2.3多元回归系数估计的直观理解

蔡诺璇

2020.10.22

韦恩图（Venn Diagram）

- 用于直观理解最小二乘法估计多元线性回归模型系数的机理

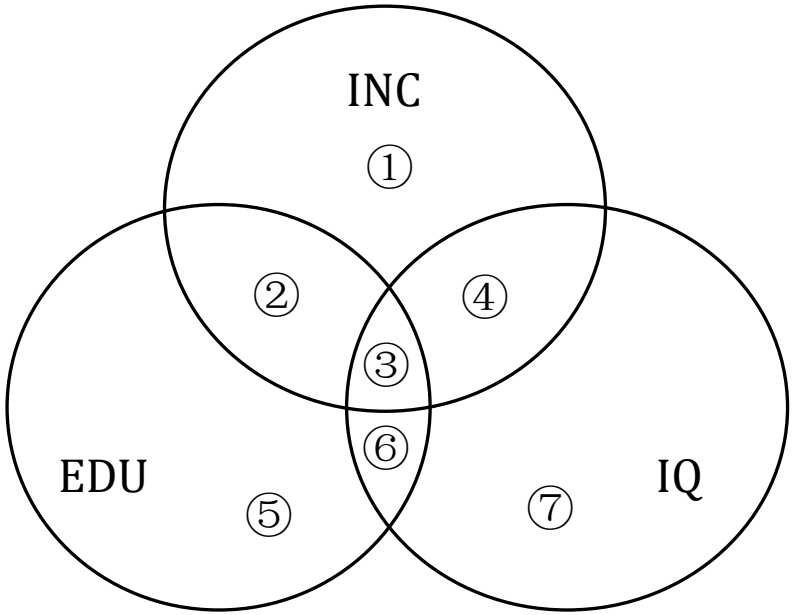
- 估计模型是

$$INC = \alpha + \beta_1 EDU + \beta_2 IQ + e$$
$$\mathbb{E}(e|EDU, IQ) = 0$$
$$\mathrm{Cov}(EDU, IQ) \neq 0$$

- 下面我们用这个模型来说明各情形

（一）INC、EDU和IQ的变化都是相关的

- 两两相交的部分指的是两个变量共同变化的部分，表明两个变量存在一定的线性相关关系

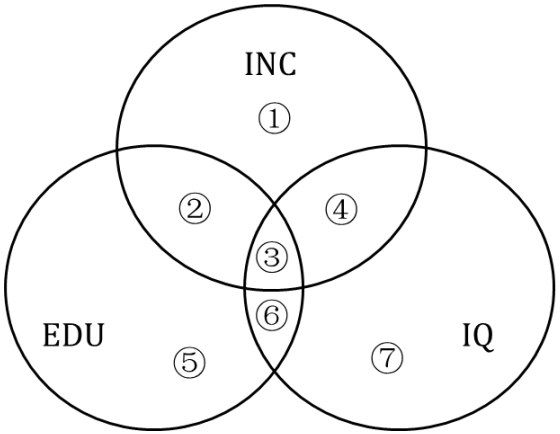


（一）INC、EDU和IQ的变化都是相关的

- （1）二元回归：

$$INC = \gamma_0 + \gamma_1 EDU + u, \quad \mathbb{E}(u|EDU) = 0$$

- 此时γ_1反映的是INC和EDU的相关性，及②和③的信息
- ①和④是INC变化与EDU无关的部分，即残差项的变化



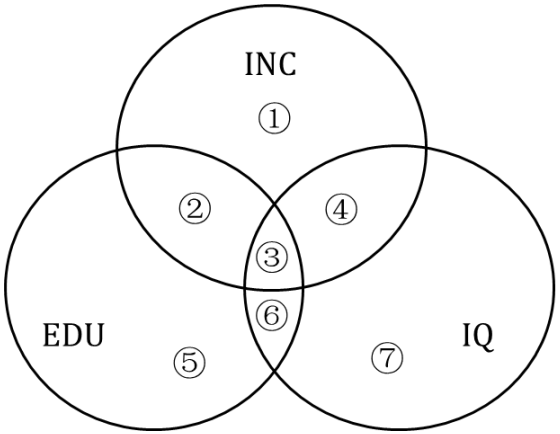
（一）INC、EDU和IQ的变化都是相关的

- （2）多元回归：

$$INC = \alpha + \beta_1 EDU + \beta_2 IQ + e$$

$$\mathbb{E}(e|EDU, IQ) = 0$$

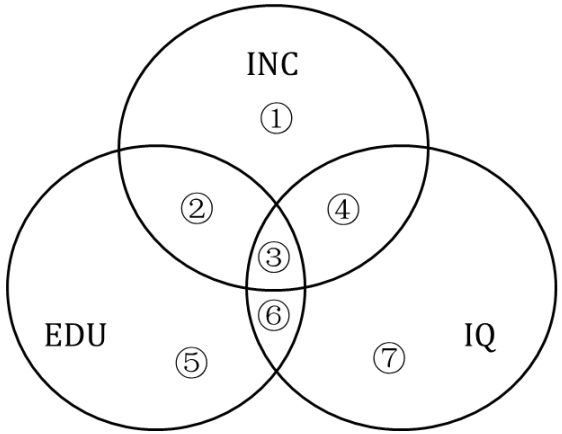
- 此时β_1反映的只有②的信息，β_2反映的只有④的信息，而信息③（反映了共同的影响）被同时舍去
- 此时，回归系数反映了EDU和IQ各自对INC的影响，即因果关系



（一）INC、EDU和IQ的变化都是相关的

- 结合第一节的结论：$\gamma_1 = \beta_1 + \beta_2\phi_1$
- 对应于韦恩图各部分：

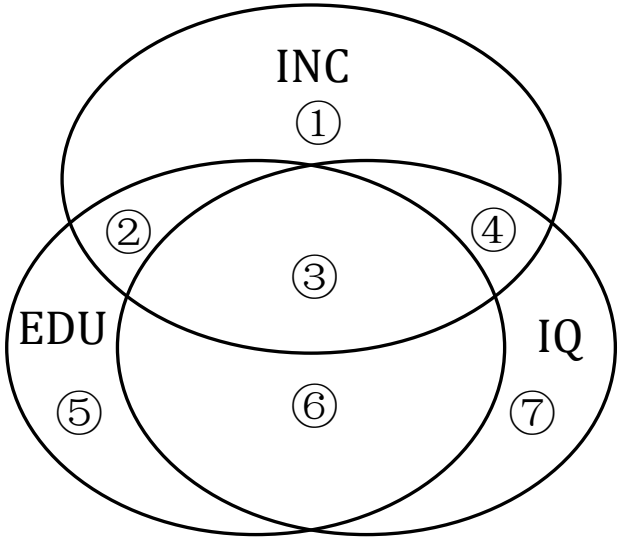
β_1指的是②； $\beta_2\phi_1$指的是③



（二）EDU和IQ高度共线性

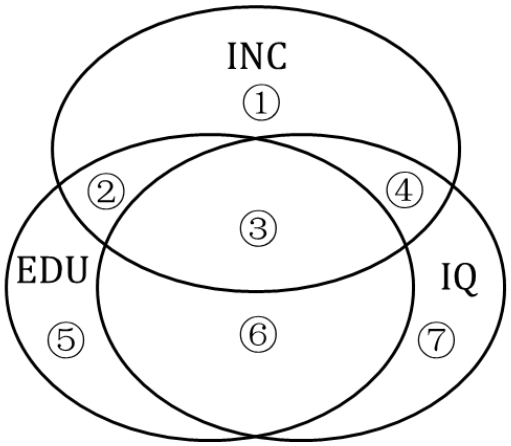
- EDU和IQ高度共线性指

EDU和IQ的重合面积③非常大，变量独有的②和④特别小



（二）EDU和IQ高度共线性

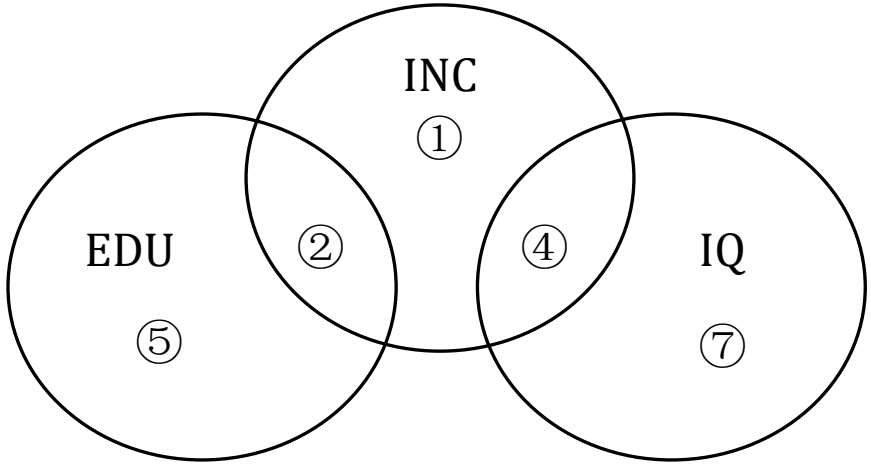
- 易知，由于两个变量可供估计的独立信息很少（②和④的面积很小），回归结果得到的系数β_1和β_2很难体现出统计显著
- 此时，我们将无法估计出β_1和β_2



（三）EDU和IQ不相关

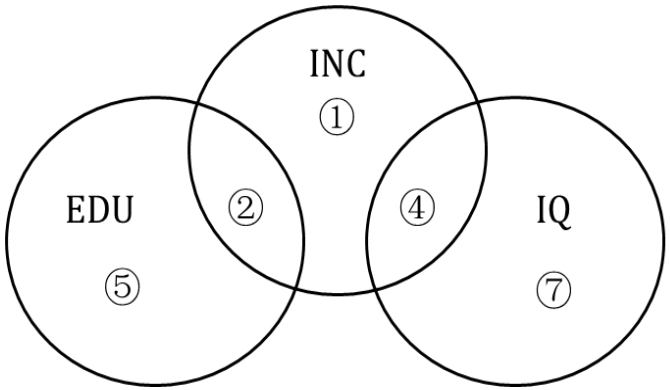
- EDU和IQ不相关指

INC和EDU、IQ均相交，但EDU和IQ不相交



（三）EDU和IQ不相关

- 此时，面积③不存在
- INC只对EDU回归的结果γ_1与INC同时对EDU和IQ回归的结果的EDU的系数β_1相同，即均为②
- 结论：当两个解释变量不相关时，缺失其中一个并不会影响另一个的回归系数

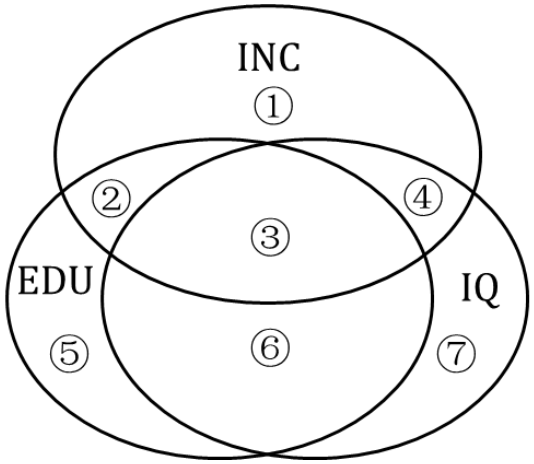


R^2和系数显著性

- 对于模型 $INC = \alpha + \beta_1 EDU + \beta_2 IQ + e$

其拟合系数R^2指的是(②+③+④)/(①+②+③+④)

- 对于情形二，即解释变量存在高度共线性问题时，R^2值很大，但解释变量各自的系数却不显著
- 所以我们在实证工作中通常要先检验解释变量之间是否存在共线性的问题



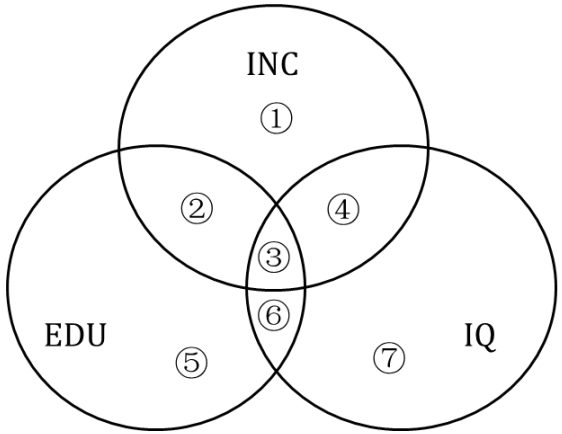
2.4多元线性回归分解

回归分解法

- 模型：

$$Y = \hat{\alpha} + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_i X_i + \cdots + \hat{\beta}_k X_k + \hat{e}$$

- 上式中任何一个解释变量 X_i 的系数 $\hat{\beta}_i$ 都可以分解为两步进行求解



回归分解法

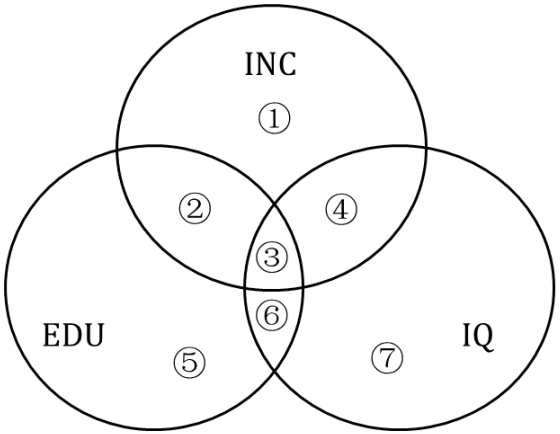
- 第一步：将X_i作为因变量，其他不包含X_i的解释变量作为自变量，用最小二乘法得到它们的线性函数关系残差项$\widetilde{X}_i=\hat{v}_i$:

$$X_i=\hat{\gamma}_0+\hat{\gamma}_1X_1+\cdots+\hat{\gamma}_{i-1}X_{i-1}+\hat{\gamma}_{i+1}X_{i+1}+\cdots+\hat{\gamma}_kX_k+\hat{v}_i$$

- 第二步：将Y作为因变量，$\widetilde{X}_i$作为自变量，用最小二乘法得到下面方程的系数β_i:

$$Y=\hat{a}+\hat{\beta}_i\widetilde{X}_i+\hat{\varepsilon}$$

即$\hat{\beta}_i=\dfrac{\mathrm{Cov}(Y,\hat{v}_i)}{\mathrm{Var}(\hat{v}_i)}$



证明：

将$Y = \hat{\alpha} + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_i X_i + \cdots + \hat{\beta}_k X_k + \hat{e}$代入

$\frac{\mathrm{Cov}(Y,\hat{v}_i)}{\mathrm{Var}(\hat{v}_i)}$中，有：

$$\frac{\mathrm{Cov}(Y, \hat{v}_i)}{\mathrm{Var}(\hat{v}_i)}$$

$$= \frac{\mathrm{Cov}(\hat{\alpha} + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_i X_i + \cdots + \hat{\beta}_k X_k + \hat{e}, \hat{v}_i)}{\mathrm{Var}(\hat{v}_i)}$$

$$= \frac{\mathrm{Cov}(\hat{\beta}_i X_i + \hat{e}, \hat{v}_i)}{\mathrm{Var}(\hat{v}_i)}$$

证明：

考虑

$$\mathrm{Cov}(X_i, \hat{v}_i)$$
$$= \mathrm{Cov}(\hat{\gamma}_0 + \hat{\gamma}_1 X_1 + \cdots + \hat{\gamma}_{i-1} X_{i-1} + \hat{\gamma}_{i+1} X_{i+1} + \cdots + \hat{\gamma}_k X_k + \hat{v}_i, \hat{v}_i)$$
$$= \mathrm{Cov}(\hat{v}_i, \hat{v}_i)$$
$$= \mathrm{Var}(\hat{v}_i)$$

证明：

考虑

$$\mathrm{Cov}(\hat{e}, \hat{v}_i)$$

$$= \mathrm{Cov}(\hat{e}, X_i - \hat{\gamma}_0 - \hat{\gamma}_1 X_1 - \cdots - \hat{\gamma}_{i-1} X_{i-1} - \hat{\gamma}_{i+1} X_{i+1} - \cdots - \hat{\gamma}_k X_k)$$

$$= 0$$

所以

$$\frac{\mathrm{Cov}(\hat{\beta}_i X_i + \hat{e}, \hat{v}_i)}{\mathrm{Var}(\hat{v}_i)}$$

$$= \hat{\beta}_i$$

2.5内生性和因果关系

内生性问题

- 对于给定的线性回归模型

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + e$$

如果干扰项和解释变量是相关的，即

$$\mathbb{E}(e | X_1, X_2, \cdots, X_k) \neq 0$$

那么可以说这个线性模型存在内生性问题

- “内生性问题”指当干扰项和解释变量相关时，我们无法识别解释变量的因果关系系数的情形

偏差

- “内生性问题会造成最小二乘法估计系数有偏”
- 在因果关系分析中，偏差指相关关系系数对因果关系系数的偏差
- 模型2：$INC = \alpha + \beta_1 EDU + \varepsilon$，$\mathbb{E}(\varepsilon|\mathrm{EDU}) \neq 0$
- 模型3：$INC = \gamma_0 + \gamma_1 EDU + u$，$\mathbb{E}(u|EDU) = 0$
- 我们期望估计出β_1，但实际是对γ_1的估计，而 $\gamma_1 = \beta_1 + \beta_2\phi_1$，$\beta_2\phi_1 \neq 0$。所以$\widehat{\gamma_1}$是对因果关系系数$\beta_1$的有偏估计

内生性来源

- 来源一：遗漏解释变量

考虑模型：$INC = \alpha + \beta_1 EDU + \beta_2 IQ + e$

$$\mathbb{E}(e|EDU, IQ) = 0，\ \mathrm{Cov}(EDU, IQ) \neq 0$$

若遗漏了解释变量IQ，即使用$INC = \alpha + \beta_1 EDU + v$进行回归，则

$$\mathbb{E}(v|EDU) = \mathbb{E}(\beta_2 IQ + e|EDU) = \beta_2 \mathbb{E}(IQ|EDU) \neq 0$$

- 遗漏解释变量会造成内生性原因是遗漏变量和未遗漏解释变量之间存在相关性

内生性来源

- 来源二：测量误差
- （1）解释变量的测量误差

考虑模型：

$$Y^* = \beta_0 + \beta_1 X^* + e，\ \mathbb{E}(e|X^*) = 0$$

当解释变量X^*存在测量误差，即$X = X^* + \mathrm{u}$，同时

$$\mathrm{Cov}(u, X^*) = 0, Cov(u, Y^*) = 0, \mathbb{E}(u|X^*) = 0$$

此时模型变成

$$Y^* = \beta_0 + \beta_1(X - u) + e = \beta_0 + \beta_1 X + v$$

$$v = -\beta_1 u + e$$

内生性来源

$$\begin{aligned}&\mathrm{Cov}(X, v)\\&= \mathrm{Cov}(X^* + \mathrm{u}, -\beta_1 u + e)\\&= \mathrm{Cov}(u, -\beta_1 u)\\&= -\beta_1 \mathrm{Var}(u) \neq 0\end{aligned}$$

- 解释变量存在测量误差时，会造成内生性问题
- 原因：解释变量存在测量误差时会造成干扰项里面存在测量误差，从而导致干扰项和观测的解释变量相关（均包含了测量误差）

内生性来源

- （2）被解释变量的测量误差

考虑模型：

$$Y^* = \beta_0 + \beta_1 X^* + e，\ \mathbb{E}(e|X^*) = 0$$

当解释变量Y^*存在测量误差，即$Y = Y^* + u$，同时

$$\mathrm{Cov}(u, X^*) = 0, \mathrm{Cov}(u, Y^*) = 0, \mathbb{E}(u|X^*) = 0$$

此时模型变成

$$Y = \beta_0 + \beta_1 X^* + e + u = \beta_0 + \beta_1 X^* + v$$

$$v = e + u$$

内生性来源

$$\begin{aligned}&\mathrm{Cov}(X^*, v)\\&= \mathrm{Cov}(X^*, e + u)\\&= 0\end{aligned}$$

- 当被解释变量存在测量误差时，不会造成内生性问题
- 但是由于误差项（噪音）变大，回归结果的显著性会有所降低

内生性来源

- 来源三：互为因果
- 若两个变量互为因果，则任何一方都可以作为对方的解释变量，此时任何一个单方面的回归都存在内生性问题

内生性来源

- 考虑如下情形

$$Y_1 = \beta_1 X_1 + \phi_1 Y_2 + e_1 \qquad (1)$$

$$Y_2 = \beta_2 X_2 + \phi_2 Y_1 + e_2 \qquad (2)$$

且有

$$\mathbb{E}(e_i | X_1, X_2) = 0 \qquad i = 1,2$$

$$\mathrm{Cov}(e_1, e_2) = 0$$

将式（2）代入式（1）中，可以得到

$$Y_1 = \frac{\beta_1}{1-\phi_1\phi_2} X_1 + \frac{\beta_2\phi_1}{1-\phi_1\phi_2} X_2 + \frac{e_1}{1-\phi_1\phi_2} + \frac{e_2\phi_1}{1-\phi_1\phi_2} \qquad (3)$$

内生性来源

由式子（3）

$$\mathrm{Cov}(Y_1, e_2)$$

$$= \mathrm{Cov}\left(\frac{\beta_1}{1-\phi_1\phi_2}X_1 + \frac{\beta_2\phi_1}{1-\phi_1\phi_2}X_2 + \frac{e_1}{1-\phi_1\phi_2} + \frac{e_2\phi_1}{1-\phi_1\phi_2}, e_2\right)$$

$$= \mathrm{Cov}\left(\frac{e_2\phi_1}{1-\phi_1\phi_2}, e_2\right)$$

$$= \frac{\phi_1}{1-\phi_1\phi_2}\mathrm{Var}(e_2) \neq 0$$

- 所以模型(2)存在内生性问题，模型（1）同理可证

条件期望函数

- 条件期望值$\mathbb{E}(Y \mid X = x)$是指给定一个$X = x$值，对应的Y的期望值（即Y的概率平均值）
- Y的期望值是关于X的函数，即$\mathbb{E}(Y \mid X = x) = f(x)$，我们将其称为条件期望函数
- 计算：

$$\mathbb{E}(Y \mid X = x) = \sum_{y} yP(X = x, Y = y)$$

条件期望函数

- 残差项

Y和条件期望值$E(Y \mid X)$之间的差异称为残差$\hat{e}$

总可以把Y分解为：

$Y = \mathbb{E}(Y \mid X) + \hat{e}$，其中$\mathbb{E}(\hat{e} \mid X) = 0$

- 证明：

$$\mathbb{E}(\hat{e} \mid X)$$
$$= \mathbb{E}(Y - \mathbb{E}(Y \mid X) \mid X)$$
$$= \mathbb{E}(Y \mid X) - \mathbb{E}(Y \mid X)$$
$$= 0$$