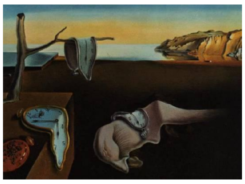

Journal of Statistical Software

June 2008, Volume 26, Book Review 1. *http://www.jstatsoft.org/*

Reviewer: Hadley Wickham
Rice University

Data Manipulation with R

Phil Spector
Springer-Verlag, Carey, NC, 2008.
ISBN 978-0-387-74730-9. 154 pp. USD 54.95 (P).

Introduction

This slim volume provides a solid introduction to many of the most useful functions and packages for importing, manipulating and processing data in R. The R project provides an environment for statistical computing and data analysis. There are many books on statistics in R, and a few on programming in R, but this is the first book devoted to the first part of a data analysis: getting the data into R and into a form that you can work with.

Book contents

The book covers fundamental R data structures, data import and export, and data processing. Chapter 1 starts the book with an introduction to the basic data structures of R (vectors, matrices, lists, and data frames), and tools to describe and to convert among them.

Chapter 2 discusses how to get data into and out of R: from text files (CSV, tab delimited, and fixed width), **Excel** (with the **RODBC** or **gdata** packages), binary files, and from the output of other statistical packages (SPSS, SAS, etc., with the **foreign** package). It also covers using connections to access data over Web or in a zip file; techniques for generating data from random numbers, sequences and permutations; and output to text and binary files. Chapter 3 continues this theme with a discussion of connecting R to databases, to both extract and update data, and includes a brief introduction to SQL. The chapter focuses on **MySQL**, an open-source database, but the ideas are applicable to any client-server database.

Chapters 4, 5 and 7 describe useful operations on dates, factors and strings. Chapter 6, which describes subsetting, seems a little out of place here, and might be best read after Chapter 1. These chapters are brief, but cover the essentials: the various date formats, how to parse strings into dates, and how to extract various components; creating and manipulating factors; extracting parts of strings, and a brief introduction to regular expressions.

Chapters 8 and 9 conclude the book with tools to aggregate and reshape data. As well as just describing useful functions, Spector also suggests some good ways to think about the problems. Chapter 8 includes aggregation by hand with loops, the `apply()` functions and `split()`; and with pre-built functions like `aggregate()`, `by()`, and the **reshape** package. Chapter 9 covers tools for manipulating columns of a data frame (`transform()`, `ifelse()`, `recode()`); combining data frames (`cbind()`, `rbind()`, `merge()`); and reshaping with the **reshape** package.

Conclusion

Overall, this is a useful well-written book, and I think would make a good second book for the R user. There are a few technical errors: character vectors are *smaller* than factors, not larger; and `methods()` should be used to find methods of a function, not `apropos()`. The SQL chapter would have benefited from a discussion of **SQLite**, which is much easier to set up than **MySQL**, and in particular **sqldf** provides a particularly smooth transition between data frames and database tables.

The code in the book is well laid out and easy to follow, although more spaces after commas would aid readability, and prefer the convention of using `<-` for assignment. The book currently lacks a Website or supporting package where all code and data can be downloaded. Hopefully this will change in the future.

Reviewer:

Hadley Wickham
Rice University
Department of Statistics
Houston, TX, United States of America
E-mail: h.wickham@gmail.com
URL: http://had.co.nz/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 26, Book Review 1 *Published:* 2008-06-18
June 2008