

6089 ECONOMETRIA
A.A. 2008 – 2009
Lezione Info4

In un celebre studio del 1995, l'economista di Berkeley David Card usa la distanza tra la più vicina università e la vecchia residenza degli individui come strumento per gli anni di studio effettivamente conseguiti, in modo da risolvere il problema di endogeneità dell'istruzione nella stima di un'equazione salariale. In questa esercitazione, stimeremo il rendimento atteso dell'istruzione in termini di futuro salario, usando i dati e la strategia di Card insieme ad altre variabili strumentali introdotte a titolo di esempio.

1. Operazioni preliminari

- a) Create la directory "C:\econometria" e copiatevi il file card.dta.
- b) Nella finestra di comando di Stata digitate "**cd c:\econometria**": da questo momento questa diventa la directory di riferimento per le operazioni che svolgeremo con Stata.
- c) Create un file in cui saranno contenuti tutti i passaggi svolti durante la lezione digitando "**log using info4.log, replace**".

2. Un primo sguardo ai dati e stima OLS del rendimento dell'istruzione

- a) Aprite il file card.dta: "**use card,clear**".
- b) Il comando "**describe**" ("**des**") fornisce una descrizione del contenuto del dataset (numero di osservazioni, tipo e descrizione, *label* delle variabili), mentre il comando "**summarize**" ("**sum**") fornisce alcune statistiche descrittive delle variabili.
- c) Stimate con OLS l'equazione salariale (usando il comando "**reg**"):
$$lwage_i = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 expersq_i + \beta_5 black_i + \beta_6 smsa_i + \beta_7 south_i + \lambda_r + \varepsilon_i$$
dove i parametri λ_r catturano l'effetto di *dummies* regionali riferite alla vecchia residenza.
 - Commentate il segno e la significatività dei coefficienti stimati.
 - Commentate i valori dell' R^2 e del test F.
 - Quali problemi vi aspettate in termini d'endogeneità? Commentate di nuovo la stima del rendimento dell'istruzione in questa luce.

3. Usare i risultati dei test IQ/KWW per controllare per l'abilità non osservata

- a) Disponendo dei risultati di due test standardizzati (IQ e KWW), possiamo usarli come *proxy* dell'abilità individuale non osservata: iniziamo includendo IQ tra i regressori del modello OLS stimato sopra. Come cambia la stima del rendimento dell'istruzione? (Si discuta questo punto anche alla luce della correlazione tra IQ e gli anni di studio.)
- b) Adesso, preoccupati da possibili errori di misura nel test IQ come *proxy* dell'abilità, potremmo usare il risultato nel secondo test come strumento, attraverso il comando: **ivregress 2sls lwage educ exper expersq black smsa south reg661-reg668 (IQ=KWW)**. Su quali assunzioni si basa questa strategia? Si commentino i risultati.
- c) Se vogliamo capire se lo strumento KWW è forte o debole, cosa dobbiamo fare? Lo si spieghi stimando direttamente l'equazione in forma ridotta o ripetendo la stima con "**ivregress 2sls**" e usando di seguito il comando "**estat firststage**".

- d) Possiamo applicare il test di Sargan per testare l'esogeneità dello strumento (comando: **estat override**)? Perché?

4. Usare la distanza dall'università come strumento per l'istruzione

- a) Assumiamo, adesso, di non fidarci dei test IQ/KWW usati sopra e di escluderli dal set di regressori utilizzati per stimare i rendimenti dell'istruzione. Ci troviamo di nuovo a dover affrontare un problema di endogeneità da variabile omessa. La strategia di Card può offrire una possibile soluzione. Proviamo a usare, quindi, la distanza della vecchia residenza dall'università come variabile strumentale. Abbiamo due variabili (binarie) di questo tipo: la distanza da un *college* che offre titoli di 4 anni o di 2 anni. Guardano alla semplice correlazione tra questi due possibili strumenti, la variabile endogena e la variabile dipendente, che cosa ci aspettiamo?
- b) Si stimi di nuovo il modello (con il comando "**ivregress 2sls**"):
- $$lwage_i = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 expersq_i + \beta_5 black_i + \beta_6 smsa_i + \beta_7 south_i + \lambda_r + \varepsilon_i$$
- usando rispettivamente `nearc2`, `nearc4` o entrambe le variabili come strumenti.
- Per quale motivo in questo caso è particolarmente importante includere λ_r ?
 - Commentate i valori delle diverse stime. Che idea possiamo farci sui rendimenti dell'istruzione?
 - Che cosa possiamo dire sulla debolezza degli strumenti (usando di nuovo il comando "**estat firststage**")? Come può interagire questo aspetto con la consistenza delle stime analizzate al punto precedente?
 - Nel caso sovraidentificato, cosa ci suggerisce il test di Sargan (usando di nuovo il comando "**estat override**")? Su quali assunzioni si basa questo test? Sono plausibili in questo caso?

5. Test di sovraidentificazione in presenza di più strumenti (con "storie" diverse)

- a) Assumiamo, adesso, che un *referee* ci suggerisca di usare un altro strumento per gli anni di studio: gli anni di studio della madre. Prima di applicare il suo suggerimento (i *referee* hanno sempre ragione!), che cosa possiamo dire sulla validità del nuovo strumento?
- b) Si stimi il modello (sovraidentificato) con gli anni di studio della madre come strumento aggiuntivo rispetto alla distanza dal *college*. Che cosa succede alla stima del rendimento dell'istruzione? Come possiamo valutare l'eventuale debolezza degli strumenti?
- c) Il test di Sargan sull'esogeneità di tutti gli strumenti che cosa suggerisce?
- d) Si assuma, infine, di avere più fiducia nella validità dello strumento "distanza dal *college*" rispetto allo strumento "anni di studio della madre". Sotto l'ipotesi di validità del primo, possiamo quindi testare la restrizione di esclusione del secondo. Come?

6. Operazioni di chiusura

Salvate il file dei dati con le nuove variabili costruite e chiudete il log file: **save card_new.dta, replace; log close.**