Regressione Lineare con regressori multipli

- L'idea chiave della regressione multipla è che, se sono disponibili i dati sulle variabili omesse, possiamo aggiungerle come regressori addizionali.

- In altre parole possiamo stimare l'effetto di un regressore tenendo costanti le altre variabili.

- Tornando al nostro esempio dei distretti scolastici possiamo prendere in considerazione anche il ruolo della presenza di studenti non di madrelingua inglese (PctEL).
- Ignorando tale fattore la media della distribuzione campionaria dello stimatore OLS potrebbe non essere uguale all'effetto vero sui punteggi del test di una variazione unitaria nel rapporto studenti-insegnanti.
- In altre parole, lo stimatore OLS della pendenza della retta di regressione potrebbe essere distorto.

- I dati della California supportano il nostro timore.
- Vi è una correlazione positiva (0.19) tra il rapporto studenti-insegnanti e la percentuale di studenti non di madrelingua inglese (PctEL).
- Che informazioni otteniamo da tale coefficiente di correlazione?

Distorsione da variabile omessa

- Se il regressore è correlato con una variabile omessa dall'analisi ma che influenza la variabile dipendente, lo stimatore OLS subirà una distorsione da variabile omessa.
- Tale situazione si verifica alla presenza di due condizioni
  - La variabile omessa è correlata con il regressore incluso
  - La variabile omessa contribuisce a determinare la variabile dipendente.
- In tale caso viene meno la prima ipotesi dei minimi quadrati $E(u_i|X_i)=0$.

Esempi

- Ora del test
  - Quale condizione vale? Perche?
  - Omettere tale variabile comporta una distorsione?

- L'area di parcheggio per studente
  - Quale condizione vale? Perché?
  - Omettere tale variabile comporta una distorsione?

- Come possiamo risolvere tale problema?
- Potremmo studiare l'effetto del rapporto studenti-insegnanti tenendo costanti gli altri fattori inclusa PctEL.
- Per fare ciò dovremmo concentrarci su quei distretti in cui PctEL è simile a quella del distretto del provveditore.
- In questo sottogruppo, i distretti con classi più piccole ottengono punteggi migliori?
- La tabella successiva illustra la situazione

TABLE 6.1 Differences in Test Scores for California School Districts with Low and High Student–Teacher Ratios, by the Percentage of English Learners in the District

	Student–Teacher Ratio < 20		Student–Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High STR	
	Average Test Score	*n*	Average Test Score	*n*	Difference	*t*-statistic
All districts	657.4	238	650.0	182	7.4	4.04
Percentage of English learners						
< 1.9%	664.5	76	665.4	27	−0.9	−0.30
1.9–8.8%	665.2	64	661.8	44	3.3	1.13
8.8–23.0%	654.9	54	649.7	50	5.2	1.72
> 23.0%	636.7	44	634.8	61	1.9	0.68

- I distretti sono divisi in 8 gruppi sulla base dei quartili della distribuzione di PctEL. Inoltre, sono anche divisi in due gruppi a seconda che il rapporto studenti-insegnanti sia piccolo o grande.
- La prima riga corrisponde alla regressione già vista in cui non si è differenziato per PctEL.
- Per l'intero campione il punteggio medio è 7.4 punti più alto in quei distretti che hanno un rapporto studenti-insegnanti minore. La statistica-t è 4.04. Rigettiamo l'ipotesi nulla di uguaglianza dei punteggi medi?

- La tabella mostra che i distretti con il numero minore di PctEL, i punteggi del test sono in media 1.3 inferiori rispetto ai distretti con rapporti studenti-insegnanti bassi.
- Nel secondo quartile, i distretti con rapporti studenti-insegnanti bassi hanno ottenuto punteggi di 4.3 punti più alti di quelli con rapporti maggiori.
- Come è possibile che l'effetto totale dei punteggi è il doppio dell'effetto dei punteggi all'interno di ciascun quartile?

- I distretti con il numero maggiore di PctEl tendono ad avere sia il più alto rapporto studenti-insegnanti sia i più bassi punteggi.
- Questo risultato rafforza il timore del provveditore, ed anche il nostro, che vi sia distorsione da variabile omessa nella regressione dei punteggi del test sul rapporto studenti-insegnanti.
- La nostra analisi, tuttavia, non fornisce ancora una stima utile dell'effetto sui punteggi della variazione della dimensione delle classi, tenendo costante la frazione di studenti di madrelingua.

Il Modello di Regressione Multipla

- È una semplice estensione del modello di regressione con un singolo regressore.

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

Dove

- Y_i è la i-esima osservazione della variabile indipendente, X_{1i}, X_{2i}..,sono le i-esime osservazioni di ciascuno dei K regressori e u_i è l'errore.

- La retta di regressione della popolazione è la relazione tra la Y e la X che vale in media nella popolazione:

$$E(Y|X_{1i} = x_1, X_{2i} = x_2, ..., X_{ki} = x_k) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki}$$

- β_1 è il coefficiente angolare di X_1, β_2 è il coefficiente angolare di X_2 ecc. Il coefficiente β_1 rappresenta la variazione attesa di Y_i che deriva da una variazione unitaria in X_{1i}, tenendo costanti X_{2i},…,X_{ki}.

- L'intercetta β_0 è il valore atteso di Y, quando tutte le X sono pari a zero. Può anche essere pensata come il coefficiente di un regressore, $X_{0i,}$ che è uguale ad uno per ogni i.

Lo Stimatore OLS della Regressione Multipla

- Il metodo degli OLS può essere usato anche per stimare i coefficienti del modello di regressione multipla.
- Gli stimatori OLS $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ sono quei valori di b_0, $b_1, \ldots, b_k$ che minimizzano la somma dei quadrati degli errori di previsione

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - ... - b_k X_{ki})^2$$

- I valori predetti e i residui degli OLS sono:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + ... + \hat{\beta}_k X_{ki}, \text{ con } i = 1,..,n \text{ e } \hat{u}_i = Y_i - \hat{Y}_i$$

- Gli stimatori OLS ed il residuo sono calcolati per un campione di n osservazioni. Essi sono stimatori dei veri coefficienti ignoti della popolazione e dell'errore u_i.

Esempio

- Nel caso della regressione con un singolo regressore avevamo la seguente relazione

$$\hat{TestScore} = 698.9 - 2.28STR$$

- Adesso introducendo PctEL abbiamo

$$\hat{TestScore} = 686 - 1.10STR - 0.65PctEL$$

- Il coefficiente di STR ora rappresenta l'effetto di una variazione di STR tenendo costante PctEL mentre nella regressione con un singolo regressore PctEL non è tenuto costante.

- Abbiamo raggiunto lo stesso risultato circa la distorsione da variabile omessa mostrato nella tabella precedente.
- Tuttavia, la regressione multipla ha due vantaggi
  - Fornisce una stima quantitativa dell'effetto di un decremento unitario nel rapporto studenti-insegnanti (voluto dal provveditore)
  - Si adatta facilmente al caso di più di due regressori.

Assunzioni dei minimi quadrati per la regressione multipla

1. La distribuzione condizionata di u_i date X_{1i}, X_{2i}, …,X_{ki} ha media nulla

2. (X_{1i}, X_{2i},…,X_{ki}, Y_i) con i=1,…,n sono i.i.d.

3. X_{1i}, X_{2i},…,X_{ki} e u_i hanno quattro momenti

4. Collinearità non perfetta.
    1. I regressori sono perfettamente collineari se uno dei regressori è una funzione lineare esatta degli altri.

Esempi di collinearità

- Frazione di studenti non di madrelingua inglese (FracEL) che varia tra 1 e 0.
  - Avremmo collinearità perfetta? Perché?
  - PctEL=100*FracEL

- Percentuale di studenti di madrelingua inglese PctES
  - PctES=100*X_{0i}-PctEl

- La collinearità perfetta è una caratteristica dell'intero insieme dei regressori.

- Rimedio: diversa specificazione della regressione.

La distribuzione degli stimatori OLS nella regressione multipla

- Sotto le assunzioni dei minimi quadrati appena menzionate, gli stimatori OLS sono stimatori non distorti e consistenti dei coefficienti del modello di regressione multipla della popolazione.
- Per grandi campioni la distribuzione campionaria congiunta degli stimatori OLS è ben approssimata da una distribuzione normale multivariata.
- Anche il teorema del limite centrale si applica agli stimatori OLS nel modello di regressione multipla.

- Lo stimatore OLS del j-esimo coefficiente di regressione $\hat{\beta}_j$ ha una deviazione standard, che è stimata tramite il suo errore standard SE($\hat{\beta}_j$).

- Le idee chiave (normalità in grandi campioni degli stimatori, la capacità di stimare consistentemente la deviazione standard della loro distribuzione campionaria) non cambiano all'aumentare dei regressori.

Verifica di ipotesi ed intervalli di confidenza

- Si calcoli l'errore standardizzato di $\hat{\beta}_j$, SE($\hat{\beta}_j$)
- Si calcoli la statistica t,

$$t = \frac{\hat{\beta}_j - \hat{\beta}_{j,o}}{SE(\hat{\beta}_j)}$$

- Si calcoli il valore-p=$2\Phi(-|t^{act}|)$, dove t^{act} è il valore effettivamente calcolato della statistica t.
- Si rifiuti l'ipotesi al livello di significatività 5% se il valore-p è minore di 0.05 oppure se $t^{act}|>1.96$
- L'intervallo di confidenza di livello 95% è

$$\beta_j = (\hat{\beta}_j - 1.96 * SE(\hat{\beta}_j), \hat{\beta}_j + 1.96 * SE(\hat{\beta}_j)),$$

Esempio

- Possiamo rifiutare l'ipotesi nulla che una variazione nel numero di studenti per insegnanti non abbia alcun effetto sui punteggi del test, dopo aver controllato la percentuale di studenti non di madrelingua inglese nel distretto?
- Qual è l'intervallo di confidenza di livello 95% per l'effetto sui punteggi del test di una variazione nel rapporto studenti-insegnanti controllando per la percentuale di studenti non di madrelingua inglese?

$$\widehat{TestScore} = \underset{(8.7)}{686} - \underset{(0.43)}{1.10}\,STR - \underset{(0.031)}{0.65}\,PctEL$$

```
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
------------------------------------------------------------------------------
```

$$\widehat{TestScore} = 686.0 - 1.10' \, STR - 0.65 PctEL$$

More on this printout later…

- La statistica t relativa all'ipotesi che il coefficiente di STR sia uguale a 0 è t=(-1.10-0)/0.43=-2.54
- Il valore-p=2Φ(-2.54)=1.1%
- Cosa concludiamo?
- L'intervallo di confidenza di livello 95% per il coefficiente di STR è

  -1.10±(1.96x0.43)=(-1.95, -0.26)
- Finalmente abbiamo convinto il provveditore che, con i dati a disposizione, la riduzione della dimensione delle classi aiuterà i punteggi del test nel suo distretto
- Ma….

- Come si possono finanziare le nuove assunzioni? Tagli al bilancio o aumento di stanziamento di bilancio?
- Qual è l'effetto sui punteggi di una riduzione del rapporto studenti-insegnanti, tenendo costanti le pese per studente e la percentuale di studenti non di madrelingua inglese?

$$\widehat{TestScore} = \underset{(15.5)}{649.6} - \underset{(0.48)}{0.29}\,STR + \underset{(1.59)}{3.87}\,Expn - \underset{(0.032)}{0.656}\,PctEL$$

- Tenendo costante la spesa per studente e PctEI, la variazione di STR ha un effetto minimo sui punteggi.
- La statistica t è ora t=(-0.29-0)/0.48=-0.60
  - Rifiutiamo o no?
- È anche aumentato l'errore standard di STR dopo aver aggiunto Expn. La correlazione tra questi regressori (-0.62) rende meno precisi gli stimatori OLS.
- Per testare se il contribuente è alterato (aumento delle spese per studente) abbiamo bisogno di una nuova statistica.

Verifica di ipotesi su 2 o più coefficienti

- Il contribuente alterato ipotizza che non vi sia alcun effetto sui punteggi né da parte di STR né da parte di Expn. In termini matematici abbiamo:
- H_0:β_1=0 e β_2=0 contro H_1: $\beta_1 \neq 0$ e/o $\beta_2 \neq 0$
- L'ipotesi nulla pone delle restrizioni sul valore dei due coefficienti.
- In generale, un'ipotesi congiunta è un'ipotesi che impone due o più restrizioni sui coefficienti di regressione.
- Ad esempio in una regressione con k=6 regressori, un'ipotesi nulla è che i coefficienti del secondo, quarto e quinto regressore siano pari a zero, ponendo così 3 restrizioni.
- Se una o più delle uguaglianze sotto l'ipotesi nulla è falsa allora l'ipotesi nulla congiunta è falsa.

Statistica F

- La statistica F con q=2 restrizioni combina le due statistiche t tramite la formula

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

- Sotto l'ipotesi nulla la statistica F ha distribuzione $F_{2,\infty}$ sia che le statistiche t siano correlate o meno.
- Tale statistica si può generalizzare a q restrizioni, $F_{q,\infty}$

- Il valore-p = $Pr[F_{q,\infty} > F^{act}]$
- La statistica F per l'"intera" regressione verifica l'ipotesi congiunta che tutti i coefficienti tranne l'intercetta siano nulli.
- H_0: $\beta_1=0$, $\beta_2=0$,.., $\beta_k=0$ contro H_1: $\beta_j \neq 0$ per almeno un j, con j=1,..,k
- Sotto questa ipotesi nulla, nessuno dei regressori spiega alcunché della variazione in Y_i, sebbene l'intercetta possa essere non nulla.
- Se q=1, abbiamo un'ipotesi nulla su un singolo coefficiente di regressione e la statistica F è il quadrato della statistica t.

Esempio

- Possiamo ora verificare l'ipotesi sostenuta dal contribuente alterato, controllando per PctEL.
- Bisogna calcolare la statistica F nel caso in cui $\beta_1=0$ e $\beta_2=0$ nella regressione multipla.
- Tale statistica F è 5.43. Il valore critico al 5% della distribuzione $F_{2,\infty}$ è 3.00 e 4.61 all'1%.
- Possiamo rifiutare l'ipotesi nulla al livello 1%, mostrando che sia STR sia Expn hanno effetto sui punteggi.

Verifica di restrizioni singole che coinvolgono coefficienti

- Può verificarsi il caso di una singola restrizione che coinvolge due o più coefficienti. Ad esempio,
- H_0: $\beta_1=\beta_2$ contro H_1: $\beta_1 \neq \beta_2$
- Due soluzioni
  - Il software statistico ha un comando specifico
  - Si trasforma la nostra restrizione in una restrizione su un singolo coefficiente

```
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
------------------------------------------------------------------------------
```

NOTE

```
test str expn_stu;
```

The test command follows the regression

```
 ( 1)  str = 0.0
 ( 2)  expn_stu = 0.0

F(  2,   416) =    5.43
Prob > F =    0.0047
```

There are q=2 restrictions being tested

The 5% critical value for q=2 is 3.00

Stata computes the p-value for you

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

possiamo aggiungere e sottrarre $\beta_2 X_{1i}$ ottenendo

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

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

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

La regressione può essere riscritta come

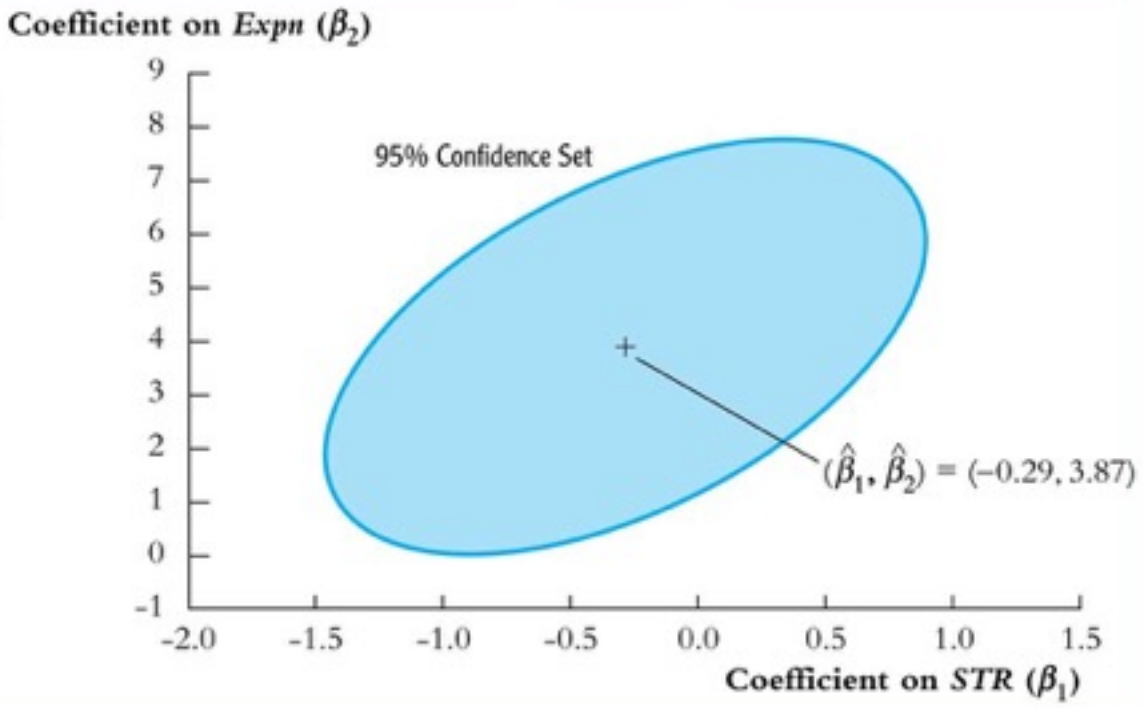
$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i$$

- L'ipotesi nulla diventa quindi $\gamma_1=0$ contro $\gamma_1\neq0$ ed è possibile usare la statistica t per tale coefficiente.

- Nella regressione multipla parliamo di regioni di confidenza per coefficienti multipli intendendo una regione che contiene i veri valori di questi coefficienti nel 95% dei campioni estratti casualmente dalla popolazione.

FIGURE 7.1 95% Confidence Set for Coefficients on *STR* and *Expn* from Equation (7.6)

The 95% confidence set for the coefficients on *STR* (β_1) and *Expn* (β_2) is an ellipse. The ellipse contains the pairs of values of β_1 and β_2 that cannot be rejected using the *F*-statistic at the 5% significance level.



- L’errore standard della regressione (SER) stima la deviazione standard dell’errore u_i ed è

$$SER = s_{\hat{u}}, \text{ dove } s_{\hat{u}}^2 = \frac{1}{n-k-1}\sum_{i=1}^{1} \hat{u}_i^2 = \frac{SSR}{n-k-1}$$

- La sola differenza con la definizione di SER già vista è che il denominatore è n-k-1 invece di n-2.
- In questo caso, stimiamo k+1 coefficienti e quindi correggiamo per i gradi di libertà.

- L'R^2 è la frazione della varianza campionaria di Y_i spiegata dai regressori. La sua definizione matematica è la stessa che per la regressione con un singolo regressore.

- Nella regressione multipla l'R^2 cresce ogni volta che viene aggiunto un regressore.

- Per questo motivo un suo aumento non necessariamente implica un miglioramento della nostra regressione. Potremmo avere una stima in eccesso della bontà della regressione.

- Per risolvere tale problema si usa l'R^2corretto o $\overline{R}^2$

Il quale non cresce all'aumentare dei regressori

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s_{\hat{u}}^2}{s_Y^2}$$

- La differenza tra questa formula e quella dell'R^2 è data dal termine (n-1)/(n-k-1) che è la solita correzione per i gradi di libertà.

- Dato che tale (n-1)/(n-k-1) è sempre maggiore di uno, $\overline{R}^2$ è sempre minore di R^2.

- L'aggiunta di un regressore ha due effetti opposti su R^2 corretto.
  - L'SSR cresce e quindi aumenta l'R^2 corretto
  - Il fattore (n-1)/(n-k-1) aumenta.

- L'R^2 corretto può essere negativo.

Quattro potenziali problemi

1. Un aumento dell'R^2, o dell'R^2 corretto, non significa necessariamente che la variabile aggiunta sia statisticamente significativa. Per essere sicuri è meglio condurre un test con la statistica t
2. Un R^2, o un R^2 corretto, elevato non implica che i regressori siano la vera causa della variabile dipendente. La relazione tra variabili può non essere causale (STR ed i parcheggi)

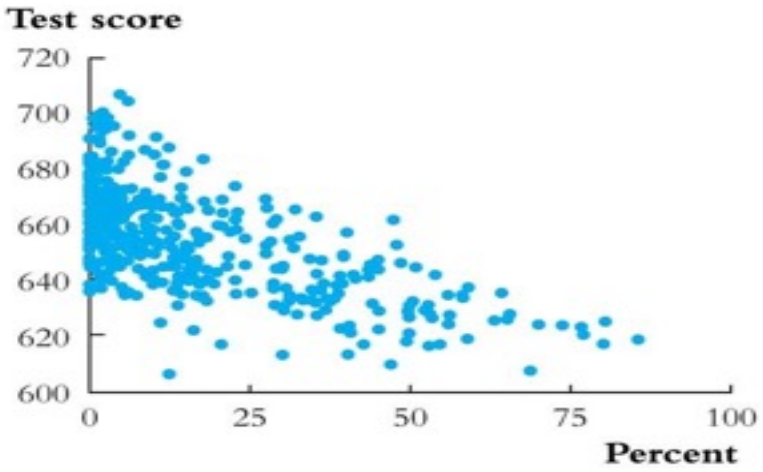
3. Un R^2, o un R^2 corretto, elevato non implica che vi sia distorsione da variabile omessa.

4. Un R^2, o un R^2 corretto, elevato non significa necessariamente che abbiamo scelto l'insieme di regressori più appropriato, né un basso R^2, o R^2 corretto, implica che ne abbiamo scelto uno inappropriato.

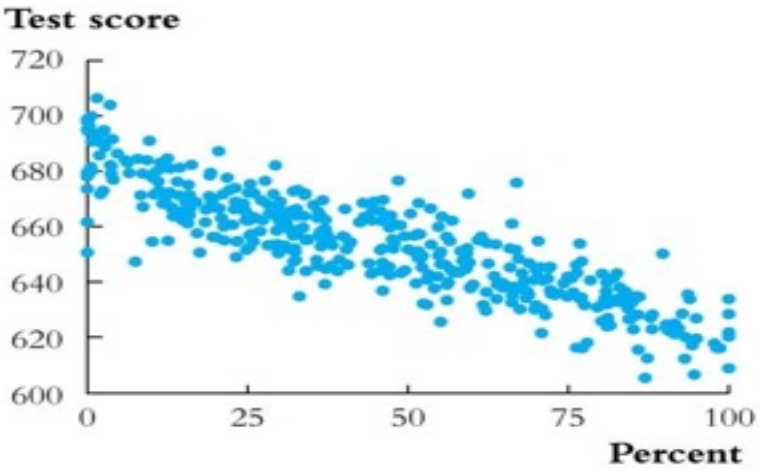
- Distorsione da variabile omessa nel caso di regressione multipla
- Specificazione di base
- Specificazioni alternative

Esempio

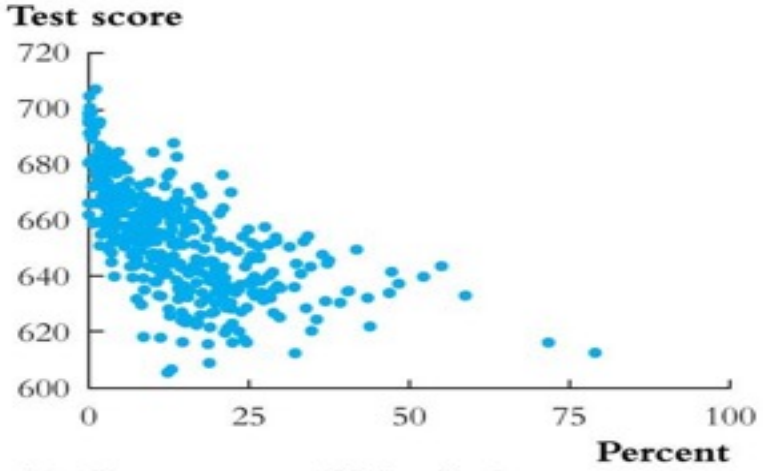
- Consideriamo l'effetto di tre variabili PctEl, la percentuale di studenti che hanno diritto al sussidio mensa totale o parziale e la percentuale di studenti nel distretto in cui la famiglia è idonea per il programma di pubblica assistenza sul reddito dello stato della California.
- Questi due nuovi indicatori misurano la frazione di bambini economicamente svantaggiati nel distretto



(a) Percentage of English language learners



(b) Percentage qualifying for reduced price lunch



(c) Percentage qualifying for income assistance

- Correlazione a) -0.64; b) -0.87; c) -0.63

TABLE 7.1 Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28** (0.52)	−1.10* (0.43)	−1.00** (0.27)	−1.31** (0.34)	−1.01** (0.27)
Percent English learners (X_2)		−0.650** (0.031)	−0.122** (0.033)	−0.488** (0.030)	−0.130** (0.036)
Percent eligible for subsidized lunch (X_3)			−0.547** (0.024)		−0.529** (0.038)
Percent on public income assistance (X_4)				−0.790** (0.068)	0.048 (0.059)
Intercept	698.9** (10.4)	686.0** (8.7)	700.2** (5.6)	698.0** (6.9)	700.4** (5.5)
Summary Statistics					
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	420	420	420	420

These regressions were estimated using the data on K-8 school districts in California, described in Appendix 4.1. Standard errors are given in parentheses under coefficients. The individual coefficient is statistically significant at the *5% level or **1% significance level using a two-sided test.

- Nel caso di molte regressioni multivariate, la migliore presentazione è quella mostrata in tabella.
- Le statistiche t possono essere facilmente calcolate.
- I risultati suggeriscono 3 conclusioni

1. Controllare per queste caratteristiche riduce l'effetto del rapporto studenti-insegnanti sui punteggi di circa la metà. L'effetto stimato non è molto sensibile alle variabili di controllo usate. Comunque STR è sempre significativa al 5% ed una sua riduzione comporta un aumento del punteggio del test di un punto circa.

2. Le variabili che rappresentano le caratteristiche degli studenti sono molto utili. La variabile STR spiega poco da sola. Si veda l'R^2 corretto. Infine il segno dei coefficienti è coerente con l'andamento osservato in figura. In altre parole, i distretti con molti studenti non di madrelingua ed i distretti con molti bambini poveri ottengono punteggi più bassi.

3. Le variabili di controllo non sono sempre statisticamente significative. In (5) il coefficiente di X_4 non è statisticamente diverso da zero. Quindi, poiché aggiungere tale variabile a (3) non comporta miglioramenti sensibili e poiché non è significativo in (5), tale variabile di controllo addizionale è ridondante in quest'analisi.