

内生性问题

司继春

上海对外经贸大学

2025年8月

内生性问题

- 在线性回归中，我们假设了 $\mathbb{E}(u|x) = 0$ ，从而得到了最小二乘估计。
- 当这一假定被违背时，即 $\mathbb{E}(u|x) \neq 0$ ，我们称为有内生性问题。
 - ① 微观、宏观中的「内生变量」与计量中的内生性的联系区别？
 - ② 内生性可能的原因：
 - 遗漏变量
 - 互为因果
 - 度量误差
 - 自选择
 -

遗漏变量

如果真实的数据生成过程为：

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \gamma q_i + v_i$$

而变量 q_i 是观测不到的，并且假设 q_i 与 x_i 之间存在着相关性：

$$q_i = \delta_0 + \delta_1 x_{1i} + \cdots + \delta_K x_{Ki} + e_i$$

那么将以上方程带入结构式，得到：

$$y_i = (\beta_0 + \gamma\delta_0) + (\beta_1 + \gamma\delta_1) x_{1i} + \cdots + (\beta_K + \gamma\delta_K) x_{Ki} + \gamma e_i + v_i$$

如果令 $u_i = \gamma e_i + v_i$ ，我们有 $\mathbb{E}(u_i | x_i) = 0$ ，因而那么如果我们忽略了变量 q_i ，那么实际得到的回归系数为

$$\beta_k^* = \beta_k + \gamma\delta_k$$

存在着偏误。

互为因果

如果存在互为因果的两个内生变量 y_1, y_2 ，结构方程：

$$y_1 = \alpha_1 y_2 + x_1 \beta_1 + u_1$$

$$y_2 = \alpha_2 y_1 + x_2 \beta_2 + u_2$$

联立两个方程，可以得到：

$$\begin{aligned} y_2 &= \alpha_2 (\alpha_1 y_2 + x_1 \beta_1 + u_1) + x_2 \beta_2 + u_2 \\ &= \alpha_1 \alpha_2 y_2 + \alpha_2 x_1 \beta_1 + x_2 \beta_2 + \alpha_2 u_1 + u_2 \\ &= \frac{\alpha_2}{1 - \alpha_1 \alpha_2} x_1 \beta_1 + \frac{1}{1 - \alpha_1 \alpha_2} x_2 \beta_2 + \frac{\alpha_2 u_1 + u_2}{1 - \alpha_1 \alpha_2} \\ &\triangleq x' \delta + v \end{aligned}$$

使用决定 y_2 、但同时不决定 y_1 的外生的扰动，即包含在 x_2 而不包含在 x_1 的变量(z)进行识别。

互为因果

供给与需求曲线

如果我们希望估计某农产品的需求曲线，假设 y_i 为成交量， x_i 为农产品的价格。假设农产品的需求曲线为

$$y_i^d = \beta_0 + \beta x_i + u_i$$

供给曲线为

$$y_i^s = \delta_0 + \delta x_i + v_i$$

均衡的成交量和价格应该使得供给需求相等，即

$$\beta_0 + \beta x_i + u_i = \delta_0 + \delta x_i + v_i$$

解得均衡的价格为

$$x_i = \frac{\delta_0 - \beta_0 + v_i - u_i}{\beta - \delta}$$

注意到 $\mathbb{C}(x_i, u_i) \neq 0$ （同时 $\mathbb{C}(x_i, v_i) \neq 0$ ），因而如果我们使用 y_i 对 x_i 做回归，并不能得到 β （或者 δ ）的一致估计。

度量误差

考虑一个一元线性回归，假设数据的真实生成过程为：

$$y_i = \beta_0^* + \beta^* x_i^* + v_i$$

其中 x_i^* 为真实值。

- 观察不到 x_i^*
- 只能观察到有误差的 $x_i = x_i^* + e_i$

如果我们直接用 y_i 对 x_i 做回归，即：

$$y_i = \beta_0 + \beta x_i + u_i$$

那么 $u_i = v_i - \beta e_i$ 。如果假设 $\mathbb{E}(e_i|x_i^*) = 0$ ，那么

$$\mathbb{C}(x_i, e_i) = \mathbb{C}(x_i^* + e_i, e_i) = \mathbb{V}(e_i)$$

因而

$$\mathbb{C}(x_i, u_i) = \mathbb{C}(x_i, v_i - \beta e_i) = -\beta \mathbb{V}(e_i) \neq 0$$

因而导致了内生性问题。

度量误差

如果我们直接使用带有度量误差的 x_i 进行回归，那么我们将得到：

$$\begin{aligned}\text{plim}\hat{\beta} &= \frac{\mathbb{C}(x_i, y_i)}{\mathbb{V}(x_i)} \\ &= \frac{\mathbb{C}(x_i^* + e_i, \beta_0^* + \beta^* x_i^* + v_i)}{\mathbb{V}(x_i^*) + \mathbb{V}(e_i)} \\ &= \beta^* \cdot \frac{\mathbb{V}(x_i^*)}{\mathbb{V}(x_i^*) + \mathbb{V}(e_i)}\end{aligned}$$

得到的 $|\hat{\beta}| < |\beta^*|$ ，即估计的系数的绝对值总是小于真实值的绝对值，存在着向中性偏误（attenuation bias）。

自选择

- 在经济学中，需要研究的很多变量通常并不是随机给定，而是行为人为最大化效用进而做出选择的结果。
- 当存在这种自选择（self-selection）问题时，如果将这些行为人自己选择的变量作为自变量，经常会存在内生性问题。

自选择

教育回报

如果我们关心是否上大学对未来收入的影响，记 $w_i = 1$ 为上过大学， $w_i = 0$ 为没上过大学，记 $y_i(0)$ 为假设该个体没上过大学的收入， $y_i(1)$ 为假设该个体上过大学的收入，且数据生成过程为：

$$y_i(1) = \gamma + x_i'\beta + u_{1i}$$

$$y_i(0) = x_i'\beta + u_{0i}$$

那么观察到的收入为：

$$\begin{aligned} y_i &= w_i y_i(1) + (1 - w_i) y_i(0) \\ &= \gamma \cdot w_i + x_i'\beta + w_i u_{1i} + (1 - w_i) u_{0i} \\ &\triangleq \gamma \cdot w_i + x_i'\beta + v_i \end{aligned}$$

自选择

教育回报

若 w_i 是完全随机分配的, 即 $w_i \perp\!\!\!\perp (u_{1i}, u_{0i})$, 那么:

$$\begin{aligned}\mathbb{C}(w_i, v_i) &= \mathbb{C}(w_i, w_i u_{1i} + (1 - w_i) u_{0i}) \\ &= \mathbb{C}(w_i, w_i u_{1i}) + \mathbb{C}(w_i, (1 - w_i) u_{0i}) \\ &= \mathbb{E}(w_i^2 u_{1i}) - \mathbb{E}(w_i) \mathbb{E}(w_i u_{1i}) - \mathbb{E}(w_i (1 - w_i) u_{0i}) \\ &= 0\end{aligned}$$

然而, 如果 w_i 不是随机分配的, 比如个体通过如下过程选择是否上大学:

$$w_i = \mathbb{1}\{y_i(1) \geq y_i(0)\} = \mathbb{1}\{\gamma + u_{1i} \geq u_{0i}\}$$

那么 w_i 与 (u_{1i}, u_{0i}) 不可能独立, 因而 $\mathbb{C}(w_i, v_i) \neq 0$ 。