in bold typeface which is both unprofessional and unacceptable for a book in this price range.

Section 2 continues with another case study deepening the examination of regression. Several imporant sub-topics are mentioned and the discussion is reasonably complete without being exhaustive at under fifty pages. Some corners have been cut: generalized linear models are mentioned, albeit without stating any transformations which may be an omission that can be remedied in subsequent printing. Similarly the discussion of R *classes* and programming does not mention either S3 or S4 despite these being the principal class implementations in R.

Section 3 revisits comparisons and some topics of the first chapter before discussing statistical power which is also examined via simulation. Section 4 on higher dimensions and multivariate analysis does a nice job in introducing the famous iris data set by actually going back to some biologically motivated variables. There is also a very useful discussion of the general difficulties in detecting higher order dependence in data. Lastly, the final case study on a model for body fat and BMI is rather nicely done.

Besides the runaway boldface mentioned earlier, a number of small errors muddy the overall picture. Page 16 refers to `norm` where it meant `rnorm`. Page 64 has 'ivertibe' instead of 'invertible'. Page A-195 has `lsf.str()` where it should be `ls.str()`. Page A-215 has 'automatiic' with two i, and page A-221 inexplicably switches font sizes in the table.

Apart from these small issues, the book is well put togther and quite enjoyable for its purpose of serving a small course on computational statistics. Given the increasing importance of this topic and the need for extended instruction on it, this book should find large and receptive audiences.

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 32, Book Review 2 *Published:* 2009-12-18
December 2009