Journal of Statistical Software

January 2013, Volume 52, Book Review 3. *http://www.jstatsoft.org/*

Reviewer: Ulrike Grömping
Beuth University of Applied Sciences Berlin

The R Primer

Claus Thorn Ekstrøm
Chapman and Hall/CRC, Boca Raton, FL, 2011.
ISBN 978-1-4398-6206-3. 299 pp. USD 39.95 (P).
`http://wwwu.rprimer.dk/`

A "primer" is supposed to be a book that covers very elementary needs. This book does so in a quite special way: It provides 142 problems with solutions, called "rules", grouped into the chapters "Data import" (includes export, 21 rules, 24 pages), "Manipulating data" (29 rules, 42 pages), "Statistical analyses" (45 rules, 130 pages), "Graphics" (30 rules, 56 pages) and "R" (17 rules, 22 pages, referring to general things like interfacing with the system, getting help etc.). Each chapter has a brief introductory section, before the first rule is presented.

The book efficiently addresses the many impediments against simply "go do it" for people who have already done statistical analyses with software other than R and want to quickly learn how to do the same things in R. Readers are assumed to be familiar with the basics of both statistics and R. Readers who are very new to R might be best off looking at Chapters 1, 2 and 5 first. In the first two chapters, the book does a good job on illustrating many kinds of importing and exporting data, as well as various data manipulation topics. For some of the latter (rules 2.1 to 2.4), the problem-solution approach seems a bit forced, I would have preferred a longer introductory section. But that's a matter of taste.

Overall, the book's rules provide useful starting points from which readers can easily find their way to further material. Generally, each rule is stand-alone, i.e., all necessary R steps for the task at hand are included. I very much like this feature, as well as the fact that the "See also" sections often refer readers to related rules. However, a reader who wants to implement an unfamiliar statistical method will usually have to consult additional statistical literature that has to be found elsewhere; it is one of the less likable features of the book that it provides very few external references only. On the plus side, this keeps the material succinct, which may have been the reason behind this decision. In many cases, the documentation of the R packages used in the book will give the appropriate references.

In the following, the diverse content of the book is illustrated by some specific examples. These cover some basic methods (contingency tables, the first few rules from the "Statistical analyses" chapter and related later rules) and a few more advanced methods (bootstrapping, partial least squares regression and ordinal logistic regression). Except for one case (ordinal logistic regression, where the code failed due to a change in **MASS**, Venables and Ripley 2002, since code creation in a relatively old version of R), the code given in the book worked without change and was easily understood.

- Contingency tables are shown in rule 2.17 ("Create a table of counts"). The rule nicely explains how tables are created with functions `table` and `ftable`. The "See also" section mentions possibilities for obtaining margins or proportions and mentions the alternative function `xtabs` for table creation. It would have been helpful to also mention function `addmargins` for adding margins to a table, and to also say something about missing value handling. Nevertheless, with the rule as a starting point, readers can find their way into the topic using the R help and its "See also" sections.

- The first rule of the statistics chapter shows a very specific way of creating a basic summary table for a data frame – function `stat.desc` from package **pastecs** (Ibanez, Grosjean, and Etienne 2012). There are various functions in contributed R packages for creating such tables, for example the functions `describe` from packages **Hmisc** (Harrell 2012) or **prettyR** (Lemon and Grosjean 2012). The function from **pastecs** would not have been my choice; I even prefer the method for data frames of function `summary` from base R that is not at all mentioned for this purpose. Wish for a future edition: at least mention some further possibilities in the "See also" section.

- Rules 3.2 and 3.3 ("Fit a linear regression model" and "Fit a multiple linear regression model") handle simple and multiple linear regression, using the `trees` dataset for illustration. That dataset contains volumes, heights and girths of black cherry trees, where the volume is to be explained by the other two. An obvious formula, also suggested in the help file for the data, is $volume = constant * height * girth^2$ (formula for a cone of the given height and average girth), which is of course not linear and would require log transformation of all variables for a linear model.

  Rule 3.2 models volume by a simple linear model with height as the only explanatory variable, rule 3.3 extends this model to also include girth. Both rules are stand-alone and refer readers to rules 3.23 and 4.18 for checking of model assumptions.

  In rule 4.18 ("Graphical model validation for linear models", again stand-alone), the model from rule 3.3 is checked with the residual plotting facilities of base R. Interpretation of the diagnostic plots produced by the method for `lm` objects of function `plot` is briefly explained. Furthermore, the standardized residuals are plotted against the variable `Height`; this latter choice is surprising, as a plot against `Girth` would have been more interesting, because it would have indicated some non-linearity.

  Rule 3.23 ("Validate a linear or generalized linear model") also investigates this same linear model (again stand-alone). With function `cumres` from package **gof** (Holst 2012), it is demonstrated that the fit is likely inadequate, particularly regarding the effect of variable `Girth`. Referring to the cone model equation (without explicitly stating it), the book log-transforms all three variables. The resulting model looks adequate under function `cumres`. I've learnt something here, I didn't know about that functionality before. The reference given in the package has been put on my "interesting to read" list.

- Rule 3.4 ("Fit a polynomial regression model") fits a cubic regression model to AIDS death numbers vs calender year. The book uses this opportunity for introducing the

I() function for inhibiting interpretation of exponentiations as modeling expressions. Also, the predict method for class lm objects is introduced for generating response values in a plot of the model function for these data. This would have been a good opportunity for mentioning function poly as an alternative to manual creation of the polynomial, and for guiding readers to do effects plots with package **effects** (Fox 2003) (using poly, an effects plot for the full polynomial effect of the calendar year could have been obtained).

- Rule 3.5 ("Fit a one-way analysis of variance") analyses the well-known orchard sprays data (OrchardSprays in R). It uses function lm and the summary method for class lm objects for investigating the effect of eight concentrations of a substance on the response. Furthermore, it refers the reader to rule 3.18 (t.test) for two groups only, to rule 3.41 for the Kruskal-Wallis test, in case variances are unequal, and to rule 4.6 ("Make a boxplot") for comparative boxplots. Applying rule 4.6 to the orchard sprays data shows that the variances are quite unequal.

  Rule 3.41 ("Compare groups using Kruskal-Wallis' test") presents the base R function kruskal.test and function kruskal_test from package **coin** (Hothorn, Hornik, van de Wiel, and Zeileis 2006); the latter uses resampling-based approximation rather than asymptotic normality for evaluating the null distribution of the test statistic. Furthermore, the rule shows how to complement the Kruskal-Wallis test with subsequent non-parametric multiple comparisons, based on function kruskalmc of package **pgirmess** (Giraudoux 2012) that implements Bonferroni-corrected asymptotic normality-based multiple comparisons of all treatment pairs or of all treatments with a control (without saying so). For treatment pairs – for which it is used in the book – the base R function pairwise.wilcox.test would have done something very similar, when choosing the p.adjust = "bonferroni" option (recalculating ranks for each comparison, which is not done by function kruskalmc). Function pairwise.wilcox.test has the further advantages that it parallels the analogous function for the parametric case and that it also handles the one-sample case; however, it does not cover comparisons to a single control. Wish for a future edition: functions pairwise.wilcox.test, pairwise.t.test, pairwise.prop.test and pairwise.table) for multiple comparisons should at least be mentioned, either here or in rule 3.38, where the p.adjust function is introduced.

- Rule 3.35 ("Non-parametric bootstrap analysis") first explains the rough idea of the non-parametric bootstrap. Bootstrapping is then demonstrated for a heritability index, based on artifical data on r offspring per sire. The rule provides a function to be handed to function boot from package **boot** (Canty and Ripley 2012) and nicely explains how to obtain confidence intervals by using function boot.ci on the resulting outcome. I remember that my initial efforts with bootstrapping something were quite cumbersome. Readers of this example might have fewer troubles getting started with bootstrapping.

- Rule 3.34 ("Use partial least squares regression for prediction") describes how to do partial least squares in R. First, the method is explained such that readers familiar with principal components analysis get a rough idea what partial least squares is about. Then, application of the method with function plsr from package **pls** (Mevik, Wehrens, and Liland 2011) is demonstrated using the example data set gasoline from that package.

- Rule 3.11 ("Fit an ordinal logistic regression model") uses function polr from package

> **MASS** for ordinal logistic regression. This is the only place I have stumbled upon where the code in the book does not work straight-away. The reason is that the explanatory variables have two missing values that change the run size of the data set when dropping variables. This now leads to an error message but didn't in the earlier version of package **MASS** with which the example was run. A correction is promised on the book's website. I nevertheless discuss the content of this section; however, beware that results arising from model comparison by the `drop` function are incorrect (invalid comparisons are made). The book describes ordinal regression and then uses function `polr` for fitting a proportional odds model of the ordinal response "exercise frequency" based on the explanatory variables gender, age and smoking status. The book nicely explains interpretation of parameters and also implements a test for the proportional odds assumption, by comparing the ordinal model to a model with unordered categories fit by function `multmom` from package **nnet** (Venables and Ripley 2002).

The diverse statistical analysis chapter is structured into unnumbered sections in a way that appears a bit arbitrary in places; for example, I would not see adjustment for multiple testing as "robust statistics", and it is a mystery to me, why the t-test comes under "Specific methods" (after many other specific regression methods) or why sample size estimation with functions like `power.t.test` comes under resampling methods. As a person who really appreciates a compelling structure, I found this somewhat distracting. Overall, it is a small matter, in particular as the book refers readers to related rules in the "See also" section of each rule. Nevertheless, in my view, the book would become better by an improvement of structure, e.g., perhaps by introducing a section for simple tests, including the t-test, nonparametrics tests, multiple testing adjustment and power and sample-size calculation.

Choice of topics in general has (of course) been influenced by the author's background – Claus Ekstrøm has mainly worked on biological and medical applications and therefore covers topics like determining the area under the curve (2.29), Bland-Altman method comparison (3.26, 3.27, Bland and Altman 1986), or the last-observation-carried-forward (LOCF) approach to missing values in repeated measurement data (2.8). Nevertheless, the book will be useful to readers from many different fields.

The book uses core R (R Core Team 2012) functionality and also about 50 add-on packages (rough count). It would be desirable for a future edition that the packages used are referenced; currently they are not even listed. Also, apparently, different sections of the book have been written under different versions of R. It would be preferrable to have all calculations in the book run under a defined latest version of the software, in order to have a documented version status and as much up-to-dateness as possible.

Do I consider this book worth reading/buying ? Yes I do! Not as the only R reference or or as a self-learning text on the R language, but as a collection of useful starting points on how to accomplish practically relevant tasks for applied statistics in R.

References

Bland JM, Altman DG (1986). "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *Lancet*, **327**, 307–310.

Canty A, Ripley BD (2012). ***boot**: Bootstrap R (S-Plus) Functions.* R package version 1.3-7, URL http://CRAN.R-project.org/package=boot.

Fox J (2003). "Effect Displays in R for Generalised Linear Models." *Journal of Statistical Software*, **8**(15), 1–27. URL http://www.jstatsoft.org/v08/i15/.

Giraudoux P (2012). ***pgirmess**: Data Analysis in Ecology.* R package version 1.5.6, URL http://CRAN.R-project.org/package=pgirmess.

Harrell Jr FE (2012). ***Hmisc**: Harrell Miscellaneous.* R package version 3.10-1, URL http://CRAN.R-project.org/package=Hmisc.

Holst KK (2012). ***gof**: Model-Diagnostics Based on Cumulative Residuals.* R package version 0.8-2, URL http://CRAN.R-project.org/package=gof.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2006). "A Lego System for Conditional Inference." *The American Statistician*, **60**(3), 257–263.

Ibanez F, Grosjean P, Etienne M (2012). ***pastecs**: Package for Analysis of Space-Time Ecological Series.* R package version 1.3-11, URL http://CRAN.R-project.org/package=pastecs.

Lemon J, Grosjean P (2012). ***prettyR**: Pretty Descriptive Stats.* R package version 2.0-6, URL http://CRAN.R-project.org/package=prettyR.

Mevik BH, Wehrens R, Liland KH (2011). ***pls**: Partial Least Squares and Principal Component Regression.* R package version 2.3-0, URL http://CRAN.R-project.org/package=pls.

R Core Team (2012). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S.* 4th edition. Springer-Verlag, New York.

Reviewer:

Ulrike Grömping
Beuth University of Applied Sciences Berlin
Department II
13353 Berlin, Germany
E-mail: groemping@bht-berlin.de
URL: http://prof.beuth-hochschule.de/groemping/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 52, Book Review 3 *Published:* 2013-01-14
January 2013