Journal of Statistical Software

October 2015, Volume 67, Book Review 5. *doi: 10.18637/jss.v067.b05*

Reviewer: James E. Helmreich
Marist College

Statistics: An Introduction Using R (2nd Edition)

Michael J. Crawley
John Wiley & Sons, Chichester, 2015.
ISBN 978-1-118-94109-6. 339 pp. USD 45.00 (P).
http://www.imperial.ac.uk/bio/research/crawley/statistics/

Introduction

Michael J. Crawley's text *Statistics: An Introduction with R* is indeed an introduction to statistical analysis as well as excellent introduction to working in and with R. However, it is not for the faint of heart. The first half covers standard introductory material (descriptive statistics, simple regression and inferential statistics) in a somewhat idiosyncratic way. The second half of the text is a good introduction to modeling (multiple regression, ANOVA and ANCOVA, various general linear models, survival analysis). The treatments are thorough and yet short – only 300 odd pages. As you can imagine he is extraordinarily succinct. There are many useful and thoughtful insights; it is well written and clear. It contains no exercises; there is an excellent appendix that serves well as a basic R tutorial. If you teach this material without a text or think you might like to try, you might find that it (paradoxically) will work as the structural backbone for a course that you fill out as you see fit. The text will serve well for an introductory course or for a second course in modeling.

Discussion

The overall structure and approach is very nice. His use of R is quite gentle, intertwined with the narrative in a way that is natural, informative, and quickly builds basic facility. In early chapters Crawley introduces a statistic, then calculates it directly 'by hand' in R. He writes a function to automate the process, and only having done so points out the built-in R version. He simultaneously discusses the meaning and use of the statistic, code to calculate it, and higher level R structures. For instance, within the first few pages using R he introduces functions, modular arithmetic, nested looping and ways to plot the results. This sounds like a lot and it is, but it works well here. The magical black-box feel of many routines in varied platforms should be dispelled by the approach. Once in a while he does not follow through –

e.g., deriving the least squares coefficients yet going directly to `lm` – but for the most part he uses this method to good effect.

The material is kept quite applied. No real discussion of the mathematics or theory is provided. Nevertheless he is careful with algebraic arguments concerning partitioning sums of squares or calculation of degrees of freedom; he derives the (two-variate) normal equations and solutions. These types of calculations are the only instances of boxed, independent asides outside of the general narrative.

The book starts out with a chapter of fundamentals, more of an annotated glossary for later reference than an introduction. Here we find general guidelines for types of analyses; discussion of hypotheses and error types; typical assumptions; experimental design; maximum likelihood estimation; power; replication; randomization; weak and strong inference; multiple comparisons; as well as a laundry list of some of the basic model types available in R. In twenty-two pages. Whew!

The presentation is rather terse – the subsection on controls reads in its entirety:

> No controls, no conclusions.

Unfortunately it is often opaque at first (student) reading:

> Pseudoreplication occurs when you analyze the data as if you had more degrees of freedom than you really have.

This is the first mention of degrees of freedom I can find in the text; these sorts of out of place sections lard the text, and can cause some confusion. While frustrating to students new to the material, it is in service of brevity, elegance of presentation, and perhaps piquing the interest of the reader for what is to come. Though at times annoying, I find it works reasonably well overall. Clearly, it is intended that some of the material will become clearer as the reader progresses through the book.

The second chapter begins the material on R, introducing dataframes and their manipulation, as well as “initial data inspection” – summary statistics and basic graphical analysis (in that order: I’d reverse it, but then that’s me). Datasets from the book’s website are used, with R commands highlighted in red, the output in blue.

The third chapter begins where most texts start: measures of central tendency. Here, as well as in later chapters, the treatment is not completely standard. A discussion of quartiles and the five number summary is omitted. After discussing the mean and median, the geometric and harmonic means are presented (with good motivating examples) as well.

Another instance of non-linearity: the next chapter on variance ends with a residual scatterplot showing heteroscedasticity long before chapters on regression or bivariate data. I found myself frequently wishing for a more thorough index. On the other hand, Crawley introduces confidence intervals very naturally and intuitively, both t and bootstrapped versions. He demonstrates sample size effects on estimation using nested loops for sampling and graphics. It is an odd, or at least unorthodox, presentation that works nicely.

In the next chapters he introduces analysis of single samples and two samples, including normality tests and both parametric and nonparametric inference techniques. The single-sample chapter finishes up with a discussion of statistics to measure skew and kurtosis. Two-sample nonparametric alternatives (rank sum and binomial test) feature prominently. He includes graphical techniques as well: to with notched box plots. The χ^2 and Fisher's exact tests for independence of two samples of counts is covered. The two-sample discussion concludes with correlation – including Pearson's product-moments correlation method of testing for differences from zero.

From here, Crawley begins simple regression by reminding the reader of the definition of slope and other rather rudimentary topics. Curiously, he spends a rather verbose three pages at a mathematical level dramatically lower than the rest of the text's narrative. That aside, the presentation is very good, and quickly goes well beyond normal material to include partitioning sums of squares; calculating standard errors for the slope and intercept, careful model checking and transformations to correct problems, polynomial regression and even non-linear regression as well as generalized additive models.

Again, whew! This recalls to mind a colleague's quip that assigning a difficult text is a desirable thing in that it allows you to be the good guy when you elaborate and clarify in the classroom. Too many modern texts are written in such a way that the roles are reversed. But I overstate: the worked examples are clear, interesting and easily replicated with the commands shown. Color is used well but sparingly to differentiate text, code, and output, and for graphics. There are no colored boxes, marginal notes, distracting highlights or gratuitous photos in the text.

The narrative continues with chapters on ANOVA, ANCOVA, multiple regression, contrasts, and a series of chapters on general linear models. In all of these the focus is on the nature of the data. Poisson regression is discussed in the chapter on count data, logistic in the chapters on proportion data and binary response data. A brief chapter on survival analysis rounds out the text.

These regression chapters are my favorite part of the book. With minor variations each chapter in the latter half of the book takes a data set, sets out the situation, does a basic analysis, finds a maximal model and then either by hand or using `step` prunes to a minimal adequate model. The implications of the model and need for transformations are assessed, and if necessary the process is repeated. He uses graphics well to assess models and composes nice multilayer plots of the data and fitted curves.

The entire set of chapters begins with an overview of types of data, error structures, link functions, deviance and AIC. This type of overview is one of the great strengths of the exposition. Crawley provides excellent characterizations at various points outlining the big picture or the general steps an analysis should take.

I have a few quibbles. The exposition is, again, at times non-linear. We frequently meet terms that are only defined later in the text. There is very little theory, which gives many of the topics a rabbit-out-of-the-hat feel to them – though when possible nice intuitive comments (e.g., for maximum likelihood techniques) are presented in lieu of the mathematics.

There are some inconsistencies. He makes a strong case for graphical analysis of regression results, providing an example of a model with a strong numeric summary that nevertheless graphical analysis shows to be flawed. The point is a good one if not consistently followed. Before this example there were several occasions where he proceeds, willy-nilly, with a numerical analysis without even a precaution to the student that they consider graphing the data first. Indeed, in Chapter 4 variance is calculated repeatedly while ignoring any plots.

Conclusion

Crawley has a real knack for outlining the general steps and processes of an analysis that is clear, concise, informative and useful. He frequently makes pithy, insightful points without overelaboration. He steps back regularly to give a broad overview of the topic. The use of R is integrated seamlessly into the exposition, and will provide the novice with a straightforward introduction to R code. The approach is not quite what I expected from an introductory text, but I am left with the strong sense that it should be.

I could easily see this used as the backbone of both a first-year course for STEM students with the mathematical maturity to take Calculus, as well as a course in applied statistics for students who have taken at least a semester of mathematical probability and statistics. It does omit exercises, and given the author's brevity will need elaboration in the classroom. However the text provides strong overall structure and would be a good student resource. Having taught the material from the second half of the text recently without a text, I find myself wishing I had known of and made this text available to my class.

Reviewer:

James E. Helmreich
Marist College
Department of Mathematics
3399 North Road
Poughkeepsie, NY 12514, United States of America
E-mail: James.Helmreich@Marist.edu
URL: http://foxweb.marist.edu/users/james.helmreich/

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 5 *Published:* 2015-01-06
doi:10.18637/jss.v067.b05