Journal of Statistical Software

July 2016, Volume 71, Book Review 2. *doi: 10.18637/jss.v071.b02*

Reviewer: David W. Zeitler
Grand Valley State University

Applied Multivariate Statistics with R

Daniel Zelterman
Springer-Verlag, Switzerland, 2015.
ISBN 978-3319140926. 393 pp. USD 99.00 (Hardcover), USD 69.99 (eBook).
http://www.springer.com/9783319140926

Overview

Daniel Zelterman's 'Applied Multivariate Statistics with R' provides a well positioned mid-level introduction to modern multivariate statistical analysis. Too many multivariate texts are either deep in linear algebra leaving many readers behind, or so superficial as to be less useful than reading help files. Zelterman's text hits a good middle ground, accessible to students with minimal background and useful enough to be worth keeping on the shelf by the computer. It mixes brief coverage of theory with good examples using moderately sized data sets that make learning the techniques accessible. As with most things using R, visualization is presented as an integral component of the technique, making the methods more easily understood and utilized.

The level of Zelterman's material is great for a multivariate offering for upper level undergraduate or masters level graduate students in applied statistics. A second course in statistics (multiple regression and two way ANOVA) using SAS or R is adequate preparation. It would however be helpful to have more exercises and consistent availability of solutions, for example all odd problems. This not only helps usability as a textbook but also improves the text for the independent learner.

Options of USD 99 hardcover and USD 69 e-text make the book comparatively accessible. Members of organizations with Springer Link access are able to download PDF and ePub versions of the text for free, though hardcopy is still preferred by many. This makes independent use or course utilization an easier decision than requiring students to purchase a text that can run well over USD 100 even in used form.

Tying a text to a specific computing language has the benefit of allowing for hands on work with the material, so having access to data and code is important. Unfortunately in this text, data files are sometimes difficult or impossible to obtain so that examples can be difficult to reproduce. For example the health study data from the http://www.cdc.gov/ link is no

longer available. Fortunately an e-mail to the author quickly cured the problem, but providing a single website with downloadable copies of the example data files and scripts would help a great deal.

The use and discussion of R code is somewhat dated. R is much more capable of big data work than indicated in chapter 1. Also, use of recent package releases such as **dplyr**, **tidyr** and **readr** would make data wrangling much simpler than indicated in the text examples. Particularly for students not familiar with R, this allows the reader to focus more on the techniques and less on the language.

Details

- *R quickly:* Chapters 2 and 3 cover basic programming and graphics with R. They are a limited but adequate introduction to the critical elements of the language. As stated above, use of the the newer data wrangling and graphics coming out of the **RStudio** group would greatly enhance the users ability to use R, especially for the larger more complex data sets likely to be encountered in multivariate work.

- *Linear algebra and statistical background:* Chapters 4 through 7 provides a solid linear algebra and normal distribution statistical background often skipped in common multivariate texts. It cannot be assumed at this level and is absolutely critical to understanding and not just blindly running the techniques. The normal distribution treatment here spanning three chapters starting from univariate, moving to bivariate and then to multivariate normality provides a good development for students with limited theoretical statistics background.

- *Factor methods:* With several chapters of background and preparation out of the way, the text logically transitions into the meat of the material with variable reduction and orthogonalization techniques. Coverage of principal component analysis (PCA) and factor analysis (FA) touches on the basic concepts including covariance versus correlation matrix based approaches, proportion of variability in components/factors, scree plots, biplots with examples from finance, astronomy and health care.

- *Multivariate multiple regression:* The logical next step is using the components or factors for modeling. Examples come from automotive data and a health survey. Examining the residuals from several regression equations for remaining multivariate structure using PCA and multivariate normality tests is a great place to start, but there is more to do. Particularly some work with partial least squares would be warranted here.

- *Supervised learning:* Discrimination and classification, often equated with predictive analytics, covers logistic regression, linear discriminant analysis, support vector machines and regression trees. The only thing missing is neural networks, but we can do without them.

- *Unsupervised learning:* Unsupervised learning is essentially clustering and covers both hierarchical and k-means methods. Like much of this text, there is a lot more that could be done here, but what is provided is an excellent start.

- *Time series:* An area often missed in multivariate texts is time series, even though most data is actually longitudinal. This section includes discussion and demonstration of both ARMA and spectral decomposition methods.
- *There's always more:* The 'Other Methods' chapter includes canonical correlation, rankings for paired data and a nice discussion about looking at the extreme values that deserves more detail. A short discussion of big and wide data once again laments that maybe we should have been using SAS, but does note that there are R implementations coming.

Wish list

This is a wish for Zelterman's text, but really applies to almost all texts covering statistical material, especially higher level methods like multivariate statistics. It is initially desirable to study statistics using relatively small manageable data, however real applications often require working with much more complex information. Data sources are often messy, with large numbers of variables and massive numbers of observations. Understanding the impact this has on methods and having some idea of how to approach the problem is becoming a much sought after skill.

Students should not be expected to make the leap from clean, easily manageable 'textbook' problems to real data on their own. Discussion of data cleaning and wrangling, particularly with big and complex data is becoming more important. In particular, examples of working with messy data and discussion of the use of R connections to database systems for managing massive data as well as relatively new R based approaches to big data analysis such as the following would be beneficial.

- *Microsoft R Server* (formerly Revolution R Enterprise, https://www.microsoft.com/en-us/server-cloud/products/r-server/) is readily available and scales R analyses to massive data sizes.
- *Tessera* (http://tessera.io/), a collaboration between Purdue University's Statistics Department and the Pacific Northwest National Laboratory, also brings R based big data analysis to Hadoop clusters.
- *Task view* (https://CRAN.R-project.org/view=HighPerformanceComputing). The CRAN high performance computing task view identifies many other approaches to managing and analyzing big data.

Conclusion

Wish list aside, Zelterman's 'Applied Multivariate Statistics with R' provided an excellent supplement to Johnson and Wichern (2007) for multivariate instruction this past semester, giving students with limited mathematics background a much more approachable and visually appealing introduction to the techniques. Next time through the course, it will be required and maybe even the primary text. I like the level of development and the coverage of topics for our audience.

References

Johnson RA, Wichern DW (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall.

Reviewer:

David Zeitler
Grand Valley State University
Statistics Department
1 Campus Drive
Allendale, MI 49401, United States of America
E-mail: zeitlerd@gvsu.edu
URL: http://faculty.gvsu.edu/zeitlerd/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
July 2016, Volume 71, Book Review 2 *Published:* 2016-07-11
doi:10.18637/jss.v071.b02