Bibliography

Hornik, K. (2003): *The R FAQ*, Version 1.7-3. http://www.ci.tuwien.ac.at/~hornik/R/ ISBN 3-901167-51-X. 26

R Development Core Team (2003a): *R Data Import/Export*. URL http://CRAN.R-project.org/manuals.html. ISBN 3-901167-53-6. 26

R Development Core Team (2003b): *R Installation and Administration*. URL http://CRAN.R-project.org/manuals.html. ISBN 3-901167-52-8. 26

R Development Core Team (2003c): *R Language Definition*. URL http://CRAN.R-project.org/manuals.html. ISBN 3-901167-56-0. 26

R Development Core Team (2003d): *Writing R Extensions*. URL http://CRAN.R-project.org/manuals.html. ISBN 3-901167-54-4. 26

Venables, W. N. and Ripley, B. D. (2000): *S Programming*. New York: Springer-Verlag. 27

Venables, W. N. and Ripley, B. D. (2002): *Modern Applied Statistics with S*. New York: Springer-Verlag, 4th edition. 27

Venables, W. N., Smith, D. M., and the R Development Core Team (2003): *An Introduction to R*. URL http://CRAN.R-project.org/manuals.html. ISBN 3-901167-55-2. 26

Uwe Ligges
Fachbereich Statistik, Universität Dortmund, Germany
ligges@statistik.uni-dortmund.de

Book Reviews

Michael Crawley: Statistical Computing: An Introduction to Data Analysis Using S-Plus

John Wiley and Sons, New York, USA, 2002
770 pages, ISBN 0-471-56040-5
http://www.bio.ic.ac.uk/research/mjcraw/statcomp/

The number of monographs related to the S programming language is sufficiently small (a few dozen, at most) that there are still great opportunities for new books to fill some voids in the literature. *Statistical Computing* (hereafter referred to as *SC*) is one such text, presenting an array of modern statistical methods in S-Plus at an introductory level.

The approach taken by author is to emphasize "graphical data inspection, parameter estimation and model criticism rather than classical hypothesis testing" (p. ix). Within this framework, the text covers a vast range of applied data analysis topics. Background material on S-Plus, probability, and distributions is included in the text. The traditional topics such as linear models, regression, and analysis of variance are covered, sometimes to a greater extent than is common (for example, a split-split-split-split plot experiment and a whole chapter on ANCOVA). More advanced modeling techniques such as bootstrap, generalized linear models, mixed effects models and spatial statistics are also presented. For each topic, some elementary theory is presented and usually an example is worked by hand. Crawley also discusses more philosophical issues such as randomization, replication, model parsimony, appropriate transformations, etc. The book contains a very thorough index of more than 25 pages. In the index, all S-Plus language words appear in bold type.

A supporting web site contains all the data files, scripts with all the code from each chapter of the book, and a few additional (rough) chapters. The scripts pre-suppose that the data sets have already been imported into S-Plus by some mechanism–no explanation of how to do this is given–presenting the novice S-Plus user with a minor obstacle. Once the data is loaded, the scripts can be used to re-do the analyses in the book. Nearly every worked example begins by `attach`ing a data frame, and then working with the variable names of the data frame. No `detach` command is present in the scripts, so the work environment can become cluttered and even cause errors if multiple data frames have common column names. Avoid this problem by quitting the S-Plus (or R) session after working through the script for each chapter.

SC contains a very broad use of S-Plus commands to the point the author says about `subplot`, "Having gone to all this trouble, I can't actually see why you would ever want to do this in earnest" (p. 158). Chapter 2 presents a number of functions (`grep`, `regexpr`, `solve`, `sapply`) that are not used within the text, but are usefully included for reference.

Readers of this newsletter will undoubtedly want to know how well the book can be used with R (version 1.6.2 base with recommended packages). According to the preface of *SC*, "The computing is presented in S-Plus, but all the examples will also work in the freeware program called R." Unfortunately, this claim is too strong. The book's web site does contain (as of March, 2003) three modifications for using R with the book. Many more modifications are needed. A rough estimate would be that 90% of the S-Plus code in the scripts needs no modification to work in R. Basic calculations and data manipulations are nearly identical in S-Plus and R. In other cases, a slight change is necessary (`ks.gof` in S-Plus, `ks.test` in R), alternate functions must be used (`multicomp` in S-Plus, `TukeyHSD` in R), and/or additional packages (`nlme`) must be loaded into R. Other topics (bootstrapping, trees) will require more extensive code modification efforts and some functionality is not available in R (`varcomp`, `subplot`).

Because of these differences and the minimal documentation for R within the text, users inexperienced with an S language are likely to find the incompatibilities confusing, especially for self-study of the text. Instructors wishing to use the text with R for teaching should not find it too difficult to help students with the differences in the dialects of S. A complete set of online complements for using R with *SC* would be a useful addition to the book's web site.

In summary, *SC* is a nice addition to the literature of S. The breadth of topics is similar to that of *Modern Applied Statistics with S* (Venables and Ripley, Springer, 2002), but at a more introductory level. With some modification, the book can be used with R. The book is likely to be especially useful as an introduction to a wide variety of data analysis techniques.

Kevin Wright
Pioneer Hi-Bred International
Kevin.Wright@pioneer.com

Peter Dalgaard: Introductory Statistics with R

Springer, Heidelberg, Germany, 2002
282 pages, ISBN 0-387-95475-9
http://www.biostat.ku.dk/~pd/ISwR.html

This is the first book, other than the manuals that are shipped with the R distribution, that is solely devoted to R (and its use for statistical analysis). The author would be familiar to the subscribers of the R mailing lists: Prof. Dalgaard is a member of R Core, and has been providing people on the lists with insightful, elegant solutions to their problems. The author's R wizardry is demonstrated in the many valuable R tips and tricks sprinkled throughout the book, although care is taken to not use too-clever constructs that are likely to leave novices bewildered.

As R is itself originally written as a teaching tool, and many are using it as such (including the author), I am sure its publication is to be welcomed by many. The targeted audience of the book is "nonstatistician scientists in various fields and students of statistics". It is based upon a set of notes the author developed for the course Basic Statistics for Health Researchers at the University of Copenhagen. A couple of caveats were stated in the Preface: Given the origin of the book, the examples used in the book were mostly, if not all, from health sciences. The author also made it clear that the book is not intended to be used as a stand alone text. For teaching, another standard introductory text should be used as a supplement, as many key concepts and definitions are only briefly described. For scientists who have had some statistics, however, this book itself may be sufficient for self-study.

The titles of the chapters are:

1. Basics
2. Probability and distributions
3. Descriptive statistics and graphics
4. One- and two-sample tests
5. Regression and correlation
6. ANOVA and Kruskal-Wallis
7. Tabular data
8. Power and computation of sample sizes
9. Multiple regression
10. Linear Models
11. Logistic regression
12. Survival analysis

plus three appendices: Obtaining and installing R, Data sets in the `ISwR` package (help files for data sets the package), and Compendium (short summary of the R language). Each chapter ends with a few exercises involving further analysis of example data. When I first opened the book, I was a bit surprised to see the last four chapters. I don't think most people expected to see those topics covered in an introductory text. I do agree with the author that these topics are quite essential for data analysis tasks for practical research.

Chapter one covers the basics of the R language, as much as needed in interactive use for most data analysis tasks. People with some experience with S-PLUS but new to R will soon appreciate many nuggets of features that make R easier to use (e.g., some vectorized graphical parameters such as `col`, `pch`, etc.; the functions `subset`, `transform`, among others). As much programming as is typically needed for interactive use (and may be a little more; e.g., the use of `deparse(substitute(...)` as default argument for the title of a plot) is also described.

Chapter two covers the calculation of probability distributions and random number generation for simulations in R. Chapter three describes descriptive statistics (mean, SD, quantiles, summary tables, etc.), as well as how to graphically display summary plots such as Q-Q plots, empirical CDFs, boxplots, dotcharts, bar plots, etc. Chapter four describes the t and Wilcoxon tests for one- and two-sample comparison as well as paired data, and the F test for equality of variances.

Chapter five covers simple linear regression and bivariate correlations. Chapter six contains quite a bit of material: from oneway ANOVA, pairwise comparisons, Welch's modified F test, Barlett's test for equal variances, to twoway ANOVA and even a repeated measures example. Nonparametric procedures are also described. Chapter seven covers basic inferences for categorical data (tests for proportions, contingency tables, etc.). Chapter eight covers power and sample size calculations for means and proportions in one- and two-sample problems.

The last four chapters provide a very nice, brief summary on the more advanced topics. The graphical presentations of regression diagnostics (e.g., color coding the magnitudes of the diagnostic measures) can be very useful. However, I would have preferred the omission of coverage on variable selection in multiple linear regression entirely. The coverage on these topics is necessarily brief, but do provide enough material for the reader to follow through the examples. It is impressive to cover this many topics in such brevity, without sacrificing clarity of the descriptions. (The linear models chapter covers polynomial regression, ANCOVA, regression diagnostics, among others.)

As with (almost?) all first editions, typos are inevitable (e.g., on page 33, in the modification to vectorize Newton's method for calculating square roots, `any` should have been `all`; on page 191, the logit should be defined as $\log[p/(1-p)]$). On the whole, however, this is a very well written book, with logical ordering of the content and a nice flow. Ample code examples are interspersed throughout the text, making it easy for the reader to follow along on a computer. I whole-heartedly recommend this book to people in the "intended audience" population.

Andy Liaw
Merck Research Laboratories
andy_liaw@merck.com

John Fox: An R and S-Plus Companion to Applied Regression

Sage Publications, Thousand Oaks, CA, USA, 2002
328 pages, ISBN 0-7619-2280-6
http://www.socsci.mcmaster.ca/jfox/Books/companion/

The aim of this book is to enable people with knowledge on linear regression modelling to analyze their data using R or S-Plus. Any textbook on linear regression may provide the necessary background, of course both language and contents of this companion are closely aligned with the authors own *Applied Regression, Linear Models and Related Metods* (Fox, Sage Publications, 1997).

Chapters 1–2 give a basic introduction to R and S-Plus, including a thorough treatment of data import/export and handling of missing values. Chapter 3 complements the introduction with exploratory data analysis and Box-Cox data transformations. The target audience of the first three sections are clearly new users who might never have used R or S-Plus before.

Chapters 4–6 form the main part of the book and show how to apply linear models, generalized linear models and regression diagnostics, respectively. This includes detailed explanations of model formulae, dummy-variable regression and contrasts for categorical data, and interaction effects. Analysis of variance is mostly explained using "Type II" tests as implemented by the `Anova()` function in package **car**. The infamous "Type III" tests, which are requested frequently on the `r-help` mailing list, are provided by the same function, but their usage is discouraged quite strongly in both book and help page.

Chapter 5 reads as a natural extension of linear models and concentrates on error distributions provided by the exponential family and corresponding link functions. The main focus is on models for categorical responses and count data. Much of the functionality provided by package **car** is on regression diagnostics and has methods both for linear and generalized linear models. Heavy usage of graphics, `identify()` and text-based menus for plots emphasize the interactive and exploratory side of regression modelling. Other sections deal with Box-Cox transformations, detection of non-linearities and variable selection.

Finally, Chapters 7 & 8 extend the introductory Chapters 1–3 with material on drawing graphics and S programming. Although the examples use regression models, e.g., plotting the surface of fitted values of a generalized linear model using `persp()`, both chapters are rather general and not "regression-only".

Several real world data sets are used throughout the book. All S code necessary to reproduce numerical and graphical results is shown and explained in detail. Package **car**, which contains all data sets used in the book and functions for ANOVA and regression diagnostics, is available from CRAN and the book's homepage.

The style of presentation is rather informal and hands-on, software usage is demonstrated using examples, without lengthy discussions of theory. Over wide parts the text can be directly used in class to demonstrate regression modelling with S to students (after more theoretical lectures on the topic). However, the book contains enough "reminders" about the theory, such that it is not necessary to switch to a more theoretical text while reading it.

The book is self-contained with respect to S and should easily get new users started. The placement of R *before* S-Plus in the title is not only lexicographic, the author uses R as the primary S engine. Differences of S-Plus (as compared with R) are clearly marked in framed boxes, the graphical user interface of S-Plus is not covered at all.

The only drawback of the book follows directly from the target audience of possible S novices: More experienced S users will probably be disappointed by the ratio of S introduction to material on regression analysis. Only about 130 out of 300 pages deal directly with linear models, the major part is an introduction to S. The online appendix at the book's homepage (mirrored on CRAN) contains several extension chapters on more advanced topics like bootstrapping, time series, nonlinear or robust regression.

In summary, I highly recommend the book to anyone who wants to learn or teach applied regression analysis with S.

Friedrich Leisch
Universität Wien, Austria
Friedrich.Leisch@R-project.org

Changes in R 1.7.0

by the R Core Team

User-visible changes

- `solve()`, `chol()`, `eigen()` and `svd()` now use LAPACK routines unless a new back-compatibility option is turned on. The signs and normalization of eigen/singular vectors may change from earlier versions.

- The 'methods', 'modreg', 'mva', 'nls' and 'ts' packages are now attached by default at startup (in addition to 'ctest'). The option `"defaultPackages"` has been added which contains the initial list of packages. See `?Startup` and `?options` for details. Note that `.First()` is no longer used by R itself.

  `class()` now always (not just when 'methods' is attached) gives a non-null class, and `UseMethod()` always dispatches on the class that `class()` returns. This means that methods like foo.matrix and foo.integer will be used. Functions `oldClass()` and `oldClass<-()` get and set the "class" attribute as R without 'methods' used to.

- The default random number generators have been changed to 'Mersenne-Twister' and 'Inversion'. A new `RNGversion()` function allows you to restore the generators of an earlier R version if reproducibility is required.

- Namespaces can now be defined for packages other than 'base': see 'Writing R Extensions'. This hides some internal objects and changes the search path from objects in a namespace. All the base packages (except methods and tcltk) have namespaces, as well as the recommended packages 'KernSmooth', 'MASS', 'boot', 'class', 'nnet', 'rpart' and 'spatial'.

- Formulae are not longer automatically simplified when `terms()` is called, so the formulae in results may still be in the original form rather than the equivalent simplified form (which may have reordered the terms): the results are now much closer to those of S.

- The tables for plotmath, Hershey and Japanese have been moved from the help pages (`example(plotmath)` etc) to `demo(plotmath)` etc.

- Errors and warnings are sent to stderr not stdout on command-line versions of R (Unix and Windows).

- The R_X11 module is no longer loaded until it is needed, so do test that `x11()` works in a new Unix-alike R installation.

New features

- `if()` and `while()` give a warning if called with a vector condition.

- Installed packages under Unix without compiled code are no longer stamped with the platform and can be copied to other Unix-alike platforms (but not to other OSes because of potential problems with line endings and OS-specific help files).

- The internal random number generators will now never return values of 0 or 1 for `runif()`. This might affect simulation output in extremely rare cases. Note that this is not guaranteed for user-supplied random-number generators, nor when the standalone Rmath library is used.

- When assigning names to a vector, a value that is too short is padded by character NAs. (Wishlist part of PR#2358)

- It is now recommended to use the 'SystemRequirements:' field in the DESCRIPTION file for specifying dependencies external to the R system.

- Output text connections no longer have a line-length limit.