工具变量真的能识别因果吗

更新：November 6, 2023

1 IV、2SLS 和 GMM 的简单回顾

1.1 工具变量法

考虑简单线性回归模型

$$Y_i = \alpha + \beta D_i + \varepsilon_i$$

其中 Y_i 为结果变量, D_i 为二元解释变量, ε_i 为随机扰动项, 假设可观测的 $\{Y_i, D_i\}_{i=1}^N$ 是 i.i.d. 随机样本. 根据计量经济学的内容可知 β 的 OLS 估计量为

$$\hat{\beta} = \frac{\sum_{i=1}^N (D_i - \overline{D})(Y_i - \overline{Y})}{\sum_{i=1}^N (D_i - \overline{D})^2}$$

当样本容量 $N \to \infty$ 时有

$$\operatorname*{plim}_{N\to\infty} \hat{\beta} = \beta + \frac{\operatorname{cov}(D_i, \varepsilon_i)}{\operatorname{var}(D_i)}$$

当 $\operatorname{cov}(D_i, \varepsilon_i) \neq 0$, 也即 D_i 和 ε_i 存在相关性时, $\hat{\beta}$ 不是 β 的一致估计量. 一般地, 对于线性回归模型

$$Y_i = \alpha + \beta D_i + X_i'\gamma + \varepsilon_i$$

这里的 X_i 为 $K\times 1$ 维列向量, 只要 $\mathbb{E}[\varepsilon_i|D_i,X_i]\neq 0$,[1] 则 β 和 γ 的 OLS 估计量不是一致的, 此时模型存在**内生性问题** (endogeneity problem), 导致模型存在内生性的解释变量称为内生变量. 通常而言, 内生性问题主要来源于以下三个方面: **遗漏变量** (omitted variable), **测量误差** (measurement error) 和**联立方程** (simultaneous equations).

为了解决内生性问题, 需要使用**工具变量** (Instrumental Variable, IV), 它是与内生变量 D_i 相关但和随机扰动项 ε_i 无关的解释变量, 也即工具变量 Z_i 满足

$$\operatorname{cov}(Z_i,D_i)\neq 0,\quad \operatorname{cov}(Z_i,\varepsilon_i)=0$$

根据这两个条件可知

$$\begin{aligned}\operatorname{cov}(Z_i,Y_i)&=\operatorname{cov}(Z_i,\alpha+\beta D_i+\varepsilon_i)\\&=\beta\operatorname{cov}(Z_i,D_i)+\operatorname{cov}(Z_i,\varepsilon_i)=\beta\operatorname{cov}(Z_i,D_i)\end{aligned}$$

也即 $\beta=\operatorname{cov}(Y_i,Z_i)/\operatorname{cov}(D_i,Z_i)$, 此时可以得到 IV 估计量

$$\hat{\beta}_{\mathrm{IV}}=\frac{\sum_{i=1}^N(Z_i-\overline{Z})(Y_i-\overline{Y})}{\sum_{i=1}^N(Z_i-\overline{Z})(D_i-\overline{D})}$$

[1] 事实上, 这一条件可以弱化为 D_i 或 X_i 与 ε_i 存在相关性.

在正则条件成立的条件下, 当 $N \to \infty$ 时有 $\sqrt{N}(\hat{\beta}_{\mathrm{IV}} - \beta) \xrightarrow{d} N(0, \sigma^2)$. 对于更一般的线性回归模型

$$Y_i = \alpha + \beta D_i + X_i'\gamma + \varepsilon_i$$

假设只有 D_i 是内生解释变量, 并且 M_i 是 D_i 的工具变量, 此时可以定义 $K+1$ 维列向量 $Z_i = [M_i, X_i']'$ 与 $W_i = [D_i, X_i']'$, 得到 IV 估计量为

$$\begin{pmatrix} \hat{\beta}_{\mathrm{IV}} \\ \hat{\gamma} \end{pmatrix} = \left(\sum_{i=1}^{N} Z_i W_i' \right)^{-1} \left(\sum_{i=1}^{N} Z_i Y_i \right)$$

它在正则条件下仍然服从渐近正态分布.

更一般地, 如果有多个内生变量, 那么 IV 估计量适用于恰好识别 (just identified) 情况, 也即工具变量和内生变量个数相等. 如果内生变量个数大于工具变量个数, 此时模型不可识别 (unidentified), 如果内生变量个数小于工具变量个数, 则称模型过度识别 (over identified).

1.2 二阶段最小二乘法

仍然考虑模型

$$Y_i = \alpha + \beta D_i + X_i'\gamma + \varepsilon_i$$

假设有 L 个 D_i 的工具变量 $M_{1i}, \cdots, M_{Li}$ 并且 $L > 1$, 此时无法直接求得 β 的 IV 估计量, 因为矩阵 $\sum Z_i W_i'$ 不可逆. 为了得到 β 的一致估计量, 一种做法是去掉多余的工具变量使得恰好识别条件成立, 但这样会损失信息, 更好的做法是使用二阶段最小二乘法 (2SLS).