Regressione lineare multipla

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 multipla
6. La distribuzione degli stimatori OLS nella regressione multipla
7. Verifica di ipotesi congiunte
8. Specificazione del modello e presentazione dei risultati

La distorsione da variabili omesse

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 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:

Le due condizioni per la distorsione da variabile omessa

1. Z è correlata con il regressore *X* (*cioè* corr(*Z*,*X*) ≠ 0)

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

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

Nell'esempio dei punteggi nei test:

1. Le comunità di immigrati tendono ad avere redditi più bassi e quindi budget scolastici inferiori e *STR* maggiori: *Z* è correlata con *X*.
2. 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*.

Di conseguenza, β_1 è distorto. In quale direzione?

- *Che cosa suggerisce il buon senso?*
- Se il buon senso non basta, c'è una formula…

Altre variabili omesse sono ad esempio l'ora del test, l'area di parcheggio per studente... Possono determinare problemi di distorsione da variabili omesse?

Formula per la distorsione da variabili omesse.

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

$$\hat{\beta}_1 \rightarrow \beta_1 + \rho_{Xu}\frac{\sigma_u}{\sigma_X}$$

dove ρ_{Xu} = corr(*X*,*u*). Se vale la prima assunzione, allora $\rho_{Xu} = 0$, ma se non vale 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 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 madrelingua si arriverebbe a sovrastimare l'effetto della dimensione delle classi. Matematicamente

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

dove β_1 e β_2 hanno entrambi segno negativo. Quindi se $\beta_2 Z_i$ è escluso e finisce nel termine d'errore, $\rho_{Xu} < 0$ e $\hat{\beta}_1$ risulta distorto verso il basso.

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 omette 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?
- 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 per un economista 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.

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

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?).

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*) ≠ 0

- Casualizzazione + gruppo di controllo significa che i gruppi di trattamento e di controllo sono omogenei tra loro e rispetto a tutte le variabili non osservate, e quindi l'effetto misurato dipende esclusivamente dal 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*.
  - Questo è un modo per "controllare" per l'effetto di *PctEL* quando si stima l'effetto di *STR*.
  - Cioè è un modo di ricondursi ad un esperimento controllato casualizzato.

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. Separare i dati in gruppi, rispetto ai valori di *PctEL* – all'interno di ogni gruppo, tutte le classi hanno lo stesso *PctEL*, perciò controlliamo per *PctEL* (*ma per ogni gruppo si avrebbero pochi dati, e che dire di altri determinanti come il reddito familiare 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

- 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

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

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:

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

Facendo la differenza membro a membro

$$\Delta Y = \beta_1 \Delta X_1 \quad \rightarrow \quad \beta_1 = \frac{\Delta Y}{\Delta X_1}$$

Lo stimatore OLS della regressione multipla

- 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 β_0 e β_1.**

Esempio: i dati dei punteggi nei test della California

Regressione di *TestScore* su *STR*:

$\widehat{TestScore}$ = 698,9 – 2,28×*STR*

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

$\widehat{TestScore}$ = 686,0 – 1,10×*STR* – 0,65*PctEL*

- Che cosa accade al coefficiente di *STR*?

Misure di bontà dell'adattamento nella regressione multipla

Reale = predetto + residuale: $y_i = \hat{y}_i + \hat{u}_i$

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

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

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.

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

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

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$
$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$

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

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{RSS}{TSS}$$

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

Concetto chiave 5.8: *l'R^2 e l'$\bar{R}^2$: cosa ci dicono e cosa non ci dicono*

L'R^2 e l'$\bar{R}^2$ ci dicono se i regressori sono idonei a prevedere, o a "spiegare" i valori della variabile dipendente nel campione di dati a disposizione. Se l'R^2 (o l'$\bar{R}^2$) tende a uno, i regressori producono delle buone previsioni della variabile dipendente in quel campione, nel senso che la varianza dei residui OLS è piccola rispetto alla varianza della variabile dipendente. Se l'R^2 (o l'$\bar{R}^2$) tende a zero, è vero il contrario.

L'R^2 e l'$\bar{R}^2$ NON ci dicono se:

1. una variabile inclusa è statisticamente significativa;
2. i regressori sono causa effettiva dei movimenti della variabile dipendente;
3. c'è una distorsione da variabile omessa;
4. abbiamo scelto il gruppo di regressori più appropriato.

Esempio del punteggio nei test:

(1) $\widehat{TestScore}$ = 698,9 – 2,28×*STR*,
R^2 = 0,05, *SER* = 18,6

(2) $\widehat{TestScore}$ = 686,0 – 1,10×*STR* – 0,65*PctEL*,
R^2 = 0,426, $\bar{R}^2$ = 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

Concetto chiave 5.4: *le assunzioni dei minimi quadrati relative al modello di regressione multipla*

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

1. u_i ha media condizionata nulla, date $X_{1i}, X_{2i}, \ldots, X_{ki}$, ovvero $E(u_i | X_{1i}, X_{2i}, \ldots, X_{ki}) = 0$;
2. $(X_{1i}, \ldots, X_{ki}, Y_i)$, con $i = 1, \ldots, n$, sono estratti indipendentemente e indenticamente distribuiti (i.i.d.) dalla propria distribuzione congiunta;
3. $(X_{1i}, \ldots, X_{ki}, u_i)$ hanno momenti quarti finiti e non nulli;
4. non vi è collinearità perfetta.

Collinearità perfetta e imperfetta

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

Esempi di collinearità perfetta

1. Inclusione ripetuta di uno stesso regressore (e.s. *STR*)
2. Inclusione nel modello della frequenza e della percentuale di studenti stranieri
3. Inclusione nel modello di una dummy per classi "non troppo piccole", ossia $D_i = 1$ se $STR_i \geq 12$ e $D_i = 0$ altrimenti
4. Inclusione nel modello della percentuale di studenti madrelingua e degli studenti non madrelingua
5. Inclusione nel modello di una dummy per le classi piccole e di una dummy per le classi grandi
6. 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 (Classi piccole, classi medie, classi grandi). Se tutte queste variabili dummy e una costante sono incluse nel modello, si avrà collinearità perfetta – si parla in questo caso di ***trappola delle variabili dummy***.

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

•*Soluzioni alla trappola delle variabili dummy*:

1.omettere la dummy per uno dei gruppi (per esempio classi piccole), oppure

2.omettere l'intercetta

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

Parametrizzazioni diverse (risposta alla precedente domanda)

•*Soluzioni alla trappola delle variabili dummy*:

1. omettere la dummy per uno dei gruppi (per esempio classi grandi),

$$TestScore_i = \beta_0 + \beta_1 \cdot el\_pct_i + \beta_2 \cdot D_{1i} + \beta_3 \cdot D_{2i}$$

classi grandi → $TestScore = \beta_0 + \beta_1 \cdot el\_pct$

classi medie → $TestScore = \beta_0 + \beta_3 + \beta_1 \cdot el\_pct$

classi piccole → $TestScore = \beta_0 + \beta_2 + \beta_1 \cdot el\_pct$

2. omettere l'intercetta

$$TestScore_i = \gamma_1 \cdot el\_pct_i + \gamma_2 \cdot D_{1i} + \gamma_3 \cdot D_{2i} + \gamma_4 \cdot D_{3i}$$

classi grandi → $TestScore = \gamma_1 \cdot el\_pct_i + \gamma_4$

classi medie → $TestScore = \gamma_1 \cdot el\_pct_i + \gamma_3$

classi piccole → $TestScore = \gamma_1 \cdot el\_pct_i + \gamma_2$

Effetti differenziali:

- $\beta_2 = \gamma_2 - \gamma_4$ rappresenta la differenza nel punteggio medio del test, dei distretti con classi piccole rispetto ai distretti con classi grandi, a parità di studenti non madrelingua
- $\beta_3 = \gamma_3 - \gamma_4$ rappresenta la differenza nel punteggio medio del test, dei distretti con classi medie rispetto ai distretti con classi grandi, a parità di studenti non madrelingua
- $\beta_2 - \beta_3 = \gamma_2 - \gamma_3$ rappresenta la differenza nel punteggio medio del test, dei distretti con classi piccole rispetto ai distretti con classi medie, a parità di studenti non madrelingua
- Inoltre $\gamma_1 = \beta_1$ e $\beta_0 = \gamma_4$

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 lo segnala – bloccandosi, o mostrando un messaggio di errore, o eliminando 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.

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 distribuzione degli stimatori OLS nella regressione multipla

Concetto chiave 5.5: *la distribuzione di $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ in grandi campioni*

Se valgono le assunzioni dei minimi quadrati (concetto chiave 5.4), gli stimatori OLS $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ sono, in grandi campioni, congiuntamente distribuiti secondo una normale e ogni $\hat{\beta}_j$ si distribuisce secondo una $N(\beta_j, \sigma^2_{\hat{\beta}_j})$, con $j = 0, \ldots, k$.

Inoltre gli stimatori OLS sono consistenti.

Verifica di ipotesi e intervalli di confidenza per un singolo coefficiente

- 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.
- $\frac{\hat{\beta}_j - E(\hat{\beta}_j)}{\hat{\sigma}_{\hat{\beta}_j}}$ è approssimativamente distribuita come N(0,1)
- Perciò le ipotesi su un singolo coefficiente β_j possono essere verificate mediante la consueta statistica t e gli intervalli di confidenza costruiti come $\hat{\beta}_j \pm z_{\alpha/2}\hat{\sigma}_{\hat{\beta}_j}$

***Esempio*: dati sulle dimensioni delle classi in California**

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

- Il coefficiente di *STR* in è l'effetto su *TestScore* di un aumento unitario in *STR*, mantenendo costante la percentuale di studenti non di madrelingua nel distretto
- L'intervallo di confidenza al 95% per il coefficiente di *STR* è

$$\hat{\beta}_j \pm z_{\alpha/2}\hat{\sigma}_{\hat{\beta}_j} = 1{,}10 \pm 1{,}96 \cdot 0{,}43 = [-1{,}95\,;-0{,}26]$$

- La statistica t per verificare l'ipotesi $\beta_{STR} = 0$ è $t = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} = \frac{-1{,}10}{0{,}43} = -2{,}54$, perciò rifiutiamo l'ipotesi nulla al livello di significatività del 5%

Verifica di ipotesi congiunte

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: $\beta_1 = 0$ ***e*** $\beta_2 = 0$

vs. H_1: ***o*** $\beta_1 \neq 0$ ***o*** $\beta_2 \neq 0$ ***o entrambi***

Verifica di restrizioni multiple su più coefficienti

- 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%)!

Due soluzioni:

- Utilizzare un valore critico diverso in questa procedura – non 1,96 ("metodo Bonferroni", utilizzato raramente nella pratica)
- Utilizzare una statistica test diversa studiata per verificare congiuntamente $\beta_1 = 0$ ***e*** $\beta_2 = 0$: la statistica *F*

La statistica *F* in condizioni di omoschedasticità pura

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

• 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 significativamente meglio, si rifiuta l'ipotesi nulla

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

<u>Regressione senza vincolo (sotto H_1):</u>

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

<u>Regressione vincolata (ossia, sotto H_0):</u>

$TestScore_i = \beta_0 + \beta_3 PctEL_i + u_i$

- Il numero di vincoli sotto H_0 è $q = 2$
- L'adattamento risulterà migliore (R^2 sarà maggiore) nella regressione non vincolata
- Di quanto dovrà aumentare R^2 affinché almeno uno dei coefficienti di *Expn* e *PctEL* sia giudicato statisticamente significativo?

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

- La statistica *F* classica rifiuta l'ipotesi nulla quando aggiungendo le due variabili, R^2 aumenta significativamente – vale a dire, quando aggiungendo le due variabili si migliora l'adattamento della regressione in maniera significativa
- Se gli errori sono omoschedastici, la statistica *F* classica ha una distribuzione in grandi campioni che è $F_{q,n-k_{unrestricted}-1}$
- La statistica *F* classica viene anche utilizzata per verificare l'ipotesi congiunta che tutti i coefficienti dei regressori siano pari a zero, contro l'alternativa che almeno uno sia diverso da zero.
- Se la statistica *F* classica viene utilizzata per verificare l'ipotesi su un unico coefficiente, il test fornisce gli stessi risultati del t-test.

Esempio:

<u>Regressione vincolata</u>:

$$\widehat{TestScore} = \underset{(1,0)}{644,7} - \underset{(0,032)}{0,671} \times PctEL_i, \quad R^2 = 0,4149, \quad SER = 14,59$$

<u>Regressione non vincolata</u>:

$$\widehat{TestScore} = \underset{(15,5)}{649,6} - \underset{(0,032)}{0,656} \times PctEL_i - \underset{(0,482)}{0,286} \times STR + \underset{(1,581)}{3,878} \times Expn_i,$$

$$R^2 = 0,4366, \quad SER = 14,35$$

Quindi

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)} = \frac{(0,4366 - 0,4149)/2}{(1 - 0,4366)/(420 - 3 - 1)} = \mathbf{8,01}$$

F_α=3,02

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

Verifica di restrizioni singole che coinvolgono coefficienti multipli

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

Considerate l'ipotesi nulla e le ipotesi alternative,

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

Questa ipotesi nulla impone una *singola* restrizione ($q = 1$) su coefficienti *multipli* – non si tratta di ipotesi congiunte con restrizioni multiple (confrontate con H_0: $\beta_1 = 0$ e $\beta_2 = 0$).

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 R, consentono di verificare le restrizioni utilizzando direttamente coefficienti multipli

Metodo 1: Riorganizzare ("trasformare") la regressione

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

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

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

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

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

oppure

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

Dove

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

- 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 H_0: $\gamma_1 = 0$ vs. H_1: $\gamma_1 \neq 0$ 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 R, per verificare $\beta_1 = \beta_2$ vs. $\beta_1 \neq \beta_2$ (bilaterale):

```
reg<-lm(testscr ~ str+el_pct+expn_stu)
linearHypothesis(reg,"str=expn_stu",vcov =
  vcovHC(reg, "HC1"))
```

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

Specificazione del modello e presentazione dei risultati

Anche in una regressione multipla possono sussistere le condizioni perchè vi sia distorsione da variabili omesse.

Concetto chiave 5.9: *la distorsione da variabile omessa nella regressione multipla*

La distorsione da variabile omessa è la distorsione dello stimatore OLS che nasce quando uno o più tra i regressori inclusi sono correlati con una variabile omessa. Perché si abbia distorsione da variabile omessa, debbono valere due condizioni:

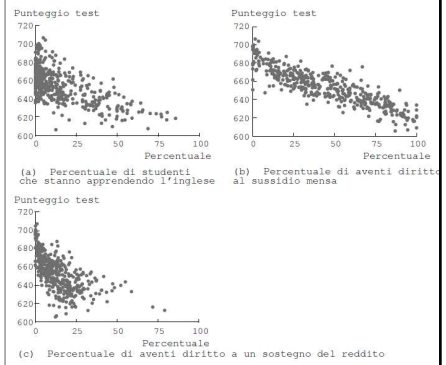
1. almeno uno dei regressori inclusi deve essere correlato con la variabile omessa;
2. la variabile omessa deve essere una determinante della variabile dipendente Y.

- Quando sono disponibili dati sulla variabile omessa, la soluzione al problema della distorsione è quella di includere la variabile omessa nella regressione
  - Solitamente non sono disponibili dati tutti questi fattori omessi (per esempio condizioni economiche delle famiglie del distretto).
    In questo caso si possono includere "variabili di controllo" correlate a questi fattori causali omessi, ma che di per sé non sono causali (per esempio, percentuale di studenti che hanno diritto al sussidio mensa totale o parziale)
- In pratica, si dovrebbe partire da una **specificazione di base** del modello. Tale specificazione dovrebbe contenere le variabili di interesse primario e le variabili di controllo suggerite dall'esperienza e dalla teoria economica.
- Si sviluppa quindi un elenco di possibili **specificazioni alternative,** utilizzando insiemi alternativi di regressori.
- Se le stime dei coefficienti di interesse sono numericamente simili nelle diverse specificazioni alternative, questo costituisce evidenza del fatto che le stime sono affidabili
- Viceversa se le stime dei coefficienti d'interesse cambiano sostanzialmente tra le varie specificazioni, ciò è spesso sintomo del fatto che la specificazione originale soffre di distorsione da variabile omessa

Esempio

- Nell'esempio sui distretti della California la variabile d'interesse per la quale si vuole misurare l'effetto causale è STR.
- Molti fattori possono influenzare il punteggio medio del test e alcuni sono correlati con il rapporto studenti-insegnanti. Ometterli dalla regressione determinerà distorsione da variabile omessa.
- Esempio: frazione di studenti non di madrelingua inglese, condizioni economiche degli studenti…

Figura 5.2: *grafici a nuvola dei punteggi del test su tre caratteristiche dello studente*



I grafici a nuvola mostrano una relazione negativa tra punteggi del 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).

Tabella 5.2: *risultati delle regressioni dei punteggi del test sul rapporto studenti-insegnanti e su altre variabili che controllano per le caratteristiche degli studenti usando i dati relativi ai distretti scolasti elementari della California*

Variabile dipendente: media dei punteggi del 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
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, descritti nell'appendice 4.1, della California. Gli errori standard sono mostrati in parentesi sotto i coefficienti. Il coefficiente è significativo al livello *5% o **1% utilizzando un test bilaterale.*

La tabella riporta tutte le informazioni d'interesse sui modelli stimati.

Es: $\widehat{TestScore} = \underset{(10,4)}{698,9} - \underset{(0,52)}{2,28} \times STR,\ \bar{R}^2 = 0,049,\ SER = 19,26,\ n = 420.$

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.