

# **Esercizi sulla regressione e lo stimatore OLS**

**Corso di Econometria**

**Professor Valerio Potì**

## Esercizi sulla regressione and lo stimatore dei minimi quadrati (OLS)

### Esercizio B.1 (verifica di ipotesi su coefficienti)

Si consideri la regressione del punteggio medio ( $TestScore_i$ ) di classe nei test INVALSI sulla dimensione delle classi ( $CS$ ):

$$TestScore_i = \beta_0 + \beta_1 CS_i + u_i$$

Un ricercatore, usando i dati appunto sulla dimensione delle classi e il punteggio medio nei test per 50 classi stima, col metodo OLS, la seguente regressione (errori standard tra parentesi):

$$\widehat{TestScore} = \underbrace{640,3}_{(23,5)} - \underbrace{4,93}_{(2,02)} \times CS, \quad R^2 = 0,11, \quad SER = 8,7$$

- Si costruisca un intervallo di confidenza di livello 95% per  $\beta_1$ , la pendenza della regressione. Possiamo rigettare l'ipotesi nulla  $H_0 : \beta_1 = 0$  al livello di significatività del 5% in un test bilaterale? E al livello del 2.5% in un test unilaterale (contro l'ipotesi alternativa che  $\beta_1 < 0$ )?
- Si calcoli il valore- $p$  di un test bilaterale per l'ipotesi nulla  $H_0 : \beta_1 = 0$ . Si rigetta l'ipotesi nulla (i) al livello di significatività del 5% e (ii) al livello di significatività dell'1%?
- Si calcoli il valore- $p$  di un test bilaterale per l'ipotesi nulla  $H_0 : \beta_1 = -5,0$ . Senza calcoli aggiuntivi, si deduca se  $-5,0$  è contenuto nell'intervallo di confidenza al 95% per  $\beta_1$
- Si costruisca un intervallo di confidenza al 90% per  $\beta_1$

### Soluzione

- (a) L'intervallo di confidenza al 95% è per  $\beta_1$  è  $\{-4,93 \pm 1,96 \times 2,02\} = \{-4,93 \pm 3,96\} = [-8,89, -0,9708]$ , ovvero

$$\{-8,889 \leq \beta_1 \leq -0,9708\}$$

Dunque, poiché  $\beta_1 = 0$  è al di fuori di questo intervallo, possiamo rigettare l'ipotesi nulla  $H_0 : \beta_1 = 0$  al livello di significatività del 5% in un test bilaterale ed al livello del 2.5% in un test unilaterale (contro l'ipotesi alternativa che  $\beta_1 < 0$ ).

- (b) La statistica  $t$  è

$$t^{act} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = -\frac{4,93}{2,02} = -2,44$$

Per calcolare il valore- $p$  del test  $H_0 : \beta_1 = 0$  contro  $H_1 : \beta_1 \neq 0$  possiamo utilizzare la distribuzione per campioni finiti oppure, se assumiamo che un campione con 50 osservazioni è grande, quella per campioni infiniti.

Se vogliamo utilizzare la distribuzione per campioni finiti, questa sarà una  $t$ -student con 50 gradi di libertà e dunque la funzione di Excel da utilizzare sarà DISTRIB.T. Avremmo, in questo caso, quanto segue:

$$\text{valore-}p = 2 \times \phi(-|t^{act}|) = 2 \times \phi(-2,44) = 2 \times 0,009 = 0,018 = 1,8\%$$

Il valore- $p$  del test  $H_0 : \beta_1 = 0$  contro  $H_1 : \beta_1 \neq 0$  secondo la distribuzione in campioni infiniti (dunque utilizzando la distribuzione asintotica), ovvero quella di una variabile normale standardizzata, è invece

$$\text{valore-}p = 2 \times \phi(-|t^{act}|) = 2 \times \phi(-2,44) = 2 \times 0,007 = 0,014 = 1,4\%$$

che si trova tramite Excel utilizzando la funzione DISTRIB.NORM.ST.

Il valore-p è dunque in entrambi i casi maggiore di 1% ma minore di 5%. Dunque, si rifiuta la ipotesi nulla al livello di significatività del 5% ma non a quello dell'1% sia secondo la distribuzione per campioni infiniti che per quella in campioni finiti.

(c) La statistica  $t$  è

$$t^{act} = \frac{\hat{\beta}_1 - (-5)}{SE(\hat{\beta}_1)} = \frac{-4,93 - (-5)}{2,02} = \frac{0,07}{2,02} = 0,034653$$

Utilizzando la distribuzione per campioni infiniti, il valore-p del test  $H_0: \beta_1 = 0$  contro  $H_1: \beta_1 \neq 0$  è dunque

$$\text{valore-p} = 2 \times \phi(-|t^{act}|) = 2 \times \phi(-0,034653) = 0,9723 \approx 97\%$$

Utilizzando la distribuzione per campioni finiti con  $n - k - 1 = 50 - 1 - 2 = 48$  gradi di libertà, il valore-p del test  $H_0: \beta_1 = 0$  contro  $H_1: \beta_1 \neq 0$  è dunque

$$\text{valore-p} = 2 \times \phi(-|t^{act}|) = 2 \times \phi(-0,034653) = 0,9724 \approx 97\%$$

Il valore-p è dunque maggiore di 10% sia secondo la distribuzione per campioni finiti che per quella per campioni infiniti. Dunque, non si rifiuta la ipotesi nulla a qualunque livello convenzionale di significatività. Pertanto, possiamo concludere che  $-5,0$  è contenuto nell'intervallo di confidenza al 95% per  $\beta_1$ .

(d) L'intervallo di confidenza è

$$\beta_0 \in \{-4,93 \pm 1,64 \times 2,02\}$$

Ovvero

$$-8,24 \leq \beta_0 \leq -1,61$$

## Esercizio B.2 (Esercizio in Gretl su regressione vs. test su differenze di medie di sotto-campioni)

### Preambolo:

Si ricordi che, quando abbiamo confrontato i distretti con dimensioni delle classi “piccole” ( $STR < 20$ ) e “grandi” ( $STR \geq 20$ ), abbiamo ottenuto i risultati seguenti:

Dimensione classe	Punteggio medio ( $\bar{Y}$ )	Deviazione standard ( $s_Y$ )	$n$
Piccola	657,4	19,4	238
Grande	650,0	17,9	182

Questi risultati ci dicono che vi è un effetto significativo della numerosità delle classi sul rendimento scolastico, infatti calcolammo la seguente statistica del test  $t$  pari

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{657,4 - 650,0}{\sqrt{\frac{19,4^2}{238} + \frac{17,9^2}{182}}} = \frac{7,4}{1,83} = 4,05$$

Poiché  $|t| = 4,05 > 1,96$ , rifiutammo (al livello di significatività del 5%) l’ipotesi nulla che le due medie coincidano.

Raggiungiamo conclusioni simili tramite la regressione di TESTSCR su STR, come mostrato dalla seguente schermata di Gretl:

$$TESTSCR_i = \alpha + \beta STR_i + u_i$$

Modello 1: OLS, usando le osservazioni 1-420  
Variabile dipendente: TESTSCR

	coefficiente	errore std.	rapporto t	p-value
const	698,933	9,46749	73,82	6,57e-242 ***
STR	-2,27981	0,479826	-4,751	2,78e-06 ***
Media var. dipendente	654,1565	SQM var. dipendente	19,05335	
Somma quadr. residui	144315,5	E.S. della regressione	18,58097	
R-quadro	0,051240	R-quadro corretto	0,048970	
F(1, 418)	22,57511	P-value (F)	2,78e-06	
Log-verosimiglianza	-1822,250	Criterio di Akaike	3648,499	
Criterio di Schwarz	3656,580	Hannan-Quinn	3651,693	

Note: SQM = scarto quadratico medio; E.S. = errore standard

### Domanda:

Nel primo approccio che abbiamo visto non si usa la regressione mentre nell’altro la si usa.

La conclusione che raggiungiamo con i due approcci è simile ma non identica in termini di statistica del test  $t$  e, dunque, del livello di significatività a cui rifiutiamo l’ipotesi nulla che non ci sia alcun effetto.

Esiste un modo di raggiungere la stessa conclusione nei due approcci?

Risposta:

Si regredisca TESTSCR su una variabile di comodo (“dummy” in inglese) che assume valore 1 se STR < 20 e valore 0 altrimenti. È come stimare l’effetto di avere classi con meno o più di 20 studenti e infatti ci dà un risultato che ci porta alle stesse conclusioni di quello che abbiamo precedentemente ottenuto senza far ricorso alla regressione:

```
Modello 3: OLS, usando le osservazioni 1-420
Variabile dipendente: TESTSCR
```

	coefficiente	errore std.	rapporto t	p-value	
const	649,979	1,38772	468,4	0,0000	***
dummy_small	7,37241	1,84347	3,999	7,52e-05	***

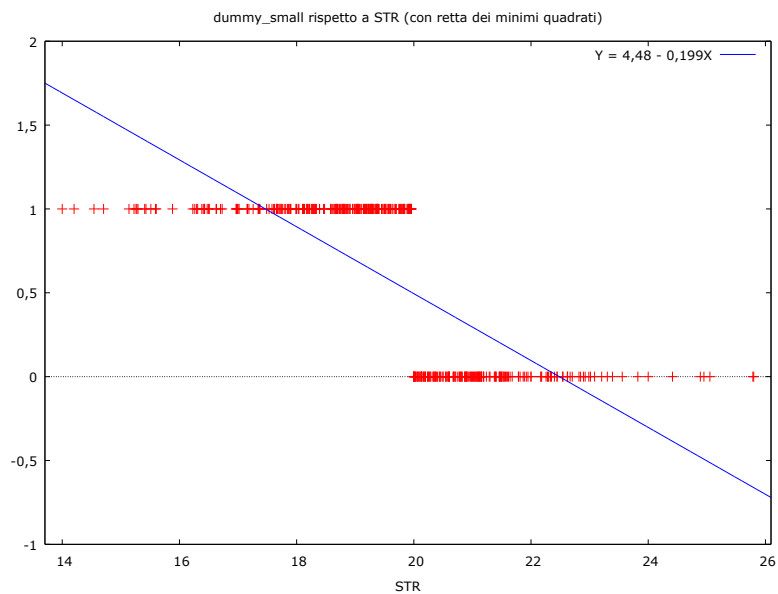
Media var. dipendente	654,1565	SQM var. dipendente	19,05335
Somma quadr. residui	146504,0	E.S. della regressione	18,72133
R-quadro	0,036852	R-quadro corretto	0,034548
F(1, 418)	15,99354	P-value (F)	0,000075
Log-verosimiglianza	-1825,410	Criterio di Akaike	3654,821
Criterio di Schwarz	3662,901	Hannan-Quinn	3658,015

Note: SQM = scarto quadratico medio; E.S. = errore standard

NB: come si crea la variabile di comodo in Gretl? Si fa così:

Aggiungi -> Definisci nuova variabile -> nella finestra di dialogo che si apre (che dice di “inserire la formula per la nuova variabile) digitare “dummy\_small = STR < 20” -> verificare che la variabile desiderata sia ora inclusa tra quelle in memoria.

Per verificare che la variabile dummy creata sia davvero quella che ci serve, conviene mostrare in un grafico a dispersione come varia al variare di STR:



Per creare questo grafico fare così:

Visualizza -> Grafico -> X-Y a dispersione -> nella finestra di dialogo che segue, selezionare dummy\_small come la variabile all’asse delle Y e STR come quella all’asse delle X.



### Esercizio B.3 (Esercizio in Gretl su regressione e test di specificazione)

Si effettui ora prima una regressione di TESTSCR su STR e EL\_PCT e poi una regressione di TESTSCR su STR, EL\_PCT, ENRL\_TOT e MEAL\_PCT, come mostrato dalla seguente schermata di Gretl:

- Si calcoli la statistica F del test che l' $R^2$  è significativamente differente da zero e la si confronti con la statistica riportata nella stampa di Gretl.
- Si sottoponga a verifica l'ipotesi nulla che l' $R^2$  del primo modello non sia significativamente più piccolo di quello del secondo modello.
- Potremmo raggiungere la stessa conclusione guardando semplicemente i test di significatività dei singoli coefficienti?

### Risposte

Le stime dei modelli considerati sono le seguenti:

Modello 1: OLS, usando le osservazioni 1-420  
Variabile dipendente: TESTSCR

	<i>Coefficiente</i>	<i>Errore Std.</i>	<i>rapporto t</i>	<i>p-value</i>	
const	686,032	7,41131	92,57	<0,0001	***
STR	-1,10130	0,380278	-2,896	0,0040	***
EL_PCT	-0,649777	0,0393425	-16,52	<0,0001	***
Media var. dipendente	654,1565	SQM var. dipendente	19,05335		
Somma quadr. residui	87245,29	E.S. della regressione	14,46448		
R-quadro	0,426431	R-quadro corretto	0,423680		
F(2, 417)	155,0136	P-value(F)	4,62e-51		
Log-verosimiglianza	-1716,561	Criterio di Akaike	3439,123		
Criterio di Schwarz	3451,243	Hannan-Quinn	3443,913		

Modello 2: OLS, usando le osservazioni 1-420  
Variabile dipendente: TESTSCR

	<i>Coefficiente</i>	<i>Errore Std.</i>	<i>rapporto t</i>	<i>p-value</i>	
const	700,959	4,78275	146,6	<0,0001	***
STR	-1,05248	0,247185	-4,258	<0,0001	***
EL_PCT	-0,131490	0,0343675	-3,826	0,0002	***
ENRL_TOT	0,000107882	0,000126871	0,8503	0,3956	
MEAL_PCT	-0,544480	0,0218673	-24,90	<0,0001	***
Media var. dipendente	654,1565	SQM var. dipendente	19,05335		
Somma quadr. residui	34238,65	E.S. della regressione	9,083103		
R-quadro	0,774908	R-quadro corretto	0,772738		
F(4, 415)	357,1727	P-value(F)	6,7e-133		
Log-verosimiglianza	-1520,134	Criterio di Akaike	3050,268		
Criterio di Schwarz	3070,469	Hannan-Quinn	3058,252		

Per quanto riguarda le risposte a ciascuna domanda, abbiamo:

(a) La statistica F del primo modello è

$$F = \frac{\frac{(R_{non\ ristretto}^2 - R_{ristretto}^2)}{q}}{\frac{(1 - R_{non\ ristretto}^2)}{(n - k_{non\ ristretto} - 1)}} = \frac{(0,426431 - 0) \div 2}{(1 - 0,426431) \div (420 - 2 - 1)} = 155,0136$$

La statistica F del primo modello è

$$F = \frac{\frac{(R_{non\ ristretto}^2 - R_{ristretto}^2)}{q}}{\frac{(1 - R_{non\ ristretto}^2)}{(n - k_{non\ ristretto} - 1)}} = \frac{(0,774908 - 0) \div 4}{(1 - 0,774908) \div (420 - 4 - 1)} = 357,1727$$

Sono gli stessi valori riportati da Gretl?

(b) La statistica F richiesta è

$$F = \frac{\frac{(R_{non\ ristretto}^2 - R_{ristretto}^2)}{q}}{\frac{(1 - R_{non\ ristretto}^2)}{(n - k_{non\ ristretto} - 1)}} = \frac{(0,774908 - 0,426431) \div 2}{(1 - 0,774908) \div (420 - 4 - 1)} = 321,24$$

Se l'ipotesi nulla è corretta, questa avrà una distribuzione F con  $q = 4 - 2 = 2$  ed  $n - k_{non\ ristretto} - 1 = 420 - 4 - 1 = 415$  gradi di libertà. Secondo Gretl, il valore- $p$  della statistica F del test è  $5,11101e-085$  (ovvero, 0,000). La ipotesi nulla, dunque, si rifiuta e concludiamo che l' $R^2$  del primo modello, probabilmente, è effettivamente più piccolo di quello del secondo e dunque, sempre probabilmente, I regressori aggiuntivi sono significativi.

(c) La conclusione di cui sopra è implicata dal fatto che il coefficiente di uno dei regressori aggiuntivi è significativo (ne è una condizione sufficiente anche se non necessaria).

Una domanda sorge spontanea: perché si usano allora i test F?