高等学校经济学类核心课程教材

计量经济学及Stata应用

陈 强 编著

高等教育出版社

高等学校经济学类核心课程教材

计量经济学及Stata应用

陈 强 编著

Jiliang Jingjixue ji Stata Yingyong

高等教育出版社·北京

内容简介

本书为既接轨现代计量经济学，又适合中国国情的本科计量经济学教材。在理论体系上，本书充分借鉴最新国际主流教材，以大样本理论为主线，并针对中国学生的知识体系进行编写。本书内容全面，包括横截面数据（多元回归、工具变量法、离散选择）、时间序列（平稳时间序列、单位根、协整），以及面板数据（随机效应、固定效应）等。

本书力图以清晰而生动的语言、较多的插图与经济意义，来直观地解释计量方法。同时结合目前欧美最为流行的 Stata 计量软件，及时地介绍相应的计算机操作与经典实例，为读者提供“一站式”服务。本书还较多地使用计算机模拟（蒙特卡罗法），作为强有力的学习工具。

本书适合高等学校经济管理类及社科类的本科生使用。先修课为微积分、线性代数与概率统计。阅读本书可使读者掌握当代实证研究的精神实质与基本方法，并学会实际处理数据的重要技能，从而为毕业论文乃至读研深造打下良好基础。

图书在版编目（CIP）数据

计量经济学及 Stata 应用／陈强编著. --北京：高等教育出版社，2015.7

ISBN 978-7-04-042751-6

Ⅰ. ①计… Ⅱ. ①陈… Ⅲ. ①计量经济学-应用软件-高等学校-教材 Ⅳ. ①F224.0-39

中国版本图书馆 CIP 数据核字（2015）第 101233 号

策划编辑 施春花	责任编辑 施春花	封面设计 杨立新	版式设计 杜微言
插图绘制 于 博	责任校对 陈 杨	责任印制 毛斯璐	

出版发行	高等教育出版社	咨询电话	400-810-0598
社 址	北京市西城区德外大街 4 号	网 址	http://www.hep.edu.cn
邮政编码	100120		http://www.hep.com.cn
印 刷	国防工业出版社印刷厂	网上订购	http://www.landraco.com
开 本	787mm×1092mm 1/16		http://www.landraco.com.cn
印 张	22.5	版 次	2015 年 7 月第 1 版
字 数	550 千字	印 次	2015 年 7 月第 1 次印刷
购书热线	010-58581118	定 价	39.00 元

物 料 号 42751-00

前　言

自从1995年我以北大硕士生身份留校，并执教“统计学”与“计量经济学”课程以来，至今20年已匆匆过去了。期间，西方国家的计量经济学突飞猛进，国内的计量经济学教学也取得了长足进步。而我自己也赴美求学多年，并于2008年加盟山东大学经济学院，再次执教“计量经济学”（本科生、硕士生与博士生）课程。

在山东大学的最初几年，我一直使用Stock and Watson的*Introduction to Econometrics*作为本科计量经济学教材，并以Wooldridge的*Introductory Econometrics: A Modern Approach*作为辅助参考，这是目前国际上最为流行的两本本科计量经济学教材。这两本教材的作者均为国际计量经济学大师，分别执教于哈佛大学、普林斯顿大学与密歇根州立大学，因此他们的教材能很好地贴近当代计量经济学的主流方法，对于细节的交代也很清晰到位。

然而，在赞叹之余，也渐渐发现这两本国际畅销教材并不太适合中国的国情。首先，这两本书的英文版均超过800页，教师无法在一学期教完，而学生甚至无法通读一遍。由于计量经济学本身技术性较强，再加上英文的语言障碍，则是难上加难。虽然，这两本教材均有不同的中译本，但面对巨大的翻译工作量，译者们往往无法长时间去润色打磨，不仅难以达到“信达雅”的理想境界，甚至与我们所熟悉的汉语阅读习惯也大相径庭。其次，由于美国大学数学课开设较晚，而这两本教材主要面向美国读者，故在正文尽量回避数学，而把矩阵形式的计量模型放到最后的章节甚至附录，导致不得不用大量的语言来解释计量理论，致使教材篇幅过长。另外，中国学生在学习计量经济学之前，一般已学过微积分、线性代数与概率统计，如果不适当地运用这些数学，则是巨大的资源浪费与效率损失。

如何才能让中国学生更轻松地掌握计量经济学的当代知识？单纯地依赖国际教材及其中译本似乎并非捷径。时下，在中国的高等教育界，国际化是一股轰轰烈烈的潮流。许多人认为，国际化就是采用英文教材，进行双语甚至全英文教学。或许，这适用于一小部分学生，比如全英文实验班，但对于大多数中国学生而言，恐怕会出现水土不服，甚至事倍功半的效果。笔者以为，国际化的一个较高境界其实是本土化，即将国际知识洋为中用，以汉语的形式最快地让绝大部分中国学生直接受益。

为此，在编著畅销研究生教材《高级计量经济学及Stata应用》（高等教育出版社，2010年第1版，2014年第2版）的基础上，我决心为广大中国学生编写一本既通俗易懂、贴近主流，又极具操作性的本科计量经济学教材。本书的主要特色可概括如下：

（1）接轨现代计量经济学。本书较多地借鉴了Baum（2006），Dougherty（2011），Gujarati and Porter（2008），Hill *et al.*（2011），Kennedy（2008），Maddala and Lahiri（2009），Stock and Watson（2012），Studenmund（2010），Wooldridge（2009）等国际流行的本科教材，其中尤以Stock and Watson（2012）与Wooldridge（2009）对本书的影响最深。另外，还借鉴了一些经典的研究生教材，比如Cameron and Trivedi（2005，2010），Greene（2012），Hayashi（2000），Poirier（1995），

Verbeek(2004), Wooldridge (2010),以期高屋建瓴。

(2) 在内容上坚持不灌水。长期以来,出于教学上的权宜之计,国内本科教材常常做一些不现实的假设。比如,假定解释变量为固定、非随机的(fixed regressors),这固然使问题简化便于理解,但现实中的经济变量几乎都是随机的。假设解释变量非随机,便无从讨论解释变量与扰动项的相关性,给后续教学造成莫大障碍。又比如,目前国际计量学界的主流方法已是大样本理论,而多数国内教材仍侧重于小样本理论。小样本理论的严苛假设(比如,严格外生性、正态分布),使得它在现实中很难应用。为此,本书坚持在内容上不灌水,自始至终假设随机解释变量,并以大样本理论为核心,将当代计量的主流方法介绍给学生。试想,如果只给学生灌水的内容,却要求他们去看最新的文献,完成高质量的毕业论文,则无异于赶着未经良好训练的士兵上战场。"工欲善其事,必先利其器",故需将最好的计量工具介绍给学生。

(3) 理论与实践紧密结合。学习计量经济学的学生,既需要了解计量经济学原理,也需要知道如何在计算机上实现,掌握处理数据的实际技能。为此,本书提供一站式服务,在讲解每个估计方法后,随即介绍相应的 Stata 计算机操作(Stata 为目前欧美最为流行的计量软件),并深入分析有趣的经典实例,便于读者迅速掌握相应的理论与操作。本书还较多地使用计算机模拟(蒙特卡罗法),作为强有力的学习工具,直观地呈现计量理论与结果。

(4) 尽量使用清晰、通俗而生动的语言。在某些方面,写作计量经济学本科教材的难度甚至超过研究生教材,因为前者无法像后者那样自由地使用矩阵等数学工具。当然,一味地回避数学显然不可取,因为最精确的理解仍然依赖于数学表达式。另外,对于必不可少的数学公式,则应辅以清晰、通俗而生动的语言,乃至插图,加以直观地解释。为了清晰地表述某个理论或思想,有时需要多番修改提炼,以找到最为直指人心的表达方式。

在本书出版之际,特别感谢以下曾教授过我统计学或计量经济学的授业恩师们(以时间先后为序):范培华、胡健颖、靳云汇、陈良焜 (北京大学);Dale Poirier (University of California, Irvine);Susan Porter-Hudak, Nader Ebrahimi, Mohsen Pourahmadi (Northern Illinois University)。没有他们的谆谆教诲,本书是绝不可能完成的。

山东大学经济学院的同事与学生们对本书的写作给予了大力支持。常东风副教授、唐明哲副教授、王永副教授、韩青博士、孔建宁博士、薛欣欣博士、博士生张博,以及硕士生李昱璇、刘春雨、卢秋全、毛会贞、孟鸽、孙丰凯、徐艳娴等参加了本书的校对,并提出了很好的修改意见,在此表示衷心感谢(当然,文责自负)。最后,特别感谢高等教育出版社的施春花编辑及其同仁们,为保证本书的高质量,他们付出了辛勤的劳动。

当然,由于笔者学识有限,对于本书的错漏之处,恳请读者及时指出,以便在网上公布勘误表,并在未来的版本中更正。联系邮箱为 qiang2chen2@126.com。

本书用到的所有数据集以及勘误表,均可在我的个人网页下载:http://econ.sdu.edu.cn/tree/content.php?id=52841。

陈强

2015 年 3 月

目　录

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write. ——H. G. Wells

There are three kinds of lies: lies, damn lies, and statistics. ——Benjamin Disraeli

1. 导　论

1.1 什么是计量经济学

"计量经济学"(econometrics[①],也译为"经济计量学"),顾名思义,是运用概率统计方法对经济变量之间的(因果)关系进行定量分析的科学。之所以把"因果"两个字加括号是因为,一方面,由于实验数据的缺乏,计量经济学常常不足以确定经济变量之间的因果关系;另一方面,大多数实证分析的目的恰恰正是要确定变量之间的因果关系(即 X 是否导致 Y),而非仅仅是相关关系。因此,在学习与应用计量经济学的过程中,很有必要时时以"因果关系"作为思考的框架与指引。

例(相关关系)　你看到街上的人们带雨伞,于是预测今天要下雨。但这只是一种相关关系,因为"人们带伞"并不是造成"下雨"的原因。

例(相关关系)　根据与流感相关的海量词条搜索记录,谷歌公司通过分析大数据(big data),可以很快地预测流行病的地域传播(参见本章附录)。但这也只是相关关系,因为上网搜索流感信息并不导致流感的传播。

由以上两例可知,如果我们只对预测感兴趣,则相关关系就足够了。然而,如果为了推断变量之间的因果关系,则计量分析必须建立在经济理论的基础之上,即在理论上存在 X 导致 Y 的作用机制。然而,即使有理论基础,因果关系常常依然不好分辨。首先,可能存在"逆向因果关系"(reverse causality)或"双向因果关系"。

例(逆向因果)　FDI(外商直接投资)促进经济增长,但 FDI 也被吸引到快速增长的地区。

例(逆向因果)　收入增加引起消费增长,而消费增长也拉动收入增加。

例(逆向因果)　经济萧条可能引起内战,但内战也会导致经济停滞。

其次,被遗漏的第三个变量(Z)也可能对这两个变量(X,Y)同时起作用,参见图 1.1。

例(遗漏变量)　某外星人来到地球,发现人类会死亡,十分不解。于是开始在全球广泛观察死亡现象,并收集了大量的数据。结果发现,许多人类躺在医院病床(X)之后死去(Y),故推断医院病床是死亡的原因。外星人认为,由于躺在医院病床上,总是发生在死亡之前,故不可能

① 其中,"econ"表示经济,"metrics"表示度量或测量,故"econometrics"的字面意思为"economic measurement",即"经济度量"。而计量经济学家则称为"econometrician"。

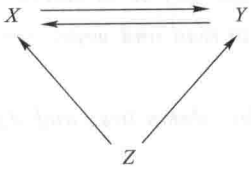

图 1.1 可能的因果关系

存在逆向因果关系。外星人于是将研究报告投稿发表于某顶尖经济学期刊,并在文末给出政策建议"珍爱生命,远离病床"①。在此例中,遗漏的变量 Z 是什么?

例(遗漏变量) 考虑决定教育投资回报率(returns to schooling)的因素:

$$\ln w_i = \alpha + \beta s_i + \varepsilon_i \tag{1.1}$$

其中,$\ln w_i$(工资对数)为"被解释变量"(dependent variable)②。s_i(schooling,教育年限)为"解释变量"(explanatory variable,regressor)、"自变量"(independent variable)或"协变量"(covariate)。ε_i 为不可观测(unobservable)的"误差项"(error term)或"随机扰动项"(stochastic disturbance),包括所有除 s_i 以外对 $\ln w_i$ 有影响的因素,以及人类行为的随机性。下标 i 表示第 i 个观测值(即个体 i)。截距项 α 与斜率 β 为待估参数。其中,β 的经济含义为教育投资的回报率,即多上一年学,未来工资能增加百分之几(参见第 4 章)。

如果用数据估计一元回归方程(1.1),其结果一般会显示,对数工资与教育年限显著正相关,而且教育投资回报率 β 还较高。然而,一个人的工资收入也与能力有关,但能力一般不能直接观测,而能力高的人通常选择接受更多教育。因此,在这个简单的回归中,教育的高回报其实包含了对能力的回报。

进一步,影响工资收入的因素还可能包括工作经验、毕业学校、人种、性别、外貌等。因此,需要尽可能多地引入"控制变量"(control variables),也就是多元回归的方法,才能较准确地估计我们"感兴趣的参数"(parameters of interest),即本例中的教育投资回报率 β。然而,现实中总有某些相关的变量无法观测,即存在"遗漏变量"(omitted variables),而这些遗漏变量统统被纳入到随机扰动项 ε_i 中了。

随机扰动项 ε_i 中还可能包含哪些其他因素呢?如果真实模型(true model)为

$$\ln w_i = \alpha + \beta s_i + \gamma s_i^2 + \varepsilon_i \tag{1.2}$$

那么 γs_i^2 也被纳入到扰动项中了(可以视为广义的遗漏变量)。如果变量测量得不准确,则测量误差也被放入扰动项中了。总之,一方面,扰动项就像是一个"垃圾桶",所有你不想要、无法把握的东西都往里面扔;另一方面,我们又希望扰动项拥有很好的性质。在很多情况下,这是自相矛盾的。西方有个谚语"The devil is in the details",意即"魔鬼就在细节中"。套用到计量经济学上来,或许可以说"The devil is in the error term",意即魔鬼就在扰动项中。计量经济学的很多玄妙之处就在于扰动项。如果真正理解了扰动项,也就加深了对计量经济学的理解。

---

① 此例来自香港大学商学院周文教授,参见 http://www2.fbe.hku.hk/staff/wzhou/。

② 之所以使用工资对数而不用工资作为被解释变量,是基于劳动经济学(labor economics)的理论模型,而且对工资对数建模也与经验数据更为吻合,参见第 4 章。

1.2 经济数据的特点与类型

由于在经济学中通常无法像自然科学那样做“控制实验”(controlled experiment),故经济数据一般不是“实验数据”(experimental data),而是自然发生的“观测数据”(observational data);比如,统计局所收集的数据。由于个人行为的随机性,所有经济变量原则上都是随机变量[①]。

在有些本科计量教材中,为了简单起见,有时假设解释变量是非随机的、固定的(fixed regressors)。这只是教学法上的权宜之计,却给更深入的理论探讨带来了不便。比如,如果解释变量为非随机,则无法考虑其与扰动项的相关性。因此,在本书中,所有变量都是随机的(即使非随机的常数,也可视为退化的随机变量)。

经济数据按照其性质,可大致分成以下三种类型。

(1) 横截面数据(cross-sectional data,简称截面数据):指的是多个经济个体的变量在同一时点上的取值。比如,2013 年中国各省(直辖市、自治区)的 GDP,参见表 1.1。

表 1.1 2013 年中国分省(直辖市、自治区)GDP 单位:亿元

省(直辖市、自治区)	GDP	省(直辖市、自治区)	GDP
北京	19 500.56	湖北	24 668.49
天津	14 370.16	湖南	24 501.67
河北	28 301.41	广东	62 163.97
山西	12 602.24	广西	14 378.00
内蒙古	16 832.38	海南	3 146.46
辽宁	27 077.65	重庆	12 656.69
吉林	12 981.46	四川	26 260.77
黑龙江	14 382.93	贵州	8 006.79
上海	21 602.12	云南	11 720.91
江苏	59 161.75	西藏	807.67
浙江	37 568.49	陕西	16 045.21
安徽	19 038.87	甘肃	6 268.01
福建	21 759.64	青海	2 101.05
江西	14 338.50	宁夏	2 565.06
山东	54 684.33	新疆	8 360.24
河南	32 155.86		

资料来源:国家统计局网站. http://data.stats.gov.cn/workspace/index?m=fsnd

(2) 时间序列数据(time series data):指的是某个经济个体的变量在不同时点上的取值。

① 你能举出哪些经济数据(变量)不是随机变量吗?

比如,1994—2013 年山东省每年的 GDP,参见表 1.2。

表 1.2 1994—2013 年山东省 GDP 单位:亿元

年份	GDP	年份	GDP
1994	3 844.5	2004	15 021.8
1995	4 953.35	2005	18 366.9
1996	5 883.8	2006	21 900.2
1997	6 537.07	2007	25 776.9
1998	7 021.35	2008	30 933.3
1999	7 493.84	2009	33 896.6
2000	8 337.47	2010	39 169.9
2001	9 195.04	2011	45 361.9
2002	10 275.5	2012	50 013.2
2003	12 078.2	2013	54 684.3

资料来源:国家统计局网站. http://data.stats.gov.cn/workspace/index?m=fsnd

(3) 面板数据(panel data):指的是多个经济个体的变量在不同时点上的取值。比如,1994—2013 年中国各省(直辖市、自治区)每年的 GDP,参见表 1.3。

表 1.3 1994—2013 年中国分省(直辖市、自治区)GDP 单位:亿元

省(直辖市、自治区)	年份	GDP
北京	1994	1 145.31
北京	1995	1 507.69
⋮	⋮	⋮
北京	2012	17 879.4
北京	2013	19 500.56
天津	1994	732.89
天津	1995	931.97
⋮	⋮	⋮
天津	2012	12 893.88
天津	2013	14 370.16
⋮	⋮	⋮
新疆	1994	662.32
新疆	1995	814.85
⋮	⋮	⋮
新疆	2012	7 505.31
新疆	2013	8 360.24

资料来源:国家统计局网站. http://data.stats.gov.cn/workspace/index?m=fsnd

本书介绍的计量经济理论包括以上三种数据类型，并使用国际上最为流行的 Stata 计量软件进行数据处理（Stata 13 版本，2013 年发布）。为此，我们将在第 2 章介绍 Stata 软件。第 3 章将回顾相关数学知识，并引入一些新概念（比如，均值独立、迭代期望定律）。有了这些铺垫之后，第 4 章将正式进入计量经济学的理论部分。

附录 A1.1　谷歌如何通过搜索记录预测流感的传播

2009 年 3 月底，一种新流感甲型 H1N1 流感（最初命名为“人感染猪流感”）在墨西哥和美国加利福尼亚州、得克萨斯州爆发，并在全球不断蔓延。这种新型病毒的基因中包含猪流感、禽流感和人流感三种流感病毒的基因片段。截至 2010 年 5 月底，出现疫情的国家和地区达到 214 个，持续一年多的疫情造成约 1.85 万人死亡。①

由于缺乏对抗这种新型流感病毒的疫苗，公共卫生专家所能做的只是减慢其传播速度，而这取决于知道这种流感出现在哪里。在美国，虽然要求医生在发现新型流感病例时告知疾控中心（Centers for Disease Control and Prevention），但人们可能患病多日才去医院，而此信息传到疾控中心也要时间，故通告新流感病例往往有一两周的延迟。然而，对于迅速传播的流行病，信息滞后两周可能带来致命的后果。

能否找到“预测”流行病的更快方法？迈尔 - 舍恩伯格与库克耶（2013，p. 2 - 4）在畅销书《大数据时代》介绍谷歌如何通过搜索记录来更快更准地预测流感的传播：

> 在甲型 H1N1 流感爆发的几周前，互联网巨头谷歌公司的工程师们在《自然》杂志上发表了一篇引人注目的论文。它令公共卫生官员们和计算机科学家感到震惊。文中解释了谷歌为什么能够预测冬季流感的传播：不仅是全美范围的传播，而且可以具体到特定的地区和州。谷歌通过观察人们在网上的搜索记录来完成这个预测，而这种方法以前一直是被忽略的。谷歌保存了多年来所有的搜索记录，而且每天都会收到来自全球 30 亿条的搜索指令，如此庞大的数据资源足以支撑和帮助它完成这项工作。
>
> 谷歌公司把 5 000 万条美国人最频繁检索的词条和美国疾控中心在 2003—2008 年间季节性流感传播时期的数据进行了比较。……他们设立的这个系统唯一关注的就是特定检索词条的使用频率与流感在时间和空间上的传播之间的联系。谷歌公司为了测试这些检索词条，总共处理了 4.5 亿个不同的数学模型。在将得出的预测与 2007 年、2008 年美国疾控中心记录的实际流感病例进行对比后，谷歌公司发现，他们的软件发现了 45 条检索词条的组合，将它们用于一个特定的数学模型后，他们的预测与官方数据的相关性高达 97%。和疾控中心一样，他们也能判断出流感是从哪里传播出来的，而且判断非常及时，不会像疾控中心那样要在流感爆发后一两周才可以做到。
>
> 所以，2009 年甲型 H1N1 流感爆发的时候，与习惯性滞后的官方数据相比，谷歌成为了一个更有效、更及时的指示标。公共卫生机构的官员获得了非常有价值的数据信息。

① 详见维基百科：http://zh.wikipedia.org/wiki/。

Computers are useless. They can only give you answers. ——Pablo Picasso

2. Stata 入门

2.1 为什么使用 Stata

Stata 软件因其操作简单且功能强大，成为目前在欧美最流行的统计与计量分析软件，拥有为数众多的用户。Stata 公司也通过定期升级软件，以适应计量经济学的迅猛发展。同时，Stata 软件留有“用户接口”，允许用户自己编写命令与函数，并上传到网上实现共享。因此，对于一些最新的计量方法，可以在线查找和下载由用户编写的 Stata 命令程序（user-written Stata commands）。这些“非官方命令”（也称“外部命令”）的使用方法与官方命令完全相同，使得 Stata 的功能如虎添翼，深受用户的喜爱。

本书使用 Stata 13 版本（2013 年 6 月发布）。对于绝大多数的命令与功能，即使你用更低的 Stata 版本（比如 Stata 11 或 Stata 12），也几乎没有差别。即使你没有任何基础，通常只需要半天时间，看完本章内容并亲自操作一遍，就可以实现 Stata 入门的要求（后续学习可随着本书而逐渐深入）。

2.2 Stata 的窗口

安装 Stata 13 后，在安装的文件夹中将出现如图 2.1 所示的 Stata 13 图标（Stata 11 或 Stata 12 的图标大同小异）。

图 2.1 Stata 13 的图标

双击此 Stata 图标，即可打开 Stata。如果想在计算机桌面创建一个开启 Stata软件的快捷方式，可以右键点击 Stata 13 的图标，然后选择“发送到”→“桌面快捷方式”，参见图 2.2。

打开 Stata 后可看到，在最上方有一排“下拉式菜单”（pull-down menu），参见图 2.3。

点击图 2.3 中的任何选项，都会弹出一个“级联式”菜单，每个选项之下还可能有子菜单。在 Stata 中运行单个命令主要有两种方式，其一为点击菜单，其二为在“命令窗口”输入命令（参见下文）。通过菜单执行命令（menu-driven）可能要点击多重菜单，通常最后还要填写一个对话框（dialog），以明确命令的参数，故一般不如在命令窗口直接输入命令更为方便有效。

在菜单之下，为一系列图标，起着快捷键的作用，参见图 2.4。只要将鼠标放在图 2.4 中的任何快捷键上，就会显示其相应的功能。这些快捷键的具体用法，将在下文逐步介绍。

在快捷键图标之下，有五个窗口，参见图 2.5。其中，左边为“历史命令窗口”（Review），记录启动 Stata 后用过的命令。中间的大窗口为“结果窗口”（Results），显示执行 Stata 命令后

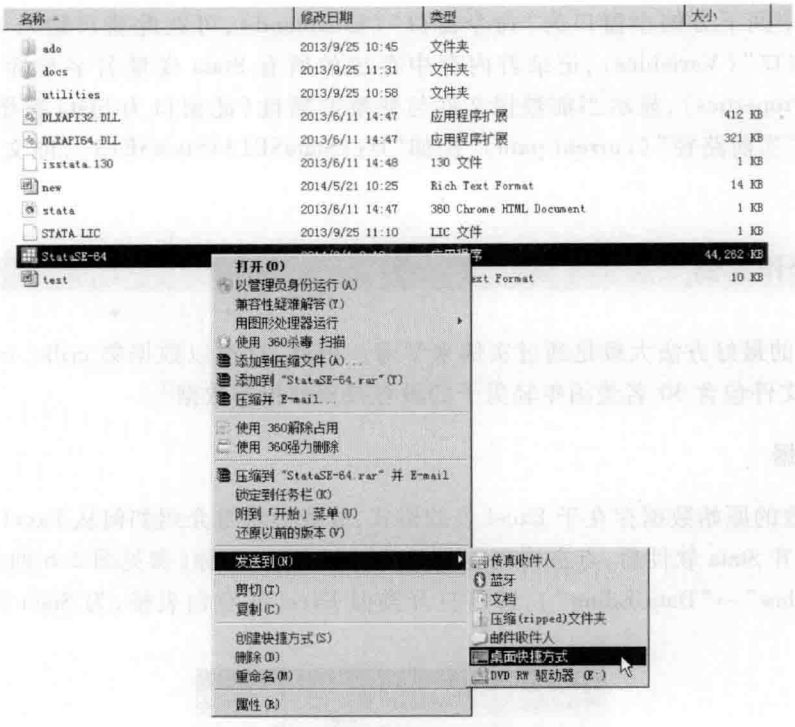

图 2.2 发送 Stata 13 到桌面快捷方式

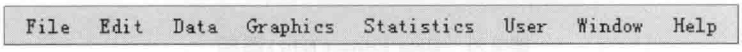

图 2.3 Stata 的下拉式菜单

图 2.4 Stata 的快捷键

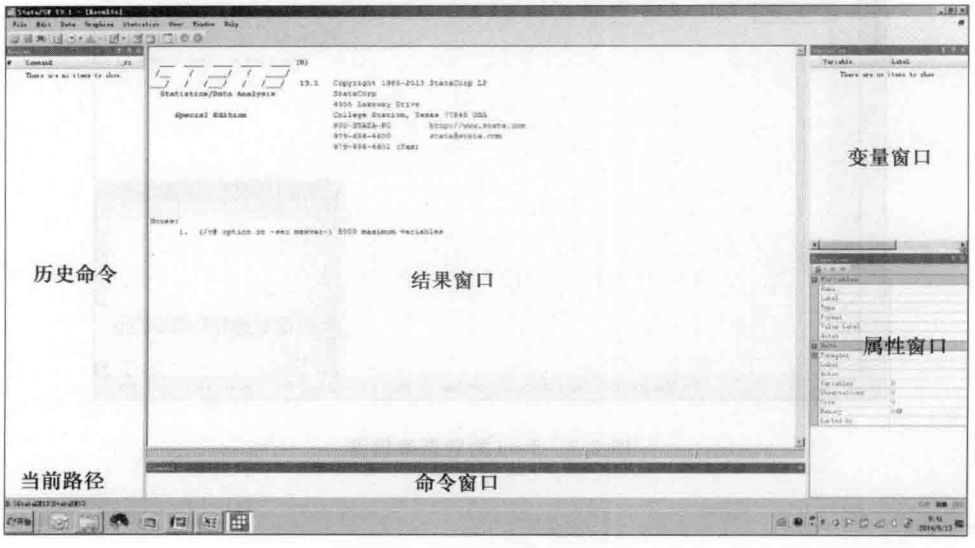

图 2.5 Stata 13 的主要窗口