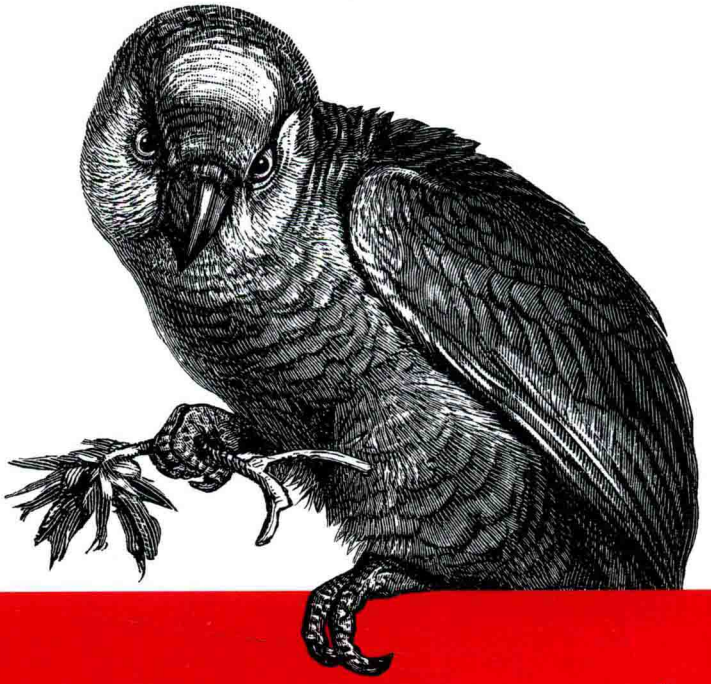
O'REILLY®

TURING 图灵程序设计丛书



R语言入门与实践

Hands-On Programming with R

将R编程所需的方方面面巧妙融合在三个精心挑选的示例中，
助你轻松掌握R语言，为成为优秀的数据科学家奠定坚实基础

[美] Garrett Grolemund 著
[新西兰] Hadley Wickham 序
冯凌秉 译

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

R语言入门与实践

Hands-On Programming with R: Write Your Own Functions and Simulations

[美] Garrett Grolemund 著
冯凌秉 译

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

O'Reilly Media, Inc.授权人民邮电出版社出版

人民邮电出版社
北　京

图书在版编目（CIP）数据

R语言入门与实践 / (美) 格罗勒芒德(Grolemund,G.) 著 ; 冯凌秉译. -- 北京 : 人民邮电出版社, 2016.6
（图灵程序设计丛书）
ISBN 978-7-115-42471-6

Ⅰ. ①R… Ⅱ. ①格… ②冯… Ⅲ. ①程序语言－程序设计 Ⅳ. ①TP312

中国版本图书馆CIP数据核字(2016)第109791号

内 容 提 要

本书通过三个精心挑选的例子，深入浅出地讲解如何使用R语言玩转数据。书中涵盖R语言编程的方方面面，内容涉及R对象的类型、R的记号体系和环境系统、自定义函数、`if else`语句、`for`循环、S3类、R的包系统以及调试工具等。本书还通过示例演示如何进行向量化编程，从而对代码进行提速并最大化地发挥R的潜能。

本书适合立志成为数据科学家的R语言初学者阅读。

◆ 著 [美] Garrett Grolemund
译 冯凌秉
责任编辑 朱 巍
执行编辑 谢婷婷
责任印制 彭志环

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路11号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷

◆ 开本：800×1000 1/16
印张：14.5
字数：338千字 2016年6月第1版
印数：1－4 000册 2016年6月北京第1次印刷
著作权合同登记号 图字：01-2015-5426号

定价：59.00元

读者服务热线：(010)51095186转600 印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广字第8052号

站在巨人的肩上

Standing on Shoulders of Giants

iTuring.cn

版权声明

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；公司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。无论是通过书籍出版、在线服务或者面授课程，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*

序

要想认真理解数据，学习编程至关重要。毋庸置疑，数据科学工作必须在计算机上完成，但你可以有两种选择：学会使用具有图形用户界面（GUI）的统计软件，或者学会一门编程语言。Garrett 和我都深信，对于每天跟数据打交道的人来说，编程是一项必需的技能。图形用户界面简单易用，但也有着根本的局限性，主要体现在它束缚了好的数据分析所应具备的以下三个属性。

- 可再现性
  可再现性指能够完全再现之前某个分析结果的能力，这是优秀科学的重要特征。

- 自动化
  当数据发生改变时（往往经常如此），能够快速更新分析结果的能力。

- 沟通
  从形式上看，代码仅仅是一段文本，十分利于沟通。对于一门编程语言的学习者来说，代码的文本特征极大地降低了学习的门槛：只要有不懂的问题，就可以把代码放到邮件列表中提问，或者放到 Google 的搜索框中搜索，或者上传到 Stack Overflow 找高手解惑，无论哪种途径都能够得到解答。

不要对“编程”二字望而生畏！只要有合适的动力，每个人都可以学好编程，而本书的目的就是充分调动你学习编程的积极性。这并不是一本编程参考书，它围绕着三个实际的编程挑战展开。如果能够顺利地掌握应对这三个挑战的技术，你将掌握关于 R 编程的基本知识，甚至还能学习一些中级技能，比如向量化编程、作用域和 S3 方法等。动手解决实际问题对于学习一门编程语言来说十分重要。学习编程不在于能够死记硬背多少函数，而在于学习函数并将之用于解决实际问题。总而言之，编程需要在实践中学习，而不是靠死记硬背。

学习编程的道路布满荆棘，你会不时地因为棘手的问题而感到沮丧。与学习其他语言一

样，熟练掌握程序语言并非一朝一夕之事。在学习编程的过程中感到沮丧是难免的，它听起来十分消极，但实际上对于学习本身来说是有所裨益的。沮丧的感觉是由大脑的懒惰造成的，它在提示你放弃做这么难的事情，去找些容易或有趣的事情来做。有过健身经历的人都知道，如果想更加健美，身体越是抗拒，你就越要鞭策自己去努力锻炼。学习编程也一样，大脑越是感到沮丧，你越应该鞭策自己去战胜沮丧。认识到这一点之后，在学习编程的过程中，如果遇到了令人沮丧的难题，要乐观积极地对待这样的感觉：因为你是在挑战自我。如果每天都能挑战自己一点点，可以预见，不用多久你就会成为一名信心满满的程序员。

本书浅显易懂，娓娓道来，又十分深入与实用。如果你想跟我或者 Garrett 学习 R 编程却没有机会与我们当面交流，这本书是不二之选。读这本书于我而言是一种享受，希望你也有同感。

——Hadley Wickham
RStudio 首席科学家

顺便说一句，Garrett 完全没有介绍自己的成果，其实有点过谦了。他写的 R 包 lubridate 是处理时间类数据的必备神器，不信大家可以试一试。

前言

这本书的初衷是教会你用 R 编程，从基本的加载数据到编写自定义函数（编写出比其他 R 用户更好的函数）。但是从内容上来说，这本书并不同于市面上其他的 R 导论类型的书籍，其更高远的目标是帮助你成为数据科学家和计算机科学家。因此这本书着重传授的 R 编程技能更贴近数据科学的要求。

本书内容分为三大块，其中每一块都对应一个实际的项目。考虑到每一个项目的内容都比较多，因此每一块又细分为几章。选择这三个项目是基于两个方面的原因。首先，它们基本涵盖了 R 语言编程的方方面面。具体来说，你将学到如何加载数据，组合与拆解数据对象，玩转 R 的系统环境，编写自己的 R 函数，以及使用 R 的所有编程工具，比如 `if else` 语句、`for` 循环、S3 类、R 的包系统以及调试工具等。这三个项目也会涉及如何编写向量化的 R 代码，这种代码风格的好处在于其速度较快，并且能够最大化地发挥 R 的潜能。

更为重要的是，数据科学中涉及很多“数据物流问题”，而这三个精心挑选的项目将教你如何解决这些问题。在和数据打交道的时候，所谓的数据物流问题指的是如何高效无误地存储、检索和操作大规模的数据。读完这本书，你会发现我们的目标不仅在于学习如何利用 R 语言进行编程，更在于如何利用编程技术解决数据科学家在工作中会遇到的各种实际问题。

并非所有程序员都需要成为数据科学家，因此这本书并不一定适合所有程序员。但是，如果你属于以下两类人之一，一定会发现这本书很有助益。

(1) 已经把 R 当作一个统计研究的工具，并且想要使用 R 编写自己的函数和模拟。
(2) 想自学编程，并且感到有必要学习一门与数据科学相关的编程语言。

R 语言的传统强项是建模与作图，市面上关于 R 语言的书籍基本都以这两点为主线。相比之下，本书最大的特色之一就在于将 R 视为一门纯粹的编程语言，而不是统计软件。因此我们将淡化这两大主线，将视角集中在编程语言特色上。这样的视角看似有些偏颇，因为 R 设计之初是作为一个开源的统计软件，其宗旨是帮助科学家解决数据分析问题。

它内置了许多精良的数据分析函数，并且在数据可视化和统计模型方面有着明显的优越性。因此，许多统计学家将 R 视为一个软件，需要哪个函数就学习哪个函数，而忽视其他的部分。

这样的视角和学习方法其实也是有道理的，数据科学家在统计模型和数据可视化上的确应该倾注大量的精力。想要从数据中得出可靠的结论，数据科学家需要具备扎实的专业知识、敏锐的洞察力和专注精神。我并不建议你在了解基本的数据科学理论和工具之前，就把大量精力放在学习计算机编程上。鉴于此，如果你想学习数据科学的基础知识，我向你推荐本书的姊妹篇 *R for Data Science* [1]。

即便如此，编程也应该成为每一个数据科学家的必修课。学会编程不仅能够让你在分析数据时游刃有余，还能够从各个方面提升你掌握数据科学的能力。关于这一点，Greg Snow 在 2006 年 5 月的 R 帮助邮件列表中有过一段十分精妙的比喻。使用 R 中现成的函数好比是乘坐公共汽车，而编写自己的 R 函数则类似于自己驾驶汽车。

> 搭乘公共汽车再容易不过了。你只需要知道该搭乘哪一路车，在哪里上车，又在哪里下车（当然，上车后你得投币或者刷卡）。但是自己驾车就远没有这么简单了：你需要一张地图或者起码是一条明确的导航信息（即使你已将路线记在脑子里）；你还需要不时地给汽车加油，了解并遵守交通规则（你得有驾照）。相比于搭乘公共汽车，自己驾车的优点是更加自由，你可以去公共汽车去不了的许多地方。并且对于某些需要换乘公共汽车的旅程来说，自驾往往能够更快地到达目的地。
>
> 使用类似 SPSS 这样的统计软件就好比是搭乘公共汽车。对于某些基本和传统的统计分析来说，SPSS 简单易用。但是，当你想做一些 SPSS 没有内置模块支持的分析时，就十分无助了。
>
> R 就好比一辆 4 驱 SUV（当然，R 不需要汽油，因此环保），车背驮了一辆山地自行车，车顶备了一艘小型皮艇，车内还有慢跑鞋、登山装备和洞穴探险装备。
>
> 你想去哪儿，R 都能做到，只要你肯花时间学习如何使用各种装备。但是，与像公共汽车的 SPSS 相比，学习 R 需要更多的时间和精力。
>
> ——Greg Snow

Greg 做这个对比的时候，是默认大家把 R 当作一门编程语言来使用的；换句话说，你知道如何用 R 语言编程以实现自己想要的分析。如果你只是使用 R 的众多包提供的现成函数的话，那么 R 对于你来说只相当于另一个版本的 SPSS；也就是说，你把它当作了一辆公共汽车，因此你能到达的地方也十分有限。

注 1：本书中文版也将由人民邮电出版社图灵公司出版。另外已经出版的《数据科学入门》也是介绍数据科学基础知识的重量级读本。——编者注

灵活性对于数据科学家来说至关重要。对于不同的数据，统计模型和数据模拟在细节上会有差异。如果不能因地制宜地使用合适的方法，你就可能设定不现实的模型假设，而使用现成的但又不太匹配手头数据的方法。

这本书的目的恰恰是要帮助你从公共汽车的乘客转变为 SUV 的驾驶员。本书面向广大的编程初学者。书中既不会讨论关于计算机科学的任何理论知识，比如大 O() 和小 o() 的区别，也不会讨论编程语言的底层细节，比如惰性求值（lazy evaluation）机制。如果想学习计算机科学的理论知识，这些高阶的细节固然重要，但是对于初学编程的人来说不是必需的。

我所准备做的，是通过三个精心挑选的例子，把 R 编程所需要的方方面面的知识都传授给你。这些例子不会过于冗长，并且易于理解。

作为 RStudio 的高级培训师，我已多次使用这些例子讲授 R 课程。我发现，实例教学有利于学习者快速理解和掌握抽象的编程概念。通过实例学习编程的另一个好处是，它提供了绝佳的练习机会。学习编程就像学习一门新的语言。通过学习书中提供的这些例子，并动手实践，你的 R 编程学习之路将变得更加平坦。

> 本书的姊妹篇叫作 *R for Data Science*，即将出版。这本书详细介绍了如何用 R 作图、建模、编写报表，等等。它更注重数据科学的实用技能，并要求读者具备一定的数据科学基础知识。本书则侧重于如何使用 R 编程。它并不要求读者具有数据科学技能，当然如果具备一定的基础知识对于学习本书的内容也是大有裨益的。如果你同时掌握了数据科学和编程技能，那么恭喜你，你已经成为一位响当当的数据科学家，不仅可以挺直腰板找老板加薪了，还在数据科学领域更具话语权。

排版约定

本书使用下列排版约定。

- 楷体
  表示新术语。

- 等宽字体（`constant width`）
  表示程序片段，以及正文中出现的变量、函数名、数据库、数据类型、环境变量、语句和关键字等。

- 加粗等宽字体（**`constant width bold`**）
  表示应该由用户输入的命令或其他文本。

- 斜体等宽字体（*`constant width italic`*）
  表示应该用用户输入的值或者根据上下文确定的值替换的文本。

该图标表示此处有提示或建议

该图标表示一般性注解

该图标表示警告或者需要额外注意

Safari® Books Online

Safari Books Online（http://www.safaribooksonline.com）是应运而生的数字图书馆。它同时以图书和视频的形式出版世界顶级技术和商务作家的专业作品。技术专家、软件开发人员、Web设计师、商务人士和创意专家等，在开展调研、解决问题、学习和认证培训时，都将Safari Books Online视作获取资料的首选渠道。

对于组织团体、政府机构和个人，Safari Books Online提供各种产品组合和灵活的定价策略。用户可通过一个功能完备的数据库检索系统访问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以及其他几十家出版社的上千种图书、培训视频和正式出版之前的书稿。要了解Safari Books Online的更多信息，我们网上见。

联系我们

请把对本书的评价和问题发给出版社。

美国：

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472

中国：

北京市西城区西直门南大街 2 号成铭大厦 C 座 807 室（100035）
奥莱利技术咨询（北京）有限公司

O'Reilly 的每一本书都有专属网页，你可以在那儿找到本书的相关信息，包括勘误表、示例代码以及其他信息。本书的网站地址是：
http://shop.oreilly.com/product/0636920028574.do

对于本书的评论和技术性问题，请发送电子邮件到：
bookquestions@oreilly.com

要了解更多 O'Reilly 图书、培训课程、会议和新闻的信息，请访问以下网站：
http://www.oreilly.com

我们在 Facebook 的地址如下：http://facebook.com/oreilly

请关注我们的 Twitter 动态：http://twitter.com/oreillymedia

我们的 YouTube 视频地址如下：http://www.youtube.com/oreillymedia

致谢

本书的写作和出版得到了很多优秀人士的支持和帮助，包括本书的两位编辑 Courtney Nash 和 Julie Steele，以及负责版面设计、校对及书页索引的 O'Reilly 团队成员。另外，Greg Snow 慷慨地允许我在前言中引用他的话，在此表示感谢。

我还要感谢 Hadley Wickham，他帮助我形成了自己思考 R 和传授 R 的方式。这本书中的很多想法和内容都来自“统计 405”这门课。我在莱斯大学读博士的时候是这门课的助教，而 Hadley 正是这门课的讲师。

此外，我代表 RStudio 举办过主题为“R 与数据科学简介”的研讨会，参与研讨会的学生和老师都给了我很多灵感与建议，在此向他们表示感谢。还要特别感谢我的助教们，他们是 Josh Paulson、Winston Chang、Jaime Ramos、Jay Emerson 和 Vivian Zhang。

感谢 JJ Allaire 以及我在 RStudio 的其他同事，他们创造并分享了著名的 RStudio IDE（RStudio 集成开发环境），极大地简化了人们使用、传授和编写 R 的方式。

最后，我要感谢我的妻子 Kristin，在本书的写作过程中给予我莫大的支持与理解。

目录

第二部分　项目 2：玩扑克牌

第三部分 项目 3：老虎机