

R Cookbook



R语言 经典实例

O'REILLY®



机械工业出版社
China Machine Press

Paul Teetor 著

李洪成 朱文佳 沈毅诚 译

013334252

TP312
937

R语言经典实例



Paul Teetor 著

李洪成 朱文佳 沈毅诚 译

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

O'Reilly Media, Inc. 授权机械工业出版社出版

机械工业出版社



北航

C1641565

TP312
937

图书在版编目 (CIP) 数据

01838422

R语言经典实例 / (美) 蒂特 (Teetor, P.) 著; 李洪成, 朱文佳, 沈毅诚译.
—北京: 机械工业出版社, 2013.4

(O'Reilly精品图书系列)

书名原文: R Cookbook

ISBN 978-7-111-42021-7

I. R… II. ①蒂… ②李… ③朱… ④沈… III. 程序语言—程序设计 IV. TP312
中国版本图书馆CIP数据核字 (2013) 第066465号

北京市版权局著作权合同登记

图字: 01-2011-7869号

©2011 by O'Reilly Media, Inc.

Simplified Chinese Edition, jointly published by O'Reilly Media, Inc. and China Machine Press, 2013. Authorized translation of the English edition, 2011 O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由O'Reilly Media, Inc. 出版2011。

简体中文版由机械工业出版社出版2013。英文原版的翻译得到O'Reilly Media, Inc.的授权。此简体中文版的出版和销售得到出版权和销售权的所有者——O'Reilly Media, Inc.的许可。

版权所有, 未得书面许可, 本书的任何部分和全部不得以任何形式复制。

封底无防伪标均为盗版

本书法律顾问

北京市展达律师事务所

书 名/ R语言经典实例

书 号/ ISBN 978-7-111-42021-7

责任编辑/ 盛思源

封面设计/ Karen Montgomery, 张健

出版发行/ 机械工业出版社

地 址/ 北京市西城区百万庄大街22号 (邮政编码 100037)

印 刷/ 藁城市京瑞印刷有限公司

开 本/ 178毫米×233毫米 16开本 26.5印张

版 次/ 2013年5月第1版 2013年5月第1次印刷

定 价/ 79.00元 (册)

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)88378991; 88361066

购书热线: (010)68326294; 88379649; 68995259

投稿热线: (010)88379604

读者信箱: hzsj@hzbook.com



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

中文版序

R软件的应用正变得越来越广泛。在雇主招聘需求所要求的技能中，以前只要求具备SPSS技能和SAS技能，现在（2012年6月写这个序言的时候）的招聘需求中对应用R软件的技能更加关注。在美国统计教育学院*Statistics.com*的在线课程中，R课程是增长最快的。R的流行经历了三个阶段。在第一个阶段，R先是由于它的开放源代码和免费而在学术界变得流行；第二个阶段，应用R的相关专业的研究生把R带到他们的工作环境中；第三个阶段，由于研究生把R的火种带到工作中，商业分析领域中非统计专业研究生的其他分析人员也开始试着应用R。从地域上讲，R最早发源于新西兰，然后传播到了美国和欧洲，现在R在中国也逐渐变得广为人知。

本书可以帮助你迅速提高R的技巧。你可以不必按照作者认为重要的冗长主题进行阅读，你可以立即专注于你的工作任务，学习如何用R来解决这些任务。这些任务都在书中有明确的标识，它们组织精炼，有的任务只有一页或者两页，便于查找，方便应用。

如果你寻求解决某个特定任务的方法，你可以在因特网上搜索。那么在本书中寻找解决方案的优势在哪里呢？其一，你可以有信心本书中提供的解决方案是合理的；其二，本书有一致的风格，你可以很容易地找到你需要的内容并应用它们。

大多数烹饪的人是根据菜谱来学习烹饪，或者应用在新情况下获取的知识来学习烹饪，没有人通过阅读食物的历史并学习它们的化学成分来学习烹饪。同样的道理，随着你的R知识的拓展，你会发现本书提供了相当有效的“菜谱”和建议。

——彼得·布如斯

美国统计教育学院*Statistics.com*总裁

译者序

本书的英文版从出版后就在亚马逊美国网站上收到了极高的评价。机械工业出版社以极快的速度引进这本书的中文版，使国内读者在原版出版一年左右的时间里读到中文版，不得不赞扬他们独到的眼光。这本书中文稿完成的时候（2012年6月），其英文版的销量还是排在专业书籍的前列，并稳居生物信息学类书籍的第一位，排在建模和仿真类书籍的第7位。

本书的作者Paul Teetor是一位统计学专家和计算机专家，他同时也是一位量化投资分析和风险管理专家。他和大芝加哥地区的对冲基金、投资组合经理一起工作，因此有丰富的投资量化分析经验。他把R软件用户在学习和应用R软件中常见的问题进行系统整理和总结，汇集成了本书。书中涵盖了从R软件的基础知识（安装、帮助系统、解决实际问题的途径、R数据结构、R编程、R的输入和输出等）到用R进行数据分析的具体方法（数据变换、概率和统计基础、R绘图、回归分析和方差分析、R常用技巧、高级数据分析方法和时间序列分析等）。全书以问题、解决方案和对解决方案的讨论与拓展为主线来组织内容。读者既可以把本书作为学习R的一本优秀教材，也可以根据自己的需要参考书中的某些具体方法，找到自己实际问题的解决方案。

R本身是一款十分优秀的统计分析软件，R的书籍和文档也是相当多的。但是缺乏一本适合R初学者的书籍，尤其是针对那些对R不甚精通但是急切需要用R来解决问题的R用户。本书以问题和解决方案的形式组织内容，脉络清晰，读者很容易找到自己需要的内容。不管是R初学者，还是熟练的R用户都能从书中找到对自己有用的内容。

本书的译者从2003年年初接触R，那时国内几乎没有多少R用户。现在情况发生了极大的变化，R成为数据分析从业者谈论最多的软件之一，许多学校的统计系在教学中也广

泛地应用R。从2008年开始，国内的R用户每年召开R用户大会，大力传播和推广R。从2011年开始国内出版社也开始出版和引进R书籍，这对国内R的推广都起到了极大的作用。

在本书的翻译过程中，得到了原作者Paul Tector的大力帮助，他解答了译者在翻译过程中遇到的各种问题。美国统计教育学院*Statistics.com*是目前世界上开设R在线课程最多的机构，十分感谢他们的总裁彼得·布如斯先生欣然为本书中文版作序。另外，十分感谢机械工业出版社的王春华老师的大力支持和帮助，感谢盛思源编辑对本书一丝不苟的校对与检查。本书的翻译工作由李洪成、朱文佳和沈毅诚共同完成，由于水平所限，中间可能会有翻译不当之处，希望读者多加指正。

——李洪成

目录

前言	1
第1章 R入门和获得帮助	7
1.1 下载和安装R软件	8
1.2 开始运行R软件	10
1.3 输入R命令	13
1.4 退出R	15
1.5 中断R正在运行的程序	16
1.6 查看帮助文档	17
1.7 获取函数的帮助文档	18
1.8 搜索帮助文档	20
1.9 查看R软件包帮助信息	21
1.10 通过网络获取帮助	23
1.11 寻找相关函数与数据包	26
1.12 查询邮件列表	27
1.13 向邮件列表提交问题	27
第2章 基础知识	30
2.1 显示内容	30
2.2 设定变量	32
2.3 列出所有变量	34
2.4 删除变量	35
2.5 生成向量	36

2.6 计算基本统计量	37
2.7 生成数列	40
2.8 向量比较	42
2.9 选取向量中的元素	43
2.10 向量的计算	46
2.11 运算符优先级问题	48
2.12 定义函数	50
2.13 减少输入，得到更多命令	52
2.14 常见错误	54
第3章 R软件导览	58
3.1 获取和设定工作目录	58
3.2 保存工作空间	59
3.3 查看历史命令记录	60
3.4 保存先前命令产生的结果	60
3.5 显示搜索路径	61
3.6 使用R包中的函数	62
3.7 使用R的内置数据集	64
3.8 查看已安装的R包列表	65
3.9 从CRAN网站安装R包	67
3.10 设定默认CRAN网站镜像	69
3.11 隐藏启动信息	70
3.12 运行脚本	70
3.13 批量运行R代码	71
3.14 获取和设定环境变量	74
3.15 找到R的主目录	75
3.16 R的客户化	76
第4章 输入与输出	80
4.1 使用键盘输入数据	81
4.2 显示更少的位数（或更多的位数）	82
4.3 将输出结果重定向到某一文件	84

4.4 显示文件列表	85
4.5 解决无法在Windows中打开文件的问题	86
4.6 阅读固定宽度数据记录	87
4.7 读取表格数据文件	88
4.8 读取CSV文件	90
4.9 写入CSV文件	92
4.10 从网络中读取表格或CSV格式数据	93
4.11 读取HTML表格数据	94
4.12 读取复杂格式数据文件	96
4.13 读取MySQL数据库中的数据	100
4.14 保存和传送目标	102

第5章 数据结构 104

5.1 对向量添加数据	111
5.2 在向量中插入数据	112
5.3 理解循环规则	113
5.4 构建因子（即分类变量）	115
5.5 将多个向量合并成单个向量以及平行因子	117
5.6 创建列表	118
5.7 根据位置选定列表元素	119
5.8 根据名称选定列表元素	121
5.9 构建一个名称/值关联表	122
5.10 从列表中移除元素	124
5.11 将列表转换为向量	125
5.12 从列表中移除取值为空值（即NULL）的元素	126
5.13 使用条件来移除列表元素	127
5.14 矩阵初始化	129
5.15 执行矩阵运算	130
5.16 将描述性名称赋给矩阵的行和列	131
5.17 从矩阵中选定一行或一列	132
5.18 用列数据初始化数据框	133
5.19 由行数据初始化数据框	134

5.20 添加行至数据框.....	136
5.21 预分配数据框	137
5.22 根据位置选择数据框的列.....	138
5.23 根据列名选定数据框的列.....	142
5.24 更便捷地选定行和列	143
5.25 修改数据框的列名	145
5.26 编辑数据框	146
5.27 从数据框中移除NA值	148
5.28 根据名称排除列.....	149
5.29 合并两个数据框.....	150
5.30 根据共有列合并数据框	151
5.31 更便捷地访问数据框内容.....	152
5.32 基本数据类型之间的转换.....	154
5.33 不同结构化数据类型间的转换	156
第6章 数据转换	159
6.1 向量分组.....	160
6.2 将函数应用于每个列表元素	161
6.3 将函数应用于每行.....	163
6.4 将函数应用于每列.....	164
6.5 将函数应用于组数据	166
6.6 将函数应用于行组.....	168
6.7 将函数应用于平行向量或列表	170
第7章 字符串和日期.....	172
7.1 获取字符串长度	174
7.2 连接字符串	175
7.3 提取子串.....	176
7.4 根据分隔符分割字符串	176
7.5 替代子串.....	178
7.6 查看字符串中的特殊字符.....	179
7.7 生成字符串的所有成对组合	179

7.8 得到当前日期	181
7.9 转换字符串为日期	181
7.10 转换日期为字符串	182
7.11 转化年、月、日为日期	183
7.12 得到儒略日期	185
7.13 提取日期的一部分	185
7.14 创建日期序列	187
第8章 概率	189
8.1 计算组合数	191
8.2 生成组合	192
8.3 生成随机数	193
8.4 生成可再生的随机数	194
8.5 生成随机样本	196
8.6 生成随机序列	197
8.7 随机排列向量	198
8.8 计算离散分布的概率	198
8.9 计算连续分布的概率	200
8.10 转换概率为分位数	201
8.11 绘制密度函数	203
第9章 统计概论	206
9.1 汇总数据	208
9.2 计算相对频数	210
9.3 因子制表和列联表创建	211
9.4 检验分类变量独立性	212
9.5 计算数据集的分位数（和四分位数）	212
9.6 求分位数的逆	213
9.7 数据转换为z分数	214
9.8 检验样本均值（t检验）	215
9.9 均值的置信区间	216
9.10 中位数的置信区间	217

9.11 检验样本比例	218
9.12 比例的置信区间.....	219
9.13 检验正态性	220
9.14 游程检验.....	222
9.15 比较两个样本的均值	223
9.16 比较两个非参数样本的位置	225
9.17 检验相关系数的显著性	226
9.18 检验组的等比例.....	228
9.19 组均值间成对比较	229
9.20 检验两样本的相同分布	230

第10章 图形 232

10.1 创建散点图	234
10.2 添加标题和标签.....	236
10.3 添加网格.....	237
10.4 创建多组散点图.....	238
10.5 添加图例.....	240
10.6 绘制散点图的回归线	242
10.7 多变量散点图的绘制	243
10.8 创建每个因子水平的散点图	244
10.9 创建条形图	246
10.10 对条形图添加置信区间	248
10.11 给条形图上色.....	249
10.12 绘制过点x和y的线	251
10.13 改变线的类型、宽度或者颜色	253
10.14 绘制多个数据集.....	254
10.15 添加垂直线和水平线	256
10.16 创建箱线图	257
10.17 对每个因子水平创建箱线图	258
10.18 创建直方图	259
10.19 对直方图添加密度估计	261
10.20 创建离散直方图.....	262

10.21 创建正态Q-Q图	264
10.22 创建其他Q-Q图	265
10.23 用多种颜色绘制变量	266
10.24 绘制函数	269
10.25 图形间暂停	270
10.26 在一页中显示多个图形	271
10.27 打开另一个图形窗口	273
10.28 在文档中绘制图形	274
10.29 改变图形参数	275

第11章 线性回归和方差分析..... 277

11.1 简单线性回归	279
11.2 多元线性回归	281
11.3 得到回归统计量	282
11.4 理解回归的汇总结果	286
11.5 运行无截距的线性回归	289
11.6 运行有交互项的线性回归	290
11.7 选择最合适的回归变量	292
11.8 对数据子集回归	295
11.9 在回归公式中使用表达式	296
11.10 多项式回归	298
11.11 转换数据的回归	299
11.12 寻找最佳幂变换	301
11.13 回归系数的置信区间	304
11.14 绘制回归残差	304
11.15 诊断线性回归	306
11.16 识别有影响的观察值	309
11.17 残差自相关检验	310
11.18 预测新值	311
11.19 建立预测区间	312
11.20 运行单因素方差分析	313
11.21 创建交互关系图	315

11.22 找到组间均值的不同	316
11.23 执行稳健方差分析	318
11.24 运用方差分析比较模型	320

第12章 有用的方法 323

12.1 查看你的数据	323
12.2 拓宽你的输出	324
12.3 输出赋值结果	325
12.4 对行和列求和	325
12.5 按列输出数据	326
12.6 对数据分级	328
12.7 找到特定值的位置	329
12.8 每隔 n 个选定一个向量元素	330
12.9 找到成对的最小值或者最大值	331
12.10 生成多个因子的组合	332
12.11 转换一个数据框	333
12.12 对数据框排序	334
12.13 对两列排序	335
12.14 移除变量属性	336
12.15 显示对象的结构	337
12.16 代码运行时间	340
12.17 抑制警告和错误消息	341
12.18 从列表中提取函数参数	342
12.19 定义你自己的二元运算符	344

第13章 高级数值分析和统计方法 347

13.1 最小化或者最大化一个单参数函数	347
13.2 最小化或者最大化多参数函数	348
13.3 计算特征值和特征向量	350
13.4 主成分分析	351
13.5 简单正交回归	352
13.6 数据的聚类	354

13.7 预测二元变量 (逻辑回归)	357
13.8 统计量的自助法	359
13.9 因子分析	361
第14章 时间序列分析	366
14.1 表示时间序列	367
14.2 绘制时序图	370
14.3 提取最老的观测值或者最新的观测值	373
14.4 选取时间序列的子集	374
14.5 合并多个时间序列	376
14.6 缺失时间序列的填充	378
14.7 时间序列的滞后	380
14.8 计算逐次差分	381
14.9 时间序列相关的计算	382
14.10 计算移动平均	383
14.11 在日历时间范围内应用函数	384
14.12 应用滚动函数	386
14.13 绘制自相关函数图	388
14.14 检验时间序列的自相关	389
14.15 绘制偏自相关函数	390
14.16 两个时间序列间的滞后相关性	391
14.17 剔除时间序列的趋势	393
14.18 拟合ARIMA模型	394
14.19 剔除ARIMA模型中不显著的系数	397
14.20 对ARIMA模型进行诊断	399
14.21 用ARIMA模型进行预测	400
14.22 均值回归的检验	402
14.23 时间序列的平滑	404

前言

R软件是进行统计分析、绘图和统计编程的强大工具。现在成千上万的人用它来进行日常的重要统计分析。R软件是一个自由、开源的软件平台，它是许多聪明、勤奋工作的人们集体工作的成果。R软件有超过两千多个软件包插件。R软件是其他所有商业统计软件包的强劲竞争对手。

但是，开始使用R软件可能感到无从下手。对于许多任务，即便是一些基本的任务，R的实现也不是很明显。当了解了R的使用方法后，简单的问题自然能得心应手地解决，但学习“如何”使用R的过程有时会让人感到发狂。

本书介绍了如何使用R软件的一些方法，其中每一个方法对应解决某个特定的问题。介绍这些方法的途径是这样的：首先给出待解决的问题，然后给出解决方案的简单介绍，之后再给出对解决方案的讨论，深入剖析解决方案，给出该方案的原理。我知道这些方法有效实用，我也知道这些方法可行，因为我本人也使用它们。

这些方法所涉及的范围较为广泛。首先从基本的任务开始介绍，然后介绍数据的输入和输出、基础统计、绘图以及线性回归。与R有关的工作都将或多或少地涉及本书介绍的方法。

通过本书的讲解，初学者能快速地了解R并获得提高。如果你对R软件有一定的了解，那么本书也能帮助你巩固已学的知识，拓宽你的思维（例如，“下一次我应该怎么使用K-S检验”）。

从严格意义上来说，本书并不是一本关于R软件的教程，但你将会从中学习到许多R软件的应用技巧。本书也不是一本关于R的参考手册，但它确实包含了许多实用的内容。本书也不是一本R软件的编程指南，但书中很多方法都可以应用到R的编程脚本中。