Regressione con Variabili Strumentali

Tre importanti minacce alla validità interna del modello di regressione sono:

- Errore da variabile omessa che è correlata con X ma non essendo osservabile non può essere inclusa nella regressione;
- Errore dovuto alla causalità simultanea (X determina Y, Y determina X);
- Errori nelle variabili (X è misurata con errore – errore di misura).
- La regressione con variabili strumentali può eliminare l'errore da queste tre fonti.

Lo Stimatore IV con un singolo regressore e un Singolo Strumento

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- La regressione IV si suddivide in due parti: una parte che potrebbe essere correlata con u, e una parte che non lo è. Isolando la parte non correlata con u, è possibile stimare β_1.
- Questo si ottiene utilizzando una ***variabile strumentale***, Z_i, che non è correlata con u_i.
- La variabile strumentale cattura i movimenti in X_i che non sono correlati con u_i, e li utilizza per stimare β_1.

Terminologia: Endogeneità e Esogeneità

Una variabile si dice ***endogena*** quando è correlata con u.
Una variabile si dice ***esogena*** quando non è correlata con u.

Nota storica: "Endogeno" letteralmente significa "determinate nell'ambito di un sistema," cioè, una variabile determinata congiuntamente con Y, cioè, una variabile soggetta a causalità simultanea. Tuttavia questa definizione è troppo ristretta. La regressione IV può essere utilizzata anche per risolvere il problema dell'errore da OV e degli errori-in variabile, non solo l'errore per causalità simultanea.

Due condizione per la validità dello strumento

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Affinché una variabile strumentale (uno "***strumento***") Z sia valido, deve soddisfare due condizioni:

1. ***Rilevanza dello strumento***: $\text{corr}(Z_i, X_i) \neq 0$
2. ***Esogeneità dello strumento***: $\text{corr}(Z_i, u_i) = 0$

Supponiamo per il momento che conosciamo lo strumento Z_i (discuteremo in seguito come trovare uno strumento valido). Come possiamo usare Z_i per stimare β_1?

Lo stimatore IV, una *X* e una *Z*

Spiegazione #1: Minimi- quadrati a due stadi (Two Stage Least Squares - TSLS -)

Come il termine implica, TSLS ha due stadi – due regressioni:

(1) Nel primo stadio, isola la parte di X che non è correlata con u:

regredire X su Z usando OLS

$$X_i = \pi_0 + \pi_1 Z_i + v_i \qquad (1)$$

- Poiché Z_i non è correlate con u_i, $\pi_0 + \pi_1 Z_i$ non è correlata u_i. Non conosciamo π_0 or π_1 ma li stimiamo, in modo…
- Da calcolare il valore predetto di X_i, $\hat{X}_i$, dove $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$, $i = 1,\ldots,n$.

(2) Sostituire X_i con $\hat{X}_i$ nella regressione di interesse:

regredire Y su $\hat{X}_i$ usando OLS:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i \qquad (2)$$

- Poichè $\hat{X}_i$ non è correlata con u_i in grandi campioni, la prima assunzione del metodo dei minimi quadrati è valida.
- Quindi β_1 può essere stimato con OLS usando la regressione (2).
- Questo argomento è valido per campioni grandi (in modo che π_0 e π_1 siano stimati correttamente usando la regressione (1)).
- Questo stimatore è chiamato "Stimatore dei minimi quadrati a due stadi" - "Two Stage Least Squares (TSLS) estimator", $\hat{\beta}_1^{TSLS}$ -.

Srimatore dei Minimi Quadrati a Due Stadi, continua.

Supporre di conoscere uno strumento valido, Z_i.

Fase 1:

Regredire X_i su Z_i, e ottenere un valore predetto $\hat{X}_i$

Fase 2:

Regredire Y_i su $\hat{X}_i$; il coefficiente su $\hat{X}_i$ è lo stimatore TSLS, $\hat{\beta}_1^{TSLS}$.

Allora $\hat{\beta}_1^{TSLS}$ è uno stimatore consistente di β_1.

Stimatore TSLS, cont., una X e una Z, cont.

Spiegazione #2: (solo) un po' di algebra

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Quindi,

$$\begin{aligned}
\text{cov}(Y_i,Z_i) &= \text{cov}(\beta_0 + \beta_1 X_i + u_i, Z_i) \\
&= \text{cov}(\beta_0,Z_i) + \text{cov}(\beta_1 X_i,Z_i) + \text{cov}(u_i,Z_i) \\
&= \quad 0 \quad + \text{cov}(\beta_1 X_i,Z_i) + \quad 0 \\
&= \beta_1 \text{cov}(X_i,Z_i)
\end{aligned}$$

dove $\text{cov}(u_i,Z_i) = 0$ (esogeneità dello strumento); perciò

$$\beta_1 = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(X_i, Z_i)}$$

Stimatore TSLS, ctd., una *X* e una *Z*, ctd.

$$\beta_1 = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(X_i, Z_i)}$$

Lo stimatore IV sostituisce queste covarianze della popolazione con le covarianze campionarie:

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}},$$

s_{YZ} e s_{XZ} sono le covarianze campionarie.

Questo è lo stimatore TSLS.

Consistenza dello stimatore TSLS

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$$

Le covarianze campionarie sono consistenti: $s_{YZ} \xrightarrow{p} \text{cov}(Y,Z)$ e $s_{XZ} \xrightarrow{p} \text{cov}(X,Z)$. Quindi,

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}} \xrightarrow{p} \frac{\text{cov}(Y,Z)}{\text{cov}(X,Z)} = \beta_1$$

- La condizione della rilevanza dello strumento, $\text{cov}(X,Z) \neq 0$, assicura che non si divida per zero.

Esempio #1: Offerta e Domanda di Burro

Regressione IV è stata sviluppata originariamente per stimare le elasticità della domanda per beni agricoli, per esempio il burro:

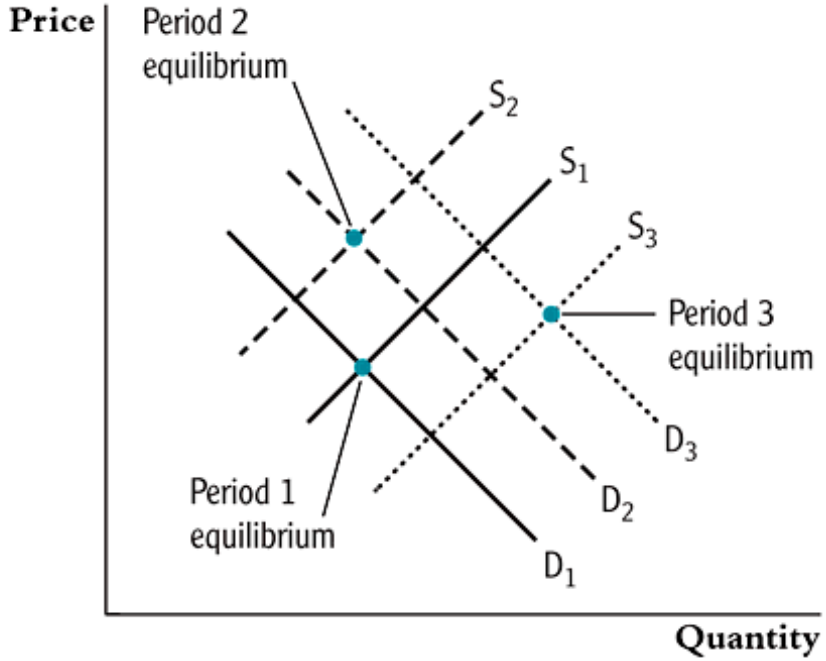
$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

- β_1 = l'elasticità del prezzo del burro = la variazione percentuale nella quantità per una variazione del 1% nel prezzo (ricorda la discussione delle specificazioni log-log)
- Dati: osservazioni sul prezzo e la quantità del burro per anni differenti
- La regressione OLS di $\ln(Q_i^{butter})$ su $\ln(P_i^{butter})$ soffre dell'errore di causalità simultanea (*perché*?)

L'errore di causalità simultanea nella regressione OLS di $\ln(Q_i^{butter})$ su $\ln(P_i^{butter})$ è dovuto al fatto che prezzo e quantità sono determinati simultaneamente dalla interazione di domanda *e* offerta

FIGURE 10.1

(a) Price and quantity are determined by the intersection of the supply and demand curves. The equilibrium in the first period is determined by the intersection of the demand curve D_1 and the supply curve S_1. Equilibrium in the second period is the intersection of D_2 and S_2, and equilibrium in the third period is the intersection of D_3 and S_3.

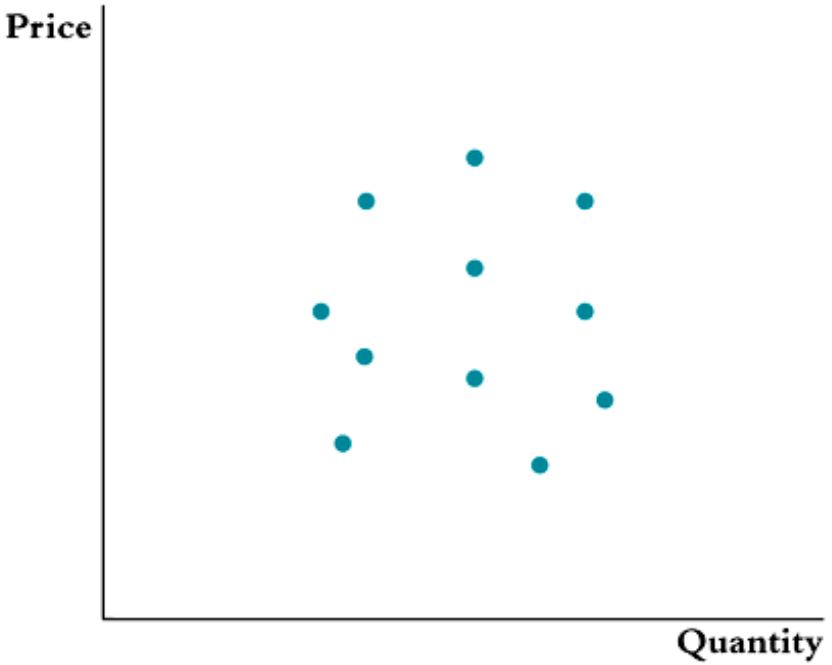


(a) Demand and Supply in Three Time Periods

Questa interazione di domanda e offerta produce…

FIGURE 10.1

(b) This scatterplot shows equilibrium price and quantity in eleven different time periods. The demand and supply curves are hidden. Can you determine the demand and supply curves from the points on the scatterplot?



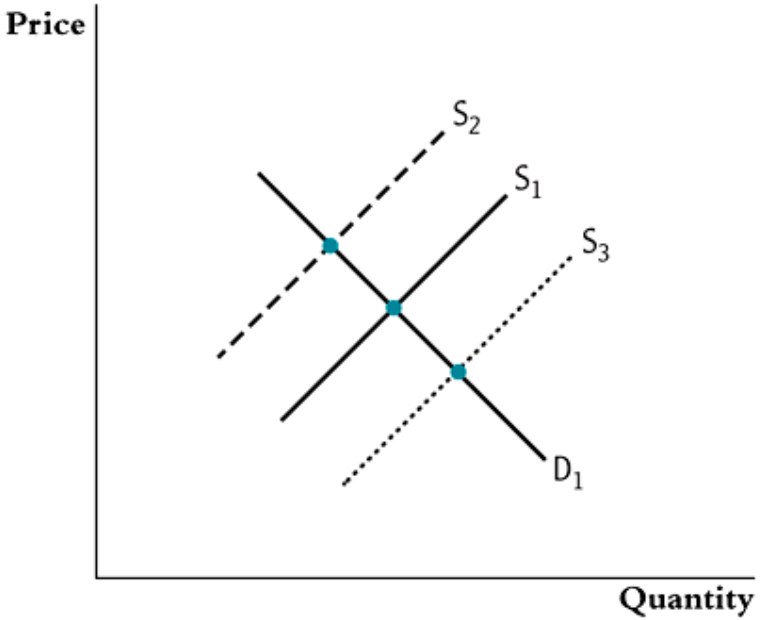
(b) Equilibrium Price and Quantity for Eleven Time Periods

Una regressione che utilizza questi dati stima la curva di domanda?

Cosa si ottiene se si avessero dei movimenti solo dell'offerta?

FIGURE 10.1

(c) When the supply curve shifts from S_1 to S_2 to S_3 but the demand curve remains at D_1, the equilibrium prices and quantities trace out the demand curve.



(c) Equilibrium Price and Quantity When Only the Supply Curve Shifts

- TSLS stima la curva di domanda isolando i movimenti nei prezzi e nelle quantità che derivano da spostamenti della curva di offerta.
- *Z* è una variabile che sposta la curva di offerta ma non la curva di domanda.

TSLS nell'esempio offerta-domanda:

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

Sia Z = quantità di pioggia nelle regioni che producono burro.

E' Z uno strumento valido?

(1) Esogeno? corr($rain_i$,u_i) = 0?

Plausibile: se piove nelle regioni che producono burro, la pioggia non dovrebbe influenzare la domanda

(2) Rilevanza? corr($rain_i$,$\ln(P_i^{butter})$) ≠ 0?

Plausibile: insufficiente pioggia determina minore pascolo e quindi meno burro.

TSLS nell'esempio offerta-domanda, cont.

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

$Z_i = rain_i$ = quantità di pioggia nelle regioni che producono burro.

Stadio 1: regredire $\ln(P_i^{butter})$ su *rain*, ottenere $\ln\left(\hat{P}_i^{butter}\right)$

$\ln\left(\hat{P}_i^{butter}\right)$ isola le variazioni in log price dovute all'offerta (almeno in parte)

Stadio 2: regredire $\ln(Q_i^{butter})$ su $\ln\left(\hat{P}_i^{butter}\right)$ La regressione che utilizza gli spostamenti della curva di offerta per individuare la curva di domanda.

Esempio #2: Valutazione dei Test e dimensione della classe

- Le regressioni per i distretti della California potrebbero avere errore OV (per esempio il coinvolgimento dei genitori).
- Questo errore potrebbe essere eliminato utilizzando una regressione IV (TSLS).
- La regressione IV richiede uno strumento valido, cioè, uno strumento che sia:
  (1) rilevante: corr(Z_i,STR_i) $\neq 0$
  (2) esogeno: corr(Z_i,u_i) = 0

Esempio #2: Valutazione dei Test e dimensione della classe, cont.

Ecco uno strumento (ipotetico):

- alcuni distretti, sono stati colpiti da un terremoto, "raddoppia" la dimensione delle classi:

  $Z_i = Quake_i = 1$ se è stato colpito dal terremoto, = 0 diversamente

- *Sono in questo caso valide le due condizioni per uno strumento valido*?
- Il terremoto agisce come se i distretti fossero assegnati in modo del tutto casuale in un esperimento perfetto controllato. Quindi la variazione in *STR* dovuto al terremoto è esogena.
- Il primo stadio di TSLS regredisce *STR* su *Quake*, quindi isola la parte di *STR* che è esogena (la parte che è scelta come se fosse assegnata)

Faremo altri esempi in seguito…

Inferenza usando TSLS

- In grandi campioni, la distribuzione campionaria dello stimatore TSLS è normale
- Per l'inferenza (test delle ipotesi, intervalli di confidenza) si procede nel solito modo, per esempio. $1.96SE$
- L'idea alla base della distribuzione normale dello stimatore TSLS in grandi campioni è che – come tutti gli altri stimatori che abbiamo considerato – è determinato da una media di variabili casuali i.i.d con media zero, alle quali possiamo applicare il CLT.

- Qui di seguito una sintesi della derivazione matematica (vedere il testo di riferimento –Appendice del capitolo sulle variabili strumentali – per maggiori dettagli)...

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}$$

Sostituiamo in $Y_i = \beta_0 + \beta_1 X_i + u_i$ e semplirichiamo:

Prima,

$$Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + (u_i - \bar{u})$$

così

$$\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z}) = \frac{1}{n-1}\sum_{i=1}^{n}[\beta_1(X_i - \bar{X}) + (u_i - \bar{u})](Z_i - \bar{Z})$$

$$= \beta_1 \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z}) + \frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{u})(Z_i - \bar{Z}).$$

Perciò

$$\hat{\beta}_1^{TSLS} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}$$

$$= \frac{\beta_1 \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z}) + \frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{u})(Z_i - \bar{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}$$

$$= \beta_1 + \frac{\frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{u})(Z_i - \bar{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}.$$

Sottraendo β_1 da entrambi i lati, otteniamo

$$\hat{\beta}_1^{TSLS} - \beta_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{u})(Z_i - \bar{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}$$

Moltriplicando per $\sqrt{n-1}$ e utilizzando l'approssimazione $\sqrt{n-1} \approx \sqrt{n}$ otteniamo:

$$\sqrt{n}(\hat{\beta}_1^{TSLS} - \beta_1) \approx \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(u_i - \bar{u})(Z_i - \bar{Z})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}$$

$$\sqrt{n}(\hat{\beta}_1^{TSLS} - \beta_1) \approx \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(u_i - \bar{u})(Z_i - \bar{Z})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}$$

- Consideriamo il numeratore: in grandi campioni,

  $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(u_i - \bar{u})(Z_i - \bar{Z})$ si distribuisce $N(0,\text{var}[(Z-\mu_Z)u])$

- Poi consideriamo il denominatore:

  $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z}) \xrightarrow{p} \text{cov}(X,Z)$ utilizzando la legge dei grandi numeri LLN

  dove $\text{cov}(X,Z) \approx 0$ perchè lo strumento è rilevante (per assunzione)

  (*Cosa succede nel caso lo strumento fosse irrilevante? Lo vedremo in seguito*.)

Considerando numeratore e denominatore, otteniamo:

$$\sqrt{n}(\hat{\beta}_1^{TSLS} - \beta_1) \approx \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(u_i - \bar{u})(Z_i - \bar{Z})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}$$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z}) \xrightarrow{p} \text{cov}(X,Z)$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(u_i - \bar{u})(Z_i - \bar{Z}) \text{ si distribuisce } N(0,\text{var}[(Z-\mu_Z)u])$$

Otteniamo che:

$\hat{\beta}_1^{TSLS}$ si distribuisce approssimativamente come una $N(\beta_1,\sigma^2_{\hat{\beta}_1^{TSLS}})$,

where $$\sigma^2_{\hat{\beta}_1^{TSLS}} = \frac{1}{n}\frac{\text{var}[(Z_i - \mu_Z)u_i]}{[\text{cov}(Z_i, X_i)]^2}.$$

Inferenza usando TSLS, continua

$$\hat{\beta}_1^{TSLS} \text{ è distribuito } N(\beta_1, \sigma^2_{\hat{\beta}_1^{TSLS}}),$$

- Per l'inferenza statistica si procede come al solito.
- La giustificazione è (come al solito) basata sulla teoria in grandi campioni
- Si assume che gli strumenti siano validi – discuteremo cosa succede se non sono validi.
- ***Importante nota sugli errori standard***:
  - Gli standard error OLS del secondo stadio della regressione non sono corretti – non tengono conto della stima nel primo stadio ($\hat{X}_i$ è stimata).
  - Invece, è necessario usare, quando si stima al computer TSLS, un commando nel software per la stima corretta degli *SE* dello stimatore TSLS.
  - Come al solito, bisogna usare SE robusto all'eteroschedasticità

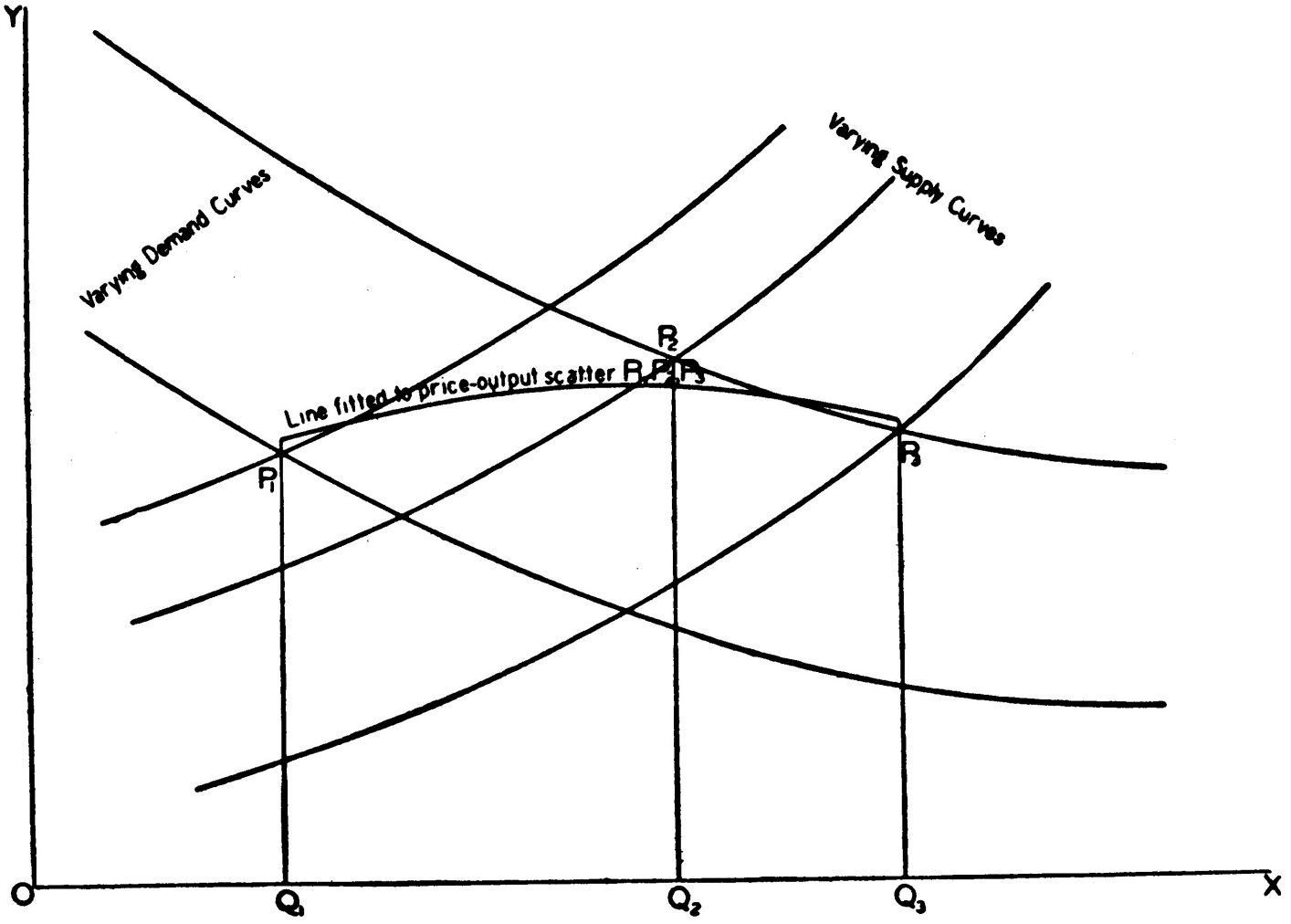
Un discorso completo:

La storia iniziale della regressione IV

- Quanti soldi potrebbero essere raccolti da una tariffa sulle importazioni di olii vegetali e animaili (burro, olio di oliva, olio di soia, ecc)?
- Per fare questo calcolo abbiamo bisogno del valore delle elasticità della domanda e dell'offerta, sia interna che estera
- Questo problema fu risolto nell'Appendice dello studio di Wright (1928), "The Tariff on Animal and Vegetable Oils."

Figura 4, p. 296, dall'Appendice B (1928):

FIGURE 4. PRICE-OUTPUT DATA FAIL TO REVEAL EITHER SUPPLY OR DEMAND CURVE.



Who wrote Appendix B of Philip Wright (1928)?

> …questa appendice si pensa sia stata scritta da suo figlio, Sewall Wright, uno statistico importante. (SW, p. 334)

Chi erano gli autori di questa storia?

	
Philip Wright (1861-1934) *obscure economist and poet* MA Harvard, Econ, 1887 Lecturer, Harvard, 1913-1917	**Sewall Wright (1889-1988)** *famous genetic statistician* ScD Harvard, Biology, 1915 Prof., U. Chicago, 1930-1954

Esempio: Domanda di Sigarette

- Quanto una ipotetica tassa sulle sigarette ridurebbe il consumo delle sigarette?
- Per rispondere, abbiamo bisogno dell'elasticità della domanda di sigarette, cioè, β_1, nella regressione,

$$\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1 \ln(P_i^{cigarettes}) + u_i$$

- Potrebbe essere lo stimatore OLS plausibilmente corretto?

  Perchè?

Esempio: Domanda di sigarette, ctd.

$$\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1 \ln(P_i^{cigarettes}) + u_i$$

Dati Panel:

- Consumo annuale di sigarette e prezzo medio pagato (incluse le tasse)
- 48 stati US, 1985-1995

Variabile strumentale proposta:

- Z_i = tassa generica sulle vendite per pacchetto nello stato = $SalesTax_i$
- E' uno strumento valido?

  (1) Rilevante? corr($SalesTax_i$, $\ln(P_i^{cigarettes})$) $\neq 0$?

  (2) Esogeno? corr($SalesTax_i$,u_i) = 0?

Per il momento utilizziamo, solo i dati di un anno 1995.

Regressione OLS del primo stadio:

$$\widehat{\ln}(P_i^{cigarettes}) = 4.63 + .031 SalesTax_i,\ n = 48$$

Regressione OLS del secondo statdio:

$$\widehat{\ln}(Q_i^{cigarettes}) = 9.72 - 1.08\ \widehat{\ln}(P_i^{cigarettes}),\ n = 48$$

Combiniamo le regressione con gli errori standard corretti, robusti all'eteroschedasticità:

$$\widehat{\ln}(Q_i^{cigarettes}) = 9.72 - 1.08\ \widehat{\ln}(P_i^{cigarettes}),\ n = 48$$

$$(1.53)\ \ (0.32)$$

STATA Esempio (molto simile in GRETL): Domanda di Sigarette, Primo stadio

Strumento = Z = *rtaxso* = tassa generica sulle vendite di sigarette (valore reale $/Per pacchetto)

X *Z*

```
. reg lravgprs rtaxso if year==1995, r;

Regression with robust standard errors                Number of obs =      48
                                                      F(  1,    46) =   40.39
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4710
                                                      Root MSE      =  .09394

------------------------------------------------------------------------------
             |               Robust
    lravgprs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      rtaxso |   .0307289   .0048354     6.35   0.000     .0209956    .0404621
       _cons |   4.616546   .0289177   159.64   0.000     4.558338    4.674755
------------------------------------------------------------------------------
```

X-hat

```
. predict lravphat;
```

Now we have the predicted values from the 1st stage

Secondo Stadio

```
        Y       X-hat
. reg lpackpc lravphat if year==1995, r;

Regression with robust standard errors                Number of obs =      48
                                                      F(  1,    46) =   10.54
                                                      Prob > F      =  0.0022
                                                      R-squared     =  0.1525
                                                      Root MSE      =  .22645

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravphat |  -1.083586   .3336949    -3.25   0.002    -1.755279   -.4118932
       _cons |   9.719875   1.597119     6.09   0.000     6.505042    12.93471
------------------------------------------------------------------------------
```

- Questi coefficienti sono le stime TSLS
- Gli errori standard sono errati perchè ignorano il fatto che il regressore è stato stimato nel primo stadio.

Combiniamo i due stadi in un singolo comando:

```
          Y          X          Z
. ivreg lpackpc (lravgprs = rtaxso) if year==1995, r;

IV (2SLS) regression with robust standard errors      Number of obs =      48
                                                      F(  1,    46) =   11.54
                                                      Prob > F      =  0.0014
                                                      R-squared     =  0.4011
                                                      Root MSE      =  .19035

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.083587   .3189183    -3.40   0.001    -1.725536   -.4416373
       _cons |   9.719876   1.528322     6.36   0.000     6.643525    12.79623
------------------------------------------------------------------------------
Instrumented:  lravgprs          This is the endogenous regressor
Instruments:   rtaxso            This is the instrumental varible
------------------------------------------------------------------------------
```

OK, la variazione degli errori standard è minima in questo caso...ma non è sempre così!

$$\dot{\ln}(Q_i^{cigarettes}) = 9.72 - 1.08\ \dot{\ln}(P_i^{cigarettes}),\ n = 48$$

$$(1.53)\ \ (0.32)$$

Sintesi della Regressione IV con una singola X e Z

- Uno strumento Z valido deve soddisfare due condizioni:
  (1) *rilevanza*: $\text{corr}(Z_i, X_i) \neq 0$
  (2) *esogeneità*: $\text{corr}(Z_i, u_i) = 0$
- TSLS procede regredendo prima X su Z ottenendo $\hat{X}$, poi regredendo Y su $\hat{X}$.
- L'idea chiave è che il primo stadio isola la parte della variazione in X che è non correlata con u
- Se lo strumento è valido, allora la distribuzione campionaria in grandi campioni dello stimatore TSLS è normale, quindi per l'inferenza si procede come al solito

Il Modello Generico della Regressione IV

- Fino a questo momento abbiamo considerato la regressione IV con un singolo regressore endogeno (X) ed uno strumento (Z).
- Abbiamo bisogno di estendere il modello:
  - regressori endogeni multipli ($X_1,\ldots,X_k$)
  - variabili esogene multiple ($W_1,\ldots,W_r$)

    Questo è necessario includerle per la minaccia solita da OV
  - variabili strumentali multiple ($Z_1,\ldots,Z_m$)

    Più (rilevanti) strumenti possono produrre una varianza minore dello stimatore TSLS: l' R^2 del primo stadio aumenta, quindi si ha più variazione in $\hat{X}$.

Esempio: domanda di sigarette

- Un'altra determinante della domanda è il reddito; non includerlo significherebbe commettere un errore per omissione di variabile
- La domanda di sigarette con una X, una W, e 2 strumenti (2 Z):

$$\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1 \ln(P_i^{cigarettes}) + \beta_2 \ln(Income_i) + u_i$$

Z_{1i} = solo componente generica della tassa sulle vendite

Z_{2i} = solo componente specifica della tassa sulle vendite di sigarette

- Altre variabili W potrebbero essere l'effetto fisso dello stato e/o effetti temporali associati ai diversi anni (*nei modelli con dati panel*)

Il modello di regressione IV generico: notazione e terminologia

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$$

- Y_i è la variabile dipendente
- $X_{1i},\ldots, X_{ki}$ sono i regressori endogeni (potenzialmente correlati con u_i)
- $W_{1i},\ldots,W_{ri}$ sono le ***variabili esogene incluse*** o ***regressori esogenei inclusi*** (non correlati con u_i)
- $\beta_0, \beta_1,\ldots, \beta_{k+r}$ sono i coefficienti non noti della regressione
- $Z_{1i},\ldots,Z_{mi}$ sono le m variabili strumentali (le ***variaibili esogene escluse***)

Il modello di regressione IV generico, continua.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$$

Abbiamo bisogno di introdurre qualche nuovo concetto e di estendere qualche concetto vecchio al modello di regressione generico IV:

- Terminologia: *identificazione* e *sovraidentificazione*
- TSLS con variabili esogene incluse
  - un regressore endogeno
  - vari regressori endogeni
- Assunzioni che sottolineano la distribuzione campionaria normale dello TSLS
  - Validità degli strumenti (rilevanza ed esogeneità)
  - Assunzioni del modello generico di regressione IV

Identificazione

- In genere, un parametro si dice ***identificato*** se differenti valori del parametro producono differenti distribuzioni dei dati.
- Nelle regressioni IV, se i coefficienti sono identificati o meno dipende dalla relazione tra numero di strumenti (m) e numero di regressori endogeni (k)
- Intuitivamente, se ci sono meno strumenti dei regressori endogeni, non è possible stimare $\beta_1,\ldots,\beta_k$
- Per esempio, supporre che $k = 1$ ma $m = 0$ (non ci sono strumenti)!

Identificazione, cont.

I coefficienti $\beta_1,\ldots,\beta_k$ si dicono:

- ***identificati esattamente*** se $m = k$.

  Ci sono abbastanza strumenti per stimare $\beta_1,\ldots,\beta_k$.

- ***sovraidentificati*** se $m > k$.

  Ci sono più strumenti che variabili endogene per stimare $\beta_1,\ldots,\beta_k$. *In questo caso, si può testare se gli strumenti sono validi* (*un test delle "restrizioni di sovraidentificazione*") –

- ***sottoidentificati*** se $m < k$.

  Ci sono pochi strumenti per stimare $\beta_1,\ldots,\beta_k$. *In questo caso, è necessario trovare più strumenti!*

Modello Generale di regressione IV: TSLS, 1 regressore endogeno

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i$$

- Strument: $Z_{1i},\ldots,Z_m$
- Primo stadio
  - Regredire X_1 su *tutti* i regressori esogeni: regredire X_1 su $W_1,\ldots,W_r,Z_1,\ldots,Z_m$ tramite OLS
  - Calcolare il valore predetto $\hat{X}_{1i}$, $i = 1,\ldots,n$
- Secondo stadio
  - Regredire Y su $\hat{X}_1,W_1,\ldots,W_r$ tramite OLS
  - I coefficienti della regressione del secondo stadio sono gli stimatori TSLS, ma gli *SE* sono sbagliati
- Per ottenere *SE* corretti, bisogna efffettuare una stima in un unico stadio

Esempio: Demanda di sigarette

$$\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1 \ln(P_i^{cigarettes}) + \beta_2 \ln(Income_i) + u_i$$

Z_{1i} = tassa generica sulle vendite

Z_{2i} = tassa specifica sulle sigarette

- Variabile endogena: $\ln(P_i^{cigarettes})$ ("una X")
- Variabili esogene incluse: $\ln(Income_i)$ ("una W")
- **Strumenti (variabili esogene escluse)**: tassa generica sulle vendite, tassa specifica sulle sigarette ("due Z")
- *E' l'elasticità della domanda di sigarette* β_1 *sovraidentificata, esattamente identificata, o sottoidentificata?*

Esempio: Domanda di sigarette, uno strumento

```
            Y        W          X           Z
. ivreg lpackpc lperinc (lravgprs = rtaxso) if year==1995, r;

IV (2SLS) regression with robust standard errors      Number of obs =      48
                                                      F(  2,    45) =    8.19
                                                      Prob > F      =  0.0009
                                                      R-squared     =  0.4189
                                                      Root MSE      =  .18957

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.143375   .3723025    -3.07   0.004    -1.893231   -.3935191
     lperinc |    .214515   .3117467     0.69   0.495     -.413375    .842405
       _cons |   9.430658   1.259392     7.49   0.000     6.894112    11.9672
------------------------------------------------------------------------------
Instrumented:  lravgprs
Instruments:   lperinc rtaxso        STATA lists ALL the exogenous regressors
                                     as instruments - slightly different
                                     terminology than we have been using
------------------------------------------------------------------------------
```

- Calcolare IV con un solo commando produce *SE* corretti
- Usare , r per ottenere *SE* robusti all'etoeroschedasticità

Esempio: Domanda di sigarette, due strumenti

Y *W* *X* Z_1 Z_2

```
. ivreg lpackpc lperinc (lravgprs = rtaxso rtax) if year==1995, r;

IV (2SLS) regression with robust standard errors      Number of obs =      48
                                                      F(  2,    45) =   16.17
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4294
                                                      Root MSE      =  .18786

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.277424   .2496099    -5.12   0.000    -1.780164   -.7746837
     lperinc |   .2804045   .2538894     1.10   0.275     -.230955    .7917641
       _cons |   9.894955   .9592169    10.32   0.000     7.962993    11.82692
------------------------------------------------------------------------------
Instrumented:  lravgprs
Instruments:   lperinc rtaxso rtax    STATA elenca TUTTI i regressori
------------------------------------------------------------------------------
```

Esogenei come "strumenti" – terminologia livemente differente a quella utilizzat
nei lucidi

Stime TSLS, Z = tassa sulle vendite ($m = 1$)

$$\widehat{\ln}(Q_i^{cigarettes}) = 9.43 - 1.14\widehat{\ln}(P_i^{cigarettes}) + 0.21\ln(Income_i)$$
$$(1.26) \quad (0.37) \qquad\qquad (0.31)$$

Stime TSLS, Z = tassa sulle vendite, tassa specifica sulle sigarette ($m = 2$)

$$\widehat{\ln}(Q_i^{cigarettes}) = 9.89 - 1.28\widehat{\ln}(P_i^{cigarettes}) + 0.28\ln(Income_i)$$
$$(0.96) \quad (0.25) \qquad\qquad (0.25)$$

- Più bassi *SE* per $m = 2$. Usando 2 strumenti si utilizza più informazione – maggiore "variazione casuale" (variabilità nella $\hat{X}_{1i}$)
- Elasticità del reddito basso (non è un bene di lusso); L'elasticità del reddito non è statisticamente differente da zero 0

- Sorprendentemente elasticità del prezzo alta

Modello Generale di regressione IV: TSLS con regressori endogeni multipli

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$$

- Strumenti: $Z_{1i},\ldots,Z_m$
- Ci sono k regressioni nel primo stadio:
  - Regredire X_1 su $W_1,\ldots, W_r, Z_1,\ldots, Z_m$ attraverso OLS
  - Calcolare i valori predetti $\hat{X}_{1i}$, $i = 1,\ldots,n$
  - Regredire X_2 su $W_1,\ldots, W_r, Z_1,\ldots, Z_m$ con OLS
  - Calcolare i valori predetti $\hat{X}_{2i}$, $i = 1,\ldots,n$
  - Ripetere per tutte le X, ottenere $\hat{X}_{1i}$, $\hat{X}_{2i},\ldots, \hat{X}_{ki}$

TSLS con regressori endogeni multipli, continua

- Secondo stadio
  - Regredire Y su $\hat{X}_{1i}$, $\hat{X}_{2i}$,…, $\hat{X}_{ki}$, W_1,…, W_r con OLS
  - I coefficienti dal secondo stadio delle regressioni sono gli stimatori TSLS, ma *SE* sono errati
- Per ottenere *SE* corretti, bisogna calcolare tutto in un unico passo
- *Cosa succederebbe nella regressione del secondo stadio se i coefficienti fossero sotto identificati (cioè, se #strumenti < #variabili endogene); per esempio, se k = 2, m = 1?*

Distribuzione campionaria dello stimatore TSLS nel modello generico di regressione IV

- Significato di strumento "valido" nel caso generale
- Le assunzioni della regressione IV
- Implicazioni: se le assunzioni della regressione IV sono valide, allora lo stimatore TSLS è distribuito normalmente, e per l'inferenza (test, intervalli di confidenza) si procede come al solito

Una serie di strumenti "validi" nel caso generico

Gli strumenti devono essere rilevanti ed esogeni:

1. Rilevanza dello strumento: *Caso speciale di una sola X*

   Almeno uno strumento deve entrare nella popolazione della regressione del primo stadio.

2. Esogeneità dello strumento

   Tutti gli strumenti sono non correlati con il termine di errore: $\text{corr}(Z_{1i},u_i) = 0,\ldots, \text{corr}(Z_m,u_i) = 0$

"Validi" strumenti nel caso generico, cont.

(1) Condizione della rilevanza dello strumento nel caso generico:

- *Caso generico, multiple X*

  Supporre che la regressione del secondo stadio potrebbe essere calcolata usando i valori predetti della *popolazione* nella regressione del primo stadio.

  Quindi: non c'è perfetta multicollinearità in questa regressione del secondo stadio (impossibile)

- *Caso speciale di una X*

  Almeno uno strumento deve entrare nella popolazione della retta di regressione del primo stadio.

Le Assunzioni della Regressione IV

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$$

1. $E(u_i | W_{1i}, \ldots, W_{ri}) = 0$
2. $(Y_i, X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}, Z_{1i}, \ldots, Z_{mi})$ sono i.i.d.
3. Le X, W, Z, e Y hanno momenti quarti, finiti
4. Le W non sono perfettamente multicollineari
5. Gli strumenti $(Z_{1i}, \ldots, Z_{mi})$ soddisfano le condizioni per la validità degli strumenti.

- #1 dice che "i regressori esogeni sono esogeni."
- #2 – #4 sono come al solito; le abbiamo discusse.
- #5.

Implicazioni: Distribuzione campionaria dello stimatore TSLS

- Se le assunzioni della regressione IV sono valide, allora lo stimatore TSLS è distribuito normalmente in grandi campioni.
- Per l'inferenza (verifica delle ipotesi, intervalli di confidenza) si procede come al solito.
- Due osservazioni sugli errori standard:
  - Gli *SE* del secondo stadio non sono corretti perchè non considerano la stima del primo stadio; per ottenere *SE* corretti, utilizzare il commando del software econometrico (GRETL o STATA) TSLS
  - Usare gli *SE*, robusti all'eteroschedasticità.
- *Tutto questo si basa sul fatto che abbiamo strumenti validi…*

Controllare la Validità degli Strumenti

Ricordare i due requisiti per la validità dello strumento:

1. *Rilevanza* (caso speciale di una X)

   Almeno uno strumento deve entrare nella popolazione della regressione del primo stadio.

2. *Esogeneità*

   Tutti gli strumenti devono essere non correlati con il termini di errore: $\text{corr}(Z_{1i},u_i) = 0,\ldots, \text{corr}(Z_{mi},u_i) = 0$

Cosa succeed se uno di questi requisiti non è soddisfatto?
Come si può controllare? E cosa bisogna fare?

Controllare l'Assunzione #1: Rilevanza dello strumento

Concentriamoci sul caso di un singolo regressere endogeno:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i$$

Regressione del primo stadio:

$$X_i = \pi_0 + \pi_1 Z_{1i} + \ldots + \pi_{mi} Z_{mi} + \pi_{m+1i} W_{1i} + \ldots + \pi_{m+ki} W_{ki} + u_i$$

- Gli strumenti sono rilevanti se almeno uno di $\pi_1,\ldots,\pi_m$ è non zero.
- Gli strumenti sono detti ***deboli*** se tutte le $\pi_1,\ldots,\pi_m$ sono zero o vicino allo zero .
- ***Strumenti deboli*** spiegano poco della variazione in X, oltre a quella spiegata dale W

Quali sono le conseguenze di strumenti deboli?

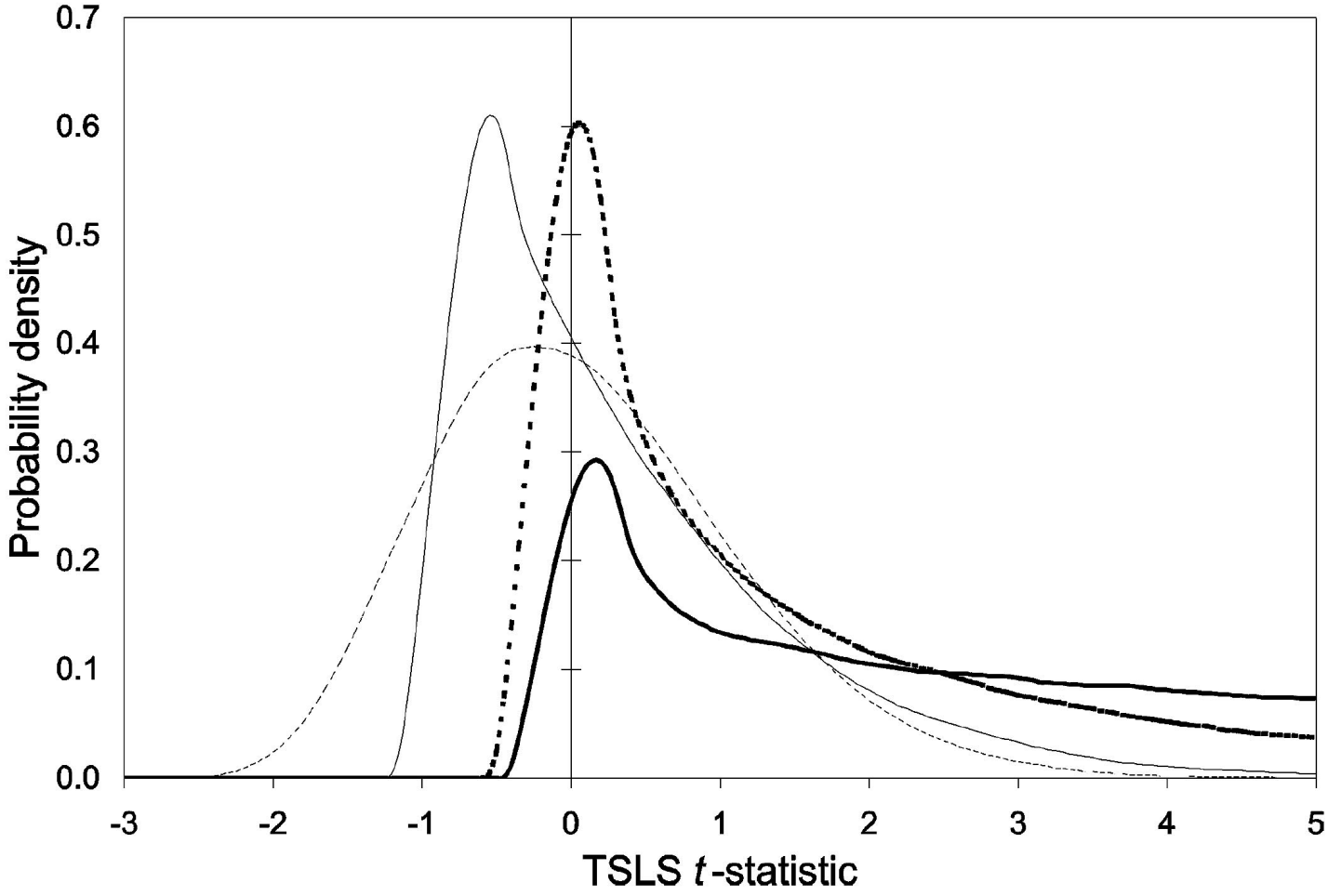
Consideraimo il caso semplice:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$X_i = \pi_0 + \pi_1 Z_i + u_i$$

- Lo stimatore IV è $\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$
- Se cov(X,Z) è zero o piccola, allora s_{XZ} sarà piccolo: Con strumenti deboli, il denominatore è prossimo allo zero.
- In questo caso, la distribuzione campionaria di $\hat{\beta}_1^{TSLS}$ (e la sua statistica t) non è approssimata in grandi campioni in modo corretto da n approssimazioni normali…

Un esempio: la distribuzione TSLS della statistica *t* con strumenti deboli



Linea nera = strumenti irrilevanti

Linea punteggiata chiara = strumenti forti

Perchè non possiamo applicare l'approssimazione normale!?!

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$$

- Se cov(X,Z) è piccola, piccole variazioni in s_{XZ} (da un campione ad un altro) possono determinare grandi variazioni in $\hat{\beta}_1^{TSLS}$
- Supporre che in un campione si abbia s_{XZ} = .00001!
- Quindi l'approssimazione normale per grandi -n è un'approssimazione molto debole della distribuzione campionaria di $\hat{\beta}_1^{TSLS}$
- Una migliore approssimazione implica che $\hat{\beta}_1^{TSLS}$ è distribuito come un *tasso* di due variabili casuali normali correlate
- Se gli strumenti sono deboli, i metodi di solito utilizzati per l'inferenza non sono attendibili – spesso molto sbagliati.

Misurare praticamente la forza degli strumenti: La statistica *F* del primo stadio

- La regressione del primo stadio (una X):

  Regredisce X su $Z_1,..,Z_m,W_1,\ldots,W_k$.

- Strumenti completamente irrilevanti, → *tutti* i coefficienti su $Z_1,\ldots,Z_m$ sono zero.
- La ***statistica F del primo stadio*** verifica l'ipotesi che $Z_1,\ldots,Z_m$ non entrano nel primo stadio della regressione.
- Strumenti deboli implicano un valore basso della statistica F del primo stadio.

Controllare strumenti deboli con una sola *X*

- Calcolare la statistica *F* nel primo stadio.

  Regola pratica: Se la statistica F nel primo stadio è inferiore di 10, allora concludere che lo strumento è debole.

- In questo caso, lo stimatore TSLS sarà distorto, è l'inferenza statistica (errori standard, test delle ipotesi, intervalli di confidenza) sono fuorvianti.
- Notare che il semplice rifiuto dell'ipotesi nulla che i coefficienti delle *Z* sono zero non è sufficiente– bisogna essere certi che l'ipotesi dell'approssimazione normale sia corretta.
- Ci sono delle tecniche più sofisticate di quella di verificare se il valore di *F* sia maggiore di 10.

Cosa bisogna fare se abbiamo strumenti deboli?

- Cercare strumenti migliori (!)
- Se si hanno molti strumenti, alcuni dei quali probabilmente più deboli di altri è una buona idea eliminare i più deboli (eliminare gli strumenti non rilevanti che accrescono il valore della *F* nel primo stadio)
- Usare uno stimatore delle IV differente da TSLS
    - Ci sono molti stimatori IV disponibili quando i coefficienti sono sovraidentificati.
    - La funzione di verosimiglianza per informazione limitata *Limited information maximum likelihood* è uno stimatore meno sensibile agli strumenti deboli.
    - *Questi argomenti fanno parte di corsi più avanzati di econometria e quindi non verranno trattati*...

Controllare l'Assunzione #2: Esogeneità degli strumenti

- Esogeneità degli strumenti: ***Tutti*** gli strumenti sono non correlati con il termine di errore corr(Z_{1i},u_i) = 0,…, corr(Z_{mi},u_i) = 0
- Se gli strumenti sono correlati con il termine dell'errore, il primo stadio di TSLS non riesce a isolare la componente di X che non è correlata con il termine di errore, quindi $\hat{X}$ è correlata con u e TSLS è non consistente.
- Se ci sono più strumenti di regressori endogeni, è possible testare – *in parte* – se gli strumenti sono esogeni.

Verificare le restrizione di sovraidentificazione

Considerare il caso semplice:

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

- Supporre che ci siano due validi strumenti: Z_{1i}, Z_{2i}
- Si potrebbe calcoare due stime differenti del TSLS.
- Intuitivamente, se le 2 TSLS stime sono molto differenti l'una dall'altra, allora c'è qualcosa di sbagliato: uno o l'altro, o entrambi gli strumenti non sono validi.
- Il *J*-test odelle restrizioni di sovraidentificazione confrontano questo caso statisticamente.
- Questo test può essere applicato solo se $\#Z > \#X$ (sovraidentificazione).

Supporre che #strumenti = $m >$ # $X = k$ (sovraidentificato)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$$

Il *J*-test delle restrizioni di sovraidentificazione

1. Stimare l'equazione di interesse utilizzando TSLS e tutti gli strumenti m; calcolare il valore predetto $\hat{Y}_i$, usando il valore *attuale* X (non $\hat{X}$ utilizzato nel secondo stadio)
2. Calcolare i residui $\hat{u}_i = Y_i - \hat{Y}_i$
3. Regredire $\hat{u}_i$ su $Z_{1i},\ldots,Z_{mi}$, $W_{1i},\ldots,W_{ri}$
4. Calcolare la statistica F per verificare l'ipoteche che tutti i coefficienti su $Z_{1i},\ldots,Z_{mi}$ sono zero;
5. La ***J-statistica*** è $J = mF$
6. $J = mF$, dove $F =$ la statistica-F verifica i coefficienti $Z_{1i},\ldots,Z_{mi}$ nella regressione TSLS dei residui su $Z_{1i},\ldots,Z_{mi}$, $W_{1i},\ldots,W_{ri}$.

Distribuzione della statistica-*J*

- Sotto l'ipotesi nulla che tutti gli strument sono esogeni, *J* ha una distribuzione chi-quadro con $m–k$ gradi di libertà
- Se $m = k$, $J = 0$ (*ha senso?*)
- Se alcuni strumenti sono esogeni e altri endogeni, la statistica *J* sarà grande, e l'ipotesi nulla che tutti gli strumensi sono esogeni sarà rifiutata.

Applicazione alla domanda di sigarette

Perchè siamo interessati a conoscere l'elasticità della domanda di sigarette?

- Teoria della tassazione ottima: la tassa ottima è inversa all'elasticità: minore è la perdita se la quantià interessata è minore.
- Esternalità del fumo – ruolo dell'intervento governativo per scoraggiare il fumo
  - fumatori passivi (non monetario)
  - esternalità monetarie

Panel data set

- Consumo di sigarette annuale, prezzo medio pagato dal consumatore (incluse le tasse), reddito personale
- 48 stati continentali negli US, 1985-1995

Strategia della Stima

- Avendo dati panel, possiamo controllare per le caratteristiche non osservabili che influenzano la domanda di sigaretta, solo se sono costanti nel tempo
- Dobbiamo comunque usare lo stimatore IV che permetta di risolvere il problema dell'errore dovuto causalità simultanea che dipende dall'interazione tra domanda e offerta.

Modello a effetti-fissi della domanda di sigaretta

$$\ln(Q_{it}^{cigarettes}) = \alpha_i + \beta_1 \ln(P_{it}^{cigarettes}) + \beta_2 \ln(Income_{it}) + u_{it}$$

- $i = 1,\ldots,48$, $t = 1985, 1986,\ldots,1995$
- α_i riflette i fattori non osservabili omessi che variano tra stati ma non nel tempo, per esempio l'attidudine al fumo
- Ancora, corr($\ln(P_{it}^{cigarettes})$,u_{it}) è plausibilmente differente da zero perchè c'è relazione offerta/domanda
- Strategia di stima:
  - Usare la regressione con il metodo di dati panel per eliminare α_i
  - Usare TSLS per risolvere il problema dell'errore dovuto alla causalità simultanea

Panel data IV regressione: due approcci

(a) Il metodo con "n-1 indicatori binari"

(b) Il metodo delle "variazioni" (quando T=2)

(a) Il metodo con "n-1 indicatori binari"

Rescrivere

$$\ln(Q_{it}^{cigarettes}) = \alpha_i + \beta_1 \ln(P_{it}^{cigarettes}) + \beta_2 \ln(Income_{it}) + u_{it}$$

come

$$\ln(Q_{it}^{cigarettes}) = \beta_0 + \beta_1 \ln(P_{it}^{cigarettes}) + \beta_2 \ln(Income_{it}) + \gamma_2 D2_{it} + \ldots + \gamma_{48} D48_{it} + u_{it}$$

Strumenti:

Z_{1it} = tassa generica sulle vendite

Z_{2it} = tassa specifica sulle sigarette

Questo modello potrebbe essere stimato con il modello generido della regressione con IV:

$$\ln(Q_{it}^{cigarettes}) = \beta_0 + \beta_1 \ln(P_{it}^{cigarettes}) + \beta_2 \ln(Income_{it}) + \gamma_2 D2_{it} + \ldots + \gamma_{48} D48_{it} + u_{it}$$

- X (regressore endogeno) = $\ln(P_{it}^{cigarettes})$
- 48 W (regressori esogeni inclusi) = $\ln(Income_{it})$, $D2_{it}, \ldots, D48_{it}$
- Due strumenti = Z_{1it}, Z_{2it}
- Il modello può essere stimato con TSLS!
- Una questione emerge quando la risposta dinamica è importante (ritardo nell'aggiustamento), come nel nostro esempio – serve tempo per smettere di fumare – come si modella l'effetto ritardato?

(b) Il metodo delle "variazioni" (quando T=2)

- Un modo per modellare gli effetti di lungo periodo è di considerare le variazione ogni 10 anni, tra il 1985 e il 1995
- Riscrivere il modello di regressione in forma di "variazioni":

$$\ln(Q_{i1995}^{cigarettes}) - \ln(Q_{i1985}^{cigarettes})$$
$$= \beta_1[\ln(P_{i1995}^{cigarettes}) - \ln(P_{i1985}^{cigarettes})]$$
$$+\beta_2[\ln(Income_{i1995}) - \ln(Income_{i1985})]$$
$$+ (u_{i1995} - u_{i1985})$$

- Bisogna calcolare "variazione in 10-anni" delle variabili, per esempio:

  10-anni di variazione nel log prezzo = $\ln(P_{i1995}) - \ln(P_{i1985})$
- Poi stimare l'elasticità della domanda con TSLS usando variazioni nei 10-anni delle variabili strumentali
- Useremo questo approccio

STATA: Domanda di sigarette

Prima calcoliamo "10-anni di variazione" delle variabili

10-anni di variazione nel log del prezzo

$$= \ln(P_{it}) - \ln(P_{it-10}) = \ln(P_{it}/P_{it-10})$$

```
. gen dlpackpc = log(packpc/packpc[_n-10]);      _n-10 is the 10-yr lagged value
. gen dlavgprs = log(avgprs/avgprs[_n-10]);
. gen dlperinc = log(perinc/perinc[_n-10]);
. gen drtaxs  = rtaxs-rtaxs[_n-10];
. gen drtax   = rtax-rtax[_n-10];
. gen drtaxso = rtaxso-rtaxso[_n-10];
```

Usare TSLS per stimare l'elasticità della domanda utilizzando la specificazione "10-anni di variazioni"

```
          Y          W       X          Z
. ivreg dlpackpc dlperinc (dlavgprs = drtaxso) , r;

IV (2SLS) regression with robust standard errors      Number of obs =      48
                                                      F(  2,    45) =   12.31
                                                      Prob > F      =  0.0001
                                                      R-squared     =  0.5499
                                                      Root MSE      =  .09092

------------------------------------------------------------------------------
             |               Robust
    dlpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    dlavgprs |  -.9380143   .2075022    -4.52   0.000    -1.355945   -.5200834
    dlperinc |   .5259693   .3394942     1.55   0.128    -.1578071    1.209746
       _cons |   .2085492   .1302294     1.60   0.116    -.0537463    .4708446
------------------------------------------------------------------------------
Instrumented:  dlavgprs
Instruments:   dlperinc drtaxso
------------------------------------------------------------------------------
```

NOTA:

- *Tutte le variabili – Y, X, W, e Z – sono in 10-anni di variazione*
- *Elasticità stimata = –.94 (SE = .21) – sorprendentemente elastica!*
- *Elasticità del reddito bassa, non statisticamente differente da zero*
- *Bisogna controllare che lo strumento sia rilevante…*

Controllare la rilevanza dello strumento: calcolare *F* nel primo stadio

```
. reg dlavgprs drtaxso dlperinc , r;

Regression with robust standard errors                Number of obs =      48
                                                      F(  2,    45) =   16.84
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5146
                                                      Root MSE      =  .06334

------------------------------------------------------------------------------
             |               Robust
    dlavgprs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     drtaxso |   .0254611   .0043876     5.80   0.000      .016624    .0342982
    dlperinc |  -.2241037   .2188815    -1.02   0.311    -.6649536    .2167463
       _cons |   .5321948   .0295315    18.02   0.000     .4727153    .5916742
------------------------------------------------------------------------------

. test drtaxso;

 ( 1)  drtaxso = 0

       F(  1,    45) =   33.67
            Prob > F =    0.0000
```

*We didn't need to run "test" here because with m=1 instrument, the F-statistic is the square of the t-statistic, that is, 5.80*5.80 = 33.67*

Primo stadio F = 33.7 > 10 quindi lo strumento non è deobole

Possiamo controllare l'esogenietà dello strumento? *No…m = k*

E se abbiamo due strumenti (tassa specifica sulle sigarette e tassa generica sulle vendite)?

```
. ivreg dlpackpc dlperinc (dlavgprs = drtaxso drtax) , r;

IV (2SLS) regression with robust standard errors      Number of obs =      48
                                                      F(  2,    45) =   21.30
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5466
                                                      Root MSE      =  .09125
------------------------------------------------------------------------------
             |               Robust
    dlpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    dlavgprs |  -1.202403   .1969433    -6.11   0.000    -1.599068   -.8057392
    dlperinc |   .4620299   .3093405     1.49   0.142    -.1610138    1.085074
       _cons |   .3665388   .1219126     3.01   0.004     .1209942    .6120834
------------------------------------------------------------------------------
Instrumented:  dlavgprs
Instruments:   dlperinc drtaxso drtax
------------------------------------------------------------------------------
```

drtaxso = solo tassa generica sulle vendite

drtax = solo tassa specifica sulle sigarette

L'elasticità stimata è -1.2, più elastica del caso in cui si utilizzi solo la tassa generica sulle vendite

Con *m*>*k*, possiamo verificare le restrizioni di sovraidentificazione. Test delle restrizioni di sovraidentificazione

```
.  predict e, resid;      Calcola i valori predetti per le stima stima della
regressione (la regressione precedente TSLS)
.  reg e drtaxso drtax dlperinc;     Regress e on Z's and W's

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  3,    44) =    1.64
       Model |  .037769176     3  .012589725            Prob > F      =  0.1929
    Residual |  .336952289    44  .007658007            R-squared     =  0.1008
-------------+------------------------------           Adj R-squared =  0.0395
       Total |  .374721465    47  .007972797            Root MSE      =  .08751
------------------------------------------------------------------------------
           e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     drtaxso |   .0127669   .0061587     2.07   0.044      .000355    .0251789
       drtax |  -.0038077   .0021179    -1.80   0.079     -.008076    .0004607
    dlperinc |  -.0934062   .2978459    -0.31   0.755    -.6936752    .5068627
       _cons |    .002939   .0446131     0.07   0.948    -.0869728    .0928509
------------------------------------------------------------------------------
.  test drtaxso drtax;
 ( 1)  drtaxso = 0                  Calcola J-statistic, cioè m*F,
 ( 2)  drtax = 0                    dove F verifica se i coefficienti degli
                                    strumenti sono tutti zero
       F(  2,    44) =    2.47            so J = 2 + 2.47 = 4.93
       Prob > F =    0.0966      ** ATTENZIONE - questa utilizza errati d.f. **
```

I gradi di libertà corretti per la statistica-J sono $m–k$:

- $J = mF$, dove F = la statistica-F verifica se i coefficienti su $Z_{1i},\ldots,Z_{mi}$ nella regressione TSLS dei residui su $Z_{1i},\ldots,Z_{mi}$, $W_{1i},\ldots,W_{mi}$.
- Sotto l'ipotesi null ache tutti gli strumenti sono esogeni, J ha una distribuzione chi-quadro con $m–k$ gradi di libertà
- In questo caso, $J = 4.93$, distribuita come un chi-quadro con (gradi di libertà) d.f. = 1; il 5% valore critico è 3.84, quindi rifiutiamo ad un livello di significatività del 5%.
- In STATA:

```
.  dis "J-stat = " r(df)*r(F) "  p-value = "  chiprob(r(df)-1,r(df)*r(F));
J-stat = 4.9319853  p-value = .02636401
```

J = 2x2.47 = 4.93 *p-value dalla distribuzione chi-quadro(1)*

Controllare la rilevanza dello strumento: Calcolare *F* nel primo stadio

X *Z1* *Z2* *W*

```
. reg dlavgprs drtaxso drtax dlperinc , r;

Regression with robust standard errors                Number of obs =      48
                                                      F(  3,    44) =   66.68
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.7779
                                                      Root MSE      =  .04333

------------------------------------------------------------------------------
             |               Robust
    dlavgprs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     drtaxso |    .013457   .0031405     4.28   0.000     .0071277    .0197863
       drtax |   .0075734   .0008859     8.55   0.000     .0057879    .0093588
    dlperinc |  -.0289943   .1242309    -0.23   0.817    -.2793654    .2213767
       _cons |   .4919733   .0183233    26.85   0.000     .4550451    .5289015
------------------------------------------------------------------------------
. test drtaxso drtax;

 ( 1)  drtaxso = 0
 ( 2)  drtax = 0

       F(  2,    44) =   88.62          88.62 > 10 so instruments aren't weak
            Prob > F =    0.0000
```

Tabella sintetica dei risultati:

TABLE 10.1 Two Stage Least Squares Estimates of the Demand for Cigarettes Using Panel Data for 48 U.S. States

Dependent variable: $\ln(Q^{cigarettes}_{i,1995}) - \ln(Q^{cigarettes}_{i,1985})$

Regressor	(1)	(2)	(3)
$\ln(P^{cigarettes}_{i,1995}) - \ln(P^{cigarettes}_{i,1985})$	−0.94** (0.21)	−1.34** (0.23)	−1.20** (0.20)
$\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})$	0.53 (0.34)	0.43 (0.30)	0.46 (0.31)
Intercept	0.21 (0.13)	0.45** (0.14)	0.37** (0.12)
Instrumental variable(s)	Sales tax	Cigarette-specific tax	Both sales tax and cigarette-specific tax
First-stage *F*-statistic	33.70	107.20	88.60
Overidentifying restrictions *J*-test and *p*-value	-	-	4.93 (0.026)

These regressions were estimated using data for 48 U.S. states (48 observations on the ten-year differences). The data are described in Appendix 10.1. The *J*-test of overidentifying restrictions is described in Key Concept 10.6 (its *p*-value is given in parentheses), and the first-stage *F*-statistic is described in Key Concept 10.5. Individual coefficients are statistically significant at the *5% level or **1% significance level.

Come si interpreta il rifiuto del test *J*?

- *J*-test rifiuta l'opotesi nulla che entrambi strumenti sono esogeni
- Questo significa che o *rtaxso* è endogeno, o *rtax* è endogena, o entrambi
- *J*-test non indica quale!! E' necessario riflettere:
- Perchè *rtax* (tassa specifica sulle sigarette) potrebbe essere endogena?
  - Forze politiche: la storia del fumo o molti dei fumatori ≈ pressione politica per abbassare le tasse sulle sigarette
  - In questo caso, la tassa specifica sulle sigarette è endogena
- Questo ragionamento non si applica alla tassa generica sulle vendite
- → usare solo uno strumento, la tassa generica sulle vendite

La Domanda di Sigarette: Sintesi dei Risultati Empirici

- Usare l'elasticità stimata basata su TSLS con la tassa generica sulle vendite come unico strumento:

  Elasticità = -.94, *SE* = .21

- Questa elasticità è sorprendentemente grande (non inelastica) – un aumento nel prezzo del 1% reduce la vendita di sicarette di circa 1%. Questa elasticità è molto maggiore di quella comunemente assunta nella letteratura degli studi economici sulla salute.
- Questa è l'elasticità di lungo periodo (variazione di 10 anni). *Cosa ti aspetteresti come valore dell'elasticità di breve periodo (variazione di un anno) – più o meno elastica?*

Quali sono le minacce rimanenti per la validità interna?

- Errore da omissione di variabili ?
  - *Stimatore dei dati Panel; probabilmente OK*

- Errata specificazione della forma funzionale
  - Hmmm…necessario controllare…
  - Una questione correlate è l'interpretazione dell'elasticità: usare una differenza di 10 anni, l'interpretazione è elasticità di lungo periodo. Stime differenti si otterrebbero usando differenze nel breve periodo.

Minacce rimanenti alla validità interna, continua

- Rimane l'errore per causalità simultanea?
  - No se lo strumento "tassa generale sulle vendite" è uno strumento valido:
    - rilevante?
    - esogeno?
- Distorsione per errori nelle variabili? *Domanda interessante*: *quanto accuratamente è misurato il prezzo effettivamente pagato? E i prezzi relativi alle vendite vicino ai confini con altri paesi?*
- Distorsione per selezione campionaria? (*no, abbiamo tutti gli stati*)

Concludiamo, che questa è una stima attendibile dell'elasticità della domandi di lungo periodo sebbene alcuni problemi potrebbero ancora essere presenti.

Come otteniamo strumenti validi?

- Strumenti validi sono (1) rilevanti e (2) esogeni
- Un modo generico per trovare strumenti è cercare per la variazione esogena – una variazione che "fosse come" casualmente assegnata in un esperimento casuale controllato – che influenza la *X*.
  - La pioggia sposta la curva di offerta non non la curva di domanda; la pioggia "è come se fosse" assegnata casualmente
  - La tassa sulle vendite sposta la curva di offerta delle sigarette ma non la curva di domanda; la tassa sulle vendite è come "se fosse" assegnata casulamete

• Ecco un esempio…

Esempio: Cateterizzazione Cardiaca

La Cateterizzazione Cardiaca migliora la longevità dei pazienti che hanno subito un infarto?

Y_i = tempo di sopravvivenza (in giorni) di un paziente che ha subito un infarto

X_i = 1 se il paziente riceve la cateterizzazione cardiaca,

= 0 in caso contrario

- Le prove cliniche mostrano che *CardCath* influenza *SurvivalDays*.
- Ma è questo un trattamento effettivo "in campo medico"?

$$SurvivalDays_i = \beta_0 + \beta_1 CardCath_i + u_i$$

- E' OLS corretto? La decisione di curare un paziente con la cateterizzazione cardiaca è endogena – è (*fu*) presa nel campo medito dal tecnico EMT e dipende da u_i (caratteristiche non osservate della salute dei pazienti)
- Se i pazienti hanno caratteristiche di salute buone, allora OLS ha un errore da causalità simultanea e OLS sovrastima l'effetto CC
- Strumento proposto: distanza dall'ospedale più vicino che effettua CC – distanza al più vicino ospedale "regolare"

- Z = distanza dall'ospedale CC
  - Rilevante? Se un ospedale CC è lontano, i pazienti non verrranno presi e non saranno trattati con CC
  - Esogeno? Se la distanza dall'ospedale CC hospital non influenza la sopravvivenza, oltre che attraverso l'effetto della $CardCath_i$, allora corr(distance,u_i) = 0 quindi esogeno
  - Se la località dei pazienti è casuale, allora la distanza è "come se fosse" assegnata casulamente.
  - *Il primo stadio è un modello di probabilità lineare: la distanza influenza la probabilità di ricevere il trattamento*
- Risultati (McClellan, McNeil, Newhous, *JAMA*, 1994):
  - Stune OLS significative e gradi valori dell'effetto di CC
  - TSLS stime basse, spesso effetto non significativo

Sintesi: Regressione IV

- Uno strumetno valido ci permette di isolare la parte di X che non è correlata con u, e questa parte può essere utilizzata per stimare l'effetto di cambi in X su Y
- La regressione IV richiede strumenti validi:

  (1) *Rilevanza*: controlla via primo stadio - first-stage - F

  (2) *Esogeneità*: Verificare attraverso i test le restrizione di sovra identificazione - *over*identifying restrictions -, via la statistica-J - J-statistic-.
- Uno strumento valido isola la variazione in X che è "come se fosse" assegnata in modo casuale.
- Il requisito critico di almeno m strumenti validi non può essere testato – *you must use your head.*