R语言技术手册（影印版）

第二版

R IN A NUTSHELL

A Desktop Quick Reference

O'REILLY®
東南大學出版社

Joseph Adler 著

第二版

R语言技术手册（影印版）

R in a nutshell

Joseph Adler 著

O'REILLY®

Beijing · Cambridge · Farnham · Köln · Sebastopol · Tokyo

O'Reilly Media, Inc.授权东南大学出版社出版

图书在版编目（CIP）数据

R语言技术手册：第2版：英文/(美)艾德勒（Adler, J.）著.—影印本.—南京：东南大学出版社，2013.5

书名原文：R in a Nutshell, 2E

ISBN 978-7-5641-4203-2

I. ①R… II. ①艾… III. ①程序语言－程序设计－技术手册－英文 IV. ①TP312

中国版本图书馆CIP数据核字（2013）第097350号

江苏省版权局著作权合同登记

图字：10-2013-119号

©2012 by O'Reilly Media, Inc.

Reprint of the English Edition, jointly published by O'Reilly Media, Inc. and Southeast University Pres 2013. Authorized reprint of the original English edition, 2013 O'Reilly Media, Inc., the owner of all righ to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由O'Reilly Media, Inc. 出版2012。

英文影印版由东南大学出版社出版2013。此影印版的出版和销售得到出版权和销售权的所有者——O'Reill Media, Inc. 的许可。

版权所有，未得书面许可，本书的任何部分和全部不得以任何形式重制。

R语言技术手册 第二版（影印版）

出版发行：东南大学出版社

地　　址：南京四牌楼2号　　邮编：210096

出 版 人：江建中

网　　址：http://www.seupress.com

电子邮件：press@seupress.com

印　　刷：扬中市印刷有限公司

开　　本：787毫米×980毫米　16开本

印　　张：45.25

字　　数：886千字

版　　次：2013年5月第1版

印　　次：2013年5月第1次印刷

书　　号：ISBN 978-7-5641-4203-2

定　　价：92.00元（册）

本社图书若有印装质量问题，请直接与营销部联系。电话（传真）：025-83791830

Preface

It's been over 10 years since I was first introduced to R. Back then, I was a young product development manager at DoubleClick, a company that sold advertising software for managing online ad sales. I was working on inventory prediction: estimating the number of ad impressions that could be sold for a given search term, web page, or demographic characteristic. I wanted to play with the data myself, but we couldn't afford a piece of expensive software like SAS or MATLAB. I looked around for a little while, trying to find an open-source statistics package, and stumbled on R. Back then, R was a bit rough around the edges and was missing a lot of the features it has today (like fancy graphics and statistics functions). But R was intuitive and easy to use; I was hooked. Since that time, I've used R to do many different things: estimate credit risk, analyze baseball statistics, and look for Internet security threats. I've learned a lot about data and matured a lot as a data analyst.

R, too, has matured a great deal over the past decade. R is used at the world's largest technology companies (including Google, Microsoft, and Facebook), the largest pharmaceutical companies (including Johnson & Johnson, Merck, and Pfizer), and at hundreds of other companies. It's used in statistics classes at universities around the world and by statistics researchers to try new techniques and algorithms.

Why I Wrote This Book

This book is designed to be a concise guide to R. It's not intended to be a book about statistics or an exhaustive guide to R. In this book, I tried to show all the things that R can do and to give examples showing how to do them. This book is designed to be a good desktop reference.

I wrote this book because I like R. R is fun and intuitive in ways that other solutions are not. You can do things in a few lines of R that could take hours of struggling in a spreadsheet. Similarly, you can do things in a few lines of R that could take pages of Java code (and hours of Java coding). There are some excellent books on R, but

I couldn't find an inexpensive book that gave an overview of everything you could do in R. I hope this book helps you use R.

When Should You Use R?

I think R is a great piece of software, but it isn't the right tool for every problem. Clearly, it would be ridiculous to write a video game in R, but it's not even the best tool for all data problems.

R is very good at plotting graphics, analyzing data, and fitting statistical models using data that fits in the computer's memory. It's not as good at storing data in complicated structures, efficiently querying data, or working with data that doesn't fit in the computer's memory.

Typically, I use a scripting language like Perl, Python, or Ruby to preprocess files before using them in R. (If the files are really big, I'll use Pig.) It's technically possible to use R for these problems (by reading files one line at a time and using R's regular expression support), but it's pretty awkward. To hold large data files, I usually use Hadoop. Sometimes I use a database like MySQL, PostgreSQL, SQLite, or Oracle (when someone else is paying the license fee).

What's New in the Second Edition?

This edition isn't a total rewrite of the first book. But I have tried to improve the book in a few significant ways:

- There are new chapters on `ggplot2` and using R with Hadoop.
- Formatting changes should make code examples easier to read.
- I've changed the order of the book slightly, grouping the plotting chapters together.
- I've made some minor updates to reflect changes in R 2.14 and R 2.15.
- There are some new sections on useful tools for manipulating data in R, such as plyr and reshape.
- I've corrected dozens of errors.

此为试读,需要完整PDF请访问: www.ertongbook.com

R License Terms

R is an open-source software package, licensed under the GNU General Public License (GPL).[1] This means that you can install R for free on most desktop and server machines. (Comparable commercial software packages sell for hundreds or thousands of dollars. If R were a poor substitute for the commercial software packages, they might have limited appeal. However, I think R is *better* than its commercial counterparts in many respects.)

Capability
: You can find implementations for hundreds (maybe thousands) of statistical and data analysis algorithms in R. No commercial package offers anywhere near the scope of functionality available through the Comprehensive R Archive Network (CRAN).

Community
: There are now hundreds of thousands (if not millions) of R users worldwide. By using R, you can be sure that you're using the same software your colleagues are using.

Performance
: R's performance is comparable, or superior, to most commercial analysis packages. R requires you to load data sets into memory before processing. If you have enough memory to hold the data, R can run very quickly. Luckily, memory is cheap. You can buy 32 GB of server RAM for less than the cost of a single desktop license of a comparable piece of commercial statistical software.

Examples

In this book, I have tried to provide many working examples of R code. I deliberately decided to use new and original examples, instead of relying on the data sets included with R. I am not implying that the included examples are not good; they are good. I just wanted to give readers a second set of examples. In most cases, the examples are short and simple and I have not provided them in a downloadable form. However, I have included example data and a few of the longer examples in the `nutshell` R package, available through CRAN. To install the `nutshell` package, type the following command on the R console:

```
> install.packages("nutshell")
```

1. There is some controversy about GPL licensed software and what it means to you as a corporate user. Some users are afraid that any code they write in R will be bound by the GPL. If you are not writing extensions to R, you do not need to worry about this issue. R is an interpreter, and the GPL does not apply to a program just because it is executed on a GPL-licensed interpreter.

   If you are writing extensions to R, they might be bound by the GPL. For more information, see the GNU foundation's FAQ on the GPL: *http://www.gnu.org/licenses/gplfaq*. However, for a definite answer, see an attorney. If you are worried about a specific application, see an attorney.

How This Book Is Organized

I've broken this book into parts:

- Part I, *R Basics*, covers the basics of getting and running R. It's designed to help get you up and running if you're a new user, including a short tour of the many things you can do with R.
- Part II, *The R Language*, picks up where the first section leaves off, describing the R language in detail.
- Part III, *Working with Data*, covers data processing in R: loading data into R, transforming data, and summarizing data.
- Part IV, *Data Visualization*, describes how to plot data with R.
- Part V, *Statistics with R*, covers statistical tests and models in R.
- Part VI, *Additional Topics*, contains chapters that don't belong elsewhere: tuning R programs, writing parallel R programs, and Bioconductor.
- Finally, I included an Appendix describing functions and data sets included with the base distribution of R.

If you are new to R, install R and start with Chapter 3. Next, take a look at Chapter 5 to learn some of the rules of the R language. If you plan to use R for plotting, statistical tests, or statistical models, take a look at the appropriate chapter. Make sure you look at the first few sections of the chapter, because these provide an overview of how all the related functions work. (For example, don't skip straight to "Random forests for regression" on page 448 without reading "Example: A Simple Linear Model" on page 401.)

Conventions Used in This Book

The following typographical conventions are used in this book:

Italic
: Indicates new terms, URLs, email addresses, filenames, and file extensions.

`Constant width`
: Used for program listings as well as within paragraphs to refer to program elements such as variable or function names, databases, data types, environment variables, statements, and keywords. (When showing input and output on the R console, I use constant width text to show prompts and other information produced by the R interpreter.)

`Constant width bold`
: Shows commands or other text that should be typed literally by the user. (When showing input and output on the R console, I use constant width bold text to show you what I typed, including comments.)

`Constant width italic`
: Shows text that should be replaced with user-supplied values or by values determined by context.

This icon indicates a tip, suggestion, or general note.

This icon indicates a warning or a caution.

In this book, I will sometimes show commands that I entered on my operating system prompt (i.e., in a Bash shell on Linux), and sometimes show commands that I entered in the R console. For commands that I entered in the operating system shell, I use a `$` character to show the prompt; for commands entered in the R console, I will use `>` or `+` to show the prompt. (In either case, don't type the prompt character.)

Using Code Examples

This book is here to help you get your job done. In general, you may use the code in this book in your programs and documentation. You do not need to contact us for permission unless you're reproducing a significant portion of the code. For example, writing a program that uses several chunks of code from this book does not require permission. Selling or distributing a CD-ROM of examples from O'Reilly books does require permission. Answering a question by citing this book and quoting example code does not require permission. Incorporating a significant amount of example code from this book into your product's documentation does require permission.

We appreciate, but do not require, attribution. An attribution usually includes the title, author, publisher, and ISBN. For example: "*R in a Nutshell* by Joseph Adler. Copyright 2012 Joseph Adler, 978-1-449-31208-4."

If you feel your use of code examples falls outside fair use or the permission given above, feel free to contact us at *permissions@oreilly.com*.

Safari® Books Online

Safari Books Online (*www.safaribooksonline.com*) is an on-demand digital library that delivers expert content in both book and video form from the world's leading authors in technology and business.

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of product mixes and pricing programs for organizations, government agencies, and individuals. Subscribers have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like O'Reilly Media, Prentice Hall Professional,

Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology, and dozens more. For more information about Safari Books Online, please visit us online.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

We have a web page for this book where we list errata, examples, and any additional information. You can access this page at *http://oreil.ly/r_in_a_nutshell_2e*.

To comment or to ask technical questions about this book, send email to *bookquestions@oreilly.com*.

For more information about our books, courses, conferences, and news, see our website at *http://www.oreilly.com*.

Find us on Facebook: *http://facebook.com/oreilly*

Follow us on Twitter: *http://twitter.com/oreillymedia*

Watch us on YouTube: *http://www.youtube.com/oreillymedia*

Acknowledgments

First, I'd like to thank everyone who read the first book. I wrote *R in a Nutshell* to be useful. I tried to write the book that I wanted to read; I tried my best to share as much useful information as I could about R. That's an ambitious goal, and I wrote an imperfect book. I appreciate all the feedback, suggestions, and corrections that I have received from readers and have tried my best to improve the book in the second edition.

I'd like to thank the team at O'Reilly for their support. Tim O'Reilly has said that he follows three guiding principles: work on something that matters to you more than money, create more value than you capture, and take the long view.[2] I tried to follow these principles when writing this book. As an author, I felt like the team at O'Reilly followed these principles. My goal in writing *R in a Nutshell* was to write the best book I could write. I hope that when people read this book, they learn something new and use what they learned to solve important problems.

2. See *http://radar.oreilly.com/2009/01/work-on-stuff-that-matters-fir.html*.

Many people helped support the writing of this book. First, I'd like to thank all of my technical reviewers. These folks check to make sure the examples work, look for technical and mathematical errors, and make many suggestions on writing quality. It's not possible to write a quality technical book without quality technical reviewers: Peter Goldstein, Aaron Mandel, and David Hoaglin are the reason that this book reads as well as it does.

For the past two years, I've worked at LinkedIn, ground zero for the data revolution. I've learned a huge amount working side by side with people like DJ Patil, Monica Rogati, Daniel Tunkelang, Sam Shah, and Jay Kreps. I've had the chance to discover interesting patterns, figure out how to share them with other people, and figure out how to scale my programs to work for hundreds of millions of users. I hope the second edition of this book reflects some of the lessons that I've learned on data, and helps other people learn the same things.

I'd like to thank Randall Munroe, author of the xkcd comic. He kindly allowed us to reprint two of his (excellent) comics in this book. You can find his comics (and assorted merchandise) at *http://www.xkcd.com*.

Additionally, I'd like to thank everyone who provided or suggested improvements. Aaron Schatz of Football Outsiders (*http://www.footballoutsiders.com*) provided me with play-by-play data from the 2005 NFL season (the field goal data is from its database). Sandor Szalma of Johnson & Johnson suggested GSE2034 as an example of gene expression data. Jeremy Howard of Kaggle suggested adding glmnet.

Finally, I'd like to thank my wife, Sarah, my daughter, Zoe, and my son, Zeke. Writing a book takes a lot of time, and they were very understanding when I needed to work. They were also very understanding when I dragged them to the San Diego Zoo to look at the harpy eagles.

Table of Contents

此为试读,需要完整PDF请访问: www.ertongbook.com

此为试读,需要完整PDF请访问: www.ertongbook.com