

TURING

图灵程序设计丛书

길벗

R语言 与 数据分析 实战

[韩] 徐珉久 著
武传海 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语言与 数据分析实战

[韩] 徐珉久 著
武传海 译

人民邮电出版社
北京

图书在版编目(CIP)数据

R语言与数据分析实战/(韩)徐珉久著;武传海译

北京:人民邮电出版社,2017.1

(图灵程序设计丛书)

ISBN 978-7-115-44246-8

I. ①R… II. ①徐… ②武… 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298230号

Original Title: R 을 이용한 데이터 처리 & 분석 실무

Practical Data Processing and Analysis Using R by Minkoo Seo

Copyright © 2014 Minkoo Seo

All rights reserved.

Original Korean edition published by Gilbut Publishing Co., Ltd. Seoul, Korea

Simplified Chinese Translation Copyright ©2017 by Posts & Telecom Press

This Simplified Chinese edition arranged with Gilbut Publishing Co., Ltd.

Through Eric Yang Agency.

本书中文简体字版由 Gilbut 授权人民邮电出版社独家出版。未经出版者书面许可,不得以任何方式复制或抄袭本书内容。

版权所有,侵权必究。

内 容 提 要

本书以R语言的“编程属性”为中心,内容涵盖R语言基础理论到实际数据分析,通过分析模型和算法等更实用的示例,讲解了数据可视化、统计分析、数据挖掘、机器学习等实际业务中常用的实操技巧,以及代码生成方法。书中还收录了作者的实战经验和学习体会,可以解决数据分析过程中出现的各种问题。对R语言有一定了解但在实际运用中感到困惑的读者,可以在书中找到多种解题方法,并能够迅速应用于一线业务。

-
- ◆ 著 [韩] 徐珉久
 - 译 武传海
 - 责任编辑 陈曦
 - 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>
 - 北京昌平百善印刷厂印刷
 - ◆ 开本: 800×1000 1/16
 - 印张: 29.75
 - 字数: 717千字 2017年1月第1版
 - 印数: 1-3 000册 2017年1月北京第1次印刷
 - 著作权合同登记号 图字: 01-2016-3941号

定价: 89.00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

大数据应用、大数据分析都是现在十分热门的话题，这些内容涉及各个领域，甚至在广告文案中也可以看到。但对于如何应用大数据，人们的概念依然相当模糊。在许多人眼里，只要引入大数据技术，一切问题就能迎刃而解。然而，如果你是一名软件工程师，就必须对数据分析、假设、模型设置、检验、应用等所有过程具备正确理解，并能够灵活运用。

大数据应用与分析中，最重要的是如何看待大数据、如何树立假设检验，以及如何应用模型。要想拥有这样的能力，除了掌握基本的编程语言之外，还要通过对多种数据的分析实战不断积累经验，加深对相应知识的理解、认识，最终才能找到答案。R 不仅是一种编程语言，还提供了数据分析环境，为我们学习数据分析提供了强大的支持与工具。

目前，韩国国内已经有许多关于 R 的翻译图书，也有很多学习课程。但是，本书特色在于作者是谷歌的软件工程师，书中内容是作者多年学习与应用 R 的经验总结。本书从软件工程师角度讲解 R 语言基本语法、函数等内容，如同作者本人坐在身旁专门为你讲解一样。书中还有“提示”格式的内容，用于介绍实际编程中需要了解的知识。

本书后半部分介绍了应用于实际业务的统计分析、数据挖掘以及与机器学习相关的分析模型与算法，重点讲解基本的模型与算法，并配以精选示例，使各位能够看到实际运行的结果。其他图书讲解分析模型与算法时，往往容易陷入理论的泥潭，但本书站在软件工程师角度进行讲解，简单明了，具有很强的实用性。

与其他外版翻译书和参考书不同，本书不仅对 R 语言进行了介绍，还讲解了分析模型与算法的理论知识，并结合精心挑选的示例，使读者深入理解基本知识的同时，学习并掌握具体的应用方法与技巧。如果你是软件工程师——即使不从事大数据相关工作——我强烈建议阅读本书，相信会有很多收获。

——谷歌韩国技术部经理 朴勇灿

KDnuggets (<http://www.kdnuggets.com>) 是数据挖掘专业的知名网站，每年都会进行题为“在过去的 12 个月中，哪些数据分析、数据挖掘、数据科学软件与工具的实际应用最广泛？”的问卷调查。结果表明，多年来，有一种软件与 RapidMiner 一直占据着榜单首位，这款软件就是本书讲解的主题——R 语言。实际项目中广泛应用 R 语言的原因在于，其使用方便、功能丰富、支持多种环境、容易扩展，并且是免费的。

本书内容丰富多样，还给出了多种示例，涵盖从 R 的基本知识到使用 R 进行数据处理与挖掘的各种方法，相信能够为各位提供大量帮助。如果你是刚刚入门统计与数据挖掘的学生，本书将帮助你所学知识快速应用于实战，更好地完成研究课题；如果你是一名数据分析从业者，那本书将是一本不可多得的参考书，能够帮助你深化理解与认识，进一步提高数据分析水平。

——谷歌韩国软件工程师 姜在浩

数据分析的起点——R 编程！

Web、移动应用、社交网络、检索、大数据是贯穿当今时代的关键词，将其串联在一起的另一个关键词就是基于数据分析与数据的决策。将分析与决策应用于网页的典型案例是美国总统奥巴马募集 6000 万美元选举资金的事情^①，工作人员制作了两种设计风格的网页，并分析（称为 A/B 测试）使用哪种设计能够吸引更多选民。这种分析（A/B 测试）不仅可以用于 App 营销，也可以用于开发 App。另一个数据分析的例子是分析社交网络图的结构，或者更改社交网络网站的页面组成以观察用户反应。在网页搜索中也进行过大量实验^②。最近，应用大数据进行数据分析备受青睐，分析对象甚至包含大数据系统本身如何快速运行。数据分析的下一步是预测分析（Predictive Analytics），它是决策的根基。

随着分析、预测、决策等话题的火爆，相信 R 语言接下来也会受到热捧。为什么这样说呢？首先，R 语言是一种专门语言，重点在于数据分析、统计分析、机器学习、数据可视化。使用 R 提供的多种包能够轻松解决分析与预测问题。同时，R 也是一种编程语言，容易扩展，适用于解决多种问题。其次，R 是一种开源软件，任何个人、企业、学校、机关都可以免费使用，无需背负沉重的经济负担。第三，R 背后有强大的社区，社区中开发的多种分析包都是免费提供的。最后，R 拥有丰富的帮助文档，相关图书的出版开始猛增。现在，几乎任何一本统计分析图书都使用 R 语言编写示例代码。毕竟，亲自动手编写并运行代码与只使用笔纸学习方法有着很大不同。

人们对 R 语言学习热情的高涨促使了本书的诞生与出版。本书是我在多年学习笔记的基础上编写而成的，这些笔记是我为了使用 R 进行机器学习而整理的，在我的个人博客（<http://mkseo.pe.kr/stats>）上可以看到。整理并挑选示例时，我参考了多种图书与资料，每当遇到问题，我都会使用谷歌搜索引擎和 StackOverflow（<http://stackoverflow.com>）寻找答案。随着资料的增多，逐渐形成了图书的形态，最后促使本书产生。书中整理了大量 R 初学者经常遇到的问题及答案。通过阅读本书，读者可以轻松学习 R 语言并掌握应用方法，不必再经历我当时学习的痛苦了。

① 此事相关报道请参考网址 <http://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/>

② *Overlapping Experiment Infrastructure: More, Better, Faster Experimentation*, Proceedings 16th Conference on Knowledge Discovery and Data Mining, 2010, ACM. <http://research.google.com/pubs/pub36500.html>

本书韩文版的顺利出版得益于 Gilbut 出版社韩东勋课长和许亨哲组长的帮助，申景根先生帮我确定了全书的行文风格与方向。此外，还要感谢 Gilbut 出版社的相关工作人员，他们为本书的出版付出了巨大努力。

感谢我的妻子。对于每个周末都要坐在电脑前的丈夫，她心里不免会有些怨言，但从未流露出来，也从未说出口，只是一直陪在我身边默默等待。谢谢你的鼓励！

最后，感谢购买本书的读者朋友们。写作本书时，我已竭尽所能，倾注大量心血，但由于自身的不足，难免会出现各类问题。如果大家在阅读过程中发现任何问题，请给我发送邮件（minkoo.seo@gmail.com），我将尽自己所能为你解答。谢谢！

徐珉久

2014年10月

本书面向的是熟悉软件开发的读者，以及关注统计、机器学习技术的朋友，希望能够帮助他们快速学习 R 语言。因此，书中讲解了我认为最重要的 R 函数与相关包，使各位能够在短时间内了解并使用。我尽量使书中代码可以独立运行，并且添加了许多代码的运行结果截图，以帮助各位轻松理解所学内容，而不必每次都要亲自运行。

本书不讲解编程基础，也不介绍相关理论，所以不会详细叙述变量、变量作用域、循环语句含义等编程语言中的基本概念。相关内容的讲解虽然涉及统计学、机器学习等基本概念，但不会对理论部分进行深入介绍。其实，一本书不可能涵盖所有内容，书中列出了相关参考资料，各位可以进一步学习。此外，由于本书旨在使各位掌握 R 语言的使用方法，所以不会涉及 R 语言的内部结构和使用 R 语言创建库等内容。

- 第 1 章介绍 R 的安装、启动、开发环境等。
- 第 2 章与第 3 章讲解数据存储中使用的多种数据类型、条件语句、循环语句等基本编程知识，以帮助各位了解使用 R 进行编程的基本方法。

数据分析中，先要读入数据，然后计算总值或平均值，再创建模型进行评估。整个过程看似简单，实则不然。创建模型前需要耗费相当长的时间对数据进行预处理，因为数据分析中得到的数据资料往往都是未经整理的原始数据。比如，即使是同样的体检数据（如 175 cm），也会随着输入者的不同而有 1.75 m、175 cm、175.0 cm、175 等形式。使用这些数据前，需要对其进行统一处理，以便后续使用。

我们有时需要从数据已有属性推导新属性，比如通过身高、体重数据计算 BMI 指数，并将其作为新属性存储到数据。

像这样，数据经过一系列处理后，才能分组计算数据的总值、平均值等基本统计量，才能帮助我们更好地了解数据特征。比如，按照班级、科目分别计算学生的平均成绩。

- 第 4 章与第 5 章介绍 R 中的基本函数与应用包，它们贯穿数据分析全过程。随着函数的增多，学习开始变得有难度，需要耗费大量时间，但这也恰恰表明其重要性。
- 第 4 章与第 5 章讲解的数据主要为数值或表格形式，而第 6 章讲解数据可视化。对数据倾向性进行判断或比较时，采用柱形图、折线图、密度图等形式，这比采用数值或表格形式更易于理解。
- 第 7~10 章主要讲解统计分析、线性回归、分类算法（Classification Algorithm）应用方法等内容，它们对整理好的数据进行分析建模。主要内容包括数据平均值是否不同、比值间

是否有差别、向数据应用 $Y=aX+b$ 等简单关系式计算 a 与 b 的值、使用机器学习算法预测数据分类。

- 第 11 章讲解建模示例，对“泰坦尼克”号生还者数据应用分类算法，创建预测乘客生还可能性的模型。这一章的示例代码包含本书全部内容，有一定难度，但是一个完整的系统。各位以后创建机器学习模型时，可以参考第 11 章的示例代码，相信会得到很好的启发。

第 1 章 ▶ 搭建 R 编程环境 1

- 1.1 为什么是 R 2
- 1.2 安装 R 2
 - 1.2.1 在 Windows 操作系统中安装 R 4
 - 1.2.2 在 Linux 系统下安装 R 6
 - 1.2.3 在 Mac OS X 中安装 R 12
- 1.3 启动 R 15
- 1.4 查看帮助 16
- 1.5 R 集成开发环境 19
- 1.6 批处理 21
- 1.7 使用包 22
- 参考资料 24

第 2 章 ▶ 数据类型 25

- 2.1 变量 26
 - 2.1.1 变量命名规则 26
 - 2.1.2 变量赋值 26
- 2.2 调用函数时指定参数 27
- 2.3 标量 28
 - 2.3.1 数值 28
 - 2.3.2 NA 29
 - 2.3.3 NULL 29
 - 2.3.4 字符串 30
 - 2.3.5 逻辑值 31
 - 2.3.6 因子 32
- 2.4 向量 34

- 2.4.1 创建向量 35
- 2.4.2 访问向量中的数据 36
- 2.4.3 向量运算 39
- 2.4.4 连续数字组成的向量 41
- 2.4.5 保存重复值的向量 42
- 2.5 列表 43
 - 2.5.1 创建列表 43
 - 2.5.2 访问列表中的数据 44
- 2.6 矩阵 45
 - 2.6.1 创建矩阵 45
 - 2.6.2 访问矩阵中的数据 48
 - 2.6.3 矩阵运算 49
- 2.7 数组 52
 - 2.7.1 创建数组 53
 - 2.7.2 访问数组数据 54
- 2.8 数据框 54
 - 2.8.1 创建数据框 55
 - 2.8.2 访问数据框 57
 - 2.8.3 实用工具函数 59
- 2.9 类型判别 61
- 2.10 类型转换 62
- 参考资料 64

第3章 ▶ R 语言编程 65

- 3.1 R 的特征 66
- 3.2 流程控制 (条件语句与循环语句) 66
 - 3.2.1 if 语句 66
 - 3.2.2 循环语句 67
- 3.3 运算 69
 - 3.3.1 数值运算 70
 - 3.3.2 向量运算 70
 - 3.3.3 NA 处理 72
- 3.4 定义函数 74
 - 3.4.1 基本定义 74
 - 3.4.2 可变长参数 75

- 3.4.3 嵌套函数 76
- 3.5 作用域 77
- 3.6 值传递 80
- 3.7 对象的不变性 81
- 3.8 模块模式 83
 - 3.8.1 队列 84
 - 3.8.2 编写队列模块 85
- |参考资料| 86

第4章 ▶ 数据操作 I：基于向量的处理与外部数据处理 87

- 4.1 鸢尾花数据集 88
- 4.2 读写文件 90
 - 4.2.1 读写 CSV 文件 90
 - 4.2.2 读写对象文件 93
- 4.3 合并数据框的行与列 94
- 4.4 apply 系列函数 96
 - 4.4.1 apply() 97
 - 4.4.2 lapply() 函数 99
 - 4.4.3 sapply() 102
 - 4.4.4 tapply() 104
 - 4.4.5 mapply() 106
- 4.5 数据分组并调用函数 107
 - 4.5.1 summaryBy() 108
 - 4.5.2 orderBy() 110
 - 4.5.3 sampleBy() 112
- 4.6 数据拆分与合并 114
 - 4.6.1 split() 115
 - 4.6.2 subset() 116
 - 4.6.3 数据合并 117
- 4.7 数据排序 119
 - 4.7.1 sort() 119
 - 4.7.2 order() 120
- 4.8 访问数据框中的列 121
 - 4.8.1 with() 121
 - 4.8.2 within() 122

- 4.8.3 attach() 与 detach() 124
- 4.9 查找符合条件的数据索引 126
- 4.10 分组运算 127
- 4.11 更易处理的数据表现形式 128
- 4.12 与 MySQL 联动 131
 - 4.12.1 安装 MySQL 及 RMySQL 131
 - 4.12.2 使用 RMySQL 访问 MySQL 数据库 140
- |参考资料| 141

第 5 章 ▶ 数据操作 II：数据处理及加工 143

- 5.1 数据处理及加工相关包 144
- 5.2 使用 SQL 处理数据 144
- 5.3 数据分析：拆分、应用、合并 146
 - 5.3.1 adply() 函数 147
 - 5.3.2 ddply() 函数 149
 - 5.3.3 轻松进行按组运算 150
 - 5.3.4 mdply() 153
- 5.4 数据结构变形与汇总 154
 - 5.4.1 melt() 155
 - 5.4.2 cast() 157
 - 5.4.3 数据汇总 158
- 5.5 数据表：更快、更方便的数据框 160
 - 5.5.1 创建数据表 160
 - 5.5.2 数据访问与分组运算 162
 - 5.5.3 使用 key 快速访问数据 164
 - 5.5.4 使用 key 合并数据表 166
 - 5.5.5 利用引用修改数据 167
 - 5.5.6 将列表转换为数据框 168
- 5.6 更好的循环语句 170
- 5.7 并行处理 172
 - 5.7.1 设置进程数 173
 - 5.7.2 plyr 并行化 174
 - 5.7.3 foreach 并行化 176
- 5.8 单元测试与调试 177
 - 5.8.1 testthat 177

- 5.8.2 使用 test_that() 进行测试分组 179
- 5.8.3 测试文件的结构 180
- 5.8.4 调试 181
- 5.9 测定代码执行时间 187
 - 5.9.1 测定命令语句执行时间 187
 - 5.9.2 代码性能测试 189
- |参考资料| 191

第6章 ▶ 绘图 193

- 6.1 散点图 194
- 6.2 图形选项 195
 - 6.2.1 坐标轴名称 196
 - 6.2.2 图形标题 197
 - 6.2.3 点的类型 197
 - 6.2.4 点的大小 198
 - 6.2.5 颜色 199
 - 6.2.6 坐标轴的取值范围 200
 - 6.2.7 图形类型 201
 - 6.2.8 线型 204
 - 6.2.9 图形排列 204
 - 6.2.10 抖动 205
- 6.3 基本图形 207
 - 6.3.1 点 207
 - 6.3.2 折线 209
 - 6.3.3 直线 211
 - 6.3.4 曲线 212
 - 6.3.5 多边形 213
- 6.4 字符串 216
- 6.5 识别图形中的数据 218
- 6.6 图例 219
- 6.7 绘制矩阵中的数据 (matplot、matlines、matpoints) 220
- 6.8 应用图形 221
 - 6.8.1 箱线图 222
 - 6.8.2 直方图 225
 - 6.8.3 密度图 227

- 6.8.4 条形图 229
- 6.8.5 饼图 230
- 6.8.6 马赛克图 232
- 6.8.7 散点图矩阵 234
- 6.8.8 透视图、等高线图 235
- 参考资料 238

第7章 ▶ 统计分析 239

- 7.1 生成随机数与分布函数 240
- 7.2 基本统计量 243
 - 7.2.1 样本均值、样本方差、样本标准差 243
 - 7.2.2 五数概括 244
 - 7.2.3 众数 246
- 7.3 样本抽取 246
 - 7.3.1 简单随机抽样 247
 - 7.3.2 考虑权值的样本抽取 248
 - 7.3.3 分层随机抽样 249
 - 7.3.4 系统抽样 251
- 7.4 列联表 252
 - 7.4.1 创建列联表 253
 - 7.4.2 求和与百分比 254
 - 7.4.3 独立性检验 256
 - 7.4.4 费舍尔精确检验 261
 - 7.4.5 McNemar 检验 262
- 7.5 拟合优度检验 265
 - 7.5.1 卡方检验 265
 - 7.5.2 夏皮罗-威尔克检验 265
 - 7.5.3 柯尔莫诺夫-斯米尔诺夫检验 266
 - 7.5.4 Q-Q 图 268
- 7.6 相关分析 271
 - 7.6.1 皮尔逊相关系数 272
 - 7.6.2 斯皮尔曼相关系数 275
 - 7.6.3 肯德尔等级相关系数 277
 - 7.6.4 相关系数检验 277
- 7.7 估计与检验 278

- 7.7.1 单样本均值 279
- 7.7.2 两独立样本均值 282
- 7.7.3 两配对样本均值 285
- 7.7.4 两样本方差 287
- 7.7.5 单样本比率 288
- 7.7.6 两样本比率 290

|参考资料| 291

第8章 ▶ 线性回归 293

- 8.1 线性回归的基本假设 294
 - 8.2 简单线性回归 295
 - 8.2.1 创建模型 295
 - 8.2.2 提取线性回归结果 296
 - 8.2.3 预测与置信区间 298
 - 8.2.4 模型评估 299
 - 8.2.5 方差分析及模型间比较 302
 - 8.2.6 模型诊断图形 304
 - 8.2.7 回归直线的可视化 306
 - 8.3 多元回归 307
 - 8.3.1 创建及评估模型 307
 - 8.3.2 分类变量 308
 - 8.3.3 多元回归模型的可视化 310
 - 8.3.4 使用函数 `l()` 312
 - 8.3.5 变量的变换 314
 - 8.3.6 交互作用 314
 - 8.4 异常值 320
 - 8.5 变量选择 321
 - 8.5.1 选择变量的方法 322
 - 8.5.2 比较所有情形 325
- |参考资料| 328

第9章 ▶ 分类算法 I：数据探索、预处理、模型评估方法 331

- 9.1 数据探索 332
 - 9.1.1 技术统计 332
 - 9.1.2 数据可视化 337

- 9.2 预处理 340
 - 9.2.1 数据变换 340
 - 9.2.2 缺失值处理 345
 - 9.2.3 变量选择 348
- 9.3 模型评估方法 358
 - 9.3.1 评估指标 358
 - 9.3.2 ROC 曲线 361
 - 9.3.3 交叉检验 365
- !参考资料! 375

第 10 章 ▶ 分类算法 II：机器学习算法 377

- 10.1 逻辑回归模型 378
- 10.2 多项逻辑回归分析 381
- 10.3 决策树 384
 - 10.3.1 决策树模型 385
 - 10.3.2 分类与回归树 386
 - 10.3.3 条件推断决策树 389
 - 10.3.4 随机森林 391
- 10.4 神经网络 396
 - 10.4.1 神经网络模型 396
 - 10.4.2 神经网络模型学习 398
- 10.5 支持向量机 402
 - 10.5.1 支持向量机模型 403
 - 10.5.2 支持向量机学习 404
- 10.6 类别不平衡 408
 - 10.6.1 向上取样、向下取样 409
 - 10.6.2 SMOTE 411
- 10.7 文档分类 413
 - 10.7.1 语料库与文档 413
 - 10.7.2 文档变换 414
 - 10.7.3 文档的矩阵表示 415
 - 10.7.4 高频词 418
 - 10.7.5 词语之间的相关关系 419
 - 10.7.6 文档分类 420
 - 10.7.7 从文件创建语料库 422

10.7.8	元数据	424
10.8	caret 包	427
!参考资料!		431
第 11 章 ▶	利用泰坦尼克数据练习机器学习	433
11.1	泰坦尼克数据格式	434
11.2	读入数据	434
11.2.1	转换数据类型	435
11.2.2	分离测试数据	437
11.2.3	准备交叉检验	438
11.3	数据探索	440
11.4	评估指标	444
11.5	决策树模型	444
11.5.1	rpart 的交叉检验	445
11.5.2	准确度评估	446
11.5.3	条件推断决策树	447
11.6	发现其他特征	448
11.6.1	使用 ticket 识别家庭	448
11.6.2	预测生还概率	449
11.6.3	添加家庭 ID	450
11.6.4	合并家庭成员的生还概率	452
11.6.5	使用家庭信息建模 ctree()	454
11.6.6	性能评估	455
11.7	交叉检验并行化	457
11.7.1	反复执行 3 次 10 层交叉检验	457
11.7.2	使用 foreach() 与 %dopar% 进行并行化	458
11.8	开发更好的算法	459
!参考资料!		460