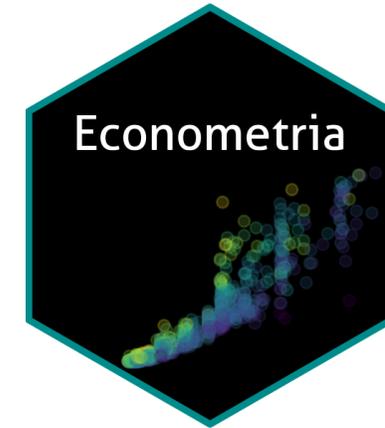


Econometria | 2023/2024



Lezione 14: Variabili strumentali

Giuseppe Ragusa

Dipartimento di Economia e Diritto
Sapienza Università di Roma

<https://gragusa.org>

Roma, maggio 2024

Sommario

1. Regressione IV: cosa e perché; minimi quadrati in due stadi
2. Il modello generale di regressione IV
3. Verifica della validità degli strumenti
 - a. Strumenti deboli e forti
 - b. Esogeneità degli strumenti
4. Applicazione: domanda di sigarette
5. Esempi: dove troviamo gli strumenti?

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

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

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:

$$\begin{aligned} \text{cov}(Y_i, Z_i) &= \text{cov}(\beta_0 + \beta_1 X_i + u_i, Z_i) = \\ &= \underbrace{\text{cov}(\beta_0, Z_i)}_0 + \text{cov}(\beta_1 X_i, Z_i) + \underbrace{\text{cov}(u_i, Z_i)}_0 = \\ &= \text{cov}(\beta_1 X_i, Z_i) = \\ &= \beta_1 \text{cov}(X_i, Z_i) \end{aligned}$$

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

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

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

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

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”

$$\begin{aligned}Y_i &= \beta_0 + \beta_1 X_i + u_i \\X_i &= \pi_0 + \pi_1 Z_i + \nu_i\end{aligned}$$

Sostituiamo X nell'equazione di Y :

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

- Visto che $\beta_1 \pi_1 = \gamma_1$, $\beta = \gamma_1 / \pi_1$
- Possiamo stimare π_1 ($\hat{\pi}_1$) e γ_1 ($\hat{\gamma}_1$) con OLS, uno stimatore di β_1 è $\hat{\beta}_1 = \hat{\gamma}_1 / \hat{\pi}_1$

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 = media voti (scala 4 punti)

X = tempo di studio (ore al giorno)

Z = 1 se il compagno ha portato un videogioco, = 0 altrimenti

- $n = 210$ studenti del primo anno al Berea College (Kentucky) nel 2001
- Y = media voti del primo semestre
- X = media ore di studi al giorno (sondaggio)
- I compagni di stanza sono stati assegnati a caso
- $Z = 1$ se il compagno di stanza ha portato un videogioco, = 0 altrimenti

Effetto dello studio sui voti

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

$$Y_i = \gamma_0 + \gamma_1 Z_i + \omega_i$$

Y = media voti (scala 4 punti)

X = tempo di studio (ore al giorno)

Z = 1 se il compagno ha portato un videogioco, = 0 altrimenti

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 = media voti (scala 4 punti)

X = tempo di studio (ore al giorno)

Z = 1 se il compagno ha portato un videogioco, = 0 altrimenti

Risultati di Stinebrinckner e Stinebrinckner:

$$\hat{\pi}_1 = -0.668$$

$$\hat{\gamma}_1 = -0.241$$

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

- Quali sono le unità? Queste stime appaiono verosimili?

Consistenza dello stimatore TSLS

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

Le covarianze campionarie sono consistenti:

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

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

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

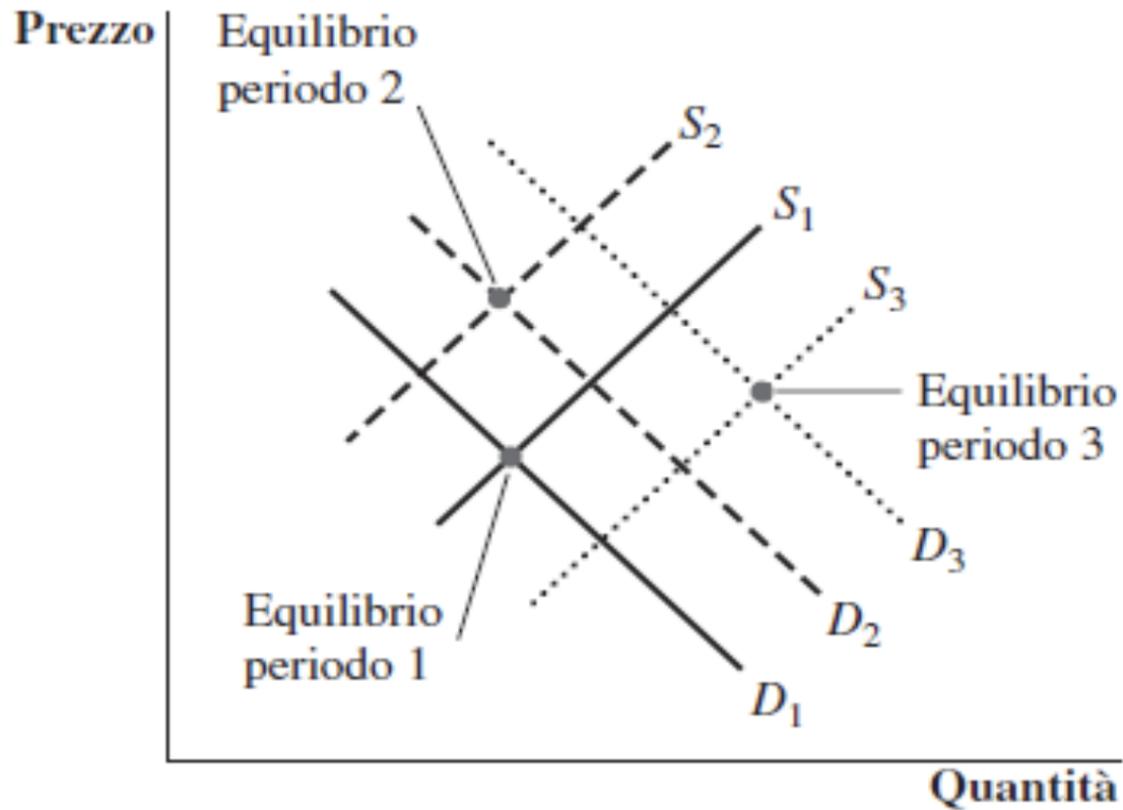
Esempio 2: offerta e domanda di burro

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

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

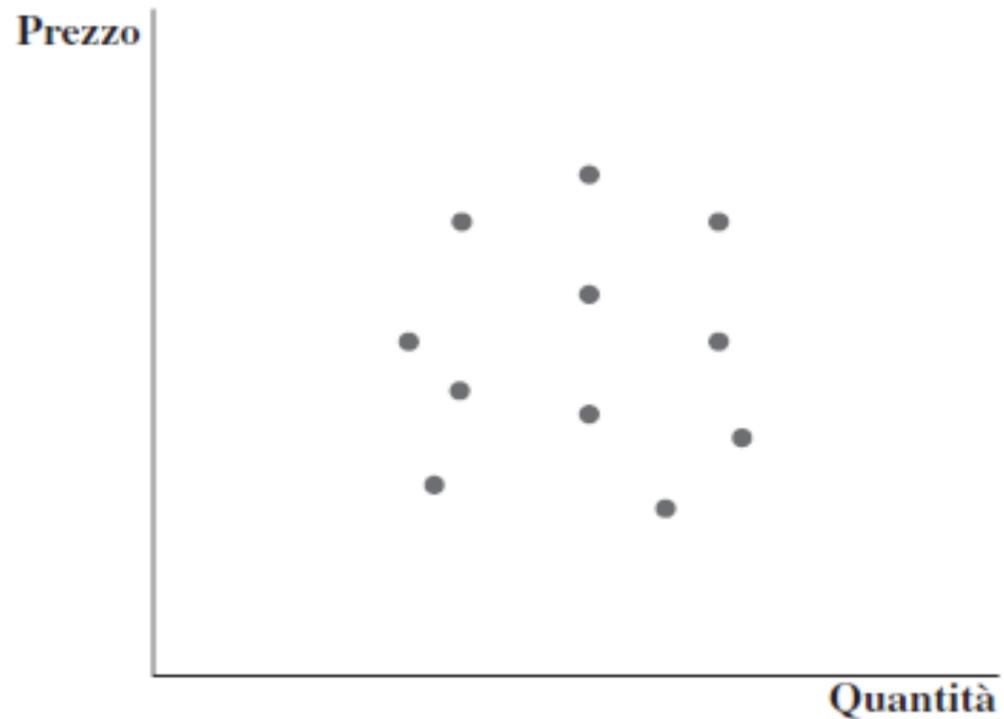
- β_1 = elasticità del burro = variazione percentuale in quantità per una variazione dell'1% in prezzo (si ricordi la discussione sulla specifica log-log)
- Dati: osservazioni su prezzo e quantità di burro per diversi anni
- La regressione OLS di $\ln(P_i^{butter})$ su $\ln(Q_i^{butter})$ soffre di distorsione 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:



(a) Domanda e offerta in tre periodi

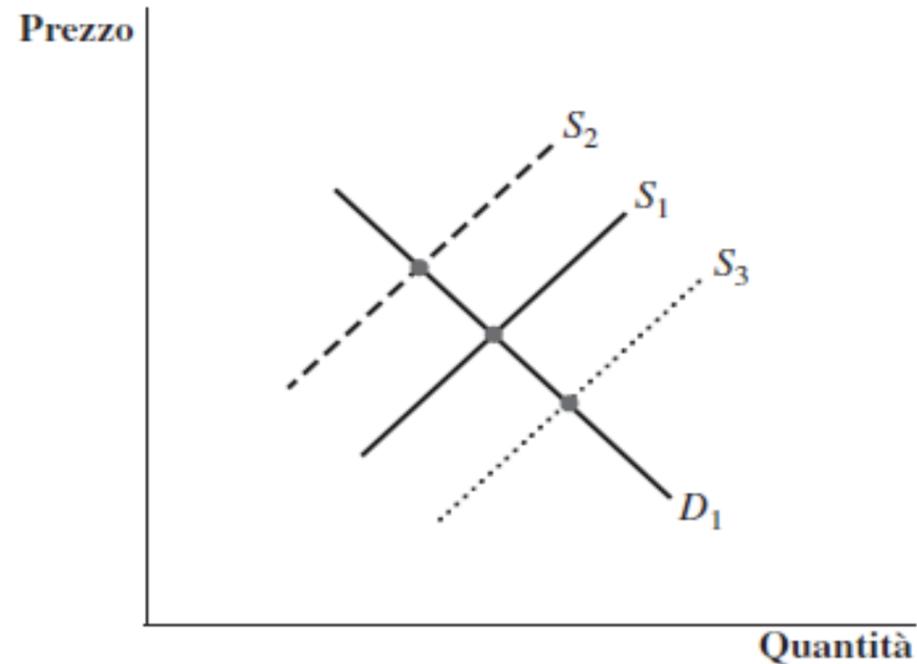
Questa interazione di domanda e offerta produce dati come...



(b) Prezzo e quantità di equilibrio per 11 periodi

Una regressione con questi dati produrrebbe la curva di domanda?

E se si spostasse solo l'offerta?



(c) Prezzo e quantità di equilibrio
quando solo la curva di offerta si sposta

- TSLS stima la curva di domanda isolando gli spostamenti di prezzo e quantità conseguenti a spostamenti dell'offerta
- Z è una variabile che sposta l'offerta ma non la domanda

TOLS 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

TOLS 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:

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

- 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.

Dimostrazione (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, $\beta_1^{TSLS} - \beta_1 = \frac{\sum_{i=1}^n u_i (Z_i - \bar{Z})}{\sum_{i=1}^n X_i (Z_i - \bar{Z})}$. Moltiplicando entrambi i membri per \sqrt{n} :

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

dove:

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

- denominatore:

$$\frac{1}{n} \sum_{i=1}^n X_i (Z_i - \bar{Z}) \xrightarrow{p} \text{cov}(X, Z) \neq 0$$

- numeratore:

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

$$\implies \beta_1^{TSLS} \xrightarrow{d} N\left(\beta_1, \sigma_{\hat{\beta}_1^{TSLS}}^2\right), \text{ dove } \sigma_{\hat{\beta}_1^{TSLS}}^2 = \frac{1}{n} \frac{[\text{var}(Z_i - \mu_Z)u_i]}{[\text{cov}(X_i, Z_i)]^2}$$

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

Inferenza con TSLS (continua)

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

- L'inferenza statistica procede nel modo consueto.
- La giustificazione è (come di consueto) basata su grandi campioni
- Tutto questo assume che gli strumenti siano validi - vedremo tra breve che cosa accade se non lo sono.
- **Nota importante sugli errori standard:**
 - Gli errori standard OLS dalla regressione del secondo stadio non sono corretti - non tengono conto della stima al primo stadio (\hat{X}_i è stimata).
 - Si utilizza invece un singolo comando apposito che calcola lo stimatore TSLS e gli errori standard corretti.
 - Come di consueto, si usano errori standard robusti all'eteroschedasticità

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 sigarette, per stato, per i 48 stati contigui USA, 1985-1995.
- Variabile strumentale proposta:

$$Z_i = \text{imposta generale sulle vendite nello stato} = SalesTax_i$$

- Pensate che questo strumento sia valido?
 1. Rilevante? $cor(SalesTax_i, \ln(P_i^{cigarettes})) \neq 0$?
 2. Esogeno? $cor(SalesTax_i, u_i) = 0$?

Domanda di sigarette (continua) - TSLS

```
1 library(Ecdat)
2 library(fixest)
3 library(dplyr)
4 data("Cigarette")
5
6 Cigarette <- Cigarette |> mutate(rprice = avgprs/cpi,
7                                rtax = tax/cpi,
8                                rtaxs = taxs/cpi,
9                                rincome = income/pop/cpi,
10                               rsaletax = (taxs - tax) / cpi)
11 C1995 <- Cigarette |> filter(year==1995)
12 ## Regressione primo stadio
13 fs <- lm(log(rprice)~rsaletax, data=C1995)
14 logrpricehat <- fitted(fs)
15 ## Regressione secondo stadio
16 ss <- lm(log(packpc)~logrpricehat, data=C1995)
17
18 ss
```

Call:

```
lm(formula = log(packpc) ~ logrpricehat, data = C1995)
```

Coefficients:

```
(Intercept)  logrpricehat
          9.720         -1.084
```

- Questi coefficienti sono le stime TSLS
- Gli errori standard sono sbagliati (anche se avessimo usato `feols`) perché ignorano la stima al primo stadio

Domanda di sigarette (continua) - IV

```
1 feols(log(packpc)~1|log(rprice)~rsaletax, data=C1995, vcov = "hetero")
```

```
TOLS estimation, Dep. Var.: log(packpc), Endo.: log(rprice), Instr.: rsaletax
Second stage: Dep. Var.: log(packpc)
Observations: 48
Standard-errors: Heteroskedasticity-robust
      Estimate Std. Error  t value  Pr(>|t|)
(Intercept)    9.71988    1.528322  6.35984  8.3462e-08 ***
fit_log(rprice) -1.08359    0.318918 -3.39769  1.4114e-03 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.186346  Adj. R2: 0.38811
F-test (1st stage), log(rprice): stat = 41.0      , p = 7.271e-8, on 1 and 46 DoF.
Wu-Hausman: stat = 0.313803, p = 0.578134, on 1 and 45 DoF.
```

- Gli errori standard sono corretti — tengono conto della stima al primo stadio

Il modello generale di regressione IV (Paragrafo 12.2)

- Finora abbiamo considerato la regressione IV con un singolo regressore endogeno (X) e un singolo strumento (Z).
- Dobbiamo estenderla a:
 - più regressori endogeni (X_1, \dots, X_k)
 - più variabili incluse esogene (W_1, \dots, W_r) o variabili di controllo, che devono essere incluse per il consueto motivo delle variabili omesse
 - più variabili strumentali (Z_1, \dots, Z_m). Più strumenti (rilevanti) possono produrre una minore varianza del TSLS: l' R^2 del primo stadio aumenta, perciò si ha maggiore variazione in \hat{X} .
- Nuovi termini: identificazione e sovraidentificazione

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 β_1, \dots, β_k
- Per esempio, supponiamo $k = 14$ ma $m = 0$ (nessuno strumento)!

Identificazione (continua)

I coefficienti β_1, \dots, β_k si dicono:

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

Ci sono esattamente gli strumenti sufficienti per stimare β_1, \dots, β_k

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

Ci sono più strumenti di quelli necessari per stimare β_1, \dots, β_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 β_1, \dots, β_k . In questo caso occorre procurarsi più strumenti!

Il modello generale di regressione IV: riepilogo della terminologia

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

- Y_i è la **variabile dipendente**
- X_{1i}, \dots, X_{ki} sono i **regressori endogeni** (**potenzialmente correlati** con u_i)
- W_{1i}, \dots, W_{ri} sono i **regressori esogeni inclusi** (**incorrelati** con u_i) o variabili di **controllo** (inclusi in modo che Z_i sia incorrelata con u_i , una volta inclusi i W)
- $\beta_0, \beta_1, \dots, \beta_{k+r}$ sono i coefficienti di regressione ignoti
- Z_{1i}, \dots, Z_{mi} sono le m **variabili strumentali** (variabili **esogene escluse**)
- I coefficienti sono **sovraidentificati** se $m > k$; **esattamente identificati** se $m = k$; **sottoidentificati** se $m < k$.

TOLS con un singolo regressore endogeno

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

- m strumenti: Z_{1i}, \dots, Z_{mi}
- **Primo stadio:**
 - Regressione di X_1 su **tutti** i regressori **esogeni**: regressione di X_1 su $W_1, \dots, W_r, Z_1, \dots, Z_m$, e un'intercetta, usando OLS
 - Calcolo dei valori predetti $\hat{X}_{1i}, i = 1, \dots, n$
- **Secondo stadio:**
 - Regressione di Y_i su $\hat{X}_{1i}, W_1, \dots, W_r$, e un'intercetta, usando OLS
 - I coefficienti di questa regressione del secondo stadio sono gli stimatori TOLS, ma gli errori standard sono sbagliati :::{.fborder} Per ottenere errori standard corretti, occorre procedere in un singolo passaggio con il software di regressione :::

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.

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

Domanda di sigarette: uno strumento

```
1 #Stima via variabili strumentali
2 cig_ivreg <- feols(log(packpc)~log(rincome)|log(rprice)~log(rincome)+rsaletax, data=C1995, vcov = "hetero")
3 summary(cig_ivreg, stage = 2)
```

TSLS estimation, Dep. Var.: log(packpc), Endo.: log(rprice), Instr.: log(rincome), rsaletax

Second stage: Dep. Var.: log(packpc)

Observations: 48

Standard-errors: Heteroskedasticity-robust

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	9.430658	1.259393	7.488260	1.9347e-09	***
fit_log(rprice)	-1.143375	0.372303	-3.071090	3.6113e-03	**
log(rincome)	0.214515	0.311747	0.688107	4.9492e-01	

... 1 variable was removed because of collinearity (log(rincome))

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

RMSE: 0.183555 Adj. R2: 0.393109

F-test (1st stage), log(rprice): stat = 22.6, p = 1.621e-7, on 2 and 45 DoF.

Wu-Hausman: stat = 1.102, p = 0.299559, on 1 and 44 DoF.

Sargan: stat = -1.066e-14, p = 1, on 1 DoF.

Domanda di sigarette: due strumenti

```
1 cig_ivreg2 <- feols(log(packpc)~log(rincome)|log(rprice)~log(rincome)+rsaletax+rtax,data=C1995, vcov = "hete")
2 summary(cig_ivreg2,stage=2)
```

TSLs estimation, Dep. Var.: log(packpc), Endo.: log(rprice), Instr.: log(rincome), rsaletax, rtax

Second stage: Dep. Var.: log(packpc)

Observations: 48

Standard-errors: Heteroskedasticity-robust

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	9.894956	0.959217	10.31566	1.9467e-13	***
fit_log(rprice)	-1.277424	0.249610	-5.11768	6.2107e-06	***
log(rincome)	0.280405	0.253890	1.10444	2.7527e-01	

... 1 variable was removed because of collinearity (log(rincome))

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

RMSE: 0.181891 Adj. R2: 0.404063

F-test (1st stage), log(rprice): stat = 163.2, p < 2.2e-16, on 3 and 44 DoF.

Wu-Hausman: stat = 3.06782, p = 0.086825, on 1 and 44 DoF.

Sargan: stat = 0.332622, p = 0.846783, on 2 DoF.

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

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

1. **Esogeneità:** $cor(Z_{1i}, u_i) = 0, \dots, cor(Z_{mi}, u_i) = 0$
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

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

1. $E(u_i | W_{1i}, \dots, W_{ri}) = 0$
 - l'assunzione 1 dice "i regressori esogeni sono esogeni"
2. $(Y_i, X_{1i}, \dots, X_{ki}, W_{1i}, \dots, W_{ri}, Z_{1i}, \dots, Z_{mi})$ 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}, \dots, Z_{mi}) sono validi.
 - Ne abbiamo parlato

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

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?

Effetto dello studio sui voti (continua)

Perché Z potrebbe essere correlata u ?

- Ecco una ipotetica possibilità: il genere. Supponiamo:
 - le donne ottengono voti migliori degli uomini, a parità di ore di studio
 - gli uomini hanno più probabilità di portare un videogioco, rispetto alle donne
 - Allora $cor(Z_i, u_i) < 0$ (i maschi hanno più probabilità di avere un compagno di stanza [maschio] che porti un videogioco, ma i maschi tendono anche ad avere voti inferiori, a parità di tempo di studio).
- È solo un altro caso di distorsione da variabili omesse. La soluzione sta nel controllare per (o includere) la variabile omessa, in questo caso il genere.
- Questa logica porta a includere W = genere come variabile di controllo nella regressione IV:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

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:

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

1. Che cosa accade se uno di questi requisiti non è soddisfatto? Come si può verificare?
Che cosa occorre fare?
2. Se si hanno più strumenti, quale si deve usare?

Verifica dell'assunzione 1: rilevanza dello strumento

Ci concentreremo su un singolo regressore incluso:

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

Regressione del primo stadio:

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

- Gli strumenti sono rilevanti se almeno uno dei π_1, \dots, π_m è diverso da zero.
- Gli strumenti si dicono **deboli** se i coefficienti π_1, \dots, π_m non sono abbastanza grandi ...
- Gli **strumenti deboli** spiegano una parte della variazione di X troppo piccola per poter spiegare accuratamente β_1

Quali sono le conseguenze di strumenti deboli?

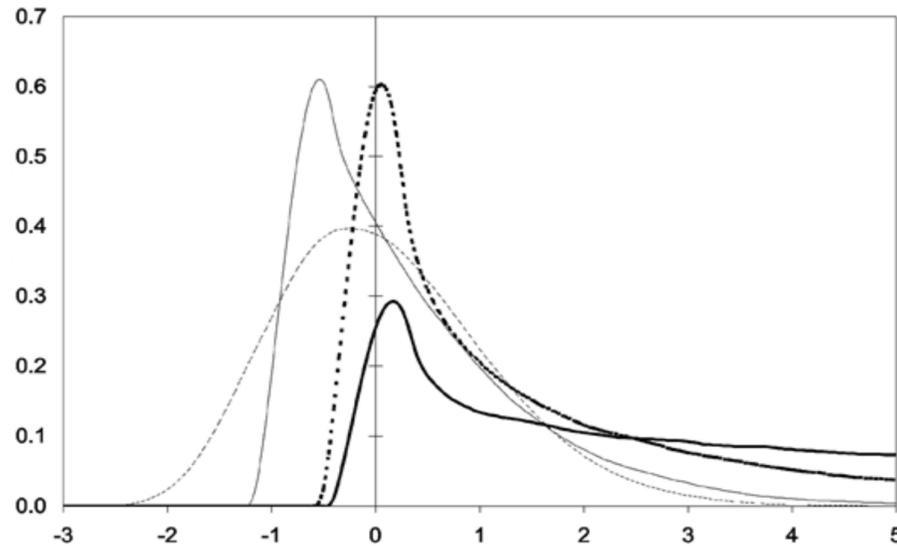
Se gli strumenti sono deboli, la distribuzione campionaria del TSLS e della sua statistica t non è normale, anche con n grande.

Consideriamo il caso più semplice:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i \\ X_i &= \pi_0 + \pi_1 Z_i + \nu_i \end{aligned}$$

- 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...

La distribuzione campionaria della statistica t del TSLS con strumenti deboli



- Linea scura = strumenti non rilevanti
- Linea chiara tratteggiata = strumenti forti

Se gli strumenti sono deboli, i consueti metodi di inferenza sono inaffidabili - potenzialmente molto inaffidabili!

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 $Z_1, \dots, Z_m, W_1, \dots, W_k$.
- Strumenti totalmente irrilevanti \Rightarrow tutti i coefficienti di Z_1, \dots, Z_m sono zero.
- La **statistica Wald del primo stadio** verifica l'ipotesi che Z_1, \dots, 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.
- Non è sufficiente respingere l'ipotesi nulla che i coefficienti delle Z siano zero, ma 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:** $Wald = F \times m$

```
1 library(car)
2 ## First stage
3 fs <- feols(log(rprice) ~ log(rincome) + rsaletax + rtax, data = C1995, vcov = "hetero")
4 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

	Df	Chisq	Pr(>Chisq)
1			
2	2	419.35	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
1 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

	Df	Chisq	Pr(>Chisq)
1			
2	2	419.35	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Verifica dell'assunzione 2: esogeneità dello strumento

- Esogeneità dello strumento: **Tutti** gli strumenti non sono correlati con il termine d'errore:

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

- 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)

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

Procedura:

1. Prima si stima l'equazione di interesse usando TSLS e tutti gli m strumenti; si calcolano i valori predetti \hat{Y}_i , usando le X effettive (non le \hat{X} usate per stimare il secondo stadio)
2. Si calcolano i residui $\hat{u}_i = Y_i - \hat{Y}_i$
3. Si esegue la regressione rispetto a $Z_{1i}, \dots, Z_{mi}, W_{1i}, \dots, W_{ri}$
4. Si calcola la statistica Wald che verifica l'ipotesi che i coefficienti di Z_{1i}, \dots, Z_{mi} siano tutti zero;
5. La **statistica J** è $J = Wald$

Il test J (continua)

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

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 esogeni sarà **rifiutata**.

J test

```
1 cig_ivreg2 <- feols(log(packpc) ~ log(rincome) |
2                   log(rprice)~log(rincome) + rsaletax + rtax, data = C1995, vcov = "hetero")
3 C1995 <- C1995 |> mutate(uhat = log(packpc) - cig_ivreg2$fitted.values)
4 Jlm <- feols(uhat~log(rincome) + rsaletax + rtax, data = C1995, vcov = "hetero")
5 J <- linearHypothesis(Jlm, c("rsaletax=0", "rtax=0"))
6 J
```

Linear hypothesis test

Hypothesis:
rsaletax = 0
rtax = 0

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

	Df	Chisq	Pr(>Chisq)
1			
2	2	0.2565	0.8796

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

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

```
[1] 0.6125548
```

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:
 $cor(Z_{1i}, u_i) = 0, \dots, cor(Z_{mi}, u_i) = 0$
- 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.

Esercizio

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

```
1 DCig <- Cigarette |> filter(year==1995 | year==1985) |>
2   arrange(state, year) |>
3   group_by(state) |>
4   mutate(Drsaletax = rsaletax - lag(rsaletax),
5           Drtax = rtax - lag(rtax),
6           Dlogrincome = log(rincome) - lag(log(rincome)),
7           Dlogrprice = log(rprice) - lag(log(rprice)),
8           Dlogpackpc = log(packpc) - lag(log(packpc))
9         )
10  summary(feols(Dlogpackpc ~ Dlogrincome |
11              Dlogrprice~Drsaletax + Drtax,
12              data = DCig), stage = 2)
```

TOLS estimation, Dep. Var.: Dlogpackpc, Endo.: Dlogrprice, Instr.: Drsaletax, Drtax

Second stage: Dep. Var.: Dlogpackpc

Observations: 48

Standard-errors: IID

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.052003	0.060495	-0.859636	3.9455e-01
fit_Dlogrprice	-1.202403	0.171193	-7.023677	9.3986e-09 ***
Dlogrincome	0.462030	0.308101	1.499604	1.4070e-01

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

RMSE: 0.088356 Adj. R2: 0.526441

F-test (1st stage), Dlogrprice: stat = 75.7, p = 5.758e-15, on 2 and 44 DoF.

Wu-Hausman: stat = 3.50149, p = 0.067972, on 1 and 44 DoF.

Sargan: stat = 4.83805, p = 0.027838, on 1 DoF.