

Macroeconomic modeling for all

应用计量经济学讲稿

作者：许文立

组织：安徽大学经济学院、东北财经大学国民经济工程实验室（北京）

西蒙弗雷泽大学（温哥华）、宏观经济研学会

时间：Sept 2, 2021

版本：2.1

Email: xuweny87@hotmail.com



老师的真正重任应该是让学生超越自己。——尤达大师，《星球大战》

特别声明

自 2017 年 7 月，我开始写《应用计量经济学讲稿》以来，它受到很多学生和老师的喜爱，甚至有一些老师将其作为授课的参考资料。这让我大受鼓舞。

我爱人在澳大利亚国立大学留学期间推荐我看 Stock and Watson (2016 的《Econometrics》。我看后爱不释手，当初的感觉就是怎么有写得这么通俗易懂的计量经济学教材。然后，正值我入职安徽大学前夕，想着为以后的学生写点授课的讲稿，不让学生还要费钱买那些“深奥难懂”的课本——唯一的作用就是在毕业季当做“礼物”送给学弟学妹们，以树立好学长的“形象”。在看完 SW 的计量经济学后，我果断参考这本教材写了《应用计量经济学讲稿》。

入职安徽大学一年后，我将这份计量经济学讲稿分享出来。

非常欢迎大家给我们提出有益意见和建议。个人和机构可以利用本讲稿进行教学活动，但请不要用于商业目的。版权和最终解释权归许文立所有。当然，文责自负。

关注“宏观经济研学会”微信公众号的人可能有所了解，我的做事风格就是，深怕没人学我会的东西。因此，我也想再次呼吁大家：为了中国的经济学研究，不要藏着掖着啦，多分享，多奉献。

老师的真正重任应该是让学生超越自己。——尤达大师，《星球大战》

许文立
18, 9, 2020

目录

1	引言	1
1.1	计量经济学是什么?	1
1.2	计量经济学的方法	2
1.3	计量经济学概念	2
1.4	数据、数据结构与数据来源	3
1.4.1	观测数据	3
1.4.2	数据结构	3
1.4.3	数据来源	3
1.5	计量软件	4
1.6	进一步阅读资料	4
2	概率与统计基础	5
2.1	概率论	5
2.1.1	单变量分布	5
2.1.1.1	基本概念	5
2.1.1.2	主要统计量	5
2.1.2	多变量分布	7
2.1.3	常用分布	9
2.1.3.1	正态分布	9
2.1.3.2	卡方分布	10
2.1.3.3	t 分布	11
2.1.3.4	F 分布	11
2.1.4	随机抽样与大样本近似	13
2.1.4.1	随机抽样与样本矩	13
2.1.4.2	大样本近似	13
2.1.5	小结	13
2.2	统计学概述	14
2.2.1	估计	14
2.2.2	假设检验	15
2.2.3	置信区间	16
2.3	贝叶斯统计概述	16
2.4	附录	16
3	一元线性回归	17
3.1	线性回归模型估计	17
3.1.1	线性回归模型	17
3.1.2	系数估计	18
3.1.3	拟合度	19
3.1.4	最小二乘的假设	21
3.2	假设检验和置信区间	21
3.2.1	回归系数的假设检验	21
3.2.2	置信区间	22

3.2.3	虚拟变量	22
3.3	STATA 教程（一）	23
4	多元线性回归	31
4.1	遗漏变量偏误	31
4.1.1	遗漏变量偏误的定义	31
4.2	多元回归模型	32
4.2.1	多元回归中的 OLS 估计量	32
4.2.2	拟合度	33
4.2.3	多元回归中的 OLS 假设	34
4.3	假设检验与置信区间	34
4.3.1	单系数假设检验与置信区间	34
4.3.1.1	单系数假设检验	34
4.3.1.2	置信区间	35
4.3.2	联合假设检验	36
4.3.2.1	F 统计量	36
4.3.3	多元回归模型设定	37
4.4	多元回归 Stata 操作示例	38
5	识别的评价框架	43
5.1	内部有效性和外部有效性框架	43
5.1.1	内部有效性	43
5.1.2	外部有效性	43
5.2	内部有效性的威胁	44
5.2.1	遗漏变量偏误	44
5.2.2	函数形式误设	45
5.2.3	测量误差	45
5.2.4	缺失数据和样本选择	45
5.2.5	双向因果	46
5.2.6	OLS 标准误的不一致性	46
5.3	宏观中的识别	46
5.4	小班教学及其 Stata 操作	46
5.4.1	外部有效性	46
5.5	总结	47
6	面板数据模型	48
6.1	面板数据	48
6.2	固定效应回归	50
6.2.1	两期“比较”	50
6.2.2	固定效应回归	51
6.2.3	例子	53
7	工具变量法	55
7.1	一元回归与单工具变量	55
7.1.1	两阶段最小二乘 TSLS	55
7.2	IV 回归	56

7.2.1	TSLS	57
7.2.2	例子：香烟需求	57
7.3	如何检验 IV 的有效性	58
7.4	哪里去寻找有效的工具变量呢？	60
7.5	Stata 命令	60
8	实验和自然实验	61
8.1	理想实验	61
8.2	事件研究	63
8.3	双重差分 (DID)	63
8.3.1	DID	63
8.3.2	更多经典 DID 的例子	64
8.3.3	未处理组与平行趋势假设	64
8.3.4	DID 在 Stata 中的实现	67
8.3.5	长期效应	70
8.3.6	交叠 DID 设计	70
8.4	断点回归设计 (RDD)	70
8.4.1	断点回归估计量	70
8.4.2	断点回归的规定动作	73
8.4.3	例子与 stata 操作	74
8.5	其他因果效应识别方法——匹配与合成控制法	77
8.5.1	匹配法	77
8.5.1.1	PSM 的 Stata 应用演示	79
8.5.2	合成控制法	79
8.5.2.1	例子及 stata 操作	82
8.6	克服计量方法选择困难症	86
9	VARs	88
10	DSGE 及 Stata 操作	89
10.1	静态 IS-LM-PC 模型	89
10.2	三方方程 DSGE 模型	89
10.3	Stata 命令	89
11	极简 CGE：一个教学式模型	90
11.1	导论	90
11.2	极简 CGE：McLure and Thirsk(1975) 模型	90
11.3	税收政策的一般均衡效应：CGE 应用	92
11.3.1	企业所得税	92
11.3.2	商品税/增值税	92
11.3.3	福利效应	94
12	如何讲好经济学故事：经济学论文的流行形式	95
12.1	好的想法	95
12.2	标题的流行形式	95
12.3	文献综述	95
12.4	经验特征或待检验命题	95

12.4.1 逻辑推演式	95
12.4.2 数理模型式	95
12.5 实证分析	95
12.6 机制/反事实的数理模拟	95
12.7 结论与政策含义	95
12.8 引言	95
12.9 摘要	95
12.10参考文献	95
12.11附录	95
A 基本数学工具	96
A.1 求和算子与描述统计量	96

第1章 引言

从经验来看，计量经济学对于老师和学生来说都是一门非常有趣的课程。¹因为现实世界太复杂，我们不能凭直觉判断事物（变量）之间的关系。例如，

1、提高香烟消费税就能有效减少抽烟吗？

吸烟有害健康！这句话更可能的含义是，吸烟对他人的危害。因为二手烟可能给其他人带去更加严重的健康问题。因此，吸烟是一个全球面临的公共健康问题，随处可见禁止吸烟标识。经济理论告诉我们，治理外部性的一种方法是征税。目前，中国烟草企业缴税包括：烟草税、消费税、增值税、城市维护建设税、教育费附加、进口关税、企业所得税。2015年烟草消费税从5%提高到11%，以期控烟。

经济理论告诉我们，烟草消费税提高，烟草价格上升，从而导致烟草需求量下降。但是经济理论不能告诉我们，消费税率提高1个百分点，烟草需求量下降多少。

2、小班教学能提高教育产出吗？

发达国家提倡小班教学，认为这样能改善教学效果，提高学生的教育产出。这几年，中国也越来越重视小班教学，例如，“应用经济学人才卓越班”，全班20多人，大部分课程都单独授课。这样每个学生都可以得到老师更多的关注（当然，肯定有一些学生不希望老师太关注他），课堂讨论也能更充分，学习效果更好，学生成绩也能提高。

但是，真的是这样的吗？小班教学就意味着要雇佣更多的老师，建筑更多的教室，购买更多的教学设备等，那么，校长（或者李院长）就会考虑这种“小班教学”是否“划算”。李院长可能想知道小班教学所带来的益处是什么？有多大？以便能与上述成本进行比较。

常识和日常经验告诉我们，小班教学确实有很大好处。但是常识不能告诉我们这个好处有多大。为了提供一个定量答案，我们必须要进行经验测量，基于数据——班级规模（学生数）与学生成绩——来分析小班教学对学生成绩的影响有多大。

3、头发长得快能促进经济增长吗？

曾经，我的老师们（也就是各位的师公们）经常告诉我们，一定要警惕“伪回归”（也就是两个变量之间本来没有因果关系，生拉硬拽的把它们拿来做回归分析）。经常提到的例子是，头发与GDP就是伪回归，头发每天在长，GDP也每天在长，你能说头发促进了经济增长吗？那个时候，小伙子血气方刚，“天下唯我独尊”，我就要这么做回归，爱咋地咋地。咻当，计量经济学不及格！

到现在，我还是这么固执，坚持认为头发跟GDP是有关系的。所以，大家放心，期末不会不及格，除非你们交白卷。为什么头发生长会促进GDP？因为头发长出来了，你要去剪头发，理发所支付的费用会核算进GDP中，因此，你头发长得快，剪头发频率高，GDP就会增长越快。（当然，要是像某些女孩子一样，头上顶个碗，对着镜子自己动手剪刘海，那就不算GDP了）

4、央行降低利率对GDP、消费、投资等的影响有多大？

央行的货币政策（利率）对投资、消费会产生影响，进而影响到经济增长。但是这个效应分别为多大呢？这就需要计量经济学去评估。

计量经济学就是定量的来回答这些看似复杂的问题。计量经济学为我们理解复杂的世界打开了一扇窗。

1.1 计量经济学是什么？

计量经济学（Econometrics）一词据说是由挪威经济学家 R. Frisch（1895-1973）²创造出来的。Frisch 在 *Econometrica* 第一卷的卷首语中写道 (Econometrica, 1933, 1, pp. 1-2):

- 经济理论与统计学和数学之间联系的进展；

¹例如，许多计量微信群、计量微信公众号等等，读者基本是老师和一些对此感兴趣的高年级本科生与研究生。

²Frisch 是计量经济学会三个主要创始人之一，也是 *Econometrica* 杂志的首任主编，同时也是 1969 年第一届诺贝尔经济学奖的共同获得者。

- 经济问题的理论定量研究和经验定量研究；
- 计量经济学与经济统计学、数学在经济学中的应用不是一回事；
- 经验显示统计学、经济理论和数学都很重要，只有它们相互结合才能对现实世界的经济关系有更好的理解；
- 正是这三者的结合构成了计量经济学。

Frisch 的这些定义在今天仍然适用，只是在某些用法方面可能发生了一些变化。**计量经济学**就是综合利用经济模型、数理统计和经济数据来分析经济问题。Stock and Watson (2015, *Introduce to Econometrics Updated 3rd*) 说：“计量经济学是利用经济理论和数理统计技术来分析经济数据。”它可以分为两类：

- (1) **计量经济理论**，或者**理论计量经济学**包括工具和方法的发展，以及对方法性质的研究；
- (2) **应用计量经济学**描述了定量经济学的发展，以及利用经济数据来应用这些模型。

1.2 计量经济学的方法

现代计量经济学的统一方法是由挪威经济学家 T. Haavelmo(1911-1999)³开创的。1944 年他在 *Econometrica* 上发表“The probability approach in econometrics”。他认为定量经济模型就是一个概率模型，因此，要在经济模型中加入随机性。那么，对经济模型的量化、估计和推断的恰当方法必须要以数理统计学为基础。这就是计量经济学的**概率方法**。

Haavelmo 的概率方法很快就被经济学专业接受，发展，并广为传播。因此，当今的经济学定量研究离不开概率方法。

但是，最接近 Haavelmo 原始想法的并不是概率方法，而是**结构方法**。通常，计量经济模型和定量分析都是在模型正确设定的假设下进行的。**结构方法**则引出了似然分析，例如极大似然估计 (MLE) 和贝叶斯估计 (BE)。但是结构方法最大的缺点是认为经济模型设定正确。

但是，更准确的是，我们应该把模型当做现实世界的一种抽象和近似。因此，推断的**准结构方法**就把模型当做一种近似，而非真实的。这种理论引出了“伪真实值”(pseudo-true value)、拟似然函数、拟 MLE 和拟似然推断。与此紧密联系的是**半参数方法**。概率经济模型是一种局部设定模型，有一些经济特征并没有被设定。这种方法发展了最小二乘 (LS)、广义矩方法 (GMM)。这也是本课程主要关注的方法。

定量结构模型的另一个分支就是**校准方法**。与准结构方法相似，校准方法把模型理解为一种近似。它们之间的区别在于，校准方法拒绝统计推断，而是用模型与数据矩匹配的方法来选择参数。这是宏观计量中的主要方法。

1.3 计量经济学概念

最常用的计量经济学概念就是**数据**、**数据集**和**样本**。它们是一系列可描述的信息，例如劳动收入、学习成绩、年龄、投资额和 GDP 等。

经济学家总是面对着有关变量的一系列重复测量值。而对于变量的不同重复测量，我们称为**观测值**。

经济学家通常用 x 、 y 和/或 z 来表示观测值。计量经济学中，通常用 y 来表示**被解释变量/因变量**，而 x 和 z 表示**解释变量/自变量**。实数用小写字母表示，例如 y ；向量用粗体小写字母表示，例如 \mathbf{x} 。例如，

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ . \\ . \\ . \\ x_k \end{pmatrix} \quad (1.1)$$

³T. Haavelmo 是 1989 年诺贝尔经济学奖得主。

加粗大写字母 \mathbf{X} 表示矩阵。

带下标 i （有时候也用 j 或其它字母表示）的变量表示观测值，例如 y_i , x_i 和 z_i 。此外，带时间下标 t 的变量表示时间序列观测值。面板数据观测值带有 it 下标。

第 i 个观测值是 (y_i, x_i, z_i) 。样本是 $(y_i, x_i, z_i): i=1, \dots, n$ 的集合。

小写希腊字母 β , θ 等表示计量模型的未知参数。加粗希腊字母 β , θ 表示系数向量。

回归模型——微观计量中最主要的模型——量化一个变量发生变化导致另一个变量的变化程度。这里一个变量变化导致另一个变量变化就是**因果效应**，例如，在红薯地里施肥会产出更多的红薯。测量因果效应的一种方法就是进行试验：在气候条件、土壤条件、土地面积等等相同的情况下，给一块地施肥 1kg ，而其余的地则不施肥。而哪块地施肥则是由抽签随机决定的。等到秋收季节，我们把红薯挖出来，施肥的红薯与没施肥的红薯之间的重量差就是施肥所带来的增产效应。这就是**随机控制实验**，没有施肥的地块是**控制组**，施肥的地块是**处理组**。

1.4 数据、数据结构与数据来源

1.4.1 观测数据

计量经济学通常就是量化一个变量对另一个变量的影响。例如，企业所得税对企业投资的影响。

从自然科学的角度来看，最理想的情形就是利用实验数据来回答这些问题。但是，经济学，或者社会科学中，做实验，要么成本很大，我们不能让一个企业缴纳 30% 税率，另一个企业只缴纳 10% 税率；要么，不道德，教育的影响，让一部分孩子不上学，这太不道德了。

因此，大多数的经济数据是可观测数据。（注意：目前的实验经济学是可以得到某些实验数据的。）例如，我们通常能收集到教育与工资的记录数据，据此，我们可以测算两个变量的联合分布。但我们并不能从观测数据中推断它们之间的因果关系。因为我们不能操纵个人教育层次和年限，来观测他的不同工资结果。

1.4.2 数据结构

五种主要的数据结构：

- (1) 截面数据
- (2) 时间序列数据
- (3) 面板数据

(4) 聚类数据：与面板数据相关。在聚类抽样中，观测值被归类——类别间相互独立，类别中相关。与面板数据的主要差别在于，聚类抽样并不显性建模误差结构。

- (5) 空间数据：根据空间指标而具有相互依赖性。

1.4.3 数据来源

目前，有许多公开的数据来源：

- (1) 国家统计局
- (2) 各种类型的统计年鉴
- (3) CGSS
- (4) 其它微观调研数据

1.5 计量软件

目前，有许多计量软件：

- (1) Stata
- (2) Eviews
- (3) R
- (4) Matlab
- (5) Python
- (6) 其它软件

1.6 进一步阅读资料

第 2 章 概率与统计基础

2.1 概率论

在本讲中，我会向大家介绍回归分析、结构分析和计量经济学中用到的核心概率与统计理论。我们生活在一个无处不随机的世界中。而概率论为量化和描述随机性提供了有用的工具。

2.1.1 单变量分布

2.1.1.1 基本概念

结果 (outcomes) 是一个随机过程中许多相互排斥的潜在结果 (results)。例如，明天某一时刻的天气可能是晴天，可能是多云，可能是阴天，也可能是狂风暴雨。这些不同的天气情况就是结果 (outcomes)，但是只有其中一个结果 (outcomes) 会发生。而且，通常每种结果都不是等可能性发生的。而**概率**就是一种结果 (outcome) 在长期内出现次数的比例。例如，在你们写作课程论文期间，电脑宕机的概率为 20%，也就是说，你们在未来写 100 篇论文的时候，会有 20 篇论文写作过程中，电脑“挂”了（这个故事除了告诉我们概率的含义外，还提醒我们要注意时刻记得保存、备份重要文档）。

所有可能结果 (outcomes) 的集合成为**样本空间**。样本空间的子集成为**事件**。例如，“写论文过程中电脑宕机不会超过一次”成为一个事件，即电脑宕机次数 0,1 是电脑宕机这个样本空间的一个子集。

随机变量分为**离散随机变量**，例如，0,1,2,3，和**连续随机变量**。计量经济学中使用的变量大部分为离散随机变量。

离散随机变量的**概率分布**是所有可能的变量值及其发生的概率列表（所有概率之和等于 1）。**累积概率分布**，**cumulative probability distribution** 就是随机变量小于等于某一特定值的概率，也称为累积分布函数，简写为 **c.d.f.** 或者**累积分布**。例如，电脑宕机的次数 M 是一个随机变量，每次宕机的概率如表 1 所示。

表 2.1: 随机变量概率

	结果 (宕机次数)				
	0	1	2	3	4
概率分布	0.8	0.1	0.06	0.03	0.01
累积概率分布	0.8	0.9	0.96	0.99	1

一个非常重要的离散分布函数是**伯努利分布 (Bernoulli distribution)**

而连续随机变量的累积概率分布与离散累积概率分布类似。连续随机变量的概率用**概率密度函数**，**probability density function** 来概述。任何两点之间的概率密度函数所形成的区域就是该随机变量落在这两点之间的概率。概率密度函数可以简写为 **p.d.f.**，或者**密度函数**，或者**密度**。

2.1.1.2 主要统计量

期望

随机变量 Y 的期望用 $E(Y)$ 表示， μ_Y ，指长期重复试验或发生的随机变量的均值。离散随机变量的期望是所有可能结果的加权平均，权数为每个结果发生的概率。

例如，上面的电脑宕机次数的期望为：

$$E(M) = 0.8 \times 0 + 0.1 \times 1 + 0.06 \times 2 + 0.03 \times 3 + 0.01 \times 4 = 0.35 \quad (2.1)$$

也就是说，电脑宕机次数的期望为 0.35 次。需要注意的是，实际电脑宕机次数肯定是一个整数，我们说“写论文期间电脑宕机了 0.35 次”没有任何意义。而公式 (1) 的计算结果表明，写论文过程中，电脑宕机的平均次

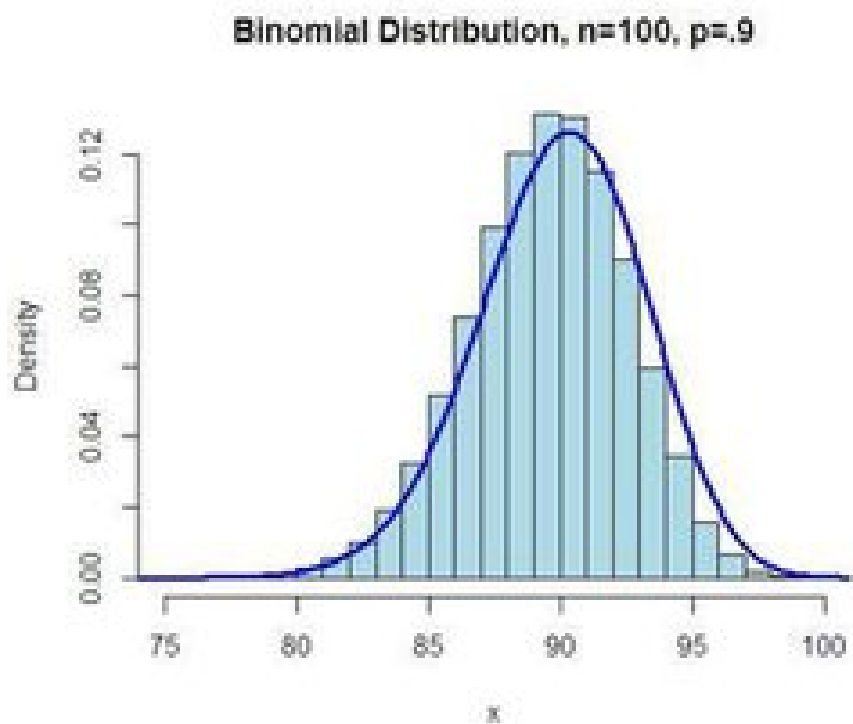


图 2.1: 伯努利分布: 来源于百度图片

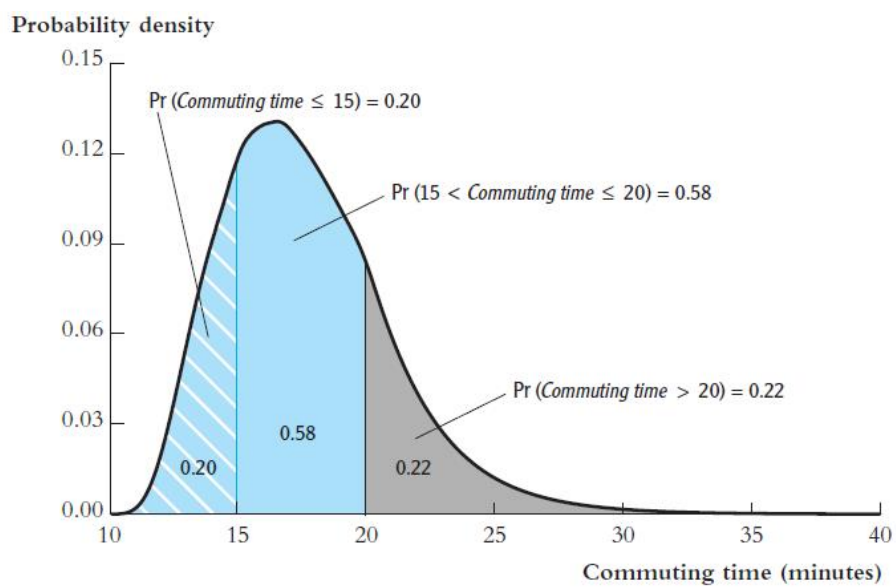


图 2.2: 概率密度函数: 来源于 Stock and Watson, 2015, pp18

数。那么，随机变量的期望计算公式为

$$E(Y) = p_1 y_1 + p_2 y_2 + \cdots + p_k y_k = \sum_{i=1}^k p_i y_i \quad (2.2)$$

标准差和方差

一个随机变量 Y 的方差用 $\text{var}(Y)$ 表示, 其计算公式为 $\text{var}(Y) = E[(Y - \mu_Y)^2]$ 。

而标准差是方差的开方, 用 σ_Y 表示。

$$\sigma_Y^2 = \text{var}(Y) = E[(Y - \mu_Y)^2] = \sum_{i=1}^k (y_i - \mu_Y)^2 p_i \quad (2.3)$$

根据公式 (3), 我们计算电脑宕机次数的方差和标准差为

$$\text{var}(Y) = 0.8 \times (0 - 0.35)^2 + 0.1 \times (1 - 0.35)^2 + 0.06 \times (2 - 0.35)^2 + 0.03 \times (3 - 0.35)^2 + 0.01 \times (4 - 0.35)^2 = 0.647 \quad (2.4)$$

$$\sigma_Y = \sqrt{\text{var}(Y)} = \sqrt{0.647} \approx 0.80 \quad (2.5)$$

均值、方差的性质

(1) $Z = a + bY$, a, b 都是常数, 那么 $E(Z) = E(a + bY) = a + bE(Y)$;

(2) $\text{var}(Z) = \text{var}(a + bY) = b^2 \text{var}(Y)$

其它分布特征

分布的特征除了均值和方差 (或标准差) 外, 还有另外两个重要的形状指标: **峰度**——测量尾部有多“厚”, 和**偏度**——测量分布非对称性程度。均值、方差、峰度和偏度都是分布的矩。

一个随机变量 Y 的分布的峰度计算公式为

$$S(Y) = \frac{E[(Y - \mu_Y)^4]}{\sigma_Y^4} \quad (2.6)$$

偏度的计算公式为

$$S(Y) = \frac{E[(Y - \mu_Y)^3]}{\sigma_Y^3} \quad (2.7)$$

2.1.2 多变量分布

大多数经济学问题都是以两个或多个随机变量的形式出现, 例如, 教育与工作收入、性别与工作收入等等。因此, 我们必须了解多个随机变量的联合概率分布、边际概率分布和条件概率分布。

联合概率分布

两个离散随机变量 (X, Y) 的联合概率分布就是两个随机变量同时取得某个值 (例如, x, y) 时的概率, 其可以写成 $\text{Pr}(X = x, Y = y)$ 。

边际概率分布

变量 Y 的边际概率分布仅仅只是 Y 概率分布的另一个名字, 它是为了区分单一变量 Y 的分布和 Y 与其他变量的联合概率分布。从联合概率分布中计算 Y 的边际概率分布, 就是把 Y 取某个特定值的所有概率相加。假设 X 取 1 个值, $Y=y$ 的边际概率分布为

$$\text{Pr}(Y = y) = \sum_{i=1}^l \text{Pr}(X = x_i, Y = y) \quad (2.8)$$

条件概率分布

给定 X 的值, 随机变量 Y 的概率分布就叫做 Y 的条件概率分布, 表示为 $\text{Pr}(Y = y|X = x)$ 。条件概率分布的计算公式为:

$$\text{Pr}(Y = y|X = x) = \frac{\text{Pr}(X = x, Y = y)}{\text{Pr}(X = x)} \quad (2.9)$$

条件期望

给定 X, Y 的条件期望, 也称为给定 X, Y 的条件均值, 是给定 X, Y 的条件分布的均值。已知 $X=x$ 条件下, Y 的条件期望为

$$E(Y|X=x) = \sum_{i=1}^k y_i Pr(Y=y_i|X=x) \quad (2.10)$$

期望迭代法则

Y 的均值是给定 X 的条件下 Y 的条件期望的加权平均, 而权重是 X 的概率分布。数学表达式为

$$E(Y) = \sum_{i=1}^k E(Y|X=x_i) Pr(X=x_i) \quad (2.11)$$

换句话说, Y 的期望就是给定 X 条件下, Y 的期望的期望

$$E(Y) = E[E(Y|X)] \quad (2.12)$$

公式 (12) 右边的内部期望是给定 X 条件下 Y 的条件期望, 而外部期望是利用 X 的边际分布计算得到。而 (12) 就是期望迭代法则。

需要注意的是, 如果给定 X 条件下 Y 的条件期望为 0, 那么, Y 的均值也为 0。证明: $E(Y|X)=0, E(Y)=E[0]=0$, 证毕。

条件方差基于 X 的 Y 的条件方差是给定 X 的条件下 Y 的概率分布的方差。公式为

$$var(Y|X=x) = \sum_{i=1}^k [y_i - E(Y|X=x)]^2 Pr(Y=y_i|X=x) \quad (2.13)$$

相互独立两个随机变量 X 和 Y , 如果在不提供一个随机变量的信息情况下, 能得出另一个随机变量的值, 那么, 称 X, Y 独立分布, 或者相互独立。尤其是, 如果给定 X 的条件下 Y 的条件分布等于 Y 的边际分布, 那么 X, Y 相互独立, 即对于所有的 x, y , 如果

$$Pr(Y=y|X=x) = PrY=y \quad (2.14)$$

那么, X 和 Y 相互独立。

把等式 (14) 代入公式 (9) 中, 得到 X 和 Y 独立的另一个等价条件:

$$Pr(X=x, Y=y) = Pr(X=x)Pr(Y=y) \quad (2.15)$$

也就是说, 两个独立随机变量的联合分布就是它们的边际分布之积。

协方差和相关 协方差是测度两个随机变量共变程度的一种指标。通俗地说就是, 你变, 我也变, 绝对值越大, 说明我们两个越“心有灵犀”。 X 和 Y 的协方差是 X 与其均值之差乘以 Y 与其均值之差的期望, 用 $cov(X, Y)$ 表示。数学公式为

$$cov(X, Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = \sum_{i=1}^k \sum_{j=1}^l (x_j - \mu_X)(y_i - \mu_Y) Pr(X=x_j, Y=y_i) \quad (2.16)$$

如果两个随机变量同方向变动, 那么, 协方差为正; 如果反方向变化, 则协方差为负; 如果相互独立, 则协方差为 0。

由于协方差的单位为 X 的单位乘以 Y 的单位, 因此, 协方差的数值难以理解。为了解决“单位”问题, 另一种表示 X 和 Y 之间相互依赖程度的测量指标就是**相关系数**, 即协方差除以标准差之积:

$$corr(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (2.17)$$

当 $corr(X, Y) = 0$, 就说 X 和 Y 不相关。相关系数总是处于 -1 和 1 之间。

如果 Y 的条件均值不依赖于 X , 那么, X 和 Y 不相关。

需要注意的是, 独立, 一定不相关; 但不相关, 不一定独立。

分布特征的性质:

$$(1) E(X+Y) = E(X) + E(Y) = \mu_X + \mu_Y$$

$$(2) var(X+Y) = var(X) + var(Y) + 2cov(X, Y) = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}$$

$$(3) E(X^2) = \sigma_X^2 + \mu_Y^2$$

$$(4) E(XY) = \sigma_{XY} + \mu_X \mu_Y$$

2.1.3 常用分布

计量经济学中最常用的概率分布是正态分布、卡方分布、t 分布和 F 分布。

2.1.3.1 正态分布

正态分布的连续随机变量有钟型概率密度形状，如图 3 所示。

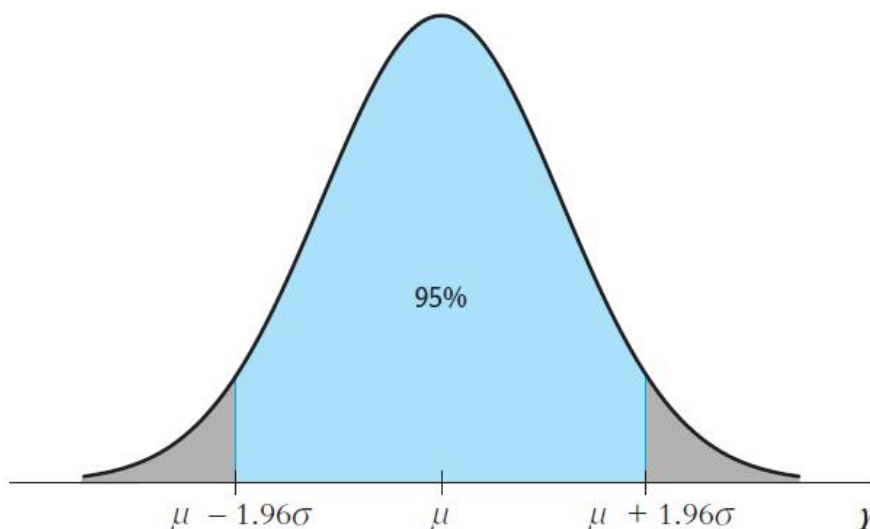


图 2.3: 正态概率密度函数：来源于 Stock and Watson, 2015, pp36

数学定义：一个连续随机变量 x_i 的概率密度函数为

$$f(x_i) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2} \quad (2.18)$$

遵循正态分布，且均值为 μ ，方差为 σ^2 。

由上述数学定义可以看出，正态分布有两个参数，均值 μ ，方差 σ ，因此，正态分布又可以表示为 $N(\mu, \sigma^2)$ 。而其中， μ 又可以叫做尺度参数 (scale parameter)， σ 称为形状参数 (shape parameter)。(注意：尺度参数和形状参数在后面的 DSGE 模型的贝叶斯估计中经常用到。大家知道有这些名称即可。)由此，可以定义**标准正态分布**，即均值为 0，方差为 1 的正态分布 $N(0, 1)$ ，通常用 Z 表示。标准正态累积分布方程用大写希腊字母表示 Φ ， $Pr(Z \leq c) = \Phi(c)$ 。标准正态分布函数的图形如图 4 所示。

从图 3 和图 4 中可以看到，正态分布的图形是在均值 μ 处对称的。从图 3 中还可以看出，随机变量值落在均值附近 $\pm 2\sigma$ 区间内的概率为 0.95。

我在前面的 1.1.2 节给出了均值和方差的性质。这些内容也可以理解为随机变量的线性转换。即 x_i 是正态随机变量，那么它的线性变换 y_i 也是正态分布。且两个正态随机变量的线性组合仍然为正态分布。

如果 x_i 是独立、同分布 (iid) 的正态随机变量，那么

$$\bar{x}_i \sim N\left(\mu_x, \frac{\sigma^2}{n}\right) \quad (2.19)$$

任何一个正态随机变量都可以通过线性变换转换成标准正态随机变量。这一过程称为**变量标准化**。这也为不同均值方差的正态分布的概率计算带来了方便。**变量标准化**就是随机变量减去均值，然后除以标准差。

例 1: 假设 $Y \sim N(1, 4)$ ，求 $Pr(Y \leq 2)$

$$\frac{(Y-1)}{\sqrt{4}} = \frac{1}{2}(Y-1)$$

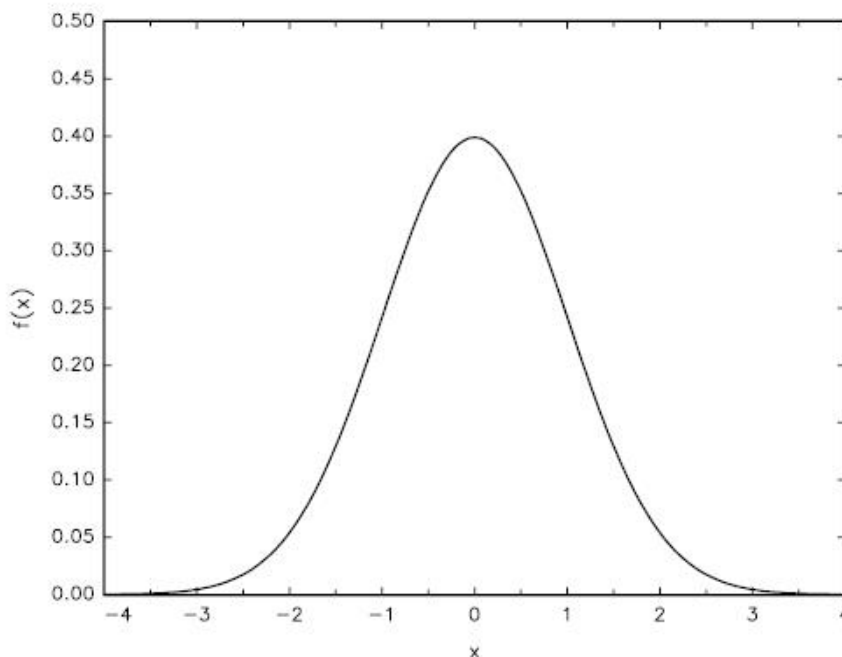


图 2.4: 标准正态分布函数: 来源于 Brown, 2010

$Y \leq 2$ 等价于 $\frac{1}{2}(Y - 1) \leq \frac{1}{2}(2 - 1)$

$Pr(Y \leq 2) = Pr[\frac{1}{2}(Y - 1) \leq \frac{1}{2}] = Pr(Z \leq \frac{1}{2}) = \Phi(0.5) = 0.691$

$\Phi(0.5) = 0.691$ 可以从临界值表中查询。

下面，我们来看看，正态分布变换成标准正态分布的正式数学过程：

(1) 首先，标准化

$$Z = \frac{\bar{x} - \mu_x}{\sqrt{\frac{\sigma_x^2}{n}}} = \frac{\bar{x}}{\sqrt{\frac{\sigma_x^2}{n}}} - \frac{\mu_x}{\sqrt{\frac{\sigma_x^2}{n}}}$$

(2) Z 的均值

$$EZ = \frac{E\bar{x}}{\sqrt{\frac{\sigma_x^2}{n}}} - \frac{\mu_x}{\sqrt{\frac{\sigma_x^2}{n}}} = \frac{\mu_x}{\sqrt{\frac{\sigma_x^2}{n}}} - \frac{\mu_x}{\sqrt{\frac{\sigma_x^2}{n}}} = 0$$

(3) Z 的方差

$$Var(Z) = E\left(\frac{E\bar{x}}{\sqrt{\frac{\sigma_x^2}{n}}} - \frac{\mu_x}{\sqrt{\frac{\sigma_x^2}{n}}}\right)^2 = E\left[\frac{n}{\sigma_x^2}(\bar{x} - \mu_x)^2\right] = \frac{n}{\sigma_x^2} \frac{\sigma_x^2}{n} = 1$$

正态分布在统计学中非常的重要。不仅是因为许多随机变量都遵循正态分布，而且更重要的是，任何样本随着其样本规模的增大，样本均值趋向于服从正态分布，这就是**中心极限定理**。

2.1.3.2 卡方分布

卡方分布是 m 个标准正态随机变量的平方和的分布，常用于检验某些类型的假设。其中， m 称为自由度。例如， Z_1, Z_2, Z_3 是标准正态随机变量，那么， $Z_1^2 + Z_2^2 + Z_3^2$ 就是一个自由度为 3 的卡方分布。一个自由度为 m 的卡方分布表示为： χ_m^2 。下面给出卡方分布的正式定义：

定义：假设 $Z_1, Z_2, Z_3, \dots, Z_n$ 是一组简单的随机样本，且服从 $Z_i \sim N(0, 1)$ ，那么，

$$\sum_{i=1}^n Z_i^2 \sim \chi_n^2 \quad (2.20)$$

其中， n 为卡方分布的自由度。

χ_n^2 的概率密度函数为

$$f_{\chi^2}(x) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, x \geq 0 \quad (2.21)$$

其中, $\Gamma(x)$ 是伽马函数。如果任意一个服从正态分布的随机变量 $x_i \sim N(\mu_x, \sigma_x^2)$, 都有

$$\sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2 \sim \chi_n^2 \quad (2.22)$$

卡方分布如图 5 所示。

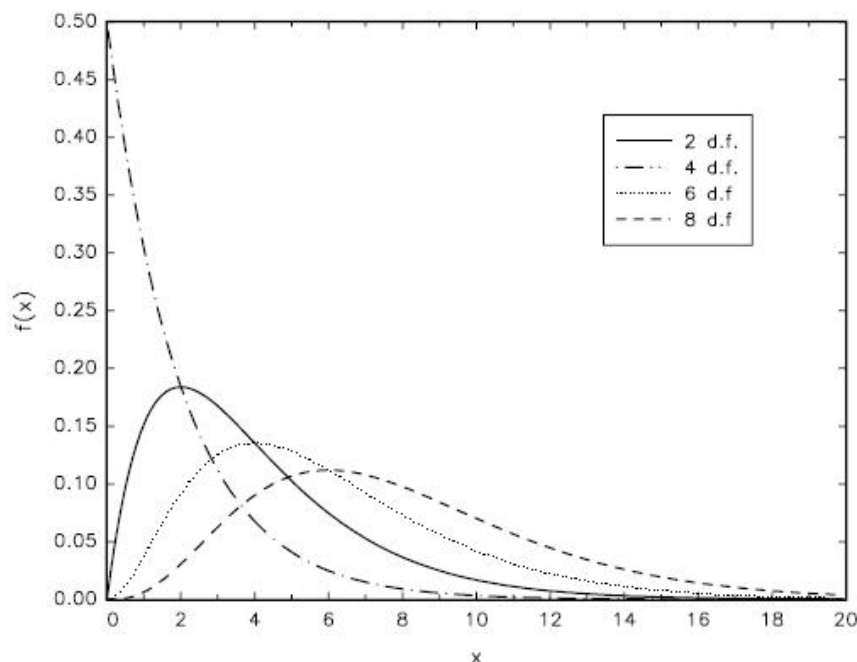


图 2.5: 卡方分布函数: 来源于 Brown, 2010

2.1.3.3 t 分布

t 分布, 也称为学生 **t 分布** 是标准正态分布与自由度 m 的卡方分布除以 m 再开方的比率。用 t_m 表示。

定义: 假设 $Z_i \sim N(0, 1)$, $Y \sim \chi_m^2$, 且 Z 和 Y 相互独立, 那么,

$$\frac{Z}{\sqrt{\frac{Y}{m}}} \sim t_m \quad (2.23)$$

其中, m 为 **t 分布** 的自由度。**t 分布** 的概率密度函数如图 6 所示。

t 分布 也是钟型图案, 类似于正态分布。但是当自由度较小 (20 或更小), 更多落在尾部, 也就是说 **t 分布** 比正态分布更扁平; 当自由度大于等于 30 时, **t 分布** 近似于正态分布, 而 t_∞ 等价于正态分布。

2.1.3.4 F 分布

自由度为 m, n 的 **F 分布**¹ 是一个自由度为 m 的卡方随机变量除以 m 比上自由度为 n 的卡方随机变量除以 n 的比值, 表示为 $F_{m,n}$ 。

定义: 假设 $Y \sim \chi_m^2$, $W \sim \chi_n^2$, 且 Y 和 W 相互独立, 那么,

$$\frac{Y/m}{W/n} \sim F_{m,n} \quad (2.24)$$

其中, m, n 是 **F 分布** 的自由度。

注意: (1) 如果 x 服从 **t 分布**, x^2 服从 **F 分布**。(2) 当分母的自由度趋向无穷时, $\frac{Y}{m} \sim F_{m,\infty}$ 。**F 分布** 的图形如图 7 所示。

¹F 分布是以伟大的统计学家 Sir Ronald A. Fisher 的名字命名的

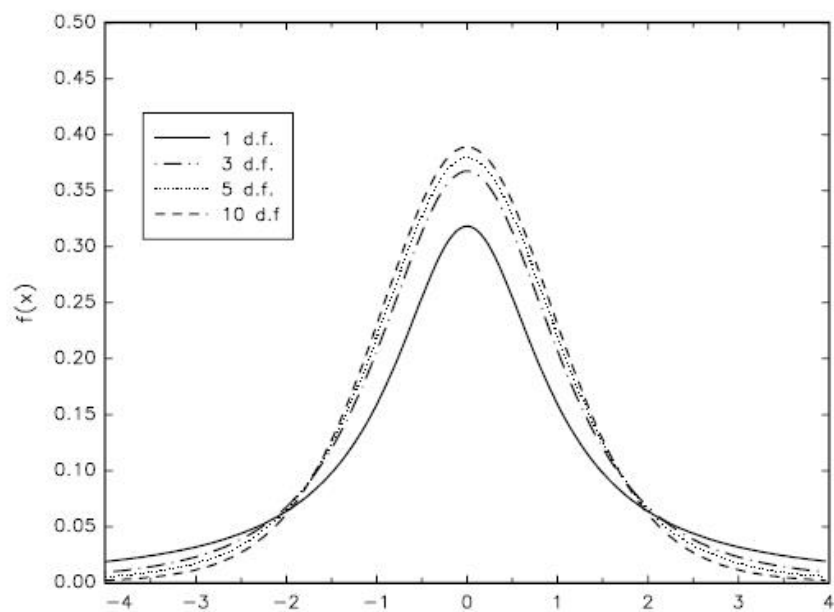


图 2.6: t 分布函数: 来源于 Brown, 2010

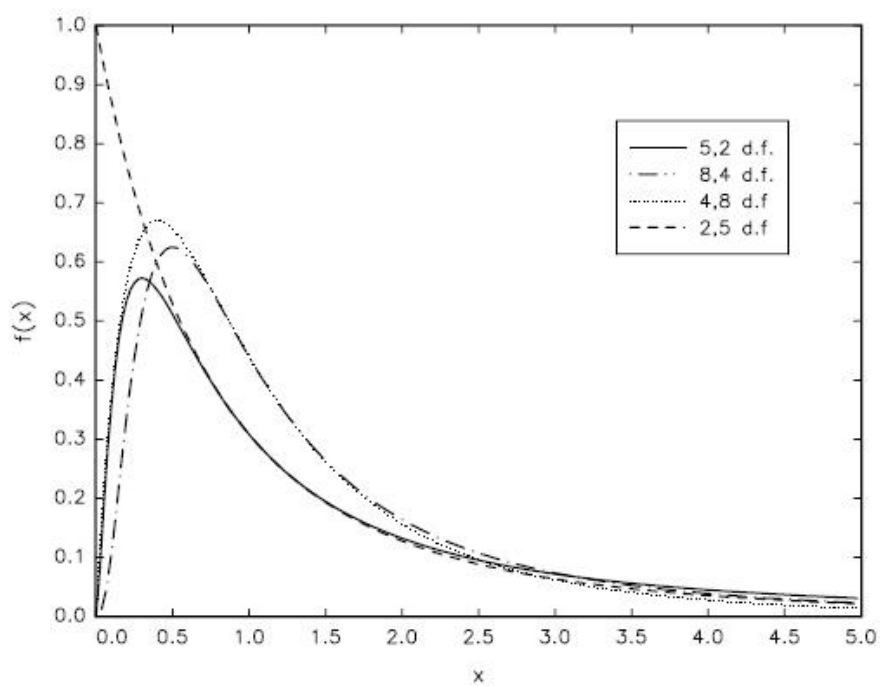


图 2.7: F 分布函数: 来源于 Brown, 2010

2.1.4 随机抽样与大样本近似

2.1.4.1 随机抽样与样本矩

随机抽样就是从更大的总体中随机抽取一个样本。这个过程为了使样本均值（见 1.3.1 节）本身成为一个随机变量。那么，就可以探讨样本均值的分布——**抽样分布**。

简单随机抽样是指从总体 N 个单位中任意抽取 n 个单位作为样本，使每个可能的样本被抽中的概率相等的一种抽样方式。

因为 Y_1, Y_2, \dots, Y_n 是从总体中随机抽取，因此，每一个样本 Y_i 的边际概率分布都相同，都与总体 Y 的分布相同。当 Y_i 有相同的边际概率分布时，我们称 Y_1, Y_2, \dots, Y_n 为**同分布**。

在简单随机抽样下，已知 Y_1 的值并不能为 Y_2 提供任何信息。因此，给定 Y_1 条件下， Y_2 的条件概率分布与 Y_2 的边际概率分布相同。也就是说，在简单随机抽样下， Y_1 的分布独立于 Y_2 的分布。当 Y_1, Y_2, \dots, Y_n 来自于相同的总体，又独立分布时，我们称为**独立同分布 (i.i.d)**。

考虑随机样本 Y_1, Y_2, \dots, Y_n ，假设 $EY_i = \mu$ ， $Var(Y_i) = \sigma^2$ 。定义 $S = Y_1 + Y_2 + \dots + Y_n$ 为样本和。那么，

$$ES = E(Y_1 + Y_2 + \dots + Y_n) = EY_1 + EY_2 + \dots + EY_n = n\mu \quad (2.25)$$

$$Var(S) = E(S - ES)^2 = E(Y_1 + Y_2 + \dots + Y_n - n\mu)^2 = E\left[\sum_{i=1}^n (Y_i - \mu)\right]^2 = n\sigma^2 \quad (2.26)$$

定义**样本均值**为 $\bar{Y} = \frac{\sum Y_i}{n}$ 。那么，

$$E\bar{Y} = E\frac{S}{n} = \frac{1}{n}ES = \frac{1}{n}n\mu = \mu \quad (2.27)$$

$$Var(\bar{Y}) = E(\bar{Y} - \mu)^2 = E\left(\frac{S}{n} - \mu\right)^2 = \frac{1}{n^2}E(S - n\mu)^2 = \frac{\sigma^2}{n} \quad (2.28)$$

2.1.4.2 大样本近似

目前，有两种方法刻画抽样分布：精确法和近似法。

精确分布又称有限抽样分布。

“近似法”利用近似式来表达抽样分布，这种方法依赖于大样本规模。抽样分布的大样本近似通常称为**渐近分布**——“渐近”是因为随着 n 趋向于无穷，近似就变成精确了。

注意：即使样本只有 30 个观测值，近似也非常精确。因为计量经济学中的观测值通常达到成百上千，因此，渐近分布能为精确抽样分布提供一个较好的近似。

当样本很大的时候，两个法则很关键：大数法则和中心极限定理。

大数法则是当样本规模很大时， \bar{Y} 以很高的概率接近于 μ_Y 。

中心极限定理是当样本规模很大时，标准化样本均值的抽样分布， $\frac{(\bar{Y} - \mu_Y)}{\sigma_{\bar{Y}}}$ ，近似服从正太分布。

因此，渐近正态分布并不依赖于 Y 的分布。渐近理论为回归分析提供了极大的简化。

2.1.5 小结

1、The probabilities with which a random variable takes on different values are summarized by the cumulative distribution function, the probability distribution function (for discrete random variables), and the probability density function (for continuous random variables).

2、The expected value of a random variable Y (also called its mean, m_Y), denoted $E(Y)$, is its probability-weighted average value. The variance of Y is $\sigma_Y^2 = E[(Y - \mu_Y)^2]$, and the standard deviation of Y is the square root of its variance.

3、The joint probabilities for two random variables X and Y are summarized by their joint probability distribution. The conditional probability distribution of Y given $X = x$ is the probability distribution of Y , conditional on X taking on the value x .

4、A normally distributed random variable has the bell-shaped probability density in Figure 4. To calculate a probability associated with a normal random variable, first standardize the variable and then use the standard normal cumulative distribution.

5、Simple random sampling produces n random observations Y_1, \dots, Y_n that are independently and identically distributed (i.i.d.).

6、The sample average, \bar{Y} , varies from one randomly chosen sample to the next and thus is a random variable with a sampling distribution. If Y_1, \dots, Y_n are i.i.d., then:

- the sampling distribution of \bar{Y} has mean μ_Y and variance $\sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n}$;
- the law of large numbers says that \bar{Y} converges in probability to μ_Y ; and
- the central limit theorem says that the standardized version of \bar{Y} , $\frac{(\bar{Y} - \mu_Y)}{\sigma_{\bar{Y}}}$, has a standard normal distribution $[N(0,1)$ distribution] when n is large.

2.2 统计学概述

统计学是应用数据来了解我们所生活世界的一门科学 (Stock and Watson, 2015)。统计工具能提供一些我们关注的总体分布特征。

我们对整个世界，或者整个中国经济、社会、人口感兴趣。但是，以目前的技术水平，我们不可能去调查 14 亿人口，因为调查总体的成本非常大。但我们又想知道总体分布特征，怎么办？统计学的主要任务就是来解决这个问题。回忆一下，前一节讲过的随机抽样。我们只需要从总体中随机抽取样本，然后，利用统计方法，结合随机样本信息来推断总体分布特征。这样也可以得到一个较为满意的近似结果。

计量经济学中使用的统计方法主要有三种：估计、假设检验与置信区间。**估计**就是从样本数据中，为一个总体分布特征——均值、方差等——计算出一个“最佳猜测”值。**假设检验**就是提出一个假设，然后用样本证据来验证假设是否为真。**置信区间**就是利用一组样本数据来估计未知总体分布特征的范围或区间。

2.2.1 估计

再次回忆一下随机抽样，从总体中随机抽取的样本 Y_1, \dots, Y_n 是独立同分布 (i.i.d.)，且与总体 Y 同分布，那么，样本均值 \bar{Y} 就能很自然地被认为是总体均值 μ_Y 。这种样本均值也称为总体均值的**估计量**²。

但是，计算样本均值 \bar{Y} 是得到总体均值估计量的唯一一种方式吗？答案是否定的。 Y_1, \dots, Y_n 都是与 Y 同分布，那么， Y_1 也可以作为总体均值的一个估计量。以此类推，事实上， μ_Y 的估计量很多。那么，我们如何判断一个估计量比另一个估计量“更好”呢？我们前面讲过，抽样随机变量和样本均值都有概率分布，那么，这个问题还可以表达成：一个估计量的合意分布特征是什么呢？

既然，我们是从样本信息中推断未知总体分布特征。那么，最合意的结果肯定是，样本估计量尽可能的接近总体分布“真值”。由此，可以给出，合意结果的三个特征：**无偏性**、**一致性**和**有效性**。注意，在后面的回归分析中，这三个特征非常非常重要。

无偏性 如果你通过重复抽样来评估一个估计量，一般来说，你会得到一个“真值”。因此，一个估计量的合意性质就是要使其抽样分布均值等于总体均值 μ_Y 。如果是这样，那么，我们就称这个估计量**无偏**。用 $\hat{\mu}_Y$ 来表示 μ_Y 的估计量。用 $E(\hat{\mu}_Y)$ 表示估计量抽样分布的均值。如果 $E(\hat{\mu}_Y) = \mu_Y$ ，那么，估计量 $\hat{\mu}_Y$ 是无偏的，反之亦然。

一致性 当样本量很大时，由样本的随机变动引起的 μ_Y 值的不确定性就非常小。也就是说， $\hat{\mu}_Y$ 落入真值 μ_Y 的一个较小区间内的概率随着样本量的增长而接近于 1。即是说， $\hat{\mu}_Y$ 是 μ_Y 的一致估计。

²估计量 (estimator) 是数据样本的一个函数；估计 (estimate) 则是估计量的数值。

有效性 如果你有两个无偏的估计量 $\hat{\mu}_Y$ 和 $\tilde{\mu}_Y$ ，那么，你会如何选择？此时，你应该选择最小方差的估计量。如果 $\hat{\mu}_Y$ 的方差比 $\tilde{\mu}_Y$ 更小，就说明 $\hat{\mu}_Y$ 比 $\tilde{\mu}_Y$ 更有效³。

下面，我们来看看样本均值 \bar{Y} 是否满足上述估计量的三个标准。

(1) 样本均值等于总体均值已经在 1.4.1 节证明 $\bar{Y} = \mu$ ，因此，样本均值是无偏的。

(2) 根据大数法则，见 1.4.2 节，样本规模越大， \bar{Y} 以很大概率接近 μ ，因此，样本均值是一致的。

(3) 那怎么判断 \bar{Y} 是有效的估计量呢？回忆一下，我在前面提到过， μ_Y 的估计量还有很多，例如 Y_1, Y_2, \dots, Y_n 。我们现在选择用 Y_1 与 \bar{Y} 进行比较。首先， Y_1 与 \bar{Y} 都是无偏估计。而 Y_1 的方差为 $\text{Var}(Y_1) = \sigma_Y^2$ 。根据 1.4.1 节， \bar{Y} 的方差为 $\frac{\sigma_Y^2}{n}$ 。只要 $n \geq 2$ ，那么， \bar{Y} 的方差就小于 Y_1 的方差，因此， \bar{Y} 是有效估计量。

综上所述，我们也把样本均值 \bar{Y} 称为最优线性无偏估计 (**Best Linear Unbiased Eestimator, BLUE**)。

此外，还有一点非常重要，那就是随机抽样的重要性。虽然我们不能实施一个完全随机的抽样，但是我们设计的抽样要尽可能降低偏误。

2.2.2 假设检验

待检验的假设成为**原假设**。假设检验就是用数据来比较原假设与另一个假设——**备择假设**。如果原假设不成立，那么，备择假设成立。在统计学中，原假设通常为总体均值等于某一特定值，用 H_0 表示，即

$$H_0 : E(Y) = \mu_{Y,0} \quad (2.29)$$

最常用的备择假设为 $H_1 : E(Y) \neq \mu_{Y,0}$ ，这种类型被称为**双向备择假设**，因为该假设允许 $E(Y)$ 要么大于特定值，要么小于特定值。

统计学理论将会告诉我们如何利用样本数据来判断是否接受 H_0 ，还是接受 H_1 。

现实中，我们不可能知道总体均值，只能用随机抽样的样本均值 \bar{Y} 代替。那么， \bar{Y} 不可能精确地等于 $\mu_{Y,0}$ 。 \bar{Y} 与 $\mu_{Y,0}$ 之间的差异，要么是因为真实均值并不等于 $\mu_{Y,0}$ （原假设为假），要么因为真实均值等于 $\mu_{Y,0}$ （原假设为真）但由于随机抽样使得 \bar{Y} 与 $\mu_{Y,0}$ 不等。这两种可能性，几乎区分不了，但我们可以计算一个概率来允许检验原假设。即利用数据来计算原假设的 p 值。

p 值，也称为显著性概率是利用样本数据计算的一个对原假设不利的概率值。也就是说，p 值越小，结果越显著。其数学定义为

$$p\text{-value} = \Pr[|\bar{Y} - \mu_{Y,0}| \geq |\bar{Y}^{act} - \mu_{Y,0}|] \quad (2.30)$$

其中， \bar{Y}^{act} 表示用实际数据计算的样本均值， \Pr_{H_0} 原假设下计算的概率。也就是说，p 值是 \bar{Y} 的分布尾部超出 $\mu_{Y,0} \pm |\bar{Y}^{act} - \mu_{Y,0}|$ 的区域。如果 p 值越大，观测到的 \bar{Y}^{act} 就与原假设一致，如果 p 较小，则拒绝原假设。

t 统计量

$$t = \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})} \quad (2.31)$$

当样本规模很大时，t 的分布近似于标准正态分布 $N(0, 1)$

在假设检验中通常犯两类错误：(1) **第一类错误**，原假设为真时却被拒绝；(2) **第二类错误**，原假设为假时却没有拒绝。

如果你选择拒绝原假设（为真）的预设概率水平（例如，5%），那么，只有 p 值小于 0.05 时才拒绝原假设。在实践中，5% 对应的标准正态分布的尾部区域是 ± 1.96 之外的区域，即简单规则为

$$|t^{act}| \geq 1.96, H_0 \quad (2.32)$$

也就是说，第一类错误的预设概率就是检验的**显著性水平**。

实践中，常用的显著性水平有：10%、5%、1%、0.1%。

³“有效性”这个术语源于，如果 $\hat{\mu}_Y$ 比 $\tilde{\mu}_Y$ 方差更小，那么， $\hat{\mu}_Y$ 能更有效的利用数据信息

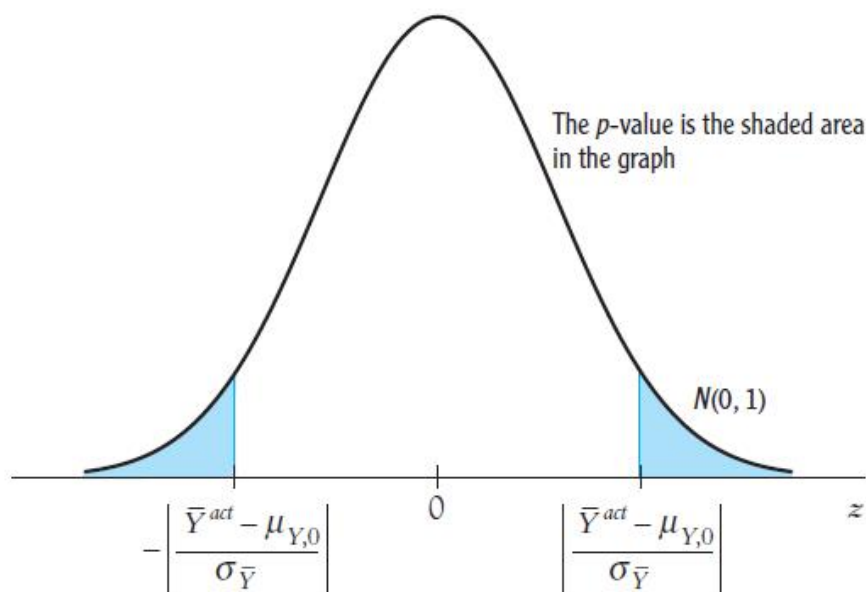


图 2.8: p 值: 来源于 Stock and Watson, 2015, pp74

2.2.3 置信区间

总体均值的 95% 置信区间就是真值有 95% 的概率落入该区间。当样本规模很大时, 90%、95%、99% 对应的置信区间为

$$90\% : \mu_Y = [\bar{Y} \pm 1.64E(\bar{Y})]$$

$$95\% : \mu_Y = [\bar{Y} \pm 1.96E(\bar{Y})]$$

$$99\% : \mu_Y = [\bar{Y} \pm 2.58E(\bar{Y})]$$

2.3 贝叶斯统计概述

贝叶斯 (T. Bayes, 1702-1763) 是英国数学家。他首先将归纳推理法用于概率论基础理论, 并创立了贝叶斯统计理论, 对于统计决策函数、统计推断、统计估算等作出了重要贡献。

贝叶斯于 1763 年在英国皇家学会学报上发表 "An essay towards solving a problem in the doctrine of chances"。该文中提出的二项分布参数推断方法后来被称为贝叶斯定理。贝叶斯公式

$$P(A|B) = \frac{P(A)P(B|A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \quad (2.33)$$

看上去贝叶斯公式只是把 A 的后验概率转换成了 B 的后验概率 + A 的边缘概率的组合表达形式, 因为很多现实问题中 $P(A|B)$ 很难直接观测, 但是 $P(B|A)$ 和 $P(A)$ 却很容易测得, 利用贝叶斯公式可以方便我们计算很多实际的概率问题。

具体可以参见:

- (1) 朱慧明, 林静. 2009, 《贝叶斯计量经济模型》, 科学出版社
- (2) Koop, G., Poirier, D. J., Tobias, J. L. (2007). Bayesian econometric methods. Cambridge University Press.
- (3) Geweke, J. (2005). Contemporary Bayesian econometrics and statistics (Vol. 537). John Wiley and Sons.
- (4) Koop, G., Korobilis, D. (2010). Bayesian multivariate time series methods for empirical macroeconomics. Foundations and Trends? in Econometrics, 3(4), 267-358.

2.4 附录

第3章 一元线性回归

2015年，政府提高香烟消费税对吸烟率的影响是什么？小班教学能提高学生测试得分吗？性别对工资的影响是什么？

其实，上述三个问题都是在问一个变量， X （包括消费税、班级规模和性别）的变化对另一个变量， Y （包括吸烟率、测试分数和工资）的影响。

线性回归模型就是把 X 和 Y 联系起来。这条回归线的斜率就是 X 变化一单位引起的 Y 的变化。因为 Y 的总体均值未知，所以这个斜率也未知。而计量经济学就是要用 X, Y 的样本数据来估计回归线的斜率。

3.1 线性回归模型估计

3.1.1 线性回归模型

回顾一下小班教学的例子。李院长还不太确定是否要缩减你们本科的班级规模。假设你们是计量经济学家或者咨询师，李院长来向你们寻求帮助。李院长说，他面临着一个选择困难：一方面，父母肯定是希望小班教学；另一方面，缩小班级规模，就要雇佣更多的老师，要支出更多的经费。因此，他问你们：如果缩小班级规模，学生的成绩会发生什么变化？

也就是说，如果李院长要改变班级规模，例如每个班级缩减10名学生，那么，学生的标准化成绩会发生什么变化？我们用希腊字母， $\beta_{ClassSize}$ ，来表示班级规模变化引起的成绩变化，数学表达式为

$$\beta_{ClassSize} = \frac{ScoreChange}{ClasssizeChange} = \frac{\Delta Score}{\Delta ClassSize} \quad (3.1)$$

其中， Δ 表示变化量；而 $\beta_{ClassSize}$ 就是由班级规模变化引起的学生成绩变化与班级规模变化的比值。如果你们运气好，知道了这个 $\beta_{ClassSize}$ ，例如，-0.5，那么，你们可以直接告诉李院长，班级规模变小，会让学生的成绩提高，且根据公式（1），提高的幅度为：

$$\Delta Score = \beta_{ClassSize} \times \Delta ClassSize \quad (3.2)$$

那么，班级规模减少10名学生，预期学生成绩会提高 $(-0.5) \times (-10) = 5$ 。也就是说，每个班级减少10名学生，预期学生成绩会提高5分。据此，公式（1）定义了班级规模与学生成绩之间直线的斜率。因此，可以把这条直线写成

$$Score = \beta_0 + \beta_{ClassSize} \times ClassSize \quad (3.3)$$

这个时候，你会不会兴奋地拿着公式（3）跑到李院长办公室，告诉他，我不仅能告诉您每个班级减少10人，学生成绩会提高多少。而且，只要您告诉我班级规模，我还能预期到学生的平均成绩会是多少。但是，李院长会说，不好意思，我对你这个方程和结果表示怀疑。因为每个班的学生本身有差异，每个班的授课老师不同，可能用的课本也不同。这些原因都可能导致学生的成绩不同，因此，公式（3）并不是对所有班级都成立。

接受了李院长的建议，回去重新修正模型，加入影响学生成绩的其他因素，得到下式

$$Score = \beta_0 + \beta_{ClassSize} \times ClassSize + OtherFactors \quad (3.4)$$

其中， $OtherFactors$ 里面包含了李院长提到的，和没提到的影响学生成绩的因素。公式（4）更一般化，因为我们关注于班级规模与学生成绩，所以才能把其它因素统统“装进” $OtherFactors$ 中。假设有 n 个班级， Y_i 表示第 i 个班级的平均成绩， X_i 表示第 i 个班级的学生人数。那么，公式（4）就可以表示为

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (3.5)$$

公式（5）称为一元线性回归模型， Y 称为因变量或被解释变量， X 称为自变量或解释变量。 $\beta_0 + \beta_1 X_i$ 称为总体回归线或总体回归方程。截距 β_0 和斜率 β_1 是总体回归线的系数，也称参数。斜率 β_1 可以理解为 X 变化一

单位， Y 的变化程度。¹

u_i 为误差项，其对应着第 i 个班级平均成绩与总体回归线预测的成绩只检测差异的所有因素。因此，误差项包含除了 X 之外所有决定因变量 Y 的因素。

3.1.2 系数估计

在实际情形中，我们不可能知道总体分布，即我们不可能知道总体回归线中的两个参数值。但是从第二讲可知，我们可以从随机抽样的样本数据中估计总体参数。同理，我们也可以用数据来估计总体回归线的斜率与截距。

如果大家有兴趣，可以去调查一下班级大小与成绩的信息，然后自己估计一下回归系数。正如第一讲中提到，这类调查往往成本巨大，可能有一些机构或者教育部门有这类调查数据，但是很遗憾没有公开。那么，我们就暂且使用一下美帝的数据样本来作为例子。数据为 1999 年加利福尼亚 420 个学区的测试分数和班级规模。表 1 中概述了这两个样本的分布。

表 3.1: 测试分数与师生比的分布

	样本量	均值	标准差	分位数		
				10%	50%	95%
学生-老师比	420	19.64	1.89	17.35	19.72	22.65
测试分数	420	654.16	19.05	630.38	654.45	685.5

由表 1 可以看到，平均每个老师带 19.64 个学生，标准差为 1.89。每个学区的分数均值为 654.16，标准差为 19.05。两个样本的散点图，如图 1 所示。分数与班级规模的相关系数为 -0.226。

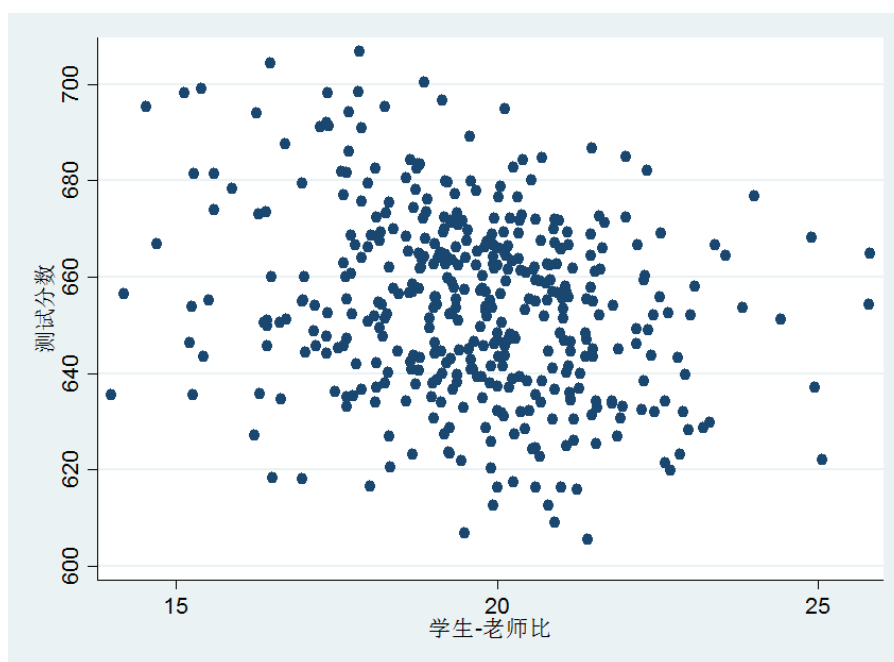


图 3.1: 学生-老师比与分数散点图

根据散点图和相关系数，我们大致可以判断基于这些数据的直线应该是向右下倾斜。只要我们画出这条线，我们就得到了斜率 β_1 的估计值。但是我们如何画出这条线呢？最常用的方法就是普通最小二乘（OLS）来拟合这些数据。

¹需要注意的是，从数学上理解，截距 β_0 是 $X=0$ 时 Y 的值，也就是总体回归线与 Y 轴的交点。但在经济学中，这个截距有时候有经济学含义，有时候则没有经济学含义，例如班级规模为 0 时，班级的平均成绩为 β_0 就不符合实际了，因此，这个时候要将其单纯理解成数学意义上的系数。

(1) **OLS 估计量** OLS 估计量使得估计的回归线尽可能的接近观测数据。而接近程度则由给定 X 条件下，预测 Y 的误差平方和来测度。

假设 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 用来表示 β_0 和 β_1 的估计量。那么，第 i 个观测值的误差为 $Y_i - \beta_0 - \beta_1 X_i$ 。那么，误差平方和为

$$\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2 \quad (3.6)$$

根据第二讲的统计学理论，存在唯一一对 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 来使得公式 (6) 最小化。由此得到的系数为 β_0 和 β_1 的 OLS 估计量。OLS 回归线称为样本回归线或样本回归函数。第 i 个观测值 Y_i 与其预测值之差为余项 (residual): $\hat{u}_i = Y_i - \hat{Y}_i$ 。

OLS 估计量的公式为

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (3.7)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (3.8)$$

OLS 预测值及残差

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad (3.9)$$

$$\hat{u}_i = Y_i - \hat{Y}_i \quad (3.10)$$

(2) 示例我们用 Stata14 来估计 OLS 回归线:

$$\hat{Y} = 698.9 - 2.28 \times X \quad (3.11)$$

Source	SS	df	MS	Number of obs	=	420
Model	7794.11004	1	7794.11004	F(1, 418)	=	22.58
Residual	144315.484	418	345.252353	Prob > F	=	0.0000
				R-squared	=	0.0512
				Adj R-squared	=	0.0490
Total	152109.594	419	363.030056	Root MSE	=	18.581

testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
str	-2.279808	.4798256	-4.75	0.000	-3.22298 -1.336637
_cons	698.933	9.467491	73.82	0.000	680.3231 717.5428

图 3.2: stata 结果

我们在 Y 上面加 hat 是为了区别它为基于 OLS 回归线的预测值。负斜率意味着班级规模越大，平均测试分数越低。

3.1.3 拟合度

我们已经估计出了班级规模对测试成绩效应的线性回归，如公式 (11)。正如李院长质疑的，我们都可能疑惑，估计的线性回归线对数据的拟合程度如何呢？

在计量经济学中， R^2 和回归标准误 (SER) 用来测量 OLS 回归线对数据的拟合程度。 $0 \leq R^2 \leq 1$ 测量的是 X_i 能解释 Y_i 的方差的比例。SER 测量的是 Y_i 离预测值有多远。

(1) R^2

根据预测值与残差的定义，可知

$$Y_i = \hat{Y}_i + \hat{u}_i \quad (3.12)$$

根据 R^2 的定义，它的数学形式可以表达为回归平方和或者解释平方和（**explained sum of squares, ESS**）与总平方和（**Total Sum of Squares, TSS**）之比。

$$ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \quad (3.13)$$

$$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (3.14)$$

那么， R^2 的公式为

$$R^2 = \frac{ESS}{TSS} \quad (3.15)$$

我们还可以这么思考： X 不能解释 Y 的方差的比例，同样可以表示出 R^2 。不能解释的部分就是残差平方和（**sum of squared residuals, SSR**），即 $SSR = \sum_{i=1}^n \hat{u}_i^2$ 。综上所述， $TSS = ESS + SSR$ 。据此，

$$R^2 = 1 - \frac{SSR}{TSS} \quad (3.16)$$

注：一元回归中的 R^2 就是 X 和 Y 的相关系数的平方。 R^2 越接近于 1，说明用 X 预测 Y 越好，即回归线拟合数据越好，反之亦然。

SER

回归标准误（SER）是回归误差标准差的估计量。它是观测值在回归线附近的分散程度的一种测量。OLS 残差为 \hat{u}_i 。那么，

$$SER = \sqrt{S_{\hat{u}}^2}, S_{\hat{u}}^2 = \frac{1}{(n-2)} \sum_{i=1}^n \hat{u}_i^2 = \frac{SSR}{(n-2)} \quad (3.17)$$

其中，OLS 残差的样本均值为 0。

例如，图 2 中的回归结果， $R^2 = 0.0512$, $SER(MSE) = 18.581$ 。这意味着，班级规模可以解释测试分数方差的 5.21%。而 $SER = 18.581$ 说明观测值在回归线附近分散较开，这也可以从图 3 中看出。

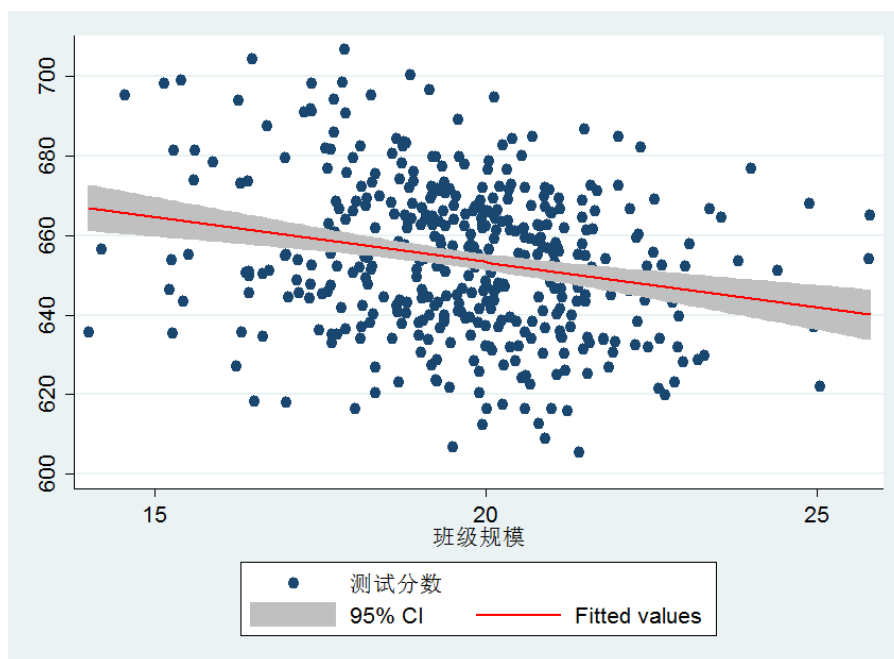


图 3.3: 回归线

注意：事实上， R^2 很小（或者 SER 很大）本身并不能说明回归的“好坏”。很小的 R^2 只是表面，除了解释变量 X 外，还有其它重要的因素影响 Y 。但是较小的 R^2 或者较大的 SER 并不能给出缺失的重要因素是什么，它们仅仅说明现有的 X 只能解释 Y 方差的较小部分。

3.1.4 最小二乘的假设

下面，我们简单的介绍一下 OLS 的三个假设。

假设一：给定 X 的条件下， u 的条件均值为 0

这个假设是说，“丢弃”到残差项 u 里的其它因素与 X 无关，即给定 X 条件下，这些因素的分布均值为 0。该假设等价于总体回归线就是给定 X 条件下的 Y 的条件均值。且该假设也意味着 $\text{corr}(X, u) = 0$ 。

假设二：(X_i, Y_i) 是独立同分布

假设三： X_i, Y_i 不可能有较大奇异值

较大的奇异值会使得 OLS 结果产生误差。这个假设就使得 X, Y 有非零的四阶矩： $0 \leq E(X_i^4) \leq \infty, 0 \leq E(Y_i^4) \leq \infty$ 。也就是说， X 和 Y 存在有限峰度。可能的来源：1、输入错误；2、单位错误。如果输入错误，就纠正它，如果不能纠正，就从样本中删除。

3.2 假设检验和置信区间

第一部分概述了一元回归系数的估计，这个部分将概述估计量有多精确地描述了抽样不确定性。

3.2.1 回归系数的假设检验

有一些人武断地说，班级规模并不会对测试分数产生影响。也就是说，总体回归线的斜率 $\beta_1 = 0$ 。下面，我们就来检验斜率是否为 0。也就是说，我们先假设 $\beta_1 = 0$ （原假设）。然后，我们来判断是否接受或者拒绝原假设。

首先，我们回顾一下 3.2 节中的总体假设检验。

原假设为 Y 的均值为某一特定值 $\mu_{Y,0}$ ，可以写成 $H_0 : E(Y) = \mu_{Y,0}, H_1 \neq \mu_{Y,0}$ 。

假设检验分三步走：

1、计算 \bar{Y} 的标准误 $SE(\bar{Y})$ ；

2、计算 t 统计量，即 $t = \frac{(\bar{Y} - \mu_{Y,0})}{SE(\bar{Y})}$ ；

3、计算 p 值，它是拒绝原假设的最低显著性水平。双边假设 p 值为 $2\Phi(-|t_{act}|)$ ，其中， t_{act} 是计算得到的 t 统计量， Φ 是累积标准正态分布。

在实践中，第三步的 p 值通常与临界值比较。例如，5% 显著性水平的双边假设对应着 $|t_{act}| > 1.96$ 。即是说，总体均值在 5% 的显著性水平下显著异于假设值。

系数的假设检验

上面已经提到过，有些人觉得小班没有效果。我们应该假设 $\beta_1 = 0$ ，那么，原假设和双边备择假设为

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 \neq 0 \quad (3.18)$$

那么，按照上述三步走：

第一步：计算 $\hat{\beta}_1$ 的标准误 $SE(\hat{\beta}_1)$ 。该标准误是 σ_{β_1} 的一个估计值。即

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}_{\beta_1}^2} \quad (3.19)$$

其中，

$$\hat{\sigma}_{\beta_1}^2 = \frac{1}{n} \times \frac{\frac{1}{n-2} \sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2}{[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2]^2} \quad (3.20)$$

第二步：计算 t 统计量

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \quad (3.21)$$

第三步：计算 p 值

$$p - value = Pr_{H_0} [|\hat{\beta}_1 - 0| > |\hat{\beta}_1^{act} - 0|] \\ = Pr_{H_0} \left[\left| \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \right| > \left| \frac{\hat{\beta}_1^{act} - 0}{SE(\hat{\beta}_1)} \right| \right] = Pr_{H_0} (|t| \geq |t^{act}|) \quad (3.22)$$

因为 t 统计量近似标准正态分布，因此

$$p - value = Pr(|Z| > |t^{act}|) = 2\Phi(-|t^{act}|) \quad (3.23)$$

如果 p 值小于 5%，即是说，在 5% 的显著性水平下拒绝原假设。5% 的显著性水平对应着 1.96 的临界值。

在实践中，我们并不用分别按照上述步骤计算出估计量和统计量，因为现在我们有计量经济学软件包，例如 Stata。我们把数据导入 stata 中，输入回归命令就可以直接得到上述三个步骤的结果，如图 2 所示。

例如，从图 2 中可以看出， β_1 的标准误为 0.48，系数为 -2.28，那么 $t = \frac{-2.28-0}{0.48} = -4.75$ 。t 统计量的绝对值大于 1.96，也就是在 5% 显著性水平下拒绝原假设。其实，我们计算的 t 统计量绝对值还要大于 2.58（1%）。

3.2.2 置信区间

从样本数据并不能得到系数的真值。但是，我们能根据 OLS 估计量和标准误构建一个包含真值的置信区间。

系数 β_1 的 95% 置信区间：

1、用 5% 显著性水平的双边假设检验不能拒绝的一系列值；

2、有 95% 的可能性包含 β_1 真值的区间

当样本规模很大时， β_1 的 95% 置信区间为

$$[\hat{\beta}_1 - 1.96SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96SE(\hat{\beta}_1)] \quad (3.24)$$

例如，班级规模与测试分数回归中的 β_1 的 95% 置信区间为 $[-2.28 \pm 1.96 \times 0.48] = [-3.22, -1.34]$

3.2.3 虚拟变量

迄今为止，我们讨论的自变量为连续型变量。还有一类回归因子为二值，即它只取两个值——0 和 1。例如，当班级规模小于 20 人时为小班，X 取值为 1，当班级规模大于等于 20 人时为大班，X 取值为 0。这样的变量也被称为指示变量、哑变量或虚拟变量。

虚拟变量回归与上述回归相同，但是对于虚拟变量回归系数的理解却有些不同。

二值因变量回归实际上就是执行了一个均值差分。假设 D_i 等于 0 或 1，取决于班级规模大小：

$$D_i = \begin{cases} 1, & X < 20 \\ 0, & X \geq 20 \end{cases}$$

总体回归方程为

$$Y_i = \beta_0 + \beta_1 D_i + u_i \quad (3.25)$$

因为 D_i 是二值，那么，不能再将 β_1 理解成斜率，因为回归方程不是一条线了。那么，我们应该如何理解 D_i 呢？当 $D_i = 0$ 时，回归方程变成

$$Y_i = \beta_0 + u_i \quad (3.26)$$

因为 $E(u_i|D_i) = 0$ ，所以 $E(Y_i|D_i = 0) = \beta_0$ 。也就是说， β_0 是大班的情况下的平均分数。类似地，当 $D_i = 1$ 时，回归方程变成

$$Y_i = \beta_0 + \beta_1 + u_i \quad (3.27)$$

因此， $E(Y_i|D_i = 1) = \beta_0 + \beta_1$ ；即是说 $\beta_0 + \beta_1$ 是小班的平均分。

综上所述， $(\beta_0 + \beta_1) - \beta_0 = \beta_1$ 就是小班和大班平均分数的差异。换句话说， $\beta_1 = E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$ 。因为 β_1 是总体均值之间的差异，因此，OLS 估计量就是两个组的 Y 的平均值之差。

假设检验和置信区间与前面内容相同。

例如，小班教学的例子中，设置学生-老师比小于 20 时虚拟变量为 1，其余为 0。回归结果如下图所示。

Source	SS	df	MS	Number of obs	=	420
Model	5605.54742	1	5605.54742	F(1, 418)	=	15.99
Residual	146504.046	418	350.488149	Prob > F	=	0.0001
				R-squared	=	0.0369
				Adj R-squared	=	0.0345
Total	152109.594	419	363.030056	Root MSE	=	18.721

testscor	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
D	7.37241	1.843475	4.00	0.000	3.748774 10.99605
_cons	649.9788	1.387717	468.38	0.000	647.2511 652.7066

图 3.4: 虚拟变量回归结果

3.3 STATA 教程（一）

Stata 是一款流行的统计软件包。目前已经更新至 stata15，更多详细信息可参见www.stata.com。本讲稿向大家介绍 Stata 以及上述回归的操作。

我使用的 Stata14 MP 版。点击桌面的“stata”图标，打开之后的界面如下图

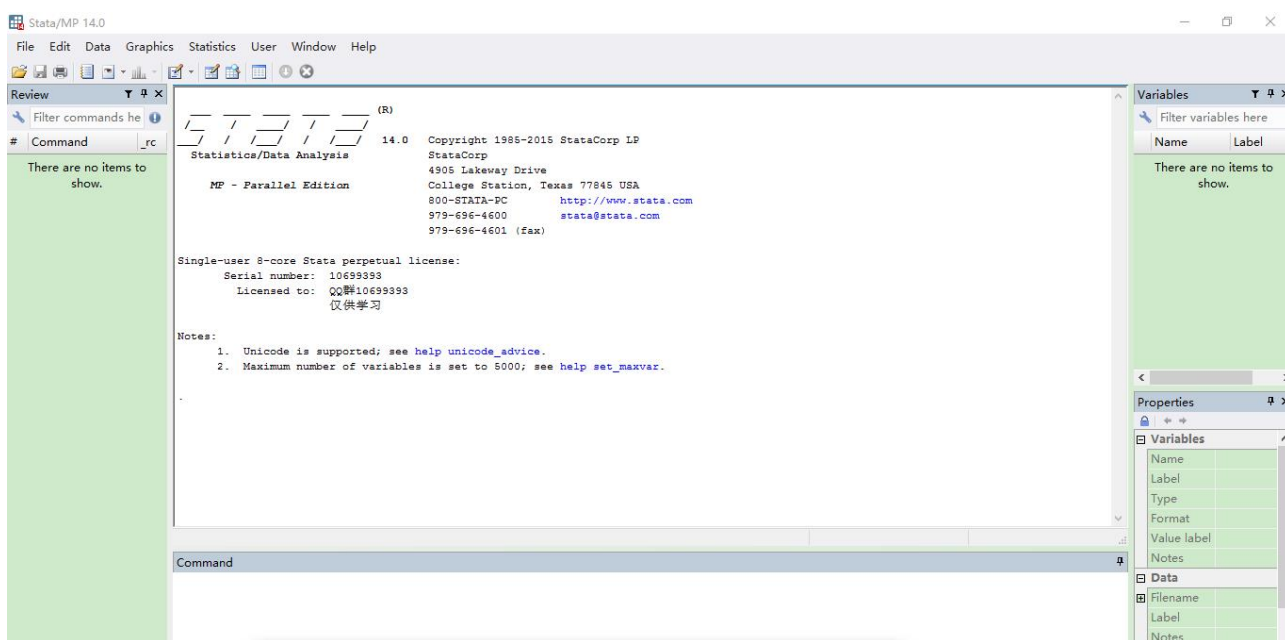


图 3.5: stata 界面

stata 面板最上面是“菜单栏”

左边窗口是“历史命令”

中间上窗口是“结果显示”

中间下窗口是“命令”

右边上窗口是“变量名”

(1) 数据输入

首先点击“菜单栏”中的“Data”—“Data Editor”，选择“Data Editor (Edit)”，就会出现如下窗口

在这个界面，我们可以手动输入数据，也可以直接从 Excel 中复制粘贴。我们输入的数据如下：

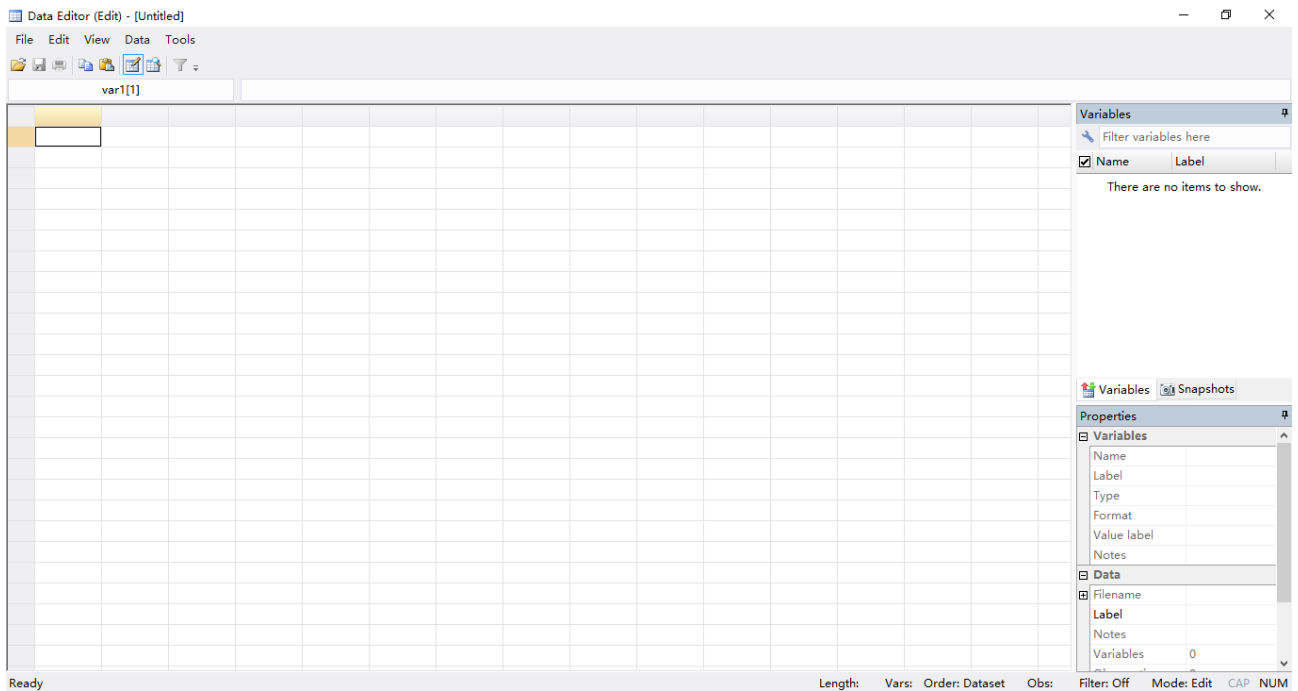


图 3.6: 数据输入界面

表 3.2: 输入数据

obs	testscr	str
1	690.8	17.889
2	661.2	21.5247
3	643.6	18.6713
⋮	⋮	⋮

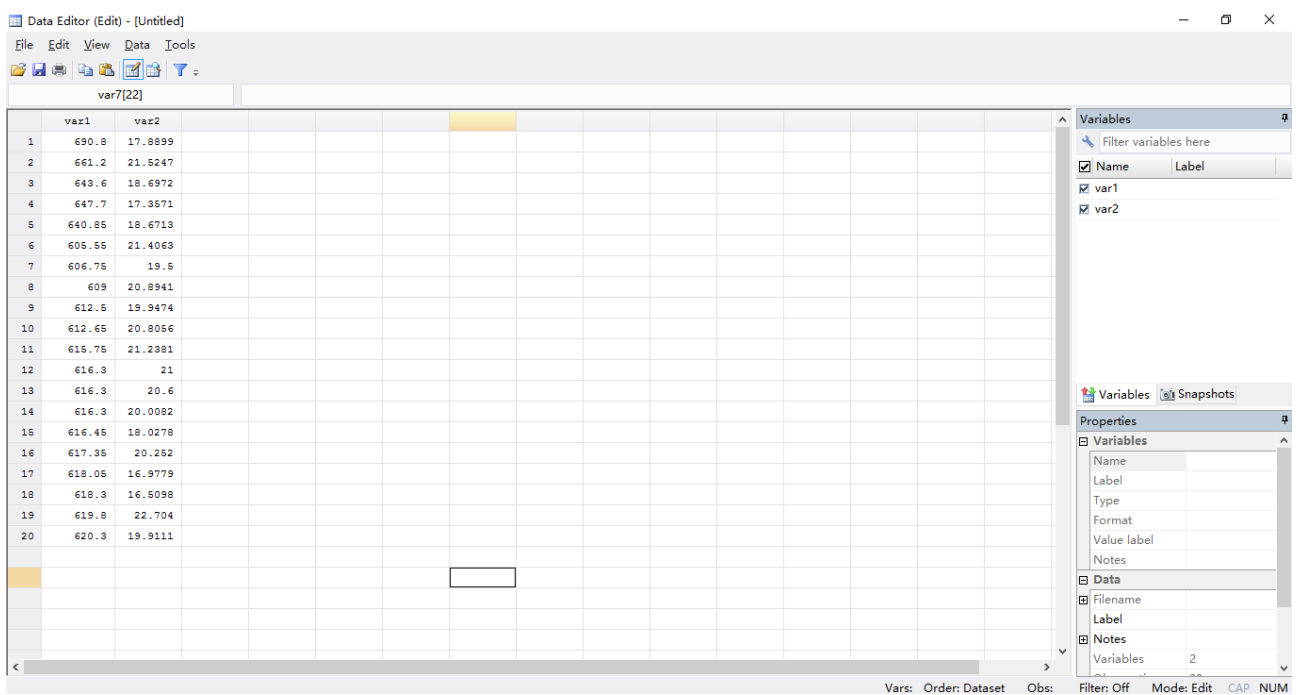


图 3.7: 数据输入界面

得到如下界面：

单击第一列的灰色方框，可以看到右侧下窗口“properties”变成

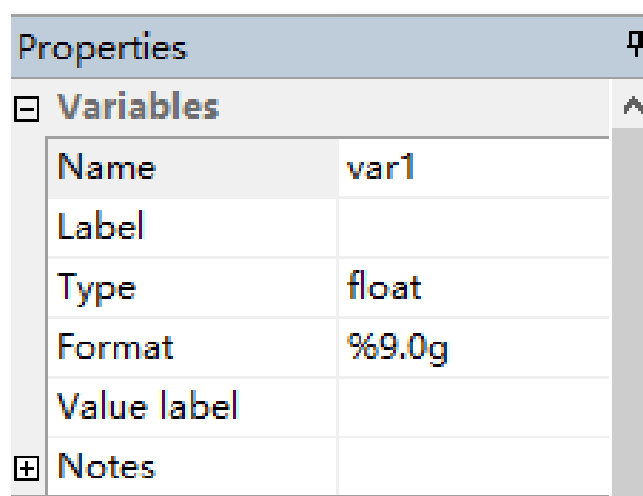


图 3.8: 数据输入界面

单击其中的“name”，修改“var1”为“testscr”。同理，也可以把“var2”修改为“str”。得到下图的数据输入结果。

与此同时，我们还可以在 stata 主面板上看到如下结果

打开 data editor (edit) 的另一种方式是点击菜单栏中的表格按钮“data editor (edit)”。

这样输入数据很麻烦，也会出很多错误。下面还会介绍另一种输入数据的方式。

通常，我们会查看一下现存的一些变量，可以输入下列命令

```
list varname1 varname2 ...
```

我们上面的变量名，所要输入的命令是

```
list testscr str
```

这个命令将会把所有变量的观测值都列示在结果窗口中。缺失数据会用“.”表示。但是一旦样本量大了，这种列示所有观测数据的方法就不适用了。要想终止列示进程，可以点击菜单栏中的“break”按钮。以后再介绍另一些检查错误的方式。你会看到如下界面

如本讲中，我们需要知道样本数据的统计特征。我们可以输入如下命令

```
sum testscr str,detail
```

我们可以得到下图

散点图的命令为

```
scatter testscr str
```

得到的图形如下

我们还想看看散点图的拟合线。命令如下

```
twoway scatter testscr str || lfit testscr str
```

得到的图如下：

而简单的回归的命令为


```
reg testscr str
```

得到的结果如下







而稳健标准误的回归命令为

```
reg testscr str,r
```

得到的结果如下

 Data Editor (Edit) - [Untitled]

File Edit View Data Tools

var8[12]

	testscr	str	
1	690.8	17.8899	
2	661.2	21.5247	
3	643.6	18.6972	
4	647.7	17.3571	
5	640.85	18.6713	
6	605.55	21.4063	
7	606.75	19.5	
8	609	20.8941	
9	612.5	19.9474	
10	612.65	20.8056	
11	615.75	21.2381	
12	616.3	21	
13	616.3	20.6	
14	616.3	20.0082	
15	616.45	18.0278	
16	617.35	20.252	
17	618.05	16.9779	
18	618.3	16.5098	
19	619.8	22.704	
20	620.3	19.9111	

图 3.9: 数据输入界面

#	Command	_rc
1	set obs 1	
2	generate var1 = 1 in 1	
3	rename var1 testscr	
4	rename testscr var1	
5	replace var1 = 620.3 in 20	
6	rename var1 testscr	
7	rename var2 str	

图 3.10: 数据输入界面

Variables		Filter variables here
Name	Label	
testscr		
str		

图 3.11: 数据输入界面

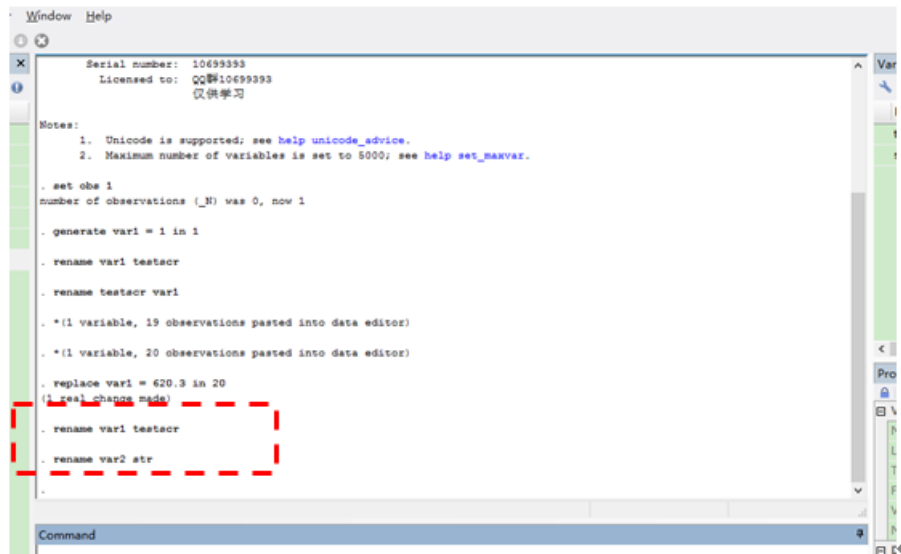


图 3.12: 数据输入界面

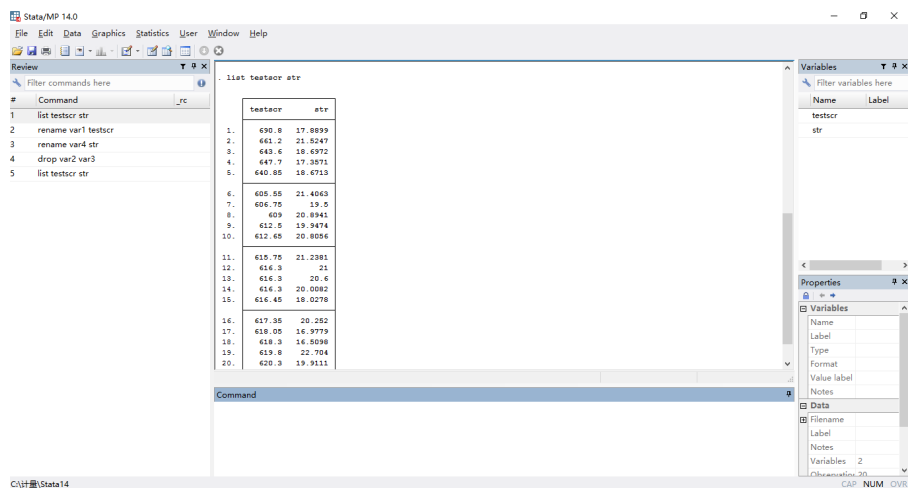


图 3.13: 观测值列表

```
. sum testscr str, detail
```

testscr					
Percentiles	Smallest				
1%	605.55	605.55			
5%	606.15	606.75			
10%	607.875	609	Obs		20
25%	614.2	612.5	Sum of Wgt.		20
50%	616.9		Mean	625.275	
			Std. Dev.	21.29214	
75%	630.575	643.6			
90%	654.45	647.7	Variance	453.3554	
95%	676	661.2	Skewness	1.797063	
99%	690.8	690.8	Kurtosis	5.649889	

str					
Percentiles	Smallest				
1%	16.5098	16.5098			
5%	16.74384	16.97787			
10%	17.16751	17.35714	Obs		20
25%	18.34955	17.88991	Sum of Wgt.		20
50%	19.97779		Mean	19.69612	
			Std. Dev.	1.694211	
75%	20.94706	21.23809			
90%	21.46546	21.40625	Variance	2.870351	
95%	22.11434	21.52466	Skewness	-.328499	
99%	22.70402	22.70402	Kurtosis	2.169139	

图 3.14: 统计量

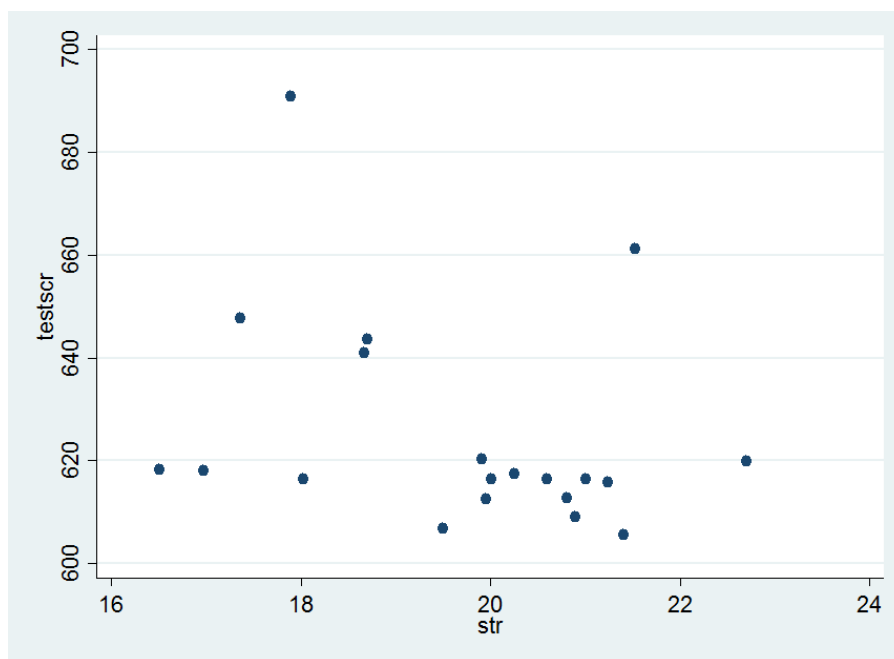


图 3.15: 散点图

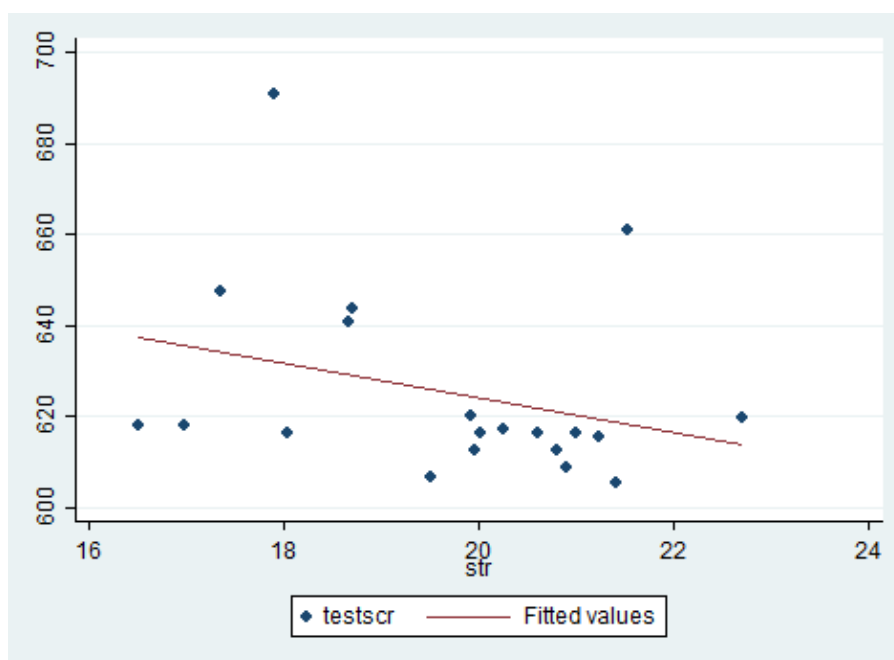


图 3.16: 拟合线

```
. reg testscr str
```

Source	SS	df	MS	Number of obs	=	20
Model	799.805171	1	799.805171	F(1, 18)	=	1.84
Residual	7813.94765	18	434.108203	Prob > F	=	0.1914
				R-squared	=	0.0929
				Adj R-squared	=	0.0425
Total	8613.75282	19	453.355412	Root MSE	=	20.835

testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
str	-3.829551	2.821335	-1.36	0.191	-9.756957	2.097855
_cons	700.7023	55.76432	12.57	0.000	583.5458	817.8588

图 3.17: 简单回归结果

```
. reg testscr str,r
```

Linear regression	Number of obs	=	20
	F(1, 18)	=	1.47
	Prob > F	=	0.2404
	R-squared	=	0.0929
	Root MSE	=	20.835

testscr	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
str	-3.829551	3.154197	-1.21	0.240	-10.45627	2.79717
_cons	700.7023	63.98125	10.95	0.000	566.2827	835.1219

图 3.18: 稳健标准误回归结果

第4章 多元线性回归

第三讲的回归结果显示：学生老师比越小，平均分数越高。但是，肯定有人怀疑这一结论，因为小班的学生可能还有其他的因素使得其平均分数更高。

在第三讲中，这些被遗漏的因素全部“丢进”了误差项 u_i 中。但是这会使得 OLS 估计量产生偏误，我们在下面的内容中将详细阐述。那么，怎么解决“遗漏变量偏误”呢？多元回归就是消除遗漏变量偏误的一种方法。

多元回归的 idea 很直观：如果那些遗漏变量的数据可用，那么，我们就能将这些变量纳入回归方程中作为回归因子，并且在保持其它变量不变的情况下，估计出一个回归因子的效应。

4.1 遗漏变量偏误

第三讲用小班教学作为例子。在这个例子中，班级规模（学生-老师比）越小，平均成绩越高。但是，仅仅考虑学生-老师比这一个因素不够，还忽略了许多重要的潜在决定因素对测试成绩的影响。这些潜在影响因素包括：学校特征（教师质量、硬件设备等等）、学生特征（家庭背景、语言差异等等）。下面，我们以语言差异为例。

大家都知道，中国方言甚多，甚至同一个城市里不同区域的方言也不相同。在湖北省高考语文试题中，经常考字的读音。湖北人普通话不标准，因为前鼻音“l”和后鼻音“n”不分，平舌“si”和翘舌“shi”不分。还有很多的地方的人“飞”和“灰”不分。因此，估计很多学生都怕读音题，尤其是南方人。

那么，如果一个班里南方人比例多，而另一个班级里北方人比例高，如果考读音题，估计南方人多的班级平均分会低于北方人多的班级。如果我们忽略这种语言差异，仅仅用学生-老师比来回归，预期班级规模对测试成绩的效应会有偏。因为南方学生在读音题上的得分可能低于北方学生。如果大班中有许多南方人，那么，学生-老师比的 OLS 回归系数可能会高估对测试分数的效应。

4.1.1 遗漏变量偏误的定义

如果一个回归元与模型中遗漏的变量有关，且这个遗漏变量还是因变量的决定因素，那么，OLS 估计量会产生遗漏变量偏误。

遗漏变量偏误产生必须同时满足两个条件：

- 1、X 与遗漏变量相关；
- 2、遗漏变量是因变量 Y 的一个决定因素。

遗漏变量偏误与第一个 LS 假设。回顾一下第三讲中有关 OLS 的三个假设，其中，第一个是 $E(u_i|X_i) = 0$ 。遗漏变量偏误就意味着这个假设不成立。

在一元回归中， u_i 包含除了 X_i 以外所有决定 Y 的因素。如果这些遗漏的因素中有一个与 X_i 相关，那么，误差项 u_i 就与 X_i 相关。因此， $E(u_i|X_i) \neq 0$ 。这个假设不成立，后果很严重：OLS 估计量有偏。即使在大样本下，偏误也不会消除，而且 OLS 估计量不是一致估计量。

因为，遗漏变量与 X_i 相关，我们定义 $\text{corr}(X_i, u_i) = \rho_{Xu}$ 。假设 LS 的第二和第三个假设仍然成立。那么，OLS 估计量就有下列极限：

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_X} \quad (4.1)$$

也就是说，随着样本规模的增大， $\hat{\beta}_1$ 以越来越高的概率趋近于 $\beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_X}$ 。

note:

1、无论样本规模大小，遗漏变量偏误问题都要引起注意。从公式 (1) 中可以看出， $\hat{\beta}_1$ 并不收敛到真值 β_1 。且偏误的大小为 $\rho_{Xu} \frac{\sigma_u}{\sigma_X}$ 。

2、偏误的大小取决于误差项与 X 的相关系数 ρ_{Xu} 。 $|\rho_{Xu}|$ 越大，偏误越大。

3、偏误的方向（也就是，系数高估还是低估）取决于误差项与 X 是正相关还是负相关。如果 $\rho_{Xu} < 0$ ，OLS 估计量就是低估，反之亦然。

例子：听莫扎特可以提高智力？！

在孩子教育问题方面，流传着这样的一个故事：让孩子每天听听莫扎特的音乐，可以提高孩子的智力。其实，这是 Rauscher et al. (1993) 在 Nature 上发表的研究成果。他们建议，听 10-15 分钟的莫扎特会暂时性提高 IQ8-9 个点。

真的存在“莫扎特效应”吗？如果存在，提高 8-9 点 IQ 是高还是低了？现在我们学了一点计量了，我们用计量经济学的语言，这个效应估计可能存在遗漏变量偏误。

4.2 多元回归模型

既然遗漏变量偏误是由于某些决定 Y，而又与 X 相关的变量没有包含在回归方程中，那么，只要这些变量数据可用，我们只要把它们纳入回归方程中就可以消除遗漏变量偏误。这就是**多元回归模型**。多元回归模型可以在保持 X_2 不变的情况下，估计出 X_1 对 Y 的效应。

总体回归线

假设只有两个自变量 X_{1i} 和 X_{2i} 。在线性多元回归模型中，自变量和因变量之间的关系由下式给出

$$E(Y_i | X_{1i} = x_1, X_{2i} = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad (4.2)$$

公式 (2) 称为**总体回归线**，或者**总体回归函数**。多元回归模型中一个或多个自变量有时候也称为**控制变量**。公式 (2) 中的系数 β_1 含义与一元回归中有些不同。在多元回归中，这个系数是保持 X_2 为常数或者控制 X_2 时， X_1 的单位变化引起 Y 的变化。这个系数也称为**局部效应**。

总体回归方程

正如一元回归，多元回归线也不能精确表示自变量与 Y 之间的关系，因为还有许多影响 Y 的因素并没有包含在多元回归线中。因此，公式 (2) 也需要包含误差项来代表其它因素。即

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1, \dots, n \quad (4.3)$$

我们可以把 β_0 理解成是值为 1 的自变量的系数。因为该自变量的值恒为 1，因此称为常自变量。类似， β_0 也被称为**常数项**。

在实践中，多元回归模型通常包含两个以上的自变量，形式如下

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i, \quad i = 1, \dots, n \quad (4.4)$$

4.2.1 多元回归中的 OLS 估计量

回忆一下，一元回归的 OLS 估计量：选择系数来最小化预测误差平方和，即选择 $\hat{\beta}_0, \hat{\beta}_1$ 来最小化 $\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$ 。 $\hat{\beta}_0, \hat{\beta}_1$ 就是 OLS 估计量。

这一思想也可以沿用至多元回归的系数估计。即

$$\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_k X_{ki})^2 \quad (4.5)$$

其中， $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ 是 OLS 估计量。而 OLS 残差用 $\hat{u}_i = Y_i - \hat{Y}_i$ 表示。

上面的 OLS 估计量公式有点麻烦。但是，幸运地是这些计算公式都已经编写进了统计软件中，例如 Stata，我们把数据输入后，软件就可以直接给出结果。

应用

第三讲中，学生-老师比对测试成绩的效应，用的美帝加利福利亚州 420 个观测样本估计的回归模型为

$$\hat{Y} = 698.9 - 2.28 \times X \quad (4.6)$$

但是，通过上面内容的讲解，我们担心这个回归模型对小班教学效应的估计不准确。因为它存在遗漏变量偏误问题。因为美帝是一个移民国家，学校里有许多学生母语是非英语，因此，其在测试分数上表现稍微差一些。我们正好也有加利福利亚州学生母语为非英语人数的数据。那么，我们就可以在上述一元回归模型中引入

非母语学生变量，从而消除遗漏变量偏误问题。得到的多元回归方程为

$$\hat{Y} = 686.0 - 1.10 \times X_1 - 0.65 \times X_2 \quad (4.7)$$

其中， X_1 表示学生-老师比 (str)， X_2 表示非英语母语学生 (elpct)。

stata 结果为

Source	SS	df	MS	Number of obs	=	420
Model	64864.3011	2	32432.1506	F(2, 417)	=	155.01
Residual	87245.2925	417	209.221325	Prob > F	=	0.0000
				R-squared	=	0.4264
				Adj R-squared	=	0.4237
Total	152109.594	419	363.030056	Root MSE	=	14.464

testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
str	-1.101296	.3802783	-2.90	0.004	-1.848797 - .3537945
el_pct	-.6497768	.0393425	-16.52	0.000	-.7271112 - .5724423
_cons	686.0322	7.411312	92.57	0.000	671.4641 700.6004

图 4.1: 多元回归结果

将一元回归方程 (6) 和多元回归方程 (7) 中，学生-老师比对测试分数的效应的 OLS 估计结果进行对比。多元回归中， β_1 的 OLS 估计量为 -1.10，这几乎是一元回归估计量的一半。也就是说，多元回归中班级规模对测试分数的效应是一元回归中估计地效应的一半。这是因为在多元回归中，-1.10 表示保持 X_2 不变时，班级规模的效应，而 -2.28 则表示班级规模与非英语母语学生都在变化时的效应。

这种对比也可以看出，一元回归存在遗漏变量偏误。估计出的班级规模效应偏大。

4.2.2 拟合度

与一元回归类似，多元回归也有三个常用的统计量来检验回归方程对数据的拟合程度，它们分别是：SER、 R^2 和调整 R^2 (\bar{R}^2)。

SER

SER 估计误差项 u_i 的标准差。在多元回归中，SER 为

$$SER = s_{\hat{u}} = \sqrt{s_{\hat{u}}^2} \text{ where } s_{\hat{u}}^2 = \frac{1}{n-k-1} \sum_{i=1}^n \hat{u}_i^2 = \frac{SSR}{n-k-1} \quad (4.8)$$

上式与第三讲中的 SER 公式的差异在分母是 $n-k-1$ ，而不是 $n-2$ 。第三讲中，除数 $n-2$ 是为了调整由估计两个系数而引起的向下偏误。而此处 $n-k-1$ 则是为了调整估计 $k+1$ 个系数 (k 个斜率和一个截距) 引起的向下偏误。 $n-k-1$ 成为自由度。当 n 很大时，自由度调整可以忽略。

R^2

回忆一下，第三讲中 R^2 定义为回归因子所能解释的 Y_i 的样本方差比例，或者 1 减回归因子不能解释的样本方差比例。即

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \quad (4.9)$$

其中，回归平方和 $ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ ，总平方和 $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$ 。

多元回归中的 R^2 定义同上。但是，特别需要注意的是，除非增加的回归量的系数为 0，否则随着回归量的增加， R^2 逐渐增大。根据 OLS，选择系数值来最小化残差平方和 SSR。(1) 如果增加的回归量的系数为 0，那么，SSR 不会随着这个增加的回归量而变化。(2) 如果增加的回归量系数不为 0，那么，增加该回归量之后的 SSR 会变小，从公式 (9) 可知， R^2 变大。

那么，增加回归量， R^2 变大，是否意味着增加回归量就提高了模型的拟合程度呢？答案是否定的。因此，就

需要纠正多元回归中的 R^2 ，为此，提出了调整的 R^2 ，即 \bar{R}^2

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \frac{SSR}{TSS} = 1 - \frac{s_a^2}{s_y^2} \quad (4.10)$$

公式 (10) 与 (9) 之间的差别就是残差平方和与总平方和之比前面成了一个因子 ($\frac{n-1}{n-k-1}$)。关于 \bar{R}^2 有三点需要注意：

第一， $\frac{n-1}{n-k-1}$ 总是大于 1，因此， $\bar{R}^2 < R^2$ ；

第二，增加一个回归量对 \bar{R}^2 有正负两个方面的影响，一方面，SSR 下降， \bar{R}^2 上升；另一方面，因子 $\frac{n-1}{n-k-1}$ 变大， \bar{R}^2 变小。因此， \bar{R}^2 变大变小取决于这两个效应谁占主导地位；

第三， \bar{R}^2 可以为负数。当增加回归量，SSR 下降的程度不足以抵补 $\frac{n-1}{n-k-1}$ 的下降，那么 \bar{R}^2 就可能为负。

示例：从上文的图 1 中可以看出， $R^2 = 0.4264$ ，而 $\bar{R}^2 = 0.4237$ ， $SE_R = 14.464$ 。将这些结果与第三讲中的一元回归结果进行对比， R^2 从 0.051 上升到 0.4264，也就是说只有学生-老师比这一个自变量时，自变量只能解释测试分数方差的 5.1%，而增加非英语母语学生这个自变量时，两个自变量可以解释测试分数方差的 42.64%。从这个意义上看，增加一个自变量确实提高了回归模型的拟合程度。因为样本量 $n = 420$ ，回归量 $k = 2$ ，因此， R^2 与 \bar{R}^2 之间的差异就非常小。

此外，SER 也从一元回归的 18.6 上升到多元回归的 14.5，这也说明拟合的更好。

提醒：虽然 \bar{R}^2 与 R^2 很有用，但是太依赖于 \bar{R}^2 就会掉进陷阱。在实际应用中，“最大化 \bar{R}^2 ”几乎不能回答任何有意义的计量或统计问题。相反，是否要增加一个变量应该基于增加这个变量可以更好的估计出我们感兴趣的因果效应。

4.2.3 多元回归中的 OLS 假设

与一元回归类似，多元回归中 OLS 也有一些假设：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \quad i = 1, \cdots, n \quad (4.11)$$

其中，

- 1、给定 $X_{1i}, X_{2i}, \cdots, X_{ki}$ 的条件下， u_i 的条件均值为 0，即 $E(u_i | X_{1i}, X_{2i}, \cdots, X_{ki}) = 0$ ；
- 2、 $(X_{1i}, X_{2i}, \cdots, X_{ki}, Y_i)$ 独立同分布 (i.i.d.)；
- 3、不可能出现较大奇异值： $X_{1i}, X_{2i}, \cdots, X_{ki}, Y_i$ 有非零有限的四阶矩；
- 4、不存在完全多重共线。

4.3 假设检验与置信区间

多元回归为消除遗漏变量偏误问题提供了一种方法。但多元回归中的 OLS 估计量也存在抽样不确定性。与一元回归不同，多元回归的假设可能包含两个或多个回归系数。检验这种“联合”假设的统计量，称为 **F 统计量**。

4.3.1 单系数假设检验与置信区间

回忆一下，一元回归系数的方差是由第三讲公式 (20) 给出的。在 LS 假设下，大数法则意味着样本均值会收敛到总体均值，因此， $\frac{\hat{\sigma}_{\hat{\beta}_1}^2}{\sigma_{\hat{\beta}_1}^2} \xrightarrow{P} 1$ 。 $\sigma_{\hat{\beta}_1}^2$ 的平方根就是 $\hat{\beta}_1$ 的标准误， $SE(\hat{\beta}_1)$ 。

这个计算标准误的方法也可以推广至多元回归。

4.3.1.1 单系数假设检验

一般来讲，我们想要检验多元回归第 j 个自变量的系数 β_j 等于某一确定值 $\beta_{j,0}$ 。这个特定的值要么来源于经济理论，要么来源于实际应用中的决策值。如果备择假设是双边假设，那么，原假设与备择假设为

$$H_0 : \beta_j = \beta_{j,0} \text{ vs. } H_1 : \beta_j \neq \beta_{j,0} \quad (4.12)$$

例如，在小班教学的例子中，原假设就是 $\beta_1 = 0$ 。我们的任务就是要用样本数据来检验原假设和备择假设。与一元回归的假设检验步骤类似，多元回归假设检验步骤如下：

第一步，计算 $\hat{\beta}_j$ 的标准误， $SE(\hat{\beta}_j)$ ；

第二步，计算 t 统计量

$$t = \frac{\hat{\beta}_j - \beta_{j,0}}{SE(\hat{\beta}_j)} \quad (4.13)$$

第三步，计算 p 值

$$p = 2\Phi(-|t^{act}|) \quad (4.14)$$

其中， t^{act} 是计算出来的实际 t 统计量。如果 p 值小于 0.05 或者 $|t^{act}| > 1.96$ ，那么就在 5% 的显著性水平下拒绝原假设。

注：我们从上面的 stata 结果可以看出，标准误、t 统计量和 p 值都是由软件自动输出的。

Source	SS	df	MS	Number of obs	=	420
Model	64864.3011	2	32432.1506	F(2, 417)	=	155.01
Residual	87245.2925	417	209.221325	Prob > F	=	0.0000
				R-squared	=	0.4264
				Adj R-squared	=	0.4237
Total	152109.594	419	363.030056	Root MSE	=	14.464

	标准误 ψ	t 统计量 ψ	p 值 ψ		
testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
str	-1.101296	.3802783	-2.90	0.004	-1.848797 - .3537945
el_pct	-.6497768	.0393425	-16.52	0.000	-.7271112 - .5724423
_cons	686.0322	7.411312	92.57	0.000	671.4641 700.6004

图 4.2: 标准误、t 统计量和 p 值

4.3.1.2 置信区间

多元回归的置信区间与一元回归相同。

例如，系数 β_j 的 95% 的双边置信区间以 95% 的概率包含 β_j 的真实值。等价地， β_j 的一系列值不能被 5% 的双边假设检验所拒绝。当样本规模很大时，95% 的置信区间为

$$95\% \text{ conf. interval} = [\hat{\beta}_j - 1.96SE(\hat{\beta}_j), \hat{\beta}_j + 1.96SE(\hat{\beta}_j)] \quad (4.15)$$

注：如果是 90% 的置信区间，就用 1.64 代替 1.96；如果是 99% 的置信区间，就用 2.58 取代 1.96。

提醒：无论是假设检验的方法和置信区间的方法都依赖于大样本正态近似于 OLS 估计量的分布。因此，要时刻记住这些量化抽样不确定性的方法仅仅在大样本下才起作用。

示例 1：图 2

由图 2 的多元回归结果可知，学生-老师比的系数为-1.10，SER 为 0.38，而原假设为 $\beta_1 = 0$ ，因此，t 值为 $\frac{-1.10-0}{0.38} = -2.89$ ，这一结果与图 2 中显示的一致。对应的 p 值为 0.4%，因为 p 值小于 5%，所以在 5% 的显著性水平下拒绝原假设（这个值甚至小于 1% 的显著性水平）。

由图 2 还可知道 95% 的置信区间为 $[-1.85, -0.35]$ 。

示例 2：加入其它控制变量

假设还有其它因素影响测试成绩，例如生均教师支出（expn）。那么，将生均教师支出加入回归方程后的结果为

从图 3 中，我们看到了十分有趣的结果：加入 expn 之后，str（学生-老师比）的效应更小了，仅为-0.286，而且 t 统计量仅为-0.6，对应的 p 值为 0.551。也就是说，总体回归中这个系数为 0 的假设不能在 10% 的显著性水平下被拒绝。其经济含义就是，保持生均教师支出和非英语母语学生不变的情况下，没有证据显示，小班教学会提高测试成绩。

Source	SS	df	MS	Number of obs	=	420
Model	66409.8837	3	22136.6279	F(3, 416)	=	107.45
Residual	85699.7099	416	206.008918	Prob > F	=	0.0000
				R-squared	=	0.4366
				Adj R-squared	=	0.4325
Total	152109.594	419	363.030056	Root MSE	=	14.353

testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
str	-.2863992	.4805232	-0.60	0.551	-1.230955 .658157
expn_str	.0038679	.0014121	2.74	0.006	.0010921 .0066437
el_pct	-.6560227	.0391059	-16.78	0.000	-.7328924 -.5791529
_cons	649.5779	15.20572	42.72	0.000	619.6883 679.4676

图 4.3: 加入生均支出的回归结果

从这个结果还可以有另一种理解或推断：教育管理部门有效地批准了教育资金。假设一种相反的结果，加入生均教师支出后，str 的系数很大，且为负。那么，教育管理部门只需要减少其它的教育支出（例如，课本、教学设备等等），而将其用于雇佣更多的教师。这样既可以保持教育支出不变，也而缩减班级规模，从而使得测试成绩提高。但是上面的回归结果却是 str 的系数很小，且统计不显著，这也就意味着将其他教育支出转移至教师支出这种资金配置不会对测试成绩的提高产生效应。从而推断出教育管理部门有效地配置了教育资金。

小贴士：从图 3 中，str 的标准误从图 2 中的 0.38 变为图 3 中的 0.48，这说明 str 和 expn 可能存在多重共线性。而多重共线性会导致 OLS 估计不精确。

4.3.2 联合假设检验

如果原假设为学生-老师比和生均支出的系数同时为 0，该如何检验呢？

我们控制非英语母语学生变量，联合假设为

$$H_0 : \beta_1 = 0, \beta_2 = 0 \text{ vs. } H_1 : \beta_1 \neq 0, \beta_2 \neq 0 \quad (4.16)$$

联合假设就是对两个或多个回归系数施加限制。只要原假设中的任何一个系数等式不成立，那么，联合原假设就为假。

那么，我们为什么不能一次检验一个系数呢？

如果我们对上述联合假设检验感兴趣，并分别用 t_1 和 t_2 来检验第一个系数为 0 和第二个系数为 0。那么，只要 t_1 和 t_2 中有一个大于 1.96 就应该拒绝原假设？

由于这个问题中涉及两个随机变量 t_1 和 t_2 ，因此，需要刻画它们的联合抽样分布。在大样本下， β_1 和 β_2 有联合正态分布，因此， t_1 和 t_2 有一个双变量正态分布。假设两个 t 统计量不相关且独立。那么，不能拒绝原假设当且仅当 $|t_1| \geq 1.96, |t_2| \geq 1.96$ 。 $Pr(|t_1| \geq 1.96, |t_2| \geq 1.96) = 0.95^2 = 0.9025$ 。因此，拒绝原假设的概率为 9.57%(1-0.9025)。

如果这两个 t 统计量相关，那么，这种情形更加复杂。但是“一次检验一个”的方法不会得到一个合意的显著性水平。而另一种法就是基于 F 统计量的检验。

4.3.2.1 F 统计量

F 统计量用于检验回归分析中的联合假设。Stata 软件可以直接输出 F 统计量。当联合原假设为两个系数为 0 时，F 统计量的公式为

$$F = \frac{1}{2} \left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1, t_2} t_1 t_2}{1 - \hat{\rho}_{t_1, t_2}^2} \right) \quad (4.17)$$

其中， $\hat{\rho}_{t_1, t_2}$ 是两个 t 统计量相关系数的估计值。

下面来看看 q 个系数的联合假设检验。大样本下, F 统计量有 F_q^∞ 分布。因此, F 统计量的临界值就能通过 $F_{q,\infty}$ 分布表查出来, 得到一个合适的显著性水平。 F 统计量计算的 p 值为

$$p = Pr[F_{q,\infty} > F^{act}] \quad (4.18)$$

上式计算出来的 p 值与 F 分布临界值对比。在给大家上课的时候, 虽然把这些步骤都告诉大家, 但是 stata 等软件直接帮我们略去了这些细节, 直接给出结果。

F 统计量检验所有系数为 0 的联合假设。即原假设和备择假设为

$$H_0: \beta_1 = 0, \beta_2 = 0, \dots, \beta_k = 0 \text{ vs. } H_1: \beta_j \neq 0, j = 1, 2, \dots, k \quad (4.19)$$

其中, 备择假设说明至少有一个 j 不等于 0。

在原假设下, 没有回归量能解释 Y 的方差。当 $q = 1$ 时, F 统计量检验单一系数的限制。联合假设就退化成一元回归系数的假设, F 统计量就是 t 统计量的平方。

示例

检验学生-老师比和生均支出的系数为 0。计算 F 统计量为 5.43。如下图所示

```
. vce
Covariance matrix of coefficients of regress model

      e(V) |      str      expn_stu      el_pct      _cons
-----+-----
      str |      .2323942
      expn_stu |      .00040035      2.499e-06
      el_pct |      -.00244872      -.00001024      .00101025
      _cons |      -6.6649192      -.02070346      .08180682      238.96038

. test str expn_stu

( 1)  str = 0
( 2)  expn_stu = 0

      F( 2, 416) =      5.43
      Prob > F =      0.0047
```

图 4.4: F 统计量

在原假设下, 大样本性质使得 F 统计量有 $F_{2,\infty}$ 分布。查 F 分布临界值表, $F_{2,\infty}$ 分 5% 的临界值为 3.00, 1% 的临界值为 4.61。而计算得到的 F 统计量为 5.43, 大于 1% 显著性水平下的临界值, 因此, 在 1% 的显著性水平下拒绝原假设。

4.3.3 多元回归模型设定

多元回归中有两个或多个自变量, 那么, 如何决定哪些变量要放进多元回归中呢? 目前, 还没有“万能灵药”来应对所有情形。但是也不用失望, 因为有许多指导性建议可用。选择一个变量作为自变量, 应该从可能的遗漏变量偏误着手。这有赖于你们对经验问题的专业知识, 由此获得一个无偏的因果效应。而不是仅仅完全依赖于统计拟合程度, 例如 R^2 和 \bar{R}^2 。

回忆一下, 本讲第一节提到的遗漏变量偏误, 必须满足两个条件: (1) 至少有一个回归量与遗漏变量相关; (2) 遗漏变量必须是因变量 Y 的决定因素。

这也就意味着给定 $X_{1i}, X_{2i}, \dots, X_{ki}, u_i$ 的条件期望为非零, 这就会打破 LS 第一个假设。因此, 即使在大样本下, 遗漏变量偏误还是会存在。即是说, 遗漏变量偏误使得 OLS 估计量非一致。

上面的例子隐含着核心解释变量 (我们希望估计的因果效应) 和控制变量。

控制变量并不是我们感兴趣的目标；它们是包含在多元回归中，保持为不变的因素；如果忽略它们就会导致感兴趣变量（核心解释变量）的因果效应遭遇遗漏变量偏误。

在 LS 第一假设上，我们来区别对待核心解释变量和控制变量。核心解释变量的 OLS 估计量是无偏的，但是控制变量的 OLS 估计量一般来讲是有偏的，因此并没有因果含义。

示例

考虑由于遗漏外部学习机会而引起的潜在遗漏变量偏误。外部学习机会非常抽象和广泛，因此，很难测量。但是这些机会与学生的经济背景有关，而经济背景可以测量。因此，经济背景就可以加入多元回归中，进而控制收入相关的遗漏因素。例如，我们在 `str` 和 `pctel` 之外，再加入受到免费午餐的学生比例 (`lchpct`)。那么回归结果为

理论和实践中的模型设定从理论来讲，如果遗漏变量数据可用，解决遗漏变量偏误，只需要在回归模型中加入遗漏变量即可。然而，在实践中，决定是否加入一个变量非常困难，并要三思而行。

从实践来看，应对遗漏变量偏误的方法：

基础解释变量集应用依靠专业判断、经济理论以及对数据的了解，并进行综合决策。包含基础解释变量的集合有时也称为**基准模型**。

第一步，基准模型应该包含主要的核心解释变量和控制变量，这些变量是根据专业判断和经济理论得到的。

第二步，有经济理论得到的变量经常没有可用的数据，那么，就需要提出许多备择模型设定，即一些备择的回归因子。

如果在备择模型中，核心解释变量的数值与基准模型相似，这就说明基准模型的估计结果可信。另一方面，如果核心解释变量的估计结果在备择模型中变化较大，那么，这说明基准模型存在遗漏变量偏误。我们将在第五讲中详细阐述遗漏变量偏误及其解决办法。

实践中， R^2 和 \bar{R}^2 能告诉我们什么？不能告诉我们什么呢？

R^2 和 \bar{R}^2 能告诉我们回归因子解释因变量的方差的程度。如果 R^2 或 \bar{R}^2 接近于 1，说明回归因子能作出对因变量的较好预测，才能够这个意义上来讲，OLS 残差方法较小。如果 R^2 或 \bar{R}^2 接近于 0，情况相反。

R^2 和 \bar{R}^2 不能告诉我们

- (1) 一个解释变量是否统计显著；
- (2) 回归因子是驱动因变量变动的真实原因；
- (3) 存在遗漏变量偏误；
- (4) 我们已经选择了最适合的解释变量集合。

4.4 多元回归 Stata 操作示例

这一节，我将利用小班教学的数据作为例子，来展示多元回归的 `stata` 操作及其结果分析与讨论。主要目的是为了说明利用多元回归如何消除遗漏变量偏误。

第一步，基准模型和备择模型的设定

从前面的讲稿内容可以看出，我们关心的班级规模（学生-老师比）对测试成绩的效应，且控制了学生的特征（例如，经济背景、非英语母语等）。除此之外，还有许多影响测试分数的潜在因素，且它们与学生-老师比相关。如果遗漏这些因素就会导致遗漏变量偏误。如果控制变量使得条件均值独立性假设成立，那么，学生-老师比的系数就是保持控制变量不变时班级规模对测试分数的效应。

我们考虑三个学生特征：非英语母语学生比例 (`elpct`)、接受免费午餐的学生比例 (`mealpct`)、有资格接受家庭收入援助的学生比例 (`calwpct`)。后面两个变量都可以刻画学生的经济背景。经济理论和转专业判断并不能告诉我们这两个变量中哪一个作为控制变量代表学生经济特征更合适。因此，我们把免费午餐学生比例作为基准回归模型，而把接受家庭收入援助的学生比例作为备择回归模型。

下面，我们分别作出测试分数 (`testscr`) 与非英语母语学生比例 (`elpct`)、接受免费午餐的学生比例 (`mealpct`)、有资格接受家庭收入援助的学生比例 (`calwpct`) 的散点图。

我们在 `stata` 命令栏中分别输入

```
scatter testscr elpct
```

```
scatter testscr mealpct
```

```
scatter testscr calwpct
```

可以得到下列三幅三点图

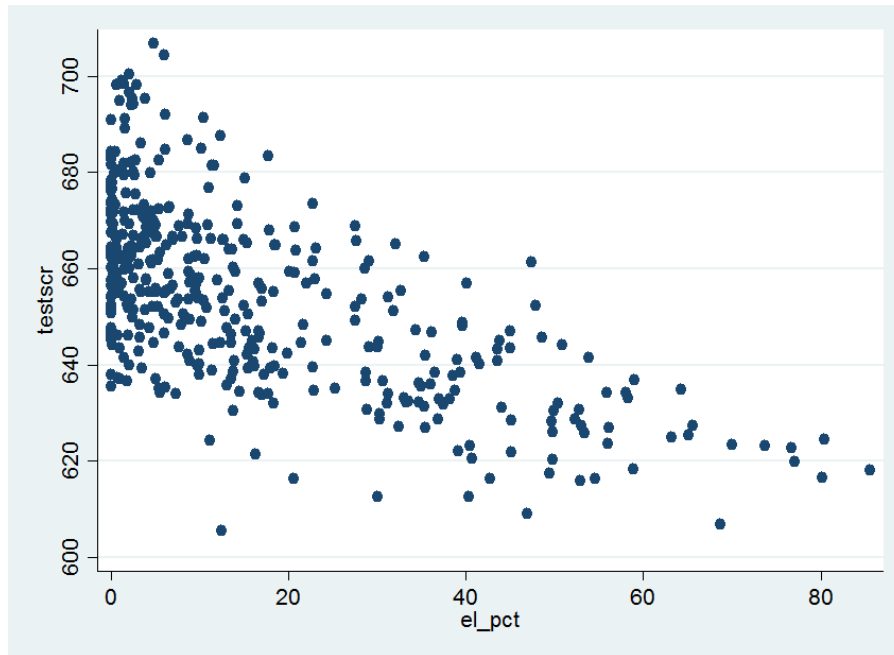


图 4.5: 测试分数与非英语母语学生比例

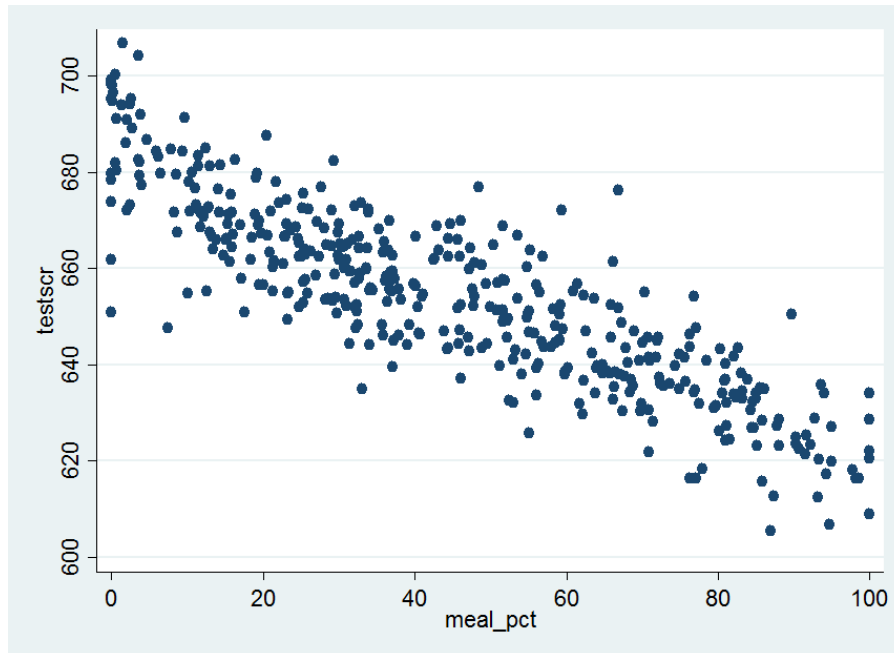


图 4.6: 测试分数与接受免费午餐的学生比例

还可以得到这些变量之间的相关系数。在 stata 命令栏输入

```
cor testscr str elpct mealpct calwpct
```

得到下列结果

由上述结果可知，接受免费午餐的学生比例（mealpct）、有资格接受家庭收入援助的学生比例（calwpct）之间的相关系数为 0.739。而测试分数与三个控制变量之间的均负相关，相关系数分别为-0.644、-0.869 和-0.627。

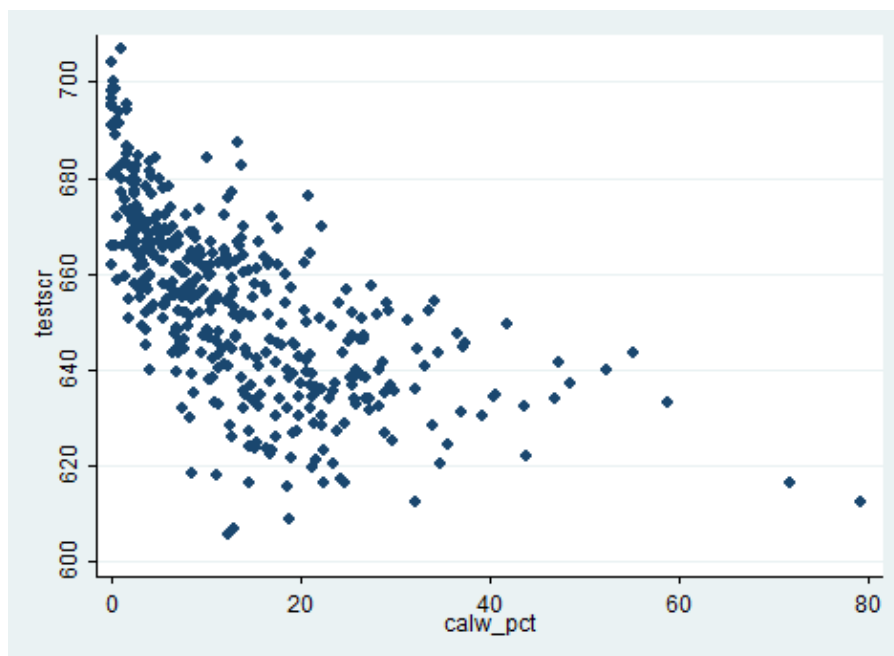


图 4.7: 测试分数与接受收入援助的学生比例

	testscr	str	el_pct	meal_pct	calw_pct
testscr	1.0000				
str	-0.2264	1.0000			
el_pct	-0.6441	0.1876	1.0000		
meal_pct	-0.8688	0.1352	0.6531	1.0000	
calw_pct	-0.6269	0.0183	0.3196	0.7394	1.0000

图 4.8: 相关系数

小贴士

我们使用的三个控制变量都是学生比例，单位是百分号，那么，这些变量的范围肯定在 0 到 100。同时，我们也可以用分数来表示这些变量，而不是百分数。那么，我们如何选择变量数值的量级或单位呢？

这个问题的答案是选择一个变量的合适量级使得回归结果更容易读取和理解。例如，测试分数对学生-老师比和非英语母语学生比例的回归结果显示，非英语母语学生比例的回归系数为-0.650。如果非英语学生比例的量级换成 $elpct/100$ 。回归模型的 R^2 和 SER 都不会变化，但是它的系数变成了-65.0。那么，在 $elpct$ 的设定中， str 保持不变， $elpct$ 的系数表示分数变化的百分点（分），而在 $elpct/100$ 的设定中， str 保持不变， $elpct/100$ 的系数表示 100 百分点的变化。尽管这两种模型设定在数学形式上是等价的，但是其 OLS 系数的含义还是前一种设定比较自然。

第二步，回归结果的呈现

基准回归模型和备择回归模型设定好了，stata 会直接给出回归结果。那么，现在的问题来了，在这么多回归结果中，如何最好的呈现出回归结果呢？回忆一下，前面的内容已经以回归方程的形式呈现出了回归结果。但是这种形式在论文中很少见，论文中多数是以表格的形式呈现出回归结果。

表 4.1: 多元回归结果

自变量	(1)	(2)	(3)	(4)	(5)
学生-老师比 X_1	-2.280*** (0.519)	-1.101** (0.433)	-0.998*** (0.270)	-1.308*** (0.339)	-1.014*** (0.269)
非英语母语学生比例 X_2		-0.650*** (0.031)	-0.122*** (0.033)	-0.488*** (0.030)	-0.130*** (0.036)
免费午餐学生比例 X_3			-0.547*** (0.024)		-0.529*** (0.038)
收入援助学生比例 X_4				-0.790*** (0.068)	-0.048 (0.059)
常数项	698.933*** (10.364)	686.032*** (8.728)	700.15*** (5.568)	697.999*** (6.920)	700.392*** (5.537)
SER	18.581	14.464	9.080	11.654	9.084
\hat{R}^2	0.049	0.424	0.773	0.626	0.773
F	19.26***	223.82***	453.48***	170.37***	361.68***
obs	420	420	420	420	420

注：括号中为异方差稳健标准误；***、**、* 分布表示 1%、5%、10% 的显著性水平。

表 1 呈现了五个回归模型的结果。第一列是自变量和统计量。从第二列至第六列，每一列代表一个多元回归模型。尽管表中没有呈现出 t 统计量，但是在实践中，括号里有时候是 t 统计量。因为在原假设下，回归系数、 t 统计量、SER 和 p 值可以相互计算得到。例如，(1) 列中， $t = \frac{-2.28-0}{0.519} = -4.39$ ，而 $4.39 > 2.58$ ，因此，该系数在 1% 的显著性水平下显著。

(2) - (5) 列包含控制变量。(2) 列结果在前文已经讲过。(3) 列是基准模型结果，一个核心解释变量——学生-老师比，两个控制变量——非英语母语学生比例和免费午餐学生比例。(4) 和 (5) 列则是备择模型结果，主要是对比学生经济特征变化的效应。(4) 列是免费午餐学生比例作为控制变量，(5) 列是在 (4) 基础上再加入收入援助项目学生比例作为控制变量。

第三步，经验结果分析

1、控制了学生特征之后，学生-老师比对测试分数的效应几乎减半。而且这个效应对模型中的控制变量并不敏感（或者十分稳健）。在所有的情形下，学生-老师比的系数均在 5% 的水平下显著。在四个带有控制变量的模型中，即从 (2) 到 (5)，保持学生特征不变时，每个老师减少一个学生的话，平均测试分数会上升 1 分左右。

2、学生特征的变量是测试分数的有效预测量。从 (1) 列看， $\hat{R}^2 = 0.049$ 说明，学生-老师比仅仅只能解释一小部分测试分数的变化。然而，当学生特征的变量加入回归模型后， \hat{R}^2 大幅度上升。学生特征变量的系数符号与散点图中呈现的模式一致：非英语母语学生比例越高、家境贫寒学生比例越高的班级，平均测试分数越低。

3、单个的控制变量并不总是显著：在（5）中，收入援助的学生比例对测试分数没有效应的原假设在 5% 的显著性水平下不能被拒绝。由于把这个控制变量增加进基准模型（3）中，对核心解释变量（学生-老师比）的系数及其标准误都没有太大影响，且这个控制变量的系数不显著，因此，收入援助的学生比例对于我们的分析目的来说是多余的控制变量。

注：上述回归结果的 **stata** 命令为

```
***** Table 1;
*****;
* Column (1);
reg testscr str, r;
dis "Adjusted Rsquared = " _result(8);
* Column (2);
reg testscr str el_pct, r;
dis "Adjusted Rsquared = " _result(8);
* Column (3);
reg testscr str el_pct meal_pct, r;
dis "Adjusted Rsquared = " _result(8);
* Column (4);
reg testscr str el_pct calw_pct, r;
dis "Adjusted Rsquared = " _result(8);
* Column (5);
reg testscr str el_pct meal_pct calw_pct, r;
dis "Adjusted Rsquared = " _result(8);
```

第5章 识别的评价框架

回归分析已经成为计量经济学领域最重要的经验研究方法。那么，自然出现的问题就是：我们如何评价基于回归分析的经验分析呢？或者如何判断采用回归分析方法的经验研究是否可信呢？

从第三讲和第四讲可以看出，一元回归会遗漏重要的回归量，从而导致我们关心的效应产生遗漏变量偏误，引入数据可用的遗漏变量，采用多元回归可以消除遗漏变量偏误。那么，如何评价我们所做的回归分析呢？

在本讲中，我会向大家介绍评价一个有用的经验研究的标准和步骤。如果发现回归分析的问题，应如何改进。

5.1 内部有效性和外部有效性框架

评价一个回归分析的有效性，要基于内部有效性和外部有效性的概念。如果一个研究关于因果效应的统计推断对于所研究的总体和环境是有效的，那么，这个研究就具有**内部有效性**；如果这个研究德尔结论能推广至其他总体和环境，那么，它也具有**外部有效性**。

其中，**研究的总体**是研究所刻画的样本来源总体；**感兴趣的总体**是从研究中得到的因果推断推广应用的总体。例如，高中各种实验班对 211、985 高校升学率的效应，是否可以推广至高校各类实验班的设立呢。

而“环境”则是指制度、法律、社会、文化和经济环境等。例如，前面回归所得的美帝小班教学的效应，是否对中国有用呢？因为美帝和中国差异还是非常大的。

5.1.1 内部有效性

内部有效性由两部分构成：

- 1、因果效应估计量是无偏和一致的。
- 2、假设检验有合意的显著性水平，并且置信区间有合意的置信水平。

在回归分析中，因果效应是利用估计的回归函数来估算的，假设检验是利用估计的回归系数和标准误来执行的。因此，内部有效性要求 OLS 估计量是无偏和一致的，标准误的计算要使得置信区间有合意的置信水平。但是实践中，有许多原因使得这个要求不能得到满足。我们第四讲中提到的遗漏变量偏误就是其中之一，因为它使得回归量与误差项相关，破坏了 OLS 第一假设。如果遗漏变量数据可用，我们可以纳入回归模型中来消除遗漏变量偏误。其它原因，我们在下面将会详细讲解。

5.1.2 外部有效性

从外部有效性的定义可知，破坏外部有效性的潜在因素来源于所研究总体和环境与感兴趣的总体和环境存在差异。

1、总体差异

所研究的总体与感兴趣的总体之间存在差异会威胁到一个研究的外部有效性。例如，医药实验总是从小白鼠开始，但是一种新药在小白鼠身上起作用，其对人类也起作用吗？毕竟小白鼠总体与人类总体存在非常大的差异。

一般来说，真实因果效应在所研究的总体和感兴趣的总体中不可能完全相同。

2、环境差异

即使所研究的总体和感兴趣的总体相同，只要环境存在差异，一个研究的结果也不可能推广到更一般化情形。

3、加利福利亚的小班教学

从三、四讲来看，加利福利亚州的小班教学确实会提高学生的平均测试成绩，也就是说，学生-老师比越小，平均测试成绩越高。但是，这一结果是在加利福利亚的初等教育学生这一总体得出的。如果想推广至中等教育，

甚至高等教育，小班教学的回归分析可能就存在外部性有效性问题，因为初等教育总体和中等教育、高等教育总体存在差异。

同样的道理，如果想把这一研究结果应用到中国来，也可能存在外部有效性问题。

综上所述，所研究的总体和环境与感兴趣的总体和环境越接近，外部有效性就越强。

4、如何判断外部有效性？

要判断外部有效性，可能就要我们非常了解所研究总体和环境与感兴趣的总体和环境。两者之间的重要差异都可引起对研究外部有效性的怀疑。

从实践的角度来说，如果我们有不同但相关总体的多个研究，那么，外部有效性就能通过对比这些研究结果来判断。一般来说，多个研究得到相似的结果可以支持外部有效性，反之亦然。

5、如何设计一个外部有效的研究

因为外部有效性来源于总体和环境之间的不可比较或者比较起来较为困难。因此，最小化外部有效性的威胁要在研究初期解决，例如，研究设计和数据收集阶段。有关研究设计的讨论可以参见 Shadish, Cook, Campbell (2002)。

5.2 内部有效性的威胁

因为回归分析的内部有效性包含两个方面的内容：OLS 估计量无偏和一致；标准误得到的置信区间有合意的置信水平。

目前，回归分析中引起估计量有偏的原因主要有五个方面：遗漏变量偏误、回归函数误设、变量的测量误差、样本选择偏误、双向因果。出现这五个方面的问题都是因为总体回归方程中回归量与误差项相关，打破了 OLS 第一假设。

5.2.1 遗漏变量偏误

回忆一下第四讲中提到的遗漏变量偏误。遗漏变量既要决定 Y ，又要与 X 相关。即使在大样本下，遗漏变量偏误也会存在，因此，OLS 估计量不具有一致性。如何最好地最小化遗漏变量偏误取决于控制的潜在遗漏变量是否可用。

1、当变量可观测或有恰当控制的变量时，遗漏变量偏误的解决办法

如果我们有遗漏变量的数据，那么，把它们包含进多元回归中即可。

但是在回归模型中，增加一个变量既有好处又有坏处：一方面，遗漏变量会导致遗漏变量偏误，增加遗漏变量会消除潜在的遗漏变量偏误；另一方面，包含一个不重要的变量（例如，它的总体回归系数为 0）会降低另一些回归因子系数估计量的精确性。换句话说，是否加入一个变量，就等同于在系数的偏误和方差之间做出取舍。实践中，通常用以下四步来判断是否要加入一个变量或一些变量：

第一步，识别出回归模型中感兴趣的系数和关键系数。例如，前面的例子中，关键系数是学生-老师比的系数。

第二步，自问“我们的回归中，重要的遗漏变量偏误来源于什么？”这就需要我们应用经济理论和专业知识。这一步应该在回归之前就完成。这一步的回归应该当做**基准回归模型**，也就是经验回归分析的开始。

第三步，扩展基准回归。如果加入的控制变量系数是统计显著的，或者如果加入遗漏变量后，感兴趣的变量系数估计量发生明显变化，那么，这些变量应该保留，反之亦然。

第四步，把我们的回归结果详细的列示在图表中。这一步是为了方便别人查看，并发现一些可疑之处，进而提出一些有益的意见和建议。

2、当适当的控制变量不可用时，遗漏变量偏误的解决办法

当遗漏变量的数据不可用时，我们就不能将其增加到回归模型中作为一个回归量。但是，实践中还有以下几种方式来解决遗漏变量偏误。每种方法都适用不同的数据类型。

第一种方法：面板数据模型，面板数据可以控制不可观测的遗漏变量，但是要求这些遗漏变量不随时间变化。

第二种方法：工具变量法，这种方法依赖一个被称为工具变量的新变量。

第三种方法：随机控制实验——DID 或者 RD 等。

在实践中，以上三种方法通常结合使用。但目前的主流面板数据模型仍只考虑了不随时间变化的不可观测遗漏变量。那么，如果这些遗漏变量随时间可变，即属于不可观测的时变特征，解决这类问题的方法如下：

第四种方法：时间差分 and 空间差分法（详见 Duranton et al., 2009; Belotti et al., 2016）、时间趋势多项式法（包括时间趋势二次型）（参见 Wolfers, 2006）、交互固定效应（参见 Bai, 2009; Kim and Oka, 2014）。

后面几讲中，我们将详细讲解前三种方法。

5.2.2 函数形式误设

我们前面讲的回归函数是线性的，但是如果真实总体回归函数是非线性的，那么，这种函数形式误设也会使得 OLS 估计量有偏。为什么这种偏误也属于遗漏变量偏误的一种类型呢？试想一下，如果总体回归函数是抛物线，那么，线性回归模型就遗漏二次项变量，这个二次项变量也当做一个回归量，就与上面的遗漏变量偏误没有本质区别。

函数形式误设通常利用图形和估计的回归函数来甄别。它能利用不同的函数形式进行纠正。例如多项式回归函数、对数形式、交互项、线性概率模型（probit 或 tobit 等）等等。

5.2.3 测量误差

假如我们在测量或者录入数据时，不小心把自变量的数据搞错了，者也会导致 OLS 估计产生偏误。这种情形称为**变量误差偏误**。

变量误差可能有许多原因：调查时，受访者提供错误答案；数据录入错误等等。从数学形式来讲， ΔX_i 表示测量误差，那么，根据测量误差修正总体回归方程 $Y_i = \beta_0 + \beta_1 X_i + u_i$

$$Y_i = \beta_0 + \beta_1 (X_i + \Delta X_i) + u_i = \beta_0 + \beta_1 X_i + \beta_1 \Delta X_i + u_i \quad (5.1)$$

从式 (1) 中可以看出，存在变量误差时，就相当于多元回归中遗漏了误差这个自变量，从而引起遗漏变量偏误。

如果 Y 存在测量误差，那么，会使得回归方差增大，但是不会引起 β 的估计偏误。例如，Y 的误差会进入随机误差项，即新的误差项 $v_i = u_i + w_i$ 。如果 w_i 是随机的，那么， $E(w_i|X_i) = 0$ ，这也意味着 $E(v_i|X_i) = 0$ ，因此， β 的估计量是无偏的，但是 $var(v_i) > var(u_i)$ ，使得 β 的估计量的方差增大。

解决方法

方法一：工具变量法，工具变量与 X 相关，但不与测量误差相关。

方法二：提出一个测量误差的数学模型，如果可能用包含这个数学公式所表示的指标来调整估计量。最简单的方法就是利用一个更精确的 X 来重新回归。也就是通常，用多个表示 X 含义的变量来回归，并将结果进行比较。

5.2.4 缺失数据和样本选择

数据缺失是经济数据集的常见现象。数据缺失是否对内部有效性造成威胁取决于数据为什么缺失。考虑三种情形：

情形一：数据缺失完全是随机的。在这种情形下，随机缺失的原因与 X 或 Y 不相关，这只会导致样本规模变小（删除缺失样本），而不会导致估计偏误。

情形二：缺失数据依赖于 X。在这种情形下，也只会导致样本规模变小，不会引起偏误。

情形三：由于样本选择过程造成的数据缺失，与 Y 相关。这种选择过程会引起误差项与回归量相关。这种 OLS 估计量偏误称为**样本选择偏误**。

解决方法：上面提到的那些方法都解决不了样本选择偏误。在实践中，常用的方法是改变样本，对多个样本进行回归，并比较回归结果。

5.2.5 双向因果

迄今为止，我们都假设因果关系只从 X 到 Y ，即 X 导致 Y 。但是，如果因果关系也从 Y 到 X 呢？也就是， X 导 Y ， Y 也可以导致 X ，这就是**双向因果关系**。双向因果也会导致误差项与回归量相关。用数学形式描述反向因果联系：

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (5.2)$$

$$X_i = \gamma_0 + \gamma_1 Y_i + v_i \quad (5.3)$$

公式 (2) 表示 X 变化引起的 Y 的效应，公式 (3) 则是反向因果。想想一下 u_i 为正，那么， Y_i 会增大。而反向因果关系表明，更高的 Y 会影响 X 的值。如果 γ_1 为正，那个 Y 越大， X 也越大，那么， X_i 与 u_i 就正相关。
$$\text{cov}(X_i, u_i) = \text{cov}(\gamma_0 + \gamma_1 Y_i + v_i, u_i) = \gamma_1 \text{cov}(Y_i, u_i) + \text{cov}(v_i, u_i) = \frac{\gamma_1 \sigma_u^2}{1 - \gamma_1 \beta_1}$$

解决方法

有两种方法可以解决双向因果引起的偏误：**工具变量回归**；**随机控制实验**。我们将在后面详细讲述。

5.2.6 OLS 标准误的不一致性

不一致的标准误会对内部有效性造成很大的伤害。即使 OLS 估计量是一致的，样本规模也很大，但是非一致的标准误会导致假设检验的显著性水平不合意，并使得置信区间在合意置信水平下没有包含真值。

引起非一致标准误，主要有两个原因：

1、异方差

有些统计软件只呈现同方差标准误。但是，如果存在异方差，标准误就不能作为可信的假设检验和置信区间的基础。

解决异方差的方法是利用异方差稳健标准误，并利用异方差稳健方差估计量来构建 F 统计量。Stata 软件提供了稳健标准误和稳健的 F 统计量。

2、序列相关

在某些情形下，总体回归误差在观测值之间是相关的。如果数据是完全随机抽样得到的，那么，序列相关就不会发生。但在实践中，经济样本往往是部分随机的。

序列相关并不会使得 OLS 估计量产生偏误或者非一致性，但是它会打破 OLS 第二个假设。这就会使得估计的 OLS 标准误产生的置信区间不在合意的置信水平下。

解决方法就是利用另一种标准误的计算公式。这些将在面板数据模型中详细阐述。

5.3 宏观中的识别

5.4 小班教学及其 Stata 操作

我们将上述外部有效性和内部有效性框架应用于小班教学效应研究。

5.4.1 外部有效性

前面两讲中使用的美帝加利福利亚 420 个学区的缩减班级规模对测试分数的影响是否能一般化到美帝其它州呢——即是说，这项研究是否具有外部有效性？那么，我们就要看看加利福利亚的学区总体及其环境是否能推广至其他州。

1.2 节给出了一种判断外部有效性的方法：比较两个或多个相同主题研究的结果。因此，我们再利用另一个州——马萨诸塞州——的初等教育标准测试得分数据的回归结果来与加利福尼亚州的回归结果进行对比。

数据比较。首先，比较变量的定义，即两项研究主要的变量定义比较接近。其次，比较两个样本的主要统计量，如表 1 所示。

表 5.1: 样本比较

	加利福尼亚		马萨诸塞	
	均值	标准差	均值	标准差
测试分数	654.16	19.05	709.83	15.13
学生-老师比	19.64	1.89	17.34	2.28
非英语母语学生比 (%)	15.77	18.29	1.12	2.90
免费午餐学生比 (%)	44.71	27.12	15.32	15.06
地区平均收入 (千元)	15.32	7.23	18.75	5.81
样本量	420		220	
年份	1999		1998	

从表 1 看出，加州的平均测试分数更低，但是由于两个州测试不同，这种直接比较没有多大意义。加州的平均学生-老师比也更大，也就是说，加州的平均班级规模更大。加州的平均收入更低，但是标准差更大。而加州的非英语母语学生比例以及免费午餐学生比例均更高。

5.5 总结

1、内部有效性与外部有效性框架

内部有效性根据因果效应的统计推断来评价；外部有效性根据是否具有—般性来评价。因此，外部有效性要在研究设计或数据收集之前完成，而经验研究中更多关注于内部有效性评价。

2、内部有效性评价

内部有效性的威胁及其解决措施：

(1) 遗漏变量偏误：(a) 对于可观测的变量，加入可能减低偏误，但是估计量的方差会增加，四步走：第一步，识别你的核心系数或感兴趣的系数；第二步，根据经济理论和专业知识、经验，自问重要的遗漏偏误来源可能有哪些；第三步，检验第二步中那些仍有疑问的控制变量系数是否为 0；第四步，把所有回归结果都呈现出来，让同行给你意见或建议。(b) 对于不可观测的遗漏变量，三种解决办法：第一，面板数据；第二，工具变量；第三，随机控制实验 (DID、RD 等)。

(2) 回归函数误设

有专门的非线性回归解决方法

(3) 测量误差

最好的办法就是得到更精确的变量值。但这通常是不可能的，那么，两种方法备选：第一，工具变量，与变量相关，但与误差无关；第二，构建测量误差的数学公式来调整估计值，也就是说换一个变量值。

(4) 样本选择偏误

(5) 反向因果或同时因果

两种方式解决：第一，工具变量；第二，随机控制实验。

(6) 估计量不一致性来源

异方差：用异方差稳健标准误和异方差稳健方差估计量构建的 F 统计量来解决；

系列相关：使用稳健标准误来解决。

第 6 章 面板数据模型

正如第五讲所述，一项经验研究可能存在的主要问题：遗漏变量偏误、变量测量误差、反向因果、模型设定错误、样本选择偏误、异方差和序列相关。第四讲呈现的多元回归方法可以消除某些遗漏变量偏误。但多元回归只能应对遗漏变量的数据可用的情形。如果遗漏变量的数据不可用，那么，多元回归中就不能包含它们，此时，OLS 估计系数就可能存在遗漏变量偏误。在特定的数据类型（面板数据）下，本讲呈现了一种方法来降低不可观测的遗漏变量偏误。

下面，我们以“醉驾”问题为例，来说明面板数据模型的原理与实际操作。使用的数据是美国 48 个州，1982-1988 年的交通事故（traffic fatalities）、酒精税（alcohol taxes）、醉驾法律（drunk driving laws）以及其他相关变量。

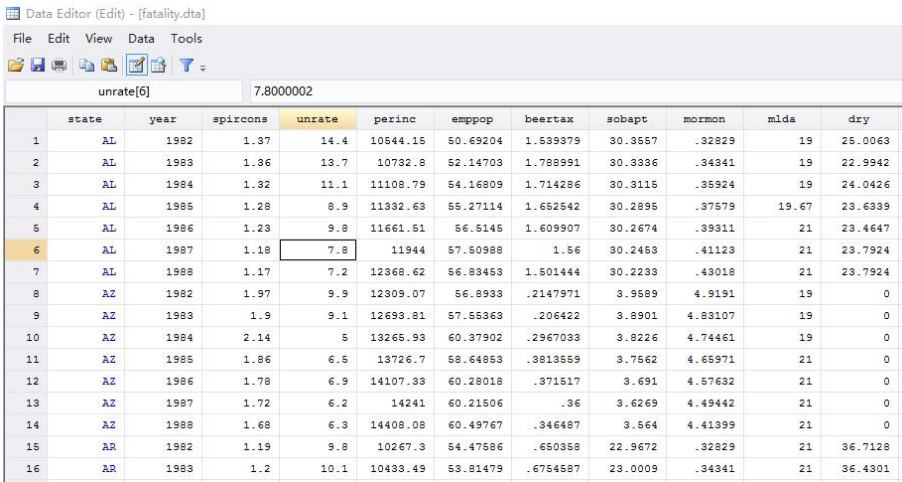
6.1 面板数据

我们在第一讲中提到过面板数据的类型。在**面板数据**中，有 n 个观测单元/个体，每个观测个体有两期及多期观测值。

在计量经济学中通常用 X_{it} 来表示面板数据，其中下标 i 表示观测个体，下标 t 表示观测时间。本讲使用的美国交通事故数据就是面板数据，它包含美国的 48 个州（ $n=48$ ），每个州有 7 年的观测值（ $T=7$ ），共 $48 \times 7 = 336$ 个样本。

与面板数据相关的两个重要概念：

（1）**平衡面板**，即样本中每一个个体，每一时期均有观测值，如图 1 所示；



Data Editor (Edit) - [fatality.dta]											
File Edit View Data Tools											
unrate[6] 7.8000002											
	state	year	spircons	unrate	perinc	emppop	beertax	sobapt	mormon	mlda	dry
1	AL	1982	1.37	14.4	10544.15	50.69204	1.539379	30.3557	.32829	19	25.0063
2	AL	1983	1.36	13.7	10732.8	52.14703	1.788991	30.3336	.34341	19	22.9942
3	AL	1984	1.32	11.1	11108.79	54.16809	1.714286	30.3115	.35924	19	24.0426
4	AL	1985	1.28	8.9	11332.63	55.27114	1.652542	30.2895	.37579	19.67	23.6339
5	AL	1986	1.23	9.8	11661.51	56.5145	1.609907	30.2674	.39311	21	23.4647
6	AL	1987	1.18	7.8	11944	57.50988	1.56	30.2453	.41123	21	23.7924
7	AL	1988	1.17	7.2	12368.62	56.83453	1.501444	30.2233	.43018	21	23.7924
8	AZ	1982	1.97	9.9	12309.07	56.8933	.2147971	3.9589	4.9191	19	0
9	AZ	1983	1.9	9.1	12693.81	57.55363	.206422	3.8901	4.83107	19	0
10	AZ	1984	2.14	5	13265.93	60.37902	.2967033	3.8226	4.74461	19	0
11	AZ	1985	1.86	6.5	13726.7	58.64853	.3813559	3.7562	4.65971	21	0
12	AZ	1986	1.78	6.9	14107.33	60.28018	.371517	3.691	4.57632	21	0
13	AZ	1987	1.72	6.2	14241	60.21506	.36	3.6269	4.49442	21	0
14	AZ	1988	1.68	6.3	14408.08	60.49767	.346487	3.564	4.41399	21	0
15	AR	1982	1.19	9.8	10267.3	54.47586	.650358	22.9672	.32829	21	36.7128
16	AR	1983	1.2	10.1	10433.49	53.81479	.6754587	23.0009	.34341	21	36.4301

图 6.1: 平衡面板

（2）**非平衡面板**，即面板中至少有一个时期、一个个体的观测值是缺失的，如图 2 所示。

在美国，每年将近 40000 高速交通事故，其中近 1/4 与醉驾有关。Levitt and Porter（2001）估计在凌晨 1 点-3 点开车的司机中，且达到法定饮酒年龄的，有 25% 的司机饮酒后开车上路，他们造成的交通事故至少是没有饮酒的司机的 13 倍。（注：中国这一情况也非常严重。2009 年全国查处酒后驾驶案件 31.3 万起，其中醉酒驾驶 4.2 万起。2010 年，全国查处醉驾达 8.7 万起。2009 年 1-8 月，共发生 3206 起，造成 1302 人死亡，其中，酒后驾车肇事 2162 起，造成 893 人死亡；醉酒驾车肇事 1044 起，造成 409 人死亡。）

下面，我们来看看美国政府实施的抑制醉驾行为的政策到底有多大效果。我们所使用的数据中包含：每年每个州的交通事故数和防止酒驾的政策（包括酒驾法律、酒精税）。交通死亡指标使用**死亡率（vfrall）**——**每万人年交通死亡人数**。政策指标是酒精税，使用**啤酒税（BeerTax）**，且经过 1988 年通胀处理的实际啤酒税。



图 6.2 展示了 Stata 数据编辑器中的面板数据集。表格显示了 1982 年至 1988 年间不同州的数据。第一行（AL, 1982）的 'spircons' 列显示为缺失值，并用红色框和箭头标注为“缺失值”。

	state	year	spircons	lnrate	perinc	emppop	beertax	sobapt	mormon	mlda
1	AL	1982		14.4	10544.15	50.69204	1.539379	30.3557	.32829	19
2	AL	1983		13.7	10732.8	52.14703	1.788991	30.3336	.34341	19
3	AL	1984	1.32	11.1	11108.79	54.16809	1.714286	30.3115	.35924	19
4	AL	1985	1.28	8.9	11332.63	55.27114	1.652542	30.2895	.37579	19.67
5	AL	1986	1.23	9.8	11661.51	56.5145	1.609907	30.2674	.39311	21
6	AL	1987	1.18	7.8	11944	57.50988	1.56	30.2453	.41123	21
7	AL	1988	1.17	7.2	12368.62	56.83453	1.501444	30.2233	.43018	21
8	AZ	1982	1.97	9.9	12309.07	56.8933	.2147971	3.9589	4.9191	19
9	AZ	1983	1.9	9.1	12693.81	57.55363	.206422	3.8901	4.83107	19
10	AZ	1984	2.14	5	13265.93	60.37902	.2967033	3.8226	4.74461	19
11	AZ	1985	1.86	6.5	13726.7	58.64853	.3813559	3.7562	4.65971	21
12	AZ	1986	1.78	6.9	14107.33	60.28018	.371517	3.691	4.57632	21
13	AZ	1987	1.72	6.2	14241	60.21506	.36	3.6269	4.49442	21
14	AZ	1988	1.68	6.3	14408.08	60.49767	.346487	3.564	4.41399	21
15	AR	1982	1.19	9.8	10267.3	54.47586	.650358	22.9672	.32829	21
16	AR	1983	1.2	10.1	10433.49	53.81479	.6754587	23.0009	.34341	21
17	AR	1984	1.22	8.9	10916.48	54.67128	.5989011	23.0346	.35924	21
18	AR	1985	1.12	8.7	11149.36	54.97712	.5773305	23.0684	.37579	21
19	AR	1986	.92	8.7	11399.38	55.56186	.5624355	23.1022	.39311	21
20	AR	1987	1.01	8.1	11537	56.33089	.545	23.1361	.41123	21
21	AR	1988	.99	7.7	11760.35	57.36695	.5245429	23.17	.43018	21

图 6.2: 非平衡面板

图 3 呈现了 1982 年交通死亡率与酒精税之间的散点图和拟合线。散点图上的每一个点都代表着 1982 年给定税率下的死亡率。从图中，可以看出死亡率与税率正相关。回归结果如下

$$vfrall = 2.01 + 0.15BeerTax1982 \quad (6.1)$$

回归结果显示，酒精税的系数为 0.15，t 统计量为 1.12。也就是说，实际酒精税对死亡率的效应为正，但是在 10% 的水平下不显著。

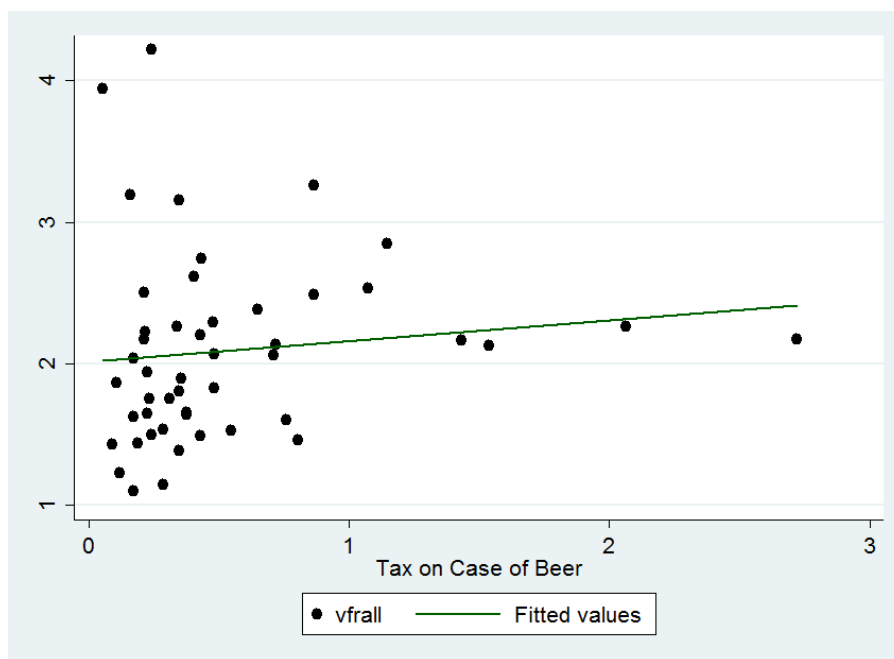


图 6.3: 1982 年交通死亡率与酒精税

从散点图和回归结果来看，这个结果似乎有点奇怪。那么，我们再来看看 1988 年的结果。其散点图和趋势线如图 4 所示。回归方程如下

$$vfrall = 1.86 + 0.44BeerTax1988 \quad (6.2)$$

1988 年回归结果显示，酒精税的系数为 0.44，且 t 统计量为 3.43，也就是说酒精税对死亡率的效应在 1% 的

水平下显著为正。

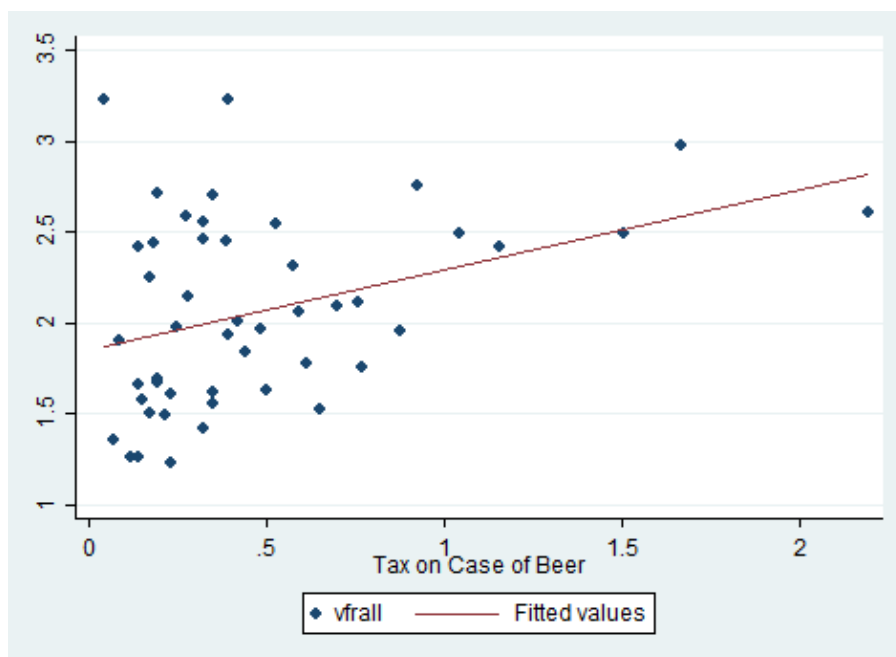


图 6.4: 1988 年交通死亡率与酒精税

虽然，1982 年和 1988 年的估计结果有所差异，但是酒精税的系数均为正。从字面来理解，实际酒精税越高，交通死亡率越高。这似乎与政策设计初衷相悖，也与常识相悖。那么，我们能得出结论：提高酒精税会增加死亡率？

答案显然是否定的！因为上述两个回归结果可能存在很大的遗漏变量偏误。还有许多影响交通死亡率，但又未包含在回归中的因素：汽车质量；公路质量；在农村开车还是在城市；道路上车流密度；对待饮酒和开车的态度等等。这些因素都可能与酒精税有关。如果它们与酒精税相关，就会导致回归结果存在遗漏变量偏误。第四讲为我们提供了一种解决有观测数据的遗漏变量问题的方法——加入其它解释变量（控制变量）。但是，上述有些因素是不可观测的——例如，对饮酒和开车的态度，那怎么办？

如果这些不可观测的遗漏变量不随时间变化，那么，我们可以使用另一种方法来降低遗漏变量偏误。这种方法就是固定效应模型。

6.2 固定效应回归

6.2.1 两期“比较”

在前面的回归中，我们分别对 1982 年和 1988 年的截面数据进行回归。那么，我们现在将两年数据结合在一起，形成一个两期的面板数据（ $n=48$ ， $T=2$ ）。这样我们就可以将 1988 年的被解释变量（死亡率）与 1982 年进行比较。也就是说，在不可观测因素恒定（在时间维度不变，在个体维度可变）时，我们关注于被解释变量“前”“后”变化。

用 Z_i 表示影响第 i 个州的交通死亡率的因素，它不随时间变化，因此没有时间下标。那么，这个包含不可观测因素的两期面板数据总体回归线为

$$vfrall_{it} = \beta_0 + \beta_1 BeerTax_{it} + \beta_2 Z_i + u_{it} \quad (6.3)$$

其中， u_{it} 表示随机误差项， $n=1, \dots, 48$ ； $T=1982, 1988$ 。

因为 Z_i 不随时间变化，也就是说，它在 1982 年和 1988 年是一样的。那么，在上述回归方程（3）中，我们可以通过分析 1982 年-1988 年死亡率的变化来消除 Z_i 的影响。从数学形式来看，1982 年和 1988 年的回归方程

分别为：

$$vfrall_{i1982} = \beta_0 + \beta_1 BeerTax_{i1982} + \beta_2 Z_i + u_{i1982} \quad (6.4)$$

$$vfrall_{i1988} = \beta_0 + \beta_1 BeerTax_{i1988} + \beta_2 Z_i + u_{i1988} \quad (6.5)$$

然后，我们将（5）式减去（4）式来消除 Z_i 的影响：

$$vfrall_{i1988} - vfrall_{i1982} = \beta_1(BeerTax_{i1988} - BeerTax_{i1982}) + (u_{i1988} - u_{i1982}) \quad (6.6)$$

从（6）式中可以很直观的看出：不可观测因素虽然影响一个州的死亡率，但是它不会在 1982 年和 1988 年间变动，因此，它们也不会对这个期间的死亡率变化产生任何影响。

也就说，通过分析被解释变量 Y 与解释变量 X 的变化量可以消除不随时间变化的不可观测因素，从而消除这种来源的遗漏变量偏误。

图 5 呈现了 1982-1988 两年的散点图和拟合线。图中，dvfrall 是 1988 年 vfrall 与 1982 年之差，dbeertax 同理。

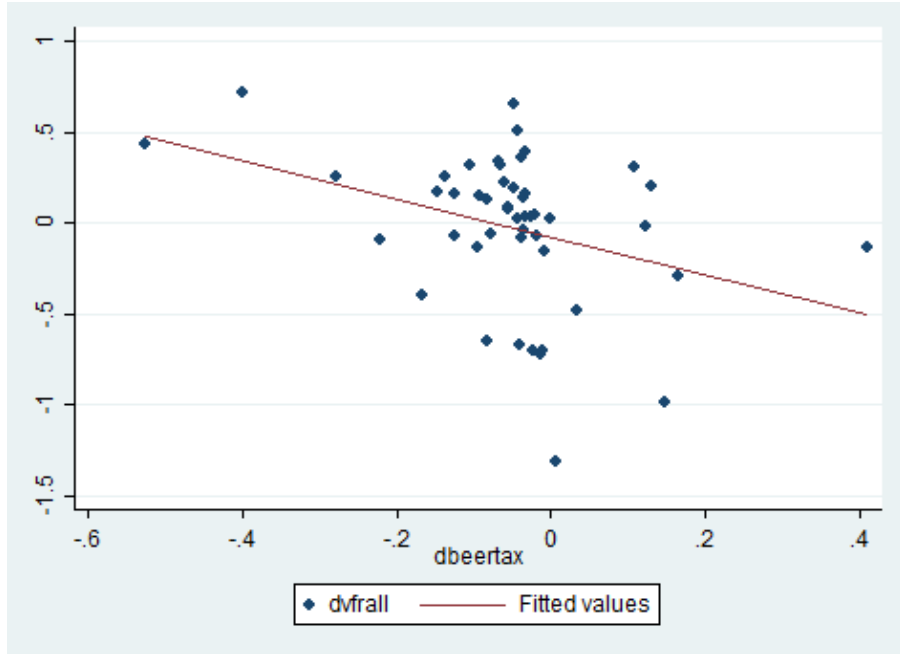


图 6.5: 1982、1988 年交通死亡率与酒精税

回归方程为

$$vfrall_{i1988} - vfrall_{i1982} = -0.072 - 1.04(BeerTax_{i1988} - BeerTax_{i1982}) \quad (6.7)$$

回归系数为-1.04，且 t 统计量为-2.93，在 1% 的水平显著。与第一节的截面数据结果相比，上述回归结果显示实际酒精税对死亡率有显著地负向影响，也符合经济理论。

上述回归只对包含两期的面板数据可用。而我们所使用的数据集包含 7 年长度。

6.2.2 固定效应回归

为了消除不随时间变化的不可观测因素带来的影响，可使用固定效应回归。在这种情形下，固定效应回归有 n 个截距，每个州都有一个。这些截距可以用一个指示变量（二值变量）来表示。而所有不随时间变化的遗漏变量影响都会被这些指示变量吸收。

我们用 Y 和 X 来分别表示死亡率和酒精税。那么，我们可以将（3）式写成

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it} \quad (6.8)$$

我们的目的就是估计出 β_1 ，即酒精税对交通死亡率的影响，且控制不可观测的地区特征 Z_i 。因为 Z_i 只在各州之间变化，而不随时间变化，因此，我们可以令 $\alpha_i = \beta_0 + \beta_2 Z_i$ 。因此，(8) 式可以写成：

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} \quad (6.9)$$

(9) 式就是**固定效应模型**。需要注意的是，上述总体回归线的斜率对于每个州都是相同的，但是每个州又对应着不同的截距。因为 α_i 考虑的是个体 i 的效应，因此，它也被称为**个体固定效应**。

我们回忆一下二值变量（虚拟变量），上述固定效应回归也可以用二值虚拟变量来表示，例如，如果 $i=1$ ，令 $D_{1i} = 1$ ，否则为 0；如果 $i=2$ ，令 $D_{2i} = 1$ ，否则为 0；等等。由于有 48 个州，因此，设置 48 个虚拟变量。但是为了避免**虚拟变量陷阱**（即完全多重共线性），我们要删除一个虚拟变量（任意删除即可）。因此，(9) 式可变形为：

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D_{2i} + \cdots + \gamma_n D_{ni} + u_{it} \quad (6.10)$$

由 (10) 式可知， $\beta_0, \beta_1, \gamma_2, \cdots, \gamma_n$ 是待估计系数。现在，我们来对比一下 (9) 式和 (10) 式：(9) 式中，第一个地区的截距为 α_1 ，在 (10) 式中，第一个地区的截距为 β_0 ，因此， $\alpha_1 = \beta_0$ ，以此类推第二个第三个等等地区的截距为 $\alpha_i = \beta_0 + \gamma_i$ 。也就是说，(9) 和 (10) 是等价的，固定效应模型有两种表达方式。而在这两种表达式中，所有地区的斜率都是一样的，个体固定效应来源于不随时间可变的不可观测地区异质性。

大家可能意识到，上述固定效应模型并没有包含可观测的控制变量。包含其他解释变量的模型同第四讲的多元回归。下面，我们继续讲解固定效应模型的估计和推断。

I. 估计和推断

(9) 式中存在不可观测因素 α_i ，因此，不能直接使用 OLS。而 (10) 从理论上讲可以直接使用 OLS 估计，但是由于其有 $k+n$ 个参数（ k 个解释变量的系数， $n-1$ 个虚拟变量系数和一个常数项），因此，一旦个体数量非常大，在实践中是很难实施 OLS 估计（在软件中输入 $k+n$ 个变量）。幸好，现在有了 Stata，它直接可以替我们跑出回归结果。下面，我们稍微看看 stata 在估计固定效应模型时的工作原理。

个体去均值算法：stata 一般会执行两步：

第一步，每个变量减去其个体层面的均值。

例如，在 (9) 式中，被解释变量 Y 的个体层面均值为 $\bar{Y}_i = \frac{1}{T} \sum_{t=1}^T Y_{it}$ ，解释变量 X 也同理。然后，用 Y 减去均值， $Y_{it} - \bar{Y}_i$ ， X 和 u 也同上。因此，可以得到

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it} \quad (6.11)$$

第二步，估计上述去均值回归方程。

我们还是利用美国 48 个州的交通死亡率和酒精税数据来看看上述固定效应回归结果：

$$v\tilde{frall}_{it} = -0.66Beer\tilde{Tax}_{it} \quad (6.12)$$

$$vfrall_{it} = -0.66BeerTax_{it} + \alpha_i \quad (6.13)$$

上文我们提到过，上述个体固定效应模型也遗漏了一些变量，除了可观测的变量之外，还有一个重要的遗漏变量就是不可观测的不随地区变化的因素。

II. 时间固定效应

类似于个体固定效应，时间固定效应是不随个体变，而随时间变化的因素。我们用 S_t 表示，那么，只包含时间固定效应的模型为

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_3 S_t + u_{it} \quad (6.14)$$

与上述个体固定效应同理，由于 S_t 不随地区变化，只随时间变化，因此，令 $\lambda_t = \beta_0 + \beta_3 S_t$ 。那么，(14) 式可以变为

$$Y_{it} = \beta_1 X_{it} + \lambda_t + u_{it} \quad (6.15)$$

同理，每一个时期都有一个截距，截距 λ_t 就是时间 t 对被解释变量 Y 的效应，也被称为**时间固定效应**。

同样的，时间固定效应模型也有两种表达方式：一种是时间虚拟变量；另一种是（15）式。

$$vfrall_{it} = -0.02BeerTax_{it} + \lambda_t \quad (6.16)$$

加入时间固定效应之后，上述回归结果并不显著。

下面，我们来看看同时加入个体固定效应和时间固定效应的模型：

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it} \quad (6.17)$$

与个体固定效应的算法相同，时间固定效应和双向固定效应的算法也可以采用“去均值算法”。

首先， Y 和 X 分别减去其个体层面均值和时间层面均值；

然后，用两次差分之后的去均值变量进行回归，用 OLS 估计系数。

另一种方式，就是我们常用的 Y 和 X 只减去个体层面的均值，然后设立时间虚拟变量，用 `fe` 命令进行回归。用这种方式得到的回归结果为

$$Y_{it} = -0.64X_{it} + \alpha_i + \lambda_t + u_{it} \quad (6.18)$$

回归结果在 10% 的水平下显著。

注：一般来讲，面板数据都需要控制个体固定效应和时间固定效应，因为这可以消除不可观测的随时间可变不随个体变化的因素或者不随个体变化随时间变化的因素所引起的遗漏变量偏误。在使用家庭（企业）层面、省（市县）层面的数据后，还要同时控制家庭固定效应、省固定效应和时间固定效应。

III. 聚类标准误

在面板数据模型中，我们假设个体之间是独立的，但是在个体层面内部，由于存在不同时间的观测值，因此它们不一定独立。这就可能存在个体内部的自相关或者序列相关问题。这就意味着误差项 u 可能也存在自相关问题。此时，截面数据回归的异方差稳健标准误就不再有效。为了消除面板数据回归中误差项自相关问题，我们要使用异方差-自相关一致性（HAC）标准误，也就是**聚类标准误**。我们上述的回归结果都是使用的聚类标准误。

注：在面板数据回归中，一定要使用“聚类标准误”，也就是个体层面的聚类

6.2.3 例子

Dependent variable: Traffic fatality rate (deaths per 10,000).							
Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Beer tax	0.36** (0.05)	-0.66* (0.29)	-0.64+ (0.36)	-0.45 (0.30)	-0.69* (0.35)	-0.46 (0.31)	-0.93** (0.34)
Drinking age 18				0.028 (0.070)	-0.010 (0.083)		0.037 (0.102)
Drinking age 19				-0.018 (0.050)	-0.076 (0.068)		-0.065 (0.099)
Drinking age 20				0.032 (0.051)	-0.100* (0.056)		-0.113 (0.125)
Drinking age						-0.002 (0.021)	
Mandatory jail or community service?				0.038 (0.103)	0.085 (0.112)	0.039 (0.103)	0.089 (0.164)
Average vehicle miles per driver				0.008 (0.007)	0.017 (0.011)	0.009 (0.007)	0.124 (0.049)
Unemployment rate				-0.063** (0.013)		-0.063** (0.013)	-0.091** (0.021)
Real income per capita (logarithm)				1.82** (0.64)		1.79** (0.64)	1.00 (0.68)
Years	1982-88	1982-88	1982-88	1982-88	1982-88	1982-88	1982 & 1988 only
State effects?	no	yes	yes	yes	yes	yes	yes
Time effects?	no	no	yes	yes	yes	yes	yes
Clustered standard errors?	no	yes	yes	yes	yes	yes	yes
F-Statistics and p-Values Testing Exclusion of Groups of Variables							
Time effects = 0			4.22 (0.002)	10.12 (< 0.001)	3.48 (0.006)	10.28 (< 0.001)	37.49 (< 0.001)
Drinking age coefficients = 0				0.35 (0.786)	1.41 (0.253)		0.42 (0.738)
Unemployment rate, income per capita = 0				29.62 (< 0.001)		31.96 (< 0.001)	25.20 (< 0.001)
\overline{R}^2	0.091	0.889	0.891	0.926	0.893	0.926	0.899
These regressions were estimated using panel data for 48 U.S. states. Regressions (1) through (6) use data for all years 1982 to 1988, and regression (7) uses data from 1982 and 1988 only. The data set is described in Appendix 10.1. Standard errors are given in parentheses under the coefficients, and p-values are given in parentheses under the F-statistics. The individual coefficient is statistically significant at the *10%, *5%, or **1% significance level.							

图 6.6: 交通死亡率与酒精税

第7章 工具变量法

正如第五讲所述，一项经验研究可能存在的主要问题：遗漏变量偏误、变量测量误差、反向因果、模型设定错误、样本选择偏误、异方差和序列相关。第四讲和第六讲分别呈现了消除某种遗漏变量偏误的方法——多元回归和面板数据模型。多元回归应对遗漏变量数据可用情形；面板数据模型引入个体固定效应和时间固定效应来消除遗漏变量数据不可用，且截面或时间单一维度变化时的遗漏变量偏误。

一来，上述两种解决方法均相当于在回归模型中增加核心解释变量 (X_{it}) 以外的自变量 ($Z_{it}, \alpha_i, \gamma_t$)；二来，变量测量误差和反向因果所引起的问题，并不能由多元回归和面板数据模型直接解决。那么，除此之外，还有没有其他方法来解决遗漏变量偏误、变量测量误差、反向因果问题呢？

工具变量 (Instrumental variables, IV) 回归就是获得 X 与 u 相关的总体回归函数未知系数一致估计量的一种常用方法。**IV** 的核心思想：把核心解释变量 X 的变动分解成两个部分：一个部分与误差项 u 相关，另一个部分与误差项 u 不相关。如果有资料、数据、信息来分离出第二部分，那么，我们就可以只关注于误差项无关的第二部分，丢弃引起 OLS 估计偏误的第一部分。这些表征 X 变动，且与 u 不相关的数据信息可能来源于一个或多个其他变量，这些变量就称为工具变量 (**IV**)。如字面意思，这些变量被作为工具来分离出 X 中与 u 无关的部分，从而确保回归系数估计量具有一致性。

下面，详细阐述工具变量法的作用原理及其应用。

7.1 一元回归与单工具变量

如第五讲所述，引起 X 与 u 相关的原因可能是遗漏变量、变量误差、反向因果。无论什么原因，如果我们有一个有效的工具变量 I ，那么，我们就可以利用工具变量估计量来估计出 X 对 Y 的效应。总回归模型为

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (7.1)$$

如果 $\text{corr}(X_i, u_i) \neq 0$ ，OLS 估计量就是非一致的。我们就可以用工具变量 I_i 来分离出 X_i 中与 u_i 无关的部分。

依据是否与误差项 u 相关，可以把变量划分成内生变量和外生变量，前者与误差项相关，后者与误差项无关。这两个名字可以追溯至多方程模型，即“内生变量”由模型内决定，“外生变量”由模型外决定。这两个变量在后面的 DSGE 模型中还会见到。

有效的工具变量必须同时满足两个条件：

(1) 工具变量相关性条件： $\text{corr}(I_i, X_i) \neq 0$

(2) 工具变量外生性条件： $\text{corr}(I_i, u_i) = 0$

工具变量相关性表明一个工具变量的变动与 X 的变动相关。工具变量外生性表明工具变量抓住了 X 中外生变化的部分。这两个条件对于工具变量回归非常重要。

7.1.1 两阶段最小二乘 TSLS

两阶段最小二乘 (TSLS) 是用来估计 IV 估计量的方法。正如该方法的名字所述，IV 估计量是通过两个阶段计算出来的。

第一阶段，把 X 分解成两个部分：可能与误差项相关的部分和与误差项无关的第二部分。

具体来说，第一阶段用 X 对工具变量 I 回归：

$$X_i = \pi_0 + \pi_1 I_i + v_i \quad (7.2)$$

公式 (2) 就把 X 分解成： v_i 和 $\pi_0 + \pi_1 I_i$ 。由于 I_i 是外生的，因此， $\pi_0 + \pi_1 I_i$ 与 u_i 无关，而剩下的 v_i 与 u_i 相关。据此，我们可以用样本数据估计出 $\hat{\pi}_0, \hat{\pi}_1$ ，然后从 OLS 回归中得到 X 的预测值 $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 I_i$ 。

第二阶段，利用与误差项无关的第二部分估计 β_1 。也就是说， Y_i 对 \hat{X}_i 回归，利用 OLS 估计系数，进而得到 TSLS 估计量， $\hat{\beta}_0^{TSLS} \hat{\beta}_1^{TSLS}$

I. 香烟需求

下面，我们使用美国 48 个州 1985-1995 年的香烟销售相关数据。我们用这些数据来估计香烟的需求弹性。工具变量——销售税：来自于一般销售税中对烟草征收的税。香烟消费量， Q_i^{cig} ，是第 i 个州人均消费香烟包数。价格， P_i^{cig} ，含税实际平均价格。

在进行 TSLS 之前，大家首先要关注工具变量是否有效，也就是说，我们选择的工具变量是否满足前面的两个条件——相关性和外生性。后面，我们会详细给出如何通过统计工具来检验工具变量的有效性。在此之前，我们先来看看这销售税是否能作为一个有效的工具变量。

首先，工具变量相关性。销售税越高，香烟的税后价格也会越高，那么，销售税与香烟价格具有相关性；

其次，工具变量外生性。一般来说，销售税在各州之间是不同的，但是这种差异并不主要由香烟需求决定，而是由于政治考虑。因此，我们也可以认为销售税是外生的。

根据上述 TSLS 的两个阶段，用 1995 年的 48 个州的数据，我们先来看看第一阶段回归：

$$\ln(\tilde{P}_i^{cig}) = 4.62 + 0.31SalesTax_i \quad (7.3)$$

回归结果均在 1% 下显著，而且如经济理论预测的，销售税越高，税后价格就越高。回归方程的 $R^2 = 0.47$ ，也就睡说，销售税变化解释了 47% 的香烟价格变动。

在第二阶段中，用 $\ln(Q_i^{cig})$ 对 $\ln(\tilde{P}_i^{cig})$ 来进行回归。回归结果是

$$\ln(\tilde{Q}_i^{cig}) = 9.72 - 1.08\ln(\tilde{P}_i^{cig}) \quad (7.4)$$

也就是说，第一阶段的预测值 $\ln(\tilde{P}_i^{cig})$ ，被用作第二阶段的回归量。但是，软件中输出的结果是 $\ln(P_i^{cig})$ ，而不是 $\ln(\tilde{P}_i^{cig})$ 。因此，TSLS 估计为

$$\ln(\tilde{Q}_i^{cig}) = 9.72 - 1.08\ln(P_i^{cig}) \quad (7.5)$$

TSLS 的估计结果显示，香烟的需求弹性是富有弹性的：价格提高 1%，香烟需求量下降 1.08%。

从第四、五、六讲我们可以知道，上述估计结果可能存在遗漏变量偏误。

7.2 IV 回归

在一般化的 IV 回归模型中，主要包括四种类型的变量：被解释变量 Y ；内生解释变量 X ；外生解释变量 W ；工具变量 I 。一般来说，可能存在多个内生解释变量，多个外生解释变量和多个工具变量。

回归系数是恰好识别，如果工具变量与内生解释变量一样多；

回归系数是过度识别，如果工具变量比内生解释变量还多；

回归系数是识别不足，如果工具变量比内生解释变量还少；

需要注意的是：IV 回归中，至少要有与内生回归因子（内生解释变量）一样多的工具变量。

在 IV 中包含外生变量或控制变量是为了确保工具变量与误差项不相关。

【小贴士】（以下内容来源于“SociologyOfDrink”微信公众号 2018-01-05 期张友浪“控制变量是否越多越好”）控制变量一般没有因果解释，但控制变量又很重要，那么，怎么选择控制变量呢？一般来说我们只需控制能够同时影响解释变量和被解释变量的变量（confounder）。但是，我们在投稿时经常会收到审稿人的意见，说这个没有控制那个也没有控制。或者自己有意无意的在回归中包含了过多的控制变量。而控制变量过多往往会造成模型损失自由度，模型不够简洁，模型过度拟合，甚至会让我们得出错误的结论。这些问题在小样本回归中尤为严重。因此，我们在回归时，真正要考虑的是，什么样的变量不需要被控制？

根据某一备选变量在因果关系中的位置，可以分以下几种情况讨论：

(1) 既不影响解释变量，也不影响被解释变量。很多社会经济变量不仅仅因为常见，而被人们要求加入模型。但如果没有可信的理论来支撑对解释变量和被解释变量的影响的话，不应将其纳入模型；

(2) 只影响解释变量的变量。这类变量与关键解释变量没有直接影响，不影响我们对解释变量影响的估计，当然无需纳入模型中控制；

(3) 只影响被解释变量的变量。这类变量与关键解释变量不存在理论上的相关性，不会造成遗漏变量偏误，无需控制；

(4) 被解释变量影响，又影响解释变量的变量，即中介变量 (mediator)。考虑到我们的关键解释变量对被解释变量的影响往往是通过一个或多个渠道 (因果链条可无限细分)，这时就要分两种情况做决定：如果该中介变量处在我么假设的因果链条中，那就应该将其去掉，因为加入这个变量会让解释变量的影响从全部影响减弱为直接影响，而间接影响则被中介变量吸收，从而削弱了我们对解释变量整体效应的估计；如果该中介变量并不处在我们假设的因果链条中，那就应该保留，这时对自变量影响的估计就会自动排除竞争性解释的影响，有助于提高估计结果的可信度；

(5) 当然，有时候，会遇到审稿人搞不清因果关系，坚持要求加入某个变量来控制。考虑到硬怼审稿人没什么好下场，因此建议，审稿人说什么就是什么，听从他们的意见，控制审稿意见要求的变量。

7.2.1 TSLS

当回归方程中只有一个内生解释变量 X 和多个外生变量时，回归方程为

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i \quad (7.6)$$

如前所述， X 可能与误差项相关， W 则不与误差项相关。

根据前文 TSLS 的两个阶段。在**第一阶段**中，用 X 与所有的外生变量——外生解释变量 W 和工具变量 I 进行回归。

$$X_i = \pi_0 + \pi_1 I_{1i} + \cdots + \pi_m I_{mi} + \pi_{m+1} W_{1i} + \cdots + \pi_{m+r} W_{ri} + v_i \quad (7.7)$$

在 TSLS 的**第二阶段**，先用 (7) 式估计的 \tilde{X}_i 来替代 (6) 式中的 X ，然后进行 OLS 估计。由此，得到的 β_0, β_1, \cdots 就是 TSLS 的估计量。

【小贴士】

1、在进行第一阶段回归时，除了工具变量外，所有的外生变量（或控制变量）都需要包含在其中。而第二阶段回归也需要包含这些外生变量。

2、当有多个内生解释变量时，第一阶段的工具变量回归是每个内生解释变量分别与它对应的工具变量进行回归，然后再将所有的估计值放入第二阶段回归中得到 TSLS 估计量。

7.2.2 例子：香烟需求

在第一节的香烟需求例子中，我们只估计了单变量。但是这可能会存在遗漏变量偏误，例如，州的收入水平可能既影响香烟需求，也影响销售税。那么，这还会使得工具变量外生性条件不被满足。因此，下面，我们在回归中包含州的收入水平。TSLS 的估计结果为

$$\ln(\tilde{Q}_i^{cig}) = 9.43 - 1.14 \ln(P_i^{cig}) + 0.21 \ln(Inc_i) \quad (7.8)$$

上面的回归系数的标准误为 0.37，是恰好识别，也就是说单一内生解释变量对应单一工具变量。除了销售税作为工具变量之外，还可以使用州的烟草特种税。因此，这种税是第二种可能的工具变量。烟草特种税 (CigTax) 提高会增加香烟的价格，因此满足相关性条件，如果它与误差项无关，那么它也满足外生性条件。下面，我们用两个工具变量来重新进行 TSLS 估计，估计结果如下：

$$\ln(\tilde{Q}_i^{cig}) = 9.89 - 1.28 \ln(P_i^{cig}) + 0.28 \ln(Inc_i) \quad (7.9)$$

上述两个 IV 的估计系数标准误为 0.25，比较 (9) 式和 (8) 式的标准误，我们可以看出，(9) 的标准误比 (8) 下降了三分之一左右。这是因为 (9) 式利用了更多的信息，两个 IV 解释了更大的价格变动。

那么，上述 IV 估计结果可信吗？很遗憾，我们不能立马回答上述问题，因为可信度依赖于 IV 是否有效。因此，IV 有效的两个条件——相关性和外生性就必须要被检验。

7.3 如何检验 IV 的有效性

我们仍然用美国 48 个州的 1985-1995 年的香烟销售数据。我们来估计长期价格弹性，因此用十年的数据来进行回归，例如，我们用香烟销售量的对数之差 $\ln(Q_{i,1995}^{cig}) - \ln(Q_{i,1985}^{cig})$ ，价格的对数之差 $\ln(P_{i,1995}^{cig}) - \ln(P_{i,1985}^{cig})$ 和收入的对数之差 $\ln(Inc_{i,1995}^{cig}) - \ln(Inc_{i,1985}^{cig})$ 。两个工具变量是 $SalsTax_{i,1995} - SalsTax_{i,1985}$ ， $CigTax_{i,1995} - CigTax_{i,1985}$ 。

结果呈现在图 1 中，每一列都是不同的回归，都是用 TSLS 估计量，唯一的差别是工具变量不同。第一列是只包含销售税这个工具变量；第二列是只包含烟草税这个工具变量；第三列是包含两个工具变量。从图 1 中的结果可以看出，三列结果的第一行均在 5% 水平下显著为负。但这个结果可信吗？这依赖于我们使用的工具变量是否有效。

(I) 我们首先来看看，工具变量的相关性。

工具变量相关性的作用类似于“样本规模的作用”。因为工具变量与内生解释变量越相关，说明 IV 回归中包含 X 的信息越多，TSLS 估计量越准确，这就类似与样本量越大，估计结果越准确。

工具变量解释 X 变动的部分较少时，这种工具变量成为弱工具变量。在上面的例子中，如果选取香烟生产企业到州的距离作为工具变量，这可能就是一个弱工具变量。尽管这个距离越远，香烟的销售价格越高，但是香烟较轻，运输成本可能在其价格中并不是主要组成部分，因此，距离的变动很可能只能解释价格变动中的一小部分。因此，生产距离可能就是一个弱工具变量。那么，如何检验一个工具变量是否是弱工具变量？如果是弱工具变量，我们应该怎么处理？

上面已经说过，工具变量的作用类似于大样本的作用。如果存在弱工具变量问题，正态分布就不是 TSLS 估计量抽样分布的良好近似，那么，TSLS 估计量就不再可信。那么，工具变量相关性程度有多大才是一个良好的分布近似呢？这个答案很复杂，但是幸运的是，我们在实践中有一些经验规则可用：

检验弱工具变量的经验规则：当只有一个内生解释变量时，检验弱工具变量的方法就是计算 TSLS 第一阶段的 F 统计量。第一阶段 F 统计量为包含在工具变量中的信息提供了一个不错的测量指标：包含的信息越多，F 统计量越大。经验规则是；如果第一阶段 F 统计量大于 10，不存在弱工具变量；如果小于 10，可能就是弱工具变量。

我们从图 1 中可以看到，三个 TSLS 的回归结果中，一阶段 F 统计量分别为 33.7, 107.2 和 88.6，因此，我们选择的工具变量不是弱工具变量。

那如果上面的一阶段 F 统计量小于 10，也就是说存在弱工具变量，我们该怎么办呢？

如果存在弱工具变量，且有一些工具变量比另一些更弱。那么，就应该舍弃那些最弱的工具变量。当我们放弃一些弱工具变量时，TSLS 估计量的标准误可能会变大，但是请记住“原始标准误没有任何意义”！

但是，如果系数恰好识别，也就是一个内生解释变量，只有一个工具变量，且是弱工具变量时，我们就不能舍弃这个弱工具变量了。即使在过度识别时，没有足够的强工具变量来取得识别效果，舍弃弱工具变量也不会有什么帮助。这种情况下，我们可以干两件事：

(1) 去寻找其他的更强的工具变量。说起来容易，做起来难！这需要我们对所研究的问题有足够的认识，并且能重写设计和收集相关数据。

(2) 仍然使用弱工具变量，但是估计方法不用 TSLS，而用其他估计方法，例如有限信息极大似然 (LIML) 估计量。

(II) 工具变量外生性

如果有一个内生解释变量，多个工具变量，那么，我们可以计算出多个 TSLS 估计量（每个工具变量计算一个）。假设有两个工具变量，那么，我们计算的两个 TSLS 估计量不同。但是如果两个工具变量都是外生的，那么，它们会十分接近。如果我们估计的两个 TSLS 估计量差异非常大，那么，我们就要非常警觉：要么其中一个工具变量不是外生的，要么两个都不是外生的。在过度识别情形下，过度识别限制检验 (J 统计量) 就是在对多个工具变量 TSLS 估计量进行比较。

总之，在过度识别情形下，我们能计算出多个 TSLS 估计量，然后比较它们是否接近，即通过 stata 计算出 J 统计量。如果是精确识别情形，我们就不能比较，实际上，这个时候的 J 统计量为 0。

Dependent variable: Traffic fatality rate (deaths per 10,000).							
Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Beer tax	0.36** (0.05)	-0.66* (0.29)	-0.64+ (0.36)	-0.45 (0.30)	-0.69* (0.35)	-0.46 (0.31)	-0.93** (0.34)
Drinking age 18				0.028 (0.070)	-0.010 (0.083)		0.037 (0.102)
Drinking age 19				-0.018 (0.050)	-0.076 (0.068)		-0.065 (0.099)
Drinking age 20				0.032 (0.051)	-0.100+ (0.056)		-0.113 (0.125)
Drinking age						-0.002 (0.021)	
Mandatory jail or community service?				0.038 (0.103)	0.085 (0.112)	0.039 (0.103)	0.089 (0.164)
Average vehicle miles per driver				0.008 (0.007)	0.017 (0.011)	0.009 (0.007)	0.124 (0.049)
Unemployment rate				-0.063** (0.013)		-0.063** (0.013)	-0.091** (0.021)
Real income per capita (logarithm)				1.82** (0.64)		1.79** (0.64)	1.00 (0.68)
Years	1982-88	1982-88	1982-88	1982-88	1982-88	1982-88	1982 & 1988 only
State effects?	no	yes	yes	yes	yes	yes	yes
Time effects?	no	no	yes	yes	yes	yes	yes
Clustered standard errors?	no	yes	yes	yes	yes	yes	yes
F-Statistics and p-Values Testing Exclusion of Groups of Variables							
Time effects = 0			4.22 (0.002)	10.12 (< 0.001)	3.48 (0.006)	10.28 (< 0.001)	37.49 (< 0.001)
Drinking age coefficients = 0				0.35 (0.786)	1.41 (0.253)		0.42 (0.738)
Unemployment rate, income per capita = 0				29.62 (< 0.001)		31.96 (< 0.001)	25.20 (< 0.001)
\bar{R}^2	0.091	0.889	0.891	0.926	0.893	0.926	0.899
These regressions were estimated using panel data for 48 U.S. states. Regressions (1) through (6) use data for all years 1982 to 1988, and regression (7) uses data from 1982 and 1988 only. The data set is described in Appendix 10.1. Standard errors are given in parentheses under the coefficients, and p-values are given in parentheses under the F-statistics. The individual coefficient is statistically significant at the +10%, *5%, or **1% significance level.							

图 7.1: 香烟需求

从图 1 中可以看出, (1) 列和 (2) 列只有一个工具变量, 因此, 不存在 J 统计量。而在第 (3) 列中, 有两个工具变量, 是过度识别情形, 因此, 可以计算 J 统计量, 其结果为 4.93, 它服从卡方分布, 5% 的临界值为 3.84, 因此, 它在 5% 的水平下拒绝两个工具变量都是外生的假设。这是因为两个工具变量的 $TSLS$ 估计量差异很大。 J 统计量拒绝原假设意味着第 (3) 列的估计是基于无效的 IV 估计, 因为 IV 外生性条件不满足。那么, 这是否就意味着我们估计的三个结果都不行呢? J 统计量拒绝原假设只意味着两个工具变量中至少有一个是内生的, 那么, 我们可以推断: 第一, 销售税是外生的, 烟草税不是, 那么, (1) 中的结果就是可靠的; 第二, 烟草税是外生的, 销售税不是, 那么, (2) 中的结果是可靠的; 第三, 两个都不是外生的, 那么, 三个结果在统计意义上都可靠。

千万要记住: 统计证据并不能告诉我们哪一种可能是正确的, 因此这就需要我们根据我们的经验以及经济理论去 **argue**。

7.4 哪里去寻找有效的工具变量呢?

在实践中, IV 回归最难的就是找到有效的工具变量。虽然如此, 但是还是有两种指导性的方法:

(I) 遵循经济理论来找工具变量。例如, IV 回归的发明者 P. Wright (1928) 通过他对农业市场的理解, 使他认识到所寻找的 IV 不是使需求曲线移动, 而是使供给曲线移动, 因此, 他就想到用农业地区的天气条件作为有效 IV 。经济理论法最成功的领域就是金融经济学。在这个领域, 有一些投资者行为的经济模型通常是非线性的, 此时, 不能使用 IV 估计。因此, 将 IV 方法扩展到非线性模型时, 这种扩展方法就是广义矩估计 (GMM)。但是经济理论太抽象, 并不总是能找到一个有效的 IV 。

(II) 构造工具变量。从这种视角出发, 我们要去寻找那些引起内生解释变量变化的随机事件, 从这些随机事件中剥离出 X 变动的外生冲击。

7.5 Stata 命令

*****; 上述回归数据和 stata 命令可以给我发邮件所要!

第 8 章 实验和自然实验

在许多领域，实验是最常用的因果效应估计方法，例如心理学和医药学。研发了一种新药，投放市场前，必须要经过临床实验检验其效力。而这种临床试验就是随机的选择一些病人来服用这种新药，而另一些病人服用无害的无效替代品（也称为“安慰剂”，这就是为什么我们经常在论文中看到“安慰剂检验”）。只有这种随机控制实验表明新药是安全有效的（可信的统计证据），新药才会投放市场。

从目前来看，随机控制实验也是经济学中最重要的关注点，现在很多学者都“绞尽脑汁”去寻找各种政策的“随机控制实验（自然实验）”。如果你找到了，恭喜你成功了一大半了！

当你幸运地发现了一个别人还没有做过的政策“随机控制实验”的时候，接下来你可以就要思考，用什么方法把政策的效应给估计出来。那么，本讲的内容对你可能是最实用的。我们在进行项目评估，也就是估计一个项目、政策、干预或“处理”的效应时，最常用的估计方法有：双重差分（DID）、断点回归设计（RDD）、倾向性匹配得分（PSM）和合成控制法。而其中前两种是最流行、最常用的方法。

8.1 理想实验

假设我们正在进行医药试验，随机选择一部分病患服用试验药品，被称为**处理组**；另一部分病患服用无害无效的安慰剂，称为**控制组**。这样，我们就可以得到两种结果。这两种**潜在结果**的差异就是试验药品对病患的（因果）效应。

在理想实验中，处理组与控制组的病患选择完全是随机的。因此，病患个体究竟被分到哪一组或是否服用试验药品，与个体的特征或其他可能影响潜在结果的因素是完全独立的。因此，解释变量（是否服用试验药品）与扰动项不相关，即 $cov(X_i, u_i) = 0$ 。这样，无论是否有遗漏变量，都不会出现遗漏变量偏误。这就是理想实验的最大优点。

I. 差分估计量

假设 $X_i = 1$ 表示服用新药， $X_i = 0$ 表示未服用新药， Y_i 表示病人的病情。那么，服用新药后的潜在结果为 $\bar{Y}_{i,treat} = E(Y_i|X_i = 1)$ ，而服用安慰剂的潜在病情为 $\bar{Y}_{i,control} = E(Y_i|X_i = 0)$ 。那么，平均处理效应（ATE）就是两种潜在结果之差，即 $E(Y_i|X_i = 1) - E(Y_i|X_i = 0)$ 。如果我们写出回归方程

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (8.1)$$

那么，(1) 式的 OLS 估计量为

$$\tilde{\beta}_1^{OLS} = \bar{Y}_{i,treat} - \bar{Y}_{i,control} \quad (8.2)$$

由此，可以看出， β_1 的 OLS 估计量是处理组样本均值与控制组样本均值之差，故也被成为“差分估计量”。

注意：虽然，在理想实验中，服用新药的病人是随机分配的，无论是否存在遗漏变量，服用新药（处理）对病情影响的估计量总是无偏、一致的。但是，在实践中，我们通常还是会加入如控制变量 W ，因为在回归中加入控制变量 W 可以提高 β_1 的 OLS 估计量的效率（Stock and Watson, 2015）。如果控制变量 W 对解释 Y 的变动有帮助，那么，回归中加入控制变量 W 可以降低 β_1 的 OLS 估计量的标准误。

II. 小班教学实验

20 世纪 80 年代初期，美国田纳西州进行了一项围棋 4 年的初级教育阶段班级缩减的实验，STAR 项目。这项实验主要是想评价小班教学的效果。这次小班教学实验划分了三种班级规模：普通规模——22-25 名学生/班/教师，无助教；小班——13-17 名学生/班/教师，无助教；普通规模，有助教。学生和老师在入学时都是**随机分配**到不同班级规模。这项实验的第一年，大约有 6400 名学生，108 个小班，101 个无助教普通班，99 个有助教普通班。4 年期间，共有 11600 名学生和 80 所学校参与小班教学实验。

这个实验与上面论述的有点差异：两个**处理组**——小班（ SC_i ）和有助教普通班（ RA_i ），一个**控制组**——无

助教普通班。因此，小班教学实验的基准回归模型为

$$Y_i = \beta_0 + \beta_1 SC_i + \beta_2 RA_i + W_i \gamma_i + u_i \quad (8.3)$$

其中， Y_i 表示考试成绩；如果第 i 个学生在小班，那么 $SC_i = 1$ ，否则等于 0；如果第 i 个学生在有助教普通班，那么 $RA_i = 1$ ，否则等于 0； W_i 表示控制变量向量。根据前面对平均处理效应的论述，相对于无助教普通班来说，小班的效应就是 β_1 ，有助教普通班的效应是 β_2 。回归方程 (3) 中， β_1 ， β_2 的 OLS 估计量可以通过差分估计量计算得到。如表 1 所示。

表 8.1: 小班教学实验的效应：无控制变量

解释变量	(1)	(2)	(3)	(4)
小班	13.90** (2.45)	29.78** (2.83)	19.39** (2.71)	19.59** (2.40)
有助教普通班	0.31 (2.27)	11.96** (2.65)	3.48 (2.54)	-0.29 (2.27)
常数项	918.04** (1.63)	1039.39** (1.78)	1157.81** (1.82)	1228.51** (1.68)
Obs.	5786	6379	6049	5967

注：(1)1-4 列分别为全样本，以及 1-3 年级样本；
(2) 括号中为标准误；
(3)***、**、* 表示 1%、5%、10% 的显著性水平。

下面，我们加入一些控制变量。回归结果如表 2 所示。

[插入表 2]

表 8.2: 小班教学实验的效应：控制变量

解释变量	(1)	(2)	(3)	(4)
小班	13.90** (2.45)	14.00** (2.45)	15.93** (2.24)	15.89** (2.16)
有助教普通班	0.31 (2.27)	-0.60 (2.25)	1.22 (2.04)	1.79 (1.96)
教师执教年数		1.47** (0.17)	0.74** (0.17)	0.66** (0.17)
男孩				-12.09** (1.67)
免费午餐				-34.70** (1.99)
黑人				-25.43** (3.50)
其他种族				-8.50 (12.52)
常数项	918.04** (1.63)	904.72** (2.22)		
学校指标	no	no	yes	yes
\bar{R}^2	0.01	0.02	0.22	0.28
Obs.	5786	5766	5766	5748

注：(1)1-4 列分别为全样本，以及 1-3 年级样本；
(2) 括号中为标准误；
(3)***、**、* 表示 1%、5%、10% 的显著性水平。

表 1 和表 2 中估计的教学实验效果应该如何理解呢？

有两种方式回答这个问题：

第一，将每一行的考试分数估计值转换成单位考试得分标准差的效应，这就使得每个年级的估计效应可以进行比较。例如，学生考试成绩的全样本标准差为 73.7，从表 1 中可以知道，小班教学的效应估计值是 13.9，那么，小班教学的单位得分标准差效应为 $13.9/73.7=0.19$ ，标准误为 $2.45/73.7=0.03$ 。

第二，将班级规模的效应系数与另外一些估计系数进行比较。

III. 准实验或自然实验

这项实验花费了田纳西州 1200 万美元。因此，实验经济学成本实在太高，对于我们普通人来说几乎不可能实施。而理想实验的最大优势在于处理组和控制组的随机配分配。因此，我们要紧紧抓住这个随机性。

在实践中，我们通常见到的是一种非实验情形，但它又具有某些随机性。我们把这种情形称为准实验或自然实验。在这些自然实验情形下，对个体的处理通常是“似乎”是随机分配的。在我们国家，最常见的自然实验就是各种改革措施或政策的试点，例如“营改增”、“省直管县”、“开发区”等等。

两种类型的自然实验：

(1) 个体是否受到“处理”，似乎是随机决定的。这种情形就可以利用全面的“差分估计量”来计算处理效应；

(2) 随机变动“似乎”只是部分决定了是否被处理。这个时候，因果效应就可以利用 IV 回归。回忆一下第七讲中构造 IV 的方法就是分离出与误差项无关的成分，而这种自然实验中的随机变动部分就提供了工具变量。

下面的问题就是，我们如何估计自然实验中的因果效应。

8.2 事件研究

8.3 双重差分 (DID)

在自然实验中，个体接受处理与否似乎是随机分配的，但是我们并不能控制这种随机性，即使我们控制了那些影响随机性的变量 W ，处理组和控制组之间的某些效应仍然不能估计出来。

有一种方法可以消除上述问题：我们不去比较处理组和控制组之间产出水平 Y 的差异，而是去比较两组产出 Y 变化的差异，和处理前后 Y 变化的差异。这个估计量是处理组和控制组之间效应变化的差异，或者时间上的差异，因此，这就是我们熟知的双重差分 (DID) 估计量。这种方法似乎与事件研究设计类似。与事件研究设计不同的是，DID 引入了永远不被处理的组群，也就是说，在数据中，我们既有处理组，也有未处理组。这似乎有点反直觉：未处理组可能与处理组存在差异。那么，我们不是引入了组群差异等额外的因素来影响识别结果吗？这样不是让事情变得更坏了吗？尽管可能会有这些问题，但是，DID 的关键在于，我们现在有了未处理组，虽然增加了其它因素，我们就可以控制这些因素。

8.3.1 DID

下面，我们来看看上海对外经贸大学的司继春博士在“知乎”上举的一个例子：<https://www.zhihu.com/question/24322044>)

现在要修一条铁路，铁路是条线，所以必然会有穿过的城市 and 没有被穿过的城市。记 $D_i = 1$ ，如果铁路穿过城市 i ； $D_i = 0$ ，如果城市 i 没有被穿过。现在我们感兴趣的问题是：铁路修好以后，被铁路穿过的城市是不是经济增长更快了？我们该怎么做呢？一开始的想法是，我们把 $D_i = 1$ 的城市的 GDP 加总，减去 $D_i = 0$ 的城市的 GDP 加总，然后两者一减，即 $E(Y_i | D_i = 1) - E(Y_i | D_i = 0)$ ，这样我们就算出了两类城市 GDP 的平均之差。如果大家还记得，这就是我们前面所讲的理想实验中的差分估计量。

那么，这样做是不是就得到了我们感兴趣的铁路效应呢？不用说这样肯定有问题。如果没有问题，那我们讲完第一节就可以打铃下课，我收工回家了。我们想想，万一被铁路穿过的城市在建铁路之前 GDP 就高呢？为了解决这个问题，我们需要观察到至少两期，第一期是建铁路之前，第二期是建铁路之后。我们先把两类城市的 GDP 做铁路修建前后两期之差，即：

$$\Delta Y_i = \frac{1}{N} \sum (Y_{i,after} - Y_{i,before}) \quad (8.4)$$

(4) 式就是第一次差分。它计算的实际上是城市 i 建铁路前后平均 GDP 的增长（如果是 GDP 取对数，就是增长率）。接下来，我们再来计算 GDP 变化的平均处理效应，也就是

$$ATE = E(\Delta Y_i | D_i = 1) - E(\Delta Y_i | D_i = 0) \quad (8.5)$$

这是第二次差分。这一步就把两类城市在修建铁路之前和之后的 GDP 增长的差异给算出来了，这就是我们要的处理效应，即修建铁路之后对城市经济的促进作用。

还可以将 DID 估计量换一个写法。记 $T=1$ ，如果时间为建铁路之后； $T=0$ ，如果时间为建铁路之前。然后，结合上面城市修建铁路与否的虚拟变量，我们可以得到下面的表 3：

表 8.3: DID 估计量

Treated	D=1	D=0
T=1	1	0
T=0	0	0

Treated 表示在某一时期，某城市是否修建了铁路。从表 3 可以看出，在 $T=0$ 时期，没有城市修建铁路，而在 $T=1$ 期，也只有 $D=1$ 的城市修建了铁路。因此， $Treated = D_i \times T$ 。我们可以写出下列回归方程：

$$Y_{it} = \alpha D_i + \beta T + \gamma D_i \times T + u_{it} \quad (8.6)$$

其中， Y_{it} 表示城市 i 在第 t 期的 GDP。我们感兴趣的是系数 γ 。

首先，我们将 (5) 式在时间维度上做一次差分：

$$\Delta Y_i = \beta + \gamma D_i + \Delta u_i \quad (8.7)$$

然后，再对 (6) 式在个体层面做一次差分，并取期望：

$$E(\Delta Y_1 - \Delta Y_0) = \gamma \quad (8.8)$$

到此，我们得到了建铁路的经济增长效应 DID 估计量 $\hat{\gamma}$ 。

这是怎么发生的呢？

- 1 分离出有高铁和没有高铁这两类城市的组内变动。因为我们分离出了城市变动，我们就可以通过**城市组**来控制其的差异，进而关闭城市因素的影响——第一次差分；
- 2 比较有高铁的城市的差异与没有高铁城市的差异。因为没有高铁城市的变动会受到时间的影响，那么，前面的差异比较就可以控制时间变动，进而通过**时间**来关闭那些由于时间变动产生的影响——第二次差分。

总而言之，我们想要的估计量 = (处理后的处理组 - 处理前的处理组) - (处理后的未处理组 - 处理前的未处理组)。这隐含意味着，未处理组的变动代表着没有发生处理时，处理组的预期变动。因此，未处理组对于 DID 非常关键，没有未处理组，就不能做 DID。

8.3.2 更多经典 DID 的例子

8.3.3 未处理组与平行趋势假设

既然未处理组这么重要，那么，它要具备什么特征，DID 才是一个好的识别策略呢？在做 DID 的时候，我们需要未处理组满足一定的条件，我们称之为**平行趋势假设**。

平行趋势假设说的是，如果没有发生处理，那么，处理组和未处理组仍然在处理时点后保持相同的变化趋势（与处理时点前的趋势相同）。

但是，很不幸，平行趋势不可观测。它是一种反事实的情形：是假设处理没有发生的时候的情形。我们来看一个不满足平行趋势假设的例子。我们想象一下，我想看看高铁站对周围酒店餐馆的效应。例如 2008 年在武汉的汉口、武昌和青山都有高铁站（汉口站、武昌站和武汉站），而在武汉黄陂区则没有高铁站（黄陂区有美丽的“花木兰故乡”等旅游景点，是武汉的后花园，欢迎大家前来旅游，我可以做导游）。这个时候，我们肯定会想着用黄陂区作为未处理组。

我们来看看 2007 年（处理前）和 2008 年（处理后）的武汉汉口和黄陂，用 DID 来识别高铁站对地区的酒店餐馆的效应。结果发现，汉口的酒店餐馆生意更差，没有黄陂的餐馆受欢迎。What 弄啥呢？这一实证结果与我们的预期结果不符呀，这怎么办，收集数据做得回归都白瞎了，我的顶刊梦碎了（欲哭无泪呀）。不要着急，不要心慌，不要放弃。我们来看看是什么原因造成了这种结果。

我们仔细分析一下我们的数据。突然发现，2008 年，黄陂区突然新开了大量的酒店和餐馆（木兰特色的）。哦哦，原来上面的实证结果——汉口和黄陂 2007/2008 年的变动包括两个方面：汉口高铁站的新建和黄陂特色旅游酒店和餐馆的新建。这个时候就很明显了，我们不能得到结论，高铁站让汉口的酒店和餐馆经营变得更差了。上述 DID 估计并不是一个很好的估计量，即我们没有很好的识别出高铁站对酒店餐馆的效应。我们应该选择一个地区，在 2008 年并没有新建很多酒店和餐馆，假设武汉江夏区。但是，理想可能很丰满，现实却很骨感。我们可能根本就找不到一个在 2007-2008 没有发生任何变化的地区，因为那段时期的武汉到处都在大挖大建（我曾经在外省读书一年，然后过年回家，没找到回家的路，汉口站前面修建了一个二环高架，我家在火车站旁边五分钟，我硬是没找到回家的路，最后让我爸妈来接我了。）。如果我们选择武汉江夏来作为未处理组，我们可能发现，高铁站可能也没有使得汉口的酒店餐馆变得更好，因为太多的大学跑到江夏去建立新校区了，因此带过去了大量的大学，因此，带动了江夏的“堕落街”兴起。因此，这个识别也不好。

记住，DID 的研究设计就是利用未处理组来代表处理组中的所有非处理变动。

那么，我们可以将上述假设用下列数学表示出来：

- 1 处理组在处理前后的差异 = 处理效应 + 其他因素引起的处理组变化
- 2 未处理组在处理前后的差异 = 其他因素引起的未处理组变化
- 3 DID 的效应 = 处理效应 + 其他因素引起的处理组变化 - 其他因素引起的未处理组变化

那么，对于 DID 来说，我们仅仅需要识别**处理效应**，也就是说，“其他因素引起的处理组变化”要抵消“其他因素引起的未处理组变化”。这就是平行趋势的内容。

那么，我们在做 DID 前，如何挑选未处理组以使得 DID 识别比较可信呢？我们可以做一下一些事情（并非必须，但大有益处）：

- 1、找不到理由说，未处理组在处理时点突然发生了变化；
- 2、处理组和未处理组在许多方面都是类似的；
- 3、处理组和未处理组的因变量在处理前有相似的变化路径。

下面，我们用图 1 来看看。

由图 1 可以清晰地看出，DID 最关键的假设是 **common trend**，也就是两个组别在不处理的情况下，Y 的趋势是一样的。那么你仍会说，铁路穿过的城市可能本身 GDP 也高，而 GDP 高的城市按照理论 GDP 增长率可能更高可能更低，所以 **common trend** 的假设可能是不对的，那怎么办？如果这个问题存在，我们可以进一步假设在控制了某些外生变量之后，**common trend** 是对的，比如上个问题，我们可以控制城市在 $t=0$ 期的 GDP level。当我们控制其他变量之后，自然不能直接减两次了，我们需要用上面说的回归式子，即对下列回归方程 run OLS:

$$Y_{it} = \alpha D_i + \beta T + \gamma D_i \times T + X' \delta + u_{it} \quad (8.9)$$

其中，X 是控制变量向量。

既然 **common trend** 是 DID 最关键的假设，那么，我们如何评价非平行趋势呢？也就是，我们可以采取一些方法和方式来检查一下平行趋势假设，看看我们采用的 DID 识别是否可信。但是，需要特别提醒的是，这些方法和方式并不是检验平行趋势是否成立。即使“通过”了这些检验，我们也不能说平行趋势就一定成立。实际上，没有检验可以证实或者证伪平行趋势假设，因为它是反事实的，我们观测不到。这些检验方法更多是建议性的证据。如果没通过这些检验，那么，平行趋势假设可行性就很低。

此外，下面我们来看看实际操作中最常用的两种方法（必做）：

方法一、画时间趋势图

如果在政策干预前有多期数据，则可分别画处理组与控制组的时间趋势图（类似于上图），并直观判断这两组的时间趋势是否平行（比如，考察是否存在 **Ashenfelter's dip**）。如果二者大致平行，则可增强对平行趋势假定

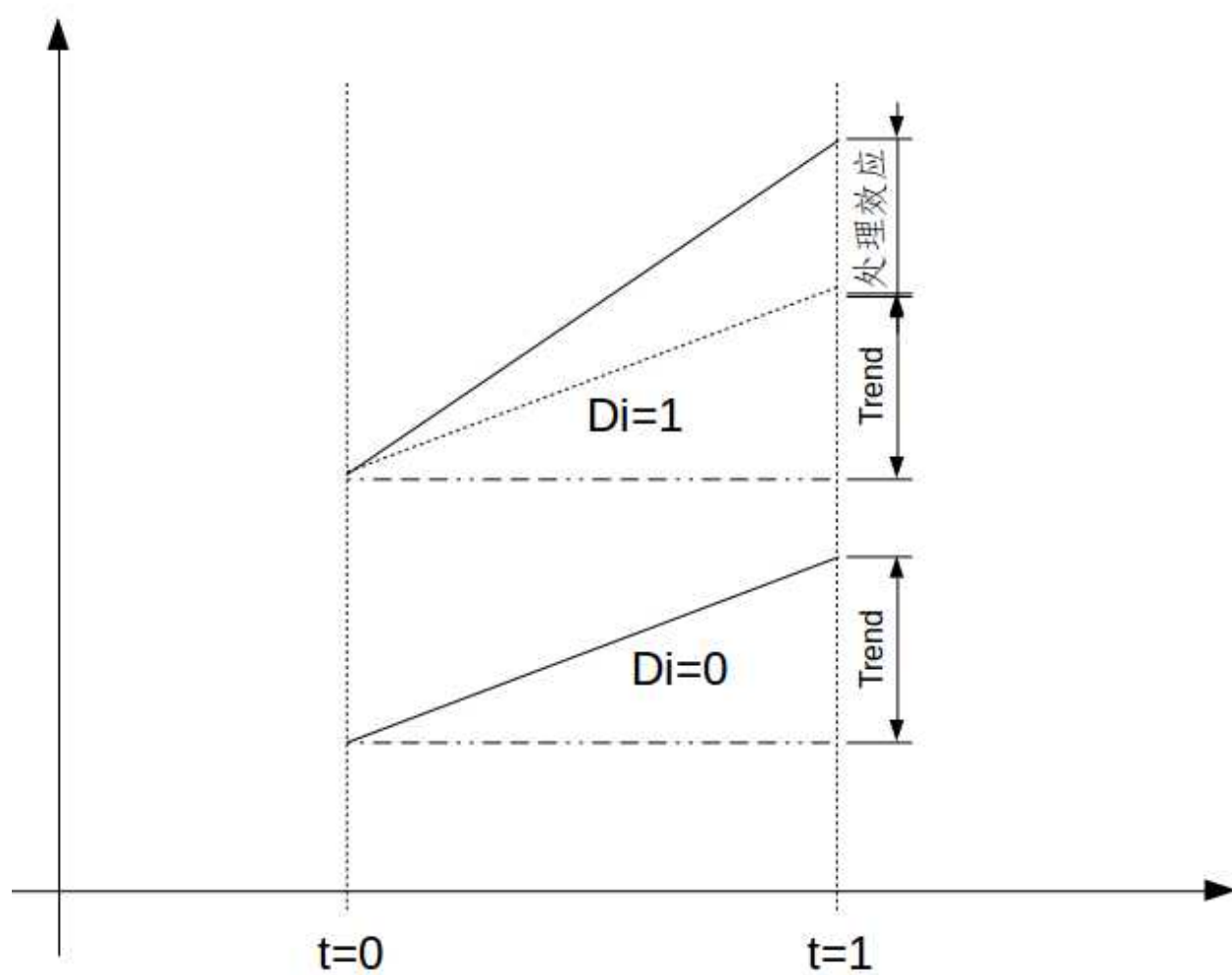


图 8.1: DID 估计量-平行趋势

的信心。然而，即使在政策干预前两组的时间趋势相同，也无法保证二者在干预后的时间趋势也相同（后者本质上不可观测，因为时间效应已与处理效应混合在一起）。另外，如果只有两期数据，则无法使用此法。

方法二、安慰剂检验

在 DID 的安慰剂检验中，例如，2008 年武汉修建了高铁站，那么，我们仅仅只用 2008 年以前的数据样本，忽略掉所有的 2008 年（处理后）的数据。然后，用 2008 年前的数据样本，我们挑选一些不同的时期，假设高铁站修建在这些时间。再然后，我们用假设的处理时点来估计 DID。如果我们仍然发现了在假设的高铁站修建时点有显著的 DID 效应，那么，这就意味着可能还有一些其它的因素干扰了平行趋势假设。

也就是说，我们估计的 DID 效应显著不等于 0（在实际处理并未发生时）可以给我们传递的信息是，处理组的非处理变化并没有完全抵消掉未处理组的非处理变化。那么，我们还需要更多的笔墨来解释为什么我们应该相信在实际处理时点，两者相互抵消了。

陈强老师于 2016-10-25 在微信公众号“计量经济学及 stata 应用”上给出更多的可操作方法：

方法三、加入更多的控制变量

从上文的讨论可知，非平行趋势可能由于遗漏变量所导致，故在回归方程中加入更多控制变量，或可缓解内生性。但此法在实践中不易实施。

方法四、假设线性时间趋势

如果假设时间趋势为线性函数，则可加入每位个体的时间趋势项：

在具体回归时，加入个体虚拟变量与时间趋势项 $t = 1, 2, \dots, T$ 的交互项即可。然而，线性时间趋势毕竟是比较强的假定，不一定能成立。故此法也不完全解决问题，但可作为稳健性检验。

方法五、三重差分法

在一定条件下，可通过引入两个控制组，进行三次差分，称为“三重差分法”（difference-in-differences-in-differences，简记 DDD），这样可以更好地控制时间趋势的差异，使得平行趋势假定更易成立。有关 DDD 的进一步介绍，参见陈强（2014，第 343 页）。

最后但也很重要的事（经常被忽略）：平行趋势意味着我们还要认真仔细地想想我们的因变量是如何测度和进行数据转换的。因为平行趋势并不仅仅是因果效应的假设，它也是处理组和未处理组在处理前的差异大小基本保持恒定，这就意味着我们还要考虑我们如何测度这个差异的问题。我们以对数转换为例，如果因变量 Y 的平行趋势成立，那么， $\ln(Y)$ 就不成立，反之亦然。

这是一件显而易见的事，但是我们当中许多人从来不会去考虑这个问题。因此，我们仔细思考一下因变量满足平行趋势的形式是什么，然后我们就使用这种形式的因变量。

8.3.4 DID 在 Stata 中的实现


要估计自然实验中的平均处理效应，如果直接在 stata 中 run (9) 式，那么，直接使用普通的面板数据命令 `xtreg` 即可。而 DID 则有专门的命令估计。厦门大学赵西亮老师的书里介绍了一种 DID 的命令，`diff`，其语法和基本选项为：

```
diff outcome-var [if] [in] [weight], period(varname)
treated(varname) [cov(varlist)
kernel id(varname) bw(#) ktype(kernel) rcs
qdid(quantile) pscore(varname) logit
support addcov(varlist) cluster(varname)
robust bs reps(int) test report nostar export(filename)]
```

`outcome-var` 是结果变量，`period(varname)` 告诉软件时期变量，`treated(varname)` 告诉软件处理变量。其他命令（也就是中括号里的命令）都是可选择的。参见赵西亮（2017）第 177 页。

操作实例：下面，我们利用 Card and Krueger（1994，AER）的数据为例，估计新泽西州最低工资调整对新泽西州快餐业就业的影响，数据为两期面板数据，主要变量有：`id` 为快餐；`t` 为时间，最低工资调整前为 0，调整后

为 1; `treated` 为分组变量, 1 为新泽西, 0 为宾夕法尼亚; `fte` 为全职就业人数, 协变量有 `bk`、`kfc`、`roys`、`wendys`。如下图 2 所示:



Name	Label
<code>id</code>	Store ID
<code>t</code>	Feb. 1992 = 0; N...
<code>treated</code>	New Jersey = 1; ...
<code>fte</code>	Output: Full Tim...
<code>bk</code>	Burger King == 1
<code>kfc</code>	Kentucky Fried C...
<code>roys</code>	Roy Rogers == 1
<code>wendys</code>	Wendy's == 1

图 8.2: 变量

首先, 安装 DID 命令: `ssc install diff, replace`

然后, 我们就可以在 `stata` 中输入 DID 命令估计回归系数。

不控制任何协变量时的结果:

```
diff fte, period(t) treated(treated) robust
```

控制协变量时的结果:

```
diff fte, period(t) treated(treated) robust cov(bk kfc roys)
```

```
. diff fte,period(t) treated(treated) robust
```

DIFFERENCE-IN-DIFFERENCES ESTIMATION RESULTS
Number of observations in the DIFF-IN-DIFF: 780

	Baseline	Follow-up	
Control:	76	76	152
Treated:	314	314	628
	390	390	

Outcome var.	fte	S. Err.	t	P> t
Baseline				
Control	20.013			
Treated	17.069			
Diff (T-C)	-2.944	1.440	-2.04	0.041**
Follow-up				
Control	17.523			
Treated	17.518			
Diff (T-C)	-0.005	1.037	-0.00	0.996
Diff-in-Diff	2.939	1.774	1.66	0.098*

R-square: 0.01
* Means and Standard Errors are estimated by linear regression
**Robust Std. Errors
Inference: * p<0.01; ** p<0.05; * p<0.1

图 8.3: DID 结果

```
. diff fte,period(t) treated(treated) robust cov( hk kfc roys)
```

DIFFERENCE-IN-DIFFERENCES WITH COVARIATES
Number of observations in the DIFF-IN-DIFF: 780

	Baseline	Follow-up	
Control:	76	76	152
Treated:	314	314	628
	390	390	

Outcome var.	fte	S. Err.	t	P> t
Baseline				
Control	21.342			
Treated	19.003			
Diff (T-C)	-2.339	1.282	-1.83	0.068*
Follow-up				
Control	18.852			
Treated	19.452			
Diff (T-C)	0.600	0.912	0.66	0.511
Diff-in-Diff	2.939	1.573	1.87	0.062*

R-square: 0.19
* Means and Standard Errors are estimated by linear regression
**Robust Std. Errors
Inference: * p<0.01; ** p<0.05; * p<0.1

图 8.4: 控制协变量的 DID 结果

8.3.5 长期效应

8.3.6 交叠 DID 设计

8.4 断点回归设计 (RDD)

如果大家关注了微信公众号“香樟经济学术圈”的话，肯定记得 2016 年的时候，“满天都是 RD”——各种 RD 经典文献解读，RD 原理介绍。社科院付明卫老师写了一篇“断点回归 (RD) 的规定动作”的推文。里面写道：

订阅了各种经济学类公号的小伙伴们，最近有没有断点回归 (RD) 设计满天飞的感觉？作为同道中人，我感觉，被推送的 RDD 论文数量，在今年六七月份明显存在一个断点：从那以后，开始井喷！看着这些推文，多少人心头默念：“论文发表不轻松，要把断点为我用！”

RDD 确实是个好方法。它等于是在断点附近的局部随机试验。这一点赖以成立的前提条件，并不难以满足。此外，跟随机试验中全域 (global) 随机性可以被检验一样，RDD 等于局部随机试验的假设，也可以通过观察前定变量的分布是否平衡来检验。从这个意义上讲，RD 方法比 IV、DiD 更接近于随机试验。随机试验是因果识别的终极杀招，越接近随机试验的方法当然越好！

8.4.1 断点回归估计量

在自然实验中，还可能出现一种情形：个体接受处理完全或部分依赖于某个可观测变量 W 是否超过某一阈值（门槛）。例如，一个学生是否要参加“短学期”依赖于他期末平均绩点 (GPA) 是否在规定阈值以下。根据前面的理想实验中平均处理效应的 idea，估计参加“短学期”的效应也是要比那些 GPA 在阈值以下（参加短学期）的学生成绩与那些 GPA 在阈值以上（不参加短学期）的学生成绩。这个有阈值限制的可观测变量 W 称为参考变量。

另外一个例子是 Lee(2008, JoE) 对美国各地区众议员选举中在位党在竞选中是否具有优势的分析。美国两大党派在选举中获得的选票份额超过对手时，该党就是在位党。Lee 以民主党选票份额与共和党选票份额之差作为参考变量 W ，间断点为 0，只要上次选举中参考变量大于 0，即意味着民主党在位，否则共和党在位。图 5 展示了数据集的散点图，如果 $W > 0$ ，民主党在位。

图 5 显示了，下一次的选举得票份额是现在两党得票之差 W 的函数。如果阈值 w_0 的唯一作用只是识别在位党派，那么，下一次选举得票份额在阈值处的“跳跃”就是竞选中在位的效应估计值。

也就是说，更一般化的分组规制是

$$D_i = \begin{cases} 1 & \text{if } x_i \geq w_0 \\ 0 & \text{if } x_i < w_0 \end{cases} \quad (8.10)$$

假设在选举前，各党派的得票份额的结果 y_i 与 x_i 之间存在如下线性关系：

$$y_i = \alpha + \beta x_i + \epsilon_i \quad (8.11)$$

我们从图 5 可以看出，在 $x_i = w_0$ 处， y_i 与 x_i 的线性关系存在一个向上跳跃 (jump) 的断点。但是，得票率 (%) 为 49.8、49.9、50、50.1、50.2 等，可以认为党派在各个方面没有系统差异，因此，这个跳跃发生的唯一原因只可能是 D_i 的处理效应，也就是在位党的优势。

图 5 也是一个分段函数，因此，我们可以引入虚拟变量来表示具有不同截距的分段函数。因此，我们可以将 (11) 式重新写成：

$$y_i = \alpha + \beta(x_i - w_0) + \delta D_i + \gamma(x_i - w_0)D_i + \epsilon_i \quad (8.12)$$

引入交互项 $(x_i - w_0)D_i$ 是为了允许在断点两侧的回归线斜率不同。对方程 (12) 进行 OLS 回归，得到的 δ

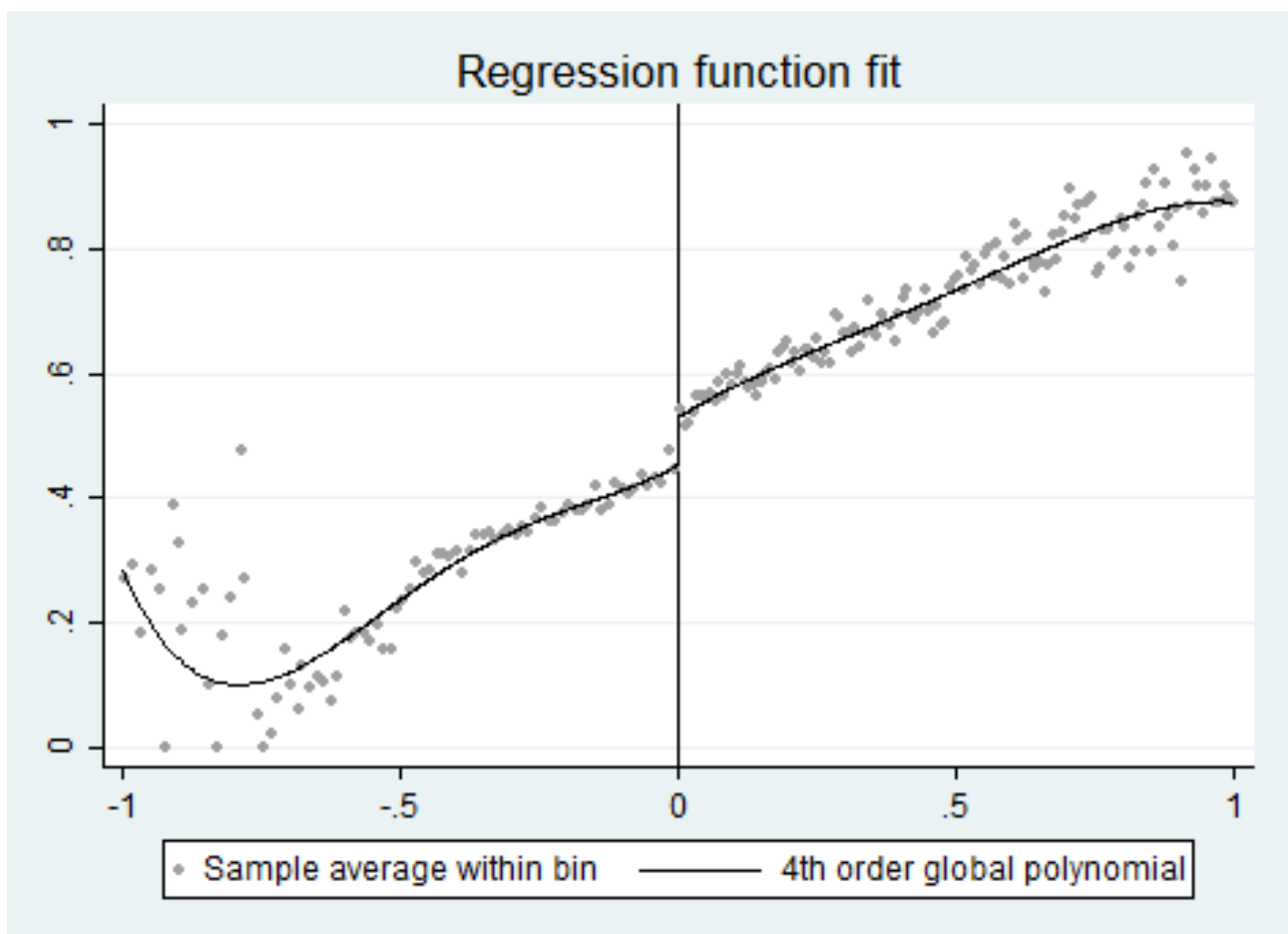


图 8.5: 断点回归设计的散点图

就是断点回归估计量，也称为局部平均处理效应 (LATE)。

在估计断点回归时，要特别注意两点：

1、方程 (12) 中包含了交互项。如果断点两侧的回归线斜率相同，则可不包含交互项。但在实践中，一般断点两侧斜率会不同，因此，如果不包含交互项，则可能导致断点右（左）侧的观测值影响对左（右）侧截距的估计，从而引起偏误；

2、在有交互项的情形下，如果方程中没有 $(x_i - w_0)$ ，而是使用的 x_i ，那么，虽然 δ 还是断点两侧的距离之差，但是并不等于这两条回归线在 $(x_i = w_0)$ 处跳跃的距离。

由于在参考变量的阈值处，结果变量的跳跃或断点，那些探讨在某一阈值处接受处理的概率的非连续性的研究被称为断点回归 (RD) 设计。它又分为精准断点回归 (sharp RD) 和模糊断点回归 (fuzzy RD)。SRD 在断点 $x_i = w_0$ 处，个体接受处理的概率从 0 跳跃到 1，而 FRD 在断点 $x_i = w_0$ 处，我们只知道个体接受处理的概率从 a 跳跃到 b ，而 $0 \leq a \leq b \leq 1$ 。

I. 精准断点回归

在 Sharp RDD 中，接受处理完全由参考变量 W 是否超过某一阈值决定：当 $W \geq 0$ 时，民主党是在位党，当 $W < 0$ 时，共和党是在位党；即用 D 表示民主党是否在位，当 $W \geq 0$ 时， $D_i = 1$ ，当 $W < 0$ 时， $D_i = 0$ 。在这种情形下，下一次获得选票份额 Y 在 $W = 0$ 处的跳跃就等于 $W = 0$ 时子样本的处理效应。

由此，我们可以看到，上述例子是一个精准断点回归。那么，我们是否还可以利用 (12) 式进行 OLS 估计呢？可以是可以，但是这存在两个问题：

1、可能存在遗漏变量偏误，例如如果回归中还有高次项 $(x_i - w_0)^2$ ；

2、断点回归可以看作是“局部随机实验”，因此从原理上看，我们应该只是用断点附近的观测值样本，但我们在实践中却是用全部样本进行回归。

为了解决上述问题，我们可以引入高次项，并限定 x 的范围 $w_0 - h \leq x_i \leq w_0 + h$ 。这里的 h 就是最优带宽。回归方程变为

$$y_i = \alpha + \beta_1(x_i - w_0) + \beta_2(x_i - w_0)^2 + \delta D_i + \gamma_1(x_i - w_0)D_i + \gamma_2(x_i - w_0)^2 D_i + \epsilon_i, \quad w_0 - h \leq x_i \leq w_0 + h \quad (8.13)$$

可是，现在我们不能确定最优带宽 h ，还是不能估计 (13) 式呀。在确定 h 时，一般是采用非参数回归来最小化均方误差 (MSE)。直观来说， h 越小，偏差越小，但是估计方差会变大；反之亦然。

针对断点回归，我们一般使用两种核回归 (kernel regression)：三角核 (triangle kernel) 与矩形核 (rectangle kernel)。

关于协变量问题

1、我们可以在 (13) 式中加入影响 Y 的协变量。虽然断点回归是局部随机实验，包不包括协变量并不影响断点回归估计量的一致性，但是加入协变量的好处为：加入协变量可以解释被解释变量 Y ，那么，就可以减低方差。使得估计更准确。但坏处是：如果加入的协变量是内生变量，与误差项相关，那么就会影响估计量。

2、实际上，断点回归有个隐含假设：协变量在断点处不存在跳跃，是连续的。如果协变量在断点处也存在跳跃，那么，我们就不能把 δ 全部归于处理效应。因此，在实践中，我们要现将所有的协变量作为被解释变量，进行断点回归，考察其分布是否在断点处存在跳跃。

此外，我们还应该注意“内生分组”问题。如果个体事先知道分组规则，并可通过自身行为来完全控制分组变量，那么，就可以自行选择进入处理组还是控制组，这就导致了随机分组失败，从而断点回归失灵。

小贴士：在实践中，我们建议同时汇报出以下情形，以确保结果稳健：

- 1、分别汇报三角核与矩形核的回归结果；
- 2、分别汇报使用不同带宽的结果；
- 3、分别汇报包含协变量与不包含协变量的结果；
- 4、进行模型设定检验时，包括检验分组变量与协变量的条件密度在断点处是否存在跳跃。

II. 模糊断点回归

在 Fuzzy RDD 中，参考变量超过阈值会影响到是否接受处理，但这不是决定处理的唯一影响因素。例如，假

设有些 GPA 在阈值以下的学生并没有参加短学期，而有些 GPA 超过阈值的学生又参加了短学期。如果临界值规则是一个决定 *treated* 非常复杂的过程的一部分，那么上述情况就可能会出现。在模糊断点回归中， X_i 一般与误差项 u_i 相关。

8.4.2 断点回归的规定动作

下面的内容来源于“香樟经济学术圈”的推文，付明卫（2016）：

第 1 步

检查配置变量（assignment variable，又叫 running variable、forcing variable）是否被操纵。画出配置变量的分布图。最直接的方法，是使用一定数量的箱体（bin），画出配置变量的历史直方图（histogram）。为了观察出分布的总体形状，箱体的宽度要尽量小。频数（frequencies）在箱体间的跳跃式变化，能就断点处的跳跃是否正常给我们一些启发。从这个角度来说，最好利用核密度估计做出一个光滑的函数曲线。McCrary（2008）为判断密度函数是否存在断点提供了一个正规的检验（命令是 `DCdensity`，介绍见陈强编著的《高级计量经济学及 Stata 应用》（第二版）第 569 页）。

第 2 步

挑选出一定数目的箱体，求因变量在每个箱体内的均值，画出均值对箱体中间点的散点图。一定要画每个箱体平均值的图。如果直接画原始数据的散点图，那么噪音太大，看不出潜在函数的形状。不要画非参数估计的连续统，因为这个方法自然地倾向于给出存在断点的印象，尽管总体中本来不存在这样的断点。需要报告由交叉验证法（Cross-validation, CV）挑选的带宽。一般而言，为了看出潜在函数的形状，不要挑选过大的带宽。但是，带宽太小也会导致看不出潜在函数的形状。比较因变量均值在断点两边的两个箱体间的变化，可以预判处理效应的大小。如果图形中都看不出因变量在断点处有跳跃，那么回归方程也不可能得到显著的结果。

第 3 步

将 Y 在每个箱体内的均值作为因变量，用处理变量、配置变量的多次项作为自变量，在断点两边分别跑回归，得到因变量的拟合值。将这些拟合值画在第 2 步的图中，并用光滑的曲线连接起来。在推文人读过的 RD 论文中，多次项一般都使用 1 到 4 次项，但没有论文解释为什么只用到 4 次项。

第 4 步

检验前定变量在断点处是否跳跃。此步和第 1 步是 RD 方法的适用性检验。此步的检验包括两项内容：1. 像前三步那样画前定变量的图。无论参数还是非参数，RD 研究都要大把的图！这些图在正式发表的论文中都必不可少！原文中说了这么句话：用 RD 做的论文，如果缺乏相关的图，十有八九是因为图显示的结果不好，作者故意不报告。2. 将前定变量作为因变量，将常数项、处理变量、配置变量多次项、处理变量和配置变量多次项的交互项作为自变量，跑回归。一个前定变量有一个回归，看所有回归中处理变量的系数估计是否都为 0。检验这种跨方程的假设，需要用似不相关回归（Seemingly Unrelated Regression, SUR）（命令是 `sureg`，用法见陈强编著的《高级计量经济学及 Stata 应用》（第二版）第 471-474 页）。在推文人读过的 RD 实证论文中（尤其是 AER2015-2016 年所有用 RD 做的论文中），均没用 SUR，只是简单的看每个回归中处理变量的系数估计均为 0。

第 5 步

检验结果对不同带宽、不同多项式次数的稳健性。尝试的其它带宽，一般是最优带宽的一半和两倍。挑选多项式的最优次数，可用赤池信息准则（Akaike's Information Criterion, AIC）。在我们尝试的包含配置变量 1 次方、2 次方、N 次方的众多方程中，AIC 取值最小的那个就是我们想要的。实操时，试到多少次为好？原文中至少试到了 6 次。我们做研究时需要试到 10 次还是 100 次呢？Gelman 和 Imbens（2014）解除了我们的这个烦恼，详见“江湖上的新动作”这一部分。

第 6 步

检验结果对加入前定变量的稳健性。如上所述，如果不能操控配置变量的假设成立，那么无论前定变量与因变量的相关性有多高，模型中加入前定变量都不应该影响处理效应的估计结果。如果加入前定变量导致处理效应的估计结果变化较大，那么配置变量可能存在排序现象，前定变量在断点处也很可能存在跳跃。实操时在确定多项式的次数后，直接在回归方程中加入前定变量。如果这导致处理效应估计值大幅变化或者导致标准误

大幅增加,那么可能意味着函数中多项式的次数不正确。另外一个检验是残差化,看相同次数的多项式模型对残差的拟合好不好。

江湖上的新动作

Thistlethwaite 和 Campbell 1960 年首次用 RD 方法做政策评估。经过近 40 年的沉寂后,20 世纪 90 年代末以来,经济学关于 RD 方法的性质、局限性等方面的理论研究有了巨大进展。关于 RD 方法本身的研究,并没有因为 Lee 和 Lemieux (2010) 的发表而停止。我把 Lee 和 Lemieux (2010) 发表后的进展称作“新招式”。据我的不完全了解,“新招式”有这些:

1. 多项式次数的选择。根据 Lee 和 Lemieux (2010),配置变量的次数要试到 N 次。但是, Gelman 和 Imbens (2014) 的 NBER 工作论文说,试到 N 次的做法要不得,最多只能搞到 2 次。至于原因,他们讲了三条,感兴趣的请参考原文。尽管他们的论文还未正式发表,但学界都已乖乖听他们的啦。AER2015-2016 年间所有用 RD 做的论文(共 6 篇)里,5 篇都只用 1 次或 2 次。

2. 最优带宽。Lee 和 Lemieux (2010) 介绍了两种确定最优带宽的方法:拇指规则法(rule of thumb)和交叉验证法(CV)。现在,江湖上有另外两种比较受关注的方法:IK 法和 CCT 法。IK 法以 Imbens 和 Kalyanaraman 两个人命名,对应着论文 Imbens 和 Kalyanaraman (2012)。这篇论文发表在 Review of Economic Studies, Lee 和 Lemieux (2010) 文中提到过此文 2009 年的 NBER 工作论文版。CCT 法以 Calonico、Cattaneo 和 Titiunik 三个人命名,对应着论文 Calonico、Cattaneo 和 Titiunik (2014a)。用非参数法做断点回归估计时的 stata 命令 rd,就是用 IK 法确定最优带宽。stata 命令 rdrobust、rdbwselect,提供 CV、IK、CCT 三种不同的最优带宽计算方法选项。然而,尽管 Calonico、Cattaneo 和 Titiunik (2014a) 2014 年发表在牛刊 Econometrica 上,AER2015-2016 年上的文章没有买它的账。AER2015-2016 年的 6 篇相关文章中,仅有 1 篇提到过 CCT,其他 5 篇就像不知道 Calonico、Cattaneo 和 Titiunik (2014a) 这篇文章。我甚为不解!难道是因为 CCT 非牛人?

3. 核密度检验。Lee 和 Lemieux (2010) 介绍了 McCrary (2008) 的核密度检验方法。Frandsen (2013) 提出了一种新的检验方法,感兴趣的请参考原文。

8.4.3 例子与 stata 操作

下面,我们使用 Lee(2008) 的数据来演示一下断点回归的 stata 操作。这个数据集中包括两个变量:vote 表示民主党的选票份额;margin 表示民主党在上次竞选中获得的选票与共和党选票份额之差。因此,margin 就是参考变量,如果 margin 大于 0,民主党就是在位党,这是一个 SRD。我们感兴趣的问题是,在位党是否会获得优势。我们将样本限制在 $\text{margin} \pm 0.5$ 之间,样本共有 4900 个。

第一步,下载安装断点回归命令。Calonico et al. (2014) 提供了一个专门进行断点回归分析的程序包 rdrobust,里面包含三个命令:rdplot——断点回归图形;rdbwselect——选择最优带宽;rdrobust——估计断点回归估计量。

findit rdrobust (查找、安装 rd 程序)

stata 会出来下列界面

点击“st0366_1 from <http://www.stata-journal.com/software/sj17-2>”,进入页面再点击“click here to install”进行安装。

第二步,画断点图。输入

rdplot vote margin, c(0) nbins(50) (画断点图)

第三步,选择最优带宽。输入

rdbwselect vote margin, c(0) kernel(uni) all (选择最优带宽)

上述命令中, kernel() 是设置核估计方法。此处选择的是矩形核。得到的结果是

第四步,估计断点回归估计量。输入:

rdrobust vote margin, c(0) kernel(uni) all (估计断点回归估计量)

以上四步就是断点回归的基本操作步骤。

下面,我们来详细介绍一下上述三个 rd 命令的基本语法格式:

```

search for rdrobust

Search of official help files, FAQs, Examples, SJs, and STBs

SJ-14-4 st0366 . . . Robust data-driven inference in reg.-discontinuity design
. . . . . S. Calonico, M. D. Cattaneo, and R. Titiunik
(help rdrobust, rdbwselect, rdplot if installed)
Q4/14 SJ 14(4):909--946
conducts robust data-driven statistical inference in
regression-discontinuity designs

Web resources from Stata and other users

(contacting http://www.stata.com)

2 packages found (Stata Journal and STB listed first)
-----

st0366_1 from http://www.stata-journal.com/software/sj17-2
SJ17-2 st0366_1. Update: Local polynomial... / Update: Local polynomial
regression-discontinuity / estimation with robust bias-corrected
confidence / intervals and inference procedures / by Sebastian Calonico,
University of Miami, / Miami, FL / Matias D. Cattaneo, University of

st0366 from http://www.stata-journal.com/software/sj14-4
SJ14-4 st0366. Robust data-driven inference... / Robust data-driven
inference in the regression- / discontinuity design / by Sebastian
Calonico, University of Miami, / Coral Gables, FL / Matias D. Cattaneo,
University of Michigan, / Ann Arbor, MI / Rocio Titiunik, University of

(click here to return to the previous screen)

(end of search)

```

图 8.6: 查找 rd 程序

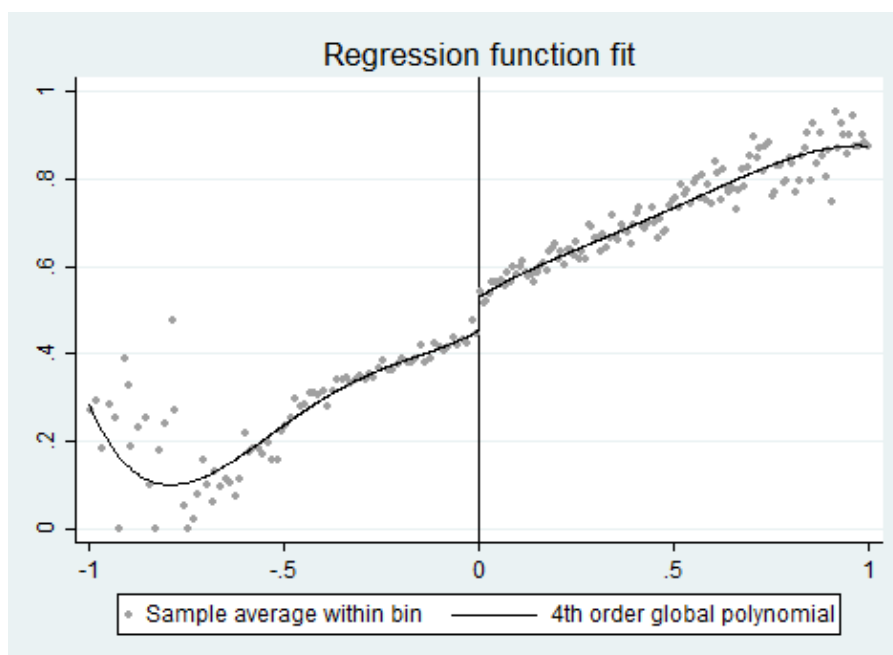


图 8.7: 断点回归设计的散点图


```
. rdbwselect vote margin, c(0) kernel(uni) all
Computing CCT bandwidth selector.
Computing IK bandwidth selector.
Computing CV bandwidth selector.

Bandwidth estimators for RD local polynomial regression
```

Cutoff c = 0		Left of c	Right of c		
Number of obs		2740	3818	Number of obs =	6558
Order loc. poly. (p)		1	1	NN matches =	3
Order bias (q)		2	2	Kernel type =	Uniform
Range of margin		1.000	1.000	Min BW grid =	0.00300
				Max BW grid =	0.99970
				Length BW grid =	0.04984

Method	h	b	rho
CCT	.1198642	.2299834	.5211865
IK	.1781139	.2221868	.8016402
CV	.30201	NA	NA

图 8.8: 最优带宽选择结果

```
. rdrobust vote margin, c(0) kernel(uni) all
Preparing data.
Computing bandwidth selectors.
Computing variance-covariance matrix.
Computing RD estimates.
Estimation completed.

Sharp RD estimates using local polynomial regression.
```

Cutoff c = 0		Left of c	Right of c		
Number of obs		698	729	Number of obs =	6558
Order loc. poly. (p)		1	1	NN matches =	3
Order bias (q)		2	2	BW type =	CCT
BW loc. poly. (h)		0.120	0.120	Kernel type =	Uniform
BW bias (b)		0.230	0.230		
rho (h/b)		0.521	0.521		

Outcome: vote. Running variable: margin.

Method	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Conventional	.06622	.01126	5.8822	0.000	.044155	.088285
Robust	-	-	4.8784	0.000	.037599	.088101

All estimates.

Method	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Conventional	.06622	.01126	5.8822	0.000	.044155	.088285
Bias-corrected	.06285	.01126	5.5828	0.000	.040785	.084915
Robust	.06285	.01288	4.8784	0.000	.037599	.088101

图 8.9: 断点回归估计结果

```
rdplot depvar indepvar [if] [in][,c(#)] p(#) kernel(#) weights() h(# #)
nbins(# #) binselect() scale(# #) ci() shade generate(id_var meanx_var meany_var cil_var cir_var)
graph_options(gphopts) hide]
```

两个必选项：

1、depvar 是结果变量、原因变量或其他协变量；

2、indepvar 是参考变量；

中括号里的全部为可选命令：

3、c(#) 用于设置断点的位置；

4、p(#) 设定多项式的阶数；

5、kernel(#) 设定核估计类型，有三种：三角核 triangular、Epanechnikov 核 epanechnikov、矩形核 uniform；

6、h(# #) 设置断点左右的带宽；

7、nbins(# #) 设定划分的区间数；

8、binselect() 设定带宽的选择方法；

9、ci() shade 画出每个区间拟合点的置信区间，shade 表示置信区间用阴影表示。

```
rdbwselect depvar indepvar [if] [in][,c(#)] p(#) q(#) deriv(#) fuzzy(fuzzyvar[sharpbw])
covs(#) kernel(#) bwselect() scaleregul(#) vce(vctype[vceopt1 vceopt2]) all]
```

最优带宽选择命令中与画图命令中有很多相同命令。需要注意的是：

1、q(#) 为偏差修正的多项式阶数；

2、deriv(#) 可以用于估计弯折回归（RKD），0 为断点回归，1 为弯折回归；

3、fuzzy(fuzzy-var[sharpbw]) 用于模糊断点回归或模糊弯折回归，fuzzyvar 是原因变量，sharpbw 表示使用结果变量的最优带宽；

4、covs(#) 引入协变量；

5、bwselect() 最优带宽的估计方法。

```
rdrobust depvar runvar [if] [in][,c(#)] p(#) q(#) deriv(#) fuzzy(fuzzyvar[sharpbw])
covs(#) kernel(#) h(# #) b(# #) rho(#) bwselect() scaleregul(#)
scalepar(#) vce(vctype[vceopt1 vceopt2]) level(#) all]
```

估计命令与带宽估计命令相似。

8.5 其他因果效应识别方法——匹配与合成控制法

8.5.1 匹配法

我们回忆一下在理想实验中，平均处理效应（ATE）等于服用新药之后的病情（ $E(Y_{1i}|D_i = 1)$ ）与没有服用新药病情（ $E(Y_{0i}|D_i = 0)$ ）之间的差异。即 $ATE = E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 0)$ 。

下面，我们来做一个简单的数学变换（不要看到“数学”两个字就怕，它们很多时候都是“纸老虎”，例如，此处只要学过中学的加减移项即可看懂）。我们将 ATE 计算式右边加上一项 $E(Y_{0i}|D_i = 1)$ ，又减去一项 $E(Y_{0i}|D_i = 1)$ ：

$$\begin{aligned} ATE &= E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 0) \\ &= [E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 1)] + [E(Y_{0i}|D_i = 1) - E(Y_{0i}|D_i = 0)] \end{aligned} \quad (8.14)$$

(10) 式显示，我们可以把 ATE 分解成两个部分（由中括号表示）：第一个部分是“参与者平均处理效应（ATT）”——项目实际参与者的平均处理效应；第二部分是选择偏差——假设参与者和未参与者的处理效应之差（实际上两类人都未参与项目）。

从政策制定者的角度来看，他们可能更关心 ATT，因为这是政策或项目实施后的毛收益，我们只需要将这个毛收益与政策成本进行比较，就能判定这个政策是否值得。但从 (10) 式可知，ATE 与 ATT 之间是存在一个

选择偏差。(需要注意的是, 这里的选择偏差与第五讲中的样本选择偏误有所不同。样本选择偏误是所选取的样本不是总体中的代表性样本所引起的偏误, 这时通常不考虑政策或项目效应。

那么, 如何解决这个问题呢? 我们前面已经提到过, 随机实验最大的优势在于其随机分配。那么, 完全随机选择处理与否自然可以消除上述选择偏差。但实践中, 很难执行这种随机实验。在这种情况下, 我们有两种方法可以尽量消除选择偏误:

1. 匹配估计量

使用条件: 假设个体根据可观测变量来选择是否参与项目

下面, 我们用一个就业培训项目为例。在对这个项目进行效应评估时, 我们除了能观测到人们是否参与了该项目 D_i 和项目实施前后的收入 Y_i , 我们还能观测到参与者的一些个体特征, 例如年龄、受教育程度、肤色、婚否、性别等——协变量。

如果个体是否参与项目完全由某些协变量 X 决定的, 那么, 我们就可以利用**匹配估计方法**来估计出处理效应。

匹配估计的思想其实非常简单: 实践中, 个体 i 参与培训了 (处理组), 他就不可能又“穿越”回到过去不参加培训。这个时候, 我们就需要在没有参加培训的那些人 (控制组) 找到某个或某些人 j , 那么怎么找呢? 上面说过, 参与项目 D_i 完全取决于可观测变量 X_i , 那么, 自然就是找那些与参与者 i 有相近 X 的未参与人 j 。我们选择到的 X_j 与 X_i 越接近, j 参与培训的概率就越接近 i 。那么, 我们就可以把 j 的收入 Y_j 近似当作 i 在没有参与培训情形下的收入, 然后把 i 的实际收入 Y_i 减去这个近似收入 Y_j , 即可得到培训的处理效应, 即匹配估计量。

我们来看看表 3 中的一个简单匹配例子, 参见陈强 (2014): 541 页。

表 8.4: 匹配估计

i	D_i	X_i	Y_i	匹配对象	\tilde{Y}_{0i}	\tilde{Y}_{1i}
1	0	2	7	5	7	8
2	0	4	8	4,6	8	7.5
3	0	5	6	4,6	6	7.5
4	1	3	9	1,2	7.5	9
5	1	2	8	1	7	8
6	1	3	6	1,2	7.5	6
7	1	1	5	1	7	5

根据匹配估计的思想, 是否参加项目 D 只依赖于 X , 而且我们要从未参加组 ($D=0$) 里为参加者 ($D=1$) 找到 X 相近的那些人, 例如, 我们要从 1-3 中为 4-7 找到 X 相近的人, 第 7 个人的 $X=1$, 而 1-3 中 X 最接近 1 的是第一个人的 $X=2$, 因此, 与第 7 个人匹配的对象就是第 1 个人。那么, 我们就应该把第 1 个人的 $Y=7$ 当做第 7 个人没有参加项时的 $\tilde{Y}=7$, 而实际参与项目的 $Y=5$, 因此, 该项目的效应就是 $5-7=-2$ 。其他人也这样匹配。

两个技术细节需要特别注意:

第一, 在寻找匹配对象时是否允许匹配对象放回。放回就是说, 当表 3 中的第 4 个人与第 2 个人匹配了, 但是在对第 6 个人进行匹配时, 第 2 个人仍然在备选之列, 也就是第 2 个人匹配之后又放回备选对象行列, 可以进行下一次匹配。而不放回, 就意味着 2 与 4 匹配之后, 2 就不能进行下次匹配, 那么 6 只能与 1 进行匹配。

第二, 是否允许匹配对象并列。也就是说, 4 与 1、2 的 X 都比较接近, 那么, 在允许并列的情形下, 我们会将 1、2 的 Y 的均值作为 4 的 \tilde{Y} 。如果不允许并列, 软件就会根据数据排列的顺序来选择匹配对象及其 \tilde{Y} , 例如, 在不允许并列时, 根据表 3 数据的排列顺序, 与 4 匹配的就是 1, 那么 Y_1 就会作为 4 的 \tilde{Y}_4 。

一般来说, 匹配估计量会存在偏差, 因为 X_i 不可能与 X_j 完全相同。那么, 在非精确匹配的情形下: (1) 一对一匹配, 偏差较大, 方差较小; (2) 一对多匹配, 偏差较小, 方差加大。**经验法则: 最好进行一对四匹配, 这样能使均方误差 (MSE) 最小。**

上面就是匹配估计法的最基本思想: 就是找到两组中 X 最接近的对象进行匹配。虽然原理很简单, 但是实际操作起来可就难了。因为在实践中, 我们通常不会像表 3 那样, D 依赖的 X 只有单一变量, 而是 X 中会包含很多个变量。也就是说, 我们要根据多个协变量同时进行比较, 例如对不同人的年龄、受教育程度、性别等同时

进行比较，这个时候就可能会遇到，两个人年龄相仿，但受教育程度差距很大，受教育程度相同，但年龄差距有很大，这个时候我们要这么比较这匹配对象是否接近呢？

这个时候，我们就需要拿出一种有效的武器——倾向得分匹配 (PSM)。

那么，PSM 的思想是什么呢？说简单也简单，说难也难！

简单是因为，我们就是要找到一个批判不同对象之间是否相似，但在多个 X 情况下，我们无从下手。那么，我们想法办把多个 X 转换成一个指标，即通过某种函数 $f(X)$ ，把多维变量变成一维变量，这个一维变量就是倾向得分 (PS)。然后，我们就可以根据这个倾向得分来进行上述匹配。

难是因为，这个转换函数 $f(X)$ 到底是什么？这个问题我们就不展开了。有兴趣者可自行查阅相关资料。

PSM 计算处理效应的步骤：

(1) 选择协变量 X 。尽量将影响 D 和 Y 的相关变量都包括在协变量中。如果协变量选择不当或太少，就会引起效应估计偏误；

(2) 计算倾向得分，一般用 logit 回归；

(3) 进行倾向得分匹配。如果倾向得分估计较为精确，那么， X 在匹配后的处理组和控制组之间均匀分布，这就是数据平衡。那么我们检验得分是否准确就需要计算 X 中每个分量的“标准化偏差”。经验法则：一般来说，标准化偏差不能超过 10%，如果超过 10%，我们就要重新返回第 (2) 步重新计算，甚至第 (1) 步重新选择匹配协变量，或者改变匹配方法。

(4) 根据匹配后的样本计算处理效应。

第三步中，得分匹配效果不好，可能要改变匹配方法：一、 k 邻近匹配；二、卡尺匹配或半径匹配；三、卡尺内最近邻匹配；四、核匹配；五、局部线性回归匹配；六、样条匹配。在实践中，并没有明确准则来限定使用哪种匹配方法。但有一些经验法则可作为参考：

(1) 如果控制组个体不多，则应该选择又放回匹配；

(2) 如果控制组有较多个体，则应该选择核匹配；

(3) 最常用的方法：尝试不同的匹配方法，然后比较它们的结果，结果相似说明很稳健。结果差异较大，就要去深挖其中的原因。

PSM 的局限性：

(1) 大样本；

(2) 要求处理组和控制组有较大的共同取值范围；

(3) 只控制了可观测的变量，如果存在不可观测的协变量，则会引起“隐性偏差”。

II. DID-PSM 估计量

使用条件：假设个体根据不可观测变量来选择是否参与项目

上面提到，如果存在根据不可观测变量进行选择时，会引起“隐性偏差”。而消除这种影响的方法很多，其中之一就是利用面板数据，且结合 DID-PSM 来计算处理效应。DID 和 PSM 原理我们在上面均详细讲过，因此，下面直接给出其 stata 操作。

除了 DID-PSM 之外，断点回归和工具变量法都可以尽量消除“隐性偏差”。

8.5.1.1 PSM 的 Stata 应用演示

下面，我们用 Dehejia and Wahba (1999) 职业培训的数据来演示 stata 的匹配操作。

从 stata13.0 开始，就提供匹配命令 `teffects` 命令，我们在 stata 中输入“`help teffects`”就可以看到命令描述：

下面，我们采用 1:1 最近邻匹配，估计培训对个人收入的效应。输入如下命令：

8.5.2 合成控制法

在上面的讲解中，我们反复多次强调，随机实验中的“随机性”最为关键，因为它可以很好地识别出我们所要进行比较的“处理组”和“控制组”。一旦“处理组”和“控制组”的 `outcomes` 被我们观测到，我们就可以利用处理组的结果减去控制组的结果得到我们感兴趣的处理效应，例如产业政策效应。

Title

[TE] **teffects** — Treatment-effects estimation for observational data

Syntax

teffects *subcommand* ... [, *options*]

<i>subcommand</i>	Description
aipw	augmented inverse-probability weighting
ipw	inverse-probability weighting
ipwra	inverse-probability-weighted regression adjustment
nnmatch	nearest-neighbor matching
overlap	overlap plots
psmatch	propensity-score matching
ra	regression adjustment

Description

teffects estimates potential-outcome means (POMs), average treatment effects (ATEs), and average treatment effects on the treated (ATTs). Regression-adjustment, inverse-probability-weighted, and matching estimators are provided, as are doubly robust estimators. **teffects** overlap plots the estimated densities of the probability of getting each treatment level. The outcomes can be continuous, binary, count, fractional, or nonnegative. The treatment model can be binary or continuous. For a brief description and example of each estimator, see [Remarks in **teffects** intro](#).

图 8.10: 匹配命令描述

但困难之处在于，实践中，我们往往只能观测到“处理组”的结果，而观测不到“如果处理组没有接受处理”的结果。这个时候，我们就需要去选择或“假象”一个或多个“控制组”（注：这里的假象，也是有理有据地假象。我们还有一个专业术语叫做“构造”）。

陈强老师说“选择控制组是一门艺术。确实，寻找适当的控制组（control group），即在各方面都与受干预地区相似却未受干预的其他地区，以作为处理组（treated group，即受到干预的地区）的反事实替身（counterfactuals）。但通常不易找到最理想的控制地区（control region），在各方面都接近于处理地区（treated region）。

比如，要考察仅在北京实施的某政策效果，自然会想到以上海作为控制地区；但上海毕竟与北京不完全相同。或可用其他一线城市（上海、广州、深圳）构成北京的控制组，比较上海、广州、深圳与北京在政策实施前后的差别，此方法也称“比较案例研究”（comparative case studies）。但如何选择控制组通常存在主观随意性（ambiguity），而上海、广州、深圳与北京的相似度也不尽相同。

因此，在上面一小节，我们简单介绍了匹配方法——通过能体现个体特征的协变量的相似度来构造出一个“控制地区”的结果。除了上述 PSM 方法之外，还有另一种构造“控制组”的方法——合成控制法（Synthetic Control Method）。

合成控制法是由 Abadie and Gardeazabal (2003) 提出来研究西班牙巴斯克地区（Basque country）恐怖活动的经济成本。其基本思想为：虽然无法找到巴斯克地区的最佳控制地区，但通常可对西班牙的若干大城市进行适当的线性组合，以构造一个更为贴切的“合成控制地区”（synthetic control region），并将“真实的巴斯克地区”与“合成的巴斯克地区”进行对比，故名“合成控制法”。合成控制法的一大优势是，可以根据数据（data-driven）来选择线性组合的最优权重，避免了研究者主观选择控制组的随意性。

Abadie et al.(2010) 用合成控制法研究了美国加州香烟控制 99 法案对加州香烟消费的影响。为了限制香烟消费，1988 年 11 月，加州政府通过了 99 法案，主要内容是将香烟消费税每包提高 25 美分，该法案 1989 年 1 月正式生效，作者主要考察这一政策对香烟消费的抑制作用到底有多大。由此，我们知道，香烟消费税在加州地区改变（提高）了，而在美国其他州并没有变化。因此，加州是“处理地区”，美国其他州是可能的“控制地区”。

此时，可能我们立马就能想起来，可以使用 PSM 呀。我们找出各个州的典型特征作为协变量，然后计算匹配得分，进行对比处理地区和匹配地区的香烟消费，得到香烟消费税提高的控烟效应。如果大家还记得上一节

最后我们给出的 PSM 方法局限性——要求大样本。大家可能就会意识到：哎呀，PSM 用在加州控烟税上可能有问题，因为这个样本中就只有加州和其他 40 多个州多年数据，这个样本似乎也不大。这种情况也是我们在写论文或者进行政策评估时经常会遇到的情况，所以用什么方法已经要深思熟虑。

我们接着来看加州控烟税的例子。为了避免其他州类似的政策对控制组产生影响，Abadie 等剔除了研究期内出台相似控烟政策的州，最后潜在控制组里剩下 38 个州。也就是说样本包括美国 39 各州，1970-2000 年的面板数据，变量包括：州年度人均香烟消费量、香烟平均零售价格、州人均收入对数、州人口结构（15-24 岁比例）、州人均啤酒消费量等等。

记 Y_{it} 为地区 i 在 t 期实际观测到的结果变量，即香烟消费量。

记 Y_{it}^I 的上标 I 表示地区 i 接受政策干预，即这个变量表示加州在提高香烟消费税后的香烟消费量。同理， Y_{it}^N 的上标 N 表示没有受到政策干预。根据上文将的处理效应，我们实际上感兴趣的是：

$$ATE = Y_{it}^I - Y_{it}^N$$

现在的问题是， Y_{it}^N 观测不到，因此，我们要估计出它。那么，怎么估计呢？即书本上的“因子模型”来估计 Y_{it}^N ：

$$Y_{it}^N = \delta_t + \theta_t Z_i + \lambda_t \mu_i + \epsilon_{it} \quad (8.15)$$

(11) 式右边第一项表示时间固定效应 (time fixed effects)。第二项表示可观测的向量（不受政策干预影响，也不随时间而变；比如，干预之前的预测变量之平均值），由于它对 Y 的效应随时间可变，因此，其系数 θ 有时间下标。第三项表示不可观测的“交互固定效应” (Interactive Fixed Effects)，即个体固定效应 μ_i 与时间固定效应 λ_t 的乘积 (Bai, 2009)。第 (4) 项为随机扰动项。

我们咋一看，这不就是面板回归吗？但是请注意，第三项“交互固定效应”是不可观测的变量。这一点很重要。

我们回忆一下合成控制法的基本思想：虽然无法找到加州地区的最佳控制地区，但通常可对美国其他的若干州进行适当的线性组合，以构造一个更为贴切的“合成控制地区” (synthetic control region)，并将“真实的加州地区”与“合成的加州地区”进行对比。也就是说，我们要利用其他州的线性组合来拟合出加州的 Y_{it}^N ，线性组合就是每个州的 Y 前面乘以一个权重 W 之和，假设每个州的权重用 w_j 表示，那么，我们将线性组合表示为

$$\sum_{j=1}^{j=N} w_j Y_{jt} = \delta_t + \theta_t \sum_{j=1}^{j=N} w_j Z_j + \lambda_t \sum_{j=1}^{j=N} w_j \mu_j + \sum_{j=1}^{j=N} w_j \epsilon_{jt} \quad (8.16)$$

接下来，我们用 (11) 式减去 (12) 式：

$$Y_{it}^N - \sum_{j=1}^{j=N} w_j Y_{jt} = \theta_t (Z_i - \sum_{j=1}^{j=N} w_j Z_j) + \lambda_t \sum_{j=1}^{j=N} (\mu_i - w_j \mu_j) + \sum_{j=1}^{j=N} (\epsilon_{it} w_j \epsilon_{jt}) \quad (8.17)$$

我们的目的是用其它 38 个州的线性组合 $\sum_{j=1}^{j=N} w_j Y_{jt}$ 来代替加州的 Y_{it}^N 。因此，我们想要 $Y_{it}^N - \sum_{j=1}^{j=N} w_j Y_{jt} = 0$ 。那么，我们只要使得 (13) 式右边的第一个括号为 0，第二个括号为 0，那么整个 (13) 式的期望就等于 0。此时，用 $\sum_{j=1}^{j=N} w_j Y_{jt}$ 合成的结果就是无偏的。但是，我们要注意， μ_i 是不可观测的变量，因此， $\lambda_t \sum_{j=1}^{j=N} (\mu_i - w_j \mu_j)$ 估计不出来，也即是说，上述估计行不通。

这怎么办呢？

我们仔细观察 (13) 式，里面可观测的变量有干预前的 Y_{it} 、 Y_{jt} 、 Z_i 、 Z_j ，那么，我们只需要找到最优的权重 w_j 使得：

$$Y_{it}^N - \sum_{j=1}^{j=N} w_j Y_{jt} = 0, Z_i - \sum_{j=1}^{j=N} w_j Z_j = 0$$

即根据可观测的经济特征与干预前结果变量所选择的合成控制 w ，也会使得合成控制的不可观测特征接近于处理地区。反之，如果无法找到 w ，使得合成控制能很好地复制 (reproduce) 处理地区的经济特征以及干预之前的结果变量，则不建议使用合成控制法。Abadie et al. (2010) 已经证明了，当干预前的时期数趋向于无穷，那么，合成控制估计量就是无偏的。

小贴士：合成控制法对样本数据的要求为政策干预以前需要很多期数据，有人认为至少需要 15 年的数据，而政策干预后需要有 5 年以上的数据。同时地区最好超过 10 个，但是又不会太多。最为重要的是，接受政策干预的地区个数极少。

8.5.2.1 例子及 stata 操作

我们来继续看看加州控烟税政策的效果。

第一步，我们在 stata 中输入下列命令：

`ssc install synth, replace` （下载并安装 synth 程序）

其中，选择项“replace”表示如有此命令更新版本，可以新命令覆盖旧命令。

命令 synth 的基本句型为：

`synth y x1 x2 x3, trunit(#) trperiod(#) counit(numlist) xperiod(numlist) mspeperiod() figure
resultperiod() nested allopt keep(filename)`

其中，“y”为结果变量（outcome variable），“x1 x2 x3”为预测变量（predictors）。

必选项“trunit(#)”用于指定处理地区（trunit 表示 treated unit）。

必选项“trperiod(#)”用于指定政策干预开始的时期（trperiod 表示 treated period）。

选择项“counit(numlist)”用于指定潜在的控制地区（即 donor pool，其中 counit 表示 control units），默认为数据集中的除处理地区以外的所有地区。

选择项“xperiod(numlist)”用于指定将预测变量（predictors）进行平均的期间，默认为政策干预开始之前的所有时期（the entire pre-intervention period）。

选择项“mspeperiod()”用于指定最小化均方预测误差（MSPE）的时期，默认为政策干预开始之前的所有时期。

选择项“figure”表示将处理地区与合成控制的结果变量画时间趋势图，而选择项“resultperiod()”用于指定此图的时间范围（默认为整个样本期间）。

选择项“nested”表示使用嵌套的数值方法寻找最优的合成控制（推荐使用此选项），这比默认方法更费时间，但可能更精确。在使用选择项“nested”时，如果再加上选择项“allopt”（即“nested allopt”），则比单独使用“nested”还要费时间，但精确度可能更高。

选择项“keep(filename)”将估计结果（比如，合成控制的权重、结果变量）存为另一 Stata 数据集（filename.dta），以便进行后续计算。更多选择项，详见 help synth。

第二步，打开数据集之后，输入下列命令：

`xtset state year`（声明面板数据）

第三步，在 stata 中输入下列合成控制法估计命令：

`synth cigsale retprice lnincome age15to24 beer cigsale(1975) cigsale(1980) cigsale(1988),
trunit(3) trperiod(1989) xperiod(1980(1)1988) figure nested keep(smoking_synth)`

估计结果如下：

图 6 显示，大多数州的权重为 0，而只有以下五个州的权重为正，即 Colorado (0.161)，Connecticut (0.068)，Montana (0.201)，Nevada (0.235) 与 Utah (0.335)，此结果与 Abadie et al. (2010) 汇报的结果非常接近（细微差别或由于计算误差）。

下图显示了加州和其他 38 个州的合成结果：

从图 7 可知，加州与合成加州的预测变量均十分接近，故合成加州可以很好地复制加州的经济特征。然后比较二者的人均香烟消费量在 1989 年前后的表现：

从上图可知，在 1989 年控烟法之前，合成加州的人均香烟消费与真实加州几乎如影相随，表明合成加州可以很好地作为加州如未控烟的反事实替身。在控烟法实施之后，加州与合成加州的人均香烟消费量即开始分岔，而且此效应越来越大。

更直观地，可打开另一 Stata 程序，调用已存的数据集 smoking_synth.dta，计算加州与合成加州人均香烟消费之差（即处理效应），然后画图。

州 [↗]	权重 [↗]	州 [↗]	权重 [↗]
Alabama [↗]	0 [↗]	Nevada [↗]	.235 [↗]
Arkansas [↗]	0 [↗]	New Hampshire [↗]	0 [↗]
Colorado [↗]	.161 [↗]	New Mexico [↗]	0 [↗]
Connecticut [↗]	.068 [↗]	North Carolina [↗]	0 [↗]
Delaware [↗]	0 [↗]	North Dakota [↗]	0 [↗]
Georgia [↗]	0 [↗]	Ohio [↗]	0 [↗]
Idaho [↗]	0 [↗]	Oklahoma [↗]	0 [↗]
Illinois [↗]	0 [↗]	Pennsylvania [↗]	0 [↗]
Indiana [↗]	0 [↗]	Rhode Island [↗]	0 [↗]
Iowa [↗]	0 [↗]	South Carolina [↗]	0 [↗]
Kansas [↗]	0 [↗]	South Dakota [↗]	0 [↗]
Kentucky [↗]	0 [↗]	Tennessee [↗]	0 [↗]
Louisiana [↗]	0 [↗]	Texas [↗]	0 [↗]
Maine [↗]	0 [↗]	Utah [↗]	.335 [↗]
Minnesota [↗]	0 [↗]	Vermont [↗]	0 [↗]
Mississippi [↗]	0 [↗]	Virginia [↗]	0 [↗]
Missouri [↗]	0 [↗]	West Virginia [↗]	0 [↗]
Montana [↗]	.201 [↗]	Wisconsin [↗]	0 [↗]

	Treated	Synthetic
retprice	89.42222	89.41464
lnincome	10.07656	9.858694
age15to24	.1735324	.1735444
beer	24.28	24.21326
cigsale(1975)	127.1	127.0633
cigsale(1980)	120.2	120.4545
cigsale(1988)	90.1	91.6356

图 8.12: 合成结果比较

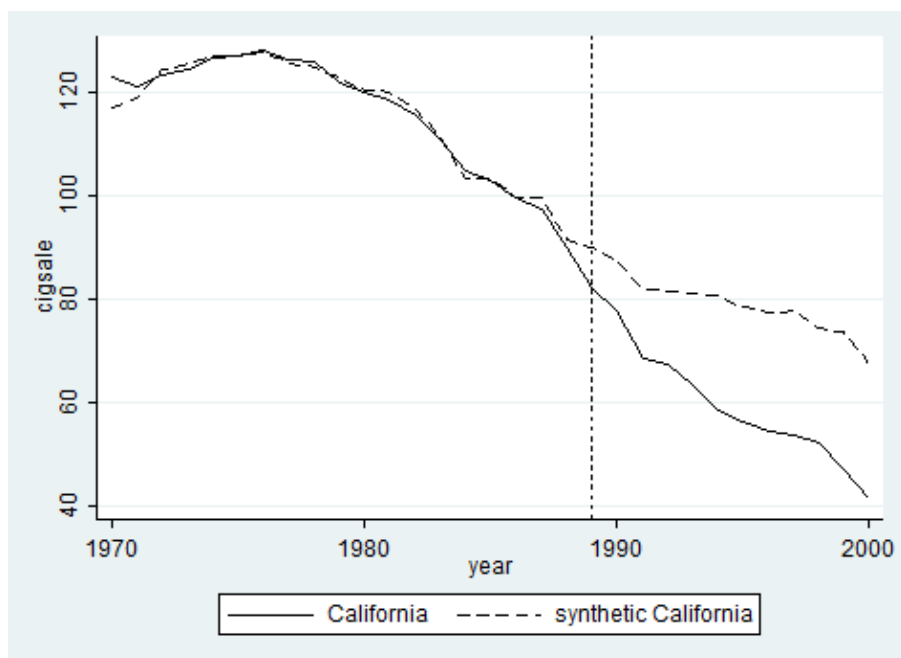


图 8.13: 合成控制结果


```
use smoking_synth.dta, clear
```

(如不打开另一 Stata 程序, 则此数据集将覆盖原有的数据集 smoking.dta)

```
gen effect = _Y_treated - _Y_synthetic
```

(定义处理效应为变量 effect, 其中“_Y_treated”与“_Y_synthetic”分别表示处理地区与合成控制的结果变量)

```
label variable _time "year"
```

```
label variable effect "gap in per-capita cigarette sales(in packs)"
```

(为了画图更漂亮, 加上时间变量与处理效应的标签, 可使用变量管理器 (variable manager) 来方便地加标签)

```
line effect _time, xline(1989, lp(dash)) yline(0, lp(dash))
```

(画处理效应的时间趋势图, 并在横轴 1989 年处与纵轴 0 处分别画虚线, 结果见下图)

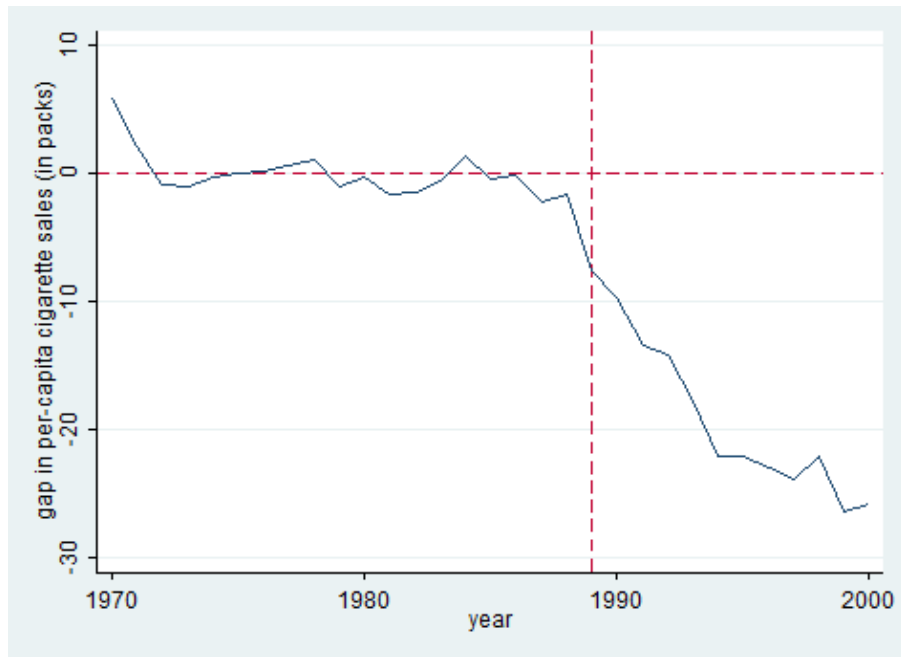


图 8.14: 趋势图

图 9 显示, 加州控烟法对于人均香烟消费量有很大的负效应, 而且此效应随着时间推移而变大。具体来说, 在 1989-2000 年期间, 加州的人均年香烟消费减少了 20 多包, 大约下降了 25% 之多, 故其经济效应十分显著 (economically significant)。

到此, 我们关心的合成控制法主要结果都呈现出来了。但是在使用合成控制法时, 如何进行稳健性检验与统计推断?

为了检验上述合成控制估计结果的稳健性, Abadie et al. (2010) 加入了更多的预测变量, 比如失业率、收入不平等、贫困率、福利转移、犯罪率、毒品相关的逮捕率、香烟税、人口密度等; 发现结果依然稳健。

另外一个担心是, 地区之间无互相影响 (no interference between units) 的假定可能不满足, 比如加州的反烟运动可能波及其他州, 烟草行业或将其他州的香烟广告预算投入到加州, 甚至从其他州走私便宜香烟到加州。Abadie et al. (2010) 根据史实对此进行了探讨, 认为这些效应均不大, 至少不可能导致上文图中如此大的处理效应。

安慰剂检验

上述结果为对控烟法处理效应的点估计。此点估计是否在统计上显著 (statistically significant)? Abadie et al. (2010) 认为, 在比较案例研究中, 由于潜在的控制地区数目通常并不多, 故不适合使用大样本理论进行统计推断。

为此, Abadie et al. (2010) 提出使用“安慰剂检验” (placebo test) 来进行统计检验, 这种方法类似于统计学中的“排列检验” (permutation test), 适用于任何样本容量。

“安慰剂”(placebo)一词来自医学上的随机实验,比如要检验某种新药的疗效。此时,可将参加实验的人群随机分为两组,其中一组为实验组,服用真药;而另一组为控制组,服用安慰剂(比如,无用的糖丸),并且不让参与者知道自己服用的究竟是真药还是安慰剂,以避免由于主观心理作用而影响实验效果,称为“安慰剂效应”(placebo effect)。

安慰剂检验借用了安慰剂的思想。具体到加州控烟法的案例,我们想知道,使用上述合成控制法所估计的控烟效应,是否完全由偶然因素所驱动?换言之,如果从 donor pool 随机抽取一个州(而不是加州)进行合成控制估计,能否得到类似的效应?

为此,Abadie et al. (2010) 进行了一系列的安慰剂检验,依次将 donor pool 中的每个州作为假想的处理地区(假设也在 1988 年通过控烟法),而将加州作为控制地区对待,然后使用合成控制法估计其“控烟效应”,也称为“安慰剂效应”。通过这一系列的安慰剂检验,即可得到安慰剂效应的分布,并将加州的处理效应与之对比。

在此有个技术细节,即在对某个州进行安慰剂检验时,如果在“干预之前”其合成控制的拟合效果很差(均方预测误差 MSPE 很大),则有可能出现在“干预之后”的“效应”波动也很大,故结果不可信。类似地,如果合成加州在干预前对于加州的拟合很差,则我们也不会相信干预之后的合成控制估计结果。

注意事项

在使用合成控制法时,需要特别注意以下几点:

1、我们是将没有实行政策的地区作为备选的合成控制组,但是如果 i 地区实施政策对 j 地区也产生了影响,那么,这个时候,我们就应该将 j 地区从备选控制组中提出掉,例如加州提高控烟税,对威斯康辛州有很大影响,那么,我们就应该将威斯康辛州从 38 个备选里去掉;

2、如果在研究期间,有一些地区受到非常大的特殊冲击,那么,这时候我们也要将其剔除;

3、尽量使得控制地区与处理地区具有相似的特征;

4、合成控制法对样本数据的要求为政策干预以前需要很多期数据,有人认为至少需要 15 年的数据,而政策干预后需要有 5 年以上的数据。同时地区最好超过 10 个,但是又不会太多。最为重要的是,接受政策干预的地区个数极少。

5、如果政策冲击的效应需要一段时间才会显现(滞后效应),则也要求干预后的期数足够大。

8.6 克服计量方法选择困难症

DID 与 PSM

我曾经看过一个最简单的描述:

DID 是比较四个点, Treated before, treated after; control before, control after;

Matching 是比较两个点: Treated, control;

DID + Matching 是用 matching 的方法来确定 treated 和 control。

断点回归

1、断点回归最大的优势就是在断点附近接近随机实验,也就是说,断点回归可以认为是一种“局部随机实验”;

2、但是,正因为断点回归只在断点附近随机性强,因此,仅能推断断点处的因果关系,不一定能推广到其他样本,也就是说结论在理论上的一般性存疑。

合成控制法与 DID

首先,根据 Abadie et al. (2010) 的因子模型(factor model),合成控制法对双向固定效应模型作了推广。具体来说,双重差分法仅允许个体固定效应与个体时间效应以相加(additive)的形式存在,隐含假设所有个体的时间趋势都相同(parallel trend assumption);而合成控制法的因子模型,则允许“互动固定效应”(interactive fixed effects),即可以存在多维的共同冲击(common shocks),而每位个体对于共同冲击的反应(factor loading)可以不同,故允许不同个体有不同的时间趋势。

其次,Abadie et al. (2015) 指出,回归法也可以视为对控制地区作了线性组合,且权重之和也为 1;而不同之处在于,合成控制法的权重必须非负,但回归法的权重可能出现负值,即出现过分外推(extrapolation)而离开

了样本数据的取值范围 (**support of the data**)。比如, 在跨国研究中, 将很不相同的国家放在一起进行回归, 就可能出现过分外推, 而导致“外推偏差” (**extrapolation bias**)。由于合成控制法的权重必须非负, 故避免了过分外推。

第 9 章 VARs

第 10 章 DSGE 及 Stata 操作

动态随机一般均衡 (DSGE) 模型是近 40 年来宏观经济学的最大进展。现在,它已经成为宏观经济和政策研究的中流砥柱,并且该方法也已经成为宏观经济学家之间交流思想观点的最重要方式 (Kehoe et al., 2018; Solis-Garcia, 2018)。作为回应 Lucas 批判 (Lucas, 1976) 的结构宏观计量模型,DSGE 模型不仅在宏观经济学专业领域广泛传播,同时,也被各国政府和国际机构用于政策分析和预测工具 (Del Negro and Schorfheide, 2013; Cai et al., 2018)。

但是,正如 JFV et al., (2016) 指出:“对于博士研究生来说不幸、对长期研究 DSGE 的研究人来说幸运的是,DSGE 模型的门槛非常高。”如果 DSGE 对于博士研究生来说门槛都高,那么,对于本科生和硕士研究生来说鸿沟更是巨大。本讲最大的目的就在于尽量降低 DSGE 的门槛,从而让研究生,甚至高年级本科生都能对 DSGE 的理论与经验含义有直观上的感知与认识。

一方面,主流中级水平的宏观经济学教科书——Blanchard (2017) 的《Macroeconomics (7th edition)》和 Mankiw (2015) 的《Macroeconomics (9th edition)》——已经用专门章节来阐述 DSGE 的基本结构和经济含义,他们均从传统的静态 IS-LM-PC 框架扩展到动态 IS, NKPC 和泰勒规则货币政策的分析框架。另一方面,Stata 15 推出了 DSGE 模块,国内大部分高校均开设了以 Stata 软件为工具的计量经济学课程,国内学者和学生 Stata 软件较为熟悉,因此,基于 Stata 软件来实现 DSGE 模型的定量政策分析与预测教学也具有事半功倍的作用。

10.1 静态 IS-LM-PC 模型

10.2 三方程 DSGE 模型

10.3 Stata 命令

第 11 章 极简 CGE：一个教学式模型

本讲稿基于 Don Fullerton and Chi L. Ta(2017,NBER):"Public Finance in a Nutshell: A Cobb Douglas Teaching Tool for General Equilibrium Tax Incidence and Excess Burden"。

11.1 导论

在经济学本科阶段，老师教授给学生的大部分经济学理论是基于单一均衡价格和数量的单一市场供需曲线图，即用局部均衡来分析经济问题，而其它市场则假定不受影响。在一学期末尾，老师稍微涉及了一点一般均衡的内容，然后复杂性又让学生们望而生畏。因此，年轻的学生们重一开始就是被吓着了——一般均衡太难，还是算了吧！那就更不用谈，构建可计算一般均衡（CGE）模型来应用于现实经济问题了。

但是，许多经济问题必须要使用一般均衡模型讨论和解释。例如，“营改增”的受益分布必须要知道其对其它市场价格的影响——资本回报、工资率以及其它产品价格的影响等等。有人可能会疑惑，前九讲的回归分析也可以用来研究“营改增”的效应呀（其实，我也用回归分析写了几篇“营改增”效应的文章了，发出来的时候可以用来作为具体例子讲解）。但是，我还是想提醒，在某些情形下，用一般均衡模型得到的“营改增”效应可能与局部均衡模型结果不一致，甚至相反。

本讲就来描述一个极简化的可计算一般均衡（CGE）模型，简化到可以用手、计算器或者 Excel 来计算结果。本讲与第十讲 DSGE 一样，并不是为了与回归模型进行比较优劣，而是为了丰富经验计量分析的方法，从而呈现出经济学中多样化的经验研究。

11.2 极简 CGE：McLure and Thirsk(1975) 模型

企业所得税可能会直接降低企业的投资回报，从而使得资本和劳动重新配置，这又会影响非企业回报、工资和产品价格等。为了研究企业所得税的受益归宿，Harberger(1962) 构建了一个封闭经济模型，其中企业生产一种产品 X，非企业部门生产另一种产品 Y。该模型包括两种要素：资本和劳动。而且假设固定要素供给，完全就业，规模报酬不变和其它完美市场假设。每一种产品的生产函数为

$$X = F(K_x, L_x)$$

完全竞争市场假设意味着利润为 0，但企业所得税应用于 X 企业所有者的资本收益，因此，Harberger 将企业所得税建模为 τ_{kx} ，且对资本 K_x 征收。

McLure and Thirsk(1975) 模型则使用了 CD 生产函数。该模型对于研究任何一个两部门（房地产部门与制造业、农业与制造业等）的税收政策效应非常有用。它也可以用于研究任何两种投入（污染型投入和非污染型投入）的税收变化的效应。例如，CD 生产函数为

$$X = AK_x^\alpha L_x^{1-\alpha}, \quad Y = BK_y^\beta L_y^{1-\beta} \quad (11.1)$$

其中，A, B 分别表示两个部门的全要素生产率，参数 $\alpha = 0.6, \beta = 0.2$ 分别表示两个部门资本收入份额。K 和 L 表示两部门的要素投入。

由于我们假设要素供给恒定，因此，家庭没有储蓄和劳动供给决策。整个经济的资源约束为：

$$K_x + K_y = K, \quad L_x + L_y = L \quad (11.2)$$

而家庭的效用函数为

$$U = X^\gamma Y^{1-\gamma} \quad (11.3)$$

其中，参数 $\gamma = 0.5$ 。家庭的预算约束为

$$P_x X + P_y Y = I \quad (11.4)$$

其中, P_X, P_Y 表示产品 X, Y 的价格, 而 I 表示家庭的收入。家庭在预算约束下选择产品 X, Y 来实现效用最大化, 从而得到产品的需求函数

$$X = \frac{\gamma I}{P_X}, Y = \frac{(1 - \gamma)I}{P_Y} \quad (11.5)$$

如果忘记了, 大家可以拿起初级微观来看看消费者理论的内容。

在大部分一般均衡模型中, 我们都会假定一种产品的价格固定不变, 而解出其它产品的相对价格。但是 McLure and Thirsk(1975) 假设名义总收入固定来锚定价格水平, 即 $I = \hat{I}$ 。因此, 如果对产品 X 征税, 那么, X 的价格 P_X 会上升, 然后, 两种产品的均衡数量会发生变化, Y 的价格 P_Y 会下降, 以至于总收入 $I = P_X X + P_Y Y$ 保持不变。在他们的模型中, $\hat{I} = 2400$, 因此, 我们可以计算得到家庭对产品 X 和 Y 的消费支出额:

$$P_X X = \gamma I = 0.5 \times 2400 = 1200 \quad (11.6)$$

$$P_Y Y = (1 - \gamma)I = (1 - 0.5) \times 2400 = 1200 \quad (11.7)$$

下面, 我们来看看家庭的收入来源。在上面的模型经济中, 家庭的收入来源于资本所得、劳动收入和政府的转移支付 (例如养老金等, 此处假设政府将全部收入转移给家庭)。那么, 家庭的收入为

$$I = RK + WL + T \quad (11.8)$$

其中, R 表示资本利率, W 表示工资率, T 表示政府税收 (转移支付)。

政府对产品 X 征税, 可以表示为 $(1 - \tau_x)P_X X$, τ_x 表示消费者消费产品 X 的一种价外税。同理, 如果我们要建模企业 X 的企业所得税, $(1 - \tau_{k,x})R_X K_X$ 。其它的税收政策也可以类似建模。

企业决定投入要素 K, L 来最大化其利润, 约束为生产函数。由此我们可以得到要素需求函数:

$$(1 - \tau_{k,x})R_X = \alpha \frac{(1 - \tau_x)P_X X}{K_X}, R_Y = \beta \frac{P_Y Y}{K_Y} \quad (11.9)$$

$$W_X = (1 - \alpha) \frac{(1 - \tau_x)P_X X}{L_X}, W_Y = (1 - \beta) \frac{P_Y Y}{L_Y} \quad (11.10)$$

这个时候, 我们就可以来分析税收 (商品税和企业所得税) 变化的效应了。如果我们只想要分析企业所得税的效应, 可以设置 $\tau_x = 0$, 其它政策同理。

在初始均衡中, 所有的税收都设置为 0。McLure and Thirsk(1975) 用 1 单位来定义每一种产品或要素作为一个单位成本, 那么, 初始的价格 $P_K^0 = P_L^0 = P_X^0 = P_Y^0 = 1$, 上标“0”表示初始均衡状态, 且没有税收。那么, 如方程 (6)、(7) 所示, 两种产品的支出额为 $P_X X = P_Y Y = 1200$, 因此, 两种产品的初始数量为 $X^0 = Y^0 = 1200$ 。根据 (9) 和 (10) 可以计算得到,

$$R_X^0 K_X^0 = \alpha P_X^0 X^0 = 0.6 \times 1200 = 720 \quad (11.11)$$

$$W_X^0 L_X^0 = (1 - \alpha) P_X^0 X^0 = 0.4 \times 1200 = 480 \quad (11.12)$$

$$R_Y^0 K_Y^0 = \beta P_Y^0 Y^0 = 0.2 \times 1200 = 240 \quad (11.13)$$

$$W_Y^0 L_Y^0 = (1 - \beta) P_Y^0 Y^0 = 0.8 \times 1200 = 960 \quad (11.14)$$

在 $R_K^0 = W_L^0 = 1$ 条件下, 这些方程可以计算得到完整的初始数量, 如表 1 所示:

表 11.1: 初始均衡配置

$L_X^0 = 480$	$L_Y^0 = 960$	$\hat{L} = 1440$
$K_X^0 = 72$	$K_Y^0 = 240$	$\hat{K} = 960$
$X^0 = 1200$	$Y^0 = 1200$	$\hat{I} = 2400$

也就是说，对于 CD 生产函数和上述参数，计算得到的初始均衡条件为，总的劳动禀赋必须为 1440，总的资本禀赋必须为 960。

11.3 税收政策的一般均衡效应：CGE 应用

上一节中的式（9）和（10）可以用来分析政策的一般均衡效应。

11.3.1 企业所得税

本节以企业所得税为例，例如，设置产品税 $\tau_x = 0$ ，而资本所得税 $\tau_{k,x}$ 为任何给定的税率。那么，式（9）之和为：

$$(1 + \tau_{k,x})R_x K_x + R_y K_y = \alpha(1 - \tau_x)P_x X + \beta P_y Y = (\alpha\gamma + \beta(1 - \gamma))I \quad (11.15)$$

根据完全竞争市场， $R_x = R_y = R$ ，且在均衡中，资本市场出清 $K_x + K_y = K$ 。那么，（15）式可以转换为

$$R\hat{K} = con - \tau_{k,x}RK_x \quad (11.16)$$

其中， $con = (\alpha\gamma + \beta(1 - \gamma))I$ 表示为常数。这就意味着企业所得税的任何变化都会降低资本收入，而且变化幅度为 $\Delta\tau_{k,x}RK_x$ 。也就是说资本所得税率的变化所带来的负担完全由资本所有者承担了。同理，式（10）之和为：

$$W_x L_x + W_y L_y = (1 - \alpha)P_x X + (1 - \beta)P_y Y = ((1 - \alpha)\gamma + (1 - \beta)(1 - \gamma))I \quad (11.17)$$

根据完全竞争市场， $W_x = W_y = W$ ，且在均衡中，资本市场出清 $L_x + L_y = L$ 。那么，（17）式可以转换为

$$W\hat{L} = con_L \quad (11.18)$$

其中， $con_L = ((1 - \alpha)\gamma + (1 - \beta)(1 - \gamma))I$ 表示为常数。由（18）式可以看出，劳动者完全不承担任何企业所得税负担。

上面，我们分析了企业所得税变化对要素市场的影响。下面，我们来看看企业所得税变化对产品市场的影响。

从家庭的产品需求函数（5）可知， $P_x X = \gamma I$ ， $P_y Y = (1 - \gamma)I$ ，也就是说，家庭对两种产品的需求是其收入的一定比例。而根据（8）式，家庭的收入为：

$$I = RK - \tau_{k,x}RK_x + WL + \tau_{k,x}RK_x \quad (11.19)$$

（19）式等号右边的第一项为税收资本所得，第二项为劳动所得，第三项为政府转移支付。由此可见，家庭的收入 $I = R\hat{K} + W\hat{L}$ 。这就意味着，家庭的收入并不受企业所得税变化的影响，因此，企业所得税变化也不影响家庭对两种产品的需求。

11.3.2 商品税/增值税

下面，我们分析一下对 X 产品征税的一般均衡效应。假设 $\tau_x = 0.30$ ， $\tau_{k,x} = 0$ 。所有的初始条件都不变。如（19）式，产品的支出只与收入有关，因此，产品 X 的总支出仍然为 $P'_x X' = \gamma I = 1200$ ，但是，此时，而企业生产 X 支付的商品税为 $\tau_x P'_x X' = 360$ 。而方程（9）和（10）对于任何商品税率都成立，因此，

$$R'K'_x = \alpha P'_x X' (1 - \tau_x) = 0.6 \times 1200 \times 0.70 = 504, R'K'_y = 240 \quad (11.20)$$

$$W'L'_x = (1 - \alpha)P'_x X' (1 - \tau_x) = 0.4 \times 1200 \times 0.70 = 336, W'L'_y = 960 \quad (11.21)$$

由此可得，企业支付的总的资本收入为 $R'\hat{K} = 504 + 240 = 744$ ，与征收商品税之前的资本收入 $R\hat{K} = 960$ 相比下降幅度为 216，也就是说，资本所有者承担的商品税负担为 $\frac{216}{960} \times 100\% = 22.5\%$ 。同理，企业支付的总的劳动收入为 $336 + 960 = 1296$ ，与征税前劳动总收入 1440 相比下降了 144，也就是说，劳动者承担的商品税负担为 $\frac{144}{1440} \times 100\% = 10\%$ 。

接下来，我们可以用上述结果来计算资本利率和工资率的变化，也就是说商品税对要素价格的扭曲程度。

$$R' \hat{K} = 744$$

$$\hat{K} = 960$$

$$R' = \frac{744}{960} = 0.775$$

这就意味着商品税造成了资本价格下降了 $\frac{1-0.775}{1} \times 100\% = 22.5\%$ 。

同理，对劳动工资的扭曲程度为：

$$W' \hat{L} = 1296$$

$$\hat{L} = 1440$$

$$W' = \frac{1296}{1440} = 0.90$$

这就意味着商品税造成了资本价格下降了 $\frac{1-0.90}{1} \times 100\% = 10\%$ 。

那么，为什么商品税对资本所有者的负担更重呢？

这是因为商品税对产品 X 征收，而企业 X 是资本密集型产业。从 CD 生产函数来看，企业 X 的生产函数中资本收入占比达到 0.6，而企业 Y 的资本收入占比只有 0.2，因此，对 X 征收产品税会降低总资本的需求，由此对资本造成的税负更重。

下面，我们来看看商品税对资源配置造成的影响。

$$K'_x = \frac{504}{0.775} = 650.32, K'_y = \frac{240}{0.775} = 309.68$$

$$L'_x = \frac{336}{0.9} = 373.33, L'_y = \frac{960}{0.9} = 1066.67$$

注意：两个部门的资本总额还是 960，劳动总额还是 1440。因为要素市场是完全竞争的，对产品 X 征收 30% 的税会使得资本和劳动从部门 X 转移至部门 Y，直到两个部门的要素回报相同。因为 X 部门是资本密集型，资本价格 R 的下降，会使得 Y 部门来租赁所有的剩余资本。

下面，我们来计算税收对产出的影响。根据表 1 中的初始均衡配置，我们可以计算得到全要素生产率参数 A、B。也就是说

$$A = \frac{X^0}{(K_x^0)^{0.6}(L_x^0)^{0.4}} = 1.96013$$

$$B = \frac{Y^0}{(K_y^0)^{0.2}(L_y^0)^{0.8}} = 1.64938$$

在上文中，我们已经计算得到征收商品税后，要素的重新配置情况。因此，可以计算得到两个部门新的产出量

$$X' = A(K'_x)^{0.6}(L'_x)^{0.4} = 1.96013 \times 650.32^{0.6} \times 373.33^{0.4} = 1020.94$$

$$Y' = A(K'_y)^{0.2}(L'_y)^{0.8} = 1.64938 \times 309.68^{0.2} \times 1066.67^{0.8} = 1373.81$$

利用产出信息，可以计算得到产品价格

$$P'_x = \frac{1200}{X'} = 1.17538$$

$$P'_y = \frac{1200}{Y'} = 0.87348$$

表 2 中呈现了所有关键变量的均衡价格和数量。初始均衡在第 (3) 列，商品税 $\tau_x = 0.30$ 在第四列，第五列为更高的商品税率 $\tau_x = 0.31$ 。

表 11.2: 关键变量的均衡值

(1) 变量	(2) 定义	(3) $\tau_x = 0$	(4) $\tau_x = 0.3$	(5) $\tau_x = 0.31$
面板 A: 配置与价格				
K_x	企业 X 投入的资本	720	650.323	674.296
L_x	企业 X 投入的劳动	480	373.333	369.368
K_y	企业 Y 投入的资本	240	309.677	312.704
L_y	企业 Y 投入的劳动	960	1066.67	1070.63
X	X 产量	1200	1020.94	1013.75
Y	Y 产量	1200	1373.81	1380.58
R	资本利率	1	0.775	0.7675
W	工资率	1	0.9	0.89667
P_x	产品 X 的价格	1	1.17538	1.18372
P_y	产品 Y 的价格	1	0.87348	0.8692
面板 B: 福利指标				
\hat{P}	总的价格水平	2	2.0265	2.02869
U	效用水平	1200	1184.306	1183.03
EB	超额负担	0	31.387	33.934
RE	转移支付	0	360	372
AER	平均超额负担 EB/RE		0.08719	0.09124
MEB	边际超额负担			0.21271

11.3.3 福利效应

下面，我们来计算商品税的福利效应，用支出函数来获得超额负担。

首先，产品的需求函数为 $X = \frac{\gamma I}{P_x}$, $Y = \frac{(1-\gamma)I}{P_y}$ ，将它们带入家庭效用函数 $U = X^\gamma Y^{1-\gamma}$ 得到间接效用函数：

$$V(P_x, P_y, I) = \left(\frac{\gamma I}{P_x} \right)^\gamma \left[\frac{(1-\gamma)I}{P_y} \right]^{1-\gamma} = \frac{I^\gamma I^{1-\gamma}}{\left(\frac{P_x}{\gamma} \right)^\gamma \left(\frac{P_y}{1-\gamma} \right)^{1-\gamma}} = \frac{I}{\hat{P}} \quad (11.22)$$

其中， $\hat{P} = \left(\frac{P_x}{\gamma} \right)^\gamma \left(\frac{P_y}{1-\gamma} \right)^{1-\gamma}$ 是“理想”价格指标，由两种产品的价格合成得到。变形间接效用函数得到支出函数：

$$I = E(\hat{P}, U) = U \times \hat{P} \quad (11.23)$$

注：更复杂的 CGE 模型可以参看“Introduction to Computable General Equilibrium Models”，Mary E. Burfisher(2011)，或者直接使用“量化经济分析平台”的全球模型和国家模型来操作，请参见《平台使用说明书》。

第 12 章 如何讲好经济学故事：经济学论文的流行形式

12.1 好的想法

12.2 标题的流行形式

12.3 文献综述

12.4 经验特征或待检验命题

12.4.1 逻辑推演式

12.4.2 数理模型式

12.5 实证分析

12.6 机制/反事实的数理模拟

12.7 结论与政策含义

12.8 引言

12.9 摘要

12.10 参考文献

12.11 附录

附录 A 基本数学工具

本附录包括了计量经济学中用到的一些基本数学，我们扼要论述了求和算子的各种性质，研究了线性和某些非线性方程的性质，并复习了比例和百分数。我们还介绍了一些在应用计量经济学中常见的特殊函数，包括二次函数和自然对数，前 4 节只要求基本的代数技巧，第 5 节则对微分学进行了简要回顾；虽然要理解本书的大部分内容，微积分并非必需，但在一些章末附录和第 3 篇某些高深专题中，我们还是用到了微积分。

A.1 求和算子与描述统计量

求和算子是用以表达多个数求和运算的一个缩略符号，它在统计学和计量经济学分析中扮演着重要作用。如果 $\{x_i : i = 1, 2, \dots, n\}$ 表示 n 个数的一個序列，那么我们就把这 n 个数的和写为：

$$\sum_{i=1}^n x_i \equiv x_1 + x_2 + \cdots + x_n \quad (\text{A.1})$$