CENGAGE Learning

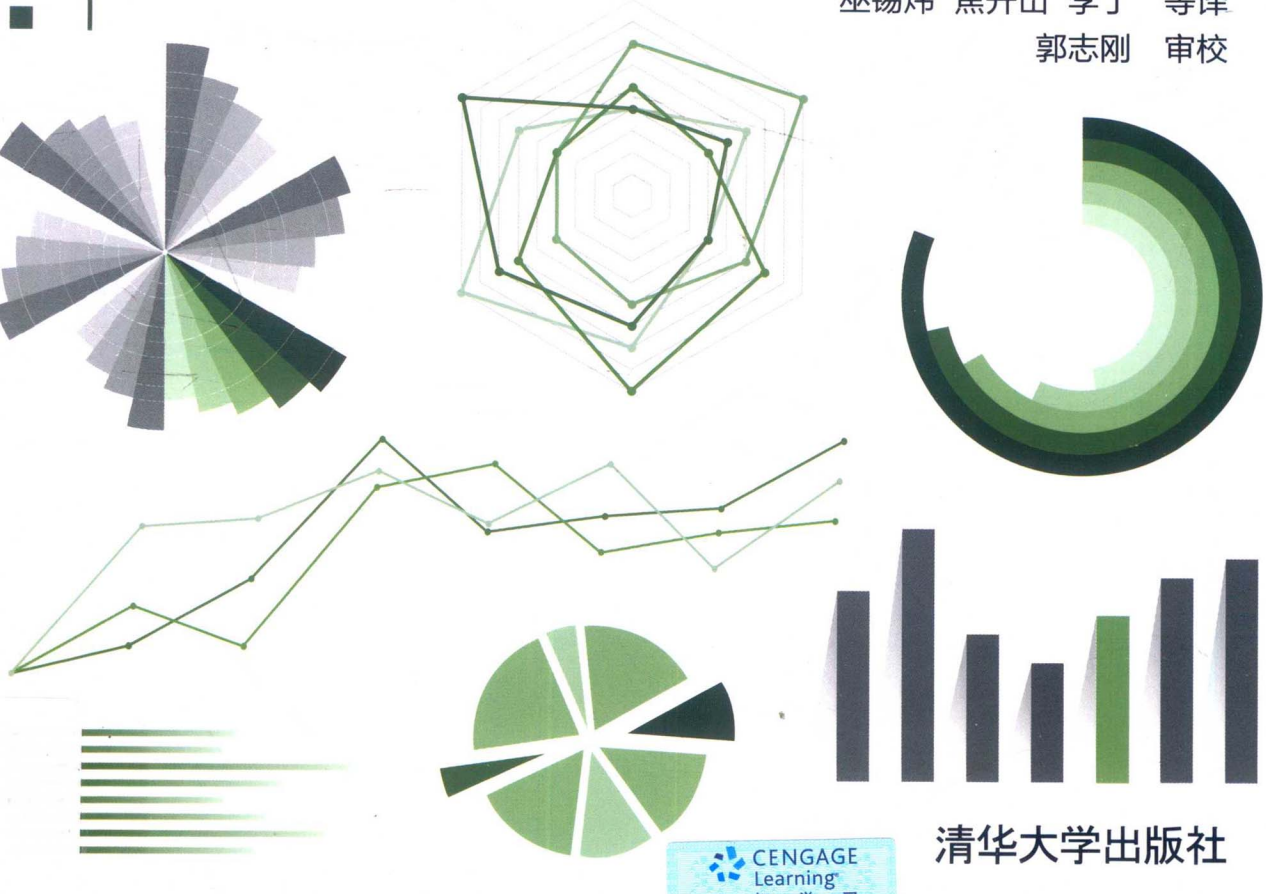
应用STATA做统计分析

更新至STATA 12（原书第8版）

Statistics with STATA

Version 12, Eighth Edition

[美] Lawrence C. Hamilton 著
巫锡炜 焦开山 李丁 等译
郭志刚 审校



清华大学出版社

CENGAGE Learning 圣智学习

应用 STATA 做统计分析

更新至 STATA 12（原书第 8 版）

[美] Lawrence C. Hamilton 著

巫锡炜 焦开山 李丁 等译

清华大学出版社

北 京

Lawrence C. Hamilton
Statistics with STATA: Version 12, Eighth Editon
EISBN: 978-0-8400-6463-9

ISBN: 978-7-302-46665-9
北京市版权局著作权合同登记号　图字：01-2016-2701
Cengage Learning Asia Pte. Ltd.
151 Lorong Chuan, #02-08 New Tech Park, Singapore 556741

图书在版编目(CIP)数据

应用 STATA 做统计分析：更新至 STATA 12：原书第 8 版/(美)劳伦斯·C. 汉密尔顿(Lawrence C. Hamilton) 著；巫锡炜，焦开山，李丁 等译. —北京：清华大学出版社，2017
书名原文：Statistics with STATA:Version 12, Eighth Editon
ISBN 978-7-302-46665-9

Ⅰ. ①应…　Ⅱ. ①劳… ②巫… ③焦… ④李…　Ⅲ. ①统计分析－应用软件　Ⅳ. ①C819

中国版本图书馆 CIP 数据核字(2017)第 036026 号

责任编辑：王　军　李维杰
装帧设计：牛艳敏
责任校对：曹　阳
责任印制：何　芊

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
印 刷 者：三河市君旺印务有限公司
装 订 者：三河市新茂装订有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　印　　张：27.75　　字　　数：746 千字
版　　次：2017 年 5 月第 1 版　　印　　次：2017 年 5 月第 1 次印刷
印　　数：1～4000
定　　价：69.80 元

产品编号：069147-01

中文版序

很高兴看到《应用 STATA 做统计分析》一书经巫锡炜、焦开山、李丁、赵联飞和王军等人的努力又一次被翻译成中文。此书的英文版一直非常成功，一版再版，所以读者现在阅读的已是其第 8 版的中文版。伴随 Stata 本身的发展，《应用 STATA 做统计分析》一书在每次修订后都会变得篇幅更长，并且覆盖更多的主题。借助此中文版，我希望初次偶读到《应用 STATA 做统计分析》一书的新读者也将会开卷有益。

熟悉更早各版本的读者们将会见到新的内容。新的介绍调查研究的一章出现在书的前面，因为社会科学领域的课程经常会涉及该主题。其他章节中的新增内容会介绍缺失值多重填补、结构方程建模、因子分在回归中的使用及混合效应建模的应用。最后一章介绍编程，内容做了简化，并围绕部分读者会觉得实用的一个主要例子(绘制多幅调查图形)来进行。

在本版的写作中，我也设法使用更有趣且最新的例子。比如，介绍时间序列分析的第 12 章使用了全球气温数据。它向读者说明了如何证实更大规模研究所得的主要结论：持续变暖的全球气温并不能为太阳辐照、火山爆发或自然变化(厄尔尼诺现象)所解释，而只有当我们考虑到持续攀升的二氧化碳浓度水平时才能得以理解。一些其他章也使用了环境主题的例子，从北极海冰到环境问题看法的调查，不同领域的读者或许都会对它们感兴趣。与这些例子相对应的数据可以从 Stata 网站的 Bookstore 处下载。

《应用 STATA 做统计分析》(1990)是第一本针对 Stata 而写的书。与 Stata 软件本身一样，此书也旨在做一些前人未做过的事情。我写这本书的目的是想为学生和研究人员弥合理论色彩浓厚的教材与 Stata 自带手册中数千页内容之间的差距。研究人员需要掌握分析其数据的各式技能。因此，《应用 STATA 做统计分析》一书从基本的主题开始，比如统计学导论课上的那些内容，或如何建立新的数据集。然后进入到中级和高级主题，诸如回归诊断、logit 模型、稳健回归、因子分析、生存分析、时间序列模型乃至编程。其中的一些可能出现在研究生的统计学课上，而另一些则可能会在开展研究项目过程中遇到。对于每一章，我都关注两个实用的问题：*我如何在 Stata 中进行该分析？所得结果告诉我什么？*我的目的是为读者写一本工作时会摆在其计算机旁的书。我经常收到 Google Scholar

发来的信息告诉我不同国家的人们确实如此，并且在他们自己发表的研究中引用本书。感谢巫锡炜、焦开山、李丁、赵联飞和王军为翻译此书所付出的努力，现在您有机会来判定它对您的用处了。

Lawrence C. Hamilton
2016 年 9 月

审校者序

看到巫锡炜、焦开山、李丁、赵联飞和王军等完成 Lawrence C. Hamilton 所著 *Statistics with STATA: Updated for Version 12, Eighth Edition* 一书的翻译工作并邀请我审校，我感到十分欣慰和高兴。欣慰的是因为他们都是我的学生，而我看到他们在博士毕业之后仍能在科研、教学之余做翻译统计学教材这种费力不讨好但却非常基础性的工作。在我看来，保持对知识、方法的不断学习和不断更新的渴望是一名研究者应当具备的基本素质，而专业文献翻译是学习和消化新知识、新方法的重要途径之一。高兴的是虽然我不再参与本书的翻译工作，但是审校过程中，他们 5 人的工作仍保持了之前的高质量，而熟悉本书的读者会看到，本版相对于以前版本在篇幅、所介绍内容和所用示例等方面做了幅度挺大的改动。

一如既往，本书保持了它的实用风格。第一，它介绍了社会研究人员常用的统计方法，从最基础的数据创建、变量改造等到诸如时间序列模型、生存分析、混合效应建模乃至结构方程建模等更复杂的建模技术。第二，它侧重 Stata 操作以及对统计分析结果的解读，从而在统计学教科书与 Stata 软件之间架起了一座桥梁。第三，更值得称道的是，原作者介绍每一统计方法时都使用自己或其他研究者所做的实际研究作为示例，从研究问题开始，到如何做数据分析，最后说明哪些分析结果回答了研究问题，娓娓道来，这几乎相当于教人如何完整实施一项研究课题，引人入胜。以我本人作为前面两版翻译组织者和本版审校者的经历，相信对定量社会研究方法感兴趣的读者朋友都能从中获益匪浅。

相比于之前我带领他们 5 位翻译的两个版本，这一版虽然在章数上有所减少，但内容上做了重新编排，结构上显得更紧凑。比如，原来有关线性回归、回归诊断、拟合曲线和稳健回归的数章被整合成两章；同时，还结合 Stata 软件 12.0 版的更新和统计方法的发展，新增了缺失值多重填补、结构方程建模以及复杂抽样设计下的调查数据等新内容，还介绍了一套非常有用的 Stata 模型拟合后续命令 **margins** 和 **marginsplot**。从我对译稿所做审读和校对工作来看，译者们不但对新增方法本身理解到位，而且对其 Stata 软件的实现也很熟练，从而确保了翻译的质量。专业文献的翻译首先要追求的是准确无误，这一版的翻译满足了这一标准。当然，就表述的精当、流畅而言，译者们仍有继续完善的空间。

郭志刚

2016 年 10 月

译者序

这是 *Statistics with STATA* 第 8 版的中译本，也是该书的第三个中文版。

前面两版都是我们几位在郭志刚教授的带领和指导下完成的。他一直用这种翻译专业文献的方式训练我们对统计方法的学习和掌握。我们都从中受益匪浅。这次我们之所以“劈腿”郭老师独立承担翻译工作，完全是他向出版社力荐的结果。这既是他对我们的认可，更是一种鼓励和帮助。感谢郭老师给予这一难得的机会！

Statistics with STATA 一书堪称 Stata 软件应用教材中的经典。自 1990 年以来，此书伴随着 Stata 软件的更新和统计方法的发展而一版再版。一本统计软件应用教材能够在图书市场存活 25 年并且越来越受读者欢迎，非经典之作而不能为，要知道这个时间差不多与 Stata 软件本身到目前已经存在 30 余年的时光一样长。当然，成就其为经典之作更重要的还在于下面两点。一是此书在形式上既兼顾必要但简洁易懂的统计学原理介绍，又从实际研究问题出发示例说明如何应用 Stata 软件完成数据分析并解读统计分析结果以回答研究问题，非常好地将让很多人觉得枯燥甚至深奥的统计学原理与看上去浩繁冗长的 Stata 软件手册融合起来。二是此书在内容上紧跟统计理论的发展和研究实践的需要，介绍大多数学科领域中最为实用的统计方法。比如，此版中既有属于“基本功”的数据管理方面的内容，结合 Stata 12.0 版的新功能，也涉及近年来日渐增多的混合效应建模、结构方程建模和缺失值多重填补等“进阶术”。所以，尽管此书不断修订再版，但始终能够让新老读者开卷有益。

翻译本身就是一件费力不讨好的事。对于 *Statistics with STATA* 这样的经典教材，翻译它更让人觉得有压力，尤其是前面还有郭老师之前的两个高质量译本。幸运的是，郭老师建议我们在之前译本的基础上完成翻译工作，甚至提供了翻译之前版本时创建的关键词中英文对照表，这大大方便了我们的翻译工作，翻译质量也有一定保证。加上他还亲自对译稿进行审校，更为翻译质量增加了一重保证。希望本次翻译仍能如之前郭老师亲自带领翻译的两个中译本那样受读者们的欢迎和好评。

本版的翻译工作从今年 4 月 11 日同清华大学出版社李万红、王军老师的第一次见面就开始启动。出于方便，由巫锡炜协调整个翻译工作。我们根据各自的兴趣和时间确定了任务分工：全书正文共 14 章，巫锡炜承担第 1、第 2、第 3、第 11、第 12、第 13 章以及书中的前言、中文版序言等内容，焦开山承担第 4、第 5、第 14 章，赵联飞承担第 6、第 7 章，李丁承担第 8、第 10 章，王军承担第 9 章。虽然各有分工，但是我们在翻译过程中相

互讨论，并对其他人的译稿提出修订意见。不过，非常遗憾的是，由于出版署名方面的一些限制，只有巫锡炜、焦开山和李丁作为译者署名出现，而具有近乎同样贡献的赵联飞和王军则被“等”取代了。

受专业水平和理解能力所限，翻译中难免有不当甚或舛误之处，恳请读者们指教和斧正！

巫锡炜、焦开山、李丁、赵联飞、王军

2016 年 10 月

前　言

《应用 STATA 做统计分析》一书旨在为学生和实际研究工作者在统计教材和 Stata 应用之间架设桥梁，以缩小两者之间的差距。为扮演这样一个中介角色，本书既不准备对某一合适教材做详细说明，也不打算尽可能地描述 Stata 的全部特征。相反，本书示范了如何使用 Stata 来完成各种各样的统计任务。每章的讨论遵循统计学概念主题展开，而并非只集中在特定的 Stata 命令上，这使得《应用 STATA 做统计分析》一书又具有与 Stata 参考手册不同的结构。比如，数据管理一章涉及了创建、导入、合并或改变数据文件结构的各种程序。有关图形、概要统计与表格，以及方差分析与其他比较方法的这几章也都包含诸多不同技术在内而又具有类似性的宽泛主题。本书将新的介绍调查数据(Survey Data)的一章放到了前面，为后续各章在恰当位置出现的更具技术性的调查数据示例提供了背景知识。

前 7 章(直到线性回归分析)为一般性主题，大体上对应了应用统计学中本科生或研究生一年级水平的课程，但是增加了深度，讨论了分析人员经常碰到的实际问题——比如，如何导入数据、绘制符合发表质量要求的图形、使用调查权重，或者解决回归中的问题。在第 8 章(高级回归)及随后各章中，我们转入高级课程或原创研究的领域。这里，读者能够找到有关 lowess 修匀、稳健回归、分位数回归、非线性回归、logit 模型、序次 logit 模型、多项 logit 模型或泊松回归的基本信息和举例说明；应用新方法进行结构方程建模(structural equation modeling)或缺失值多重填补(multiple imputation)；拟合存活时间和事件计数模型；根据因子分析或主成分结果构建和使用合成变量(composite variables)；将观测案例区分成不同的经验类型或聚类；分析简单或多元时间序列；以及拟合多层或混合效应模型。Stata 近年来一直致力于提升其一流地位，这种努力尤其体现在它现在所提供的各种各样的统计建模命令上。

本书最后介绍 Stata 编程的内容。许多读者将会发现 Stata 可以做他们想做的任何事情，因此他们不需要编写原始程序。但是，对于积极主动的少数人而言，编程能力也是 Stata 的主要吸引力之一，并且它也肯定构成了 Stata 广泛传播和快速发展的基础。第 14 章为想探索 Stata 编程的初学者开启了大门，不论是用于专业化的数据管理，还是建立一种新的统计方法以进行蒙特卡罗实验或教学。

通常，对于 Windows、Macintosh 和 Unix 等操作系统的计算机都有类似版本(“风格”)的 Stata 可以安装运行。在所有操作系统上，Stata 都使用相同的命令并形成相同的输出结

果。这些风格只是在屏幕外观、菜单和文件处理的一些细节上有些差异，这是因为 Stata 会遵循每一操作系统自己的规则——比如，Windows 系统下采用诸如“\目录\文件名”的文件设定，而在 Unix 系统下则采用“目录/文件名”的设定。本书并未示范所有三种规则，而只采用 Windows 规则，但是采用其他操作系统的用户应能发现，其实只需要稍加改变即可。

关于第 8 版的说明

笔者从 1985 年开始使用 Stata，当时还是它的首次发布年。起初，Stata 只在 MS-DOS 系统的个人电脑上运行，但其面向桌面的特点使得它明显比其主要竞争对手更现代，因为那时大多数竞争者还处于桌面革命之前，还基于主机环境、使用 80 列穿孔卡的 Fortran 语言。与认为每个用户都是一堆卡片的主机统计软件不同，Stata 将用户视为人机对话。它的互动本质以及统计程序与数据管理和制图的浑然一体支持了分析思维的自然流程，而这些方面则是其他程序所不具备的。**graph**(作图命令)和 **predict**(预测命令)很快成为倍受欢迎的命令。笔者深受其所有内容浑然一体打动，并开始写作《应用 STATA 做统计分析》的最初版本，该书对应着 Stata 第 2 版，并于 1989 年出版。Stata 在 2005 年迎来了它的 20 周年纪念，为此该年的《Stata 期刊》(*Stata Journal*)开辟了一期特刊，登载有关它发展史的文章和访谈，以及受邀而写就的《应用 STATA 做统计分析》一书的简史。

自该书第 1 版问世以来，Stata 已经发生了巨大变化。笔者在该书中就注意到，“Stata 并不是一个万能程序……但是只要是它做的事情，它就做得棒极了”。Stata 功能的扩展一直都引人注目。这一点在模型拟合程序的激增以及随后不断条理化方面显而易见。William Gould 为 Stata 建立的架构，包括其编程工具和统一的命令语法都已非常成熟，并已证明能够容纳新发展出来的统计思想。本书第 3 章广泛的作图命令、第 8 章开头提供的大量建模命令或者后续各章所介绍的新的时间序列分析、调查数据分析、多重填补或混合建模能力，都说明多年来 Stata 在这些方面日益变得丰富。比如，适用于面板数据(**xt**)、调查数据(**svy**)、时间序列数据(**ts**)、存活时间数据(**st**)或数据多重填补(**mi**)等的套装新技术开辟了更多可能领域，像一般化线性模型(**glm**)以及最大似然估计的一般程序中的可编程命令也同样做到了这点。其他重要扩展还包括矩阵编程能力的发展、大量新的数据管理特征以及诸如边际效应图(marginal plots)或结构方程建模等新的多用途分析工具。在最初版本的《应用 STATA 做统计分析》中，数据管理只是一个附带的话题；但它在本书的第 8 版中已经合乎情理地成为最长的一章。

Stata 全面的菜单和对话框系统提供了对大多数键入命令的点选式替代。不过，菜单和对话选择系列通过探索比通过阅读更易于学习，因此《应用 STATA 做统计分析》会在每章开头只提供有关菜单的一般性建议。绝大部分情况下都用命令来展示 Stata 能做什么；找到那些命令的对应菜单应非难事。相反，若你主要凭借菜单开始工作，Stata 会通过在结果窗口中呈现每一条相应的命令提供非正式训练。菜单/对话框系统通过将点选操作翻译成 Stata 命令，然后反馈给 Stata 并执行。

分析性制图是 Stata 的一大强项，这一点在每一章中都有体现。本书的许多例子都并非意在说明一种特定方法的单调图像，而都做了一些改进以满足发表或演示要求。读者或许会浏览这些图形以了解制图的潜力，这超出了 Stata 手册的内容。针对 Stata 12.0 更新的《应用 STATA 做统计分析》与之前针对 Stata 10.0 更新的该书大为不同。很多章已被重新组织，包括出现在本书前面新的介绍调查数据分析的一章。10.0 版的本书中分为 4 章的回归分析内容在这里已被更加逻辑性地整合和组织成篇幅更长的线性回归分析和高级回归两章。“高级回归”一章包含新的有关缺失值多重填补和结构方程建模(Structural Equation Modeling，SEM)的内容。主成分、因子和聚类分析一章也纳入两节新内容，介绍回归中因子得分的使用和 SEM 中测量模型的使用。分层与混合效应建模一章中新的一节呈现了一个重复测量数据分析的例子。有关编程的最后一章已被精简并围绕一个主要示例(绘制多幅调查数据图)来展开，可以证明这对于一些读者而言更有益。

本次针对 Stata 12.0 所做修订的一个目标是更新许多例子，其中一些涉及本人自 20 世纪 90 年代以来的研究，但已经过时。挑战者号航天飞机一例曾出现在最初 1989 年版的封面上，仍在 logistic 回归一章开头很好地说明基本思路。但是，该章的结尾为对 2011 年调查时收集到的人们关于气候变化的知识和观点的应答所做的加权多分类 logit 分析(weighted multinomial logit analysis)。气候调查是三个新的 2010 或 2011 调查数据集之一，这些数据集为若干章提供了重要的例子。其中一章(主成分和因子分析)以简单的行星数据开篇，但结尾则是使用 2011 年沿海环境调查数据所做的结合因子分析与回归的分析，或者类似的测量和结构方程模型。其他例子涉及物理学气候指标的时间序列。一个关于 42 个北极阿拉斯加村庄的独特数据集取自 2011 年的一篇论文，被用来示例说明混合效应建模如何可以将自然科学数据与社会科学数据结合起来。时间序列一章最后部分的 ARMAX 模型受到 2011 年一篇考察全球变暖“真实迹象”(real signal)的重要论文的启发。只要可能，都致力于使用提出大众感兴趣研究问题的例子，而不仅仅是提供一堆数字来示例说明一个技术。许多示例数据，包括书中所讨论之外的其他变量，吸引着读者自行去做进一步分析。

正如在第 1 章指出的，Stata 的帮助和搜索功能也与程序同步，得以完善。除了可以通过帮助文件获得的互动说明文档以外，可用资源还包括了 Stata 的网站、互联网及其文献搜索功能、用户社区邮件列表、网络课程、《Stata 期刊》以及 9000 多页的手册文档。《应用 STATA 做统计分析》提供了 Stata 的便捷入门，而这些其他资源将帮助你走得更远。

致谢

Stata 的设计师 William Gould 值得称赞，因为是他创建了《应用 STATA 做统计分析》所介绍的这个一流程序。Stata 公司的很多其他人员多年来贡献过他们的真知灼见。就此第 8 版而言，要特别感谢组织评阅工作的 Pat Branton 和阅读过绝大部分章节的 Kristin MacDonald。James Hamilton 为第 12 和 13 章的时间序列提出过重要建议。Leslie Hamilton 阅读并帮着修改了最终手稿的诸多部分。

本书围绕着数据分析的内容而写成。该版中新的一节对数据来源做了说明，包括存在的网页链接，或者所发表论文的索引。许多例子取自于公共资源，它们是其他研究者辛苦工作的成果。也借鉴了本人自己的研究，特别是一些新近的调查与整合自然和社会科学数据的研究。所有与本人一同开展这些项目的同事都值得称赞，包括 Mil Duncan 和 Tom Safford(CERA 农村调查)，Richard Lammers、Dan White 和 Greta Myerchin(阿拉斯加社区调查)，David Moore 和 Cameron Wake(气候环境调查)，Barry Keim 和 Cliff Brown(滑雪运动与气候环境研究)，以及 Rasmus Ole Rasmussen 和 Per Lyster Pedersen(格陵兰岛人口状况研究)。慷慨分享原始数据的其他人还有 Dave Hamilton、Dave Meeker、Steve Selvin、Andrew Smith 和 Sally Ward。

献给

Leslie、Sarah 和 Dave。

目　录

第 1 章　Stata 软件与 Stata 的资源……1
1.1　本书体例的说明……1
1.2　一个 Stata 操作的例子……2
1.3　Stata 的文件管理与帮助文件……6
1.4　搜寻信息……7
1.5　Stata 公司……8
1.6　《Stata 期刊》……9
1.7　应用 Stata 的图书……10

第 2 章　数据管理……13
2.1　命令示范……14
2.2　创建一个新的数据集……16
2.3　通过复制和粘贴创建新数据集……21
2.4　定义数据的子集：in 和 if 选择条件……22
2.5　创建和替代变量……25
2.6　缺失值编码……28
2.7　使用函数……31
2.8　数值和字符串之间的格式转换……34
2.9　创建新的分类变量和定序变量……37
2.10　标注变量下标……39
2.11　导入其他程序的数据……40
2.12　合并两个或多个 Stata 文件……43
2.13　数据分类汇总……46
2.14　重组数据结构……49
2.15　使用权数……52
2.16　生成随机数据和随机样本……53
2.17　编制数据管理程序……57

第 3 章　制图……59
3.1　命令示范……59
3.2　直方图……62
3.3　箱线图……65
3.4　散点图和叠并……68
3.5　曲线标绘图和连线标绘图……73
3.6　其他类型的二维标绘图……77
3.7　条形图和饼图……79
3.8　对称图和分位数图……82
3.9　给图形添加文本……84
3.10　使用 do 文件制图……86
3.11　读取与合并图形……87
3.12　图形编辑器……88
3.13　创造性制图……91

第 4 章　调查数据……99
4.1　命令示范……99
4.2　定义调查数据……100
4.3　设计权数……102
4.4　事后分层权数……104
4.5　调查加权的表格和图形……107
4.6　多重比较的条形图……110

第 5 章 概要统计及统计表 …… 115
5.1 命令示范 …… 115
5.2 测量变量的描述性统计 …… 117
5.3 探索性数据分析 …… 119
5.4 正态性检验和数据转换 …… 121
5.5 频数表和二维交互表 …… 124
5.6 多表和多维交互表 …… 127
5.7 均值、中位数以及其他概要统计量的列表 …… 129
5.8 使用频数权数 …… 131

第 6 章 方差分析和其他比较方法 …… 133
6.1 示范 …… 134
6.2 单样本检验 …… 135
6.3 两样本检验 …… 138
6.4 单因素方差分析 …… 140
6.5 双因素和多因素方差分析 …… 143
6.6 因素变量和协方差分析 …… 144
6.7 预测值和误差条形图 …… 147

第 7 章 线性回归分析 …… 151
7.1 命令示范 …… 151
7.2 简单回归 …… 155
7.3 相关 …… 158
7.4 多元回归 …… 161
7.5 假设检验 …… 165
7.6 虚拟变量 …… 167
7.7 交互效应 …… 170
7.8 方差的稳健估计 …… 175
7.9 预测值及残差 …… 177
7.10 其他案例统计量 …… 181
7.11 诊断多重共线性和异方差性 … 186
7.12 简单回归中的置信带 …… 188
7.13 诊断回归 …… 191

第 8 章 高级回归 …… 197
8.1 命令示范 …… 197
8.2 lowess 修匀 …… 199
8.3 稳健回归 …… 204
8.4 对 rreg 和 qreg 的更多应用 … 209
8.5 曲线回归 1 …… 212
8.6 曲线回归 2 …… 214
8.7 Box-Cox 回归 …… 219
8.8 缺失值的多重填补 …… 221
8.9 结构方程建模 …… 225

第 9 章 logistic 回归 …… 231
9.1 命令示范 …… 233
9.2 航天飞机数据 …… 234
9.3 使用 logistic 回归 …… 238
9.4 边际或条件效应标绘图 …… 241
9.5 诊断统计量与标绘图 …… 243
9.6 对序次 y 的 logistic 回归 …… 247
9.7 多项 logistic 回归 …… 249
9.8 缺失值的多重填补——logit 回归的例子 …… 256

第 10 章 生存模型与事件计数模型 … 259
10.1 命令示范 …… 260
10.2 生存时间数据 …… 262
10.3 计数时间数据 …… 264
10.4 Kaplan-Meier 存活函数 …… 266
10.5 Cox 比例风险模型 …… 268
10.6 指数回归与 Weibull 回归 … 273
10.7 泊松回归 …… 277
10.8 一般化线性模型 …… 280

第 11 章 主成分分析、因子分析和聚类分析 …… 285
11.1 命令示范 …… 286
11.2 主成分分析和主成分因子法 …… 287
11.3 旋转 …… 289
11.4 因子分 …… 292
11.5 主因子法 …… 294
11.6 最大似然因子法 …… 296

11.7　聚类分析-1 ····· 297
11.8　聚类分析-2 ····· 301
11.9　因子分在回归中的使用 ····· 305
11.10　测量与结构方程模型 ····· 312

第 12 章　时间序列分析 ····· 317
12.1　命令示范 ····· 317
12.2　修匀 ····· 319
12.3　时间标绘图的更多例子 ····· 325
12.4　最近的气候变化 ····· 328
12.5　时滞、前导和差分 ····· 331
12.6　相关图 ····· 336
12.7　ARIMA 模型 ····· 339
12.8　ARMAX 模型 ····· 346

第 13 章　多层与混合效应建模 ····· 351
13.1　命令示范 ····· 352
13.2　含随机截距的回归 ····· 354
13.3　随机截距和斜率 ····· 358
13.4　多个随机斜率 ····· 363
13.5　多层嵌套 ····· 366
13.6　重复测量 ····· 368
13.7　截面时间序列 ····· 371
13.8　混合效应 logit 回归 ····· 376

第 14 章　编程入门 ····· 383
14.1　基本概念与工具 ····· 383
14.2　程序示范：multicat(画出许多定类变量的图) ····· 393
14.3　使用 multicat ····· 396
14.4　帮助文件 ····· 400
14.5　蒙特卡罗模拟 ····· 403
14.6　用 Mata 进行矩阵编程 ····· 410

数据来源 ····· 415

参考文献 ····· 419

第 1 章

Stata 软件与 Stata 的资源

Stata 是用于 Windows、Mac 以及 Unix 操作系统上的一种功能完备的统计软件包。它的特点包括易操作、速度快，还包括一整套预先编好的分析与数据管理功能，同时也允许用户根据需要来创建自己的程序、添加更多功能。大部分操作既可以通过下拉菜单系统来完成，也可以更直接地通过键入命令来完成。初学者可以在菜单的帮助下学习使用 Stata，任何人在应用自己所不熟悉的程序时都可以由此获得帮助。Stata 的命令有很强的一致性和直观意义，可以使有经验的用户更高效地工作，这一特点还使得对更复杂或需要多次重复的任务进行编程变得十分容易。如有必要，在应用 Stata 时还可以混用菜单方法和命令方法。它还提供广泛的帮助、查找和链接功能，轻轻松松便能完成像查询某一命令语法或其他信息这类的事情。本书即为补充这些特征而著。

本书先提供一些介绍性信息，然后我们从一段 Stata 应用示范来让你领略数据分析过程，以及怎样使用分析结果。后续各章将做更详细的解释。然而，即使没有任何解释，你也可以看到有关命令多么简单明了：打开数据文件 *filename* 的命令就是 **use** *filename*，取得概要统计的命令是 **summarize**，得到相关矩阵的命令是 **correlate**，如此等等。或者，也可以通过 Data 或 Statistics 菜单上的选择来取得同样的结果。

有各种各样的资源来帮助用户学习 Stata，以解决任何难度级别的问题。这些资源并不只是来自于 Stata 公司，而且也来自于活跃的 Stata 用户群体。本章的一部分内容会介绍一些主要资源：包括 Stata 的在线帮助和印刷版文档，以及寻求技术帮助时应该给哪里写信或发电子邮件，提供包括软件更新与常见问题解答等诸多服务的 Stata 网址(www.stata.com)、互联网论坛 Statalist Internet 以及经审阅的《*Stata* 期刊》(Stata *Journal*)。

1.1 本书体例的说明

本书采用几种不同的印刷体例来标志有关文字的类型意义：

- 用户键入的命令以粗体显示。当给出完整的命令行时，将以一个英文句点作为起始点，这与在 Stata 结果窗口(Results window)或日志(输出)文件中见到的一样：

```
. correlate extent area volume temp
```

- 命令中的变量或文件名均为斜体，以强调它们是随意指定的，而并不是该命令的固定部分。
- 本书一般行文中涉及变量或文件名时也将以斜体显示，以示它们与普通文字内容的区别。
- Stata 菜单上的项将以 Arial 体表示，以“→”间隔表示随后的选项。比如，我们可以通过选择 File→Open 来打开已存在的数据集，然后找到并单击这一特定数据集的文件名。注意，一些常见菜单的动作也可以通过 Stata 主菜单工具条中的文字选项来完成：

  File　Edit　Data　Graphics　Statistics　User　Window　Help

或者单击这些文字下面相应的图标来完成。比如，选择 File → Open 与单击最左侧的开启文件夹小图标 的功能完全一样。用户还可以直接键入以下命令来实现同样的操作：

```
. use filename
```

于是，我们呈现名为 *extent* 的一个变量的概要统计指标的计算结果如下：

```
. summarize extent
```

Variable	Obs	Mean	Std. Dev.	Min	Max
extent	33	6.51697	.9691796	4.3	7.88

这些体例只适用于本书，而不适用于 Stata 程序本身。Stata 可以显示不同的屏幕字体，但是它在命令中并不使用斜体。一旦 Stata 的日志文件被导入文字处理软件，或者已将统计结果表复制并粘贴到文字处理软件，就应该将其格式改为 Courier 字体的 10 号或更小字号，这样才能将各列正确对应。

对于命令和变量名，Stata 严格区分大小写。所以 **summarize** 是一个命令，而 Summarize 和 SUMMARIZE 就不是。*Extent* 和 *extent* 将是两个不同的变量。

1.2　一个 Stata 操作的例子

作为对运行中 Stata 的一个预览，本节将介绍如何打开和分析一个以往创建的数据文件，名为 *Arctic9.dta*。这一小规模时间序列涵盖了卫星时代(1979 年到 2011 年)对 9 月份北冰洋冰情的观测。数据取自三个不同来源(见有关数据来源的附录)。变量 *extent* 是对每年 9 月份海冰密集度不低于 15%的北半球海域的卫星测量。*Area* 数字略小于 *extent*，表示海冰本身的面积。另一个变量 *tempN* 记录了北纬 64°以北平均年度表面气温。气温被表达为以摄氏度衡量的异常，即与 1951 年到 1980 年平均气温的偏差。我们有 33 个观测(年份)和 8 个变量。

如果我们想记录下这段工作，最好的方法是在工作开始时先打开一个日志文件。日志文件可以存放命令和统计结果表，但是不能存放图形。要建立一个日志文件，先从顶部菜单栏中选择 File→Log→Begin...，并为这个输出的日志文件指定文件名和文件夹。也可以