# Endogeneità e variabili strumentali

$\{y_t, x_t\}$, $t=1,\dots,n$ . Caso stazionario ed ergodico.

$y_t = \beta' x_t + \varepsilon_t$ , $E[\varepsilon_t | x_t] = 0$

$\hat{\beta} = (X'X)^{-1} X' y$

Dimostrazione della consistenza/normalità asintotica

$$\hat{\beta} - \beta = (X'X)^{-1} X' \varepsilon = \left(\frac{1}{n}\sum x_t x_t'\right)^{-1}\left(\frac{1}{n}\sum x_t \varepsilon_t\right)$$

$$\left.\begin{array}{l} \frac{1}{n}\sum x_t x_t' \xrightarrow{p} E(x_t x_t') =: \Sigma_{xx} \\ \frac{1}{n}\sum x_t \varepsilon_t \xrightarrow{p} E(x_t \varepsilon_t) =: \Sigma_{x\varepsilon} = 0 \end{array}\right\} \Rightarrow \hat{\beta} - \beta \xrightarrow{p} 0$$

$$\frac{1}{\sqrt{n}}\sum x_t \varepsilon_t \xrightarrow{d} N(0, \sigma^2 \Sigma_{xx})$$

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 \Sigma_{xx}^{-1})$$

La condizione cruciale è $\Sigma_{x\varepsilon} := E[x_t \varepsilon_t] = 0$
Se $\Sigma_{x\varepsilon} \neq 0$, $\hat{\beta}$ non è più consistente:

$$\hat{\beta} - \beta \xrightarrow{P} \Sigma_{xx}^{-1} \Sigma_{x\varepsilon} \neq 0$$

Se $\Sigma_{x\varepsilon} \neq 0$, c'è un problema di correlazione tra regressore e componente stocastica (Problema di endogeneità)

## Esempio 1: regressore autocorrelato in un modello dinamico

$$y_t = \beta' x_t + \varepsilon_t, \qquad x_t := \begin{pmatrix} w_t \\ y_{t-1} \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

ovvero

$$y_t = \beta_1 w_t + \beta_2 y_{t-1} + \varepsilon_t .$$

Nel caso stazionario/ergodico, $\hat{\beta}$ è C.A.N. se $E[\varepsilon_t x_t] = 0$

Supponiamo che $w_t$ sia un AR(1):

$$w_t = \alpha w_{t-1} + u_t$$

dove $u_t$ è correlato con $\varepsilon_t$. Ovvero $\begin{pmatrix} u_t \\ \varepsilon_t \end{pmatrix} \sim \underline{iid}\left(0, \begin{pmatrix} \sigma_u^2 & \sigma_{u\varepsilon} \\ & \sigma_\varepsilon^2 \end{pmatrix}\right)$

Vediamo a cosa corrisponde $E(x_t \varepsilon_t)$ in questo caso.

$$E[x_t \varepsilon_t] = E\left[\begin{pmatrix} w_t \\ y_{t-1} \end{pmatrix} \varepsilon_t\right] = \begin{bmatrix} E[w_t \varepsilon_t] \\ E[y_{t-1} \varepsilon_t] \end{bmatrix} = \begin{bmatrix} E[(\alpha w_{t-1} + u_t)\varepsilon_t] \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} \alpha E[w_{t-1} \varepsilon_t] + E[u_t \varepsilon_t] \\ 0 \end{bmatrix} = \begin{bmatrix} \alpha \cdot 0 + \sigma_{u\varepsilon} \\ 0 \end{bmatrix} = \begin{bmatrix} \sigma_{u\varepsilon} \\ 0 \end{bmatrix} \neq 0$$

Lo stimatore OLS non è più consistente:

$$\hat{\beta} - \beta \xrightarrow{p} \Sigma_{xx}^{-1} \begin{bmatrix} \sigma_{u\varepsilon} \\ 0 \end{bmatrix} \neq 0$$

esercizio: verificare che la seguente regressione:

$$y_t = \gamma_1 w_{t-1} + \gamma_2 y_{t-1} + e_t$$

può essere stimata consistentemente con OLS.

Esempio 2 : un sistema di equazioni simultanee

$$\begin{cases} C_t = \alpha_0 + \alpha_1 Y_t + \varepsilon_t & 0 < \alpha_1 < 1 \\ Y_t = C_t + I_t \end{cases}$$

Ipotizziamo che $I_t$ non sia correlato con $\varepsilon_t$ : $Cov(I_t, \varepsilon_t) = 0$

Quali sono le proprietà di OLS di $C_t$ su $Y_t$ ? Si può stimare $\alpha_1$ consistentemente?

Ricaviamo $Y_t$ :

$$Y_t = C_t + I_t = \alpha_0 + \alpha_1 Y_t + \varepsilon_t + I_t$$

$$(1 - \alpha_1) Y_t = \alpha_0 + I_t + \varepsilon_t$$

$$Y_t = \frac{\alpha_0}{1-\alpha_1} + \frac{1}{1-\alpha_1} I_t + \frac{1}{1-\alpha_1} \varepsilon_t$$

da cui:

$$Cov(Y_t, \varepsilon_t) = Cov\left[\left(\frac{\alpha_0}{1-\alpha_1} + \frac{1}{1-\alpha_1} I_t + \frac{1}{1-\alpha_1}\varepsilon_t\right) \varepsilon_t\right]$$

$$= \frac{1}{1-\alpha_1} \cancel{Cov(I_t, \varepsilon_t)} + \frac{1}{1-\alpha_1} Cov(\varepsilon_t, \varepsilon_t) = \frac{1}{1-\alpha_1} \sigma_\varepsilon^2 , \quad \sigma_\varepsilon^2 := V(\varepsilon_t)$$

La stima OLS della regressione di $C_t$ su $I_t$ non è consistente per $\alpha_0$ e $\alpha_1$.
(Problema di endogeneità di $Y_t$)

$$\hat{\alpha}_1 = \frac{\hat{Cov}(C_t, Y_t)}{\hat{V}(Y_t)} = \frac{\hat{Cov}(\alpha_0 + \alpha_1 Y_t + \varepsilon_t, Y_t)}{\hat{V}(Y_t)} = \alpha_1 + \frac{\hat{Cov}(\varepsilon_t, Y_t)}{\hat{V}(Y_t)}$$

$$\hat{Cov}(\varepsilon_t, Y_t) \xrightarrow{p} Cov(\varepsilon_t, Y_t) = \frac{\sigma_\varepsilon^2}{1-\alpha_1}$$

$$\hat{V}(Y_t) \xrightarrow{p} V(Y_t) = \frac{1}{(1-\alpha_1)^2} \sigma_I^2 + \frac{1}{(1-\alpha_1)^2} \sigma_\varepsilon^2 , \quad \sigma_I^2 := V(I_t)$$

da cui

$$\hat{\alpha}_1 - \alpha_1 \xrightarrow{p} \frac{\dfrac{\sigma_\varepsilon^2}{\cancel{1-\alpha_1}}}{\dfrac{1}{(1-\alpha_1)^{\cancel{2}}}(\sigma_I^2 + \sigma_\varepsilon^2)} = \frac{\sigma_\varepsilon^2}{\sigma_I^2 + \sigma_\varepsilon^2}(1-\alpha_1) = \frac{1}{1 + \dfrac{\sigma_I^2}{\sigma_\varepsilon^2}}(1-\alpha_1) > 0$$

## Esempio 3: Variabili omesse / Eterogeneità non osservabile.

$w_i$ : salario individuo $i$-esimo $\quad (i = 1, 2, \dots, n)$
$ED_i$ : numero di anni di studio

Viene specificata la seguente regressione

$$w_i = \beta_0 + \beta_1 ED_i + \varepsilon_i$$

Immaginiamo che il modello "vero" sia

$$w_i = \beta_0 + \beta_1 ED_i + \beta_2 AB_i + e_i \, , \qquad E[e_i \mid ED_i, AB_i] = 0$$

dove $AB_i$ = abilità individuale (individuo $i$-esimo)

Se viene stimato il primo modello (senza $AB_i$), notiamo che:

$$w_i = \beta_0 + \beta_1 ED_i + \varepsilon_i \, , \quad \varepsilon_i = e_i + \beta_2 AB_i \, .$$

Abbiamo:

$$Cov(ED_i, \varepsilon_i) = Cov(ED_i, e_i + \beta_2 AB_i)$$

$$= \underbrace{Cov(ED_i, e_i)}_{0} + \beta_2 \underbrace{Cov(ED_i, AB_i)}_{\sigma_{ED,AB}}$$

$$= 0 + \beta_2 \, \sigma_{ED,AB}$$

Se $\beta_2 \neq 0$ e se $\sigma_{ED,AB} \neq 0$, allora c'è un problema di correlazione tra regressore ed errore:

$$\hat{\beta} - \beta \overset{p}{\nrightarrow} 0$$

## Esempio 4: errori di misura.

$$C_t = K\, Y_t^* + \varepsilon_t \quad , \quad Y_t^* = \text{reddito permanente.}$$

$Y_t^*$ non è osservato. Lo sostituisco con una proxy, $Y_t$, correlata con $Y_t^*$ ma affetta da errore di misura:

$$Y_t = Y_t^* + y_t \longrightarrow \text{errore di misura}$$

Ipotizziamo che $E[y_t] = 0$, $Cov(y_t, Y_t^*) = 0$, $Cov(y_t, C_t) = 0$

Cosa succede se cerchiamo di stimare $K$ tramite la regressione OLS di $C_t$ su $Y_t$?

$$C_t = K\, Y_t^* + \varepsilon_t = K(Y_t - y_t) + \varepsilon_t$$

$$C_t = K Y_t + \underbrace{\varepsilon_t - y_t \cdot K}$$

$$C_t = K Y_t + \qquad u_t$$

La stima OLS di $C_t$ su $Y_t$ è inconsistente poiché $Y_t$ e $u_t$ sono tra loro correlati.

$$Cov(Y_t, u_t) = Cov(Y_t^* + y_t, \varepsilon_t - y_t \cdot k)$$

$$= Cov(y_t, -y_t \cdot K) = -K\, V(y_t) = -K \sigma_y^2$$

Inoltre:

$$V(Y_t) = V(Y_t^* + y_t) = V(Y_t^*) + V(y_t) = \sigma_{Y^*}^2 + \sigma_y^2$$

Quindi:

$$\hat{K} = \frac{\widehat{Cov}(C_t, Y_t)}{\hat{V}(Y_t)} \xrightarrow{p} \frac{Cov(C_t, Y_t)}{V(Y_t)} = \frac{Cov(K Y_t + u_t, Y_t)}{V(Y_t)}$$

$$= K + \frac{Cov(u_t, Y_t)}{V(Y_t)} = K - K \frac{\sigma_y^2}{\sigma_{Y^*}^2 + \sigma_y^2} = K - K \frac{1}{\frac{\sigma_{Y^*}^2}{\sigma_y^2} + 1} \neq K$$

$\frac{\sigma^2_{y^*}}{\sigma^2_y}$ è detto signal to noise ratio.

Soluzione – Stimatore delle variabili strumentali [IV]

$$\underset{1\times1}{y_t} = \underset{1\times K}{\beta'} \underset{K\times1}{X_t} + \underset{1\times1}{\varepsilon_t} \qquad E[\underset{K\times1}{X_t} \underset{1\times1}{\varepsilon_t}] \neq \underset{K\times1}{0}$$

Immaginiamo di avere un altro vettore di variabili, $\underset{m\times1}{z_t}$, con le seguenti proprietà:

$E[z_t \varepsilon_t] = 0$ incorrelazione con $\varepsilon_t$

$Cov(z_t, x_t) \neq 0$ correlazione con $x_t$

Se valgono queste proprietà, $z_t$ si dice strumento.

Usiamo $z_t$ per costruire uno stimatore consistente di $\beta$.

Caso in cui $m = K$.

$$E[z_t \varepsilon_t] = 0$$

$$E[z_t (y_t - x_t' \beta)] = 0$$

$$E[z_t y_t] - E[z_t x_t'] \beta = 0$$

$$\beta = E[z_t x_t']^{-1} E[z_t y_t] \approx \frac{Cov(z_t, y_t)}{Cov(z_t, x_t)}$$

Uno stimatore di $\beta$ [metodo dei momenti] è

$$\hat{\beta}_{IV} = \left(\frac{1}{n} \sum_{t=1}^{n} z_t x_t'\right)^{-1} \left(\frac{1}{n} \sum_{t=1}^{n} z_t y_t\right) = (Z'X)^{-1} Z'y$$

dove richiediamo $Z'X$ abbia rango pieno.

Proprietà:

$$\hat{\beta}_{IV} = (Z'X)^{-1} Z'y = (Z'X)^{-1} Z'(X\beta + \varepsilon)$$

$$= \cancel{(Z'X)^{-1} Z'X}\beta + (Z'X)^{-1} Z'\varepsilon$$

$$= \beta + \left(\frac{Z'X}{n}\right)^{-1}\frac{Z'\varepsilon}{n} = \beta + \underbrace{\left(\frac{1}{n}\sum z_t x_t'\right)^{-1}}_{\downarrow p \ \Sigma_{zx}}\underbrace{\left(\frac{1}{n}\sum z_t \varepsilon_t\right)}_{\downarrow p \ \Sigma_{z\varepsilon} = 0} \xrightarrow{p} 0$$

esercizio: dimostrare che $\sqrt{n}(\hat{\beta}_{IV} - \beta) \xrightarrow{d} N\left(0, \Sigma_{zx}^{-1}\Sigma_{zz}\Sigma_{xz}^{-1}\sigma_\varepsilon^2\right)$

Nella pratica, il problema è trovare gli strumenti.

Nell'esempio 1: $x_t = \begin{pmatrix} w_t \\ y_{t-1} \end{pmatrix}$ $\quad \varepsilon_t$ v.i.i.d $\quad Cov(w_t, \varepsilon_t) \neq 0$

$$z_t = \begin{pmatrix} w_{t-1} \\ y_{t-1} \end{pmatrix}$$

$w_{t-1}$ è infatti incorrelata con $\varepsilon_t$ e correlata con $w_t$ se $\alpha \neq 0$

$(w_t = \alpha w_{t-1} + u_t)$

Caso $m > K$ (altra interpretazione)

$y_t = \underset{K\times 1}{x_t'}\beta + \varepsilon_t$ $\qquad E[x_t \varepsilon_t] \neq 0$ , $E[z_t \varepsilon_t] = 0$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \mathrm{Cov}(\underset{m\times 1}{z_t}, x_t) \neq 0$

Costruzione stimatore IV

① proiettare $X$ su $Z$, e prendere i valori stimati

$$\hat{B} = (Z'Z)^{-1} Z'X$$

$$\hat{X} = Z\hat{B} = Z(Z'Z)^{-1}Z'X = P_Z X \quad \text{(valori stimati)}$$

② regredire $y$ su $\hat{X}$

$$\hat{\beta}_{IV} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$$

$$= (X'P_Z X)^{-1}(X'P_Z y)$$

$$= (X'Z(Z'Z)^{-1}Z'X)^{-1} X'Z(Z'Z)^{-1}Z'y$$

(Calcolabile se $Z$ ha rango $m$ e se $X$ ha rango $K$)

Esercizio: dimostrare che $\hat{\beta}_{IV} \xrightarrow{p} \beta$ e che

$$\sqrt{n}\left(\hat{\beta}_{IV} - \beta\right) \xrightarrow{d} N(0, V)$$

(trovare $V$)