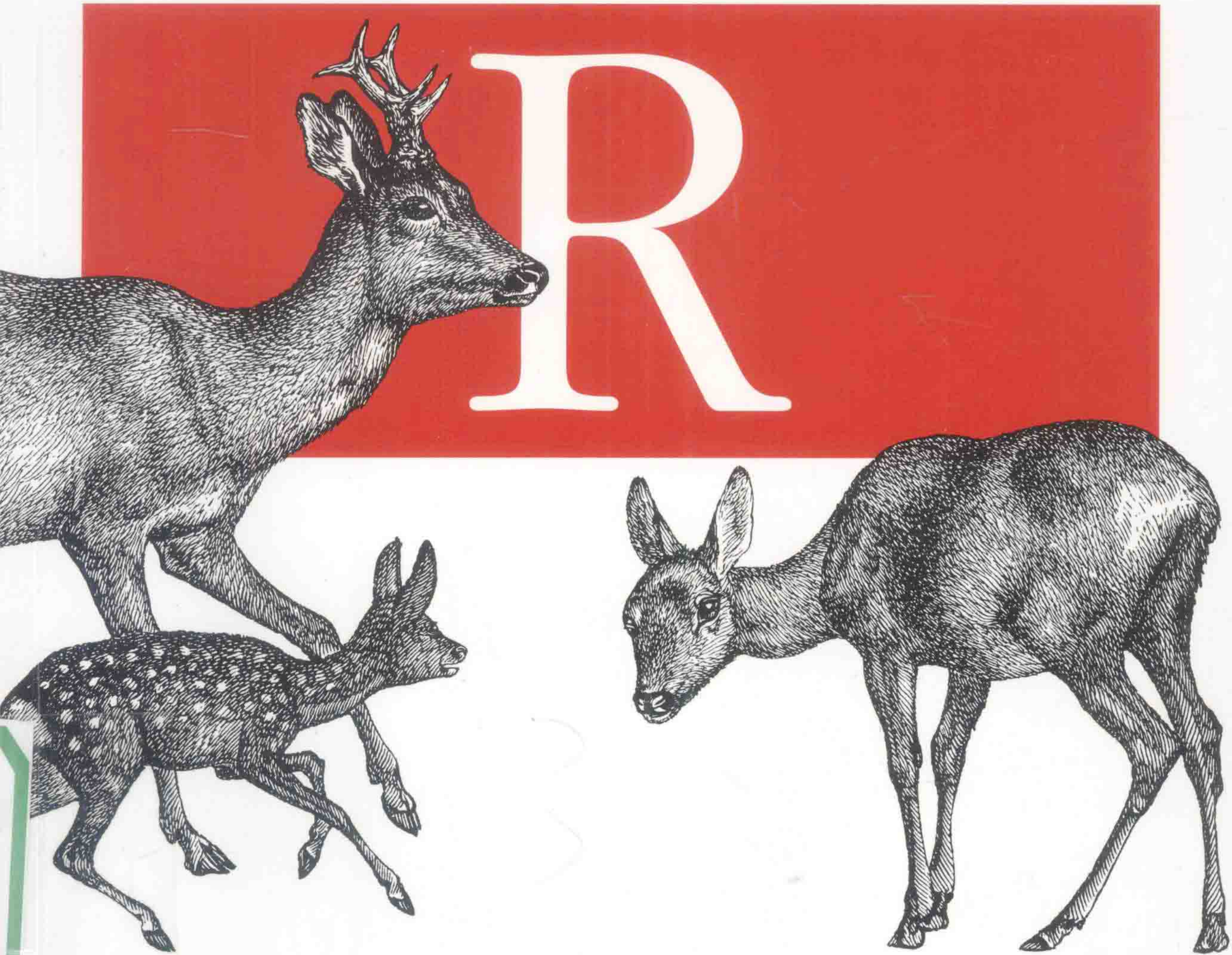
学习R语言（影印版）

Learning R



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

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

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 Grolemund 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.

About the Author

Richard Cotton is a data scientist with a background in chemical health and safety, and has worked extensively on tools to give nontechnical users access to statistical models. He is the author of the R packages `assertive` (for checking the state of your variables) and `sig` (to make sure your functions have a sensible API). He runs The Damned Liars statistics consultancy.

Colophon

The animals on the cover of *Learning R* are roe deer (*Capreolus capreolus*), a species of deer found throughout much of Europe, Scandinavia, and the Mediterranean region. *Roe* is derived from the Old English word for "spotted," though other translations show that it might be an ancient word for "red."

The roe deer is rather small (averaging 3–4 feet long and around 2 feet tall at the shoulders), with long graceful legs. Male roe have short antlers with just a few branches, but are otherwise nearly the same size as female roe. These deer also have very short tails with a white rump patch, which they flash when alarmed by something. In the summer, their coats are red, but fade to gray or pale brown in the winter. Their fawns are born in summer after a 10-month gestation period, and have white spots for their first six weeks of life.

Woodland is the preferred habitat of the roe deer, though they also graze in grasslands and sparse forests. Occasionally they can be found near farms, but tend to avoid fields where livestock have been kept, perhaps because the grass is trampled and less clean. They eat grass, shoots, leaves, and berries, and are most active at twilight.

In the Austrian books upon which the classic Disney film *Bambi* was based, Bambi was a roe deer—but Disney changed the character to a white-tailed deer, which was more familiar to North American audiences.

The cover image is from Cassell's *Natural History*. The cover font is Adobe ITC Garamond. The text font is Adobe Minion Pro; the heading font is Adobe Myriad Condensed; and the code font is Dalton Maag's Ubuntu Mono.

O’Reilly Media, Inc.介绍

O’Reilly Media通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自1978年开始，O’Reilly一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O’Reilly的发展充满了对创新的倡导、创造和发扬光大。

O’Reilly为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了Make杂志，从而成为DIY革命的主要先锋；一如既往地通过多种形式缔结信息与人的纽带。O’Reilly的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O’Reilly现在还将先锋专家的知识传递给普通的计算机用户。无论是书籍出版、在线服务还是面授课程，每一项O’Reilly的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O’Reilly Radar博客有口皆碑。”

——Wired

“O’Reilly凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——Business 2.0

“O’Reilly Conference是聚集关键思想领袖的绝对典范。”

——CRN

“一本O’Reilly的书就代表一个有用、有前途、需要学习的主题。”

——Irish Times

“Tim是位特立独行的商人，他不光放眼于最长远、最广阔的视野，并且切实地按照Yogi Berra的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去，Tim似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——Linux Journal

出版说明

随着计算机技术的成熟和广泛应用，人类正在步入一个技术迅猛发展的新时期。计算机技术的发展给人们的工业生产、商业活动和日常生活都带来了巨大的影响。然而，计算机领域的技术更新速度之快也是众所周知的，为了帮助国内技术人员在第一时间了解国外最新的技术，东南大学出版社和美国 O'Reilly Meida, Inc. 达成协议，将陆续引进该公司的代表前沿技术或者在某专项领域享有盛名的著作，以影印版或者简体中文版的形式呈献给读者。其中，影印版书籍力求与国外图书“同步”出版，并且“原汁原味”展现给读者。

我们真诚地希望，所引进的书籍能对国内相关行业的技术人员、科研机构的研究人员和高校师生的学习和工作有所帮助，对国内计算机技术的发展有所促进。也衷心期望读者提出宝贵的意见和建议。

最新出版的影印版图书，包括：

- 《学习 R 语言》（影印版）
- 《游戏设计之快乐理论 第 2 版》（影印版）
- 《深入浅出 C# 第 3 版》（影印版）
- 《数据科学》（影印版）
- 《Java 网络编程 第 4 版》（影印版）
- 《高性能浏览器网络》（影印版）
- 《行为变化设计》（影印版）
- 《深入浅出 PMP 第 3 版》（影印版）
- 《深入浅出 JavaScript 编程》（影印版）
- 《深入浅出 iPhone 和 iPad 开发 第 3 版》（影印版）
- 《挖掘社交网络 第 2 版》（影印版）
- 《精通 Perl 第 2 版》（影印版）
- 《基于数据分析的网络安全》（影印版）

Table of Contents