Use R!

Use R!

Albert: Bayesian Computation with R

Cook/Swayne: Interactive and Dynamic Graphics for Data Analysis: With R and GGobi

Hahne/Huber/Gentleman/Falcon: Bioconductor Case Studies

Paradis: Analysis of Phylogenetics and Evolution with R

Pfaff: Analysis of Integrated and Cointegrated Time Series with R

Sarkar: Lattice: Multivariate Data Visualization with R

Spector: Data Manipulation with R

Phil Spector

Data Manipulation with R

Springer

Phil Spector
Statistical Computing Facility
Department of Statistics
University of Califonia, Berkeley
Berkeley, California 94720
spector@stat.berkeley.edu

ISBN 978-0-387-74730-9 e-ISBN 978-0-387-74731-6
DOI: 10.1663/978-0-387-74731-6

Library of Congress Control Number: 2008921862

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

Preface

The R language provides a rich environment for working with data, especially data to be used for statistical modeling or graphics. Coupled with the large variety of easily available packages, it allows access to both well-established and experimental statistical techniques. However techniques that might make sense in other languages are often very inefficient in R, but, due to R's flexibility, it is often possible to implement these techniques in R. Generally, the problem with such techniques is that they do not scale properly; that is, as the problem size grows, the methods slow down at a rate that might be unexpected. The goal of this book is to present a wide variety of data manipulation techniques implemented in R to take advantage of the way that R works, rather than directly resembling methods used in other languages. Since this requires a basic notion of how R stores data, the first chapter of the book is devoted to the fundamentals of data in R. The material in this chapter is a prerequisite for understanding the ideas introduced in later chapters.

Since one of the first tasks in any project involving data and R is getting the data into R in a way that it will be usable, Chapter 2 covers reading data from a variety of sources (text files, spreadsheets, files from other programs, etc.), as well as saving R objects both in native form and in formats that other programs will be able to work with. Chapter 3 addresses the issue of relational databases, since large datasets are often stored in such databases. Some guidance in setting up and using databases to work with large datasets is also included in this chapter.

Chapter 4 covers the topic of dates and times in R. While some work can be done using a simple character representation of this type of data, a wider range of operations are available when dates and times are converted to an internal form that allows for comparisons and other manipulations. There are a variety of mechanisms for storing dates and times in R, and this chapter is presented to encourage users of such data to convert them to the appropriate type as early as possible.

While factors are undeniably valuable in data modeling and graphics, they often "get in the way" when performing more basic operations on data. Chapter 5 addresses how to convert objects to and from factors, along with guidelines on how to avoid factor conversions when necessary.

Chapter 6 explores the many ways that subscripting in R can be used to access and modify data. Subscripts (especially logical subscripts), are one of the most powerful tools in R. Many operations that normally require loops or complex programs can be solved elegantly and efficiently in R by using the power of subscripting.

Although R is usually thought of as a language for working with numbers, more and more data is appearing in the form of character strings instead of numbers. Along with basic functions for breaking apart and putting together character strings, R provides a complete implementation of regular expressions; coupled with vectorization, most character data problems can be solved simply and efficiently. Chapter 7 addresses those areas of R focused on character data.

Since most analyses, both model-based and graphical, operate on data frames, the final two chapters of the book directly address working with data frames. Chapter 8 discusses aggregation techniques, where the contents of a data frame are summarized, often broken down by groups. Chapter 9 covers the somewhat related issue of transforming and reshaping data frames. Emphasis is on methods that take advantage of R's power, and which will scale up appropriately as the size of data they operate on increases.

One aspect of this book that may seem unfamiliar is the use of the equal sign (`=`) as an assignment operator rather than the more traditional "gets" operator (`<-`). I find using the equal sign more natural than the other notation, so I've used it in all the examples. The one situation where this causes problems (assigning a value to a variable as part of a function call) is discussed in Section 8.7.

While the focus of the book is using the functions and methods that are built in to base R, a number of packages from CRAN (the Comprehensive R Archive Network) are introduced in the text. These are packages that I've personally found useful in my own work, and omission of other packages is by no means meant to imply that those packages aren't useful. In fact, with the wide variety of new packages contributed by the R community, and serious R programmer would be well advised to visit the R project homepage (`http://r-project.org` or preferably an appropriate mirror site) to check for new packages. Another valuable resource on this page is the R Newsletter, which often provides in-depth information on using some of the new packages.

I'd like to express my most sincere gratitude to both the original developers of the S language, the R Core development team, and the entire R community for creating such a wonderful language and inspiring its users to come up with new and exciting ways of using it.

Contents