计量经济分析

第三章　一元线性回归

徐秋华，xuqh@swufe.edu.cn

西南财经大学，金融学院

大纲

1. 一元线性回归模型
2. 普通最小二乘估计
3. 回归系数的解释
4. 拟合优度
5. 回归系数估计量的小样本性质
6. Monte Carlo 模拟
7. 虚拟解释变量
8. 潜在结果与因果推断

一元线性回归模型

一元线性回归

- 一元线性回归模型通过建立两个变量之间的“线性”关系研究其中一个变量对另一个变量的影响。
- 例如，受教育年限对对数工资的影响：

$$\log wage_i = \alpha + \beta \times edu_i + \varepsilon_i, \quad i = 1, \cdots, n.$$

- 这里关注的两个变量分别是受教育年限与工资，因为上式中的工资取了对数，所以 *edu* 和 *wage* 之间并非是线性关系。
- 上述模型仍然被称为是一元线性回归模型，这是因为一元线性回归模型中的“线性”是指回归方程右边相对于参数 α 和 β 呈线性关系。

更一般地，一元线性回归模型可以表示为：

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \cdots, n, \tag{1}$$

其中，

- Y_i 被称为因变量（dependent variable）、被解释变量（explained variable）或响应变量（response variable），在线性回归模型中，Y_i 通常是连续变量；
- X_i 被称为自变量（independent variable）、解释变量（explanatory variable）或回归元（regressor），X_i 既可以是连续变量，也可以是离散变量；
- ε_i 被称为误差项（error term）或扰动项（disturbance），它包含了除 X_i 之外所有其他可能影响 Y_i 的因素；
- β_0 和 β_1 被称为回归系数（coefficients）或模型参数（parameters），其中，β_0 被称为常数项（constant）或截距（intercept），β_1 被称为斜率（slope）；
- 下标 i 表示样本中的第 i 个观测（observation）。对于横截面数据，i 表示第 i 个横截面个体；对于时间序列数据，i 表示第 i 个时刻。下标 i 的取值范围是从 1 到 n 的正整数，n 通常被称为样本量（sample size）。

一元线性回归模型

- 计量经济学中的模型是指对解释变量和被解释变量的联合分布施加的一组假设。
- 式 (1) 要求回归方程相对于回归系数是线性的。
- 除此之外，为了推导回归系数估计量的理论性质，还需要施加一些其他的假设，这其中就包括对误差项 ε_i 与 X_i 之间关系的设定。

具体地，一元线性回归模型需要满足的假设由下面的定义给出：

一元线性回归模型

假设 3.1（线性假设）：

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad (i = 1, 2, \cdots, n).$$

一元线性回归模型

假设 3.2（随机样本）：$\{(X_i, Y_i) : i = 1, 2, \cdots, n\}$ 是从总体中抽取的一个样本量为 n 的随机样本。

假设 3.3（解释变量存在样本变异）：解释变量 X 的样本观测值 $\{X_i : i = 1, 2, \cdots, n\}$ 不取某一固定的常数。

假设 3.4（解释变量外生）：

$$E(\varepsilon_i \mid X_i) = 0, \quad (i = 1, 2, \cdots, n). \tag{2}$$

假设 3.5（条件同方差）：

$$E(\varepsilon_i^2 \mid X_i) = \sigma^2 > 0, \quad (i = 1, 2, \cdots, n). \tag{3}$$

关于假设的说明

- 假设 3.2 要求不同个体的观测值是从总体中随机抽取的，许多横截面数据满足此要求，但是它不适用于时间序列数据。
- 随机样本假设并非是推导回归系数估计量统计性质需要施加的最弱假设。后面介绍多元线性回归时会放松随机样本假设。
- 假设 3.3 保证了回归系数估计量是可计算的。直观上，一元线性回归是利用解释变量和被解释变量的样本观测值去寻找一条最适合描述两者之间关系的直线，β_0 和 β_1 分别是这条直线的截距和斜率。

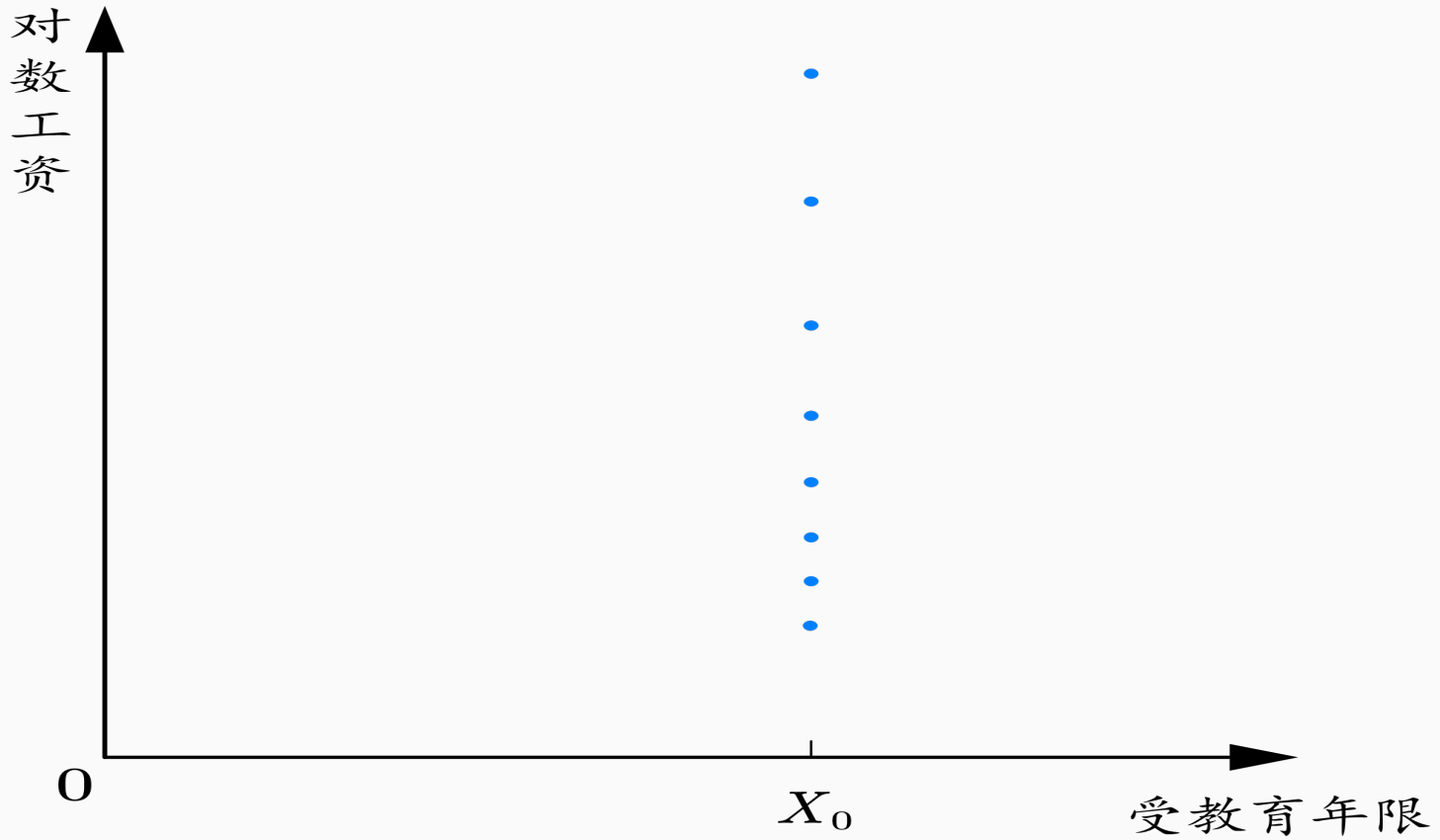


图 1： 解释变量不存在样本变异时无法估计回归系数

如图 1 所示，若解释变量不存在样本变异，则所有样本观测值分布在一条垂直于横轴的直线上。因此，$X = X_0$ 是描述 X 与 Y 之间关系的最适合直线，但此直线不存在截距并且斜率为无穷大。

假设 3.4

假设 3.4 要求 ε_i 均值独立于 X_i，即 $E(\varepsilon_i \mid X_i)$ 与 X_i 无关。同时，此假设进一步要求 $E(\varepsilon_i \mid X_i) = 0$。由 SLIE 可以推出 $E(\varepsilon_i) = 0$。又因为两变量均值独立蕴涵了两变量不相关，于是，$\mathrm{Cov}(X_i, \varepsilon_i) = 0$。在假设 3.4 下，容易验证：

$$E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i, \tag{4}$$

因此，一元线性回归模型实际上将被解释变量的条件期望设置为线性的形式。因此，(β_0, β_1) 为 Y_i 对常数项和 X_i 做投影的线性投影系数。

式 (4) 通常被称为总体回归线 (population regression line, PRL)。图 2 展示了样本观测值（图中的实心圆点）、总体回归线和误差项彼此之间的关系。

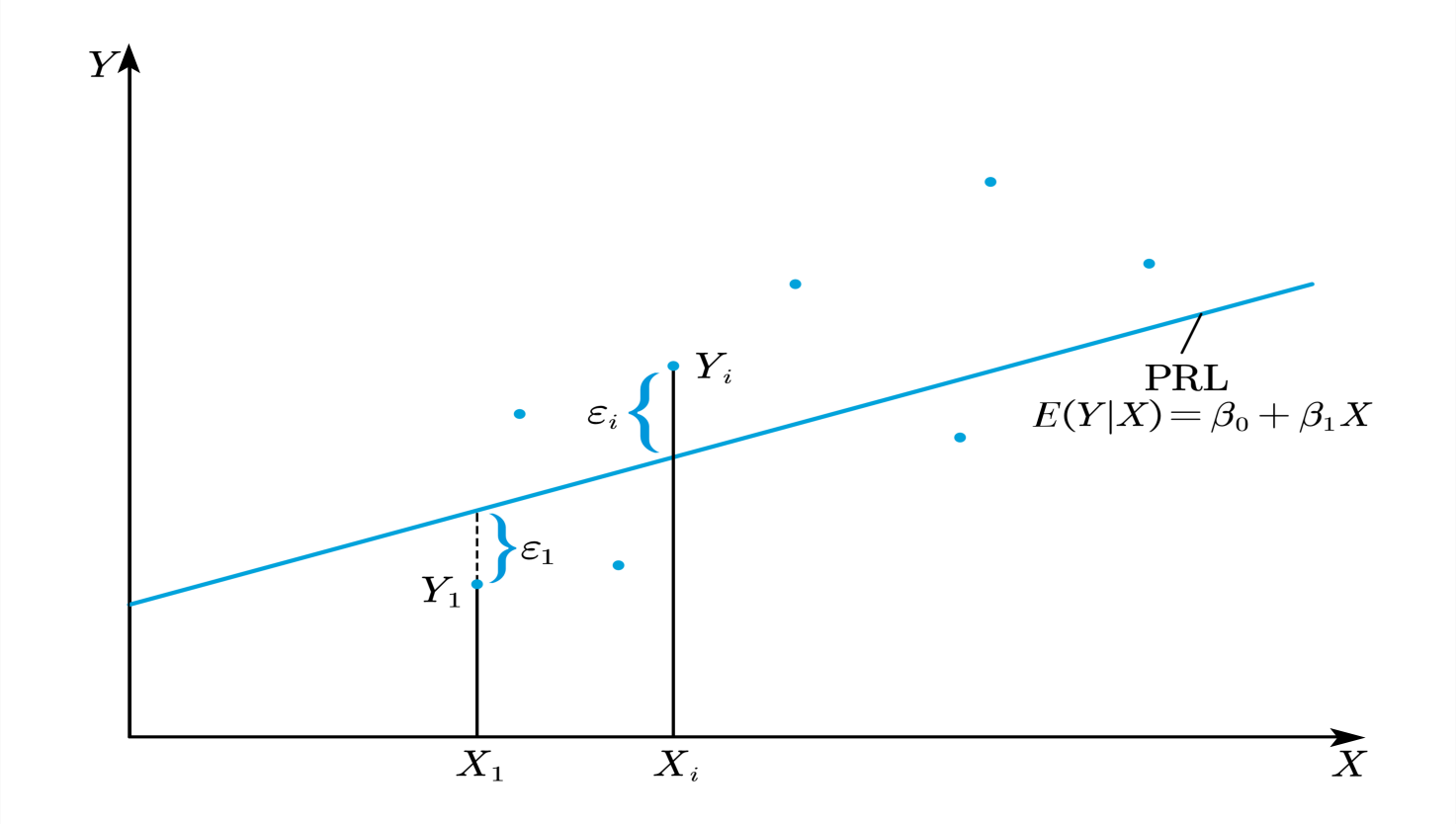


图 2: 样本观测值、总体回归线和误差项

- 假设 3.4 在推断 X 与 Y 之间的因果关系过程中发挥了重要的作用。
- 为了理解这一点，不妨假设 X 为二值变量，只能取值 0 或者 1。
- 考虑大学教育 (X) 对对数工资 (Y) 的影响，$X = 1$ 表示受过大学教育，$X = 0$ 表示没有受过大学教育。
- 除了大学教育之外，许多其他因素也可以影响工资，例如：个人能力、工作经验、父母受教育程度和父母收入等。所有这些影响因素都包含在误差项 ε 之中。
- 理想的随机化实验可以随机地分配大学教育，然后观察其对未来工资的影响。此时，X 的分配，即哪些个体 $X = 1$ 以及哪些个体 $X = 0$，完全独立于个人能力等其他因素。
- 而对于非实验数据，假设 3.4 意味着 X 的分配与个人能力等其他因素不相关，因此，这是对随机化实验的一种近似。
- 由式 (4) 可知，线性回归是为条件期望建模，因此，X 不需要独立于 ε 中包含的其他因素，而只要求 ε 的期望与 X 无关，即 $E(\varepsilon \mid X)$ 为不依赖于 X 的常数。

注意到假设 3.4 要求这一常数的取值为 0。事实上，在包含常数项的回归模型中，假设 $E(\varepsilon \mid X)$ 为 0 并不是一个过于强的假设。这是因为若存在非零常数 c 使得 $E(\varepsilon_i \mid X_i) = c$，总可以通过重新定义回归中的常数项和误差项从而使假设 3.4 成立。具体地，

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = (\beta_0 + c) + \beta_1 X_i + (\varepsilon_i - c) = \beta_0^* + \beta_1 X_i + \varepsilon_i^*,$$

其中，$\beta_0^* = \beta_0 + c$，$\varepsilon_i^* = \varepsilon_i - c$。重新定义的误差项 ε_i^* 满足假设 3.4。因为回归分析关注的是参数 β_1，所以这里的模型变换对回归分析并无影响。

假设 3.5

假设 3.5 常被称为条件同方差（conditional homoskedasticity）假设，这是因为根据假设 3.4，ε_i 的条件方差可以表示为：

$$\mathrm{Var}(\varepsilon_i \mid X_i) = E(\varepsilon_i^2 \mid X_i) - E(\varepsilon_i \mid X_i)^2 = E(\varepsilon_i^2 \mid X_i).$$

此外，由 SLIE 可知误差项的无条件方差也是 σ^2。事实上，条件同方差是一个很强的假设，在很多情况下并不满足。

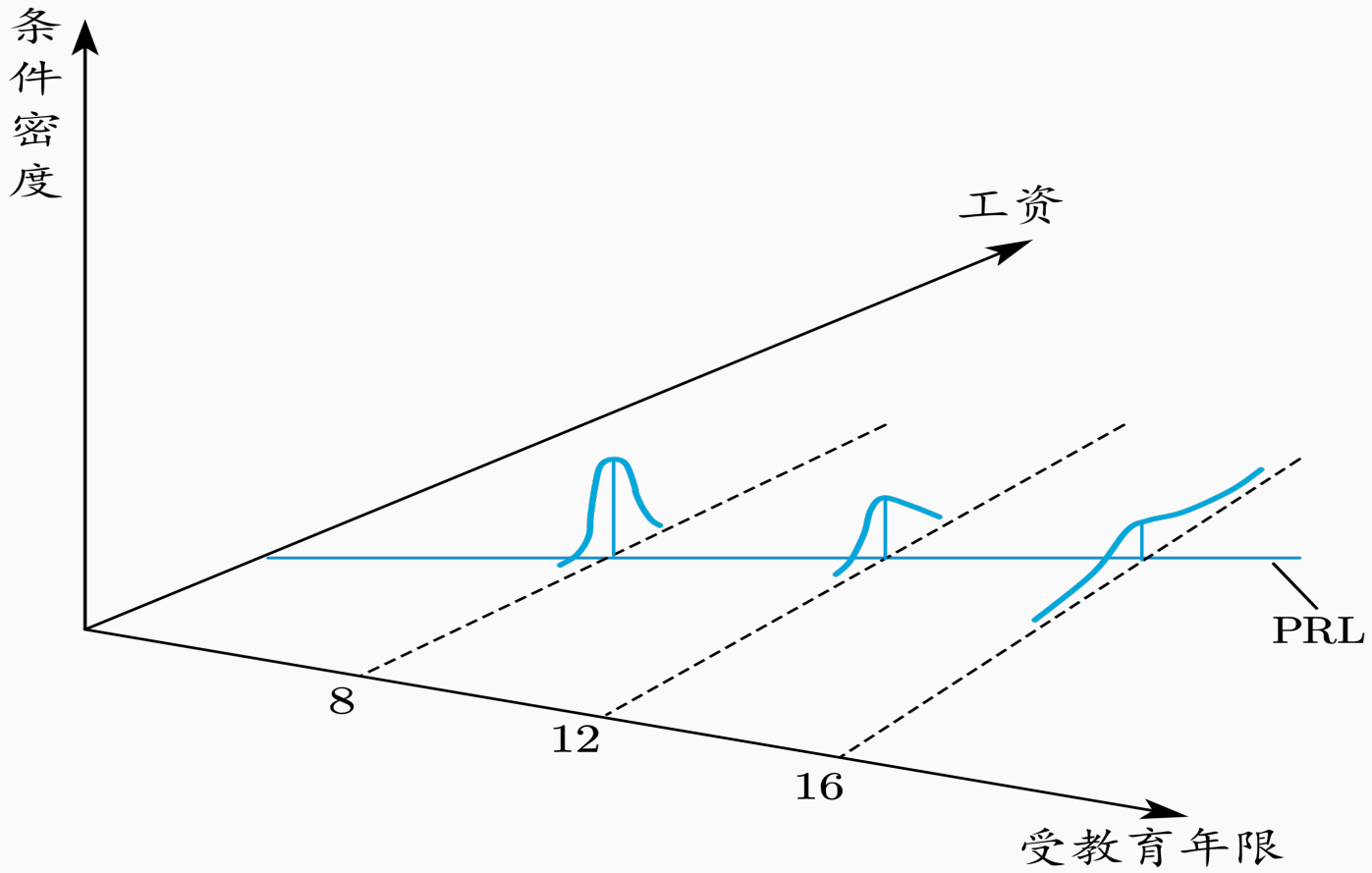


图 3: 条件异方差的示例

- 更一般地，$E(\varepsilon_i^2 \mid X_i)$ 通常是 X_i 的函数，这种误差项条件方差与解释变量取值有关的情况在计量经济学中被称为条件异方差 (conditional heteroskedasticity)。
- 异方差是实证分析数据表现出的常态情况，同方差才是特殊情况。
- 例如，总体回归线是对条件期望的最优线性近似，即使数据真的是同方差，但条件期望是非线性的，回归模型的误差项也将是异方差的。具体地，

$$\begin{aligned} E(\varepsilon_i^2 \mid X_i) &= E[(Y_i - \beta_0 - \beta_1 X_i)^2 \mid X_i] \\ &= E\left\{[Y_i - E(Y_i \mid X_i) + E(Y_i \mid X_i) - \beta_0 - \beta_1 X_i]^2 \mid X_i\right\} \\ &= \mathrm{Var}(Y_i \mid X_i) + [E(Y_i \mid X_i) - \beta_0 - \beta_1 X_i]^2 . \end{aligned} \tag{5}$$

即使 $\mathrm{Var}(Y_i \mid X_i) = \sigma^2$，如果条件期望是非线性的，第二项将不等于 0，而是 X_i 的函数。

此外，前文提到回归模型的被解释变量通常为连续变量。即使条件期望函数是线性的，但同时被解释变量是离散的，则误差项也会表现出条件异方差。例如，假设 Y 为虚拟变量，则有：

$$
\begin{aligned}
& E(\varepsilon_i^2 \mid X_i) \\
= \; & E\left\{[Y_i - E(Y_i \mid X_i)]^2 \mid X_i\right\} = E\left\{[Y_i - \Pr(Y_i = 1 \mid X_i)]^2 \mid X_i\right\} \\
= \; & [1 - \Pr(Y_i = 1 \mid X_i)]^2 \Pr(Y_i = 1 \mid X_i) \\
& + \Pr(Y_i = 1 \mid X_i)^2 [1 - \Pr(Y_i = 1 \mid X_i)] \\
= \; & \Pr(Y_i = 1 \mid X_i)[1 - \Pr(Y_i = 1 \mid X_i)],
\end{aligned}
$$

其中，第二个等号利用了虚拟变量的期望等于它取 1 的概率；第三个等号利用了离散随机变量条件期望的定义。

传统计量教材通常假设条件同方差，而将条件异方差当成是一种问题进行处理。而在实际的实证分析当中，研究者大多从一开始就假设异方差的存在，进而使用对异方差稳健（robustness）的计量方法进行统计推断。

普通最小二乘估计

在假设 3.4 下，总体回归线和线性投影 $\mathscr{P}\{Y \mid (1, X)\}$ 等价，因此，β_0 和 β_1 是线性投影系数。根据线性投影的定义，β_0 和 β_1 是如下最优化问题的最优解：

$$\min_{b_0, b_1} E(Y - b_0 - b_1 X)^2.$$

令 $\{(X_i, Y_i) : i = 1, 2, \cdots, n\}$ 是样本量为 n 的一个样本。为了估计回归系数，可以利用样本均值近似总体期望得到如下最优化问题：

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

由于样本量是固定的，上式中的 $1/n$ 可以省略。因此，β_0 和 β_1 的估计量可以通过解如下最优化问题求得：

$$\min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2. \tag{6}$$

上式中的目标函数常被称为残差平方和（residual sum of squares），记为 $RSS(\boldsymbol{b})$。通过最小化 $RSS(\boldsymbol{b})$ 得到回归系数估计量的估计方法被称为普通最小二乘（ordinary least squares, OLS）估计。

根据最优化的一阶条件，最小二乘估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 满足：

$$\frac{\partial}{\partial b_0} RSS(\boldsymbol{b}) = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0, \tag{7}$$

$$\frac{\partial}{\partial b_1} RSS(\boldsymbol{b}) = -2\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0. \tag{8}$$

由式 (7) 可得：

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \tag{9}$$

其中，$\bar{Y}$ 和 $\bar{X}$ 分别表示被解释变量和解释变量的样本均值，即 $\bar{Y} = (1/n)\sum_{i=1}^{n} Y_i$，$\bar{X} = (1/n)\sum_{i=1}^{n} X_i$。由式 (8) 可得：

$$\frac{1}{n}\sum_i X_i Y_i - \hat{\beta}_0 \bar{X} - \hat{\beta}_1 \frac{1}{n}\sum_i X_i^2 = 0.$$

将式 (9) 代入上式得到：

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{\frac{1}{n-1}\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n-1}\sum_i (X_i - \bar{X})^2} \equiv \frac{s_{XY}}{s_X^2}, \tag{10}$$

其中，s_{XY} 表示 X_i 和 Y_i 的样本协方差；s_X^2 表示 X_i 的样本方差。假设 3.3 保证了 s_X^2 大于 0。

回忆线性投影系数的结论：当 $\boldsymbol{X} = (1, X_1)'$，$Y$ 可以表示为：

$$Y = \alpha + \beta_1 X_1 + e,$$

其中，α 和 β_1 分别是常数项 1 和 X_1 对应的线性投影系数。容易证明：

$$\beta_1 = \frac{\mathrm{Cov}(Y, X_1)}{\mathrm{Var}(X_1)}, \quad \alpha = E(Y) - \beta_1 E(X_1). \tag{11}$$

对比式 (10) 不难发现，普通最小二乘估计实际上是在估计线性投影系数。

事实上，在假设 3.4 下，误差项 ε_i 即为线性投影误差，则有：$E(X_i\varepsilon_i) = 0$ 和 $E(\varepsilon_i) = 0$。根据矩估计的思想，期望可以利用样本均值代替，则线性投影系数估计量（也就是 OLS 估计量）应该满足：

$$\begin{cases} \dfrac{1}{n}\sum_i X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0, \\ \dfrac{1}{n}\sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0. \end{cases} \tag{12}$$

这恰好就是式 (7) 和 (8) 的一阶条件。若将 1 看成是特殊的解释变量，常数项 β_0 为其系数，并记 $\boldsymbol{X}_i = (1, X_i)'$，则 ε_i 满足的两个矩条件可以合并为 $E(\boldsymbol{X}_i\varepsilon_i) = \boldsymbol{0}$，称为 ε_i 与 $\boldsymbol{X}_i$ 正交。OLS 估计量可以由解最优化问题 (6) 求得，也可以直接由正交条件的样本形式 (12) 解出，这两种方式是彼此等价的。

定义回归的拟合值（fitted value）为 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$，残差（residual）为 $e_i = Y_i - \hat{Y}_i$。式 (12) 可以表示为：

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i e_i = \mathbf{0},$$

或，等价地，

$$\sum_{i=1}^{n} \boldsymbol{X}_i e_i = \mathbf{0}. \tag{13}$$

此式被称为正则方程（normal equation），它表明：在样本中，残差与解释变量正交。因此，OLS 估计的核心是正交条件：在总体中，误差项和解释变量在 $E(\boldsymbol{X}_i \varepsilon_i) = \mathbf{0}$ 的意义上彼此正交，其样本形式 $\sum_{i=1}^{n} \boldsymbol{X}_i e_i = \mathbf{0}$ 说明 OLS 估计量满足残差与解释变量正交。

$\hat{\beta}_0 + \hat{\beta}_1 X_i$ 通常被称为样本回归线（sample regression line, SRL），它是对总体回归线的估计。图 4 展示了总体回归线、样本回归线、拟合值、误差和残差之间的关系。

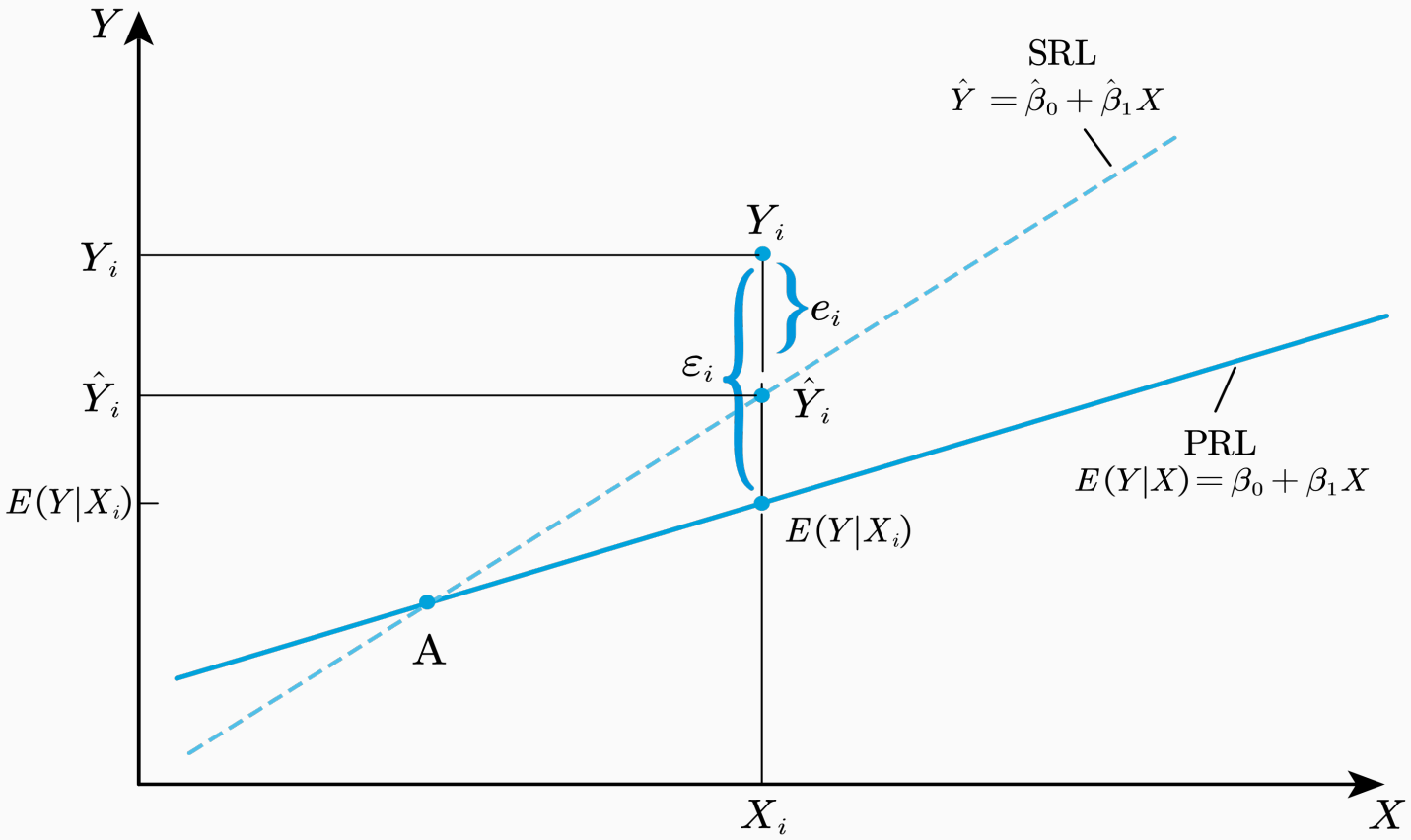


图 4: 样本回归线和总体回归线

回归系数的解释

由式 (1)，Y 的变化满足：

$$\Delta Y = \beta_1 \Delta X + \Delta \varepsilon,$$

其中，Δ 表示 “变化”。若 $\Delta \varepsilon = 0$，则 $\beta_1 = \Delta Y / \Delta X$。因此，$\beta_1$ 衡量了固定其他影响 Y 的因素不变之后，X 变化一单位引起的 Y 的变化。由于 β_1 未知，利用样本数据对其进行 OLS 估计可以得到样本回归线：

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

由上式可知，$\hat{\beta}_0$ 是把 $X = 0$ 代入到样本回归线之后得到的 Y 的预测值。在很多实证分析中，$X = 0$ 是没有实际意义的（例如：X 表示年龄）。因此，实证分析中通常对 $\hat{\beta}_0$ 不感兴趣。同时，由上式还可以得到：

$$\hat{\beta}_1 = \Delta \hat{Y} / \Delta X.$$

这表明 $\hat{\beta}_1$ 衡量了（固定其他因素不变）X 变化一单位引起的 $\hat{Y}$ 的变化。

考虑 [Wooldridge, 2019] 提供的数据集 WAGE1.dta，该数据集收集了 526 人的受教育年限（educ）和小时工资（wage）等变量的数据。利用如下 Stata 代码将 wage 对 educ 进行回归：

```
. use WAGE1.dta, clear
. reg wage educ
```

回归结果可以总结为如下方程：

$$\widehat{wage} = -0.90 + 0.54 \times educ.$$

于是，增加一年的教育，时薪预计会增加 0.54 美元。

当线性回归模型中包含有可观测变量的非线性变换时，初学者要特别留意如何解释回归系数。下面讨论几种常见的情况：

- 对数-线性模型：此类模型具有如下形式：

$$\log(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

注意到 $\log(Y + \Delta Y) - \log(Y) \approx \Delta Y/Y$，所以 $\Delta Y/Y \approx \beta_1 \Delta X$。因此，若 X 变化一单位，即 $\Delta X = 1$，则 $\Delta Y/Y$ 变化 β_1。等价地，X 变化一单位引起 Y 变化 $(100 \times \beta_1)\%$。例如，若将对数工资 (lwage) 对 educ 进行回归得到如下 SRL：

$$\widehat{lwage} = 0.58 + 0.08 \times educ.$$

上述回归结果可以解释为：每增加一年教育，时薪预计会增加 8%。在对数-线性模型中，$100 \cdot \beta_1$ 通常被称为 Y 关于 X 的“半弹性 (semi-elasticity)”。

- **对数-对数模型**：此类模型具有如下形式：

$$\log(Y_i) = \beta_0 + \beta_1 \log(X_i) + \varepsilon_i.$$

类似地，若 $\Delta\varepsilon = 0$，则有：

$$\log(Y + \Delta Y) - \log(Y) = \beta_1[\log(X + \Delta X) - \log(X)].$$

两端同时取近似得到：$\Delta Y/Y \approx \beta_1 \Delta X/X$。于是，$\beta_1$ 可近似表示为：

$$\beta_1 \approx \frac{\Delta Y/Y}{\Delta X/X} = \frac{100 \times \Delta Y/Y}{100 \times \Delta X/X} = \frac{Y \text{ 的百分比变化}}{X \text{ 的百分比变化}}.$$

上式表明 β_1 可以解释为 Y 相对于 X 的弹性，即 X 变化 1% 引起的 Y 的百分比变化。

考虑如下例子：

$$\widehat{\log(salary)} = 3.824 + 0.238 \times \log(sale),$$

其中，$\log(salary)$ 表示公司高管薪酬对数，$\log(sale)$ 表示公司销售收入对数。上述回归结果可以解释为：销售收入增加 1% 预计会使高管薪酬增加 0.238%。

- **线性-对数模型**：此类模型具有如下形式：

$$Y_i = \beta_0 + \beta_1 \log(X_i) + \varepsilon_i.$$

类似地，若 $\Delta\varepsilon = 0$，则有：

$$\Delta Y = \beta_1[\log(X + \Delta X) - \log(X)] \approx \beta_1(\Delta X/X).$$

因此，X 变化 1% 引起的 Y 的变化是 $0.01\beta_1$。

考虑如下例子：

$$\widehat{TestScore} = 62.7 + 48.37 \times \log(Income),$$

其中，*TestScore* 为学生学习成绩，*Income* 为家庭年人均收入。上述回归结果可以解释为：家庭收入增加 1% 预计使学生成绩增加 0.4837 分。

拟合优度

所谓拟合优度是指样本回归线对样本观测值的拟合程度。在计量经济学中，通常用（中心化）R^2 作为衡量拟合优度的指标。

（中心化）R^2

（中心化）R^2，也称为可决系数（coefficient of determination），衡量的是除常数项之外的解释变量对 Y 的解释能力，其定义如下：

$$R^2 \equiv 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}. \tag{14}$$

为了更好地理解 R^2 的定义，需要借助回归的平方和分解公式（如何证明）：如果线性回归中含有常数项，则有：

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2. \tag{15}$$

在式 (15) 中，$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ 称为总平方和 (total sum of squares, TSS)，$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ 称为回归平方和 (explained sum of squares, ESS)。因此，式 (15) 的结论也可以表示为：

$$\text{TSS} = \text{ESS} + \text{RSS},$$

其中，TSS 衡量了被解释变量的整体样本变异 (sample variation) 程度，ESS 衡量了可由 X 解释的 Y 的样本变异程度。根据式 (13)，解释变量与残差正交，于是，RSS 是 Y 的样本变异中无法由 X 解释的部分。由式 (15)，R^2 可以表示成：

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}. \tag{16}$$

根据上式，R^2 是由 X 解释了的 Y 的样本变异占其总样本变异的比例。

- 由式 (15) 易知，$0 \leqslant R^2 \leqslant 1$
- R^2 是 Y_i 和 $\hat{Y}_i$ 之间样本相关系数的平方
- 对于一元线性回归来说，R^2 也是 Y_i 和 X_i 之间样本相关系数的平方
- 若 $R^2 = 1$，由式 (14)，残差平方和 RSS 等于 0。因此，对任意的 $i = 1, \cdots, n$，$e_i = Y_i - \hat{Y}_i = 0$，此时，所有的观测值 (X_i, Y_i) 都落在 SRL 上
- 若 $R^2 = 0$，则 ESS 为 0，对任意的 $i = 1, \cdots, n$，$\hat{Y}_i = \bar{Y}$，此时，SRL 与 X 轴平行，$\hat{\beta}_1 = 0$，被解释变量的样本变异无法由 X 解释

- 虽然拟合优度指标衡量了解释变量的拟合能力，但是，若回归分析的目的是为了揭示解释变量和被解释变量之间的因果关系，则不用过度关注拟合优度。

- 具有较低 R^2 的线性回归模型的系数估计量依然可能是解释变量对被解释变量因果效应的合理估计。

- 关于 R^2 的更多讨论可以阅读一篇知乎上的文章“为什么计量经济学家不看 R-square”，网址为：https://zhuanlan.zhihu.com/p/19931167。

回归系数估计量的小样本性质

- 讨论小样本性质之前，先要明确的是 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 都应该被视为随机变量
- 当固定样本 $\{(X_i, Y_i) : i = 1, 2, \cdots, n\}$ 之后，根据式 (9) 和 (10) 可以计算 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的具体取值
- 这里要区分“估计值（estimate）”和“估计量（estimator）”

假设 $\{(X_i, Y_i) : i = 1, 2, \cdots, n\}$ 是从总体中抽取的随机样本，以 β_1 的估计量 $\hat{\beta}_1$ 为例，$\hat{\beta}_1$ 是此随机样本的函数：

$$\hat{\beta}_1 = h((X_1, Y_1), \cdots, (X_n, Y_n)),$$

其中，$h(\cdot)$ 是已知的，且不含有未知的参数。因此，β_1 的估计量是由 $h(\cdot)$ 定义的计算规则。式 (10) 给出的计算规则为：

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}.$$

因为 (X_i, Y_i) 是随机变量，所以，$\hat{\beta}_1$ 也是随机变量。

若给定随机变量 (X_i, Y_i) 的具体实现值 (x_i, y_i)，可以利用上式计算 $\hat{\beta}_1$ 的具体取值：

$$\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2},$$

这一取值被称为 β_1 的“估计值”。为了节约符号，将随机变量 (X_i, Y_i) 和其具体实现值 (x_i, y_i) 统一记为 (X_i, Y_i)；将 β_1 的估计量和估计值统一记为 $\hat{\beta}_1$。结合上下文可以对估计量和估计值进行区分，例如：$E(\hat{\beta}_1)$ 中的 $\hat{\beta}_1$ 表示估计量；$\hat{\beta}_1 = 0.2$ 中的 $\hat{\beta}_1$ 表示估计值。

小样本性质

小样本性质又称为有限样本 (finite sample) 性质，是指无论样本量 n 取值为多少都成立的统计性质。当然，为了满足假设 3.3，必须要求 $n \geqslant 2$。所以，更确切地说，一元线性回归系数估计量的小样本性质是指对任意不小于 2 的样本量 n 都成立的统计性质。

OLS 估计量的无偏性

若假设 3.1-3.4 成立，则 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 分别是 β_0 和 β_1 的无偏估计量，即

$$E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1.$$

考虑从总体中重复地抽取样本量为 n 的样本，每一次抽样计算一次系数的估计值，这样经过大量重复抽样得到不同的系数估计值。无偏性表明这些系数估计值的均值近似等于系数的真实值。

- OLS 估计量无偏性成立需要假设 3.1-3.4，
- 从证明过程中可以看到，这其中最重要的假设是 3.4。
- 在一元线性回归中，假设 3.4 经常是不成立的，导致此假设不成立的常见原因之一是遗漏变量。
- 解决遗漏变量问题是使用多元线性回归的主要原因之一。

OLS 估计量的方差（同方差情况）

若假设 3.1-3.5 成立，则以 $(X_1, \cdots, X_n)$ 为条件 $\hat{\beta}_1$ 和 $\hat{\beta}_0$ 的条件方差分别是：

$$\mathrm{Var}(\hat{\beta}_1 \mid X_1, \cdots, X_n) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \equiv \frac{\sigma^2}{\mathrm{TSS}_X}, \tag{17}$$

$$\mathrm{Var}(\hat{\beta}_0 \mid X_1, \cdots, X_n) = \frac{\sigma^2 n^{-1} \sum_{i=1}^n X_i^2}{\sum_{i=1}^n (X_i - \bar{X})^2}. \tag{18}$$

从式 (17) 可以看出，影响 $\hat{\beta}_1$ 的方差的因素有误差项的方差 σ^2 和解释变量的总样本变异 TSS_X。直观上，影响 Y 的不可观测因素的波动幅度越大，精确估计 β_1 就越困难；同时，解释变量的样本变异性越强，越容易估计 β_1。

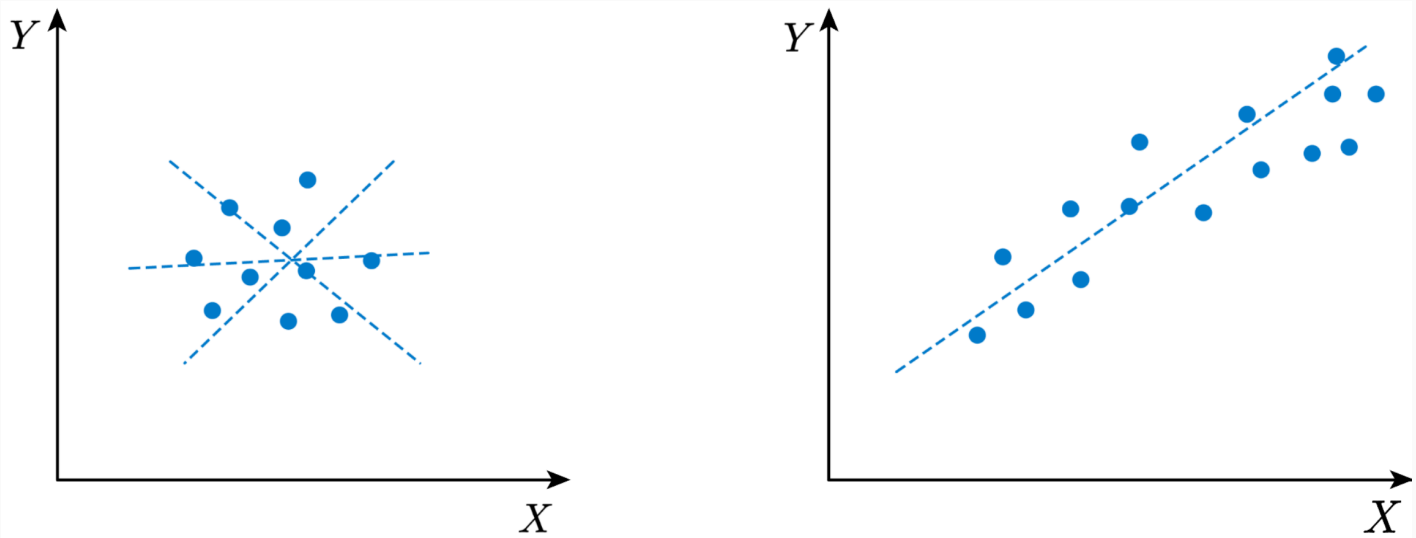


图 5: 解释变量样本变异越强越利于估计 β_1

- 为使上述结论成立需要施加假设 3.5，因此，式 (17) 只适用于同方差的情况。
- 在同方差假设下，OLS 估计量的方差形式简洁，并且可以证明此种形式的方差具有有效性，即在某一类估计量中其方差最小（Gauss-Markov 定理）。
- 然而，前面曾指出异方差是“常态”，因此，如今的实证分析很少基于式 (17) 进行统计推断，在牺牲有效性的同时更加注重稳健性。

- 除了对回归系数进行估计之外，计量经济学的另一项任务是对回归系数进行假设检验。
- 例如，检验受教育年限对工资的影响是否显著（significant），即考虑原假设 $\mathbf{H}_0 : \beta_1 = 0$，其备择假设为 $\mathbf{H}_1 : \beta_1 \neq 0$。
- 为了构造检验 $\mathbf{H}_0$ 的统计量，需要对式 (17) 给出的方差进行估计，也就是估计误差项的方差 σ^2。
- 注意到 $E(\varepsilon_i^2) = \sigma^2$，最直接的想法是用 $n^{-1} \sum_i \varepsilon_i^2$ 估计 σ^2。然而，ε_i 是无法观测的，不能直接用来构造估计 σ^2 的估计量。

如果用残差 e_i 替换误差项 ε_i，那么，$n^{-1}\sum_i e_i^2$ 是否是一个“理想的估计量”呢？若“理想的估计量”应该满足无偏性，则 $n^{-1}\sum_i e_i^2$ 不是一个理想的估计量。这是因为 OLS 残差的自由度不是 n。为了计算残差，需要求解 (12)，这就损失了两个自由度。也就是说，如果知道了 $n-2$ 个残差，根据式 (12)，可以求得另外两个残差。因此，一元线性回归的残差的自由度是 $n-2$。通过调整自由度可以得到如下 σ^2 的一个无偏估计量：

$$s^2 \equiv \frac{\sum_i e_i^2}{n-2} = \frac{\text{RSS}}{n-2}.$$

误差项方差的无偏估计

若假设 3.1-3.5 成立，则：$E(s^2) = \sigma^2$。

Monte Carlo 模拟

- Monte Carlo 模拟，也称为统计模拟方法，是指依赖重复随机抽样来获得数值结果的一类计算方法。
- 20 世纪 40 年代末，数学家斯塔尼斯拉夫 · 乌拉姆（Stanislaw Ulam）和约翰 · 冯 · 诺伊曼（John von Neumann）在美国 Los Alamos 国家实验室从事核武器项目时发明了 Monte Carlo 模拟方法。
- 为了保密，两人的工作需要一个代号，他们的同事尼古拉斯 · 梅特罗波利斯（Nicholas Metropolis）建议取名为 Monte Carlo。
- Monte Carlo 名字来源于摩纳哥的 Monte Carlo 赌场，Ulam 的叔叔常去那里赌博。
- 在计量经济学中，Monte Carlo 模拟常被用来考察估计量或检验统计量的小样本性质。
- Monte Carlo 方法的优势在于研究者可以根据需要设定 DGP 并且按此过程生成数据用于模拟实验，这使得研究者可以在已知和可控的统计环境下分析估计量或检验统计量的小样本表现。

通过模拟验证无偏性

考虑如下 DGP：

$$Y = 3 + 2 \times X + \varepsilon,$$

其中，$X \sim \mathcal{N}(0, 3^2)$，$\varepsilon \sim \mathcal{N}(\mu, \sigma^2)$，$\mu$ 和 σ 可以根据模拟的需要进行设定。这里只考虑斜率系数的估计。为了验证无偏性，需要从上述 DGP 中重复抽样，假设每次抽样的样本量为 $n = 100$，重复抽样的次数为 $d = 1000$。具体的模拟过程为：

1. 按照 $\mathcal{N}(0, 3^2)$ 分布生成 100 个随机数作为解释变量 X；
2. 根据设定的 μ 和 σ，按照 $\mathcal{N}(\mu, \sigma^2)$ 分布生成 100 个随机数作为误差项；
3. 根据 DGP 计算 Y，进而得到样本 $\{(X_i, Y_i) : i = 1, \cdots, 100\}$；
4. 利用第 3 步得到的样本进行 OLS 估计，得到斜率系数的估计值 $\hat{\beta}_1$；
5. 将步骤 1-4 重复 1000 次，得到斜率系数的 1000 个估计值

$$\{\hat{\beta}_1^{(1)}, \cdots, \hat{\beta}_1^{(1000)}\}$$

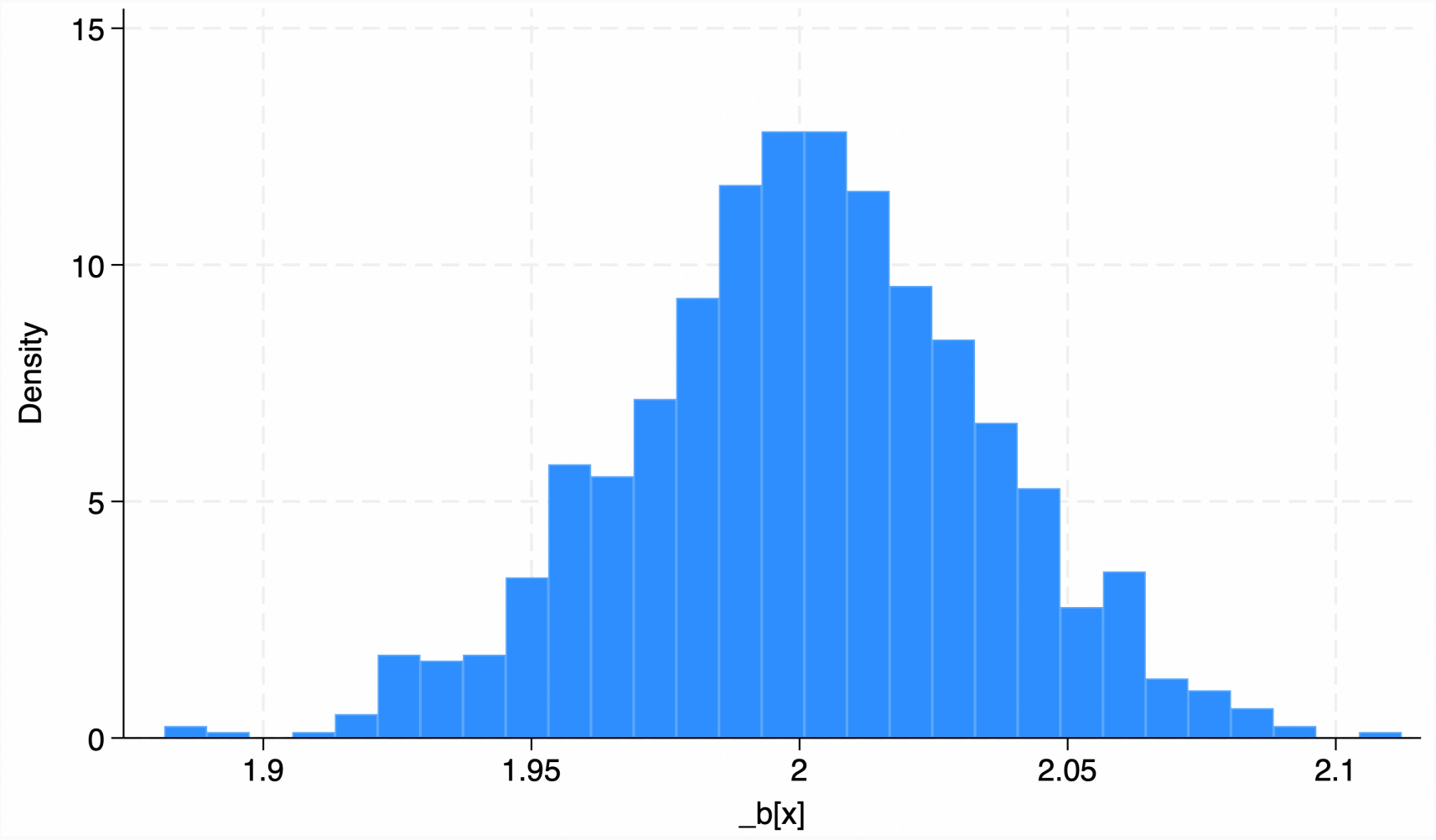
6. 计算斜率系数估计值的均值，并绘制直方图。

```
. clear all
. set seed 20221227
. program define ulrm, rclass //定义函数ulrm, 函数返回结果保存在r
    ()中
. version 18.0                  //设置Stata版本
. syntax [, obs(integer 1) mu(real 0) sigma(real 1)]
                                //设置函数ulrm的选项
.    clear
.    set obs `obs'              //将选项obs设置为样本量
.    gen x = rnormal(0,3)
.    gen u = rnormal(`mu',`sigma') //按照选项设置的mu和sigma生成误
    差项
.    gen y = 3 + 2*x + u
.    reg y x
.    end                                //函数定义结束

. simulate beta=_b[x], reps(1000): ulrm, obs(100) mu(0) sigma(1)
                                //调用函数ulrm进行1000次模拟
. sum                           //汇报斜率估计值的描述统计
. hist beta                     //绘制直方图
```

表 1： 1000 次模拟得到的斜率估计值的描述统计

变量	样本量	均值	标准差	最小值	最大值
beta	1000	2.001245	0.0334709	1.881541	2.112275



虚拟解释变量

考虑如下一元线性回归：

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{19}$$

其中，X 为二值的虚拟变量。例如，Y 表示工资，X 表示是否具有大学学历：$X = 1$ 表示具有大学学历；$X = 0$ 表示不具有大学学历。只要假设 3.1-3.5 成立，即便 X 是虚拟变量，前文提到的 OLS 估计量的小样本性质也依然是成立的。这里主要关注如何解释系数 β_1。

若假设 3.4 成立，β_1 为线性投影系数，当 X 为虚拟变量时，可以证明式 (11) 等价于：

$$\beta_1 = E(Y \mid X = 1) - E(Y \mid X = 0). \tag{20}$$

上式也可以直接由假设 3.4 得出。事实上，以 X 为条件对式 (19) 两端同时取条件期望可以推出：

$$E(Y \mid X) = \beta_0 + \beta_1 X + E(\varepsilon \mid X) = \beta_0 + \beta_1 X.$$

将 $X = 1$ 和 $X = 0$ 分别代入上式，可得：

$$E(Y \mid X = 1) = \beta_0 + \beta_1, \quad E(Y \mid X = 0) = \beta_0.$$

两式做差即可得到 (20)。

从总体的角度看，根据 X 的不同取值，总体中的个体被分为两组，分别对应着 $X = 1$ 和 $X = 0$。式 (20) 说明 β_1 是两组 Y 的期望之差。仍然以大学学历对工资的影响为例，β_1 是具有大学学历的群体与不具有大学学历的群体的工资期望之差。

X 是虚拟变量并不影响利用式 (9) 和 (10) 计算回归系数的估计值，但是，根据式 (20)，β_1 还可以按照如下方式估计：

$$\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0, \tag{21}$$

其中，$\bar{Y}_1$ 是样本中 $X_i = 1$ 的个体的 Y_i 的样本均值；$\bar{Y}_0$ 是样本中 $X_i = 0$ 的个体的 Y_i 的样本均值。

除了从式 (20) 出发得到 $\hat{\beta}_1$，还可以利用只含有常数项的回归的性质得到 $\hat{\beta}_1$。事实上，若只考虑 $X_i = 1$ 的样本，则有：

$$Y_i = \beta_0 + \beta_1 + \varepsilon_i.$$

于是，$\widehat{\beta_0 + \beta_1} = \bar{Y}_1$。同理，$\hat{\beta}_0 = \bar{Y}_0$。因此，

$$\widehat{\beta_0 + \beta_1} - \hat{\beta}_0 = \hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0.$$

式 (21) 给出的估计量被称为均值差分（Difference-in-Mean, DIM）估计量。

再次考虑 [Wooldridge, 2019] 提供的数据集 WAGE1.dta，定义 educ 不小于 16 年为具有大学学历，据此生成虚拟变量 bachelor，将 wage 对 bachelor 进行回归。此外，分别计算具有大学学历和不具有大学学历的个体的工资均值，并将两个均值做差后与回归系数估计值进行比较。使用的 Stata 程序如下：

```
. use WAGE1.dta, clear
. gen bachelor = 1 if educ >= 16   //若educ不小于16定义bachelor为1
. replace bachelor = 0 if educ < 16 //若educ小于16定义bachelor为0
. reg wage bachelor
. sum wage if bachelor == 1     //汇报具有大学学历的子样本的工资描
    述统计
. gen y1 = r(mean)              //将此子样本工资的均值赋值给y1
. sum wage if bachelor == 0     //汇报不具有大学学历的子样本的工资
    描述统计
. gen y0 = r(mean)              //将此子样本工资的均值赋值给y0
. dis y1-y0                     //显示两个均值之差^^I
```

第 4 行命令的回归结果可以总结为如下方程：

$$\widehat{wage} = 5.189 + 3.759 \times bachelor.$$

于是，具有大学学历的平均工资预计比不具有大学学历的平均工资高 3.759 美元。第 9 行命令输出的结果与回归命令估计的斜率系数完全一致。但是，使用回归的好处是可以得到 DIM 估计量的标准误（standard error）。[1]

[1] 标准差的估计量被称为标准误。

潜在结果与因果推断

为了讨论问题的方便，这里只分析二值虚拟变量和被解释变量之间的因果关系，这种情况下的因果效应常被称为 X 对 Y 的处置效应（treatment effect）。

为了强调虚拟变量 X 的特殊性，下文将使用字母 D 来代替 X。虚拟变量处置效应分析的重要应用领域是政策评估（policy evaluation）。

- $D = 1$ 的个体是政策实施的对象，属于处理组（treatment group）
- $D = 0$ 的个体不受政策影响，属于控制组（control group）

评估政策的效果就是对处置效应进行估计和检验。

- 前面曾经强调反事实比较在推断因果过程中起到了关键的作用。
- 真实结果与反事实结果是同一个体在两种不同的政策处置状态下呈现出的结果，
- 反事实比较实现了“控制其他因素不变”情况下的因果效应分析。
- 定义因果效应的核心是如何体现反事实的比较。目前，在使用统计或计量的手段分析因果关系时，定义因果效应的主流方法是引入潜在结果（potential outcome）框架。
- 该框架最早由 [Neyman, 1923] 在完全随机化实验的情况下提出，并由 [Rubin, 1974] 在非随机的观测研究中进一步推广完善。

个体 i 在处置变量 D_i 下的潜在结果定义为：

$$\text{潜在结果} = \begin{cases} Y_{1i} & \text{如果 } D_i = 1, \\ Y_{0i} & \text{如果 } D_i = 0. \end{cases}$$

这里需要注意，对任意的个体 i，不可能同时观测到 Y_{1i} 和 Y_{0i}。但是，这并不影响定义潜在结果。Y_{1i} 是假如 $D_i = 1$ 时会出现的结果，Y_{0i} 是假如 $D_i = 0$ 时会出现的结果，这与个体 i 的真实处置状态 D_i 取值是 0 还是 1 无关。

潜在结果 (Y_{0i}, Y_{1i}) 和观测结果 Y_i 之间的区分是计量因果推断的核心所在。

有了潜在结果的定义，处置 D_i 对个体 i 的处置效应可以容易地定义为：

$$\tau_i = Y_{1i} - Y_{0i}. \tag{22}$$

τ_i 是在“控制其他因素不变”的情况下比较处置实施与否导致的结果差异。τ_i 的定义也充分体现了反事实比较的思想，这是因为两个潜在结果中的一个在处置实施之后会成为观测结果，而另一个就是反事实的结果。

由于每一个个体 i 的处置效应 τ_i 很难估计，在实证分析时，研究者通常关注的是平均处置效应（average treatment effect, ATE），其定义为：

$$\tau_{ate} = E[\tau_i] = E[Y_{1i} - Y_{0i}]. \tag{23}$$

此外，研究者还可能关注处理组的平均处置效应（average treatment effect on the treated, ATET）和控制组的平均处置效应（average treatment effect on the untreated, ATENT），它们的定义分别为：

$$\tau_{atet} = E[Y_{1i} - Y_{0i} \mid D_i = 1], \tag{24}$$

$$\tau_{atent} = E[Y_{1i} - Y_{0i} \mid D_i = 0]. \tag{25}$$

一元线性回归是否可以用来推断因果关系？

考虑如下回归方程：

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i.$$

根据上一节的分析，β_1 可以表示为：

$$\beta_1 = E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0).$$

并且，最小二乘估计量 $\hat{\beta}_1$ 通过 DIM 的方式估计 β_1。

β_1 是否具有处置效应的解释？

以 ATET 为例，β_1 可以进一步表示为：

$$
\begin{aligned}
\beta_1 &= E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0) \\
&= E(Y_{1i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 0) \\
&= E(Y_{1i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 1) + E(Y_{0i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 0) \\
&= \tau_{atet} + E(Y_{0i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 0).
\end{aligned}
$$

由上式可知 β_1 并不等于 τ_{atet}，而是在 ATET 的基础上加上了 $E(Y_{0i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 0)$，这两个期望的差代表了样本选择偏误 (selection bias)。

若要一元线性回归可以用来推断因果关系，它需要满足什么条件呢？为了回答这个问题，需要建立观测结果和潜在结果之间的关系，即潜在结果模型（potential outcome model, POM）：

$$Y_i = Y_{0i} + D_i(Y_{1i} - Y_{0i}). \tag{26}$$

同质处置效应

对任意个体 i，$Y_{1i} - Y_{0i} = \tau$。此时，POM 可以表示为：

$$Y_i = Y_{0i} + \tau D_i. \tag{27}$$

现将上式转换为一元线性回归的形式，为此，定义 $\beta_0 = E(Y_{0i})$，则有：

$$Y_i = \beta_0 + \tau D_i + \varepsilon_{0i},$$

其中，$\varepsilon_{0i} = Y_{0i} - E(Y_{0i}) = Y_{0i} - \beta_0$，$\tau$ 为 τ_{ate}。显然假设 3.1 和 3.2 是可以满足的。同时，只要样本里既有处理组个体又有控制组个体，假设 3.3 也是可以满足的。于是，只要 $E(\varepsilon_{0i} \mid D_i) = 0$，$\tau$ 的 OLS 估计量 $\hat{\tau}$ 就是 ATE 的无偏估计。而 $E(\varepsilon_{0i} \mid D_i) = 0$ 等价于 $E(Y_{0i} \mid D_i) = E(Y_{0i})$，即 Y_{0i} 均值独立于 D_i。

异质处置效应

对于更一般的情况，τ_i 可以表示为：

$$\tau_i = Y_{1i} - Y_{0i} = E(Y_{1i}) + \varepsilon_{1i} - [E(Y_{0i}) + \varepsilon_{0i}] = \tau_{ate} + [\varepsilon_{1i} - \varepsilon_{0i}],$$

其中，$\varepsilon_{1i} = Y_{1i} - E(Y_{1i})$。由式 (26)，容易得到：

$$Y_i = Y_{0i} + D_i\left[\tau_{ate} + (\varepsilon_{1i} - \varepsilon_{0i})\right] = \beta_0 + \tau_{ate} \times D_i + \varepsilon_i,$$

其中，$\varepsilon_i = D_i(\varepsilon_{1i} - \varepsilon_{0i}) + \varepsilon_{0i}$。为了保证 OLS 估计量无偏，需要 $E(\varepsilon_i \mid D_i) = 0$，即

$$E(\varepsilon_i \mid D_i) = E\left[D_i(\varepsilon_{1i} - \varepsilon_{0i}) + \varepsilon_{0i} \mid D_i\right] = 0.$$

使上式成立的充分条件是 $E(\varepsilon_{1i} \mid D_i) = 0$ 且 $E(\varepsilon_{0i} \mid D_i) = 0$，也就是 Y_{0i} 和 Y_{1i} 均值独立于 D_i。

再议 R^2

- 在大学学历对工资的影响的例子中，能由是否具有大学学历这一虚拟变量解释的工资样本变异占工资总样本变异的比例通常很低
- 但是只要假设 3.4 成立，这样的一元线性回归依然可以用来估计大学学历对工资的因果效应
- 从因果推断的角度看，拟合优度低的回归不一定就是没有用的回归，拟合优度高的回归也不一定就能确保估计出因果效应

Questions?

References i


Neyman, J. S. (1923).
On the application of probability theory to agricultural experiments. Essay on principles.
Annals of Agricultural Sciences, 10:1–51.


Rubin, D. B. (1974).
Estimating causal effects of treatments in randomized and nonrandomized studies.
Journal of Educational Psychology, 66(5):688–701.


Wooldridge, J. M. (2019).
***Introductory Econometrics: a Modern Approach*.**
Cengage Learning, 7 edition.