Journal of Statistical Software

January 2016, Volume 69, Book Review 3. doi: 10.18637/jss.v069.b03

Reviewer: Christophe Lalanne
Paris-Diderot University

---

**Statistical Data Analytics. Foundations for Data Mining, Informatics, and Knowledge Discovery**

Walter W. Piegorsch
John Wiley & Sons, Chichester, 2015.
ISBN 1-118-61965-0. 464 pp. EUR 94.50 (P).
http://www.wiley.com/WileyCDA/WileyTitle/productCd-111861965X.html

---

This book is not a book on data mining, as the (sub)title may suggest. It sets out the principles of exploratory data analysis, statistical inference (Part 1) and basic building blocks for supervised and unsupervised learning (Part 2) with the R statistical package.

In Chapter 2, the basic of probability theory and statistical distributions are discussed and illustred with R on real-world data sets. Chapters 3 and 4 are dedicated to data manipulation and data visualization, using R base graphics. The author emphasises the use of Tukey's five-number summary for descriptive statistics and outlier analysis, and of smoothing techniques for data with low signal-to-noise ratio. Most of the classical univariate and bivariate graphical displays are discussed at length, with relevant details about R's internals on, e.g., kernel density estimation or histogram binning. Chapter 5 deals with parameter estimation and classical test of hypotheses for the comparison of means and proportions. Interestingly, this chapter is not limited to maximum likelihood point and interval estimation. The author also briefly discusses the method of moments or the weighted least-squares approach, as well as simultaneous or bootstrap confidence intervals, among others.

The second part of the book focuses on statistical modeling, including prediction and classification as well as unsupervised learning techniques. The linear model is covered in Chapter 6 and 7, while generalized linear models are covered in Chapter 8. Dedicated sections are reserved to residual analysis and influence measures (although the author does not mention R's `lm.influence()` function which provides several indicators discussed in Section 6.3), model selection, and regularized approaches such as the lasso or ridge regression (**glmnet** package) to deal with high-dimensional or multicollinear data. Beyond logistic regression, the author also demonstrates the use of log-linear models to analyze two-way cross-tabulated data (although no references are made to the **vcd**/**vcdExtra** and **gnm** packages). As in some of the preceding chapters, the author alternates between R's built-in commands and user-written functions. For example, in the case of the trend test for binary data, both the `prop.trend.test()` function and the Cochran-Armitage test with Tarone's skewness adjustment are illustrated in R. Logistic regression analysis is also used for the purpose of classification tasks (Chapter 9), as

a prelude to linear discriminant analysis of binary outcomes. Other classification techniques are discussed carefully in this chapter, including decision trees and support vector machines. These models are applied on the same data set, which allows to compare their predictive performance. This chapter is also used to remind the reader that proper cross-validation schemes are required to avoid overfitting the data, as in the case of regularized regression. It is worth noting that other attractive approaches such as boosting, ensemble learning, or the random forest algorithm are not covered in detail.

Chapters 10 and 11 are dedicated to unsupervised learning. Principal components and factor analysis are presented as useful tools for dimensionality reduction. In the case of exploratory factor analysis, illustrations rely on the maximum likelihood approach as implemented in `factanal()`, although the author discusses other approaches that are implemented in the **psych** package. The last chapter is about cluster analysis. It is limited to hierarchical and $k$-means clustering, though, and important topics such as how to assess the stability of cluster solutions or to use model-based approaches (e.g., **mclust**) are not discussed. Canonical correlation analysis is also briefly presented as an extension of principal components analysis in the case of two-block data structure. Unfortunately, the author does not really indicate how these techniques could be used in conjunction with supervised models developed in the preceding chapters (e.g., principal components or reduced rank regression, sparse canonical correlation analysis or partial least-squares regression). Association rules and the apriori algorithm (**arules** package) are also briefly discussed at the end of this chapter.

To sum up, this book provides a nice overview of core statistical techniques for applied research using the R statistical package. Standard modeling techniques such as ordinary least-squares regression or classification trees are discussed along with regularized regression or support vector machines, which renders this book attractive for both practitioners and graduate students interested in predictive modeling using R.

## Reviewer:

Christophe Lalanne
Paris-Diderot University
URC ECO – Unité de Recherche Clinique en Economie de la Santé, AP-HP
Hôpital Hotel-Dieu
1, Place du Parvis Notre-Dame
75004 Paris, France
E-mail: ch.lalanne@gmail.com
URL: http://aliquote.org/

*Journal of Statistical Software* http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
January 2016, Volume 69, Book Review 3 *Published:* 2016-01-29
doi:10.18637/jss.v069.b03