Esercitazione 7

Regressione con variabili strumentali

Econometria I

Sapienza Università di Roma

June 10, 2025

Card (1995)

College Proximity

Abstract

- A convincing analysis of the causal link between schooling and earnings requires an exogenous source of variation in education outcomes.
- This paper explores the use of college proximity as an exogenous determinant of schooling.
- Men who grew up in local labor markets with a nearby college have significantly higher education and earnings than other men.
- The education and earnings gains are concentrated among men with poorly-educated parents – men who would otherwise stop schooling
- IV estimates of the return to schooling are higher than conventional OLS estimates

Dataset “card”

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

- $wage$: sono le retribuzioni orarie **in centesimi** di dollari
- $educ$: sono gli anni di istruzione
- $exper$: anni di esperienza
- $smsa$: residenza in area metropolitana (dummy)
- $black$: se la persona è nera (dummy)
- $south$: se l’individuo risiede al sud
- $nearc4$: dummy uguale ad 1 se l’individuo vive vicino a un college di 4 anni
- $nearc2$: dummy uguale ad 1 se l’individuo vive vicino a un college di 2 anni

Regressione OLS (Esercizio)

Regressione del salario orario in centesimi di dollari sugli anni di istruzione e gli anni di esperienza.

```r
library(fixest)
wage_educ <- feols(wage ~ educ + exper, data = card, vcov = "hetero")
logwage_educ <- feols(log(wage) ~ educ + exper, data = card, vcov = "hetero")
```

Cosa cambia se il salario viene espresso in dollari? (wage / 100)

```r
wage_educdoll <- feols(wage/100 ~ educ + exper, data = card, vcov = "hetero")
logwage_educdoll <- feols(log(wage/100) ~ educ + exper, data = card, vcov =
"hetero")
```

Regressione OLS (Esercizio)

	Wage	Log Wage
(Intercept)	-375.588	4.666
	(42.504)	(0.065)
educ	55.055	0.093
	(2.480)	(0.004)
exper	25.142	0.041
	(1.513)	(0.002)
Num.Obs.	3010	3010
R2	0.181	0.181
Std.Errors	Heteroskedasticity-robust	Heteroskedasticity-robust

- Un anno aggiuntivo di istruzione è associato ad un aumento in media del salario di 55.055 centesimi di dollari (Un anno aggiuntivo di istruzione è associato ad aumento del salario, in media, di circa il 9.3%) a parità di anni esperienza.

Regressione OLS (Esercizio)

	Wage	Log Wage
(Intercept)	-3.756	0.061
	(0.425)	(0.065)
educ	0.551	0.093
	(0.025)	(0.004)
exper	0.251	0.041
	(0.015)	(0.002)
Num.Obs.	3010	3010
R2	0.181	0.181
Std.Errors	Heteroskedasticity-robust	Heteroskedasticity-robust

- Nella prima regressione i coefficienti e errori standard sono divisi per 100.

Regressione OLS (Esercizio)

- Nella regressione "Log Wage" i coefficienti associati a $educ$ e $exper$ rimangono invariati. L'intercetta è uguale a $\beta_0 - log(100)$ cioè $4.666 - log(100) = 4.666 - 4.605 = 0.061$
- Questo perché $log(\frac{wage}{100}) = log(wage) - log(100)$
- Se avessi moltiplicato come nel caso dell'**Esercitazione 4** per 140 (traformazione costante da ore a mesi) otterrei $\beta_0 + log(140)$
- Gli errori standard rimangono invariati (nella Regressione "Log Wage")
- L'R^2 rimane invariato in entrambe le regressioni

Regressione con variabili strumentali

Validità

- La variabile strumentale (o “strumento”) Z deve soddisfare le seguenti condizioni:
  1. **Rilevanza:** $cor(Z_i, X_i) \neq 0$
  2. **Esogeneità:** $cor(Z_i, u_i) = 0$
- Nel caso di Card (1995):
  1. Vicinanza al college deve essere associata a maggiori anni di istruzione
  2. La vicinanza al college deve essere incorrelata con l’errore. La vicinanza al college deve influenzare il salario (futuro) solo indirettamente attraverso gli anni di istruzione

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

Uno strumento e una variabile endogena

```
iv_card <- feols(log(wage) ~ 1 | educ ~ nearc4, data = card, vcov = "hetero")
iv_card
```

```
TSLS estimation - Dep. Var.: log(wage)
                  Endo.    : educ
                  Instr.   : nearc4
Second stage: Dep. Var.: log(wage)
Observations: 3,010
Standard-errors: Heteroskedasticity-robust
            Estimate Std. Error  t value   Pr(>|t|)
(Intercept) 3.767472   0.346742 10.86535  < 2.2e-16 ***
fit_educ    0.188063   0.026143  7.19373 7.9229e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

TSLS

```
RMSE: 0.556673   Adj. R2: -0.574414
F-test (1st stage), educ: stat = 63.9, p = 1.838e-15, on 1 and 3,008 DoF.
              Wu-Hausman: stat = 48.5, p = 4.141e-12, on 1 and 3,007 DoF.
```

- Usiamo ~ 1 perché non abbiamo altre variabili
- Stima un modello IV in cui $educ$ è endogena, strumentata con $nearc4$, e l'unica variabile esplicativa (oltre a $educ$) è una costante.
- Nel primo stadio regredisce l'endogena sullo strumento ($educ$ su $nearc4$)
- Nel secondo stadio regredisce la variabile dipendente sui valori predetti del primo stadio (log(wage) su $\widehat{educ}$)
- Gli errori standard tengono conto della stima nel primo stadio

TSLS (Primo Stadio da `iv_card`)

```r
summary(iv_card, stage = 1)
```

```
TSLS estimation - Dep. Var.: educ
                  Endo.    : educ
                  Instr.   : nearc4
First stage: Dep. Var.: educ
Observations: 3,010
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept) 12.698015   0.090220 140.74510  < 2.2e-16 ***
nearc4       0.829019   0.106694   7.77005 1.0684e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 2.64848   Adj. R2: 0.02048
F-test (1st stage): stat = 63.9, p = 1.838e-15, on 1 and 3,008 DoF.
```

TSLS (Primo Stadio)

```
fs_card <- feols(educ ~ nearc4, data = card, vcov = "hetero")
fs_card
```

```
OLS estimation, Dep. Var.: educ
Observations: 3,010
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept) 12.698015   0.090220 140.74510  < 2.2e-16 ***
nearc4       0.829019   0.106694   7.77005 1.0684e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 2.64848   Adj. R2: 0.02048
```

TSLS (Secondo Stadio)

```
card$educ_hat <- predict(fs_card)
feols(log(wage) ~ educ_hat, data = card, vcov = "hetero")
```

```
OLS estimation, Dep. Var.: log(wage)
Observations: 3,010
Standard-errors: Heteroskedasticity-robust
            Estimate Std. Error  t value  Pr(>|t|)
(Intercept) 3.767472   0.272493 13.82595 < 2.2e-16 ***
educ_hat    0.188063   0.020545  9.15350 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.437744   Adj. R2: 0.026448
```

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

Derivazione diretta

Lo stimatore della regressione con variabili strumentali può essere ottenuto in questo modo:

$$\beta_1^{TSLS} = \frac{cov(Z, Y)}{cov(Z, X)}$$

- Notate come $cov(Z, X)$ è ciò che stimiamo nel primo stadio. Se fosse uguale a zero non potremmo stimare β_1^{TSLS}

Nel nostro caso utilizzando i dati:

```
cov(card$nearc4, card$lwage)/cov(card$nearc4, card$educ)
```

```
[1] 0.1880626
```

“Forma Ridotta”

Definizioni

- Il termine “Forma Ridotta” proviene dalla tradizione dei modelli ad equazioni simultanee (SEM): nel modello in forma ridotta le endogene sono espresse come funzione delle esogene.

Il libro definisce la forma ridotta di X, che di fatto coincide con il primo stadio:

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

Sostituendo X_i nella seguente:

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

Oteniamo “forma ridotta” per Y:

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

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 =** $\boldsymbol{m \times F}$
- Alcuni studi suggeriscono valori critici più alti o altri test (Montiel Olea and Pflueegger, 2013; Kleibergen-Paap rk statistics)

Rilevanza dello strumento

```
library(car)
linearHypothesis(fs_card , "nearc4=0")
```

```

Linear hypothesis test:
nearc4 = 0

Model 1: restricted model
Model 2: educ ~ nearc4

  Res.Df Df  Chisq Pr(>Chisq)
1   3009
2   3008  1 60.374  7.845e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rilevanza dello strumento

```r
library(car)
linearHypothesis(fs_card , "nearc4=0", test="F")
```

```

Linear hypothesis test:
nearc4 = 0

Model 1: restricted model
Model 2: educ ~ nearc4

  Res.Df Df      F    Pr(>F)
1   3009
2   3008  1 60.374 1.068e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rilevanza dello strumento (2 strumenti)

Calcoliamo il primo stadio della regressione:

```
fs_card_overid <- feols(educ ~ nearc4 + nearc2 + exper + black +smsa + south
+ married, data = card, vcov = "hetero")
```

Rilevanza dello strumento (2 strumenti)

```r
linearHypothesis(fs_card_overid, c("nearc4=0", "nearc2=0"))
```

```


Linear hypothesis test:
nearc4 = 0
nearc2 = 0

Model 1: restricted model
Model 2: educ ~ nearc4 + nearc2 + exper + black + smsa + south + married

  Res.Df Df  Chisq Pr(>Chisq)
1   2997
2   2995  2 19.088  7.162e-05 ***
---
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 $Z_1, ..., Z_m$ e le variabili esogene $W_1, ..., W_r$
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 (χ_{m-k})

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

J di Sargan

```
iv_card_overid <- feols(log(wage) ~ exper + black +smsa + south + married |
educ ~ nearc4 + nearc2, data = card, vcov = "hetero")
```

```
library(car)
library(dplyr)
card <- card |> mutate(uhat = log(wage) - iv_card_overid$fitted.values)
Jlm <- feols(uhat~ nearc4 + nearc2 + exper + black +smsa + south + married,
data = card, vcov = "iid")
```

J Test

```r
linearHypothesis(Jlm, c("nearc4=0", "nearc2=0"))
```

```

Linear hypothesis test:
nearc4 = 0
nearc2 = 0

Model 1: restricted model
Model 2: uhat ~ nearc4 + nearc2 + exper + black + smsa + south + married

  Res.Df Df  Chisq Pr(>Chisq)
1   2997
2   2995  2 9.3421   0.009362 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

J di Sargan

! Attenzione

Il test di sovraidentificazione restituisce una statistica $J = 9.3421$.

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 **3.84** $(\alpha = 0.05)$

Quindi 9.3421 > 3.84. Rigetto H_0. Il p-value corretto è:

```r
pchisq(9.3421, df = 1, lower.tail = FALSE)
```

```
[1] 0.002239488
```

o in modo equivalente `1 - pchisq(J$Chisq[2], df = 1)` quindi rifiutiamo l'ipotesi nulla di esogeneità degli strumenti. **Rigetto H_0. Almeno uno degli strumenti è esogeno**