北京大學
PEKING UNIVERSITY

Notes for *Intermediate Econometrics*

Lectured by Dandan Zhang Noted by J.Y. Chang

2022 Spring

Introduction

- 计量经济学探讨用统计技术研究经济数据的方法，目的是通过对微观数据的分析寻找经济学问题的因果解答。本课程的教学内容包括经典普通最小二乘法、多元线性回归模型、工具变量方法、和面板数据分析方法等。本课注重培养利用统计软件（STATA）实际操作微观数据并分析现实经济问题的能力。通过这门课的学习，学生们将初步掌握如何进行实证研究，以及如何评判实证研究中存在的问题。
- 通过这门课的学习，学生们将能够解释并使用数学公式表达计量模型，阐述计量分析中可能存在的问题，使用计量软件 STATA 分析微观数据，解释回归结果并进行必要的统计检验。

Contents

1. Introduction (We7 Chapter 1)
2. Probability Theory, Estimation and Hypothesis Testing (We7 Appendix B&C)
3. The Bivariate Linear Regression Model (We7 Chapter 2)
4. The Multivariate Linear Regression Model (We7 Chapter 3)
5. Inference (We7 Chapter 4)
6. Further Issues (We7 Chapter 6)
7. Dummy Varables (We7 Chapter 7)
8. Heteroscedasticity (We7 Chapter 8)
9. More on Specification and Data Issues (We7 Chapter 9)
10. Instrument Variables (We7 Chapter 15)
11. Pooled Cross-Sectional and Panel Data (We7 Chapter 14)

Reference

- *Introductory Econometrics: A Modern Approach*, Jeffery M. Wooldridge, 7th Edition, 2019.
- *Introduction to Econometrics*, James H. Stock and Mark W. Watson, 4th Edition, 2019.

1. Introduction

2022.02.26

---

1. What is Econometrics?

- **理论分析 [*Theoretical analysis*] v.s. 经验分析 [*Empirical analysis*]**:
  - 经济学家利用理论分析得到了变量之间的关系，如微观经济学第一个理论模型：消费者选择理论。计量经济学家利用数据验证假设、并计算因果效应[*Causal effects*]的量级。
  - 经济学理论用于研究每个变量是否有影响；计量经济学研究影响的方向与规模。
- **计量经济学的应用**
  - 利用时间序列预测未来走向：失业率、经济增长率等
  - 评估政策干预的影响
  - 研究经济关系

---

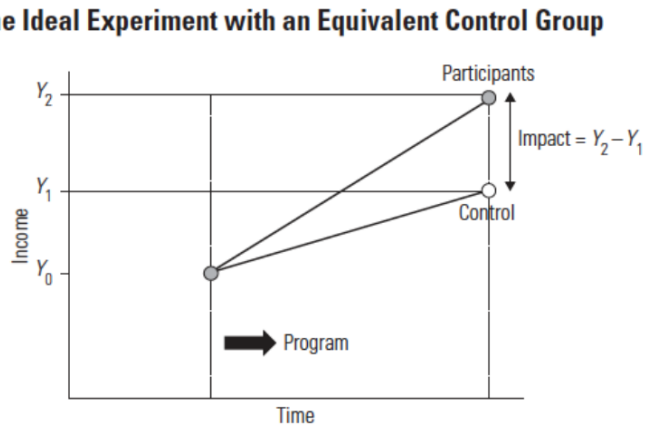
2. Impact Evaluation

2.1 What is causality?

- $$X \to Y$$
- $$\text{Cause} \to \text{Effect}$$
- **Causality v.s. Correlation?** [鸡叫与天亮]
  - **案例 1:** 去医院使得人更不健康?
    - 因为总去医院的人本身身体就更不好，所以结论“去医院会导致健康状态下降”是错误的⇒ 去医院和不去医院的人是不可比的⇒ 选择性偏误 [*selection bias*]。
    - *Treatment effect* =去医院的人的健康状态 - 没去医院的人的健康状态. [***Biased***?]
      - $$E(Y_{1i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 0)$$
      - 其中$E(Y_{1i} \mid D_i = 1)$是该去医院也的确去了的那些人；
      - $E(Y_{0i} \mid D_i = 0)$是不该去医院的人且实际上也没有去的那些人。
    - *Causality* = 去医院的人的健康状态 - 应该去医院但没去时的健康状态。
      - $$E(Y_{1i} \mid D_i = 1) - E(Y'_{0i} \mid D_i = 1)$$
      - $E(Y'_{0i} \mid D_i = 1)$是无法观测的反事实状态 [*counterfactual situation*]。
    - $E(Y_{1i} \mid D_i = 1) - E(Y'_{0i} \mid D_i = 1) + E(Y'_{0i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 0) = E(Y_{1i} \mid D_i = 1) - E(Y_{0i} \mid D_i$
    - 其中$E(Y'_{0i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 0)$ 这一部分即为选择性偏误。
  - **其他案例**:
    - 空气污染越严重，预期寿命越长？经济发展程度同时影响空气污染水平与预期寿命→ 遗漏变量偏误
    - 戒烟的人，健康情况反而恶化？一个健康状态出问题的人才会戒烟→ 反向因果
- **为什么要准确地识别因果关系?**
  - 因果关系是一个真实的关系，我们希望能够抛开现象看到本质，所以需要识别。
  - 具有政策含义，知道了因果关系才能去改变Y.

2.2 Experimental Studies

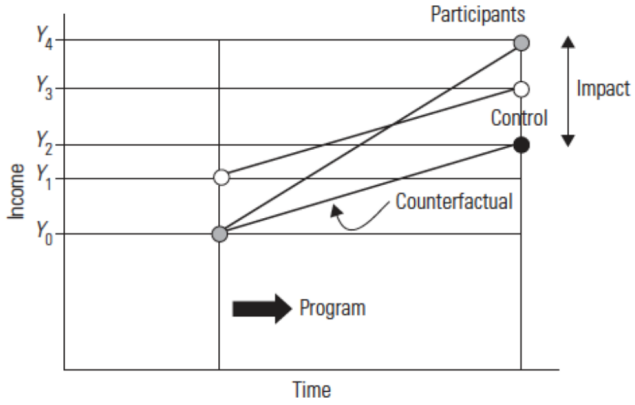
- **案例:** 临床试验
  - ⇒ 样本随机分成实验组[*treatment group*]和对照组[*control group*].

- ⇒ 比较结果，始别干预的因果效应。
- ⇒ 避免安慰剂效应，双盲实验。
- 

- **问题:** 社会科学研究中很难有理想实验
  - 伦理问题：不能让相同条件的一部分学生上大学、一部分不上
  - 难以实现随机分配：精准扶贫往往针对的是更加贫困的村民
  - 安慰剂效应 [*Placebo effects*]
  - 在完美实验不可得的情况下，就需要借助计量方法完善非实验研究 [*non-experimental*].

2.3 Impact Evaluation: Non-Experimental Studies

> Q：**在理想实验不可得的情况下，应如何进行研究?**

- **Counterfactual Situation**：
  - 反事实情况无法观测→ 如何构建反事实情境?
  - 双重差分法
  - 

    - 在这个图中，灰色点-白色点的部分即为选择性偏误$E(Y_{1i}|D_i=1)-E(Y_{0i}|D_i=0)$（因为两个本来起点就不一样）。
    - 因此，我们需要做的是找到反事实⇒→ 如何构造反事实？通过观察对照组（白色部分）进行模拟，画出平行线。
    - 怎么知道这个假设是否正确？可以进行一个test，比如收集一些在项目实行之前的数据，观察是否平行。

2.4 Causal (Ceteris Paribus) Effect

- 计量经济学的根本目标就是始别因果关系。
- *ceteris paribus*- which means "*other (relevant) factors being equal*" - plays an important role in causal analysis. [控他条件]
- 计量经济学方法能够使研究者尽可能接近一个控他实验。
- 实证研究中的关键问题：我是否固定住了所有相关变量?

3. Studying Economic Relationships

3.1 Case Study

- 犯罪的经济学模型: Economic Model of Crime (Becker, 1968)

> - $$\begin{aligned}\text{crime} &= \beta_0 + \beta_1 \text{ wage }_m + \beta_2 \text{ othnic } + \beta_3 \text{ freqarr } + \beta_4 \text{ freqconv} \\ &+ \beta_5 \text{ avgsen } + \beta_6 \text{ age } + u\end{aligned}$$

3.2 Steps in Empirical Economic Analysis

1. 形成问题
2. 通过直觉与经济学理论模型建立初步计量模型
3. 细化计量模型
4. 参数估计与检验
5. 解释与政策含义

---

4. The Structure of Economic Data

- 截面数据 [*Cross-sectional data*]
- 时间序列数据 [*Time series data*]
- 合并的截面数据 [*Pooled cross-sectional data*]
- 追踪数据/ 面板数据 [*Panel data*]

4.1 Cross-Sectional Data: [Main Focus]

- 在某一特定的时点上，在研究总体中随机抽样得到数据。
- **案例**：人口普查数据

4.2 Time Series Data: [Not Covered]

- 对于某一特定指标的长期追踪，自变量是时间
- **案例**：原油价格、GDP、消费者价格指数等
- 在时间序列分析中，**季节模式**[*seasonal pattern*]是一个重要的变量。

4.3 Pooled Cross-Sectional Data:

- 将n年的截面数据放在一起，每一年的数据都**重新抽样、相互独立**。
- 既有截面数据、又有时间序列数据的特征。
- 当单次调查样本量极小时，使用合并的截面数据可以扩大样本量。

4.4 Panel (or Longitudinal) Data:

- 只在基年baseline做抽样，之后不再抽样，而是每个周期*revisit*。
- **优势**：既能够控制不随时间变化的误差项、又能够考虑到**滞后变量** [*lag variable*]的影响。
- **缺陷**：随着时间的流逝，你能再次联系到这个样本的可能性越来越少[去世或者搬家]，出现**数据磨损**，进而丧失代表性。

2.1 Probability Theory

2022.03.05

---

1. Random Variables and Probability Distributions

- 可参考：Basics of Prob and Stats_ShiyuanLi.pdf

1.1 实验与随机变量

- 如果一项试验能够在相同条件下重复进行；而且每一次试验至少有两种可能结果，但在每一次试验前都无法确定哪个可能结果会得以实现。则称之为随机试验.
- 随机变量为随机结果的数值。比如正面朝上规定为1，反面朝上规定为0
- 大X表示变量，x表示X的一个可能取值
- 只有0和1两种取值的变量被称为伯努利变量或二分变量 [*Bernoulli variables/ binary variables*]

1.2 离散型随机变量

- 最简单的离散型随机变量 [*discrete random variables*]即为二分变量
  - $$P(X=1)=\theta \quad P(X=0)=1-\theta$$
- 若有多种可能结果，则更具有一般性的情况为
  - $$p_j=P(X=x_j), \quad j=1,2,\dots,k, \quad \sum_j p_j=1$$
- **概率密度函数** [*probability density function* (pdf)]
  - $$p_j=f(x_j), \quad j=1,2,\dots,k$$
  - 对于离散变量而言，它的概率密度函数就是抽到这个数值的概率，介于0和1之间：
    - $$0 \le P(X=x_j) \le 1$$
  - 所有值的概率之和为1
    - $$\sum_{j=1}^{k} f(x_j)=1$$
- **累积分布函数** [*cumulative distribution function* (cdf)]
  - 给定x，小于x的所有可能情况的概率的加总构成了累积分布函数
    - $$F(x) \equiv \sum_{x_j \le x} f(x)=P(X \le x)$$
  - 图像：对于离散变量来说，就是阶梯型上升到1

1.3 连续型随机变量

- 连续型随机变量 [*Continuous random variables*]在整个实数轴上取值。
- 看区间取值可能性，而不是一个特定点的取值可能性。一个点的概率值没有意义。
- **概率密度函数**：
  - 概率密度函数是非负函数
    - $$f(x) \ge 0 \quad \text{so that} \quad \int_{-\infty}^{+\infty} f(x)dx=1$$

    - $$P(a \leq X \leq b) = \int_a^b f(x)dx \geq 0$$
  - 图像：$P(a \leq X \leq b)$指代的是曲线下的面积
- **累积分布函数**
  - $$F(x) \equiv \int_{-\infty}^{x} f(x)dx = P(X \leq x) \quad \text{with} \quad f(x) = \frac{\Delta F(x)}{\Delta x}$$
  - pdf是cdf的微分，cdf是pdf的积分
- **性质**
  - $0 \leq F(x) \leq 1$
  - if $x_2 > x_1$, then $F(x_2) \geq F(x_1)$
  - $F(+\infty) = 1$
  - $F(-\infty) = 0$
  - For any number c, $P(X > c) = 1 - F(c)$
  - For any numbers $a < b$, $P(a < X \leq b) = F(b) - F(a)$[牛顿-莱布尼茨公式]

---

2. Joint Distributions

> Q：**如果我们想知道两个以上变量的分布呢?**

2.1 Joint distributions

- 比如性别与婚姻，x有两个取值男女，y有两个取值已婚未婚。联合分布就是要考虑四种可能性发生概率的分布（四种可能性的概率相加为1）
- **联合概率密度函数**：
  - $$f_{X,Y}(x,y) = P(X = x, Y = y)$$
  - **性质**：
    - $$f_{X,Y}(x,y) \geq 0 \text{ and } \sum_x \sum_y f_{X,Y}(x,y) = 1 \text{ in the discrete case}$$
    - $$f_{X,Y}(x,y) \geq 0 \text{ and } \int_x \int_y f_{X,Y}(x,y)dydx = 1 \text{ in the continuous case}$$
  - 给定区间，则：
    - $$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x,y)dydx$$
- **累积联合概率密度函数**：
  - $F(z,w) = P(X \leq z, Y \leq w)$ are
    - $$F(z,w) = \sum_{x \leq z} \sum_{y \leq w} f_{X,Y}(x,y) = 1 \text{ in the discrete case}$$
    - $$F(z,w) = \int_{-\infty}^{z} \int_{-\infty}^{w} f_{X,Y}(x,y)dydx \text{ in the continuous case}$$
- 如果两个变量相互独立：
  - $$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$
  - $f_X(x)$ 和 $f_Y(y)$ 是边缘概率密度函数 [*marginal probability density functions*]
  - $$f_X(x) = \begin{cases} \sum_y f_{X,Y}(x,y) & \text{in the discrete case} \\ \int_y f_{X,Y}(x,s)ds & \text{in the continuous case} \end{cases}$$

- $$f_Y(y) = \begin{cases} \sum_x f_{X,Y}(x,y) & \text{in the discrete case} \\ \int_x f_{X,Y}(s,y)ds & \text{in the continuous case} \end{cases}$$

2.2 Conditional Distributions

- 条件概率密度函数 [*Conditional probability density function*] 是给定x之后y的分布情况；联合分布是在所有样本中的分布情况.

- $$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \text{ with } f_X(x) > 0$$

- **性质:**

- $$f_{Y|X}(y|x) \geq 0 \quad \text{and} \quad \begin{cases} \sum_y f_{Y|X}(y|x) = 1 & \text{in the discrete case} \\ \int_{-\infty}^{\infty} f_{Y|X}(y|x)dy = 1 & \text{in the continuous case} \end{cases}$$

- **条件、联合和边缘概率分布的关系**:

- $$\text{conditional} = \frac{\text{joint}}{\text{marginal}}$$

- $$f_{X,Y}(y,x) = f_X(x)f_{Y|X}(y|x)$$

---

3. Features of Probability Distributions

> Q：**我们主要关注变量的哪些方面?**

- **集中趋势**
  - 均值 (Expected value)
    - 一般来说sample用mean，population 用expected value
  - 中位数(Median)
    - 将pdf分为两个相等的部分
- **离散趋势**
  - 方差 (Variance)
    - squared difference of X from its expected value
  - 标准差 (Standard deviation)
    - square root of the variance
- **变量关联性**
  - 协方差 (Covariance)
    - measure of *linear dependence* between two random variables
  - 相关系数 (Correlation coefficient)
    - *unit-free* measure of linear dependence between two random variables

3.1 Expected Value

- 期望是所有可能取值的加权平均:

- $$E(X) = \mu_X = \sum_x xf(x)$$

- $$E(X) = \mu x = \int_{-\infty}^{\infty} xf(x)dx$$

- **计算法则**:
  - $E(a) = a$ [扔一亿次只有一个数的骰子依然是这个数]
  - $E(X+Y) = E(X) + E(Y) = \mu_X + \mu_Y$
  - $E(aX+b) = aE(X) + b = a\mu_X + b$

3.2 Variance

- 为衡量离散趋势，我们考虑使用平方差 $(X-\mu_X)^2$
  - 消除正负号的影响
  - 距离越远值越大
- 我们无法用一个随机变量去衡量随机变量的特征，因此取均值，得到方差：
  - $$Var(x)=\sigma_x^2\equiv E[(X-\mu_x)^2]=\begin{cases}\sum_x f(x)(x-\mu_x)^2 & \text{if discrete}\\ \int_{-\infty}^{\infty} f(x)(x-\mu_x)^2dx & \text{if continuous}\end{cases}$$
  - 简而言之，观测值与平均值的偏差和的平均值，就被定义为方差。
- **计算法则**：
  - $$Var(aX+b)=Var(aX)=a^2Var(X)=a^2\sigma_X^2$$
- **标准差**：
  - $$sd(X)=\sigma_X\equiv\sqrt{Var(X)}$$

3.3 Standardizing a Random Variable

- We may define a new random variable Z by substracting its mean and dividing by its standard deviation:
  - $$Z=\frac{X-\mu}{\sigma}$$
  - which can be written as Z = aX + b with $a\equiv(1/\sigma)$ and $b\equiv-(\mu/\sigma)$
- $\Rightarrow$ Z即为**标准化随机变量** [*standardized random variable*].
  - $$E(Z)=E(aX+b)=aE(X)+b=(\mu/\sigma)-(\mu/\sigma)=0$$
  - $$Var(Z)=Var(aX+b)=a^2Var(X)=(\sigma^2/\sigma^2)=1$$
  - 所谓的标准，就是均值为0方差为1

3.4 Skewness and Kurtosis

- X**的n阶中心矩** [*nth central moment*]
  - $$\mu_n=E[(X-\mu_X)^n]$$
  - 所谓的n阶矩，即为$E(X^n)$。在此基础上，将n阶中心距定义为减去均值后的n次方
- **1阶中心矩**：
  - $E[(X-\mu_X)^1]=E(X)-\mu X=0$
  - 这也是为什么方差要定义为二次方的原因，否则会抵消为0
- **2阶中心矩**：
  - $E[(X-\mu_X)^2]$，即为方差的定义
- **3阶中心矩**：
  - $E[(X-\mu_X)^3]$ 被定义为**偏度** [*skewness*]
  - 偏度衡量对称性。如果完全对称，则偏度为0。
  - 尾巴在左边→ 分布左偏→ 偏度为负
  - 尾巴在右边→ 右偏→ 偏度为正
- **4阶中心矩**：
  - $E[(X-\mu_X)^4]$ 被定义为**峰度** [*kurtosis*]
  - 峰度越小越高耸，尾巴越窄
  - 峰度越大尾巴越厚

3.5 Covariance

- 从单个变量的关系→ 考虑两个变量的关系：协方差
- 协方差 [*covariance*] 衡量了两个随机变量之间的**线性**关系

> - $$Cov(X,Y) = \sigma XY \equiv E[(X-\mu_x)(Y-\mu_Y)]$$
> - $$= \begin{cases} \sum_x \sum_y f(x,y)(x-\mu_x)(y-\mu_y) & \text{if discrete} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)(x-\mu_x)(y-\mu_y)dxdy & \text{if continuous} \end{cases}$$

- 协方差的另一个公式：乘积的期望-期望的乘积

> - $$Cov(X,Y) = E(XY) - \mu_X\mu_Y$$

- 如果两个变量线性独立，则

> - $$E(XY) = E(X)E(Y) = \mu_X X \mu_Y \longrightarrow Cov(X,Y) = 0$$

- 协方差为0只能推出线性独立，无法推出两个随机变量分布独立/均值独立。
- **计算法则**：
  - $Cov(X,X) = Var(X)$: 自己和自己的协方差就是方差。
  - $Cov(aX+b, cY+d) = acCov(X,Y)$
  - $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
  - $Var(aX+bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$

3.6 Correlation Coefficient

- 相关系数 [*correlation coefficient*]是去量纲的协方差：

> - $$\text{Corr}(X,Y) = \rho_{XY} = \frac{\text{Cov}(X,Y)}{sd(X)sd(Y)} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

- 相关系数在[-1,1]之间取值，正负性与协方差一致。
- 两个变量线性独立，则相关系数为0.

3.7 Conditional Expectation ❖

- 给定x，计算Y的期望，即可得到条件期望：$E(Y \mid X=x)$
- 条件期望告诉我们Y的期望如何随着x取值的变化而变化，而这就是计量经济学主要回答的问题→ 随着教育程度的提高，工资的期望如何变化?

3.8 The Law of Iterated Expectations

- 迭代期望定律的基本原理：**总体的均值=各组组内均值的加权平均**
- 定理的用途：**已知条件期望（各组的组内均值），可以求出无条件期望（总体的均值）**

> - $$E[E(Y \mid X)] = E(Y)$$

- 五个教育程度组的平均工资再做一个期望，就等于不分组的情况下的所有人的平均工资。

3.9 Independence and Correlation

- **条件最强-分布独立 independent**： X 中的任何信息(是否上过大学)都无法预测 Y 取任意值(收入高低)的概率

> - $$f_{XY}(x,y) = f_x(x)f_y(y)$$

- **条件适中- 均值独立 mean independent**：X 中的任何信息(是否上过大学)都无法预测 Y 的**平均值**(收入的**平均**水平)

> - $$E(Y|X) = E(Y)$$
> - 不同教育水平的群体中，收入均值没有差异

- **条件最弱- 线性独立| 不相关 linear independent | uncorrelated**：协方差与相关系数为0，仅代表X与Y不存在线性关系，不意味着X与Y没有其他非线性关联。

> - $$E(XY) = E(X)E(Y)$$

- [分布]独立可以推出均值独立，进而可以推出线性独立，但反之不成立。

3.10 Conditional Variance

- 计算公式：
  - $$Var(Y \mid X = x) = E(Y^2 \mid X = x) - [E(Y \mid X = x)]^2$$
- Example: Var(SAVING|INCOME) = 400 + .25*INCOME*
  - ⇒ 随着收入的提高，储蓄的方差变大（异方差）
  - 收入越高的人，储蓄的变化范围更大。
- 若两个随机变量相互独立，则 $Var(Y \mid X) = Var(Y)$

2.2 Estimation and Hypothesis Testing

2022.03.05

---

1. Statistical Inference and Estimation

- 通过样本估计总体参数、推断总体即为统计推断
- 估计量[*Estimator*] 是样本的函数。将不同的观测值带入到估计量代表的函数中，就能得到估计值[*Estimate*]（常数）。估计值是估计量的一个实现值。
- 估计量是计算参数估计值的法则，不依赖于实际的样本值。我们只能评价估计量的好坏、无法评价估计值的好坏。
- 估计量的评价指标：
  - 无偏性 [*Unbiasedness*]
  - 有效性 [*Efficiency*]
  - 一致性 [*Consistency*]

---

2. Unbiasedness, Efficiency and Consistency

> Q：我们以什么标准衡量估计量的优劣？

2.1 Unbiasedness

- 如果期望落在真值上，即样本的均值恰好等于总体参数，就是无偏的
  - $$E(W)=\theta$$
- 不是说一定保证最后的估计值是无偏的，而是推断的这个过程保证是冲着真值来的，我们就认可。
- $$E(\bar{Y})=\mu$$
  - 和的期望=期望的和，注意灵活使用
- $$S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2$$
  - 若希望保证总体方差σ^2无偏，则样本方差的分母应为$n-1$
  - 为什么样本方差（sample variance）的分母是 n-1？

2.2 Efficiency

- 在无偏的情况下选择方差小 [更集中]的，因为落在真值附近的概率更大，分散性较小
- 然而，光有有效性是不够的，比如用常数0做估计量，稳定得不行，但是离真实估计量十万八千里。
- ⇒ 如何得到抽样方差 $\mathrm{Var}(\bar{Y})$ [*Sampling Variance*]?
  - 抽样分布的方差即为抽样方差。抽样方差是常数，而非随机变量
  - $$\mathrm{Var}(\bar{Y})=\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(Y_i\right)=\frac{1}{n^2}n\sigma^2=\frac{\sigma^2}{n}$$
    - 因为这里假定每一次抽取都是相互独立的，因此和的方差等于方差的和。
  - 当样本量扩大时，抽样方差降低；当样本足够多到就是一个整体时，抽样方差就是0了。

2.3 Consistency

- 如果随着样本量的增大，估计量越来越接近总体参数，则这个估计量是一致的。最极端的情况，如果总体中的所有人都进入到样本中，则这个样本得到的估计量的取值就等于总体参数。

> - $$P(|W_n - \theta| > \varepsilon) \to 0 \quad \text{as } n \to \infty$$
> - $$\operatorname{plim}(W_n) = \theta \text{ (the probability limit of } W_n \text{ is } \theta \text{)}$$

- 估计量的一致性在中级计量经济学的框架中，是评价估计量好坏最重要，最基本的标准。如果不满足一致性，即便获得了总体的信息，基于不一致的估计量的估计结果，依然有可能与总体参数相去甚远。
- 无偏不等于一致: 以抛硬币为例，无论投多少次的结果都是0或1，统计量并没有在试验次数变多的时候无限接近真实值，这也就是说无偏是不能得出一致的。
- 无偏性与n的大小无关，即无论样本量多小，其均值应该都是接近于真值的；但是一致性则和n的大小有关，随着样本量的增大，其方差应该越来越小，估计量越来越接近于真值，是一个渐进的概念。无偏性的要求更加严格。

3. The Bivariate Linear Regression Model

2022.03.19

---

1. The Basic Concept

1.1 Definition of a Simple Regression Model

> Q：**如何用x 解释y？y 如何随着x 的变化而变化？**

- → 建立简单线性回归模型[*simple linear regression model/ bivariate linear regression model/ two variable linear regression model*].
- **形式**：
  - $$y = \beta_0 + \beta_1 x + u \tag{3.1}$$
- **术语解释**：
  - β_0 是截距系数[*intercept parameter*]，又称常数项[*constant term*]
  - β_1 是斜率系数[*slope paramete*]，衡量在其他变量不变的情况下，y和x的关系→ $\Delta y = \beta_1 \Delta x \quad \text{if } \Delta u = 0$
  - u 被称为残差项或干扰项[*error term/disturbance*]，代表了除x以外其他所有影响y的因素。
  - 线性 [*linearity*]: 参数之间为线性关系，y与x的关系未必是线性的[如可以取对数]→ See Section 3.4.
- **案例**：教育回报
  - $$wage = \beta_0 + \beta_1 educ + u$$
    - 在控制其他变量的情况下，增加一年的教育程度，工资会增加β_1
    - 误差项中包括工作经验、能力、职业等。
  - **问题**：
    - β_1 如何衡量“递增回报”[*increasing returns*]？→ See Section 3.4
      - 在小学多上了一年（一年级和二年级），和在大学的时候多上了一年（大三与大四）对工资的影响不一样，后者影响更大；
      - 毕业年比非毕业年的影响更大（有了毕业证）→ 并非线性变化
    - 这个回归模型能否在控他的情况下，衡量x如何影响y？→ See Section 3.5

1.2 Zero Conditional Mean Assumption

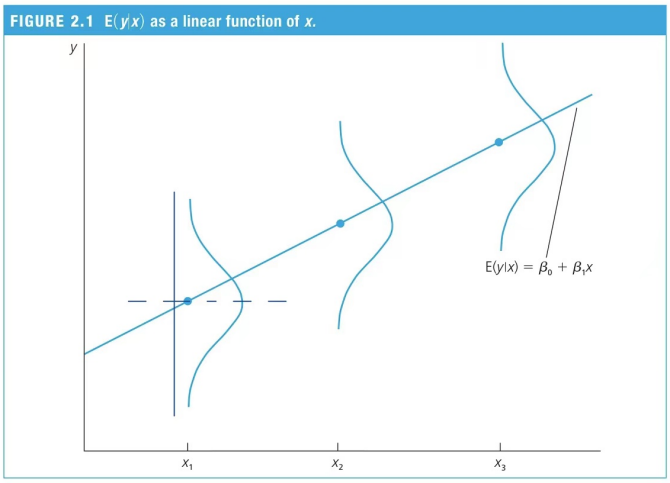
> Q：**在建立ceteris paribus模型之前，需要做出什么假定？**

- **假定1**：假定残差期望为0
  - $$E(u) = 0 \tag{3.2}$$
  - 这一假定并不严格。只要方程中包括截距项β_0，就总能通过调整截距，使得总体中残差项的均值为0 → $E(u) = k \neq 0$ 只会对截距造成影响。截距参数没有经济学含义，所以问题不大。
  - 推导：
    - $$\begin{aligned} y &= \beta_0 + \beta_1 x + u \\ &= \beta_0 + k + \beta_1 x + u - k \\ &= \beta_0^* + \beta_1 x + u^* \end{aligned}$$
    - $$\Rightarrow E(u^*) = E(u - k) = E(u) - k = k - k = 0$$
- **假定2**：条件期望零值假设[*Zero Conditional Mean Assumption*]

- 在假定1的基础上，我们需要对u和x的关系做出更加严格的假定。
- u和x的相关系数为零只能得到u和x线性独立 [uncorrelated=linear independence]; u 同时应该不与x的任何函数[如x^2]相关 .
- ⇒ 得到更加严格的假定
  - $$E(u \mid x) = E(u) = 0 \tag{3.3}$$
- 当假定[3.3]成立时，u 与x以及x任意形式的函数均不相关→ 均值独立
- 案例：
  - $E(ability \mid educ) = E(abiliy)$：对于任何一个教育水平的人而言，其平均能力相同，同时与总体中的能力的期望相等 [$E(abil|8) = E(abil|6)$]。换言之，x中的任何信息都无法预测u的平均值，教育水平的差异无法体现能力的差异。
    - 如果随着教育水平的增长，能力也增长，则零条件均值假设不成立。
    - 一般而言，能力越高的人会接受越多的教育，如果我们无法保证对于所有的教育水平来说个人能力都是相同的，那就不能应用简单回归。
  - 比如吃药和健康状态的关系，在理想实验中，treatment应该与其他所有与health status相关的因素独立，比如不会有一组是好医生，另一组是兽医。所以理想实验不需要计量。
- **总结**：
  - 零条件均值假设包括两部分：
    - 假定$E(u) = 0$，这个假定无非是定义了截距
    - $E(u|x) = E(u)$，u均值独立于x
  - 通过零条件均值假设，可以得出在总体中，u与x不仅线性独立，同时均值独立→ u的平均值与x值无关且为0

1.3 Conditional Mean of the Dependent Variable Y

> Q：在零条件均值假设的基础上，**Y的条件均值是什么**?

- 综合[3.1]与[3.3]，得到**总体回归函数** [*population regression function*, PRF]
  - $$E(y \mid x) = \beta_0 + \beta_1 x \tag{3.4}$$
- **推导**：
  - $E(y \mid x) = E(\beta_0 + \beta_1 x + u \mid x) = \beta_0 + \beta_1 E(x \mid x) + E(u \mid x) = \beta_0 + \beta_1 x$
- **解读**：x增加一单位，y的期望值增加β_1
- y 的条件均值是x的线性函数→ 得到**总体回归线** [*population regression line*]
- **分布图像**：对于任意给定的x, y的分布都以 $E(y \mid x)$ 为中心：
  - 

---

2. Ordinary Least Squares (OLS)

2.1 Deriving β0 and β1

> Q：如何从样本信息推断总体参数?

- 由于β_0 和 β_1 都是未知的，因此需要利用样本中的信息$\{(x_i, y_i) : i = 1, \cdots, n\}$推断总体的参数$\beta$。
  - 每一个观测到的样本都可以写成 $\Rightarrow y_i = \beta_0 + \beta_1 x_i + u_i$ [带i的都是样本]
- **三种主流方法**:
  - 矩估计 [Method of moments (MoM)]
    - 通过假设得到了两个矩条件，再将样本带入到两个条件中计算得到
  - 最小二乘法 [*Ordinary LeaEst Squares* (OLS)]※
    - 不依赖任何假设，目的在于**将残差项最小化**
  - Maximum Likelihood (ML) [not discussed here]

2.2 Method of Moments

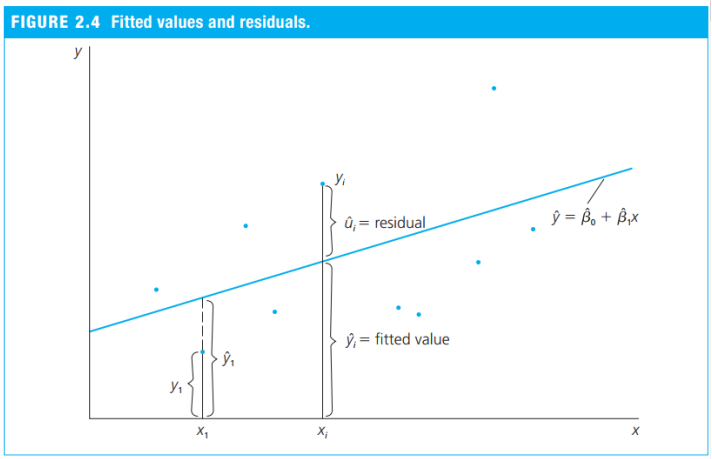
- 由假设[3.3]均值独立可以推出x和u线性独立
  - $$Cov(x, u) = E(xu) - E(x)E(u) = E(xu) = E(x)E(u) = 0 \tag{a}$$
  - $$\Rightarrow E(u) = Cov(x, u) = E(xu) = 0 \tag{b}$$
- $u = y - \beta_0 - \beta_1 x$，利用$x$, y, β_0, β_1替代$u \Rightarrow$ moment restrictions
  - $$E(y - \beta_0 - \beta_1 x) = 0 \tag{c}$$
  - $$E[x(y - \beta_0 - \beta_1 x)] = 0 \tag{d}$$
- 带入样本数据:
  - $$n^{-1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{e}$$
  - $$n^{-1} \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{f}$$
- 改写
  - $$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \tag{g}$$
  - $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{h}$$
- 将(h)带入(f)
  - $$\sum_{i=1}^{n} x_i [y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i] = 0 \tag{i}$$
- 移项，得到
  - $$\sum_{i=1}^{n} x_i (y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} x_i (x_i - \bar{x}) \tag{j}$$
- 由summation的基本代数性质可知,
  - $$\sum_{i=1}^{n} x_i (y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{k}$$
  - $$\sum_{i=1}^{n} x_i (x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{l}$$
- 则斜率的估计量为
  - $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3.5}$$
  - provided that $\sum_{i=1}^{n} (x_i - x)^2 > 0$

- 注：β_1 =协方差除以方差，同时需要保证方差不为0，x必须有变化。如果所有人都是大学毕业，就无法判断教育对于工资的影响。

2.3 OLS Estimation

> Q：如何将实际值与估计值的距离最小化?

2.3.1 概念

- 残差的定义：
  - 实际值与回归线上的估计值的垂直距离被定义为**残差**[*residuals*].
    - $$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$
  - if $\hat{u}_i > 0$, the line underpredicts y_i.
  - if $\hat{u}_i < 0$, the line overpredicts y_i.
  - if $\hat{u}_i = 0$, the data point lies on the OLS line. [none of the data points must actually lie on the line.]
  - u_i和$\hat{u}_i$的区别：前者是真值，后者是前者的估计值
    - 我们无法知道总体中的β [真值]，因此只能通过样本得到$\hat{\beta}$ [估计值]
    - $u_i = y_i - \beta x_i$
    - $\hat{u}_i = y_i - \hat{\beta} x_i$
- 样本回归函数 [*sample regression function*, SRF]
  - $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$
  - 样本回归函数又名OLS回归线[*OLS regression line*]，是对总体回归函数$E(y \mid x) = \beta_0 + \beta_1 x$ 的估计形式[estimated version]。
  - $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是未知参数 β_0 和 β_1 的估计量
  - 残差$\hat{u}_i$代表误差项u_i的估计值
  - $\hat{y}_i$ 代表给定 x_i的情况下，得到的预测值/拟合值[*predicted/fitted values*]。每个拟合值都在OLS回归线上。
  - 总体回归线斜率的估计值即为x提高1单位，y 的拟合值 $\hat{y}$ 的变动程度
    - $$\hat{\beta}_1 = \frac{\Delta \hat{y}}{\Delta x}$$
- 图示：
  - 

- 代码：生成 $\hat{y}_i$ 和 $\hat{u}_i$
  - ```
    reg y x
    predict yhat, xb
    predict uhat, residual
    ```
    ```

2.3.2 推导

- **核心**：找到 β_0 和 β_1 的值，使残差平方和 $\sum_{i=1}^{n} \hat{u}_i{}^2$最小化

> - $$\min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} \widehat{u}_i^2 = \min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i \right) \right]^2 \tag{a}$$

- 计算一阶导数：

> - $$\frac{\partial \sum_{i=1}^{n} u_i^2}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{b}$$
> - $$\frac{\partial \sum_{i=1}^{n} \hat{u}_i^2}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \tag{c}$$

- 由于 $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

> - $$\Rightarrow \sum_{i=1}^{n} (y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i) x_i = 0 \tag{d}$$
> - $$\sum_{i=1}^{n} x_i (y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} x_i (x_i - \bar{x}) \tag{e}$$

- 由summation的代数性质可知

> - $$\sum_{i=1}^{n} x_i (x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \text{ and } \sum_{i=1}^{n} x_i (y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{f}$$
> - $$\Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{g}$$

- $\hat{\beta}_1$ 的另一种表达方式：

> - $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - x)(y_i - y)}{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - x)^2}$$
> - $$= \frac{\widehat{cov(x,y)}}{\widehat{Var(x)}} = \frac{\hat{\sigma} xy}{\hat{\sigma}^2 x} \tag{3.6}$$
>
> - 因为是总体[而不是样本]的方差与协方差，是一个估计值。因此需要加一个hat。
> - 只有在$\widehat{Var(x)} > 0$时才能估计参数，即x必须有变动

- 为什么选择最小化平方和?
  - 最小化绝对值不容易解
  - 二次函数值越大增长越快，所以对大的偏离的惩罚更强一些→ 我们能够“惩罚”离预测值更远的离群值。

---

3. Algebraic Properties of OLS Statistics

3.1 Algebraic Properties

1. 样本残差均值为0
   - 由 2.3.2-(b) 可知，样本中，**OLS残差的均值为0**

> - $$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{3.7}$$
> - $$E(\hat{u}_i) = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i = 0$$

- 注意：这是OLS估计量的固有性质，并非假设。

2. **自变量与残差估计值线性独立**
   - 由 2.3.2-(c) 可知，自变量与残差估计值的协方差为0→ $Cov(x_i, \hat{u}_i) = 0$:
     - $$\sum_{i=1}^{n} x_i \hat{u}_i = \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{3.8}$$

3. **均值点始终在回归线上**
   - $(\bar{x}, \bar{y})$ 始终在OLS回归线上
     - $$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \tag{3.9}$$

3.2 Implications

- **因变量分解**:
  - 我们可以将观测值y_i分解成拟合值和残差两部分，即:
    - $$y_i = \hat{y}_i + \hat{u}_i \tag{3.10}$$
  - 由[3.7]可知，y的估计值分布的均值和y的实际分布的均值是重合的。
    - $$\frac{1}{n} \sum_{i=1}^{n} \hat{u}_i = \bar{u} = 0 \quad \Rightarrow \quad \bar{\hat{y}} = \bar{y}$$
- **y的拟合值与残差估计值线性独立**
  - 由定义可知
    - $$\operatorname{Cov}\left(\hat{y}_i, \hat{u}_i\right) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right) \left(\hat{u}_i - \bar{\hat{u}}\right) \tag{a}$$
  - 由Summation的性质可知
    - $$\operatorname{Cov}\left(\hat{y}_i, \hat{u}_i\right) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right) \hat{u}_i = \frac{1}{n-1} \sum_{i=1}^{n} \hat{y}_i \hat{u}_i \tag{b}$$
  - 带入拟合值方程
    - $$\begin{aligned} \operatorname{Cov}\left(\hat{y}_i, \hat{u}_i\right) &= \frac{1}{n-1} \sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) \hat{u}_i \\ &= \frac{1}{n-1} \hat{\beta}_0 \sum_{i=1}^{n} \hat{u}_i + \frac{1}{n-1} \hat{\beta}_1 \sum_{i=1}^{n} x_i \hat{u}_i \end{aligned} \tag{c}$$
  - 由[3.7]和[3.8]可知
    - $$\operatorname{Cov}\left(\hat{y}_i, \hat{u}_i\right) = 0 \tag{3.11}$$
- **SST SSR SSE**
  - 总平方和SST [*Total sum of squares*]:
    - $$SST \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3.12}$$
    - 衡量了样本中因变量y_i的全部变动。
    - $\frac{SST}{n-1}$即为y_i的方差 [y的样本方差]
  - 解释平方和SSE [*Explained sum of squares*]:
    - $$SSE \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{3.13}$$
    - 衡量了因变量拟合值的变动$\hat{y}_i$ (use the fact that $\bar{\hat{y}} = \bar{y}$).
  - 残差平方和SSR [*Residual sum of squares*]:

- $$SSR \equiv \sum_{i=1}^{n} \hat{u}_i^2 \tag{3.14}$$
- 衡量了残差的估计值的变动 $\hat{u}_i$.

- 关系
  - $$SST = SSE + SSR \tag{3.15}$$
  - $$\begin{aligned}\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}\left[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y}_i)\right]^2 \\ &= \sum_{i=1}^{n}\left[\widehat{u}_i+(\hat{y}_i-\bar{y})\right]^2 \\ &= \sum_{i=1}^{n}\widehat{u}_i^2+2\sum_{i=1}^{n}\widehat{u}_i\left(\hat{y}_i-\bar{y}\right)+\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2 \\ &= SSR+2\sum_{i=1}^{n}\widehat{u}_i\left(\hat{y}_i-\bar{y}\right)+SSE\end{aligned}$$

3.3 Goodness-of-Fit

> Q：**如何衡量x能在多大程度上解释y?**

- **概念**:
  - $$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \tag{3.16}$$
  - 由于$SSE \leq SST$，$R^2 > 0$ 恒成立
- **解读**:
  - $R^2 = 1$代表完美拟合[perfect fit]，样本中y的所有变化都可以被x解释.
  - $R^2 = 0$ 代表完全没有解释力，x不能解释y的任何变化 [model with constant but no regressor].
  - 在对 R^2进行解读时，需要转化为百分数→ R^2%的y的变化可以被x解释
- **问题**:
  - 对于截面数据分析[*cross-sectional analysis*]而言，R^2 一般较小；时间序列分析[time series analysis] R^2 一般较大.
    - 因为昨天的你和今天的你一定是高度相关的，所以时间序列的数据的R^2更高。
  - 即便增加的解释项几乎没有解释力，由于SSE不会下降，R^2依然会提高（至少保持不变）
  - 仅仅评价了数据的拟合程度，即y能在多大程度上被x解释。因果关系[ceteris paribus]无法利用R^2衡量
  - 在评价回归模型的时候，不应过分在意R^2

---

4. Units of Measurement and Functional Form

4.1 Units of Measurement and Functional Form

> Q: **刻度的变化如何影响回归方程?**

- 改变单位会改变待估参数的值，但不会影响对结果的解读。
- R^2不会因为单位的变化而变化
- 案例: CEO 工资与股权回报
  - Baseline: 工资以 1,000 dollars为单位，roe为股权回报百分比
    - $$\widehat{salary} = 963.191 + 18.501roe$$
  - 因变量以美元为单位，$salardol = 1000 * salary$

    - $$\Rightarrow \widehat{salardol} = 963191 + 18501roe$$
    - 全部系数同步扩大
  - 自变量缩小一百倍，$roedec = roe = 100$
    - $\Rightarrow \widehat{salary} = 963.191 + 1850.1roedec$
  - $y * c \Rightarrow \hat{\beta}_0 * c, \hat{\beta}_1 * c$
  - $x * c \Rightarrow \hat{\beta}_0$ 不变, $\hat{\beta}_1 / c$

4.2 Incorporating Nonlinearities in Simple Regression

- 将简单回归模型改写为
  - $$f(y) = \beta_0 + \beta_1 g(x) + u$$
- 其中，$f(\cdot)$ & $g(\cdot)$ 可以是非线性方程
- **加入对数**
  -

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

- **案例**：Wage v.s. Log(Wage)
  - level-level regression function:
    - $$\widehat{wage} = -0.90 + 0.54educ$$
    - 教育程度增加一年，每小时多赚54美分。
  - Log-level regression function:
    - $$\widehat{log(wage)} = 0.584 + 0.083educ$$
    - 教育程度增加一年，每小时多赚8.3%.
  - 注：因为两个模型因变量不同，R^2不可比。

---

5. Unbiasedness of OLS

> Q: **OLS估计量的无偏性需要满足何种假定?**

5.1 The Gauss-Markov Assumptions for Simple Regression

- **SLR.1 Linear in Parameters**
  - 因变量y与自变量x和误差项u相关，得到总体回归方程
  - $$y = \beta_0 + \beta_1 x + u$$
- **SLR.2 Random Sampling**
  - 通过随机抽样形成样本，得到样本回归方程
  - $$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, 2, \dots, n$$
- **SLR.3 Sample Variation in the Explanatory Variable**
  - 解释变量x非常数
- **SLR.4 Zero Conditional Mean ※**
  - $$E(u \mid x) = 0$$

- **SLR.5 Homoskedasticity**

> - $$\operatorname{Var}(u \mid x)=\sigma^2$$
> - 给定任何一个x，误差项的方差均相同。

5.2 Unbiasedness of OLS

- 满足**SLR.1-4**，即可推导出无偏性。
- 何为无偏性？

> - $$E\left(\widehat{\beta}_0\right)=\beta_0 \quad E\left(\widehat{\beta}_1\right)=\beta_1$$

- 给定$y_i-\bar{y}=\left(\beta_0+\beta_1 x_i+u_i\right)-\left(\beta_0+\beta_1 \bar{x}\right)=\beta_1\left(x_i-\bar{x}\right)+u_i$，

> - $$\begin{aligned}\Rightarrow E\left(\widehat{\beta}_1\right) & =E\left(\frac{\sum_{i=1}^n \beta_1\left(x_i-\bar{x}\right)^2+\left(x_i-\bar{x}\right) u_i}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}\right) \\ & =\beta_1+E_x\left(\frac{\sum_{i=1}^n\left(x_i-\bar{x}\right) E\left(u_i \mid x_i\right)}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}\right) \\ & =\beta_1+\frac{\sum_{i=1}^n\left(x_i-\bar{x}\right)}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2} E\left(u_i \mid x_i\right) \\ & =\beta_1\end{aligned}$$

---

6. Variances of the OLS Estimators

6.1 The Sampling Variances of the Estimators

- 给定同方差假设，可得

> - $$\sigma^2=E\left(u^2\right)=\operatorname{Var}(u)$$
> - $$\operatorname{Var}(y \mid x)=\sigma^2$$

- 满足**SLR.1-4**，即可得到估计量的抽样方差
  - 给定$\widehat{\beta}_1-\beta_1=\frac{\sum_{i=1}^n\left(x_i-\bar{x}\right) u_i}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}$
  - $$\begin{aligned}\Rightarrow \sigma_{\widehat{\beta}_1}^2 & =\operatorname{Var}\left(\widehat{\beta}_1\right)=E\left[\left(\widehat{\beta}_1-E\left(\widehat{\beta}_1\right)\right)^2\right]=E\left[\left(\widehat{\beta}_1-\beta_1\right)^2\right] \\ & =E\left[\left(\frac{\sum_{i=1}^n\left(x_i-\bar{x}\right) u_i}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}\right)^2\right] \\ & =\frac{E\left(u^2 \mid x\right)}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2} \\ & =\frac{\sigma^2}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}\end{aligned}$$
- $$\operatorname{Var}\left(\widehat{\beta}_1\right)=\frac{\sigma^2}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}=\frac{\sigma^2}{SST_x}$$

- 如果残差的方差变大，抽样方差变大；如果解释变量的方差变大，抽样方差变小→ 大样本情况下，估计的准确性更高。

6.2 Estimating the Error Variance

> Q: 如何估计误差项的方差？

- 由于我们不知道u_i，因此只能通过$\hat{u}_i$估计误差项方差。

- 我们使用OLS估计法本身就使用了$\sum_{i=1}^{n} \widehat{u}_i = 0, \quad \sum_{i=1}^{n} x_i \widehat{u}_i = 0$两个条件，因此自由度为$n-2$:

> - $$\widehat{\sigma}^2 = \frac{1}{(n-2)} \sum_{i=1}^{n} \widehat{u}_i^2 = \frac{SSR}{(n-2)}$$

6.3 Standard Error of the Regression and the Estimated Coefficients

- 由于我们无法知道真值σ，只能通过$\hat{\sigma}$估计系数的标准差$\mathrm{s}d\left(\widehat{\beta}_1\right) = \frac{\sigma}{\sqrt{SST_x}}$。
- 因此，标准误即为系数标准差的估计量

> - $$\operatorname{se}\left(\widehat{\beta}_1\right) = \frac{\widehat{\sigma}}{\sqrt{SST_x}}$$

4. The Multivariate Linear Regression Model

2022.03.26

---

1. Motivation of Multiple Regression

1.1 Motivation

> Q：为什么要做多元回归?

- 简单回归的缺陷
  - 零条件均值假设要求所有其他影响y的因素都与x不相关，而这并不现实。[比如教育 工资以及能力]
  - 难以得到在控制其他变量的情况下x和y的关系。
- 多元回归的优势
  - 可以控制更多变量
  - 可以放入更多的相关变量
  - 更好地推断因果 [向causality更近一步]
  - 能够解释更多的变动
  - 更好的预测
  - 可利用更多的方程形式 [如交互项 二次项]

1.2 The Model with Two Independent Variables

- 普通形式方程
  - $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$
  - β_0 是截距
  - β_1 衡量x_1的变化如何影响y的变化，holding x_2 fixed;
  - β_2 衡量x_2的变化如何影响y的变化，holding x_1 fixed.
  - 注意：放入方程中的x_1和x_2可以相关
- 含有平方项的模型
  - $$cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$$
  - 由于inc和inc^2同时变化，因此在控制某一项不变没有意义。
  - β_1相当于求一阶导；β_2相当于求二阶导
  - 由于该方程的本质是一元二次方程，因此具有类似的性质：
    - 拐点[极值]是$-\frac{b}{2a}$，即$-\frac{\beta_1}{2\beta_2}$
    - β_2决定了图像的开口向上还是向下
  - 因变量的变化为
    - $$\frac{\Delta cons}{\Delta inc} = \beta_1 + 2\beta_2 inc$$
    - 衡量的是边际消费倾向
  - 这种模型在实证研究中得到了广泛应用：以kuznets curve为例，该曲线描述了收入分配状况随经济发展过程而变化的曲线，是发展经济学中重要的概念。库兹涅茨曲线表明在经济发展过程开始的时候，尤其是在国民人均收入从最低上升到中等水平时，收入分配状况先趋于恶化，继而随着经济发展，逐步改善，最后达到比较公平的收入分配状况，呈颠倒过来的U的形状。
- 模型假设
  - x_1 x_2 以及 u 的关系
    - $$E(u \mid x_1, x_2) = 0$$

- 换言之，x_1x_2以外的因素与x_1x_2都不相关。
- 例如，如果内生能力为u，则该假设意味着教育和经验和能力无关[均值独立]，否则估计出的参数有偏。

1.3 The Model with k Independent Variables

- **多元线性回归模型**
  - $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u$$
    - β_0 为截距
    - $\beta_1, \ldots, \beta_k$ 是与 $x_1, \ldots, x_k$ 相关的参数
    - 误差项 u 包括模型中没有的所有因素
    - 模型包括 $k+1$ 个未知总体参数
- **关键假设**
  - $$E(u \mid x_1, x_2, \ldots, x_k) = 0$$
  - 这一假设意味着所有误差项都与解释变量无关，同时方程形式是正确的。

---

2. Mechanics and Interpretation of Ordinary Least Squares

2.1 Obtaining the OLS Estimators

- **OLS回归线**:
  - $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$
- **参数估计策略**：残差平方和最小化
  - $$\sum_{i=1}^{n} {\hat{u}_i}^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik})^2$$
- **k+1阶导数**:
  - $$\begin{aligned}
\frac{\partial \sum_{i=1}^{n} \widehat{u}_i^2}{\partial \widehat{\beta}_0} &= -2 \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots - \widehat{\beta}_k x_{ik} \right) = 0 \\
\frac{\partial \sum_{i=1}^{n} \widehat{u}_i^2}{\partial \widehat{\beta}_1} &= -2 \sum_{i=1}^{n} x_{i1} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots - \widehat{\beta}_k x_{ik} \right) = 0, \\
\frac{\partial \sum_{i=1}^{n} \widehat{u}_i^2}{\partial \widehat{\beta}_2} &= -2 \sum_{i=1}^{n} x_{i2} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots - \widehat{\beta}_k x_{ik} \right) = 0 \\
&\vdots \\
\frac{\partial \sum_{i=1}^{n} \widehat{u}_i^2}{\partial \widehat{\beta}_k} &= -2 \sum_{i=1}^{n} x_{ik} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots - \widehat{\beta}_k x_{ik} \right) = 0
\end{aligned}$$
- **重要性质**
  - $\sum \hat{u}_i = 0$，样本的残差均值为0，
    - $$\bar{y} = \bar{\hat{y}}$$
  - 自变量及其OLS残差的协方差为0，即
    - $$cov(x_{ij}, \hat{u}_i) = 0$$
    - $$cov(\hat{y}_i, \hat{u}_i) = 0$$
  - 点 $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, \bar{y})$ 总在回归线上：
    - $$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \cdots + \hat{\beta}_k \bar{x}_k$$

- **系数解读**:
  - OLS回归线兼样本回归函数:
    - $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$
  - $\hat{\beta}_0$是所有自变量都为0时y的预测值
  - 对于每一个样本而言，给定全部信息，预测值为
    - $\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \ldots + \widehat{\beta}_k x_{ik}$
  - 每个样本的残差估计值为
    - $$\hat{u_i} = y_i - \hat{y}_i$$
  - 局部影响 [*partial effect*]
    - $$\Delta\hat{y} = \widehat{\beta}_1 \Delta x_1 + \widehat{\beta}_2 \Delta x_2 + \ldots + \widehat{\beta}_k \Delta x_k$$
    - 控制其他因素不变，即$\Delta x_2 = \Delta x_3 = \ldots = \Delta x_k = 0$
      - $$\Rightarrow \quad \Delta\hat{y} = \widehat{\beta}_1 \Delta x_1$$

2.2 A "Partialling Out" Interpretation of Multivariate Regression

- 做回归时，如何"Partialling Out" 其他因素的影响?
  - 1st Stage: reg x_{i1} on $x_{i2} \ldots x_{ik}$，得到$\hat{r_{i1}}$
    - $$x_{i1} = \widehat{\gamma}_1 + \widehat{\gamma}_2 x_{i2} + \widehat{\gamma}_3 x_{i3} + \ldots + \widehat{\gamma}_k x_{ik} + \hat{r}_{i1} = \widehat{x}_{i1} + \hat{r}_{i1}$$
    - 得到的残差$\hat{r_{i1}}$即为x_1中与$x_2 \ldots x_k$不相关的部分，相当于将x_1提纯，仅留下与其他变量不相关的部分。
  - 2nd Stage: reg y_i on $\hat{r_{i1}}$，得到β_1
    - $$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2}$$
  - 两步法的STATA CODE
    - ```
      reg x1 x2 x3 xk
      predict rhat, residual
      reg y rhat
      ```
- **推导**:
  - 通过第一步的辅助回归 [*auxiliary regression*]，得到$x_{i1} = \widehat{x}_{i1} + \hat{r}_{i1}$，代入主回归的一阶条件中，得到
    - $$\frac{\partial \sum_{i=1}^{n} \widehat{u}_i^2}{\partial \widehat{\beta}_1} = -2 \sum_{i=1}^{n} \left(\widehat{x}_{i1} + \hat{r}_{i1}\right) \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots - \widehat{\beta}_k x_{ik}\right) = 0$$
  - 由于主回归残差与解释变量不相关，而$\hat{x_{i_1}}$是$x_{i2} \ldots x_{ik}$的线性方程，因此$\sum_{i=1}^{n} \widehat{x}_{i1} \widehat{u}_i = 0$，得到
    - $$\sum_{i=1}^{n} \hat{r}_{i1} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots \widehat{\beta}_k x_{ik}\right) = 0$$
  - 辅助回归的残差和辅助回归的解释变量$x_{i2} \ldots x_{ik}$不相关，同时$\hat{\beta}_0$为常数、残差和为0，得到
    - $$\Rightarrow \sum_{i=1}^{n} \hat{r}_{i1} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots \widehat{\beta}_k x_{ik}\right) = \sum_{i=1}^{n} \hat{r}_{i1} \left(y_i - \widehat{\beta}_1 x_{i1}\right) = 0$$
  - 代入$x_{i1} = \widehat{x}_{i1} + \hat{r}_{i1}$，由于辅助回归的拟合值与残差不相关，$\sum_{i=1}^{n} \widehat{x}_{i1} \hat{r}_{i1} = 0$，则
    -
      ```

$$\sum_{i=1}^{n} \hat{r}_{i1}\left(y_i-\widehat{\beta}_1\hat{r}_{i1}\right)=0$$

$$\Rightarrow \widehat{\beta}_1=\frac{\sum_{i=1}^{n}\hat{r}_{i1}y_i}{\sum_{i=1}^{n}\hat{r}_{i1}^2}$$

2.3 Comparison of Bivariate and Multivariate Regression Estimates

- 给定二元回归方程$\tilde{y}=\widetilde{\beta}_0+\widetilde{\beta}_1x_1$与多元回归方程$\hat{y}=\widehat{\beta}_0+\widehat{\beta}_1x_1+\widehat{\beta}_2x_2$
- 辅助回归：得到$\tilde{\delta_1}$
  - $$x_{i2}=\delta_0+\delta_1x_{i1}+v_i\quad i=1,\dots,n$$
- 则多元回归系数$\hat{\beta}_1$与一元回归系数$\tilde{\beta}_1$的关系为
  - $$\widetilde{\beta}_1=\widehat{\beta}_1+\widehat{\beta}_2\tilde{\delta}_1$$
  - 换言之，简单回归方程中$x_1\to y$的影响本质上是$x_1\to y$实际上的直接影响加上$x_1\to x_2\to y$的间接影响
- **推导**：
  - $$\tilde{\beta}_1=\frac{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_1\right)\left(y_i-\bar{y}\right)}{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_1\right)^2}$$
  - 用多元回归模型的y进行替代，分子为
    - $$\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_1\right)\left(\hat{\beta}_0+\hat{\beta}_1x_{i1}+\hat{\beta}_2x_{i2}+\hat{u}_i-\hat{\beta}_0-\hat{\beta}_1\bar{x}_1-\hat{\beta}_2\bar{x}_2\right)$$
    - $$=\hat{\beta}_1\left[\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_1\right)^2\right]+\hat{\beta}_2\left[\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_1\right)\left(x_{i2}-\bar{x}_2\right)\right]+\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_1\right)\hat{u}_i$$
  - 由于残差与被解释变量不相关，$\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_1\right)\hat{u}_i=0$, 原式可化简为
    - $$\tilde{\beta}_1=\frac{\hat{\beta}_1\left[\sum_{i=1}^{n}\left(x_{i1}-\bar{x}\right)^2\right]+\hat{\beta}_2\left[\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_1\right)\left(x_{i2}-\bar{x}_2\right)\right]}{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}\right)^2}$$
  - 由于$\tilde{\delta_1}=\frac{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_1\right)\left(x_{i2}-\bar{x}_2\right)}{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}\right)^2}$，则易得
    - $$\widetilde{\beta_1}=\widehat{\beta}_1+\widehat{\beta}_2\tilde{\delta}_1$$
- **在什么情况下**$\widetilde{\beta_1}=\widehat{\beta_1}$**？**
  - $\hat{\beta}_2=0$，即 $x_2\nRightarrow y$
  - $\tilde{\delta_1}=0$，即 $x_2\nLeftrightarrow x_1$
  - 即，需要控制住的是既与y有关，又与关键自变量x_1相关的变量。
- **拓展**：
  - 首先正常跑回归模型，得到$\hat{\beta}_j$
  - 其次，去掉自变量x_k，得到$\tilde{\beta}_j$
  - 将自变量x_k与其他所有解释变量做回归，得到$\tilde{\delta}_j$
  - $$\Rightarrow\widetilde{\beta}_j=\widehat{\beta}_j+\widehat{\beta}_k\tilde{\delta}_j$$

2.4 Goodness-of-Fit

- $R^2\equiv\frac{SSE}{SST}$，分子分母同乘SSE：
  - $$\frac{SSE}{SST}=\frac{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}=\frac{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}\frac{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{\hat{y}}\right)^2}{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{\hat{y}}\right)^2}$$
- 给定$\bar{y}=\bar{\hat{y}}$，则

- $$\sum_{i=1}^{n}\left(\hat{y}_i-\bar{\hat{y}}\right)^2=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{\hat{y}}\right)\left[\left(y_i-\bar{y}\right)+\left(\hat{y}_i-y_i\right)\right]$$

- 由于$\hat{y}_i-y_i=-\hat{u}_i$，$\hat{u}_i$与$\hat{y}_i$不相关，同时$\sum_{i=1}^{n}\hat{u}_i=0$，则

  - $$\sum_{i=1}^{n}\left(\hat{y}_i-\bar{\hat{y}}\right)\left[\left(y_i-\bar{y}\right)-u\right]=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)\left(y_i-\bar{y}\right)$$

- 即，

  - $$\frac{SSE}{SST}=\frac{\left[\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\left(\hat{y}_i-\bar{\hat{y}}\right)\right]^2}{\left[\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2\right]\left[\sum_{i=1}^{n}\left(\hat{y}_i-\bar{\hat{y}}\right)^2\right]}=\left(\frac{\operatorname{Cov}\left(y_i,\hat{y}_i\right)}{\sqrt{\operatorname{Var}\left(y_i\right)}\sqrt{\operatorname{Var}\left(\hat{y}_i\right)}}\right)^2$$

- 综上所述，

  - $$R^2\equiv\frac{SSE}{SST}=1-\frac{SSR}{SST}=\left(\frac{\operatorname{Cov}\left(y_i,\hat{y}_i\right)}{\sqrt{\operatorname{Var}\left(y_i\right)}\sqrt{\operatorname{Var}\left(\hat{y}_i\right)}}\right)^2$$

---

3. The Expected Value of the OLS Estimators

3.1 Assumptions for Multiple Linear Regression

> Q: 使用多元回归模型需要满足哪些假定?

- **MLR.1 Linear in Parameters**
  - 总体模型/真实模型的定义：
  - $$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k+u$$
- **MLR.2 Random Sampling**
  - 样本模型
  - $$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_kx_{ik}+u_i$$
  - 其中，OLS估计量$\widehat{\beta}_0,\widehat{\beta}_1,\ldots,\widehat{\beta}_k$是真实参数$\beta_0,\beta_1,\ldots,\beta_k$的一种估计量
- **MLR.3 No Perfect Collinearity**
  - 不能有解释变量始终为常数。
  - 解释变量之间不能存在线性依存关系→ 自变量可以相关但不能是完全相关，否则会导致完全共线问题[*perfect collinearity*]
  - 样本量不能小于待估参数的数量，即$n\geq k+1$
  - **案例**：
    - $\log(\text{wage})=\beta_0+\beta_1\,\text{north}+\beta_2\,\text{west}+\beta_3\,\text{south}+\beta_4\,\text{east}+u$
    - $\log(\text{cons})=\beta_0+\beta_1\log(\text{inc})+\beta_2\log\left(\text{inc}^{\,2}\right)+u$
  - 完全共线的解决方法：去掉β_0或去掉一个自变量。
- **MLR.4 Zero Conditional Mean**
  - 残差项与自变量应该无关，即$E\left(u\mid x_1,x_2,\ldots,x_k\right)=0$
  - **假设4不成立的四种情况**：
    - **方程形式误设**：该加**对数**没加对数；该加**交互项**没加交互项；该加**平方项**没加平方项
    - **重要变量遗失**：一个与因变量以及自变量**同时相关**的变量未被纳入模型之中。
    - **测量误差**：**自变量**存在测量误差，如社会期许偏差，导致测不准。
    - **反向因果**：自变量同时被y所影响
  - 如果MLR.4成立，则称该变量外生 [*exogenous*]；如果MLR.4不成立，自变量与残差相关，则称该变量内生 [*endogenous*]

3.2 Unbiasedness of the OLS Estimators

- **Theorem 3.1: Unbiasedness of OLS**
- 当假设MLR.1-MLR.4成立时，OLS估计量$\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$是总体参数$\beta_0, \beta_1, \ldots, \beta_k$的无偏估计，即
  - $$E\left(\widehat{\beta}_j\right)=\beta_j, \quad j=0,1, \ldots, k$$
- **推导**:
  - 1st stage：分解$\hat{\beta}_j$，得到真值与随机扰动项
    - 已知 $\hat{\beta}_j=\frac{\sum_{i=1}^n \hat{r}_{ij} y_i}{\sum_{i=1}^n \hat{r}_{ij}^2}$ ，且根据MLR.1有 $y_i=\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik}+u_i$，则不难得到
      - $$\hat{\beta}_j=\frac{\sum_{i=1}^n \hat{r}_{ij}\left(\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik}+u_i\right)}{\sum_{i=1}^n \hat{r}_{ij}^2}$$
    - 由于$\hat{r_{ij}}$与x_j以外的解释变量不相关，$\sum_{i=1}^n \hat{r_{ij}} x_i=0$；同时$\beta_0$为常数，则进一步化简，得到
      - $$\hat{\beta}_j=\frac{\sum_{i=1}^n \hat{r}_{ij} y_i}{\sum_{i=1}^n \hat{r}_{ij}^2}=\frac{\sum_{i=1}^n \hat{r}_{ij}\left(\beta_j x_{ij}+u_i\right)}{\sum_{i=1}^n \hat{r}_{ij}^2}$$
    - 由于$\hat{r_{ij}}$与拟合值x_j不相关，同时$x_{ij}=\hat{x_{ij}}+\hat{r_{ij}}$则
      - $$\hat{\beta}_j=\frac{\sum_{i=1}^n \hat{r}_{ij}\left(\beta_j x_{ij}+u_i\right)}{\sum_{i=1}^n \hat{r}_{ij}^2}=\frac{\sum_{i=1}^n \hat{r}_{ij}\left(\beta_j \hat{x_{ij}}+\beta_j \hat{r_{ij}}+u_i\right)}{\sum_{i=1}^n \hat{r}_{ij}^2}=\beta_j+\frac{\sum_{i=1}^n \hat{r}_{ij} u_i}{\sum_{i=1}^n \hat{r}_{ij}^2}$$
  - 2nd stage：求期望
    - 在给定X后，$\hat{r_{ij}}$为常数，结合迭代法则：
    - $$E\left(\hat{\beta}_j \mid X\right)=\beta_j+\frac{\sum_{i=1}^n \hat{r}_{ij} E\left(u_i \mid X\right)}{\sum_{i=1}^n \hat{r}_{ij}^2}=\beta_j$$

3.3 Including Irrelevant Variables in a Regression Model

> Q: 在回归模型中加入无关变量是否影响无偏性？

- 如果加入了无关变量，即模型存在过度识别的问题[*overspecified*]
  - 其他变量的参数$\hat{\beta}$的无偏性不受影响
  - OLS估计量的方差变大，估计的有效性下降。
- **推导**:
  - 假定真实模型为$\tilde{y}=\tilde{\beta}_0+\tilde{\beta}_1 x_1+\tilde{\beta}_2 x_2$
  - 过度识别的模型为$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\hat{\beta}_3 x_3$，其中$\hat{\beta}_3=0$
  - 则$\tilde{\beta}_1=\hat{\beta}_1+\hat{\beta}_3 \tilde{\delta}_1$
  - 由于$\hat{\beta}_1$无偏，且给定自变量后$\tilde{\delta}_1$为常数，则
    - $$\begin{aligned} E\left(\tilde{\beta}_1\right) & =E\left(\hat{\beta}_1\right)+E\left(\hat{\beta}_3 \tilde{\delta}_1\right) \\ & =E\left(\hat{\beta}_1\right)+\tilde{\delta}_1 E\left(\hat{\beta}_3\right)=E\left(\hat{\beta}_1\right)=\beta_1 \end{aligned}$$

3.4 Omitted Variable Bias

> Q: 回归模型中如果遗漏有关变量是否影响无偏性？

- 如果一个相关的变量未被纳入模型中，则该模型识别不足 [*underspecified*]
  - 遗漏变量会导致OLS估计量有偏
- **推导**:
  - 假定真实模型为$y=\beta_0+\beta_1 x_1+\beta_2 x_2+u$
  - 识别不足的模型为$\tilde{y}=\widetilde{\beta}_0+\widetilde{\beta}_1 x_1$

- 则$\widetilde{\beta}_1 = \widehat{\beta}_1 + \widehat{\beta}_2\tilde{\delta}_1$
- 由于$\hat{\beta}_1\hat{\beta}_2$无偏，且给定自变量后$\tilde{\delta}_1$为常数，则
  - $$E\left(\widetilde{\beta}_1\right) = E\left(\widehat{\beta}_1 + \widehat{\beta}_2\tilde{\delta}_1\right) = E\left(\widehat{\beta}_1\right) + E\left(\widehat{\beta}_2\right)\tilde{\delta}_1 = \beta_1 + \beta_2\tilde{\delta}_1$$

- **遗漏变量偏误**:
  - $$\Rightarrow \text{Bias}\left(\widetilde{\beta}_1\right) = E\left(\widetilde{\beta}_1\right) - \beta_1 = \beta_2\tilde{\delta}_1$$
  - 如果遗漏变量不影响因变量$\beta_2 = 0$，或x_1与x_2不相关，则参数依然是无偏的。
- **偏误方向**:
  -

	$\text{Corr}\,(x_1, x_2) > 0$	$\text{Corr}\,(x_1, x_2) < 0$
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

  - $E(\tilde{\beta}_1) > \beta_1$，**上偏** [*upward bias*]
  - $E(\tilde{\beta}_1) < \beta_1$，**下偏** [*downward bias*]
  - 如果相较于真值，估计量更靠近0，则低估[*underestimate*]了变量的影响；不向0偏的，则高估[*overestimate*]了变量的影响。
  - 案例:
    - 能力正向影响教育水平与工资，上偏，远离0，则遗漏能力会高估教育对工资的正面影响
    - 教育负向影响吸烟、正向影响健康水平，下偏，远离0，则遗漏教育会高估吸烟对健康的负面影响、
- **推广**:
  - 如果模型中有多个自变量，遗漏变量x_3与x_1相关，与x_2不相关，$\tilde{\beta}_1$一定有偏，$\tilde{\beta}_2$是否有偏?
  - 除非x_1和x_2无关，否则$\tilde{\beta}_2$有偏。
  - 证明:
    - 辅助回归中$\tilde{x}_3 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1 + \tilde{\delta}_2 x_2$，由于$x_1 x_2$有关，因此尽管$\tilde{\delta}_2$不会太大，但$\tilde{\delta}_2 \neq 0$.
    - 则$\tilde{\beta}_2 = \widehat{\beta}_2 + \widehat{\beta}_3\tilde{\delta}_2$中的$\widehat{\beta}_3\tilde{\delta}_2 \neq 0$，参数有偏。
  - 同理，如果回归方程中有k个变量，只要遗漏变量与其中一个x_j有关，则所有与x_j相关的变量的系数均有偏。

4.The Variance of the OLS Estimators

> Q: **如何衡量抽样分布的分散性，即参数估计的有效性?**

4.1 Assumption MLR.5: Homoscedasticity

- 为衡量参数估计的有效性，需要引入新的假设"同方差假设" [*Homoscedasticity*]:
  - 给定解释变量的值，误差项u的方差相同。即x变化，方差也不会发生变化
  - $$\text{Var}\,(u \mid x_1, \ldots, x_k) = \sigma^2$$
- **异方差性**:
  - 随着教育程度的提高，收入分布的方差扩大，则为异方差性。在截面数据中，异方差性十分常见。
  - 参见Lecture 8
- 根据MLR.5，不同观测值的残差不相关，即 $\text{cov}\,(u_i, u_j) = 0 (i \neq j)$$
- 给定MLR.1-MLR.4，我们可以得到
  - $$E(y \mid \vec{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$
- 加入MLR.5，我们可以得到
  - $$\text{Var}(y \mid \vec{x}) = \sigma^2$$

- 即，给定一个样本所有自变量的信息，y的方差是相同的

4.2 Sampling Variances of the OLS Estimators

- **Theorem 3.2 Sampling Variances of the OLS Slope Estimators**
- 给定MLR.1-MLR.5，多元回归模型参数的方差为：
  - $$\operatorname{Var}\left(\widehat{\beta}_j\right)=\frac{\sigma^2}{SST_j\left(1-R_j^2\right)} \quad \text{for } j=1,2,\ldots,k$$
    - $SST_j=\sum_{i=1}^n\left(x_{ij}-\bar{x}_j\right)^2$为$x_j$的总变动
    - R_j^2 是reg x_j on 其他变量，得到的R^2
  - 简单线性回归模型中，参数的方差为$\operatorname{Var}\left(\widehat{\beta}_j\right)=\frac{\sigma^2}{SST_x}$，即为$R_j=0$时的特例。[*reg on nothing*]
- **推导**：
  - 分解$\hat{\beta}_1$
    - $$\hat{\beta}_1=\beta_1+\frac{\sum_{i=1}^n \hat{r}_{i1}u_i}{\sum_{i=1}^n \hat{r}_{i1}^2}$$
  - 两边同时求方差，与无偏性的证明类似，
    - $$Var(\hat{\beta}_1)=\frac{\sum_{i=1}^n \hat{r}_{i1}^2\operatorname{Var}\left(u_i \mid X\right)}{\left(\sum_{i=1}^n \hat{r}_{i1}^2\right)^2}=\frac{\sum_{i=1}^n \hat{r}_{i1}^2\sigma^2}{\left(\sum_{i=1}^n \hat{r}_{i1}^2\right)^2}=\frac{\sigma^2}{\sum_{i=1}^n \hat{r}_{i1}^2}$$
  - 因 $\sum_{i=1}^n \hat{r}_{i1}^2$ 是 x_1 对 $x_2,\cdots,x_k$ 进行回归的残差平方和, 所以
    - $$\sum_{i=1}^n \hat{r}_{i1}^2=\mathrm{SSR}=\mathrm{SST}_1\left(1-R_1^2\right)$$
- **OLS统计量方差的组成部分**
  - σ^2：误差项方差越大，统计量方差越大。总体特征，我们只能通过加入更多的相关变量减少误差项方差。
  - SST_j：样本变动越大，统计量方差越小→ 可以通过扩大样本量扩大SST
  - R_j^2：R_j^2代表了x_j解释其他自变量变动的能力；自变量之间的线性关系越强，统计量方差越大。
    - $\operatorname{Var}\left(\widehat{\beta}_j\right) \rightarrow \infty$, when $R_j^2 \rightarrow 1$
    - 如果$R_j^2=1$，则会违背MLR.3 → 完全共线问题
    - R_j^2距离1越近，则多重共线问题越严重。[*multicollinearity*]

4.3 Multicollinearity

- 多重共线并没有一个明确的指标进行界定。
- 通过**收集更多数据**以及**去掉部分变量**，多重共线性问题会得到缓解→ 存在变量数量与估计效率之间的权衡
  - 一元回归模型的估计量方差一定小于多元回归模型。
- 但我们最关注的是无偏性而非有效性，因此必须优先保证相关变量处在模型之中、再追求效率。
- 注意：部分控制变量之间高度相关不影响关键自变量估计的有效性。

---

5. Estimating the Error Variance

> Q: **由于误差项的方差是未知的，我们如何对其进行估计？**

- 由于在估计$k+1$个参数时已经损失了$k+1$个自由度，因此利用$\hat{u}_i$估计误差项方差的公式为：
  - $$\widehat{\sigma}^2=\frac{1}{(n-k-1)}\sum_{i=1}^n \widehat{u}_i^2=\frac{1}{(n-k-1)}SSR$$

- **Theorem 3.3: Under GM assumptions MLR.1-5**
  - $$E\left(\hat{\sigma}^2\right)=\sigma^2$$
- **OLS参数的标准差:**
  - $$\operatorname{sd}\left(\widehat{\beta}_j\right)=\sqrt{\operatorname{Var}\left(\widehat{\beta}_j\right)}=\sqrt{\frac{\sigma^2}{SST_j\left(1-R_j^2\right)}}$$
- 用σ^2代替真值，得到**OLS参数的标准误**:
  - $$\operatorname{se}\left(\widehat{\beta}_j\right)=\sqrt{\frac{\widehat{\sigma}^2}{SST_j\left(1-R_j^2\right)}}$$
- 注：如果不满足MLR.5，即存在**异方差**问题，则该标准误不正确。

---

6. Efficiency of OLS: The Gauss-Markov Theorem

- **Theorem 3.4: Gauss-Markov Theorem**
- 在假设MLR. 1-5成立的情况下，OLS参数 $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$ 是真实参数 $\beta_0, \beta_1, \ldots, \beta_k$ 的最优无偏估计量 [**B**est **L**inear **U**nbiased **E**stimators (“BLUEs”)]
- 所谓的“Best”，指的是相比于其他任何参数，OLS参数最有效率、方差最小
  - $$\operatorname{Var}(\widetilde{\beta}) \geq \operatorname{Var}(\widehat{\beta})$$
- 高斯-马尔科夫定理证明了OLS是估计多元回归模型的最优策略→ 只要MLR.1-5成立，我们就不必选择其他无偏估计。
- 如果高斯-马尔科夫定理不成立，则OLS不是BLUE的估计量:
  - MLR.4 不成立→ OLS估计量有偏
  - MLR.5 不成立→ OLS估计量不再是最有效的估计量

5. Inference

2022.04.09

---

1. Sampling Distributions of the OLS Estimators

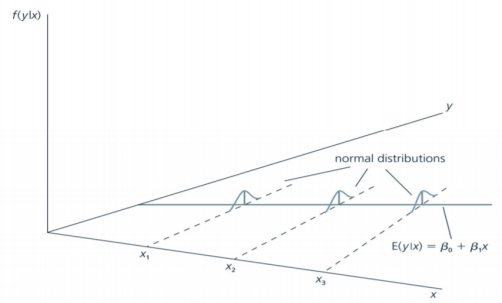
1.0 Overview

- 高斯-马尔科夫假设共包含MLR.1-MLR.5五个假设
- MLR.1-4可以推导出期望的无偏性(定理3.1)
- MLR.1-5可以推导出OLS参数的方差(定理3.2)、对误差项的估计(定理3.3)以及"BLUE"(定理3.4)
- 为进行假设检验，必须假定样本分布的形态→ 引入**MLR.6 正态性假设**。

1.1 Normality Assumption

- OLS参数的抽样分布依赖于误差项的分布→ 对总体误差的分布做出假设
- **MLR.6 Normality Assumption**
  - 由MLR.4可知：$E(u \mid x_1, \ldots, x_k) = E(u) = 0$
  - 由MLR.5可知：$\text{Var}(u \mid x_1, \ldots, x_k) = \text{Var}(u) = \sigma^2$
  - 在此基础上，假定总体误差项u与自变量无关，同时服从于均值为0、方差为σ^2的正态分布：
    - $$u \sim N(0, \sigma^2)$$
  - MLR.6 是最强的假设。
    - 将u视为一系列影响y且无法观察到的因素的和，进而使用中心极限定理[Central Limit Theorem, CLT]，证明残差项符合正态分布→ 然而，不同的因素可能有不同的分布形态，如工资并非正态分布、而性别、地区等变量同样如此；仅取几个值的分布，如生孩子的个数，更接近泊松分布而非正态分布。
    - 同时，中心极限定理假定非观察到的变量可以简单相加→ 然而，u可能是一系列未观察到的变量的复杂函数。
    - 然而，当样本量足够大时，非正态并非一个严重的问题。→ 可以假定正态分布

1.2 Classical Linear Model Assumption

- 高斯-马尔可夫假定+MLR.6=经典线性模型假设 [CLM]
- 基于MLR.1-MLR.6得到的回归模型即为经典线性回归模型 [*classical linear regression model*]
- 在该假定下，给定$\vec{x}$，y服从于正态分布
  - $$y \mid (x_1, \ldots, x_k) \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k, \sigma^2)$$
- MLR.5与MLR.6针对的是误差项，但由于y的变动来源于误差项的扰动，因此对误差项做出的假设可以推广到因变量之中。
- 基于MLR.1-4，因变量的期望落在总体回归线上。
- 同方差正态分布二维图像
  - 

1.3 Normality Sampling Distributions for OLS estimators

- **Theorem 4.1 Normal Sampling Distributions**
- 在MLR.1-MLR.6成立的情况下，$\widehat{\beta}_j \sim N\left(\beta_j, \operatorname{Var}\left(\widehat{\beta}_j\right)\right)$，因此
  - $$\frac{\widehat{\beta}_j-\beta_j}{sd\left(\widehat{\beta}_j\right)} \sim N(0,1)$$
- 证明：
  - $$\hat{\beta}_j=\beta_1+\sum_{i=1}^n \frac{\hat{r}_{ij}}{\hat{r}_{ij}^2} u_i$$
  - 根据MLR.6，u_i服从正态分布。根据中心极限定理，它的和也服从正态分布。
- 同时，$\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$的线性组合同样服从正态分布。

---

2. Testing Hypotheses about a Single Population Parameter: The t Test

2.1 t Distribution for Standardized Estimators

- **Theorem 4.2 t Distribution for Standardized Estimators**
- 在MLR.1-6成立的情况下，
  - $$\frac{\widehat{\beta}_j-\beta_j}{\operatorname{se}\left(\widehat{\beta}_j\right)} \sim t_{n-k-1}$$
  - 注意：自由度为$n-k-1$
- 定理4.1说明以标准差[未知常数]为分母，服从正态分布；定理4.2说明以标准误[随机变量]为分母，服从t分布。

2.2 t Test

> Q: 如何对单个参数进行检验？

- 案例：
  - 回归方程$\log(\text{ wage }) = \beta_0 + \beta_1 \text{ educ } + \beta_2 \text{ exper } + \beta_3 \text{ tenure } + u$
  - 零假设$H_0: \beta_2 = 0$若成立，意味着你的工资不取决于你在接受这个工作之前工作了多少年。如果原假设成立，则会降低换工作的积极性。
- **t统计量**[*t statistic/ t ratio*]
  - 我们一般使用t值验证原假设是否成立；同时，我们通常将a_j设定为0
    - $$t_{\widehat{\beta}_j} \equiv \frac{\widehat{\beta}_j-\beta_j}{\operatorname{se}\left(\widehat{\beta}_j\right)}=\frac{\widehat{\beta}_j-a_j}{\operatorname{se}\left(\widehat{\beta}_j\right)}$$
    - t值越大，距离a_j的距离越远，越能够拒绝原假设。
  - 标准误恒为正数，因此t值与$\hat{\beta}_j$同号。
- **原假设与备择假设**
  - 我们检验的是未知的总体参数，而非已知估计值，因此原假设[*null hypothesis*]并非$\hat{\beta}_1 = 0$，而是
    - $H_0: \beta_1 = 0$
  - 备择假设[*Alternative hypotheses*]可分为三类：
    - $H_1: \beta_j > a_j$ [*one-sided hypothesis*]
    - $H_1: \beta_j < a_j$ [*one-sided hypothesis*]
    - $H_1: \beta_j \neq a_j$ [*two-sided hypothesis*]

- **显著性水平与临界值**
  - 显著性水平[*Significance level*]α指代的是犯错误（弃真）的可能性，因此越小越好。
  - 常见的显著性水平包括10%，5%以及1%
  - 在确定了显著性水平后，t分布的第$1-\alpha$百分位的值即为临界值[*Critical value*]。临界值取决于
    - 显著性水平→ 一般设定为5%
    - 备择假设→ 确定单边/双边检验→ 单边检验更容易显著，因此默认选择双边检验。
    - 自由度：$n-k-1$
- **t检验的步骤**
  1. 写出原假设与备择假设
  2. 决定显著性水平并找到对应的临界值：n>120时。5%的置信水平下，单边检验临界值为1.645，双边检验为1.96]
  3. 基于样本数据，计算t值。
  4. 比较t值与临界值，决定是否拒绝原假设：如果t值的绝对值更大，则可以拒绝原假设
  - 如果我们拒绝原假设，则一般表述为"在α%的显著性水平上，自变量统计显著"
  - 如果我们无法拒绝原假设，则一般表述为"在α%的显著性水平上，自变量统计不显著"
- **p-values**
  - p值[*P-values*]指代的是双边检验中t值所对应的百分比
- **经济显著与统计显著**
  - 统计显著性[*Statistical Significance*]仅仅由t值决定。
  - 经济显著性[*Economic Significance*]则由待估参数的大小决定。
  - 我们希望找到一个既统计显著、又具有经济意义的解释变量。

---

3. Confidence Intervals

> Q: 如何进行区间估计?

- 建构置信区间？[*confidence interval* (CI)]
  - $$\widehat{\beta}_j \pm c \cdot \operatorname{se}\left(\widehat{\beta}_j\right)$$
- 95%置信区间的含义：抽样无数次，落在置信区间的概率为95%。
- 当 $n-k-1>120$时，95%的置信区间为$\left[\widehat{\beta}_j-1.96 \cdot \operatorname{se}\left(\widehat{\beta}_j\right), \widehat{\beta}_j+1.96 \cdot \operatorname{se}\left(\widehat{\beta}_j\right)\right]$

---

4. Testing Hypotheses about a Single Linear Combination of the Parameters

> Q: 如何对若干参数的单个线性组合进行检验?

- **零假设与备择假设**:
  - $H_0: \beta_1-\beta_2=0$
  - $H_1: \beta_1-\beta_2<0$
- **构建t值**
  - $$t=\frac{\widehat{\beta}_1-\widehat{\beta}_2}{\operatorname{se}\left(\widehat{\beta}_1-\widehat{\beta}_2\right)}$$
  - $$\begin{aligned}\operatorname{Var}\left(\widehat{\beta}_1-\widehat{\beta}_2\right) &= \operatorname{Var}\left(\widehat{\beta}_1\right)+\operatorname{Var}\left(\widehat{\beta}_2\right)-2 \operatorname{Cov}\left(\widehat{\beta}_1, \widehat{\beta}_2\right) \\ \Rightarrow \operatorname{se}\left(\widehat{\beta}_1-\widehat{\beta}_2\right) &= \sqrt{\operatorname{Var}\left(\widehat{\beta}_1-\widehat{\beta}_2\right)}.\end{aligned}$$
  - 计算协方差的STATA CODE

```
matrix list e(V)
```

- **替代方案**
  - 定义新参数：$\theta_1 = \beta_1 - \beta_2$
  - 假设：$H_0 : \theta_1 = 0$ against $H_1 : \theta_1 < 0$
  - t值：$t = \frac{\hat{\theta}_1}{\text{se}\left(\hat{\theta}_1\right)}$
  - 回归方程
    - $$\begin{aligned}\log(\text{ wage }) &= \beta_0 + (\theta_1 + \beta_2)jc + \beta_2 \text{ univ } + \beta_3 \text{ exper } + u \\ &= \beta_0 + \theta_1 jc + \beta_2(jc + \text{ univ }) + \beta_3 \text{ exper } + u\end{aligned}$$
    - 直接看θ_1的p值即可判断是否统计显著。
  - jc前的系数θ_1衡量的是二者的差额；前面的原$\beta_1 = \theta_1 + \beta_2$；$totcoll$前的系数$\beta_2$的解读不发生变化→ 变量名字变了的系数不变，变量名字没变的系数变了。

---

5. Testing Multiple Linear Restrictions

> Q: **如何检验一组自变量是否会影响因变量?**

- 原假设：$H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0.$
- 备择假设：$H_1 : H_0$ is not true.只要有一个不为0即可。
- 如果这一组自变量高度相关，存在多重共线问题，则会**使得分别的t检验失效**→ 使用**F检验**，观察SSR的变化是否足够大。
- 代入原假设，得到受限模型[*restricted model*]，跑回归得到SSR_r
- 将所有变量代入模型，得到非受限模型[*unrestricted model*]，跑回归得到SSR_{ur}
- **构建F值**
  - $$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \sim F_{q,n-k-1}$$
    - q为受限制的变量的个数，$n-k-1$为分母的自由度。
    - F值恒为正数。
  - 由于$SSR_r = SST\left(1 - R_r^2\right)$ 且 $SSR_{ur} = SST\left(1 - R_{ur}^2\right)$，则
    - $$F = \frac{\left(R_{ur}^2 - R_r^2\right)/q}{(1 - R_{ur}^2)/(n-k-1)} = \frac{\left(R_{ur}^2 - R_r^2\right)/q}{(1 - R_{ur}^2)/df_{ur}}$$
    - 注意：使用该公式**必须保证因变量相同**。
- 如果能够拒绝原假设，则相关变量**联合显著** [*jointly significant*]
- 如果不能够拒绝原假设，则相关变量**联合不显著** [*jointly insignificant*]
- STATA CODE

```
regress y x1 x2 x3 x4 x5
test x3=x4=x5=0
```

- 性质：对单个自变量进行F检验等同于进行双边t检验。

> Q:**如何检验模型整体显著性?**

- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$；$H_1$：至少有一个$\beta_j$显著不为0。
- 限制性模型：$y = \beta_0 + u$
- F值的计算公式为

- $$F=\frac{R^2/k}{(1-R^2)/(n-k-1)}$$

> Q: 如果原假设更复杂?

- $H_0:\beta_1=1,\beta_2=0,\beta_3=0,\beta_4=0$；$H_1$: not H_0
- 限制性模型：$y=\beta_0+x_1+u$，系数固定，无法估计，因此将模型改写为$z=y-x_1$
  - 由于因变量发生了改变，因此只能使用SSR计算
- STATA CODE

```
gen z=y-x1
reg z
reg y x1 x2 x3 x4
test x1-1=x2=x3=0
```

6. Further Issues

2022.04.16

---

1. Effects of Data Scaling on OLS Statistics

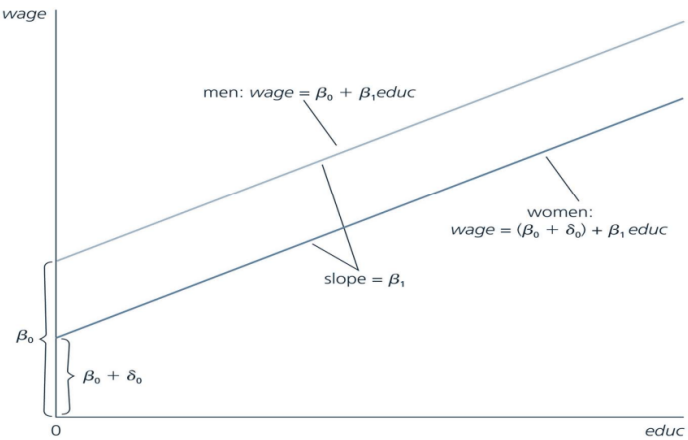
- 我们已经讨论过改变单位对OLS斜率截距系数的影响
  - $y*c \rightarrow$所有系数均$*c$
    - $cy = cax + cb$
  - $x_2 * c \rightarrow \hat{\beta}_2 * \frac{1}{c}$，其他系数不变。
    - $y = a \cdot \frac{1}{c} \cdot cx + b$
  - R^2不变

> Q: **变量单位的变化如何影响显著性**?

- **改变因变量的单位**：令y 扩大c倍
  - t值：不变
  - p值：不变
  - 置信区间：扩大c倍
  - SSE SSR SST：扩大c^2倍
  - MSE MSR MST：扩大c^2倍
  - F值 [*overall significance*]：不变
  - $\hat{\sigma}$：扩大c倍
- **改变自变量的单位**：令x 缩小c倍
  - 系数、标准误与置信区间**扩大**c倍
  - 其他指标不发生改变

---

2. Standardized Coefficients

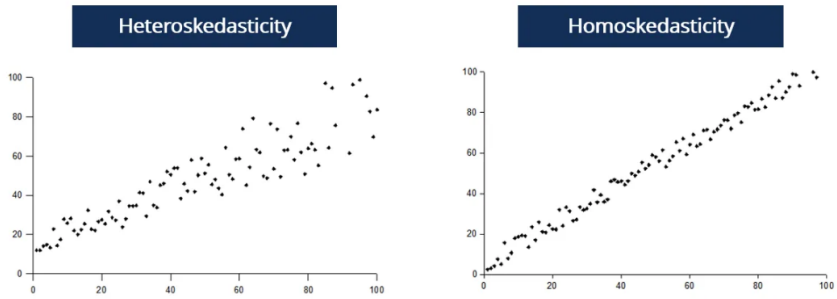
- 将系数**标准化**的方法：$\frac{\text{原始值-均值}}{\text{标准差}}$
- 将OLS方程$y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \ldots + \widehat{\beta}_k x_{ik} + \widehat{u}_i$标准化，得到
  - $$z_y = \hat{b}_1 z_1 + \hat{b}_2 z_2 + \ldots + \hat{b}_k z_k + \text{ error}$$
  - $$\widehat{b_j} = \left(\frac{\widehat{\sigma}_j}{\widehat{\sigma}_y}\right)\widehat{\beta}_j \quad \text{for } j = 1, \ldots, k$$
    - 相当于因变量的单位缩小了$\widehat{\sigma}_y$倍，自变量的单位缩小了$\widehat{\sigma}_j$倍。
- STATA CODE
  - ```
    reg y x1 x2,beta
    ```
- **系数解释**：x增加一个标准化，y变化$\hat{b}$个标准差
- 将系数标准化去掉了单位，使得不同自变量的影响力可以进行比较：**系数越大，影响越大**。
- 标准化回归方程没有截距项。
- 标准化本质上等同于换单位，因此不影响统计显著性→ t值不变。
- 为区分起见，标准化后的变量名称前 + “z”
    ```

3. Functional Forms

> Q: **如何在线性的框架衡量x与y的非线性关系**?

- 对x或y取**自然对数**
- 将x**平方** [*quadratic*]
- 对自变量做**交互项**

3.1 Logarithmic Functions

- 在一个log-level模型中，$\%\Delta y \approx 100\Delta \log(y)$随着$\log(y)$的扩大而变得更不准确。
- 更精确的百分比变化为$\%\Delta \hat{y} = 100 \cdot \left[\exp\left(\widehat{\beta}_j \Delta x_j\right) - 1\right]$
- 然而，由于增加或减少一个单位带来的系数变化有差异，而粗略估计的数值处于二者之间，因此如果没有特殊要求，使用近似数值即可。
- **优势**:
  - 解释更有经济学含义
  - 可以去掉单位
  - 使分布更接近于正态，缓解异方差以及左偏/右偏问题
  - 缩小变量的取值范围，使模型对离群值更不敏感
- **适用范围**
  - 大的正数，如企业销售量、工资、人数等等经常取对数
  - 与年限有关的，如教育年度、年龄等等不取对数。[日常生活中通常说教育增加1年，而非教育程度增加10%]
  - 与比率有关的概念，如失业率、基尼系数等等，是否取对数均可，但解读上有所不同
    - level：失业率增加1个百分点、犯罪率增加β个百分点
    - log：失业率增加1%，犯罪率增加β%→ 一般不取对数
- **局限性**
  - 如果变量有零值或负值，则无法取对数。
  - → 如果仅有少量零值存在，则可以取$\log(1+y)$

3.2 Quadratic Functions

- $y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$ 的拐点为
  - $$\Rightarrow x = -\frac{\widehat{\beta}_1}{2\widehat{\beta}_2}$$
- 抛物线图像的形成未必完全代表实际:
  - 存在outlier，扭转了原本的趋势 [部分德高望重、经验丰富的老同志工资很低]→ 可以忽略
  - 存在遗漏变量，导致抛物线两边的样本存在系统性差异 [如文革影响了老同志们的教育情况、进而影响工资]

3.3 Models with Interaction Terms

- 二次项本质是自己与自己的交互
- 交互项模型
  - $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$$
- 在交互项模型中β_2 是当$x_1 = 0$ 时x_2 对y 的偏效应→ 系数解释可能没实际意义→ 重新设定方程
- **Demean**处理：**交互项减去均值**
  - $$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + u$$
  - **系数解读**:

        - $\delta_1 = \beta_1 + \beta_3\mu_2$ 或 $\hat{\delta}_1 = \widehat{\beta}_1 + \widehat{\beta}_3\bar{x}_2$：当$x_2$取均值时，$x_1$对$y$的影响
        - $\delta_2 = \beta_2 + \beta_3\mu_1$ 或 $\hat{\delta}_2 = \widehat{\beta}_2 + \widehat{\beta}_3\bar{x}_1$：当$x_1$取均值时，$x_2$对$y$的影响
        - → 更加有意义，同时也更容易显著。
    - **推导**：两种方程的效应相等，则
        - $$\beta_1 + \beta_3 x_2 = \delta_1 + \beta_3 x_2 - \beta_3\mu_2 \Rightarrow \delta_1 = \beta_1 + \beta_3\mu_2$$
    - 注：如果我们只关心x_1 [出勤率] 对y [成绩]的影响，可以只对x_2 [原始GPA]demean
- 交互项很可能影响显著性→ 多重共线→ 如果t检验不显著，可以尝试F检验

---

4. More on Goodness-of-Fit

- **调整后R^2**
    - 只要加入新变量，R^2就会增加，但这并不意味着模型的拟合度增加。
    - 调整后R^2 通过调整自由度对这一问题进行了修正，k增加，$\bar{R}^2$下降：
        - $$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$$
- **关系**：
    - 公式：
        - $$\bar{R}^2 = 1 - \frac{\left(1-R^2\right)(n-1)}{(n-k-1)}$$
    - $\bar{R}^2$ 通常小于R^2。当样本量较小而变量数较多时，差距拉大；当样本量足够大时，二者差距并不明显。
    - $\bar{R}^2$有可能是负值→ 说明模型拟合度极差。
- **局限性**：
    - 不能告诉我们变量是否统计显著
    - 不能告诉我们X是否是导致Y的真正原因
    - 不能告诉我们是否存在遗漏变量偏误
- **应用**：选择模型
    - 如果两个方程为非嵌套模型 [*Non-Nested Models*]，即两个模型均包括对方所没有的变量、并非子集的关系，则可以通过R^2 的大小进行判断。
        - 注：如果两个模型为嵌套模型，则使用F检验。
    - 如果两个方程形式存在差异 [对数vs平方项]，则可以通过比较$\bar{R}^2$判断哪个模型更符合实际
        - 注：该方法仅适用于自变量形式存在差异。如果因变量的形式存在差异[对数vs非对数]，则不能通过R^2进行比较。

7. Dummy Variables

2022.04.30

---

1. A Single Dummy Independent Variable

1.1 Describing Qualitative Information

> Q: **如何描述定性型信息**，如性别、婚姻状态、种族、职业等。

- **解决方法**：设置二元变量/虚拟变量[*binary variable/ dummy variable*]。因为通常选择0和1作为两个取值[方便解释回归模型的系数]，因此又称01变量。
  - 注：一般而言，变量名即为取值为1时所代表的状态。
    - variable name = female, so 1= women 0=men
    - variable name = married, so 1= married 0=unmarried
  - 在建立回归模型时，定量型变量前面的系数为β，而定性型变量前面的系数为δ.
- **案例**：工资、性别与教育程度
  - $$\text{wage} = \beta_0 + \delta_0 \text{ female} + \beta_1 \text{ educ} + u$$
  - 总体回归线$E(\text{wage}|\text{female,educ}) = \beta_0 + \delta_0 female + \beta_1 educ$
    - female=1 $E(\text{wage}|\text{female,educ}) = \beta_0 + \delta_0 + \beta_1 educ$
    - female=0 $E(\text{wage}|\text{male,educ}) = \beta_0 + \beta_1 educ$
  - 则 δ_0=$E(\text{wage}|\text{female,educ}) - E(\text{wage}|\text{male,educ})$
  - δ_0代表男性和女性平均工资的差，即工资的性别差距[*gender wage gap*]，代表了给定教育程度的情况下，不同性别工资的差异有多大。
  - $\delta_0 < 0$ 经常被解读为歧视[*discrimination*]，但是discrimination和gender wage gap之间存在差异→ 什么情况下可以把差异定义为歧视？
    - 在经济学中，歧视指的是两个人**所有其他方面一样**，但是被给予了不同的待遇，而这**仅仅**是因为性别。即，歧视必须保证其他所有特征都是一样的。
    - 如果仅仅是控制了部分特征，就是wage gap。
    - 歧视包含价值判断，因此更加严格，从差别得到歧视是非常困难的。
  - 在这个例子中，只有零条件均值假设[假设4]成立时，才能说这里是歧视。如果假设4不成立，则**无法**得到“歧视”这个结论。
    - 女性所承担的家庭责任[生育]有损于工作经验的积累, 容易导致女性在劳动力市场处于不利的地位；
    - 职业的性别隔离是影响收入性别差异的重要机制, 工种的不同以及工作性质的差异也会影响到工资。
- **Graph** of $wage = \beta_0 + \delta_0 female + \beta_1 educ$ for $\delta_0 < 0$
  - 

- δ_0即为两条回归线的截距差
- 两条总体回归线平行
- 虚拟变量取0的变量，截距为β_0
- 虚拟变量取1的变量，截距为$\beta_0 + \delta_0$

1.2 Dummy Variable Trap

- 如果有截距项，且男性和女性都设置了虚拟变量，则存在**完全共线**的问题。换言之，如果产生g个分组，还包括截距，会导致**虚拟变量陷阱**[*Dummy Variable Trap*]。
- **解决方法**：
  1. ~~去掉截距项，保留两个虚拟变量~~
  2. 去掉一个虚拟变量
- 为什么去掉一个虚拟变量更好？
  - δ_0 代表着组间差，而我们关心的也是不同组别的差距，因此δ_0更具有经济学含义
    - 如果我们选择去掉截距项，得到的则是两个性别的总体回归线的截距长度。与基准组的差异难以解读与检验
  - R^2的计算方式不同。
- 被省略掉的分组取值为0，视为基准组/对照组。[*base group/ benchmark group/comparison group*]

1.3 Interpretation

- **regress y on dummy other variable**
  - 虚拟变量前的系数代表了，在给定其他变量的情况下，不同组别的平均差异。
    - 例：在给定其他变量的情况下，男女平均的工资差异是-1.81美金 [女性比男性低1.81美金]。
  - 虚拟变量的t值即代表不同组别的差距是否显著。
- **regress y on dummy ONLY**
  - 代表了在没有控制任何变量的情况下，检验两个组别的平均工资差以及是否显著 [等同于做一个t检验比较均值]
  - 系数代表了取值为1时的因变量均值-取值为0时的因变量均值，即不同组别的差值。
  - 截距代表了男性的平均工资[自变量取值为0]
- **两个回归的比较**
  - 系数=-2.51代表在不控制任何变量的情况下，女性的工资比男性低2.51美金[raw wage gap]；而-1.81指的是在加入了控制变量后双方的平均工资差[gender wage gap/difference]；如果能够穷尽其他所有控制变量，则得到的系数衡量的就是discrimination。
  - 为什么两个回归得到的系数缩小了？
    - 说明educ exper tenure可以在一定程度上解释男女的工资差. [比如男女的平均受教育程度不一样，因此影响了平均工资]

---

2. Dummy Variables and Logged Dependent Variables

> Q: **问题：当因变量取对数时如何解释系数**？

- 如果被解释变量取log，则虚拟变量的系数=$\log\widehat{(y_1)} - \log\widehat{(y_0)}$ =因变量的对数差。如果不追求精确的百分比变化，则$100 \cdot \delta_0$%即可以代表不同群体因变量的百分比变化[不考虑base group].
- 如果想计算更加精确的百分比变化，则$\frac{\widehat{(y_1)} - \widehat{(y_0)}}{\widehat{(y_0)}} = e^{\delta_0} - 1 \rightarrow 100 \cdot (e^{\delta_0} - 1)\%$
  - 注意，在计算该函数时，系数取值的正负很重要。
  - slide18说明了女性和男性的工资差异占男性收入的百分比；slide19说明了女性和男性的工资差异占女性收入的百分比，因为女性平均工资低，所以百分比的数值更大→ base group不同，得到的数值不同

---

3. Using Dummy Variables for Multiple Categories

3.1 Use Several Dummy Variables in the Same Equation

> Q: 如何将多分类变量纳入模型之中?

- 如果我们有n个分组，则我们建立$n-1$个虚拟变量 [防止完全共线问题].
- 被去掉的分类就是base group，其他所有组均和被省略的组比较.

```
reg y x1 i.dummy
```

> Q: 问题：如何研究多种虚拟变量？[以婚姻溢价为例]

- **方法1**：直接将不同虚拟变量纳入回归模型之中
  - 局限性：假定对于男性与女性而言，婚姻溢价是相同的→ 相当于未分组，结果可能不显著
- **方法2**：将两个虚拟变量相乘建立新虚拟变量，形成n个分组，设置$n-1$个新虚拟变量
  - 得到的系数代表着其他组别和base group[单身男]的差额
  - 回归得到的总截距代表base group的截距
  - 通过加减系数，可以对任意两个组别进行比较。但是这种方法无法得到显著性，为解决显著性问题必须更改base group重新回归。
  - STATA CODE：

```
generate marrmale=married* (1-female)
generate marrfem=married* female
generate singfem=(1-married)* female
```

```
reg income female#marriage
```

- **方法3**：设置n个新虚拟变量，去掉截距项β_0
  - 无法得到差值，系数难以解释和检验
  - R^2 的计算方式有所不同

3.2 Using Ordinal Information

> Q: 问题：如何将定序变量纳入模型之中?

- **方法1**：建立线性模型
  - 如果建立最简单的回归模型，则得到的系数是等级提高 1 个级别时因变量的变化。即，假定关系是线性的。
    - 假定学历“从小学上升到初中”以及“从研究生上升到博士”对收入的影响一样
  - 选择线性模型还是非线性模型取决于研究的目的，如果只是将多分类变量作为控制变量，则选择线性模型即可。
- **方法2**：建立非线性模型
  - 但在现实中，两个变量的变化可能是非线性的，即自变量从0到1对因变量的影响和自变量从3-4对因变量的影响是不同的。把自变量作为连续变量计算出的β无法捕捉这种非线性关系。
    - 这种非线性关系的一个案例是边际效益递减
    - 换言之，$\delta_4 \neq 4\delta_1$
  - 如果有n个分类，则建立$n-1$个虚拟变量
    - $MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 +$ other factors.

    - 如果我们关心的不是多上一年学工资如何变化，而是多拿到一个学历之后工资如何变化 则不能用线性模型 而应该使用虚拟变量。
- 案例：长相和工资
    - homely指代的是没法出门的程度的不好看（。
    - 因为在两个极端的人比较少，所以进行了重新分组→ 分组不要分的太细
    - 一般和处于中间的类别进行比较

---

4. Interactions Involving Dummy Variables ❖❖

4.1 Interactions between Dummy Variables

> Q: 如何处理虚拟变量之间的交互?

- 建立回归模型
    - $$\log(\text{ wage }) = \beta_0 + \beta_1 \text{ female } + \beta_2 \text{ married } + \beta_3 \text{ female } \cdot \text{ married } + \text{ other factors}$$
- 系数解读
    - 回归得到的β_0指代的是单身男的截距[female=married=0]
    - 单身女性的截距是$\beta_0 + \beta_1$ [female=1 married=0]
    - 已婚男性的截距是$\beta_0 + \beta_2$ [female=0 married=1]
    - 已婚女性的截距式$\beta_0 + \beta_1 + \beta_2 + \beta_3$ [female=1 married=1 femalemarried=1]
    - β_1是单身女性和单身男性的差别
    - β_2是已婚男性和单身男性的差别
    - $\beta_1 + \beta_2 + \beta_3$是已婚女性和单身男性的差别
    - 系数解读可能会设置陷阱!
- 建立新虚拟变量与建立交互项的比较
    - 建立交互项的优势:
        - 不需要手动生成新的虚拟变量，操作简单
        - 如果我们不关注系数，而仅仅将其当作控制变量，则直接使用交互项即可
    - 建立新虚拟变量的优势:
        - 可以直接得到已婚女性和单身男性的差别以及显著性，而不需要计算$\beta_1 + \beta_2 + \beta_3$
            - 两个方程中的β_1和β_2不变
        - 如果我们关注系数，特别是关注已婚女和单身男的差异时，则建立新虚拟变量会更加方便

4.2 Interacting Dummies with Continuous Variables

> Q: 如何处理虚拟变量和连续变量的交互?

- 建立回归模型
    - $$log(wage) = \beta_0 + \delta_0 \text{ female } + \beta_1 \text{ educ } + \delta_1 \text{ female } \cdot \text{ educ } + u$$
- 系数解读
    - 在虚拟变量和连续变量之间做交互，使得不同分组可以得到不同斜率的回归线→ 交互项相当于一个开关
    - 本质上等同于有n个分组，建立了n个斜率和截距均不同的总体回归线→ 建立交互项与建立不同回归方程的比较
        - 如果想要研究的是不同组别的差异，则建立交互项可以直接得到差值的大小并检验显著性。
    - 对于男性而言，截距是β_0，斜率是β_1
    - 对于女性而言，截距是$\beta_0 + \delta_0$，斜率是$\beta_1 + \delta_1$
    - 其中，δ_0衡量了两个性别在截距上[收入]的差异；δ_1衡量了两个性别在斜率上[教育回报]的差异。
- 交互项图像

- 

  - 因为女性的平均收入低于男性，因此两个图像中δ_0均小于0
  - 在左图中，女性的教育回报小于男性，则斜率更小，$\delta_1 < 0$
  - 在有图中，女性的教育回报大于男性，则斜率更大，$\delta_1 > 0$
  - $\delta_1 > 0$ 说明随着教育程度的增加，男女之间的工资差异缩小→ 如果增加女性受教育的机会，则男女之间的工资差异会缩小。实证研究表明，右图更符合实际情况。但是在最高的教育水平上，女性依然仅仅是catch up而不是overtake，即停留在了交点之前。

4.3 Hypothesis Test for Different Slopes and Interpretation

> Q: **如何对斜率和截距进行假设检验?**

- **原假设**：$H_0 : \delta_0 = 0, \quad \delta_1 = 0$
- **检验方法**：t检验或F检验

  - ```
    regress lwage female educ femaleeduc otherfactors
    ```

- **系数解读**
  - educ前面的系数仅仅代表男性的教育回报[8.2%]，女性的教育回报需要加上femaleeduc前的系数[7.64%]. femaleeduc的系数代表着教育回报差异是否经济显著，t值代表是否统计显著→ 无法拒绝原假设
  - 注意：female的t值显著下降→ 建立交互项后形成了**多重共线性**问题，标准误提高
  - 解决方法：demean→ 将交互项由$female \cdot educ$改为$female \cdot (edu - \overline{educ})$→ 仅改变female前的系数与标准误→ 一般虚拟变量和连续变量做交互、均需要对连续变量做demean处理

4.4 Testing for Differences in Regression Functions across Groups

> Q: **如何对不同群体的回归方程总体进行假设检验?** 比如想检验男性和女性的工资回归方程是否存在系统差异

- 建立**完全交互模型** [fully interacted model]，**每一项**都建立交互
  - 男性：$\log(\text{wage}) = \beta_0 + \beta_1 \text{ educ } + \beta_2 \text{ exper}$
  - 女性：$\log(\text{wage}) = \beta_0 + \beta_1 \text{ educ } + \beta_2 \text{ exper } + \delta_0 \text{ female} + \delta_1 \text{ educ} \cdot \text{female} + \delta_2 \text{ exper} \cdot \text{female}$
- **检验方法1：F检验**
  - 原假设：$H_0 : \delta_0 = \delta_1 = \ldots = \delta_n = 0, H_1 : \text{not } H_0$
  - stata代码

    - ```
      regress lwage educ exper female femaleeduc femaleexper
      test female=femaleeduc=femaleexper=0
      ```

  - 相当于检验联合显著性。所有的交互项衡量的都是男性和女性斜率的差异。

    - 注意：为什么不能用ttest而需要使用联合检验？存在多重共线的问题→ femaleedu和edu天然相关，因此交互项不显著是非常正常的。
- **检验方法2：Chow检验**
    - 如果回归方程包括几十个变量，则依次建立交互项进行F检验十分繁琐→ 使用Chow检验分析不同群体之间是否存在显著差异。
    - 基本原理：按照不同组别分别做回归，此后与合并的回归做比较
    - 非限制模型：对不同组别分别做回归 [比如按照性别分成两个组]
        - $y = \beta_{g,0} + \beta_{g,1}x_1 + \beta_{g,2}x_2 + \ldots + \beta_{g,k}x_k + u$
        - 两个方程的总样本数为n，每个方程有$k+1$个待估系数，两个回归方程加起来，自由度为$n-2(k+1)$
        - 非限制模型的总残差和为两个回归方程的残差和之和：$SSR_{ur} = SSR_1 + SSR_2$
    - 限制模型：假定不同组别拥有相同的系数，整合在一起做回归
        - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$
        - 自由度为$n-k-1$
        - 残差和$SSR_r = SSR_P$
        - 注：两个模型中，用于分组的变量[female]均不参与回归。
    - 构建F值：
        - $$F = \frac{[SSR_P - (SSR_1 + SSR_2)]}{SSR_1 + SSR_2} \cdot \frac{[n-2(k+1)]}{k+1}$$
        - 注意：这里只能使用SSR计算，不能用R^2
    - 缺陷：
        - Chow检验只能对模型整体[所有系数]进行检验，不能仅仅检验某一个变量的交互。因此，如果只希望检验某个变量的组间差异，则应该选择更加灵活的F-test。

---

5. A Dummy Dependent Variable: the Linear Probability Model

> Q: **如何处理定性型因变量**?

- **系数解读**：
    - $$\begin{aligned} E(y \mid x) &= P(y=1 \mid x)E(y=1) + P(y=0 \mid x)E(y=0) \\ &= P(y=1 \mid x) \cdot 1 + P(y=0 \mid x) \cdot 0 \\ &= P(y=1 \mid x) \\ &\Rightarrow P(y=1 \mid x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \end{aligned}$$
    - 其中，$P(y=1 \mid x)$为应答概率[*response probability*]，$P(y=1 \mid x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ 为线性可能性模型[*linear probability model* (LPM)]。
    - LPM模型估计的因变量是给定x，预测$y=1$的概率性
        - $$\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_k x_k$$
    - 估计的系数β代表x变化1单位，$y=1$的可能性改变了多少。
        - $$\Delta P(y=1 \mid x) = \beta_j \Delta x_j$$
- **案例**：已婚女性的劳动力参与
    - $\text{inlf} = \beta_0 + \beta_1 \text{ nwifeinc} + \beta_2 \text{ educ} + \beta_3 \text{ exper} + \beta_4 \text{ exper}^2 + \beta_5 \text{ age} + \beta_6 \text{ kidslt 6} + \beta_7 \text{ kidsge 6} + u$
    - 男性的工资增加1000美金，女性的劳动参与会下降0.34个**百分点**[*percentage point*]
    - 教育程度增加一年，女性劳动参与的可能性会增加3.799个**百分点**。
    - 多一个6岁以下的小孩，参与劳动的概率下降约四分之一
- **问题**：
    - 如果有四个孩子，参与劳动力市场的可能性就是负数→ 在一些极端值时会出现一些不正常的预测，可能性在0和1区间之外。LPM对于**在均值附近**的自变量拟合得更好。
    - LPM模型必然存在**异方差问题**：

- 当y是二元变量时，其条件方差为：
  - $$\mathrm{Var}(y \mid x) = P(y=1 \mid x)[1 - P(y=1 \mid x)]$$
- 因为方差是x的方程，取决于x的取值，因此必然存在异方差问题。
- 当MLR.5不成立时，t检验和F检验可能不够可靠，也会影响BLUE.
- 推导：
  - 假定n_1个y取1，n_2个y取0，则
  - $$\bar{y} = \frac{n_1}{n} = p(y=1 \mid x)$$
  - $$\begin{aligned}\mathrm{Var}(y) &= \frac{1}{n}\sum (y-\bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n_1}(1-P)^2 + \frac{1}{n}\sum_{i=1}^{n_2}(0-P)^2 \\ &= \frac{n_1}{n}(1-P)^2 + \frac{n_2}{n}P^2 \\ &= P(1-P)^2 + (1-P)P^2 \\ &= P(1-P)\end{aligned}$$

- **替代方案**：使用logit和Probit模型。由于替代方案得到的结果与LPM差异不大，因此尽管存在问题，但是LPM的应用依然十分广泛。

---

6. More on Policy Analysis and Program Evaluation

- **案例1**：项目评估破损率
  - 如果给谁不给谁grant是完全随机的，每个企业得到grant的概率是一样的，直接回归即可。如果不是随机的安排，而是先到先得，则会构成了**自选择问题** [self-selection]，进而导致内生性问题。
  - 积极参与报名的企业本身具有更高的破损率、规模也更小→ 需要控制住企业的规模和企业的业绩，消除实验组与对照组在干预前的差异。
- **案例2**：下岗女工再就业问题
  - 存在自选择问题，$E(u \mid \text{participate}=1) \neq E(u \mid \text{participate}=0)$，是否参加项目是内生变量。
- **案例3**：贷款审批中的种族歧视问题
  - 与前两个案例不同，第三个案例中的“种族”属性我们无法选择，因此不存在自选择问题
  - 但我们无法直接将系数上的差异完全归咎于歧视：不同种族的人之间可能存在系统性差异，比如黑人和白人的教育水平不同。
  - 因此，需要控制在不同种族之间存在系统性差异、同时会影响因变量的因素。

8. Heteroscedasticity

2022.05.07

---

1. Estimating Robust Standard Errors

1.1 Consequences of Heteroscedasticity for OLS

> Q: 放松同方差性假设后，**异方差性会对OLS造成什么影响**?

- 同方差性假设：假定误差项u的条件方差为常数，$\operatorname{Var}(u \mid x_1, \ldots, x_k) = \sigma^2$
- **异方差性**：当x变化时，误差项的方差也随之发生变化。
- **估计量的性质**：
  - **一致性**和**无偏性不变**→ 由MLR.1-4推导得到，与MLR.5无关
    - 异方差性不影响计量经济学的本质与最终目标
  - 有效性受影响：OLS估计量不再具有方差最小的性质，不再是BLUE
- **拟合优度**：R^2和$\overline{R}^2$不受影响
- **假设检验**：异方差性会影响标准误，进而影响置信区间以及t值→t检验与F检验不可靠。
  - $Var(\widehat{\beta_j})$有偏，影响标准误。

1.2 Robust Standard Errors

> Q: **如何在异方差性假定下使用OLS估计**?

- **定义**：
  - 稳健性估计不对模型形式做更改，只调整对于系数标准误的估计，得到**稳健标准误** [*Robust Standard Errors*]。即便MLR.5不成立，在**大样本**情况下，稳健标准误**依然不受**总体异方差的影响。
  - 在保证MLR.1-4依然成立的情况下，不同的自变量得到的σ不同→ $\operatorname{Var}(u_i \mid x_{i1}, \ldots, x_{ik}) = \sigma_i^2$
- **推导**：
  - 在满足MLR.1-4的情况下，利用“**partialling out**”法表示估计量
    - $$\begin{aligned}\widehat{\beta}_j &= \frac{\sum_{i=1}^n \hat{r}_{ij} y_i}{\sum_{i=1}^n \hat{r}_{ij}^2} \\ &= \frac{\sum_{i=1}^n \hat{r}_{ij}\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i\right)}{\sum_{i=1}^n \hat{r}_{ij}^2} \\ &= \beta_j + \frac{\sum_{i=1}^n \hat{r}_{ij} u_i}{\sum_{i=1}^n \hat{r}_{ij}^2}.\end{aligned}$$
  - 和的方差=方差的和，则
    - $$\begin{aligned}\Rightarrow \operatorname{Var}\left(\widehat{\beta}_j\right) &= \operatorname{Var}\left(\frac{\sum_{i=1}^n \hat{r}_{ij} u_i}{\sum_{i=1}^n \hat{r}_{ij}^2}\right) = \operatorname{Var}\left(\frac{\sum_{i=1}^n \hat{r}_{ij} u_i}{SSR_j}\right) \\ &= \frac{\sum_{i=1}^n \hat{r}_{ij}^2 \sigma_i^2}{SSR_j^2}\end{aligned}$$
  - 由于真实的u_i不可知，因此当样本量足够大时，White(1980)提出可以用$\hat{u}_i$进行近似替代，则：
    - $$\widehat{\operatorname{Var}\left(\widehat{\beta}_j\right)} = \frac{\sum_{i=1}^n \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}$$

- MacKinnon and White (1985)对自由度进行了调整，得到**适用于小样本**情况的**异方差稳健标准误***[heteroscedastic-robust standard errors*]的计算公式
  - $$\operatorname{se}\left(\widehat{\beta}_j\right)=\sqrt{\widehat{\operatorname{Var}}\left(\widehat{\beta}_j\right)}=\sqrt{\frac{n}{n-k-1}\frac{\sum_{i=1}^n \hat{r}_{ij}^2 \widehat{u}_i^2}{SSR_j^2}}$$

- **STATA CODE**
  - ```
    regress y x1 x2, robust
    ```

- 相比于普通标准误，Robust标准误一般会**变大**，**显著性会差**。
- 无论总体中是否存在异方差问题，稳健的标准误都是有效的。
- 既然稳健标准误如此有效，为何我们还使用OLS标准误呢？
  - 如果假设5成立，无论样本量多大，所有的检验都是正确的；
  - 但是稳健的标准误只适用于大样本环境，如果样本量不大，分布会有一定的偏差。
- 因此，我们通常在**大样本截面数据**的情况下使用稳健标准误。
  - 如果使用**截面数据**，一定要加robust：截面数据个体的差异性很大，异方差性不可避免。

---

2. Testing for Heteroscedasticity

- 假设5成立时，使用OLS依然是最好的方法→

> Q: **如何检验是否存在异方差问题**？

2.1 Breusch-Pagan test

- 在进行检验之前需要保证MLR.1-4是成立的。

- **原理**：

  - 如果同方差假设成立，则$E\left(u^2\right)=\sigma^2$，即$u^2$的变动不与$x$相关，则以$u^2$为因变量建立回归模型：
    - $$u^2=\delta_0+\delta_1 x_1+\delta_2 x_2+\ldots+\delta_k x_k+\nu$$
    - $$H_0: \delta_1=\delta_2=\ldots=\delta_k=0$$
  - 由于我们无法获知真实的残差，因此用残差的估计值进行估计：
    - $$\widehat{u}_i^2=\delta_0+\delta_1 x_1+\delta_2 x_2+\ldots+\delta_k x_k+\text{ error}$$
  - 换言之，将异方差的检验转变为新**多元回归模型的联合显著性检验**→ 如果无法拒绝原假设，说明同方差假设成立；如果能够拒绝原假设，说明u^2的变化与x有关，因此存在异方差问题。
  - 如果我们怀疑异方差性仅仅由于几个特定的自变量，我们可以只regress $\widehat{u}_i^2$ on 特定的x. 有过特定的x只有一个，则我们通过t值即可判断。
  - F统计量的计算：使用R^2
    - $$F=\frac{R_{\widehat{u}^2}^2/k}{\left(1-R_{\widehat{u}^2}^2\right)/(n-k-1)} \sim F_{k, n-k-1}$$

- **步骤**：

1. 使用OLS估计量，得到每个观测值残差的估计值及其平方。
    ```

2. 用残差估计值的平方为因变量做回归，$\widehat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k +$ error，得到$R^2_{\widehat{u}^2}$
3. 计算F值与对应的p值，如果无法拒绝原假设，则说明MLR.5成立，否则存在异方差问题

- **STATA CODE**：两种方法

> - ```
>   estat hettest,iid rhs
>   ```
> - ```
>   reg y x1 x2 x3
>   predict uhat,residual
>   generate uhatsq=uhat*uhat
>   reg uhatsq x1 x2 x3
>   ```

- **图形法**：scatter y x

> - 
>
> - 如果随着x的变动，y变动的范围扩大，说明存在异方差问题。

2.2 White test

- **原理**：
  - White Test在B-P test基础上修正，加入了自变量的平方项以及两两之间的交互项，要求更加严格，因此更容易检查出异方差的问题。
  > - $$\begin{aligned}\widehat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 \\ + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + \text{ error}\end{aligned}$$
  - 由于项数实在是太多了，因此可以选择用x的线性方程$\hat{y}$与非线性方程$\hat{y}^2$进行进一步简化：
  > - $$\widehat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + \text{ error}$$
- **步骤**：
  1. 通过OLS回归得到$\hat{u}$与$\hat{y}$，计算$\hat{u}^2$与$\hat{y}^2$
  2. 做回归$\widehat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 +$ error，得到$R^2_{\widehat{u}^2}$
  3. 计算F值与对应的p值，如果无法拒绝原假设，则说明MLR.5成立，否则存在异方差问题
- **STATA CODE**：两种方法

> - ```
>   estat imtest,white
>   ```
> - ```
>   reg y x1 x2 x3
>   predict uhat,residual
>   predict yhat,xb
>   generate uhatsq=uhat*uhat
>   generate yhatsq=yhat*yhat
>   ```

```
reg uhatsq yhat yhatsq
test yhat yhatsq
```

- 注意：应首先检查内生性问题，如是否存在方程形式误设，再检查异方差性。

---

3. Weighted Least Squares (WLS) Estimation

> Q: **诊断出异方差后，如何处理**?

- 如果我们已知方差的形式[一个由自变量构成的函数]，则处理异方差问题的一个方法是使用**加权最小二乘法** [*weighted least squares* (WLS)]重新设定其形式。
- **原理**:
  - 假设方差的形式为:
    - $$\operatorname{Var}\left(u_i \mid x_{i1}, \ldots, x_{ik}\right)=\sigma^2 h\left(x_{i1}, \ldots, x_{ik}\right)$$
  - 其中，$h_i=h\left(x_{i1}, \ldots, x_{ik}\right)$是$x$的函数
    - h_i必须为正数→ 因为方差为正数
    - h_i必须随着观测值的不同而有所变动→ 如果是常数，则符合同方差假设。
    - 假定h_i已知，而总体中的σ^2未知→ 用样本中的$\hat{\sigma}^2$进行替代。
  - 在原回归方程$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_k x_{ik}+u_i$的基础上，除以$\sqrt{h_i}$:
    - $$\begin{aligned} y_i / \sqrt{h_i}=\beta_0 / \sqrt{h_i}+\beta_1\left(x_{i1} / \sqrt{h_i}\right)+\beta_2\left(x_{i2} / \sqrt{h_i}\right)+\ldots \\ +\beta_k\left(x_{ik} / \sqrt{h_i}\right)+\left(u_i / \sqrt{h_i}\right) \end{aligned}$$
    - $$\Rightarrow y_i^*=\beta_0 x_{i0}^*+\beta_1 x_{i1}^*+\ldots+\beta_k x_{ik}^*+u_i^*$$
    - $$\Rightarrow E\left(u_i^*\right)=E\left(u_i / \sqrt{h_i}\right)=\frac{1}{\sqrt{h_i}} E\left(u_i\right)=0$$
    - $$\Rightarrow \operatorname{Var}\left(u_i^*\right)=E\left(\left(u_i^*\right)^2\right)=E\left(\left(u_i / \sqrt{h_i}\right)^2\right)=E\left(u_i^2\right) / h_i=\left(\sigma^2 h_i\right) / h_i=\sigma^2$$
  - 换言之，通过重新设定模型，新的回归方程满足同方差假设。残差的平方以$1/h_i$加权，则方差越大，权重越小，最终实现平衡。
- **广义最小二乘法** [*generalized least squares* (GLS)]
  - 将回归方程变形得到的OLS参数是广义最小二乘法估计量的一种。
  - 被用于解决异方差问题的GLS估计量又被称为WLS估计量，因为估计量β_j^* 能够将加权的残差平方和最小化。
  - GLS估计量满足MLR.1-5，因此WLS得到的结果比OLS估计量更加有效→ SE降低，显著性增加。
  - 注意：系数代回原方程解释。
- STATA 代码
  - 因为手动设置了常数项，因此需要加上nonconstant

```
gen ystar=y/sqrt(hi)
gen x1star=x1/sqrt(hi)
gen x2star=x2/sqrt(hi)
gen constant=1/sqrt(hi)
reg ystar x1star x2star constant, nonconstant
```

  - aw=average weight；注意，这里的aw不需要开根号

```
reg y x1 x2 [aw=1/hi]
```

4. Feasible Generalized Least Squares

> 不知道具体的hi时，如何对异方差问题进行修正?

- 通常我们无法得知h_i的具体形式，因此需要估计$\hat{h_i}$，进而得到**可行的GLS估计量**[*feasible GLS* (FGLS) *estimator*]。
- **如何估计**？设定新方程：
  - $$\operatorname{Var}\left(u_i \mid x_{i1}, \ldots, x_{ik}\right)=\sigma^2 h_i=\sigma^2 \exp\left(\delta_0+\delta_1 x_{i1}+\delta_2 x_{i2}+\ldots+\delta_k x_{ik}\right)$$
  - 原因：保证是正数，同时是x的函数
  - 注：$\exp(x)=e^x$
- **如何估计参数**?
  - 残差的平方的真值为：
    - $$\Rightarrow u^2=\sigma^2 \exp\left(\delta_0+\delta_1 x_1+\delta_2 x_2+\ldots+\delta_k x_k\right) \nu$$
  - 取对数，得到
    - $$\log\left(u^2\right)=\alpha_0+\delta_1 x_1+\delta_2 x_2+\ldots+\delta_k x_k+e$$
  - 为了得到参数，$\operatorname{reg} \log\left(\widehat{u}^2\right)$ on $x_1, \ldots, x_k$，得到
    - $$\hat{g}_i \equiv \widehat{\log\left(\widehat{u}_i^2\right)}$$
    - $$\Rightarrow \widehat{h}_i=\exp\left(\hat{g}_i\right)$$
  - 如果变量太多，也可以$\log\left(\widehat{u}^2\right)$ on $\hat{y}, \hat{y}^2$，得到$\hat{g}_i$，进而得到$\widehat{h}_i=\exp\left(\hat{g}_i\right)$
- **STATA 代码**：
  - ```
    reg y x1 x2
    predict uhat,r
    gen logu=log(uhat^2)
    reg logu x1 x2
    predict ghat,xb
    gen hhat=exp(ghat)
    gen ystar=y/sqrt(hhat)
    gen x1star=x1/sqrt(hhat)
    gen x2star=x2/sqrt(hhat)
    gen constant=1/sqrt(hhat)
    reg ystar x1star x2star constant, nonconstant
    reg y x1 x2 [aweight = 1/hhat] //简约版
    ```
- **注意**：如果OLS和WLS估计量相差甚远，说明MLR.4不成立，模型存在内生性问题。
- 如果估计到的$\hat{h_i}$有问题，依然无法保证同方差怎么办？→ **双保险：计算稳健标准误**

5. Heteroscedasticity in the Linear Probability Model

- LPM模型一定存在异方差问题，同时我们知道它异方差的形式：$\operatorname{Var}(y \mid x)=p(x)[1-p(x)]$，因此
  - $$\widehat{h}_i=\hat{y}_i\left(1-\hat{y}_i\right)$$
- **步骤**：
  1. 跑OLS回归模型，得到$\hat{y}$
    ```

2. 确认所有拟合值$\hat{y}_i$都乖乖待在[0,1]区间内。如果有**小部分**数字不在区间内，则将负值改为0.001；将大于1的值改为0.999.
3. 建立$\widehat{h}_i = \hat{y}_i \left(1 - \hat{y}_i\right)$
4. 按照$1/\widehat{h}$的权重重新跑回归

- **STATA 代码**:

> - ```
>   reg y x1 x2
>   predict yhat,xb
>   tab yhat //检查取值是否在[0,1]区间内
>   gen newyhat=yhat if yhat>=0
>   replace newyhat=0.001 if yhat<0
>   replace newyhat=0.999 if yhat>=1
>   gen hhat=newyhat*(1-newyhat)
>   reg y x1 x2 [aweight = 1/hhat]
>   ```

9. More on Specification and Data Issues

2022.05.14

---

1. Functional Form Misspecification

> Q：**如何解决内生性中方程形式误设的问题**？→ 违背MLR.4

1.1 What is Functional Form Misspecification?

- x和y并非线性关系，却按照线性模型估计参数，使得残差项中包含了x的非线性部分，u和x相关，违背MLR.4，导致内生性问题
- 方程形式误设是遗失变量偏误的特殊情况，即x的非线性部分被遗漏了[*omitted*]。
- **三种形式**
  - 省略了**平方项**x^2
  - 省略了**交互项**x_1x_2
  - 未采用**对数**形式log

1.2 How to detect Misspecified Functional Form?

- **F检验**
  - 首先进行简单线性回归，其次将每一个显著变量的平方项都加入回归模型中，进行**F检验**。如果平方项也联合显著，则应该被放入模型中。
  - 按照F检验的思路加入更多平方项的弊端
    - 损失自由度，同时容易引发多重共线性问题，影响显著性
    - 系数的解释变得更加困难
    - 加入平方项无法解决所有非线性问题
  - → 在MLR.4已经满足的条件下，我们不应该加入显著的平方项→ 引入RESET[*regression specification error test*]法
- **RESET**：**General Test**
  - **步骤**
    1. 做初始回归$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$，得到拟合值 $\hat{y}$.
    2. 对初始回归方程进行拓展，加入 $\hat{y}^2$与 $\hat{y}^3$，得到新回归方程
       - $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + \text{ error}$$
       - 与White Test的区分：一个因变量是y，一个因变量是$\hat{u}^2$
    3. 对原始方程与拓展方程进行**F检验**
       - $H_0 : \delta_1 = \delta_2 = 0$
       - 自由度分别为2和$(n-k-1)-2 = n-k-3$
       - 如果联合不显著，无法拒绝原假设，则说明原始模型设定正确。否则存在方程模型误设问题。
  - **STATA 代码**
    - ```
      regress y x1 x2 x3
      predict yhat, xb
      generate yhatsq=yhat*yhat
      generate yhatcub=yhat^3
      regress y x1 x2 x3 yhatsq yhatcub
      test yhatsq yhatcub
      ```
      ```

- **局限性**
  - RESET法只能检验**是否**存在方程形式误设，无法确认哪一部分出了问题&问题的方向。
  - 无法检测更加一般性的遗漏变量偏误。
  - 如果方程设定正确，则无法检测**异方差**问题。
  - → RESET法只能检验方程形式误设

- **Davidson-MacKinnon test**
  - D-M test适用于确认是否应该使用对数形式
  - 对原始方程$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$取对数，得到新方程
    - $$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + u$$
  - 由于两个模型为非嵌套模型，因此我们不能使用F检验
  - **步骤**
    - 运行第一个模型，得到拟合值$\hat{y}$
    - 运行第二个模型，得到拟合值$\hat{\hat{y}}$
    - 估计下述两个模型
      - $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta_1 \hat{\hat{y}}$$
      - $$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + \theta_2 \hat{y} + error$$
      - 注意：两个y必须**统一形式**，不能一个是level一个是log
    - 如果只有θ_1显著，要拒绝水平模型
      - θ_1显著说明对数部分是显著的，模型2胜出。
    - 如果只有θ_2显著，要拒绝对数模型。
      - θ_2显著说明线性部分是显著的，对数部分不显著，模型1胜出。
    - 如果θ_1和θ_2都显著，说明两个模型都不好
    - 如果θ_1和θ_2都不显著，两个模型均可，选择R^2更大的方程
  - **局限性**
    - 如果x不同，则不能使用
    - 如果θ_1和θ_2都显著，这个检验没有帮助。
    - 即便拒绝了水平模型，也不能证明对数模型是正确的。其他形式的模型误设也会导致水平模型被拒绝。
    - 无法适用于 y 和 $log(y)$ 的比较。

---

2. Proxy Variables

> Q：**如何解决内生性中重要变量遗失的问题**？→ 违背MLR.4

2.1 Definition

- **代理变量** [*proxy variable*]是**与无法观测到的变量相关**的可测量的变量，可以在一定程度上缓解遗漏变量偏误。
  - 如IQ是能力的代理变量
- $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$
  - 其中x_3^* 无法被观测，x_3 是其代理变量。
- x_3^*与x_3的关系
  - $$x_3^* = \delta_0 + \delta_3 x_3 + \nu_3$$
  - δ_0 的存在使得x_3^*与x_3可以用不同的单位衡量。
  - δ_3 衡量了x_3^*与x_3之间的关系
  - ν_3 代表着x_3^*与x_3之间的随机扰动→ 二者并非完全一致
  - 换言之，相当于用一个方程替代了无法被观测到的变量。

- 代入原方程，得到$y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + u + \beta_3\nu_3$，即
  - $$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e$$
  - 其中$\alpha_0 = \beta_0 + \beta_3\delta_0$, $\alpha_3 = \beta_3\delta_3$, $e = u + \beta_3\nu_3$,。
  - 其中，e又被称为*composite error term*，一部分由代表原方程的误差项，另一部分则是用代理变量衡量真实变量时带来的误差。

2.2 Unbiasedness Requirements

- 为保证β_1和β_2依然无偏，至少应该满足：
- **假设1**：**多余性假设**
  - u 应该与 x_1, x_2, x_3^* 以及 x_3 无关，即
    - $$E(u \mid x_1, x_2, x_3^*) = E(u \mid x_1, x_2, x_3^*, x_3) = 0$$
  - 如果x_3^*真实存在且能被测量，那么x_3就不应该出现在总体模型中，即x_3是多余的。换言之，x_3 对 y 的影响必须通过x_3^*，否则它就不是一个好的代理变量。
  - x_3不能包含超过x_3^*的信息，信息不能太多。
- **假设2**：**好的代理变量假设**
  - ν_3需要与 $x_1, x_2,$ 以及 x_3 不相关，即
    - $$E\left(x_3^* \mid x_1, x_2, x_3\right) = E\left(x_3^* \mid x_3\right) = \delta_0 + \delta_3 x_3$$
    - $$E(\nu_3 \mid x_3) = E(\nu_3 \mid x_1, x_2, x_3) = 0$$
  - 换言之，给定x_3后，x_3^*的期望不取决于x_1和x_2的值→ 用x_3解释x_3^*后，剩下的部分无法用$x_1 x_2$解释。
  - x_3又必须包括充分的信息，使其能够解释x_3^*，信息不能太少→ 代理变量可遇不可求。
- 综合假设1和假设2，综合误差项e与x_1, x_2, x_3无关，即$E(e \mid x_1, x_2, x_3) = 0$，满足MLR.1-4，参数无偏。
- 注意，无偏的参数是β_1, β_2, α_0和α_3，而非β_0和β_3。
- **注意事项**：
  - 由于这两个假设是否成立无法被检验，因此很难说服别人这是一个好的代理变量。
  - 如果假设1和假设2无法满足，则我们无法得到无偏（或一致）的估计量:
    - 如果$x_3^* = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \nu_3$，即与所有的自变量相关，则我们得到的总体模型为
      - $$y = (\beta_0 + \beta_3\delta_0) + (\beta_1 + \beta_3\delta_1)x_1 + (\beta_2 + \beta_3\delta_2)x_2 + \beta_3\delta_3 x_3 + u + \beta_3\nu_3$$
    - 显然，参数有偏。

2. 3 Using Lagged Dependent Variables as Proxy Variables

- 使用滞后（过去）的因变量，解决**所有无法观测且不随时间变化**的遗漏变量。
- 案例：失业率与执法投入对犯罪率的影响
  - $$\text{crime} = \beta_0 + \beta_1 \text{ unem} + \beta_2 \text{ expend} + \beta_3 \text{ crime}_{-1} + u.$$
  - 本质上类似于做了一个自然实验，相当于找到了两个在上一期犯罪率完全相同的城市进行实验。则β_2衡量的是在两个犯罪率和当前失业率相同的城市中，扩大执法投入对犯罪率的影响。
  - 历史上犯罪率高的城市更有可能扩大执法投入，反向因果，直接做回归的结果就是执法力度越高的城市犯罪率反而越高→ 在加入历史犯罪率后，执法力度与犯罪率呈现反向变动的趋势，更符合直觉。加入历史数据是有道理的。

---

3. Measurement Error

> Q：**如何解决内生性中测量误差的问题？** → MLR.4

- 尽管有的变量可以被观测到，但在测量过程上存在误差[*measured with error*]，进而导致测量误差[*measurement error*]。

- 相比于因为无知而产生的误差 [比如不了解自己家的收入]，刻意撒谎而产生的误差问题更加严重。[如企业家们上报的数据和实际决策时使用的数据不相符]

3.1 Measurement Error in the Dependent Variable

- 基于真实信息的模型为 $y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$，满足G-M假定
- 而观察值y与真值y^*之间的测量误差定义为
  - $$e_0 = y - y^*$$
- 实际估计的模型为 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u + e_0$
  - 误差项发生变动
- **无偏性变动**
  - 如果$E(e_0) \neq 0$，那么 β_0 有偏。但通常而言， β_0有偏的影响不大。
  - 如果e_0在统计学意义上**与自变量无关**，即 $E(e_0 \mid x_1, \ldots, x_k) = 0$，说明这种测量上的误差具有随机性，而非系统性差异（如从事不法活动的企业家更容易瞒报收入）因此不影响OLS参数的无偏性与一致性→ **随机报告的误差**
  - 如果e_0和u无关，$\text{Var}(u + e_0) = \sigma_u^2 + \sigma_e^2 > \sigma_u^2$，估计的有效性变差，因此t值下降，显著性下降。
  - 如果e_0在统计学意义上与自变量有关，则估计出的参数有偏。
    - 比如：低教育程度的人倾向于高报工资，高教育程度的人倾向于低报工资，进而会低估教育回报。

3.2 Measurement Error in an Explanatory Variable

- 基于真实信息的模型为$y = \beta_0 + \beta_1 x_1^* + u$，满足MLR.1-4.
- 观察值x_1和真值x_1^*之间的测量误差定义为
  - $$e_1 = x_1 - x_1^*$$
- **假设**
  - 为方便进一步推导，做出两个假设
  - 假设1：在不损失一般性的情况下，我们始终可以假设测量误差的均值为0，即$E(e_1) = 0$
  - 假设2：类似于代理变量的**多余性假设**，我们可以假设：
    - $$E(u \mid x_1) = E(u \mid x_1^*) = E(u \mid x_1^*, x_1) = 0 \Rightarrow E(y \mid x_1^*, x_1) = E(y \mid x_1^*)$$
    - 即观测值只通过真值起作用，如果能观测到真值，则无需观测值。
- 则我们得到的估计模型为
  - $$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$$
- **无偏性变动**
  - 由假设2可知，联合误差项中的u与x_1无关，则无偏性取决于e_1和x_1是否相关。
  - $\text{Cov}(x_1, e_1) = 0$
    - 相当于引入了一个随机报告的误差，$\text{Cov}(x_1, u - \beta_1 e_1) = 0$。
    - 不影响待估参数的无偏性以及OLS参数的特性，但影响其有效性。
      - $$\text{Var}(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2 \sigma_{e_1}^2$$
  - $\text{Cov}(x_1, e_1) \neq 0$
    - 首先引入**经典的变量存在误差假设**[*Classical errors-in-variable* (CEV)]，即假设真实值和测量误差之间无关，$\text{Cov}(x_1^*, e_1) = 0$
    - 因为x_1和e_1的均值为0，则 $\text{Cov}(x_1, e_1) = E(x_1 e_1)$
    - 由于$x_1 = x_1^* + e_1$，则 $E(x_1 e_1) = E(x_1^* e_1 + e_1^2)$
    - 和的期望=期望的和，$D(X) = E\left(X^2\right) - E^2(X)$，则得到
      - $$\text{Cov}(x_1, e_1) = E(x_1 e_1) = E(x_1^* e_1) + E\left(e_1^2\right) = 0 + \sigma_{e_1}^2 = \sigma_{e_1}^2$$
    - 即：当CEV假设成立时，自变量与误差的协方差为$\sigma_{e_1}^2$

        - 换言之，只有满足$\text{Cov}(x_1^*, e_1) = -\sigma_{e_1}^2$这个严苛的条件时，$\text{Cov}(x_1, e_1) = 0$才能够成立。因此第一种情况$\text{Cov}(x_1, e_1) = 0$十分少见。→ 自变量如果存在测量误差，估计出的参数**极大概率有偏且不一致**。
- **偏差方向判断**
    - 基于上述讨论，在CEV假设成立的情况下，显然
        - $$\text{Cov}(x_1, u - \beta_1 e_1) = -\beta_1 \text{Cov}(x_1, e_1) = -\beta_1 \sigma_{e_1}^2$$
    - 推导：
        - plim=probability limitation，即$n \to \infty$时参数的取值。
        - 将$\hat{\beta}_1$分解成真值与随机扰动部分
            - $$\hat{\beta}_1 = \beta_1 + \frac{\sum (x_1 - \bar{x}_1) \cdot error}{\sum (x_1 - \bar{x}_1)^2}$$
        - 则两边取plim，得到
            - $$\text{plim}\, \widehat{\beta}_1 = \beta_1 + \frac{\text{Cov}(x_1,\ \text{error})}{\text{Var}(x_1)}$$
            - 注意，这里不再是样本的方差，而是总体的方差。
        - 带入$errorterm = u - \beta_1 e_1$，同时和的方差=方差的和，得到
            - $$\begin{aligned} \text{plim}\, \widehat{\beta}_1 &= \beta_1 + \frac{\text{Cov}(x_1, u - \beta_1 e_1)}{\text{Var}(x_1)} \\ &= \beta_1 - \frac{\beta_1 \sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \\ &= \beta_1 \left(1 - \frac{\sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2}\right) \\ &= \beta_1 \left(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2}\right). \end{aligned}$$
            - $$\text{plim}\, \widehat{\beta}_1 = \beta_1 \left(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2}\right) = \beta_1 \frac{\text{Var}(x_1^*)}{\text{Var}(x_1)}$$
            - $$\Rightarrow \text{Var}(x_1) > \text{Var}(x_1^*) \Rightarrow \frac{\text{Var}(x_1^*)}{\text{Var}(x_1)} < 1$$
        - 显然，**OLS估计量小于真值**，β_1向0偏：**衰减性偏差** [*attenuation bias*]
            - 偏差越严重，估计出来的参数越靠近0
    - 该结论可以推广到多元回归模型中。更多的自变量会让事情变得更加复杂。偏差的方向和程度不易推导得到。
- 由于自变量之间的相关性，测量误差会导致所有的估计量都不具备一致性。
- 解决方法：工具变量→ 工具变量可以解决除方程形式误设以外的所有内生性问题。

---

4. Missing Data, Nonrandom Samples, and Outliers

> Q：**如何解决MLR.2不满足时存在的问题？**

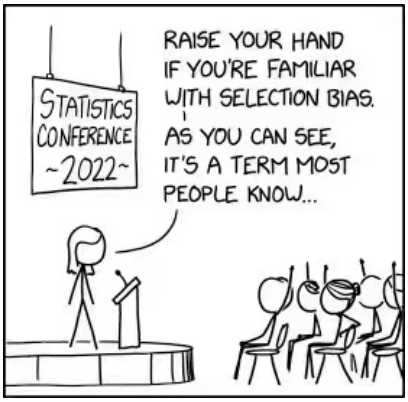
- 违背MLR.2:随机抽样假定、导致样本无法代表总体的三种情境：**Missing data, Nonrandom, samples Outliers**

4.1 Missing Data

- **数据缺失的统计影响**

- 如果数据随机缺失 [*missing at random*]，相当于n变小，仅仅是样本规模变小了。这会影响估计的精确性与有效性，但不会导致有偏。
- 如果数据系统性缺失，如高收入的人很难调查到、同时更有可能不报告收入，则会导致**非随机样本**。

4.2 Nonrandom Samples

- **自变量非随机**
  - 自变量出现非随机样本问题被定义为**外生样本选择** [*exogenous sample selection*]，估计量无偏，问题不大。
  - **CASE 1**：假设我们只有35岁以上的人的样本。我们估计的总体回归线是以age为自变量，saving为因变量的直线，则斜率在35岁以下的部分和在35岁以上的部分是相同的。自变量的非随机样本可被近似地视为切了一刀、做了一个垂线，但垂线本身并不影响斜率。
    - 对于以自变量划界的任何一个子集而言，$E($ saving $|$ income, age, size $)$ 都相同，因此不影响总体回归线的估计。
  - **CASE 2**：IQ高的人更愿意报告自己的IQ，即随着IQ的增加，数据缺失的概率越小。但影响样本选择的因素[**自变量本身**, 即IQ]与残差项相互独立 [$E(u \mid x)=0$]，样本选择问题是外生的，OLS参数无偏。
  - **CASE 3**：假设IQ信息的缺失与教育程度 [**模型的另一个自变量**] 有关，即教育程度越高越倾向报告IQ，与情况2同理，$E(u \mid x)=0$，则OLS参数无偏。
  - **CASE 4**：假设影响IQ信息缺失的因素是考试中心与家的距离，而距离显然与因变量 [wage] 无关，因此OLS参数无偏，仅仅会使样本规模缩小。[**随机缺失**]
  - 换言之：影响信息缺失的因素如果与误差项独立，则样本选择是外生样本选择。
- **因变量非随机**
  - 因变量出现非随机样本问题被定义为**内生样本选择** [*endogenous sample selection*]→ 导致选择性偏差 [*selection bias*]
  - **Case 1**：以income为因变量，同时income高于250万的人被排除在样本之外。由于 $E(wealth \mid educ, exper, age) < E(wealth < 250000 \mid educ, exper, age)$，则OLS参数有偏且不连续。
    - 相当于横着切了一刀，会导致总体回归线的斜率变化。
  - **Case 2**：假设影响IQ信息缺失的因素是情商EQ，EQ**不在自变量中，但与y有关**，则样本选择就变成了y 的样本选择 [*sample selection*]
  - **解决方法**：Tobit Model
- **抽样策略**引发的非随机样本 [*Sampling Schemes*]
  - **分层抽样**[*Stratified Sampling*]是一种常见的数据收集方法，但如果分层方案基于因变量，则同样会导致内生样本选择偏误，需要进一步处理。
    - 比如：如果研究罪犯群体，选择120个罪犯、300个普通人，这一罪犯比例显然与总体不相符。
  - 注：如果基于自变量，导致的外生样本选择偏误不需要处理。如果基于因变量，则需要通过进阶计量方法处理（参考文献）。
- **样本选择性偏误** [*Sample selection bias*]
  - 工资回归线所基于的样本是有工作的人，而是否进入劳动力市场是有系统性选择的，换言之，只有决定去工作的人才能被选择到。
  - 然而，是否去工作的影响因素可能同样会影响潜在的工资，进而导致内生性样本选择问题。
  - **解决方法**：Heckman Model
  - 选择性偏误与自选择问题的**区别**：
    - 选择性偏误指的是该样本是否进入模型并非随机。
      - 

      - 研究子女数量与母亲工资的关系→ 样本为生育后依然工作的母亲，不包括全职妇女。

        - 自选择是指样本属于干预组还是控制组并非随机，是个人选择的结果。
            - 研究老兵的工资回报→ 决定参军和不参军的人本身就存在区别→ 通过控制变量解决。

4.3 Outliers

- 离群点/奇异值指的是如果drop掉它，会对回归结果产生巨大影响的样本[*influential observations*]。
- **产生原因**:
    - 录入问题→ 数据清理
    - 总体中的部分成员异常奇葩，而奇葩确实到处都有。
- **处理方法**:
    - Method 1: Drop掉，只要在5%之内都可以去掉。
    - Method 2: 转换函数形式→ 取对数
    - Method 3: 不使用OLS，使用LAD法 [*Least Absolute Deviations*]
        - 取绝对值而非平方和
        - LAD法对离群值更不敏感→ 不常用

10. Instrument Variables

2022.05.28

---

1. Omitted Variables in a Simple Regression Model

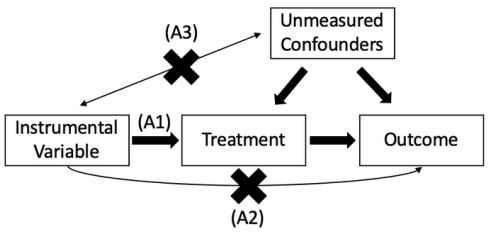
> Q：**如何解决遗漏变量问题?**

1. 佛系地**什么也不做**，仅仅指出存在偏误：只适用于向0偏，且结果依然显著的情况→ 不会推翻结论
2. 代理变量：可遇不可求且难以证明
3. 面板数据 [*Panel data*]：通过不同期数据处理掉不随时间变化的变量→ 无法处理地很干净，同时面板数据可及性较差。
4. 工具变量 [*Instrumental variables approach* (IV)]

- 注：代理变量是寻找遗漏变量的代理[IQ与能力]；而工具变量则寻找与遗漏变量完全不相关的因素 [出生季度与能力]，从而将自变量提纯。

1.1 Two Assumptions for IV

> Q：**使用工具变量法需要满足哪些假定**?

- $\text{cov}(z,u)=0$ **外生性假设** [*Instrument Exogeneity*]
  - 在控制x之后，z无法直接影响y
  - z与遗漏变量无关
  - 难以检验，依赖经济学直觉：如果工具变量能够**直接影响**y，或工具变量能够**影响遗漏变量进而影响**y，则外生性假设无法成立。[两条不成立路径]
- $\text{cov}(z,x)\neq 0$ **相关性假设** [*Instrument relevance*]
  - z 必须与x有关，即z在解释x的波动时有重要作用。
  - 可以检验：以x为因变量，z为自变量跑一个简单回归，检验斜率系数是否显著不为0。
    - $$x=\pi_0+\pi_1 z+\nu$$
    - $\pi_1=\frac{\text{cov}(z,x)}{\text{var}(z)}$，只要$\pi_1\neq 0$，则 $\text{cov}(z,x)\neq 0$
- 满足上述两个条件的变量即为合格的x的工具变量 [*valid*]
  - 

1.2 What Could Be Valid Instruments?

- 判断教育对工资的影响，一个重要的遗漏变量是能力→
- **父母的教育水平**：
  - 父母教育水平与子女的教育水平有关，但父母的能力一方面会通过遗传影响子女的能力[u]
  - 另一方面也可能通过社会经济地位直接影响子女的工资回报[y]
  - → 只满足相关性假设、但不满足外生性假设的工具变量
- **兄弟姐妹的数量**：

    - 孩子的数量会影响子女的教育水平 [资源分配理论]
    - 越穷的家庭孩子越多，因此家里有很多小孩可能意味着父母的能力较低，进而影响孩子的能力。
    - 兄弟姐妹数量可能也会影响工资水平→ 只满足相关性假设、但不满足外生性假设的工具变量
- 孩子的出生季度：
    - 出生季度和能力几乎没有关系，但是出生在下半年的同级学生因为年龄不足以进入劳动力市场，因此更有可能取上高中。换言之，第四季度出生的学生的平均教育水平高于第一季度，教育和出生季度有关。(Angrist and Krueger, 1991)
    - 因为季度和教育水平的相关性并不强，因此必须使用非常大的样本（25万）从而得到显著的结论。
    - 尽管显著性存在差异，但是IV和OLS法得到的系数并没有太大的差异→ 教育回报的内生性问题并不严重。
- 是否住在大学附近：
    - 在四年制大学附近居住，耳濡目染，因此教育水平更高。
    - 但是小时候的居住选择和个人能力以及成人之后的工资水平不相关。
    - 自选择问题：学区问题[附中附小]，且住在大学附近的家长很有可能就是大学教授

1.3 Identification with Instrumental Variables

- 距估计法
    - 对简单回归方程$y=\beta_0+\beta_1 x+u$进行分解，不难得到
        - $$\operatorname{cov}(z, y)=\beta_1 \operatorname{cov}(z, x)+\operatorname{cov}(z, u)$$
    - 在外生性和相关性成立的情况下，
        - $$\beta_1=\frac{\operatorname{cov}(z, y)}{\operatorname{cov}(z, x)}$$
        - $$\begin{gathered}\widehat{\beta}_1=\frac{\sum_{i=1}^n\left(z_i-\bar{z}\right)\left(y_i-\bar{y}\right)}{\sum_{i=1}^n\left(z_i-\bar{z}\right)\left(x_i-\bar{x}\right)} \\ \widehat{\beta}_0=\bar{y}-\widehat{\beta}_1 \bar{x}\end{gathered}$$
    - OLS是IV估计量的一个特例，如果x是外生的，则可被视为自己的工具变量，即$z=x$，参数估计等同于OLS法。
- 两阶段OLS法 [2SLS]
    - 第一阶段: reg x on z
        - $$x_i=\pi_0+\pi_1 z_i+\nu_i$$
        - $$\Rightarrow \hat{x}_i=\hat{\pi}_0+\hat{\pi}_1 z_i$$
        - 这一阶段的回归结果显著才能进行下一步。相当于将x外生化。
    - 第二阶段：reg y_i on $\hat{x}_i$
        - $$y_i=\beta_0+\beta_1 \hat{x}_i+u_i$$
        - $$\hat{\beta}_1=\frac{\sum\left(\hat{x}_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sum\left(\hat{x}_i-\bar{x}\right)^2}$$
        - $\because \hat{x}_i=\hat{\pi}_0+\hat{\pi}_1 z_i \quad \therefore \hat{x}_i-\bar{x}=\hat{\pi}_1\left(z_i-\bar{z}\right)$，即
            - $$\hat{\beta}_1=\frac{\sum \hat{\pi}_1\left(z_i-\bar{z}\right)\left(y_i-\bar{y}\right)}{\sum \hat{\pi}_1\left(z_i-\bar{z}\right)\left(\hat{x}_i-\bar{x}\right)}$$
        - 又 $\hat{x}_i-\bar{x}=x_i-\hat{\nu}_i-\bar{x}$, 且$\nu$与$z$不相关，则
            - $$\hat{\beta}_1=\frac{\sum\left(z_i-\bar{z}\right)\left(y_i-\bar{y}\right)}{\sum\left(z_i-\bar{z}\right)\left(x_i-\bar{x}\right)}$$
- IV估计量的性质
    - 在两个假设都满足的情况下，IV估计量具有一致性，$\Rightarrow \operatorname{plim}\left(\widehat{\beta}_1\right)=\beta$
        -

$$p\lim \hat{\beta}_1^{IV} = \beta_1 + \frac{\operatorname{cov}(z,u)}{\operatorname{cov}(z,x)}$$

- 外生性假设→ 分子为0；相关性假设→ 分母不为0，满足一致性。

- IV估计量**一定有偏**，因此只能在**大样本情况**下使用工具变量法
  - $$\hat{\beta}_1 = \beta_1 + \frac{\sum (z_i - \bar{z})(u_i - \bar{u})}{\sum (z_i - \hat{z})(x_1 - \bar{x})}$$
  - $$E\left(\hat{\beta}_1^{IV}\right) = \beta_1 + E_x\left[E\left(\frac{\sum (z_i - \bar{z})(u_i - \bar{u})}{\sum (z_i - \bar{z})(x_i - \bar{x})}\middle| \begin{matrix} z_i \\ x_i \end{matrix}\right)\right]$$
  - 由于MLR.4不成立，x_i与u相关，则参数有偏。

1.4 Statistical Inference with IV

- 为得到IV估计量的标准误，同样需要设置同方差假设。但与OLS估计量不同，该假设**设定于工具变量**z而非内生解释变量x上。
  - $$E\left(u^2 \mid z\right) = \sigma^2 = \operatorname{Var}(u)$$
- **IV估计量的渐进方差**[*asymptotic variance*]为
  - 未知的总体参数：
    - $$\frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2}$$
    - σ^2为残差项u的总体方差，σ_x^2为x的总体方差，$\rho_{x,z}^2$ 是x和z总体相关系数的平方。
  - 估计量：
    - $$\frac{\widehat{\sigma}^2}{SST_x \cdot R_{x,z}^2}$$
  - 其中，总体方差的估计量用残差拟合值$\widehat{u}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$衡量，
    - $$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{u}_i^2$$
  - 自变量的总体方差用样本方差衡量，$\frac{SST_x}{n}$
  - 为估算 $\rho_{x,z}^2$，可以通过第一步回归，即reg x_i on z_i得到的 $R_{x,z}^2$ 衡量
- OLS估计量的标准误与IV估计量渐进标准误 [*asymptotic standard error*]的比较
  - $$\operatorname{SE}\left(\hat{\beta}_1^{\text{OLS}}\right) = \frac{\hat{\sigma}}{\sqrt{SST_x}}$$
  - $$\operatorname{SE}\left(\hat{\beta}_1^{\text{IV}}\right) = \frac{\hat{\sigma}}{\sqrt{SST_{\hat{x}}}} = \frac{\hat{\sigma}}{\sqrt{SST_x \cdot R_{x,z}^2}}$$
  - 由于R^2恒小于1，则**IV的方差始终大于OLS的方差**。同时，如果x和z的相关性很弱，R^2太小，则IV估计量的抽样方差将很大→ 必须选择与x高度相关的工具变量。在从相关到因果的过程中，不得不放弃一些显著性。
- R^2 of IV Estimation
  - IV估计得到的R^2**可能为负数**，缺乏实际含义。因此使用工具变量时一般**不报告**R^2。
  - 同理，在进行F检验时，也**不能**直接使用工具变量估计法的R^2

---

2. IV Estimation of the Multiple Regression Model

> Q：**如何将该结论推广到多元回归模型**?

- 如果一个回归方程中**既有内生的解释变量又有外生的解释变量**，则将这一方程定义为**结构方程**[*structural equation*]

> - $$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$
> - y_1是被解释变量，y_2是内生的解释变量，z_1是外生的解释变量。
> - 如：$\log(\text{ wage }) = \beta_0 + \beta_1 \text{ educ } + \beta_2 \text{ exper } + u_1$，其中遗漏变量是能力。

- 为了得到一致的估计量，我们需要在结构方程中引入工具变量 z_2，提供第三个矩条件→

> - $$E(u_1) = 0, \operatorname{Cov}(z_1, u_1) = 0, \text{ and } \operatorname{Cov}(z_2, u_1) = 0$$

- 根据这三个矩条件，我们可以得到

> - $$\begin{aligned} \sum_{i=1}^{n}\left(y_{i1} - \widehat{\beta}_0 - \widehat{\beta}_1 y_{i2} - \widehat{\beta}_2 z_{i1}\right) &= 0 \\ \sum_{i=1}^{n} z_{i1}\left(y_{i1} - \widehat{\beta}_0 - \widehat{\beta}_1 y_{i2} - \widehat{\beta}_2 z_{i1}\right) &= 0 \\ \sum_{i=1}^{n} z_{i2}\left(y_{i1} - \widehat{\beta}_0 - \widehat{\beta}_1 y_{i2} - \widehat{\beta}_2 z_{i1}\right) &= 0 \end{aligned}$$

- 多元回归方程中有效的工具变量需要与内生的解释变量相关

> - $$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \nu_2, \quad \pi_2 \neq 0$$
> - 相当于两步回归方程的第一步，将内生变量进行外生化处理，即用外生变量表示内生变量。
> - 与一元回归时不同，该方程中**需要将工具变量以外的外生变量放入模型中**→ 保证工具变量和其他外生解释变量不太像，即在控制了外生解释变量后，工具变量依然对内生变量有影响。
> - 以y_2为因变量的方程被称为**简约方程式** [*reduced form equation*]

---

3. Two Stage Least Squares

> Q：**如何用多个工具变量表示一个内生变量?**

3.1 Definition of 2SLS

- 若内生解释变量的有效工具变量不止一个，即z_2和z_3，则结构方程 $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \nu_1$ 的条件包括：
  - z_2和z_3不应出现在结构方程中→ z_2和z_3本身不会影响y
  - z_2和z_3与误差项u_1无关
  - 上述两个要求为外生性假设服务。
- 由于$z_1 z_2$和z_3均与u_1无关，内生变量的最佳形式应为z_j的线性组合，即y_2^*

> - $$y_2^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3$$
> - 注：$\pi_2 \neq 0, \pi_3 \neq 0$
> - reg 内生变量 on **所有的外生变量**

- 与做t检验的一元回归不同，由于两个工具变量之间可能存在一定的共线性，因此使用F检验：

> - $H_0: \pi_2 = 0 \text{ and } \pi_3 = 0$
> - 通常而言，F值应该**大于10**，否则相关性假设不成立；如果只有一个工具变量，则希望t值**大于2.**

- 通过这种方式，得到式子

> - $$\begin{aligned} \sum_{i=1}^{n}\left(y_{i1} - \widehat{\beta}_0 - \widehat{\beta}_1 y_{i2} - \widehat{\beta}_2 z_{i1}\right) &= 0 \\ \sum_{i=1}^{n} z_{i1}\left(y_{i1} - \widehat{\beta}_0 - \widehat{\beta}_1 y_{i2} - \widehat{\beta}_2 z_{i1}\right) &= 0 \\ \sum_{i=1}^{n} \hat{y}_{i2}^*\left(y_{i1} - \widehat{\beta}_0 - \widehat{\beta}_1 y_{i2} - \widehat{\beta}_2 z_{i1}\right) &= 0 \end{aligned}$$

- $$\hat{y}_2^* = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3$$

- 由很多工具变量构成的IV估计量又被称为两阶段最小二乘估计量 [*two stage least squares* (2SLS) *estimator*]
  - 1st stage：回归简约式方程，得到拟合值$\hat{y}_2$
  - 2nd stage：OLS回归，reg y_1 on $\hat{y}_2$和z_1
  - 一步法stata命令：两种方法
    - ```
      ssc install ivreg2
      ivreg2 y1 (y2=z2 z3) z1, first, robust
      ivreg2 y1 (y2=z2 z3) z1
      ```
    - ```
      ivregress 2sls y1 (y2=z2 z3) z1, first, robust
      ivregress 2sls y1 (y2=z2 z3) z1
      ```
  - 相比于一步法，手动两步法会系统地低估标准误、增大显著性→ 我们需要避免手动使用两步法

3.2 The (Asymptotic) Variance of the 2SLS Estimator

- **2SLS估计量的渐进方差**：
  - $$\frac{\sigma^2}{\widehat{SST}_{\hat{y}_2}\left(1-\hat{R}^2_{\hat{y}_2}\right)}$$
  - $\sigma^2 = \text{Var}\,(u_1), \widehat{SST}_2$为 $\hat{y}_2$的总变动SST，$\hat{R}^2_2$ 是 reg $\hat{y}_2$ on 结构方程中所有外生变量[**不包括工具变量!**]的R^2
- **2SLS与OLS方差比较**：
  - $\hat{y}_2$的变动幅度小于y_2，则$\widehat{SST}_2$更小
  - 显然reg $\hat{y}_2$ on 其他外生变量的R^2要大于reg y_2 on 其他外生变量的R^2
  - → **2SLS的方差一定大于OLS估计量的方差**，多元回归模型的估计量方差一定大于一元回归。

3.3 Multiple Endogenous Explanatory Variables

> Q: 如何处理多个内生解释变量的情况?

- $$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + \beta_4 z_2 + \beta_5 z_3 + u_1$$
- ***Order condition***：工具变量的个数必须≥结构模型中内生变量的个数

---

4. Testing for Endogeneity

- 相比于OLS，2SLS的标准误更大，因此需要先检验模型是否存在内生性。
- 将怀疑具有内生性的变量y_2 与结构方程中所有的外生变量与工具变量进行回归，得到$\hat{\nu}_2$
  - $$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + \nu_2$$
- 将$\hat{\nu}_2$加入结构方程中，检验$\hat{\nu}_2$的显著性。如果系数δ_1显著不为0，则y_2确实是内生的
  - $$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{\nu}_2 + \text{ error}$$
  - 之所以在第一步中加入工具变量，是为了防止在这一步的回归方程中出现完全共线。

11. Pooled Cross-Sectional and Panel Data

2022.06.04

---

1. Pooled Cross-Sections

1.1 What is it?

- 合并的截面数据并非面板数据：合并的截面数据每次都是重新抽样，每一个截面都是独立的，而面板数据则是追踪同一个个体。
- **应用1**：可以观察变量随时间变化的趋势，比如1990年时适龄妇女的生育意愿和2020年时适龄妇女的生育意愿是否具有显著差异。
  - $$\begin{aligned}\text{Kids} &= \beta_0 + \beta_1 \text{ educ} + \beta_2 \text{ age} + \beta_3 \text{age}^2 + \beta_4 \text{ black} + \beta_5 \text{ east} \\ &+ \beta_6 \text{ northcen} + \beta_7 \text{ west} + \beta_8 \text{ farm} + \beta_9 \text{ othrural} + \beta_{10} \text{ town} + \beta_{11} \text{ smcity} \\ &+ \beta_{12} y74 + \beta_{13} y76 + \beta_{14} y78 + \beta_{15} y80 + \beta_{16} y82 + \beta_{17} y84 + u\end{aligned}$$
  - 挑战：进行简单的对比可能会忽略群体内部构成变动带来的影响→ 可能1990年和2020年女性的生育意愿本身并没有发生变化，但是城乡人口比例变化、受教育程度比例变化使得两个年份的生育孩子的数量完全不同→ 加入一些人口学特征作为控制变量。
    - 农村的女性生孩子的意愿不变，但是农村女性的数量变少，所以女性生育率下降。
  - 注：72年是base group
  - 如果年份前的系数显著，则说明时间本身会影响生孩子的数量。如果不显著，则说明更有可能是群体构成的变化导致的生育率下降。
- **应用2**：观察 x 和 y 的关系是否随着时间变化而变化，如教育回报的变化趋势、性别歧视的变化趋势等。
  - $$\begin{aligned}\log(\text{ wage }) &= \beta_0 + \delta_0 y85 + \beta_1 \text{ educ} + \delta_1 \text{ educ} \times y85 + \beta_2 \text{ exper} \\ &+ \beta_3 \text{ exper }^2 + \beta_4 \text{ union} + \beta_5 \text{ female} + \delta_5 \text{ female} \times y85 + u\end{aligned}$$
  - δ_1识别的是两年的教育回报的差异。如果δ_1显著为正，说明教育回报提高。
  - δ_5识别的是两年的男女工资差异的变化。由于β_5一定为负数，如果δ_5显著为正，说明85年时男女的工资差异缩小。
  - δ_0识别的是78年和85年的工资变化。平均工资提高，$\delta_0 > 0$

1.2 The Chow Test for Structural Change across Time

- 由于年份是虚拟变量，因此检验方式和虚拟变量的Chow Test一致。
- **步骤**
  - 对限制性模型做回归，假设不同年份的系数相同，得到SSR_r。
  - 一共T个时间段，对每一个时间段做回归，得到SSR_{ur}
    - $$SSR_{ur} = SSR_1 + SSR_2 + \cdots + SSR_T$$
  - 计算F值，
    - $$\frac{(SSR_r - SSR_{ur})}{SSR_{ur}} \cdot \frac{n - T - Tk}{(T-1)k}$$
- 如果F值显著，说明随着时间的变化，回归方程本身也发生了变化。

1.3 Policy Analysis with Pooled Cross-Sectional Data

- 我们无法同时观测到结果以及反事实结果。
  -

	$D_i = 1$ (participants)	$D_i = 0$ (non-participants)
$y_{ti} + \Delta_i$	observable	unobservable (counterfactual)
y_{ti}	unobservable (counterfactual)	observable

- 我们可以通过回归模型估计因果关系，但是我们只能估计平均因果效应，无法衡量对于每一个个体而言的因果效应。
- **建立反事实情况的三种方法**：
  - ***Cross-section comparison*** 同一时点，分成实验组和控制组→ 横比。
  - ***Before and after comparison*** 同一组，比较干预前与干预后→ 纵比。
    - 比如二胎放开政策，不存在部分人放开，部分人不放开。
  - ***Difference-in-differences estimation*** 双重差分法[DiD]
- **变量设定**
  - y 代表因变量，x代表既影响因变量、有影响是否参与干预的自变量。
  - D：虚拟变量，1参加0不参加。
  - T：虚拟变量，1代表干预后的t，0代表未干预的t'

1.4 Cross-Section Comparison

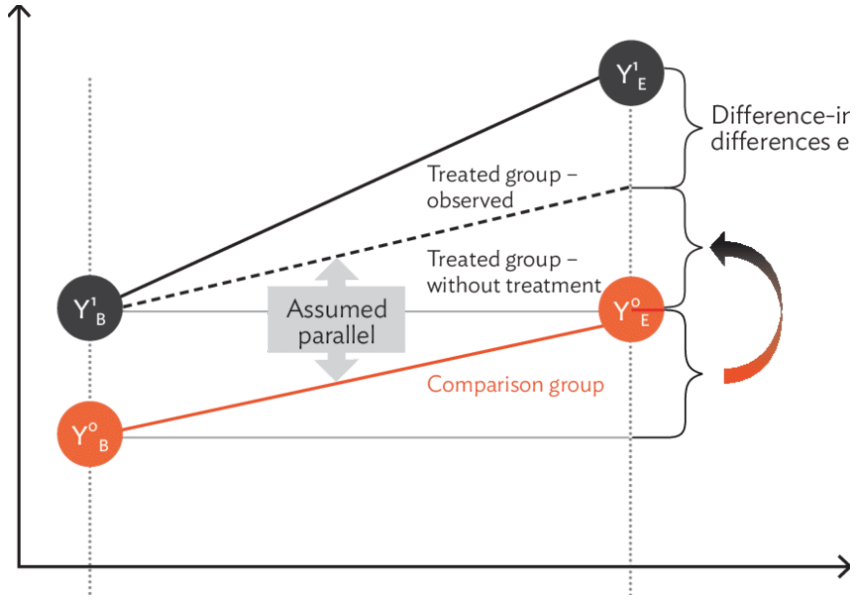
- 在使用这个方法之前，需要满足**识别假设** [*Identification assumption*]:
  - $$E\left(y_t \mid x_1, \ldots, x_k, D=1\right)=E\left(y_t \mid x_1, \ldots, x_k, D=0\right)$$
  - 注：$D=1$时，y_t 是无法观测的。因为我们需要用$D=0$ 时的因变量值构建反事实情况，因此要求二者相同。
  - 即，假设**实验组和控制组干预前状态相同**。
- **回归模型**：
  - $$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_k x_{ik}+\alpha D_i+u_i$$
  - $$\widehat{\alpha}=E\left(y_t+\Delta \mid x_1, \ldots, x_k, D=1\right)-E\left(y_t \mid x_1, \ldots, x_k, D=0\right)$$
  - 如果$\hat{\alpha}$显著不为零，则干预有效果。
- **问题**：识别假设不成立的两种情况
  - 自选择[*Self-selection*]/ 未观察到的异质性[*unobserved heterogeneity*]
  - 存在一些无法观察到的因素影响了是否被干预→ 关键在于控制不同组之间、影响y 的差异

1.5 Before-After Comparison

- **识别假设**：
  - $$E\left(y_t \mid x_1, \ldots, x_k, D=1\right)=E\left(y_{t'} \mid x_1, \ldots, x_k, D=1\right)$$
  - 前后比较，只有$D=1$一个组。$E\left(y_t \mid x_1, \ldots, x_k, D=1\right)$是无法观测到的反事实状态 [实际上观察到的是$y_t+\Delta_i$]
  - 即如果没有干预，两个时点的因变量是不变的→ 假设**因变量不随时间的变化而变化**。基于这一假定，我们可以通过上一期的数据构建当期的反事实情况。
- **回归模型**：
  - $$y_{it}=\beta_0+\beta_1 x_{it1}+\beta_2 x_{it2}+\ldots+\beta_k x_{itk}+\gamma T_{it}+u_{it}$$
    - 固定随时间变化，同时影响因变量的自变量。
  - $$\widehat{\gamma}=E\left(y_t+\Delta \mid x_1, \ldots, x_k, D=1\right)-E\left(y_{t'} \mid x_1, \ldots, x_k, D=1\right)$$
  - 如果$\hat{\gamma}$显著，则说明干预有效果。
- **问题**：
  - 需要面板数据或合并的截面数据，对数据的要求比较高。
  - 对商业周期敏感

- **Ashenfelter's Dip**：政策的执行会影响人的预期。比如研究提供小额贷款对就业的影响，有的人听说了有这个政策，在t'的时间点上的行为已经发生了变化，为了参加这个政策而不会认真的找工作，因此高估了政策的影响。同理，开放二胎政策也会使得样本产生"马上放开"的预期，从而使得政策开始前生二胎的现象减少，进而高估开放二胎对生育意愿的影响。
  - 如果能自选数据，不应选择与政策执行过近的时间点。

1.6 Difference-in-Differences (DiD) Estimation ❖❖

- 识别假设
  - $$\begin{aligned} &E\left(y_t+\Delta \mid x_1, \ldots, x_k, D=1\right)-E\left(y_t \mid x_1, \ldots, x_k, D=1\right) \\ &=E\left(y_t+\Delta-y_{t'} \mid x_1, \ldots, x_k, D=1\right)-E\left(y_t-y_{t'} \mid x_1, \ldots, x_k, D=0\right) \end{aligned}$$
  - 该假定本质上是一种平行假定，即假定参与政策的人，如果没有参与政策[反事实情况]的变化趋势[斜率]与没参加政策的人的**变化趋势相同**[斜率相同、平行]。参加政策的人以及未参加政策的人在因变量上的均值差异在每个时间点上都是相同的。→ 通过**平移**构建反事实。
  - 

    - C-S比较：$Y_E^1-Y_E^0$
    - B-A比较：$Y_E^1-Y_B^1$
    - DID: $(Y_E^1-Y_B^1)-(Y_E^0-Y_B^0)$ 或$(Y_E^1-Y_E^0)-(Y_B^1-Y_B^0)$
- 优势：
  - 相比于B-A比较，DID放松了对于时间假定，即**允许样本在不同的时点上存在差异**；相比于C-S比较，DID放松了对于群体差异的假定，即**允许对照组和实验组在干预开始前存在差异**。
  - 消除了未被观察到的异质性的影响
- 回归模型
  - $$y_{it}=\beta_0+\beta_1 x_{it1}+\beta_2 x_{it2}+\ldots+\beta_k x_{itk}+\delta_1 D_i+\delta_2 T_{it}+\delta_3\left(D_i \cdot T_{it}\right)+u_{it}$$
  - 系数解读：
    - $$\begin{aligned} \hat{\delta}_3=&\left[E\left(y_t+\Delta \mid x_1, \ldots, x_k, D=1\right)-E\left(y_{t'} \mid x_1, \ldots, x_k, D=1\right)\right] \\ &-\left[E\left(y_t \mid x_1, \ldots, x_k, D=0\right)-E\left(y_{t'} \mid x_1, \ldots, x_k, D=0\right)\right] \end{aligned}$$
	实施前	实施后	Difference
干预组	$\beta_0+\delta_1$	$\beta_0+\delta_1+\delta_2+\delta_3$	$\delta_2+\delta_3$
对照组	β_0	$\beta_0+\delta_2$	δ_2
Difference	δ_1	$\delta_1+\delta_3$	δ_3 (DiD)
    - 通过"B-A效应-C-S效应"或"C-S效应-B-A效应"得到的$\hat{\delta}_3$ 衡量了政策干预的效果。
- 问题：假设是否成立?
  - 需要面板数据或合并的截面数据，对数据的要求高。
  - 存在扰动使得参与者与非参与者变化的趋势不一致，违背假定。
  - Ashenfelter's Dip：影响预期
- 案例：垃圾焚烧厂的建立对房价的影响

    - 对照组和实验组的选择：以垃圾焚烧厂为中心画圈，圈之内的是实验组。80年建垃圾场，选择了78年和81年的数据。
    - 分别跑回归，发现是否在垃圾场附近对房价的影响系数变大，同时R^2也变大。
        - $$\text{rprice}^{t}=\gamma_0^t+\gamma_1^t\ \text{nearinc}\ +u^t, t=1978\ \text{or}\ 1981$$
    - 一起跑回归
        - $$\text{rprice}\ =\beta_1+\beta_2 y81+\beta_3\ \text{nearinc}\ +\delta_1 y81\times\ \text{nearinc}\ +u$$
        - $$\hat{\delta}_1=\left(\overline{\text{eprice}}_{81,\text{near}}-\overline{\text{eprice}}_{81,\text{far}}\right)-\left(\overline{\text{eprice}}_{78,\text{near}}-\overline{\text{eprice}}_{78,\text{far}}\right)$$
    - 问题：
        - 该回归假定如果圈内不修建垃圾焚烧厂，房价的变化和圈外是一致的→ 如果不满足呢？比如圈内房价相对下降也有可能是因为同期圈外修建了公路，和垃圾场关系不大。
        - → 需要对分析的数据与问题有清晰的了解
- **自然实验**
    - 外生事件[如政策变动]会导致**自然实验** [*Natural experiment (quasi-experiment)*]
    - 为评估自然实验的影响，应该按照基期-现期、实验组-对照组分为四类，评估平均效应。
        - $$\hat{\delta}=(\bar{y}_{2,T}-\bar{y}_{2,c})-(\bar{y}_{1,T}-\bar{y}_{1,c})$$

---

2. Panel Data

2.1 What is it?

- 面板数据是对于同一个体的反复观察。中级计量经济学我们研究最简单的情况，即**平衡面板**[*balanced panel*]，即每个个体每一期的数据都是完备的，不存在**数据磨损**[*attrition*]。
- **优势**：
    - 降低了对于其他数据的需求
    - 控制无法观察到的异质性
    - 能够确认因果方向
    - 有助于发现政策的长期效果
- **局限性**：
    - 数据收集困难
    - 加剧测量误差的影响
    - 无法解决选择性问题，数据磨损可能会使其更加恶化→ 留下来的样本和丢失的样本可能存在系统性差异
    - 如果关注的变量不随时间变化，如30岁之后人的受教育水平基本不变，用面板数据就很没有必要。

2.2 Two Period Panel Data Analysis

- **回归模型**：*fixed effects model/unobserved effect model*
    - $$y_{it}=\beta_0+\beta_1 x_{it}+\delta_0 d2_t+\alpha_i+u_{it}$$
    - $d2_t$是一个虚拟变量，如果是第2期取1，第1期取0.
    - 误差项分为两部分：
        - **不随时间变量的变量** α_i：如出生地、性别、种族
            - 又称：**固定效应**[*fixed effect*]/未观察到的异质性 [*unobserved heterogeneity*]
            - 去除固定效应是面板数据的主要优势→ 如何去掉？
        - **随时间变化的变量** u_{it}：如子女数量、年收入
- **去除固定效应的两个技巧**
    - Differencing→ **一阶差分模型** [*First-differenced model*]
    - Demeaning→ **固定效应模型** [*Fixed effect model*]

2.3 First-differenced (FD) model

- 对两期数据写成两个模型，二者相减则能够消除不随时间变化的因素。

> - $$y_{i2} = \beta_0 + \beta_1 x_{i2} + \delta_0 + \alpha_i + u_{i2}, (t = 2)$$ $$y_{i1} = \beta_0 + \beta_1 x_{i1} + \alpha_i + u_{i1}, (t = 1)$$
> - $$(y_{i2} - y_{i1}) = \delta_0 + \beta_1 (x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$$

- 得到一阶差分模型

> - $$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$
> - 所有未观测到的、不随时间变化的变量全部消掉。

- **使用OLS的关键假设**：
  - $\text{Cov}(\Delta u_i, \Delta x_i) = 0$，在两期随时间变化的特征$u$不应和自变量相关。
  - Δx_i 必须要有变动
- **系数解读**
  - δ_0 衡量的是两期的变化
  - β_1衡量的是x变化1单位，y的变化→ 按照原方程取解释

2.4 Fixed Effects (FE) Estimation

- 取均值，得到方程

> - $$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + \delta_0 \overline{d2} + \alpha_i + \bar{u}_i$$

- ***Demean***，得到方程

> - $$(y_{it} - \bar{y}_i) = \beta_1 (x_{it} - \bar{x}_i) + \delta_0 \left(d2_t - \overline{d2}\right) + (u_{it} - \bar{u}_i)$$

- 固定效应模型同样要求关键自变量和因变量需要随时间变化。
- 注意：固定效应估计的样本量为nT，而一阶差分模型的样本量为$n(T-1)$。如果$T = 2$，两种方法得到的结果完全相同。如果$T = 3$，FE更有效。