

作者是R语言社区领袖，金融和大数据领域跨界专家，英文版和繁体版将在美国和中国台湾发行

从金融投资学理论、R语言数据处理、量化投资策略3个维度讲解如何将R语言技术应用于金融市场的实操，填补领域空白

数据分析与决策

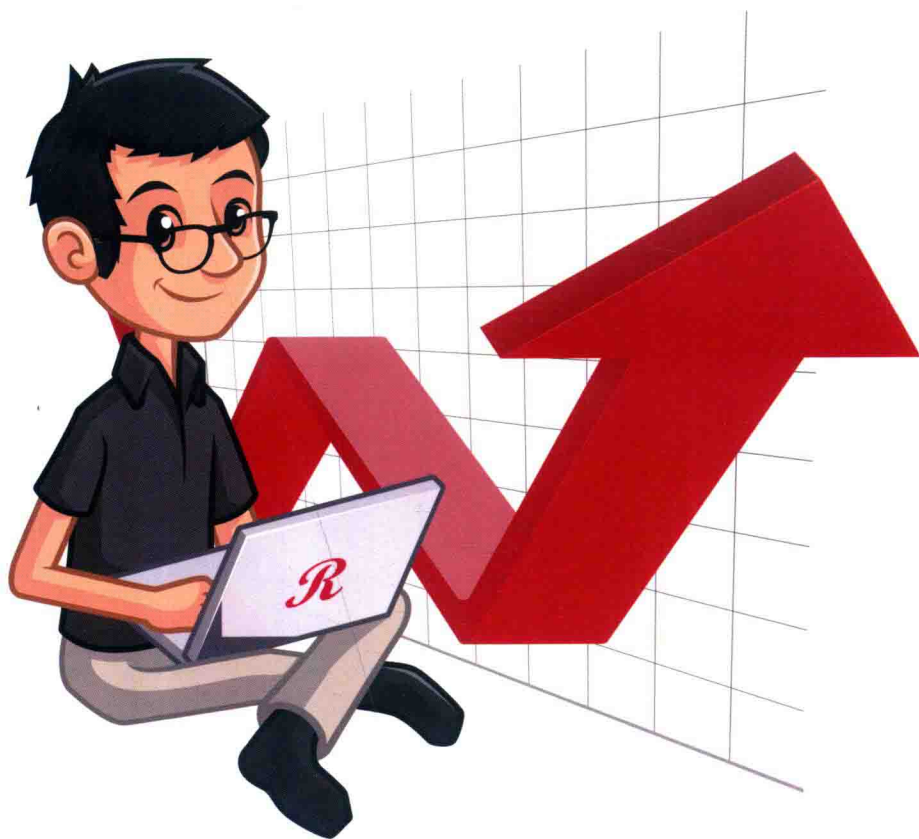
技术丛书

R for Programmers: Quantitative Investment

R的极客理想

量化投资篇

张丹◎著



机械工业出版社
China Machine Press

R for Programmers: Quantitative Investment

R的极客理想

量化投资篇

张丹◎著



机械工业出版社
China Machine Press

图书在版编目 (CIP) 数据

R 的极客理想——量化投资篇 / 张丹著. —北京: 机械工业出版社, 2017.11
(数据分析与决策技术丛书)

ISBN 978-7-111-58297-7

I. R… II. 张… III. 程序语言—程序设计 IV. TP312

中国版本图书馆 CIP 数据核字 (2017) 第 258815 号

R 的极客理想——量化投资篇

出版发行: 机械工业出版社 (北京市西城区百万庄大街 22 号 邮政编码: 100037)

责任编辑: 何欣阳

责任校对: 殷虹

印 刷: 三河市宏图印务有限公司

版 次: 2018 年 1 月第 1 版第 1 次印刷

开 本: 186mm×240mm 1/16

印 张: 22

书 号: ISBN 978-7-111-58297-7

定 价: 79.00 元

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) 88379426 88361066

投稿热线: (010) 88379604

购书热线: (010) 68326294 88379649 68995259

读者信箱: hzit@hzbook.com

版权所有·侵权必究

封底无防伪标均为盗版

本书法律顾问: 北京大成律师事务所 韩光/邹晓东

华章IT
HZBOOKS | Information Technology



这是一本以中国金融二级市场为背景，理论与实战紧密结合的书籍，实用性非常强，是量化投资必读之书。

我是通过阅读张丹的博客——“R 的极客理想”的系列文章而与他结缘。精炼的 R 语言编程风格，让我看到了一个极客对于技术的不懈追求。我买了他的两本书，《R 的极客理想——工具篇》和《R 的极客理想——高级开发篇》，后来我们进行了邮件交流，并通过电话深入地讨论了一些量化交易的问题。随后一直保持联系，也在筹划一起做全球市场的量化交易。我可以感觉到他是一个有想法的年轻人，愿意和他人分享知识，并且有非常扎实的跨学科知识积累和中国金融市场的交易经验。

本书涵盖了几个主题，包括金融市场与金融理论、R 语言数据处理与高性能计算、金融策略实战、量化投资策略案例。

以下是我认为对大家非常有帮助的知识点：

- R 语言数据处理和调用 C++ 的章节，对于提升 R 的性能是非常有用的。
- Docker 架构现在在金融业非常受欢迎，书中介绍的 Docker 的用法，为应用程序的自动化部署提供了解决方案。
- 可转债交易监控系统，是使用 R 语言的一个很好的例子，它提供了一个原型的建模方法，按照这个思路可以扩展到对一般债券交易建立模型。
- 均衡回归和追涨杀跌的量化交易模型，是量化交易的关键要素，每个人都应该知道它的原理。
- 基金会系统的设计和实现，对于资金运作和持仓管理非常有用。

坦白说，在一本书中，能够提供如此多解决实际问题的方法，是很少见的。我也从这本书中学到了很多东西，强烈建议每个人都把这本书放在自己的书架上，作为量化投资必读之书。

陈琪龙 华盛顿大学博士

铨智金融科技合伙人 (www.quanffett.com)

序 二 *Foreword*

作为金融统计学科的老师，经常有学生兴冲冲地跑上门来问我如何通过金融统计赚钱，我总是回答他们：如果我会这个，我早就辞职发财去了！于是大家哈哈一笑。

学习金融、统计、数学、计量能不能赚钱？我想大概是能的，否则也不会有文艺复兴这样的公司和 James Simons 这样的大神了。

本书到底可以带给读者什么呢？是知识，堪比真金白银的知识。

首先，是金融量化的基础知识。你可以从本书第一部分学习到金融量化的基础知识和金融市场的交易规则。书中深入浅出地剖析了资本资产的定价模型，并将 R 语言与金融市场相结合，对于金融初学者来说，是入门的精品书籍。

其次，是掌握并运用工具的知识。不管是想通过数据分析发财的，还是像我这样准备写几篇论文的，要进入这一领域，最基本的要求就是能够“玩得动”数据。金融数据以海量著称，动辄就是 GB、TB 的数据量。如何高效地处理大规模的数据，就成了数据分析从业人员的基本功。本书第二部分，就是针对“效率”而作。第 3 章对 R 语言中常见的数据操作进行整理和归纳，得到一套高效数据处理方法。第 4 章讲解了如何通过使用第三方软件大幅提升 R 语言的计算效率。至于 R 语言的专业知识，请参见作者的另外两本著作《R 的极客理想——工具篇》和《R 的极客理想——高级开发篇》。

最后，让技术落地，通过市场来检验投资方法的正确性。本书第三部分，教读者如何编写交易策略，并让市场来检验交易策略的有效性。跟着书中的 4 个实践案例来操作，相信读者很快就能掌握金融建模的方法，以及对应的 R 语言代码。之后，你就可以大胆地试验自己的投资想法是否有效了。

黄达

复旦大学管理学院统计学系

为什么要写这本书

本书是《R 的极客理想》系列丛书的第三本，是将 R 语言与金融量化投资相结合的一本书。本书主要的写作目的是把 R 语言的技术和实际的金融量化案例结合起来，让读者能亲身体会如何把知识变成真正的生产力。

传统的交易员凭借多年的交易训练，每日人工盯盘，观察市场的变化。一个好的交易员，可以同时观测几个金融市场的几十个交易品种。随着金融产品的发展，中国市场 A 股股票已经达到 3000 多支，债券有 7000 多支，公募基金接近 4000 支，还有多种金融衍生品。如此大量的金融产品，已经不能依靠个人之力消化和分析了。

通过计算机对全市场进行扫描，从而发现不合理的定价和交易机会，这样可以极大地提高交易员的工作效率。一种理想化的设计是让程序来为我们赚钱，而我们就可以去做自己喜欢的事情了。也就是说，让技术变现，解放我们的生活。

本书中的原创观点和方法，都是基于理论研究和实践的成果。实际上，长久以来我也在找这样一本书，能够把书本上的理论模型与实际业务相结合，但并没有找到，或者并没有符合中国市场的实际案例应用，所以只能自己动手写一本。本书有点像自己的笔记，我也会经常翻看，让自己的头脑始终保持思路清晰。

本书的主要特色

本书撰写的主要思路是从 IT 人的角度，通过技术来切入金融市场，进行量化投资。发挥 IT 人专注学习、乐于分享的精神，借助互联网快速传播知识，打破传统的金融壁垒。发挥“极客”的创造力，让知识变成生产力，让更多有理想的 IT 人，能够有机会进入金融行业，推动金融行业的改革和创新。

但这不是一本简单易懂的书，因为量化投资是跨学科领域。你需要有多学科的知识储备，才能胜任量化投资的工作。要理解和掌握本书的相关内容，可能需要多本书籍的相关知识做支撑。

阅读本书，不但需要你有 R 语言的使用经验，更需要有对金融市场知识的理解。本书主要介绍了三部分内容，涵盖金融市场、统计知识和 IT 技术。金融市场部分，包括中国金融二级市场环境的介绍、交易工具的使用、金融产品的交易规则、国内机构的投研思路和策略，以及基金会计等金融行业的基础知识。统计知识部分，包括时间序列、一元线性回归、多元线性回归、自回归等统计和计量的模型算法。IT 技术部分，就是 R 语言相关的编程技术、金融量化程序包的使用、金融数据处理、金融数据模型的构建、量化策略的实现思路、R 语言代码的编写等。

本书的一大特色是使用了很多真实案例，以中国的金融市场为背景，让你可以体会到市场所带来的波动，国家宏观政策对市场的影响，散户思维与专业投资者的差异，量化思路与主观思路对于市场的不同理解。

本书是我在实际投资研究中的总结，从金融理论模型、市场特征检验，到数学公式、R 语言建模，再到历史数据回测、会计资产核算，最后进行实盘交易。通过 R 语言，可以很简单地实现我们的投资想法。类似的投资想法其实谁都有，利用 IT 人的技术优势，可以真正地将想法与实际操作结合起来。

本书所涉及的金融产品包括股票、期货、债券、基金、现金管理等，跨越多个金融市场的多种金融标的物；涉及的交易模型和交易策略，有基于市场技术指标的量价策略，有基于统计理论的套利策略，有基于金融产品规则的事件性策略，有针对全市场扫描的选股策略，也有高频交易的择时策略。相信本书会令你感受到金融市场的魅力，以及技术优势能给我们带来的价值。

要想深入理解本书的内容，你可能需要像我一样，不仅有技术的积累，还要真正地去金融市场实践，多与行业内的人进行沟通，不断地学习和思考。

本书读者对象

本书适合以下所有 R 语言工作者：

- ☐ 有计算机背景的软件工程师
- ☐ 数据分析背景的数据科学家
- ☐ 金融行业从业者，如券商研究员、分析师、基金经理、宽客 (Quant)
- ☐ 金融、统计、数据科学、计算机专业的学生

如何阅读本书

全书一共 6 章，分为三个部分：金融市场与金融理论、R 语言数据处理与高性能计算、

金融策略实战，书中每一章都是一个大的知识体系。

第一部分是金融市场与金融理论（第1章和第2章），从了解金融开始，建立对金融量化认识的基本思路。第1章为全书开篇，主要介绍了利用R语言做量化投资的思路和方法。第2章主要介绍了金融经典理论模型和R语言的实现方法。

第二部分是R语言数据处理与高性能计算（第3章和第4章），详细介绍了利用R语言进行数据处理的必备工具和使用方法。第3章以R语言数据处理技术为核心，介绍了用R语言处理各种类型的数据的方法。第4章讲解了如何通过3种外部技术来让R语言的性能达到生产环境的要求。

第三部分是金融策略实战（第5章和第6章），结合R语言技术和金融市场规则，解决金融量化领域的实际问题。第5章让读者了解低风险的投资市场和投资方法，第6章介绍了完整的从理论到实践的投资研究方法。

本书有很多综合运用的知识，建议顺序阅读全部章节。书中的一些技术实现，用到了笔者前两本书《R的极客理想——高级开发篇》和《R的极客理想——工具篇》中的知识，有兴趣的读者可以一起阅读。

勘误和资源

由于水平有限，加之编写时间仓促，书中难免会出现一些错误或者不准确的地方，恳请读者批评和指正。为此，我为这系列书创建了网站（<http://fens.me/book>），方便与读者进行沟通。书中的全部源代码都可以从网站下载，如果在阅读过程中遇到问题，也可以在网站中留言，我将尽量为你提供满意的解答。如果你有任何意见和建议，欢迎发邮件至 bsspirit@gmail.com，期待能够得到你的反馈。

致谢

感谢在我最失意的时候，帮助我度过难关的朋友——北京千庄智金科技有限责任公司总经理张颂、量子金服CEO刘亚非、民生银行同事许斌。

感谢所有R语言的读者，以及社区的各位朋友，让我们通过R语言认识，并一起把知识传播下去。

感谢天善智能创始人梁勇为本书提供的赞助和推广。同时，感谢铨智金融科技合伙人陈琪龙博士和复旦大学黄达先生为本书作序。

感谢机械工业出版社华章公司的副总编杨福川和编辑李艺帮助我审阅全部书稿，让本书得以出版。

特别感谢我的爱人一直鼓励我，最终让我走出了失意的阴影。感谢我的爸爸、妈妈，感谢你们对我工作上的支持和生活上的照顾！

谨以此书献给我最亲爱的家人以及众多的 R 语言爱好者们！

张丹

序一	
序二	
前言	
第一部分 金融市场与金融理论	
第 1 章 金融市场概述	2
1.1 R 语言为量化而生	2
1.1.1 为什么是 R 语言	3
1.1.2 跨界结合	4
1.1.3 R 语言量化工具包	5
1.1.4 实战应用	6
1.1.5 量化交易平台系统架构	11
1.2 算法, 如何改变命运	13
1.2.1 算法在各个行业的应用	14
1.2.2 投身于哪个行业好	15
1.2.3 金融最靠谱	15
1.3 FinTech 金融领域的风口	18
1.3.1 大起大落	19
1.3.2 互联网已经在并购阶段	20
1.3.3 寻找好的行业风口	21
1.3.4 Gartner 技术成熟曲线	21
1.3.5 FinTech 金融领域的风口	22
1.4 国内量化投资工具介绍	23
1.4.1 量化交易概况工具	23
1.4.2 证券期货客户端	26
1.4.3 金融数据库	31
1.4.4 互联网在线策略平台	32
1.4.5 量化工具软件	34
1.4.6 API 程序工具	36
1.5 国内低风险交易策略	37
1.5.1 企业债	37
1.5.2 可转债	39
1.5.3 逆回购和正回购	41
1.5.4 现金管理	42
1.5.5 分级基金 A	43
1.5.6 期货	45
第 2 章 金融理论模型	46
2.1 R 语言解读资本资产定价模型	
CAPM	46
2.1.1 故事背景	47
2.1.2 资本市场线	48
2.1.3 资本资产定价模型	52
2.1.4 用 R 构建投资组合模型	54
2.1.5 Beta VS Alpha	60
2.2 R 语言解读一元线性回归模型	60
2.2.1 一元线性回归介绍	61
2.2.2 数据集和数学模型	62
2.2.3 回归参数估计	64

2.2.4	回归方程的显著性检验	66	3.2.2	data.table 包的使用	112
2.2.5	残差分析和异常点检测	67	3.2.3	data.table 包性能对比	121
2.2.6	模型预测	71	3.3	R 语言高效的管道操作 magrittr	126
2.3	R 语言解读多元线性回归模型	72	3.3.1	magrittr 介绍	126
2.3.1	多元线性回归介绍	73	3.3.2	magrittr 包的基本使用	127
2.3.2	多元线性回归建模	73	3.3.3	magrittr 包的扩展功能	132
2.3.3	模型优化	78	3.4	R 语言字符串处理包 stringr	134
2.3.4	案例：黑色系期货日 K 线		3.4.1	stringr 介绍	135
	数据验证	82	3.4.2	stringr 的 API 介绍	135
2.4	R 语言解读自回归模型	85	3.5	R 语言中文分词包 jiebaR	151
2.4.1	自回归模型介绍	85	3.5.1	jiebaR 包介绍	152
2.4.2	用 R 语言构建自回归模型	86	3.5.2	5 分钟上手 jiebaR	152
2.4.3	模型识别 ACF/PACF	88	3.5.3	分词引擎	154
2.4.4	模型预测	92	3.5.4	配置词典	156
			3.5.5	停止词过滤	160
			3.5.6	关键词提取	161
第二部分 R 语言数据处理与高性能计算					
第 3 章 R 语言数据处理 96					
3.1	掌握 R 语言中的 apply 函数族	96			
3.1.1	apply 的家族函数	97			
3.1.2	apply 函数	98			
3.1.3	lapply 函数	101			
3.1.4	sapply 函数	102			
3.1.5	vapply 函数	104			
3.1.6	mapply 函数	105			
3.1.7	tapply 函数	106			
3.1.8	rapply 函数	108			
3.1.9	eapply 函数	109			
3.2	超高性能数据处理包 data.table	111			
3.2.1	data.table 包介绍	112			
第 4 章 R 语言高性能计算 164					
4.1	OpenBlas 让 R 的矩阵计算加速	164			
4.1.1	OpenBlas 介绍	165			
4.1.2	R 和 OpenBlas 的安装	165			
4.1.3	让 R 语言加速	169			
4.2	R 语言跨界调用 C++	171			
4.2.1	Rcpp 的简单介绍	172			
4.2.2	5 分钟上手 Rcpp	172			
4.2.3	数据类型转换	176			
4.3	当 R 语言遇上 Docker	186			
4.3.1	当 R 遇上 Docker	187			
4.3.2	用 Docker 来管理 R 的程序	188			

第三部分 金融策略实战

第5章 债券和回购 196

5.1 了解国债 196

5.1.1 国债基本介绍 197

5.1.2 国债的意义 198

5.1.3 记账式国债 200

5.1.4 国债 101308 200

5.1.5 国债的历史表现 202

5.2 企业债和企业债套利 205

5.2.1 什么是企业债? 206

5.2.2 什么是公司债? 207

5.2.3 企业债和公司债的区别 209

5.2.4 企业债统计分析 209

5.2.5 企业债举例 213

5.2.6 企业债交易操作 214

5.3 可转债套利实践 216

5.3.1 可转债介绍 216

5.3.2 可转债操作 218

5.3.3 负溢价率套利策略 219

5.4 金融无风险交易工具逆回购 231

5.4.1 逆回购简单介绍 231

5.4.2 逆回购的品种有哪些? 232

5.4.3 逆回购交易 233

5.4.4 正回购操作 236

5.4.5 央行的公开市场操作 237

第6章 量化投资策略案例 241

6.1 均值回归, 逆市中的投资机会 241

6.1.1 均值回归原理 242

6.1.2 均值回归模型和实现 245

6.1.3 量化选股 257

6.2 R语言构建追涨杀跌量化交易模型 262

6.2.1 什么是追涨杀跌 262

6.2.2 追涨杀跌的建模和实现 265

6.2.3 模型优化 275

6.3 R语言构建配对交易量化模型 279

6.3.1 什么是配对交易 279

6.3.2 配对交易的模型 280

6.3.3 用R语言实现配对交易 284

6.4 基金会计系统设计和实现 293

6.4.1 基金会计系统介绍 294

6.4.2 资产核算 300

6.4.3 净值份额核算 300

6.4.4 计算案例 305

6.4.5 会计系统架构 307

6.5 用数据解读摩羯智投 313

6.5.1 摩羯智投介绍 313

6.5.2 数据收集 315

6.5.3 数据建模分析 317

6.5.4 结论 328

结束语 329

附录A Docker环境安装 330



第一部分 *Part 1*

金融市场与金融理论

- 第1章 金融市场概述
 - 第2章 金融理论
- 

金融市场概述

本章为全书开篇，主要介绍如何使用 R 语言做量化投资的思路和方法。量化投资是跨学科知识结合的一个方向，包括 R 语言技术层面的知识、基础学科的应用和金融市场的情况。R 语言社区提供了丰富的金融工具包，可以让我们快速构建量化投资的体系结构。本章内容是以我个人的从业经验为基础，从数据的角度观察中国的金融市场，发现机会，找到风口。

1.1 R 语言为量化而生

问题

为什么用 R 语言进行量化投资？



引言

做数据分析的朋友，一定知道 R 语言。R 语言是一门统计语言，在数据分析领域具有非常明显的优势。金融本身是重视数据的行业，而 R 的最大优势就是数据分析，所以将 R 语言与金融相结合，用 R 来做量化投资的策略，真的很配，不仅顺手而且方便，用了你就会知道。

本章以“R 语言为量化而生”作为本书的开篇，主要强调了 R 语言在量化投资领域有着广泛的应用场景，而且是重要的量化投资工具。

1.1.1 为什么是 R 语言

R 语言是一门面向数据的编程语言，早期只是在统计领域被统计学家所使用。近年来随着大数据技术的发展，以及 R 语言自身生态的快速壮大，不仅在统计领域，在各大行业领域中都能看到 R 语言的身影，包括互联网、数据科学、人工智能、机器学习、生物医药、游戏、电子商务、全球地理科学、数据可视化等。

R 语言不仅能把数据分析做好，而且能通过数据连接到我们每个人的日常生活，让我们能够有更多的思考。比如，去超市购物，你购买的东西会记录在你的购物小票中，如果把这些购物小票都收集起来，通过 Apriori 关联规则的频繁项集算法，就可以分析出你的购物习惯。

再比如，我们经常会收到各种推销的骚扰电话，有时候是卖房，有时候是银行贷款，有时候是股票推荐，有时候是代开发票，还有时候是“猜猜我是谁”，此时我们可以收集自己的通话记录，通过电话号码、号码注册位置、通话时长来分析一下，哪些是有用通话，哪些是浪费时间的通话，从而计算出我们每一天会浪费多少时间，通过 Bayes 算法把有效通话和无效通话进行二元分类，从而阻止骚扰电话的入侵。

如果我们不仅能够收集自己的通话记录，还能把亲戚朋友或更多人的通话记录都收集起来，组建成一个数据库，再把统计出来的骚扰电话统一标记，做成黑名单，然后把黑名单公开出来，就可以让更多的人免受骚扰电话的影响。

人们的衣食住行都会产生数据，产生数据的地方，就需要做数据分析。通过数据分析，可以提高生活的效率，为我们节约宝贵的个人时间。

在人类社会，人们除了衣食住行的基本需求外，更多是对钱的需求。我们每天都会跟钱打交道，会有消费的需求，月底会有工资收入，当收入大于消费，资金积累到一定阶段的时候，我们就会有投资的需求。中国人的投资一般就几种，买房、买黄金、买股票、买保险、买理财。投资时，很多人是不理性的，心血来潮就买了，也不仔细观察行情，买完很容易被套住，像“中国大妈抢购黄金”“中国大妈是救市的主力军”等事件都一度被新闻媒体热炒。

在金融市场中最不缺的就是机会，不管投资什么，如果能在交易的时候多思考一下，结果说不定就会不一样。运用我们数据分析的知识和对金融市场的理解，通过 R 语言

进行数据分析，抓住机会，就会让我们辛辛苦苦赚到的钱得到保值和增值。

1.1.2 跨界结合

R 语言与其他编程语言的最大不同在于，R 语言是面向数据的，这一点很重要，甚至影响了我们的思维模式。我写了 10 多年的 Java 程序，程序员的思维模式在我的脑子里根深蒂固。我习惯用面向对象的思想来建模，把现实世界中的物体、行为、连接都用面向对象的方法来表达。我也习惯了 IT 项目的开发模式，从架构设计到开发的自上而下的设计思路，再到完全由需求驱动的敏捷开发，以及为业务随需而变的开发路径。作为程序员，我有理想，我自认为技术能改变世界，而这一切都是程序化思维在影响着我。如果你还不熟悉什么是面向对象，请参考《R 的极客理想——高级开发篇》第 4 章。

后来，当我接触了 R 语言，通过另一扇门理解世界的时候，我发现世界是不一样的。在没有程序的世界里，你可以通过文字来表达你做事的方法，你可以感性地认识世界，也可以让数据来说出这个世界的意义。你可以天天看新闻联播，了解国家大事；也可以看微博、朋友圈的爆料，及时了解身边的消息。这并不是程序员的思维，而是大众化社会的想法。在这个社会里，程序员只是小众群体。

改变世界的方法，除了技术还有别的，比如，政策、资本、市场、法律、人事等，从公司运作的层面看，似乎哪个都比技术更重要。也许程序员并不关心这些，但面对市场时，再牛的技术也会被其他因素所制约。

我经历过创业，一个很深刻的教训就是，要让自己跳出原有的思维定式，通过多维的视角看世界、看自己。R 语言帮我打开了通过数据看世界的窗口。我的能力提升不再局限于如何把程序写得多么漂亮，或是在时间复杂度上用 $O(\log(N))$ 解决 $O(N^2)$ 的问题，而是开始关注怎么收集数据，怎么提高数据质量，来解决实际的业务问题。当然，由于程序员与生俱来的职业强迫性，我也会要求自己熬夜一晚上把 50 行代码优化成 40 行。

由上可见，问题被转化了，虽说术业要有专攻，但作为一个有理想的极客，全面掌握跨领域的知识才是我们的目标。我创业时经常被投资人问到，做量化投资的项目有多难？招几个金融背景的人才，再招几个 IT 背景的人才，捣鼓个项目，不是很快就做出来了吗？其实不然，不同行业背景的人是很难交流和沟通的，尤其是金融和 IT 行业：一种是封闭不愿意分享，一种是开放毫无保留；一种是先把话说出来，一种是先把事情做起来；一种是所有人都要为我服务，一种是只专注于自己的兴趣。所以，如果不是能够跨界通吃的人，是做不好量化项目的。

我们要立志做个跨学科的人，这才是最核心的点，是不可被替代的。从 IT 技术出发，你需要扎实的编程功底及架构思想、科学的项目管理方法、严谨的产品设计逻辑；你还需要具备统计、数学等基础学科的知识，以及数据挖掘与机器学习的数据处理经验；最重要的，不管是量化投资或其他的领域，你要能结合业务，理解市场的规则（如图 1-1 所示）。

这听起来很难，实际上也确实很难！虽然很难，但不光是对你，对所有人都是这么难。只要你能坚持下来，一定能通过知识改变命运。