

数量
经济
学
系列
丛书

R Programming with Applications to
Financial Quantitative Analysis

R软件

及其在金融定量分析中的应用

许启发 蒋翠侠 编著

清华大学出版社

Q
U
A
N
T
I
T
A
T
I
V
E

E
C
O
N
O
M
I
C
S

数量经济学系列丛书

R软件 及其在金融定量分析中的应用

许启发 蒋翠侠 编著

清华大学出版社
北京

内 容 简 介

金融定量分析主要以金融理论为指导,以数理方法为手段,以计算机软件为工具,分析金融系统中的各种数量关系,预测金融发展变动规律,为金融决策提供智力支持。本教材旨在阐明如何使用 R 软件开展金融定量分析,由三个部分组成:第一部分主要阐述 R 软件基础及基于 R 软件的计算等问题,为金融定量分析提供理论方法与计算工具准备;第二部分主要阐述基于 R 软件金融数据读取、整理以及金融收益计算等问题,为金融定量分析提供数据原材料;第三部分主要讨论了金融定量分析的核心内容并给出 R 软件的实现,包括:波动率估计、风险值计算、组合投资、资产定价、风险分散、羊群效应、微观金融等。本书配备了大量金融案例与 R 软件代码,可供读者直接使用或二次开发。

本教材可以作为金融学、统计学、数量经济学、金融数学等专业高年级本科生和相关领域研究生的教科书,也可以为相关领域的研究人员、大学老师、从业人员提供研究参考。

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 软件及其在金融定量分析中的应用/许启发,蒋翠侠编著.--北京:清华大学出版社,2015

(数量经济学系列丛书)

ISBN 978-7-302-39403-7

I. ①R… II. ①许… ②蒋… III. ①金融—定量分析—统计分析—应用软件 IV. ①F83-39

中国版本图书馆 CIP 数据核字 2015 第 031506 号



责任编辑:高晓蔚

封面设计:傅瑞学

责任校对:王凤芝

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

质 量 反 馈:010-62772015, zhiliang@tup.tsinghua.edu.cn

课 件 下 载: <http://www.tup.com.cn>, 010-62770175-4903

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:185mm×260mm 印 张:29.25 插 页:1 字 数:670 千字

(附光盘 1 张)

版 次:2015 年 5 月第 1 版

印 次:2015 年 5 月第 1 次印刷

印 数:1~4000

定 价:49.00 元

金融定量分析有助于深化金融市场运行规律的认识,是金融工程与金融风险管理的必备工具。随着理论研究深入与金融实践发展,开展金融定量分析所依赖的数理方法、计算工具与数据基础基本具备,市场也迫切需要大量的金融定量分析高级专门人才。无疑,一部优秀著作的出版,可以为此作出贡献。

如今,在金融定量分析领域,关于其数理基础、分析模型、计算软件等,出版了一批优秀的作品。特别是 R 软件的发展,掀起了使用 R 软件开展金融定量分析的热潮。迄今,国内尚无一本著作,介绍基于 R 软件的金融定量分析,相信本书的出版能够弥补这一缺憾。

本书由合肥工业大学管理学院许启发、蒋翠侠两位老师编著,旨在阐明如何使用 R 软件开展金融定量分析,可以作为金融学、统计学、数量经济学、金融数学等专业高年级本科生和相关领域研究生的教科书,也可以为相关领域的研究人员、大学老师、从业人员提供研究参考。

本书系统地介绍了金融定量分析的数理方法与模型,这些模型与方法涵盖了金融定量分析的核心内容,既包含经典又涉及前沿,便于读者对本领域研究全貌的把握。本书以问题为导向,通过案例分析,有效地将理论联系实际;通过鲜活的数据与定量计算,有助于读者认识现实金融市场的运行规律。本书提供了大量的 R 软件代码,既能够深化读者对金融定量分析理论与方法的掌握,又能够方便读者迅速掌握其计算技巧。

总之,本书将模型方法、案例分析、编程计算有机地融为一体,基于 R 软件细致地阐述了整个金融定量分析的全过程,包括:数据收集、数据整理、数据分析、数据存储等。本书集中了作者在本领域长期的研究成果与研究经验,许多 R 软件编程思路可供读者借鉴,更为可贵的是大量金融案例与计算代码可供读者直接使用,能够引领读者迅速成长为本领域的专业人士。

天津大学管理与经济学部

张世英

2014 年 10 月

随着互联网与大数据技术的发展,金融定量分析受到空前的关注,收益计算、组合投资、量化投资、资产定价、信用评估、风险分析等理论与方法都取得了巨大进展,在金融实践中发挥了越来越重要的作用。

金融定量分析主要以金融理论为指导,以数理方法为手段,以计算机软件为工具,分析金融系统中的各种数量关系,预测金融发展变动规律,为金融决策提供智力支持。在金融定量分析教学与研究中,需要解决三个方面的问题。第一,数理方法选择,主要涉及统计学、经济计量学、时间序列分析、运筹学、数据挖掘、机器学习等学科理论与方法。关于数理方法的介绍,要么散落于各个教程,要么并非针对金融定量分析展开。第二,计算工具使用,主要有 EViews、MATLAB、STATA、SAS、S-Plus 以及 R 软件等。R 软件由于其功能强大、更新及时、代码公开、免费获取等,受到越来越多用户的青睐,R 的出版物不断增加。迄今,国内尚无一本著作介绍基于 R 软件的金融定量分析。第三,金融数据获取。随着金融数据库建设、互联网平台发展,金融数据资源可获得性显著增加,基本能够满足金融定量分析的要求。

作者在金融定量分析领域开展过一定的前期研究,在 R 软件使用方面已有 7 个年头,觉得有必要将这方面的工作总结出版。本书写作过程以金融定量分析中涉及的问题为导向,阐述数理方法、进行案例分析、给出 R 软件代码。在 R 软件代码编写过程中,尽量使用 R 软件中自带的包,在没有 R 包的情况下,我们自定义 R 函数解决。本书具有以下 5 个显著特点。

(1) 在数理方法介绍方面,注重深入浅出。在第 2 章与第 3 章分别介绍了金融定量分析的数理基础,包括统计学、经济计量学、时间序列分析、优化理论与方法、人工智能、高维数据分析等,旨在介绍基本理论与方法,为进一步地给出金融定量分析模型作铺垫。在第 6 章至第 12 章,阐述了直接用于金融定量分析的各种模型,包括: GARCH 模型、SV 模型、金融风险模型、组合投资决策模型、资本资产定价模型、协整与协同持续、神经网络分位数回归模型等,注重阐明其提出背景、基本思路、模型表示、求解算法等,而不特别强调其理论推导过程。

(2) 在金融案例选择方面,强调案例实效。以金融定量分析中涉及的问题为导向,进行案例选择,涵盖了收益计算、组合投资、资产定价、风险计量、风险传染、羊群效应、基金风格、基金绩效、信用评估等宏观与微观金融定量分析的核心内容。这些案例设计与安排,既有金融定量分析中的经典之作,也有作者的最新研究成果,可以在丰富教学内容的同时,也为开展相关研究工作提供一个借鉴。整个案例分析强调理论联系实际,从问题提出、数据搜集、软件处理、结果输出、结果解读等环节展开,便于读者理解与模仿。

(3) 在数据选取使用方面,体现现实针对性。本书除一小部分数据来源于必要的数值模拟外,绝大部分数据都来源于实际,并非虚构。这些实际数据主要来源于金融数据库、互联网资源等渠道,体现了金融市场最新活动规律,具有现实针对性。使用这些数据,一方面能够帮助读者熟悉与掌握金融定量分析的基本原理与方法,把握 R 软件进行金融定量计算的技巧;另一方面能够帮助读者认识现实金融市场的运行规律,达到研究与解决实际问题的目的。

(4) 在计算程序编写方面,突出思路清晰。本书配备了大量的 R 软件计算代码,粗略合计为 6600 行、字节量近 8M。这些代码,有一部分已经在本书中予以解释说明,还有一部分在本书的附配光盘中,需要读者调阅。在 R 软件代码编写过程中,强调结构化编程思路,按照功能模块组织程序,每一函数只实现一个功能模块。R 软件代码的编写,融入了作者多年的编程经验,也包含了最新研究进展,便于读者直接使用或二次开发。

(5) 在网络资源提供方面,力求全面有效。全书共提供了两类网络资源,一类为数据网络资源,不仅列举了重要的金融数据库和网站信息,而且通过 R 软件代码演示了如何从不同的金融数据库和金融网站中读取金融数据资料;另一类为 R 软件网络资源,较为全面地介绍了与金融定量方法相对应的 R 网站、R 包和 R 函数。能够方便读者了解金融数据来源、定量方法选择、编程计算实现等,迅速掌握从数据搜集到定量计算的最佳途径。

本书在结构安排上,可以划分为三个部分:第一部分由第 1 章至第 3 章组成,主要阐述 R 软件基础及基于 R 软件的计算等问题,为金融定量分析提供理论方法与计算工具作准备;第二部分由第 4 章、第 5 章组成,主要阐述基于 R 软件的金融数据读取、整理以及金融收益计算等问题,为金融定量分析提供数据原材料;第三部分由第 6 章至第 12 章组成,讨论了金融定量分析的核心内容并给出 R 软件的实现,包括:波动率估计、风险值计算、组合投资、资产定价、风险分散、羊群效应、微观金融等。

本书章节编写安排为:第 1 章至第 6 章由许启发教授编写;第 7 章由许启发教授及硕士生陈士俊共同编写;第 8 章、第 9 章、第 12 章由蒋翠侠副教授编写;第 10 章由康宁讲师编写;第 11 章由博士生刘曦编写;许启发教授对全书进行了统稿。另外,感谢王长久、张金秀、徐金菊、刘玉叶、周莹莹、邓锴、张可新、俞奕涵、郭志峰、卓杏轩、蔡超、李辉艳等研究生同学,在素材搜集、图形制作、代码编写、数据处理、书稿校对等方面付出了大量劳动。本书的编写过程,参阅了大量的文献资料,在参考文献中予以列出,对这些文献的作者表示衷心的感谢!本书的编写与出版,还要衷心感谢清华大学出版社高晓蔚老师的大力帮助与支持!

由于作者水平有限,书中肯定存在不足与纰漏,敬请广大读者不吝指正。欢迎大家将意见与建议发送到邮箱: xuqifa1975@126.com。我们将在适当时机予以改进。

作者

2014 年 10 月于合肥工业大学



| | |
|------------------------|----|
| 第 1 章 R 软件基础 | 1 |
| 1.1 工作环境 | 1 |
| 1.1.1 R 的历史与发展 | 1 |
| 1.1.2 R 的资源 | 2 |
| 1.1.3 RGui | 4 |
| 1.1.4 RStudio | 6 |
| 1.2 数据操作 | 8 |
| 1.2.1 对象 | 8 |
| 1.2.2 基本类型 | 9 |
| 1.2.3 向量 | 10 |
| 1.2.4 数组与矩阵 | 11 |
| 1.2.5 列表与数据框 | 15 |
| 1.2.6 因子 | 19 |
| 1.2.7 表达式 | 20 |
| 1.2.8 对象的运算 | 21 |
| 1.3 常用命令 | 28 |
| 1.3.1 工作目录与 R 内存 | 28 |
| 1.3.2 保存与加载 | 29 |
| 1.3.3 显示命令 | 30 |
| 1.3.4 挂接命令 | 31 |
| 1.4 图形制作 | 32 |
| 1.4.1 绘图函数 | 32 |
| 1.4.2 绘图参数 | 36 |
| 1.4.3 制图案例 | 37 |
| 1.5 编程计算 | 41 |
| 1.5.1 函数定义 | 41 |
| 1.5.2 函数调用 | 41 |
| 1.5.3 函数调试 | 42 |
| 1.6 常用程序包 | 43 |
| 1.6.1 标准包 | 43 |

| | | |
|--------------|---------------------------|------------|
| 1.6.2 | 安装包 | 44 |
| 1.6.3 | 常用包 | 47 |
| 1.7 | 习题 | 62 |
| 1.8 | 参考文献 | 62 |
| 第 2 章 | 基于 R 软件的传统计算 | 64 |
| 2.1 | 统计分析 | 64 |
| 2.1.1 | 多元回归分析 | 64 |
| 2.1.2 | 逐步回归分析 | 72 |
| 2.1.3 | 聚类分析 | 76 |
| 2.1.4 | 因子分析 | 81 |
| 2.2 | 经济计量分析 | 87 |
| 2.2.1 | 数据测量层次 | 87 |
| 2.2.2 | 二元选择模型 | 88 |
| 2.2.3 | 计数数据模型 | 90 |
| 2.2.4 | 广义线性模型 | 91 |
| 2.3 | 时间序列分析 | 93 |
| 2.3.1 | ARMA 模型 | 94 |
| 2.3.2 | VAR 模型 | 96 |
| 2.3.3 | 脉冲响应 | 97 |
| 2.3.4 | 方差分解 | 98 |
| 2.3.5 | Granger 因果 | 98 |
| 2.3.6 | 案例分析 | 99 |
| 2.4 | 优化理论与方法 | 103 |
| 2.4.1 | 问题提出 | 103 |
| 2.4.2 | 线性规划 | 104 |
| 2.4.3 | 目标规划 | 105 |
| 2.4.4 | 非线性规划 | 106 |
| 2.5 | 习题 | 112 |
| 2.6 | 参考文献 | 112 |
| 第 3 章 | 基于 R 软件的现代计算 | 113 |
| 3.1 | 人工智能方法 | 113 |
| 3.1.1 | 人工神经网络 | 113 |
| 3.1.2 | 支持向量机 | 120 |
| 3.2 | 高维数据分析 | 125 |
| 3.2.1 | 问题提出 | 126 |
| 3.2.2 | LASSO 回归 | 127 |
| 3.3 | 习题 | 133 |

| | |
|-----------------------------|------------|
| 3.4 参考文献 | 133 |
| 第4章 金融数据整理与预处理 | 135 |
| 4.1 金融数据库 | 135 |
| 4.1.1 金融数据与金融数据库 | 135 |
| 4.1.2 国外金融数据库概况 | 135 |
| 4.1.3 国内金融数据库概况 | 137 |
| 4.1.4 金融数据库数据主要内容 | 138 |
| 4.2 金融数据格式 | 140 |
| 4.2.1 xls、xlsx 格式 | 140 |
| 4.2.2 csv 格式 | 140 |
| 4.2.3 txt 格式 | 141 |
| 4.2.4 XML 格式 | 141 |
| 4.2.5 HTML 格式 | 141 |
| 4.2.6 从其他统计软件导入 | 141 |
| 4.2.7 关系型数据库 | 142 |
| 4.2.8 DBF 格式 | 142 |
| 4.3 金融数据的导入 | 142 |
| 4.3.1 从控制台输入数据 | 143 |
| 4.3.2 上市公司财务报表信息读取 | 145 |
| 4.3.3 股票数据的读取 | 146 |
| 4.4 金融数据的预处理 | 147 |
| 4.4.1 时间序列数据预处理 | 147 |
| 4.4.2 截面数据预处理 | 151 |
| 4.5 习题 | 154 |
| 4.6 参考文献 | 155 |
| 第5章 金融资产收益计算 | 156 |
| 5.1 收益率定义 | 156 |
| 5.1.1 常用收益率 | 156 |
| 5.1.2 红利收益率 | 157 |
| 5.1.3 超额收益率 | 157 |
| 5.2 股票类资产收益率计算 | 158 |
| 5.2.1 单个股票收益率计算 | 158 |
| 5.2.2 多个股票收益率计算 | 159 |
| 5.2.3 资产组合收益率计算 | 159 |
| 5.3 债券类资产收益率计算 | 160 |
| 5.3.1 三种收益计算 | 160 |
| 5.3.2 债券久期与凸度计算 | 163 |

| | | |
|--------------|------------------------------|------------|
| 5.3.3 | 债券绩效评价 | 165 |
| 5.4 | 收益率的分布及其特征 | 166 |
| 5.4.1 | 分布函数与数字特征 | 166 |
| 5.4.2 | 常用分布函数 | 169 |
| 5.4.3 | 多元收益率统计 | 172 |
| 5.5 | 习题 | 175 |
| 5.6 | 参考文献 | 176 |
| 第 6 章 | 金融波动模型 | 177 |
| 6.1 | GARCH 类模型 | 177 |
| 6.1.1 | ARCH 模型 | 178 |
| 6.1.2 | GARCH 模型 | 180 |
| 6.1.3 | GARCH 模型扩展 | 187 |
| 6.1.4 | 多元 GARCH 模型 | 189 |
| 6.2 | SV 类模型 | 194 |
| 6.2.1 | 基本 SV 模型 | 194 |
| 6.2.2 | 扩展 SV 模型 | 196 |
| 6.2.3 | 多元 SV 模型 | 197 |
| 6.2.4 | 案例分析 | 198 |
| 6.3 | 高频波动模型 | 205 |
| 6.3.1 | 金融高频数据及其特征 | 205 |
| 6.3.2 | “已实现”方差模型 | 207 |
| 6.3.3 | ACD 模型 | 210 |
| 6.3.4 | 案例分析 | 212 |
| 6.4 | 习题 | 228 |
| 6.5 | 参考文献 | 228 |
| 第 7 章 | 极值、分位数与 VaR(ES) | 230 |
| 7.1 | VaR 与 ES 的计算 | 230 |
| 7.1.1 | VaR | 231 |
| 7.1.2 | ES | 232 |
| 7.1.3 | RiskMetrics 模型与 VaR 和 ES 的计算 | 233 |
| 7.1.4 | GARCH 模型与 VaR 和 ES 的计算 | 235 |
| 7.2 | 分位数回归与 VaR(ES) 计算 | 238 |
| 7.2.1 | 线性分位数回归 | 239 |
| 7.2.2 | 非线性分位数回归 | 244 |
| 7.2.3 | 基于分位数回归的 VaR 和 ES 的计算 | 249 |
| 7.3 | VaR(ES) 的极值方法 | 250 |
| 7.3.1 | 区间极大值模型 | 250 |

| | |
|--------------------------------|------------|
| 7.3.2 阈值模型····· | 255 |
| 7.4 习题····· | 260 |
| 7.5 参考文献····· | 260 |
| 第8章 金融组合投资决策分析····· | 262 |
| 8.1 均值-方差分析····· | 262 |
| 8.1.1 变量及其含义····· | 262 |
| 8.1.2 均值-方差模型····· | 264 |
| 8.2 均值-VaR(CVaR、CDaR)模型····· | 277 |
| 8.2.1 均值-VaR 模型····· | 278 |
| 8.2.2 均值-CVaR 模型····· | 280 |
| 8.2.3 均值-CDaR 模型····· | 285 |
| 8.3 均值-高阶矩模型····· | 292 |
| 8.3.1 高阶矩风险及其计算····· | 292 |
| 8.3.2 基于 M-V-S-K 的组合投资选择····· | 293 |
| 8.4 大规模组合投资决策模型····· | 299 |
| 8.4.1 两个重要模型····· | 299 |
| 8.4.2 模型求解····· | 299 |
| 8.4.3 数值模拟····· | 300 |
| 8.4.4 R 包与案例分析····· | 301 |
| 8.5 习题····· | 308 |
| 8.6 参考文献····· | 308 |
| 第9章 金融资产定价分析····· | 310 |
| 9.1 CAPM 及其应用····· | 310 |
| 9.1.1 标准的 CAPM 模型与分散化投资····· | 310 |
| 9.1.2 高阶矩 CAPM····· | 316 |
| 9.2 APT 及其应用····· | 323 |
| 9.2.1 因子模型····· | 323 |
| 9.2.2 APT 模型····· | 327 |
| 9.3 期权定价模型及其应用····· | 333 |
| 9.3.1 布朗运动与维纳过程····· | 333 |
| 9.3.2 期权定价原理····· | 336 |
| 9.3.3 二叉树期权定价····· | 338 |
| 9.3.4 B-S 期权定价····· | 340 |
| 9.3.5 隐含波动与波动微笑····· | 347 |
| 9.4 习题····· | 349 |
| 9.5 参考文献····· | 350 |

| | |
|---------------------------------|-----|
| 第 10 章 金融风险共同趋势分析 | 351 |
| 10.1 收益序列间的共同趋势 | 351 |
| 10.1.1 单位根检验 | 352 |
| 10.1.2 协整与误差校正模型 | 362 |
| 10.1.3 单方程协整关系的估计与检验 | 365 |
| 10.1.4 系统方程协整关系的估计与检验 | 370 |
| 10.2 风险序列间的共同趋势 | 379 |
| 10.2.1 协同持续建模 | 379 |
| 10.2.2 案例分析 | 381 |
| 10.3 习题 | 385 |
| 10.4 参考文献 | 385 |
| 第 11 章 金融市场羊群效应 | 387 |
| 11.1 基于均值回归羊群效应分析 | 387 |
| 11.1.1 参数模型 | 388 |
| 11.1.2 非参数均值模型 | 389 |
| 11.1.3 案例分析 | 390 |
| 11.2 基于分位数回归羊群效应分析 | 394 |
| 11.2.1 参数分位数模型 | 394 |
| 11.2.2 非参数分位数模型 | 395 |
| 11.2.3 案例分析 | 396 |
| 11.3 基于神经网络分位数回归羊群效应分析 | 400 |
| 11.3.1 模型表示 | 400 |
| 11.3.2 模型估计 | 401 |
| 11.3.3 模型选择 | 402 |
| 11.3.4 R 包 QRNN 主要函数 | 402 |
| 11.3.5 案例分析 | 403 |
| 11.4 习题 | 406 |
| 11.5 参考文献 | 406 |
| 第 12 章 微观金融定量分析 | 407 |
| 12.1 破产概率预测 | 407 |
| 12.1.1 模型与方法 | 407 |
| 12.1.2 案例分析 | 410 |
| 12.2 证券投资基金风格分析 | 414 |
| 12.2.1 证券投资基金概述及投资风格分析方法 | 414 |
| 12.2.2 基于均值回归的证券投资基金风格分析 | 415 |
| 12.2.3 基于分位数回归的证券投资基金风格分析 | 416 |

| | |
|----------------------------------|-----|
| 12.2.4 案例分析 | 417 |
| 12.3 证券投资基金绩效评价 | 424 |
| 12.3.1 证券投资基金绩效评价概述 | 425 |
| 12.3.2 证券投资基金的绩效评价指标与方法 | 426 |
| 12.3.3 基金经理人的绩效评价 | 436 |
| 12.4 基于 Rhadoop 的大数据金融定量分析 | 437 |
| 12.4.1 大数据 | 437 |
| 12.4.2 Rhadoop 平台部署及初步应用 | 442 |
| 12.5 习题 | 451 |
| 12.6 参考文献 | 451 |

R 软件基础

本章重点介绍 R 软件的工作环境、基本操作、常用命令、图形绘制与编程计算等,最后介绍了 R 软件常用的软件包。

1.1 工作环境

R 软件是一个有着强大的数据分析、统计计算以及图形展示功能的优秀的开源软件,是一种简捷方便且功能强大的计算机编程语言。R 软件具有有效的数据处理和数据存储机制,擅长数组和数据矩阵操作,提供了全面完整的数据分析中间工具,支持各种数据的可视化、多样化输出,拥有简洁高效、全面完整的彻底面向对象的统计语言。此外,R 软件还提供了与其他编程语言和数据库之间的接口,具备了良好的可扩展性、技术传承性与共同协作性。R 软件拥有丰富的网站、书刊、社区及会议交流等资源,为 R 软件开发者及使用者提供了一个完全免费、自由开放、高度灵活、适用范围广泛、行业支持丰富的统计工具。

R 软件的一个显著优点是其营造了一个富有弹性、灵活互动的数学计算环境,具备了多种统计学及数据分析处理功能,提供了大量的 R 软件包与计算函数。R 的所有功能均是通过 R 软件包来实现的,函数均内嵌在软件包之中,开发者和使用者可以通过下载安装软件包来进行统计计算,增加新功能。通过众多 R 志愿者的贡献,使得涉及学界和业界各个领域、各个学科的前沿理论方法都能迅速地转换为 R 计算,其相应的计算机程序都会以 R 软件包的形式得以实现,并存储在 R 官网、R 的 CRAN(the comprehensive R archive network)社区、R 的开发者平台以及遍布世界各地的镜像之中,以供 R 的开发者 and 使用者下载使用。可以毫不夸张地说,如果你掌握了 R 软件,你将与世界最优秀的统计学家、计量经济学家同步,可以开展最前沿的研究工作。

1.1.1 R 的历史与发展

R 在建立之初只是一个普通的统计学教学软件,是由新西兰奥克兰大学的统计系教授 Robert Gentleman 和 Ross Ihaka 于 1993 年根据 AT&T 贝尔实验室的 Rick Becker、John Chambers 和 Allan Wilks 所开发的 S 语言为规范编写而成,因而 R 与 S 语言极为相似,可视为是 S 语言的一种实现或形式。因为共同创立了 R 的 Robert Gentleman 和 Ross Ihaka 两位教授的名字的首字母都是 R,故将该种语言称为 R。1995 年,R 开始启动开源模式,这意味着 R 将采用社区合作的方式进行软件开发,R 的所有源代码都可以从不同的渠道免费获取,从而便于对源代码的检验与利用。同年,随着 R 开发进程的不断推进,出现了 R 程序

版本的归档问题,维也纳工业大学的 Kurt Hornik 通过在维也纳建立一个 R 程序的归档解决了该问题,同时在全世界各地也出现了 R 程序的镜像。用户通过 R 帮助邮件列表可以获取免费的关于 R 的帮助及建议的信息,通过可搜索的邮件列表文档可以得到关于 R 邮件列表的更多详细信息,因而有效克服了开源软件缺少支持的主要缺陷。1997 年,S 语言的创始人之一 Chambers 也加入 R 的创建队伍之中,使得 R 的发展速度得以快速提升。与此同时,在瑞士苏黎世理工学院 Martin Machler 教授的帮助和推动下,R 正式成为自由软件基金会旗下的 GNU 项目开源软件,使得后继使用者能够在前人所编写的工具包代码的基础上进行拷贝、模仿、修改与创新。这些工作突出了 R 的技术传承优势,提高了创作效率,扩展了全球化合作模式,加快了 R 的发展速度。现在,R 是由一个 R 核心开发小组对其进行开发、维护与更新,该小组由十几个来自世界各地不同机构与单位的成员组成。他们本着自愿主动、努力工作、认真负责的原则与态度致力于 R 的管理工作,使得用户可以通过世界各地的 CRAN 镜像下载 R 资源,方便快捷地获取最新信息和使用说明,分享使用全球优秀的统计成果。

如今,R 越发受到 statisticians 们的青睐,得到学术界与业界人士的认可,俨然成为数据处理最重要的工具之一,成为统计学科工具软件之首。首先,R 已被广泛应用于统计学、经济学、金融学、社会学、物理学、生物学等众多学科以及财经、气象、制药、环境保护等众多领域的经济计量、财经分析、人文科学研究、人工智能研发以及用户行为分析等方面的统计计算和可视化分析。其次,R 与著名的 Oracle 公司、Revolution Analytics 公司以及 JMP、Mathematica、MATLAB、EViews、SPSS、STATISTICA 和 SAS 等商业统计软件均存在数据接口,这些软件也纷纷向 R 提供各种服务,使得 R 的计算与作图功能得以不断强大,操作与使用平台得以拓宽。

截至 2014 年 7 月,R 所拥有的软件包数目已经超过 5700 个,既有官方注册开发者,也有非官方注册开发者。2012 年,根据 KD Nuggets 所进行的“使用何种编程或统计类语言进行分析和数据挖掘”以及“过去 12 个月中你在实际项目中使用的数据挖掘或分析工具”两个问卷调查结果均表明,R 的使用率在所有的分析工具中独占鳌头,超过了 SQL、JAVA、SAS、Excel 等统计分析工具。2013 年,在 TIOBE 发布的编程软件排行榜中,R 位居第 24 位,在统计类软件中位次最高。与此同时,Rexer Analytics 5th 数据挖掘者调查报告结果显示,R 稳居工具类首位且一直保持着强劲的上升势头。

1.1.2 R 的资源

R 发展至今,已经形成了层次丰富、形式多样、来源广泛的资源信息体系,主要包括 R 开源平台、R 官网及社区、R 出版物、R 会议、R 学习交流网站以及 R 搜索引擎与 R 博客等。

1.1.2.1 R 开源平台

R 的开源平台使得使用者可以自由获取源代码,可以自主运行、修改、拷贝和重新发布源代码,与此同时,通过采用补丁开发与合并的开发模式还能够保持源代码的完整性。在 R 开源平台之中,个人、团体与特定软件均可自由获取与使用该资源,进而扩大了 R 的使用范围、增加了 R 的使用用户。

1.1.2.2 R 官网及社区

从 R 的官网(<http://www.r-project.org>)中可以获得 R 的核心开发小组所发布的实时信息,可以了解 R 的更新,数据包发布,相关图书、期刊、通讯及会议等各方面的信息,可以下载 R 的安装程序、各种外挂程序包、相关 R 文档、源代码及其他资源。同时,R 使用者还可以通过 R 主页享受预定邮件服务,众多的 R 志愿者们贡献了他们大量的时间来回答 R 用户的问题,将解答邮件发往 R 邮件列表之中并进行存档,旨在通过电子邮件的渠道为用户提供帮助。R 的使用与软件包密切相关,R 的 CRAN(网址 <http://cran.r-project.org>)镜像网站上有数量庞大的软件包,这些软件包均由世界各地的 R 志愿者们所提供,涉及各个学科领域,覆盖面极其广泛,使用者可以免费下载使用。

1.1.2.3 R 出版物

R 出版物主要包括 R 刊物及 R 书籍,R 的官方电子杂志是 *The R Journal* (原名 *R News*),该杂志主要介绍了 R 的最新特征、CRAN 最新动态以及关于各个 R 软件包的介绍文件、帮助性文件以及相应的开发技术文件等,使得 R 使用者可以更加方便、快捷、简单、高效地使用 R。此外,R 还拥有大量的专业出版物,其中包括 J STAT SOFTW 等 120 多部的英文版本出版物,涉及各个学科领域的 R 应用。随着 R 的不断发展,也涌现出了大量关于 R 的参考书籍,以帮助读者学习和使用 R。R 的项目网站上包含了大量的与 R 相关的参考书籍,用户可以免费下载。如介绍许多操作技巧、有助于初学者入门的 *An Introduction to R*^[1] 以及被视为快速上手指南和操作参考书的 *R in a Nutshell* 等书籍均是学习 R 的重要资源。一些著作介绍基于 R 的统计分析、经济计量分析、金融定量分析,如: Dalgaard (2008)^[2]、Kleiber 等(2008)^[3]、Daróczy 等(2013)^[4]等。国内也有一些非常优秀的教材介绍 R 的使用,与统计与计量密切相关的有:陈毅恒等(2006)^[5]、汤银才(2008)^[6]、薛毅等(2007)^[7]等。通过这些出版物对 R 进行推广与普及,使得 R 开发者和使用者可以更加深入了解 R、学习 R、使用 R,促进了 R 的完善与发展。

迄今,国内尚无一本著作,专门介绍基于 R 软件的金融定量分析,但愿本书能够弥补这一缺憾。

1.1.2.4 R 会议

R 会议组织的主要目的是为 R 的开发者与使用者提供一个面对面交流的机会,通过这样一个场合,使得使用者更深入地了解 R,开发者也更懂得用户的需求与意见。目前有许多围绕 R 召开的会议,其中最具有影响力的 R 会议包括“ues! 会议”与“DSC 会议”。“ues! 会议”是由 R 官方机构举办的年度会议,由 R 开发者和使用者共同参与,旨在促进两者间的沟通与交流。“DSC 会议”只在奇数年份召开,并且只有 R 的开发者参与,重点讨论 R 在统计计算方面的应用问题。在国内,由中国人民大学发起的“中国 R 软件会议”对 R 在学界及业界的普及与应用起着积极的促进作用。

1.1.2.5 R 学习交流网站

为了便于 R 开发者和使用者获得 R 的相关信息,了解 R 的相关动态,分享彼此之间的

使用心得与进行彼此之间的经验交流,各国都组织 R 学习交流站,其中 R 官方网站所承认列举的 R 学习交流站就有 20 多个,比较知名的有 www.cookbook-r.com、R Video Tutorial (Dan Goldstein) 以及 R Podcasts (Andrew Jackson) 等。在国内,最初由中国人民大学统计系谢益辉组建的统计之都 (<http://cos.name>) 是重要的 R 学习交流网站,通过一大批志愿者的共同努力为 R 开发者及使用者提供了一个可以互相交流、互相合作的自由开放平台。

1.1.2.6 R 的搜索引擎

除了 R 网站、社区外,还可以通过 R 专用搜索引擎、博客等网络途径来获取 R 资源。R 专用搜索引擎是指通过一些链接网址和浏览器在网络上搜索与 R 相关的帮助信息。R 的主要搜索引擎包括 Rseek (<http://rseek.org>), 其利用 Google 搜索引擎的强大功能专注于搜索与 R 相关的网站信息,并将搜索结果进行分类整理,组织成一个非常直观简便的页面显示形式,进而在有效避免无用信息泛滥的同时提高信息搜索效率; Stack Overflow 即 Q&A (<http://stackoverflow.com>), 这是一个具有搜索功能的 Q&A 网站,也是一个面向问题的网站,用户通过在该网站上提问,就会得到其他不同用户的回答,用户再通过对这些答案进行评价投票,从而使得最佳答案随着时间的推移而逐渐移至顶部,进而形成一个功能强大的 Q&A 对话数据库,旨在快捷高效地解决用户所提出的问题。用户在 Q&A 网站上所提的问题侧重于 R 的程序设计、代码编写、数据结构以及图形绘制等方面; Stack Exchange (<http://stats.stackexchange.com>), 也是面向问题的搜索网站,该网站的一个专门的统计分析 (Statistical Analysis) 区是具有搜索功能的 Q&A 网站,相对于 Stack Overflow 网站着重于编程问题,该网站偏向于统计问题的提问。

1.1.2.7 R 博客

获取 R 资源的另一有效渠道是阅读关于 R 的博客,通过阅读 R 博客可以关注关于 R 及其数据的讨论,可以学习 R 的基础知识,可以了解 R 的发展前沿,可以同最新研究成果保持同步。R 博客的数量之大、覆盖面之广、信息量之多使之成为 R 用户的学习乐园。比较知名的 R 博客有: 由 Tal Galili 建立的 R-bloggers (<http://www.r-bloggers.com>)、Planet R 群组 (<http://www.planet-r.org>) 以及中文 R 语言资源聚合平台——R 客 (<http://r-ke.info>) 等。这些博客中收录了各地的优秀 R 技术博客,并提供相应的全文链接、R 相关事件与新闻链接以及 R 其他资源链接。通过订阅相应的 RSS 摘要,得以和其他站点共享 R 的相关信息,使用户可以从众多的网站上获得自己感兴趣的有用信息。

R 的这些资源平台是相辅相成、互为补充的,通过连贯的运行机制使之成为一个动态的有机整体,为 R 开发者及使用者提供一个丰富的资源使用环境。

1.1.3 RGui

RGui (R graphic user's interface) 是 R 核心开发团队发布的一个官方版的 R 软件,R 的开发者与使用者可以从 R 的官方网站 (<http://www.r-project.org/index.html>) 免费下载,安装使用,RGui 可以在 Linux、Mac OS X、Windows 三个平台上运行,用户可根据实际需要下载与自己使用的平台相对应的版本。RGui 可通过输入命令语句或程序代码以实现数据集的建立、数据的分析以及统计作图等操作。RGui 的界面比较简单,由操作界面、输出界面