Journal of Statistical Software

MMMMMM YYYY, Volume VV, Book Review II. http://www.jstatsoft.org/

Reviewer: Hadley Wickham
Rice University

A First Course in Statistical Programming with R

W. John Braun and Duncan J. Murdoch
Cambridge University Press, Cambridge, England, 2007.
ISBN 9780521694247. 174 pp. USD 50.00 (paperback).
`http://www.cambridge.org/0521872650`

Introduction

"A First Course in Statistical Programming with R" gives a solid basic introduction to programming in R. It covers similar content to "An Introduction to R", at a more introductory level, written in a more accessible and friendly manner, with many worked examples. The book covers all foundational aspects of R in a clear and approachable style. Topics are described succinctly, and illustrated with worked examples. Each section comes with a set of exercises to allow you to practice what you have learned.

The book starts off strong with a nice definition of statistical programming that includes computations to aid statistical analysis, statistical graphics and simulation. Unfortunately, by this definition, the book covers little statistical programming and it would be better titled "A First Course in Programming with R". Few examples and exercises are statistical in nature, usually having a more finance or computer science flavour, and the topic sections of the book on simulation, computational linear algebra and numerical optimisation are rather dated, focussing on very traditional statistical computing material.

Book contents

The book has seven chapters, which, with the exception of the introduction, are described below. A short appendix gives a brief review of random variables and their distributions.

Chapter 2 introduces the basics of the R language. It covers everything you might expect: starting and quitting, creating variables, listing variables, accessing variables, calling functions, missing values, vectors, sequences, subsetting, accessing help, boolean algebra, basic statistical summaries, working directory and input and output. This chapter covers a lot of material in little space, and only skims the surface of many important topics. For example, I would have liked to see more discussion of subsetting. It is such an important part of effective

R use, and students take some time to learn and internalise strategies for effective subsetting, particularly using logical subsetting with data frames.

I particularly liked Section 2.4 on boolean algebra as it made clear the connection to set theory, which many students otherwise miss. However, the discussion was restricted to the classical two-value algebra, completely neglecting the complications that missing values add to the system: students do not expect the comparison `NA == NA` to yield `NA`.

Chapter 3, "Programming statistical graphics" covers the basics of the built-in high level graphics functions: bar, pie, histogram, box, scatterplot, and qq plot, includes a succinct introduction to low-level graphics, and points the interested reader to other high-level graphics engines. The coverage is adequate, but brief, and more discussion about how to actually use the plots would have been helpful. I was disappointed that only small "textbook" datasets were used, a shame as with R there is no reason not to use large, modern (and interesting!) datasets.

Chapter 4, "Programming with R" is the highlight of the book. It covers the basics of flow control (`if`, `for`, `while`, `repeat` and `break`) and how to write functions. Section 4.2 is excellent, giving solid advice on naming functions and variables, and on the use of comments. I loved Sections 4.4 and 4.5 which outline a basic strategy for writing functions and gives advice on how to debug and maintain code. These are important programming topics that are rarely covered in statistics texts. Section 4.6 on efficiency gives a brief but thorough introduction to efficiency including a heuristic description of expected running time and the ideas behind big-O notation.

There are a few minor flaws in this chapter. The authors suggest using `fix` to modify a function: this is not reproducible, and makes me wonder how students are writing functions. Why are they not using a text editor? The examples—the sieve of Eratosthenes, merge sort, Newton's root finding algorithm and the bisection algorithm—are traditional computer science examples. Why couldn't they be more statistical? Some discussion of how to test functions with stochastic output would also be useful for the statistical setting.

Chapter 5, "Simulation", introduces (but does not define) Monte Carlo simulation, and shows how to draw random numbers from the uniform, Bernoulli, binomial, Poisson, exponential and normal distributions. The chapter concludes with Monte Carlo integration and the rejection and importance sampling methods. Some interesting examples would have really made this chapter come to life and would have illustrated why these techniques are so important to statisticians.

Chapters 6 and 7 cover linear algebra and numerical optimisation respectively. The linear algebra chapter covers the creation and subsetting of matrices, extracting important components, matrix multiplication and inversion, calculation of the determinant, trace, eigenvalues and eigenvectors, condition numbers and various matrix decompositions (svd, Cholesky and QR). Unfortunately there is no motivation as to why we would want to do any of those things. Chapter 7 covers numerical optimisation in much same way. We learn about the golden section method, Newton-Raphson, and Nelder-Mead, but we don't learn how we should choose between them, or what we would use them for in statistics. There's not even a simple maximum likelihood problem! The chapter finishes with linear programming and its extensions to integer and quadratic programming.

Conclusion

I would cautiously recommend this book for use in an introductory course, with the caveat that the instructor would need to be assertive in selecting interesting and relevant data problems to illustrate the tools described in the book. The book fails to provide motivation for many of the topics, so it would be crucial for the instructor to provide motivation in the form of challenging statistical and data analysis problems.

Reviewer:

Hadley Wickham
Rice University
Department of Statistics
Houston, TX
E-mail: hadley@rice.edu
URL: http://had.co.nz

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Book Review II *Published:* yyyy-mm-dd
MMMMMM YYYY