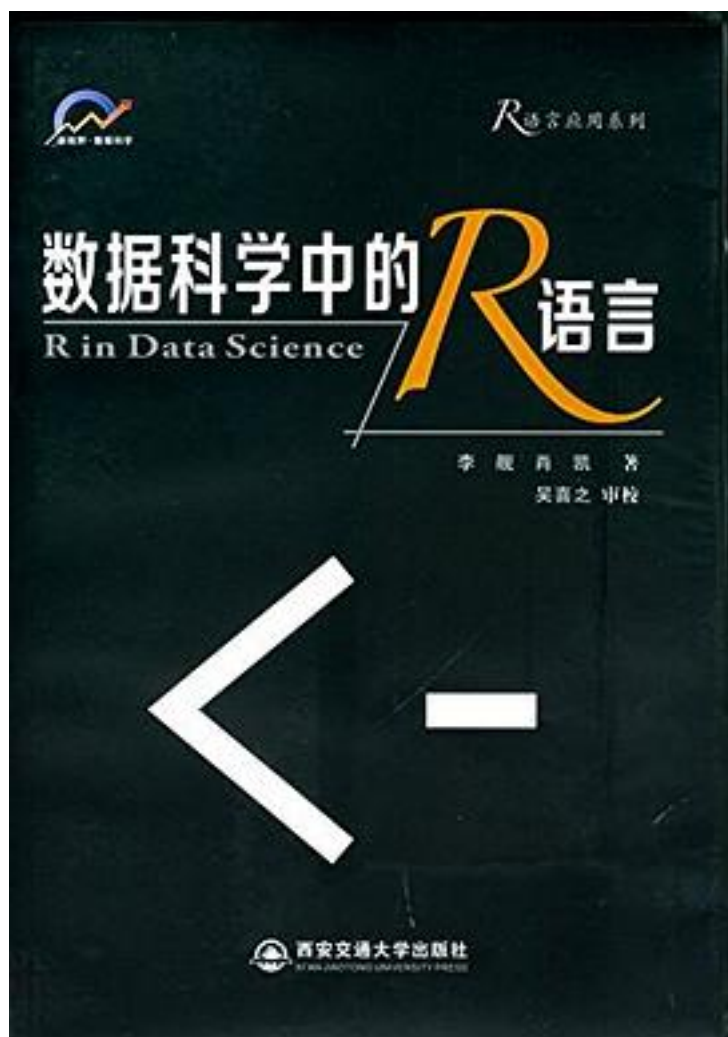


数据科学中的R语言



[数据科学中的R语言_下载链接1](#)

著者:李舰

出版者:西安交通大学出版社

出版时间:2015-7-3

装帧:平装

isbn:9787560570822

数据科学是将数据转化为行动的艺术，是综合了统计学、计算机科学和领域知识的新兴

学科。数据科学与很多传统学科的最大区别在于其为应用而生，因此本书一切从实际应用出发，以R语言为核心工具，介绍了各类分析方法的实现及其在各领域的应用情况。

如果是初学者，可以从“编程篇”开始阅读，了解R语言的基本特点。如果是程序员，在迅速掌握语法后，可以深入研究“模型篇”，学习如何用统计学和其他流行的分析方法来描绘复杂的数据世界。如果关心数据价值的落地，“应用篇”必不负你所望，金融、制药、文本挖掘、社交网络、地理信息、高性能计算、可重复研究，这些最热门的行业和应用领域的案例将会为你带来数据应用的最真实的体验。

本书适用于所有对数据应用感兴趣的读者，也适合作为统计和数据分析的实务操作型课程的教材。

作者介绍:

李舰，从2003年起，一直专注于数据科学在行业里的应用。擅长R语言的工程开发与分析建模，是 Rweibo、Rwordseg、tmcn等 R包的作者。现任堡力山（PMI）集团副总，曾任 Mango Solutions 中国区数据总监。

肖凯，一个喜欢折腾数据的人，资深R语言用户，参加过多次中国R语言社区大会的演讲。“数据科学中的R和Python”博客博主，目前供职于一号店商务智能部，任数据挖掘工程师。

目录: 序言

前言

第1章 引言：数据科学与R

1.1 数据科学简介

1.1.1 什么是数据科学？

1.1.2 如何成为数据科学家？

1.2 R 语言简介

1.2.1 什么是R？

1.2.2 如何学习R 语言？

1.2.3 R 的安装和配置

1.2.4 R 的常用编辑器

1.2.5 R 的第一步

1.3 如何使用本书？

1.3.1 排版和代码环境

1.3.2 测试环境

1.3.3 本书相关资源

第一部分 编程篇

第2章 数据对象

2.1 基本对象

2.1.1 向量入门

2.1.2 向量的生成

2.1.3 向量的计算

2.2 复合对象

2.2.1 矩阵

2.2.2 数据框

2.2.3 列表

2.3 特殊对象

2.3.1 缺失值与空值

2.3.2 连接

2.3.3 公式

- 2.3.4 表达式
- 2.3.5 环境
- 2.3.6 函数
- 第3章 数据操作
- 3.1 向量化操作
- 3.2 数据转换整理
- 3.2.1 取子集和编码转换
- 3.2.2 长宽格式互转
- 3.2.3 数据的拆分和合并
- 3.3 输入与输出
- 3.3.1 控制台的输入和输出
- 3.3.2 文本文件
- 3.3.3 表格型文件
- 3.3.4 其他外部文件
- 3.4 时间相关数据的处理
- 3.4.1 时间类数据处理
- 3.4.2 时间序列类数据
- 3.4.3 时间数据处理实例
- 第4章 控制语句与函数
- 4.1 控制语句
- 4.1.1 条件判断
- 4.1.2 循环
- 4.2 函数
- 4.3 函数式编程
- 4.4 工程开发的相关函数
- 4.4.1 程序调试
- 4.4.2 异常处理
- 第5章 面向对象
- 5.1 对象导论
- 5.1.1 面向对象的思想
- 5.1.2 面向对象编程的特性
- 5.1.3 R 的内置对象
- 5.2 S3
- 5.2.1 初识S3
- 5.2.2 面向对象的实现
- 5.3 S4
- 5.3.1 类的定义
- 5.3.2 对象的实例化
- 5.3.3 泛型函数和多态
- 5.4 引用对象
- 第二部分 模型篇
- 第6章 统计模型与回归分析
- 6.1 线性回归
- 6.1.1 回归模型和经典假设
- 6.1.2 参数估计
- 6.1.3 模型预测
- 6.1.4 离散自变量的情况
- 6.2 模型的诊断
- 6.2.1 非正态性
- 6.2.2 非线性
- 6.2.3 异方差
- 6.2.4 自相关
- 6.2.5 异常值
- 6.2.6 多重共线性

- 6.3 线性回归的扩展
 - 6.3.1 非线性回归
 - 6.3.2 非参数回归
 - 6.3.3 Logistic 回归
- 第7章 其他统计分析方法
 - 7.1 假设检验
 - 7.2 多元分析
 - 7.2.1 主成分分析
 - 7.2.2 对应分析
 - 7.2.3 多元分析的可视化
 - 7.3 时间序列
 - 7.4 随机模拟
 - 7.4.1 随机变量与分布
 - 7.4.2 蒙特卡洛方法
- 第8章 数据挖掘和机器学习
 - 8.1 一般挖掘流程
 - 8.2 聚类
 - 8.2.1 层次聚类
 - 8.2.2 K 均值聚类
 - 8.2.3 基于密度的聚类
 - 8.2.4 自组织映射
 - 8.3 分类
 - 8.3.1 决策树模型
 - 8.3.2 贝叶斯分类
 - 8.3.3 最近邻分类
 - 8.3.4 神经网络分类
 - 8.3.5 支持向量机分类
 - 8.3.6 集成学习与随机森林
- 第9章 最优化方法
 - 9.1 无约束非线性规划
 - 9.2 线性规划
 - 9.2.1 整数规划
 - 9.2.2 Rglpk 简介
 - 9.3 约束非线性规划
 - 9.4 遗传算法
- 第10章 数据可视化
 - 10.1 R 语言可视化简介
 - 10.1.1 什么是数据可视化
 - 10.1.2 R 语言的可视化环境
 - 10.1.3 ggplot2 入门
 - 10.2 分布的特征
 - 10.3 比例的构成
 - 10.4 时间的变化
 - 10.5 R 与交互可视化
- 第三部分应用篇
- 第11章 R在热门行业中的应用
 - 11.1 R 与金融分析
 - 11.1.1 金融数据获取和操作
 - 11.1.2 资产特征描述
 - 11.1.3 最优资产组合
 - 11.1.4 期权定价计算
 - 11.2 R 与新药研发
 - 11.2.1 新药研发简介
 - 11.2.2 药动学和药效学

- 11.2.3 建模和模拟
- 第12章 R与互联网文本挖掘
 - 12.1 网络数据获取
 - 12.1.1 XML 与XPath
 - 12.1.2 RCurl 抓取网页
 - 12.1.3 Rweibo 与OAuth
 - 12.2 中文文本处理
 - 12.2.1 文本处理
 - 12.2.2 正则表达式
 - 12.2.3 中文分词
 - 12.3 文本挖掘
 - 12.3.1 文本对象
 - 12.3.2 基本操作
 - 12.3.3 分析方法
- 第13章 大数据时代下的R
 - 13.1 地理信息数据
 - 13.1.1 空间数据对象
 - 13.1.2 R 与GIS 的结合
 - 13.1.3 互联网地理信息
 - 13.2 社交网络数据
 - 13.2.1 R 与网络数据
 - 13.2.2 R 与Gephi 的结合
 - 13.3 图像数据
 - 13.3.1 图像数据的处理
 - 13.3.2 图像识别
- 第14章 可重复的数据分析
 - 14.1 基于Sweave 的报告
 - 14.1.1 LATEX 与Sweave
 - 14.1.2 R 的Vignettes
 - 14.2 基于knitr 的报告
 - 14.2.1 Markdown 简介
 - 14.2.2 knitr 和LATEX
 - 14.2.3 报告中的图片
 - 14.2.4 xtable 与表格生成
 - 14.2.5 slidify 与幻灯片
 - 14.3 基于Office 的报告
 - 14.3.1 R2PPT
 - 14.3.2 ReporteRs
- 第15章 R与其他系统的交互
 - 15.1 R 与Excel
 - 15.1.1 安装DCOM 环境
 - 15.1.2 安装RExcel
 - 15.1.3 RExcel 的使用
 - 15.2 R 与数据库
 - 15.2.1 DBI 和RSQLite
 - 15.2.2 RODBC 简介
 - 15.3 R 与JAVA
 - 15.3.1 安装Java 环境
 - 15.3.2 Java 调用R
 - 15.3.3 R 调用Java
 - 15.4 R 与Microsoft Visual Studio
 - 15.4.1 R 与VB
 - 15.4.2 R 与C#
- 第16章 R与高性能运算

16.1 性能的度量与函数编译
16.2 代数运算库的优化
16.2.1 不同优化版本的实现
16.2.2 性能对比
16.3 超出内存的限制
16.3.1 内存管理机制
16.3.2 内存性能的优化
16.3.3 内存外运算
16.4 并行计算
16.4.1 Rmpi 与显式并行
16.4.2 parallel 包的应用
16.4.3 RHadoop 简介
后记
编后记
参考文献
索引
• • • • • ([收起](#))

[数据科学中的R语言_下载链接1](#)

标签

R

数据科学

数据分析

统计

数据挖掘

编程

统计学

大数据

评论

突出作者经验的章节是本书最有价值的部分。顺便缅怀一下自己也曾有过的使用R一手一脚摸索数据的年月

能看出两位作者的功底和积累

但看过几本类似的书之后就不由得需要想想这类书的动机或者说立意了

个人感觉从业者需要的不外乎两类内容（1）按照功能介绍各个R包的工具目录手册

（2）按照业界遇到的问题整体性全流程的演示操作

目前看到的书大部分都是两类各占一点儿

相对比较面向用户的就是吴喜之的《复杂数据统计方法》

希望未来这类书能更面向用户吧 而不是面向稿费或者行业知名度

结构设计很用心，编辑专业严谨，内容涉猎前沿、涵盖很广，文采飞扬，干货堆积，是目前中文R与数学科学书籍的典范。可以更加完美的地方有：在数据操作能加上tidyr+dplyr的组合，插图彩图化，R的工程化产品应用的例子再多一点。

数据科学是统计学、计算机科学、领域知识的综合，讲了各方面需要用到的技术，还涵盖了最优化方法（最近刚考完记忆犹新），R语言的用法介绍得也很全面

结束R，开始Python

解释了我好多的疑问。也可以当工具书使用。

初学者拿来入门很好。。。每章都是讲了比较实在的事儿，就是浅了些

一下午读了一半，这个作者写的不错，适合初学者

有些部分比如向量化计算是不错啦,但是时间序列那部份对我来讲有些浅了,还有数值模拟没有

很完善，把R中常用的不常用的系统的整理了出来

54. @04172016.
干货满满，很实在.作者是吴喜之老师的弟子，又有丰富的数据分析和R编程经验.

挑个错儿！ P43，2.2.3列表，“下面仍是用age、city、sex三个向量来构建列表”，明明一直在用城市和温度的数据做例子，哪来的年龄和性别？作者编着编着，编懵了？

应用篇有助于开眼界，介绍了很实用的软件集成和性能优化方法，很多已脱出数据科学，是软件工程的范畴，恰好补盲区。一阶半可以直接看后半本，不满一阶看前半本。最不满的是书品相不够好，皱巴巴的。

总感觉写的有点浅，覆盖的面倒是很全。我要是能把它当睡前读物就好了哈哈，前提得有睡前读书的习惯。

如获至宝。

R语言不错的教材

只能算是一般，写得太宽泛了，为了涉猎广，内容写得有点儿像蜻蜓点水的感觉，真的只能给你一个缥缈的轮廓~~~

中规中矩吧

模型的介绍还可以，但是R语言。。。生态不行啊

[数据科学中的R语言_下载链接1_](#)

书评

[数据科学中的R语言_下载链接1_](#)