

# Esercitazione 10

## *Regressione con variabili strumentali (Parte 2)*

### Econometria I

*Sapienza Università di Roma*

June 9, 2025

# Angrist e Evans (1998)

# Dataset "labsup"

```
library(tidyverse)
library(wooldridge)
library(modelsummary)
data("labsup", package = "wooldridge")
```

- ▶ *hours*: ore di lavoro settimanali della madre
- ▶ *kids*: numero di figli
- ▶ *educ*: anni di istruzione
- ▶ *age*: età in anni compiuti
- ▶ *black*: dummy
- ▶ *hispan*: dummy
- ▶ *samesex*: **variabile dummy =1 se i primi due figli sono dello stesso genere**
- ▶ *multi2nd*: **variabile dummy =1 se il secondo parto è gemellare**

# Regressione OLS (Esercizio)

Regressione delle ore lavorate (settimanali) sul numero di figli

```
library(fixest)
ols_labsup <- feols(hours ~ kids + educ + age + black + hispan, data =
labsup, vcov = "hetero")
```

- Come si interpreta il coefficiente associato a *kids* e *educ*?

# Regressione OLS (Esercizio)

	Hours
(Intercept)	12.646 (1.653)
kids	-2.254 (0.116)
educ	0.509 (0.037)
age	0.388 (0.030)
black	1.456 (1.346)
hispan	-5.011 (1.347)
Num.Obs.	31857
Std.Errors	Heteroskedasticity-robust

# Regressione OLS (Esercizio)

- ▶ Avere un figlio in più è associato ad una riduzione, in media, di 2.254 ore settimanali lavorate. A parità delle altre variabili
- ▶ Un anno aggiuntivo di istruzione è associato ad un aumento, in media, di 0.509 ore lavorate. A parità delle altre variabili.

# Regressione con variabili strumentali

## Validità

- ▶ La variabile strumentale (o “strumento”)  $Z$  deve soddisfare le seguenti condizioni:
  1. **Rilevanza:**  $\text{cor}(Z_i, X_i) \neq 0$
  2. **Esogeneità:**  $\text{cor}(Z_i, u_i) = 0$
- ▶ Nel caso di Angrist e Evans (1999), lo strumento *samesex*:
  1. Avere i primi due figli dello stesso genere deve essere associato ad un maggiore numero di figli
  2. Avere i primi due figli dello stesso genere non deve influenzare direttamente l'offerta di lavoro (la composizione dei figli, due maschi o due femmini non dovrebbe essere associato a diversa offerta di lavoro. Ad esempio a un maschio e una femmina)

# Regressioni con variabili strumentali

## *Validità*

La prima condizione può essere testata (come vedremo nel primo stadio). La seconda riguarda la covarianza tra  $Z$  e l'errore non osservato  $u$ . Generalmente non possiamo testare questa assunzione e in molti casi assumiamo  $Cov(Z, u) = 0$  basandoci sul ragionamento (ad esempio teoria). Testeremo le “restrizioni da sovraidentificazione”.



# TSLS

```
iv_labsup <- feols(hours ~ educ + age + black + hispan | kids ~ samesex, data  
= labsup, vcov = "hetero")
```

- ▶ Stima un modello IV in cui *kids* è la variabile endogena, strumentata con *samesex*.
- ▶ Nel primo stadio regredisce l'endogena sullo strumento (*kids* su *samesex*) e le altre variabili esogene
- ▶ Nel secondo stadio regredisce la variabile dipendente sui valori predetti del primo stadio (hours su  $\widehat{kids}$  e le altre variabili)
- ▶ Gli errori standard tengono conto della stima nel primo stadio

# TSLS Risultati (*samesex* strumento)

	Hours
(Intercept)	18.443 (6.499)
fit_kids	-5.071 (3.056)
educ	0.256 (0.276)
age	0.548 (0.176)
black	1.551 (1.378)
hispan	-5.111 (1.378)
Num.Obs.	31857
Std.Errors	Heteroskedasticity-robust

# TSLS (Primo Studio)

```
summary(iv_labsup , stage = 1)
```

```
TSLS estimation - Dep. Var.: kids
                  Endo.    : kids
                  Instr.    : samesex
```

```
First stage: Dep. Var.: kids
```

```
Observations: 31,857
```

```
Standard-errors: Heteroskedasticity-robust
```

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	2.024325	0.077092	26.258620	< 2.2e-16	***
samesex	0.069651	0.010297	6.764490	1.3609e-11	***
educ	-0.089685	0.001994	-44.970520	< 2.2e-16	***
age	0.056773	0.001425	39.845729	< 2.2e-16	***
black	0.032522	0.064401	0.504989	6.1357e-01	
hispan	-0.037187	0.064513	-0.576431	5.6433e-01	

# TSLS (Primo Studio)

---

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

RMSE: 0.918735 Adj. R2: 0.115896

F-test (1st stage): stat = 45.8, p = 1.359e-11, on 1 and 31,851 DoF.

# TSLS (Primo Studio)

```
fs_labsup <- feols(kids ~ samesex + educ + age + black + hispan, data =  
labsup, vcov = "hetero")  
fs_labsup
```

OLS estimation, Dep. Var.: kids

Observations: 31,857

Standard-errors: Heteroskedasticity-robust

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	2.024325	0.077092	26.258620	< 2.2e-16	***
samesex	0.069651	0.010297	6.764490	1.3609e-11	***
educ	-0.089685	0.001994	-44.970520	< 2.2e-16	***
age	0.056773	0.001425	39.845729	< 2.2e-16	***
black	0.032522	0.064401	0.504989	6.1357e-01	
hispan	-0.037187	0.064513	-0.576431	5.6433e-01	

---

# TSLS (Primo Studio)

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
RMSE: 0.918735   Adj. R2: 0.115896
```

# TSLS (Secondo Studio)

```
labsup$kids_hat <- predict(fs_labsup)
feols(hours ~ kids_hat + educ + age + black + hispan, data = labsup, vcov =
"hetero")
```

OLS estimation, Dep. Var.: hours

Observations: 31,857

Standard-errors: Heteroskedasticity-robust

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	18.442906	6.471201	2.849997	0.00437476	**
kids_hat	-5.071274	3.045439	-1.665203	0.09588215	.
educ	0.256421	0.275049	0.932272	0.35120302	
age	0.548170	0.175308	3.126892	0.00176824	**
black	1.551443	1.345024	1.153469	0.24872669	
hispan	-5.110861	1.345937	-3.797251	0.00014658	***
---					

# TSLS (Secondo Stadio)

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
RMSE: 18.9    Adj. R2: 0.057635
```

- ▶ Con questa procedura otteniamo gli stessi risultati di `iv_labsup`. Ma gli errori standard **non** sono corretti (non tengono conto della stima nel primo stadio)



# Rilevanza dello strumento

- ▶ Calcoliamo la statistica  $F$  per la verifica dell'ipotesi che i coefficienti degli strumenti siano tutti 0 nel **primo stadio** della regressione TSLS
- ▶ Una statistica  $F < 10$  indica che gli strumenti sono deboli (Staicker and Stock, 1997; Stock and Yogo, 2005),
- ▶ Non è la statistica  $F$  complessiva, ma testiamo che congiuntamente i coefficienti degli strumenti siano uguali a zero
- ▶ Se la statistica Wald del primo stadio è minore di  $m \times 10$ , allora l'insieme degli strumenti è debole. **Nota: Wald =  $m \times F$**
- ▶ Alcuni studi suggeriscono valori critici più alti o altri test (Montiel Olea and Pfluegger, 2013; Kleibergen-Paap rk statistics)

# Rilevanza dello strumento

```
library(car)
linearHypothesis(fs_labsup , "samesex=0")
```

Linear hypothesis test:  
samesex = 0

Model 1: restricted model

Model 2: kids ~ samesex + educ + age + black + hispan

	Res.Df	Df	Chisq	Pr(>Chisq)
1	31852			
2	31851	1	45.758	1.338e-11 ***

---

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

# Rilevanza dello strumento

- ▶ La statistica Wald è pari a  $48.758 > a 1*10$

# Rilevanza dello strumento (F)

```
linearHypothesis(fs_labsup , "samesex=0", test="F")
```

Linear hypothesis test:

samesex = 0

Model 1: restricted model

Model 2: kids ~ samesex + educ + age + black + hispan

	Res.Df	Df	F	Pr(>F)
1	31852			
2	31851	1	45.758	1.361e-11 ***

---

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

## multi2nd (secondo parto gemellare) come strumento

```
iv_labsup <- feols(hours ~ educ + age + black + hispan | kids ~ multi2nd,  
data = labsup, vcov = "hetero")  
iv_labsup
```

```
TSLS estimation - Dep. Var.: hours  
                  Endo.      : kids  
                  Instr.     : multi2nd
```

```
Second stage: Dep. Var.: hours
```

```
Observations: 31,857
```

```
Standard-errors: Heteroskedasticity-robust
```

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	13.109306	3.353528	3.90911	9.2829e-05	***
fit_kids	-2.478705	1.420373	-1.74511	8.0976e-02	.
educ	0.488897	0.132255	3.69663	2.1884e-04	***
age	0.400960	0.085949	4.66510	3.0971e-06	***

## multi2nd (secondo parto gemellare) come strumento

```
black      1.463484    1.347660    1.08594 2.7751e-01
hispan     -5.019231    1.349187   -3.72019 1.9942e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 18.8    Adj. R2: 0.068733
F-test (1st stage), kids: stat = 189.4      , p < 2.2e-16 , on 1 and 31,851
DoF.
                Wu-Hausman: stat =    0.022933, p = 0.879633, on 1 and 31,850
DoF.
```

Rilevanza ed esogeneità in questo caso:

- **Rilevanza:** secondo parto gemellare deve essere associato ad un maggiore numero di figli

## multi2nd (secondo parto gemellare) come strumento

- ▶ **Esogeneità:** Avere un secondo parto gemellare non deve influenzare direttamente l'offerta di lavoro della madre

# Rilevanza dello strumento (2 strumenti)

Calcoliamo il primo stadio della regressione:

```
fs_labsup_overid <- feols(kids ~ samesex + multi2nd + educ + age + black +  
hispan, data = labsup, vcov = "hetero")
```

Testiamo la rilevanza di entrambi gli strumenti: *multi2nd* e *samesex*

$234.72 > 2 \times 10$



# Rilevanza dello strumento (2 strumenti)

```
linearHypothesis(fs_labsup_overid, c("samesex=0", "multi2nd=0"))
```

Linear hypothesis test:

samesex = 0

multi2nd = 0

Model 1: restricted model

Model 2: kids ~ samesex + multi2nd + educ + age + black + hispan

	Res.Df	Df	Chisq	Pr(>Chisq)
1	31852			
2	31850	2	234.72	< 2.2e-16 ***

---

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

# Restrizioni da Sovraidentificazione

## *J di Sargan*

Quando il numero di strumenti disponibili  $m$  è maggiore del numero di variabili endogene  $k$ , il modello è sovraidentificato.

- ▶  $H_0$ : tutti gli strumenti sono esogeni
- ▶  $H_1$ : almeno uno degli strumenti è endogeno

# J di Sargan

## Procedura

1. Stimiano la regressione TSLS
2. Si ottengono i residui della regressione TSLS:  
$$\hat{u}_i = Y_i - \hat{Y}_i$$
3. Si esegue una regressione dei residui  $\hat{u}_i$  sugli strumenti  $Z_1, \dots, Z_m$  e le variabili esogene  $W_1, \dots, W_r$
4. Se gli strumenti fossero esogeni i coefficienti degli strumenti nella regressione di  $\hat{u}_i$  dovrebbero essere uguali a zero (statisticamente non significativi)
5. Si calcola come  $J = mF$ . Quindi  $J = Wald$
6. In grandi campioni si distribuisce come una chi-quadrato con  $m - k$  gradi di libertà ( $\chi_{m-k}$ )

$m - k$  è il grado di sovraidentificazione ( $k$  sono i regressori endogeni)

# J di Sargan

```
iv_labsup_overid <- feols(hours ~ educ + age + black + hispan | kids ~  
  samesex + multi2nd, data = labsup, vcov = "hetero")
```

```
labsup <- labsup |> mutate(uhat = hours - iv_labsup_overid$fitted.values)  
Jlm <- feols(uhat ~ samesex + multi2nd + educ + age + black + hispan, data =  
  labsup, vcov = "iid")
```

# J di Sargan

```
linearHypothesis(Jlm, c("samesex=0", "multi2nd=0"))
```

Linear hypothesis test:

samesex = 0

multi2nd = 0

Model 1: restricted model

Model 2: uhat ~ samesex + multi2nd + educ + age + black + hispan

	Res.Df	Df	Chisq	Pr(>Chisq)
1	31852			
2	31850	2	0.583	0.7471

# J di Sargan

## ! Attenzione

Il test di sovraidentificazione restituisce una statistica  $J = 0.583$ .

La statistica  $J$  si distribuisce come una  $\chi^2_{(m-k)}$ . Nel nostro caso  $m - k = 1$ . Il valore da confrontare **non** è 6 ma **3.814** ( $\alpha = 0.05$ )

Quindi  $0.583 < 3.814$ . Non rigetto  $H_0$ . Il p-value corretto è:

```
pchisq(0.583, df = 1, lower.tail = FALSE)
```

```
[1] 0.4451388
```

quindi non rifiutiamo l'ipotesi nulla di esogeneità degli strumenti

- Non rigetto  $H_0$ . Tutti gli strumenti sono esogeni.

# Regressione IV con due strumenti

	Hours
(Intercept)	14.152*** (3.109)
fit_kids	-2.985* (1.283)
educ	0.443*** (0.120)
age	0.430*** (0.079)
black	1.481 (1.351)
hispan	-5.037*** (1.352)
Std.Errors	Heteroskedasticity-robust

+ p < 0.1, \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001