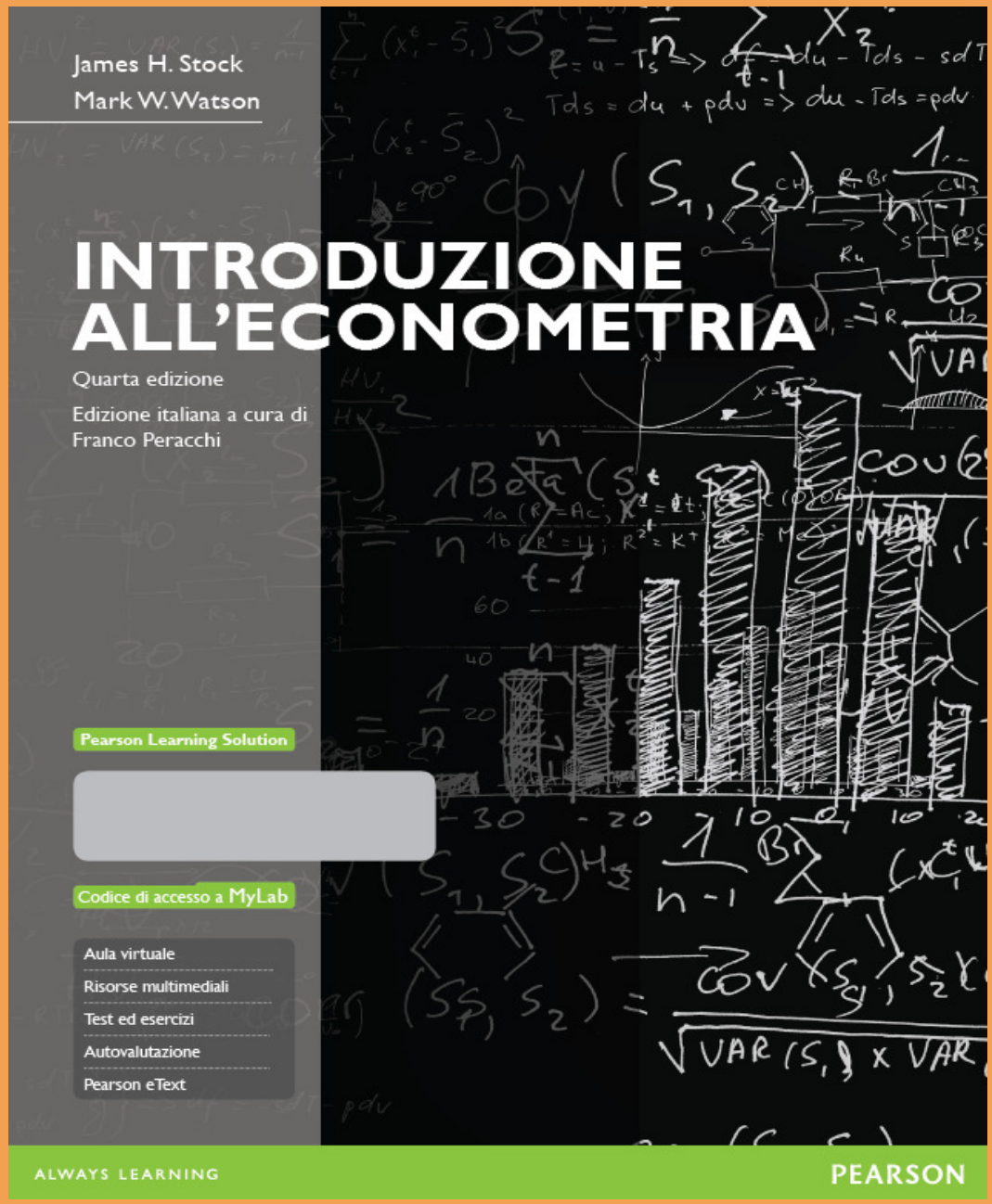
Capitolo 7

Verifica di ipotesi e intervalli di confidenza nella regressione multipla



Sommario

1. Verifica di ipotesi e intervalli di confidenza per un singolo coefficiente
2. Verifica di ipotesi congiunte su più coefficienti
3. Altri tipi di ipotesi che implicano più coefficienti
4. Variabili di interesse, variabili di controllo e come decidere quali variabili includere in un modello di regressione

Verifica di ipotesi e intervalli di confidenza per un singolo coefficiente (Paragrafo 7.1)

- Per verifica di ipotesi e intervalli di confidenza nella regressione multipla si segue la stessa logica utilizzata per la pendenza in un modello a singolo regressore.
- $\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}}$ è approssimativamente distribuita come N(0,1) (TLC).
- Perciò le ipotesi su β1 possono essere verificate mediante la consueta statistica-t e gli intervalli di confidenza costruiti come $\{\hat{\beta}_1 \pm 1{,}96 \times \text{SE}(\hat{\beta}_1)\}$.
- Lo stesso per β2,…, βk.

Esempio: dati sulle dimensioni delle classi in California

1. $\widehat{TestScore} = 698{,}9 - 2{,}28 \times STR$
   $\qquad\qquad (10{,}4) \quad (0{,}52)$
2. $\widehat{TestScore} = 686{,}0 - 1{,}10 \times STR - 0{,}650 PctEL$
   $\qquad\qquad (8{,}7) \quad (0{,}43) \qquad\qquad (0{,}031)$

- Il coefficiente di *STR* in (2) è l'effetto medio su *TestScore* del cambio di unità in *STR*, mantenendo costante la percentuale di studenti non di madrelingua nel distretto
- Il coefficiente di *STR* si dimezza
- L'intervallo di confidenza al 95% per il coefficiente di *STR* in (2) è $\{-1{,}10 \pm 1{,}96 \times 0{,}43\} = (-1{,}95, -0{,}26)$
- Il test della statistica-t $\beta_{STR} = 0$ è $t = -1{,}10/0{,}43 = -2{,}54$, perciò rifiutiamo l'ipotesi nulla al livello di significatività del 5%

Errori standard nella regressione multipla in STATA

```
reg testscr str pctel, robust;
Regression with robust standard errors                Number of obs =     420
                                                      F(  2,   417) =  223.82
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4264
                                                      Root MSE      =  14.464

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472    -2.54   0.011    -1.95213   -.2504616
       pctel |  -.6497768   .0310318   -20.94   0.000    -.710775   -.5887786
       _cons |   686.0322   8.728224    78.60   0.000    668.8754     703.189
------------------------------------------------------------------------------
```

$$\widehat{TestScore} = 686{,}0 - 1{,}10 \times STR - 0{,}650 PctEL$$
$$(8{,}7) \quad (0{,}43) \qquad (0{,}031)$$

Utilizziamo gli errori standard robusti all'eteroschedasticità – esattamente per lo stesso motivo del caso di un singolo regressore.

Verifica di ipotesi congiunte (Paragrafo 7.2)

Sia *Expn* = spese per studente e si consideri il modello di regressione:

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

L'ipotesi nulla per cui "le risorse scolastiche non contano", e l'alternativa per cui invece contano, corrisponde a:

$$H_0: \beta_1 = 0 \textbf{ e } \beta_2 = 0$$

$$\text{vs. } H_1: \textbf{o } \beta_1 \neq 0 \textbf{ o } \beta_2 \neq 0 \textbf{ o entrambi}$$

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

Verifica di ipotesi congiunte (continua)

- H_0: $\beta_1 = 0$ ***e*** $\beta_2 = 0$
- vs. H_1: ***o*** $\beta_1 \neq 0$ ***o*** $\beta_2 \neq 0$ ***o entrambe***
- Un'***ipotesi congiunta*** specifica un valore per due o più coefficienti, ossia impone una restrizione su due o più coefficienti.
- In generale, un'ipotesi congiunta implicherà q restrizioni. Nell'esempio precedente, $q = 2$ e le due restrizioni sono $\beta_1 = 0$ e $\beta_2 = 0$.
- Un'idea di "buon senso" è quella di rifiutare se l'una o l'altra delle statistiche-t supera 1,96 in valore assoluto.
- ma questa verifica "coefficiente per coefficiente" non è valida: la verifica risultante ha un tasso di rifiuto troppo elevato sotto l'ipotesi nulla (più del 5%)!

Perché non possiamo verificare coefficiente per coefficiente?

Perché il tasso di rifiuto sotto l'ipotesi nulla non è il 5%. Calcoleremo la probabilità di rifiutare in modo non corretto l'ipotesi nulla usando la verifica del "buon senso" basata sulle due statistiche-*t* singole. Per semplificare il calcolo, supponete che siano distribuite in modo indipendente (non è vero in generale – lo è solo in questo esempio). Siano t_1 e t_2 le statistiche-*t*:

$$t_1 = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \quad \text{e } t_2 = \frac{\hat{\beta}_2 - 0}{SE(\hat{\beta}_2)}$$

La verifica "coeff. per coeff." è:

rifiuta H_0: $\beta_1 = \beta_2 = 0$, se $|t_1| > 1{,}96$ e/o $|t_2| > 1{,}96$

Qual è la probabilità che questa verifica "coeff. per coeff." rifiuti H_0, quando H_0 è effettivamente vero? (*Dovrebbe* essere 5%.)

***Supponete che t_1 e t_2 siano indipendenti* (per questo esempio).**

La probabilità di rifiutare in modo non corretto l'ipotesi nulla mediante la verifica "coeff. per coeff."

$= \Pr_{H_0}[|t_1| > 1{,}96 \text{ e/o } |t_2| > 1{,}96]$

$= 1 - \Pr_{H_0}[|t_1| \leq 1{,}96 \text{ e } |t_2| \leq 1{,}96]$

$= 1 - \Pr_{H_0}[|t_1| \leq 1{,}96] \times \Pr_{H_0}[|t_2| \leq 1{,}96]$

(poiché t_1 e t_2 sono indipendenti per assunzione)

$= 1 - (0{,}95)^2$

$= 0{,}0975 = 9{,}75\%$ – che ***non*** è il 5% desiderato!!

La *dimensione* di una verifica è l'effettivo tasso di rifiuto sotto l'ipotesi nulla.

- La dimensione della verifica del "buon senso" non è 5%!
- In effetti, la sua dimensione dipende dalla correlazione tra t_1 e t_2 (e quindi dalla correlazione tra $\hat{\beta}_1$ e $\hat{\beta}_2$).

Due soluzioni:

- Utilizzare un valore critico diverso in questa procedura – non 1,96 (questo è il "metodo Bonferroni" – vedi Appendice 7.1) (in ogni caso, questo metodo è utilizzato raramente nella pratica)
- Utilizzare una statistica di test diversa studiata per verificare subito *sia* β_1 sia β_2: la statistica *F* (questa è pratica comune)

La statistica *F*

La statistica *F* verifica tutte le parti di un'ipotesi congiunta in un colpo solo.

Formula per il caso speciale dell'ipotesi congiunta $\beta_1 = \beta_{1,0}$ e $\beta_2 = \beta_{2,0}$ in una regressione con due regressori:

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right)$$

dove $\hat{\rho}_{t_1,t_2}$ stima la correlazione tra t_1 e t_2.

Rifiuta quando *F* è grande (quanto grande?)

La verifica della statistica *F* β_1 e β_2:

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right)$$

- La statistica *F* è grande quando t_1 e/o t_2 è grande
- La statistica *F* corregge (nel modo giusto) per la correlazione tra t_1 e t_2.
- La formula per più di due β è brutta a vedersi, a meno che non si utilizzi l'algebra matriciale.
- Ciò fornisce alla statistica *F* una buona distribuzione approssimata in grandi campioni, ossia…

Distribuzione in grandi campioni della statistica *F*

Si consideri il *caso speciale* che t_1 e t_2 siano indipendenti, perciò $\hat{\rho}_{t_1,t_2} \xrightarrow{p} 0$; in grandi campioni la formula diventa

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right) \cong \frac{1}{2}(t_1^2 + t_2^2)$$

- Sotto l'ipotesi nulla, t_1 e t_2 hanno distribuzioni normali standard che, in questo caso speciale, sono indipendenti
- La distribuzione in grandi campioni della statistica *F* è la distribuzione della media dei quadrati di due variabili casuali standard distribuite in modo indipendente.

La distribuzione chi-quadrato

La distribuzione ***chi-quadrato*** con q gradi di libertà (χ^2_q) è definita come distribuzione della somma dei quadrati di q variabili casuali normali standard indipendenti.

In grandi campioni, *F* è distribuita come χ^2_q /*q* =F(q,infinito)

Valori critici in grandi campioni selezionati di χ^2_q/*q*

q	5% del valore critico	
1	3,84	(*perché*?)
2	3,00	(il caso q=2 precedente)
3	2,60	
4	2,37	
5	2,21	

Calcolo del valore-p mediante la statistica F:

valore-*p* = probabilità nella coda destra della distribuzione χ^2_q/q oltre la statistica *F* effettivamente calcolata (ossia F_{act} è l'estremo inferiore della coda)

Implementazione in STATA

Utilizzare il comando "test" dopo la regressione

Esempio: Verificare l'ipotesi congiunta che i coefficienti di *STR* e delle spese per studente (*expn_stu*) siano entrambi zero, a fronte dell'alternativa che almeno uno dei sia diverso da zero.

Esempio di verifica F, dati sulle dimensioni delle classi della California:

```
reg testscr str expn_stu pctel, r;
Regression with robust standard errors                Number of obs =     420
                                                      F(  3,    416) =  147,20
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0,4366
                                                      Root MSE      =  14.353
------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -.2863992   .4820728    -0.59   0.553    -1.234001     .661203
    expn_stu |   .0038679   .0015807     2.45   0.015     .0007607    .0069751
       pctel |  -.6560227   .0317844   -20.64   0.000    -.7185008   -.5935446
       _cons |   649.5779   15.45834    42.02   0.000     619.1917    679.9641
------------------------------------------------------------------------------
```

NOTA

```
test str expn_stu;
```

Il comando "test" segue la regressione

```
 ( 1)  str = 0,0
 ( 2)  expn_stu = 0.0
       F(  2,    416) =    5,43
       Prob > F =    0,0047
```

Vi sono q=2 restrizioni in verifica

Il 5% del valore critico per q=2 è 3,00

Stata calcola per voi il valore-p

Ulteriori informazioni sulla statistica *F*.

Esiste una formula semplice per la statistica F, valida solo in condizinoi di omoschedasticità (perciò non molto utile), che tuttavia può aiutare a comprendere che cosa fa la statistica F.

La statistica *F* in condizioni di omoschedasticità pura

Quando gli errori sono omoschedastici, esiste una formula semplice per il calcolo della statistica *F* in presenza di "omoschedasticità pura":

•Eseguire due regressioni, una sotto l'ipotesi nulla (regressione "vincolata") e una sotto l'ipotesi alternativa (regressione "senza vincolo").

•Confrontare gli adattamenti delle regressioni – gli R^2 – se il modello "non vincolato" si adatta sufficientemente meglio, rifiutare l'ipotesi nulla

Regressione "vincolata" e "non vincolata"

Esempio: i coefficienti di *STR* e *Expn* sono zero?

Regressione senza vincolo (sotto H_1):

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

Regressione vincolata (ossia, sotto H_0):

$$TestScore_i = \beta_0 + \beta_3 PctEL_i + u_i$$ (*perché*?)

- Il numero di vincoli sotto H_0 è $q = 2$ (*perché*?).
- L'adattamento risulterà migliore (R^2 sarà maggiore) nella regressione non vincolata (*perché*?)

Di quanto dovrà aumentare R^2 affinché i coefficienti di *Expn* e *PctEL* siano giudicati statisticamente significativi?

Formula semplice per la statistica F classica:

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted}) / q}{(1 - R^2_{unrestricted}) / (n - k_{unrestricted} - 1)}$$

dove:

$R^2_{restricted}$ = R^2 per la regressione vincolata

$R^2_{unrestricted}$ = R^2 per la regressione non vincolata

q = numero di restrizioni sotto l'ipotesi nulla

$k_{unrestricted}$ = numero di regressori nella regressione non vincolata.

- Più grande è la differenza tra l'R^2 vincolato e non vincolato, maggiore è il miglioramento dell'adattamento aggiungendo le variabili in questione – maggiore è la F in presenza di omoschedasticità pura.

***Esempio*:**

Regressione vincolata:

$\widehat{TestScore}$ = 644,7 –0,671*PctEL*, $R^2_{restricted}$ = 0,4149

(1,0) (0,032)

Regressione non vincolata:

$\widehat{TestScore}$ = 649,6 – 0,29*STR* + 3,87*Expn* – 0,656*PctEL*

(15,5) (0,48) (1,59) (0,032)

$R^2_{unrestricted}$ = 0,4366, $k_{unrestricted}$ = 3, q = 2

Quindi $$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)}$$

$$= \frac{(0,4366 - 0,4149)/2}{(1 - 0,4366)/(420 - 3 - 1)} = \mathbf{8,01}$$

Nota: *F robusta all'eteroschedasticità* = **5,43**…

La statistica F classica – riepilogo

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted}) / q}{(1 - R^2_{unrestricted}) / (n - k_{unrestricted} - 1)}$$

- La statistica *F* classica rifiuta quando aggiungendo le due variabili si aumenta R^2 di "quanto basta" – vale a dire, quando aggiungendo le due variabili si migliora l'adattamento della regressione di "quanto basta"
- Se gli errori sono omoschedastici, la statistica *F* classica ha una distribuzione in grandi campioni che è χ^2_q/q.
- Se invece gli errori sono eteroschedastici, la distribuzione in grandi campioni della statistica *F* classica non è χ^2_q/q

La distribuzione *F*

A volte in riferimento alla regressione si parla di distribuzione "*F*".

Se le quattro assunzioni dei minimi quadrati per la regressione multipla valgono ***e se***:

5. u_i è omoschedastico, ossia var($u|X_1,\ldots,X_k$) non dipende dalle X
6. $u_1,\ldots,u_n$ sono normalmente distribuiti

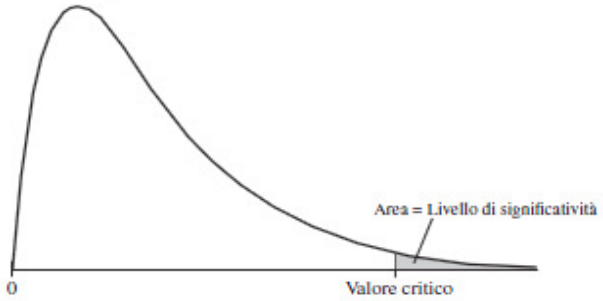
allora la statistica *F* classica ha la distribuzione "$F_{q,n-k-1}$", dove q = numero delle restrizioni e k = numero dei regressori sotto l'alternativa (modello non vincolato).

- **La distribuzione *F* è per la distribuzione χ^2_q /*q* ciò che la distribuzione t_{n-1} è per la distribuzione *N*(0,1)**

La distribuzione $F_{q,n-k-1}$:

- La distribuzione *F* è tabulata in molti punti
- Per $n \to \infty$, la statistica $F_{q,n-k-1}$ tende asintoticamente alla distribuzione χ^2_q/q:
- **Le distribuzioni $F_{q,\infty}$ e χ^2_q/q sono identiche.**
- Per *q* non troppo grande e $n \geq 100$, la distribuzione $F_{q,n-k-1}$ e la distribuzione χ^2_q/q sono sostanzialmente identiche.
- Molti pacchetti di regressione (tra cui STATA) calcolano il valore-*p* della statistica *F* mediante la distribuzione *F*
- Incontrerete la distribuzione *F* in lavori pubblicati di carattere empirico.

Tavola 4 Valori critici della distribuzione $F_{m,\infty}$



	Livello di significatività		
Gradi di libertà	**10%**	**5%**	**1%**
1	2,71	3,84	6,63
2	2,30	3,00	4,61
3	2,08	2,60	3,78
4	1,94	2,37	3,32
5	1,85	2,21	3,02
6	1,77	2,10	2,80
7	1,72	2,01	2,64
8	1,67	1,94	2,51
9	1,63	1,88	2,41
10	1,60	1,83	2,32
11	1,57	1,79	2,25
12	1,55	1,75	2,18
13	1,52	1,72	2,13
14	1,50	1,69	2,08
15	1,49	1,67	2,04
16	1,47	1,64	2,00
17	1,46	1,62	1,97
18	1,44	1,60	1,93
19	1,43	1,59	1,90
20	1,42	1,57	1,88
21	1,41	1,56	1,85
22	1,40	1,54	1,83
23	1,39	1,53	1,81
24	1,38	1,52	1,79
25	1,38	1,51	1,77
26	1,37	1,50	1,76
27	1,36	1,49	1,74
28	1,35	1,48	1,72
29	1,35	1,47	1,71
30	1,34	1,46	1,70

Questa tavola contiene il 90-esimo, 95-esimo e 99-esimo percentile della distribuzione $F_{m,\infty}$. Questi rappresentano i valori critici per test con livello di significatività del 10%, 5% e 1%.

Un'altra digressione: breve storia della statistica...

- La teoria della statistica F classica in presenza di omoschedasticità pura e le distribuzioni $F_{q,n-k-1}$ si poggiano su assunzioni troppo forti per essere plausibili (i guadagni hanno distribuzione normale?)
- Queste statistiche risalgono agli albori del XX secolo… quando le serie di dati erano piccole e i calcolatori erano persone…
- La statistica F e la distribuzione $F_{q,n-k-1}$ erano innovazioni importanti: una formula facile da calcolare, un unico insieme di tabelle che poteva essere pubblicato una volta, quindi applicato in molti casi, e una giustificazione precisa e matematicamente elegante.

Breve storia della statistica (continua)

- Le assunzioni forti erano un prezzo minimo da pagare per questa innovazione.
- Ma con i moderni computer e i grandi campioni possiamo utilizzare la statistica F robusta all'eteroschedasticità e la distribuzione $F_{q,\infty}$, che richiede soltanto le quattro assunzioni dei minimi quadrati (e non le assunzioni n. 5 e n. 6)
- Questa eredità storica persiste nel software moderno, in cui lo standard dell'omoschedasticità pura (e la statistica F) sono il default, e in cui i valori-p vengono calcolati mediante la distribuzione $F_{q,n-k-1}$.

Riepilogo: la statistica *F* classica e la distribuzione *F*

- Sono giustificate solo sotto condizioni molto forti – troppo forti per essere realistiche.
- Dovreste utilizzare la statistica *F* robusta all'eteroschedasticità robusta, con χ^2_q/q valori critici (ossia $F_{q,\infty}$).
- Per $n \geq 100$, la distribuzione *F* è essenzialmente la distribuzione χ^2_q/q.
- Per *n* piccolo, a volte i ricercatori utilizzano la distribuzione *F* perché ha valori critici più grandi e in tal senso è più prudente.

Riepilogo: verifica di ipotesi congiunte

- L'approccio "coefficiente per coefficiente" che prevede il rifiuto se l'una o l'altra statistica *t* supera 1,96 rifiuta più del 5% delle volte sotto l'ipotesi nulla (la dimensione supera il livello di significatività desiderato)
- La statistica *F* robusta all'eteroschedasticità è integrata in STATA (comando "test"); questa verifica tutte le restrizioni *q* allo stesso tempo.
- Per *n* grande, la statistica *F* ha distribuzione χ^2_q/q (= $F_{q,\infty}$)
- La statistica *F* classica è storicamente importante (e così anche nella pratica) e può aiutare l'intuizione, ma non è valida in presenza di eteroschedasticità

Verifica di restrizioni singole su coefficienti multipli (Paragrafo 7.3)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1,\ldots,n$$

Considerate l'ipotesi nulla e le ipotesi alternative,

$$H_0: \beta_1 = \beta_2 \quad \text{vs.} \quad H_1: \beta_1 \neq \beta_2$$

Questa ipotesi nulla impone una *singola* restrizione (q = 1) su coefficienti *multipli* – non si tratta di ipotesi congiunte con restrizioni multiple (confrontate con β_1 = 0 e β_2 = 0).

Verifica di restrizioni singole su coefficienti multipli (continua)

Ecco due metodi per la verifica di restrizioni singole su coefficienti multipli:

1.Riorganizzare ("trasformare") la regressione
Riorganizzare i regressori in modo che la restrizione diventi una restrizione su un singolo coefficiente in una regressione equivalente; oppure,

2.Eseguire la verifica direttamente
Alcuni software, tra cui STATA, consentono di verificare le restrizioni utilizzando direttamente coefficienti multipli

Metodo 1: Riorganizzare ("trasformare") la regressione

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$
$$H_0: \beta_1 = \beta_2 \quad \text{vs.} \quad H_1: \beta_1 \neq \beta_2$$

Sommare e sottrarre $\beta_2 X_{1i}$:

$$Y_i = \beta_0 + (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) + u_i$$

oppure

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i$$

Dove

$$\gamma_1 = \beta_1 - \beta_2$$
$$W_i = X_{1i} + X_{2i}$$

Riorganizzare la regressione (continua)

(a) Equazione originale:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

$$H_0: \beta_1 = \beta_2 \quad \text{vs.} \quad H_1: \beta_1 \neq \beta_2$$

(b) Equazione riorganizzata ("trasformata"):

$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i$

dove $\gamma_1 = \beta_1 - \beta_2$ e $W_i = X_{1i} + X_{2i}$

Quindi

$H_0: \gamma_1 = 0$ vs. $H_1: \gamma_1 \neq 0$

- Queste due regressioni ((a) e (b)) hanno lo stesso R^2, gli stessi valori previsti e gli stessi residui.
- Il problema di verifica è ora semplice: verificare se $\gamma_1 = 0$ nella regressione (b).

Metodo 2: Eseguire la verifica direttamente

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

H_0: $\beta_1 = \beta_2$ vs. H_1: $\beta_1 \neq \beta_2$

Esempio:

$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$

In STATA, per verificare $\beta_1 = \beta_2$ vs. $\beta_1 \neq \beta_2$ (bilaterale):

```
regress testscore str expn pctel, r
test str=expn
```

I dettagli dell'implementazione di questo modello sono specifici del software.

Regioni di confidenza per coefficienti multipli (Paragrafo 7.4)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1,\ldots,n$$

Qual è una regione di confidenza per β_1 e β_2?

Una ***regione di confidenza di livello*** 95% è:

- Una "funzione a più valori" dei dati che contiene il o i coefficienti reali nel 95% dei campioni ripetuti ipotetici.
- In modo equivalente, la regione dei valori dei coefficienti che non può essere rifiutata al livello di significatività del 5%.

Si può trovare una regione di confidenza del 95% come regione di (β_1, β_2) che non può essere rifiutata al livello del 5% mediante una verifica-*F* (*perché non combinare semplicemente i due intervalli di confidenza al 95%?*).

Regioni di confidenza (continua)

- Sia $F(\beta_{1,0}, \beta_{2,0})$ la verifica della statistica *F* (robusta all'eteroschedasticità) che verifica l'ipotesi che $\beta_1 = \beta_{1,0}$ e $\beta_2 = \beta_{2,0}$:

- Regione di confidenza al 95% = $\{\beta_{1,0}, \beta_{2,0}: F(\beta_{1,0}, \beta_{2,0}) < 3{,}00\}$
- 3,00 è il valore critico al 5% della distribuzione $F_{2,\infty}$
- Questa regione ha tasso di copertura del 95% perché la verifica su cui è basata (la verifica che "inverte") ha dimensione del 5%
- *Nel 5% dei casi la verifica rifiuta in modo non corretto l'ipotesi nulla quando questa è vera, quindi non lo fa il 95% dei casi; pertanto, la regione di confidenza costruita come valori non rifiutati contiene il valore vero per il 95% delle volte (nel 95% di tutti i campioni).*

La regione di confidenza basata sulla statistica F è un'ellisse:

$$\{\beta_1, \beta_2: \; F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}^2_{t_1,t_2}}\right) \leq 3{,}00\}$$

Ora

$$F = \frac{1}{2(1-\hat{\rho}^2_{t_1,t_2})} \times \left[t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2\right]$$

$$= \frac{1}{2(1-\hat{\rho}^2_{t_1,t_2})} \times$$

$$\left[\left(\frac{\hat{\beta}_2 - \beta_{2,0}}{SE(\hat{\beta}_2)}\right)^2 + \left(\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right)^2 + 2\hat{\rho}_{t_1,t_2}\left(\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right)\left(\frac{\hat{\beta}_2 - \beta_{2,0}}{SE(\hat{\beta}_2)}\right)\right]$$

Questa è una forma quadratica in $\beta_{1,0}$ e $\beta_{2,0}$ – così il confine della regione $F = 3{,}00$ è un'ellisse.

Regione di confidenza basata sull'inversione della statistica F

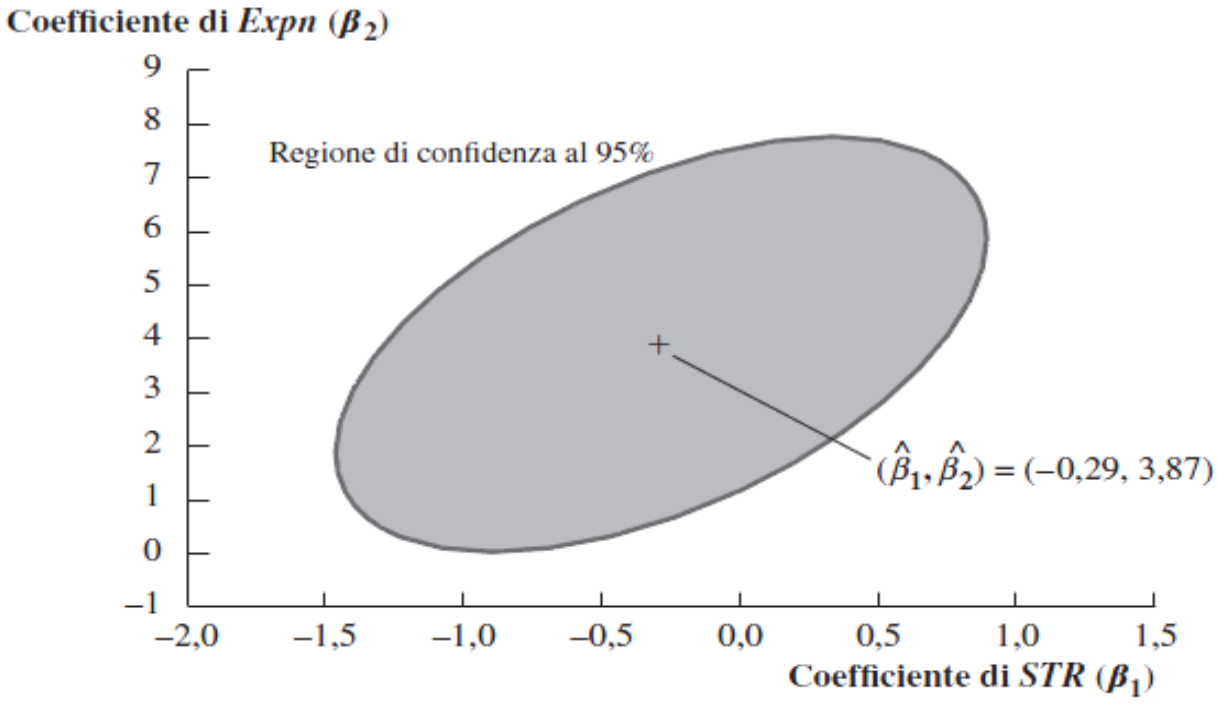


Figura 7.1

Regione di confidenza al 95% per i coefficienti di *STR* ed *Expn* dell'Equazione (7.6).

La regione di confidenza al 95% per i coefficienti di *STR* (β_1) e di *Expn* (β_2) è un'ellisse che contiene le coppie di valori di β_1 e β_2 che non possono essere rifiutati usando la statistica *F* al livello di significatività del 5%.

Specificazione della regressione: variabili di interesse, variabili di controllo e indipendenza in media condizionata (Paragrafo 7.5)

Vogliamo ottenere una stima non distorta dell'effetto sui punteggi nei test della modifica della dimensione della classe, tenendo costanti i fattori al di fuori del controllo del consiglio scolastico – quali opportunità di apprendimento esterne (musei e così via), coinvolgimento dei genitori nell'istruzione (letture a casa con la madre?) e così via.

Se potessimo eseguire un esperimento, assegneremmo casualmente studenti (e insegnanti) a classi di dimensione diversa. Allora STR_i sarebbe indipendente da tutti i fattori che rientrano in u_i, perciò $E(u_i|STR_i) = 0$ e lo stimatore OLS della pendenza nella regressione di $TestScore_i$ su STR_i sarebbe uno stimatore non distorto dell'effetto casuale desiderato.

Con dati non sperimentali, tuttavia, u_i dipende da fattori supplementari (musei, coinvolgimento dei genitori, conoscenza dell'inglese e così via).

•Se potete osservare questi fattori (per esempio *PctEL*), includeteli nella regressione.

•Ma solitamente non siete in grado di osservare tutti questi fattori omessi (per esempio il coinvolgimento dei genitori nei compiti a casa).
In questo caso potete includere "variabili di controllo" correlate a questi fattori causali omessi, ma che di per sé non sono causali.

Variabili di controllo nella regressione multipla

Una **variabile di controllo *W*** è una variabile correlata e che controlla per un fattore causale omesso nella regressione di *Y* su *X*, ma che di per sé non ha un effetto causale su *Y*.

Variabili di controllo: un esempio dai dati dei punteggi nei test della California

$$\widehat{TestScore} = 700{,}2 - 1{,}00STR - 0{,}122PctEL - 0{,}547LchPct,\ \overline{R}^2 = 0{,}773$$
$$(5{,}6)\quad (0{,}27)\quad (0{,}033)\quad (0{,}024)$$

PctEL = percentuale di studenti non di madrelingua nel distretto

LchPct = percentuali di studenti che ricevono un pasto gratuito/sovvenzionato (ne hanno diritto solo gli studenti di famiglie con reddito basso)

- Quale variabile è la variabile di interesse?
- Quali variabili sono variabili di controllo? Ci sono componenti causali? Che cosa controllano?

Esempio di variabili di controllo (continua)

$$\widehat{TestScore} = 700{,}2 - 1{,}00STR - 0{,}122PctEL - 0{,}547LchPct, \quad \overline{R}^2 = 0{,}773$$
$$(5{,}6) \quad (0{,}27) \quad (0{,}033) \quad (0{,}024)$$

- *STR* è la variabile di interesse
- *PctEL* probabilmente ha un effetto causale diretto (la scuola è più difficile per chi non è di madrelingua!). Ma è anche una variabile di controllo: le comunità di immigranti tendono a essere meno benestanti e spesso hanno minori opportunità di apprendimento esterno e *PctEL* è correlata a tali variabili causali omesse. *PctEL è sia una variabile causale sia una variabile di controllo.*
- *LchPct* potrebbe avere un effetto causale (consumare il pasto aiuta l'apprendimento); è inoltre correlata e controlla per le opportunità di apprendimento esterne legate al reddito.
  PctEL è sia una possibile variabile causale sia una variabile di controllo.

Variabili di controllo (continua)

1. Tre affermazioni intercambiabili sui fattori che determinano l'efficacia di una variabile di controllo:

I. Una variabile di controllo efficace è una che, se inclusa nella regressione, rende la condizione di errore non correlata alla variabile di interesse.

II. Tenendo costante la o le variabili di controllo, la variabile di interesse viene assegnata casualmente "così com'è".

III. Tra gli individui (unità) con lo stesso valore della o delle variabili di controllo, la variabile di interesse è non correlata ai determinanti omessi di *Y*

Variabili di controllo (continua)

2. **Le variabili di controllo non devono essere causali e i loro coefficienti in generale non hanno un'interpretazione causale.** Per esempio:

$$\widehat{TestScore} = \underset{(5,6)}{700,2} - \underset{(0,27)}{1,00}STR - \underset{(0,033)}{0,122}PctEL - \underset{(0,024)}{0,547}LchPct, \quad \overline{R}^2 = 0,773$$

- Il coefficiente di *LchPct* ha un'interpretazione causale? In questo caso, dovremmo essere in grado di ampliare i punteggi nei test (e di parecchio anche!) eliminando semplicemente il programma della mensa scolastica, in modo che *LchPct* = 0! (L'eliminazione del programma di mensa scolastica ha un effetto causale ben definito: possiamo realizzare un esperimento randomizzato per misurare l'effetto causale di questo intervento).

La matematica delle variabili di controllo: indipendenza in media condizionata

- Poiché il coefficiente di una variabile di controllo può essere distorto, la prima assunzione dei minimi quadrati ($E(u_i|X_{1i},\ldots,X_{ki}) = 0$) non deve valere. Per esempio, il coefficiente su *LchPct* è correlato a determinanti non misurati dei punteggi nei test, quali le opportunità di apprendimento esterne, perciò è soggetta a distorsione da variabili omesse. Ma il fatto che *LchPct* sia correlata a queste variabili omesse è precisamente ciò che la rende una buona variabile di controllo!

- Se la prima assunzione dei minimi quadrati non vale, allora che cosa vale?

- Ci occorre una dichiarazione matematica di ciò che renda efficace una variabile di controllo. È l'**indipendenza in media condizionata**: data la variabile di controllo, la media di u_i non dipende dalla variabile di interesse

Indipendenza in media condizionata (continua)

Sia X_i la variabile di interesse e sia W_i la o le variabili di controllo. W è una variabile di controllo efficace se vale l'indipendenza in media condizionata:

$$E(u_i|X_i, W_i) = E(u_i|W_i) \quad \text{(indipendenza in media condizionata)}$$

Se W è una variabile di controllo, allora l'indipendenza in media condizionata sostituisce la prima assunzione dei minimi quadrati – in pratica è la versione di tale assunzione che è rilevante per le variabili di controllo.

Indipendenza in media condizionata (continua)

Considerate il modello di regressione,

$Y = \beta_0 + \beta_1 X + \beta_2 W + u$

dove X è la variabile di interesse e W è una variabile di controllo efficace, cosicché vale l'indipendenza in media condizionata:

$E(u_i|X_i, W_i) = E(u_i|W_i)$.

Inoltre, supponete che le assunzioni dei minimi quadrati n. 2, n. 3 e n. 4 valgano. Quindi:

1. β_1 ha un'interpretazione causale.
2. $\hat{\beta}_1$ è non distorto
3. Il coefficiente della variabile di controllo, $\hat{\beta}_2$, è in generale distorto.

La matematica dell'indipendenza in media condizionata

Sotto l'indipendenza in media condizionata:

1. β_1 ha un'interpretazione causale.

Matematica: la variazione prevista in Y risultante da una variazione in X, mantenendo (una singola) costante W, è:

$$E(Y|X = x+\Delta x, W=w) - E(Y|X = x, W=w)$$
$$= [\beta_0 + \beta_1(x+\Delta x) + \beta_2 w + E(u|X = x+\Delta x, W=w)]$$
$$- [\beta_0 + \beta_1 x + \beta_2 w + E(u|X = x, W=w)]$$
$$= \beta_1 \Delta x + [E(u|X = x+\Delta x, W=w) - E(u|X = x, W=w)]$$
$$= \beta_1 \Delta x$$

dove la riga finale segue dall'indipendenza in media condizionata:

$E(u|X = x+\Delta x, W=w) = E(u|X = x, W=w) = E(u|W=w)$.

La matematica dell'indipendenza in media condizionata (continua)

Sotto l'indipendenza in media condizionata:

2. $\hat{\beta}_1$ è non distorto

3. $\hat{\beta}_2$ è in generale distorto

Matematica: considerate il modello di regressione

$$Y = \beta_0 + \beta_1 X + \beta_2 W + u$$

dove u soddisfa l'assunzione dell'indipendenza in media condizionata. Per comodità, supponete che $E(u|W) = \gamma_0 + \gamma_2 W$ (ossia, che $E(u|W)$ sia lineare in W). Allora, sotto l'indipendenza in media condizionata,

La matematica dell'indipendenza in media condizionata (continua)

$$E(u|X, W) = E(u|W) = \gamma_0 + \gamma_2 W. \qquad (*)$$

Sia

$$v = u - E(u|X, W) \qquad (**)$$

cosicché $E(v|X, W) = 0$. Combinando (*) e (**) si ricava,

$$u = E(u|X, W) + v$$
$$= \gamma_0 + \gamma_2 W + v, \text{ dove } E(v|X, W) = 0 \qquad (***)$$

Ora sostituite (***) nella regressione,

$$Y = \beta_0 + \beta_1 X + \beta_2 W + \qquad (+)$$

Cosicché

$$Y = \beta_0 + \beta_1 X + \beta_2 W + u \qquad (+)$$

$$= \beta_0 + \beta_1 X + \beta_2 W + \gamma_0 + \gamma_2 W + v \quad \text{da (***)}$$

$$= (\beta_0 + \gamma_0) + \beta_1 X + (\beta_2 + \gamma_2) W + v$$

$$= \delta_0 + \beta_1 X + \delta_2 W + v \qquad (++)$$

- Poiché $E(v|X, W) = 0$, l'equazione (++) soddisfa la prima assunzione dei minimi quadrati, perciò gli stimatori OLS di δ_0, β_1 e δ_2 in (++) sono non distorti.
- Poiché i regressori in (+) e (++) sono gli stessi, i coefficienti OLS nella regressione (+) soddisfano $E(\hat{\beta}_1) = \beta_1$ e $E(\hat{\beta}_2) = \delta_2 = \beta_2 + \gamma_2 \neq \beta_2$ in generale.

$E(\hat{\beta}_1) = \beta_1$

e

$E(\hat{\beta}_2) = \delta_2 = \beta_2 + \gamma_2 \neq \beta_2$

Riepilogando, se *W* è tale per cui l'indipendenza in media condizionale è soddisfatta, allora:

- Lo stimatore OLS dell'effetto di interesse, $\hat{\beta}_1$, è non distorto.

- Lo stimatore OLS del coefficiente della variabile di controllo, $\hat{\beta}_2$,è distorto. Questa distorsione nasce dal fatto che la variabile di controllo è correlata alle variabili omesse nella condizione di errore, cosicché $\hat{\beta}_2$ è soggetto a distorsione da variabili omesse.

Implicazioni per la selezione delle variabili e "*specificazione del modello*"

1. Identificate la variabile di interesse
2. Pensate agli effetti causali omessi che potrebbero risultare in distorsione delle variabili omesse
3. Se potete, includete tali effetti causali omessi o, in caso contrario, includete le variabili correlate a essi per fungere da variabili di controllo. Le variabili di controllo sono efficaci se l'assunzione dell'indipendenza in media condizionata vale in modo plausibile (se *u* è non correlata a *STR* una volta incluse le variabili di controllo). Ciò risulta in un modello "base" o "benchmark".

Specificazione del modello (continua)

4. Specificate anche una gamma di modelli alternativi plausibili, che includano variabili candidate aggiuntive.

4. Stimate il modello base e le specificazioni alternative plausibili ("controlli di sensibilità").
   - Una variabile candidata cambia il coefficiente di interesse (β_1)?
   - Una variabile candidata è statisticamente significativa?
   - Usate il giudizio e non una ricetta meccanica…
   - Non cercate semplicemente di massimizzare R^2!

Digressione sulle misure di un adattamento…

È facile cadere nella trappola di massimizzare R^2 e $\overline{R}^2$, ma ciò riduce la visibilità sull'obiettivo reale, uno stimatore non distorto dell'effetto della dimensione della classe.

- Un elevato R^2 (o $\overline{R}^2$) significa che i regressori spiegano la variazione in Y.
- Un elevato R^2 (o $\overline{R}^2$) *non* significa che avete eliminato la distorsione delle variabili omesse.
- Un elevato R^2 (o $\overline{R}^2$) *non* significa che avete uno stimatore non distorto di effetto causale (β_1).
- Un elevato R^2 (o $\overline{R}^2$) *non* significa che le variabili incluse siano statisticamente significative – ciò deve essere determinato mediante le verifiche di ipotesi.

Analisi del set di dati sul punteggio nei test (Paragrafo 7.6)

1. Identificate la variabile di interesse:
   STR
2. Pensate agli effetti causali omessi che potrebbero risultare in distorsione da variabili omesse

La lingua madre degli studenti, le opportunità do apprendimento esterne, il coinvolgimento dei genitori, la qualità degli insegnanti (se lo stipendio degli insegnanti è correlato al benessere del distretto) – la lista è lunga!

3. Se potete, includete tali effetti causali omessi o, in caso contrario, includete le variabili correlate a essi per fungere da variabili di controllo. Le variabili di controllo sono efficaci se l'assunzione dell'indipendenza in media condizionata vale in modo plausibile (se *u* è non correlata a *STR* una volta incluse le variabili di controllo). Ciò risulta in un modello "base" o "benchmark".

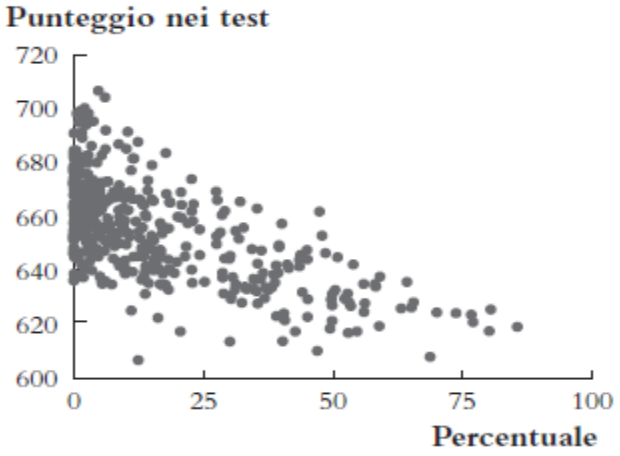
Molte delle variabili causali omesse sono difficili da misurare, perciò dobbiamo trovare le variabili di controllo. Queste includono PctEL (sia una variabile di controllo sia un fattore causale omesso) e misure del benessere del distretto.

4. Specificate anche una gamma di modelli alternativi plausibili, che includano variabili candidate aggiuntive.

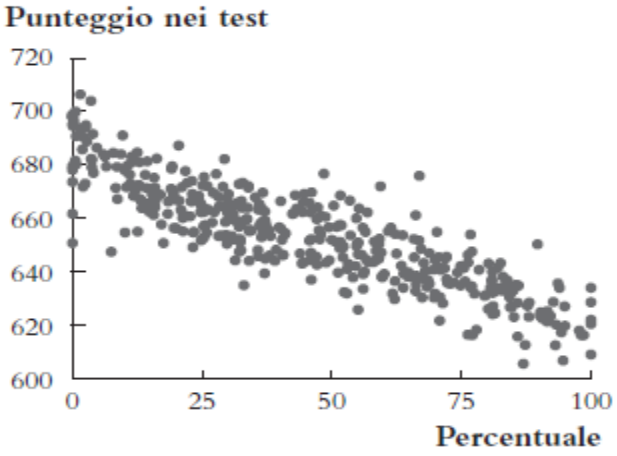
Non è chiara quale delle variabili relative al reddito controlli al meglio i molteplici fattori causali omessi, quali le opportunità di apprendimento esterno, perciò le specificazioni delle alternative comprendono regressioni con variabili di reddito diverse. Le specificazioni delle alternative considerate qui sono solo un punto di partenza e non la parola finale!

5. Stimate il modello base e le specificazioni alternative plausibili ("controlli di sensibilità").

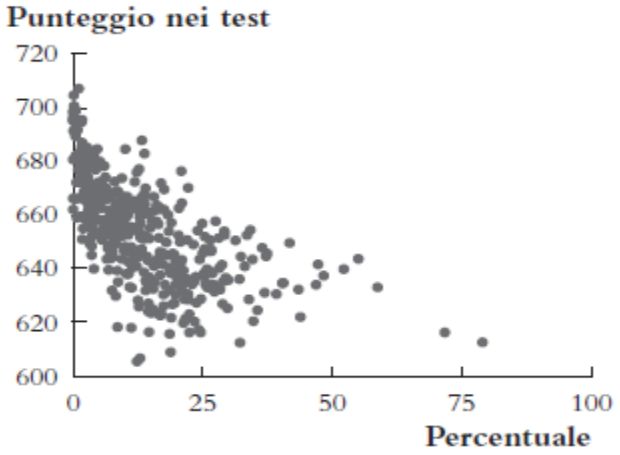
Punteggi nei test e dati socioeconomici della California…



(a) Percentuale di studenti non di madrelingua



(b) Percentuale di aventi diritto al sussidio mensa



(c) Percentuale di aventi diritto a un sostegno del reddito

Figura 7.2

Grafici a nuvola del punteggio nei test su tre caratteristiche dello studente.

I grafici a nuvola mostrano una relazione negativa tra punteggio nei test e (*a*) percentuale di studenti che apprendono l'inglese (correlazione = – 0,64), (*b*) percentuale di studenti aventi diritto al sussidio mensa (correlazione = – 0,87) e (*c*) percentuale di aventi diritto a un sostegno del reddito (correlazione = – 0,63).

Digressione sulla presentazione dei risultati della regressione

- Abbiamo numerose regressioni e desideriamo presentarle. È scomodo e difficile leggere regressioni scritte in forma di equazione, perciò tradizionalmente si riportano in formato tabulare.
- I risultati di una tabella di regressione comprendono:
  - coefficienti di regressione stimati
  - errori standard
  - misure di adattamento
  - numero di osservazioni
  - statistica *F* rilevante, se esistente
  - Qualsiasi altra informazione pertinente.
- Trovate queste informazioni nella tabella seguente:

Tabella 7.1 Risultati delle regressioni del punteggio nei test usando i dati relativi ai distretti scolastici elementari della California.

Variabile dipendente: media del punteggio nei test nel distretto.

Regressore	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
Rapporto studenti/insegnanti (X_1)	–2,28** (–0,52)	–1,10* (0,43)	–1,00** (0,27)	–1,31** (0,34)	–1,01** (0,27)
% studenti non di madrelingua (X_2)		–0,650** (0,031)	–0,122** (0,033)	–0,488** (0,030)	–0,130** (0,036)
% aventi diritto al sussidio mensa (X_3)			–0,547** (0,024)		–0,529** (0,038)
% studenti nel programma di assistenza pubblica (X_4)				–0,790** (0,068)	0,048 (0,059)
Intercetta	698,9** (10,4)	686,0** (8,7)	700,2** (5,6)	698,0** (6,9)	700,4** (5,5)
Statistiche descrittive					
SER	18,58	14,46	9,08	11,65	9,08
$\overline{R}^2$	0,049	0,424	0,773	0,626	0,773
n	420,0	420,0	420,0	420,0	420,0

Queste regressioni sono state stimate utilizzando i dati relativi ai distretti scolastici K-8 della California, descritti nell'Appendice 4.1. Gli errori standard robusti all'eteroschedasticità sono riportati tra parentesi sotto i coefficienti. Il coefficiente è significativo al livello del *5% o dell'**1% utilizzando un test bilaterale.

Riepilogo: regressione multipla

- La regressione multipla consente di stimare l'effetto su *Y* di una variazione in X_1, tenendo costanti le altri variabili incluse.
- Se potete misurare una variabile, potete evitare la distorsione della variabile omessa da tale variabile includendola.
- Se non potete misurare la variabile omessa, potreste comunque essere in grado di controllarne l'effetto includendo una variabile di controllo.
- Non esiste una ricetta semplice per decidere quali variabili appartengono a una regressione – usate il vostro giudizio.
- Un approccio è specificare un modello base – affidandosi a un ragionamento *a priori* – quindi esplorare la sensibilità delle stime chiave nelle specificazioni delle alternative.