治理技术专题

定量政治分析方法

Quantitative Analysis II

苏 毓 淞

清华大学社会科学院政治学系

第三讲 内生性问题 Endogeneity

内生性 (Endogeneity)

- 在一个回归方程式中会存在两种类型的变量：内生型和外生型变量。
- **内生型变量** (Endogenous Variables)：以其他变量为函数的变量。例如：因变量即是以自变量为函数的变量。
- **外生型变量** (Exdogenous Variables)：不以其他变量为函数的变量。例如：通常的自变量属之。
- **滞后变量** (lagged variable) 通常应该会是外生型变量，但是如果存在序相关(serial correlation) 时，则还是内生型变量。
  - 序相关：$y_t = \rho y_{t-1}$

内生性 (Endogeneity)

- 我们的世界不仅仅由一个方程式所解释，是由无数的方程式所构成的。
- 在联立方程式 (Simultaneous Equations) 中，往往存在内生性的问题，相同的变量出现在一个以上的联立方程式中。
- 经典案例：供给、需求曲线。

$$\begin{cases} Q_S = \alpha_0 + \alpha_1 P + u \\ Q_D = \beta_0 + \beta_1 P + v \end{cases}$$

从联立方程式推导内生性

- 如果自变量存在内生性的问题，它就会与回归模型的余数项相关，这违反了线性回归的假设，据此而得的估计值是有偏的，例如以下的联立方程式 (1) 和 (2)：

$$\begin{cases} Y = \alpha_0 + \alpha_1 X + \alpha_2 W_1 + u \\ X = \beta_0 + \beta_1 Y + \beta_2 W_2 + \beta_3 W_3 + v \end{cases} \tag{1}$$

- 将式 (1) 第一行代入第二行得到式 (2)：

$$X = \beta_0 + \beta_1 (\alpha_0 + \alpha_1 X + \alpha_2 W_1 + u) + \beta_2 W_2 + \beta_3 W_3 + v \tag{2}$$

- 从式 (2) 得出，X 是以 u 和其他变量为函数的变量，$\text{cov}(X, u) \neq 0$，所以式 (1) 得出的 α_1 一定是有偏的。
- 这就是因变量和自变量互为因果，因果倒置的问题。

如何找寻工具变量？

- 想象力很重要！
- 工具变量 Z 之于果变量 Y 应为外生：

$$\text{cor}(Z, Y) = 0$$

$$\text{cor}(Z, u) = 0 \qquad \text{[更为正确的表达]}$$

- 工具变量 Z 之于内生变量 X 应为内生：

$$\text{cor}(Z, X) \neq 0$$

- 思考：如果 $\text{cor}(Y, X) \neq 0$，而 $\text{cor}(Z, X) \neq 0$，那 $\text{cor}(Z, Y) = 0$ 怎么可能存在？
- 所以我们最幸运 (at best) 的情况下是找到与 Y **弱相关**的工具变量 Z，然后保证 $\text{cor}(Z, u) = 0$ 即可 (u 是估计 Y 方程的余数)，正所谓：千军易得（解释变量），一将难求（工具变量）。

两阶段最小二乘法 (Two-Stage Least Square, 2SLS)

- 联立方程：

$$\begin{cases} Y = \alpha_0 + \alpha_1 X + \alpha_2 W_1 + u \\ X = \beta_0 + \beta_1 Y + \beta_2 W_2 + \beta_3 W_3 + v \end{cases}$$

- 两阶段最小二乘法基本原理：
  1. 使用工具变量 Z 估计一个干净、新的内生变量 $\widehat{X}$:

$$X = \beta_0 + \beta_1 W_2 + \beta_2 W_3 + \beta_3 Z + v$$
$$X = \widehat{X} + \hat{v}$$

  2. 使用新产生的变量 $\widehat{X}$，重新估计 X 和 Y 的关系：

$$Y = \alpha_0 + {\alpha_1}'(\widehat{X} + \hat{v}) + \alpha_2 W_1 + u$$
$$Y = \alpha_0 + {\alpha_1}'\widehat{X} + \alpha_2 W_1 + (u + {\alpha_1}'\hat{v})$$

两阶段最小二乘法 (Two-Stage Least Square, 2SLS)

- 当样本量大时，$\alpha_1{}'$ 的估计值具有一致性 (consistent)。
- 但是 $\alpha_1{}'$ 并不是无偏的，不过当样本量大时，它的偏差是可忽略的，它与 OLS 估计值是一致的。
- 以上使用 OLS 估计得的 $\alpha_1{}'$ 仍然需要修正标准误。
- STATA 提供的程序（`ivreg`）会自动修正标准误。
- R 软件 `systemfit` 包里的 `systemfit()` 函数也会自动修正标准误。

识别性问题 (identification problem)

- 使用 2SLS，必须满足识别条件：
  - 不可识别：工具变量个数少于内生解释变量。
  - 恰好识别：工具变量个数等于内生解释变量。
  - 过度识别：工具变量个数大于内生解释变量。

识别性问题一般性判别方法

- 内生性与外生性个数算法：
  - $M =$ 模型中内生性解释变量的个数。
  - $m =$ 单一方程式中内生性解释变量的个数。
  - $K =$ 模型中外生性变量的个数。
  - $k =$ 单一方程式中外生性解释变量的个数。
- 在一个 M 次联立方程式模型中，某个方程式可识别的最低条件是模型必须存在 $M-1$ 个变量 (无论是外生或内生)。
- 在一个 M 次联立方程式模型中，要识别某个方程式，被排除的外生变量个数不得少于该方程式中内生变量的个数 $m-1$：

$$K-k \geq m-1$$

识别性问题 (identification problem)

- 使用 2SLS，必须满足识别条件：
  - 不可识别：$K - k < m - 1$
  - 恰好识别：$K - k = m - 1$
  - 过度识别：$K - k \geq m - 1$

检验 Z 与 X 的内生性

- 也就是检验 Z 工具变量的有效性，无法证明无效，但可以证明它有效。
  - H_0:Z 为弱工具变量 (Z can only weakly predict X)
  - H_a:Z 不是弱工具变量（不能说它就是强工具变量，只能说 Z can well predict X)
- Z 之于 X 如果是弱工具变量，则可能解决不了太大的（内生性）问题。
- 检验方法如下：
  - 找寻工具变量 Z，用以估计干净的 X，称之为 $\widehat{X}$。
  - 将联立方程式 (1) 中加入 $\widehat{X}$，也就是把 Y 同时对 X 和 $\widehat{X}$ 进行回归。
  - 如果此时的 $\widehat{X}$ 的回归系数仍然是统计显著的，或者 $F > 10$，则可以拒绝 Z 为弱工具变量的原假设。

检验 Z 的外生性 (Exclusion restriction)

- 工具变量方法最为关键的检验！
  - 进行 2SLS 回归分析，计算余数 (残差)ϵ。
  - 将余数 ϵ 与其他外生性变量进行回归，并获得 R^2。
  - 计算检验统计量 nR^2，n 是样本量。
  - 使用 χ^2 分布检验 nR^2 统计量，自由度为工具变量数 − 被工具变量解释的内生变量数。
  - 如果 nR^2 数值大则拒绝 Z 是外生性的假设。
  - 如果模型是恰好识别，因为自由度为 0，无法使用这个检验，所以必须在过度识别才能使用这个检验方法。
  - H_0：所有的工具变量都是外生的。
  - 如果拒绝了 H_0 则表示至少有一个工具变量不是外生的。
- 在这个检验，如果 Z 是外生的，我们估算的 nR^2 会趋近于 0，也就是无法拒绝 H_0。

检验 X 的内生性 (Hausman Test)

- 使用工具变量是因为 X 之于 Y 是内生的，如果不是，就没必要使用工具变量了，这也要检验（莫名其妙，必须找到解药，才知道是不是中毒！)。
- 检验的原假设 H_0：所有的解释变量均为外生变量。如果 H_0 成立，则 OLS 与 2SLS 的估计值都是一致的（注意：2SLS 的估计仍然不是无偏的)。
- 检验统计量 $\left(\hat{\beta}_{\text{2SLS}} - \hat{\beta}_{\text{OLS}}\right) \sim \chi^2(m)$
- STATA 命令：`hausman`
- 在这个检验，如果 X 之于 Y 存在内生性，则这个检验会统计显著，拒绝 H_0，说明 IV 回归与原来的回归显著不同，原来的方程的确有内生性问题导致的估计偏误。
- 这个检验的前提必须是工具变量是“对”的。相当于吃解药去证明有没有中毒。所以所以工具变量的选择要进行一系列的论证，主要靠“说故事”来使别人信服。

检验 X 的内生性 (Hausman Test)

$$\left(\hat{\beta}_{\text{2SLS}} - \hat{\beta}_{\text{OLS}}\right)' \times \left[\text{var}\left(\hat{\beta}_{\text{2SLS}}\right) - \text{var}\left(\hat{\beta}_{\text{OLS}}\right)\right]^{-1} \times \left(\hat{\beta}_{\text{2SLS}} - \hat{\beta}_{\text{OLS}}\right)$$