

# ECONOMETRIA

## TERZO PROBLEM SET

Soluzioni della seconda parte

### Esercizio 6

Consideriamo il modello in cui regrediamo la variabile  $y$  (che rappresenta il salario di un individuo) sulla variabile  $x_1$  (anni di studio) e  $x_2^*$  (abilità individuale):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^* + u$$

dove la variabile abilità non è osservabile e quindi è misurata con una proxy tale che:

$$x_2^* = \theta_0 + \theta_2 x_2 + v_2$$

a) Cosa rappresentano  $\theta_2$  e  $v_2$ ?

*Il parametro  $\theta_2$  rappresenta la relazione tra  $x_2^*$  e  $x_2$ ; se  $x_2^*$  e  $x_2$  sono positivamente correlati,  $\theta_2 > 0$ .*

*Se  $\theta_2 = 0$ ,  $x_2$  non è una proxy adatta per  $x_2^*$ .*

*La intercetta  $\theta_0$  può essere positiva ( $\theta_0 > 0$ ) o negativa ( $\theta_0 < 0$ ) e semplicemente permette che  $x_2^*$  e  $x_2$  siano misurati in scala diversa.*

*$v_2$  rappresenta il termine di errore.*

b) Scrivete l'equazione che stimereste in questo caso. (Suggerimento: sostituite la seconda equazione nella prima).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 (\theta_0 + \theta_2 x_2 + v_2) + u$$

$$y = \beta_0 + \beta_2 \theta_0 + \beta_1 x_1 + \beta_2 \theta_2 x_2 + \beta_2 v_2 + u$$

$$y = (\beta_0 + \beta_2 \theta_0) + \beta_1 x_1 + (\beta_2 \theta_2) x_2 + (\beta_2 v_2 + u)$$

*Considerate che:*

$$\pi_0 = \beta_0 + \beta_2 \theta_0$$

$$\pi_2 = \beta_2 \theta_2$$

$$\varepsilon = \beta_2 v_2 + u$$

$$y = \pi_0 + \beta_1 x_1 + \pi_2 x_2 + \varepsilon$$

*Adesso è possibile ottenere degli stimatori non distorti di  $\pi_0$ ,  $\beta_1$  e  $\pi_2$ .*

c) Quali sono le condizioni che garantiscono la correttezza degli stimatori OLS del modello al punto (b)? Spiegate.

*L'assunzione necessaria per garantire che questa soluzione fornisca stimatori consistenti può essere riassunta*

*in due assunzioni su  $u$  e  $v_2$ :*

$$1) \text{ cov}(x_1, u) = 0 \text{ ; e } \text{ cov}(x_2, u) = 0$$

*L'errore  $u$  deve essere incorrelato con  $x_1$  e  $x_2^*$  (assunzione di base del modello OLS).*

*In aggiunta,  $u$  deve essere incorrelato con  $x_2$ .*

*Quindi  $x_2^*$  deve influenzare  $y$  direttamente e non attraverso  $x_2$ .*

$x_2$  influenza  $y$  solo attraverso  $x_2^*$

Questa asunzione richiede che  $x_2$  sia una buona proxy per  $x_2^*$ .

2)  $cov(v_2, x_1) = 0$  ; e  $cov(x_2, \varepsilon) = 0$

L'errore  $v_2$  deve essere incorrelato con  $x_1, x_2$

d) Assumete adesso che:

$$x_2^* = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + v_2$$

Spiegate se gli stimatori OLS sono distorti in questo caso e calcolate la distorsione.

Si noti che:  $y = \beta_0 + \beta_1 x_1 + \beta_2(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + v_2) + u$

$y = \beta_0 + \beta_2 \theta_0 + \beta_1 x_1 + \beta_2 \theta_1 x_1 + \beta_2 \theta_2 x_2 + \beta_2 v_2 + u$

$y = (\beta_0 + \beta_2 \theta_0) + (\beta_1 + \beta_2 \theta_1) x_1 + (\beta_2 \theta_2) x_2 + (\beta_2 v_2 + u)$

$p \lim(\hat{\beta}_1) = \beta_1 + \beta_2 \theta_1$

$p \lim(\hat{\beta}_2) = \beta_2 \theta_2$

Quindi, se  $\beta_2 > 0$ , distorsione asintotica positiva (stimatore inconsistente),

a patto che l'abilità individuale abbia una correlazione (parziale) positiva con l'istruzione ( $\theta_2 > 0$ ).

e) Dimostrate che  $Cov(x_2^*, \varepsilon)$  è diversa da zero, dove  $\varepsilon$  è l'errore del modello stimato al punto (b).

Semplicemente:  $Cov(x_2^*, \varepsilon) = E(x_2^*, \varepsilon) = E[(x_2 + v_2)(\beta_2 v_2 + u)] = \beta_2 \sigma_{v_2}^2$

## Esercizio 7

Nel caso in cui :

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

dove  $E(\varepsilon) = 0$ ;  $cov(x_1, \varepsilon) = E(x_1 \varepsilon) \neq 0$

a) Quali sono le condizioni che una variabile strumentale  $z$  deve soddisfare?

i)  $cov(z, \varepsilon) = E(z \varepsilon) = 0$  CONDIZIONE DI ESOGENEITA'

ii)  $dy/dz = \beta_1(dx/dz)$  RESTRIZIONE DI ESCLUSIONE

iii)  $cov(z, x_1) \neq 0$  CONDIZIONE DI RILEVANZA

b) Dimostrate che lo stimatore IV (di variabili strumentali) è consistente.

$$\hat{\beta}_{IV} = \frac{\widehat{cov}(z, y)}{\widehat{cov}(z, x_1)}$$

$$P \lim(\hat{\beta}_{IV}) = P \lim\left(\frac{\widehat{cov}(z, y)}{\widehat{cov}(z, x_1)}\right) = \frac{cov(z, y)}{cov(z, x_1)}$$

Considerando che:

$$\begin{aligned}
y - E(y) &= \beta_1 [x_1 - E(x_1)] + \varepsilon \\
[z - E(z)] [y - E(y)] &= \beta_1 [z - E(z)][x_1 - E(x_1)] + \varepsilon [z - E(z)] \\
\text{cov}(z, y) &= \beta_1 \text{cov}(z, x_1) + \text{cov}(z, \varepsilon) \\
\beta_1 &= \frac{\text{cov}(z, y)}{\text{cov}(z, x_1)} - \frac{\text{cov}(z, \varepsilon)}{\text{cov}(z, x_1)} \\
\text{con } \text{cov}(z, \varepsilon) &= 0 \\
\beta_1 &= \frac{\text{cov}(z, y)}{\text{cov}(z, x_1)} \Rightarrow P \lim (\widehat{\beta}_{1IV}) = \beta_1
\end{aligned}$$

c) Cosa accade se  $\text{cov}(z, \varepsilon) \neq 0$ . Che problema si verifica?

$$\begin{aligned}
P \lim (\widehat{\beta}_{1IV}) &= P \lim \left( \frac{\widehat{\text{cov}}(z, y)}{\widehat{\text{cov}}(z, x_1)} \right) = \frac{\text{cov}(z, y)}{\text{cov}(z, x_1)} = \\
&= \beta_1 + \frac{\text{cov}(z, \varepsilon)}{\text{cov}(z, x_1)} = \\
&= \beta_1 + \frac{\frac{\text{cov}(z, \varepsilon)}{\sigma_z \sigma_\varepsilon} \sigma_\varepsilon}{\frac{\text{cov}(z, x_1)}{\sigma_z \sigma_x} \sigma_x} = \\
&= \beta_1 + \frac{\rho_{z\varepsilon} \sigma_\varepsilon}{\rho_{zx} \sigma_x}
\end{aligned}$$

In questo caso si vede che anche se  $\rho_{z\varepsilon}$  è piccolo, la distorsione dello stimatore IV può essere molto grande se  $\rho_{zx}$  è piccolo e quindi la condizione per cui  $\text{cov}(z, x_1) \neq 0$  è verificata debolmente. Quindi possiamo definire questo uno strumento debole. Inoltre, possiamo confrontare la distorsione (asintotica) di  $\widehat{\beta}_{1IV}$  con quella dello stimatore OLS:

$$\begin{aligned}
P \lim (\widehat{\beta}_1) &= \beta_1 + \frac{\text{cov}(x_1, \varepsilon)}{\text{var}(x_1)} = \\
&= \beta_1 + \frac{\text{cov}(x_1, \varepsilon) \sigma_\varepsilon}{\sigma_\varepsilon \sigma_x \sigma_x} = \\
&= \rho_{x\varepsilon} \frac{\sigma_\varepsilon}{\sigma_x}
\end{aligned}$$

Affinché la distorsione asintotica di cui soffre lo stimatore IV sia inferiore di quella di cui soffre l'OLS non basta che  $|\rho_{zx}| < |\rho_{x\varepsilon}|$ , ma deve essere che  $|\frac{\rho_{z\varepsilon}}{\rho_{zx}}| < \rho_{x\varepsilon}$  ovvero poiché  $|\rho_{zx}| < 1$ ,  $|\rho_{z\varepsilon}|$  deve essere sufficientemente inferiore a  $|\rho_{x\varepsilon}|$ .

### Esercizio 8

a) Scrivete lo stimatore IV multivariato nel caso di esatta identificazione e di sovraidentificazione. Spiegate perché è importante la condizione d'ordine.

*La condizione d'ordine è necessaria per l'identificazione e richiede che il numero delle variabili esogene sia almeno uguale al numero delle variabili endogene incluse. Questa condizione garantisce che vi sia un numero di equazioni uguale al numero di parametri da stimare nel sistema che identifica lo stimatore IV. Se  $L=K$ , abbiamo esatta identificazione, se  $L>K$ , abbiamo sovraidentificazione.*

*$L=K$  esatta identificazione*

$$\widehat{\beta}_{IV} = (Z'X)^{-1}Z'Y$$

*$L>K$  sovraidentificazione*

$$\begin{aligned}\widehat{\beta}_{2SLS} &= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'Y \\ \text{dove } \widehat{X} &= Z\widehat{\Pi} = Z(Z'Z)^{-1}Z'X = P_zX\end{aligned}$$

*sono i valori fittati della regressione OLS sul modello  $X = Z\Pi + U$*

b) Dimostrate che nel caso  $L=K$

$$\widehat{\beta}_{IV} = \widehat{\beta}_{2SLS}$$

*Per dimostrare questo passaggio, dobbiamo partire considerando il residuo del primo stadio:*

$$\begin{aligned}\widehat{u} &= M_zX \\ M_z &= (I - Z(Z'Z)^{-1}Z') = I - P_z \\ M_zZ &= 0\end{aligned}$$

*Di conseguenza:*

$$\begin{aligned}\widehat{X}'X &= \widehat{X}'(\widehat{X} + \widehat{u}) = \\ &= \widehat{X}'(\widehat{X} + M_zX) = \\ &= \widehat{X}'\widehat{X} + X'Z(Z'Z)^{-1}Z'M_zX = \\ &= \widehat{X}'\widehat{X} = \\ &= \widehat{\Pi}'Z'Z\widehat{\Pi} = \\ &= X'Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'X = \\ &= X'Z(Z'Z)^{-1}Z'X\end{aligned}$$

$$\begin{aligned}\widehat{\beta}_{2SLS} &= (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z(Z'Z)^{-1}Z')Y \\ \widehat{\beta}_{2SLS} &= \widehat{\beta}_{IV} = (Z'X)^{-1}Z'Y\end{aligned}$$

c) Dimostrate la distorsione di  $\widehat{\beta}_{2SLS}, \widehat{\beta}_{IV}$

*Si noti che:*

$$\begin{aligned}E(\widehat{\beta}_{IV}|X, Z) &= \\ &= E[(\widehat{X}'\widehat{X})^{-1}\widehat{X}'y|X, Z] = \\ &= E[(\widehat{X}'\widehat{X})^{-1}\widehat{X}'(X\beta + \varepsilon)|X, Z] = \\ &= \beta + (\widehat{X}'\widehat{X})^{-1}\widehat{X}'E(\varepsilon|X, Z)\end{aligned}$$

### Esercizio 9

Consideriamo il seguente modello stimato in STATA:

$$lw = \beta_0 + \beta_1s + \beta_2tenure + \beta_3 \exp r + \beta_4iq + \varepsilon$$

dove  $lw$ = logaritmo dei salari;  $s$ =anni di scuola completati dall'individuo;  $tenure$ =anni di tenure;  $expr$ = anni di esperienza;  $iq$ =indicatore del quoziente di intelligenza.

a) Che problema può esserci nell'uso della variabile  $iq$ ? Come è possibile risolverlo? Osservando la tabella delle variabili disponibili, cercate una soluzione la problema.

```

obs:          758                Wages of Very Young Men, Zvi
                                Griliches, J.Pol.Ec. 1976
vars:         27                31 Oct 2004 14:12
size:        68,978 (99.3% of memory free)
-----
variable name  storage  display  value  variable label
              type    format   label
-----
rns            float   %9.0g           residency in South
rns80          float   %9.0g
mrt            float   %9.0g           marital status = 1 if married
mrt80          float   %9.0g
smsa           float   %9.0g           reside metro area = 1 if urban
smsa80         float   %9.0g
med            float   %9.0g           mother's education, years
iq             float   %9.0g           iq score
kww            float   %9.0g           score on knowledge in world of
                                work test

year           float   %9.0g
age            float   %9.0g
age80          float   %9.0g
s              float   %9.0g           completed years of schooling
s80            float   %9.0g
expr           float   %9.0g           experience, years
expr80         float   %9.0g
tenure         float   %9.0g           tenure, years
tenure80       float   %9.0g
lw             float   %9.0g           log wage
lw80           float   %9.0g
_1year_67      byte    %8.0g           year==67
_1year_68      byte    %8.0g           year==68
_1year_69      byte    %8.0g           year==69
_1year_70      byte    %8.0g           year==70
_1year_71      byte    %8.0g           year==71
_1year_73      byte    %8.0g           year==73
_est_iv        byte    %8.0g           esample() from estimates store
-----

```

*Nel caso di uso della variabile iq, possono esserci problemi di misurazione e quindi dobbiamo strumentare la variabile che ha problemi di endogeneità con almeno una o più variabili esogene. Dalla tabella, possiamo usare come variabili esogene, oltre a quelle già presenti nella regressione, la variabile kww che è un indicatore di un test di conoscenza e possiamo usare anche la variabile med, che indica gli anni di istruzione della madre.*

b) Commentate il seguente output di STATA

```

. ivreg lw s tenure expr (iq=kww med ), first
First-stage regressions
-----
Source |      SS      df      MS                Number of obs =   758
-----+-----
Model | 41951.1677      5 8390.23354          F( 5, 752) =   64.09
Residual | 98448.1581     752 130.915104          Prob > F      =  0.0000
-----+-----
Total | 140399.326     757 185.468066          R-squared     =  0.2988
                                           Adj R-squared =  0.2941
                                           Root MSE    =  11.442

-----+-----
iq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
s | 2.480376   .2177002    11.39  0.000   2.053004   2.907749
tenure | .2372269   .2602616     0.91  0.362  -.2736987   .7481525
expr | -.4446049   .2109757    -2.11  0.035  -.8587763  -.0304335
kww | .3100834   .0637504     4.86  0.000   .1849335   .4352334
med | .4114398   .1626099     2.53  0.012   .0922165   .7306632
_cons | 55.11403   3.050712    18.07  0.000   49.12511   61.10296

-----+-----

Instrumental variables (2SLS) regression

Source |      SS      df      MS                Number of obs =   758
-----+-----
Model | 26.0174883      4 6.50437208          F( 4, 753) =   76.06
Residual | 113.268662     753 .150423189          Prob > F      =  0.0000
-----+-----
Total | 139.28615     757 .183997556          R-squared     =  0.1868
                                           Adj R-squared =  0.1825
                                           Root MSE    =   .38784

-----+-----
lw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
iq | .0178162   .0059301     3.00  0.003   .0061746   .0294578
s | .0520467   .019293      2.70  0.007   .0141722   .0899213
tenure | .0297485   .0090933     3.27  0.001   .0118973   .0475997
expr | .0442787   .0074261     5.96  0.000   .0297004   .058857
_cons | 3.007415   .3826309     7.86  0.000   2.256265   3.758565

-----+-----
Instrumented:  iq
Instruments:  s tenure expr kww med
-----+-----

```

*In questo output abbiamo una regressione a due stadi, nel primo stadio la variabile strumentata, iq, è regredita sui due strumenti scelti e sulle altre variabili esogene della regressione. In questa regressione, tutte le variabili, tranne tenure sembra che siano in grado di spiegare la variabile iq. Inoltre la F statistic ha un valore elevato (oltre a 10, regola del pollice di Watson) e quindi le variabili strumentali non sono strumenti deboli. Nel secondo stadio, i residui della regressione del primo stadio sono usati in sostituzione della variabile iq. Tutti i coefficienti sono significativi.*

c) Commentate l'output del test di Sargan. A cosa serve tale test?

```

. overid

Tests of overidentifying restrictions:
Sargan N*R-sq test      0.236  Chi-sq(1)    P-value = 0.6271
Basmann test           0.234  Chi-sq(1)    P-value = 0.6284

```

*Il test di Sargan serve per verificare la validità degli strumenti in eccesso (L-K) considerando che K strumenti sono validi. In questo caso, non possiamo rifiutare l'ipotesi nulla, quindi gli strumenti in eccesso sono validi.*

d) Commentate l'output del test di Hausman. A cosa serve tale test?

```
. hausman iv .
```

---- Coefficients ----				
	(b) iv	(B) .	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
iq	.0178162	.0038838	.0139323	.0058275
s	.0520467	.0947161	-.0426694	.0180552
tenure	.0297485	.0362904	-.0065419	.0045474
expr	.0442787	.0390324	.0052463	.003694

```
-----
b = consistent under Ho and Ha; obtained from ivreg
B = inconsistent under Ha, efficient under Ho; obtained from regress

Test: Ho: difference in coefficients not systematic

      chi2(4) = (b-B)'[(V_b-V_B)^(-1)](b-B)
            =      5.72
      Prob>chi2 =      0.2214
```

*Il test di Hausman serve per testare l'esogeneità della variabile iq. L'ipotesi nulla è che non ci sia correlazione tra la variabile iq e l'errore. In questo caso, l'ipotesi nulla non viene rifiutata, quindi è possibile utilizzare un semplice modello OLS al posto di una stima a 2 stadi perché entrambi gli stimatori sono consistenti ma OLS è più efficiente.*