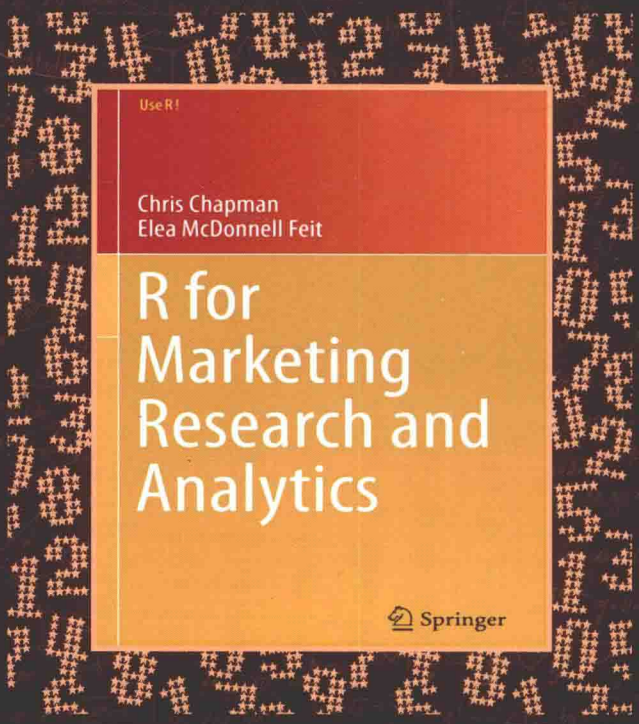
HZ BOOKS 华章IT
Springer
数据科学与工程技术丛书

R语言市场研究分析

克里斯·查普曼（Chris Chapman）
谷歌公司
[美] 著
埃里亚·麦克唐奈·费特（Elea McDonnell Feit）
德雷克塞尔大学
林荟 译



R FOR MARKETING RESEARCH AND ANALYTICS

机械工业出版社
China Machine Press

科学与工程技术丛书

R FOR MARKETING RESEARCH AND ANALYTICS

R语言市场研究分析

[美] 克里斯·查普曼（Chris Chapman） 谷歌公司 著

埃里亚·麦克唐奈·费特（Elea McDonnell Feit） 德雷克塞尔大学

林荟 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

R 语言市场研究分析 /（美）克里斯·查普曼（Chris Chapman），（美）埃里亚·麦克唐奈·费特（Elea McDonnell Feit）著；林荟译 . —北京：机械工业出版社，2016.10
（数据科学与工程技术丛书）
书名原文：R for Marketing Research and Analytics

ISBN 978-7-111-54990-1

I. R… II. ① 克… ② 埃… ③ 林… III. 程序语言－程序设计－应用－市场分析
IV. F713.52-39

中国版本图书馆 CIP 数据核字（2016）第 230643 号

本书版权登记号：图字：01-2016-1875

R 语言市场研究分析

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：谢晓芳　　责任校对：董纪丽
印　　刷：中国电影出版社印刷厂　　版　　次：2016 年 10 月第 1 版第 1 次印刷
开　　本：185mm × 260mm　1/16　　印　　张：21（含 2 面彩插）
书　　号：ISBN 978-7-111-54990-1　　定　　价：89.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

华章IT
HZBOOKS | Information Technology

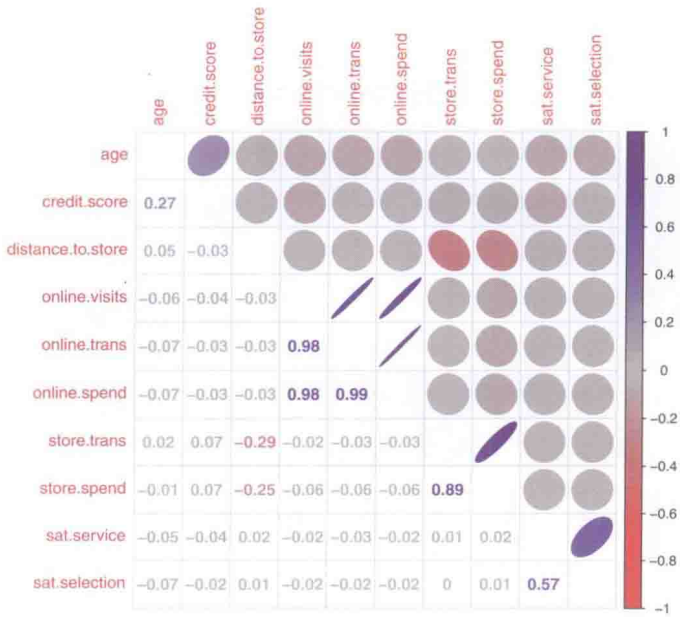


图 4.9　通过 corrplot 包中的 corrplot.mixed() 函数得到相关矩阵图形。这是很好地对相关性进行可视化的方法。相关性接近 0 的用灰色圆点表示（我们自己定义的颜色），相关性大于 0 的用椭圆表示，相关性越大椭圆越窄，蓝色代表正相关，红色代表负相关

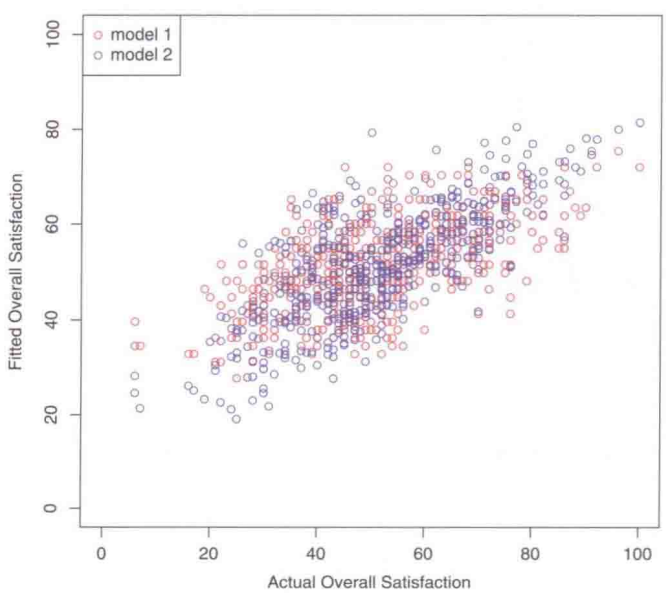


图 7.7　m1 和 m2 模型拟合值和真实值比较图

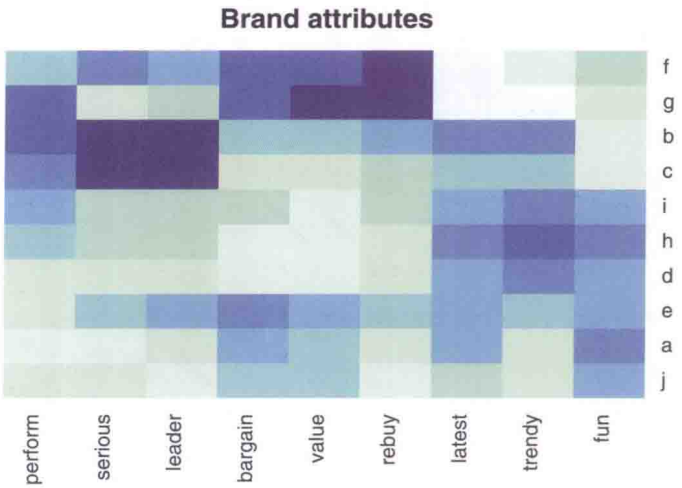


图 8.2　各品牌对应感知描述项的评分均值。品牌 f 和 g 相似——在“再次购买”（rebuy）和“价值”（value）上评分高，但是在“新潮”（latest）和有趣（fun）上评分低。其他类似的品牌组是 b/c、i/h/d 和 a/j

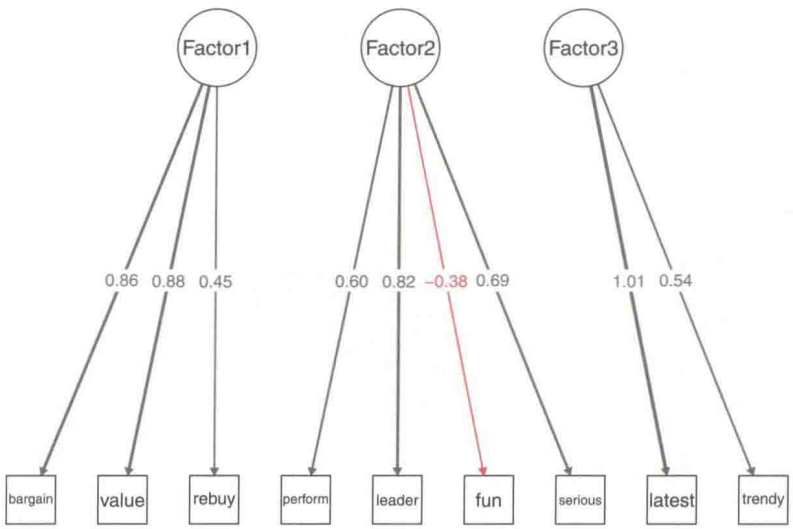


图 8.9　因子分析结果的路径图，我们可以清楚地看到 3 个因子，以及它们的载荷（载荷绝对值小于 0.3 的没有显示）。该图是通过 semPlot 包中的 semPaths() 函数生成的

中文版序

我们非常高兴看到本书中文版面世。本书英文版的主要目的是深入介绍 R 在市场研究中的应用。中文版是本书英文版之外的第一个版本，中文版和英文版上市的时间只相隔 1 年，所以书中的知识都是最新的。感谢译者的翻译！

R 的应用在不断飞速发展。如本书第 1 章提到的，在写书时已经有 6000 多个 R 包可供使用。时至今日，R 包的数目已经增长到 8000 个——平均每天增加 3 个包以上。与此同时 R 在市场营销方面的学术论文、会议报告以及工业界的使用也在不断增长。R 的前景非常广阔，现在是时候学习这门语言了。

我们特别感谢本书的译者林荟和所有为本书中文版问世做出努力的人！我们还要感谢机械工业出版社的工作人员。如果本书能对你们有所帮助，便是对所有这些努力最好的回报。

最重要的是，希望你们能像我们一样喜欢 R 并且将其用于自己的工作。学习 R 这样的语言并不容易，我们希望本书能使你的学习轻松些，帮助中国的市场营销人员掌握 R。

——Chris Chapman，Elea McDonnell Feit

2016 年 2 月

译 者 序

本书适合寻找R入门书籍的读者以及那些想将数据科学应用到市场研究分析中的读者。书中说明了如何载入数据，通过可视化技术探索数据，用统计模型分析数据，并且对模型结果给出了商业解释。书中涉及了基础的分析技能、可视化和一些高级别的分析，所以对普通读者和专业读者而言，本书都是绝佳的指南。2013年我从研究生物统计转而进入杜邦公司从事专门的市场分析，从一个商业数据分析从业者的角度看，这本书确实给我非常大的帮助。

本书有以下特点：

- 这是第一本成功介绍将一些现代统计分析技术应用到市场研究分析的书。它不同于之前那些介绍市场分析中用到的传统多元技术的书。虽然近年来有几本和模型应用相关的非常优秀的书籍问世，如James、Witten、Hastie和Tibshirani的《统计学习导论》，Kuhn和Johnson的《应用预测建模》，但是这些书并不是专门针对市场研究的。本书的针对性是其一大优点。
- 作者并非从学术的角度解决一些虚假的市场问题。书中的例子都是现实市场分析中经常遇到的问题。作者使用的是模拟数据集，乍一看让人感觉本书可能会脱离实际，毕竟用一些伪造的数据可以很容易给出模型效果很好的假象。其实不是这样的。由于两位作者在此之前都有着数十年的商业分析从业经验，因此书中数据集的抽取都非常巧妙，能够很好地反映作者在实践中遇到的真实问题。
- 除了传统的多元模型之外，书中还介绍了近年来逐渐流行的贝叶斯方法。虽然贝叶斯方法在市场分析中当前还不是主流，但我相信该方法的应用会越来越广泛。书中还专门介绍了相对较新的分析技术，如随机森林和朴素贝叶斯。
- 作者还在适当的地方对模型的应用进行了延伸。比如在讲到因子分析时，作者讨论了如何使用因子分析结果来绘制消费者“认知图”，这在很多讨论因子分析的文献中极少看到。这也充分反映了作者丰富的实践经验，以及本书以具体实践为导向的特点。
- 本书覆盖的方法比较全面。基本涉及了市场分析中从初级数据探索到高级数据建模过程中可能用到的各种技术。
- 本书没有很多数学公式，深入浅出。这使得本书适合于那些没有很强的数学基础但又想学习一些高级分析方法的市场研究人员。
- 对于R新手来说，本书是一本很好的入门指南。和单纯的R指南不同，本书提供了一个应用的语境，使得读者能够在应用中学习，极大地增强了学习效果。本书

不仅讲到了基本的 R 数据操作，还介绍了一些常用的有效可视化方法。

书中没有过多介绍现在流行的有效机器学习模型，关于这点，之前讲到的两本书《统计学习导论》和《应用预测建模》是极好的补充，如果能系统学习这 3 本书，就具备成为一个数据科学家的硬性技术条件了。

机械工业出版社的王春华编辑对本书的翻译工作给予了支持和帮助。在此对所有为本书中文版问世做出努力的人表示感谢！限于译者水平，书中难免有错误和不妥之处，恳请读者批评指正。

——林荟

前　言

我们将会帮助你在市场研究和分析中使用 R。

R 是市场分析师的绝佳选择。它拟合统计模型的能力无与伦比，对于大型和小型数据集，它可扩展，能以不同形式分析来自不同系统的数据。R 生态系统包括大量现存以及正在兴起的统计方法和可视化技术。但是 R 在市场营销中的应用程度不如其在统计、计量经济、心理和生物信息领域。希望通过大家的努力能改变现状！

本书是为两类人设计的：想要学习 R 的市场研究从业人员和分析师，想要了解如何将 R 应用于市场营销的其他领域的学生和研究人员。

阅读本书需要哪些预备知识？很简单，对 R 在市场营销中的应用感兴趣，对基础统计模型（如线性回归）有概念性的了解，并且愿意亲自动手实践学习。本书对已有一定编程经验并希望学习 R 的分析师特别有帮助。我们会在第 1 章中介绍另外一些使用 R 的原因（以及一些可能不需要使用 R 的原因）。

动手实践部分非常重要。我们将在前 7 章循序渐进地介绍（相关知识）并且让读者自行实践书中的案例（代码）；本书不是食谱类型的参考书。我们会在第一部分花一些时间（尽量少）介绍 R 基础知识，然后在第二部分介绍现实中的市场营销问题以及如何应用 R。第三部分包含一些高阶市场营销问题。每章都展示了 R 的分析能力。希望读者在每章中都能学到新鲜有趣的知识。

本书有如下特点：

- 本书围绕市场营销组织内容。不是给出泛泛的示例，而是结合介绍的方法给出市场营销案例。
- 我们假定读者有基础统计知识和少量的数学知识。本书是为分析实践者设计的，因此并不会过多地介绍方程和统计模型背后的数学细节（但我们会给出相应参考书目）。
- 这是一本讲解统计概念和 R 代码的教科书。它旨在让读者明白我们在干什么以及学会如何避免在应用统计和 R 时的问题。对比市面上其他参考书和“食谱类”指南，我们的目标在于让本书具有可读性并且能够满足不同读者的需求。
- 应用章节阐明了渐进的建模过程。我们并没有提供“答案”，而是展示一个分析师在现实工作中可能按何种方式逐步展开分析。其中比较了不同模型的统计可靠性和实用性。
- 可视化内容是核心分析的一部分。我们并没有将可视化当作独立的话题，而是相信它是数据探索和建模的一个部分。
- 你从中学到的不仅仅是 R。除了核心模型外，本书还涵盖了一些或许对有经验的

分析师来说也很陌生的有用模型，如结构方程模型、交易分析。

- 本书同时介绍了传统方法和贝叶斯方法。核心模型和传统（频率学派）模型一起介绍。但在后面的章节中会介绍线性模型和联合分析中的贝叶斯方法。
- 大部分分析用模拟数据实践 R，并额外提供了关于市场数据结构的信息。根据个人意愿，可以改变模拟数据，看其对统计模型的影响。
- 在合适的时候我们会给出选学的编程内容或模型知识，读者可根据自身情况选择阅读或跳过。这些小节用 * 标注。

本书没有包括什么？首先，本书介绍 R 在市场营销中的应用但并不讲述市场营销方面的研究。我们会讨论很多市场营销话题但会忽略 R 中那些重复用到相同分析方法的话题。如前所述，我们从概念上介绍统计模型且并不关注数学细节。由于篇幅原因，本书省略了一些复杂的话题，包括顾客终身价值模型和计量经济时间序列模型。总之，本书全面展示了市场营销研究示例和分析方法。如果掌握了本书，你将能在许多市场营销领域应用 R。

为什么我们可以教这些知识？从 1997 年开始，我们使用 R 及其前身 S 语言近 30 年，这是我们主要的分析平台。我们用 R 做各种市场分析，从简单的数据总结到复杂的分析（需要自己编写成千上万行的代码）和新模型。

我们也有丰富的 R 教学经验。本书源自于笔者在美国营销协会（AMA）、埃默里大学市场营销学院和高级研究方法论坛（ART Forum）几年来的课程讲义。我们也在 Sawtooth 软件会议上和沃顿商学院对学生和业界人士进行 R 教学。感谢许多学生的反馈意见，我们相信他们的经验会对你们有益。

关于下载数据

本书对应的 .R 代码文件中的下载数据使用的是短链接地址，如“http://goo.gl/UDv12g”。在一些国家和地区可能无法访问该链接，会出现这样的错误提示：“Error in file(file, "rt"): cannot open the connection”。如果出现这种情况，请尝试对应的完整链接地址，如“http://r-marketing.r-forge.r-project.org/data/rintro-chapter2.csv”。下面是短链接和对应完整链接的表格。

章　号	完整链接地址	短链接地址
2	http://r-marketing.r-forge.r-project.org/data/rintro-chapter2.csv	http://goo.gl/UDv12g
3	http://r-marketing.r-forge.r-project.org/data/rintro-chapter3.csv	http://goo.gl/QPDdMl
4,9	http://r-marketing.r-forge.r-project.org/data/rintro-chapter4.csv	http://goo.gl/PmPkaG
5、6、11、12	http://r-marketing.r-forge.r-project.org/data/rintro-chapter5.csv	http://goo.gl/qw303p
7	http://r-marketing.r-forge.r-project.org/data/rintro-chapter7.csv	http://goo.gl/HKnl74
8	http://r-marketing.r-forge.r-project.org/data/rintro-chapter8.csv	http://goo.gl/IQl8nc
9	http://r-marketing.r-forge.r-project.org/data/rintro-chapter9.csv	http://goo.gl/J8MH6A
9	http://r-marketing.r-forge.r-project.org/data/rintro-chapter9conjoint.csv	http://goo.gl/G8knGV
10	http://r-marketing.r-forge.r-project.org/data/rintro-chapter10pies.csv	http://goo.gl/yT0XwJ
10	http://r-marketing.r-forge.r-project.org/data/rintro-chapter10sat.csv	http://goo.gl/MhghRq
12	http://fimi.ua.ac.be/data/retail.dat	http://goo.gl/O495RV
12	http://r-marketing.r-forge.r-project.org/data/retail.dat	http://goo.gl/FfjDAO
13	http://r-marketing.r-forge.r-project.org/data/rintro-chapter13conjoint.csv	http://goo.gl/5xQObB

致谢

我们特别感谢为本书的问世做出贡献的人。首先是这些年来我们教过的所有学生，他们提供了有价值的反馈。我们希望他们的经验对你们有益。

在市场营销学术领域和实践者社区，Ken Deal、Fred Feinberg、Shane Jensen、Jake Lee、Dave Lyon 和 Bruce McCullough 提供了宝贵意见。

Chris 在谷歌科研社区的同事对本书的一些部分提供了许多建议。我们感谢如下人的鼓励和建议：Mario Callegaro、Marianna Dizik、Rohan Gifford、Tim Hesterberg、Shankar Kumar、Norman Lemke、Paul Litvak、Katrina Panovich、Marta Rey-Babarro、Kerry Rodden、Dan Russell、Angela Schörgendorfer、Steven Scott、Bob Silverstein、Gill Ward、John Webb 和 Yori Zwols。

Springer 的员工和编辑帮助我们顺利展开工作，其中尤其要感谢 Hannah Bracken、Jon Gurstelle 和“Use R!”系列丛书编辑。

本书的大部分是在公共图书馆和大学图书馆完成的。我们感谢其为我们提供场所以及大量的文献资源。本书部分是在晴朗的日子里于新奥尔良公共图书馆、纽约公共图书馆、纽约神学院的小克里斯托弗·凯勒图书馆、加州大学圣地亚哥分校的吉赛尔图书馆，华盛顿大学苏塞罗和艾伦图书馆、森尼维尔公共图书馆完成的，尤其是东京都中心图书馆，我们在那里写下了第一句话、第一行代码、全书大纲以及后续许多内容。

家人对我们在周末和夜晚编写本书给予了支持，他们还忍受了对门外汉来说关于 R 的过多讨论。谢谢 Cristi、Maddie、Jeff 和 Zoe。

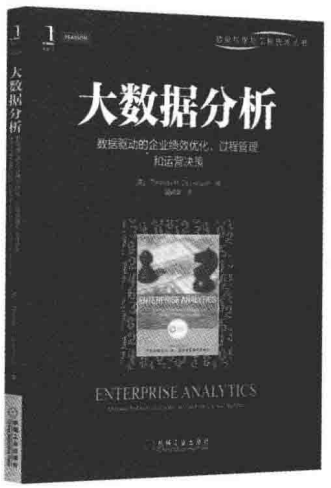
最重要的是，我们感谢各位读者。很高兴你们决定研究 R，且希望你们的努力有所收获。让我们开始吧！

——Chris Chapman、Elea McDonnell Feit

推荐阅读

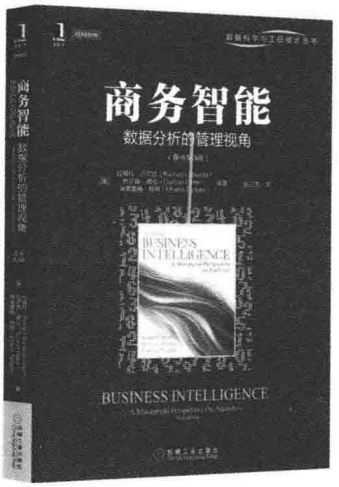
机器学习与R语言

作者：Brett Lantz ISBN：978-7-111-49157-6 定价：69.00元



大数据分析：数据驱动的企业绩效优化、过程管理和运营决策

作者：Thomas H. Davenport ISBN：978-7-111-49184-2 定价：59.00元



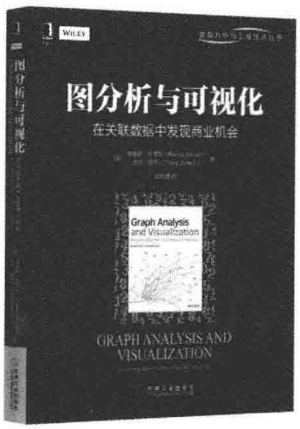
商务智能：数据分析的管理视角（原书第3版）

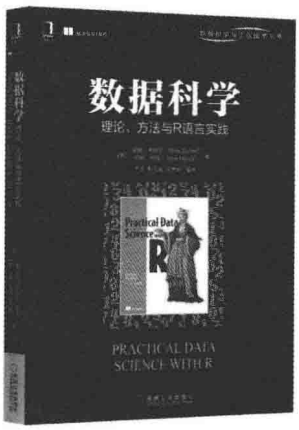
作者：Ramesh Sharda 等 ISBN：978-7-111-49439-3 定价：69.00元

统计学习导论——基于R应用

作者：Gareth James 等 ISBN：978-7-111-49771-4 定价：79.00元

推荐阅读





图分析与可视化：在关联数据中发现商业机会

作者：理查德·布莱斯 ISBN：978-7-111-52692-6 定价：119.00元

本书将图与网络理论从实验室带到真实的世界中，深入探讨如何应用图和网络分析技术发现新业务和商业机会，并介绍了各种实用的方法和工具。作者Richard Brath和David Jonker运用高级专业知识，从真正的分析人员视角出发，通过体育、金融、营销、安全和社交媒体等领域的引人入胜的真实案例，全面讲解创建强大的可视化的过程。

基于R语言的自动数据收集：网络抓取和文本挖掘实用指南

作者：西蒙·蒙策尔特 等 ISBN：978-7-111-52750-3 定价：99.00元

本书由资深社会科学家撰写，从社会科学研究角度系统且深入阐释利用R语言进行自动化数据抓取和分析的工具、方法、原则和最佳实践。作者深入剖析自动化数据抓取和分析各个层面的问题，从网络和数据技术到网络抓取和文本挖掘的实用工具箱，重点阐释利用R语言进行自动化数据抓取和分析，能为社会科学研究者与开发人员设计、开发、维护和优化自动化数据抓取和分析提供有效指导。

数据科学：理论、方法与R语言实践

作者：尼娜·朱梅尔 等 ISBN：978-7-111-52926-2 定价：69.00元

本书讨论如何应用R程序设计语言和有用的统计技术处理日常的业务情况，并通过市场营销、商务智能和决策支持领域的示例，阐述了如何设计实验（比如A/B检验）、如何建立预测模型以及如何向不同层次的受众展示结果。

目　录