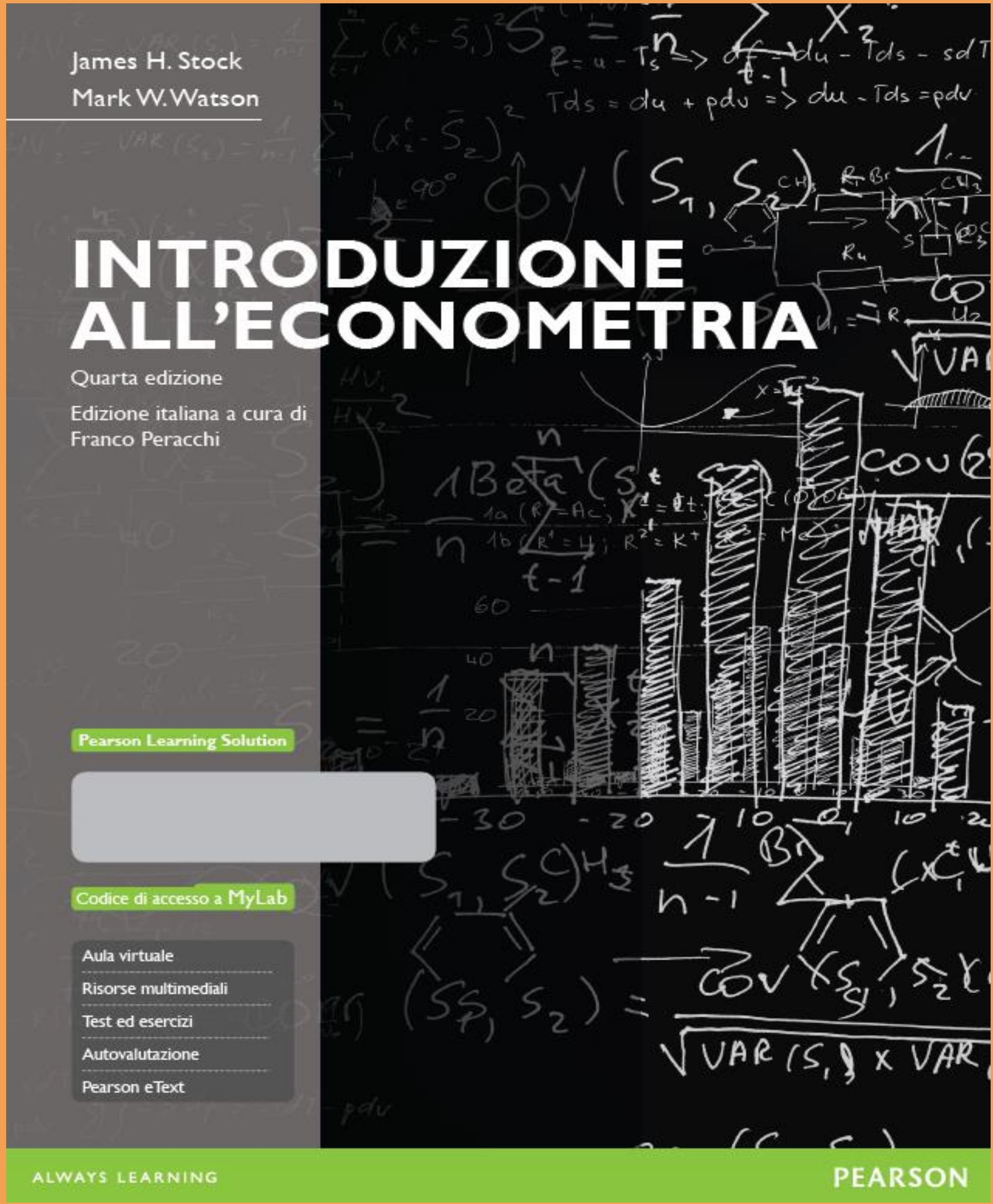
Capitolo 6

Regressione lineare con regressori multipli



Sommario

1. La distorsione da variabili omesse
2. Il modello di regressione multipla.
3. Lo stimatore OLS della regressione multipla
4. Misure di bontà dell'adattamento nella regressione multipla
5. Le assunzioni dei minimi quadrati per la regressione multpla
6. La distribuzoine degli stimatori OLS nella regressione multipla
7. Distribuzione campionaria dello stimatore OLS

La distorsione da variabili omesse (Paragrafo 6.1)

L'errore *u* si verifica a causa di fattori, o variabili, che influenzano *Y* ma non sono inclusi nella funzione di regressione. Ci sono sempre variabili omesse.

Talvolta l'omissione di queste variabili può portare a una distorsione dello stimatore OLS.

La distorsione da variabili omesse (continua)

La distorsione dello stimatore OLS che si verifica a seguito di un fattore, o variabile, omesso è detta **distorsione da variabile omessa**. Affinché si verifichi tale distorsione, la variabile omessa "*Z*" deve soddisfare due condizioni:

<u>Le due condizioni per la distorsione da variabile omessa</u>

1.Z è un determinante di *Y* (cioè *Z* è parte di *u*); **e**

2.Z è correlata con il regressore *X*
(*cioè* corr(*Z*,*X*) $\neq$ 0)

Entrambe *le condizioni devono verificarsi affinché l'omissione di Z porti a distorsione da variabile omessa*.

La distorsione da variabili omesse (continua)

Nell'esempio dei punteggi nei test:

1. Il livello di conoscenza della lingua inglese (se lo studente è di madrelingua o meno) verosimilmente influisce sui punteggi nei test standardizzati: *Z* è un determinante di *Y*.
2. Le comunità di immigrati tendono a una minore affluenza e quindi hanno budget scolastici inferiori e *STR* maggiori: *Z* è correlata con *X*.

Di conseguenza, $\hat{\beta}_1$ è distorto. In quale direzione?

- *Che cosa suggerisce il buon senso?*
- Se il buon senso vi fa difetto, c'è una formula…

La distorsione da variabili omesse (continua)

Formula per la distorsione da variabili omesse: si ricordi l'equazione

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}v_i}{\left(\frac{n-1}{n}\right)s_X^2}$$

Dove $v_i = (X_i - \bar{X})u_i \approx (X_i - \mu_X)u_i$. Sotto la prima assunzione dei minimi quadrati,

$$E[(X_i - \mu_X)u_i] = \mathrm{cov}(X_i,u_i) = 0.$$

Ma se $E[(X_i - \mu_X)u_i] = \mathrm{cov}(X_i,u_i) = \sigma_{Xu} \neq 0$?

La distorsione da variabili omesse (continua)

Sotto le assunzioni dei minimi quadrati #2 e #3 (cioè anche se la prima assunzione dei minimi quadrati non è vera),

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\xrightarrow{p} \frac{\sigma_{Xu}}{\sigma_X^2}$$

$$= \left(\frac{\sigma_u}{\sigma_X}\right) \times \left(\frac{\sigma_{Xu}}{\sigma_X \sigma_u}\right) = \left(\frac{\sigma_u}{\sigma_X}\right)\rho_{Xu}$$

dove ρ_{Xu} = corr(X,u). Se vale la prima assunzione, allora ρ_{Xu} = 0, ma se non vale abbiamo….

Formula della distorsione da variabili omesse:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \left(\frac{\sigma_u}{\sigma_X}\right)\rho_{Xu}$$

- Se una variabile omessa *Z* è ***contemporaneamente***:

1. una determinante di *Y* (cioè se è contenuta in *u*); ***e***
2. correlata con *X*,

allora $\rho_{Xu} \neq 0$ e lo stimatore OLS $\hat{\beta}_1$ è distorto e inconsistente.

- Per esempio, i distretti scolastici con pochi studenti non di madrelingua (1) ottengono punteggi migliori nei test standardizzati e (2) hanno classi più piccole (budget più elevati), perciò ignorando l'effetto di avere molti studenti non di madrelingua si arriverebbe a sovrastimare l'effetto della dimensione delle classi. *Si verifica questo nei dati riferiti alla California*?

Tabella 6.1 Differenza tra i punteggi nei test dei distretti scolastici della California con bassi e alti rapporti studenti/insegnanti (STR), per percentuali diverse di studenti non di madrelingua inglese nel distretto.

	Rapporto studenti-insegnanti < 20		Rapporto studenti-insegnanti ≥ 20		Differenza tra punteggi, basso v/s alto STR	
	Media punteggi	n	Media punteggi	n	Differenza	Statistica t
Tutti i distretti	**657,4**	**238**	**650,0**	**182**	**7,4**	**4,04**
Percentuale di studenti non di madrelingua inglese						
< 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 con meno studenti non di madrelingua ottengono migliori punteggi nei testi.
- I distrettti con una minore percentuale di studenti non di madrelingua hanno classi più piccole.
- Tra i distretti con percentuali di studenti non di madrelingua comparabili, l’effetto della dimensione delle classi è piccolo (si ricordi che complessivamente la “differenza di punteggio nei test” = 7.4).

Causalità e analisi di regressione

- L'esempio dei punteggi nei test/*STR*/percentuale di studenti non di madrelingua mostra che, se una variabile omessa soddisfa le due condizioni della distorsione da variabili omesse, allora lo stimatore OLS nella regressione che omtte tale variabile è distorto e inconsistente. Perciò, anche se *n* è grande, $\hat{\beta}_1$ non sarà vicino a β_1.

- Ciò fa sorgere una domanda più profonda: come definiamo β_1? Ovvero, che cosa vogliamo stimare, precisamente, quando eseguiamo una regressione?

Che cosa vogliamo stimare, precisamente, quando eseguiamo una regressione?

Esistono (almeno) tre possibili risposte a questa domanda:

1. Vogliamo stimare la pendenza di una retta attraverso un diagramma a nuvola come semplice riepilogo dei dati a cui non associamo un significato sostanziale.

   Questo può essere utile talvolta, ma non è molto interessante a livello intellettuale e non rientra nell'obiettivo di questo corso.

2. Vogliamo effettuare previsioni del valore di *Y* per una unità che non appartiene all'insieme dei dati, per cui conosciamo il valore di *X*.

Realizzare previsioni è importante per gli economisti, ed è possibile ottenere previsioni eccellenti utilizzando i metodi di regressione senza la necessità di conoscere gli effetti causali. Torneremo a questo tema più avanti nel corso.

3. Vogliamo stimare l'effetto causale su *Y* di una variazione in *X*.

Ecco perché siamo interessati all'effetto della dimensione delle classi. Si supponga che il consiglio scolastico decida una riduzione di 2 studenti per classe. Quale sarebbe l'effetto sui punteggi nei test? Questa è una domanda causale (qual è l'effetto causale sui punteggi nei test di STR?) perciò dobbiamo stimare questo effetto causale.

A parte la discussione dell'attività di previsione, lo scopo di questo corso è la stima di effetti causali mediante metodi di regressione.

Che cos'è, precisamente, un effetto causale?

- La "causalità" è un concetto complesso!

- In questo corso adottiamo un approccio pratico alla definizione di causalità:

Un effetto causale è definito come un effetto misurato in un esperimento controllato casualizzato ideale.

Esperimento controllato causalizzato ideale

- *Ideale*: i soggetti seguono tutti il protocollo di trattamento – perfetta compliance, nessun errore nei report, ecc.!
- *Casualizzato*: i soggetti della popolazione di interesse sono assegnati casualmente a un gruppo di trattamento o di controllo (così non ci sono fattori di confusione)
- *Controllato*: la disponibilità di un gruppo di controllo permette di misurare l'effetto differenziale del trattamento
- *Esperimento*: il trattamento è assegnato nell'esperimento: i soggetti non hanno scelta, perciò non vi è "causalità inversa" in cui i soggetti scelgono il trattamento che ritengono migliore.

Tornando alla dimensione delle classi:

Si immagini un esperimento controllato casualizzato ideale per misurare l'effetto sui punteggi nei test della riduzione di *STR*…

- In tale esperimento gli studenti sarebbero assegnati casualmente alle classi, che avrebbero dimensioni diverse.

- Poiché gli studenti sono assegnati casualmente, tutte le loro caratteristiche (e quindi gli u_i) sarebbero distribuiti in modo indipendente da STR_i.

- Quindi, $E(u_i|STR_i) = 0$ – cioè la prima assunzione dei minimi quadrati vale in un esperimento controllato casualizzato.

In che modo i nostri dati osservazionali differiscono da questa situazione ideale?

- Il trattamento non è assegnato in modo casuale
- Si consideri *PctEL* – la percentuale di studenti non di madrelingua – nel distretto. Verosimilmente soddisfa i due criteri per la distorsione da variabili omesse: $Z = PctEL$ è:
  1. un determinante di Y; ***e***
  2. correlata con il regressore X.
- Quindi i gruppi “di controllo” e “di trattamento” differiscono in modo sistematico, perciò corr(STR,$PctEL$) $\neq 0$

- Casualizzazione + gruppo di controllo significa che qualsiasi differenza tra i gruppi di trattamento e di controllo è casuale – non sistematicamente correlata al trattamento

- Possiamo eliminare la differenza di *PctEL* tra il gruppo di classi grandi (di controllo) e quello di classi piccole (di trattamento) esaminando l'effetto della dimensione delle classi tra i distretti con lo stesso valore di *PctEL*.

  – Se soltanto la differenza sistematica tra i gruppi di classi grandi e piccole è in *PctEL*, allora torniamo all'esperimento controllato casualizzato – all'interno di ciascun gruppo di *PctEL*.

  – Questo è un modo per "controlare" per l'effetto di *PctEL* quando si stima l'effetto di *STR*.

Tornando alla distorsione da variabili omesse

Tre modi per superare la distorsione da variabili omesse

1. Eseguire un esperimento controllato casualizzato in cui il trattamento (*STR*) sia assegnato casualmente: allora *PctEL* è ancora un determinante di *TestScore*, ma *PctEL* è incorrelato con *STR*. (*Questa soluzione è raramente praticabile.*)

2. Adottare l'approccio "a tabulazione incrociata", con gradazioni più fini di *STR* e *PctEL* – all'interno di ogni gruppo, tutte le classi hanno lo stesso *PctEL*, perciò controlliamo per *PctEL* (*ma presto si esauriranno i dati, e che dire di altri determinanti come il reddito famigliare e il livello di istruzione dei genitori*?)

3. Usare una regressione in cui la variabile omessa (*PctEL*) non è più omessa: includere *PctEL* come regressore aggiuntivo in una regressione multipla.

Il modello di regressione multipla (Paragrafo 6.2)

- Si consideri il caso di due regressori:

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

- Y è la *variabile dipendente*
- X_1, X_2 sono le due *variabili indipendenti* (*regressori*)
- (Y_i, X_{1i}, X_{2i}) denotano l'i-esima osservazione su Y, X_1 e X_2.
- β_0 = intercetta della popolazione ignota
- β_1 = effetto su Y di una variazione in X_1, tenendo X_2 costante
- β_2 = effetto su Y di una variazione in X_2, tenendo X_1 costante
- u_i = errore di regressione (fattori omessi)

Interpretazione dei coefficienti nella regressione multipla

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

Si consideri di variare X_1 di ΔX_1 tenendo X_2 costante:

Retta di regressione della popolazione ***prima*** della variazione:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Retta di regressione della popolazione ***dopo*** la variazione:

$$Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2$$

Prima: $Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$

Dopo: $Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$

Differenza: $\Delta Y = \beta_1 \Delta X_1$

Quindi:

$\beta_1 = \dfrac{\Delta Y}{\Delta X_1'}$ **tenendo X_2 costante**

$\beta_2 = \dfrac{\Delta Y}{\Delta X_2'}$ **tenendo X_1 costante**

β_0 = valore predetto di Y quando $X_1 = X_2 = 0$.

Lo stimatore OLS della regressione multipla (Paragrafo 6.3)

- Con due regressori, lo stimatore OLS risolve:

$$\min_{b_0, b_1, b_2} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_{1i} + b_2 X_{2i})]^2$$

- Lo stimatore OLS minimizza la differenza quadratica media tra i valori attuali di Y_i e il valore predetto in base alla retta stimata.
- Questo problema di minimizzazione si risolve usando l'analisi matematica
- **Così si ottengono gli stimatori OLS di $\boldsymbol{\beta_0}$ e $\boldsymbol{\beta_1}$.**

Esempio: i dati dei punteggi nei test della California

Regressione di *TestScore* su *STR*:

$$\widehat{TestScore} = 698{,}9 - 2{,}28 \times STR$$

Ora includiamo la percentuale di studenti non di madrelingua nel distretto (*PctEL*):

$$\widehat{TestScore} = 686{,}0 - 1{,}10 \times STR - 0{,}65 PctEL$$

- Che cosa accade al coefficiente di *STR*?
- (*STR*, *PctEL*) = 0,19)

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

$$\widehat{TestScore} = 686{,}0 - 1{,}10 \times STR - 0{,}65 PctEL$$

Più avanti torneremo su questo stampato…

Misure di bontà dell'adattamento nella regressione multipla (Paragrafo 6.4)

Reale = predetto + residuale: $Y_i = \hat{Y}_i + \hat{u}_i$

SER = deviazione standard di $\hat{u}_i$ (con correzione per gr. lib.)

$RMSE$ = deviazione standard di $\hat{u}_i$ (senza correzione per gr. lib.)

R^2 = frazione della varianza di Y spiegata da X

$\bar{R}^2$ = "R^2 corretto" = R^2 con una correzione per gradi di libertà che corregge per l'incertezza della stima; $\bar{R}^2 < R^2$

SER e RMSE

Come nella regressione con un unico regressore, *SER* e *RMSE* sono misure della dispersione delle *Y* attorno alla retta di regressione:

SER = $$\sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2}$$

RMSE = $$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}$$

R^2 e $\bar{R}^2$ (R^2 corretto)

L'R^2 è la frazione della varianza spiegata – stessa definizione della regressione con singolo regressore:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

dove $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2$, $SSR = \sum_{i=1}^{n}\hat{u}_i^2$, $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

- L'R^2 aumenta sempre quando si aggiunge un altro regressore (*perché?)* – un problema per una misura di "adattamento"

R^2 e $\bar{R}^2$ (continua)

L' $\bar{R}^2$ (l'"R^2 corretto") corregge questo problema "penalizzandovi" per l'inserimento di un altro regressore – l' $\bar{R}^2$ non aumenta necessariamente quando si aggiunge un altro regressore.

R^2 corretto : $\bar{R}^2$ = $1-\left(\frac{n-1}{n-k-1}\right)\frac{SSR}{TSS}$

Si noti che $\bar{R}^2 < R^2$, tuttavia se n è grande i due saranno molto vicini.

Misure di bontà dell'adattamento (continua)

Esempio del punteggio nei test:

$$(1)\quad \widehat{TestScore} = 698{,}9 - 2{,}28 \times STR,$$
$$R^2 = 0{,}05,\ SER = 18{,}6$$

$$(2)\quad \widehat{TestScore} = 686{,}0 - 1{,}10 \times STR - 0{,}65 PctEL,$$
$$R^2 = 0{,}426,\quad \bar{R}^2 = 0{,}424,\ SER = 14{,}5$$

- *Che cosa vi dice questo – precisamente – riguardo la bontà dell'adattamento della regressione (2) rispetto alla regressione (1)?*
- *perché l'R^2 e l'$\bar{R}^2$ sono così vicini in (2)?*

Le assunzioni dei minimi quadrati per la regressione multipla (Paragrafo 6.5)

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

1. La distribuzione di u condizionata alle X ha media nulla, cioè $E(u_i|X_{1i} = x_1,\ldots, X_{ki} = x_k) = 0$.

2. $(X_{1i},\ldots,X_{ki},Y_i)$, $i = 1,\ldots,n$, sono i.i.d.

3. Gli outlier sono improbabili: $X_1,\ldots, X_k$, e Y hanno momenti quarti: $E(X_{1i}^4) < \infty,\ldots, E(X_{ki}^4) < \infty$, $E(Y_i^4) < \infty$.

4. Non vi è collinearità perfetta.

Assunzione 1: la media condizionata di *u* date le *X* incluse è zero.

$$E(u|X_1 = x_1,\ldots, X_k = x_k) = 0$$

Ha la stessa interpretazione del caso della regressione con un singolo regressore.

- La non validità di questa condizione porta a distorsione da variabili omesse; nello specifico, se una variabile omessa

1. appartiene all'equazione (cioè è in *u*) ***e***
2. è correlata con una *X* inclusa

- allora questa condizione non vale e vi è distorsione da variabili omesse.
- La soluzione migliore, se possibile, è quella di includere la variabile omessa nella regressione.
- Una seconda soluzione, correlata alla precedente, è quella di includere una variabile che controlli per la variabile omessa (cfr. Capitolo 7)

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

È soddisfatta automaticamente se i dati sono raccolti mediante campionamento casuale semplice.

Assunzione 3: gli outlier sono rari (momenti quarti finiti)

È la stessa assunzione descritta per il caso di un regressore singolo. Come in quel caso, l'OLS può essere sensibile agli outlier, perciò occorre controllare i dati (diagrammi a nuvola!) per assicurarsi che non vi siano valori "impazziti" (refusi o errori di codifica).

Assunzione 4: Non vi è collinearità perfetta

La ***collinearità perfetta*** si ha quando uno dei regressori è funzione lineare esatta degli altri.

Esempio: si supponga di includere due volte *STR*, per errore:

```
regress testscr str str, robust
Regression with robust standard errors               Number of obs =     420
                                                     F(  1,    418) =   19.26
                                                     Prob > F      =  0.0000
                                                     R-squared     =  0.0512
                                                     Root MSE      =  18.581
-------------------------------------------------------------------------
        |               Robust
testscr |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
--------+----------------------------------------------------------------
    str |  -2.279808   .5194892    -4.39    0.000    -3.300945   -1.258671
    str |  (dropped)
  _cons |    698.933   10.36436    67.44    0.000     678.5602    719.3057
-------------------------------------------------------------------------
```

La ***collinearità perfetta*** si ha quando uno dei regressori è funzione lineare esatta degli altri.

- Nella regressione precedente, β_1 è l'effetto su *TestScore* di una variazione unitaria in *STR*, tenendo *STR* costante (???)
- Torneremo alla collinearità perfetta (e imperfetta) tra breve, con altri esempi…
-
- *Con le assunzioni dei minimi quadrati, ora possiamo derivare la distribuzione campionaria di* $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$.

La distribuzione degli stimatori OLS nella regressione multipla (Paragrafo 6.6)

Sotto le quattro assunzioni dei minimi quadrati,

- La distribuzione campionaria di $\hat{\beta}_1$ ha media β_1
- var($\hat{\beta}_1$) è inversamente proporzionale a *n*.
- Al di là di media e varianza, la distribuzione esatta (*n*-finita) di $\hat{\beta}_1$ è molto complessa; ma per *n* grande...
  - $\hat{\beta}_1$ è consistente: $\hat{\beta}_1 \xrightarrow{p} \beta_1$ (legge dei grandi numeri)
  - $\dfrac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\mathrm{var}(\hat{\beta}_1)}}$ è approssimata da una distribuzione *N*(0,1) (TLC)
  - Queste proprietà valgono per $\hat{\beta}_1, \ldots, \hat{\beta}_k$

Concettualmente, non vi è nulla di nuovo!

Collinearità perfetta e imperfetta (Paragrafo 6.7)

La ***collinearità perfetta*** si ha quando uno dei regressori è una funzione lineare esatta degli altri.

Altri esempi di collinearità perfetta

1. Dal caso precedente: includete *STR* due volte,
2. Eseguite la regressione di *TestScore* su una costante, *D*, e *B*, dove: $D_i = 1$ se $STR \leq 20$, $= 0$ altrimenti; $B_i = 1$ se $STR > 20$, $= 0$ altrimenti, perciò $B_i = 1 - D_i$ e vi è collinearità perfetta.
3. Ci sarebbe collinearità perfetta se l'intercetta (costante) fosse esclusa da questa regressione? Questo esempio è un caso speciale di…

La trappola delle variabili dummy

Si supponga di avere un insieme di più variabili binarie (dummy) che sono mutuamente esclusive ed esaustive – cioè esistono più categorie e ogni osservazione ricade in una di esse e solo in una (Matricole, Studenti del secondo anno, Junior, Senior, Altri). Se includete tutte queste variabili dummy *e* una costante, avrete collinearità perfetta – si parla talvolta di ***trappola delle variabili dummy***.

•*Perché vi è collinearità perfetta in questo caso*?

•*Soluzioni alla trappola delle variabili dummy*:

1.omettere uno dei gruppi (per esempio Senior), oppure

2.omettere l'intercetta

•*Quali sono le implicazioni di (1) o (2) per l'interpretazione dei coefficienti?*

Collinearità perfetta (continua)

- La collinearità perfetta solitamente riflette un errore nelle definizioni dei regressori, o una stranezza nei dati
- Se avete collinearità perfetta, il software statistico ve lo farà sapere – bloccandosi, o mostrando un messaggio di errore, o "scaricando" arbitrariamente una delle variabili
- La soluzione alla collinearità perfetta consiste nel modificare l'elenco di regressori.

Collinearità imperfetta

La collinearità imperfetta è ben diversa dalla collinearità perfetta, nonostante la somiglianza dei nomi.

La ***collinearità imperfetta*** si verifica quando due o più regressori sono altamente correlati.

•Perché si usa il termine "collinearità"? Se due regressori sono altamente correlati, allora il loro diagramma a nuvola apparirà molto simile a una retta – sono "co-lineari" – ma a meno che la correlazione sia esattamente ± 1, tale collinearità è imperfetta.

Collinearità imperfetta (continua)

La collinearità imperfetta implica che uno o più dei coefficienti di regressione sarà stimato in modo impreciso.

- L'idea: il coefficiente di X_1 è l'effetto di X_1 tenendo costante X_2; ma se X_1 e X_2 sono altamente correlati, vi è una ridottissima variazione in X_1 quando X_2 è mantenuta costante – perciò i dati non contengono molte informazioni su ciò che accade quando X_1 cambia e X_2 no. In questo caso, la varianza dello stimatore OLS del coefficiente di X_1 sarà grande.

- La collinearità imperfetta (correttamente) genera grandi errori standard per uno o più dei coefficienti OLS.

- La matematica? Cfr. il volume stampato, Appendice 6.2

Prossimo argomento: test di ipotesi e intervalli di confidenza…