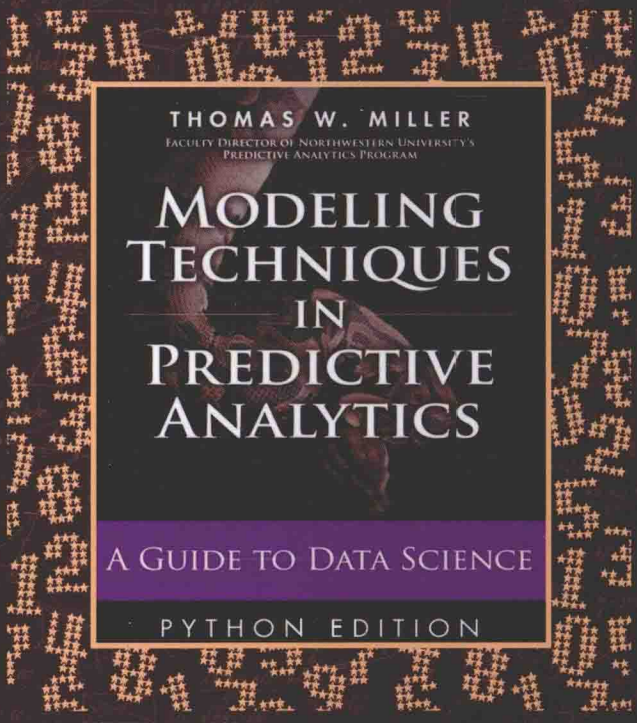
HZ BOOKS 华章IT
Pearson
数据科学与工程技术丛书

预测分析建模

Python与R语言实现

[美] 托马斯 W. 米勒（Thomas W. Miller）著
程豪 译



MODELING TECHNIQUES IN PREDICTIVE ANALYTICS WITH PYTHON AND R

A GUIDE TO DATA SCIENCE

机械工业出版社
China Machine Press

数据科学与工

MODELING TECHNIQUES
IN PREDICTIVE ANALYTICS WITH PYTHON AND R
A GUIDE TO DATA SCIENCE

预测分析建模
Python与R语言实现

[美] 托马斯 W. 米勒（Thomas W. Miller）著

程豪 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

预测分析建模：Python与R语言实现 /（美）托马斯·W. 米勒（Thomas W. Miller）著；程豪译 . —北京：机械工业出版社，2016.9
（数据科学与工程技术丛书）
书名原文：Modeling Techniques in Predictive Analytics with Python and R: A Guide to Data Science

ISBN 978-7-111-54887-4

I. 预… II. ① 托… ② 程… III. 程序语言—程序设计 IV. TP312

中国版本图书馆CIP数据核字（2016）第221060号

本书版权登记号：图字：01-2015-7584

预测分析建模：Python与R语言实现

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：陈佳媛　　责任校对：董纪丽
印　　刷：三河市宏图印务有限公司　　版　　次：2016年10月第1版第1次印刷
开　　本：185mm×260mm　1/16　　印　　张：18.5
书　　号：ISBN 978-7-111-54887-4　　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

译 者 序

作为开源的面向对象的脚本语言，R 与 Python 具有免费获得、简单易学、功能强大的共性。随着多年的实践、发展和稳定，R 与 Python 各自包含了一组完善、易懂的标准库，能够轻松解决很多现实问题。正如本书所言，在计算机密集型应用领域，Python 给予一种从 C、C++ 和 Fortran 调用编译程序的能力。而 R 能够完成当前用 Python 无法实现的建模和绘图任务。通过调用 R 软件包，用户可以处理数据分析、统计建模、统计制图和缺失数据等诸多问题，正如 R 用户从通用的 Python 语言中获益一样。

本书特别关注了众多统计领域中预测分析方向，加上用 R 与 Python 同时编程，为本书的独创性增色不少。作者通过涉及不同学科和应用领域的预测分析问题，为预测分析和数据科学提供一种综合性指南。秉承这种思想，我承担了本书的翻译工作。我希望能通过自己的努力，将这本实用性极强的 R 与 Python 综合教材推荐给更多的读者。无论您是 R 或 Python 初学者，还是 R 与 Python 高手，本书都可以为您在业界提供参考和帮助。

借此机会，不妨浅谈译书过程中的一些体悟。2015 年 9 月伊始，我暂时离开我的母校中国人民大学，由国家公派到美国哥伦比亚大学联合培养。本书的翻译也发生在这个重要的求学期间。出于对 R 与 Python 语言的热爱，我希望在满足学业要求的同时，利用周末尝试更多的挑战。翻译过程中，我感受到作者浓郁的文艺气质、渊博的专业积淀和灵动的思维韵律。在有限的时间里，我认真踏实地扮演了不同学习、工作任务中的角色，增加了人生的厚度。与音乐一样，知识的研发和传播没有国界。因此，翻译不仅是知识表达语言的转换，更是一次学习和交流的机会。与原作者对话，高山仰止，受益匪浅；与读者对话，高山流水，闻过则喜。

在此，非常感谢机械工业出版社的各位领导和编辑。感谢王春华编辑将本书推荐给我。感谢陈佳媛编辑对翻译内容的审读。作为我们的第二次合作，两位编辑一如既往的职业操守和工作态度，让我由衷钦佩。由于身在国外，很多事情需要朋友和同学的帮助和支持。感谢我的挚爱刘钰洁同学。正是她承担了必要的沟通联络工作，才顺利衔接了翻译工作的不同环节。感谢我的朋友范超、王婷和赵建喜对一些翻译内容提供的建议。感谢程悦同学在本书最终校对阶段提供的帮助和支持。这里，我要特别感谢美国哥伦比

亚大学的韦颖老师在科研上对我的指导，感谢我的导师中国人民大学易丹辉教授对我的关心和支持。感谢我的班主任尹建鑫老师，以及全体博士同学。

最后，我要特别感谢伟大的父母。作为人生中最能够包容且给予我最大支持的他们，让我有更强大的动力，去修缮和提高自己。感谢最爱的爷爷奶奶，跨洋的联系与问候让我倍感安心与温暖。感谢身边所有的亲朋好友。

介于个人时间与水平有限，如有纰漏，向您致歉，还望海涵。同时还请各位读者予以反馈，不吝赐教！

程豪

2016 年 3 月 15 日

前　言

“好吧！好吧！除了更好的卫生设备、医药、教育、葡萄酒、公共秩序、水利、公路和淡水系统和公共医疗——罗马人还为我们做过什么？”

——出自《布莱恩的一生》(1979 年) 中 John Cleese 的对白

20 世纪 70 年代末，我在明尼苏达大学攻读博士学位。在此期间，我学习了一门统计学编程课程。上课伊始，老师说：“课程作业不限编程语言，只要自己独立完成即可。”

当时，我已经熟练掌握 Fortran 语言，同时自学了 Pascal。我正在研究一种结构化的编程方式——不仅仅是 GO TO 语句。因此，我将老师的话信以为真，用 Pascal 语言完成了第一次作业。班里的其余 14 名同学用统计专业通用的 Fortran 语言。

当我提交作业的时候，老师看了看问我：“这是什么编程语言？”

“Pascal，”我回答，“您说过，我们可以选择任意一种编程语言，只要独立完成就好。”

老师回应说：“Pascal。我不会 Pascal，只会 Fortran。”

如今，数据科学世界汇聚了熟练使用 Python 语言的信息技术专业人士和熟练使用 R 语言的统计学者。他们之间有很多地方值得相互学习。对于数据分析科学家来讲，掌握多种编程语言是一种相当大的优势。

Python 有时被称为“黏合语言”，它为科学编程和研究提供了丰富的开源环境。在计算机密集型应用领域，Python 给予一种从 C、C++ 和 Fortran 调用编译程序的能力。我们可以用 Cython 将 Python 转换为优化的 C 语句。我们可以用 R 解决当前用 Python 无法实现的建模和绘图问题。通过调用 R 软件包，我们能够处理非线性估计、贝叶斯分层建模、时间序列分析、多变量方法、统计制图和缺失数据，正如 R 使用者能从通用的 Python 语言中获益一样。

现如今，数据与算法当道。欢迎来到一个崭新的世界，一个快节奏、数据密集的世界，一个开源的环境。在这个环境中，通过分析技术和思想交流可以获得一个具有竞争力却稍纵即逝的优势。

很多有关预测分析和数据科学的书都在讨论策略与管理。还有一些书关注方法和模型。其余则讨论信息技术 (和代码)。本书是一部同时兼顾三者的罕见著作，很受业界

管理者、建模人士和程序员的青睐。

在获得具有竞争力的优势过程中，我们意识到了分析的重要性。我们通过提供建模技术的现有资源和参考指南，来帮助研究者和分析师。我们能够向程序员展示如何建立一个解决真实问题的代码基础。我们图文并茂地为管理者解释模型结果，以及数据和模型的意义。

随着收集和存储的数据容量增大、可用于分析的数据类型增多、数据产生和分析需求的速度加快，数据分析的重要性与日俱增。获得具有竞争力的优势意味着为信息管理和分析提供一套新体系，意味着业界问题处理方式的改变。

由于涉及很多学科和应用领域，数据科学的文献资料浩如烟海。相关的开源代码层出不穷。事实上，提供一部预测分析和数据科学的综合性指南将成为一项挑战。

我们关注的是实际问题和真实数据。在每一章加入一些特定应用领域和业界问题的案例，并提供有效的解决方法。通过展示建模技术和编程工具，我们将抽象的概念转换为具体的例子。这些详实的案例有助于读者的理解。

我们的宗旨是提供一种适合于很多读者的预测分析和数据科学方面的综述。本书省略了数学部分。有关具体的细节和方法导论，请统计学者和建模人士查阅参考文献。我们用通俗易懂的语言讲述方法，使用数据的可视化展示业界问题的解决方案。

了解本书的宗旨后，一些读者可能会想知道我是经典学派还是贝叶斯学派。在明尼苏达大学统计学院读书时，我对两大学派都心生敬意。我非常崇拜经验贝叶斯学者和将机器学习与传统统计学相结合的研究者。在建模和推断方面，我则是一个实用主义者。我会做有效的研究工作，并做出通俗易懂的解释。

本书之所以必要，是因为世界各地成千上万的专家将时间和想法贡献给开放源代码事业。开放源代码的增加及其难度的进一步降低，确保了先进的解决方法一定会在多年以后出现。精灵跑出明灯，能手走出幕后——火箭科学不再如往常。秘密正在被揭晓。本书就是此过程的一部分。

本书的绝大部分数据来自公开数据源。美国职棒大联盟的晋级和上座率数据由 Erica Costello 提供。计算机选择研究数据由 Sharon Chamberlain 提供。“匿名银行”的呼叫中心数据由 Avi Mandelbaum 和 Ilan Guedj 提供。电影信息获得了互联网电影数据库的使用许可。IMDb 电影评论数据由 Andrew L. Mass 和他在斯坦福大学的同事一起管理。其中一些例子出自佛罗里达州坦帕市的 ToutBay，NCR Comten，Hewlett-Packard 公司，纽约的 Site Analytics 公司，威斯康星州麦迪逊的 Sunseed Research 和麦迪逊的 Union Cab Cooperative 的工作人员。

我们在一个开源的环境中分享代码。我们所做的工作就是编译程序。在这个环境中，每个人都可以浏览现有程序，一些人还可以调试程序。为了促进学生学习，所有程序都包括了方便深入分析的详细注释和建议。所有数据集和计算机程序都可从本书的网

站上下载：http://www.ftpress.com/miller/。

本书的最初计划是将R版本转换为Python版本。然而，当我只用Python撰写本书时，我对两种编程语言产生了更加深远的敬意。我见证了一些问题用Python处理起来很容易，而另外一些问题则更适合用R来处理。而且，对于从事数据实践的科学家来讲，在使用Python进行建模和绘图时，R软件包的调用成为一种明显的优势。因此，本书同时给出Python和R代码示例，提供了一部独特的双语数据科学指南。

在过去的几年间，我受到了很多人的影响。很感激那些优秀的思想家，出色的人，还有老师和导师。遗憾的是，尤西纽斯学院的哲学家Gerald Hahn Hinkle和语言学家Allan Lake Rice，还有明尼苏达大学的哲学家Herbert Feigl，他们永远离开了我们。此外，我非常感谢明尼苏达大学的心理测验学者David J. Weiss和俄勒冈大学的经济学者Kelly Eakin。德高望重的老师是我一生的财富。

感谢Michael L. Rothschild、Neal M. Ford、Peter R. Dickson和Janet Christopher。在威斯康星麦迪逊分校和A C. 尼尔森中心一起进行市场调查的那段时间里，他们给我提供了非常重要的支持。

我住在距离道奇体育场北面4英里[⊖]的加利福尼亚洲，在伊利诺伊州埃文斯市的西北大学任教，兼任佛罗里达州坦帕市数据科学公司ToutBay的产品研发指导。这些都为我提供了良好的互联网连接环境。

我很庆幸自己完成了美国西北大学专业进修学院的远程教育。感谢Glen Fogerty给予了我在西北大学预测分析编程专业授课并承担领导角色的机会。感谢管理这一研究项目的同事们和工作人员，同时感谢让我获益良多的同学们和老师们。

ToutBay是一家新兴的数据科学公司。Greg Blence是联合创始人之一，我很期待接下来的发展。感谢Greg让我加盟并扎根于实际问题。迄今为止，只有学术和数据科学模型引领着我们。为了有所作为，我们最终必须实现我们的想法和模型，并分享给大家。

TEXnology公司的Amy Hendrickson编辑了本书的文字、表格和图片，取得了开源的又一次胜利。感谢Donald Knuth和TEX/LATEX提供了很好的排版和出版系统。

感谢本书R版本的读者和审校者，他们是Suzanne Callender、Philip M. Goldfeder、Melvin Ott和Thomas P. Ryan。Lorena Martin为本书R版本的修订版提供了很多的反馈和建议。Candice Bradley兼任了审校者和文字编辑，Roy L. Sanford对统计模型和程序提供了技术支持。感谢Jeanne Glasser Levine编辑，和Pearson/FT出版社（是他们让这本书最终面世）。当然，任何写作问题和错误，以及疏漏仅是我个人的责任。

⊖ 1英里＝1.609千米——编辑注

我的好朋友 Brittney 和他的女儿 Janiya 只要在时间允许的情况下都会来陪伴我。还有我的儿子 Daniel，无论是逆境还是顺境，他总是在我身边，是我一生的朋友。他们的信任和支持让我无以为报。

Thomas W. Miller

加利福尼亚州格伦代尔市

目　录

第 1 章
分析与数据科学

Maguire 先生说："我只想告诉你一个词语，就一个词语。"
Ben 说："恩，请讲。"
Maguire 先生说："你在听吗？"
Ben 说："是的，我在听。"
Maguire 先生说："君子不器。"

——电影《毕业生》（1967 年）中 Walter Brooke 饰演 Maguire 先生，
Dustin Hoffman 饰演 Ben（Benjamin Braddock）

虽然获得哲学学位并非职业生涯中的最佳选择（除非你想成为一名哲学老师，但是这样的就业机会并不多），但我认为学习哲学的那几年非常重要。上大学时，我撰写过一篇关于英国哲学家 Bertrand Russell 的文章。在明尼苏达大学读研时，我学习了一位真正伟大的哲学家 Herbert Feigl 的课程，阅读了探索科学与真理的认识论。我最喜欢的哲学是逻辑经验论。

尽管那段"思考思考"（think about think）（这是 Feigl 对哲学的定义）远远在我们身后，但是在专业训练的早些年，我培养了自己敏锐判断真实与空谈的能力。

模型是一系列事物的代表，是对真实事物的展现和描述。数据科学的经典模型是将（一组）变量与（另一组）变量建立关系的尝试。它虽然有限、不够精确，但却有用，可以帮助我们理解这个世界。正是基于数据的原因，模型并非空谈。

预测分析汇聚了管理、信息技术和建模。它专为如今数据密集型世界而设计。预测分析是一门数据科学，一种对于商业上获得成功、非盈利组织和政府的建设非常重要的涉猎多学科的技能。无论是预测销售规模还是市场份额，找到一个好的零售点还是投资机会，辨别消费群体还是目标市场，评估新产品的市场潜力还是它与竞品之间存在的风险，预测分析中的建模方法都提供了解决这些问题的关键。

预测分析领域的数据科学家使用专业术语——会计、金融、营销、管理。他们了解包括数据结构、算法、面向对象编程在内的信息技术。他们理解统计模型、机器学习和数学规划。数据科学是方法学杂论，引用了很多学科理论，将实证研究转换为管理人员可以理解的文字和图片。

同统计学中的很多分支一样，预测分析包括找到变量间有价值的关系，并用模型表示。

响应变量是要预测的目标变量。解释变量或预测变量是可观测的、可操作的，或可控制的且与响应变量相关的。

回归方法有助于我们预测一个有意义的响应变量，比如销售数量、股票价格或投资回报。分类方法有助于我们预测一个分类型响应变量。哪种品牌会被购买？消费者是否会购买这个产品？账户持有人是否会偿还这笔贷款？这桩银行交易是真实的还是骗人的？

预测问题包括潜在预测变量的宽度或个数，以及数据集中观测的长度或数量。商务、营销、投资分析中潜在预测变量的个数会造成最大的困难。成千上万的潜在预测变量与响应变量间存在弱关系。在计算机的帮助下，成百上千或成千上万的模型可以在数据的一部分子集中作拟合，在另一部分子集中作测试，提供每一个预测变量的评价。预测模型包括发现预测变量的最优子集。拟合度较好的模型优于拟合度较差的模型。简单模型优于复杂模型。

考虑预测分析中三种主要的研究和建模方法：传统的、数据自适应的和模型依赖性的，如图 1.1 所示。学术研究统计推断和建模的传统方法是以理论或模型设定开始的。统计推断中采用经典或贝叶斯方法。传统方法，比如线性回归和 logistic 回归，对线性预测变量估计参数。模型构建包括模型拟合和模型诊断。在使用传统模型作预测之前，我们首先进行模型验证。

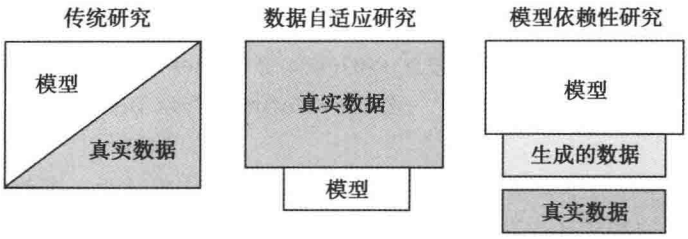


图 1.1　数据与模型研究

使用数据自适应方法时，我们从数据开始，并搜索这些数据以发现有用的预测变量。在分析之前，我们不人为设定理论和假定。这是机器学习的世界，有时也叫作统计学习或数据挖掘。数据自适应方法适用于有效数据，代表变量间的非线性关系和交互作用。数据决定了模型。

数据自适应方法是数据驱动的。正如传统模型，在使用模型作预测之前，我们首先验证数据自适应模型。

模型依赖性研究是第三种方法。它以模型详述开始，使用模型生成数据、预测或建议。作为运筹学的主要途径，模拟和数学规划方法都是模型依赖性研究的例子。当使用模型依赖性或模拟方法时，模型通过对比生成数据和真实数据得以改进。我们好奇模拟的顾客、企业和市场是否和真实的顾客、企业和市场行为相似。一种验证方式是对真实数据进行对比。

通常，模型和方法的结合最有效。考虑一个金融研究领域的应用。一位共同基金的管理者正在为基金投资组合寻找额外的股份。一位金融工程师使用一个数据自适应模型（可能是神经网络）搜索成千上万的性能指标和个股，识别一组个股来进一步研究。然后，通过操作识别出来的这组个股，金融工程师使用一种基于理论的方法（CAPM，资本资产定价模型）来识别出更小的一组个股，并将其推荐给基金管理者。作为最后一步，使用模型依赖性研究（数学规划），工程师识别出投资组合中所有个股中风险最小的投资。

数据可能由观测单元、时间和空间组成。可观测的或截面单元可能是单个消费者、业务或任何其他的收集和组织数据的基础。数据以秒、分、小时、天等时间单位来组织。经常以经度和纬度来定义空间或位置。

考虑周一（时间点）走进加利福尼亚州格伦代尔杂货商店的消费者（分析单元），忽略这些商店的空间位置——这就是截面数据。试想，我们关注其中一家店，记录连续（6 个月每天）到这家商店的消费者数目——这是时间序列数据。然后我们记录格伦代尔所有杂货商店的消费者数目——这是纵向数据或面板数据。为了完成研究，我们用经度和纬度定位这些商店，进而获得空间或时空数据。除了考虑到这家商店的消费者数目，还可以考虑测量其他任何数据结构。我们观察商店销售量、消费者或附近居民的人口特征、格伦代尔街道的交通流量，由此可以移动到多元时间序列和多元方法。我们收集的数据的组织形式影响使用模型的结构。

在本书中，我们考虑一些业界问题，涉及很多类型的模型，包括截面、时序和空间数据模型。无论数据结构和相关模型如何，预测是统一的主题。我们使用已有数据预测未知数据，发现预测结果具有不稳定性。这是一个外推和预测的过程。模型验证对这个过程很重要。

为了预测，我们可能使用经典方法或贝叶斯方法，或者可能完全免除了传统统计学以及依赖机器学习的算法。我们要做有效的研究[⊖]。预测分析方法建立在一个简单的前提下：

模型的价值在于预测的质量。

我们借鉴统计学量化的不稳定性。一方面，可以用置信区间、带一定标准误的点估计、显著性检验和 P 值——这些经典方法；另一方面，我们有后验概率分布、概率区间、预测区间、贝叶斯因素和主观（也许是扩散）先验——贝叶斯统计的方法。一些评价指标比如赤池信息量准则（AIC）或贝叶斯信息准则（BIC）有助于我们判断模型，以便在拟合度和简洁度间找到平衡。

方法的核心是**训练与测试的结合**。我们将样本数据划分为训练集和测试集。用训练集建立模型，在测试集上评价。简单地将数据划分为两到三部分，如图 1.2 所示。

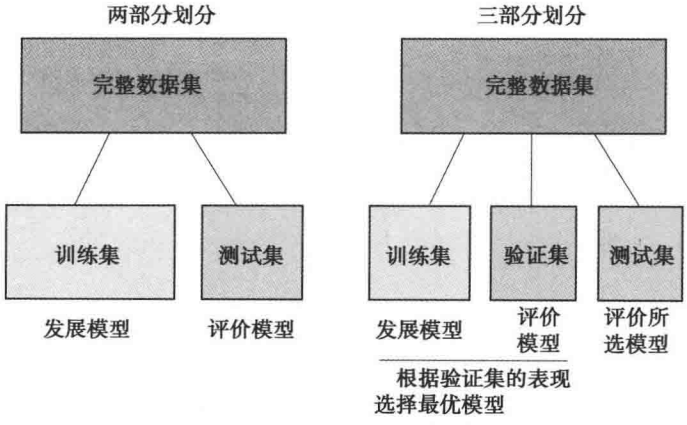


图 1.2　模型评价的训练与测试法

⊖ 统计文献中，Seymour Geisser（1929—2004）介绍了一种被描述为贝叶斯预测推断（Geisser 1993）的方法。贝叶斯统计是在贝叶斯理论创造者 Reverend Thomas Bayes（1706—1761）后命名的。在强调成功预测的同时，我们同意 Geisser 的观点。然而，我们的方法纯粹是凭经验的，而且绝不取决于经典或贝叶斯思维。

将样本随机分成训练集和测试集，尤其是在处理小数据集时，因此我们有时通过一些随机划分和对性能指标的平均进行统计实验。在训练和测试主题上存在一些扩展和变化。

训练和测试主题上的一次变化是多折交叉验证，如图 1.3 所示。我们将样本数据划分为近似等样本量的 M 折，并进行一系列测试。对于图中的 5 折交叉验证，我们将首先通过 E 训练集合 B，在集合 A 上作测试。而后通过 E 训练集合 A 和 C，在集合 B 上作测试。以此类推，直到 5 折中的每一个被用作测试集。通过平均这些测试结果评价表现。在留一法交叉验证中，多折交叉验证合乎逻辑的主题是在样本中观测数量和测试集一样多。

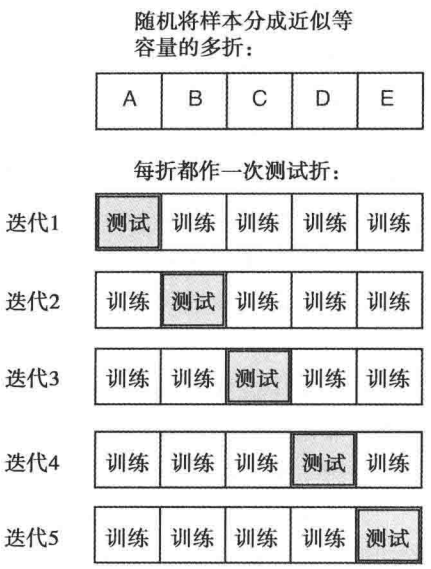
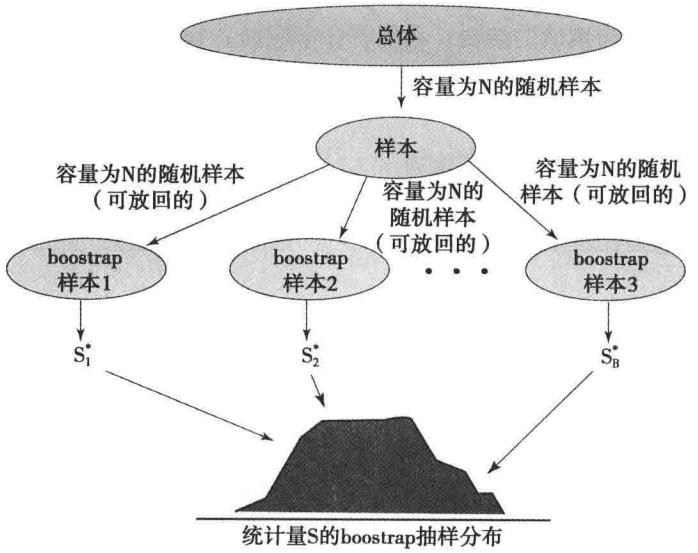


图 1.3　多折交叉验证的训练与测试

训练与测试的另一种变化是 boostrap 相关的方法。如果样本近似总体，那么从样本中抽取的样本（重抽样）也近似总体。boostrap 的流程如图 1.4 所示，是可放回的重复抽样。换句话说，我们从样本中重复抽取很多随机样本，计算每一个统计量的值。统计中的 boostrap 分布近似样本分布。boostrap 的价值何在？在于不用对总体分布作假定。我们从每个样本数据中估计标准误，并计算概率。boostrap 也可能用来改进留一法的交叉验证过程的预测误差估计。交叉验证和 boostrap 方法可参见 Davison 和 Hinkley（1997），Efron 和 Tibshirani（1993），以及 Hastie、Tibshirani 和 Friedman（2009）。



数量S_1^*，S_2^*，…，S_B^*表示从B个boostrap样本计算统计量S。

图 1.4　Boostrap 重抽样的训练与测试

数据可视化是数据科学研究的关键。本书的例子展示了在发现、识别和设计方面数据可视化的重要性。我们使用探索性数据分析（发现）和统计建模（识别）的方法。我们使用图

表与管理者沟通结果。

没有比安斯库姆四重奏（Anscombe Quartet）更能生动说明统计图表和数据可视化的重要性了。请见表 1.1 中的数据集，由 Anscombe（1973）制作。看看这些制表数据，普通读者会发现第四个数据集与其他三个明显不同。前三个数据集怎么样？ x 与 y 之间的关系模式存在明显不同吗？

表 1.1　安斯库姆四重奏（Anscombe Quartet）数据

Set Ⅰ		*Set* Ⅱ		*Set* Ⅲ		*Set* Ⅳ	
*x*1	*y*1	*x*2	*y*2	*x*3	*y*3	*x*4	*y*4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.985	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

当用 x 对 y 作回归时，模型得出相似的统计结果。响应变量 y 的均值为 7.5，解释变量 x 的均值为 9。四个数据集的回归分析几乎相同。每个数据集的拟合回归方程是 $\hat{y}=3+0.5x$。四个模型中每一个模型的响应变量方差的比例占 0.67。

追从 Anscombe（1973），我们认为统计结果没有告知数据的特点和规律。我们必须把眼光放远，不仅仅看数据表、回归系数和统计检验的结果。图 1.5 说明了数据中蕴含的故事。可以看出，四个安斯库姆（Anscombe）数据彼此差异很大。

安斯库姆四重奏数据说明我们必须理解数据。安斯库姆四重奏数据的 Python 和 R 程序见本章末的代码清单 1.1 和代码清单 1.2。

可视化工具有助于我们了解和熟悉数据。我们探索数据，发现数据规律，识别观测间的聚类情况和异常点或离群点。有时强调变量间的关系能发现数据中潜在的维度。

Tukey（1977）、Tukey 和 Mosteller（1977）的经典著作论述了解释性数据的绘图方法。Cook（1998），Cook 和 Weisberg（1999），以及 Fox 和 Weisberg（2011）介绍了回归方面的绘图方法。Tufte（1990，1997，2004，2006）、Few（2009）和 Yau（2011，2013）在其著作中说明了统计图表和数据可视化。Wilkinson（2005）回顾了人类感知和绘图方法的相关内容，还提供了理解统计图表的一个概念化结构。Cairo（2013）提供了信息图的通用综述。Heer、Bostock 和 Ogievetsky（2010）证实了网络分布的现代可视化技术。当处理大型数据集时，可能需要特殊的方法，比如部分透明化绘图法和 hexbin 图（Unwin、Theus 和 Hofmann 2006，Carr、Lewin-Koh 和 Maechler 2014，Lewin Koh 2014）。

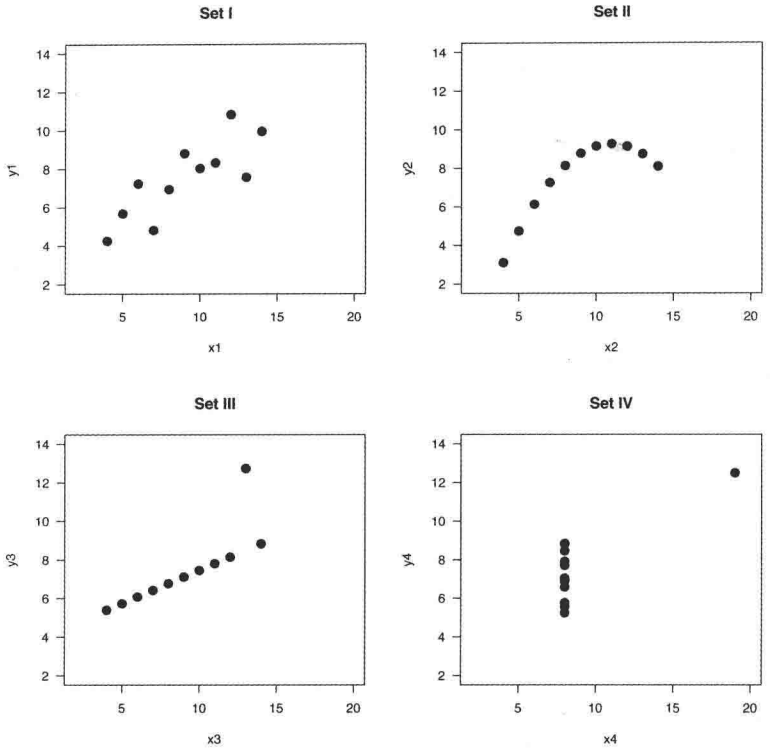


图 1.5　数据可视化的重要性：安斯库姆四重奏

Python 和 R 提供了丰富的数据可视化编程环境，还包括万维网的可视化应用接口。Chun（2007）、Beazley（2009），以及 Beazley 和 Jones（2013）评论了 Python 编程环境。Matloff（2011）和 Lander（2014）提出了有用的 R 导论。Murrell（2011）提供了一个 R 制图法综述，被 Sarkar（2008，2014）讨论的点阵图形建立在更早期系统 S-Plus Trellis（Cleveland 1993；Becker 和 Cleveland 1996）的概念结构基础上。Wilkinson（2005）的“图表语法”方法已经被 Python ggplot（Lamp 2014）软件包和 R ggplot2 软件包（Wickham 和 Chang 2014）实现，Chang（2013）提供了 R 编程的案例。Cairo（2013）、Zeileis、Hornik，以及 Murrell（2009，2014）提供了统计图表颜色的建议。Ihaka 等人（2014）展示了如何用 R 中的 hue、chroma 和 luminance 指定颜色。

数据科学家所做工作如下：

- **搜索相关研究**。第一步是信息搜索，查阅已有成果，并仔细研读相关文献。我们收集了很多研究领域的学者和从业人员，以及对预测分析和数据科学研究人士的著作成果。
- **准备文本和数据**。文本是没有结构或者有部分结构的。数据经常是凌乱或缺失的。我们从文本中提取特征，定义步骤，为分析和建模准备文本和数据。
- **熟悉数据**。我们研究的目的是作探索性数据分析和数据可视化。寻找数据聚类情况，发现离群点，识别共同的维度、规律和趋势。
- **预测数量问题**。可以预测产品销售的数量和金额，金融证券或房地产的价格。回归技术在这些预测问题中非常有用。

- **预测二分类问题**。很多商业问题是分类问题。具体来说，我们可以用分类方法预测一个人是否会购买某个产品，是否拖欠贷款，是否访问网页。
- **完成检验**。我们用诊断图检验模型，观察在一个数据集上训练得到的模型在另一个数据集上效果如何。我们用数据划分、交叉验证或者 bootstrap 方法进行训练和测试。
- **玩假设情景的游戏**。通过关键变量查看我们预测发生的变化。在模拟交易市场玩万一的游戏。作数学规划模型的敏感性或压力测试。观察输入变量的值是如何影响输出值、结局和预测结果评估预测的不确定性。
- **全部解释**。数据和模型有助于我们理解世界。我们将获得的结论转换为别人能理解的解释。用简洁明了的方式呈现项目结果。这些呈现得益于良好的数据可视化技术。

现在开始。

代码清单 1.1　安斯库姆四重奏程序（Python）

```
# The Anscombe Quartet (Python)
# demonstration data from
# Anscombe, F. J. 1973, February. Graphs in statistical analysis.
#  The American Statistician 27: 1721.

# prepare for Python version 3x features and functions
from __future__ import division, print_function

# import packages for Anscombe Quartet demonstration
import pandas as pd  # data frame operations
import numpy as np  # arrays and math functions
import statsmodels.api as sm  # statistical models (including regression)
import matplotlib.pyplot as plt  # 2D plotting

# define the anscombe data frame using dictionary of equal-length lists
anscombe = pd.DataFrame({'x1' : [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5],
    'x2' : [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5],
    'x3' : [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5],
    'x4' : [8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8],
    'y1' : [8.04, 6.95,  7.58, 8.81, 8.33, 9.96, 7.24, 4.26,10.84, 4.82, 5.68],
    'y2' : [9.14, 8.14,  8.74, 8.77, 9.26, 8.1, 6.13, 3.1,  9.13, 7.26, 4.74],
    'y3' : [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73],
    'y4' : [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91, 6.89]})

# fit linear regression models by ordinary least squares
set_I_design_matrix = sm.add_constant(anscombe['x1'])
set_I_model = sm.OLS(anscombe['y1'], set_I_design_matrix)
print(set_I_model.fit().summary())

set_II_design_matrix = sm.add_constant(anscombe['x2'])
set_II_model = sm.OLS(anscombe['y2'], set_II_design_matrix)
print(set_II_model.fit().summary())

set_III_design_matrix = sm.add_constant(anscombe['x3'])
set_III_model = sm.OLS(anscombe['y3'], set_III_design_matrix)
print(set_III_model.fit().summary())

set_IV_design_matrix = sm.add_constant(anscombe['x4'])
set_IV_model = sm.OLS(anscombe['y4'], set_IV_design_matrix)
print(set_IV_model.fit().summary())

# create scatter plots
fig = plt.figure()
set_I = fig.add_subplot(2, 2, 1)
```