多元线性回归

李德山

[1] 西南科技大学 经济管理学院

[2]School of Economics and Management
Southwest University of Science and Technology

2021 年 3 月 18 日

主要内容

1. 线性回归模型与因果推断

主要内容

1. 线性回归模型与因果推断
2. 多元线性回归 OLS 估计量

主要内容

1. 线性回归模型与因果推断
2. 多元线性回归 OLS 估计量
3. 古典线性模型的假定

主要内容

1. 线性回归模型与因果推断
2. 多元线性回归 OLS 估计量
3. 古典线性模型的假定
4. 多元线性回归系数的直观理解

主要内容

1. 线性回归模型与因果推断
2. 多元线性回归 OLS 估计量
3. 古典线性模型的假定
4. 多元线性回归系数的直观理解
5. 假设检验与置信区间

主要内容

1. 线性回归模型与因果推断
2. 多元线性回归 OLS 估计量
3. 古典线性模型的假定
4. 多元线性回归系数的直观理解
5. 假设检验与置信区间
6. 小样本 OLS 和大样本 OLS

线性回归模型与因果推断

- 我们首先回顾一下受教育年限对收入水平影响的例子。

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

其中，EDU 为核心解释变量（处置变量）；IQ 为控制变量；e 为干扰项。

- 如果回归方程中遗漏了许多重要的潜在的对收入影响的因素，比如能力、智商、家庭背景等。在这里，智商（IQ）是一个可观测变量。智商对收入有影响；同时，智商越高，受教育程度也越高。要单独估计 EDU 和 INC 的因果影响，我们必须控制住 IQ 对 INC 的影响。

线性回归模型与因果推断

- 遗漏变量偏误产生必须同时满足两个条件：（1）核心解释变量与遗漏变量相关；（2）遗漏变量是被解释变量的一个决定因素。
- 在一元回归中，误差项包含了除了核心解释变量之外所有决定被解释变量的因素，如果这些遗漏的因素中有一个与核心解释变量有关，那么 $E(u|x) \neq 0$。即使在大样本条件下，偏误也无法消除，而且 OLS 估计量不是一致估计量。
- 存在遗漏变量本身并不要紧，关键在于，遗漏变量不能与方程的解释变量相关。
- 解决遗漏变量偏差的方法主要有：加入控制变量；随机实验与自然实验；IV；面板数据模型。

线性回归模型与因果推断

- 由上一章的方程（20）可知，

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})e_i}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1}$$

- 同时结合最小二乘的两个重要假设（独立同分布；较大异常值是不太可能出现），可得：

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \xrightarrow{p} \sigma_x^2$$

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})e_i \xrightarrow{p} \mathrm{cov}(x_i, e_i) = \rho_{xe}\sigma_x\sigma_e$$

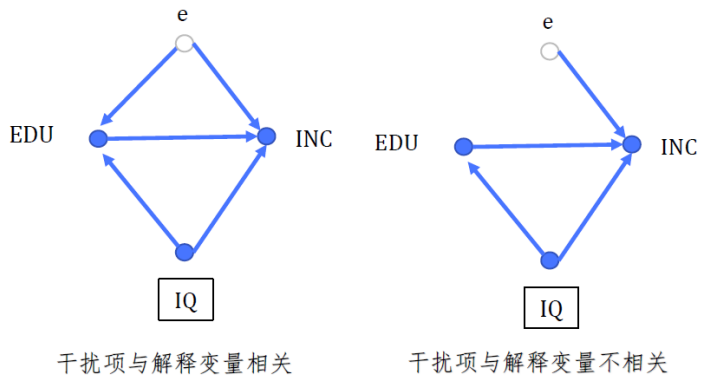
- 代入到方程（1）中，可得:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{xe}\frac{\sigma_x}{\sigma_e} \tag{2}$$

- 式（2）概括了有关遗漏变量偏差的几点思想：
  1. 无论样本容量多大，遗漏变量偏误问题都存在。β 的估计值并不收敛到真值！
  2. 偏误的大小取决于解释变量与误差项之间的相关系数。相关系数的绝对值越大，偏误就越大。
  3. 偏误的方向（系数高估还是低估）取决于解释变量与误差项是正相关还是负相关。相关系数小于 0，OLS 估计量就是被低估，反之。

线性回归模型与因果推断

- 如果干扰项包含混淆变量，则无法识别 EDU 对 INC 的因果影响有多大。如干扰项中包含了个人的人格特征（勤奋）。越勤奋的人受教育程度可能越高，同时越勤奋的人收入也可能越高。



- 左图中，如果干扰项包含了混淆变量，该混淆变量又同时与 EDU 和 INC 相关，即使控制了 IQ，EDU 到 INC 仍然有两条路径：
  1. 因果路径 $EDU \rightarrow INC$
  2. 混淆路径 $EDU \leftarrow e \rightarrow INC$
- 右图中，控制了 IQ 后，干扰项 e 与 EDU 就不存在混淆路径了。这时候可以识别出二者的因果关系。

线性回归模型与因果推断

- 回归中通常使用的条件是：**干扰项条件均值独立于解释变量**，即：

$$E(e|EDU, IQ) = E(e) = c$$

- 上式的意思是，对于给定任意受教育程度和智商水平的人，干扰项的平均值都是一样的（等于常数 c）。例如，人格特征会影响收入，由于人格特征无法观测而进入干扰项，如果干扰项条件均值独立于解释变量成立，则意味着对于给定任意受教育程度和智商水平的个体，他们的人格特征平均值都等于常数 c。因此，当教育水平或智商水平发生变化时，人格特征的平均值并不会发生变化。在这种情况下，收入平均值就可以归因于教育或智商的变化。
- 由于线性回归模型里包含常数项，所以我们可以把上式的常数 c 并入常数项，使得 $E(e|EDU, IQ) = 0$。
- 综上，线性回归模型要满足上一章我们讲到的两个重要假设：线性关系假设；干扰项条件均值为 0 假设。

线性回归模型与因果推断

- 对线性函数两边取期望值，得到**线性条件期望函数**，CEF：

$$CEF: E(INC|EDU, IQ) = \alpha + \beta_1 EDU + \beta_2 IQ$$

  - 将条件期望函数分别对 EDU 和 IQ 求偏导，可得：

$$\frac{d(E(INC|EDU, IQ))}{dEDU} = \beta_1$$

$$\frac{d(E(INC|EDU, IQ))}{dIQ} = \beta_2$$

- 如果干扰项均值独立于解释变量的假设不成立，干扰项的存在如何影响因果关系的估计？
  - 假设我们只观察到了 EDU 和 INC：

$$INC = \alpha + \beta_1 EDU + \varepsilon$$

$$\varepsilon = \beta_2 IQ + e$$

线性回归模型与因果推断

- 此时，EDU 和 IQ 存在相关性，那么：

$$E(\varepsilon|EDU) = E(\beta_2 IQ + e|EDU) = \beta_2 E(IQ|EDU) \neq 0$$

- 在这种情况下，干扰项不满足条件均值为 0 的假设。此时的线性模型为：

$$INC = \alpha + \beta_1 EDU + \varepsilon, E(\varepsilon|EDU) \neq 0$$

- 由于误以为干扰项与解释变量不相关，我们就“扭曲”了干扰项，进而“扭曲”了 EDU，INC 和干扰项的关系。在判断错误的情况下，三者之间的关系为：

$$INC = \gamma_0 + \gamma_1 EDU + u, E(u|EDU) = 0$$

- 我们将“扭曲”的干扰项 u 称为伪干扰项。

线性回归模型与因果推断

- 通过 CEF 来理解 γ_1 和 β_1 的关系：

$$E(INC|EDU) = \gamma_0 + \gamma_1 EDU$$

- 求偏导：

$$\frac{dE(INC|EDU)}{dEDU} = \gamma_1$$

- γ_1 反映了 INC 的期望如何随 EDU 变化，但并没有控制 IQ 不变。
- 计算 γ_1 和 β_1 的关系：

$$E(INC|EDU) = E(\alpha+\beta_1 EDU+\beta_2 IQ+e|EDU) = \alpha+\beta_1 EDU+\beta_2 E(IQ|EDU$$

- 对 EDU 求偏导并结合 γ_1，可得：

$$\frac{dE(INC|EDU)}{dEDU} = \beta_1 + \beta_2 \frac{dE(IQ|EDU)}{dEDU} = \gamma_1$$

线性回归模型与因果推断

- 假设受教育程度和智商之间存在线性相关关系：

$$E(IQ|EDU) = \phi_0 + \phi_1 EDU$$

  - 求偏导可得：

$$\frac{dE(IQ|EDU)}{dEDU} = \phi_1$$

- 综上，可得：

$$\gamma_1 = \beta_1 + \beta_2 \phi_1 \tag{3}$$

  - 由此可见，γ_1 只是反映了 EDU 和 INC 的相关性，它包含了受教育程度对收入的因果影响 β_1，以及受教育程度与智商的相关性 ϕ_1 乘以智商对收入的因果影响 β_2。

线性回归模型与因果推断

- 在图中可以更直观的看到，在没有控制 IQ 的情况下，EDU 和 INC 之间的相关性（γ_1）包含了因果路径 $(EDU \xrightarrow{\beta_1} INC)$ 和混淆路径（$EDU \xleftarrow{\phi_1} \varepsilon \xrightarrow{\beta_2} INC$）产生的相关性。

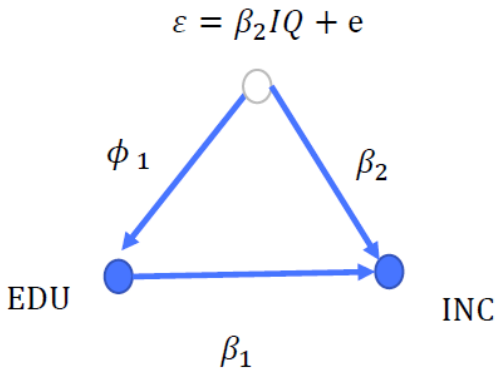


Figure: 干扰项包含混淆变量时的路径图

线性回归模型与因果推断

- 相关关系回归函数并非完全没有价值。如果我们想预测受教育程度相差 ΔEDU 的两组人，平均收入差异多大，在这种情况下，我们只关注预测收入的差别，并不在乎这些差别是否由受教育程度造成的，那么我们可以使用相关回归函数中的 γ_1，得到平均收入差异 $\Delta INC = \gamma_1 \Delta EDU$。从这个意义上来讲，模型本身没有对错之分，只是要清楚表达和认识我们所关注的问题。
- 由于干扰项是无法观测的，一个模型的系数是否反映因果关系在本质上是没法检验的。

多元线性回归 OLS 估计量

- 在实践中，多元线性回归模型的一般形式如下：

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \mu_i$$

- 多元线性回归可以用矩阵形式表示。即：

$$y_i = \sum_{k=0}^{K} \beta_k x_{ik}, x_{i0} = 1$$

x_{ik} 的第一个下标表示个体 i(共有 n 个个体，即样本容量为 n), 而第二个下标表示第 k 个解释变量。

- 乘积之和可以写为两个向量的内积。定义列向量

$$\boldsymbol{x_i} = \begin{pmatrix} 1 & x_{i1} & \cdots & x_{iK} \end{pmatrix}'$$

参数向量

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 & \beta_1 & \cdots & \beta_K \end{pmatrix}'$$

多元线性回归 OLS 估计量

- 则

$$\sum_{k=0}^{K} \beta_k x_{ik} = \boldsymbol{x_i}' \boldsymbol{\beta}$$

- 那么, 多元回归模型可以写成:

$$y_i = \begin{pmatrix} 1 & x_{i1} & \cdots & x_{iK} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} + \mu_i = \boldsymbol{x_i}' \boldsymbol{\beta} + \mu_i$$

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} \boldsymbol{x_1}' \\ \boldsymbol{x_2}' \\ \vdots \\ \boldsymbol{x_n}' \end{pmatrix}}_{x} \boldsymbol{\beta} + \underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}}_{\mu} = \boldsymbol{X\beta} + \mu$$

多元线性回归 OLS 估计量

- $\boldsymbol{X}$ 为数据矩阵。

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1K} \\ 1 & x_{21} & \cdots & x_{2K} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & \cdots & x_{nK} \end{bmatrix}_{n\times(K+1)}$$

- 多元回归的系数估计的原理也是最小化预测误差平方和。即，

$$\min_{\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k} \sum_{i=1}^{n} \mu_i^2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \cdots - \hat{\beta}_k x_{ik}\right)^2 \tag{4}$$

  - 其中，$\hat{\beta}_0$，$\hat{\beta}_1$，...,$\hat{\beta}_k$ 是 OLS 估计量，OLS 残差为 $\hat{\mu}_i = y_i - \hat{y}_i$。

多元线性回归 OLS 估计量

- 在几何上，一元回归寻找最佳拟合的回归直线，使得观测值到该回归直线的距离之平方和最小。二元回归寻找最佳拟合的回归平面。多元回归则寻找最佳拟合的回归超平面。

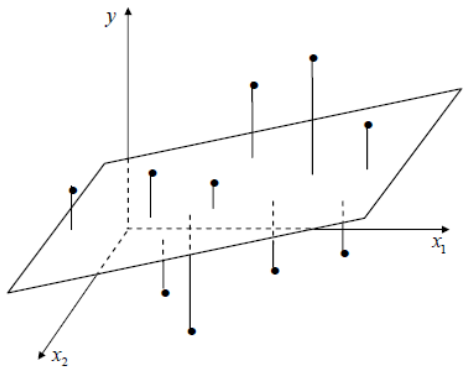


Figure: 二元线性回归几何解释

多元线性回归 OLS 估计量

- 对方程（4）求一阶条件并整理即可得到 OLS 估计量（略，方法同上一章）。
- 多元回归模型的正规方程组可以写成矩阵形式：

$$\underbrace{\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1K} & x_{2K} & \cdots & x_{nK} \end{pmatrix}}_{X'} \underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}}_{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{5}$$

- 上述方程组可以简洁写成：

$$\boldsymbol{X'\mu} = 0 \tag{6}$$

- 可以看出，残差向量与每一个解释变量正交。
- 残差向量可以写为：$\boldsymbol{\mu = y - X\hat{\beta}}$

多元线性回归 OLS 估计量

- 代入方程（6）中，可得

$$\boldsymbol{X}'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})=\mathbf{0} \tag{7}$$

- 乘开并移项可知，最小二乘估计量 $\hat{\beta}$ 满足：$(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}}=\boldsymbol{X}'\boldsymbol{y}$
- 假设 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 存在，求解 OLS 估计量：

$$\hat{\boldsymbol{\beta}}\equiv(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{8}$$

- 同样可以推导可得：$\hat{\boldsymbol{y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}$
- 拟合值向量与残差向量正交（$\hat{\boldsymbol{y}}'\boldsymbol{\mu}=(\boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\mu}=\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{\mu}=\hat{\boldsymbol{\beta}}'\cdot\mathbf{0}=0$）
- 被解释变量 $\boldsymbol{y}$ 可以分解为相互正交的拟合值 $\hat{\boldsymbol{y}}$ 与残差之和，见下图。
  $\boldsymbol{y}=\hat{\boldsymbol{y}}+\boldsymbol{\mu}$（由于 $\boldsymbol{\mu}=\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}=\boldsymbol{y}-\hat{\boldsymbol{y}}$）

多元线性回归 OLS 估计量

- 拟合值 $\hat{\boldsymbol{y}}$ 可视为被解释变量 $\boldsymbol{y}$ 向解释变量超平面 $\boldsymbol{X}$ 的投影（projection）。由于拟合值为解释变量的线性组合，故拟合值向量 $\hat{\boldsymbol{y}}$ 正好在超平面 $\boldsymbol{X}$ 上。根据 OLS 的正交性，残差向量与拟合值向量正交。

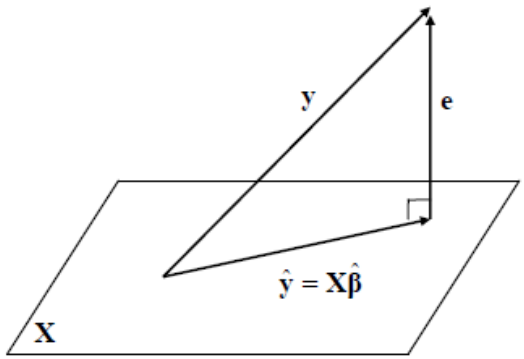


Figure: 最小二乘法的正交性

多元线性回归 OLS 估计量

- **回归标准误**（SER）是误差项 μ_i 的标准差估计。因此，SER 度量了 Y 的分布在回归线周围的离散程度。多元回归 SER 为

$$SER = s_{\hat{\mu}} = \sqrt{s^2_{\hat{\mu}}} = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{\mu}_i^2} = \sqrt{\frac{SSR}{n-k-1}}$$

  - 其中，SSR 为残差平方和。与一元回归模型的 SER 公式差异在分母上，这里是 $n-k-1$，而不是 $n-2$。此处是为了调整估计 $k+1$ 个系数（k 个斜率和一个截距）引起的向下偏误。$n-k-1$ 为自由度。当样本容量很大时，自由度调整可忽略。

多元线性回归 OLS 估计量

- 对于多元回归，在回归方程有常数项的情况下，由于 OLS 的正交性，平方和分解公式依然成立。

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{TSS}=\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{ESS}+\underbrace{\sum_{i=1}^{n}\mu_i^2}_{RSS} \tag{9}$$

- 根据平方和分解公式，可定义**拟合优度**。

$$0\leq R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\sum_{i=1}^{n}\mu_i^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{ESS}{TSS}=1-\frac{RSS}{TSS}\leq 1 \tag{10}$$

- 拟合优度的缺点是，如果增加解释变量的数量，R^2 只增不减。一般情况下增加新的解释变量会降低 RSS，这就意味着增加新的解释变量后拟合优度通常会增大。除非增加的解释变量的系数为 0。

多元线性回归 OLS 估计量

- 由于增加新的解释变量后 R^2 会增大，R^2 增大并不意味着增加一个变量就提高了模型的拟合程度。因此，引入调整拟合优度（或校正拟合优度）。此时，增加新的解释变量后，$\bar{R}^2$ 不一定增大。

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s^2_{\hat{\mu}}}{s^2_y} \tag{11}$$

1. $(n-1)/(n-k-1)$ 总是大于 1，因此调整的 $\bar{R}^2$ 总是小于 R^2。
2. 增加一个解释变量对 $\bar{R}^2$ 有两种相反的作用。一方面，SSR 降低使 $\bar{R}^2$ 增大。另一方面，$(n-1)/(n-k-1)$ 也会增大。所以 $\bar{R}^2$ 是增大还是减少取决于这两种作用的强弱。
3. $\bar{R}^2$ 可以为负数。当增加解释变量，SSR 下降的程度不足以抵补 $(n-1)/(n-k-1)$ 的下降时，$\bar{R}^2$ 就可能为负。

多元线性回归 OLS 估计量

- 虽然 $\bar{R}^2$ 比 R^2 很有用，但是太依赖与 $\bar{R}^2$ 就会掉进陷阱。在实际应用中，“最大化 $\bar{R}^2$”几乎不能回答任何有意义的计量或统计问题。相反，是否要增加一个变量应该基于增加这个变量后是否能让你更好地估计感兴趣的因果效应。

古典线性模型的假定

假设 1：关于参数是线性的（linearity）

- 线性假设的含义是每个解释变量对 y 的边际效应为常数，比如 $\frac{\partial y_i}{\partial x_{i2}} = \beta_2$（忽略扰动项）。
- 如果边际效应可变，可加入平方项或交互项。只要将平方项或交互项也视为解释变量，则仍符合线性模型的假定。
- 线性假定的本质要求是，回归函数是参数 (β_0 β_1 $\cdots$ β_K) 的线性函数。

古典线性模型的假定

假设 2：独立同分布

- $(x_{1i}, x_{2i}, \cdots, x_{ki}, y_i)$ 是独立同分布的随机变量，即它们是 $i.i.d.$ 的。
- 如果数据是通过简单随机抽样收集的，则这个假设自然成立。

古典线性模型的假定设

假设 3：零条件均值

- $\mathrm{E}(\mu_i|\boldsymbol{X}) = \mathrm{E}(\mu_i|x_1, x_2, \cdots, x_n) = 0$。
- 严格外生性意味着，在给定 $\boldsymbol{X}$ 的情况下，扰动项的条件期望为 0。
- 扰动项均值独立与所有解释变量的观测值，而不仅仅是同一观测数据 $\boldsymbol{x_i}$ 中的解释变量。这也就意味着，扰动项与所有个体的解释变量都不相关。
- 此假定很强。
- 均值独立仅要求 $\mathrm{E}(\mu_i|X) = c$，c 为常数，不一定为 0。
- 根据 $\mathrm{E}(\mu_i|X) = 0$，可证明扰动项的无条件期望也为 0。

$$\mathrm{E}(\mu_i) = E_X \underbrace{E(\mu_i|X)}_{=0} = E_X(0) = 0$$

古典线性模型的假定

假设 4：不太可能出现较大异常值

- 同一元回归情形一样。这也提醒我们 OLS 估计量对较大异常值很敏感。

古典线性模型的假定

假设 5：不存在“严重多重共线性”“strict multicollinearity”

- 即数据矩阵 $\boldsymbol{X}$ 满列秩（full column rank）。数据矩阵的各列向量为线性无关，即不存在某个解释变量为另一解释变量的倍数，或可由其他解释变量线性表示的情形
- 如果 $\boldsymbol{X}$ 不是满列秩，则 $(X'X)^{-1}$ 不存在，无法估计 OLS 估计量。

多元线性回归系数的直观理解

- 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。

情形 1：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。
- 回归模型：$INC = \gamma_0 + \gamma_1 EDU + u, E(u|EDU) = 0$。
- 此时 γ_1 反映的是 INC 和 EDU 的相关性，即②和③的信息。
- ①和④是 INC 变化与 EDU 无关的部分，即 INC 对 EDU 回归的残差项的变化。

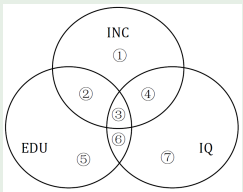


Figure: 情形 1

多元线性回归系数的直观理解

情形 1：INC、EDU 和 IQ 的变化都是相关的

- 回归模型：$INC = \alpha + \beta_1 EDU + \beta_2 IQ + \varepsilon, E(\varepsilon|EDU) = 0$
- 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(反映的是 EDU 和 IQ 共同变化的部分) 被舍去了。这些变化对 INC 的影响究竟是由 EDU 还是由 IQ 造成的，是不确定的。故被舍去。
- ②是 EDU 中与 IQ 不相关，但与 INC 相关的部分。④是 IQ 变化中与 EDU 不相关，但与 INC 相关的部分。
- 此时，回归系数反映了 EDU 和 IQ 各自对 INC 的影响，即因果关系。

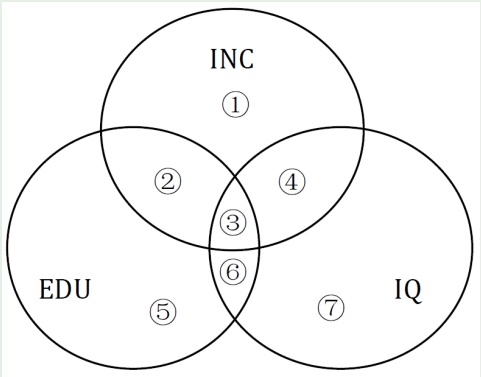


Figure: 情形 1

多元线性回归系数的直观理解

情形 1：INC、EDU 和 IQ 的变化都是相关的

- 根据之前推导的结果，$\gamma_1 = \beta_1 + \beta_2\phi_1$。
- 对应为韦恩图：β_1 对应的是②,$\beta_2\phi_1$ 对应的是③，γ_1 是利用②和③的信息进行估计。
- 另外，我们可以将多元回归分解为两步。第一步将 EDU 对 IQ 回归并取残差项 ν, 这个残差项的信息就是②和]⑤；第二步将 INC 对残差项 ν 回归，这一步使用的是（①,②,③,④和②,⑤）的交集信息，即②。这两个步骤和多元回归来估计 EDU 的系数是一致的。
- 思考：另外一种做法？

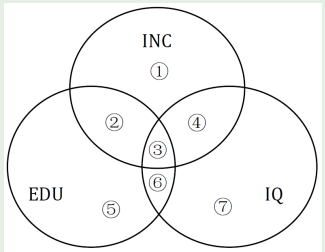


Figure: 情形 1

多元线性回归系数的直观理解

情形 2：EDU 和 IQ 严重共线性

- 在这种情况下，能提供给我们分别估计 EDU 和 IQ 对 INC 影响的信息很少，原因是 EDU 和 IQ 几乎总是一起变动，无法区分是哪个变量在影响 INC。
- 由于两个变量可供估计的独立信息很少，回归结果使得得到的系数 β_1 和 β_2 很难体现出统计显著。
- ③越大，共线性越严重。这将使得几乎没有②和④的信息来估计 β_1 和 β_2。

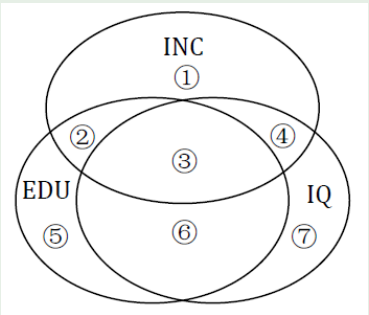


Figure: 情形 2

多元线性回归系数的直观理解

情形 3：EDU 和 IQ 不相关

- 此时，③为空集。EDU 和 IQ 不相关，但它们分别都会影响 INC。
- 在这种情况下，INC 对 EDU 和 IQ 同时进行回归或 INC 只对 EDU 回归，用这两种方法得到的 EDU 系数是一样的。这是因为，这两种情况下都用同样的②的信息来估计 EDU 的影响。
- 当两个解释变量不相关，缺失其中一个并不会影响另一个的回归系数。

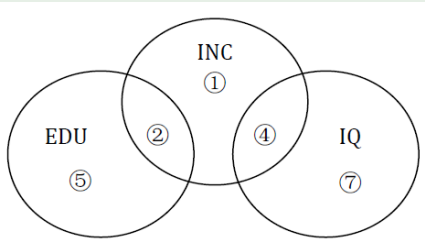


Figure: 情形 3

多元线性回归系数的直观理解

- 我们用韦恩图来解释拟合优度和系数显著性。
- 在情形 1 中，尽管在估计 EDU 和 IQ 系数的过程中舍去了 EDU 和 IQ 共同变化的部分③, 但如果我们想要知道的是 INC 的变化有多少是由 EDU 和 IQ 共同决定的（拟合优度 R^2），我们使用的是②,③,④部分的面积与整个 INC 的面积的比值。
- 计算 R^2 时，我们并不需要去掉③的信息，因为③也反映了 INC 的变化是由 EDU 和 IQ 决定。
- 在情形 2 中，③很大，②和④很大，这种情况会导致 R^2 值很大，但系数却并不显著。

多元线性回归系数的直观理解

- 目前还没有“万能灵药”来决定哪些变量要放进回归模型中，但是选择具体回归方式的出发点是考虑所有可能的遗漏变量偏差来源。这取决于你对实证问题的了解程度，目的是获得感兴趣的因果效应的无偏估计，而不是完全依赖于拟合优度。
  - 控制变量不是我们感兴趣的目标变量，而是为了保持对应因素不变而引入回归模型中，如果忽略会导致所感兴趣变量的因果效应估计遭受遗漏变量偏差。控制变量的 OLS 估计量一般来讲是有偏的，因此并没有因果含义。
- 在实践上，应对遗漏变量偏误的方法有：
  1. 第一步，基准模型应该包含主要的核心解释变量和控制变量，这些变量是根据专业知识和经济理论得到的。
  2. 第二步，如果难以获得经济理论所建议的数据，就需要列出候选的备选设定形式，即备选的回归变量集。如果在备选模型中，核心解释变量的系数估计值与基准模型的结果相近，那就说明基准模型的估计结果是可信的。

假设检验和置信区间

- 单个系数的假设检验和置信区间

- 在一元回归模型中，我们利用样本均值替代相应的期望值得到了 OLS 估计量的方差估计 $\hat{\sigma}^2_{\beta_1}$。在 OLS 假设条件下，由大数定律可知这些样本均值收敛于对应的总体值。$\hat{\sigma}^2_{\beta_1}$ 的平方根就是 $\hat{\beta}_1$ 的标准误 SE（$\hat{\beta}_1$）。这个计算标准误的方法可以推广到多元回归模型。

- 就标准误而言，一元和多元没有什么不同。其关键思想：估计量的大样本正态性及抽样分布标准差具有一致估计量。

- 原假设与备择假设：$H_0 : \beta_j = \beta_{j,0}; H_1 : \beta_j \neq \beta_{j,0}$

- 一元回归假设检验的基本步骤也适用于多元回归。
  第一步，计算 β_j 的标准误，$SE(\beta_j)$;
  第二步，计算 t 统计量，$t = \frac{\hat{\beta}_1 - \beta_{j,0}}{SE(\hat{\beta}_1)}$;
  第三步，计算 p 值,$p = 2\Phi(-|t^{act}|)$。

假设检验和置信区间

- 多元回归中单个系数的置信区间构造方法也与一元回归模型时一样。
- 无论是假设检验的方法还是质询区间的构造方法都依赖于 OLS 估计量分布的大样本正态近似。

- **联合假设**一般是指对两个或两个以上回归系数施加限制的假设。只要原假设中的任何一个系数等式不成立，那么联合原假设就为假。
- 原假设与备择假设：
  $H_0 : \beta_j = 0, \beta_m = 0, \cdots,$ q 个约束;$H_1 : \beta_j \neq 0 / \beta_m \neq 0$

- F **统计量**用于检验回归系数的联合假设。
  - 当 $q = 2$ 个约束的 F 统计量。$F = \frac{1}{2}(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}^2_{t_1,t_2}})$
  - 当有 q 个约束的 F 统计量（详细见斯托克，第五版，p548）。在大样本条件下，原假设成立时有：F 统计量服从 $F_{q,\infty}$ 分布。F 统计量的临界值可以通过查表得到。

假设检验和置信区间

- Stata 等软件可以直接帮我们计算出结果。但是大多数软件的默认状态为计算同方差标准误。因此，需要你选择“稳健”选项后才会利用异方差稳健标准误计算 F 统计量。
- 利用 F 统计量分布的大样本 $F_{q,\infty}$ 近似计算其 p 值。

$$p_{-value} = \Pr[F_{q,\infty} > F^{act}]$$

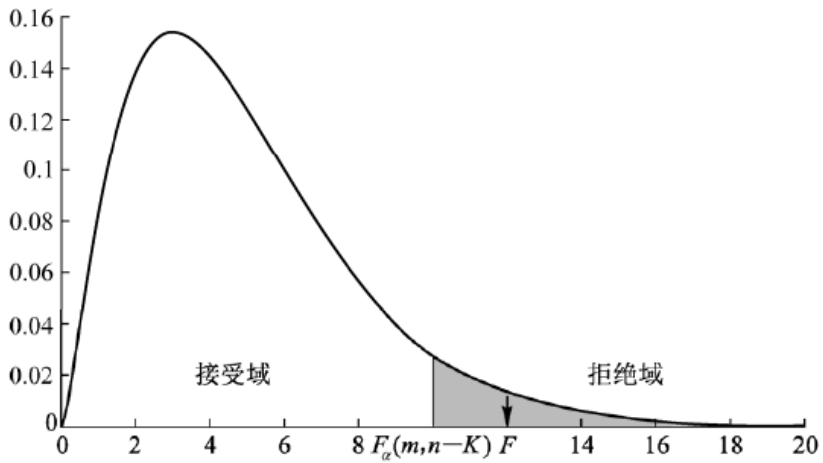


Figure: F 检验

小样本 OLS 和大样本 OLS

- 在经典线性回归模型的假设条件下，OLS 估计量具有以下性质（证略）：

- 1. 线性性：OLS 估计量 $\hat{\beta}$ 为线性估计量。
- 2. 无偏性：$E(\hat{\beta}|X) = \beta$, 即 $\hat{\beta}$ 不会系统地高估或低估 β 。
- 3. 估计量 $\hat{\beta}$ 的方差为

$$
\begin{aligned}
&Var(\hat{\beta}|X) = Var(\hat{\beta} - \beta|X) = Var(A\varepsilon|X) \\
&= AVar(\varepsilon|X)A' = A\sigma^2 I_n A' \\
&= \sigma^2 AA' = \sigma^2 (X'X)^{-1} X'X (X'X)^{-1} = \sigma^2 (X'X)^{-1}
\end{aligned}
$$

这里，β 为常数，$A \equiv (X'X)^{-1}X'$，$Var(\varepsilon|X) = \sigma^2 I_n$(球形扰动项：扰动项满足“同方差”、“无自相关”的性质)。

-

$$Var(\varepsilon|X) = \sigma^2 I_n = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$$

如果协方差矩阵 $Var(\varepsilon|X)$ 的主对角元素都等于 σ^2, 即满足“同方差”。如果不完全相等，则存在“条件异方差”。协方差矩阵的非对角元素都为 0，则不同个体的扰动项之间无“自相关”或“序列相关”，反之，则存在自相关。

- 4. 高斯 -马尔科夫定理（Gauss-Markov Theorem）: 最小二乘法是最佳线性无偏估计（BLUE）, 即在所有线性的无偏估计中，最小二乘法的方差最小。如果不满足球形扰动项，则高斯 -马尔科夫定理不成立。
- 5. 方差的无偏估计：$E(s^2|X) = \sigma^2$

小样本 OLS 和大样本 OLS

- **大样本理论**也称“渐进理论”（asymptotic theeory），研究的是当样本容量趋向无穷大时统计量的性质。

- （1）小样本理论的假设过强。首先，小样本理论的严格外生性假设要求解释变量与所有扰动项正交（不相关）。但是在时间序列模型中，这意味着解释变量与扰动项的过去、现在与未来值全部正交！大样本理论只要求解释变量与同期（同方程）的扰动项不相关。另外，小样本理论假定扰动项与为正态分布，而大样本理论无此限制。现实是很多经济变量的分布并不是正态分布。

- （2）在小样本理论的框架下，须研究统计量的精确分布，但常难以推导。大样本理论只要研究统计量的渐进分布就可以了，而渐进分布比较容易推导（可用大数定律和中心极限定理）。

- （3）使用大样本理论的代价是要求样本容量大，一般认为至少样本大于等于 30，最好在 100 以上。由于现代的数据集越来越大，经常上千上万。

小样本 OLS 和大样本 OLS

- 大样本理论的一些概念：随机序列的收敛；以概率收敛；以分布收敛；大数定律；中心极限定理；渐进正态分布；渐进方差；渐进有效；随机过程；遍历性；渐进独立定理；鞅差分序列的中心极限定理，等等。（详细见陈强，高级计量经济学及 Stata 应用，p47-p57）
- 大样本 OLS 假定；OLS 的大样本性质（详细见陈强，高级计量经济学及 Stata 应用，p57-p61）

参考文献：

Stock and Watson, Introduction to Econometrics, 3rd edition, 2011.

陈强. 计量经济学及 Stata 应用 [M]. 高等教育出版社，2015.