Econometria | 2022/2023

Lezione 12: Variabili Strumentali

Giuseppe Ragusa
https://gragusa.org

Roma, maggio 2023

Sommario

Regressione IV: perché?

Tre importanti minacce alla validità interna sono:

1. Distorsione da variabili omesse per una variabile correlata con X ma inosservata (perciò non può essere inclusa nella regressione) e per cui vi sono variabili di controllo inadeguate;
2. Distorsione da causalità simultanea (X causa Y, Y causa X);
3. Distorsione da errori nelle variabili (X è misurata con errore)

Tutti e tre i problemi comportano $E(u|X) \neq 0$.

- La regressione con variabili strumentali può eliminare la distorsione quando $E(u|X) \neq 0$ – usando una variabile strumentale (IV), Z.

Lo stimatore IV con un singolo regressore e un singolo strumento (Paragrafo 12.1)

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

- La regressione IV divide X in due parti:
  - una che potrebbe essere correlata con u, e
  - una che non lo è.

Isolando la parte che non è correlata con u, è possibile stimare β_1.

- Per fare questo si utilizza una variabile strumentale, Z_i, che è correlata con X_i ma incorrelata con u_i

Terminologia: endogeneità ed esogeneità

- Una variabile endogena è una variabile correlata con u.
- Una variabile esogena è una variabile incorrelata con u.

Nella regressione IV ci concentriamo sul caso in cui X è endogena ed esiste uno strumento, Z, esogeno.

Digressione sulla terminologia: "endogeno" significa letteralmente "determinato all'interno del sistema". Se X è congiuntamente determinata con Y, allora una regressione di Y su X è soggetta a distorsione da causalità simultanea. Ma questa definizione di endogeneità è troppo stretta perché sia possibile usare la regressione IV per risolvere i problemi di distorsione da variabili omesse e da errori nelle variabili, quindi usiamo la definizione più ampia fornita sopra.

Due condizioni per avere uno strumento valido

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

Perché una variabile strumentale (uno "strumento") Z sia valida, deve soddisfare due condizioni:

1. Rilevanza: $cor(Z_i, X_i) \neq 0$
2. Esogeneità: $cor(Z_i, u_i) = 0$

Supponiamo per ora di avere un tale Z_i (vedremo più avanti come trovare variabili strumentali); come possiamo usarlo per stimare β_1?

Lo stimatore IV con una X e una Z

Spiegazione 1: minimi quadrati in due stadi (TSLS)

Ci sono due stadi - due regressioni:

1. Si isola la parte di X che non è correlata con u mediante la regressione di X su Z usando gli OLS:

$$X_i = \pi_0 + \pi_1 Z_i + \nu_i \quad (1)$$

- Poiché Z_i non è correlato con u_i, $\pi_0 + \pi_1 Z_i$ non è correlato con u_i. Non conosciamo π_0 o π_1 ma li abbiamo stimati, perciò…
- Si calcolano i valori predetti di X_i, $\hat{X}_i$

$$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i, \quad i = 1, \ldots, n$$

Minimi quadrati in due stadi (continua)

2. Si sostituisce X_i con $\hat{X}_i$ nella regressione di interesse e si esegue la regressione di Y su usando gli OLS:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i$$

- Poiché $\hat{X}_i$ è incorrelato con u_i, la prima assunzione dei minimi quadrati vale per la regressione (2). (Ciò richiede che n sia grande in modo che π_0 e π_1 siano stimati con precisione)
- Quindi, in grandi campioni, β_1 può essere stimato con gli OLS usando la regressione (2)
- Lo stimatore risultante è detto stimatore dei minimi quadrati in due stadi (TSLS), $\hat{\beta}_1^{TSLS}$

Minimi quadrati in due stadi: riepilogo

Supponiamo che Z_i, soddisfi le due condizioni per uno strumento valido:

1. Rilevanza: $cor(Z_i, X_i) \neq 0$
2. Esogeneità: $cor(Z_i, u_i) = 0$

Minimi quadrati in due stadi:

Stadio 1: Regressione di X_i su Z_i (inclusa intercetta), ottenendo i valori predetti $\hat{X}_i$

Stadio 2: Regressione di Y_i su (inclusa intercetta); il coefficiente di $\hat{X}_i$ è lo stimatore TSLS, $\hat{\beta}_1^{TSLS}$.

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

Lo stimatore IV, una X e una Z (continua)

Spiegazione 2: derivazione algebrica diretta

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

Allora:

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

dove $cov(u_i, Z_i) = 0$ per l'esogeneità dello strumento; quindi:

$$\frac{cov(Y_i, Z_i)}{}$$

Lo stimatore IV, una X e una Z (continua)

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

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

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

s_{YZ} e s_{XZ} sono covarianze campionarie. Questo è lo stimatore TSLS - con una derivazione diversa!

Lo stimatore IV, una X e una Z (continua)

Spiegazione 3: derivazione dalla "forma ridotta"

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$
$$X_i = \pi_0 + \pi_1 Z_i + \nu_i$$

Sostituiamo X nell'equazione di Y:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i = \\ &= \beta_0 + \beta_1 \left[\pi_0 + \pi_1 Z_i + \nu_i\right] + u_i = \\ &= \underbrace{\left[\beta_0 + \beta_1 \pi_0\right]}_{\gamma_0} + \underbrace{\left(\beta_1 \pi_1\right)}_{\gamma_1} Z_i + \underbrace{\left[\beta_1 \nu_i + u_i\right]}_{\omega_i} = \\ &= \gamma_0 + \gamma_1 Z_i + \omega_i \end{aligned}$$

- Visto che $\beta_1 \pi_1 = \gamma_1, \beta = \gamma_1 / \pi_1$

Effetto dello studio sui voti

Stinebrickner, Ralph and Stinebrickner, Todd R. (2008) “The Causal Effect of Studying on Academic Performance,” The B.E. Journal of Economic Analysis & Policy: Vol. 8

$$X_i = \pi_0 + \pi_1 Z_i + \nu_i$$
$$Y_i = \gamma_0 + \gamma_1 Z_i + \omega_i$$

$$Y = \text{media voti (scala 4 punti)}$$
$$X = \text{tempo di studio (ore al giorno)}$$
$$Z = 1 \text{ se il compagno ha portato un videogioco, } = 0 \text{ altrimenti}$$

- $n = 210$ studenti del primo anno al Berea College (Kentucky) nel 2001

Effetto dello studio sui voti

$$
\begin{aligned}
X_i &= \pi_0 + \pi_1 Z_i + \nu_i \\
Y_i &= \gamma_0 + \gamma_1 Z_i + \omega_i
\end{aligned}
$$

$$
\begin{aligned}
Y &= \text{media voti (scala 4 punti)} \\
X &= \text{tempo di studio (ore al giorno)} \\
Z &= 1 \text{ se il compagno ha portato un videogioco, } = 0 \text{ altrimenti}
\end{aligned}
$$

Pensate che Z_i (indica se un compagno ha portato un videogioco) sia uno strumento valido?

1. È rilevante (correlato con X)?
2. È esogeno (incorrelato con u)?

Effetto dello studio sui voti (continua)

$$X_i = \pi_0 + \pi_1 Z_i + \nu_i$$
$$Y_i = \gamma_0 + \gamma_1 Z_i + \omega_i$$

$$Y = \text{media voti (scala 4 punti)}$$
$$X = \text{tempo di studio (ore al giorno)}$$
$$Z = 1 \text{ se il compagno ha portato un videogioco, } = 0 \text{ altrimenti}$$

Risultati di Stinebrinckner e Stinebrinckneri:

$$\hat{\pi}_1 = -0.668$$
$$\hat{\gamma}_1 = -0.241$$
$$\hat{\beta}_1^{IV} = \frac{\hat{\gamma}_1}{\hat{\pi}_1} = \frac{-0.241}{-0.668} = 0.360$$

- Quali sono le unità? Queste stime hanno senso nel mondo reale?

Consistenza dello stimatore TSLS

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

Le covarianze campionarie sono consistenti:

$$s_{YZ} \xrightarrow{p} cov(Y, Z) \text{ e } s_{XZ} \xrightarrow{p} cov(X, Z)$$

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

- La condizione di rilevanza dello strumento, $cov(X, Z) \neq 0$, assicura che non stiamo dividendo per zero.

Esempio 2: offerta e domanda di burro

La regressione IV è stata sviluppata in origine per stimare l'elasticità della domanda per beni agricoli, per esempio il burro:

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

- $\beta_1 =$ elasticità del burro = variazione percentuale in quantità per una variazione dell'1% in prezzo (si ricordi la discussione sulla specifica log-log)
- Dati: osservazioni su prezzo e quantità di burro per diversi anni
- La regressione OLS di $\ln(P_i^{butter})$ su $\ln(Q_i^{butter})$ soffre di distorsione da causalità simultanea (perché?)

La distorsione da causalità simultanea nella regressione OLS di $\ln(Q_i^{butter})$ su $\ln(P_i^{butter})$ nasce perché prezzo e quantità sono determinati dall'interazione di domanda e offerta:

Questa interazione di domanda e offerta produce dati come…

Prezzo

E se si spostasse solo l’offerta?

Prezzo |

TSLS nell'esempio di domanda e offerta:

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

Sia $Z = pioggia$ nelle aree di produzione lattiera. Z è uno strumento valido?

1. Rilevante? $cor(rain_i, \ln(P_i^{butter})) \neq 0$?

   Plausibilmente: pioggia insufficiente significa meno pascolo quindi meno burro e quindi prezzi più alti

2. Esogeno? $cor(rain_i, u_i) = 0$?

   Plausibilmente: la pioggia nelle aree di produzione lattiera non dovrebbe influenzare la domanda di burro

TSLS nell'esempio di domanda e offerta (continua)

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

$Z_i = rain_i =$ pioggia nelle aree di produzione lattiera

Stadio 1: regressione di $\ln(P_i^{butter})$ su $rain_i \Rightarrow \ln(\widehat{P_i^{butter}})$

$\ln(\widehat{P_i^{butter}})$ isola le variazioni nel log del prezzo conseguenti all'offerta (o almeno a parte di essa)

Stadio 2: regressione di $\ln(Q_i^{butter})$ su $\ln(\widehat{P_i^{butter}})$

Controparte dell'uso degli spostamenti della curva di offerta per tracciare la curva di domanda.

Esempio 3: punteggi nei test e dimensioni delle classi

Le regressioni per punteggi nei test/dimensioni delle classi in California potrebbero avere distorsione da variabili omesse (per esempio per interessamento dei genitori).

- In linea di principio questa distorsione può essere eliminata dalla regressione IV (TSLS).
- La regressione IV richiede uno strumento valido, cioè che sia:

1. rilevante: $cor(Z_i, STR_i) \neq 0$
2. esogeno: $cor(Z_i, u_i) = 0$

Esempio 3: punteggi nei test e dimensioni delle classi (continua)

Ecco uno strumento ipotetico: - alcuni distretti, colpiti casualmente da un terremoto, “raddoppiano” le classi:

$$Z_i = Quake_i = 1 \text{ se colpito da terremoto}, \quad = 0 \text{ altrimenti}$$

- Valgono le due condizioni per la validità dello strumento?
- Il terremoto crea una situazione come se i distretti rientrassero in un esperimento con assegnazione casuale. Quindi, la variazione in STR conseguente al terremoto è esogena.
- Il primo stadio del TSLS prevede la regressione di STR su $Quake$, isolando così la parte esogena di STR (la parte “come se” fosse assegnata casualmente)

Inferenza con TSLS

- In grandi campioni, la distribuzione campionaria dello stimatore TSLS è normale
- L'inferenza (verifiche di ipotesi, intervalli di confidenza) procede nel modo consueto, ovvero $\pm 1.96SE$
- Il concetto alla base della distribuzione normale in grandi campioni dello stimatore TSLS è che - come tutti gli altri stimatori che abbiamo considerato – comporta variabili casuali i.i.d. con media nulla, a cui possiamo applicare il TLC.

- Di seguito riportiamo i calcoli abbozzati (si veda l'Appendice 12.3 per i dettagli)…

$$\beta_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})} = \frac{\sum_{i=1}^{n} Y_i(Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i(Z_i - \bar{Z})}$$

Sostituendo $Y_i = \beta_0 + \beta_1 X_i + u_i$ e semplificando:

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

quindi…

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

ovvero:

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

- denominatore: $\frac{1}{n}\sum_{i=1}^{n} X_i(Z_i - \bar{Z}) \xrightarrow{p} cov(X, Z) \neq 0$
- numeratore: $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} u_i(Z_i - \bar{Z})$ ha distribuzione $N\left(0, [var(Z - \mu_Z)u]\right)$ (TLC)
- quindi: β_1^{TSLS} ha distribuzione approssimata $N\left(\beta_1, \sigma^2_{\hat{\beta}_1^{TSLS}}\right)$

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

NB: $cov(X, Z) \neq 0$ perché lo strumento è RILEVANTE!

Inferenza con TSLS (continua)

β_1^{TSLS} ha distribuzione approssimata $N\left(\beta_1, \sigma^2_{\hat{\beta}_1^{TSLS}}\right)$

- L’inferenza statistica procede nel modo consueto.

Esempio 4: domanda di sigarette

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

Perché lo stimatore OLS di β_1 è probabilmente distorto?

- Data set: dati panel sul consumo annuo e i prezzi medi (comprese le imposte) delle

Domanda di sigarette (continua) - TSLS

```r
library(Ecdat)
library(estimatr)
library(dplyr)
data("Cigarette")

Cigarette <- Cigarette |> mutate(rprice = avgprs/cpi,
                    rtax = tax/cpi,
                    rtaxs = taxs/cpi,
                    rincome = income/pop/cpi,
                    rsaletax = (taxs - tax) / cpi)
C1995 <- Cigarette |> filter(year==1995)
## Regressione primo stadio
fs <- lm_robust(log(rprice)~rsaletax, data=C1995)
logrpricehat <- fitted(fs)
## Regressione secondo stadio
ss <- lm_robust(log(packpc)~logrpricehat, data=C1995)

iv_robust(log(packpc)~log(rprice)|rsaletax, data=C1995)
```

```
             Estimate Std. Error   t value     Pr(>|t|)  CI Lower   CI Upper DF
(Intercept)  9.719877  1.5634715  6.216856 1.368798e-07  6.572772 12.8669821 46
log(rprice) -1.083587  0.3261725 -3.322128 1.757196e-03 -1.740138 -0.4270355 46
```


- Questi coefficienti sono le stime TSLS

Domanda di sigarette (continua) - IV

```
## Stima a variabili strumentali
iv_robust(log(packpc)~log(rprice)|rsaletax, data=C1995)
```

```
             Estimate Std. Error   t value     Pr(>|t|)  CI Lower   CI Upper DF
(Intercept)  9.719877  1.5634715  6.216856 1.368798e-07  6.572772 12.8669821 46
log(rprice) -1.083587  0.3261725 -3.322128 1.757196e-03 -1.740138 -0.4270355 46
```

Il modello generale di regressione IV (Paragrafo 12.2)

- Finora abbiamo considerato la regressione IV con un singolo regressore endogeno (X)

Identificazione

- In generale si dice che un parametro è identificato se diversi valori del parametro producono distribuzioni diverse dei dati.
- Nella regressione IV, il fatto che i coefficienti siano identificati dipende dalla relazione tra il numero di strumenti (m) e il numero di regressori endogeni (k)
- Intuitivamente, se ci sono meno strumenti che regressori endogeni, non possiamo stimare $\beta_1, \ldots, \beta_k$
- Per esempio, supponiamo $k = 14$ ma $m = 0$ (nessuno strumento)!

Identificazione (continua)

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

- esattamente identificati se $m = k$.

  Ci sono esattamente gli strumenti sufficienti per stimare $\beta_1, \ldots, \beta_k$

- sovraidentificati se $m > k$.

  Ci sono più strumenti di quelli necessari per stimare $\beta_1, \ldots, \beta_k$. In questo caso si può verificare se gli strumenti sono validi (test delle "restrizioni sovraidentificanti") - torneremo sul tema in seguito

- sottoidentificati se $m < k$.

  Ci sono troppo pochi strumenti per stimare $\beta_1, \ldots, \beta_k$. In questo caso occorre procurarsi più strumenti!

Il modello generale di regressione IV: riepilogo della terminologia

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

- Yi è la variabile dipendente

TSLS con un singolo regressore endogeno

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

- m strumenti: $Z_{1i}, \ldots, Z_m$

W come variabili di controllo

- In molti casi le W sono incluse allo scopo di controllare per fattori omessi, cosicché, una volta incluse le W, Z è incorrelata con u.
- Tecnicamente, la condizione perché le W siano variabili di controllo effettive è che la media condizionata degli u_i non dipenda da Z_i, date W_i:

$$E(u_i|W_i, Z_i) = E(u_i|W_i)$$

Questa è la versione IV dell'assunzione dell'indipendenza in media condizionata del Capitolo 7.

Ecco il punto chiave: in molte applicazioni occorre includere variabili di controllo (W) affinché Z sia verosimilmente esogena (incorrelata con u). (Per i dettagli si veda l'Appendice 12.6)

Esempio 4: ancora la domanda di sigarette

Si supponga che il reddito sia esogeno (è plausibile?), e di voler anche stimare l'elasticità:

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

- Una variabile endogena: $X_{1i} = \ln(P_i^{Cigarettes})$
- Una variabile esogena inclusa: $W_{1i} = \ln(Income_i)$
- Due strumenti:
  1. Z_{1i} = imposta generale sulle vendite
  2. Z_{2i} = imposta specifica sulle sigarette

Esempio: domanda di sigarette, un solo strumento

```r
#Stima via variabili strumentali
cig_ivreg <- iv_robust(log(packpc) ~ log(rprice) + log(rincome) | log(rincome) + rsaletax, data = C1995)
cig_ivreg
```

```
              Estimate Std. Error    t value     Pr(>|t|)   CI Lower
(Intercept)   9.4306583  1.2844876  7.3419610 3.178604e-09  6.8435674
log(rprice)  -1.1433751  0.3811180 -3.0000555 4.388885e-03 -1.9109862
log(rincome)  0.2145153  0.3163935  0.6780016 5.012422e-01 -0.4227339
               CI Upper DF
(Intercept)  12.0177492 45
log(rprice)  -0.3757641 45
log(rincome)  0.8517644 45
```

Esempio: domanda di sigarette, due strumenti

```r
cig_ivreg2 <- iv_robust(log(packpc) ~ log(rprice) + log(rincome) |
                        log(rincome) + rsaletax + rtax,
                        data = C1995)
summary(cig_ivreg2)
```

```
Call:
iv_robust(formula = log(packpc) ~ log(rprice) + log(rincome) |
    log(rincome) + rsaletax + rtax, data = C1995)

Standard error type:  HC2

Coefficients:
             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
(Intercept)    9.8950     0.9777  10.120 3.569e-13   7.9257  11.8642 45
log(rprice)   -1.2774     0.2547  -5.015 8.739e-06  -1.7904  -0.7644 45
log(rincome)   0.2804     0.2547   1.101 2.768e-01  -0.2326   0.7934 45

Multiple R-squared:  0.4294 ,   Adjusted R-squared:  0.4041
F-statistic:  15.5 on 2 and 45 DF,  p-value: 7.55e-06
```

Risultati stime IV uno vs due strumenti

- Errori standard minori per $m = 2$. Con 2 strumenti si hanno più informazioni, più "variazione come se casuale"
- Bassa elasticità al reddito (non è un bene di lusso); elasticità al reddito non significativamente diversa da zero a livello statistico
- Elasticità al prezzo sorprendentemente elevata

Assunzioni generali per la validità di uno strumento

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

1. Esogeneità: $cor(Z_{1i}, u_i) = 0, \ldots, cor(Z_{mi}, u_i) = 0$
2. Rilevanza: caso generale, più X
   a. almeno uno strumento deve entrare nella controparte della regressione del primo stadio; e
   b. i W non sono perfettamente collineari.

Assunzioni della regressione IV

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

1. $E(u_i | W_{1i}, \ldots, W_{ri}) = 0$
   - l'assunzione 1 dice "i regressori esogeni sono esogeni"
2. $(Y_i, X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}, Z_{1i}, \ldots, Z_{mi})$ sono i.i.d.
   - l'assunzione 2 non è nuova
3. X, W, Z, Y hanno momenti quarti finiti non nulli
   - l'assunzione 2 non è nuova
4. Gli strumenti $(Z_{1i}, \ldots, Z_{mi})$ sono validi.
   - Ne abbiamo parlato

Sotto le assunzioni 1-4, il TSLS e la sua statistica t hanno distribuzione normale

Esempio 1: effetto dello studio sui voti (continua)

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

- Y = media voti del primo semestre
- X = media di ore di studio al giorno
- Z = 1 se il compagno di stanza ha portato un videogioco, = 0 altrimenti

I compagni di stanza sono stati assegnati a caso

Conoscete un motivo per cui Z potrebbe essere correlata con u - anche se è assegnata casualmente? Che cos'altro entra nel termine d'errore, quali sono altri determinanti dei voti, oltre al tempo speso studiando?

Esempio 1: effetto dello studio sui voti (continua)

Perché Z potrebbe essere correlata u?

- Ecco una ipotetica possibilità: il genere. Supponiamo:

Verifica della validità degli strumenti (Paragrafo 12.3)

Ricordiamo i due requisiti per strumenti validi:

1. Rilevanza (caso speciale di una sola X)

Almeno uno strumento deve spiegare la variabile endogena nella regressione del primo stadio.

2. Esogeneità

Tutti gli strumenti devono essere incorrelati con il termine d'errore:

$$cor(Z_{1i}, u_i) = 0, \ldots, cor(Z_{mi}, u_i) = 0$$

Che cosa accade se uno di questi requisiti non è soddisfatto? Come si può verificare? Che cosa occorre fare?

Verifica dell'assunzione 1: rilevanza dello strumento

Ci concentreremo su un singolo regressore incluso:

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

Regressione del primo stadio:

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

- Gli strumenti sono rilevanti se almeno uno dei $\pi_1, \ldots, \pi_m$ è diverso da zero.

Quali sono le conseguenze di strumenti deboli?

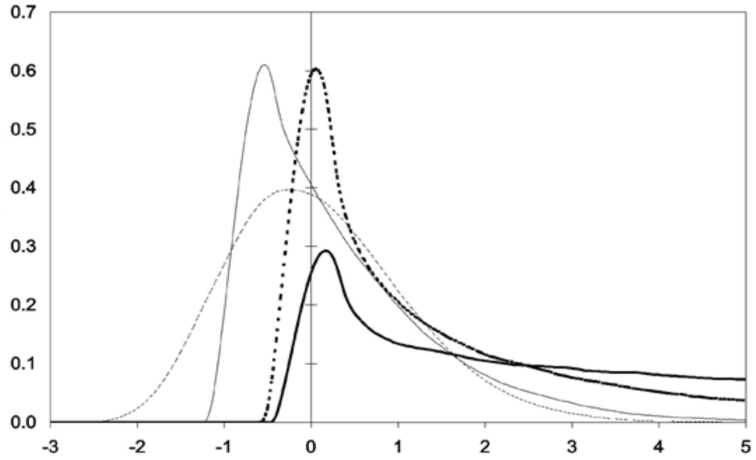
Se gli strumenti sono deboli, la distribuzione campionaria del TSLS e della sua statistica t non è normale, anche con n grande.

Consideriamo il caso più semplice:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$
$$X_i = \pi_0 + \pi_1 Z_i + \nu_i$$

- Lo stimatore IV è $\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$
- Se $cov(X, Z)$ è zero o vicino, allora s_{XZ} sarà piccolo: con strumenti deboli, il denominatore è quasi zero.
- In questo caso, la distribuzione campionaria di (e la sua statistica t) non è ben approssimata da una normale per n grande…

Esempio: la distribuzione campionaria della statistica t del TSLS con strumenti deboli



Linea scura = strumenti non rilevanti

Linea chiara tratteggiata = strumenti forti

Misurazione della forza degli strumenti in pratica: la statistica $Wald$ del primo stadio

- La regressione del primo stadio (una sola X):
- Regressione di X su $Z_1, \ldots, Z_m, W_1, \ldots, W_k$.
- Strumenti totalmente irrilevanti $\Rightarrow$ tutti i coefficienti di $Z_1, \ldots, Z_m$ sono zero.
- La statistica Wald del primo stadio verifica l'ipotesi che $Z_1, \ldots, Z_m$ non entrino nella regressione del primo stadio.
- Strumenti deboli implicano un valore basso della statistica Wald del primo stadio.

Verifica di strumenti deboli con una singola X

- Si calcola la statistica Wald del primo stadio.

  Regola empirica: se la statistica Wald del primo stadio è minore di $m \times 10$, allora l'insieme di strumenti è debole.

- In questo caso, lo stimatore TSLS sarà distorto, e le inferenze statistiche (errori standard, verifiche di ipotesi, intervalli di confidenza) possono essere fuorvianti.

Verifica di strumenti deboli

- Perché confrontare la statistica $Wald$ del primo stadio con 10m?
- Non è sufficiente respingere l'ipotesi nulla che i coefficienti delle Z siano zero - serve un contenuto predittivo sostanziale per una buona approssimazione normale.
- Se la $Wald$ è minore di $10 \times m$, la distorsione relativa è superiore al 10%, cioè il TSLS può avere una distorsione sostanziale (si veda l'Appendice 12.5).
- Nota: $Wald = F \times m$

```r
library(car)
## First stage
fs <- lm_robust(log(rprice) ~ log(rincome) + rsaletax + rtax, data = C1995)
linearHypothesis(fs, c("rsaletax=0", "rtax=0"))
```

```
Linear hypothesis test

Hypothesis:
rsaletax = 0
rtax = 0

Model 1: restricted model
Model 2: log(rprice) ~ log(rincome) + rsaletax + rtax

  Res.Df Df  Chisq Pr(>Chisq)
1     46
2     44  2 392.81  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
linearHypothesis(fs, c("rsaletax=0", "rtax=0"), test="F")
```

```
Linear hypothesis test

Hypothesis:
rsaletax = 0
rtax = 0

Model 1: restricted model
Model 2: log(rprice) ~ log(rincome) + rsaletax + rtax
```



```
  Res.Df Df     F    Pr(>F)
1     46
2     44  2 196.4 < 2.2e-16 ***
```

Verifica dell'assunzione 2: esogeneità dello strumento

- Esogeneità dello strumento: Tutti gli strumenti non sono correlati con il termine d'errore: $cor(Z_{1i}, u_i) = 0, \ldots, cor(Z_{mi}, u_i) = 0$
- Se gli strumenti sono correlati con il termine d'errore, il primo stadio del TSLS non può isolare una componente di X incorrelata con il termine d'errore, perciò $\hat{X}$ è correlata con u e il TSLS è inconsistente.
- Se ci sono più strumenti che regressori endogeni, è possibile verificare - parzialmente - l'esogeneità dello strumento.

Verifica di restrizioni di sovraidentificazione

Consideriamo il caso più semplice:

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

- Supponiamo che vi siano due strumenti validi: Z_{1i}, Z_{2i}
- Allora potremmo calcolare due stime TSLS separate.
- Intuitivamente, se queste due stime TSLS sono molto diverse tra loro, ci dev'essere qualcosa di sbagliato: uno strumento o l'altro (o entrambi) devono essere non validi.
- Il test J di restrizioni sovraidentificanti esegue questo confronto in un modo statisticamente preciso.
- Si può fare soltanto se il numero di Z è maggiore del numero di X (sovraidentificazione).

Il test J di restrizioni di sovraidentificazione

Supponiamo che il numero di strumenti = $m >$ numero di $X = k$ (sovraidentificazione)

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

Il test J è il test di Anderson-Rubin, usando lo stimatore TSLS al posto del valore ipotizzato $\beta_{1,0}$. Procedura:

1. Prima si stima l'equazione di interesse usando TSLS e tutti gli m strumenti; si calcolano

Il test J (continua)

$J = Wald$, dove $Wald$ = la statistica Wald che verifica i coefficienti di $Z_{1i}, \ldots, Z_{mi}$ in una regressione dei residui TSLS rispetto a $Z_{1i}, \ldots, Z_{mi}, W_{1i}, \ldots, W_{ri}$

Distribuzione della statistica J

- Sotto l'ipotesi nulla che tutti gli strumenti siano esogeni, J ha una distribuzione χ^2 con $m-k$ gradi di libertà
- Se $m = k$, $J = 0$ (ha senso?)
- Se alcuni strumenti sono esogeni e altri sono endogeni, la statistica J sarà grande, e l'ipotesi nulla che tutti gli strumenti sono esogenisarà rifiutata.

J test

```r
cig_ivreg2 <- iv_robust(log(packpc) ~ log(rprice) + log(rincome) |
                          log(rincome) + rsaletax + rtax, data = C1995)
C1995 <- C1995 |> mutate(uhat = log(packpc) - cig_ivreg2$fitted.values)
Jlm <- lm_robust(uhat~log(rincome) + rsaletax + rtax, data = C1995)
J <- linearHypothesis(Jlm, c("rsaletax=0", "rtax=0"))
J
```

```
Linear hypothesis test

Hypothesis:
rsaletax = 0
rtax = 0

Model 1: restricted model
Model 2: uhat ~ log(rincome) + rsaletax + rtax

  Res.Df Df  Chisq Pr(>Chisq)
1     46
2     44  2 0.3025     0.8596
```

I p-value è errato perché è calcolato usando $m = 2$ gradi di libertá anzichè $m - k = 1$.

```r
## Calcolo pvalue con df corretto
pchisq(J$Chisq[2], df = 1, lower.tail = FALSE)
```

```
[1] 0.5823442
```

Verifica della validità degli strumenti: riepilogo

Questo riepilogo considera il caso di una singola X. I due requisiti per la validità degli strumenti sono:

1. Rilevanza

- Almeno uno strumento deve entrare nella controparte della regressione del primo stadio.
- Se gli strumenti sono deboli, allora lo stimatore TSLS è distorto e la statistica t ha una distribuzione non normale
- Per verificare strumenti deboli con un singolo regressore endogeno incluso, si verifica la statistica F del primo stadio:
  - Se $Wald > 10 \times m$, gli strumenti sono forti - si usa il TSLS
  - Se $Wald < 10 \times m$, gli strumenti sono deboli.

Verifica della validità degli strumenti: riepilogo

2. Esogeneità

- Tutti gli strumenti devono essere incorrelati con il termine d'errore: $cor(Z_{1i}, u_i) = 0, \ldots, cor(Z_{mi}, u_i) = 0$
- Possiamo eseguire una verifica parziale di esogeneità: se $m > 1$, possiamo verificare l'ipotesi nulla che tutti gli strumenti siano esogeni contro l'alternativa che al massimo $m - 1$ siano endogeni (correlati con u)
- Si usa il test J, realizzato usando i residui TSLS.
- Se il J respinge l'ipotesi, allora almeno alcuni degli strumenti sono endogeni, perciò occorre prendere una decisione difficile e scartare alcuni (o tutti) gli strumenti.

Exercise

Calcolate e interpretate la statistica del primo statio e il test J della seguente stima basata sulle differenze.

```
DCig <- Cigarette |> filter(year==1995 | year==1985) |>
  arrange(state, year) |>
  group_by(state) |>
  mutate(Drsaletax = rsaletax - lag(rsaletax),
         Drtax = rtax - lag(rtax),
         Dlogrincome = log(rincome) - lag(log(rincome)),
         Dlogrprice = log(rprice) - lag(log(rprice)),
         Dlogpackpc = log(packpc) - lag(log(packpc))
         )
summary(iv_robust(Dlogpackpc ~ Dlogrprice + Dlogrincome|
                        Dlogrincome + Drsaletax + Drtax,
                      diagnostic=FALSE,
                      data = DCig))
```

```
Call:
iv_robust(formula = Dlogpackpc ~ Dlogrprice + Dlogrincome | Dlogrincome +
    Drsaletax + Drtax, data = DCig, diagnostics = FALSE)

Standard error type:  HC2

Coefficients:
```



```
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
(Intercept)   -0.052    0.06511 -0.7987 4.287e-01  -0.1831   0.07914 45
Dlogrprice    -1.202    0.20850 -5.7669 6.919e-07  -1.6223 -0.78246 45
```