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:

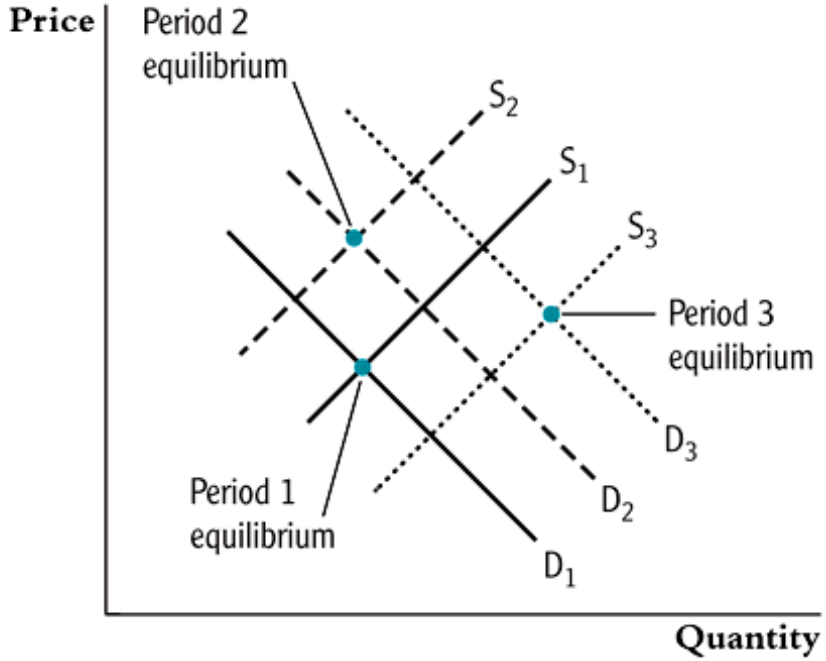
$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

- 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.

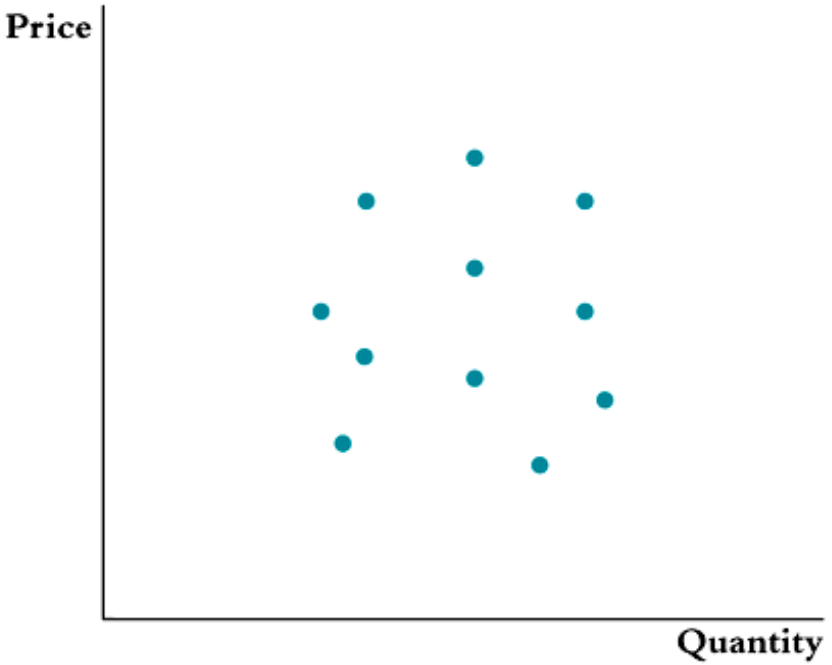


(a) Demand and Supply in Three Time Periods

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?



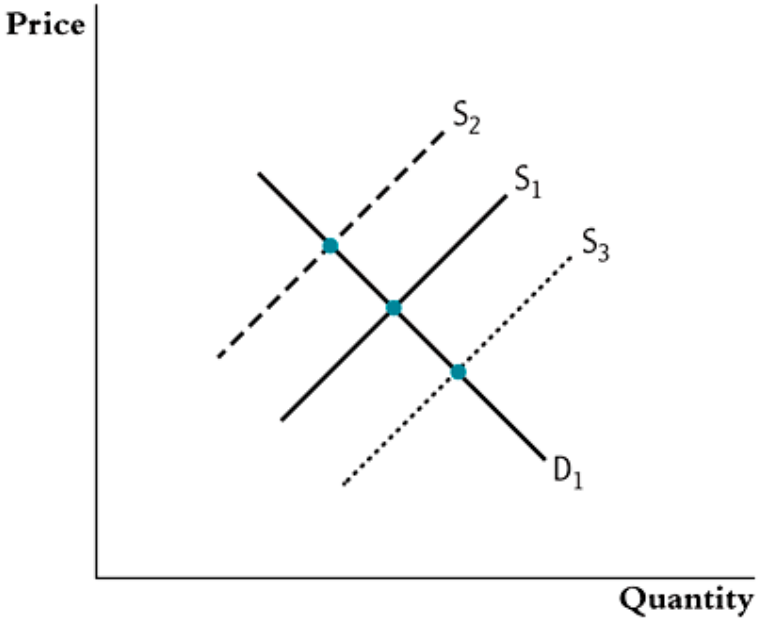
(b) Equilibrium Price and Quantity for Eleven Time Periods

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.

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 $\hat{X}$.

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

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

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.