随机对照实验 + 二变量回归模型

RCT + Bivariate OLS model

杨点溢

Department of Government
London School of Economics and Political Science

February 23, 2024

本节课内容

随机对照实验

- 我们上节课提到了选择偏差（Selection Bias）是因果推断的一个大问题
- 避免选择偏差需要确保实验分组 (Treatment Status)独立于(independent from) 潜在结果 (potential outcomes) 和其他背景特征（background attributes）
  - $Y_i(0), Y_i(1), X \perp D_i$
- 其中最完美的方式就是随机分配(Random Assignment) 实验组 (Treatment Group) 和对照组 (Control Group)
  - 这便是随机对照试验(Randomised Controlled Trials)，又称RCT.

举例：RAND 医保实验 (RB-9174-HHS)

研究医保对健康的影响

- 招募一些没有医保的人作为样本
- 随机选择一部分人给与医保（实验组），其他人为对照组
  - 3958 个样本被分到了 14 个医保套餐之一
  - 有些医保非常全面 (comprehensive)，有的只涵盖重大疾病 (catestrophic coverage)
    - 理赔 95% 的医药费（有上限）
- 在一段时间后对比实验组和对照组的健康平均值
- 只要样本足够大，这个平均值的差就是医保的平均影响
- 复杂的问题 (Complications)
  - 随机分配的方式 (Assignment Rule)
  - 脱离实验 (Attrition)
    - 实验从 1974 年进行到 1982 年
- 需要对比实验组和对照组是否“想象”
  - 至少在可观察特征 (Observable Characteristics) 上相像
  - 平衡性检验 (Balance Check)→Table 1

对比平均值 Comparing means

Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
			A. Health			
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)

- 有保险（4.01）和没保险 (3.70) 之间的差别会不会只是侥幸呢？
- 还是两者之间真的有区别的（有处理影响）？

本节课内容

1. 回顾随机对照实验 RCT
2. 假说检验 Hypothesis Testing
3. 回到 RAND 医保实验
4. 可能影响实验有效性的威胁
5. 线性回归 Linear Regression
6. 衡量回归模型的拟合程度
7. 回归模型的假设
8. OLS 的特性

假说检验 Hypothesis Testing

- 在量化研究之中我们经常进行假说检验 (Hypothesis Testing)
  - 有没有影响；有没有关系；有没有增加/减少

假说检验 Hypothesis Testing

- 在量化研究之中我们经常进行假说检验 (Hypothesis Testing)
  - 有没有影响；有没有关系；有没有增加/减少
- 零假说 (Null Hypothesis) H_0: 默认的设定
  - 没有影响，没有关系，没有增加/减少
  - 通常要估计的参数 parameter=0 （或是其他默认值）

假说检验 Hypothesis Testing

- 在量化研究之中我们经常进行假说检验 (Hypothesis Testing)
  - 有没有影响；有没有关系；有没有增加/减少
- 零假说 (Null Hypothesis) H_0: 默认的设定
  - 没有影响，没有关系，没有增加/减少
  - 通常要估计的参数 parameter=0 （或是其他默认值）
- 另类假说 (Alternative Hypothesis) H_1: 要证明的设定
  - 有影响，有关系，有增加/减少
  - 通常要估计的参数 parameter≠0 （或是其他默认值）
    - 双尾测试 (two-tailed)

假说检验 Hypothesis Testing

- 在量化研究之中我们经常进行假说检验 (Hypothesis Testing)
  - 有没有影响；有没有关系；有没有增加/减少
- 零假说 (Null Hypothesis) H_0: 默认的设定
  - 没有影响，没有关系，没有增加/减少
  - 通常要估计的参数 parameter=0 （或是其他默认值）
- 另类假说 (Alternative Hypothesis) H_1: 要证明的设定
  - 有影响，有关系，有增加/减少
  - 通常要估计的参数 parameter≠0 （或是其他默认值）
    - 双尾测试 (two-tailed)
  - 在理论背书的情况下，也有可能是参数 parameter> 或者 <0（其他默认值）.
    - 单尾测试 (one-tailed)
    - 需要非常令人信服的理由（一般不推荐做）
  - 双尾要比单尾更保守

假说检验 Hypothesis Testing

- 在量化研究之中我们经常进行假说检验 (Hypothesis Testing)
  - 有没有影响；有没有关系；有没有增加/减少
- 零假说 (Null Hypothesis) H_0: 默认的设定
  - 没有影响，没有关系，没有增加/减少
  - 通常要估计的参数 parameter=0 （或是其他默认值）
- 另类假说 (Alternative Hypothesis) H_1: 要证明的设定
  - 有影响，有关系，有增加/减少
  - 通常要估计的参数 parameter≠0 （或是其他默认值）
    - 双尾测试 (two-tailed)
  - 在理论背书的情况下，也有可能是参数 parameter> 或者 <0（其他默认值）.
    - 单尾测试 (one-tailed)
    - 需要非常令人信服的理由（一般不推荐做）
  - 双尾要比单尾更保守
- 在假设零假说成立的情况下，如果观察到的估计值比目前更极端的可能性低于显著性(significance level)，我们就有证据拒绝 (reject)H0，支持 H1.
  - 显著性（α）等级一般为 5%(0.05)。

t 检验 t-test

t 统计量（statistic）的公式：

- $t = \frac{\text{estimator} - \text{hypothesised value}}{\text{standard error of the estimator}}$

或者以 μ 为估计量（estimator），

- $t = \frac{\hat{\mu} - \tilde{\mu}}{\sigma_{\hat{\mu}}}$

如果我们是对比两个估计量（即 $\tilde{\mu}$ 也是估计量）的话，t 统计量公式如下：

- $t = \frac{\hat{\mu} - \tilde{\mu}}{\sqrt{\sigma_{\hat{\mu}}^2 + \sigma_{\tilde{\mu}}^2}}$

中央极限定理 Central Limit Theorem

- 当 n 足够大时，任意 Y 的平均值/期望值 $\bar{Y}$ 的分部都接近正态分布(Normal Distribution).
  - 不取决于 Y 是如何分部的。

$$\bar{Y} \sim N(\mu, \frac{\sigma_Y^2}{n})$$

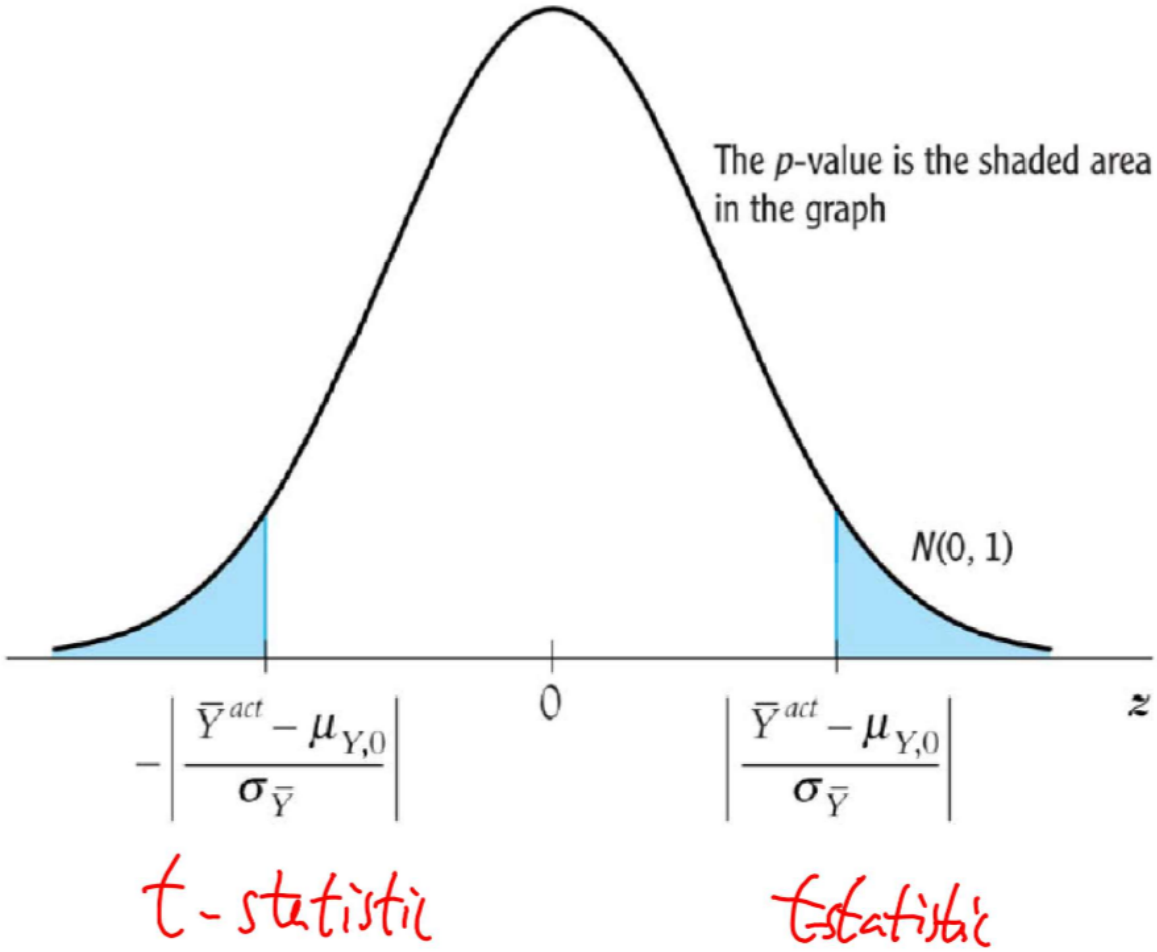
- 那么 t 估计量，在 $E(Y_i) = \mu$ 的假设下，也呈现正态分布：

$$t = \frac{\bar{Y} - \mu}{\sigma_Y / \sqrt{n}}$$

- 这个信息非常有用!

计算 P 值 p-value

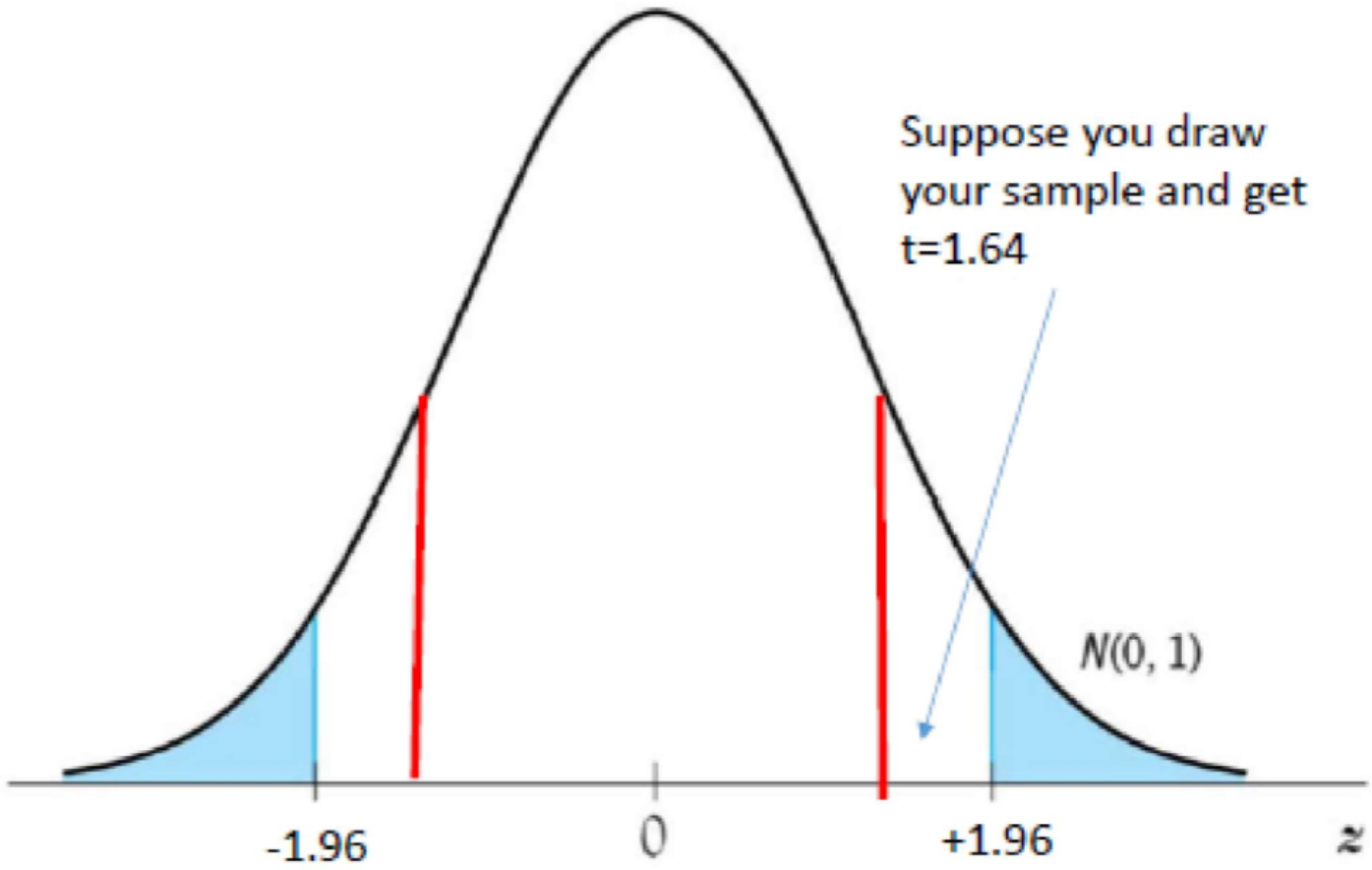
- P 值 (p-value): 假设 H_0 正确的情况下，观察到更极端的情况（t 统计量）的概率



显著性 Significance Level

- 国际标准 5% $\rightarrow$ 在 $|t| > 1.96$ 时拒绝 H0
- 5% 是我们测试的显著性(significance level)：（在 H0 正确的情况下）错误拒绝 H0 的几率
  - 假阳 (False Positive)/type 1 error/α
  - 如果 H0 是对的，那么平均每 20 次测验就会有一次 t>1.96
- 如果 t>1.96 那么我们说我们有显著的统计证据支持 H1.
- 有时候你会看到不同的显著性标准
  - $1\% \rightarrow |t| > 2.58$
  - $10\% \rightarrow |t| > 1.64$
- p 值给你的信息更多：不仅仅是拒绝还是接受 H0/H1，更告诉你需要在哪个显著性下可以拒绝 H0。
  - 比如 p 值为 0.06: 过不了 5% 测验，但是可以过 10%。

例子：10%



- 那么我们不能再 5% 的显著性下拒绝 H0
- 我们的 p 值为 10

置信区间 Confidence Interval

- 与其问我们观察到的现实符不符合 H0，我们也可以问我们的观察符合哪些H0
  - 这便是置信区间Confidence Interval
- 比如说，μ 的 95% 置信区间 $\rightarrow$ 就是众多的随机取样中，μ 的实际值会存在与 95% 个这样的区间

$$\bar{Y} \pm 1.96 \times SE(\bar{Y})$$

本节课内容

回到 RAND 医保实验 (Brook et al., 2006)

- 我们首先要做的事：
  - 确保实验组和对照组的相似性 $\rightarrow$ 平衡测验 (balance test)
  - 对比实验前两组的 Y(结果 outcome) 是不是一样的 $\rightarrow$ 没有选择偏差
- 衡量处理影响 (treatment effect)
  - 对比试验后两组的 Y(结果 outcome) 是不是一样的
  - 做 t 检验
    - 假说 Hypotheses:
    - $H_0 : \mu_T = \mu_C$ vs $H_0 : \mu_T \neq \mu_C$
    - R/STATA 可以自动做 t 检验并总结结果

Rand 实验： 平衡检测/1

Demographic characteristics and baseline health in the RAND HIE

	Means	Differences between plan groups			
	Catastrophic plan (1)	Deductible − catastrophic (2)	Coinsurance − catastrophic (3)	Free − catastrophic (4)	Any insurance − catastrophic (5)
A. Demographic characteristics					
Female	.560	−.023	−.025	−.038	−.030
		(.016)	(.015)	(.015)	(.013)
Nonwhite	.172	−.019	−.027	−.028	−.025
		(.027)	(.025)	(.025)	(.022)
Age	32.4	.56	.97	.43	.64
	[12.9]	(.68)	(.65)	(.61)	(.54)
Education	12.1	−.16	−.06	−.26	−.17
	[2.9]	(.19)	(.19)	(.18)	(.16)
Family income	31,603	−2,104	970	−976	−654
	[18,148]	(1,384)	(1,389)	(1,345)	(1,181)
Hospitalized last year	.115	.004	−.002	.001	.001
		(.016)	(.015)	(.015)	(.013)

Rand 实验：平衡检测/2

B. Baseline health variables

General health index	70.9	−1.44	.21	−1.31	−.93
	[14.9]	(.95)	(.92)	(.87)	(.77)
Cholesterol (mg/dl)	207	−1.42	−1.93	−5.25	−3.19
	[40]	(2.99)	(2.76)	(2.70)	(2.29)
Systolic blood	122	2.32	.91	1.12	1.39
pressure (mm Hg)	[17]	(1.15)	(1.08)	(1.01)	(.90)
Mental health index	73.8	−.12	1.19	.89	.71
	[14.3]	(.82)	(.81)	(.77)	(.68)
Number enrolled	759	881	1,022	1,295	3,198

Rand 实验：效果 (医疗服务使用)

Health expenditure and health outcomes in the RAND HIE

	Means	Differences between plan groups			
	Catastrophic plan (1)	Deductible − catastrophic (2)	Coinsurance − catastrophic (3)	Free − catastrophic (4)	Any insurance − catastrophic (5)
A. Health-care use					
Face-to-face visits	2.78	.19	.48	1.66	.90
	[5.50]	(.25)	(.24)	(.25)	(.20)
Outpatient expenses	248	42	60	169	101
	[488]	(21)	(21)	(20)	(17)
Hospital admissions	.099	.016	.002	.029	.017
	[.379]	(.011)	(.011)	(.010)	(.009)
Inpatient expenses	388	72	93	116	97
	[2,308]	(69)	(73)	(60)	(53)
Total expenses	636	114	152	285	198
	[2,535]	(79)	(85)	(72)	(63)

Rand 实验：效果 (健康情况)

B. Health outcomes

General health index	68.5	−.87	.61	−.78	−.36
	[15.9]	(.96)	(.90)	(.87)	(.77)
Cholesterol (mg/dl)	203	.69	−2.31	−1.83	−1.32
	[42]	(2.57)	(2.47)	(2.39)	(2.08)
Systolic blood pressure (mm Hg)	122	1.17	−1.39	−.52	−.36
	[19]	(1.06)	(.99)	(.93)	(.85)
Mental health index	75.5	.45	1.07	.43	.64
	[14.8]	(.91)	(.87)	(.83)	(.75)
Number enrolled	759	881	1,022	1,295	3,198

来跑一跑代码试试做简单的测试吧！

本节课内容

实验（内部）有效性的威胁

1. 随机化失败 ⇒ 永远自己做随机
2. 溢出效应 ⇒ 实验设计要注意
3. 不服从 (Non-compliance)
   - 指处理组没接受处理或对照组得到处理
   - 但我们还是可以计算处理意向的影响 (Intention-to-Treat or ITT)
   - 我们也可以通过工具变量 (IV) 来计算出遵从者平均处理效应 (complier/local average treatment effect)
4. 缺失/磨损 (Attrition)
   - 指部分实验对象离开实验
   - 如果有差别磨损（拥有特定特征的人更容易离开实验）会影响实验结果
   - ⇒ 紧密追踪、给对象继续参与实验的激励、经常联系。

本节课内容

1. 回顾随机对照实验 RCT
2. 假说检验 Hypothesis Testing
3. 回到 RAND 医保实验
4. 可能影响实验有效性的威胁
5. **线性回归 Linear Regression**
6. 衡量回归模型的拟合程度
7. 回归模型的假设
8. OLS 的特性

线性回归 Linear Regression：动机

- 探究 X 的 1 单位变化对 Y 的影响
- 但因果推断有点复杂，我们先总结一下描述性关系
  - X 的变化伴随着Y 怎样变化？
- 但哪怕只探索描述性关系，我们仍然需要选择好的... 来了解人口参数
  - 样本 Sample；
  - 估计量 Estimator;
  - 测试 Tests
- 那刚才做的 t-test 为什么不能用线性回归来代替呢？
  - 也不是不行!
  - 对影响的测量是一致的，但是标准误差（se）的计算有小差别
    - 为什么有小差别呢？卖个关子

人口回归线 Population Regression Line

- 考试分数如何随着班级大小变化？(Stock and Watson, 2014, pp.155-160)
- How do test scores change with a change in class size?

$$TestScore = \beta_0 + \beta_1 ClassSize$$

- 注意：我们在强加（impose）一个线性 (linear) 关系

$$\beta_1 = \frac{\Delta TestScore}{\Delta ClassSize}$$

- β_1 是 *TestScore* 伴随 *ClassSize* 变化 1 个单位的变化
  - the change in *TestScore* corresponding to a unit change in *ClassSize*
- 我们需要一个样本sample 的数据来估计(estimate)β_1

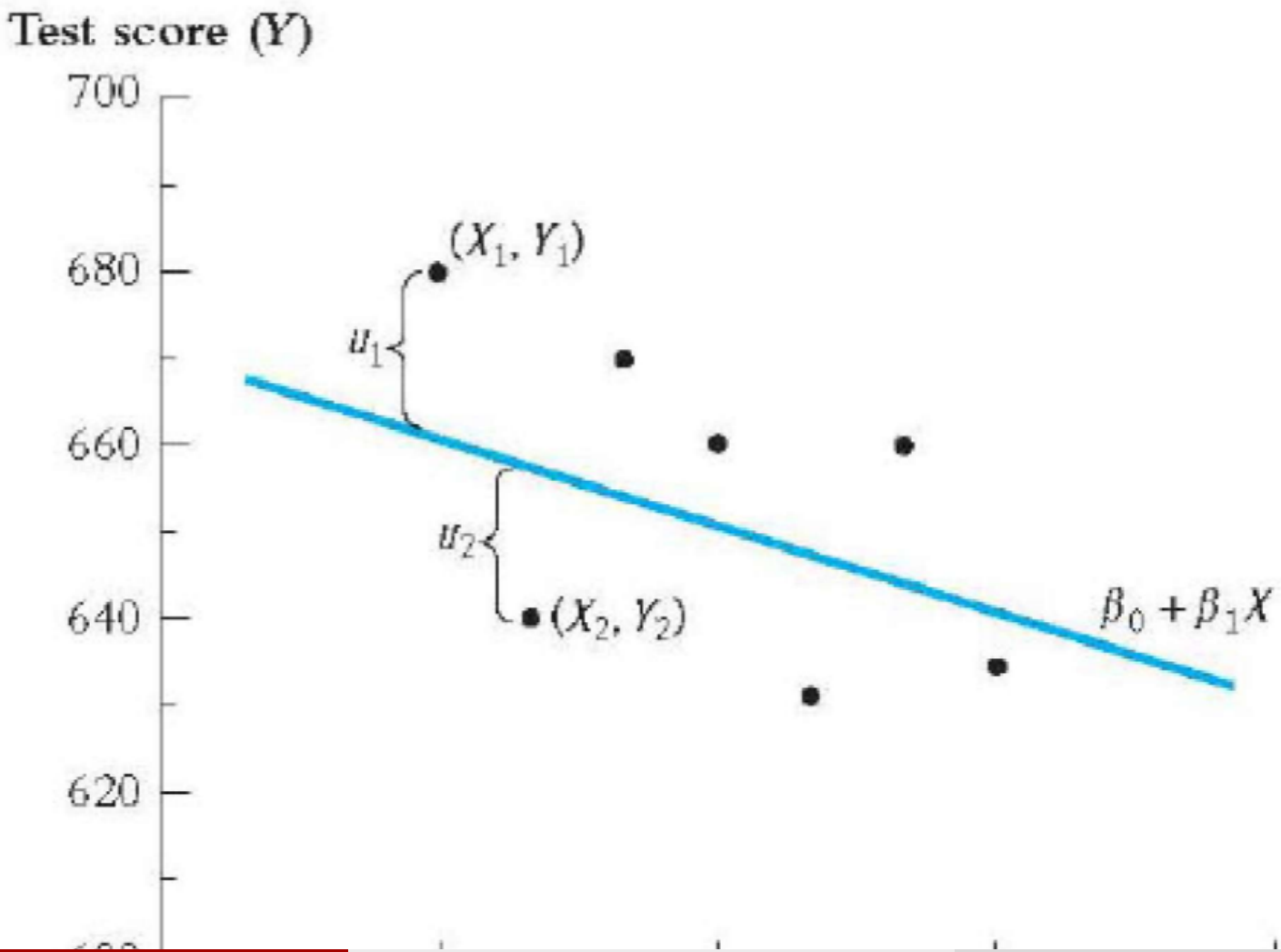
人口回归模型 Population Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

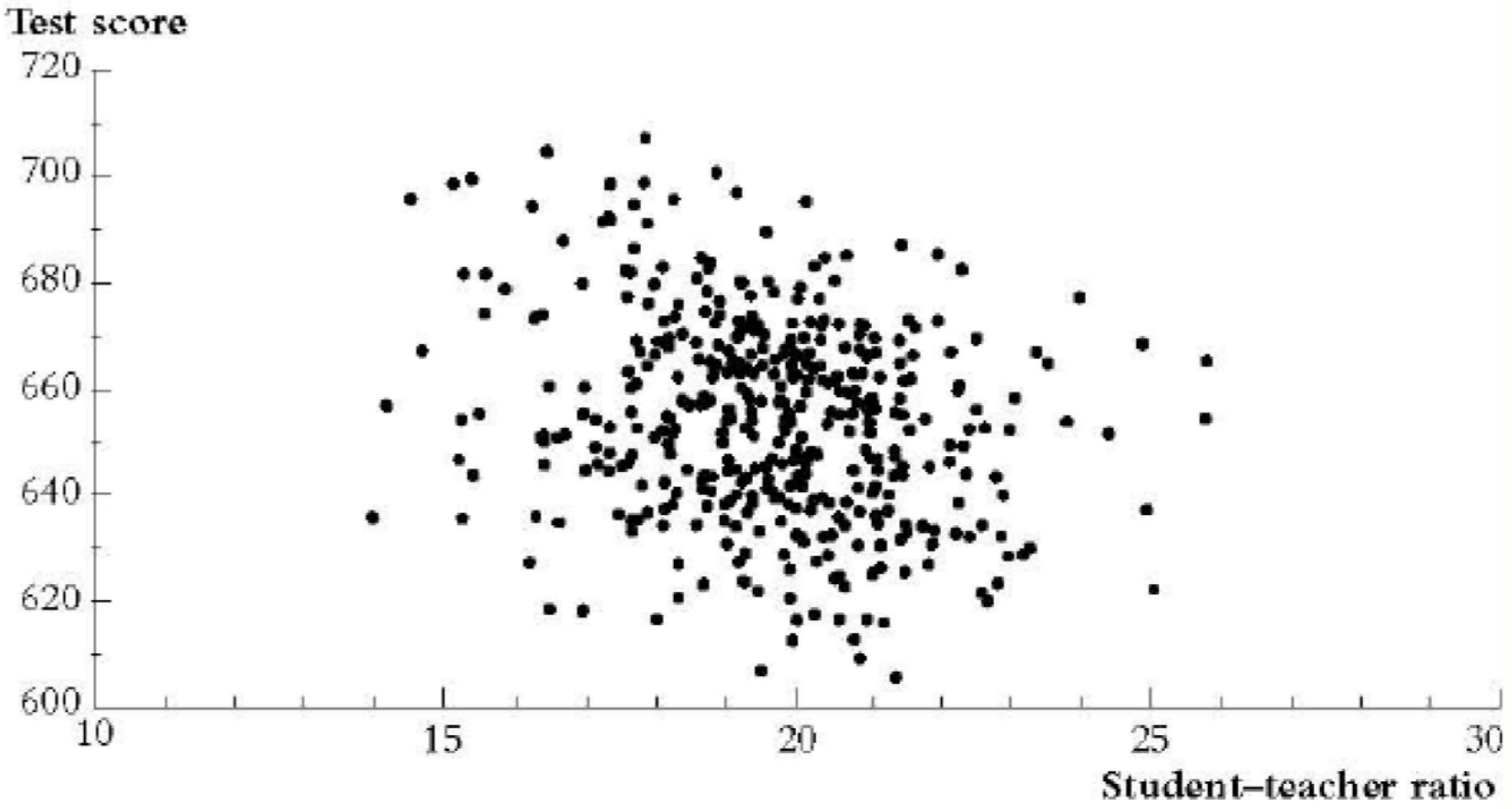
- β_i =intercept(截距)
- β_1 =slope（斜率）
- Y_i =dependent variable（因变量）
- X_i =independent variable（自变量）
  - 也叫 explanatory variable（解释变量）
  - 或者 regressor (回归量)
- u_i =error term（误差项）
  - 包含其他影响 Y_i 的因素和 Y_i 中可能的测量误差（measurement errors）

举个栗子：7 个观察值 (observations) (Stock and Watson, 2014, p.159)

1 个观察值 $=1$ 对 (X_i, Y_i)



真实的散点图 Scatter Plot：师生比与考试分数 (Stock and Watson, 2014, p.161)



最小二乘法 Least squares estimators

- 最小化估计值和观察值的差的平方和 (Minimise the sum of squares between the observations and the estimate)
  - 因此最小化了观察值到估计值的平均竖直距离 ^{2} (hence minimise the average squared distance between the observations and the estimate)
- 样本的平均值 $\bar{Y}$ 是 μ_Y 的一个最小二乘估计量，因为他是以下的解：(The sample mean $\bar{Y}$ is a least square estimator of μ_Y. It solves)

$$\min_m \sum (Y_i - m)^2$$

- OLS(Ordinary Least Squares) 对 β_0 和 β_1 是在求解

$$\min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

OLS 估计量，预测值和残差 Residuals

The OLS estimators of the slope β_1 and the intercept β_0 are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \tag{4.7}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{4.8}$$

The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, i = 1, \ldots, n \tag{4.9}$$

$$\hat{u}_i = Y_i - \hat{Y}_i, i = 1, \ldots, n. \tag{4.10}$$

拟合值 Fitted Values 和残差 Residuals

对于每一个数据点 Data Point 我们可以区分

- 拟合值 Fitted values

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- 和残差 Residual

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

- 因此回归 regression 将每个数据点分成了两部分:

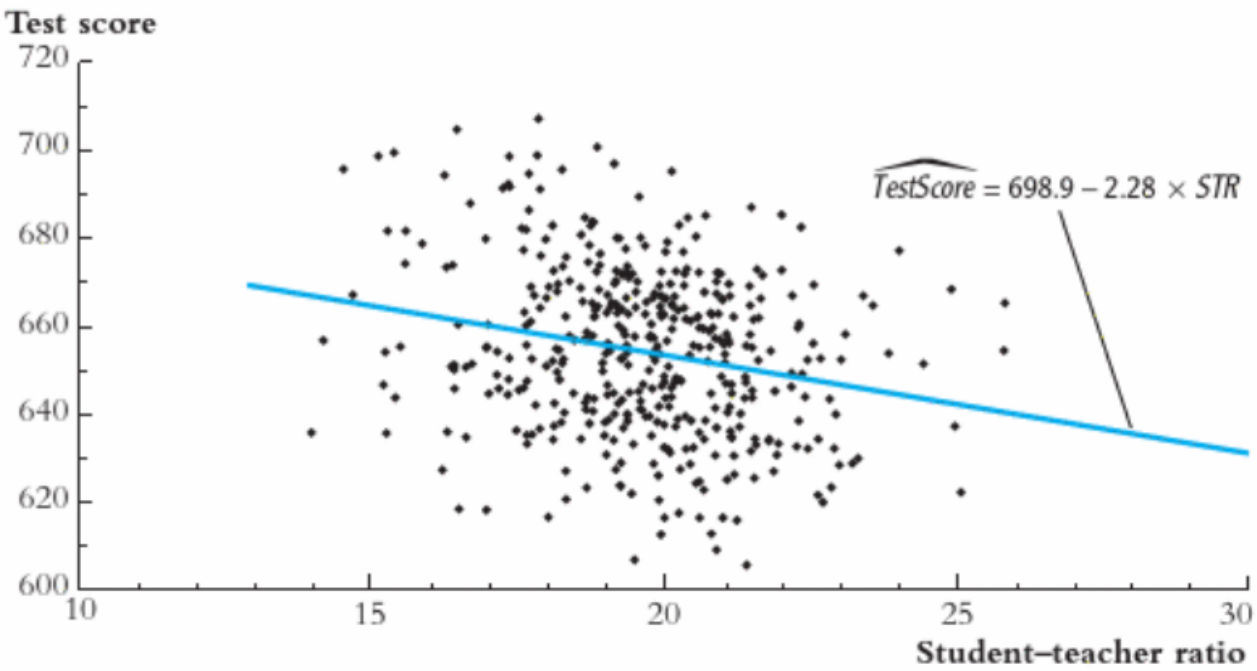
$$Y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 X_i}_{\text{fitted value}} + \underbrace{\hat{u}_i}_{\text{residual}}$$

- 回归线与残差 residual 是垂直的 (orthogonal)
  - 或者说独立的 independent

回到散点图 Scatter Plot

FIGURE 4.3 The Estimated Regression Line for the California Data

The estimated regression line shows a negative relationship between test scores and the student–teacher ratio. If class sizes fall by one student, the estimated regression predicts that test scores will increase by 2.28 points.



本节课内容

衡量拟合程度 Measures of Fit/1

- 回归 (regression) 的 R^2 衡量 Y_i 的多少方差 (Variance) 可以被 X_i 解释。
- 定义 $Y_i = \hat{Y}_i + \hat{u}_i$，则：

$$R^2 = \frac{\text{Explained sum of squares } (ESS)}{\text{Total sum of squares } (TSS)} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

或者

$$R^2 = 1 - \frac{\text{Sum of squares of the residuals } (SSR)}{TSS}$$

其中 $SSR = \sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}})^2$，因为 $\bar{\hat{u}} = \sum_{i=1}^{n} \frac{1}{n}\hat{u}_i = 0$

- $0 \leq R^2 \leq 1$

衡量拟合程度 Measures of Fit/2

回归的标准误差（Standard Error of the Regression）也是衡量观察值相对于回归线的离散程度的一个指标 (a measure of the spread of the observations around the regression line)

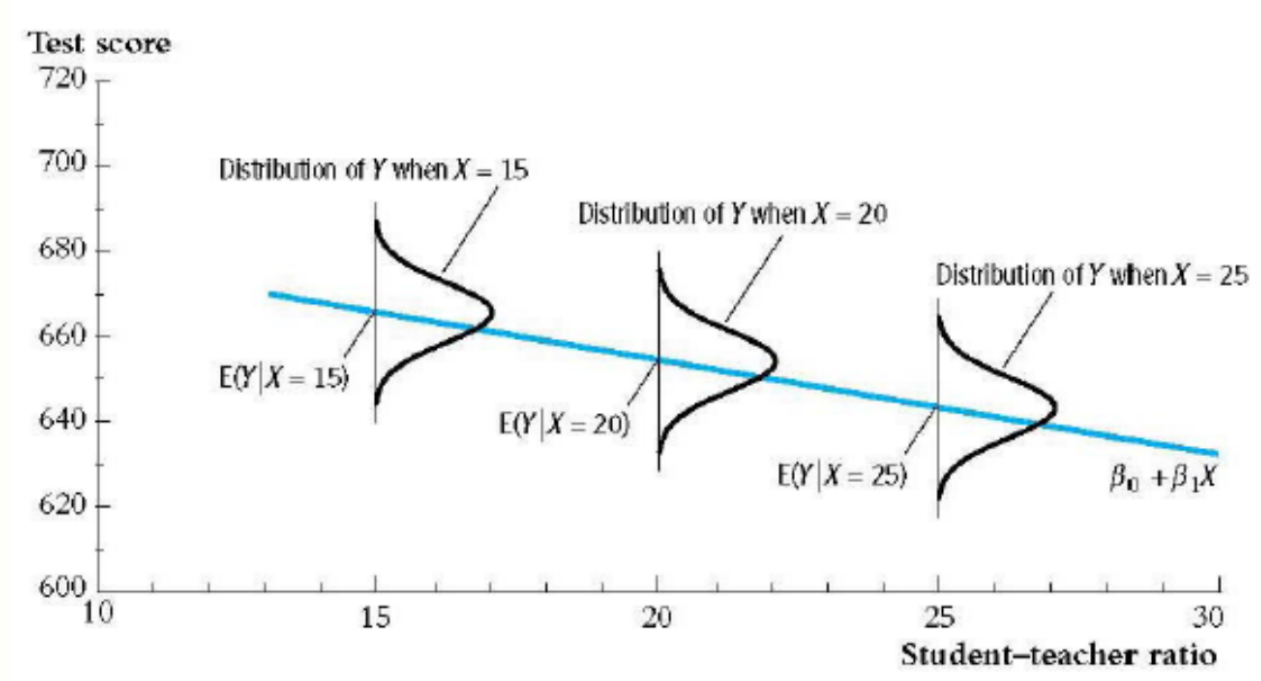
$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}})^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2} = \sqrt{\frac{SSR}{n-2}}$$

- 低 $R^2 \rightarrow X$ 只解释了 Y 的变化的一小部分
- 高 $SER \rightarrow$ 观察值（数据点）离回归线分散得比较远
- 低 R^2 和高 SER 不见得代表回归做的“不好”。这仅仅代表别的因素也解释 Y 的变化
- 而且，目前还不明确为什么我们要把尽量多的解释 Y 的因素涵盖在回归模型中（之后会讲到）

本节课内容

OLS 假设 1: $E(u_i|X_i) = 0$

1. $E(u_i|X_i) = 0$
   - 给定 X 时 u 的条件分布均值为零
   - the conditional distribution of u given X has zero mean



   - 其他影响 Y(并且被包含在 u 里) 的因素跟 X 不相关 (uncorrelated)
   - 相当于是说我们假设回归线是 $E(Y|X)$, 即给定 X 时的 Y 的条件平均值

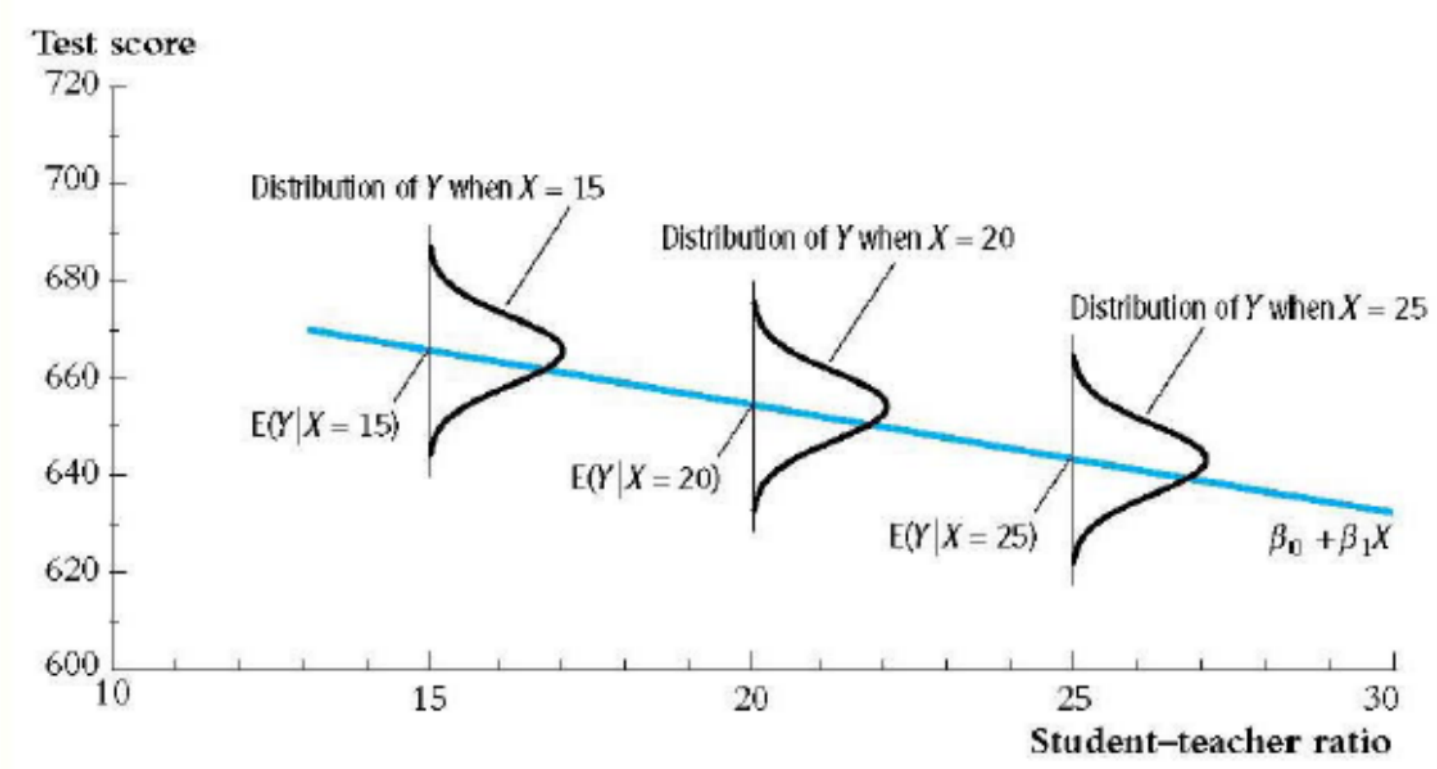
OLS 假设 2：iid

2. $(X_i, Y_i), i = 1, 2, ..., n$ 为独立且相同分布 (independently and identically distributed, iid)。

- 如果所有观察值都是抽取自同一个人口 $\rightarrow$ 相同分布 (identically distributed for all $i = 1, 2, ..., n$)
- 如果所有观察值都是随机抽取的 $\rightarrow (X_i, Y_i)$ 的值是独立分部 (independently distributed) 的
- iid 通常满足于随机的截面数据 (random cross-sectional data). 通常不满足于包含时间维度 (time dimension) 的数据。
  - 时间序列 (time-series) 数据和面板数据 (paneld ata) 包含时间维度

OLS 假设 3：较少异常值

3. 异常值的可能性比较低



- OLS 受异常值影响较大（残差的平方）
- 别的线性回归（如最小绝对偏差 least absolute deviations(LAD)）受异常值影响可能较小

本节课内容

1. 回顾随机对照实验 RCT
2. 假说检验 Hypothesis Testing
3. 回到 RAND 医保实验
4. 可能影响实验有效性的威胁
5. 线性回归 Linear Regression
6. 衡量回归模型的拟合程度
7. 回归模型的假设
8. OLS 的特性

OLS 估计量的抽样分布（sampling distribution）

- 回顾：估计量是随机抽取数据点的函数 (an estimator is a function of a random draw of data points)。
- 比如说，样本平均值 $\bar{Y}$ 取决于 (depends on) 我们抽取的样本
  - 我们知道 $E(\bar{Y}) = \mu$，而且 $V(\bar{Y}) = \frac{\sigma_Y^2}{n}$
- 因此，我们可以证明：

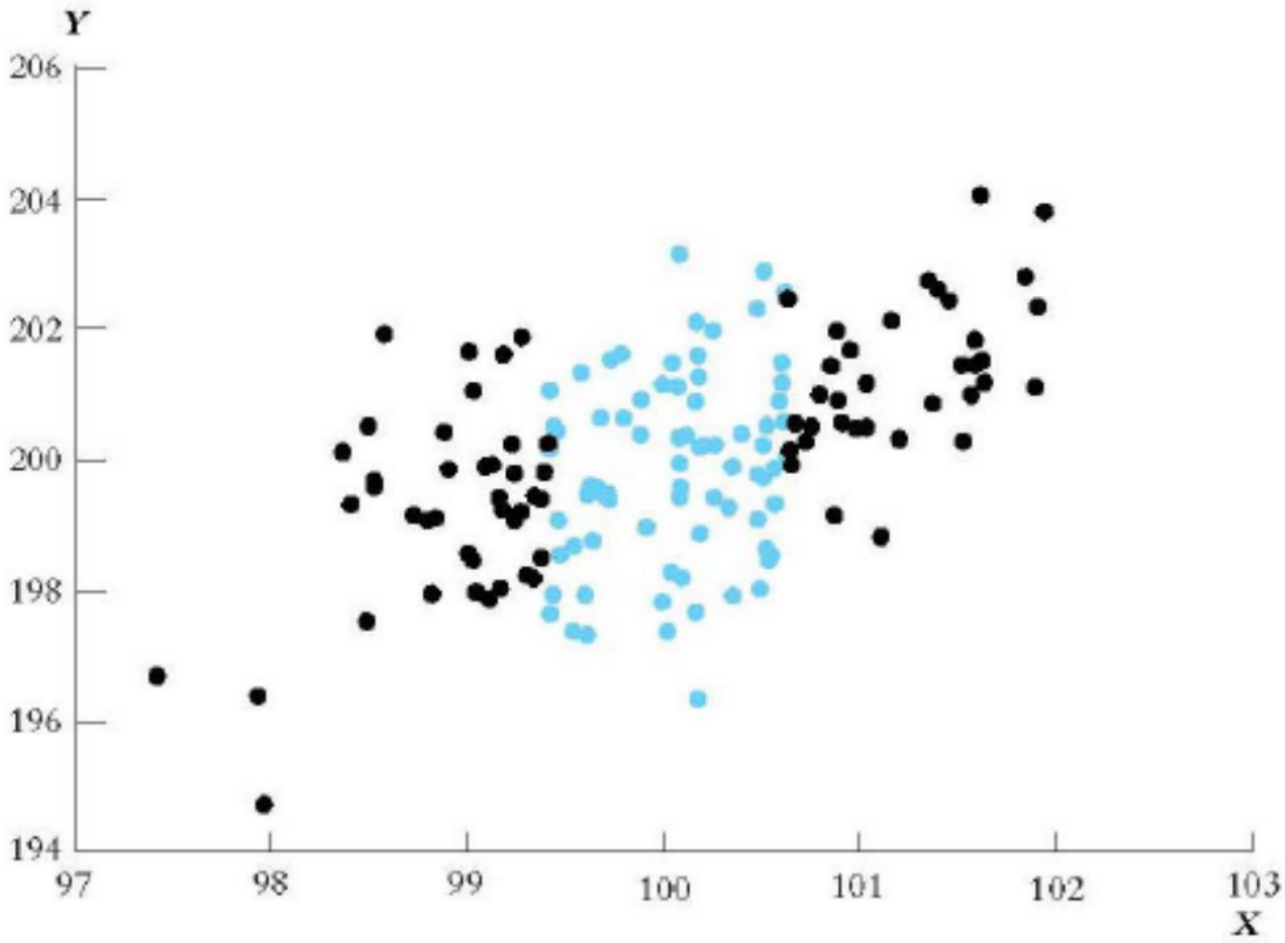
$$E(\hat{\beta}_1) = \beta_1 \text{OLS 估计量是无偏的 (unbiased)}$$

$$var(\hat{\beta}_1) = \frac{var[(X_i - \mu_X)u_i]}{n(\sigma_X^2)^2}$$

$$var(\hat{\beta}_1) \text{和样本大小 (sample size) 成反比}$$

X 方差 (Variance) 大是个好事

- 使用黑色数据点比使用蓝色数据点可以更准确地估计回归线
- The regression line can be more accurately estimated with the black data points than with the blue ones



大样本分部 (Large Sample Distribution)

通过中央极限定理 (CLT) 我们知道当 n 很大时，

$$\hat{\beta}_1 \sim N(\beta_1, \frac{var[(X_i - \mu_X)u_i]}{n(\sigma_X^2)^2})$$

$$\Rightarrow \sigma_X^2 \text{越大}，\ var(\hat{\beta}_1) \text{越小}。$$

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 (\hat{\beta}_1 \text{ 是一致的 (consistent)})$$

$$\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{var(\hat{\beta}_1)}} \sim N(0, 1)$$

假说检验 Hypothesis Testing

- 零假说 Null Hypothesis：

$$H_0 : \beta_1 = \tilde{\beta}_1$$

- 另类假说 Alternetive Hypothesis (双尾 Two-tailed)：

$$H_1 : \beta_1 \neq \tilde{\beta}_1$$

假说检验 Hypothesis Testing

- 零假说 Null Hypothesis：

$$H_0 : \beta_1 = \tilde{\beta}_1$$

- 另类假说 Alternetive Hypothesis (双尾 Two-tailed)：

$$H_1 : \beta_1 \neq \tilde{\beta}_1$$

- 我们通常判断 $\beta_1 = 0$ 或者 $\beta_1 \neq 0$，即 X 和 Y 有没有关联 (related)。

假说检验 Hypothesis Testing

- 零假说 Null Hypothesis：

$$H_0 : \beta_1 = \tilde{\beta}_1$$

- 另类假说 Alternetive Hypothesis (双尾 Two-tailed)：

$$H_1 : \beta_1 \neq \tilde{\beta}_1$$

- 我们通常判断 $\beta_1 = 0$ 或者 $\beta_1 \neq 0$，即 X 和 Y 有没有关联 (related)。
- 但是有时 H_1 也有可能是单尾的，比如：

$$H_1 : \beta_1 > \tilde{\beta}_1 \text{ or } H_1 : \beta_1 < \tilde{\beta}_1$$

假说检验 Hypothesis Testing

- 零假说 Null Hypothesis：

$$H_0 : \beta_1 = \tilde{\beta}_1$$

- 另类假说 Alternetive Hypothesis (双尾 Two-tailed)：

$$H_1 : \beta_1 \neq \tilde{\beta}_1$$

- 我们通常判断 $\beta_1 = 0$ 或者 $\beta_1 \neq 0$，即 X 和 Y 有没有关联 (related)。
- 但是有时 H_1 也有可能是单尾的，比如：

$$H_1 : \beta_1 > \tilde{\beta}_1 \text{ or } H_1 : \beta_1 < \tilde{\beta}_1$$

- 但是需要非常好的理由

t 统计量

- 回顾：t 统计量的通用公式：

$$t = \frac{\text{estimator} - \text{hypothesized value}}{\text{standard error of the estimator}}$$

- 在 OLS 中这代表着

$$t = \frac{\hat{\beta}_1 - \tilde{\beta}_1}{SE(\hat{\beta}_1)}$$

t 统计量

- 回顾：t 统计量的通用公式：

$$t = \frac{\text{estimator} - \text{hypothesized value}}{\text{standard error of the estimator}}$$

- 在 OLS 中这代表着

$$t = \frac{\hat{\beta}_1 - \tilde{\beta}_1}{SE(\hat{\beta}_1)}$$

- 使用 t 统计量我们可以计算 p 值，建立置信区间 (confidence intervals)，进行统计测试。。。就像我们之前做的那样

二元 (Binary) 解释变量

- 有时候 X 是二元 (Binary) 的：
  - $X = 1$ 代表女生，$X = 0$ 代表男生
  - $X = 1$ 代表实验组，$X = 0$ 代表对照组
  - $X = 1$ 代表是非洲，$X = 0$ 代表不是非洲

二元 (Binary) 解释变量

- 有时候 X 是二元 (Binary) 的：
  - $X = 1$ 代表女生，$X = 0$ 代表男生
  - $X = 1$ 代表实验组，$X = 0$ 代表对照组
  - $X = 1$ 代表是非洲，$X = 0$ 代表不是非洲
- 这种二元的变量又叫虚拟变量 (dummy variables)

二元 (Binary) 解释变量

- 有时候 X 是二元 (Binary) 的：
  - $X = 1$ 代表女生，$X = 0$ 代表男生
  - $X = 1$ 代表实验组，$X = 0$ 代表对照组
  - $X = 1$ 代表是非洲，$X = 0$ 代表不是非洲
- 这种二元的变量又叫虚拟变量 (dummy variables)
- 之前我们说 $\hat{\beta}_1 =$ 斜率*slope*，那么当 X 是二元的时候，又代表什么呢？

解读 X 为二元变量时的系数 (coefficient)

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

解读 X 为二元变量时的系数 (coefficient)

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- 当 $X = 0$ 时，

$$Y_i = \beta_0 + u_i \Rightarrow E(Y_i|X = 0) = \beta_0$$

解读 X 为二元变量时的系数 (coefficient)

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- 当 $X = 0$ 时，
$$Y_i = \beta_0 + u_i \Rightarrow E(Y_i|X = 0) = \beta_0$$
- 当 $X = 1$ 时，
$$Y_i = \beta_0 + \beta_1 X_i + u_i \Rightarrow E(Y_i|X = 1) = \beta_0 + \beta_1$$

解读 X 为二元变量时的系数 (coefficient)

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- 当 $X = 0$ 时，
$$Y_i = \beta_0 + u_i \Rightarrow E(Y_i|X = 0) = \beta_0$$
- 当 $X = 1$ 时，
$$Y_i = \beta_0 + \beta_1 X_i + u_i \Rightarrow E(Y_i|X = 1) = \beta_0 + \beta_1$$
- $\Rightarrow$

$$\beta_1 = E(Y_i|X = 1) - E(Y_i|X = 0) = \text{difference in group means}$$

解读 X 为二元变量时的系数 (coefficient)

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- 当 $X = 0$ 时，
$$Y_i = \beta_0 + u_i \Rightarrow E(Y_i|X = 0) = \beta_0$$
- 当 $X = 1$ 时，
$$Y_i = \beta_0 + \beta_1 X_i + u_i \Rightarrow E(Y_i|X = 1) = \beta_0 + \beta_1$$
- $\Rightarrow$

$$\beta_1 = E(Y_i|X = 1) - E(Y_i|X = 0) = \text{difference in group means}$$

- 这也是为什么我们用两个方法得到了同样的结果

解读 X 为二元变量时的系数 (coefficient)

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

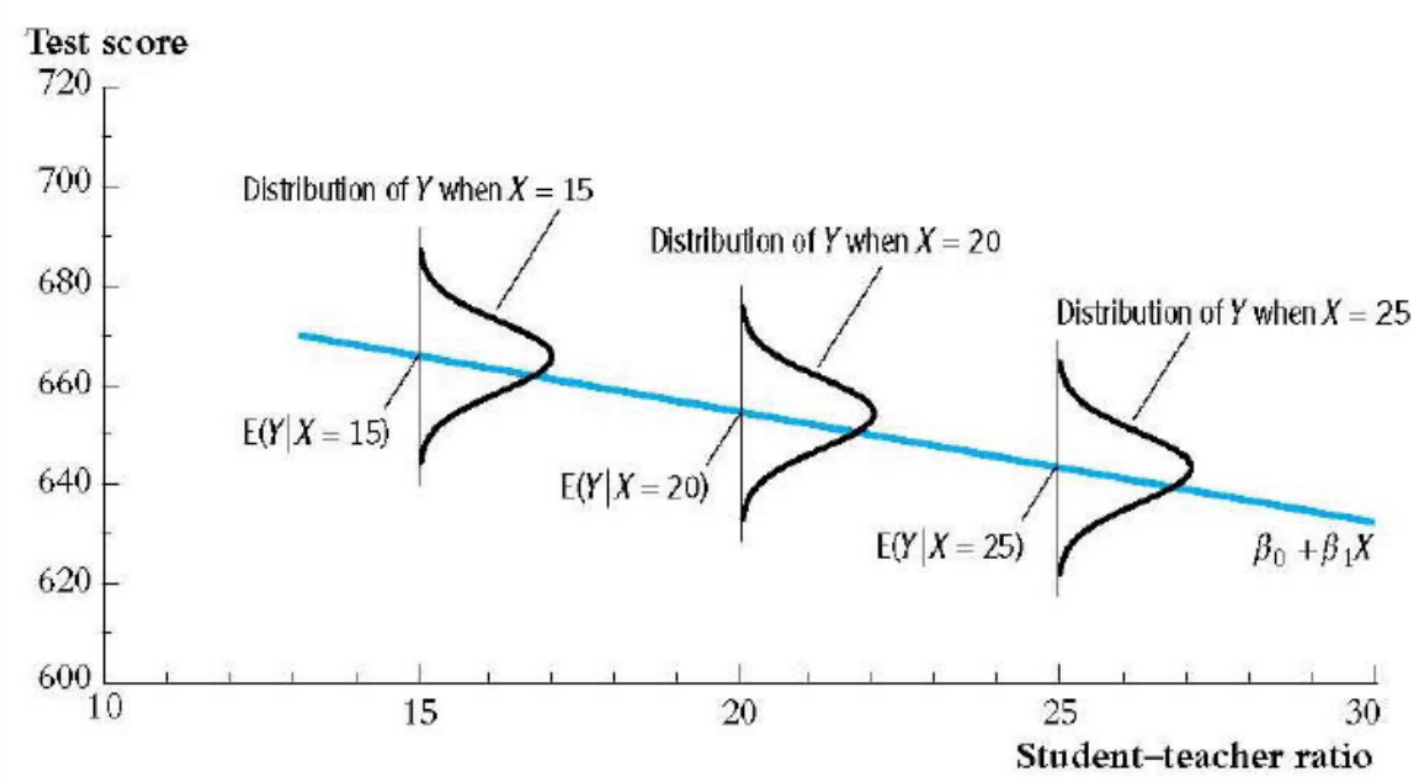
- 当 $X = 0$ 时，
$$Y_i = \beta_0 + u_i \Rightarrow E(Y_i|X = 0) = \beta_0$$
- 当 $X = 1$ 时，
$$Y_i = \beta_0 + \beta_1 X_i + u_i \Rightarrow E(Y_i|X = 1) = \beta_0 + \beta_1$$
- $\Rightarrow$

$$\beta_1 = E(Y_i|X = 1) - E(Y_i|X = 0) = \text{difference in group means}$$

- 这也是为什么我们用两个方法得到了同样的结果
  - $SE(\hat{\beta}_1)$ 的解读是一样的，但是计算方法不同!

同方差性 Homoskedasticity

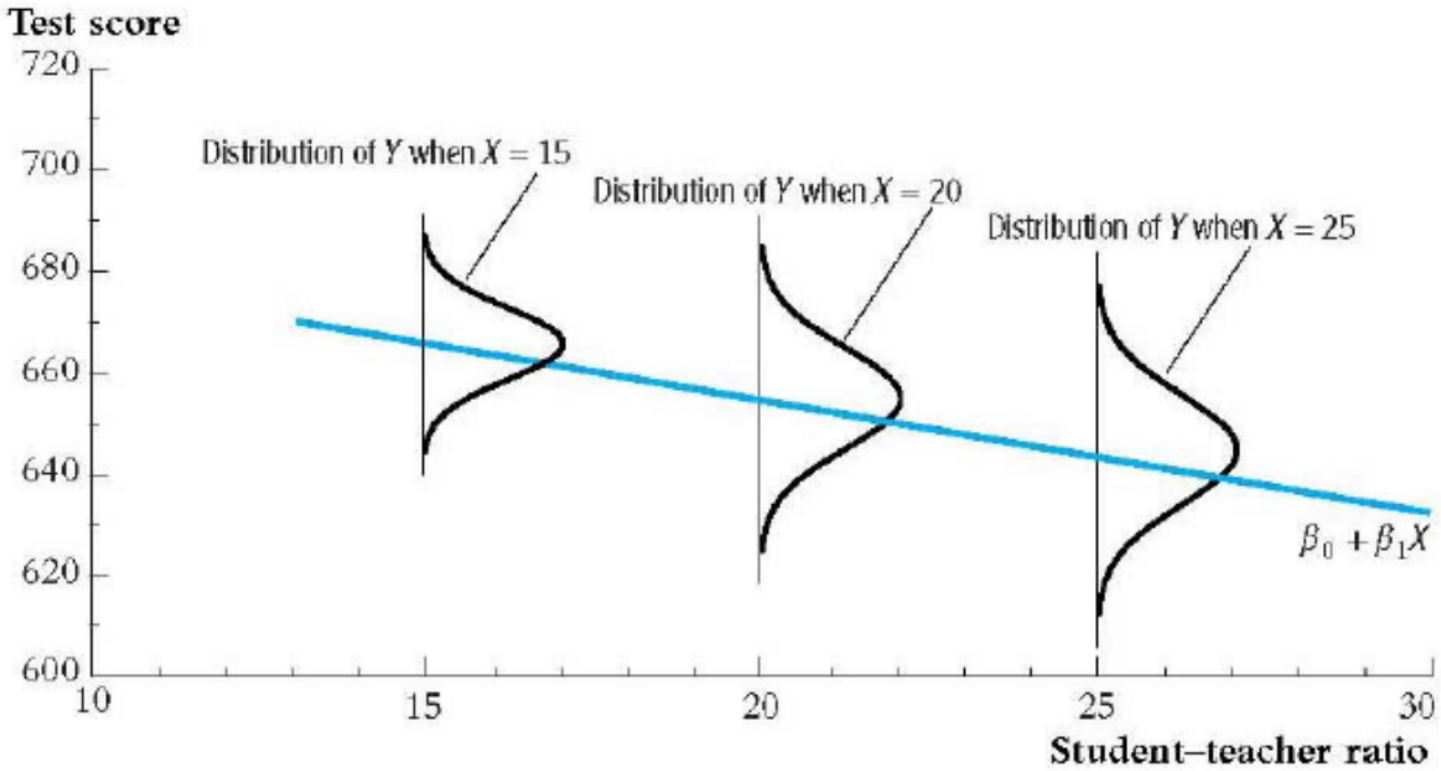
如果 $var(u|X = x)$不变(Constant)，即残差关于 X 的条件方差不随着 X 的变化而变化 (the variance of the conditional distribution of u given X does not depend on X)，那么 u 可以说是同方差的 (Homoskedastic)



- 对于我们 R 里的两元 X 呢？

异方差性 Heteroskedasticity

如果 $var(u|X = x)$变化(not constant)，即残差关于 X 的条件方差随着 X 的变化而变化 (the variance of the conditional distribution of u given X changes with X)，那么 u 可以说是异方差的 (Heteroskedastic)



两种标准误 Two types of SEs

- 无论用哪种 se，$\hat{\beta}_1$ 都是无偏 (unbiased) 且一致 (consistent) 的。
- 但是计算 $SE(hat\beta_1)$ 的公式可能不同
- (异方差性) 稳健标准误 heteroskedasticity-robust SE

$$\sqrt{\frac{var[(X_i - \mu_X)u_i]}{n(\sigma_X^2)^2}}$$

- 在同方差性的假设下，简化为仅同方差性 (homoskedasticity-only) 标准误：

$$\sqrt{\frac{\sigma_u^2}{n\sigma_X^2}}$$

实操

- 如果 errors 误差是同方差的 (homoskedastic)→ 你可以用简化的 se
  - R/Stata 默认的 se 都是这个
  - 使用 Breusch-Pagan test 测试同方差性.
- 如果 errors 是异方差的，而你用了同方差 se→ 计算的 SE 是错误的，统计检测结果也是不牢靠的
- 但是如果 se 是同方差的，你仍然可以使用稳健 se
  - 只要 n 足够大
- ⇒ 基本上都使用稳健标准误
  - Stata: reg ..., robust
  - R: lm_robust() or feols(,vcov="HC1")

References I

Brook, R. H., Keeler, E. B., Lohr, K. N., Newhouse, J. P., Ware, J. E., Rogers, W. H., Davies, A. R., Sherbourne, C. D., Goldberg, G. A., Camp, P., Kamberg, C., Leibowitz, A., Keesey, J., & Reboussin, D. (2006). *The health insurance experiment: A classic rand study speaks to the current health care reform debate*. RAND Corporation.

Stock, J., & Watson, M. (2014). *Econometrics, Update PDF Ebook, Global Edtion*. Pearson Education, Limited. Retrieved July 19, 2023, from http://ebookcentral.proquest.com/lib/londonschoolecons/detail.action?docID=5174962