DOI: 10.3969/j.issn.2095-1787.2016.03.015

BOOK REVIEW

The not so new statistics with R

Hector A. *The New Statistics with R: An Introduction for Biologists*, Oxford University Press, Oxford UK, 2015, 199 pp. $49.95 (paperback) ISBN 978-0-19-872906-8

Péter SÓLYMOS

Alberta Biodiversity Monitoring Institute and Department of Biological Sciences, CW 405, Biological Sciences Building, University of Alberta, Edmonton, Alberta, T6G 2E9, Canada

I started reading "*The New Statistics with R*" from Andy Hector with mixed expectations. The title suggested something new and exciting, whereas the table of contents revealed that the 'newest' thing it covers are the generalized linear mixed-effects models (GLMMs) that have been around for some decades now (see e.g. Breslow & Clayton, 1993). Chapter 1 makes it clear that this is a linear-model centered introduction to statistical analysis using the R programming language and environment. The material in the book evolved from course notes and is aimed at biologists dealing mostly with data from designed experiments. The term 'New' in the title, as it turns out, is in part marketing by the publisher, but in part refers to a fresh approach to long-established methods, including the use of information criteria, multimodel inference, and greater emphasis on estimates and intervals for statistical inference instead of *P*-values.

The following few chapters walk the reader through the analysis of variance (Chapter 2), *t*-test (Chapter 3), and linear regression (Chapter 4). These chapters jump into the middle of things, without giving much consideration to basic statistical concepts. Chapters include R code and text explaining the ideas behind the code. The data sets being analyzed are well chosen and interesting. Little text boxes show how to express statistical results in scientific publications, which is a nice extra for readers mostly interested in applying these techniques in their own research. The typesetting of the text could have used a bit more editing. For example syntax highlight using colors is often inconsistent or missing, in-line code (variable and function names) and R package names are often indistinguishable from the rest of the text. This makes scanning the book for commands really hard.

Another grudge that I have is the heavy dependence on highly hyped R packages, such as 'ggplot2' and 'reshape2'. Such high level functions build unnecessary dependencies for R novices instead of focusing on programmatic structures that link data, models, and interpretation relying on base R functionality alone. According to my own experience, backwards compatibility (and thus future reproducibility) of code relying on these high level packages can be compromised.

My favourite was Chapter 5: 'Comparisons Using Estimates and Intervals'. Scientists are expected to criticize their own findings by reporting uncertainty around their estimates. This is often expressed as lower and upper bounds of an interval (in tables) or a corresponding error bar (in figures). The main message here is that there are multiple such intervals with very distinct meanings, including standard deviations, standard errors, confidence intervals, least significant differences, and prediction intervals. This chapter is where the book truly shines, and teaches tricks on how

Reviewer address: Péter Sólymos, Alberta Biodiversity Monitoring Institute and Department of Biological Sciences, CW 405, Biological Sciences Building, University of Alberta, Edmonton, Alberta, T6G 2E9, Canada. E-mail: solymos@ualberta.ca

to quickly gain insight from simple figures. Or reversely, what is the right way of making such figures to facilitate understanding. Things become more intricate when considering multiple effects and interactions among them (Chapter 6: Interactions) or interactions between discrete and continuous explanatory variables (Chapter 7: Analysis of Covariance: ANCOVA).

Maximum likelihood is introduced briefly in Chapters 8 and 9 in connection with generalized linear models (GLMs) for non-continuous (discrete) outcomes. The strength of these chapters is that it teaches why and how to check the basic assumptions, including residual plots. Checking the assumptions becomes more difficult as one moves towards mixed-effects models (Chapter 10 and 11). The explanation of variance components and random effects is brief but adequate. Although GLMMs have been around for quite a long time, it is still a hot research topic in the statistical literature because estimation for such models is less straightforward, and methods are often only approximations. Therefore, these chapters can only provide a glimpse of the possible approaches. Links to relevant literature are not so diverse, there is no chapter called References, papers and books are mentioned only as part of the text. In a next edition I would certainly consider bringing in a more diverse set of references and a proper references section, thus providing stepping-stones for readers willing to go beyond the scope of this book.

Contrary to its title, Hector′s book is not too ambitious or novel, but delivers on the topics it sets out as the main focus. The book can be a great introduction for biologists just starting to learn R with the intent of analyzing their own data. There are certainly other books targeting the exact same audience (e.g. Aho, 2013). Nevertheless, I think the well thought through approach of explaining intervals and their application sets this book slightly apart from the rest of the pack.

References

Aho K A, 2013. *Foundational and Applied Statistics for Biologists Using R*. Boca Raton: CRC Press.

Breslow N E and Clayton D G, 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88: 9-25.