异步图书 www.epubit.com.cn

高效数据分析师实用指南

Packt

R语言编程指南

Learning
R Programming

任坤 著
王婷 赵孟韬 王泽贤 译

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

R语言编程指南

Learning
R Programming

任坤 著
王婷 赵孟韬 王泽贤 译

人民邮电出版社
北京

图书在版编目（CIP）数据

R语言编程指南 / 任坤著 ; 王婷, 赵孟韬, 王泽贤译. -- 北京 : 人民邮电出版社, 2017.10
ISBN 978-7-115-46264-0

Ⅰ. ①R… Ⅱ. ①任… ②王… ③赵… ④王… Ⅲ. ①程序语言－程序设计－指南 Ⅳ. ①TP312-62

中国版本图书馆CIP数据核字(2017)第193676号

版权声明

◆ 著　　任　坤
译　　王　婷　赵孟韬　王泽贤
责任编辑　胡俊英
责任印制　焦志炜

◆ 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
北京市艺辉印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：34
字数：676 千字　　2017 年 10 月第 1 版
印数：1－2 400 册　　2017 年 10 月北京第 1 次印刷

著作权合同登记号　图字：01-2016-9391 号

定价：99.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

内容提要

R 是一个开源、跨平台的科学计算和统计分析软件包，它提供了丰富多样的统计功能和强大的数据分析功能。随着数据科学的快速发展，R 已经成为数据分析领域非常流行的语言。

本书通过 15 章内容，向读者全面讲解了 R 的基础知识和编程技巧。本书不仅介绍了 R 的安装、基本对象、工作空间管理、基本表达式、基本对象操作、字符串的使用等基础内容，还对数据处理、R 的内部机制、元编程、面向对象编程、数据库操作、数据操作进行了讲解，同时也涉及高性能计算、网页爬虫和效率提升等重要主题。

本书面向数据领域的从业人员，尤其适合想要通过学习 R 编程及相关工具提升数据处理效率的读者阅读，也适合计算机或统计相关专业的学生参考使用。通过阅读本书，读者将全面掌握 R 的相关特性及其在数据处理和分析方面的应用，极大地提升自己的专业技能。

序

我毕业后一直在量化投资的一线工作，每天大量的工作都是以R语言为主要工具来研究金融数据，期间也在GitHub上开发和维护着几个R扩展包，每年也参加几场R语言会议。在这个过程中，我接触了不少还在学校的初学者，或者已经步入数据相关工作的研究人员，也有一些发来邮件寻求帮助的世界各地的用户。我有这样一种感觉，我们的同学、数据研究者经常有丰富的想法，但原始数据的形式与这些想法常常有相当大的距离。许多用户是因为对工具和编程本身不够熟悉而难以自由地操作数据，因而在面对稍显复杂的问题时便止步不前。如果是这样的原因放慢了我们探索数据世界的脚步，岂不是太可惜了？于是，我萌生了一个想法，写一本关于R语言编程的书。2015年10月，Packt出版社的编辑邀请我写一本面向初学者和专业用户的R语言图书，这正合我的想法！经过一年的时间，便有了本书。

本书与其他R语言图书有一个重要的不同：该书更倾向于帮助读者系统化地认识R作为一门编程语言的设计和行为，通过许多例子和实验帮助读者弄清楚R语言中各种常用数据结构的行为，以及所有这些行为背后统一的设计原则和行为准则。对于许多初学者以及其他编程语言的用户来说，R语言是难以预料、充满怪癖的，至少不是十全十美的。但是，当了解到这些统一的行为准则后，你可能会惊叹R语言本身的一致性，以及表达数据和逻辑的灵活性。这些特性允许我们高效地进行数据探索、分析、可视化、报告等。本书将用一半的篇幅来介绍基本的R语言和对象，然后探索R语言的高级特性，让读者更加深入地理解其行为，形成一个整体的知识脉络。此时，当你写出一个R语言表达式，就能立刻猜想会发生什么，即使和想象的不一样，也能很快找到问题所在。打好了这个基础后，我们会介绍数据相关的主要技术，包括关系性与非关系性数据库，实现快速数据操作的扩展包等。掌握了语言和对象的特性以及流行的扩展工具后，我们就可以随时根据问题选择工具，因而生产力就能大幅提升，可以将主要经历投入在思考和解决业务问题，而非一知半解、

绞尽脑汁地去找代码中的纰漏且摸不着头脑。最后，本书介绍了一系列工具，涵盖数据研究的多个方面，读者可以根据自己的需要继续学习。

本书原版为英文版，对于国内的读者可能阅读中文版更加方便。为了保证翻译的质量，我推荐厦门大学经济学院和王亚南经济研究院（WISE）的研究生学弟、学妹们翻译本书。他们来自于一个自发组织的数据科学小组 WISER Club，经受过严格的学术训练，参与过多种数据研究项目，对 R 语言相关的应用已经有相当的经验，并且对推广数据科学充满热情。

希望本书能让你更加深入地了解 R 语言和相关工具，更加自信、自由地探索数据的海洋。

——任坤

作者简介

任坤 在量化交易中使用 R 以及 C++和 C#已有近 4 年的时间，他一直致力于开发有用的但社区尚未提供的 R 包（每天工作 8～10 小时）。他为其他作者开发的扩展包做出过很多贡献，指出其中存在的问题并给出改进建议。他也是中国 R 语言大会的重要嘉宾，在 R 会议上做过多次演讲。在众多社交媒体中，任坤也受到了广泛关注。此外，他对很多项目都做出了很大贡献，从其 GitHub 账户可见一斑：

- https://github.com/renkun-ken
- https://cn.linkedin.com/in/kun-ren-76027530
- https://renkun.me
- https://renkun.me/formattable
- https://renkun.me/pipeR
- https://renkun.me/rlist

我要感谢我的爱人，她一年前就鼓励我开始编写这本书。同时还要感谢两位编辑，Rohit Kumar Singh 和 Vivek Pala，以及对本书做出贡献的所有人。他们的工作使本书的编写和出版都非常顺利。

技术评阅者简介

Kelly Black 佐治亚大学数学系的成员。他的研究方向是随机微分方程，在很多场景中都在使用统计软件，包括利用蒙特卡罗模拟进行数据分析、教育评估等。

我很感谢 Izzat Contractor 的友善和耐心，是他的帮助和指导使本书最终得以出版。

译者简介

本书译者王婷、赵孟韬、王泽贤是 WISER CLUB 的成员。WISER CLUB 是由经济学院、WISE 硕士研究生和本科生联合自发组织的数据科学互助学习组织，依托于厦门大学经济学院与王亚南经济研究院（WISE）强大的计量和统计背景以及丰富的数据科学资源。它旨在让更多的人了解、学习数据科学知识，分享数据科学在学术界和业界的最新动态，并且与学术界和业界合作向在校同学提供广阔的发展前景与机会。

本书凝结了 WISER CLUB 全体成员的努力付出。除三位主要译者外，感谢于海悦、杨琬妮对第 7 章和第 15 章以及王柳盈对第 11 章的翻译做出的重要贡献，感谢胡帆、洪祺琳、林嘉文、段孙蓬等 WISER CLUB 全体同学的支持与付出。

特别感谢贾茹学姐、黄耀鹏和邓光宏学长对校对工作的大力支持。

前言

R 是为统计计算、数据分析和可视化而设计的。近年来，它已成为数据科学和统计学中最受欢迎的语言。R 语言编程很大程度上涉及数据处理。对于不熟悉 R 语言的人来说，用 R 进行编程可能是一个挑战。

作为一种动态语言，R 可以灵活地使用不同的数据结构，不像 C++、Java 和 C#这类编译型语言那么严格。当我开始使用 R 处理和分析数据时，发现它的行为很古怪，不可预测，而且有时非常不稳定。

一些数据分析项目并没有在构建模型上做很多工作。相反，数据清洗、处理和可视化花了更多时间。事实上，在代码运行报错或返回的结果很奇怪时，找到问题的根源才是最耗时的。处理编程问题比处理专业领域内的问题更令人受挫，尤其是在遇到错误，但搞了几小时仍然毫无头绪的时候。

但是，随着项目的增多，也积累了更多的经验，我逐渐了解了对象和函数的行为，并且发现 R 比我想象的更优雅、更稳定。这就是我编这本书的原因，以分享我对 R 语言编程逐步深入的认识过程。

通过阅读本书，你会对 R 编程语言及其大量的工具有一个普遍一致的理解，并将学习到提高效率的最佳实践方法，更深入地了解如何使用数据，并且对如何在 R 中编程，以及用正确的技巧解决问题等更有信心。

本书主要内容

第 1 章“快速入门”讨论了一些有关 R 的基础内容，包括如何部署 R 环境，如何在

RStudio 中编写代码。

第 2 章“基本对象”介绍基本的 R 对象及其性质。

第 3 章“工作空间管理”介绍工作目录、R 环境和扩展包库的管理方法。

第 4 章“基本表达式”介绍 R 语言的基本表达式：赋值、条件和循环。

第 5 章“基本对象操作”讨论每个数据分析师都应该了解的基本函数，以便在 R 中使用基本对象。

第 6 章“字符串的使用”讨论与字符串相关的 R 对象，以及一些字符串操作技术。

第 7 章“数据处理”解释一些简单的读写数据的函数，并通过一些使用基本对象和函数的实际案例进行演示。

第 8 章“R 的内部机制”通过介绍惰性计算、环境、函数和词法作用域，探讨 R 的计算模式。

第 9 章“元编程”介绍元编程技术以帮助理解语言对象和非标准化求值。

第 10 章“面向对象编程”阐释 R 中众多的面向对象编程系统：S3、S4、RC 和社区提供的 R6。

第 11 章“数据库操作”介绍在 R 中如何使用 SQLite 和 MySQL 等流行的关系型数据库，以及 MongoDB 和 Redis 等非关系型数据库。

第 12 章“数据操作”介绍如何使用 data.table 和 dplyr 处理关系型数据，以及使用 rlist 处理非关系型数据的技术。

第 13 章“高性能计算”讨论 R 的性能问题和提高计算性能的常用方法。

第 14 章“网页爬虫”讨论网页、CSS 和 XPath 选择器的基本结构，以及如何使用 rvest 包从简单的网页中抓取数据。

第 15 章“效率提升”演示了如何利用 R Markdown 和 shiny app 结合交互式图形来提高数据分析报告和展示的效率。

阅读本书还需要什么

运行书中的示例代码，需要安装 R 3.3.0 或更高版本，推荐使用 RStudio 开发环境。

对于第 11 章，运行示例代码需要一个可用的 MongoDB 服务器和一个 Redis 实例。

对于第 13 章，在 Windows 操作系统下需要安装 Rtools 3.3 来创建 Rcpp 代码，在 Linux 或 macOS 操作系统下，则需要 gcc 工具链。

本书的目标读者

本书主要面向从事数据相关项目并希望提高工作效率的读者，但可能不适合对编程语言和相关工具一无所知的人阅读。

本书也适用于想要系统地学习 R 编程语言、相关技术和推荐的扩展包及其实际应用的专业数据分析师。

书中的一些章节对于初学者来说比较高深，尽管阅读这些章节并不要求你是计算机专家或者专业的数据分析师，但我认为对基础编程概念有一些了解并具有数据处理的基本经验，会有助于对本书内容的理解。

约定

在本书中，你会发现一些用于区分不同信息的文本样式。以下是这些样式的示例及其含义的解释。

文本中包含的代码、数据库表名、文件夹名、文件名、文件扩展名、路径、虚拟的网址（URL）、用户输入和推特名称用代码体显示，如下所示：“`apply( )`函数也支持数组输入和矩阵输出。”

内联代码（变量和函数名）和代码块的样式设置如下：

```
x <- c(1, 2, 3)
class(x)
## [1] "numeric"
typeof(x)
## [1] "double"
str(x)
##  num [1:3] 1 2 3
```

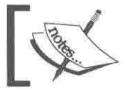
当某个函数名或变量名被选中时，相同的函数名或变量名就会高亮显示：

```
x <- rnorm(100)
y <- 2 * x + rnorm(100) * 0.5
```

```
m <- lm(y ~ x)
coef(m)
```

第一次出现的**术语**和**重要词汇**会以粗体显示。

小技巧：
警告或重要的提示出现在这样的框中。



提示：
提示和技巧以这种形式出现。

读者反馈

我们欢迎并重视读者的反馈。请让我们知道你对本书的看法——喜欢或不喜欢什么。读者的反馈对我们很重要，它有助于我们推出对读者更有价值的产品。

若想反馈给我们，可直接发送邮件到 feedback@packtpub.com，并在邮件主题中注明书名。

如果你精通某一主题，并有兴趣编写或撰稿，请前往作者指南网页：www.packtpub.com/authors。

用户支持

作为 Packt 图书的拥有者，我们准备了很多服务来最大化你的消费权益。

下载示例代码

你可以在 http://www.packtpub.com 登录账户并下载本书对应的示例代码。如果你是在其他地方购买的本书，可以访问 http://www.packtpub.com/support 进行注册，我们会将代码文件直接发到你的邮箱。

你可以按照以下步骤下载代码文件。

1. 使用电子邮箱和密码在网站登录或注册。
2. 将鼠标指针悬停在顶部的 **SUPPORT** 选项卡上。

3. 单击 **Code Downloads & Errata** 按钮。

4. 在 **Search** 对话框中输入书名。

5. 选择想要下载的代码文件对应的书名。

6. 从下拉菜单中选择购买本书的途径。

7. 单击 **Code Download** 按钮。

也可以访问 Packt 出版社网站上的网页，单击 **Code Files** 按钮下载代码文件，还可以在搜索框中输入书名来访问该网页。请注意，需要登录你的 Packt 账户才能执行以上步骤。

文件下载好之后，请务必使用以下软件的最新版本进行解压：

- WinRAR/7-Zip for Windows
- Zipeg/iZip/UnRarX for Mac
- 7-Zip/PeaZip for Linux

本书的代码同时托管在 GitHub 上：https://github.com/PacktPublishing/learningrprogramming。我们的 GitHub 主页上还有更多图书和视频资源：https://github.com/PacktPublishing/。快来浏览一下吧！

勘误

尽管我们已经尽全力确保本书内容的准确性，但是错误还是在所难免。如果你在我们的某本书中发现了错误（可能是文字或代码错误），请告知，我们将不胜感激。这样，你可以帮助其他读者避免困惑，并帮助我们改进本书的后续版本。如果您发现任何错误，请打开 http://www.packtpub.com/submit-errata 网页，选择相应图书，单击勘误表的提交表单链接，输入勘误表的详细信息。一旦您的勘误经过验证，我们将接受您提交的内容，并将勘误上传到网站，或追加到该题目现有勘误表的下面。

若想查看之前提交的勘误表，请转到 https://www.packtpub.com/books/content/support，并在搜索栏中输入书名，所需信息将显示在勘误部分的下方。

盗版

网络上的盗版问题是所有媒体一直面对的问题。在 Packt，我们非常严肃认真地保护版

权和许可。如果你在网络上发现有关作品的任何形式的盗版版本，请立即向我们提供网址或者网站名称，以便我们及时采取补救措施。

请通过 copyright@packtpub.com 与我们联系，并附上可疑盗版资料的链接。

衷心地感谢你帮助我们保护作者的权益以及我们为你带来宝贵知识的能力。

问题

如果您对本书的任何方面有任何问题，可以通过 questions@packtpub.com 与我们联系，我们会尽全力为你解决。

目录