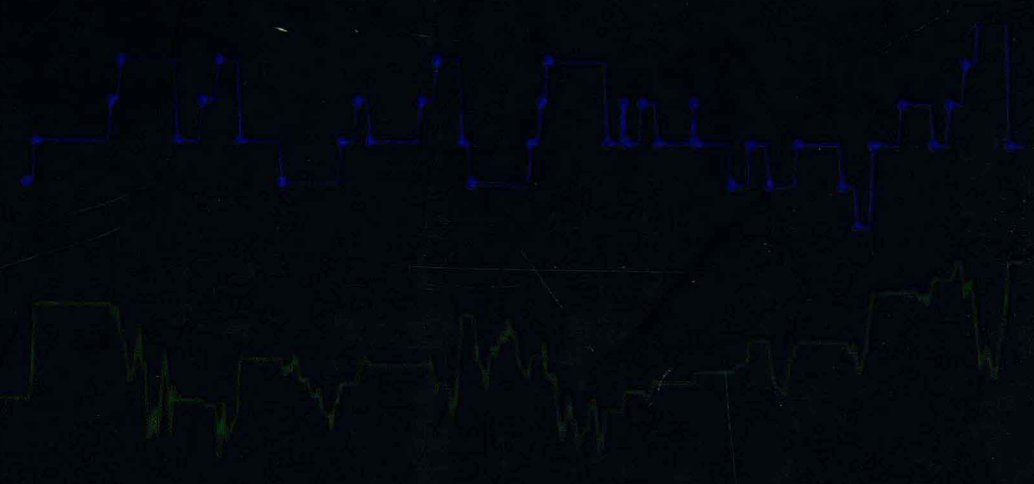


不能善用R的宽客，不是真正的宽客

量化投资

朱晓斌 著

以R为工具



R是一个具有强大的统计分析和绘图功能的软件系统

R是完全免费的软件，并由全世界的统计学家共同维护

R易学易懂易操作，资源丰富，是宽客入门级的必备利器



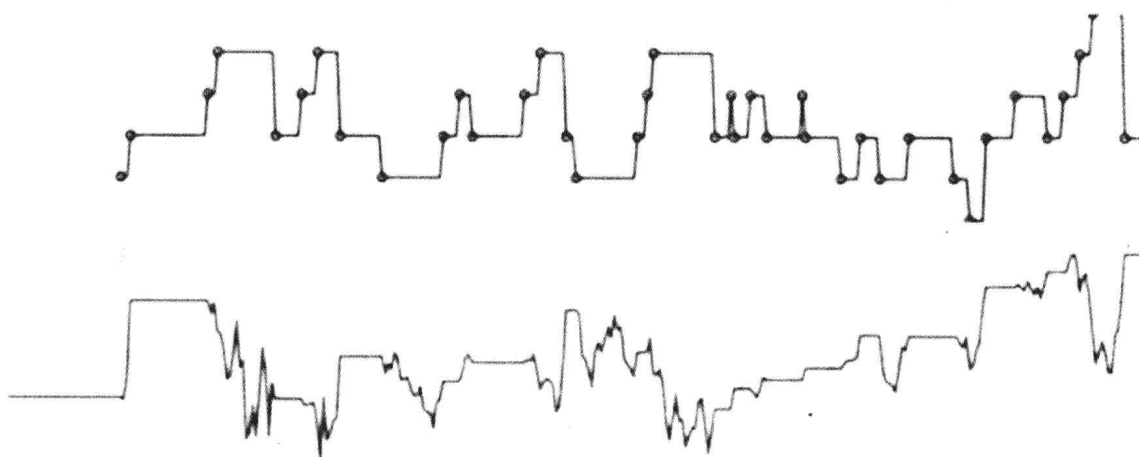
中国电力出版社
CHINA ELECTRIC POWER PRESS

不能善用R的宽客，不是真正的宽客

量化投资

朱晓斌 著

以 **R** 为工具



内 容 提 要

本书包含了相关的量化投资技术的理论、模型和思想，并利用 R 众多程序包中的内置函数，将抽象的金融模型通过 R 的数据处理和图形形式进行解释、验证和求解，旨在使读者既熟悉当前量化投资技术的理论背景，又能够熟练使用 R 处理量化投资过程中的定量计算与分析问题。

本书既可作为数学、统计、金融、经济、信息等专业本科生及研究生的一本跨学科的参考书，也可作为金融机构和经济研究部门工作人员及对投资原理感兴趣的读者的实用指南。

图书在版编目 (CIP) 数据

量化投资：以 R 为工具/朱晓斌著. —北京：中国电力出版社，2012.9
ISBN 978-7-5123-3527-1

I. ①量… II. ①朱… III. ①投资—研究 IV. ①F830.59

中国版本图书馆 CIP 数据核字 (2012) 第 225758 号

中国电力出版社出版、发行

北京市东城区北京站西街 19 号 100005 <http://www.cepp.sgcc.com.cn>

责任编辑：李耀东

责任校对：朱丽芳 责任印制：邹树群

航远印刷有限公司印刷·各地新华书店经售

2013 年 1 月第 1 版·2013 年 1 月北京第 1 次印刷

787mm×1092mm 16 开本·31.25 印张·592 千字

定价：68.00 元

敬告读者

本书封底贴有防伪标签，刮开涂层可查询真伪

本书如有印装质量问题，我社发行部负责退换

版 权 专 有 翻 印 必 究

前言

对于很多投资者来说，量化投资是比较专业的词汇，一想到量化投资，就会联想到数学模型，或者是西蒙斯^❶的神秘“黑箱子”。其实投资者在日常也会接触到量化，每个投资者都有自己的偏好，比如喜欢在投资股票的时候，去关注 MACD、KDJ 等指标，认为对投资有一定的帮助和指引。实际上这就是投资者从最浅显的角度接触统计概率——量化投资技术的基础。

如果你对眼下各种层出不穷的名词概念和天花乱坠的宣传口号感到困惑，或者有意进入数量化投资领域（尤其是金融、经济、统计、计算机等专业的学生），那么本书正是为你们所写。当然，如果你已经或者即将投身到这一在发达市场已经风生水起而在国内才刚刚起步的新兴行业，那么，希望本书也能给你带来实际的帮助，它不是各种量化投资理论模型的堆砌，而是力求将实用的量化投资技术与 R 实现紧密联系，不仅介绍了各种量化投资和交易算法的实际背景、统计模型和计算方法，还结合 R 给出了相应的解决问题的步骤，并对计算结果进行了详细分析。

证券投资，从其决策和执行过程的角度看，大致可以分为定性投资和定量投资两类。前者是以深入的基本面分析研究为核心基础，辅以对上市公司的调研，和管理层的交流，及各类研究报告。其组合决策过程是基金经理在综合了所有信息后，凭主观判断及直觉来精选个股，构建组合，以产生超额收益。沃伦·巴菲特是定性投资方法的专家，而他的老师本杰明·格雷厄姆所著的《证券分析——原理和技巧》一书则被价值投资者奉为圭臬。量化投资则是将投资思想通过具体指标、参数的设计体现到具体的模型中，让模型对市场进行不带任何情绪的跟踪。这种跟踪使得基金投资的广度和深度都得到了很大的拓展。同时，由于避免了人为因素的干扰，量化投资的另一个优势是可以在很大程度上保证投资业绩具备可持续性和可复制性。量化方法更关注“数字”背后的意义，依靠计算机的帮助，分析数据中的统计特征，从而挖掘出内在的价值。

从 1971 年巴克莱投资管理公司发行世界上第一只指数基金至今，定量投资在

❶ 詹姆斯·西蒙斯是量化投资领域的传奇人物，其管理的大奖章基金，从 1989 年到 2006 年的 17 年间，平均年收益率达到了 38.5%，远超巴菲特的 20%。他的投资经历被总结为用公式打败市场，而其投资技术（那些模型和策略）则被人们誉为赚钱的“黑箱子”。

海外的发展已有 40 多年。1970 年定量投资在海外全部投资中占比为零，而到 2009 年，定量投资在全部投资中占 30% 以上。其中指数类投资几乎全部使用定量技术，主动投资中大约有 20% ~ 30% 使用定量技术。从 2000 年到 2007 年，美国定量投资总规模翻了 4 倍多，与此相比，美国共同基金总规模（定性与定量之和）仅翻了 1.5 倍。在我国证券市场，过去 20 年对基本面的研究占据了市场的主流地位，然而随着我国资本市场市值的迅速扩大、上市公司数量的急剧增加，以及 QDII 陆续出海，如何在众多的境内外上市公司中迅速、有效地选择投资目标，降低调研和投资的成本，更科学地分配规模庞大的资产，成为机构投资者面对的新问题。在此背景下，量化投资目前在基金、保险资产、QFII、QDII 等机构投资者中的应用大大增加，在基本面投资的基础上应用数量化策略正在成为投资经理共同关心的问题。

量化投资的基础是计量经济学^❶，后者是统计学和经济学的交叉学科，它把各种数理统计方法应用于经济数据，以分析各个经济变量之间的关联关系，验证经济学所提出和构造的数理模型，并得到各种数值结果。

今天，市场上关于计量经济和数理统计的书已非常多，这类书主要是以数理统计的理论为基础，但对于计算涉及的较少，基本是以手工计算为主，目的是为了帮助读者理解相应的统计方法，可操作性不强。关于统计计算的书也不少。目前，这类书一般是讲算法（这一点与数值分析或计算方法差不多），而没有相应的软件做支撑，有些内容是数值分析内容的重复，统计味不足。

结合软件讲统计的书，目前最多的是结合 SAS、SPSS 或 EViews，这些庞大而昂贵的软件提供了菜单式的人机界面以使操作“傻瓜化”，副作用则是在数值计算和输出方面给用户制造了一个新的“黑箱子”。这类书基本上是软件的使用说明书，虽然谈到一些统计概念，但涉及的不多。

作为一款优秀的统计软件，“R”同时也是一种数学计算环境。它最初由奥克兰大学统计学系的 Ross Ihaka 和 Robert Gentleman 编写，自 1997 年开始由一个核心团队开发，这个团队的成员大部分来自大学机构，除了这些作者之外，R 还拥有一大批贡献者，他们为 R 编写代码、修正程序缺陷和撰写文档。迄今为止，R 中的程序包已经是数以千计，各种统计前沿理论方法的相应计算机程序都会在短时间内以软件包的形式得以实现，这种速度是其他统计软件无法比拟的。除此之外，R 还有一个重要的特点，那就是它是免费的、开源的！

投资艺术正在向投资科学演变，这个演变正在悄然发生并将持续下去。演变的方向虽然明确，但节奏在不断变化。随着越来越多新一代科学型投资经理人的成长，

❶ “计量经济学” (Econometrics) 一词，是挪威经济学家弗里希 (R. Frisch) 在 1926 年仿照“生物计量学”一词提出的。“计量”的意思是“以统计方法做定量研究”，所以“量”字应读作“亮”，而不读作“良”。

他们将更多地依靠科学的分析、程序及结构化而不是凭借直觉、忠告和奇思怪想。当然，这并不意味着超凡的个人投资洞察力会过时，而是意味着投资经理们可以用更系统的方式获取和运用他们的天赋。

本书既不是单纯的关于数理统计或计量经济的教科书，也不只是一本关于 R 软件的使用手册，而是一本将两者相结合并着重讲述它们在数量投资领域具体应用的参考书。它包含了相关的量化投资技术的理论、模型和思想，并利用 R 众多程序包中的内置函数，将抽象的金融模型通过 R 的数据处理和图形形式进行解释、验证和求解，旨在使读者既熟悉当前量化投资技术的理论背景，又能够熟练使用 R 处理量化投资过程中的定量计算与分析问题。既可作为数学、统计、金融、经济、信息等专业本科生及研究生的一本跨学科的参考书，也可作为金融机构和经济研究部门工作人员以及对投资原理感兴趣的读者的实用指南。

最后，相信本书能激发起读者们对量化投资这一新兴领域的浓厚兴趣。

好吧，让我们现在就开始。

排版说明

一、印刷格式

斜体 (*Italic*) 用来表示公式中的变量、URL、文件名、路径名以及 R 包名称。

等宽字体 (**Constant Width**) 用来表示文本中的变量名、R 函数名 (本质上也是变量)。

等宽黑体 (**Constant Width Bold**) 用来突出代码的某些部分。

二、R 程序代码中的记号

表示一段 R 程序的开始。在 R 中可以用 `options(prompt = '>')` 设置程序行的起始符号，

1. “>” 默认为 “>”；后文中只要遇到这个标记，则说明该标记之后的程序语句可以直接在 R 中运行（读者在运行书中的范例时不要把这个符号也写入命令行中）。
2. “+” 续行符。当一段程序在某一行中没有完整显示出来时，就会折到下一行，此时 R 会以 “+” 表示程序语句上不完整，正在继续。
3. “[n]” 其中 “n” 表示一个整数，中括号括上一个整数表示 R 程序输出的行号，如 [1] 表示这是第 1 行输出。
4. “#” 表示 R 程序注释，即不会被执行的语句（只是为了增强程序的可读性而做的“标记”）。

并非本书中所有的 R 程序代码都使用了 “>”、“+” 和 “[n]” 符号，它们是交互模式下的提示符，通常出现在：（1）代码行数较少，一般少于 5 行；（2）代码中包含了命令的输出内容。

目录

前言

第 1 章 R 简介	1
1.1 R 是什么	1
1.2 交互模式和编写脚本	2
1.3 安装 R	4
1.4 获取帮助	4
1.5 包	5
1.5.1 标准包	6
1.5.2 捐献包和 CRAN	6
1.5.3 命名空间	6
第 2 章 与数据打交道	7
2.1 简单的算术操作和向量运算	7
2.1.1 向量和赋值	7
2.1.2 向量运算	8
2.1.3 生成正则序列	9
2.1.4 逻辑向量	10
2.1.5 缺损值	11
2.1.6 字符向量	12
2.1.7 向量索引：选择和修改一个数据的子集	13
2.2 其他重要的数据类型	14
2.2.1 数组和矩阵	14
2.2.2 列表和数据框	15

2.2.3	类和泛型函数	16
2.2.4	S4 类	19
2.3	与日期打交道	20
2.4	合并数据框	22
2.5	从文件中读取数据	22
2.5.1	<code>read.table()</code> 函数	23
2.5.2	从 Excel 文件中读取	24
2.6	处理超大数据文件	24
2.6.1	用 <code>scan()</code> 读取数据字段	25
2.6.2	利用 Unix 工具	25
2.6.3	用磁盘代替内存	26
2.6.4	访问数据库	26
2.7	发布系统指令	27
2.8	从剪贴板读取数据	28
2.9	直接编辑数据	28
2.10	例子：从大文件中提取数据	29

第 3 章 横截面回归..... 30

3.1	最小二乘法	30
3.2	从回归中提取统计量	32
3.3	异方差	33
3.3.1	异方差 <i>Breusch-Pagan</i> 检验	33
3.3.2	异方差（自回归）稳健性协方差矩阵	33
3.4	线性假设检验（Wald 和 F）	33
3.5	加权和广义最小二乘法	34
3.6	带因子 / 分组的模型	35

第 4 章 特殊回归..... 37

4.1	固定 / 随机效应模型	37
-----	-------------------	----

4.1.1	固定效应	37
4.1.2	随机效应	38
4.2	定性响应	39
4.2.1	Logit/Probit.....	39
4.2.2	多项式 Logit.....	39
4.2.3	顺序 Logit/Probit.....	40
4.3	Tobit 和删失 (Censored) 回归.....	40
4.4	分位数回归	40
4.5	稳健回归 -M 估计.....	41
4.6	非线性最小二乘法	41
4.7	单结构方程的两步回归	42
4.8	方程组	42
4.8.1	似无关回归	42
4.8.2	方程组两步回归	43
4.9	例子	43
4.9.1	蒙特卡洛 (Monte Carlo) 模拟	43
4.9.2	最大似然估计	44

第 5 章 时间序列回归 45

5.1	差分和滞后	45
5.2	滤波	47
5.2.1	Canned AR 和 MA filters	47
5.2.2	手动滤波	48
5.2.3	Hordrick-Prescott 滤波.....	48
5.2.4	卡尔曼 (Kalman) 滤波.....	48
5.3	ARIMA/ARFIMA	49
5.4	ARCH/GARCH.....	50
5.4.1	基本 GARCH—garch().....	50
5.4.2	高级 GARCH—garchFit()	51

5.5	相关图	51
5.6	预测值	52
5.7	时间序列检验	53
5.7.1	自相关 <i>Durbin-Watson</i> 检验	53
5.7.2	自相关 <i>Box-Pierce</i> 和 <i>Breush-Godfrey</i> 检验	53
5.7.3	单位根 <i>Dickey-Fuller</i> 检验	53
5.8	向量自回归 (VAR)	54
5.9	谱分析	54
5.9.1	傅里叶变换	54
5.9.2	窗函数	55
5.9.3	周期图	56
5.10	小波	60
5.11	例子: 时间序列分解和预测	64
5.11.1	时间序列分解	64
5.11.2	时间序列拟合	65

第 6 章 绘图 67

6.1	绘制经验分布	68
6.2	等高图	69
6.3	添加图例和资料	71
6.4	添加箭头、文本和标记	72
6.5	多图绘制	72
6.6	保存图形	75
6.7	添加希腊字母和数学符号	75
6.8	其他图形包	76

第 7 章 统计 77

7.1	常见的统计分布	77
-----	---------------	----

7.2	检验一个数据集的分布	78
7.3	单样本和双样本检验	82
7.4	P 值	86
7.5	从数据中采样	86

第 8 章 R 中的数学 88

8.1	矩阵运算	88
8.1.1	矩阵相乘	88
8.1.2	线性方程和求逆	89
8.1.3	特征值和特征向量	89
8.1.4	奇异值分解和行列式	90
8.1.5	最小二乘法拟合和 QR 分解	90
8.2	数值优化	91
8.2.1	无约束极小化	91
8.2.2	线性约束极小化	92
8.3	数值积分	93

第 9 章 编程 94

9.1	编写函数	94
9.1.1	一个简单的例子	94
9.1.2	定义新的二元操作符	95
9.1.3	参数命名和默认值	96
9.1.4	... 参数	96
9.1.5	在函数中赋值	97
9.2	循环和避免循环	97
9.3	控制	100
9.3.1	条件控制: if 语句	100
9.3.2	循环控制: for 循环, repeat 和 while	100

9.3.3 一个例子	101
9.4 检测你的程序运行时间	102
9.5 调试	103
9.5.1 用 <code>traceback</code> 显示错误命令	103
9.5.2 运用 <code>debug</code> 一步步排查	104
9.5.3 运用 <code>trace</code> 在函数中插入代码	107
9.5.4 通过 <code>recover</code> 浏览函数调用情况	108

第 10 章 改变配置 111

第 11 章 保存你的工作 114

第 12 章 投资中的数量化技术 117

第 13 章 估值和选股 119

13.1 相对估值模型	120
13.1.1 股票价格倍数	120
13.1.2 企业价值倍数	121
13.2 绝对估值模型	123
13.2.1 现金流贴现模型	123
13.2.2 红利贴现模型	124
13.2.3 股权自由现金流模型	124
13.2.4 公司自由现金流模型	125
13.2.5 经济增加值 (EVA) 贴现法	125
13.2.6 现金流量贴现法的指标计算	126
13.3 数量化选股	127
13.3.1 简介	127
13.3.2 基本面 (Foundmental) 选股	128
13.3.3 多因素 (Multiple-Factor Regression) 模型	131
13.3.4 动量反转 (Momentum Contrarian) 选股	133
13.3.5 GARP 模型	134

第 14 章 资产配置和组合优化 136

- 14.1 战略资产配置 137
 - 14.1.1 马科维茨资产配置模型 137
 - 14.1.2 均值 -LPM 资产配置模型 139
 - 14.1.3 VaR 约束下的资产配置模型 140
 - 14.1.4 基于贝叶斯估计的资产配置模型 141
- 14.2 战术资产配置 142
 - 14.2.1 行业轮动策略 143
 - 14.2.2 风格轮动策略 144
 - 14.2.3 (可转移) Alpha 策略 146
 - 14.2.4 投资组合保险策略 147

第 15 章 基于数据挖掘的股票预测 150

- 15.1 神经网络 151
- 15.2 支持向量机 153
- 15.3 小波分析 156

第 16 章 行为金融指导下的投资策略 157

- 16.1 反向投资策略与动量交易策略 158
- 16.2 捕捉并集中投资策略 159
- 16.3 小盘股策略 160
- 16.4 成本平均策略 162
- 16.5 时间分散化策略 162
- 16.6 设立止损点的交易策略 163

第 17 章 程序化交易和算法交易 165

- 17.1 程序化交易 166

17.1.1	指数套利交易策略	167
17.1.2	数量程序交易策略	167
17.1.3	动态对冲交易策略	168
17.1.4	配对交易策略	168
17.1.5	久期平均交易策略	169
17.2	算法交易	170
17.2.1	定义	170
17.2.2	主要算法	171
17.3	交易成本分析	174
17.4	交易模型和模拟	176

第 18 章 绩效评估技术..... 178

18.1	风险调整收益分析	178
18.1.1	夏普比率	179
18.1.2	特雷纳指数	179
18.1.3	詹森指数	180
18.1.4	M2 指数	181
18.1.5	信息比率	182
18.1.6	RAROC 指标	182
18.2	投资管理人的能力分析	183
18.2.1	经验模型	183
18.2.2	T-M 和 H-M 模型	183
18.3	业绩持续性分析	185
18.3.1	回归系数法	185
18.3.2	列联表法	185
18.4	超额收益归因分析	187
18.4.1	资产配置超额收益率贡献	187
18.4.2	证券选择收益贡献	187
18.4.3	行业选择的收益贡献	187

18.4.4	行业内个股选择收益贡献	187
18.5	基金评价体系	188
18.5.1	晨星公司的评级体系	188
18.5.2	银河基金业绩评价体系	189

第 19 章 实践：几个量化投资模拟的例子 191

19.1	预备：获取数据	191
19.1.1	从网络获取股票数据	191
19.1.2	从数据库获取股票数据	192
19.1.3	获取和查看上市公司的财务报表	198
19.1.4	获取上市公司股息数据	198
19.1.5	根据股息调整股票价格	199
19.1.6	获取股票分割数据	199
19.2	用 LOWESS 平滑价格曲线	199
19.3	交易前成本分析	203
19.4	布林带策略绩效分析	215
19.5	海龟交易系统	218
19.6	配对交易策略绩效分析	227
19.7	数据挖掘预测股票收益	234

第 20 章 相关程序包参考 251

20.1	TTR.....	251
20.2	BLOTTER	294
20.3	QUNTSTRAT	314
20.4	PERFORMANCEANALYTICS	344

附录一 R 中实现交易模拟的工具链 463

附录二 CRAN 任务视图：计量经济学 464

附录三 CRAN 任务视图：实证金融学 468

附录四 CRAN 任务视图：时间序列分析 472

附录五 CRAN 任务视图：机器学习和统计学习 478

参考文献..... 482