

Marching onto the road of Big Data

# R 语言

## 迈向大数据之路

R 的起源、现在与未来

RStudio 视窗完整解说

R 的信息结构完整解说

R 与其他软件交流

数据分析与统计绘图

R 语言高阶低阶绘图

【全书包含650个实例】

请至清华大学出版社网站  
下载本书范例

洪锦魁  
蔡桂宏 著



清华大学出版社

# R语言

## 迈向大数据之路

洪锦魁◎著  
蔡桂宏

清华大学出版社  
北京

## 内 容 简 介

DOS 时代用汇编语言，Windows 时代倡导 Windows 编程，Internet 时代是 HTML 的天下，进入大数据时代，R 语言必须掌握！

本书作者作为一名历经四个时代的老程序员，深知学习编程的痛苦与欢乐，结合多年的开发经验完成此书。

本书将从无到有地教读者 R 语言的使用，同时学习本书并不需要统计学基础，在学习编程的过程中，就掌握了一些必要的统计知识。本书完整讲解了几乎所有 R 语言语法与使用技巧，通过丰富的程序案例讲解，让你事半功倍。

本书代码请到清华大学出版社网站下载。

本书封面贴有清华大学出版社防伪标签，无标签者不得销售。  
版权所有，侵权必究。侵权举报电话：010-62782989 13701121933

### 图书在版编目(CIP)数据

R 语言——迈向大数据之路 / 洪锦魁，蔡桂宏著. — 北京：清华大学出版社，2016  
ISBN 978-7-302-43005-6

I. ①R… II. ①洪… ②蔡… 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 031099 号

责任编辑：栾大成  
装帧设计：杨玉兰  
责任校对：徐俊伟  
责任印制：李红英

出版发行：清华大学出版社

网 址：<http://www.tup.com.cn>，<http://www.wqbook.com>

地 址：北京清华大学学研大厦 A 座 邮 编：100084

社 总 机：010-62770175 邮 购：010-62786544

投稿与读者服务：010-62776969，[c-service@tup.tsinghua.edu.cn](mailto:c-service@tup.tsinghua.edu.cn)

质量反馈：010-62772015，[zhiliang@tup.tsinghua.edu.cn](mailto:zhiliang@tup.tsinghua.edu.cn)

印 刷 者：北京富博印刷有限公司

装 订 者：北京市密云县京文制本装订厂

经 销：全国新华书店

开 本：188mm×260mm 印 张：31.25 字 数：654 千字

版 次：2016 年 6 月第 1 版 印 次：2016 年 6 月第 1 次印刷

印 数：1~3000

定 价：69.00 元

---

产品编号：067985-01

# 前言

在 DOS 时代，我写了 Assembly Language。

在 Windows 时代，我写了 Windows Programming Using C 和 Visual Basic。

在 Internet 时代，我写了 HTML。

写了许多的书，曾经也想退休……但仍在职场。

今天是 Big Data 时代，我完成了 R。

在 DOS 时代，我在撰写 Assembly Language 时，完成了汇编语言语法以及完整的 DOS 和 BIOS 应用的相关写作，我深知，这本书是当时最完整的汇编语言教材，我的心情是愉快的。

在 Windows 时代，我在撰写 Windows Programming 时，完成了几乎所有 Windows 组件的重新设计的写作，当初愉快的心情再度涌上心头。

在 Internet 时代，我在撰写 HTML，完成了各类网页功能的几乎所有组件设计的写作，内心有了亢奋。

现在是 Big Data 时代，若想进入这个领域，R 可说是最重要的程序语言，目前 R 语言的参考数据不多，现有几本 R 语言教材均是统计专家所撰写的，内容叙述在 R 语言部分着墨不多，这也造成了目前大多数人无法完整学习 R 语言，就进入 Big Data 的世界，即使会用 R 语言作数据分析，对于 R 的使用也无法全面了解。很多年以来，除了软件改版的书我不再写新书，因缘，我进入了这个领域，完成了这本 R 语言著作，这本书的最大特色包括以下几点。。

- (1) 从无到有一步一步教导读者 R 语言的使用。
- (2) 学习本书不需要有统计基础，但在无形中本书已灌输了统计知识给你。
- (3) 完整讲解所有 R 语言语法与使用技巧。
- (4) 丰富的程序实例与解说，让你事半功倍。

坦白说，当年撰写汇编语言时的那种心情愉快亢奋的感觉再度涌上心头，因为我知道这将是目前 R 语言最完整的教材。

最后预祝读者们学习顺利！

编者

## 特别提示

本书作者为台湾著名跨界资深程序员，虽然本书经过了较为细致的本地化工作，但是仍有极个别位置（主要是图片）存在个别繁体字，见谅！

# 目录

## Chapter 01 基本概念

1-1 Big Data 的起源 .....	2
1-2 R 语言之美 .....	2
1-3 R 语言的起源 .....	2
1-4 R 的运行环境 .....	5
1-5 R 的扩展 .....	5
1-6 本书的学习目标 .....	5
本章习题 .....	6

## Chapter 02 第一次使用 R

2-1 第一次启动 R .....	8
2-1-1 在 Mac OS 下启动 R .....	8
2-1-2 在 Mac OS 下启动 RStudio .....	8
2-1-3 在 Windows 环境中启动 R 和 RStudio .....	9
2-2 认识 RStudio 环境 .....	10
2-3 第一次使用 R .....	12
2-4 R 语言的对象设定 .....	15
2-5 Workspace 窗口 .....	16
2-6 结束 RStudio .....	18
2-7 保存工作成果 .....	19
2-7-1 使用 save () 函数保存工作成果 .....	19
2-7-2 使用 save.image () 函数保存 Workspace .....	20
2-7-3 下载之前保存的工作 .....	20
2-8 历史记录 .....	21
2-9 程序注释 .....	22
本章习题 .....	24

## Chapter 03 R 的基本数学运算

3-1 对象命名原则 .....	28
3-2 基本数学运算 .....	28

3-2-1	四则运算	28
3-2-2	余数和整除	29
3-2-3	次方或平方根	29
3-2-4	绝对值	30
3-2-5	exp () 与对数	30
3-2-6	科学符号 e	31
3-2-7	圆周率与三角函数	32
3-2-8	四舍五入函数	32
3-2-9	近似函数	33
3-2-10	阶乘	34
3-3	R 语言控制运算的优先级	34
3-4	无限大 Infinity	35
3-5	Not a Number (NaN)	36
3-6	Not Available (NA)	37
	本章习题	39

## Chapter 04 向量对象运算

4-1	数值型的向量对象	44
4-1-1	建立规则型的数值向量对象应使用序列符号	44
4-1-2	简单向量对象的运算	45
4-1-3	建立向量对象函数 seq ()	46
4-1-4	连接向量对象函数 c ()	47
4-1-5	重复向量对象函数 rep ()	48
4-1-6	numeric () 函数	48
4-1-7	程序语句跨行的处理	49
4-2	常见向量对象的数学运算函数	50
4-3	考虑 Inf、-Inf、NA 的向量运算	53
4-4	R 语言的字符串数据的属性	54
4-5	探索对象的属性	55
4-5-1	探索对象元素的属性	55
4-5-2	探索对象的结构	56
4-5-3	探索对象的数据类型	57
4-6	向量对象元素的存取	57
4-6-1	使用索引取得向量对象的元素	57
4-6-2	使用负索引挖掘向量对象内的部分元素	58



4-6-3	修改向量对象元素值 .....	59
4-6-4	认识系统内建的数据集 letters 和 LETTERS .....	60
4-7	逻辑向量 (Logical Vector) .....	61
4-7-1	基本应用 .....	61
4-7-2	对 Inf、-Inf 和缺失值 NA 的处理 .....	63
4-7-3	多组逻辑表达式的应用 .....	64
4-7-4	NOT 表达式 .....	65
4-7-5	逻辑值 TRUE 和 FALSE 的运算 .....	65
4-8	不同长度向量对象相乘的应用 .....	66
4-9	向量对象的元素名称 .....	67
4-9-1	建立简单含元素名称的向量对象 .....	67
4-9-2	names () 函数 .....	67
4-9-3	使用系统内建的数据集 islands .....	68
	本章习题 .....	71

## Chapter 05 处理矩阵与更高维数据

5-1	矩阵 Matrix .....	78
5-1-1	建立矩阵 .....	78
5-1-2	认识矩阵的属性 .....	79
5-1-3	将向量组成矩阵 .....	81
5-2	取得矩阵元素的值 .....	82
5-2-1	矩阵元素的取得 .....	82
5-2-2	使用负索引取得矩阵元素 .....	83
5-3	修改矩阵的元素值 .....	84
5-4	降低矩阵的维度 .....	86
5-5	矩阵的行名和列名 .....	87
5-5-1	取得和修改矩阵对象的行名和列名 .....	88
5-5-2	dimnames () 函数 .....	89
5-6	将行名或列名作为索引 .....	90
5-7	矩阵的运算 .....	91
5-7-1	矩阵与一般常数的四则运算 .....	91
5-7-2	行 (Row) 和列 (Column) 的运算 .....	93
5-7-3	转置矩阵 .....	94
5-7-4	%*% 矩阵相乘 .....	94
5-7-5	diag () .....	95

5-7-6	solve ()	96
5-7-7	det ()	97
5-8	三维或高维数组	97
5-8-1	建立三维数组	97
5-8-2	identical () 函数	98
5-8-3	取得三维数组的元素	98
5-9	再谈 class () 函数	99
	本章习题	101

## Chapter 06 因子 Factor

6-1	使用 factor () 或 as.factor () 函数建立因子	108
6-2	指定缺失的 Levels 值	109
6-3	labels 参数	109
6-4	因子的转换	110
6-5	数值型因子在转换时常见的错误	110
6-6	再看 levels 参数	111
6-7	有序因子 (Ordered Factor)	112
6-8	table () 函数	113
6-9	认识系统内建的数据集	114
	本章习题	116

## Chapter 07 数据框 Data Frame

7-1	认识数据框	120
7-1-1	建立第一个数据框	120
7-1-2	验证与设置数据框的列名和行名	121
7-2	认识数据框的结构	121
7-3	取得数据框的内容	122
7-3-1	一般取得	122
7-3-2	特殊字符 \$	123
7-3-3	再看取得的数据	123
7-4	使用 rbind () 函数增加数据框的行数据	124
7-5	使用 cbind () 函数增加数据框的列数据	125
7-5-1	使用 \$ 符号	126
7-5-2	一次加多个列数据	126
7-6	再谈转置函数 t ()	127
	本章习题	128



## Chapter 08 串行 List

8-1 建立串行.....	134
8-1-1 建立串行对象——对象元素不含名称.....	134
8-1-2 建立串行对象——对象元素含名称.....	134
8-1-3 处理串行内对象元素的名称.....	135
8-1-4 获得串行的对象元素个数.....	136
8-2 获得串行内对象的元素内容.....	136
8-2-1 使用“\$”符号取得串行内对象的元素内容.....	136
8-2-2 使用“[[ ]]”符号取得串行内对象的元素内容.....	137
8-2-3 串行内对象的名称也可当索引值.....	137
8-2-4 使用“[]”符号取得串行内对象的元素内容.....	138
8-3 编辑串行内对象的元素值.....	139
8-3-1 修改串行元素的内容.....	139
8-3-2 为串行增加更多元素.....	141
8-3-3 删除串行内的元素.....	144
8-4 串行合并.....	145
8-5 解析串行的内容结构.....	146
本章习题.....	148

## Chapter 09 进阶字符串的处理

9-1 语句的分割.....	154
9-2 修改字符串的大小写.....	154
9-3 unique () 函数的使用.....	155
9-4 字符串的连接.....	155
9-4-1 使用 paste () 函数常见的失败实例 1.....	155
9-4-2 使用 paste () 函数常见的失败实例 2.....	156
9-4-3 字符串的成功连接与 collapse 参数.....	156
9-4-4 再谈 paste () 函数.....	157
9-4-5 扑克牌向量有趣的应用.....	158
9-5 字符串数据的排序.....	158
9-6 搜索字符串的内容.....	159
9-6-1 使用索引值搜索.....	160
9-6-2 使用 grep () 函数搜索.....	160
9-7 字符串内容的更改.....	161
9-8 正则表达式 (Regular Expression).....	162

9-8-1 搜索具有可选择性.....	162
9-8-2 搜索分类字符串.....	163
9-8-3 搜索部分字符可重复的字符串.....	163
本章习题.....	164

### Chapter 10 日期和时间的处理

10-1 日期的设置与使用.....	170
10-1-1 as.Date () 函数.....	170
10-1-2 weekdays () 函数.....	170
10-1-3 months () 函数.....	171
10-1-4 quarters () 函数.....	171
10-1-5 Sys.localeconv () 函数.....	171
10-1-6 Sys.Date () 函数.....	172
10-1-7 再谈 seq () 函数.....	172
10-1-8 使用不同格式表示日期.....	173
10-2 时间的设置与使用.....	173
10-2-1 Sys.time () 函数.....	174
10-2-2 as.POSIXct () 函数.....	174
10-2-3 时间也是可以作比较的.....	175
10-2-4 seq () 函数与时间.....	175
10-2-5 as.POSIXlt () 函数.....	175
10-3 时间序列.....	177
本章习题.....	180

### Chapter 11 编写自己的函数

11-1 正式编写程序.....	184
11-2 函数的基本组成.....	184
11-3 设计第一个函数.....	185
11-4 函数也是一个对象.....	186
11-5 程序代码的简化.....	187
11-6 return () 的功能.....	188
11-7 省略函数的大括号.....	189
11-8 传递多个函数参数的应用.....	190
11-8-1 设计可传递两个参数的函数.....	190
11-8-2 函数参数的默认值.....	191

11-8-3 3点参数“...”的使用 .....	192
11-9 函数也可以作为参数 .....	194
11-9-1 正式实例应用 .....	194
11-9-2 以函数的程序代码作为参数传送 .....	195
11-10 局部变量和全局变量 .....	195
11-11 通用函数 (Generic Function) .....	196
11-11-1 认识通用函数 print () .....	197
11-11-2 通用函数的默认函数 .....	198
11-12 设计第一个通用函数 .....	198
11-12-1 优化转换百分比函数 .....	199
11-12-2 设计通用函数的默认函数 .....	200
本章习题 .....	202

## Chapter 12 程序的流程控制

12-1 if 语句 .....	208
12-1-1 if 语句的基本操作 .....	208
12-1-2 if ... else 语句 .....	210
12-1-3 if 语句也可有返回值 .....	212
12-1-4 if ... else if ... else if ... else .....	213
12-1-5 嵌套式 if 语句 .....	214
12-2 递归式函数的设计 .....	215
12-3 向量化的逻辑表达式 .....	217
12-3-1 处理向量数据时 if ... else 产生的错误 .....	217
12-3-2 ifelse () 函数 .....	217
12-4 switch 语句 .....	219
12-5 for 循环 .....	221
12-6 while 循环 .....	224
12-7 repeat 循环 .....	225
12-8 再谈 break 语句 .....	226
12-9 next 语句 .....	227
本章习题 .....	228

## Chapter 13 认识 apply 家族

13-1 apply () 函数 .....	234
13-2 sapply () 函数 .....	236
13-3 lapply () 函数 .....	238

13-4	tapply () 函数	238
13-5	iris 鸢尾花数据集	240
	本章习题	242

## Chapter 14 输入与输出

14-1	认识文件夹	248
14-1-1	getwd () 函数	248
14-1-2	setwd () 函数	248
14-1-3	file.path () 函数	248
14-1-4	dir () 函数	248
14-1-5	list.files () 函数	249
14-1-6	file.exists () 函数	250
14-1-7	file.rename () 函数	250
14-1-8	file.create () 函数	250
14-1-9	file.copy () 函数	250
14-1-10	file.remove () 函数	251
14-2	数据输出 cat () 函数	251
14-3	读取数据 scan () 函数	253
14-4	输出数据 write () 函数	256
14-5	数据的输入	257
14-5-1	读取剪贴板数据	257
14-5-2	读取剪贴板数据 read.table () 函数	258
14-5-3	读取 Excel 文件数据	259
14-5-4	认识 CSV 文件以及如何读取 Excel 文件数据	260
14-5-5	认识 delim 文件以及如何读取 Excel 文件数据	262
14-6	数据的输出	263
14-6-1	writeClipboard () 函数	263
14-6-2	write.table () 函数	264
14-7	处理其他数据	265
	本章习题	272

## Chapter 15 数据分析与处理

15-1	复习数据类型	276
15-2	随机抽样	276
15-2-1	将随机抽样应用于扑克牌	277

15-2-2	种子值 .....	277
15-2-3	模拟骰子 .....	279
15-2-4	比重的设置 .....	279
15-3	再谈向量数据的抽取并以 islands 为实例 .....	280
15-4	数据框数据的抽取——对重复值的处理 .....	282
15-4-1	重复值的搜索 .....	284
15-4-2	which () 函数 .....	285
15-4-3	抽取数据时去除重复值 .....	285
15-5	数据框数据的抽取——对 NA 值的处理 .....	287
15-5-1	抽取数据时去除含 NA 值的行数据 .....	287
15-5-2	na.omit () 函数 .....	288
15-6	数据框的字段运算 .....	289
15-6-1	基本数据框的字段运算 .....	289
15-6-2	with () 函数 .....	290
15-6-3	identical () 函数 .....	290
15-6-4	将字段运算结果存入新的字段 .....	290
15-6-5	within () 函数 .....	291
15-7	数据的分割 .....	291
15-7-1	cut () 函数 .....	292
15-7-2	分割数据时直接使用 labels 设定名称 .....	292
15-7-3	了解每一人口数分类有多少州 .....	293
15-8	数据的合并 .....	293
15-8-1	之前的准备工作 .....	294
15-8-2	merge () 函数使用于交集合并的情况 .....	295
15-8-3	merge () 函数使用于并集合并的情况 .....	296
15-8-4	merge () 函数参数 “all.x = TRUE” .....	296
15-8-5	merge () 函数参数 “all.y = TRUE” .....	297
15-8-6	match () 函数 .....	297
15-8-7	%in% .....	298
15-8-8	match () 函数结果的调整 .....	299
15-9	数据的排序 .....	299
15-9-1	之前的准备工作 .....	299
15-9-2	向量的排序 .....	300
15-9-3	order () 函数 .....	301
15-9-4	数据框的排序 .....	301

15-9-5	排序时增加次要键值的排序 .....	302
15-9-6	混合排序与 <code>xtfrm()</code> 函数 .....	304
15-10	系统内建数据集 <code>mtcars</code> .....	305
15-11	<code>aggregate()</code> 函数 .....	307
15-11-1	基本使用 .....	307
15-11-2	公式符号 Formula Notation .....	307
15-12	建立与认识数据表格 .....	308
15-12-1	认识长格式数据与宽格式数据 .....	309
15-12-2	<code>reshapes2</code> 扩展包 .....	309
15-12-3	将宽格式数据转成长格式数据 <code>melt()</code> 函数 .....	310
15-12-4	将长格式数据转成宽格式数据 <code>dcast()</code> 函数 .....	312
	本章习题 .....	315

## Chapter 16 数据汇总与简单图表制作

16-1	之前的准备工作 .....	320
16-1-1	下载 MASS 扩展包与 <code>crabs</code> 对象 .....	320
16-1-2	准备与调整系统内建 <code>state</code> 相关对象 .....	320
16-1-3	准备 <code>mtcars</code> 对象 .....	322
16-2	了解数据的唯一值 .....	322
16-3	基础统计知识与 R 语言 .....	323
16-3-1	数据的集中趋势 .....	323
16-3-2	数据的离散程度 .....	325
16-3-3	数据的统计 .....	328
16-4	使用基本图表认识数据 .....	331
16-4-1	绘制直方图 .....	331
16-4-2	绘制密度图 .....	334
16-4-3	在直方图内绘制密度图 .....	336
16-5	认识数据汇总函数 <code>summary()</code> .....	337
16-6	绘制箱形图 .....	338
16-7	数据的相关性分析 .....	341
16-7-1	<code>iris</code> 对象数据的相关性分析 .....	341
16-7-2	<code>stateUSA</code> 对象数据的相关性分析 .....	343
16-7-3	<code>crabs</code> 对象数据的相关性分析 .....	344
16-8	使用表格进行数据分析 .....	345
16-8-1	简单的表格分析与使用 .....	345
16-8-2	从无到有建立一个表格数据 .....	345



16-8-3	分别将矩阵与表格转成数据框	347
16-8-4	边际总和	347
16-8-5	计算数据的占比	348
16-8-6	计算行与列的数据占比	349
本章习题		350
<b>Chapter 17 正态分布</b>		
17-1	用直方图检验 crabs 对象	356
17-2	用直方图检验 beaver2 对象	357
17-3	用 QQ 图检验数据是否服从正态分布	359
17-4	shapiro.test () 函数	361
本章习题		363
<b>Chapter 18 数据分析——统计绘图</b>		
18-1	分类数据的图形描述	368
18-1-1	条形图与 barplot () 函数	368
18-1-2	圆饼图与 pie () 函数	371
18-2	量化数据的图形描述	372
18-2-1	点图与 dotchart () 函数	373
18-2-2	绘图函数 plot ()	376
18-3	在一个页面内绘制多张图表的应用	391
18-4	将数据图存盘	393
18-5	新建窗口	395
本章习题		397
<b>Chapter 19 再谈 R 的绘图功能</b>		
19-1	绘图的基本设置	404
19-1-1	绘图设备	404
19-1-2	绘图设置	407
19-1-3	layout () 函数的设置	418
19-2	高级绘图	421
19-2-1	曲线绘图 curve ()	421
19-2-2	绘图函数 coplot ()	423
19-2-3	3D 绘图函数	426
19-3	低级绘图——附加图形于已绘制完成的图形	429
19-3-1	points () 函数与 text () 函数	429

19-3-2	lines ()、arrows () 与 segments () 函数 .....	432
19-3-3	polygon () 函数绘制多边形 .....	434
19-3-4	abline () 直线、legend () 图例、title () 抬头与 axis () .....	438
19-4	交互式绘图 .....	443
	本章习题 .....	446

### Appendix A 下载和安装 R

A-1	下载 R 语言 .....	456
A-2	下载 RStudio .....	458

### Appendix B 使用 R 的补充说明

B-1	获得系统内建的数据集 .....	460
B-2	看到陌生的函数 .....	461
B-3	看到陌生的对象 .....	461
B-4	认识 CRAN .....	463
B-5	搜索扩展包 .....	463
B-6	安装与加载扩展包 .....	464
B-7	阅读扩展包的内容 .....	465
B-8	更新扩展包 .....	466
B-9	搜索系统目前的扩展包 .....	466
B-10	卸载扩展包 .....	467
B-11	R-Forge .....	467

### Appendix C 本书习题答案

### Appendix D 函数索引表

# 01

CHAPTER

## 基本概念

- 1-1 Big Data 的起源
- 1-2 R 语言之美
- 1-3 R 语言的起源
- 1-4 R 的运行环境
- 1-5 R 的扩展
- 1-6 本书的学习目标