内生性问题

司继春

上海对外经贸大学

2025年6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。

评估遗漏变量的偏误

- 根据以上分析，遗漏变量的误差取决于遗漏变量对于y的影响大小，以及遗漏变量与解释变量之间的相关性。
- 一般而言，以上两种相关性并不容易观测，因而评估遗漏变量对系数造成的偏误大小是比较困难的。
- 一种做法是可以根据可观测变量对可能的由不可观测变量导致的偏误进行评估，比如在加入和不加入控制变量两种情况下，核心解释变量的系数变化可以部分代表估计量的稳定性。
- 这一做法并不严谨，更严谨的推论通常需要更多的假设
- 一般而言，仅仅通过比较加入控制变量前后系数的变化并不足以验证遗漏变量所可能带来偏误的大小。Oster（2019）指出使用以上方法需要额外考虑R^2。

评估遗漏变量的偏误

- 为了说明这一点，我们考虑如下例子：

$$Y = \beta W + X + C$$

- 我们不妨将Y解释为收入，W为教育，而X和C分别为两种不同的能力。简单起见，假设$\beta = 0$。由于W与X和C存在相关性，所以如果我们考虑如下两种回归：
  - 直接使用Y对W回归，得到$\mathring{\beta}$；
  - 使用Y对W和X做回归，得到$\hat{\beta}$
- 那么以上两个估计量都会存在偏差。
- 一种想法是，可以通过比较$\hat{\beta}$和$\mathring{\beta}$，如果系数变化不大，那么意味着结果可能对遗漏变量是相对稳健的。

评估遗漏变量的偏误

- 然而Oster（2019）指出，这种做法是有问题的，关键在于$\mathbb{V}(X)$和$\mathbb{V}(C)$的大小。
- 如果可观测的X的方差远小于不可观测的C的方差，从而加入X的回归的R^2并不会比没有加入时增加太多，此时比较$\hat{\beta}$和$\mathring{\beta}$会发现系数是稳定的，并不会有太大的变化。
- 然而，$\hat{\beta}$和$\mathring{\beta}$实际上都存在着遗漏变量带来的偏差，从而比较$\hat{\beta}$和$\mathring{\beta}$发现系数稳定，如果不考虑R^2的变化，很有可能得到的结论是误导性的。
- Oster（2019）使用模拟的办法展示了这一问题

评估遗漏变量的偏误

- Oster（2019）考虑了更一般的设定，对于模型：

$$y_i = \beta w_i + x_i'\psi + q_i + u_i$$

其中w_i为核心解释变量，x_i为一系列控制变量，而q_i为不可观测的部分

- 记$r_i = x_i'\psi$为可观测部分，不失一般性假设q_i与x_i的每一部分均不相关（即可以将q_i看做是遗漏变量中与x_i无关的部分）。
- 定义

$$\delta = \frac{\sigma_{qw}/\sigma_q^2}{\sigma_{rw}/\sigma_r^2}$$

其中$\sigma_{rw} = \mathbb{C}(r_i, w_i), \sigma_{qw} = \mathbb{C}(q_i, w_i), \sigma_r^2 = \mathbb{V}(r_i), \sigma_q^2 = \mathbb{V}(q_i)$，因此$\delta$可以被解释为不可观测变量中与$w$相关的比例与可观测控制中与$w$相关的比例的比值。

评估遗漏变量的偏误

- 现在，如果记w对x的回归系数为η_j，如果额外假设：

$$\frac{\eta_j}{\eta_{j'}} = \frac{\psi_j}{\psi_{j'}} \tag{1}$$

- Oster（2019）得到估计量：

$$\beta^* \approx \hat{\beta} - \delta \left(\mathring{\beta} - \hat{\beta} \right) \frac{R^2_{\max} - R^2}{R^2 - \mathring{R}^2}$$

其中$R^2_{\max}$为理论上可以达到的最大的R^2，$\mathring{R}^2$为使用y对w回归的R^2，而R^2为加入x作为控制变量后的R^2

- 可以得到β^*为一致估计量：

$$\beta^* \xrightarrow{p} \beta$$

评估遗漏变量的偏误

- 由此可见，$\hat{\beta}$的偏差不仅仅取决于$\mathring{\beta} - \hat{\beta}$，还取决于三个不同的$R^2$：
  - $R^2_{\max} - R^2$可以被看做是不可观测部分的部分
  - 而$R^2 - \hat{R}^2$为可观测的控制变量部分
  - 如果不可观测的部分方差远远小于可观测的控制变量部分的方差，那么遗漏变量的偏差相应的可能会比较小。

评估遗漏变量的偏误

- 注意到如果我们知道了δ和$R^2_{\max}$，我们就可以直接计算出一致估计量β^*。
- 而如果放弃假设式(1)，也有类似的结论。
- Oster（2019）建议可以通过两种办法评估遗漏变量所带来的的偏误可能的大小：
  - 给定$\delta, R^2_{\max}$，直接计算出β^*，如果β^*与$\hat{\beta}$的符号相反，意味着$\hat{\beta}$可能不够稳健
  - 计算$\beta^* = 0$时所对应的δ，δ越大则$\hat{\beta}$更加稳健

评估遗漏变量的偏误

- 然而为了使用以上方法，我们必须给出δ和$R^2_{\max}$的具体取值。
  - 对于δ，Oster（2019）建议$\delta = 1$是一个通常来讲比较合适的选择；
  - 而对于$R^2_{\max}$，虽然由于y可能存在天然的误差等原因，实际上能够达到的最大的R^2多数情况下应该是小于1的，不过设定$R^2_{\max} = 1$给出了一个更加稳健的结果。
- 当然，实际应用时也可以根据现实情况对δ和$R^2_{\max}$做出相应调整。

评估遗漏变量的偏误

教育汇报的遗漏变量问题

我们使用WAGE2.dta，使用Mincer方程估计教育回报问题：

```
use datasets/WAGE2.dta
gen exper2=exper^2
reg lwage educ exper exper2, r
```

为了评估遗漏变量可能带来的偏误，可以使用psacalc命令

```
psacalc beta educ
```

评估遗漏变量的偏误

教育汇报的遗漏变量问题

其中beta选项代表需要计算R^*，而educ表示核心解释变量；也可以使用delta和rmax选项修改δ和$R^2_{\max}$的默认值：

```
psacalc beta educ, rmax(0.8) delta(2)
```

如果需要计算\delta，可以使用：

```
psacalc delta educ
```

此外，mcontrol选项可以用于将一些变量在估计$\hat{\beta}$和$\mathring{\beta}$时都用作控制变量，比如我们将变量exper及exper2在所有回归中都控制，仅仅比较IQ加入之前和之后的结果差异：

```
reg lwage educ exper exper2 IQ, r
psacalc beta educ, mcontrol(exper exper2)
```

评估遗漏变量的偏误

教育汇报的遗漏变量问题

- 值得注意的是，β^*的计算是通过解一个一元三次方程组实现的，因而可能出现多组解的情况，该命令默认使用与现有估计差距最小的作为beta汇报，同时也会汇报其他的解，在使用时可能需要甄别。
- 如果需要使用reghdfe，可以使用psacalc2命令：https://github.com/ArthurHowardMorris/psacalc_supports_reghdfe

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：

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

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$\beta$（或者$\delta$）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$\left|\hat{\beta}\right| < |\beta^*|$，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：

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

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：

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

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$\mathbb{C}(w_i, v_i) \neq 0$。