2014.5.13

第二十四讲 工具变量法

内生性

模型 $y = a + bx + \varepsilon,$

内生性(endogeneity)：ε与x不独立

外生性(exogeneity)： ε与x不独立

例1. 回归方程中没有控制相关变量/丢失了某些变量

假设正确模型为

$$y = a + bx + cz + \delta, \delta \perp x;$$

其中 z 与 x 相关。如果我们没有测量,或者没有在上述模型中控制z，那么工作模型：

$$y = a + bx + \varepsilon,$$

中 $\varepsilon = cz + \delta$ 与 x 不独立（假设E(z) = 0）。

例2. 响应变量是某个或某些自变量的原因(reverse causation)。

假设正确模型为

$$y = a + bx + \varepsilon \ , \varepsilon \ \perp x;$$

但实际操作中工作模型取为:

$$x = c + dy + \delta$$

从正确模型我们知道

$$x = -a/b - y/b - \varepsilon/b,$$

所以工作模型中 $\delta = -\varepsilon/b$与y有关。

例3. 自变量带误差模型 (Error in Variable，EV模型)

正确模型： $y = a + bx_0 + \varepsilon_0$， $x_0 \perp \varepsilon_0$,

假设 对于x_0的测量有误差，即我们只能测量到

$$x = x_0 + \delta, \text{其中} x_0 \perp \delta.$$

则

$$y = a + b(x - \delta) + \varepsilon_0 = a + bx + (\varepsilon_0 - b\delta) \hat{=} a + bx + \varepsilon$$

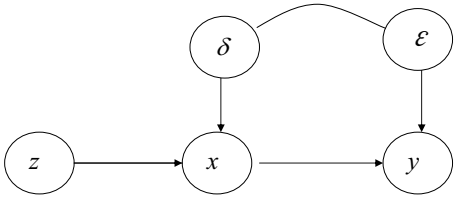
显然x与$\varepsilon = \varepsilon_0 - b\delta$相关.

例4. 结构方程模型 (structural equation models, SEM)

假设z以如下形式影响x ： $x = cz + \delta,\ \delta \sim (0, \tau^2)$

假设x以如下形式影响 y： $y = bx + \varepsilon,\ \varepsilon \sim (0, \sigma^2)$

但δ, ε不独立,因此x与ε不独立. z与(δ, ε)独立。目标是估计b

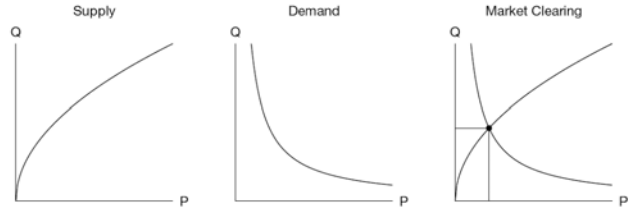


如果对模型: $y_i = bx_i + \varepsilon_i, E(\varepsilon_i) = 0, \text{var}(\varepsilon_i) = \sigma^2$

应用OLS, 得到的LS估计$b_{OLS} = \dfrac{\sum x_i y_i}{\sum x_i^2}$是有偏的.

例5. 联立方程模型(simultaneous equations model)

商品供、求的数量和价格分别为Q和P，

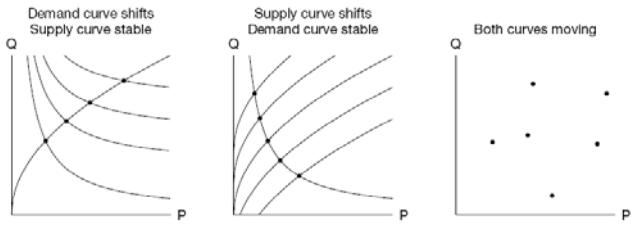


市场上供应和需求一直在变化：

如果需求增加，则价格上涨；而价格上涨导致需求减少

第三个图表示在自由市场(free market)供求双方达成一致/均衡:

供求数量相同(market clearing)。



第三个图表示市场动态变化时,在不同的时间(比如季节)达成均衡的点。

假设某种商品在自由市场上的(均衡)数量和价格分别为Q_t和P_t,

满足如下方程 (省略下标t):

$$\begin{cases} \text{Supply}: & Q = a + bP + \beta' C + \varepsilon \\ \text{Demand}: & Q = c + dP + \gamma' S + \delta \end{cases}$$

Q:数量，P:价格，C:成本，S：替代品价格

其中，ε与δ一般不独立；ε与C独立，δ与S独立

解方程得

$$P = \frac{a - c + (\beta' C - \gamma' S) + (\varepsilon - \delta)}{d - b},$$

$$Q = \frac{ad - bc + (d\beta' C - b\gamma' S) + (d\varepsilon - b\delta)}{d - b}$$

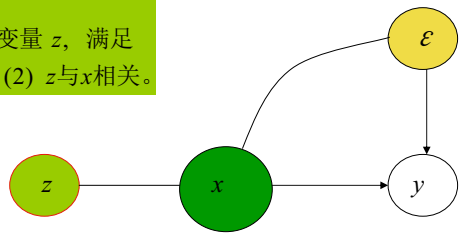
P与ε，δ有关，称为内生变量。Q也是内生的。

与此相对，与模型内变量无关的变量称为外生变量或政策性变量

工具变量法 (Instrumental Variable method)

模型 $y = a + bx + \varepsilon$， x与ε不独立（内生，endogeneous）

工具变量法：
假设存在工具变量 z，满足
(1) z与ε独立；(2) z与x相关。



可以利用外生变量 z，将x中与ε相关的部分过滤掉：
分解$x = \hat{x} + x^{\perp}$, 其中$\hat{x} = \mathrm{P}_z \mathrm{x}$，然后使用 $\hat{x}$代替x.

工具变量最小二乘法 (IVLS) 或 两阶段最小二乘法(2SLS, S=Stage)

模型： $Y = X_{n\times p}\beta_{p\times 1} + \varepsilon, \mathrm{E}(\varepsilon) = 0, \mathrm{var}(\varepsilon) = \sigma^2 \mathrm{I}_n$, ε与X不独立

假设存在q个工具变量(外生变量), 与ε独立, 与X相关.
设$Z_{n\times q}$为q个工具变量的n个观察，假设
(1) $q \geq p$,
(2) Z列满秩，$Rank(Z'X) = p$

分解$X = \hat{X} + X^{\perp}$, 其中$\hat{X} = \mathrm{P}_Z X$。重写模型：

$$Y = X\beta + \varepsilon = \hat{X}\beta + \left\{X^{\perp}\beta + \varepsilon\right\} \doteq \hat{X}\beta + \delta$$

其中$\delta = X^{\perp}\beta + \varepsilon$与$\hat{X}$独立。这是因为
$X^{\perp}\beta$与$\hat{X}$正交，ε只与Z有关故与$\hat{X}$独立。

工具变量法分两步 (2 - Stage)：

第一步. X向Z空间投影，得拟合$\hat{X} = P_Z X$;
第二步. OLS拟合 $Y = \hat{X}\beta + \delta$，得

$$\hat{\beta}_{IVLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y = \left(X'P_Z X\right)^{-1}X'P_Z Y$$

第一步：lm(X~Z) $\Rightarrow$ $\hat{X}$ (理解为X各列对Z回归)
第二步：lm(Y~$\hat{X}$) $\Rightarrow$ $\hat{\beta}_{IVLS}$

注：当$p = q$时，$\hat{\beta}_{IVLS} = (Z'X)^{-1}Z'Y$
这可以看作模型 $Y = X\beta + \varepsilon$ 两边同乘Z'(pre - conditioning)
并舍弃$Z'\varepsilon \approx 0$，解方程$Z'Y = Z'X\beta$得$\hat{\beta}_{IVLS} = (Z'X)^{-1}Z'Y$

性质：
(a) IVLS估计是渐近无偏的：

$$E(\hat{\beta}_{IVLS} \mid Z) \approx \beta \ (\text{当} n \to \infty, E(\hat{\beta}_{2SLS} \mid Z) \to \beta)$$

$$(b)\ \mathrm{var}(\hat{\beta}_{IVLS} \mid Z) = \sigma^2(\hat{X}'\hat{X})^{-1} = \sigma^2\left(X'Z(Z'Z)^{-1}Z'X\right)^{-1}$$

$$\hat{\sigma}^2 = \| Y - X\hat{\beta}_{IVLS} \|^2 / (n - p) \quad (?)$$

工具变量的例子

例**1.** 观察研究抽烟(x)是否导致健康状况(y)下降。

工具变量z可取为z = 烟草税率.

例**2.** 家庭调查中，“income”数据通常很不准确（有测量误差），因此

$$y = \beta_0 + \beta_1 \times Income + \beta_2 \times w + \varepsilon$$

y = 教育支出，ε中含Income的测量误差,所以 Income是内生变量。

Z = 开具支票个数, 与Income有关，但应与ε无关

例**3.** 模型

$$w\text{age} = \beta_0 + \beta_1 \times schooling + \varepsilon$$

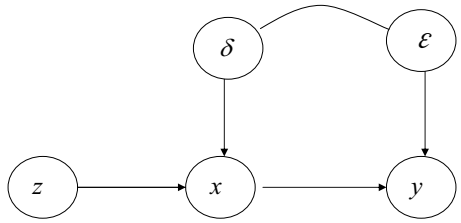
中,ε含有与w*age*,schooling都有关的因素,比如*ability*.

z = ?有人曾建议使用离住地最近的学校的距离。

例3. 假设z以如下形式影响x ： $x = cz + \delta,\ \delta \sim (0, \tau^2)$,

假设x以如下形式影响 y : $y = bx + \varepsilon,\ \varepsilon \sim (0, \sigma^2)$

δ, ε不独立。但z与(δ, ε)独立



z是外生的, z 与x相关。

所以为了估计b，z可用来做工具变量.

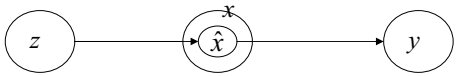
应用两阶段LS:

(1) 拟合 $x_i = cz_i + \delta_i$, 得

$$\hat{c} = \frac{\sum z_i x_i}{\sum z_i^2}, \quad \hat{x}_i = \hat{c} z_i$$

(2) 拟合$y_i = b\hat{x}_i + \varepsilon_i^*$,得

$$\hat{b}_{\text{IVLS}} = \frac{\sum \hat{x}_i y_i}{\sum \hat{x}_i^2} = \frac{\sum z_i y_i}{\sum z_i x_i}$$



爷爷身高　　$\hat{x}$ 父亲身高的预测值　　儿子身高