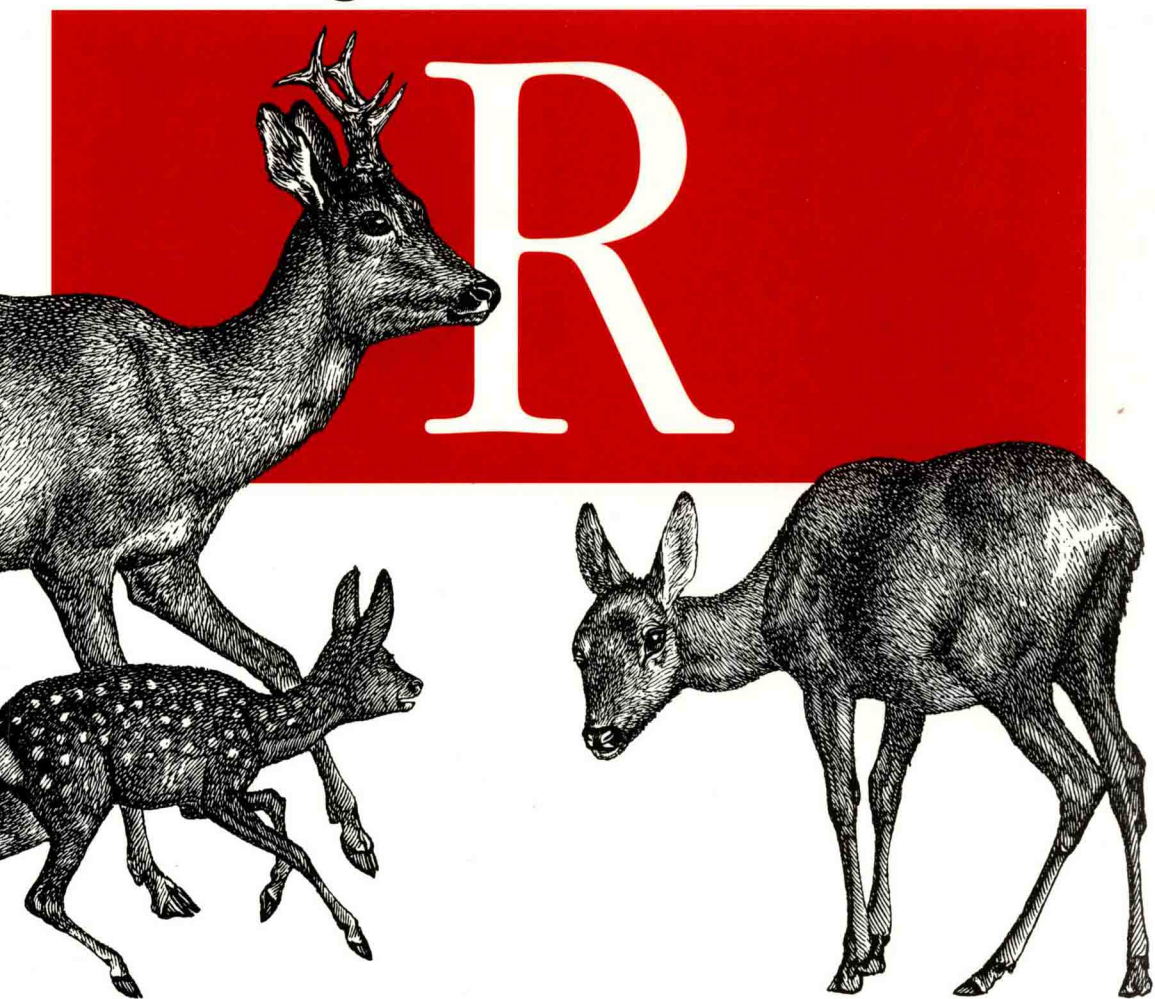


学习R语言 (影印版)

Learning



O'REILLY®

東南大學出版社

Richard Cotton 著

学习R语言 (影印版)

Learning R

Richard Cotton 著

常州大学图书馆
藏书章

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

O'Reilly Media, Inc. 授权东南大学出版社出版

南京 东南大学出版社

图书在版编目 (CIP) 数据

学习 R 语言：英文 / (美) 科顿 (Cotton, R.) 著. —影
印本. —南京：东南大学出版社，2014.9

书名原文：Learning R

ISBN 978-7-5641-4906-2

I. ①学… II. ①科… III. ①程序语言—程序设计—
英文 IV. ①TP312

中国版本图书馆 CIP 数据核字 (2014) 第 083283 号

江苏省版权局著作权合同登记

图字：10-2013-370 号

©2013 by O'Reilly Media, Inc.

Reprint of the English Edition, jointly published by O'Reilly Media, Inc. and Southeast University Press, 2014. Authorized reprint of the original English edition, 2013 O'Reilly Media, Inc., the owner of all rights 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 2013。

英文影印版由东南大学出版社出版 2014。此影印版的出版和销售得到出版权和销售权的所有者——O'Reilly 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。

学习 R 语言 (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：常州市武进第三印刷有限公司

开 本：787 毫米 × 980 毫米 16 开本

印 张：25

字 数：490 千字

版 次：2014 年 9 月第 1 版

印 次：2014 年 9 月第 1 次印刷

书 号：ISBN 978-7-5641-4906-2

定 价：67.00 元

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

学习R语言 (影印版)

Learning R

Preface

R is a programming language and a software environment for data analysis and statistics. It is a GNU project, which means that it is free, open source software. It is growing exponentially by most measures—most estimates count over a million users, and it has over 4,000 add-on packages contributed by the community, with that number increasing by about 25% each year. The Tiobe Programming Community Index of language popularity places it at number 24 at the time of this writing, roughly on a par with SAS and MATLAB.

R is used in almost every area where statistics or data analyses are needed. Finance, marketing, pharmaceuticals, genomics, epidemiology, social sciences, and teaching are all covered, as well as dozens of other smaller domains.

About This Book

Since R is primarily designed to let you do statistical analyses, many of the books written about R focus on teaching you how to calculate statistics or model datasets. This unfortunately misses a large part of the reality of analyzing data. Unless you are doing cutting-edge research, the statistical techniques that you use will often be routine, and the modeling part of your task may not be the largest one. The complete workflow for analyzing data looks more like this:

1. Retrieve some data.
2. Clean the data.
3. Explore and visualize the data.
4. Model the data and make predictions.
5. Present or publish your results.

Of course at each stage your results may generate interesting questions that lead you to look for more data, or for a different way to treat your existing data, which can send you back a step. The workflow can be iterative, but each of the steps needs to be undertaken.

The first part of this book is designed to teach you R from scratch—you don't need any experience in the language. In fact, no programming experience *at all* is necessary, but if you have some basic programming knowledge, it will help. For example, the book explains how to comment your code and how to write a for loop, but doesn't explain in great detail what they are. If you want a really introductory text on how to program, then *Python for Kids* by Jason R. Briggs is as good a place to start as any!

The second part of the book takes you through the complete data analysis workflow in R. Here, some basic statistical knowledge is assumed. For example, you should understand terms like *mean* and *standard deviation*, and what a bar chart is.

The book finishes with some more advanced R topics, like object-oriented programming and package creation. Garrett Grolemund's *Data Analysis with R* picks up where this book leaves off, covering data analysis workflow in more detail.

A word of warning: this isn't a reference book, and many of the topics aren't covered in great detail. This book provides tutorials to give you ideas about what you can do in R and let you practice. There isn't enough room to cover all 4,000 add-on packages, but by the time you've finished reading, you should be able to find the ones that you need, and get the help you need to start using them.

What Is in This Book

This is a book of two halves. The first half is designed to provide you with the technical skills you need to use R; each chapter is a short introduction to a different set of data types (for example, Chapter 4 covers vectors, matrices, and arrays) or a concept (for example, Chapter 8 covers branching and looping).

The second half of the book ramps up the fun: you get to see real data analysis in action. Each chapter covers a section of the standard data analysis workflow, from importing data to publishing your results.

Here's what you'll find in Part I, The R Language:

- Chapter 1, *Introduction*, tells you how to install R and where to get help.
- Chapter 2, *A Scientific Calculator*, shows you how to use R as a scientific calculator.
- Chapter 3, *Inspecting Variables and Your Workspace*, lets you inspect variables in different ways.
- Chapter 4, *Vectors, Matrices, and Arrays*, covers vectors, matrices, and arrays.

- Chapter 5, *Lists and Data Frames*, covers lists and data frames (for spreadsheet-like data).
- Chapter 6, *Environments and Functions*, covers environments and functions.
- Chapter 7, *Strings and Factors*, covers strings and factors (for categorical data).
- Chapter 8, *Flow Control and Loops*, covers branching (if and else), and basic looping.
- Chapter 9, *Advanced Looping*, covers advanced looping with the apply function and its variants.
- Chapter 10, *Packages*, explains how to install and use add-on packages.
- Chapter 11, *Dates and Times*, covers dates and times.

Here are the topics covered in Part II, The Data Analysis Workflow:

- Chapter 12, *Getting Data*, shows you how to import data into R.
- Chapter 13, *Cleaning and Transforming*, explains cleaning and manipulating data.
- Chapter 14, *Exploring and Visualizing*, lets you explore data by calculating statistics and plotting.
- Chapter 15, *Distributions and Modeling*, introduces modeling.
- Chapter 16, *Programming*, covers a variety of advanced programming techniques.
- Chapter 17, *Making Packages*, shows you how to package your work for others.

Lastly, there are useful references in Part III, Appendixes:

- Appendix A, *Properties of Variables*, contains tables comparing the properties of different types of variables.
- Appendix B, *Other Things to Do in R*, describes some other things that you can do in R.
- Appendix C, *Answers to Quizzes*, contains the answers to the end-of-chapter quizzes.
- Appendix D, *Solutions to Exercises*, contains the answers to the end of chapter programming exercises.

Which Chapters Should I Read?

If you have never used R before, then start at the beginning and work through chapter by chapter. If you already have some experience with R, you may wish to skip the first chapter and skim the chapters on the R core language.

Each chapter deals with a different topic, so although there is a small amount of dependency from one chapter to the next, it is possible to pick and choose chapters that interest you.

I recently discussed this matter with Andrie de Vries, author of *R For Dummies*. He suggested giving up and reading his book instead!¹

Conventions Used in This Book

The following font conventions are used in this book:

Italic

Indicates new terms, URLs, email addresses, file and pathnames, and file extensions.

Constant width

Used for code samples that should be copied verbatim, as well as within paragraphs to refer to program elements such as variable or function names, data types, environment variables, statements, and keywords. Output from blocks of code is also in constant width, preceded by a double hash (##).

Constant width italic

Shows text that should be replaced with user-supplied values or by values determined by context.

There is a style guide for the code used in this book at <http://4dpiecharts.com/r-code-style-guide>.



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



This icon indicates a warning or caution.

Goals, Summaries, Quizzes, and Exercises

Each chapter begins with a list of goals to let you know what to expect in the forthcoming pages, and finishes with a summary that reiterates what you've learned. You also get a quiz, to make sure you've been concentrating (and not just pretending to read while watching telly). The answers to the questions can be found within the chapter (or at the

1. Andrie's book covers much the same ground as *Learning R*, and in many ways is almost as good as this work, so I won't be offended if you want to read it too.

end of the book, if you want to cheat). Finally, each chapter concludes with some exercises, most of which involve you writing some R code. After each exercise description there is a number in square brackets, denoting a generous estimate of how many minutes it might take you to complete it.

Using Code Examples

Supplemental material (code examples, exercises, etc.) is available for download at <http://cran.r-project.org/web/packages/learningr>.

This book is here to help you get your job done. In general, if example code is offered with this book, you may use it 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: "*Learning R* by Richard Cotton (O'Reilly). Copyright 2013 Richard Cotton, 978-1-449-35710-8."

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 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/learningR>.

To comment or 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

Many amazing people have helped with the making of this book, not least my excellent editor Meghan Blanchette, who is full of sensible advice.

Data was donated by several wonderful people:

- Bill Hogan of AMD found and cleaned the Alpe d'Huez cycling dataset, and pointed me toward the CDC gonorrhoea dataset. He wanted me to emphasize that he's disease-free, ladies.
- Ewan Hunter of CEFAS provided the North Sea crab dataset.
- Corina Logan of the University of Cambridge compiled and provided the deer skull data.
- Edwin Thoen of Leiden University compiled and provided the Obama vs. McCain dataset.
- Gwern Branwen compiled the hafu dataset by watching and reading an inordinate amount of manga. Kudos.

Many other people sent me datasets; there wasn't room for them all, but thank you anyway!

Bill Hogan also reviewed the book, as did Daisy Vincent of Marin Software, and JD Long. I don't know where JD works, but he lives in Bermuda, so it probably involves triangles. Additional comments and feedback were provided by James White, Ben Hanks, Beccy Smith, and Guy Bourne of TDX Group; Alex Hogg and Adrian Kelsey of HSL; Tom Hull, Karen Vanstaen, Rachel Beckett, Georgina Rimmer, Ruth Wortham, Bernardo Garcia-Carreras, and Joana Silva of CEFAS; Tal Galili of Tel Aviv University; Garrett Grolemond of RStudio; and John Verzani of the City University of New York. David Maxwell of CEFAS wonderfully recruited more or less everyone else in CEFAS to review my book.

John Verzani also deserves much credit for helping conceive this book, and for providing advice on the structure.

Sanders Kleinfeld of O'Reilly provided great tech support when I was pulling my hair out over character encodings in the manuscript. Yihui Xie went above and beyond the call of duty helping me get `knitr` to generate AsciiDoc. Rachel Head single-handedly spotted over 4,000 bugs, typos, and mistakes while copyediting.

Garib Murshudov was the lecturer who first taught me R, back in 2004.

Finally, Janette Bowler deserves a medal for her endless patience and support while I've been busy writing.

Table of Contents

Preface.....	xiii
--------------	------

Part I. The R Language

1. Introduction.....	3
Chapter Goals	3
What Is R?	3
Installing R	4
Choosing an IDE	5
Emacs + ESS	5
Eclipse/Architect	6
RStudio	6
Revolution-R	7
Live-R	7
Other IDEs and Editors	7
Your First Program	8
How to Get Help in R	8
Installing Extra Related Software	11
Summary	11
Test Your Knowledge: Quiz	12
Test Your Knowledge: Exercises	12
2. A Scientific Calculator.....	13
Chapter Goals	13
Mathematical Operations and Vectors	13
Assigning Variables	17
Special Numbers	19
Logical Vectors	20
Summary	22

Test Your Knowledge: Quiz	22
Test Your Knowledge: Exercises	23
3. Inspecting Variables and Your Workspace	25
Chapter Goals	25
Classes	25
Different Types of Numbers	26
Other Common Classes	27
Checking and Changing Classes	30
Examining Variables	33
The Workspace	36
Summary	37
Test Your Knowledge: Quiz	37
Test Your Knowledge: Exercises	37
4. Vectors, Matrices, and Arrays	39
Chapter Goals	39
Vectors	39
Sequences	41
Lengths	42
Names	42
Indexing Vectors	43
Vector Recycling and Repetition	45
Matrices and Arrays	46
Creating Arrays and Matrices	46
Rows, Columns, and Dimensions	48
Row, Column, and Dimension Names	50
Indexing Arrays	51
Combining Matrices	51
Array Arithmetic	52
Summary	54
Test Your Knowledge: Quiz	55
Test Your Knowledge: Exercises	55
5. Lists and Data Frames	57
Chapter Goals	57
Lists	57
Creating Lists	57
Atomic and Recursive Variables	60
List Dimensions and Arithmetic	60
Indexing Lists	61
Converting Between Vectors and Lists	64

Combining Lists	65
NULL	66
Pairlists	70
Data Frames	70
Creating Data Frames	71
Indexing Data Frames	74
Basic Data Frame Manipulation	75
Summary	77
Test Your Knowledge: Quiz	77
Test Your Knowledge: Exercises	78
6. Environments and Functions.....	79
Chapter Goals	79
Environments	79
Functions	82
Creating and Calling Functions	82
Passing Functions to and from Other Functions	86
Variable Scope	89
Summary	91
Test Your Knowledge: Quiz	91
Test Your Knowledge: Exercises	91
7. Strings and Factors.....	93
Chapter Goals	93
Strings	93
Constructing and Printing Strings	94
Formatting Numbers	95
Special Characters	97
Changing Case	98
Extracting Substrings	98
Splitting Strings	99
File Paths	100
Factors	101
Creating Factors	101
Changing Factor Levels	103
Dropping Factor Levels	103
Ordered Factors	104
Converting Continuous Variables to Categorical	105
Converting Categorical Variables to Continuous	106
Generating Factor Levels	107
Combining Factors	107
Summary	108

Test Your Knowledge: Quiz	108
Test Your Knowledge: Exercises	108
8. Flow Control and Loops	111
Chapter Goals	111
Flow Control	111
if and else	112
Vectorized if	114
Multiple Selection	115
Loops	116
repeat Loops	116
while Loops	118
for Loops	120
Summary	122
Test Your Knowledge: Quiz	122
Test Your Knowledge: Exercises	122
9. Advanced Looping	125
Chapter Goals	125
Replication	125
Looping Over Lists	127
Looping Over Arrays	132
Multiple-Input Apply	135
Instant Vectorization	136
Split-Apply-Combine	136
The plyr Package	138
Summary	141
Test Your Knowledge: Quiz	141
Test Your Knowledge: Exercises	141
10. Packages	143
Chapter Goals	143
Loading Packages	144
The Search Path	146
Libraries and Installed Packages	146
Installing Packages	148
Maintaining Packages	150
Summary	150
Test Your Knowledge: Quiz	151
Test Your Knowledge: Exercises	151
11. Dates and Times	153

Chapter Goals	153
Date and Time Classes	154
POSIX Dates and Times	154
The Date Class	155
Other Date Classes	156
Conversion to and from Strings	156
Parsing Dates	156
Formatting Dates	157
Time Zones	158
Arithmetic with Dates and Times	160
Lubridate	161
Summary	165
Test Your Knowledge: Quiz	165
Test Your Knowledge: Exercises	166

Part II. The Data Analysis Workflow

12. Getting Data	169
Chapter Goals	169
Built-in Datasets	169
Reading Text Files	170
CSV and Tab-Delimited Files	170
Unstructured Text Files	175
XML and HTML Files	175
JSON and YAML Files	176
Reading Binary Files	179
Reading Excel Files	179
Reading SAS, Stata, SPSS, and MATLAB Files	181
Reading Other File Types	181
Web Data	182
Sites with an API	182
Scraping Web Pages	184
Accessing Databases	185
Summary	188
Test Your Knowledge: Quiz	189
Test Your Knowledge: Exercises	189
13. Cleaning and Transforming	191
Chapter Goals	191
Cleaning Strings	191
Manipulating Data Frames	196

Adding and Replacing Columns	196
Dealing with Missing Values	197
Converting Between Wide and Long Form	198
Using SQL	200
Sorting	201
Functional Programming	202
Summary	204
Test Your Knowledge: Quiz	205
Test Your Knowledge: Exercises	205
14. Exploring and Visualizing	207
Chapter Goals	207
Summary Statistics	207
The Three Plotting Systems	211
Scatterplots	212
Take 1: base Graphics	213
Take 2: lattice Graphics	218
Take 3: ggplot2 Graphics	224
Line Plots	230
Histograms	238
Box Plots	249
Bar Charts	253
Other Plotting Packages and Systems	260
Summary	261
Test Your Knowledge: Quiz	261
Test Your Knowledge: Exercises	262
15. Distributions and Modeling	263
Chapter Goals	263
Random Numbers	264
The sample Function	264
Sampling from Distributions	265
Distributions	266
Formulae	267
A First Model: Linear Regressions	268
Comparing and Updating Models	271
Plotting and Inspecting Models	276
Other Model Types	280
Summary	282
Test Your Knowledge: Quiz	282