Journal of Statistical Software

October 2015, Volume 67, Book Review 2. doi: 10.18637/jss.v067.b02

Reviewer: Thomas Rusch
WU (Vienna University of Economics and Business)

R for Marketing Research and Analytics

Chris Chapman, Elea McDonnell Feit
Springer-Verlag, Switzerland, 2015.
ISBN 978-3-319-14435-1. 454 pp. USD 64.99 (P).
http://www.springer.com/book/9783319144351

This book is another addition to Springer's popular *Use R!* series. True to the scope of this series, "R for Marketing Research and Analytics" introduces the reader to carrying out statistical analyses in a marketing and business analytics context with R. The author's self-proclaimed goal is "We are here to help you learn R for marketing research and analytics". They have designed their book for two audiences (a) practicing marketing researchers and analysts who want to learn R and (b) students or researchers (from other fields) who want to review selected marketing topics in an R context. The preface states the prerequisites to be simply an interest in learning R for marketing and conceptual familiarity with basic statistical models. It also states that the book is particularly helpful for readers with programming experience and interest to learn R. I will come back to the intended audiences at the end of the review. The book introduces a large – not to say comprehensive – number of statistical methods and how they can be utilized in R, all within a context of marketing use cases.

Throughout the book's preface and introduction, the authors repeatedly emphasize that readers who are willing to actually engage in hands-on learning will benefit most from the book. Thus the book's concept is less to serve as a cookbook or reference book, but as a book that allows to learn the material "the hard way", by actually typing the commands and analyzing data. In my opinion this is a worthwhile concept and the book is very well suited for this purpose as it provides ample opportunity for hands-on learning and many ways to benefit from the didactic approach that entails.

Overall the book consists of 13 chapters, that are organized in three parts, and four appendices. The parts are "Part 1: Basics of R", the obligatory getting-started tutorial (pages 2–44), "Part 2: Fundamentals of Data Analysis", a collection of methods for data analysis that may be considered the basic toolbox of a marketing analyst (pages 45–193) and "Part 3: Advanced Marketing Applications" which presents various advanced statistical methods in a marketing context (pages 195–400). Appendix A (pages 403–401) is dedicated to R versions, text editors, IDEs and GUIs; Appendix B (pages 411–422) gives tips on scaling up R (using data from non-R native sources, large data sets, memory profiling and speeding code up, reproducibility and automatic reporting); Appendix C (pages 423–430) lists and explains the R packages used

in the book and Appendix D (pages 431–433) describes the supplementary materials for the book.

The first chapter is titled "Welcome to R" and introduces the reader to R, its philosophy and some of its history. Here the authors discuss reasons for using R but also acknowledge limitations of doing so. Chapter 2, "An Overview of the R Language", is a fairly standard introduction for using R as can be found in most books of the series. It is correct, written reasonably clear and to the point. Some R cracks might find a few slight errors to quibble about but overall I think it is exactly how such an introduction should be. Although this chapter gives little more than the bare bones of using R, this is intended as the authors intersperse the subsequent chapters with "language briefs" that come back to specific ways of achieving certain things in R. For example, the reader does not find a discussion of loops in this chapter but later in a language brief. In this chapter I particularly like that the authors include a worked real-world example to illustrate what one can do after the book has been worked through.

With chapter 3 "Describing Data" the authors turn to data analysis with R. All data analysis chapters start with the generation of one or more simulated data sets that will then be used to illustrate the key concepts of each chapter. In Chapter 3 it is the concept of summarizing and describing univariate data, both numerically and visually. They introduce descriptive statistics, the `*apply` family, `by` and `aggregate`, univariate plots such as barplots, histograms, boxplots, density plots, qq-plots, maps and ways of customizing plots. In Chapter 4 "Relationship between continuous variables" the authors turn to bivariate descriptives, again numerical and visual. For visual description they introduce scatterplots and scatterplot matrices, multiple plots, generalized pair plots (so, also for discrete variables). For numerical description they discuss covariance, Pearson correlation, polychoric correlation, and tests and plots for the correlation measures. They also describe the concept and usefulness of data transformations (but not the dangers) and also include a subsection on Box-Cox transformations.

Chapter 5 is called "Comparing Groups: Tables and Visualizations" and includes how to use R for contingency tables and cross tabulations, as well as trellis plots (with **lattice**). Here R formulas are explained and used for the first time. Chapter 6 "Comparing Groups: Statistical Tests" then introduces the standard suite of classical statistical tests, such as the χ^2 test, the binomial test, t-test, ANOVA. They spend some time with the meaning and interpretation of confidence intervals and p-values here too. Additionally, Bayes factors as an alternative to classical tests are discussed. Chapter 7 is the last of the "fundamentals" and is called "Identifying Drivers of Outcomes: Linear Models". In it, naturally, simple and multiple linear regression models are explained and how to fit them in R. The authors discuss visualisation, fitting, model checking, model comparison, prediction and even overfitting. They also discuss how to fit linear models in a Bayesian setting, a pattern that will be repeated in other chapters.

Chapter 8 "Reducing Data Complexity" marks the first of the "advanced" chapters and features principal component analysis (PCA), exploratory factor analysis, multidimensional scaling and biplots. This is followed suit by Chapter 9 "Additional Linear Modeling Topics" which is a sort of eclectic chapter as it contains a number of techniques that are only loosely related. Here the authors discuss the notion of collinearity, how to diagnose the case with variance inflation factors and how to use PCA to deal with the issue. Logistic regression is also introduced here as well as doubledecker plots and mosaic plots. Then the authors turn to hierarchical or multilevel or mixed-effect models that they also use to introduce hierarchical Bayes models. Of all chapters in the book this might be my least favourite because of its organization and

catch-all nature. I think it would have been better to have logistic regression merged with the later chapter on choice modeling and have collinearity be discussed in chapters on linear models and PCA. This would have freed Chapter 9 to be all about multilevel models (in both denominations), which deserve a broader discussion, as marketing data are very often multilevel. This could also have helped in getting an additional, neglected aspect into the book: repeated measurements and time series.

With Chapter 10 and the subsequent ones, the authors turn to marketing core tasks. The authors introduce the – for marketing research – rather vital field of structural equation models (SEM) in Chapter 10, "Confirmatory Factor Analysis and Structural Equation Modeling". Here confirmatory factor analysis, covariance-based SEM and partial least squares based SEM are described and it is shown how they can be fitted in R. Chapter 11 is on "Segmentation: Clustering and Classification". Here the authors illustrate a broad array of techniques: hierarchical clustering, k-means, model-based clustering with Gaussian mixtures and latent class analysis on the clustering side and naive Bayes and random forests on the classification side. Perhaps single classification and regression or decision trees would have fitted into this chapter well, too. Chapter 12 then turns to "Association rules for Market Basket Analysis", describing association rules and Chapter 13 "Choice Modeling" describes various choice models based on multinomial logit models, also in a mixed formulation and also in a Bayesian flavour. This chapter also contains some comments on optimal design for choice experiments in marketing experiments but, unfortunately, this is not discussed in similar breadth as most other topics. I think it would have been worthwhile to expand on that topic a bit.

Overall the book has many strong aspects. It is for the most part well organized and has a clear didactic flow. The didactic concept is certainly well thought through and the experience is enhanced by the "key points" summaries that are available at the end of each chapter. Together with the motivating example and the emphasis on "try it yourself" the book allows for sustainable learning.

The authors take care to guide the reader through the difficult task of data analysis of marketing data with R. They are very thorough in giving good practical advice and often include words of caution, discuss common problems and pitfalls, and acknowledge limitations of procedures. This is a nice change from the hyperbolic language one finds in some business analytics or data science texts. It is well written, in a colloquial and friendly tone. The reader often has the feeling that the authors talk directly to her. The book features practically no technical statistical language which might be just what is necessary for a subset of the intended audience, while the lack of technicality will at the same time likely not bother the rest. The authors try to avoid jargon and if they are forced to use it, they do a good job in explaining it.

The examples and data sets are motivating. The idea of introducing simulated data sets at the beginning of chapter has the reader generate their own data and allows her to change parameters in the simulation to see the effect by herself. This is a good idea, but I think that throwing more real-world data into the mix would have been better. Marketing and business data are often messy and this is not reflected in a nice, simulated data set. Steps of data pre-processing, checking and cleaning are vital in the modern business analytics world, so the challenge of fashioning a clean and usable data set from the messy, merged, perhaps even unstructured data that are found in a modern business environment could have been more readily acknowledged and acted upon with some less then pretty example data. I believe that the "hands-on" concept of the book would be supported if the hands get a little dirty in the

process.

From a statistical point of view, the book's emphasis on exploratory statistics and data visualisation is definitely a strong suit. Other textbooks for data analysis for marketing neglect this aspect. The process of iterating through description and visualisation, data transformation and modeling is well described and reflected in the didactic approach. Additionally, I like that different approaches to statistical inference are presented next to each other, even if a bit asymmetrical. I think the authors could have dared to give Bayesian methods more room (and less asterisks) especially when they say that they themselves are partial to them. Still, overall the authors display a pragmatic, unpatronizing and laid-back approach to all kinds of statistical inference which I find refreshing. Sometimes the explanations of methods and statistical concepts may not be completely satisfactory to everyone (for example the definition of the 95% CI on p. 140 which had me trip semantically and I'm still not sure I have gotten up already). Those examples are few and far between though, and I mostly mention it to also have said something critical.

The book is accessible to people with some background in marketing research and analytic methods, e.g., undergraduate to masters level at University or practical experience in marketing research. It should also work well for experienced marketing researchers and analysts who are familiar with the methods and would like to use R. I think it also well suited as an accompanying book to a lecture series or course on marketing research methods on the undergrad and masters level (Part 2) and masters or PhD level (Part 3). Thus a reader who is familiar with or has some guidance to the methods will benefit greatly from the book. However, the book is less suited for self-studying when the reader has no familiarity with the methods. Most readers will need more than this one book for getting started with marketing research and analytics in R. Additional literature should be consulted for readers who want to learn about conducting marketing research from scratch with little or no statistical and methodological background. This is because the book faces the same problem similar books have (which the authors acknowledge): One cannot give a thorough introduction to so many, partially very advanced methods and at the same time be a tutorial on how to use them and still stay on 500 pages. Many chapters alone could be expanded to 500 pages! Thus the book must naturally offer only an overview and comprises tutorials for people who at least know what they want to do and that they want to do it with R.

I find the book to be a very welcome addition to the *Use R!* series and the marketing research and business analytics world. I can wholeheartedly recommend it for the described audiences and beyond and will certainly use this book in class and consulting.

Reviewer:

Thomas Rusch
WU (Vienna University of Economics and Business)

Competence Center for Empirical Research Methods
1020 Vienna, Austria
E-mail: thomas.rusch@wu.ac.at
URL: http://www.wu.ac.at/methods/team/dr-thomas-rusch/en/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
October 2015, Volume 67, Book Review 2 *Published:* 2015-10-06
doi:10.18637/jss.v067.b02