Econometria I

Corso di laurea magistrale in Finanza e Assicurazioni

Slides per lezioni

secondo semestre 2020

Marco Ventura

Richiami di probabilità

cap 2 SW

Libro di testo:

Stock H.J., Watson M.W.:

Introduzione all'Econometria, quarta edizione Pearson eds. (edizione italiana a cura di F. Peracchi)

Breve panoramica del corso

- L'economia suggerisce importanti relazioni, spesso con implicazioni politiche, ma praticamente mai fornisce dimensioni quantitative.
  - Qual è l'effetto *quantitativo* della riduzione delle dimensioni delle classi sui risultati degli studenti?
  - In che modo un anno in più di istruzione può influire sul reddito?
  - Qual è l'elasticità al prezzo delle sigarette?
  - Qual è l'effetto sulla crescita del prodotto interno lordo di un aumento di 1 punto percentuale nei tassi di interesse stabilito dalla Fed?
  - Qual è l'effetto sui prezzi delle case dei miglioramenti di tipo ambientale?

Questo corso tratta dell'uso dei dati per misurare relazioni ed anche effetti causali

- Idealmente vorremmo un esperimento
  - Quale sarebbe un esperimento per stimare l'effetto della dimensione delle classi sui punteggi nei test standardizzati?
- Ma quasi sempre abbiamo a disposizione soltanto dati osservazionali (non sperimentali).
  - rendimenti dell'istruzione
  - prezzi delle sigarette
  - politica monetaria
- La maggior parte del corso affronta le difficoltà che derivano dall'uso di dati non sperimentali per stimare effetti causali
  - effetti perturbativi (fattori omessi)
  - causalità simultanea
  - "la correlazione non implica causalità"

In questo corso:

- Prenderete confidenza con i principali metodi di analisi econometrici,
- imparerete a valutare l'analisi di regressione effettuata da altri – questo significa che sarete in grado di leggere e comprendere articoli su temi economici di carattere empirico;
- farete un esperienza pratica con l'analisi di regressione nelle serie di esercizi.
- vi focalizzerete sulle applicazioni – si ricorre alla teoria solo ove necessario per comprendere i motivi alla base dei metodi;

SE IL TEMPO LO CONSENTE:

- apprenderete metodi per stimare effetti causali usando dati non sperimentali;
- apprenderete l'uso di alcuni strumenti che possono essere impiegati per altri scopi, per esempio la previsione utilizzando serie di dati temporali;

Richiami di probabilità e statistica (Capitoli 2, 3)

- **Problema empirico:** Dimensione della classe e risultato dell'istruzione
  - Domanda: qual è l'effetto sui punteggi nei test (o su un'altra misura di risultato) della riduzione della dimensione delle classi di uno studente per classe? E di 8 studenti per classe?
  - Dobbiamo utilizzare i dati per rispondere (esiste un modo per rispondere a questa domanda *senza* dati?)

I dati dei punteggi nei test della California

Tutti i distretti scolastici K-6 e K-8 della California
(*n* = 420)

Variabili:

- Punteggi nei test del quinto anno (*Stanford-9 achievement test*, combinazione di matematica e lettura), media del distretto
- Rapporto studenti/insegnanti (STR) = numero di studenti nel distretto diviso per numero di insegnanti a tempo pieno equivalente

Primo sguardo ai dati: *(dovreste già sapere come interpretare questa tabella)*

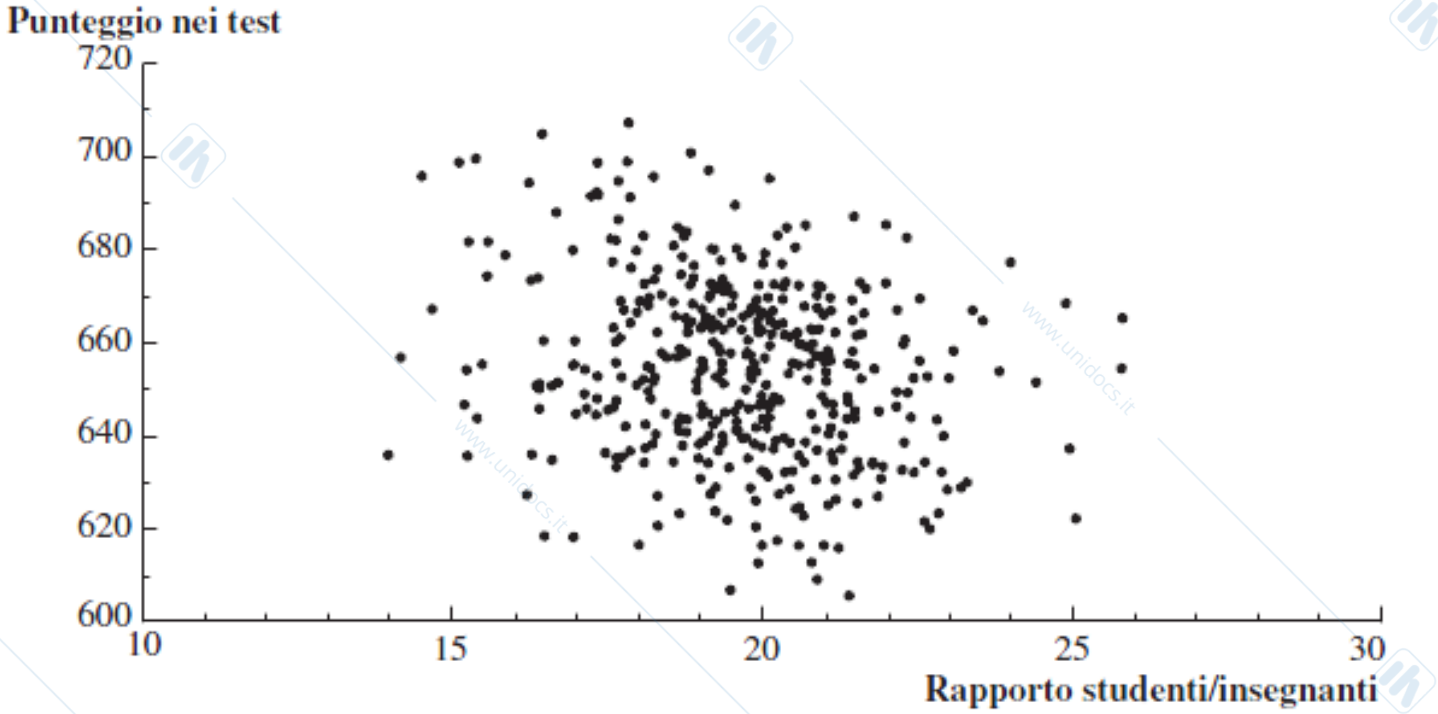
Tabella 4.1 Sintesi della distribuzione del rapporto studenti/insegnanti e del punteggio nei test relativa al quinto grado d'istruzione (quinta elementare) per 420 distretti K-8 in California nel 1998.

			Percentile						
	Media	**Deviazione standard**	10%	25%	40%	50% (mediana)	60%	75%	90%
Rapporto studenti/insegnanti	19,6	1,9	17,3	18,6	19,3	19,7	20,1	20,9	21,9
Punteggio nei test	654,2	19,1	630,4	640,0	649,1	654,5	659,4	666,7	679,1

Questa tabella non ci dice nulla sulla relazione tra punteggi nei test e *STR*.

I distretti con classi più piccole ottengono punteggi più elevati nei test?

Diagramma a nuvola di punteggio nei test e STS



Che cosa mostra questa figura?

Dobbiamo ottenere evidenza numerica che indichi se i distretti con basso STR hanno punteggi nei test più alti – ma come?

1. Confrontare i punteggi nei test nei distretti con basso STR a quelli con alto STR ("***stima***")
2. Sottoporre a verifica l'ipotesi "nulla" che i punteggi medi nei test nei due tipi di distretti siano gli stessi, contro l'ipotesi "alternativa" che siano diversi ("***test di ipotesi***")
3. Stimare un intervallo per la differenza nei punteggi medi nei test, nei distretti con alto vs basso STR ("***intervallo di confidenza***")

Analisi dei dati iniziali: confrontare i distretti con dimensioni delle classi "piccole" (STR < 20) e "grandi" (STR ≥ 20) :

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

1. ***Stima*** di Δ = differenza tra medie dei gruppi
2. ***Verifica dell'ipotesi che*** Δ = 0
3. Costruire un ***intervallo di confidenza*** per Δ

1. Stima

$$\bar{Y}_{piccola} - \bar{Y}_{grande} = \frac{1}{n_{piccola}} \sum_{i=1}^{n_{piccola}} Y_i - \frac{1}{n_{grande}} \sum_{i=1}^{n_{grande}} Y_i$$

= 657,4 – 650,0

= 7,4

È una differenza da considerare grande nel mondo reale?

- Deviazione standard tra i distretti = 19,1
- Differenza tra 60-esimo and 75-esimo percentili della distribuzione dei punteggi nei test = 667,6 – 659,4 = 8,2
- È una differenza sufficientemente grande da risultare importante per discussioni sulla riforma della scuola, per i genitori o per un comitato scolastico?

2. Verifica di ipotesi

Test di differenza tra medie: calcolare la *statistica-t*,

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)}$$

- dove $SE(\bar{Y}_s - \bar{Y}_l)$ è l'errore standard di $\bar{Y}_s - \bar{Y}_l$,

i pedici *s* e *l* indicano distretti con STR "small" e "large", e

$$s_s^2 = \frac{1}{n_s - 1}\sum_{i=1}^{n_s}(Y_i - \bar{Y}_s)^2$$

Calcolare la statistica-*t* per la differenza tra medie

Dim	$\bar{Y}$	s_Y	n
piccola	657,4	19,4	238
grande	650,0	17,9	182

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

|t| > 1,96, perciò si rifiuta l'ipotesi nulla che le due medie coincidano (al livello di significatività del 5%)

3. Intervallo di confidenza

Un intervallo di confidenza al 95% per la differenza tra medie è

$$(\bar{Y}_s - \bar{Y}_l) \pm 1{,}96 \times SE(\bar{Y}_s - \bar{Y}_l)$$

$$= 7{,}4 \pm 1{,}96 \times 1{,}83 = (3{,}8 \; ; \; 10{,}98)$$

Due affermazioni equivalenti:

1. L'intervallo di confidenza al 95% per Δ non include 0;
2. L'ipotesi che Δ = 0 è rifiutata al livello del 5%.

E ora…

- I meccanismi di stima, verifica di ipotesi e intervalli di confidenza dovrebbero risultare familiari
- Questi concetti si estendono direttamente a regressione e relative varianti
- Prima di passare alla regressione, tuttavia, rivedremo alcuni elementi della teoria alla base di stima, verifica di ipotesi e intervalli di confidenza:
  - Perché queste procedure funzionano, e perché utilizzare proprio queste invece di altre?
  - Rivedremo i fondamenti teorici di statistica ed econometria

Richiami di teoria statistica

1. **Quadro di riferimento probabilistico per l'inferenza statistica**
2. Stima
3. Verifica
4. Intervalli di confidenza

Quadro di riferimento probabilistico per l'inferenza statistica

a) Popolazione, variabile casuale e distribuzione
b) Momenti di una distribuzione (media, varianza, deviazione standard, covarianza, correlazione)
c) Distribuzione condizionata e media condizionata
d) Distribuzione di un campione di dati estratto a caso da una popolazione: Y_1, …, Y_n

(a) Popolazione, variabile casuale e distribuzione

Popolazione

- Il gruppo o l'insieme di tutte le possibili unità di interesse (distretti scolastici)
- Considereremo le popolazioni infinitamente grandi (∞ è un'approssimazione di "molto grande")

Variabile casuale Y

- Rappresentazione numerica di un risultato casuale (punteggio medio nei test del distretto, STR del distretto)

Distribuzione di Y

- Le probabilità di diversi valori di Y che si verificano nella popolazione, per esempio $\Pr[Y = 650]$ (quando Y è discreta)

- oppure: le probabilità di insiemi di questi valori, per esempio $\Pr[640 \leq Y \leq 660]$ (quando Y è continua).

(b) Momenti di una distribuzione: media, varianza, deviazione standard, covarianza, correlazione

media = valore atteso (aspettativa) di Y

= $E(Y)$

= μ_Y

= valore medio di Y su ripetute realizzazioni di Y

varianza = $E(Y - \mu_Y)^2$

= σ_Y^2

= misura della dispersione quadratica della distribuzione

deviazione standard = $\sqrt{\text{varianza}} = \sigma_Y$

Momenti (continua)

asimmestria = $\dfrac{E\left[\left(Y-\mu_Y\right)^3\right]}{\sigma_Y^3}$

= misura di asimmetria di una distribuzione

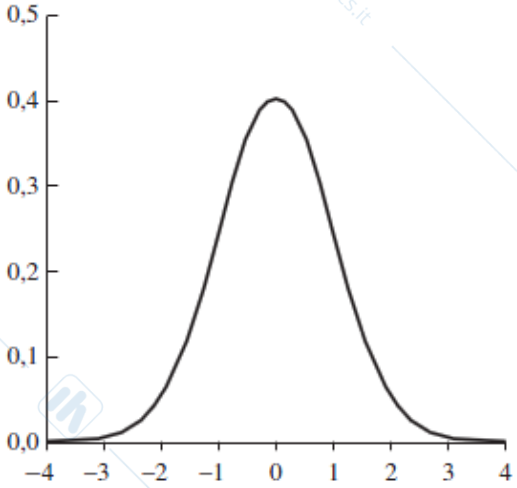
- *asimmetria* = 0: la distribuzione è simmetrica
- *assimmetria* > (<) 0: la distribuzione ha una coda lunga destra (sinistra)

curtosi = $\dfrac{E\left[\left(Y-\mu_Y\right)^4\right]}{\sigma_Y^4}$

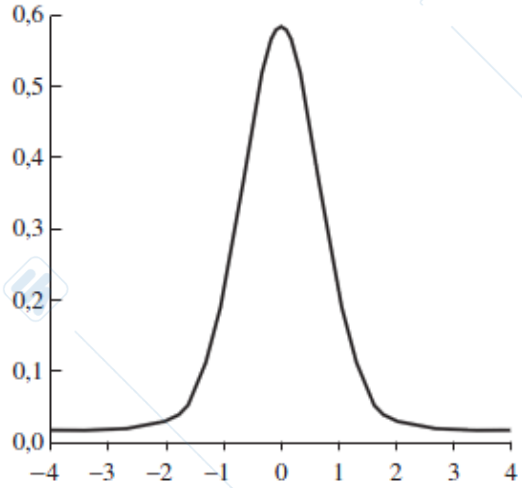
= misura di massa nelle code

= misura di probabilità di valori grandi

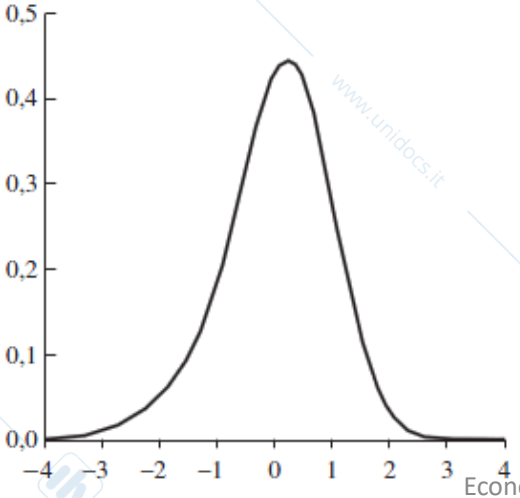
- *curtosi* = 3: distribuzione normale
- *assimmetria* > (<) 3: code pesanti "***leptocurtica***" **("platicurtica")**



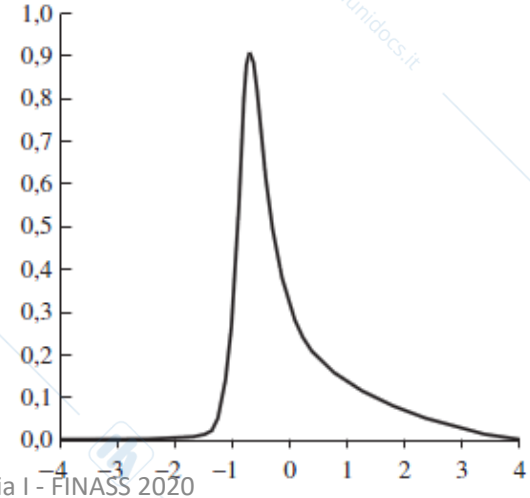
(a) Asimmetria = 0, Curtosi = 3



(b) Asimmetria = 0, Curtosi = 20



(c) Asimmetria = –0,1, Curtosi = 5



(d) Asimmetria = 0,6, Curtosi = 5

2 variabili casuali: distribuzioni congiunte e covarianza

- Le variabili casuali X e Z hanno una ***distribuzione congiunta***
- La ***covarianza*** tra X e Z è
  - $\text{cov}(X,Z) = E[(X - \mu_X)(Z - \mu_Z)] = \sigma_{XZ}$
- La covarianza è una misura dell'associazione lineare tra X e Z; le sue unità sono unità di X ×unità di Z
- $\text{cov}(X,Z) > 0$ significa una relazione positiva tra X e Z
- Se X e Z sono indipendentemente distribuite, allora $\text{cov}(X,Z) = 0$ (**ma non vale il vice versa**!! **Eccezione la N**)
- La covarianza di una variabile casuale con se stessa è la sua varianza:
  $\text{cov}(X,X) = E[(X - \mu_X)(X - \mu_X)] = E[(X - \mu_X)^2] = \sigma_X^2$

La covarianza tra *Punteggio nei test* e *Rapporto studenti/insegnanti* è negativa:

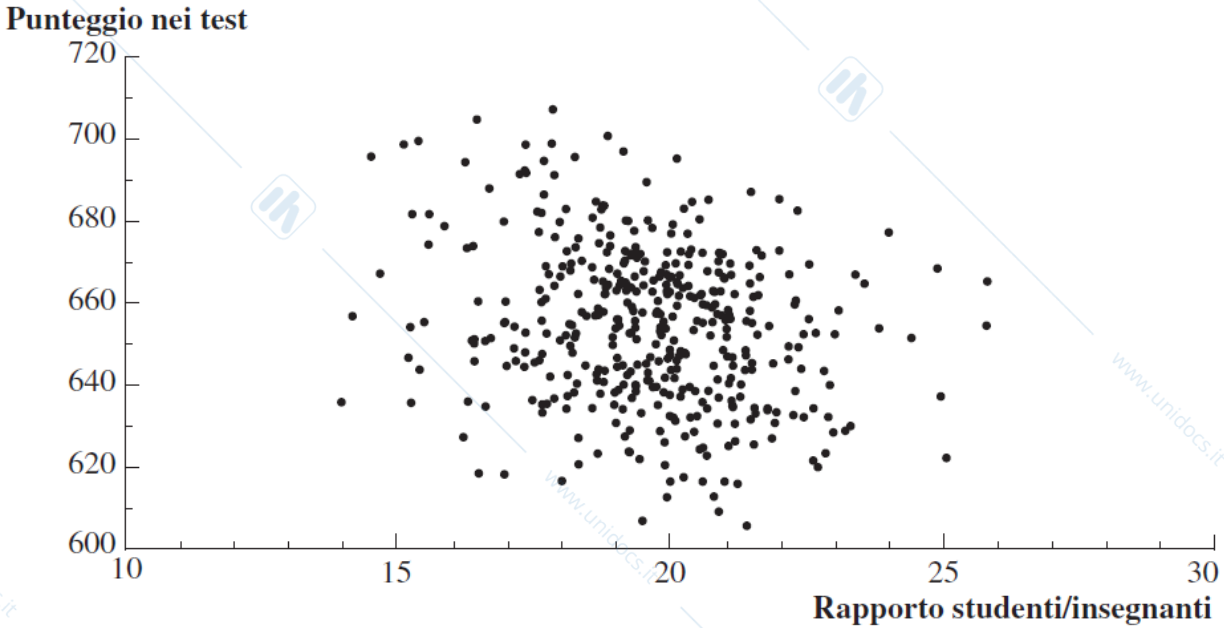


Figura 4.2

Diagramma a nuvola del punteggio nei test e del rapporto studenti/insegnanti (dati relativi ai distretti scolastici della California).

Dati per i 420 distretti scolastici della California. C'è una debole relazione negativa tra il rapporto studenti/insegnanti e il punteggio nei test: la correlazione campionaria è pari a –0,23.

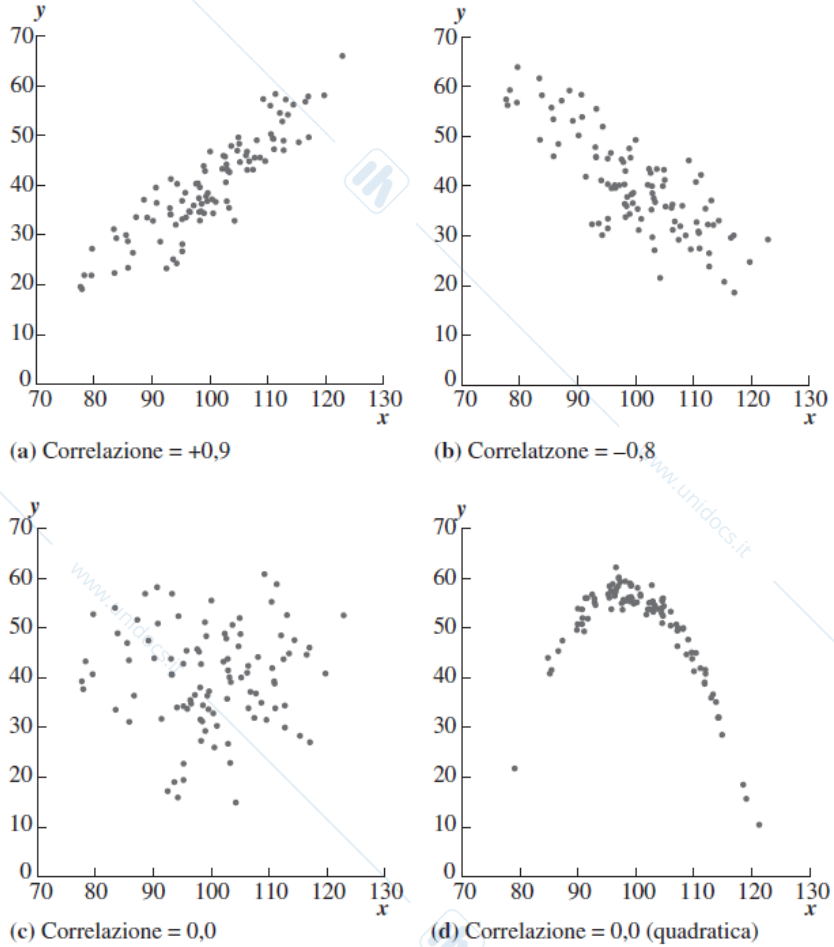
E così la ***correlazione***…

Il ***coefficiente di correlazione*** è definito in termini di covarianza:

$$\text{corr}(X,Z) = \frac{\text{cov}(X,Z)}{\sqrt{\text{var}(X)\,\text{var}(Z)}} = \frac{\sigma_{XZ}}{\sigma_X \sigma_Z} = r_{XZ}$$

- $-1 \leq \text{corr}(X,Z) \leq 1$
- $\text{corr}(X,Z) = 1$ significa associazione lineare positiva perfetta
- $\text{corr}(X,Z) = -1$ significa associazione lineare negativa perfetta
- $\text{corr}(X,Z) = 0$ significa che non c'è associazione lineare

Il coefficiente di correlazione misura l'associazione lineare



(a) Correlazione = +0,9

(b) Correlatzone = –0,8

(c) Correlazione = 0,0

(d) Correlazione = 0,0 (quadratica)

(c) Distribuzione condizionata e media condizionata

Distribuzione condizionata

- La distribuzione di *Y* dato il valore (o i valori) di un'altra variabile casuale *X*
- Es: la distribuzione dei punteggi nei test dato STR < 20

Valore atteso condizionato e momento condizionato

- *media condizionata* = media della distribuzione condizionata
  - = $E(Y|X = x)$ (***concetto e notazione importanti***)

- *varianza condizionata* = varianza della distribuzione condizionata

- *Esempio*: $E(Punteggio\ test|STR < 20)$ = media dei punteggi nei test tra i distretti con dimensioni delle classi piccole

La differenza in media è la differenza tra le medie di due distribuzioni condizionate:

Media condizionata (continua)

$\Delta = E(\textit{Punteggio test}|STR < 20) - E(\textit{Punteggio test}|STR \geq 20)$

Altri esempi di media condizionata:

- Salari di tutte le lavoratrici femmine (Y = salari, X = genere)
- Tasso di mortalità di pazienti che ricevono una cura sperimentale (Y = vivo/morto; X = trattato/non trattato)
- Se $E(X|Z)$ = costante, allora corr(X,Z) = 0 (tuttavia non vale necessariamente il vice versa)

La media condizionata è un termine utilizzato per il concetto familiare di media di gruppo

Distribuzioni marginali e condizionate: compiti a casa

Un gruppeto di 2 studenti prepari la presentazione della tavola 2.2 del libro di testo soffermandosi sul calcolo e il concetto di **distribuzione marginale**

Un gruppeto di 2 max 3 studenti prepari la presentazione della tavola 2.3 del libro di testo soffermandosi sul calcolo e il concetto di **media e varianza condizionata**.

(d) Distribuzione di un campione di dati estratto a caso da una popolazione: $Y_1,\ldots, Y_n$

Assumeremo un campionamento casuale semplice

- Scegliere a caso un individuo (distretto, unità) dalla popolazione

Casualità e dati

- Prima della selezione del campione, il valore di Y è casuale perché l'individuo selezionato è casuale
- Una volta selezionato l'individuo e osservato il valore di Y, Y è soltanto un numero – non casuale
- Il dataset è (Y_1, Y_2,…, Y_n), dove Y_i = valore di Y per l'i-esimo individuo (distrettto, unità) del campione

Distribuzione di $Y_1,\ldots,Y_n$ sotto campionamento casuale semplice

- Poiché gli individui n. 1 e 2 sono selezionati a caso, il valore di Y_1 non contiene informazioni per Y_2. Quindi:
  - Y_1 e Y_2 sono ***indipendentemente distribuiti***
  - Y_1 e Y_2 provengono dalla stessa distribuzione, cioè Y_1, Y_2 sono ***identicamente distribuiti***
  - Ovvero, sotto campionamento casuale semplice, Y_1 e Y_2 sono indipendentemente e identicamente distribuiti (***i.i.d.***).
  - Più in generale, sotto campionamento casuale semplice, $\{Y_i\}$, $i = 1,\ldots, n$, sono i.i.d.

Questo quadro consente rigorose inferenze statistiche sui momenti di distribuzione utilizzando un campione di dati tratto dalla popolazione stessa…

1. Quadro probabilistico per inferenza statistica
2. **Stima**
3. Verifica
4. Intervalli di confidenza

Stima

$\bar{Y}$ è lo stimatore naturale della media. Ma:

a) quali sono le proprietà di $\bar{Y}$?

b) Perché dovremmo usare $\bar{Y}$ anziché un altro stimatore?
   - Y_1 (prima osservazione)
   - forse pesi non uniformi – non media semplice $\bar{Y}$
   - mediana(Y_1,…, Y_n)

Il punto di partenza è la distribuzione campionaria di $\bar{Y}$

Econometria I

Corso di laurea magistrale in Finanza e Assicurazioni

Slides per lezioni

secondo semestre 2020

Marco Ventura

Richiami di statistica

cap 2-3 SW

Questo quadro consente rigorose inferenze statistiche sui momenti di distribuzione utilizzando un campione di dati tratto dalla popolazione stessa…

1. Quadro probabilistico per inferenza statistica
2. **Stima**
3. Verifica
4. Intervalli di confidenza

Stima

$\bar{Y}$ è lo stimatore naturale della media. Ma:

a) quali sono le proprietà di $\bar{Y}$?

b) Perché dovremmo usare $\bar{Y}$ anziché un altro stimatore?
   - Y_1 (prima osservazione)
   - forse pesi non uniformi – non media semplice $\bar{Y}$
   - mediana(Y_1,…, Y_n)

Il punto di partenza è la distribuzione campionaria di $\bar{Y}$

(a) La distribuzione campionaria di $\overline{Y}$ (par 2.5)

$\overline{Y}$ è una variabile casuale e le sue proprietà sono determinate dalla ***distribuzione campionaria*** di $\overline{Y}$

- Gli individui nel campione sono estratti a caso.
- Quindi i valori di $(Y_1,\ldots,Y_n)$ sono casuali
- Quindi funzioni di $(Y_1,\ldots,Y_n)$, come $\overline{Y}$, sono casuali: se si fosse estratto un campione diverso, avrebbero assunto valori differenti
- La distribuzione di $\overline{Y}$ su diversi possibili campioni di dimensione *n* si chiama ***distribuzione campionaria*** di $\overline{Y}$.
- La media e la varianza di $\overline{Y}$ sono la media e la varianza della sua distribuzione campionaria, $E(\overline{Y})$ e $\text{var}(\overline{Y})$.
- Il concetto di distribuzione campionaria è alla base di tutta l'econometria.

La distribuzione campionaria di $\overline{Y}$ (continua)

Esempio: Si supponga che *Y* assuma il valore 0 o 1 (variabile casuale di ***Bernoulli***) con la distribuzione di probabilità

Pr[*Y* = 0] = 0,22, Pr(*Y* =1) = 0,78

Allora

$E(Y) = p \times 1 + (1 - p) \times 0 = p = 0{,}78$

$\sigma_Y^2 = E[Y - E(Y)]^2 = p(1 - p)$

$= 0{,}78 \times (1 - 0{,}78) = 0{,}1716$

La distribuzione campionaria di $\overline{Y}$ dipende da *n*.

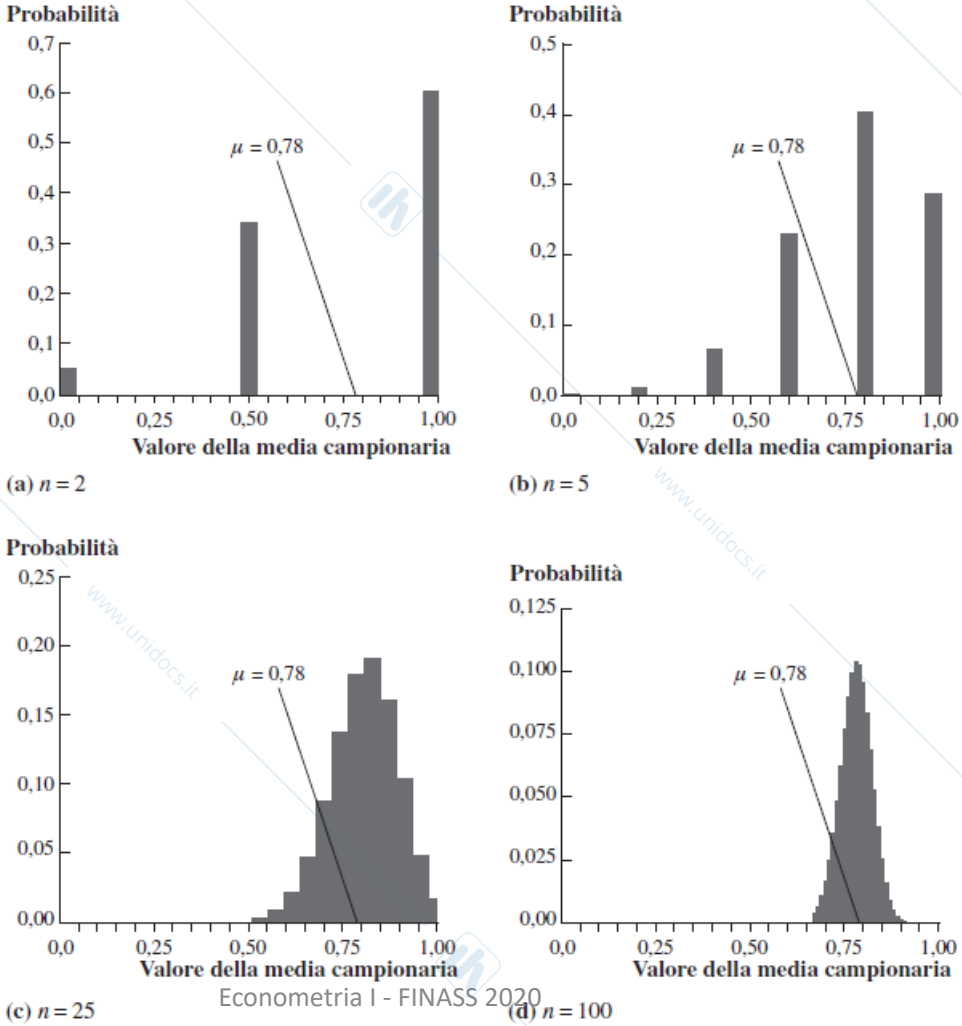
La distribuzione campionaria di $\overline{Y}$ (continua)

Si consideri *n* = 2. La distribuzione campionaria di $\overline{Y}$ è

- Pr($\overline{Y}$ = 0) = 0,22² = 0,0484 ➔ si deve verificare 2 volte che esca Y=0
- Pr($\overline{Y}$ = ½) = 2×0,22×0,78 = 0,3432 ➔ 1 volta deve uscire Y=1 e una volta Y=0. Può accadere prima 0 poi 1 o vv, per questo x2
- Pr($\overline{Y}$ = 1) = 0,78² = 0,6084 ➔ si deve verificare 2 volte che esca Y=1

Similmente per n che assume altri valori; si veda la figura seguente

Distribuzione campionaria di $\overline{Y}$ quando Y è di Bernoulli (p = 0,78):



La distribuzione campionaria di $\overline{Y}$ (continua)

La varianza della distribuzione di $\overline{Y}$ si reduce all'aumentare di n, così nella figura la distribuzione diventa più concentrate intorno a μ=0.78

Cose che vogliamo sapere sulla distribuzione campionaria:

- Qual è la media di $\overline{Y}$?
  - Se $E(\overline{Y}) = \mu = 0{,}78$, allora $\overline{Y}$ è uno stimatore ***non distorto*** di μ o corretto
- Qual è la varianza di $\overline{Y}$?
  - In che modo var($\overline{Y}$) dipende da n (famosa formula $1/n$)
- Si avvicina a μ quando n è grande?
  - Legge dei grandi numeri: $\overline{Y}$ è uno stimatore ***consistente*** di μ
- $\overline{Y} - \mu$ assume forma a campana per n grande... questo è vero in generale?
  - In effetti, $\overline{Y} - \mu$ è approssimato da una distribuzione normale per n grande (teorema limite centrale)

Media e varianza della distribuzione campionaria di $\bar{Y}$

- Caso generale – cioè, per Y_i i.i.d. da qualsiasi distribuzione, non solo di Bernoulli:
- media: $E(\bar{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \mu_Y = \mu_Y$

- Varianza: $\text{var}(\bar{Y}) = E[\bar{Y} - E(\bar{Y})]^2$

$$= E[\bar{Y} - \mu_Y]^2$$

$$= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) - \mu_Y\right]^2$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n} (Y_i - \mu_Y)\right]^2$$

Quindi var($\bar{Y}$) = $E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu_Y)\right]^2$

$$= E\left\{\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu_Y)\right]\times\left[\frac{1}{n}\sum_{j=1}^{n}(Y_j-\mu_Y)\right]\right\}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left[(Y_i-\mu_Y)(Y_j-\mu_Y)\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{cov}(Y_i,Y_j)$$

se j≠ i la cov=0 perché Y_i e Y_j sono indipendenti, quindi restano solo i termini in cui i=j cioè le var

Quindi $\text{var}(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{cov}(Y_i, Y_j)$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2$$

$$= \frac{\sigma_Y^2}{n}$$

Media e varianza della distribuzione campionaria di $\overline{Y}$ (continua)

$$E(\overline{Y}) = \mu_Y$$

$$\text{var}(\overline{Y}) = \frac{\sigma_Y^2}{n}$$

Implicazioni:

1. $\overline{Y}$ è uno stimatore non distorto di μ_Y (cioè $E(\overline{Y}) = \mu_Y$)
2. var($\overline{Y}$) è inversamente proporzionale a n
    1. la dispersione della distribuzione campionaria è proporzionale a $1/\sqrt{n}$
    2. Quindi l'incertezza campionaria associata con $\overline{Y}$ è proporzionale a $1/\sqrt{n}$ (grandi campioni, meno incertezza, ma legge con radice quadrata)

Distribuzione campionaria di $\bar{Y}$ quando *n* è grande

Per piccoli campioni, la distribuzione di $\bar{Y}$ è complicata, ma se *n* è grande, la distribuzione campionaria è semplice!

1. All'aumentare di *n*, la distribuzione di $\bar{Y}$ diventa più strettamente centrata su μ_Y (*legge dei grandi numeri*)

2. Inoltre, la distribuzione di $\bar{Y} - \mu_Y$ diventa normale (*teorema limite centrale*)

Legge dei grandi numeri:

Uno stimatore è ***consistente*** se la probabilità che ricada entro un intervallo del vero valore della popolazione tende a uno all'aumentare della dimensione del campione.

Se $(Y_1,\ldots,Y_n)$ sono i.i.d. e $\sigma_Y^2 < \infty$, allora $\overline{Y}$ è uno stimatore consistente di μ_Y, cioè

$$\Pr[|\overline{Y} - \mu_Y| < \varepsilon] \rightarrow 1 \text{ per } n \rightarrow \infty$$

che si può scrivere $\overline{Y} \xrightarrow{p} \mu_Y$

("$\overline{Y} \xrightarrow{p} \mu_Y$" significa "$\overline{Y}$ converge in probabilità a μ_Y").

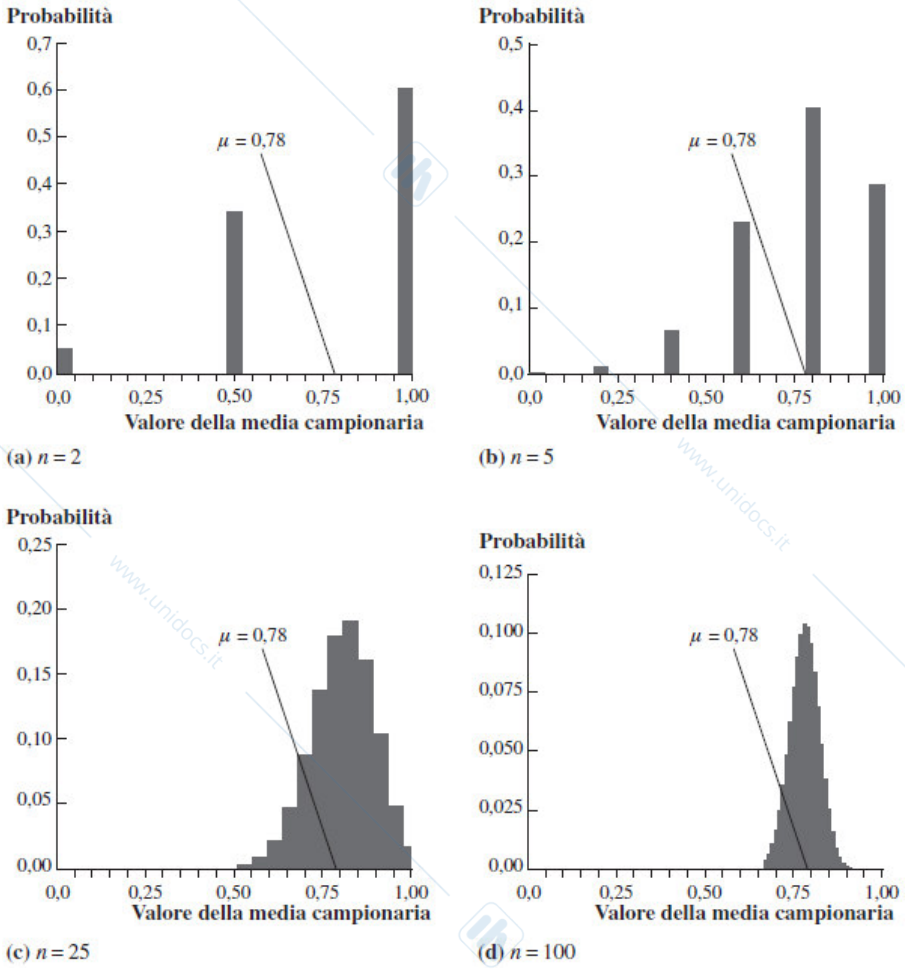
(*NB*: per $n \rightarrow \infty$, $\text{var}(\overline{Y}) = \frac{\sigma_Y^2}{n} \rightarrow 0$, il che implica che $\Pr[|\overline{Y} - \mu_Y| < \varepsilon] \rightarrow 1$.)

Teorema limite centrale (TLC):

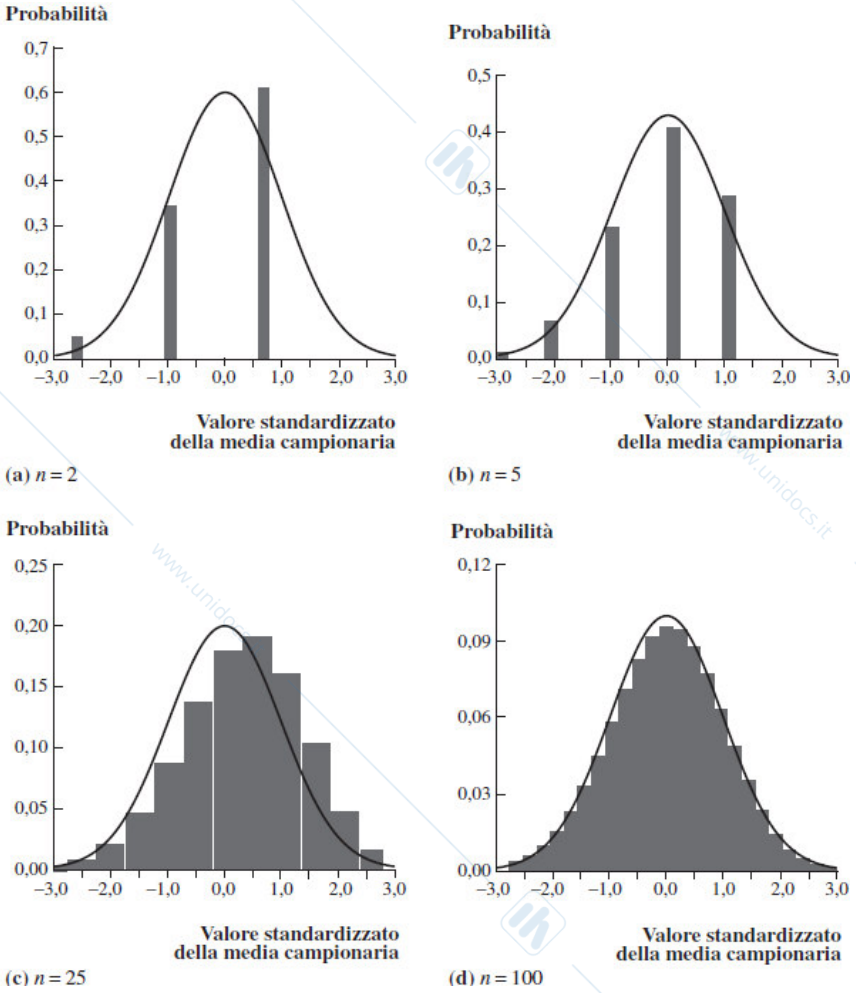
Se $(Y_1,\ldots,Y_n)$ sono i.i.d. e $\sigma_Y^2 < \infty$, allora quando n è grande la distribuzione di $\overline{Y}$ è bene approssimata da una distribuzione normale.

- $\overline{Y}$ è distribuita approssimativamente come $N(\mu_Y, \frac{\sigma_Y^2}{n})$ ("distribuzione normale con media μ_Y e varianza σ_Y^2/n")
- $\sqrt{n}\,(\overline{Y} - \mu_Y)/\sigma_Y$ è distribuita approssimativamente come $N(0,1)$ (normale standard)
- **Cioè, $\overline{Y}$"standardizzata" = $\frac{\overline{Y} - E(\overline{Y})}{\sqrt{\text{var}(\overline{Y})}} = \frac{\overline{Y} - \mu_Y}{\sigma_Y / \sqrt{n}}$è distribuita approssimativamente come $N(0,1)$**
- **Più grande è n, migliore è l'approssimazione.**

Distribuzione campionaria di $\overline{Y}$ quando Y è di Bernoulli, $p = 0{,}78$:



Stesso esempio: distribuzione campionaria di $\frac{\bar{Y} - E(\bar{Y})}{\sqrt{\text{var}(\bar{Y})}}$



Riepilogo: distribuzione campionaria di $\bar{Y}$

Per $Y_1,\ldots,Y_n$ i.i.d. con $0 < \sigma_Y^2 < \infty$,

- La distribuzione campionaria esatta (campione finito) di $\bar{Y}$ ha media μ_Y ("$\bar{Y}$ è uno stimatore non distorto di μ_Y") e varianza σ_Y^2 /*n*

- Al di là di media e varianza, la distribuzione esatta di $\bar{Y}$ è complessa e dipende dalla distribuzione di *Y* (la distribuzione della popolazione)
- Quando *n* è grande, la distribuzione campionaria si semplifica:
  - $\bar{Y} \xrightarrow{p} \mu Y$ (Legge dei grandi numeri)
  - $\frac{\bar{Y} - E(\bar{Y})}{\sqrt{\text{var}(\bar{Y})}}$ è approssimata da $N(0,1)$ (TLC)

(b) Perché usare $\bar{Y}$ per stimare μ_Y?

- $\bar{Y}$ è non distorto: $E(\bar{Y}) = \mu_Y$
- $\bar{Y}$ è consistente: $\bar{Y} \xrightarrow{p} \mu_Y$
- $\bar{Y}$ è lo stimatore "dei minimi quadrati" di μ_Y; $\bar{Y}$ risolve

$$\min_m \sum_{i=1}^{n} (Y_i - m)^2$$

quindi $\bar{Y}$ minimizza la somma dei "residui" quadrati

(cfr. anche Appendice 3.2)

$$\frac{d}{dm}\sum_{i=1}^{n}(Y_i - m)^2 = \sum_{i=1}^{n}\frac{d}{dm}(Y_i - m)^2 = -2\sum_{i=1}^{n}(Y_i - m)$$

Si pone la derivata a zero e si denota il valore ottimale di m con $\hat{m}$:

$$\sum_{i=1}^{n} Y = \sum_{i=1}^{n} \hat{m} \Rightarrow \quad \sum_{i=1}^{n} Y = n\hat{m} \quad \Rightarrow \quad \hat{m} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Perché usare $\bar{Y}$ per stimare μ_Y *(continua)*

- $\bar{Y}$ ha una varianza minore di tutti gli altri *stimatori lineari non distorti*: si consideri lo stimatore $\hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n} a_i Y_i$, dove gli $\{a_i\}$ sono tali che $\hat{\mu}_Y$ è non distorto; allora var($\bar{Y}$) ≤ var($\hat{\mu}_Y$)

  (dimostrazione: Capitolo 17 del volume stampato)

- $\bar{Y}$ non è l'unico stimatore di μ_Y – vi viene in mente un caso in cui potrebbe essere preferibile utilizzare la mediana?

1. Quadro di riferimento probabilistico per l'inferenza statistica
2. Stima
3. **Verifica di ipotesi**
4. Intervalli di confidenza

Verifica di ipotesi (par 3.2)

Il problema della ***verifica di ipotesi*** (per la media):

prendere una decisione preliminare in base all'evidenza disponibile che un'ipotesi nulla è vera, o che è vera, invece, un'ipotesi alternativa. Cioè verificare

- H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) > \mu_{Y,0}$ (monodirezionale, >)
- H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) < \mu_{Y,0}$ (monodirezionale, <)
- H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) \neq \mu_{Y,0}$ (bidirezionale)

Terminologia per la verifica di ipotesi statistiche

P-value = probabilità di ricavare una statistica (per es. $\overline{Y}$) sfavorevole all'ipotesi nulla almeno quanto il valore effettivamente calcolato con i dati, supponendo che l'ipotesi nulla sia corretta.

Il ***livello di significatività*** di un test è una probabilità predeterminata di rifiutare in modo errato l'ipotesi nulla, quando invece è corretta.

Calcolo del p-value in base a $\overline{Y}$:

valore-*p* = $\Pr_{H_0}[|\overline{Y} - \mu_{Y,0}| > |\overline{Y}^{act} - \mu_{Y,0}|]$

Dove $\overline{Y}^{act}$ è il valore di $\overline{Y}$ effettivamente osservato (non casuale)

Calcolo del valore-p (continua)

- Per calcolare il valore-*p* è necessario conoscere la distribuzione campionaria di $\bar{Y}$, che è complessa se *n* è piccolo.
- Se *n* è grande, si può usare l'approssimazione normale (TLC):

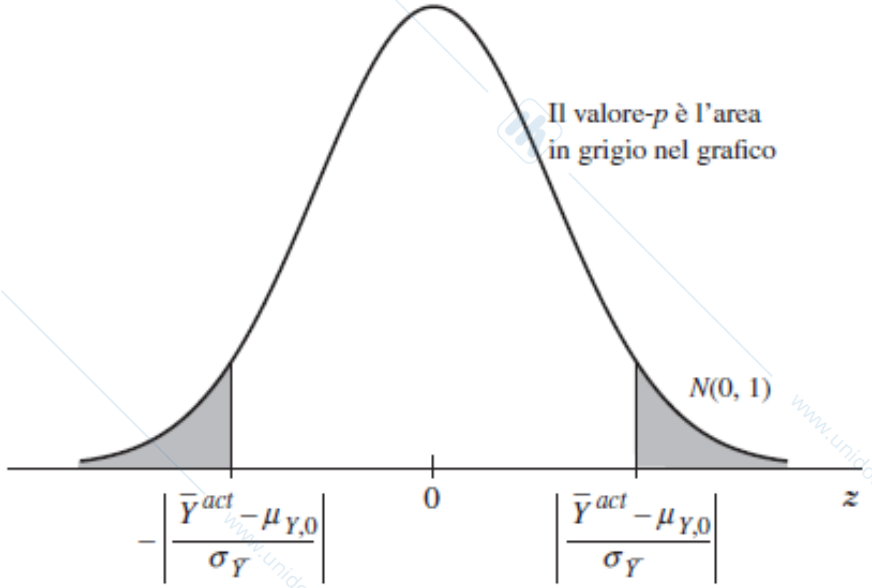
$$\text{valore-}p = \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|] \ ,$$

$$= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}\right|\right]$$

$$= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right|\right]$$

$\cong$ probabilità sotto code $N(0,1)$ sin+destra
dove $\sigma_{\bar{Y}}$ = dev. std della distribuzione di $\bar{Y} = \sigma_Y / \sqrt{n}$.

Calcolo del valore-p con σ_Y nota:



- Per *n* grande, valore-*p* = probabilità che una variabile casuale *N*(0,1) ricada al di fuori di $|(\bar{Y}^{act} - \mu_{Y,0}) / \sigma_{\bar{Y}}|$
- In pratica, $\sigma_{\bar{Y}}$ è ignota e deve essere stimata

Stimatore della varianza di Y:

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \text{“varianza campionaria di } Y\text{”}$$

Se $(Y_1,\ldots,Y_n)$ sono i.i.d. e $E(Y^4) < \infty$, allora

$$s_Y^2 \xrightarrow{p} \sigma_Y^2$$

Perché si applica la legge dei grandi numeri?

- Perché s_Y^2 è una media campionaria; cfr. Appendice 3.3
- Nota tecnica: si assume $E(Y^4) < \infty$ perché la media non è di Y_i, ma del suo quadrato; cfr. Appendice 3.3.

Calcolo del valore-p con σ_Y^2 stimato

$$\textit{valore-p} = \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]$$

$$= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}\right|\right]$$

$$\cong \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{s_Y / \sqrt{n}}\right|\right] \text{ (}n\text{ grande)}$$

Quindi

$$\text{valore-}p = \Pr_{H_0}[|t| > |t^{act}|] \quad \textbf{(} \sigma_Y^2 \textbf{ stimato)}$$

$\cong$ probabilità sotto code normali al di fuori di $|t^{act}|$

dove $t = \dfrac{\bar{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}}$ (la consueta statistica t)

Che collegamento c'è tra il valore-*p* e il livello di significatività?

- Il livello di significatività è specificato in anticipo. Per esempio, se tale livello è del 5%,
  - si rifiuta l'ipotesi nulla se $|t| \geq 1{,}96$.
  - in modo equivalente, la si rifiuta se $p \leq 0{,}05$.
  - Il valore-*p* è detto talvolta ***livello di significatività marginale***.
  - Spesso è meglio comunicare il valore-*p* che limitarsi a indicare se un test rifiuta o no – il valore-*p* contiene più informazioni di un semplice risultato "sì/no" in riferimento a un test.

A questo punto potreste chiedervi...
Che ne è della tabella-*t* e dei gradi di libertà?

Digressione: la distribuzione *t* di Student

Se Y_i, $i = 1,\ldots,n$ sono i.i.d. $N(\mu_Y, \sigma_Y^2)$, allora la statistica *t* ha la distribuzione *t* di Student *con n* – 1 gradi di libertà.

I valori critici della distribuzione *t* di Student sono elencati in tutti i libri di statistica. Ricordate la procedura?

1. Calcolare la statistica *t*
2. Calcolare i gradi di libertà, *n* – 1
3. Cercare il valore critico al 5%
4. Se la statistica *t* supera (in valore assoluto) questo valore critico, rifiutare l'ipotesi nulla.

Commenti su questa procedura e sulla distribuzione *t* di Student

1. La teoria della distribuzione *t* è stata uno dei primi trionfi della statistica matematica. È davvero sorprendente: se *Y* è i.i.d. e normale, allora è possibile conoscere la distribuzione *esatta*, *a campione finito* della statistica *t* – è la *t* di Student.

Perciò si possono costruire intervalli di confidenza (usando il valore critico *t* di Student) che hanno *esattamente* lo stesso tasso di copertura, indipendentemente dalla dimensione del campione. Questo risultato è stato molto utile in tempi in cui il "calcolatore" era una posizione lavorativa, la raccolta di dati era costosa e il numero di osservazioni si aggirava attorno alla decina.

Tuttavia...

Commenti sulla distribuzione *t* di Student *(continua)*

2. Se la dimensione del campione è modesta (qualche dozzina) o grande (centinaia o più), la differenza tra la distribuzione *t* e i valori critici N(0,1) è trascurabile. Riportiamo di seguito alcuni valori critici al 5% per test bidirezionali:

gradi di libertà (n – 1)	valore critico distribuzione *t* al 5%
10	2,23
20	2,09
30	2,04
60	2,00
∞	1,96

Commenti sulla distribuzione *t* di Student *(continua)*

3. Perciò la distribuzione *t* di Student è di interesse soltanto quando la dimensione del campione è molto piccola; ma in quel caso, affinché sia corretta, è necessario assicurarsi che la distribuzione di *Y* sia normale. Per dati economici, l'assunzione di normalità è raramente credibile. Ecco le distribuzioni di alcuni dati economici.
   - Pensate che i guadagni abbiano distribuzione normale?
   - Supponete di avere un campione di $n = 10$ osservazioni da una di queste distribuzioni – vi sentireste di usare la distribuzione *t* di Student?

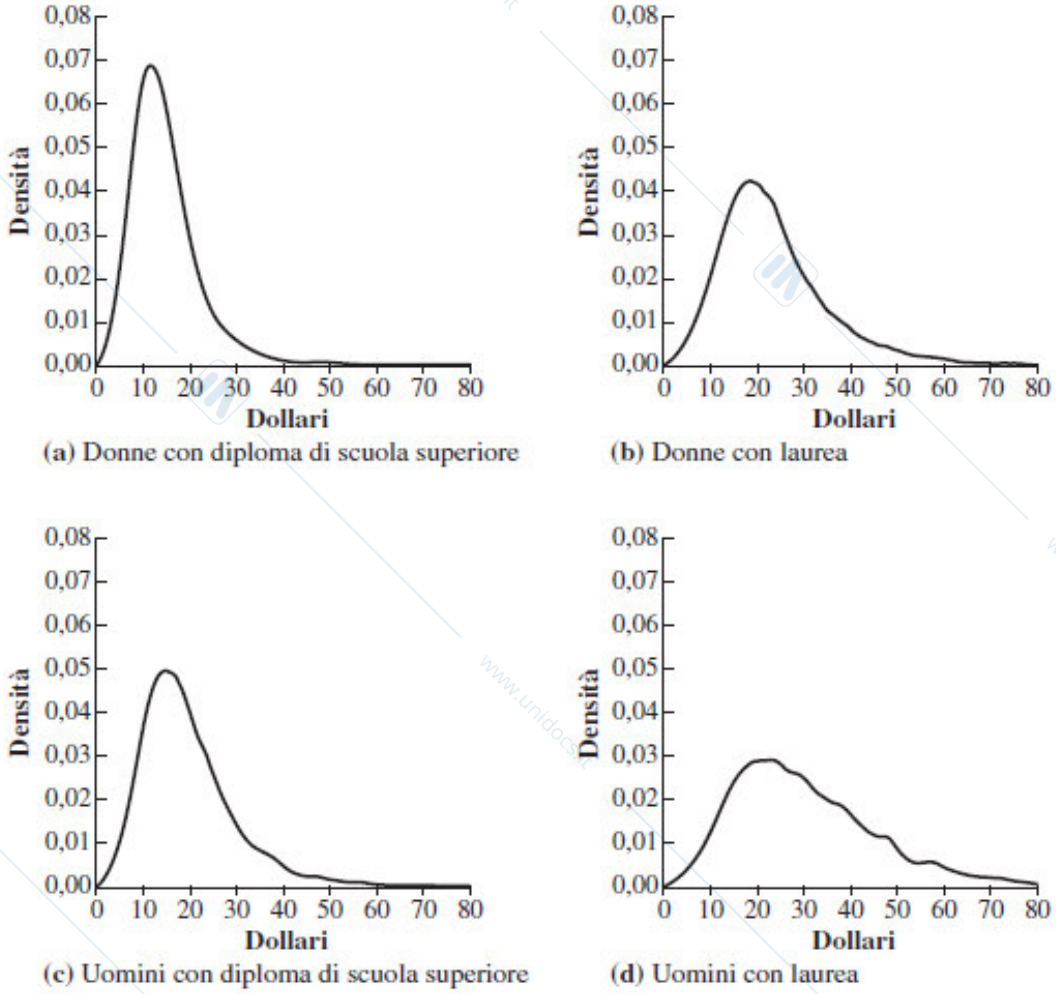


Figura 2.4

Distribuzione condizionata delle retribuzioni orarie medie dei lavoratori statunitensi a tempo pieno nel 2004, dati il livello d'istruzione e il sesso.

Le quattro distribuzioni delle retribuzioni sono per uomini e donne, per coloro che hanno solo un diploma di scuola superiore (a e c) e coloro che hanno una laurea (b e d).

Commenti sulla distribuzione *t* di Student *(continua)*

4. Forse non lo sapete. Considerate la statistica *t* che verifica l'ipotesi che due medie (gruppi *s*, *l*) siano uguali:

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)}$$

Anche se la distribuzione di *Y* nei due gruppi è normale, questa statistica non ha una distribuzione *t* di Student! Perchè lo stimatore della var nel denominatore non produce un chi2

Esiste una statistica che verifica questa ipotesi e ha distribuzione normale, la statistica *t* "a varianza aggregata" – cfr. il Paragrafo 3.6 del volume stampato – tuttavia essa è valida soltanto se le varianze delle distribuzioni normali sono le stesse nei due gruppi. Pensate che questo sia vero, per esempio, per i salari di uomini vs donne?

Solitamente l'ipotesi di uguaglianza della varianza tra due gruppi non è verificata, così nella pratica inferenza sulla differenza in media di due gruppi è basata sulla 3.19

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)} \qquad (3.19)$$

usata congiuntamente con l'approssimazione N per campioni grandi.

Anche se la disribuzione *t* è raramente applicabile in economia molto software la usano per calcolare intervalli di cofidenza e p-value. Questo non pone problem nella pratica perchè la differenza tra t e N è trascurabile per grandi campioni. Per n>15 la differenza nei p-value non eccede 0.01 e per n>80 non eccede 0.002

La distribuzione *t* di Student – Riepilogo

- L'ipotesi che Y abbia distribuzione $N(\mu_Y, \sigma_Y^2)$ è raramente plausibile nella pratica (reddito? numero di figli?)
- per $n > 30$, la distribuzione t e $N(0,1)$ sono molto vicine (al crescere di n, la distribuzione t_{n-1} converge a $N(0,1)$)
- La distribuzione t è un artefatto che risale ai tempi in cui le dimensioni dei campioni erano piccole e i "calcolatori" erano persone
- Per motivi storici, il software statistico utilizza generalmente la distribuzione t per calcolare valori-p ma questo è irrilevante quando la dimensione del campione è moderata o grande.
- Per questi motivi, in questo corso ci concentreremo sull'approssimazione con n grande data dal limite centrale (TLC)
    1. Quadro probabilistico per l'inferenza statistica
    2. Stima
    3. Verifica
    4. **Intervalli di confidenza**

Intervalli di confidenza

- Un ***intervallo di confidenza al 95%*** per μ_Y è un intervallo che contiene il valore vero di μ_Y nel 95% dei campioni ripetuti.

- *Digressione*: qual è la casualità qui? I valori di $Y_1,\ldots,Y_n$ e quindi qualsiasi funzione degli stessi – incluso l'intervallo di confidenza - differirà da un campione all'altro. Il parametro della popolazione, μ_Y, non è casuale; semplicemente, non lo conosciamo.

Intervalli di confidenza (continua)

Un intervallo di confidenza al 95% può sempre essere costruito come insieme di valori dei μ_Y non rifiutati da un test di ipotesi con un livello di significatività del 5%.

$$\{\mu_Y: \left|\frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}}\right| \leq 1{,}96\} = \{\mu_Y: -1{,}96 \leq \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}} \leq 1{,}96\}$$

$$= \{\mu_Y \in (\bar{Y} - 1{,}96\frac{s_Y}{\sqrt{n}} \ , \ \bar{Y} + 1{,}96 \frac{s_Y}{\sqrt{n}})\}$$

Questo intervallo di confidenza si basa sugli n-grande e sui risultati che $\bar{Y}$ *è approssimata da una distribuzione normale e* $s_Y^2 \xrightarrow{p} \sigma_Y^2$

Riepilogo:

Dalle due ipotesi di:

1. campionamento casuale semplice di una popolazione, cioè $\{Y_i, i = 1,\ldots,n\}$ sono i.i.d.
2. $0 < E(Y^4) < \infty$

abbiamo sviluppato, per grandi campioni (*n* grande):

- Teoria della stima (distribuzione campionaria di $\overline{Y}$)
- Teoria della verifica di ipotesi (distribuzione con *n* grande della statistica *t* e calcolo del valore-*p*)
- Teoria degli intervalli di confidenza (costruita invertendo la statistica testa)

Le ipotesi (1) e (2) sono plausibili nella pratica? **Sì**

Torniamo alla domanda politica di partenza:

Qual è l'effetto sui punteggi nei test della riduzione della dimensione delle classi di uno studente per classe?

Abbiamo risposto a questa domanda?

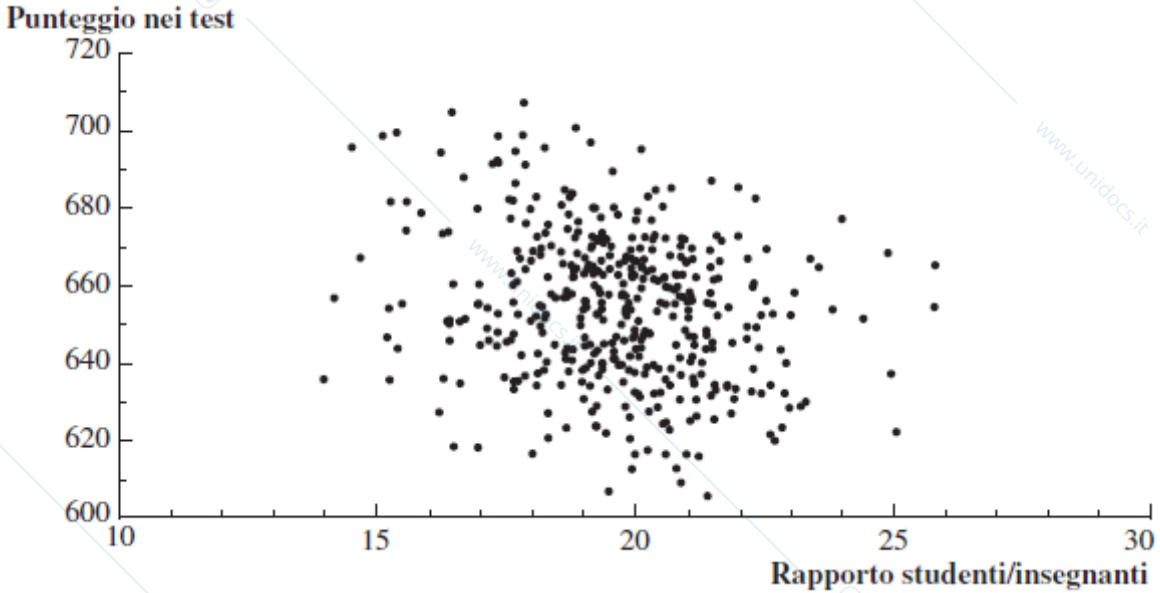


Figura 4.2

Diagramma a nuvola del punteggio nei test e del rapporto studenti/insegnanti (dati relativi ai distretti scolastici della California).

Dati per i 420 distretti scolastici della California. C'è una debole relazione negativa tra il rapporto studenti/insegnanti e il punteggio nei test: la correlazione campionaria è pari a –0,23.

Econometria I

Corso di laurea magistrale in Finanza e Assicurazioni

Slides per lezioni

secondo semestre 2020

Marco Ventura

Cap. 4 SW

Sommario del cap. 4

1. Il modello di regressione lineare
2. Lo stimatore dei minimi quadrati ordinari (OLS) e la retta di regressione campionaria
3. Misure di bontà della regressione campionaria
4. Le assunzioni dei minimi quadrati
5. La distribuzione campionaria dello stimatore OLS

La regressione lineare consente di stimare la pendenza della retta di regressione.

- La pendenza della retta di regressione è l'effetto atteso su *Y* di una variazione unitaria in *X*.
- Il nostro scopo ultimo è quello di stimare l'effetto causale su *Y* di una variazione unitaria in *X* – ma per ora ci limitiamo a considerare il problema dell'adattamento di una retta ai dati su due variabili *Y* e *X*.
- REM: correlazione non implica la causalità

Il problema dell'inferenza statistica per la regressione lineare è, a livello generale, identico a quello della stima della media o delle differenze tra medie. L'inferenza statistica, o econometrica, sulla pendenza comporta:

- Stima:
  - Come tracciare una retta attraverso i dati per stimare la pendenza della regressione?
    - Risposta: minimi quadrati ordinari (OLS).
  - Quali sono vantaggi e svantaggi dei minimi quadrati ordinari?
- Verifica di ipotesi:
  - Come verificare se la pendenza è zero?
- Intervalli di confidenza:
  - Come costruire un intervallo di confidenza per la pendenza?

Il modello di regressione lineare (Paragrafo 4.1)

La ***retta di regressione***:

$$TestScore = \beta_0 + \beta_1 STR$$

β_1 = pendenza della retta di regressione

$$= \frac{\Delta \text{Test score}}{\Delta STR}$$

= variazione nel punteggio nei test per una variazione unitaria in *STR*

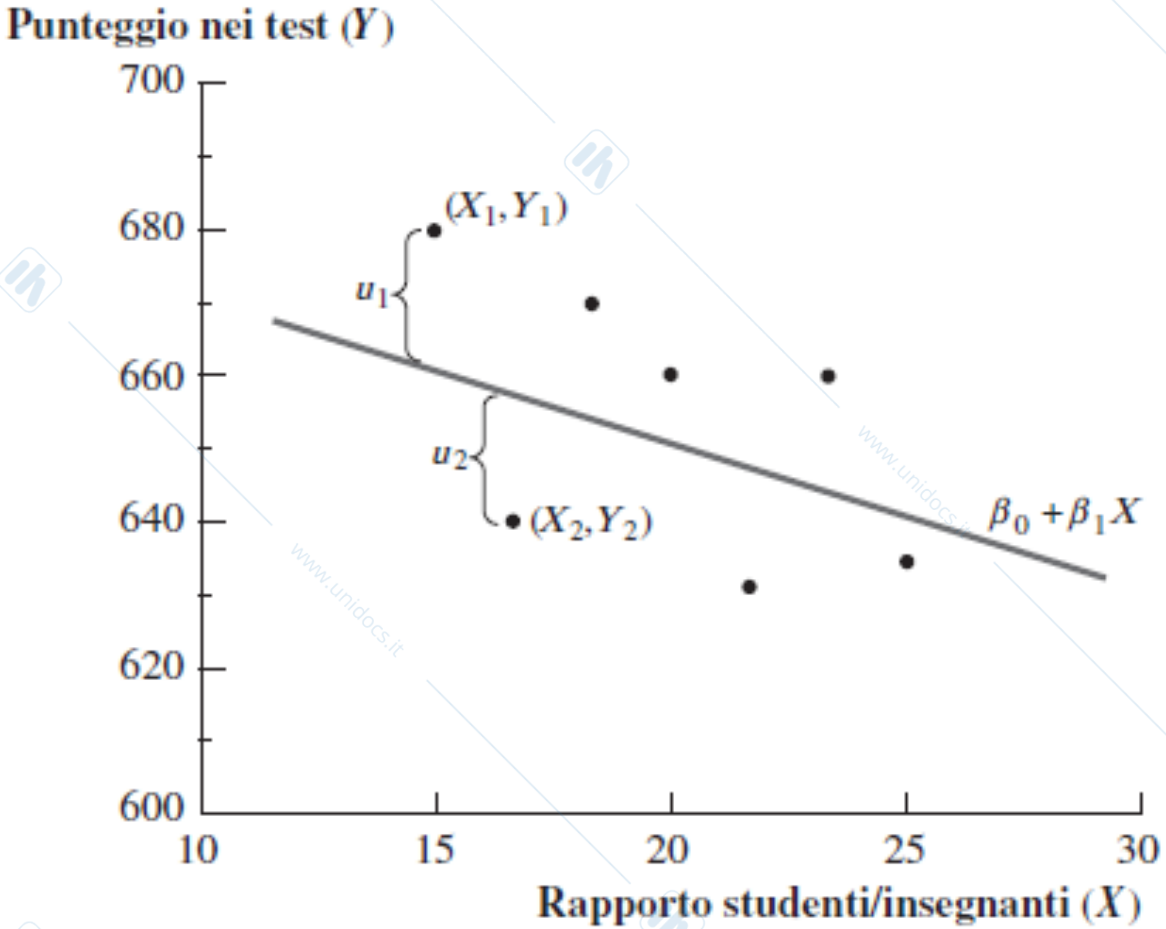
- *Perché β_0 e β_1 sono parametri di "popolazione"*?
- Vorremmo conoscere il valore di β_1.
- Non conosciamo β_1, perciò dobbiamo stimarlo utilizzando i dati.

Il modello di regressione lineare

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1,\ldots,\ n$$

- Abbiamo n osservazioni, $(X_i,\ Y_i)$, $i = 1,..,\ n$.
- X è la ***variabile indipendente*** o ***regressore***
- Y è la ***variabile dipendente***
- β_0 = ***intercetta***
- β_1 = ***pendenza***
- u_i = ***errore*** *di regressione*
- L'errore di regressione è costituito da fattori omessi. In generale questi fattori omessi sono altri fattori, diversi dalla variabile X, che influenzano Y. L'errore di regressione include anche l'errore nella misura di Y.

Il modello di regressione in un'immagine: *Osservazioni su Y e X (n = 7); la retta di regressione; l'errore di* regressione (il "termine d'errore"):



Lo stimatore OLS (Paragrafo 4.2)

Come possiamo stimare β_0 e β_1 dai data?

Si ricordi che lo stimatore OLS di μ_Y:, $\overline{Y}$, è dato da

$$\min_m \sum_{i=1}^{n} (Y_i - m)^2$$

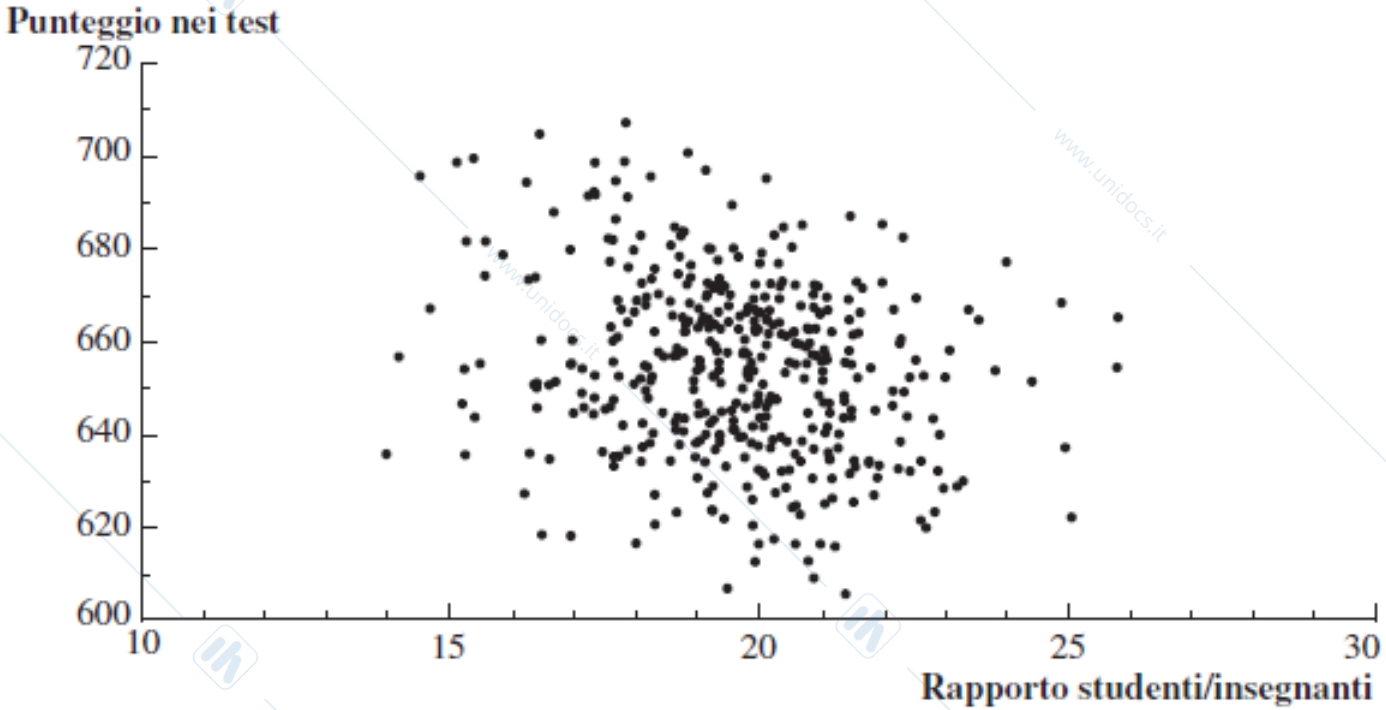
Per analogia, **ci concentreremo sullo stimatore dei minimi quadrati (OLS, "*ordinary least squares*") dei parametri ignoti β_0 e β_1.** Lo stimatore OLS è dato da

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

Meccanismo dei minimi quadrati ordinari

La retta di regressione: $TestScore = \beta_0 + \beta_1 STR$

$$\beta_1 = \frac{\Delta \text{TestScore}}{\Delta STR} = ??$$



Lo stimatore OLS: $\min_{b_0,b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$

- Lo stimatore OLS minimizza la differenza quadratica media tra i valori reali di Y_i e la previsione ("valori predetti") basata sulla retta stimata.
- Questo problema di minimizzazione si può risolvere con il calcolo differenziale (App. 4.2).
- **Il risultato sono gli stimatori OLS di β_0 e β_1.**

CONCETTO CHIAVE 4.2 **Stimatore OLS, valori predetti e residui**

Gli stimatori OLS della pendenza β_1 e dell'intercetta β_0 sono:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2} \tag{4.7}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \tag{4.8}$$

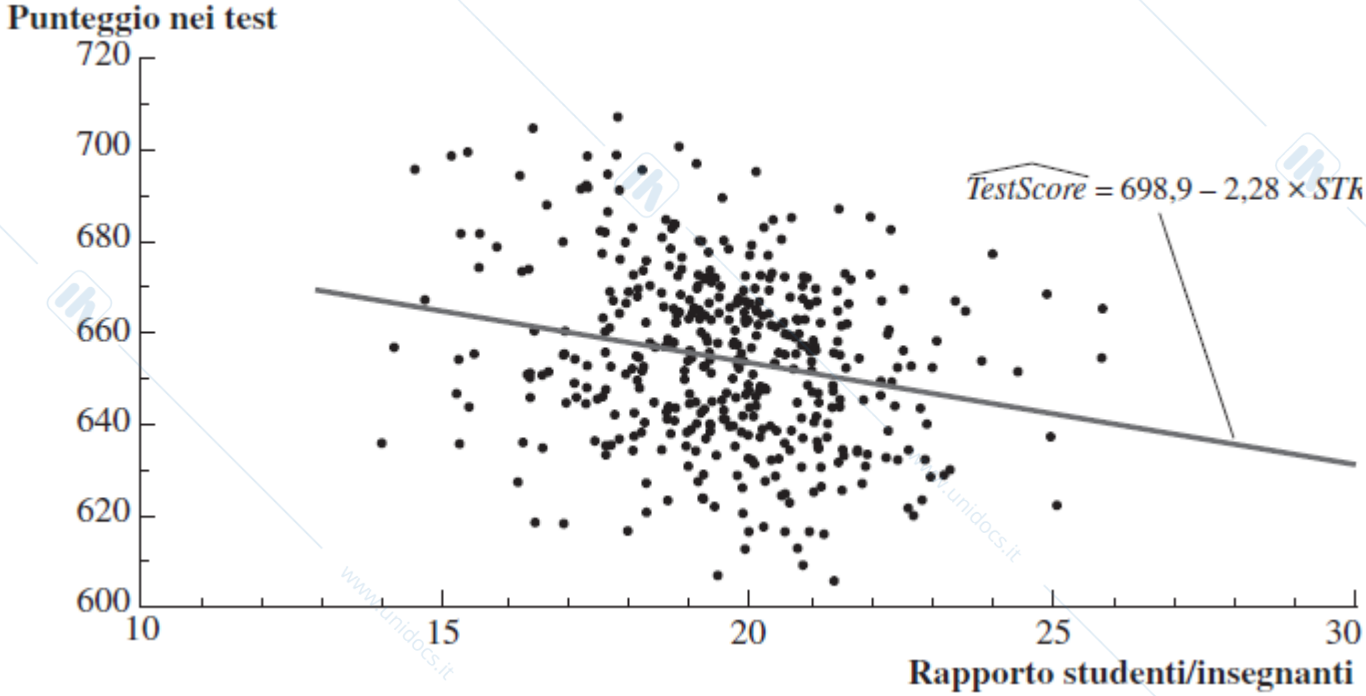
I valori predetti $\hat{Y}_i$ e i residui $\hat{u}_i$ sono:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \; i = 1, \ldots, n \tag{4.9}$$

$$\hat{u}_i = Y_i - \hat{Y}_i, \; i = 1, \ldots, n. \tag{4.10}$$

L'intercetta stimata ($\hat{\beta}_0$), la pendenza stimata ($\hat{\beta}_1$) e il residuo ($\hat{u}_1$) sono calcolati per un campione di n osservazioni di X_i e Y_i, con $i = 1,...,n$. Queste sono stime dell'intercetta β_0, della pendenza β_1 e dell'errore u_i nella popolazione.

Applicazione ai dati dei punteggi nei test della California *Punteggio nei test – Dimensioni delle classi*



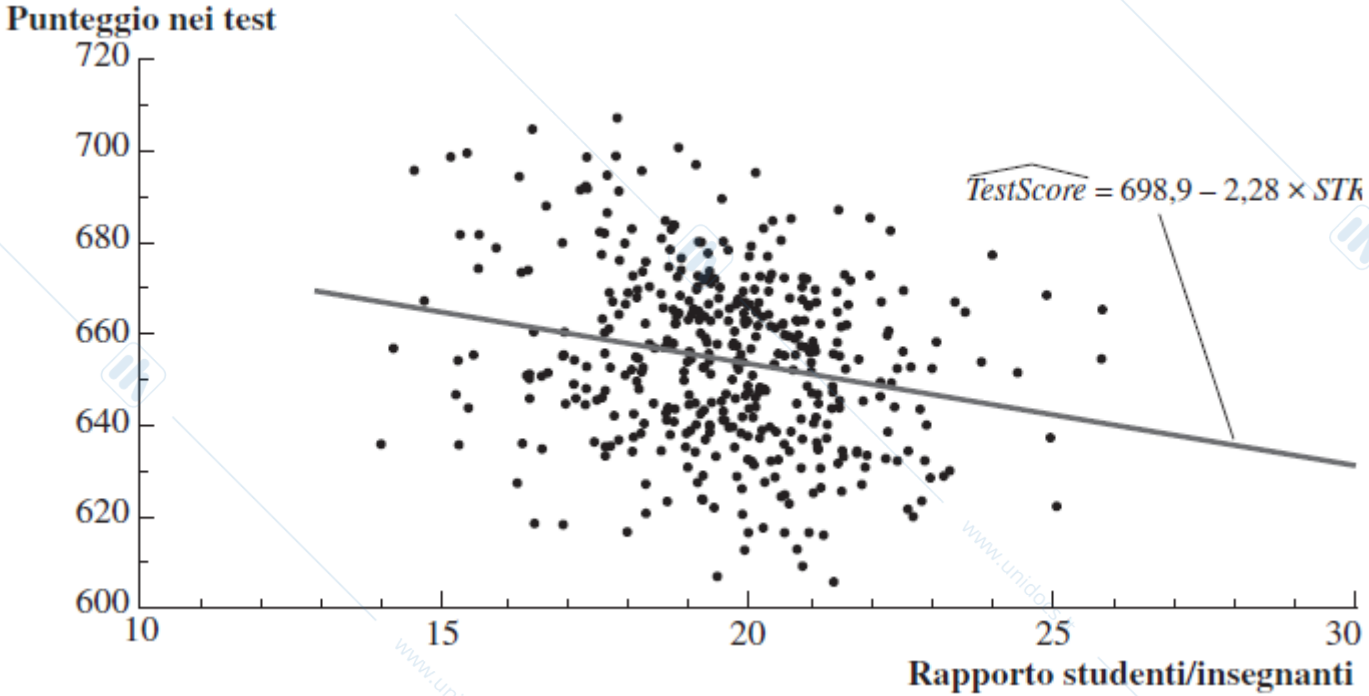
- Pendenza stimata = $\hat{\beta}_1$ = – 2,28
- Intercetta stimata = $\hat{\beta}_0$ = 698,9
- Retta di regressione stimata: $\widehat{TestScore}$ = 698,9 – 2,28 × *STR*

Interpretazione delle stime di pendenza e intercetta

- $\widehat{TestScore} = 698{,}9 - 2{,}28 \times STR$
- I distretti con uno studente in più per insegnante in media ottengono punteggi nei test inferiori di 2,28 punti.

- Cioè $\dfrac{\Delta \text{TestScore}}{\Delta STR} = -2{,}28$
- L'intercetta (letteralmente) significa che, secondo questa retta stimata, i distretti con zero studenti per insegnante otterrebbero un punteggio nei test stimato in 698,9. Ma questa interpretazione dell'intercetta non ha senso – estrapola la linea al di fuori dell'intervallo dei dati – in questo caso, l'intercetta non ha significato dal punto di vista economico.

Valori predetti e residui



Uno dei distretti nella banca dati è Antelope, CA, con *STR* = 19,33 e *TestScore* = 657,8

valore predetto: $\hat{Y}_{Antelope} = 698{,}9 - 2{,}28 \times 19{,}33 = 654{,}8$

residuo: $\hat{u}_{Antelope} = 657{,}8 - 654{,}8 = 3{,}0$

Regressione OLS: output di STATA

```
regress testscr 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
       _cons |   698.933   10.36436     67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

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

(discuteremo più avanti la parte rimanente di questo output)

Misure di bontà dell'adattamento (Paragrafo 4.3)

Due statistiche di regressione forniscono misure complementari della bonta dell'adattamento della regressione ai dati:

- L'***R*² della regressione** misura la frazione della varianza di *Y* spiegata da *X*; è priva di unità e può variare tra zero (nessun adattamento) e uno (perfetto adattamento)
- L'***errore standard della regressione* (*SER*)** misura la dimensione di un tipico residuo di regressione nelle unità di *Y*.

L'R^2 della regressione è la frazione della varianza campionaria di Y_i "spiegata" dalla regressione.

$Y_i = \hat{Y}_i + \hat{u}_i$ = stima OLS + residuo OLS

→ var camp. (Y) = var camp. ($\hat{Y}_i$) + var camp. ($\hat{u}_i$) (*perché?*)

→ somma dei quadrati (TSS) = SS "spiegata" (ESS) + SS "residua"

Definizione di R^2: $$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- $R^2 = 0$ significa $ESS = 0$ explained sum of square
- $R^2 = 1$ significa $ESS = TSS$
- $0 \leq R^2 \leq 1$
- Per la regressione con una singola X, R^2 = il quadrato del coefficiente di correlazione tra X e Y

L'errore standard della regressione (SER)

Il *SER* misura la dispersione della distribuzione di *u*. È (quasi) la deviazione standard campionaria dei residui OLS:

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

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

La seconda uguaglianza vale perché $\bar{\hat{u}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0$.

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

Il *SER*:

ha le unità di *u*, che sono le unità di *Y*

misura la "dimensione" media del residuo OLS (l'"errore" medio della retta di regressione OLS)

La ***radice dell'errore quadratico medio*** (*RMSE, Root Mean Squared Error*) è strettamente legata al *SER*:

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

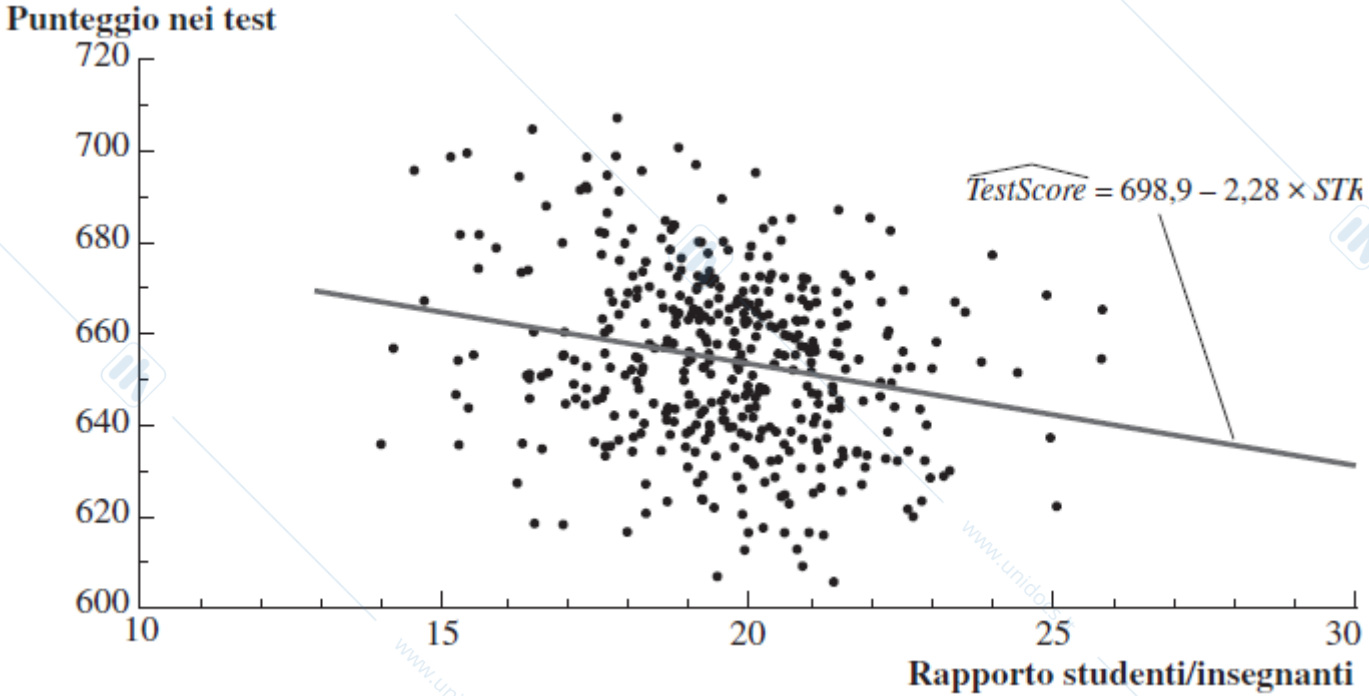
Misura la stessa cosa del *SER* – la differenza sta nel fattore $1/n$ anziché $1/(n-2)$.

Nota tecnica: perché dividere per *n*–2 anziché per *n*–1?

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

- La divisione per *n*–2 è una correzione "dei gradi di libertà" – esattamente come la divisione per *n*–1, con la differenza che per il *SER* sono stati stimati due parametri (β_0 e β_1, da $\hat{\beta}_0$ e $\hat{\beta}_1$, mentre in s_Y^2 ne è stato stimato solo uno (μ_Y, da Y).
- Quando *n* è grande non importa se si utilizzi *n*, *n*–1 o *n*–2 – anche se la formula convenzionale utilizza *n*–2 quando c'è un singolo regressore.
- Per i dettagli, cfr. il Paragrafo 17.4

Esempio di R^2 e *SER*



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

STR spiega soltanto una piccola frazione della variazione nei punteggi nei test. Ha senso questo? Significa che STR non è importante in senso politico?

Le assunzioni dei minimi quadrati (Paragrafo 4.4)

- Quali sono, precisamente, le proprietà della distribuzione campionaria dello stimatore OLS? Quando lo stimatore sarà non distorto? Qual è la sua varianza?

- Per rispondere a queste domande dobbiamo fare alcune assunzioni sulla relazione tra *Y* e *X* e su come sono ottenute (lo schema di campionamento)

- Queste assunzioni – sono tre – sono note come assunzioni dei minimi quadrati.

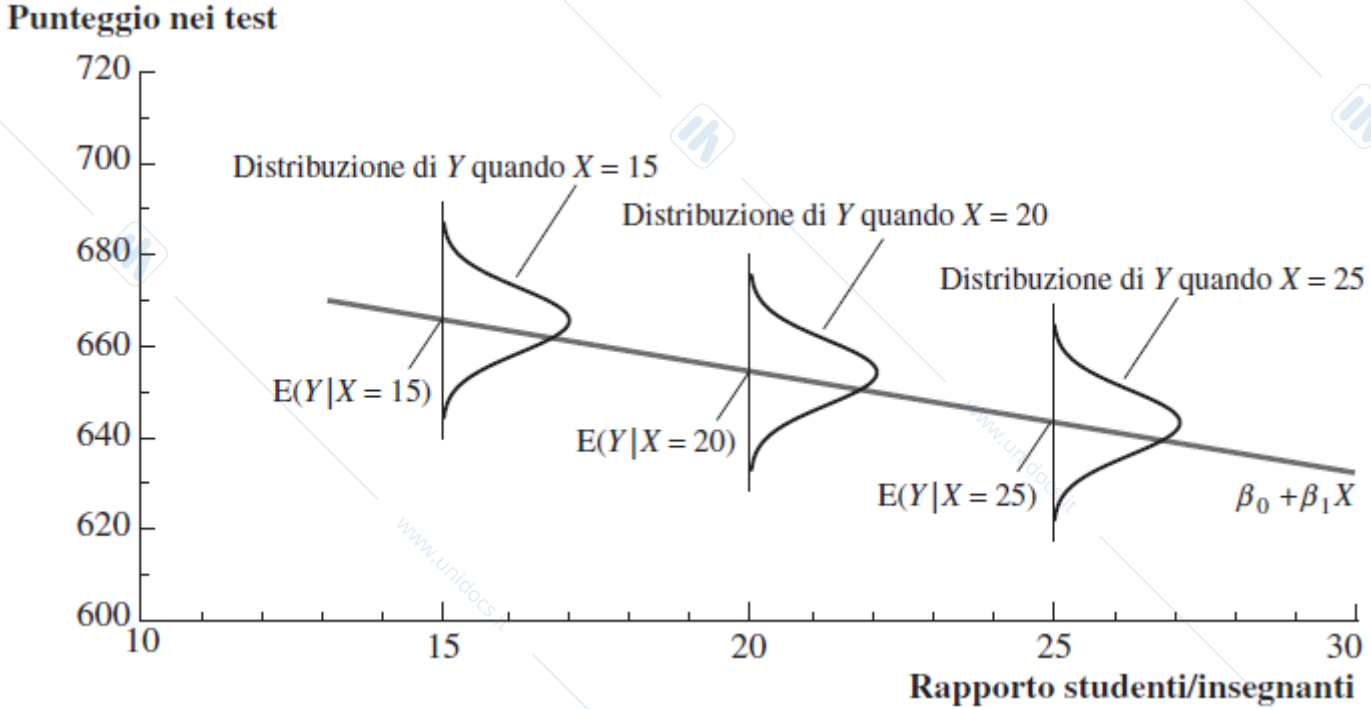
Le assunzioni dei minimi quadrati

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1,\ldots,\ n$$

1. La distribuzione di u condizionata a X ha media nulla, cioè $E(u|X = x) = 0$.
   - *Questo implica che $\hat{\beta}_1$ è non distorto*
2. (X_i, Y_i), $i = 1,\ldots,n$, sono i.i.d.
   - *Questo è vero se (X, Y) sono ottenuti mediante campionamento casuale*
   - *Questo fornisce la distribuzione campionaria di $\hat{\beta}_0$ e $\hat{\beta}_1$*
3. Gli outlier in X e/o Y sono rari.
   - *Tecnicamente, X e Y hanno momenti quarti finiti*
   - *Gli outlier possono risultare in valori privi di senso di $\hat{\beta}_1$*

Assunzione dei minimi quadrati n. 1: $E(u|X = x) = 0$.

Per ogni dato valore di X, la media di u è zero:



Esempio: $TestScore_i = \beta_0 + \beta_1 STR_i + u_i$, u_i = altri fattori

- Quali sono alcuni di questi "altri fattori"?
- $E(u|X=x) = 0$ è plausibile per questi altri fattori?

Correlazione e media condizionata

Dal par 2.27 sappiamo che se una vbl ha media nulla condiziona a un'altra vbl, allora le 2 vbl hanno cov=0 e sono incorrelate.

$E(u|X)=0 \rightarrow corr(u,X)=0$

Non vale invece il contrario. Se $corr(u,X)=0$ non implica $E(u|X)=0$.

Se $corr(u,X)\neq 0$ implica $\rightarrow E(u|X) \neq 0$

Vale la pena discutere l'hp circa $E(u|X)=0$ in termini di $corr(u,X)$, se X e u sono correlate allora l'assunzione circa la media condizionata è violata

Assunzione dei minimi quadrati n. 1 (continua)

- Un riferimento per riflettere su questa assunzione è quello di considerare un esperimento controllato casualizzato ideale:
- *X* è assegnato casualmente a persone (studenti assegnati casualmente a classi di dimensioni diverse; pazienti assegnati casualmente a trattamenti medici). La casualizzazione è svolta dal computer – senza utilizzare informazioni sull'individuo.
- Poiché *X* è assegnata casualmente, tutte le altre caratteristiche individuali – gli aspetti riassunti da *u* – sono distribuite indipendentemente da *X*, perciò *u* e *X* sono indipendenti
- Quindi, in un esperimento controllato casualizzato ideale, $\boldsymbol{E(u|X = x) = 0}$ (cioè vale l'assunzione 1)
- In esperimenti reali, o con dati non sperimentali, dovremo riflettere bene sul fatto che $E(u|X = x) = 0$ valga o meno.

Assunzione dei minimi quadrati n. 2: (X_i, Y_i), $i = 1,\ldots,n$ sono i.i.d.

Questo si verifica automaticamente se l'unità (individuo, distretto) è campionata mediante campionamento casuale semplice:

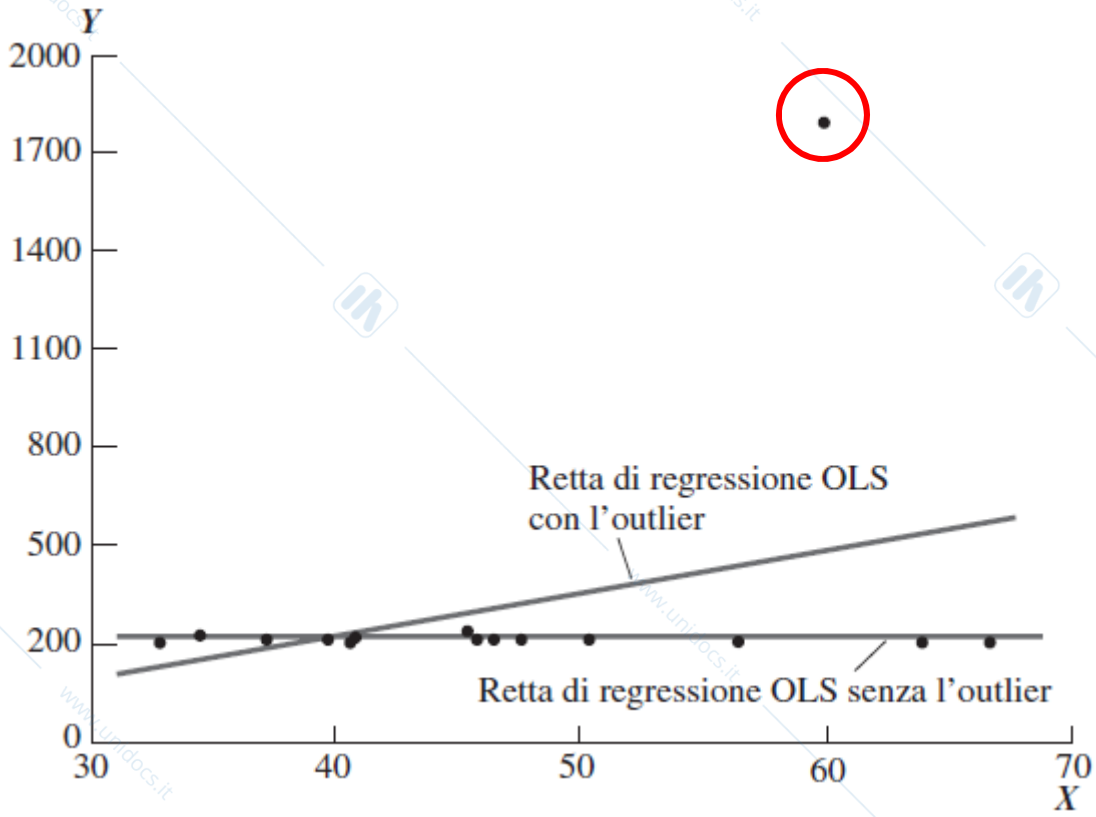
- Le unità sono scelte dalla stessa popolazione, perciò (X_i, Y_i) sono *identicamente distribuite* per ogni $i = 1,\ldots, n$.
- Le unità sono scelte a caso, perciò i valori di (X, Y) per unità diverse sono *indipendentemente distribuite*.

I campionamenti **non** i.i.d. si incontrano principalmente quando si registrano dati nel tempo per la stessa unità (dati panel e serie temporali) – affronteremo tale complicazione quando tratteremo i dati panel.

Assunzione dei minimi quadrati n. 3: *gli outlier sono rari*
Proposizione tecnica: $E(X^4) < \infty$ e $E(Y^4) < \infty$

- Un outlier è un valore estremo di *X* o *Y*
- A livello tecnico, se *X* e *Y* sono limitate, allora hanno momenti quarti finiti (i punteggi nei test standardizzati soddisfano questa condizione, come anche *STR*, reddito familiare, ecc.)
- La sostanza di questa assunzione è che un outlier può influenzare fortemente i risultati, perciò dobbiamo escludere i valori estremi.
- Esaminate i dati! Se avete un outlier, si tratta di un refuso? Non appartiene al dataset? Perché è un outlier?

Lo stimatore OLS può essere sensibile a un outlier:



- *Il punto isolato è un outlier in X o Y?*
- In pratica, gli outlier sono spesso distorsioni dei dati (problemi nella codifica o nella registrazione). Talvolta sono osservazioni che non dovrebbero stare nel dataset. Tracciate i vostri dati!

Distribuzione campionaria degli stimatori OLS (Paragrafo 4.5)

Lo stimatore OLS è calcolato da un campione di dati. Un campione diverso porta a un valore diverso di $\hat{\beta}_1$. Questa è l'origine della "incertezza campionaria" di $\hat{\beta}_1$. Vogliamo:

- quantificare l'incertezza campionaria associata a $\hat{\beta}_1$
- usare $\hat{\beta}_1$ per verificare ipotesi quali $\beta_1 = 0$
- costruire un intervallo di confidenza per β_1
- Tutti questi punti richiedono di determinare la distribuzione campionaria dello stimatore OLS. Due passaggi…
  - Quadro di riferimento probabilistico per la regressione lineare
  - Distribuzione dello stimatore OLS

Quadro di riferimento probabilistico per la regressione lineare

Il quadro di riferimento probabilistico per la regressione lineare è riepilogato dalle tre assunzioni dei minimi quadrati.

Popolazione

- Il gruppo di interesse (esempio: tutti i possibili distretti scolastici)

Variabili casuali*: *Y*, *X

- Esempio: (*TestScore*, *STR*)

***Distribuzione congiunta di* (*Y*, *X*). Assumiamo:**

- La funzione di regressione è lineare
- $E(u|X) = 0$ (prima assunzione dei OLS, vedi slide 23)
- X, Y hanno momenti quarti finiti non nulli (terza assunzione)

La raccolta dei dati mediante campionamento casuale semplice implica:

- $\{(X_i, Y_i)\}$, $i = 1,\ldots, n$, sono i.i.d. (seconda ass., slide 23)

Distribuzione campionaria di $\hat{\beta}_1$

- Come $\bar{Y}$, $\hat{\beta}_1$ ha una distribuzione campionaria.
- Qual è $E(\hat{\beta}_1)$?
  - Se $E(\hat{\beta}_1) = \beta_1$, allora lo stimatore OLS non è distorto – ottima cosa!
- Qual è var($\hat{\beta}_1$)? (misura di incertezza campionaria)
  - Dobbiamo derivare una formula per poter calcolare l'errore standard di $\hat{\beta}_1$.
- Qual è la distribuzione di $\hat{\beta}_1$ in piccoli campioni?
  - È molto complessa, in generale
- Qual è la distribuzione di $\hat{\beta}_1$ in grandi campioni?
  - In grandi campioni, $\hat{\beta}_1$ ha distribuzione normale.

Media e varianza della distribuzione campionaria di $\hat{\beta}_1$

Un po' di algebra preliminare:

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

$$\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{u}$$

perciò $Y_i - \bar{Y} = \beta_1 (X_i - \bar{X}) + (u_i - \bar{u})$

Quindi

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

sostituendo $Y_i - \bar{Y}$

$$= \frac{\sum_{i=1}^{n} (X_i - \bar{X})[\beta_1 (X_i - \bar{X}) + (u_i - \bar{u})]}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

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

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

Ora $\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^{n}(X_i - \bar{X})u_i - \left[\sum_{i=1}^{n}(X_i - \bar{X})\right]\bar{u}$

$$= \sum_{i=1}^{n}(X_i - \bar{X})u_i - \left[\left(\sum_{i=1}^{n}X_i\right) - n\bar{X}\right]\bar{u}$$

$$= \sum_{i=1}^{n}(X_i - \bar{X})u_i$$

Sostituiamo $\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^{n}(X_i - \bar{X})u_i$ nella espressione per $\hat{\beta}_1 - \beta_1$:

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

perciò

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

Ora possiamo calcolare E($\hat{\beta}_1$) e var($\hat{\beta}_1$):

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

$$= E\left\{E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] X_1,\ldots,X_n\right\}$$

$$= E\left[\frac{\sum_{i=1}^{n}(X_i-\bar{X})E(u_i|X_1,\ldots,X_n)}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right] = 0$$

= 0 **poiché $E(u_i|X_i=x) = 0$ per l'assunzione OLS 1**

- Quindi l'assunzione 1 implica che $E(\hat{\beta}_1) = \beta_1$
- Cioè $\hat{\beta}_1$ **è uno stimatore non distorto di β_1.**
- Per i dettagli cfr. Appendice 4.3

Ora calcoliamo var($\hat{\beta}_1$):

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

Ponendo $v_i = (X_i - \bar{X})u_i$; e ricordando che

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = (n-1)s_x^2$$

Possiamo riscrivere $\hat{\beta}_1 - \beta_1$ nel seguente modo

Ora calcoliamo var($\hat{\beta}_1$):

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

Se *n* è grande, $s_X^2 \approx \sigma_X^2$ e $\frac{n-1}{n} \approx 1$, perciò

$$\hat{\beta}_1 - \beta_1 \approx \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\sigma_X^2}$$

(cfr. Appendice 4.3). Quindi

$$\hat{\beta}_1 - \beta_1 \approx \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\sigma_X^2}$$

perciò $\text{var}(\hat{\beta}_1 - \beta_1) = \text{var}(\hat{\beta}_1)$

$$= \text{var}\left[\left(\frac{1}{n}\sum_{i=1}^{n} v_i\right)\right] \Big/ (\sigma_X^2)^2 \quad = \frac{\frac{1}{n^2} n\, var(v_i)}{(\sigma_x{}^2)^2}$$

$$= \frac{\frac{1}{n} var(v_i)}{(\sigma_x{}^2)^2}$$

NOTARE: l'ultima uguaglianza usa ass. 2 dei OLS, (X_i, Y_i) sono iid e quindi la Var della somma di v_i è la somma delle $\text{var}(v_i)$. Slide 27

$$\text{var}(\hat{\beta}_1) = \frac{\text{var}(v_i)/n}{(\sigma_X^2)^2} = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{(\sigma_X^2)^2}$$

Riepilogo

1. $\hat{\beta}_1$ è non distorto: $E(\hat{\beta}_1) = \beta_1$ – proprio come $\bar{Y}$!

2. $\text{var}(\hat{\beta}_1)$ è inversamente proporzionale a n – proprio come $\bar{Y}$!

Qual è la distribuzione campionaria di $\hat{\beta}_1$ *?*

Determinare la distribuzione campionaria esatta è complicato – dipende dalla distribuzione di (Y, X) – ma quando n è grande otteniamo alcune buone (e semplici) approssimazioni:

1) Poiché var($\hat{\beta}_1$) $\propto 1/n$ e $E(\hat{\beta}_1) = \beta_1$, $\hat{\beta}_1 \xrightarrow{p} \beta_1$
2) Quando n è grande, la distribuzione campionaria di $\hat{\beta}_1$ è ben approssimata da una distribuzione normale (TLC)

Ricordiamo il ***TLC***: sia $\{v_i\}$, $i = 1,\ldots, n$ i.i.d. con $E(v) = 0$ e var(v) = σ^2. Allora, quando n è grande, la distribuzione di $\frac{1}{n}\sum_{i=1}^{n} v_i$ è approssimata da $N(0, \sigma_v^2 / n)$.

Approssimazione per *n* grande della distribuzione di $\hat{\beta}_1$:

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

- Quando n è grande, $v_i = (X_i - \bar{X})u_i \approx (X_i - \mu_X)u_i$, che è i.i.d. (*perché*?) e var(v_i) < ∞ (*perché*?). Perciò, per il TLC, la distribuzione di $\frac{1}{n}\sum_{i=1}^{n} v_i$ è approssimata da $N(0, \sigma_v^2 / n)$.
- Quindi, per n grande, la distribuzione di $\hat{\beta}_1$ è approssimata da

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_v^2}{n(\sigma_X^2)^2}\right), \text{ dove } v_i = (X_i - \mu_X)u_i$$

Maggiore è la varianza di *X*, minore è la varianza di $\hat{\beta}_1$

Calcoli

$$\text{var}(\hat{\beta}_1 - \beta_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{(\sigma_X^2)^2}$$

Dove $\sigma_X^2 = \text{var}(X_i)$. La varianza di *X* appare (al quadrato) al denominatore – perciò aumentando la dispersione di *X* diminuisce la varianza di β_1.

Ragionamento intuitivo

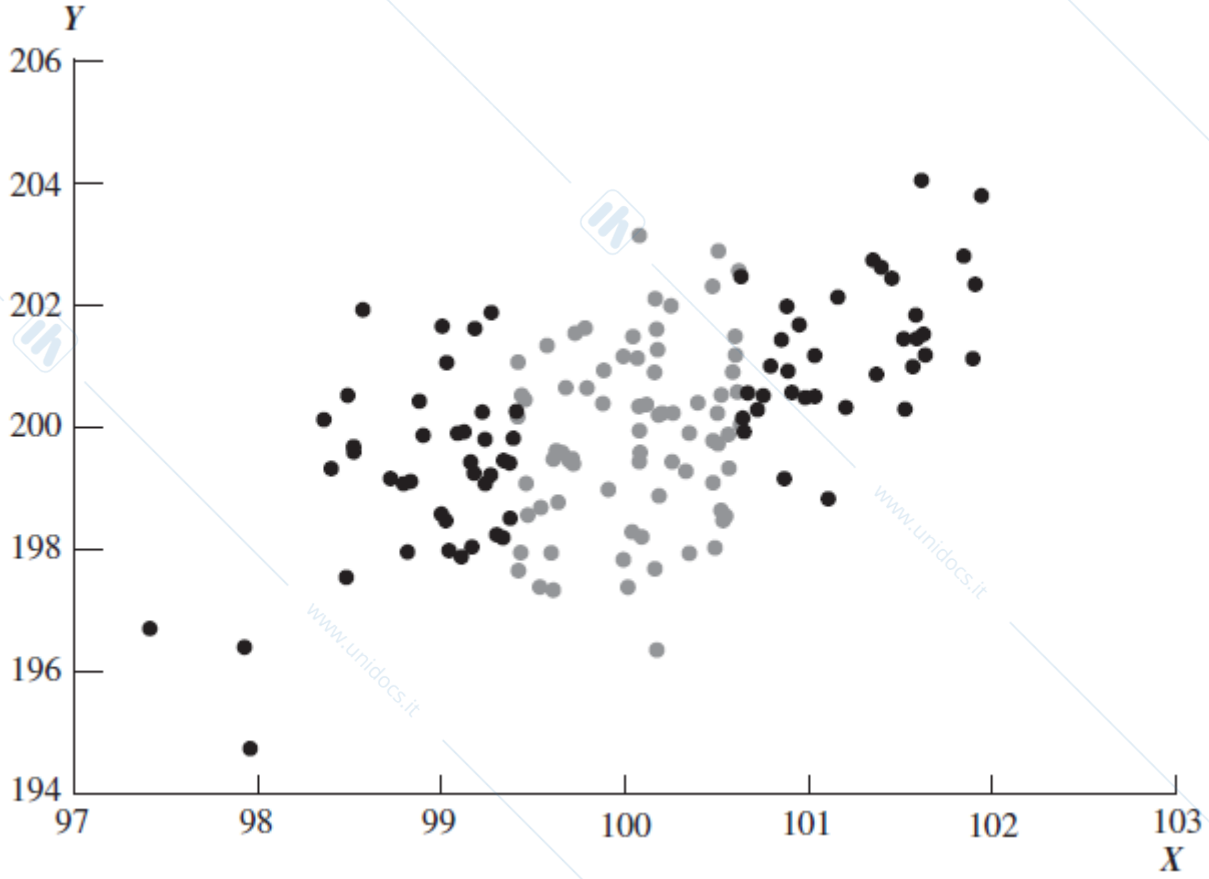
Se vi è più variazione in *X*, allora vi sono più informazioni nei dati che si possono utilizzare per l'adattamento della retta di regressione. Lo si vede meglio in una figura…

La figura riporta una nuvola di 150 osservazioni su Y e X. Le 75 grige sono state generate con minore disporsione, cioè più vicine a $\bar{X}$.

Se dovessimo disegnare una retta che passi il più vicino possible ai punti neri o a quelli grigi sarebbe più facile disegnare quella che passa attraverso i punti neri che hanno varianza maggiore.

Similmente maggiore è la varianza di X, più preciso risulta $\widehat{\beta_1}$

Maggiore è la varianza di *X*, minore è la varianza di $\hat{\beta}_1$



Il numero di punti neri e grigi è lo stesso. Quali consentono di ottenere una retta di regressione più accurata?

Riepilogo della distribuzione campionaria di $\hat{\beta}_1$:

Se valgono le tre assunzioni dei minimi quadrati, allora

- La distribuzione campionaria esatta (campione finito) di $\hat{\beta}_1$ ha:
  - $E(\hat{\beta}_1) = \beta_1$ (cioè $\hat{\beta}_1$ è non distorto)
  - $\text{var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{\sigma_X^4} \quad \infty \frac{1}{n}$.
- A parte media e varianza, la distribuzione esatta di $\hat{\beta}_1$ è complessa e dipende dalla distribuzione di (X, u)
- $\hat{\beta}_1 \xrightarrow{p} \beta_1$ (cioè $\hat{\beta}_1$ è consistente)
- Quando n è grande, $\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}} \sim N(0,1)$ (CLT)
- *Segue in parallelo la distribuzione campionaria di* $\bar{Y}$.

CONCETTO CHIAVE 4.4 **Le distribuzioni di β_1 e β_0 in grandi campioni**

Se valgono le ipotesi dei minimi quadrati presentate nel Concetto chiave 4.3, la distribuzione campionaria congiunta di $\hat{\beta}_0$ e $\hat{\beta}_1$ è approssimativamente normale in grandi campioni. La distribuzione in grandi campioni di $\hat{\beta}_1$ è $N(\beta_1, \sigma^2_{\hat{\beta}_1})$, dove la varianza di questa distribuzione, $\sigma^2_{\hat{\beta}_1}$, è

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{var[(X_i - \mu_X)u_i]}{[\text{var}(X_i)]^2}. \tag{4.21}$$

La distribuzione in grandi campioni di $\hat{\beta}_0$ è $N(\beta_0, \sigma^2_{\hat{\beta}_0})$, dove

$$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{\text{var}(H_i u_i)}{[\text{E}(H_i^2)]^2}, \text{ dove } H_i = 1 - \left[\frac{\mu_X}{\text{E}(X_i^2)}\right]X_i. \tag{4.22}$$

Ora siamo pronti ad affrontare verifiche di ipotesi e intervalli di confidenza…

Econometria I

Corso di laurea magistrale in Finanza e Assicurazioni

Slides per lezioni

secondo semestre 2020

Marco Ventura

Cap. 5 SW

Capitolo 5 SW

Regressione con un singolo regressore: verifica di ipotesi e intervalli di confidenza

Sommario

1. L'errore standard di $\hat{\beta}_1$
2. Verifiche di ipotesi concernenti β_1
3. Intervalli di confidenza per β_1
4. La regressione quando X è variabile binaria
5. Eteroschedasticità e omoschedasticità
6. Efficienza OLS e distribuzione t di Student

Panoramica sul percorso che ci attende…

Vogliamo conoscere la pendenza della retta di regressione. Disponiamo dei dati di un campione, perciò sussiste l'incertezza della campionatura. Per raggiungere l'obiettivo si procede in cinque passaggi:

1. Definire la popolazione oggetto di interesse
2. Fornire uno stimatore di questa popolazione
3. Derivare la distribuzione campionaria dello stimatore (ciò richiede alcune assunzioni). In grandi campioni questa distribuzione campionaria sarà normale per il TLC.
4. La radice quadrata della varianza stimata della distribuzione campionaria è l'errore standard (SE) dello stimatore
5. Utilizzare *SE* per costruire statistiche-*t* (per le verifiche di ipotesi) e intervalli di confidenza.

Oggetto di interesse: β_1 in

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1,\ldots,n$$

$\beta_1 = \Delta Y/\Delta X$, per una variazione autonoma in X

Stimatore: lo stimatore OLS $\hat{\beta}_1$.

La distribuzione campionaria di $\hat{\beta}_1$:

Per derivare la distribuzione in grandi campioni di $\hat{\beta}_1$, effettuiamo le seguenti assunzioni:

Le assunzioni dei minimi quadrati:

1. $E(u|X = x) = 0$.

2. (X_i, Y_i), $i = 1,\ldots,n$, sono i.i.d.

3. Gli outlier sono rari ($E(X^4) < \infty$, $E(Y^4) < \infty$.

La distribuzione campionaria di $\hat{\beta}_1$ (continua)

Sotto le assunzioni dei minimi quadrati, per *n* grande, la distribuzione di $\hat{\beta}_1$ è approssimata da

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_v^2}{n(\sigma_X^2)^2}\right)$$, dove $v_i = (X_i - \mu_X)u_i$

Verifica di ipotesi ed errore standard di $\hat{\beta}_1$ (Paragrafo 5.1)

L'obiettivo è di verificare un'ipotesi, quale $\beta_1 = 0$, utilizzando i dati – per determinare sperimentalmente che l'ipotesi (nulla) sia corretta o meno.

Impostazione generale

Ipotesi nulla e alternativa **bilaterale**:

$$H_0: \ \beta_1 = \beta_{1,0} \ \text{ vs. } H_1: \ \beta_1 \neq \beta_{1,0}$$

dove $\beta_{1,0}$ è il valore ipotizzato nella situazione nulla.

Ipotesi nulla e alternativa **unilaterale**:

$$H_0: \ \beta_1 = \beta_{1,0} \ \text{ vs. } H_1: \ \beta_1 < \beta_{1,0}$$

***Soluzione generale*: costruire la statistica-*t* e calcolare il valore-*p* (o confrontare il valore critico (0,1) *N*)**

- ***In generale***: $$t = \frac{\text{stimatore} - \text{valore ipotizzato}}{\text{errore standard dello stimatore}}$$

dove l'*SE* dello stimatore è la radice quadrata di uno stimatore della varianza dello stimatore.

- ***Per verificare la media di Y***: $t = \dfrac{\bar{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}}$
- ***Per verificare β_1***, $t = \dfrac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$,

dove *SE*($\hat{\beta}_1$) = la radice quadrata di uno stimatore della varianza della distribuzione campionaria di $\hat{\beta}_1$

Formula per SE($\hat{\beta}_1$)

Si ricordi l'espressione per la varianza di (n grande):

$$\text{var}(\hat{\beta}_1) = \frac{\text{var}[(X_i - \mu_x)u_i]}{n(\sigma_X^2)^2} = \frac{\sigma_v^2}{n(\sigma_X^2)^2}$$, dove $v_i = (X_i - \mu_X)u_i$.

Lo stimatore della varianza di $\hat{\beta}_1$ sostituisce i valori di popolazione ignnoti di σ_v^2 e σ_X^2 con gli stimatori ricavati dai dati:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\text{stimatore di } \sigma_v^2}{(\text{stimatore di } \sigma_X^2)^2} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{v}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$$

dove $\hat{v}_i = (X_i - \bar{X})\hat{u}_i$.

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{v}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$$, dove $\hat{v}_i = (X_i - \bar{X})\hat{u}_i$

$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}$ = errore standard di $\hat{\beta}_1$

È leggermente complicato, tuttavia:

- lo è meno di quanto sembri. La varianza var(v) è stimata dal numeratore, mentre [var(X)]^2 è stimata dal denominatore.
- Perché la correzione dei gradi di libertà n – 2? Perché sono stati stimati due coefficienti (β_0 e β_1).
- $SE(\hat{\beta}_1)$ viene calcolato dal software di regressione
- Il software di regressione ha questa formula memorizzata, perciò non dovete eseguire voi il calcolo.

Riepilogo:
Per verificare H_0: $\beta_1 = \beta_{1,0}$ v. H_1: $\beta_1 \neq \beta_{1,0}$,

- Costruire la statistica-*t*

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}}$$

- Si rifiuta al livello di significatività del 5% se |t| > 1,96
- Il valore *p* è *p* = Pr[|*t*| > |t^{act}|] = probabilità nell'area delle code normale |t^{act}|; si rifiuta al livello di significatività del 5% se il valore *p* è < 5%.

- Questa procedura si affida all'approssimazione di *n* grande che $\hat{\beta}_1$ sia distribuito normalmente; in generale *n* = 50 è grande abbastanza per un'approssimazione eccellente.

Esempio: *Punteggi nei test* (*TestScore*) e *STR*, dati della California

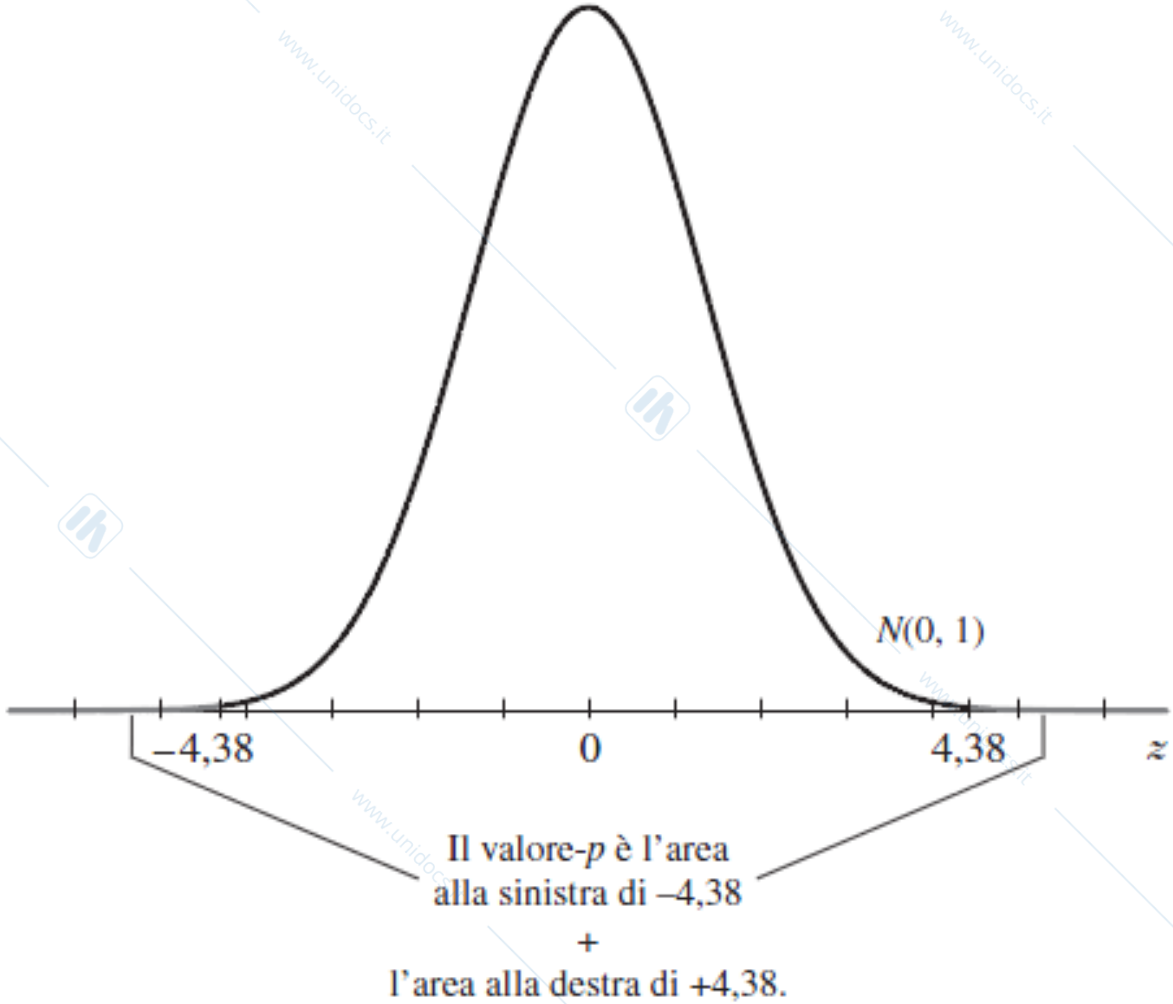
Regressione lineare stimata $\widehat{TestScore} = 698{,}9 - 2{,}28 \times STR$

Il software di regressione segnala gli errori standard:

$$SE(\hat{\beta}_0) = 10{,}4 \qquad SE(\hat{\beta}_1) = 0{,}52$$

verifica statistica-t $\beta_{1,0} = 0 = \dfrac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \dfrac{-2{,}28 - 0}{0{,}52} = -4{,}38$

- Il livello di significatività bilaterale dell'1% è 2,58, perciò rifiutiamo l'ipotesi nulla al livello di significatività dell'1%.
- In alternativa, possiamo calcolare il valore *p*…



Il valore *p* basato sull'approssimazione normale standard con *n* grande alla statistica *t* è 0,00001 (10^{-5})

Intervalli di confidenza per β_1 (Paragrafo 5.2)

Si ricordi che un intervallo di confidenza al 95% equivale a:

- la serie di punti che non può essere rifiutata al livello di significatività del 5%;
- una funzione polidroma (un intervallo funzione dei dati) che contiene il reale valore del parametro il 95% delle volte nei campioni ripetuti.

Poiché la statistica t per β_1 è N(0,1) in grandi campioni, la costruzione di un intervallo di confidenza al 95% per β_1 equivale al caso della media campionaria:

$$\text{intervallo di confidenza al 95\% per } \beta_1 = \{\hat{\beta}_1 \pm 1{,}96 \times SE(\hat{\beta}_1)\}$$

Esempio di intervallo di confidenza*: *TestScore* e *STR
Retta di regressione stimata: $\widehat{TestScore} = 698{,}9 - 2{,}28 \times STR$

$SE(\hat{\beta}_0) = 10{,}4$ $\qquad$ $SE(\hat{\beta}_1) = 0{,}52$

Intervallo di confidenza al 95% per $\hat{\beta}_1$:

$$\{\hat{\beta}_1 \pm 1{,}96 \times SE(\hat{\beta}_1)\} = \{-2{,}28 \pm 1{,}96 \times 0{,}52\}$$
$$= (-3{,}30, -1{,}26)$$

Le due affermazioni seguenti sono equivalenti *(perché?)*

- L'intervallo di confidenza al 95% non include lo zero;
- L'ipotesi $\beta_1 = 0$ è rifiutata al livello del 5%

Un modo conciso (e tradizionale) per la notazione delle regressioni:

Porre gli errori standard tra parentesi sotto i coefficienti stimati ai quali sono pertinenti.

$$\widehat{TestScore} = 698{,}9 - 2{,}28 \times STR,\ R^2 = 0{,}05,\ SER = 18{,}6$$
$$(10{,}4)\ \ (0{,}52)$$

Questa espressione fornisce molte informazioni

- La retta di regressione stimata è

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

- L'errore standard di $\hat{\beta}_0$ è 10,4
- L'errore standard di $\hat{\beta}_1$ è 0,52
- R^2 è 0,05; l'errore standard della regressione è 18,6

Regressione OLS: lettura output STATA

```
. regress testscr 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    -4.38   0.000    -3.300945   -1.258671
       _cons |   698.933   10.36436     67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

quindi:

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

(10,4) (0,52)

$t\ (\beta_1 = 0)$ = **–4,38**, valore-p = **0,000** (bilaterale)

L'intervallo conf. bilaterale al 95% per β_1 è (–3,30, –1,26)

Riepilogo di inferenza statistica in merito a β_0 e β_1

Stima:

- Gli stimatori OLS $\hat{\beta}_0$ e $\hat{\beta}_1$
- $\hat{\beta}_0$ e $\hat{\beta}_1$ hanno approssimativamente distribuzioni campionarie normali in grandi campioni

Verifica:

- H_0: $\beta_1 = \beta_{1,0}$ v. $\beta_1 \neq \beta_{1,0}$ ($\beta_{1,0}$ è il valore di β_1 sotto H_0)
- $t = (\hat{\beta}_1 - \beta_{1,0})/SE(\hat{\beta}_1)$
- valore-p = area sotto la normale standard al di fuori di t^{act} (n grande)

Intervalli di confidenza:

- l'intervallo di confidenza al 95% per β_1 è $\{\hat{\beta}_1 \pm 1{,}96 \times SE(\hat{\beta}_1)\}$
- Questo è l'insieme di β_1 che non è rifiutato al livello del 5%
- L'IC al 95% contiene il β_1 reale nel 95% di tutti i campioni.

La regressione quando *X* è una variabile binaria (Paragrafo 5.3)

A volte un regressore è binario:

- X = 1 se classe piccola, = 0 altrimenti
- X = 1 se femmina, = 0 se maschio
- X = 1 se trattato (farmaco sperimentale), = 0 altrimenti

I regressori binari sono a volte chiamati variabili "dummy".

Fin qui β_1 è stato chiamato "pendenza" ma ciò non ha senso se la variabile X è binaria.

Come interpretiamo la regressione con un regressore binario?

Interpretazione delle regressioni con un regressore binario

$Y_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_i + u_i$, dove X è binaria ($X_i = 0$ o 1):

Quando $X_i = 0$, $Y_i = \beta_0 + u_i$

- la media di Y_i è β_0
- cioè $E(Y_i|X_i=0) = \beta_0$

Quando $X_i = 1$, $Y_i = \beta_0 + \beta_1 + u_i$

- la media di Y_i è $\beta_0 + \beta_1$
- cioè $E(Y_i|X_i=1) = \beta_0 + \beta_1$

quindi:

$$\beta_1 = E(Y_i|X_i=1) - E(Y_i|X_i=0)$$
$$= \text{differenza tra medie}$$

Esempio: Sia $D_i = \begin{cases} 1 \text{ se } STR_i \leq 20 \\ 0 \text{ se } STR_i > 20 \end{cases}$

regressione OLS: $\widehat{TestScore} = \mathbf{650,0} + \mathbf{7,4} \times D$

(1,3) (**1,8**)

Tabulazione delle medie:

Dimensione classe	Punteggio medio ($\bar{Y}$)	Dev. std. (s_Y)	N
Piccola ($STR > 20$)	657,4	19,4	238
Grande ($STR \geq 20$)	**650,0**	17,9	182

Differenza tra medie: $\bar{Y}_{\text{small}} - \bar{Y}_{\text{large}} = 657,4 - 650,0 = \mathbf{7,4}$

Errore standard $SE = \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}} = \sqrt{\frac{19,4^2}{238} + \frac{17,9^2}{182}} = \mathbf{1,8}$

Riepilogo: regressione quando la variabile X_i è binaria (0/1)

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

- β_0 = media di Y quando $X = 0$
- $\beta_0 + \beta_1$ = media di Y quando $X = 1$
- β_1 = differenza tra medie, $X = 1$ meno $X = 0$
- SE($\hat{\beta}_1$) ha l'interpretazione consueta
- statistica-t, intervalli di confidenza costruiti come di consueto
- Questo è un altro modo (facile) per eseguire l'analisi della differenza tra medie
- La formulazione della regressione è particolarmente utile quando abbiamo regressori supplementari (*e li avremo molto presto*)

Eteroschedasticità e omoschedasticità, ed errori standard per l'omoschedasticità pura (Paragrafo 5.4)

1. Di che cosa si tratta...?
2. Conseguenze dell'omoschedasticità
3. Implicazioni per il calcolo degli errori standard

Che cosa significano questi due termini?

Se var($u|X=x$) è costante – ossia se la varianza della distribuzione di u condizionata a X non dipende da X – allora u è detto ***omoschedastico***. In caso contrario, u è ***eteroschedastico***.

Esempio: etero/omoschedasticità nel caso di un regressore binario (ossia confronto tra medie)

- Errore standard quando le varianze sono ***ineguali***:

$$SE = \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}$$

- Errore standard quando le varianze sono ***uguali***:

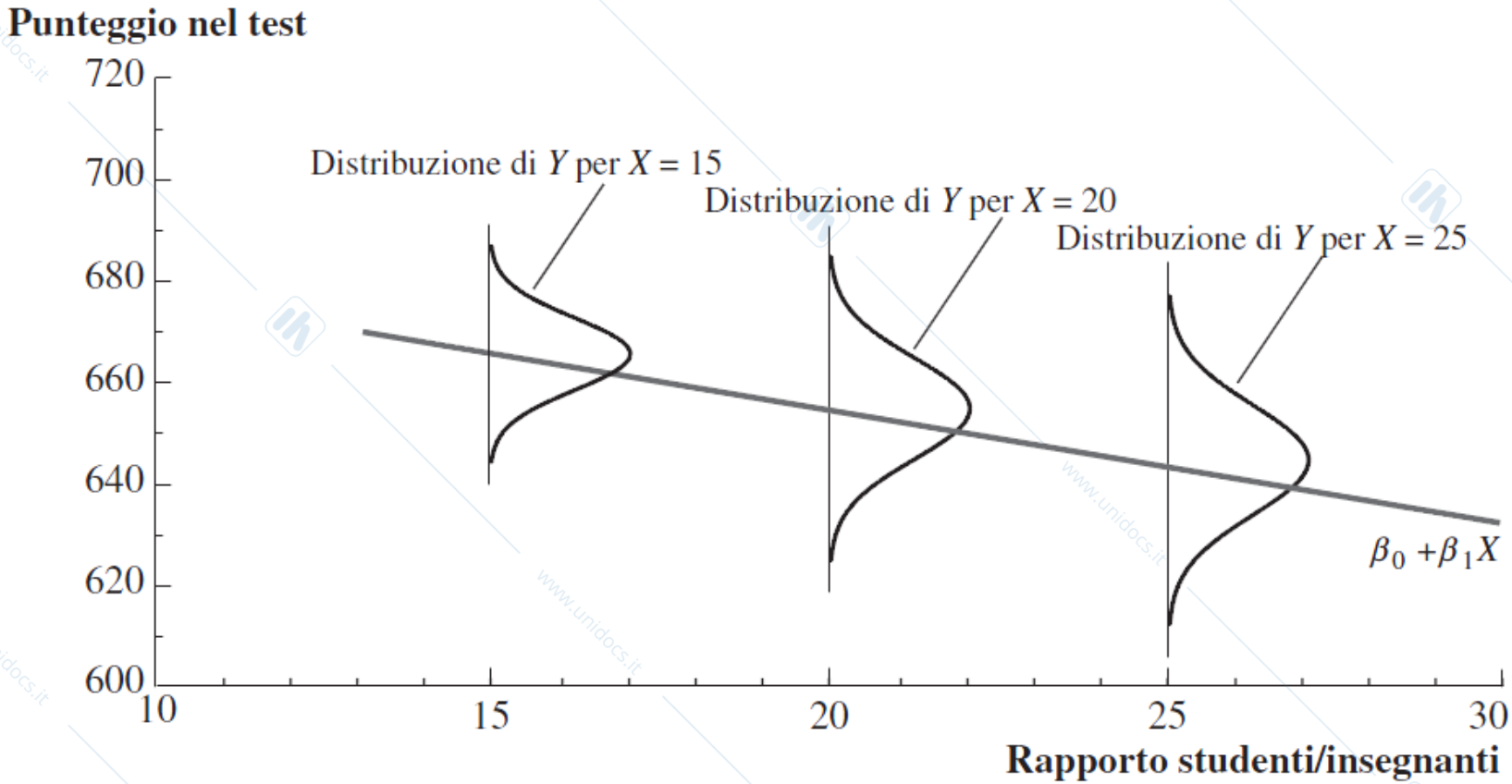
$$SE = s_p \sqrt{\frac{1}{n_s} + \frac{1}{n_l}}$$

Dove $s_p^2 = \frac{(n_s - 1)s_s^2 + (n_l - 1)s_l^2}{n_s + n_l - 2}$ (SW, Sez. 3.6)

s_p = "stimatore aggregato di σ^2" quando $\sigma_l^2 = \sigma_s^2$

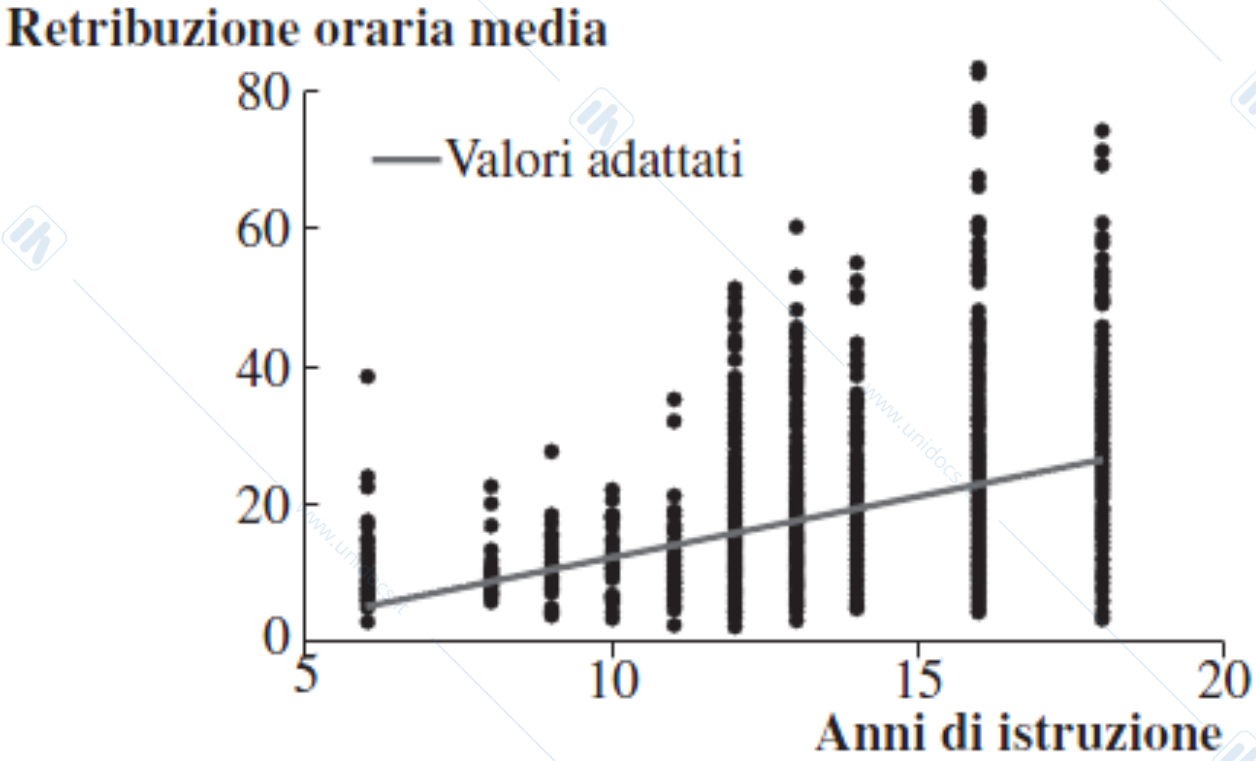
- Varianze **uguali** = **omo**schedasticità
- Varianze **ineguali** = **etero**schedasticità

Omoschedasticità in un'immagine:



- $E(u|X=x) = 0$ (u soddisfa l'assunzione dei minimi quadrati n. 1)
- La varianza di u ***non*** dipende da x

Un esempio con dati reali dall'economica del lavoro: retribuzione oraria media rispetto agli anni di istruzione (fonte dati: Current Population Survey):

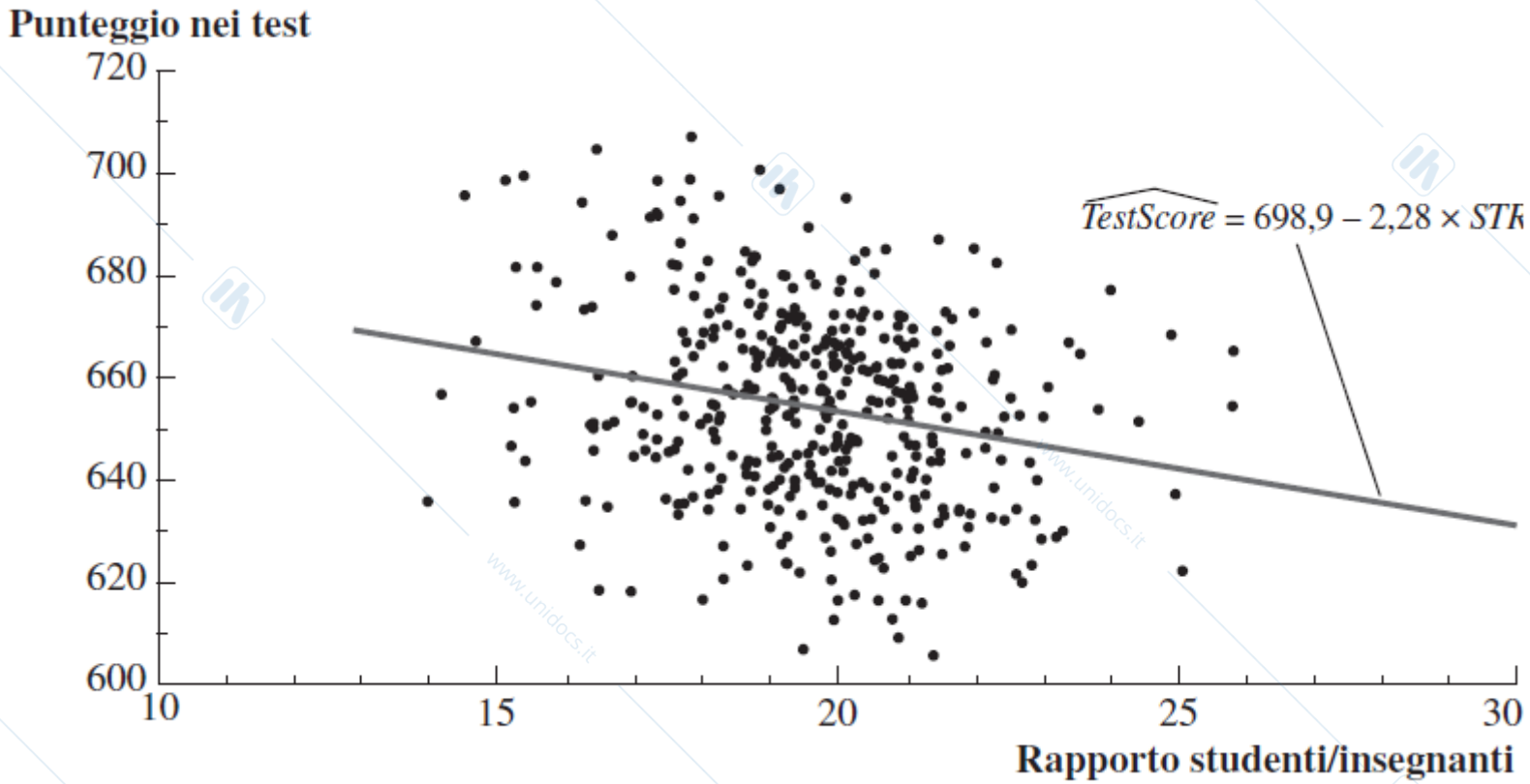


Eteroschedasticità o omoschedasticità?

In media lavoratori più istruiti hanno retribuzioni più elevate. Ma tra i più istruiti (laureate) la dispersion della retribuzione può essere più elevate. Si tratta di una domanda empirica e solo I dati possono rispondere

La figura sembra suggerire che retribuzioni più elevate hanno dispersione maggiore

Dati dimensione classe:



Eteroschedasticità o omoschedasticità?

Finora abbiamo assunto (senza dirlo) che *u* poteva essere eteroschedastico.

Ricordiamo le tre assunzioni dei minimi quadrati:

1. $E(u|X = x) = 0$
2. (X_i, Y_i), $i = 1,\ldots,n$, sono i.i.d.
3. Gli outlier sono rari

Eteroschedasticità e omoschedasticità concernono var($u|X=x$). Poiché non abbiamo assunto esplicitamente gli errori omoschedastici, abbiamo ammesso implicitamente l'eteroschedasticità.

Che cosa succede se gli errori sono in effetti omoschedastici?

- Si può dimostrare che l'OLS ha la varianza minore tra gli stimatori lineari in *Y*… un risultato chiamato teorema di Gauss-Markov, sul quale ritorneremo a breve.
- La formula per la varianza di $\hat{\beta}_1$ e per l'errore standard OLS si semplifica: se var($u_i|X_i=x$)=σ_u^2 , allora

$$\text{var}(\hat{\beta}_1) = \frac{\text{var}[(X_i-\mu_x)u_i]}{n(\sigma_X^2)^2} \text{(formula generale)}$$

$$= \frac{\sigma_u^2}{n\sigma_X^2} \text{(semplificazione se } u \text{ è omoschedastico)}$$

Nota: var($\hat{\beta}_1$) è inversamente proporzionale a var(*X*): maggiore dispersione in *X* significa più informazioni su $\hat{\beta}_1$ – abbiamo discusso di questo in precedenza, ma risulta più chiaro da questa formula.

Derivazione di var($\hat{\beta}_1$) con errori omoschedastici

- var($\hat{\beta}_1$) = $\dfrac{\text{var}[(X_i - \mu_x)u_i]}{n(\sigma_X^2)^2}$ (formula generale)

$= \dfrac{\sigma_u^2}{n\sigma_X^2}$ (semplificazione se *u* è omoschedastico)

Dalle slides cap 4, pag 35 abbiamo che

$$\hat{\beta} - \beta = \frac{\sum(X_i - \bar{X})(u_i - \bar{u})}{\sum(X_i - \bar{X})^2}$$

Da cui

$$\hat{\beta} - \beta = \frac{\sum(X_i - \bar{X})u_i}{\sum(X_i - \bar{X})^2}$$

Definendo $w_i = \frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2}$ con $\sum w_i = 0$ e $\sum w_i(X_i - \bar{X}) = 1$

$$E(\hat{\beta} - \beta)^2 = E\left[\sum_i (w_i u_i)\right]^2 = E\sum_i \left(w_i^2 u_i^2\right)$$

$$E\left[\sum_i w_i^2 E(u_i^2 | X)\right] = E\left[\sum_i w_i^2 \sigma_u^2)\right] =$$

$$\sigma_u^2 \sum_i w_i^2 = \sigma_u^2 \frac{\sum_i (X_i - \bar{X})^2}{[\sum(X_i - \bar{X})^2]^2} = \frac{\sigma_u^2}{\sum(X_i - \bar{X})^2}$$

- Insieme a questa formula di omoschedasticità pura per la varianza di $\hat{\beta}_1$, abbiamo errori standard di omoschedasticità pura:

  Formula errore standard omoschedasticità pura:

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}}.$$

  Alcune persone (per esempio i programmatori di Excel) trovano più semplice la formula dell'omoschedasticità pura – ma ciò è sbagliato a meno che gli errori non siano realmente omoschedastici.

Abbiamo ora due formule per gli errori standard per $\hat{\beta}_1$.

- ***Errori standard per l'omoschedasticità pura*** – sono validi solo se gli errori sono omoschedastici.
- Gli errori standard consueti – per differenziare i due, è convenzione chiamarli ***errori standard robusti all'eteroschedasticità***, poiché sono validi a prescindere dall'eteroschedasticità o meno degli errori.
- Il principale vantaggio degli errori standard per l'omoschedasticità pura è che la formula è più semplice. Lo svantaggio, però, è che la formula è corretta solo se gli errori sono omoschedastici.

Implicazioni pratiche…

- La formula dell'omoschedasticità pura per l'errore standard di $\hat{\beta}_1$ e la formula "robusta all'eteroschedasticità" sono diverse – quindi, in generale, *si ottengono errori standard diversi utilizzando formule differenti*.
- Gli errori standard per l'omoschedasticità pura sono l'impostazione predefinita nei software di regressione – a volte l'unica impostazione (per esempio in Excel). Per ottenere gli errori standard "robusti all'eteroschedasticità" generali occorre modificare l'impostazione di default.
- **Se non si modifica l'impostazione di default e vi è eteroschedasticità, gli errori standard (e la statistica-*t* e gli intervalli di confidenza) saranno errati – generalmente, gli *SE* per l'omoschedasticità pura sono troppo piccoli.**

Errori standard robusti all'eteroschedasticità in STATA

```
regress testscr 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
       _cons |   698.933   10.36436     67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

- Se si utilizza l'opzione "**, robust**", STATA calcola gli errori standard robusti all'eteroschedasticità
- In caso contrario, STATA calcola gli errori standard per l'omoschedasticità pura

Il punto essenziale:

- Se gli errori sono omoschedastici o eteroschedastici e si utilizzano errori standard robusti all'eteroschedasticità, va bene
- Se gli errori sono eteroschedastici e si utilizza la formula dell'omoschedasticità pura per gli errori standard, gli errori standard saranno errati (lo stimatore dell'omoschedasticità pura della varianza di $\hat{\beta}_1$ è incoerente in presenza di eteroschedasticità).
- Le due formule coincidono (quando *n* è grande) nel caso speciale di omoschedasticità
- Quindi si dovrebbero sempre utilizzare errori standard robusti all'eteroschedasticità.

Fondamenti teorici dei minimi quadrati ordinari (Paragrafo 5.5)

lo stimatore OLS è non distorto e consistente; abbiamo una formula per gli errori standard robusti all'eteroschedasticità e possiamo costruire intervalli di confidenza e statistiche di test.

Una buona ragione per utilizzare i minimi quadrati ordinari è anche l'impiego universale, percià gli altri saranno in grado di capire ciò che fate. In effetti, l'OLS è il linguaggio dell'analisi di regressione, e se utilizzate uno stimatore diverso, parlerete un linguaggio differente.

Eppure potreste ancora chiedervi…

- Tutto quanto detto è davvero una buona ragione per utilizzare OLS? Non esistono altri stimatori che potrebbero essere migliori – in particolare che potrebbero avere una varianza inferiore?
- Inoltre, che ne è stato della distribuzione *t* di Student?

Ora risponderemo a queste domande – ma per farlo abbiamo bisogno di assunzioni più forti delle tre relative ai minimi quadrati che abbiamo già visto.

Le assunzioni dei minimi quadrati estese

Consistono nelle tre assunzioni dei minimi quadrati, più altre due:

1. $E(u|X = x) = 0$.
2. (X_i, Y_i), $i = 1,\ldots,n$, sono i.i.d.
3. Gli outlier sono rari ($E(Y^4) < \infty$, $E(X^4) < \infty$).
4. u è omoschedastico
5. u ha distribuzione $N(0,\sigma^2)$

- Le assunzioni 4 e 5 sono più restrittive – perciò si applicano a un numero inferiori di casi pratici. Tuttavia, facendo queste assunzioni, determinati calcoli matematici si semplificano e si possono dimostrare risultati più robusti – che valgono se tali assunzioni aggiuntive sono vere.
- Iniziamo con una discussione sull'efficienza dello stimatore OLS

Efficienza dello stimatore OLS, parte I: il teorema di Gauss-Markov

Nelle assunzioni dei minimi quadrati ordinari estese 1-4 (le tre di base, più l'omoschedasticità), $\hat{\beta}_1$ ha la varianza minima tra *tutti gli stimatori lineari* (stimatori che sono funzioni lineari di $Y_1,\ldots, Y_n$). Questo è il ***teorema di Gauss-Markov***.

Commenti

- Il teorema di GM è dimostrato nell'Appendice 5.2

Il teorema di Gauss-Markov (continua)

- $\hat{\beta}_1$ è uno stimatore lineare, cioè può essere scritto come funzione lineare di $Y_1,\ldots, Y_n$:

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

dove $w_i = \dfrac{(X_i - \bar{X})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

- Il teorema di G-M dice che, fra tutte le scelte possibili di $\{w_i\}$, i pesi OLS portano alla minima var($\hat{\beta}_1$)

Efficienza dello stimatore OLS, parte II:

- In <u>tutte e cinque</u> le assunzioni dei minimi quadrati estese – compresa la distribuzione normale degli errori – $\hat{\beta}_1$ ha la varianza più piccola di *<u>tutti</u>* gli estimatori consistenti (funzioni lineari *o* non lineari di $Y_1,\ldots,Y_n$), per $n \to \infty$.

- Questo è un risultato assai sorprendente – afferma che, se (in aggiunta alle assunzioni dei minimi quadrati 1-3) gli errori sono omoschedastici e normalmente distribuiti, OLS è la scelta migliore di qualsiasi altro stimatore consistente. E poiché uno stimatore che non sia consistente è una scelta scadente, ciò afferma che l'OLS è davvero la miglior scelta che si possa fare – se valgono tutte e cinque le assunzioni dei minimi quadrati estese.

Alcuni aspetti critici di OLS

I risultati precedenti sono impressionanti, tuttavia tali risultati – e lo stimatore OLS – hanno limitazioni importanti.

1. Il teorema di GM non è poi così avvincente:
   - La condizione di omoschedasticità spesso non regge (l'omoschedasticità è speciale)
   - Il risultato vale solo per gli stimatori lineari – solo un piccolo sottoinsieme di stimatori (ulteriori informazioni a breve)

2. Il risultato di ottimalità più robusto ("parte II" precedente) richiede errori normali omoschedastici – cosa non plausibile nelle applicazioni (si pensi ai dati delle retribuzioni orarie!)

Limitazioni di OLS (continua)

3. L'OLS è più sensibile agli outlier rispetto ad altri stimatori. Nel caso di stima della media della popolazione, se sussistono outlier, la mediana è preferibile alla media poiché la prima è meno sensibile agli outlier – ha varianza minore rispetto all'OLS in presenza di outlier. In modo analogo, nella regressione l'OLS può essere più sensibile agli outlier e, se vi sono outlier, altri stimatori possono essere più efficienti (hanno varianza minore). Uno di essi è lo stimatore delle minime deviazioni assolute (MDA):

$$\min_{b_0,b_1} \sum_{i=1}^{n} \left| Y_i - (b_0 + b_1 X_i) \right|$$

Praticamente tutte le analisi di regressione reali si utilizza l'OLS – ed è quello che faremo anche in questo corso.

Inferenza se *u* è omoschedastico e normalmente distribuito: la distribuzione *t* di Student (Paragrafo 5.6)

Ricordiamo le cinque assunzioni dei minimi quadrati estese:

1. $E(u|X = x) = 0$.
2. (X_i, Y_i), $i = 1,\ldots,n$, sono i.i.d.
3. Gli outlier sono rari ($E(Y^4) < \infty$, $E(X^4) < \infty$).
4. u è omoschedastico
5. u ha distribuzione $N(0,\sigma^2)$

Se tutte le cinque assunzioni valgono, allora:

- $\hat{\beta}_0$ e $\hat{\beta}_1$ sono normalmente distribuiti *per tutti gli n* (!)
- la statistica-*t* ha una distribuzione *t* di Student con $n - 2$ gradi di libertà – questo vale esattamente *per tutti gli n* (!)

Normalità della distribuzione campionaria sotto le assunzioni 1–5:

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

$$= \frac{1}{n}\sum_{i=1}^{n} w_i u_i \text{ , dove } w_i = \frac{(X_i - \bar{X})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Qual è la distribuzione di una media ponderata di normali?

Sotto le assunzioni 1 – 5:

$$\hat{\beta}_1 - \beta_1 \sim N\left(0, \frac{1}{n^2}\left(\sum_{i=1}^{n} w_i^2\right)\sigma_u^2\right) \qquad (*)$$

Sostituendo w_i in (*) si produce la formula della varianza dell'omoschedasticità pura.

Inoltre, sotto le assunzioni 1 – 5, nell'ipotesi nulla la statistica *t* ha una distribuzione *t* di Student con *n – 2* gradi di libertà

- Perché $n - 2$? Perché abbiamo stimato 2 parametri, β_0 e β_1
- Per $n < 30$ i valori critici t possono essere un po' più grandi dei valori critici $N(0,1)$
- Per $n > 50$ o simile, la differenza nelle distribuzioni t_{n-2} e $N(0,1)$ è trascurabile. Ricordiamo la tabella *t* di Student:

gradi di libertà	*valore critico distribuzione-t al 5%*
10	2,23
20	2,09
30	2,04
60	2,00
∞	1,96

Implicazioni pratiche:

- Se $n < 50$ ***e*** credete davvero che, per la vostra applicazione, u sia omoschedastico e normalmente distribuito, utilizzate t_{n-2} invece dei valor critici $N(0,1)$ per le verifiche di ipotesi e gli intervalli di confidenza.

- Nella maggior parte delle applicazioni econometriche, non vi è alcun motivo di ritenere che u sia omoschedastico e normale – solitamente vi sono ottime ragioni per credere che né l'una né l'altra assunzione valga.

- Fortunatamente, nelle applicazioni moderne $n > 50$, così possiamo affidarci ai risultati per n grande presentati in precedenza, basati sul teorema limite centrale, per eseguire verifiche di ipotesi e costruire intervalli di confidenza usando l'approssimazione normale per n grande.

Riepilogo e valutazione (Paragrafo 5.7)

- Il problema iniziale:

  Si supponga che vengano assunti nuovi insegnanti cosicché il rapporto studenti/insegnanti si riduca di uno studente per classe. Qual è l'effetto dell'intervento di questa politica ("trattamento") sui punteggi nei test?
- La nostra analisi di regressione utilizzando i dati della California risponde in modo convincente?

  Non proprio – i distretti con *STR* basso tendono a essere quelli con molte altre risorse e famiglie a reddito più elevato, che offrono ai figli maggiori opportunità di apprendimento al di fuori della scuola… ciò suggerisce che corr(u_i, STR_i) > 0, quindi $E(u_i|X_i) \neq 0$.
- Sembra che abbiamo omesso alcuni fattori o variabili dalla nostra analisi e ciò ha distorto i risultati...

DERIVATION OF $\hat{\alpha}$

$$\min_{\alpha,\beta} \sum_{i=1}^{N} u^2 = \sum_{i=1}^{N} (y_i - \alpha - \beta x_i)^2$$

$$\frac{d\,(y_i - \alpha - \beta x_i)^2}{d\beta} = -2 \sum_{i=1}^{N} (y_i - \alpha - \beta$$

$$\frac{d\,(y_i - \alpha - \beta x_i)^2}{d\alpha} = -\sum_{i=1}^{N} (y_i - \alpha - \beta$$

$$\sum (y_i - \alpha - \beta x_i) x_i = 0$$

$$\sum y_i x_i - \alpha \sum x_i - \beta \sum x_i^2 = 0$$

$\uparrow \hat{\alpha}$

PLUG IN HERE

$$\sum y_i x_i - \sum x_i \left(\bar{y} - \beta \bar{x} \right) - \beta \sum x_i^2$$

$\uparrow$

UNBIASEDNESS OF $\hat{\alpha}$

Unbiasedness of alpha comes from the linearity of $\hat{\beta}$

$$\hat{\beta} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{\sum_i (X_i - \bar{X})Y_i - \bar{Y}\sum_i (X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} = \frac{\sum_i (X_i - \bar{X})Y_i}{\sum_i (X_i - \bar{X})^2}$$

Because $\bar{Y}\sum_i (X_i - \bar{X}) = 0$

By setting

$$\frac{(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} = w_i$$

with $\sum_i w_i = 0$ and $\sum_i w_i (X_i - \bar{X}) = 1$

We can rewrite $\hat{\beta} = \sum_i w_i Y_i$

since w_i does not contain terms in Y the estimator is linear

$$\sum_i w_i = 0;\ \sum_i w_i (X_i - \bar{X}) = 1 = \sum_i w_i X_i$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$= \bar{Y} - \bar{X}\sum_i w_i Y_i$$

replacing $\bar{Y}$ and grouping terms

$$= \sum_i Y_i \left(\frac{1}{n} - \bar{X} w_i\right)$$

$$= \sum_i (\alpha + \beta X_i + u_i)\left(\frac{1}{n} - \bar{X} w_i\right)$$

$$= \alpha - \alpha\bar{X}\sum_i w_i + \beta\bar{X} - \beta\bar{X}\sum_i X_i w_i + \sum_i u_i \left(\frac{1}{n} - \bar{X} w_i\right)$$

$$= \alpha + \sum_i u_i \left(\frac{1}{n} - \bar{X} w_i\right)$$

$\mathrm{E}(\hat{\alpha} - \alpha) = E\left[\sum_i u_i \left(\frac{1}{n} - \bar{X} w_i\right)\right]$ applying the law or iterated expectations

Variance of $\hat{\alpha}$ under homoscedasticity assumption

$$E(\hat{\alpha}-\alpha)^2 = E\left\{\sum_i \left(\frac{1}{n}-\bar{X}\,w_i\right)E[u_i|X]\right\}^2 = E\left\{\sum_i \left(\frac{1}{n}-\bar{X}\,w_i\right)^2 \underbrace{E[u_i|X]^2}_{\sigma_u^2}\right\}$$

$$= \sigma_u^2\left\{\sum_i \left(\frac{1}{n}-\bar{X}\,w_i\right)^2\right\} = \sigma_u^2 \sum_i \left(\frac{1}{n^2}+\bar{X}^2 w_i^2 - \frac{2}{n}\bar{X}\,w_i\right) = \sigma_u^2\left[\frac{1}{n}+\bar{X}^2\sum_i w_i^2 - \frac{2}{n}\bar{X}\,\underbrace{\sum_i w_i}_{0}\right]$$

$$var(\hat{\alpha}) = \sigma_u^2\left(\frac{1}{n}+\bar{X}^2\sum_i w_i^2\right) \text{ where } \sum_i w_i^2 = \frac{\sum_i (X_i-\bar{X})^2}{[\sum_i (X_i-\bar{X})^2]^2} = \frac{1}{\sum_i (X_i-\bar{X})^2}$$

$$var(\hat{\alpha}) = \sigma_u^2\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum_i (X_i-\bar{X})^2}\right) = \sigma_u^2\left(\frac{1}{n}+\frac{\bar{X}^2}{n\,var(X)}\right)$$

Econometria I

Corso di laurea magistrale in Finanza e Assicurazioni

Slides per lezioni

secondo semestre 2020

Marco Ventura

Cap 6 SW

Struttura

1. Distorsione da variabili omesse
2. Causalità e analisi di regressione
3. Regressione multipla e OLS
4. Misure di adattamento
5. 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:

Le due condizioni per la distorsione da variabile omessa

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*) ≠ 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,

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

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

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. *(664 vs 636,7)*
- I distretti con una minore percentuale di studenti non di madrelingua hanno classi più piccole. *(Tra quelli <20 studenti ci sono 76 distretti vs 27)*
- 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). *(leggi orizzontalmente le righe, le t non sono significative)*

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

- Il senso di questa tavola è che se non tengo conto degli studenti non di madrelingua la differenza è elevata (7,4), se invece stratifico su questa variabile la differenza si riduce considerevolmente.

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?

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 su 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 casualizzato 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 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 "controllare" 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 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 (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

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

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:

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

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:

$$\beta_1 = \frac{\Delta Y}{\Delta X_1}$$, **tenendo X_2 costante**

$$\beta_2 = \frac{\Delta Y}{\Delta X_2}$$, **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 β_0 e β_1.**

Esempio: i dati dei punteggi nei test della California

Regressione di *TestScore* su *STR*:

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

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*?
- corr(*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 output …

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; $\bar{R}^2 < R^2$

K= numero di regressori esclusa la costante

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' $\bar{R}^2$ ("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.

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

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

$R^2 = 0{,}05,\ SER = 18{,}6$

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

$R^2 = 0{,}426,\ \bar{R}^2 = 0{,}424,\ SER = 14{,}5$

- *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 (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: $(X_{1i},\ldots,X_{ki},Y_i)$, $i = 1,\ldots,n$, 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* $\hat{\beta}_1$, $\hat{\beta}_2$, …, $\hat{\beta}_k$.

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)
  - $\dfrac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}}$ 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.

<u>Altri esempi di collinearità perfetta</u>

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?*

La trappola delle variabili dummy

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

➔ In questo caso l'intercetta stima il valore medio di Y per il Gruppo D=0 e il coefficiente di D misura la differenza in media in Y dei due gruppi (Gruppo 1 – Gruppo 0)

2. omettere l'intercetta

➔ Il coefficient del Gruppo D=0 è pari all'intercetta del caso precedente, per stimare la differenza tra i due gruppi dobbiamo fare la differenza dei due coefficienti.

PREFERITE LA SOLUZIONE 1

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 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...

Econometria I

Corso di laurea magistrale in Finanza e Assicurazioni

Slides per lezioni

secondo semestre 2020

Marco Ventura

Cap. 7 SW

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 β_1 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 $\beta_2, \ldots, \beta_k$.

Esempio: dati sulle dimensioni delle classi in California

1. $\widehat{TestScore} = 698{,}9 - 2{,}28 \times STR$
   $\quad (10{,}4) \quad (0{,}52)$

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

- Il coefficiente di *STR* in (2) è l'effetto su *TestScore* della variazione unitaria in *STR*, mantenendo costante PctEL
- 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 al livello di significatività del 5%

Errori standard nella regressione multipla in STATA

```
reg testscr str pctel, robust;
Regression with robust standard errors
```

NB

```
Number of obs =     420
F(  2,   417) =  223.82
Prob > F      =  0.0000
R-squared     =  0.4264
Root MSE      =  14.464
```

testscr	Coef.	Robust 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:

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

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

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" può sembrare 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*:

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

La verifica "coeff. per coeff." è:

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

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 *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.

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 nelle code della distribuzione χ^2_q/q oltre la statistica *F* effettivamente calcolata.

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:

- 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):

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

- 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:

$$\widehat{TestScore} = 644{,}7 - 0{,}671 PctEL, \qquad R^2_{restricted} = 0{,}4149$$
$$(1{,}0) \quad (0{,}032)$$

Regressione non vincolata:

$$\widehat{TestScore} = 649{,}6 - 0{,}29 STR + 3{,}87 Expn - 0{,}656 PctEL$$
$$(15{,}5) \quad (0{,}48) \qquad (1{,}59) \qquad (0{,}032)$$

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

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

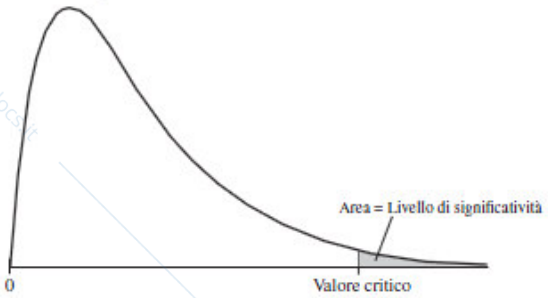
allora la statistica *F* classica ha la distribuzione "$F_{q,n-k-1}$", dove q = numero delle restrizioni e k = numero dei regressori sotto l'alternativa (modello non vincolato).

- **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 $F_{m,n}$.



	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 $F_{m,n}$. 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 (= $F_{q,\infty}$) **sia con omo che eterochedasticità**
- 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 $\beta_1 = 0$ e $\beta_2 = 0$).

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

$$H_0: \gamma_1 = 0 \quad \text{vs.} \quad H_1: \gamma_1 \neq 0$$

- 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

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

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

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

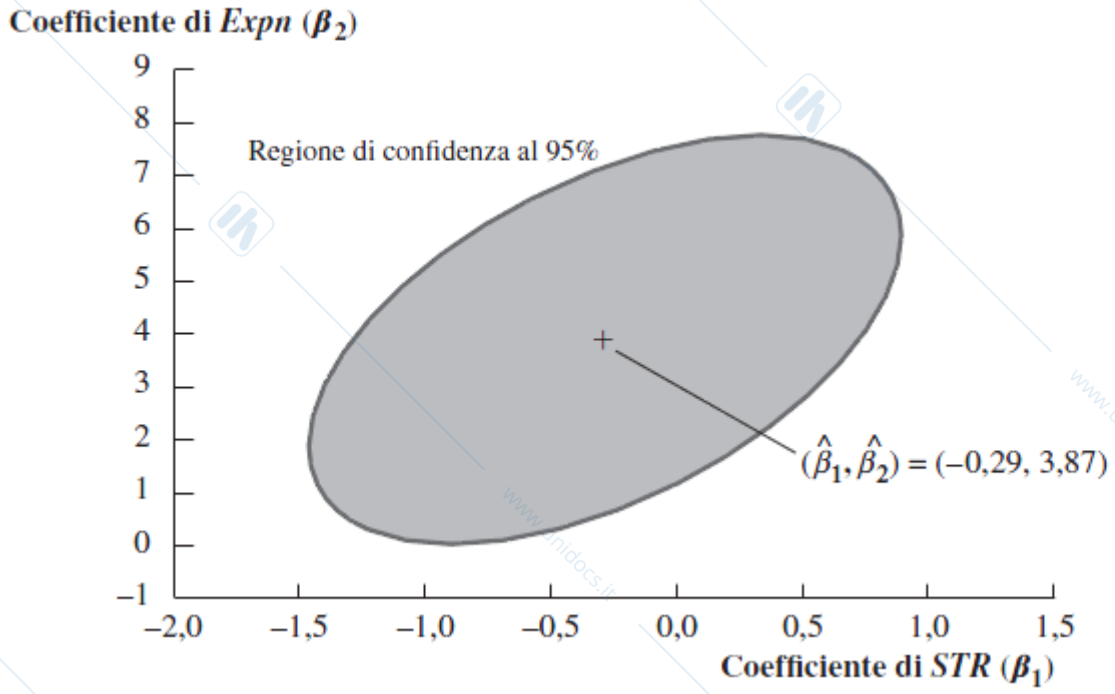


Figura 7.1

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

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

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)

- *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 variabile di controllo, i.e. a parità di varibile 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:

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

- Il coefficiente di *LchPct* ha un'interpretazione causale? Ha segno (-) quindi potremmo pensare di eliminare i buoni pasto per aumentare il punteggio dei test.

  NO! qualche bambino resterebbe affamato producendo un effetto contrario sull'apprendimento.

La matematica delle variabili di controllo: indipendenza in media condizionata (very IMP!!)

- Poiché il coefficiente di una variabile di controllo può essere distorto, la prima assunzione dei minimi quadrati ($E(u_i|X_{1i},\ldots,X_{ki}) = 0$) può non 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

Sia X_i la variabile di interesse (causale) 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.

Sotto l'hp di indipendenza in media condizionata il coefficiente di X_i ha lettura causale, ma il coefficient di W_i no.

Il concetto di indipendenza in media condizionata richiede che una volta controllato per W, allora X può essere considerata come se fosse assegnata casualmente, nel senso che la media di *u* non dipende più da essa (X).

L'inclusione di W rende X non correlata col termine di errore, perciò OLS può stimare l'effetto causale su Y di X.

W rimane correlata col termine di errore e il suo coefficiente, può essere distorto dunque non ha una lettura causale

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 (very IMP!!)

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:

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

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,

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

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,

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

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 sensitività").
   - 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 $\bar{R}^2$, ma ciò riduce la visibilità sull'obiettivo reale, uno stimatore non distorto dell'effetto della dimensione della classe.

- Un elevato R^2 (o $\bar{R}^2$) significa che i regressori spiegano la variazione in *Y*.
- Un elevato R^2 (o $\bar{R}^2$) *non* significa che avete eliminato la distorsione delle variabili omesse.
- Un elevato R^2 (o $\bar{R}^2$) *non* significa che avete uno stimatore non distorto di effetto causale (β_1).
- Un elevato R^2 (o $\bar{R}^2$) *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".

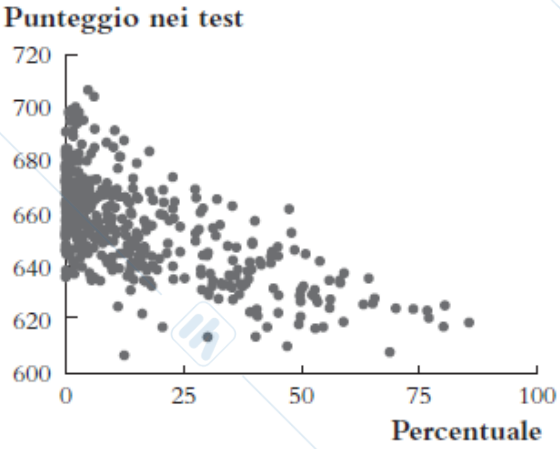
   Molte delle variabili causali omesse sono difficili da misurare, perciò dobbiamo trovare le variabili di controllo. Queste includono PctEL (sia una variabile di controllo sia un fattore causale omesso) e misure del benessere del distretto.

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

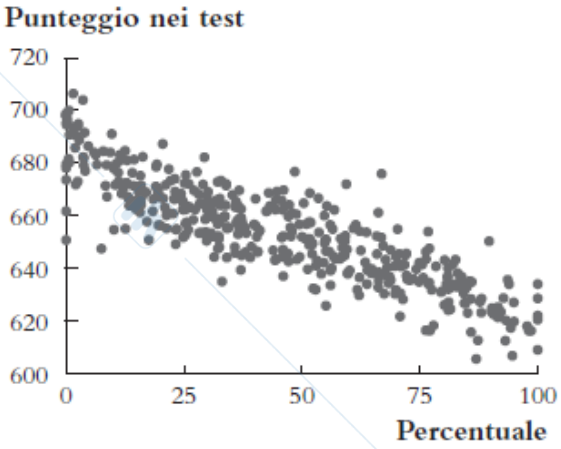
   Non è chiara quale delle variabili relative al reddito controlli al meglio i molteplici fattori causali omessi, quali le opportunità di apprendimento esterno, perciò le specificazioni delle alternative comprendono regressioni con variabili di reddito diverse. Le specificazioni delle alternative considerate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

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

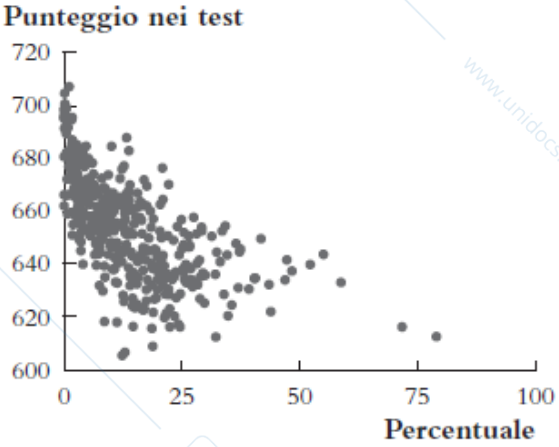
Punteggi nei test e dati socioeconomici della California...



(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
$\bar{R}^2$	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.

Takashi Yamano

Lecture Notes on Advanced Econometrics

Lecture 4: Multivariate Regression Model in Matrix Form

In this lecture, we rewrite the multiple regression model in the matrix form. A general multiple-regression model can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i \qquad \text{for } i = 1, \ldots, n.$$

In matrix form, we can rewrite this model as

$$\begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ & & & & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \\ u_n \end{bmatrix}$$

n x 1 n x (k+1) (k+1) x 1 n x 1

$$Y = X\beta + u$$

We want to estimate β.

Least Squared Residual Approach in Matrix Form

(Please see Lecture Note A1 for details)

The strategy in the least squared residual approach is the same as in the bivariate linear

Notice here that $u'u$ is a scalar or number (such as 10,000) because u' is a 1 x n matrix and u is a n x 1 matrix and the product of these two matrices is a 1 x 1 matrix (thus a scalar). Then, we can take the first derivative of this object function in matrix form. First, we simplify the matrices:

$$u'u = (Y' - \beta'X')(Y - X\beta)$$

$$= Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta$$

$$= Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

Then, by taking the first derivative with respect to β, we have:

$$\frac{\partial(u'u)}{\partial\beta} = -2X'Y + 2X'X\beta$$

From the first order condition (F.O.C.), we have

$$-2X'Y + 2X'X\hat{\beta} = 0$$

$$X'X\hat{\beta} = X'Y$$

Notice that I have replaced β with $\hat{\beta}$ because $\hat{\beta}$ satisfy the F.O.C, by definition.

Multiply the inverse matrix of $(X'X)^{-1}$ on the both sides, and we have:

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad (1)$$

$$X'(Y - X\hat{\beta}) = 0$$

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ & & & \\ x_{1k} & x_{2k} & \dots & x_{nk} \end{bmatrix} \left(\begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ & & & \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \\ \hat{\beta}_k \end{bmatrix} \right) = 0$$

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ & & & \\ x_{1k} & x_{2k} & \dots & x_{nk} \end{bmatrix} \left(\begin{bmatrix} y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_{11} \dots - \hat{\beta}_k x_{1k} \\ y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_{21} \dots - \hat{\beta}_k x_{2k} \\ \\ y_n - \hat{\beta}_0 - \hat{\beta}_1 x_{n1} \dots - \hat{\beta}_k x_{nk} \end{bmatrix} \right) = 0$$

(k+1) x n n x 1

This is the same as the first order conditions, k+1 conditions, we derived in the previous lecture note (on the simple regression model):

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - b_k x_{ik}) = 0$$

$$\sum_{i=1}^{n} x_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - b_k x_{ik}) = 0$$

$$\sum_{i=1}^{n} x_{ik} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - b_k x_{ik}) = 0$$

Example 4-1 **: A bivariate linear regression (k=1) in matrix form**

As an example, let's consider a bivariate model in matrix form. A bivariate model is

$$\begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \\ u_n \end{bmatrix}$$

From (1), we have

$$\hat{\beta} = (X'X)^{-1} X'Y \qquad (2)$$

Let's consider each component in (2).

$$X'X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ & \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

This is a 2 x 2 square matrix. Thus, the inverse matrix of $X'X$ is,

$$(X'X)^{-1} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}$$

$$= \frac{1}{} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -n\bar{x} \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} = \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

Thus the OLS estimators are:

$$\hat{\beta} = (X'X)^{-1} X'Y = \frac{1}{n(\sum_{i=1}^{n} x_i - \bar{x})^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

$$= \frac{1}{n(\sum_{i=1}^{n} x_i - \bar{x})^2} \begin{bmatrix} n\bar{y}\sum_{i=1}^{n} x_i^2 & -n\bar{x}\sum_{i=1}^{n} x_i y_i \\ -n\bar{x}\, n\bar{y} & + n\sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

$$= \frac{1}{(\sum_{i=1}^{n} x_i - \bar{x})^2} \begin{bmatrix} \bar{y}\sum_{i=1}^{n} x_i^2 & -\bar{x}\sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \end{bmatrix}$$

$$= \frac{1}{(\sum_{i=1}^{n} x_i - \bar{x})^2} \begin{bmatrix} \bar{y}\sum_{i=1}^{n} x_i^2 - \bar{y}\bar{x}^2 + \bar{y}\bar{x}^2 - \bar{x}\sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} (x_i y_i - \bar{x}\bar{y}) \end{bmatrix}$$

$$= \frac{1}{(\sum_{i=1}^{n} x_i - \bar{x})^2} \begin{bmatrix} \bar{y}(\sum_{i=1}^{n} x_i^2 - \bar{x}^2) - \bar{x}(\sum_{i=1}^{n} x_i y_i - \bar{y}\bar{x}) \\ \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \end{bmatrix}$$

This is what you studied in the previous lecture note.

End of Example 4-1

Unbiasedness of OLS

In this sub-section, we show the unbiasedness of OLS under the following assumptions.

Assumptions:

E 1 (Linear in parameters): $Y = X\beta + u$
E 2 (Zero conditional mean): $E(u \mid X) = 0$
E 3 (No perfect collinearity): **X has rank k.**

From (2), we know the OLS estimators are

$$\hat{\beta} = (X'X)^{-1} X'Y$$

We can replace **y** with the population model (**E 1**),

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1} X'(X\beta + u) \\ &= (X'X)^{-1} X'X\beta + (X'X)^{-1} X'u \\ &= \beta + (X'X)^{-1} X'u\end{aligned}$$

By taking the expectation on the both sides of the equation, we have:

$$E(\hat{\beta}) = \beta + (X'X)^{-1} E(X'u)$$

Next, we consider the variance of the estimators.

Assumption:

E 4 (Homoskedasticity): $Var(u_i \mid X) = \sigma^2$ and $Cov(u_i, u_j) = 0$, thus $Var(u \mid X) = \sigma^2 I$.

Because of this assumption, we have

$$E(uu') = E\begin{bmatrix} u_1 \\ u_2 \\ \\ u_n \end{bmatrix} \begin{bmatrix} u_1 & u_2 & & u_n \end{bmatrix} = \begin{bmatrix} E(u_1u_1) & E(u_1u_2) & E(u_1u_n) \\ E(u_2u_1) & E(u_2u_2) & E(u_2u_n) \\ \\ E(u_nu_1) & E(u_nu_2) & E(u_nu_n) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ \\ 0 & 0 & \sigma^2 \end{bmatrix} = \sigma^2 I$$

n x 1 1 x n n x n n x n n x n

Therefore,

$$\begin{aligned} Var(\hat{\beta}) &= Var[\beta + (X'X)^{-1}X'u] \\ &= Var[(X'X)^{-1}X'u] \\ &= E[(X'X)^{-1}X'uu'X(X'X)^{-1}] \\ &= (X'X)^{-1}X'E(uu')X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2 I\ X(X'X)^{-1} \qquad \text{(E4: Homoskedasticity)} \end{aligned}$$

$$Var(\hat{\beta}) = \sigma^2 (X'X)^{-1} \qquad (3)$$

Example 4-2: Step by Step Regression Estimation by STATA

In this sub-section, I would like to show you how the matrix calculations we have studied are used in econometrics packages. Of course, in practices you do not create matrix programs: econometrics packages already have built-in programs.

The following are matrix calculations with STATA using data called, NFIncomeUganda.dta. Here we want to estimate the following model:

$$\ln(income)y_i = \beta_0 + \beta_1 female_i + \beta_2 edu_i + \beta_3 edusq_i + u_i$$

All the variables are defined in Example 3-1. Descriptive information about the variables are here:

```
. su;

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      female |       648    .2222222    .4160609          0          1
         edu |       648    6.476852    4.198633         -8         19
       edusq |       648    59.55093    63.28897          0        361
-------------+--------------------------------------------------------
   ln_income |       648    12.81736    1.505715   7.600903   16.88356
```

First, we need to define matrices. In STATA, you can load specific variables (data) into matrices. The command is called **mkmat**. Here we create a matrix, called ***y***, containing the dependent variable, *ln_nfincome*, and a set of independent variables, called **x**,

```
. matrix xx=x'*x;
. mat list xx;
symmetric xx[4,4]
         female      edu     edusq     const
female      144
   edu      878    38589
 edusq     8408   407073   4889565
 const      144     4197     38589       648

. matrix ixx=syminv(xx);
. mat list ixx;
symmetric ixx[4,4]
            female         edu       edusq       const
female    .0090144
   edu    .00021374   .00053764
 edusq  -.00001238  -.00003259   2.361e-06
 const  -.00265043   -.0015892   .00007321   .00806547
```

Here is $X'Y$:

```
. matrix xy=x'*y;
. mat list xy;
xy[4,1]
        ln_nfincome
female    1775.6364
```

```
. mat list bhat;
bhat[4,1]
           ln_nfincome
female      -.59366458
   edu       .04428822
 edusq       .00688388
 const       12.252496

. ** Estimating standard error for b hat;
. matrix e=y-x*bhat;
. matrix ss=(e'*e)/(648-1-3);
. matrix kk=vecdiag(ixx);

. mat list ss;
symmetric ss[1,1]
             ln_nfincome
ln_nfincome    1.8356443

. mat list kk;
kk[1,4]
       female         edu       edusq       const
r1   .0090144   .00053764   2.361e-06   .00806547
```

Let's verify what we have found.

```
. reg ln_nfincome female edu edusq;

      Source |       SS       df       MS              Number of obs =     648
```

```
-------------+----------------------------------------------------------------
      female |  -.5936646   .1286361    -4.62   0.000    -.8462613   -.3410678
         edu |   .0442882   .0314153     1.41   0.159    -.0174005     .105977
       edusq |   .0068839   .0020818     3.31   0.001      .002796    .0109718
       _cons |    12.2525   .1216772   100.70   0.000     12.01356    12.49143
------------------------------------------------------------------------------

end of do-file
```

Lecture 5: OLS Inference under Finite-Sample Properties

So far, we have obtained OLS estimations for $E(\hat{\beta})$ *and* $Var(\hat{\beta})$. But we need to know the shape of the full sampling distribution of $\hat{\beta}$ in order to conduct statistical tests, such as *t*-tests or F-tests. The distribution of OLS estimator $\hat{\beta}$ depends on the underlying distribution of the errors. Thus, we make the following assumption (again, under finite-sample properties).

Assumption

E 5 (Normality of Errors): $u_{n\times 1} \sim N(0_{n\times 1}, \sigma^2 I_{n\times n})$

Note that $N(0_{x\times 1}, \sigma^2 I_{n\times n})$ indicates a multivariate normal distribution of *u* with mean $\mathbf{0}_{\text{n x 1}}$ and the variance-covariance matrix $\sigma^2 I_{n\times n}$.

Remember again that only assumptions E1-3 are necessary to have unbiased OLS estimators. In addition, assumption 4 is needed to show that the OLS estimators are the best linear unbiased estimator (BLUE), the Gauss-Markov theorem. We need assumption 5 to conduct statistical tests.

Normality of $\hat{\beta}$

Under the CLM assumptions (E1-5), $\hat{\beta}$ (conditional on **X**) is distributed as multivariate normal with mean B and variance-covariance matrix $\sigma^2(X'X)^{-1}$.

$$\hat{\beta} \sim N[\beta, \sigma^2(X'X)^{-1}]$$

This is a multivariate normal distribution, which means each element of $\hat{\beta}$ is normally distributed:

$$\hat{\beta}_k \sim N[\beta_k, \sigma^2(X'X)^{-1}{}_{kk}]$$

$(X'X)^{-1}{}_{kk}$ is the k-th diagonal element of $(X'X)^{-1}$. Let's denote the k-th diagonal element of $(X'X)^{-1}$ as S_{kk}. Then,

$$\sigma^2(X'X)^{-1} = \sigma^2 \begin{pmatrix} S_{11} & . & . & . & . \\ . & S_{22} & . & . & . \\ & & & & \\ . & . & . & . & S_{kk} \end{pmatrix} = \begin{pmatrix} \sigma^2 S_{11} & . & . & . & . \\ . & \sigma^2 S_{22} & . & . & . \\ & & & & \\ . & . & . & . & \sigma^2 S_{kk} \end{pmatrix}$$

This is the variance-covariance matrix of the OLS estimator. On the diagonal, there are variances of the OLS estimators. Off-the diagonal, there are covariance between the estimators. Because each OLS estimator is assumed to be normally distributed, we can obtain a standard normal distribution of an OSL estimator by subtracting the mean and dividing it by the standard deviation:

$$s^2 = \frac{\hat{u}'\hat{u}}{n-(k+1)}$$

$\hat{u}'\hat{u}$ is the sum of squared errors. (Remember $\hat{u}'\hat{u}$ is a product of a (1 x n) matrix and a (n x 1) matrix, which gives a single number.) Therefore by replacing σ^2 with s^2, we have

$$t_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{s^2 S_{kk}}} \quad .$$

This ratio has a *t*-distribution with (n-k-1) degree of freedom. It has a *t*-distribution because it is a ratio of a variable that has a standard normal distribution (the nominator in the parenthesis) and a variable that has a chi-squared distribution divided by (n-k-1).

The standard error of $\hat{\beta}_k$, se($\hat{\beta}_k$), is $\sqrt{s^2 S_{kk}}$.

Testing a Hypothesis on $\hat{\beta}_k$

In most cases we want to test the null hypothesis

H_0: $\beta_k = 0$

with the t-statistics

t-test: $(\hat{\beta}_k - 0)/ \text{se}(\hat{\beta}_k) \sim t_{\text{n-k-1.}}$

When we test the null hypothesis, the t-statistics is just a ratio of an OLS estimator over its standard error.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + u_i$$

Sometimes we want to test to see whether a group of variables jointly has effects of y. Suppose we want to know whether independent variables x_3, x_4, and x_5 jointly have effects on y.

Thus the null hypothesis is

$$H_0: \beta_3 = \beta_4 = \beta_5 = 0.$$

The null hypothesis, therefore, poses a question whether these three variables can be excluded from the model. Thus the hypothesis is also called exclusion restrictions. A model with the exclusion is called **the restricted model**:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

On the other hand, the model without the exclusion is called **the unrestricted model**:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + u_i$$

We can generalize this problem by changing the number of restrictions from three to q. The joint significance of q variables is measured by how much the sum of squared residuals (SSR) increases when the q-variables are excluded. Let denote the SSR of the restricted and unrestricted models as SSR_r and SSR_{ur}, respectively. Of course the SSR_{ur} is smaller than the SSR_r because the unrestricted model has more variables than the restricted model. But the question is how much compared with the original size of SSR. The F-statistics is defined as

F-test: $$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

The F-statistics can be re-defined as

F-test: $$F \equiv \frac{(R_{ur}^2 - R_r^2)/q}{(1-R_{ur}^2)/(n-k-1)}.$$

Again, because the unrestricted model has more variables, it has a larger R-squared than the restricted model. (Thus the numerator is always positive.) The numerator measures the loss in the explanatory power, per one restriction, when moving from the unrestricted model to the restricted model. This change is divided by the unexplained variation in y by the unrestricted model, standardized by the degree of freedom.

If the decrease in explanatory power is relatively large, then the set of q-variables is considered a jointly significant in the model. (Thus these q-variables should stay in the model.)

Lecture 6: OLS Asymptotic Properties

Consistency (instead of unbiasedness)

First, we need to define consistency. Suppose W_n is an estimator of θ on a sample of Y_1, Y_2, …, Y_n of size n. Then, W_n is a consistent estimator of θ if for every $e > 0$,

$$P(|W_n - \theta| > e) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

This says that the probability that the absolute difference between W_n and θ being larger than e goes to zero as n gets bigger. Which means that this probability could be non-zero while n is not large. For instance, let's say that we are interested in finding the average income of American people and take small samples randomly. Let's assume that the small samples include Bill Gates by chance. The sample mean income is way over the population average. Thus, when sample sizes are small, the probability that the difference between the sample and population averages is larger than e, which is any positive number, can be non-zero. However, the difference between the sample and population averages would be smaller as the sample size gets bigger (as long as the sampling is properly done). As a result, as the sample size goes to infinity, the probability that the difference between the two averages is bigger than e (no matter how small e is) becomes zero.

In other words, we say that θ is the probability limit of W_n:

$$plim\ (W_n) = \theta.$$

Under the finite-sample properties, we say that W_n is <u>unbiased</u>, $E(W_n) = \theta$. Under the asymptotic properties, we say that W_n is <u>consistent</u> because W_n converges to θ as n gets

$$\hat{\beta} = \beta + (X'X)^{-1} X'u$$

In the matrix form, we can examine the probability limit of OLS

$$p\lim \hat{\beta} = \beta + \left(\frac{1}{n} X'X\right)^{-1} p\lim\left(\frac{1}{n} X'u\right)$$

Here, we assume that

$$p\lim \frac{1}{n} X'X = Q.$$

This assumption is not a difficult one to make since the law of large numbers suggests that the each component of $\frac{1}{n} X'X$ goes to the mean values of $X'X$. And also we assume that Q^{-1} exists. From E2, we have

$$p\lim\left(\frac{1}{n} X'u\right) = 0 .$$

Thus,

$$p\lim \hat{\beta} = \beta$$

Thus, we have shown that the OLS estimator is consistent.

Nest, we focus on the asymmetric inference of the OLS estimator. To obtain the asymptotic distribution of the OLS estimator, we first derive the limit distribution of the OLS estimators by multiplying $\sqrt{n}$ (note: we multiply $\sqrt{n}$ (scaling) on $\hat{\beta} - \beta$ to obtain non-zero yet finite variance asymptotically; see Cameron and Trivedi; also $\sqrt{n}$ will be

The probability limit of $\sqrt{n}(\hat{\beta}-\beta)$ goes to zero because of the consistency of $\hat{\beta}$. The limit variance of $\sqrt{n}(\hat{\beta}-\beta)$ is

$$\sqrt{n}(\hat{\beta}-\beta)\cdot\sqrt{n}(\hat{\beta}-\beta)' = \left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{\sqrt{n}}X'u\right)\left(\frac{1}{\sqrt{n}}X'u\right)'\left(\frac{1}{n}X'X\right)^{-1}$$

$$= \left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'uu'X\right)\left(\frac{1}{n}X'X\right)^{-1}$$

From E4, the probability limit of uu' goes to $\sigma^2 I$, and we assumed *plim* of $\frac{1}{n}X'X$ is Q. Thus,

$$= Q^{-1}\left(\frac{\sigma^2}{n}X'X\right)Q^{-1}$$

$$= \sigma^2 Q^{-1}QQ^{-1}$$

$$= \sigma^2 Q^{-1}$$

Therefore, the limit distribution of the OLS estimator is

$$\sqrt{n}(\hat{\mathrm{B}}-\mathrm{B}) \sim^{d} N[0,\ \sigma^2 Q^{-1}].$$

From this, we can obtain the asymptotically distribution of the OLS estimator by multiplying $\sqrt{n}$ and manipulating:

***Example 6-1*: Consistency of OLS Estimators in Bivariate Linear Estimation**

A bivariate model: $y_i = \beta_0 + \beta_1 x_{i1} + u_i$ and $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

To examine the biasedness of the OLS estimator, we take the expectation

$$E(\hat{\beta}_1) = \beta_1 + E\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})\, u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Under the assumption of zero conditional mean (SLR 3: E(u|x) = 0), we can separate the expectation of x and u:

$$E(\hat{\beta}_1) = \beta_1 + \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})\, E(u_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

Thus we need the SLR 3 to show the OLS estimator is unbiased.

Now, suppose we have a violation of SLR 3 and cannot show the unbiasedness of the OLS estimator. We consider a consistency of the OLS estimator.

$$p\lim \hat{\beta}_1 = p\lim \beta_1 + p\lim\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})\, u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

$$p\lim \hat{\beta}_1 = \beta_1 + \frac{p\lim\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})u_i\right]}{p\lim\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]}$$

Lecture 7: OLS Further Issues

In this lecture, we will discuss some practical issues related to OLS estimations, such as functional forms and interpretations of several types of variables. For details, please read Wooldridge chapter 6 and 7.

Measurement Error in the Dependent Variable

Let y^* denote the variable that we would like to explain:

$$y^*_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i$$

However, we can only observe y which is a measured variable of y^* with measurement errors.

$$e_0 = y - y^*$$

By replacing y^* with y and e_0, we get

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + e_0 + u_i$$

Thus, if $e_0 + u$ satisfy the OLS assumptions (such as $E(e_0 + u|X)=0$), then OLS estimators are unbiased (or consistent). But the variance of the disturbance is larger by $Var(e_0)$ with the measurement error (e_0) than without.

Note, however, that the measurement error in the dependent variable could be correlated with independent variables [Cov (x_k e_0)≠0]. In that case, the estimators will be biased.

Assumption 1: Cov (x_k, e_k) = 0

Under this assumption, the error term (u - $\beta_k e_k$) has zero mean and uncorrelated with the independent variables. Thus the estimators are unbiased (consistent). The error variance, however, is bigger by ($\beta_k e_k$)^2.

Assumption 2: Cov (x^*_k, e_k) = 0

This assumption is called the Classic Errors-in-Variables (CEV) assumption. Because e_k = x_k - x^*_k, x_k and e_k must be correlated under the assumption 2:

$$\text{Cov}(x_k, e_k) = \text{E}(x_k e_k) = \text{E}(x^*_k e_k) + \text{E}(e^2_k) = \sigma^2_{ek}$$

Thus, we have the omitted variables problem, which gives inconsistent estimators of all independent variables.

The Attenuation Bias

For **a bivariate regression model** it is easy to show the exact bias cased by the CEV. Now, x_1 is measured with the measurement errors, instead of x_k. In a bivariate regression model, the least square estimator can be written as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)y_i}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}$$

$$= \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(u_i - \beta_k e_i)}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}$$

$$p\lim(\hat{\beta}_1) = \beta_1 + \frac{\text{cov}(x_1, u - \beta_1 e_1)}{Var(x_1)}$$

$$= \beta_1 + \frac{(-\beta_1 \sigma_{e1}^2)}{Var(x_1^* + e_1)}$$

$$= \beta_1 \left(1 - \frac{\sigma_{e1}^2}{\sigma_{x_1^*}^2 + \sigma_{e1}^2}\right)$$

$$= \beta_1 \left(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e1}^2}\right) < \beta_1$$

Thus, the plim($\hat{\beta}_1$) is always closer to zero (or biased toward zero) than β_1. This is called the (famous) **attenuation bias** in OLS due to classical errors-in-variables.

For a multivariate regression model, the probability limit of $\hat{\beta}_1$ is

$$p\lim(\hat{\beta}_1) = \beta_1 \left(\frac{\sigma_{r_1^*}^2}{\sigma_{r_1^*}^2 + \sigma_{e1}^2}\right) < \beta_1$$

where r_1^* *is the population error in the equation*

$$x_1^* = \alpha_0 + \alpha_1 x_2 + \ldots + \alpha_{k-1} x_k + r_1^*$$

Again the implication is the same as before. The estimated coefficient of the variable with measurement errors is biased toward zero (or less likely to reject the null hypothesis).

Data Scaling

If you scale up/down the dependent variable by ∀, the OLS estimators and standard errors will be also scaled up/down by ∀, but not *t*-statistics. Thus the significance level remains the same as before scaling.

Scaling up/down a independent variable:

$$y = \hat{\beta}_0 + (\hat{\beta}_1 / \alpha)(\alpha x_1) + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k \ .$$

If you scale up/down one independent variable, the estimated coefficient of the independent variable will be scale down/up by the same scale. Again the *t*-statistics (or significance level) does not change.

Logarithmic Forms

For a small change in *x*, a change in log(*x*) times 100, 100)log(*x*), is approximately close to a percentage change in *x*, $\Delta x / \bar{x}$. Therefore, we can interpret the following cases using percentage changes:

(1) log-log: $\log(y) = \beta_k \log(x_k) + \ldots \qquad \hat{\beta}_k = \dfrac{\Delta y / \bar{y}}{\Delta x / \bar{x}}$

One percent change in x_k changes *y* by $(100\,\hat{\beta}_k)$ percent. $\hat{\beta}_k$ is an elasticity.

(2) log-level: $\log(y) = \beta_k x_k + \ldots \qquad \hat{\beta}_k = \dfrac{\Delta y / \bar{y}}{\Delta x}$

(4) level-level: $y=\beta_k x_k+\ldots$ $\hat{\beta}_k=\dfrac{\Delta y}{\Delta x}$

One unit change in x_k changes y by $\hat{\beta}_k$.

When a change in log is not small, the approximation between a change in log(x) and a change x may not be accurate. For instance, the log-level model gives us

$$\log(\hat{y}')-\log(\hat{y})=\hat{\beta}_k. \qquad (8\text{-}1)$$

If the change in log is small, then there is no problem of interpreting this as "one unit of x_k changes y by ($100\,\hat{\beta}_k$) percent," because $\log(\hat{y}')-\log(\hat{y})\approx(y'-y)/y$. But when a change in log is not small, the approximation may not be approximate. Thus we need to transform (8-1) as:

$$\log(\hat{y}')-\log(\hat{y})=\log(\hat{y}'/y)=\hat{\beta}_k$$

$$(\hat{y}'/y)=\exp(\hat{\beta}_k)$$

$$(\hat{y}'/y)-1=\exp(\hat{\beta}_k)-1$$

$$(y'-\hat{y})/y=\exp(\hat{\beta}_k)-1$$

$$\%\hat{\Delta}y=100[\exp(\hat{\beta}_k)-1]$$

Thus one unit in x_k changes y by $100[\exp(\hat{\beta}_k)-1]$ percentage.

Quadratic Form

$\hat{\beta}_1 > 0$ *and* $\beta_2 < 0$ "an increase in *x* increases *y* with a diminishing rate "

$\hat{\beta}_1 < 0$ *and* $\beta_2 > 0$ "an increase in *x* decreases *y* with a diminishing rate "

Turning Point:

At the turning point, the first derivative of y with respect to x_1 is zero:

$$\partial y / \partial x_1 = \hat{\beta}_1 + 2\hat{\beta}_2 x_1 = 0$$

Thus the value of x_1 at the turning this point is

$$x_1^* = -\hat{\beta}_1 / (2\hat{\beta}_2)$$

Interaction Terms

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 (x_1 x_2) + \ldots + \hat{\beta}_k x_k$$

The impact of x_1 on *y* is

$$\partial y / \partial x_1 = \hat{\beta}_1 + \hat{\beta}_3 x_2$$

A Dummy Variable

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k \qquad where\ x_1 = 0\ or\ 1$$

A group of observations with x_1=0 is called **a base, benchmark, or reference group**.

Interaction Terms with Dummies

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 + \ \ldots \ + \hat{\beta}_k x_k \qquad where\ x_1 = 0\ or\ 1$$

When x_2 is a continuous variable, $\hat{\beta}_3$ measures a difference in the effect of x_2 on y between a group with x_1=0 and a group with x_1=1, or a difference in slopes of x_2:

$$\partial y / \partial x_2 = \hat{\beta}_2 \qquad when\ x_1 = 0$$
$$\partial y / \partial x_2 = \hat{\beta}_2 + \hat{\beta}_3 \qquad when\ x_1 = 1$$

Multicollinearity

From the previous lecture, we know that the variance of OLS estimators is $Var(\hat{\mathrm{B}}) = \sigma^2 (X'X)^{-1}$. The variance of an estimator, $\hat{\beta}_k$, is $\sigma^2 S_{kk}$. This can be written as

$$Var(\hat{\beta}_k) = \frac{\sigma^2}{(1 - R_k^2)\sum_{i-1}^{n}(x_{ik} - \bar{x}_k)}$$

(See Wooldridge pp94 or Greene pp57) R_k^2 is the R-squared in the regression of x_k against all other variables. In other words, R_k^2 is the proportion of the total variation in x_k that can be explained by the other independent variables.

If two variables are **perfectly correlated**, then R_k^2 will be one for both of the perfectly correlated variables, and the variance of those two variables will not be measured. Obviously, you need to drop one of the two perfectly correlated variables. In STATA, STATA drops one of perfectly correlated variables automatically. So if you see a dropped variable in STATA outputs, you should suspect that you have included perfectly correlated variables without realizing.

Even if two or more variables are not perfectly correlated, if they are highly correlated (high R_k^2), the variance of estimators will be large. This problem is called **multicollinearity**.

When in doubt, conduct a F-test on variables that you suspect causing multicollinearity.
A typical symptom of multicollinearity is

- A high joint significance and low individual significance
  (a high F-statistics but low t-statistics)

A simple solution is to keep them in the model. If your main focus is on a variable which is not the part of multicollinearity, then it is not a serious problem to have multicollinearity in your model. You could drop one of highly correlated variables, but by doing so may create an omitted variable problem. Remember that an omitted variable problem can cause biases in on all of estimators. This could be a more serious problem than multicollinearity.

OLS in Matrix Form

1 The True Model

- Let X be an $n \times k$ matrix where we have observations on k independent variables for n observations. Since our model will usually contain a constant term, one of the columns in the X matrix will contain only ones. This column should be treated exactly the same as any other column in the X matrix.
- Let y be an $n \times 1$ vector of observations on the dependent variable.
- Let ϵ be an $n \times 1$ vector of disturbances or errors.
- Let β be an $k \times 1$ vector of unknown population parameters that we want to estimate.

Our statistical model will essentially look something like the following:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix}_{n\times k} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_n \end{bmatrix}_{k\times 1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}$$

This can be rewritten more simply as:

$$y = X\beta + \epsilon \tag{1}$$

This is assumed to be an accurate reflection of the real world. The model has a systematic component ($X\beta$) and a stochastic component (ϵ). Our goal is to obtain estimates of the population parameters in the β vector.

2 Criteria for Estimates

Our *estimates* of the population parameters are referred to as $\hat{\beta}$. Recall that the criteria we use for obtaining our estimates is to find the estimator $\hat{\beta}$ that minimizes the sum of squared residuals

The sum of squared residuals (RSS) is $e'e$.[2]

$$\begin{bmatrix} e_1 & e_2 & \dots & \dots & e_n \end{bmatrix}_{1\times n} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ \vdots \\ e_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} e_1 \times e_1 + e_2 \times e_2 + \ldots + e_n \times e_n \end{bmatrix}_{1\times 1} \tag{3}$$

It should be obvious that we can write the sum of squared residuals as:

$$\begin{aligned} e'e &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \end{aligned} \tag{4}$$

where this development uses the fact that the transpose of a scalar is the scalar i.e. $y'X\hat{\beta} = (y'X\hat{\beta})' = \hat{\beta}'X'y$.

To find the $\hat{\beta}$ that minimizes the sum of squared residuals, we need to take the derivative of Eq. 4 with respect to $\hat{\beta}$. This gives us the following equation:

$$\frac{\partial e'e}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0 \tag{5}$$

To check this is a minimum, we would take the derivative of this with respect to $\hat{\beta}$ again – this gives us $2X'X$. It is easy to see that, so long as X has full rank, this is a positive definite matrix (analogous to a positive real number) and hence a minimum.[3]

[2] It is important to note that this is very different from ee' – the variance-covariance matrix of residuals.

[3] Here is a brief overview of matrix differentiaton.

$$\frac{\partial a'b}{\partial b} = \frac{\partial b'a}{\partial b} = a \tag{6}$$

when a and b are K×1 vectors.

$$\frac{\partial b'Ab}{\partial b} = 2Ab = 2b'A \tag{7}$$

when A is any symmetric matrix. Note that you can write the derivative as either 2Ab or $2b'A$.

$$\frac{\partial 2\beta'X'y}{\partial b} = \frac{\partial 2\beta'(X'y)}{\partial b} = 2X'y \tag{8}$$

and

From Eq. 5 we get what are called the 'normal equations'.

$$(X'X)\hat{\beta} = X'y \tag{10}$$

Two things to note about the $(X'X)$ matrix. First, it is always square since it is $k \times k$. Second, it is always symmetric.

Recall that $(X'X)$ and $X'y$ are known from our data but $\hat{\beta}$ is unknown. If the inverse of $(X'X)$ exists (i.e. $(X'X)^{-1}$), then pre-multiplying both sides by this inverse gives us the following equation:[4]

$$(X'X)^{-1}(X'X)\hat{\beta} = (X'X)^{-1}X'y \tag{11}$$

We know that by definition, $(X'X)^{-1}(X'X) = I$, where I in this case is a $k \times k$ identity matrix. This gives us:

$$\begin{aligned} I\hat{\beta} &= (X'X)^{-1}X'y \\ \hat{\beta} &= (X'X)^{-1}X'y \end{aligned} \tag{12}$$

Note that we have not had to make any assumptions to get this far! Since the OLS estimators in the $\hat{\beta}$ vector are a linear combination of existing random variables (X and y), they themselves are random variables with certain straightforward properties.

3 Properties of the OLS Estimators

The primary property of OLS estimators is that they satisfy the criteria of minimizing the sum of squared residuals. However, there are other properties. These properties do not depend on any assumptions - they will always be true so long as we compute them in the manner just shown.

Recall the normal form equations from earlier in Eq. 10.

$$(X'X)\hat{\beta} = X'y \tag{13}$$

Now substitute in $y = X\hat{\beta} + e$ to get

$$\begin{aligned} (X'X)\hat{\beta} &= X'(X\hat{\beta} + e) \\ (X'X)\hat{\beta} &= (X'X)\hat{\beta} + X'e \\ X'e &= 0 \end{aligned} \tag{14}$$

What does $X'e$ look like?

$$\begin{bmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ X_{k1} & X_{k2} & \dots & X_{kn} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} X_{11} \times e_1 + X_{12} \times e_2 + \ldots + X_{1n} \times e_n \\ X_{21} \times e_1 + X_{22} \times e_2 + \ldots + X_{2n} \times e_n \\ \vdots \\ \vdots \\ X_{k1} \times e_1 + X_{k2} \times e_2 + \ldots + X_{kn} \times e_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} \quad (15)$$

From $X'e = 0$, we can derive a number of properties.

1. **The observed values of X are uncorrelated with the residuals.**

   $X'e = 0$ implies that for every column x_k of X, $x_k'e = 0$. In other words, each regressor has zero sample correlation with the residuals. Note that this does not mean that X is uncorrelated with the disturbances; we'll have to assume this.

   If our regression includes a constant, then the following properties also hold.

2. **The sum of the residuals is zero.**

   If there is a constant, then the first column in X (i.e. X_1) will be a column of ones. This means that for the first element in the $X'e$ vector (i.e. $X_{11} \times e_1 + X_{12} \times e_2 + \ldots + X_{1n} \times e_n$) to be zero, it must be the case that $\sum e_i = 0$.

3. **The sample mean of the residuals is zero.**

   This follows straightforwardly from the previous property i.e. $\overline{e} = \frac{\sum e_i}{n} = 0$.

4. **The regression hyperplane passes through the means of the observed values ($\overline{X}$ and $\overline{y}$).**

   This follows from the fact that $\overline{e} = 0$. Recall that $e = y - X\hat{\beta}$. Dividing by the number of observations, we get $\overline{e} = \overline{y} - \overline{x}\hat{\beta} = 0$. This implies that $\overline{y} = \overline{x}\hat{\beta}$. This shows that the regression hyperplane goes through the point of means of the data.

5. **The predicted values of y are uncorrelated with the residuals.**

   The predicted values of y are equal to $X\hat{\beta}$ i.e. $\hat{y} = X\hat{\beta}$. From this we have

These properties always hold true. You should be careful not to infer anything from the residuals about the disturbances. For example, you cannot infer that the sum of the disturbances is zero or that the mean of the disturbances is zero just because this is true of the residuals - this is true of the residuals just because we decided to minimize the sum of squared residuals.

Note that we know nothing about $\hat{\beta}$ except that it satisfies all of the properties discussed above. We need to make some assumptions about the true model in order to make any inferences regarding β (the true population parameters) from $\hat{\beta}$ (our estimator of the true parameters). Recall that $\hat{\beta}$ comes from our sample, but we want to learn about the true parameters.

4 The Gauss-Markov Assumptions

1. $y = X\beta + \epsilon$

   This assumption states that there is a linear relationship between y and X.

2. X is an $n \times k$ matrix of full rank.

   This assumption states that there is no perfect multicollinearity. In other words, the columns of X are linearly independent. This assumption is known as the identification condition.

3. $E[\epsilon|X] = 0$

$$E\begin{bmatrix} \epsilon_1|X \\ \epsilon_2|X \\ \vdots \\ \epsilon_n|X \end{bmatrix} = \begin{bmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{17}$$

   This assumption - the zero conditional mean assumption - states that the *disturbances* average out to 0 for any value of X. Put differently, no observations of the independent variables convey any information about the expected value of the disturbance. The assumption implies that $E(y) = X\beta$. This is important since it essentially says that we get the mean function right.

4. $E(\epsilon\epsilon'|X) = \sigma^2 I$

   This captures the familiar assumption of homoskedasticity and no autocorrelation. To see why, start with the following:

which is the same as:

$$E(\epsilon\epsilon'|X) = E\begin{bmatrix} \epsilon_1^2|X & \epsilon_1\epsilon_2|X & \dots & \epsilon_1\epsilon_n|X \\ \epsilon_2\epsilon_1|X & \epsilon_2^2|X & \dots & \epsilon_2\epsilon_n|X \\ \vdots & \vdots & \vdots & \vdots \\ \epsilon_n\epsilon_1|X & \epsilon_n\epsilon_2|X & \dots & \epsilon_n^2|X \end{bmatrix} \tag{19}$$

which is the same as:

$$E(\epsilon\epsilon'|X) = \begin{bmatrix} E[\epsilon_1^2|X] & E[\epsilon_1\epsilon_2|X] & \dots & E[\epsilon_1\epsilon_n|X] \\ E[\epsilon_2\epsilon_1|X] & E[\epsilon_2^2|X] & \dots & E[\epsilon_2\epsilon_n|X] \\ \vdots & \vdots & \vdots & \vdots \\ E[\epsilon_n\epsilon_1|X] & E[\epsilon_n\epsilon_2|X] & \dots & E[\epsilon_n^2|X] \end{bmatrix} \tag{20}$$

The assumption of homoskedasticity states that the variance of ϵ_i is the same (σ^2) for all i i.e. $\text{var}[\epsilon_i|X] = \sigma^2 \ \forall \ i$. The assumption of no autocorrelation (uncorrelated errors) means that $\text{cov}(\epsilon_i, \epsilon_j|X) = 0 \ \forall \ i \neq j$ i.e. knowing something about the disturbance term for one observation tells us nothing about the disturbance term for any other observation. With these assumptions, we have:

$$E(\epsilon\epsilon'|X) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} \tag{21}$$

Finally, this can be rewritten as:

$$E(\epsilon\epsilon'|X) = \sigma^2\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \sigma^2 I \tag{22}$$

Disturbances that meet the two assumptions of homoskedasticity and no autocorrelation are referred to as spherical disturbances. We can compactly write the Gauss-Markov assumptions about the disturbances as:

$$\Omega = \sigma^2 I \tag{23}$$

where Ω is the variance-covariance matrix of the disturbances i.e. $\Omega = E[\epsilon\epsilon']$.

5 The Gauss-Markov Theorem

The Gauss-Markov Theorem states that, conditional on assumptions 1-5, there will be no other linear and unbiased estimator of the β coefficients that has a smaller sampling variance. In other words, the OLS estimator is the Best Linear, Unbiased and Efficient estimator (BLUE). How do we know this?

Proof that $\hat{\beta}$ is an unbiased estimator of β.

We know from earlier that $\hat{\beta} = (X'X)^{-1}X'y$ and that $y = X\beta + \epsilon$. This means that

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'(X\beta + \epsilon) \\ \hat{\beta} &= \beta + (X'X)^{-1}X'\epsilon \end{aligned} \tag{24}$$

since $(X'X)^{-1}X'X = I$. This shows immediately that OLS is unbiased so long as either (i) X is fixed (non-stochastic) so that we have:

$$\begin{aligned} E[\hat{\beta}] &= E[\beta] + E[(X'X)^{-1}X'\epsilon] \\ &= \beta + (X'X)^{-1}X'E[\epsilon] \end{aligned} \tag{25}$$

where $E[\epsilon] = 0$ by assumption or (ii) X is stochastic but independent of ϵ so that we have:

$$\begin{aligned} E[\hat{\beta}] &= E[\beta] + E[(X'X)^{-1}X'\epsilon] \\ &= \beta + (X'X)^{-1}E[X'\epsilon] \end{aligned} \tag{26}$$

where $E(X'\epsilon) = 0$.

Proof that $\hat{\beta}$ is a linear estimator of β.

From Eq. 24, we have:

$$\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon \tag{27}$$

Since we can write $\hat{\beta} = \beta + A\epsilon$ where $A = (X'X)^{-1}X'$, we can see that $\hat{\beta}$ is a linear function of the disturbances. By the definition that we use, this makes it a linear estimator (See Greene (2003, 45).

Proof that $\hat{\beta}$ has minimal variance among all linear and unbiased estimators

See Greene (2003, 46-47)

6 The Variance-Covariance Matrix of the OLS Estimates

We can derive the variance-covariance matrix of the OLS estimator, $\hat{\beta}$.

$$\begin{aligned} E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] &= E[((X'X)^{-1}X'\epsilon)((X'X)^{-1}X'\epsilon)'] \\ &= E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}] \end{aligned} \tag{28}$$

where we take advantage of the fact that $(AB)' = B'A'$ i.e. we can rewrite $(X'X)^{-1}X'\epsilon$ as $\epsilon'X(X'X)^{-1}$. If we assume that X is non-stochastic, we get:

$$E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] = (X'X)^{-1}X'E[\epsilon\epsilon']X(X'X)^{-1} \tag{29}$$

From Eq. 22, we have $E[\epsilon\epsilon'] = \sigma^2 I$. Thus, we have:

$$\begin{aligned} E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] &= (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1} \\ &= \sigma^2 I(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned} \tag{30}$$

We estimate σ^2 with $\hat{\sigma}^2$, where:

$$\hat{\sigma}^2 = \frac{e'e}{n-k} \tag{31}$$

To see the derivation of this, see Greene (2003, 49).

What does the variance-covariance matrix of the OLS estimator look like?

$$E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] = \begin{bmatrix} \text{var}(\hat{\beta}_1) & \text{cov}(\hat{\beta}_1,\hat{\beta}_2) & \dots & \text{cov}(\hat{\beta}_1,\hat{\beta}_k) \\ \text{cov}(\hat{\beta}_2,\hat{\beta}_1) & \text{var}(\hat{\beta}_2) & \dots & \text{cov}(\hat{\beta}_2,\hat{\beta}_k) \\ \vdots & \vdots & \vdots & \vdots \\ \text{cov}(\hat{\beta}_k,\hat{\beta}_1) & \text{cov}(\hat{\beta}_k,\hat{\beta}_2) & \dots & \text{var}(\hat{\beta}_k) \end{bmatrix} \tag{32}$$

As you can see, the standard errors of the $\hat{\beta}$ are given by the square root of the elements along the main diagonal of this matrix.

6.1 Hypothesis Testing

Recall Assumption 6 from earlier, which stated that $\epsilon|X \sim N[0, \sigma^2 I]$. I had stated that this

we are also saying that the OLS estimator is also distributed multivariate normal i.e.

$$\hat{\beta} \sim N[\beta, \sigma^2(X'X)^{-1}] \tag{34}$$

but where the mean is β and the variance is $\sigma^2(X'X)^{-1}$. It is this that allows us to conduct the normal hypothesis tests that we are familiar with.

7 Robust (Huber of White) Standard Errors

Recall from Eq. 29 that we have:

$$\begin{aligned} \text{var} - \text{cov}(\hat{\beta}) &= (X'X)^{-1}X'E[\epsilon\epsilon']X(X'X)^{-1} \\ &= (X'X)^{-1}(X'\Omega X)(X'X)^{-1} \end{aligned} \tag{35}$$

This helps us to make sense of White's heteroskedasticity consistent standard errors.[5]

Recall that heteroskedasticity does not cause problems for estimating the coefficients; it only causes problems for getting the 'correct' standard errors. We can compute $\hat{\beta}$ without making any assumptions about the disturbances i.e. $\hat{\beta}_{OLS} = (X'X)^{-1}X'y$. However, to get the results of the Gauss Markov Theorem (things like $E[\hat{\beta}] = \beta$ etc.) and to be able to conduct hypothesis tests ($\hat{\beta} \sim N[\beta, \sigma^2(X'X)^{-1}]$), we need to make assumptions about the disturbances. One of the assumptions is that $E[ee'] = \sigma^2 I$. This assumption includes the assumption of homoskedasticity – $\text{var}[\epsilon_i|X] = \sigma^2 \ \forall \ i$. However, it is not always the case that the variance will be the same for all observations i.e. we have σ_i^2 instead of σ^2. Basically, there may be many reasons why we are better at predicting some observations than others. Recall the variance-covariance matrix of the disturbance terms from earlier:

$$E(\epsilon\epsilon'|X) = \Omega = \begin{bmatrix} E[\epsilon_1^2|X] & E[\epsilon_1\epsilon_2|X] & \dots & E[\epsilon_1\epsilon_n|X] \\ E[\epsilon_2\epsilon_1|X] & E[\epsilon_2^2|X] & \dots & E[\epsilon_2\epsilon_n|X] \\ \vdots & \vdots & \vdots & \vdots \\ E[\epsilon_n\epsilon_1|X] & E[\epsilon_n\epsilon_2|X] & \dots & E[\epsilon_n^2|X] \end{bmatrix} \tag{36}$$

If we retain the assumption of no autocorrelation, this can be rewritten as:

$$E(\epsilon\epsilon'|X) = \Omega = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \tag{37}$$

assumption to make. Our OLS standard errors will be incorrect insofar as:

$$X'E[\epsilon\epsilon']X \neq \sigma^2(X'X) \tag{38}$$

Note that our OLS standard errors may be too big or too small. So, what can we do if we suspect that there is heteroskedasticity?

Essentially, there are two options.

1. **Weighted Least Squares:** To solve the problem, we just need to find something that is proportional to the variance. We might not know the variance for each observation, but if we know something about where it comes from, then we might know something that is proportional to it. In effect, we try to model the variance. Note that this only solves the problem of heteroskedasticity if we assume that we have modelled the variance correctly - we never know if this is true or not.

2. **Robust standard errors** (White 1980): This method treats heteroskedasticity as a nuisance rather than something to be modelled.

How do robust standard errors work? We never observe disturbances (ϵ) but we do observe residuals (e). While each individual residual (e_i) is not going to be a very good estimator of the corresponding disturbance (ϵ_i), White (1980) showed that $X'ee'X$ is a consistent (but not unbiased) estimator of $X'E[\epsilon\epsilon']X$.[6]

Thus, the variance-covariance matrix of the coefficient vector from the White estimator is:

$$\text{var} - \text{cov}(\hat{\beta}) = (X'X)^{-1}X'ee'X(X'X)^{-1} \tag{39}$$

rather than:

$$\begin{aligned} \text{var} - \text{cov}(\hat{\beta}) &= X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1} \\ &= (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1} \end{aligned} \tag{40}$$

from the normal OLS estimator.

White (1980) suggested that we could test for the presence of heteroskedasticity by examining the extent to which the OLS estimator diverges from his own estimator. White's test is to regress the squared residuals (e_i^2) on the terms in $X'X$ i.e. on the squares and the cross-products of the independent variables. If the R^2 exceeds a critical value ($nR^2 \sim \chi_k^2$), then heteroskedasticity causes

8 Partitioned Regression and the Frisch-Waugh-Lovell Theorem

Imagine that our true model is:

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon \tag{41}$$

In other words, there are two sets of independent variables. For example, X_1 might contain some independent variables (perhaps also the constant) whereas X_2 contains some other independent variables. The point is that X_1 and X_2 need not be two variables only. We will estimate:

$$y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + e \tag{42}$$

Say, we wanted to isolate the coefficients associated with X_2 i.e. $\hat{\beta}_2$. The normal form equations will be:[7]

$$\begin{matrix}(1)\\(2)\end{matrix} \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix} \tag{43}$$

First, let's solve for $\hat{\beta}_1$.

$$\begin{aligned} (X_1'X_1)\hat{\beta}_1 + (X_1'X_2)\hat{\beta}_2 &= X_1'y \\ (X_1'X_1)\hat{\beta}_1 &= X_1'y - (X_1'X_2)\hat{\beta}_2 \\ \hat{\beta}_1 &= (X_1'X_1)^{-1}X_1'y - (X_1'X_1)^{-1}X_1'X_2\hat{\beta}_2 \\ \hat{\beta}_1 &= (X_1'X_1)^{-1}X_1'(y - X_2\hat{\beta}_2) \end{aligned} \tag{44}$$

8.1 Omitted Variable Bias

The solution shown in Eq. 44 is the set of OLS coefficients in the regression of y on X_1, i.e, $(X_1'X_1)^{-1}X_1'y$, minus a correction vector $(X_1'X_1)^{-1}X_1'X_2\hat{\beta}_2$. This correction vector is the equation for omitted variable bias. The first part of the correction vector up to $\hat{\beta}_2$, i.e. $(X_1'X_1)^{-1}X_1'X_2$, is just the regression of the variables in X_2 done separately and then put together into a matrix on all the variables in X_1. This will only be zero if the variables in X_1 are linearly unrelated (uncorrelated or orthogonal) to the variables in X_2. The correction vector will also be zero if $\hat{\beta}_2 = 0$ i.e. if X_2 variables have no impact on y. Thus, you can ignore all potential omitted variables that are either (i) unrelated to the included variables or (ii) unrelated to the dependent variable. Any omitted variables that do not meet these conditions will change your estimates of $\hat{\beta}_1$ if they were to be included.

matrix, and β_1 and β_2 are the true coefficient vectors from a full regression including both X_1 and X_2.

8.2 The Residual Maker and the Hat Matrix

Before going any further, I introduce some useful matrices. Note that:

$$\begin{aligned} e &= y - X\hat{\beta} \\ &= y - X(X'X)^{-1}X'y \\ &= (I - X(X'X)^{-1}X')y \\ &= My \end{aligned} \tag{46}$$

where M is called the residual maker since it makes residuals out of y. M is a square matrix and is idempotent. A matrix A is idempotent if $A^2 = AA = A$.

$$\begin{aligned} MM &= (I - X(X'X)^{-1}X')(I - X(X'X)^{-1}X') \\ &= I^2 - 2X(X'X)^{-1}X' + X(X'X)^{-1}X'X(X'X)^{-1}X' \\ &= I - 2X(X'X)^{-1}X' + X(X'X)^{-1}X' \\ &= I - X(X'X)^{-1}X' \\ &= M \end{aligned} \tag{47}$$

This will prove useful. The M matrix also has the properties that $MX = 0$ and $Me = e$.

A related matrix is the hat matrix (H) which makes $\hat{y}$ our of y. Note that:

$$\hat{y} = y - e = [I - M]y = Hy \tag{48}$$

where:

$$H = X(X'X)^{-1}X' \tag{49}$$

Greene refers to this matrix as P, but he is the only one that does this.

8.3 Frisch-Waugh-Lovell Theorem

So far we have solved for $\hat{\beta}_1$.

Now we insert this into (2) of Eq. 43. This gives us

$$\begin{aligned}
X_2'y &= X_2'X_1(X_1'X_1)^{-1}X_1'y - X_2'X_1(X_1'X_1)^{-1}X_1'X_2\hat{\beta}_2 + X_2'X_2\hat{\beta}_2 \\
X_2'y - X_2'X_1(X_1'X_1)^{-1}X_1'y &= X_2'X_2\hat{\beta}_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2\hat{\beta}_2 \\
X_2'y - X_2'X_1(X_1'X_1)^{-1}X_1'y &= [X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2]\hat{\beta}_2 \\
X_2'y - X_2'X_1(X_1'X_1)^{-1}X_1'y &= [(X_2' - X_2'X_1(X_1'X_1)^{-1}X_1')X_2]\hat{\beta}_2 \\
X_2'y - X_2'X_1(X_1'X_1)^{-1}X_1'y &= [X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2]\hat{\beta}_2 \\
(X_2' - X_2'X_1(X_1'X_1)^{-1}X_1')y &= [X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2]\hat{\beta}_2 \\
X_2'(I - X_1(X_1'X_1)^{-1}X_1')y &= [X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2]\hat{\beta}_2 \\
\hat{\beta}_2 &= [X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2]^{-1}X_2'(I - X_1(X_1'X_1)^{-1}X_1')y \\
&= (X_2'M_1X_2)^{-1}(X_2'M_1y) \qquad (51)
\end{aligned}$$

Recall that M is the residual maker. In this case, M_1 makes residuals for regressions on the X_1 variables: M_1y is the vector of residuals from regressing y on the X_1 variables and M_1X_2 is the matrix made up of the column by column residuals of regressing each variable (column) in X_2 on all the variables in X_1.

Because M is both idempotent and symmetric, we can rewrite Eq. 51 as

$$\hat{\beta}_2 = (X_2^{*'}X_2)^{-1}X_2^{*'}y^* \qquad (52)$$

where $X_2^* = M_1X_2$ and $y^* = M_1y$.

From this it is easy to see that $\hat{\beta}_2$ can be obtained from regressing y^* on X_2^* (you'll get good at spotting regressions i.e. equations of the $(X'X)^{-1}X'y$ form. The starred variables are just the residuals of the variables (y or X_2) after regressing them on the X_1 variables.

This leads to the **Frisch-Waugh-Lovell Theorem**: In the OLS regression of vector y on two sets of variables, X_1 and X_2, the subvector $\hat{\beta}_2$ is the set of coefficients obtained when the residuals from a regression of y on X_1 alone are regressed on the set of residuals obtained when each column of X_2 is regressed on X_1.

We'll come back to the FWL Theorem when we look at fixed effects models.

8.4 Example

Imagine we have the following model

Step 2: regress X_2 on X_1 and obtain residuals (e2) i.e. first column of M_1X_2.
Step 3: regress X_3 on X_1 and obtain residuals (e3) i.e. second column of M_1X_2.
Step 4: regress e1 on e2 and e3 i.e. regress M_1y on M_1X_2.
Step 5: the coefficient on e2 will be $\hat{\beta}_2$ and the coefficient on e3 will be $\hat{\beta}_3$.

Steps 2 and 3 are called partialing out or netting out the effect of X_1. For this reason, the coefficients in multiple regression are often called partial regression coefficients. This is what it means to say we are holding the X_1 variables constant in the regression.

So the difference between regressing Y on both X_1 and X_2 instead of on just X_2 is that in the first case we first regress both the dependent variables and all the X_2 variables separately on X_1 and then regress the residuals on each other, but in the second case we just regress y on the X_2 variables.

Econometria I

Corso di laurea magistrale in Finanza e Assicurazioni

Slides per lezioni

secondo semestre 2020

Marco Ventura

Cap. 8 SW

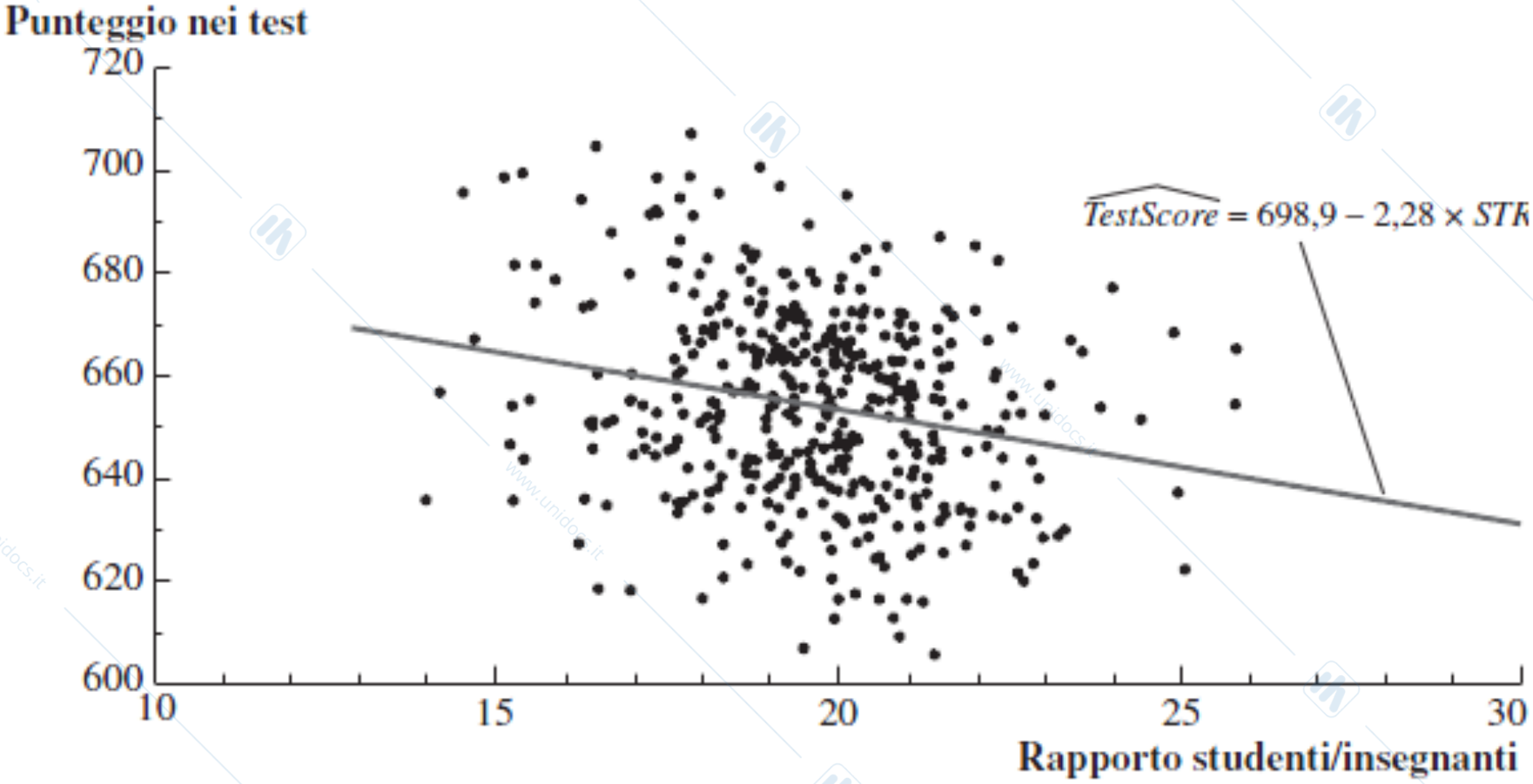
Sommario

1. Funzioni di regressione non lineari – note generali
2. Funzioni non lineari a una variabile
3. Funzioni non lineari a due variabili: interazioni
4. Applicazione al dataset dei punteggi nei test della California

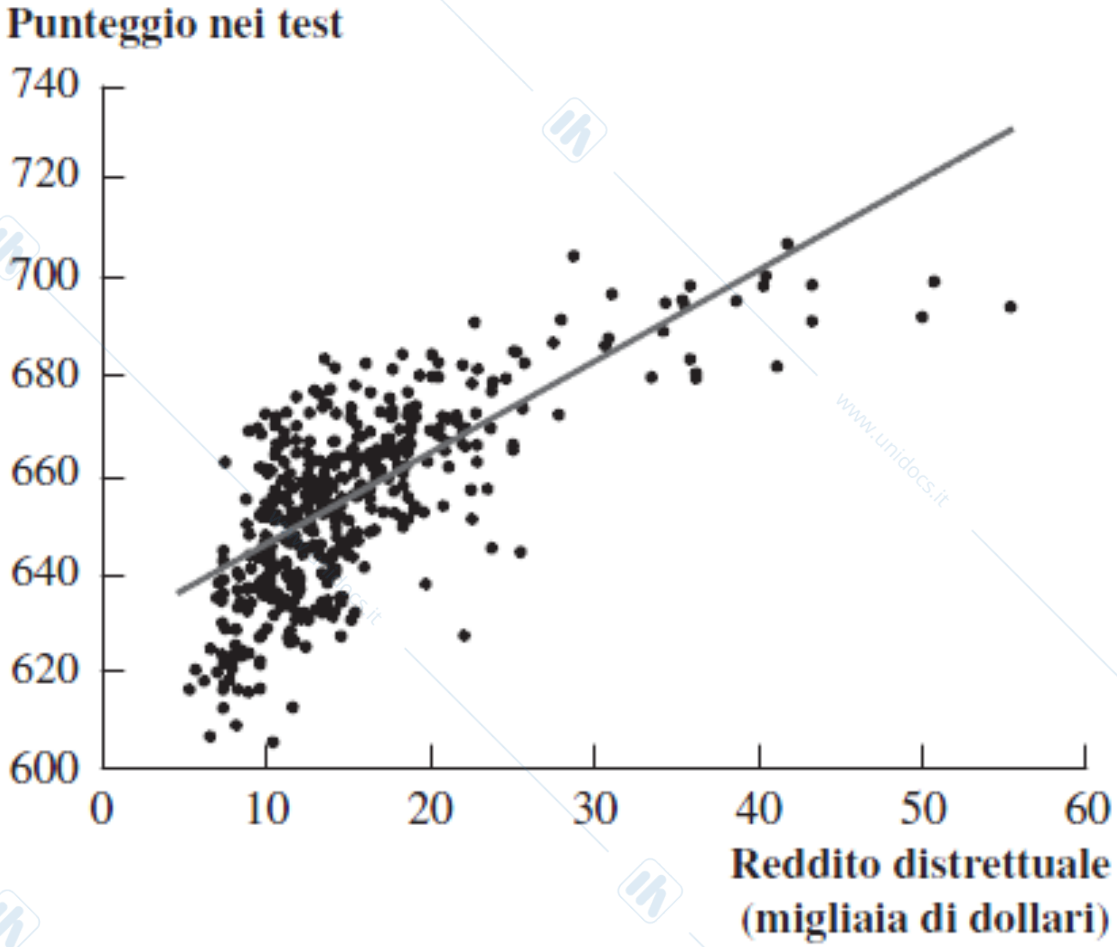
Funzioni di regressione non lineari

- Le funzioni di regressione viste finora erano lineari rispetto alla variabile *X*
- Ma l'approssimazione lineare non è sempre la migliore
- Il modello di regressione multipla può gestire funzioni di regressione non lineari in una o più *X*.

La relazione tra punteggio nei test e rapporto studenti/insegnanti sembra lineare (forse)...



Ma la relazione tra punteggio nei test e reddito distrettuale sembra non lineare…



Funzioni di regressione non lineari – concetti generali (Paragrafo 8.1)

Se una relazione tra *Y* e *X* è **non lineare**:

- L'effetto su *Y* di una variazione in *X* dipende dal valore di *X* – ovvero, l'effetto marginale di *X* non è costante
- Una regressione lineare è mal specificata: la forma funzionale è errata
- Lo stimatore dell'effetto su *Y* di *X* è distorto: in generale non è corretto nemmeno sulla media
- La soluzione consiste nell'applicare una funzione di regressione che sia non lineare in *X*

La formula generale per una funzione di regressione non lineare

$$Y_i = f(X_{1i}, X_{2i},\ldots, X_{ki}) + u_i,\ i = 1,\ldots, n$$

Assunzioni

1. $E(u_i|\ X_{1i}, X_{2i},\ldots, X_{ki}) = 0$ (identica); implica che f è il valore atteso di Y condizionato alle X.
2. $(X_{1i},\ldots, X_{ki}, Y_i)$ sono i.i.d. (identica).
3. Gli outlier sono rari (stessa idea; la condizione matematica precisa dipende dalla f in esame).
4. Assenza di multicollinearità perfetta (stessa idea; la formulazione precisa dipende dalla f in esame).

La variazione in Y associata a una variazione in X_1, mantenendo $X_2,\ldots, X_k$ costanti è:

$$\Delta Y = f(X_1 + \Delta X_1, X_2,\ldots, X_k) - f(X_1, X_2,\ldots, X_k)$$

CONCETTO CHIAVE 8.1 **L'effetto atteso su Y di una variazione di X_1, nel modello di regressione non lineare (8.3)**

La variazione attesa di Y, ΔY, associata alla variazione di X_1, ΔX_1, tenendo costanti $X_2,\ldots,X_k$, è la differenza tra il valore della funzione di regressione della popolazione prima e dopo la variazione in X_1, tenendo costanti $X_2,\ldots,X_k$. In altri termini, la variazione attesa di Y è la differenza:

$$\Delta Y = f(X_1 + \Delta X_1, X_2,\ldots,X_k) - f(X_1, X_2,\ldots,X_k). \quad (8.4)$$

Lo stimatore di tale differenza ignota è la differenza tra i valori predetti in questi due casi. Sia $\hat{f}(X_1, X_2,\ldots,X_k)$ il valore predetto di Y basato sullo stimatore $\hat{f}$ della funzione di regressione della popolazione. Allora, la variazione predetta di Y è

$$\Delta\hat{Y} = \hat{f}(X_1 + \Delta X_1, X_2,\ldots,X_k) - \hat{f}(X_1, X_2,\ldots,X_k) \quad (8.5)$$

Funzioni non lineari di un'unica variabile indipendente (Paragrafo 8.2)

Vedremo due approcci complementari:

1. Polinomiali in *X*

   La funzione di regressione della popolazione viene approssimata da una quadratica, una cubica o una polinomiale di grado più alto

2. Trasformazioni logaritmiche

   Le *Y* e/o le *X* vengono trasformate prendendone il logaritmo, che ne dà un'approssimazione "percentuale" utile in molte applicazioni

1. Polinomiali in *X*

Approssimiamo la funzione di regressione della popolazione con una polinomiale:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_r X_i^r + u_i$$

- È proprio il modello di regressione lineare multipla – salvo che i regressori sono potenze di *X*!
- Per stima, verifica delle ipotesi, ecc. si procede come nel modello di regressione multipla con OLS
- I coefficienti sono difficili da interpretare, ma la funzione risultante è interpretabile

Esempio: la relazione tra punteggio nei test e reddito distrettuale

$Income_i$ = reddito distrettuale medio nel distretto i^{esimo} (migliaia di dollari pro capite)

Approssimazione quadratica:

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 (Income_i)^2 + u_i$$

Approssimazione cubica:

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 (Income_i)^2 + \beta_3 (Income_i)^3 + u_i$$

Stima dell'approssimazione quadratica in STATA

```
generate avginc2 = avginc*avginc;
reg testscr avginc avginc2, r;
```

Crea il regressore quadratico

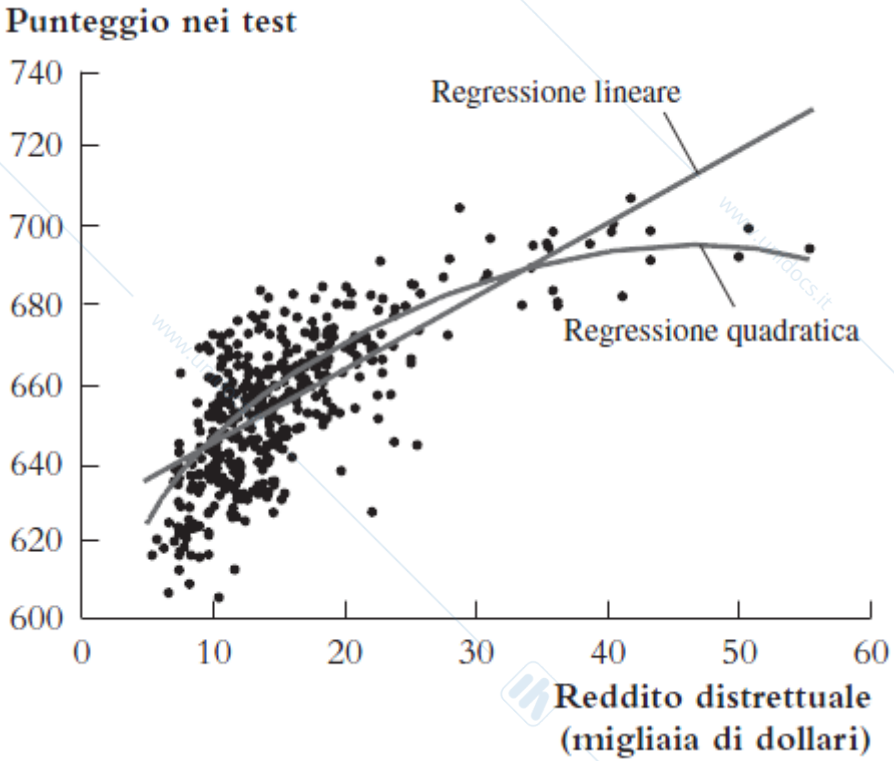
```
Regression with robust standard errors                Number of obs =     420
                                                      F(  2,   417) =  428.52
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5562
                                                      Root MSE      =  12.724

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avginc |   3.850995   .2680941    14.36   0.000      3.32401    4.377979
     avginc2 |  -.0423085   .0047803    -8.85   0.000     -.051705   -.0329119
       _cons |   607.3017   2.901754   209.29   0.000     601.5978    613.0056
------------------------------------------------------------------------------
```

Verifica l'ipotesi di linearità confrontandola con l'alternativa che la funzione di regressione sia quadratica….

Interpretazione della funzione di regressione stimata:

(a) Rappresentiamo graficamente i valori della stima

$$\widehat{TestScore} = 607{,}3 + 3{,}85 Income_i - 0{,}0423 (Income_i)^2$$
$$(2{,}9) \quad (0{,}27) \quad (0{,}0048)$$



***Interpretazione della funzione di regressione stimata*:**

(b) Calcoliamo gli "effetti" per diversi valori di *X*

$$\widehat{TestScore} = 607{,}3 + 3{,}85 Income_i - 0{,}0423 (Income_i)^2$$
$$\qquad\qquad (2{,}9) \quad (0{,}27) \qquad\qquad (0{,}0048)$$

Variazione predetta in *TestScore* per una variazione del reddito da \$5.000 pro capite a \$6.000 pro capite:

$$\begin{aligned}\Delta \widehat{TestScore} &= 607{,}3 + 3{,}85 \times 6 - 0{,}0423 \times 6^2 \\ &\quad - (607{,}3 + 3{,}85 \times 5 - 0{,}0423 \times 5^2) \\ &= 3{,}4\end{aligned}$$

$$\widehat{TestScore} = 607{,}3 + 3{,}85Income_i - 0{,}0423(Income_i)^2$$

"Effetti" attesi in base ai diversi valori di *X*:

Variazione del reddito ($1000 pro capite)	Δ $\widehat{TestScore}$
da 5 a 6	3,4
da 25 a 26	1,7
da 45 a 46	0,0

L'"effetto" di un cambiamento del reddito sull'apprendimento scolastico è maggiore per i redditi più bassi (forse un beneficio marginale decrescente con l'aumento dei budget delle scuole?)

Attenzione! Qual è l'effetto di una variazione da 65 a 66?

Non estrapolate al di fuori dell'intervallo dei dati!

Stima dell'approssimazione cubica in STATA

```
gen avginc3 = avginc*avginc2;
reg testscr avginc avginc2 avginc3, r;

Regression with robust standard errors
```

Crea il regressore cubico

```
Number of obs =     420
F(  3,   416) =  270.18
Prob > F      =  0.0000
R-squared     =  0.5584
Root MSE      =  12.707
```

testscr	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
avginc	5.018677	.7073505	7.10	0.000	3.628251	6.409104
avginc2	-.0958052	.0289537	-3.31	0.001	-.1527191	-.0388913
avginc3	.0006855	.0003471	1.98	0.049	3.27e-06	.0013677
_cons	600.079	5.102062	117.61	0.000	590.0499	610.108

Verifica dell'ipotesi nulla di linearità, contro l'alternativa che la funzione di regressione della popolazione sia quadratica e/o cubica, ovvero sia una polinomiale di grado fino a 3:

H_0: coefficienti di popolazione per $Income^2$ e $Income^3$ = 0
H_1: almeno uno di questi coefficienti è diverso da zero.

```
test avginc2 avginc3;  Eseguire il comando di test dopo aver eseguito la
regressione

 ( 1)  avginc2 = 0.0
 ( 2)  avginc3 = 0.0

       F(  2,    416) =   37.69
       Prob > F =    0.0000
```

L'ipotesi che la funzione di regressione della popolazione sia lineare viene rigettata al livello di significatività dell'1% contro l'alternativa che sia una polinomiale di grado fino a 3.

Riepilogo: funzioni di regressione polinomiali

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 \quad +\ldots+ \beta_r X_i^r + u_i$$

- Stima: via OLS dopo aver definito nuovi regressori
- I coefficienti hanno interpretazioni complicate
- Per interpretare la funzione di regressione stimata:
  - rappresentare graficamente i valori predetti come funzione di x
  - calcolare gli scarti predetti $\Delta Y/\Delta X$ per i diversi valori di x
- Le ipotesi sul grado r possono essere verificate tramite test t e F sugli appropriati blocchi di variabili.
- Scelta del grado r
  - rappresentare i dati graficamente, effettuare i test t e F, verificare la sensibilità e gli effetti stimati, giudicare.
  - *In alternativa usare il criterio di scelta del modello (più avanti)*

2. Funzioni logaritmiche di *Y* e/o *X*

- ln(*X*) = è il logaritmo naturale di *X*
- Le trasformazioni logaritmiche permettono di modellare le relazioni in termini "percentuali" (come l'elasticità) invece che linearmente.

Ecco perché: $\ln(x+\Delta x) - \ln(x) = \ln\left(1+\frac{\Delta x}{x}\right) \cong \frac{\Delta x}{x}$

Ricordate il calcolo del differenziale totale di y=ln(x) è

$$\mathrm{dy} = \frac{1}{x}dx$$

Numericamente:

quando x=100 e Δx=1 Δx/x=0,01

$$\ln(x+\Delta x)-\ln(x)= \ln\left(1+\frac{\Delta x}{x}\right) = \ln(1{,}01) = 0{,}00995 \cong 0{,}01;$$

quando x=100 e Δx=5 Δx/x=0,05

$$\ln\left(1+\frac{\Delta x}{x}\right) = 0{,}04879 \cong 0{,}05 \text{ (circa)}$$

Le tre specificazioni di regressione logaritmica:

Caso	Funzione di regressione della popolazione
I. lineare-log	$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$
II. log-lineare	$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$
III. log-log	$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$

- L'interpretazione del coefficiente pendenza è diversa in ciascun caso.
- L'interpretazione si trova applicando la regola generale "prima e dopo": predire la variazione in Y per una data variazione in X."
- Ogni caso ha una diversa interpretazione naturale (per piccole variazioni in X)

I. Funzione di regressione della popolazione lineare-logaritmica

Calcolare Y "prima" e "dopo" aver modificato la X:

$$Y = \beta_0 + \beta_1 \ln(X) \qquad \text{("prima")}$$

Ora cambiamo X: $\quad Y + \Delta Y = \beta_0 + \beta_1 \ln(X + \Delta X)$ ("dopo")

Sottrarre ("dopo") – ("prima"): $\quad \Delta Y = \beta_1 [\ln(X + \Delta X) - \ln(X)]$

ora $\quad \ln(X + \Delta X) - \ln(X) \cong \dfrac{\Delta X}{X}$,

quindi $\quad \Delta Y \cong \beta_1 \dfrac{\Delta X}{X}$

o $\quad \beta_1 \cong \dfrac{\Delta Y}{\Delta X / X}$ (per piccole ΔX)

Più intuitivamente, data l'equazione

$$Y = \beta_0 + \beta_1 \ln(X)$$

per calcolare ΔY possiamo scrivere

$$\Delta Y = \beta_1 \Delta \ln(X)$$

Ove $\Delta \ln(X) = [\ln(X + \Delta X) - \ln(X)] = \frac{\Delta X}{X}$
l'ultima uguaglianza sfrutta l'approssimazione che conosciamo

quindi $\Delta Y \cong \beta_1 \frac{\Delta X}{X}$

Ripartiamo da $\Delta Y \cong \beta_1 \frac{\Delta X}{X}$

Questa espressione può essere letta come: ΔY è la varaizione di Y a seguito di una variazione percentuale dell'1% di X.

Che detta in altre parole può essere riespressa come ***un incremento dell'1% in X (moltiplicare X per* 1,01*) è associato a una variazione di 0,*01β_1 *in Y***

O ancora:

Un aumento di 1% in X si associa ad un aumento di ***0,*01β_1 *in Y***

(1% incremento in *X* --> 0,01 incremento in ln(*X*)
--> 0,01β_1 incremento in *Y*)

Esempio: TestScore su ln(Income)

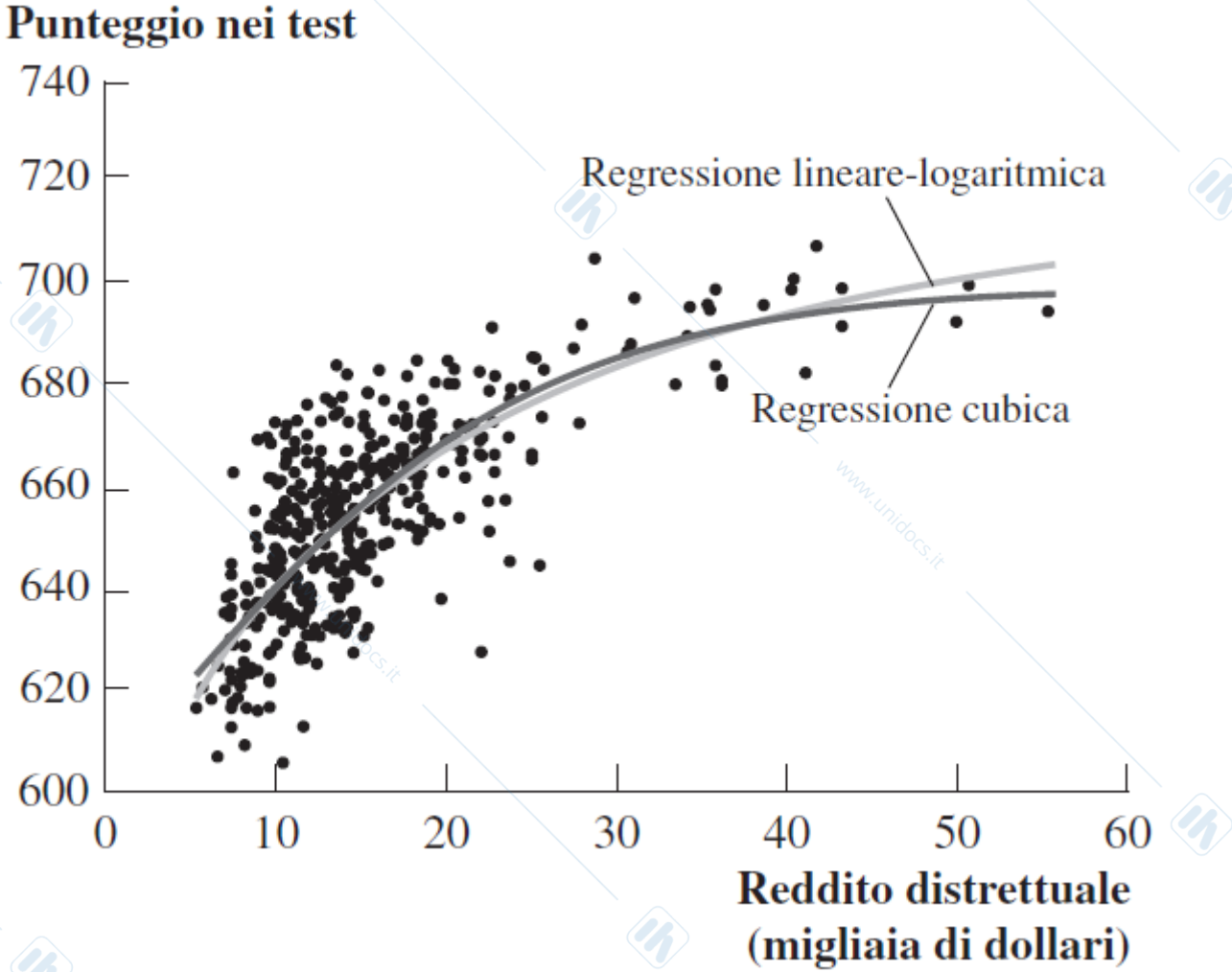
- Definiamo innanzitutto il nuovo regressore, ln(*Income*)
- Il modello è ora lineare su ln(*Income*), quindi possiamo stimare il modello lineare-log tramite OLS:

$$\widehat{TestScore} = 557{,}8 + 36{,}42 \times \ln(Income_i)$$
$$\qquad\qquad\quad (3{,}8) \quad (1{,}40)$$

quindi un incremento dell'1% in *Income* è associato a un aumento di 0,36 nel punteggio nei test.

- Si applicano tutti i soliti meccanismi di regressione: errori standard, intervalli di confidenza, R^2.
- Come confrontare tutto questo con il modello cubico?

Le funzioni di regressione lineare-logaritmica e cubica



Le due funzioni sono molto simili. Si può usare R2 che vale 0,561 per la logaritmica e 0,555 per la cubica.

La specificazione logaritmica ha un lieve vantaggio in termini di R2 e per giunta è più parsimoniosa adottiamo la logaritmica

II. Funzione di regressione della popolazione log-lineare

$$\ln(Y) = \beta_0 + \beta_1 X \qquad \text{(b)}$$

Variamo X: $\ln(Y + \Delta Y) = \beta_0 + \beta_1(X + \Delta X)$ (a)

Sottraiamo (a) – (b): $\ln(Y + \Delta Y) - \ln(Y) = \beta_1 \Delta X$

da cui $$\frac{\Delta Y}{Y} \cong \beta_1 \Delta X$$

o $$\beta_1 \cong \frac{\Delta Y / Y}{\Delta X} \quad \text{(per } \Delta X \textit{ piccole)}$$

Se X varia di 1 unità, $\Delta X=1$ ➔ $\Delta Y/Y$ varia di β_1

Così che β_1 può essere letto come la variazione % di Y a seguito di una variazione unitaria di X

detta semi-elasticità

Caso log-lineare (continua)

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$$

per piccole ΔX, $\beta_1 \cong \frac{\Delta Y/Y}{\Delta X}$

- ora $100 \times \frac{\Delta Y}{Y}$ = percentuale di variazione in Y, quindi ***una variazione in X di un'unità* ($\Delta X = 1$) *si associa a una variazione di* $100\beta_1$% *in Y*.**
- 1 unità di incremento in X → β_1 incremento in $\ln(Y)$
  → $100\beta_1$% incremento in Y
- *Nota*: quali sono le unità di u_i e SER?
  - deviazioni frazionali (proporzionali)
  - per esempio $SER = 0{,}2$ significa…

Caso log-lineare (continua)

$$\widehat{\ln(earnings)} = 2.805 + 0.0087 \times (\text{Age}_i)$$
$$(0.018) \quad (0.0004)$$

Secondo questa regressione ci si attende che le retribuzioni aumentino di 0.087% [(100*0,0087)%] per ogni anno di età in più.

III. Funzione di regressione della popolazione log-log

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i \qquad \text{(b)}$$

Variamo X: $\ln(Y + \Delta Y) = \beta_0 + \beta_1 \ln(X + \Delta X)$ (a)

Sottraiamo: $\ln(Y + \Delta Y) - \ln(Y) = \beta_1 [\ln(X + \beta X) - \ln(X)]$

Da cui $$\frac{\Delta Y}{Y} \cong \beta_1 \quad \frac{\Delta X}{X}$$

O $$\beta_1 \cong \frac{\Delta Y / Y}{\Delta X / X}$$ (per piccole ΔX)

Caso log-log (continua)

$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$

per piccole ΔX,

$$\beta_1 \cong \frac{\Delta Y / Y}{\Delta X / X}$$

Ora $100 \times \frac{\Delta Y}{Y}$ = variazione percentuale in Y, e $100 \times \frac{\Delta X}{X}$ = variazione percentuale in X, per cui ***una variazione dell'1% in X produce una variazione del β_1% in Y***.

Nella specifica log-log, β_1 ha l'interpretazione di un coefficiente di elasticità.

Esempio: ln(TestScore) su ln(Income)

- Per prima cosa definiamo una nuova variabile dipendente, ln(*TestScore*) ***e*** il nuovo regressore, ln(*Income*)
- Il modello ora è una regressione lineare di ln(*TestScore*) su ln(*Income*) che può essere stimata mediante OLS:

$$\widehat{\ln(TestScore)} = 6{,}336 + 0{,}0554 \times \ln(Income_i)$$
$$(0{,}006)\ \ (0{,}0021)$$

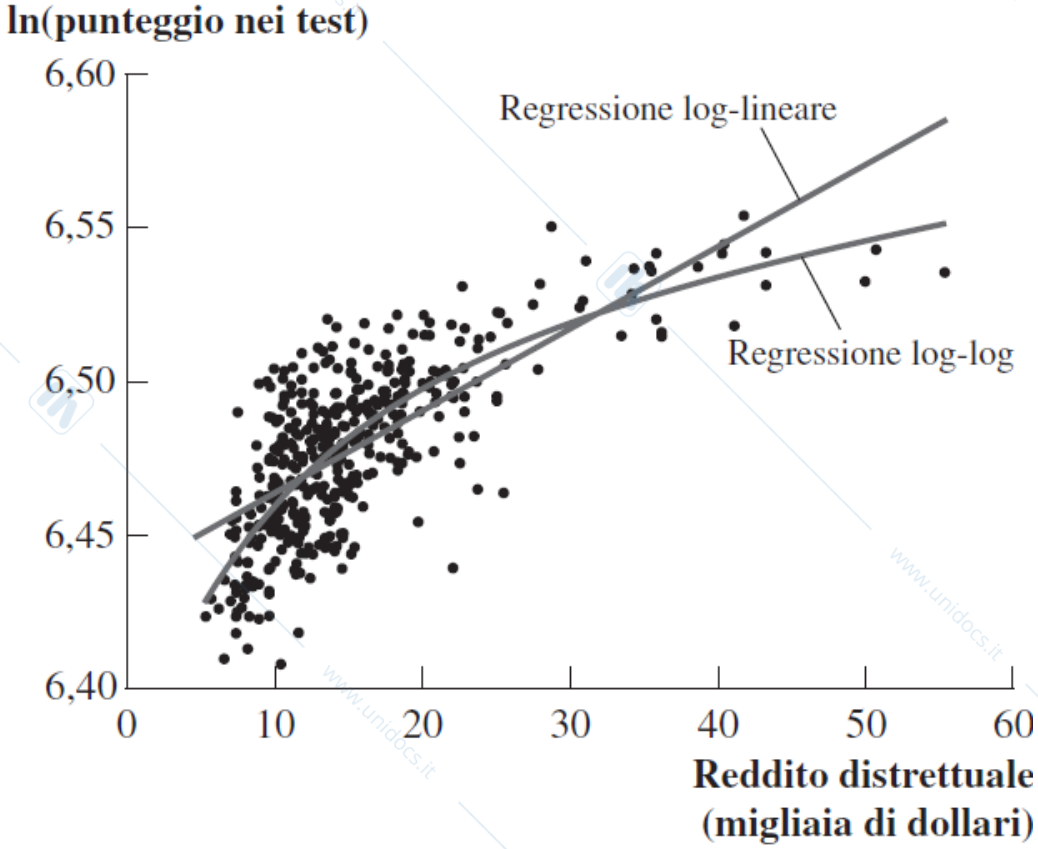
A un aumento dell'1% in *Income* si associa un aumento dello 0.0554% in *TestScore* (*Income* aumenta di un fattore 1,01, *TestScore* di un fattore 1,000554)

Esempio: ln(TestScore) su ln(Income) (continua)

$$\widehat{\ln(TestScore)} = 6{,}336 + 0{,}0554 \times \ln(Income_i)$$
$$(0{,}006)\ \ (0{,}0021)$$

- Per esempio, supponiamo che il reddito salga da 10,000$ a 11,000$, o del 10%. Quindi *TestScore* cresce approssimativamente di 0,0554 × 10% = 0,554%. Se *TestScore* = 650, questo corrisponde a un aumento di 0,00554 × 650 = 3,6 punti.
- Come si confronta rispetto al modello log-lineare?

Le specifiche log-lineare e log-log:



- *Notate l'asse verticale*
- *Niente sembra adattarsi meglio della cubica o lineare-log, almeno in base all'aspetto visivo (il confronto formale è dificile perché le variabili dipendenti differiscono)*

Riepilogo: trasformazioni logaritmiche

- Tre casi, differiscono in base alla o alle variabili Y e/o X trasformate in logaritmi.
- La regressione diventa lineare sulla(e) nuova(e) variabile(i) $\ln(Y)$ e/o $\ln(X)$, mentre i coefficienti possono essere stimati attraverso l'OLS.
- I test di ipotesi e gli intervalli di confidenza possono essere implementati e interpretati "nel solito modo"
- L'interpretazione di β_1 differisce caso per caso.

La scelta della specificazione (forma funzionale) dev'essere guidata dal ragionamento – quale interpretazione ha più senso nella vostra applicazione? – da test e dall'analisi grafica dei valori predetti

Altre funzioni non lineari (e minimi quadrati non lineari) (Appendice 8.1)

Le funzioni di regressione precedenti hanno delle limitazioni…

- Polinomiali: il punteggio nei test può decrescere all'aumentare del reddito, possono produrre una pendenza negative per alcuni valori del reddito,
- Lineare-log: il punteggio aumenta con il reddito, ma senza limite
- Questa è una funzione non lineare in cui la *Y* cresce sempre con *X* e c'è un massimo valore di *Y* (asintoto):

$$Y = \beta_0 - \alpha e^{-\beta_1 X}$$

β_0, β_1 e α sono parametri sconosciuti. Viene chiamata curva di crescita esponenziale negativa. L'asintoto per $X \rightarrow \infty$ è β_0.

Crescita esponenziale negativa

Vogliamo stimare i parametri di

$$Y_i = \beta_0 - \alpha e^{-\beta_1 X_i} + u_i$$

o

$$Y_i = \beta_0\left[1 - e^{-\beta_1 (X_i - \beta_2)}\right] + u_i \qquad (*)$$

dove $\alpha = \beta_0 e^{\beta_2}$ (perché vogliamo farlo?)

Compariamo il modello (*) con quelli lineare-log e cubico:

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_2 X_i^3 + u_i$$

I modelli lineare-log e polinomiale sono *lineari nei parametri* β_0 e β_1 – mentre il modello (*) no.

Minimi quadrati non lineari

- I modelli i cui parametri sono lineari possono essere stimati tramite OLS.
- I modelli non lineari in uno o più parametri possono essere stimati con i minimi quadrati non lineari (NLS) ma non tramite gli OLS.
- Il problema NLS per la specificazione proposta:

$$\min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n} \{Y_i - f(X_1, X_2, \dots, X_k, \beta_1, \beta_2, \dots \beta_k)\}^2$$

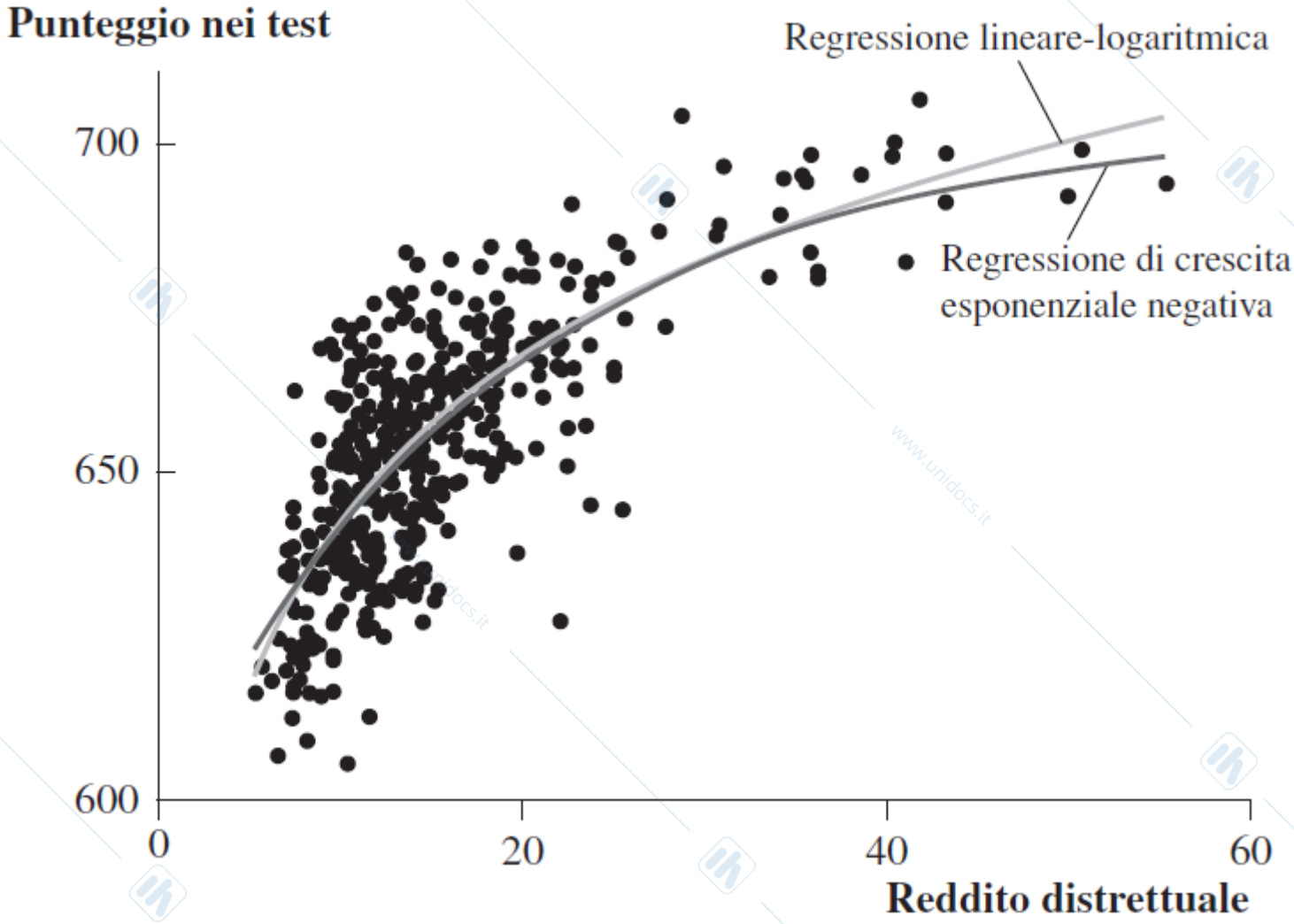
$$\min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n} \left\{Y_i - \beta_0\left[1 - e^{-\beta_1(X_i-\beta_2)}\right]\right\}^2$$

È un problema di minimizzazione non lineare (un problema di "hill-climbing"). Come risolverlo?

- Tirare a indovinare e verificare …
- Implementazione in STATA…

```
. nl (testscr = {b0=720}*(1 - exp(-1*{b1}*(avginc-{b2})))), r

(obs = 420)
Iteration 0:  residual SS =  1.80e+08                    .
Iteration 1:  residual SS =  3.84e+07                    .
Iteration 2:  residual SS =   4637400                    .
Iteration 3:  residual SS =  300290.9      STATA sta "scalando la collina"
Iteration 4:  residual SS =  70672.13      (minimizzando l'SSR)
Iteration 5:  residual SS =  66990.31                    .
Iteration 6:  residual SS =   66988.4                    .
Iteration 7:  residual SS =   66988.4                    .
Iteration 8:  residual SS =   66988.4

Nonlinear regression with robust standard errors      Number of obs =       420
                                                      F(  3,    417) = 687015.55
                                                      Prob > F      =    0.0000
                                                      R-squared     =    0.9996
                                                      Root MSE      =   12.67453
                                                      Res. dev.     =   3322.157
------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          b0 |   703.2222   4.438003   158.45   0.000     694.4986    711.9459
          b1 |   .0552339   .0068214     8.10   0.000     .0418253    .0686425
          b2 |  -34.00364    4.47778    -7.59   0.000    -42.80547    -25.2018
------------------------------------------------------------------------------
 (SEs, P values, CIs, and correlations are asymptotic approximations)
```

Crescita esponenziale negativa; *RMSE* = 12,675
Linear-log; *RMSE* = 12,618



Minimi quadrati non lineari (cont)

- Sotto condizioni generali delle forma della *f*

Y=f(X_1, X_2,…, X_k, β_1, β_2,…, β_k) lo stimatore NLS ha due proprietà fondamentali incomune con OLS: è consistente e si distribuisce normalmente in grandi campioni.

- L'inferenza circa I parametri può procedure come al solito: la t può essere costruita come di consueto

$$\widehat{TestScore} = 703{,}2\ [1-e^{-0.0552(income+34)}]$$
$$(4{,}44) \qquad (0{,}0068) \quad (4{,}48)$$

Interazioni tra variabili indipendenti (Paragrafo 8.3)

- Forse ridurre la dimensione di una classe è più efficace in alcune circostanze che in altre…
- Forse classi più piccole sono migliori se ci sono molti allievi non di madrelingua, che richiedono attenzioni individuali
- Ovvero, $\frac{\Delta TestScore}{\Delta STR}$ può dipendere da *PctEL*
- Più in generale, $\frac{\Delta Y}{\Delta X_1}$ può dipendere da X_2
- Come modellare queste "interazioni" tra X_1 e X_2?
- Consideriamo prima delle *X binarie*, poi delle *X continue*

(a) Interazioni tra due variabili binarie

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

- D_{1i}, D_{2i} sono binarie
- β_1 è l'effetto che si ha cambiando $D_1=0$ in $D_1=1$. In questa specificazione, *questo effetto non dipende dal valore di* D_2.
- Per far sì che la modifica di D_1 dipenda da D_2, si inserisce il "termine d'interazione" $D_{1i} \times D_{2i}$ come regressore:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$$

Interpretazione dei coefficienti

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$$

Regola generale: confrontare i vari casi

$$E(Y_i | D_{1i}=0, D_{2i}=d_2) = \beta_0 + \beta_2 d_2 \qquad \text{(b)}$$

$$E(Y_i | D_{1i}=1, D_{2i}=d_2) = \beta_0 + \beta_1 + \beta_2 d_2 + \beta_3 d_2 \qquad \text{(a)}$$

sottrarre (a) – (b): ie dY/dD_{1i}

$$E(Y_i | D_{1i}=1, D_{2i}=d_2) - E(Y_i | D_{1i}=0, D_{2i}=d_2) = \beta_1 + \beta_3 d_2$$

- L'effetto di D_1 dipende da d_2 (quel che volevamo)
- β_3 = incremento dell'effetto di D_1, quando $D_2 = 1$

Esempio: *TestScore, STR, allievi non di madrelingua*

Sia

$HiSTR = \begin{cases} 1 \text{ se } STR \geq 20 \\ 0 \text{ se } STR < 20 \end{cases}$ e $HiEL = \begin{cases} 1 \text{ se } PctEL \geq 10 \\ 0 \text{ se } PctEL < 10 \end{cases}$

$$\widehat{TestScore} = \underset{(1,4)}{664,1} - \underset{(2,3)}{18,2}HiEL - \underset{(1,9)}{1,9}HiSTR - \underset{(3,1)}{3,5}(HiSTR \times HiEL)$$

- "Effetto" di *HiSTR* quando *HiEL* = 0 è –1,9
- "Effetto" di *HiSTR* quando *HiEL* = 1 è –1,9 – 3,5 = –5,4

$dY/dHiSTR|_{HiEL=1}$ *derivata wrt Histr calcolata a HiEL=1*

Esempio: *TestScore, STR, allievi non di madrelingua (continua)*

La lettura di questa stima è dunque:

- Si stima che la riduzione della dimensione della classe abbia un effetto maggiore quando la percentuale degli allievi non di madrelingua è elevata

- Questa interazione non è statisticamente significativa: t = 3.5/3.1

Esempio: *TestScore, STR, allievi non di madrelingua (continua)*

Siano

$$HiSTR = \begin{cases} 1 \text{ se } STR \geq 20 \\ 0 \text{ se } STR < 20 \end{cases} \quad \text{e} \quad HiEL = \begin{cases} 1 \text{ se } PctEL \geq 10 \\ 0 \text{ se } PctEL < 10 \end{cases}$$

$$\widehat{TestScore} = \underset{(1,4)}{664,1} - \underset{(2,3)}{18,2}HiEL - \underset{(1,9)}{1,9}HiSTR - \underset{(3,1)}{3,5}(HiSTR \times HiEL)$$

- Valori predetti $\widehat{TestScore}$ per diverse combinazione delle dummy

	STR basso	***STR elevato***
HiEL basso	664,1	662,2
HiEL elevato	645,9	640,5

Esempio: *TestScore, STR, allievi non di madrelingua (continua)*

- Valori predetti $\widehat{TestScore}$ per diverse combinazione delle dummy

	STR basso	***STR elevato***
HiEL basso	664,1	662,2=664,1-1,9
HiEL elevato	645,9=664,1-18,2	640,5=664,1-18,2-1,9-3,5

(b) Interazioni tra variabili continue e binarie

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

- D_i è binaria, X è continua
- Come specificato prima, l'effetto su Y di X (tenendo costante D) = β_2, che non dipende da D
- Per far sì che l'effetto di X *dipenda da* D, includiamo il "termine d'interazione" $D_i \times X_i$ come regressore:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 (D_i \times X_i) + u_i$$

Interazioni tra variabili continue e binarie: le due rette di regressione

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 (D_i \times X_i) + u_i$$

Osservazione con $D_i = 0$ (il gruppo "$D = 0$"):

$$Y_i = \beta_0 + \beta_2 X_i + u_i$$ ***Retta di regressione con D=0***

Osservazione con $D_i = 1$ (il gruppo "$D = 1$"):

$$Y_i = \beta_0 + \beta_1 + \beta_2 X_i + \beta_3 X_i + u_i$$
$$= (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_i + u_i$$ ***Retta di regressione con D=1***

Interpretazione dei coefficienti

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 (D_i \times X_i) + u_i$$

Regola generale: confrontare i diversi casi

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 (D \times X) \qquad \text{(b)}$$

Ora cambiamo X:

$$Y + \Delta Y = \beta_0 + \beta_1 D + \beta_2 (X + \Delta X) + \beta_3 [D \times (X + \Delta X)] \qquad \text{(a)}$$

sottrarre (a) – (b):

$$\Delta Y = \beta_2 \Delta X + \beta_3 D \Delta X \quad \text{o} \quad \frac{\Delta Y}{\Delta X} = \beta_2 + \beta_3 D$$

- L'effetto di X dipende da D (quel che volevamo)
- β_3 = incremento dell'effetto di X, quando $D = 1$

Interpretazione dei coefficienti (continua)

Analizziamo separatamente i diversi casi:

Caso a) $Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i$ no interazione

Quando D=0 la retta è $\beta_0 + \beta_1 X_i$

Quando D=1 $(\beta_0 + \beta_2) + \beta_1 X_i$ (figura a)

Caso b) $Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (D_i X_i) + u_i$

la D entra da sola e nell'interazione

Quando D=0 la retta è $\beta_0 + \beta_1 X_i$

Quando D=1 $(\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i$ (figura b)

Interpretazione dei coefficienti (continua)

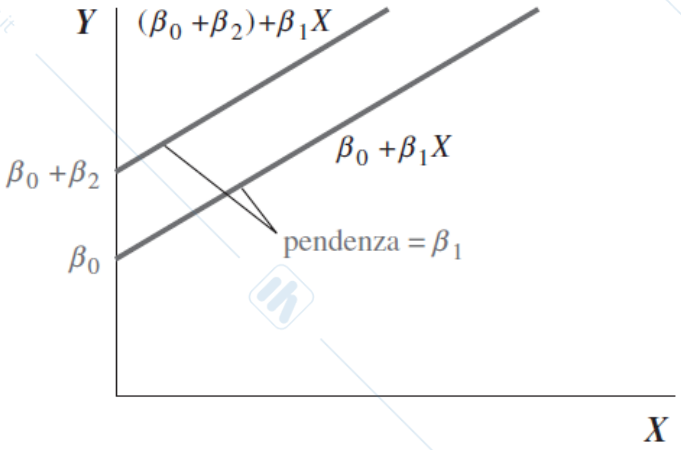
Caso c) $Y_i = \beta_0 + \beta_1 X_i + \beta_2 (D_i X_i) + u_i$

la D entra solo nell'interazione

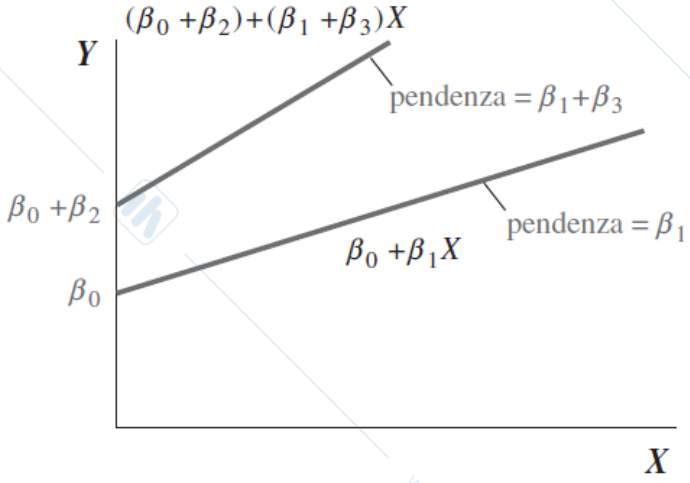
Quando D=0 la retta è $\beta_0 + \beta_1 X_i$

Quando D=1 $\beta_0 + (\beta_1 + \beta_2) X_i$ (figura c)

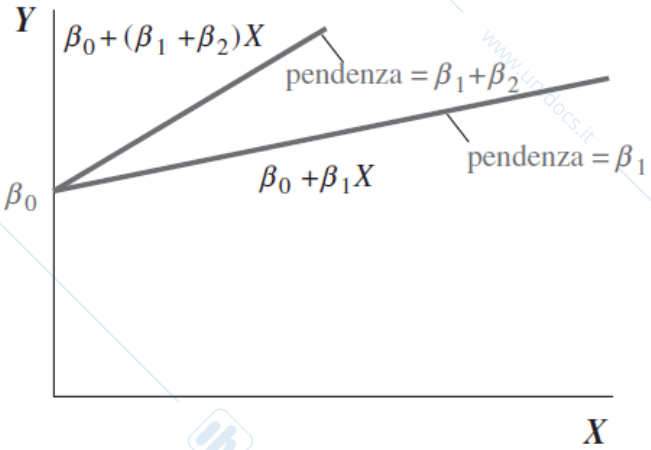
Interazioni tra variabili continue e binarie (continua)



(a) Intercette diverse, pendenze uguali



(b) Intercette diverse, pendenze diverse



(c) Intercette uguali, pendenze diverse

Esempio: *TestScore, STR, HiEL* (=1 se *PctEL* ≥ 10)

$$\widehat{TestScore} = 682{,}2 - 0{,}97STR + 5{,}6HiEL - 1{,}28(STR \times HiEL)$$
$$(11{,}9)\ (0{,}59) \qquad (19{,}5) \qquad (0{,}97)$$

Quando *HiEL* = 0:

$$\widehat{TestScore} = 682{,}2 - 0{,}97STR$$

- Quando *HiEL* = 1,

$$\widehat{TestScore} = 682{,}2 - 0{,}97STR + 5{,}6 - 1{,}28STR$$

$$\widehat{TestScore} = 687{,}8 - 2{,}25STR$$

- Due rette di regressione: una per ciascun gruppo *HiSTR*.
- Si stima che una riduzione della dimensione della classe abbia un effetto maggiore quanto più è ampia la percentuale degli studenti non di madrelingua ➔ perchè si sommano i coefficienti b1 e b3

Esempio (continua): verifica delle ipotesi

$$\widehat{TestScore} = 682{,}2 - 0{,}97STR + 5{,}6HiEL - 1{,}28(STR \times HiEL)$$
$$(11{,}9) \quad (0{,}59) \quad (19{,}5) \quad (0{,}97)$$

- Le due rette di regressione hanno la stessa **pendenza** ← → il coefficiente su *STR* × *HiEL* è zero: $t = -1{,}28/0{,}97 = -1{,}32$
- Le due rette di regressione hanno lo stesso **punto di intercetta** ← → il coefficiente di *HiEL* è zero: $t = -5{,}6/19{,}5 = 0{,}29$
- Le due rette di regressione coincidono ← → il coefficiente di *HiEL* = 0 ***e*** quello di *STR* × *HiEL* = 0: $F = 89{,}94$ (valore-p < 0,001) **!!**
- Scartiamo le ipotesi congiunte ma non quelle individuali (come può essere?)

Esempio (continua): verifica delle ipotesi

- Scartiamo le ipotesi congiunte ma non quelle individuali (come può essere?)

- Le due variabili (HiEL e HiELxSTR) sono fortemente correlate e questo dà luogo ad alti SE per i singoli coefficienti.

- Infine è possibile testare che tutti i coefficienti di STR entrino significativamente costruendo un test congiunto F per STR e l'interazione. F=5,64 p-val=0,004 ➔ I coefficienti di STR sono significative all'1%

(c) Interazioni tra due variabili continue

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

- X_1, X_2 sono continue
- Come specificato, l'effetto di X_1 non dipende da X_2
- Come specificato, l'effetto di X_2 non dipende da X_1
- Per far sì che l'effetto di X_1 dipenda da X_2, includiamo il "termine d'interazione" $X_{1i} \times X_{2i}$ come regressore:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$$

Interpretazione dei coefficienti nella interazione tra vbl continue. Variazione di X_1

Regola generale: comparazione dei vari casi

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) \qquad \text{(a)}$$

Ora facciamo variare X_1 di ΔX_1

$$Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2 + \beta_3 [(X_1 + \Delta X_1) \times X_2] \qquad \text{(b)}$$

Sottraiamo (b) – (a):

$$\Delta Y = \beta_1 \Delta X_1 + \beta_3 X_2 \Delta X_1 \quad \text{o} \quad \frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2$$

- L'effetto di X_1 dipende da X_2 (quel che volevamo)
- β_3 = incremento dell'effetto di X_1 a seguito dell'aumento di un'unità di X_2

Interpretazione dei coefficienti nella interazione tra vbl continue. Variazione di X_1 (continua)

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2$$

Supponiamo β_3 positive

ln(y)=log(retribuzioni)

X_1 =anni di esperienza lavorativa

X_2 = grado di istruzione

L'effetto su log(Y) di un anno di esperienza in più cresce di un ammontare pari a β_3 per ogni in più di istruzione del lavoratore

β_0 è per tutti, in più il salario aumenta di β_3 per ogni anno di istruzione in più

Interpretazione dei coefficienti nella interazione tra vbl continue. Variazione di X_1 (continua)

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2$$

Questo spiega il senso della frase

β_3 = incremento dell'effetto di X_1 a seguito dell'aumento di un'unità di X_2

Interpretazione dei coefficienti nella interazione tra vbl continue. Variazione di X_2

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i \qquad \text{(a)}$$

Ora facciamo variare X_2 di ΔX_2

$$Y + \Delta Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_2 + \Delta X_2) + \beta_3 [(X_2 + \Delta X_2) \times X_1] \qquad \text{(c)}$$

Sottraiamo (c) – (a): (stiamo applicando il concetto chiave 1)

$$\Delta Y = \beta_2 \Delta X_2 + \beta_3 X_2 \Delta X_2 \quad \text{o} \quad \frac{\Delta Y}{\Delta X_2} = \beta_2 + \beta_3 X_1$$

- L'effetto di X_2 dipende da X_1 (quel che volevamo)
- β_3 = incremento dell'effetto di X_2 a seguito dell'aumento di un'unità di X_1

Interpretazione dei coefficienti nella interazione tra vbl continue. Variazione di X_1 e X_2

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i \qquad \text{(a)}$$

facendo variare sia X_1 che X_2

$$Y+\Delta Y = \beta_0 + \beta_1 (X_1+\Delta X_1) + \beta_2 (X_2+\Delta X_2) + \beta_3 [(X_2+\Delta X_2) \times (X_1+\Delta X_2)] \qquad \text{(d)}$$

Sottraendo (d) – (a):

$$\Delta Y = (\beta_1+\beta_3 X_2)\, \Delta X_1 + (\beta_2+\beta_3 X_1) \Delta X_2 + \beta_3 \Delta X_2 \Delta X_1$$

- Il primo termine è l'effetto della ΔX_1 tenuto costante X_2
- Il secondo termine similmente per ΔX_2 con X_1 costante
- Il terzo termine è l'effetto aggiuntivo di una variazione sia di X_1 che di X_2

Esempio: *TestScore, STR, PctEL*

$$\widehat{TestScore} = \underset{(11,8)}{686,3} - \underset{(0,59)}{1,12}STR - \underset{(0,37)}{0,67}PctEL + \underset{(0,019)}{0,0012}(STR \times PctEL),$$

L'effetto stimato della riduzione della dimensione della classe è non lineare, perché la dimensione dell'effetto stesso dipende da *PctEL*:

$$\frac{\Delta TestScore}{\Delta STR} = -1,12 + 0,0012PctEL$$

PctEL	$\frac{\Delta TestScore}{\Delta STR}$
0	$-1,12$
20%	$-1,12+0,0012 \times 20 = -1,10$

Esempio (continua): verifica delle ipotesi

$$\widehat{TestScore} = \underset{(11,8)}{686,3} - \underset{(0,59)}{1,12}STR - \underset{(0,37)}{0,67}PctEL + \underset{(0,019)}{0,0012}(STR \times PctEL),$$

- Il coefficiente di $STR \times PctEL$ è = 0?

t = 0,0012/0,019 = 0,06 → non si rifiuta al 5%

- Il coefficiente di STR è = 0?

t = –1,12/0,59 = –1,90 → non si rifiuta al 5%

- I coefficienti di **entrambi** STR ***e*** $STR \times PctEL$ sono = 0?

F = 3,89 (valore-p = 0,021) → si rifiuta al 5% (!!)

(Perché? Multicollinearità alta ma imperfetta, vedi prima)

Applicazione: effetti non lineari del rapporto studenti/insegnanti sui punteggi nei test (Paragrafo 8.4)

Le specificazioni non lineari ci permettono di esaminare dettagli meno evidenti della relazione tra punteggi nei test e STR, quali:

1. Ci sono effetti non lineari della riduzione della dimensione della classe sui punteggi nei test? (Una riduzione da 35 a 30 ha lo stesso effetto di una riduzione da 20 a 15?)
2. Ci sono interazioni non lineari tra *PctEL* e *STR*? (Le classi piccole sono più efficaci quando ci sono molti studenti non di madrelingua?)

Strategia per la domanda #1 (effetti diversi per STR diversi?)

- Stimare funzioni lineari e non lineari di *STR*, mantenendo costanti le rilevanti variabili demografiche
  - *PctEL*
  - *Income* (si ricordi la relazione non lineare tra punteggio nei test e reddito)
  - *LunchPCT* (pranzo libero /sovvenzionato)
- Verificare se aggiungendo dei termini non lineari si ha una differenza quantitativa "economicamente rilevante" (l'importanza "economica" o "reale" è diversa e quindi statisticamente significativa)
- Verificare se i termini non lineari sono significativi

Strategia per la domanda #2 (interazioni tra *PctEL* e *STR*?)

- Stimare le funzioni lineari e non lineari di *STR*, con l'interazione di *PctEL*.
- Se la specificazione è non lineare (con *STR*, STR^2, STR^3), allora occorre aggiungere interazioni con tutti i termini, in modo che la risultante forma funzionale possa essere diversa, al variare del livello di *PctEL*.
- Utilizzare una specificazione con interazione binaria-continua aggiungendo $HiEL \times STR$, $HiEL \times STR^2$ e $HiEL \times STR^3$.

Qual è una buona specificazione di "base"?

- La relazione *Punteggio nei test – Reddito*:
- La specificazione logaritmica si comporta meglio verso gli estremi del campione, specialmente per valori di reddito alti.

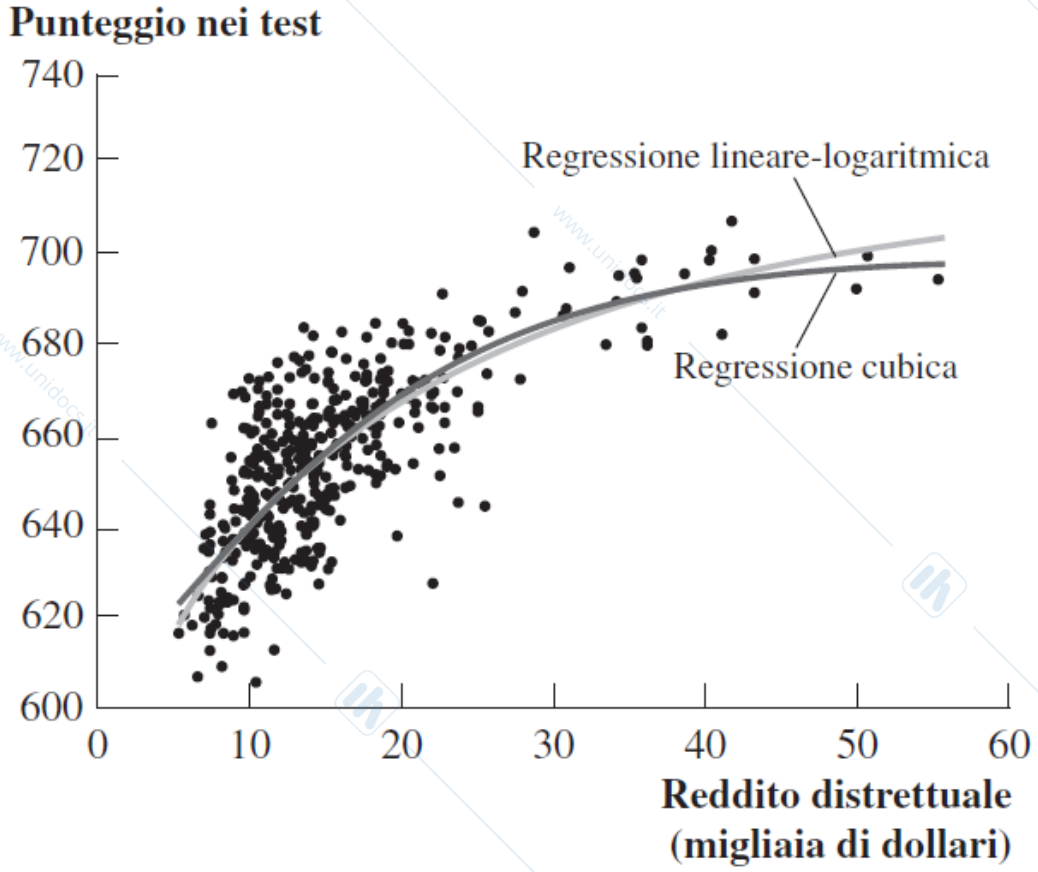


Tabella 8.3 Modelli di regressione non-lineari per il punteggio nei test.

Variabile dipendente: media del punteggio nei test del distretto; 420 osservazioni

Regressori	(1)	(2)	(3)	(4)	(5)	(6)	(7)
STR	– 1,00** (0,27)	– 0,73** (0,26)	– 0,97 (0,59)	– 0,53 (0,34)	64,33** (24,86)	83,70** (28,50)	65,29** (25,26)
STR^2					– 3,42** (1,25)	– 4,38** (1,44)	– 3,47** (1,27)
STR^3					0,059** (0,021)	0,075** (0,024)	0,060** (0,021)
% studenti non madrelingua	– 0,122** (0,033)	– 0,176** (0,034)					– 0,166** (0,034)
% studenti non madrelingua ≥ 10% (Binario, *HiEL*)			5,64 (19,51)	5,50 (9,80)	– 5,47 (1,03)	816,1* (327,7)	
HiEL × *STR*			– 1,28 (0,97)	– 0,58 (0,50)		– 123,3* (50,2)	
HiEL × STR^2						6,12* (2,54)	
HiEL × STR^3						– 0,101* (0,043)	
% aventi diritto al sussidio mensa	– 0,547** (0,024)	– 0,398** (0,033)		– 0,411** (0,029)	– 0,420** (0,029)	– 0,418** (0,029)	– 0,402** (0,033)
Reddito medio nel distretto (logaritmo)		11,57** (1,819		12,12** (1,80	11,75** (1,78	11,80** (1,78	11,51** (1,81
Intercetta	700,2** (5,6)	658,6** (8,6)	682,2** (11,9)	653,6** (9,9)	252,0 (163,6)	122,3 (185,5)	244,8 (165,7)

- Poichè la (1) manca del (log) reddito per prima cosa controlliamo che la stima sia distorta
- Nella (2) l'inclusione del (log) reddito cambia molto la stima di STR ➔ distorta, teniamo il reddito. Si usa il log perchè è emerso dal par 8.3 che è la miglior specificazione
- Nella (3) si mette HiEL binario da solo e interagito e senza il reddito, ➔ NS
- Nella (4) si rimettono le variabili economiche STR cambia ma l'interazione di HiEL ancora NS
- La (5) controlla per effetti quadratici e cubici di STR e li testa con F sotto alla tavola ➔ rifiuta la nulla

- La (6) controlla se l'effetto di STR dipende anche dallla presenza di studenti non di madre lingua, per questo interagisce HiEL con STR e i sui quadrati e cubi
- Sotto la tavola si testa le interazioni ➔ rifiuto la nulla
- La (7) è una variante della (5) in cui si usa la vbl continua anzichè binaria per la percentuale di student non di madre lingua.

Verifica di ipotesi congiunte:

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Statistiche *F* e valori-*p* per le ipotesi congiunte							
(a) tutte le variabili *STR* e le interazioni = 0			5,64 (0,004)	5,92 (0,003)	6,31 (< 0,001)	4,96 (< 0,001)	5,91 (< 0,001)
(b) STR^2 e $STR^3 = 0$					6,17 (< 0,001)	5,81 (0,003)	5,96 (0,003)
(c) *HiEL* × *STR*, *HiEL* × STR^2, *HiEL* × $STR^2 = 0$						2,96 (0,046)	
SER	9,08	8,64	15,88	8,63	8,56	8,55	8,57
$\bar{R}^2$	0,773	0,794	0,305	0,795	0,798	0,799	0,798

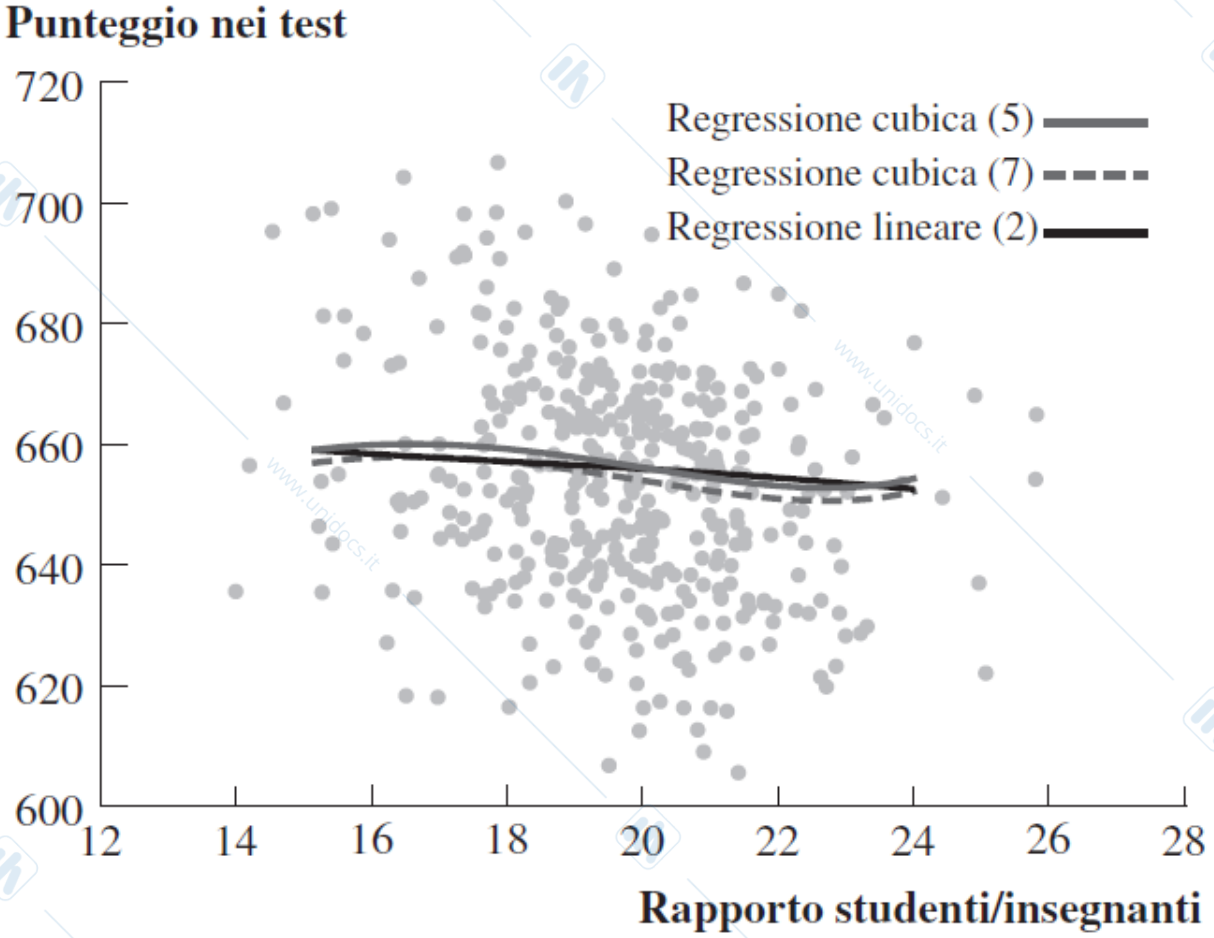
Queste regressioni sono state stimate utilizzando i dati sui distretti scolastici K-8 della California, descritti nell'Appendice 4.1. Gli errori standard sono riportati in parentesi sotto i coefficienti e i valori-*p* sono riportati in parentesi sotto le statistiche *F*. I coefficienti sono singolarmente significativi al livello del *5% o dell'**1%.

Che cosa potete concludere sulla domanda #1?

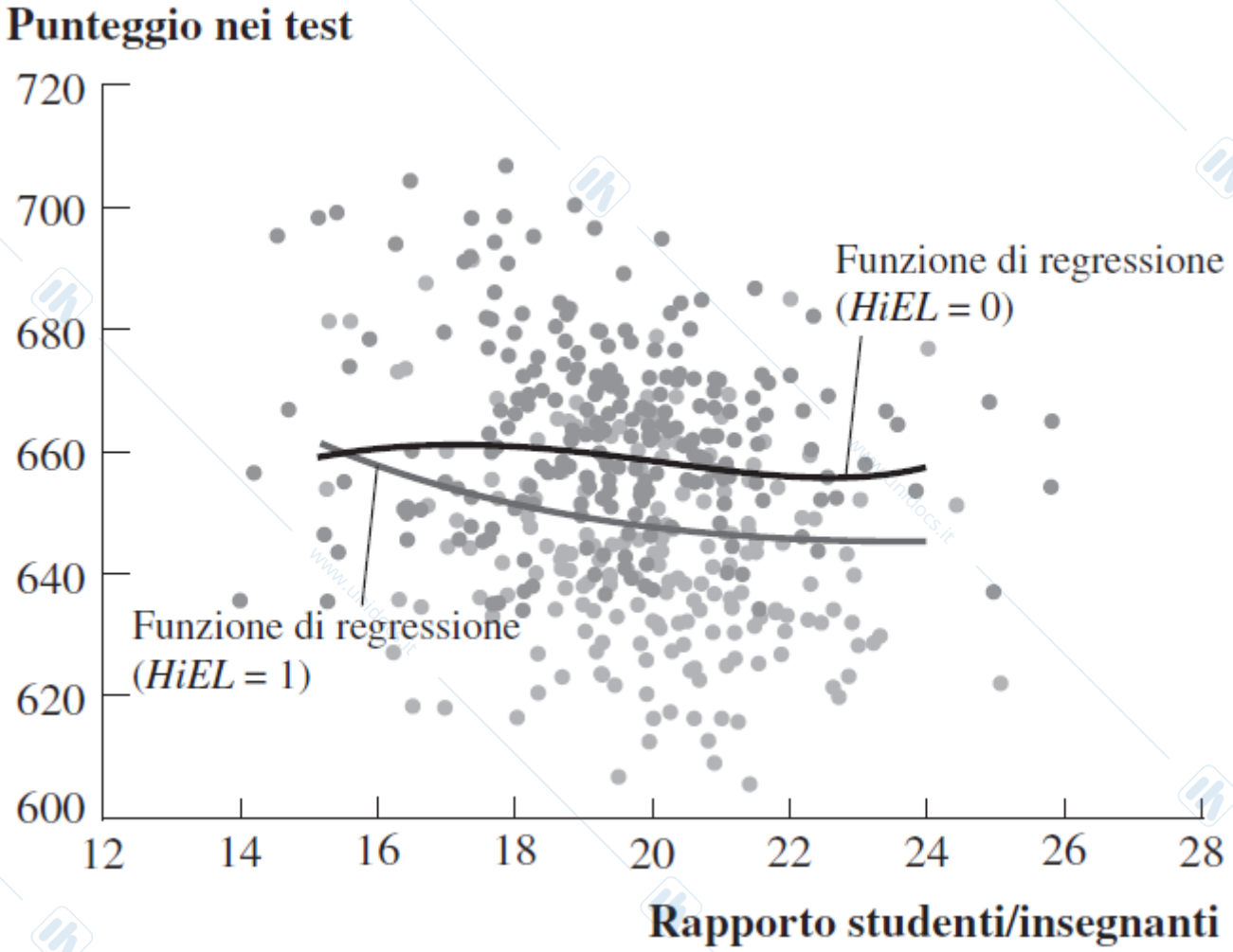
E sulla domanda #2?

Interpretazione delle funzioni di regressione per via grafica:

Per prima cosa, confrontate le specificazioni lineari e non lineari:



Quindi confrontate le regressioni con le interazioni:



Riepilogo: funzioni di regressione non lineari

- Utilizzando funzioni di variabili indipendenti come ln(X) o $X_1 \times X_2$, possiamo riformulare una vasta famiglia di funzioni di regressione lineare come regressioni multiple.
- La stima e l'inferenza procedono in modo analogo al modello di regressione lineare multiplo.
- L'interpretazione dei coefficienti è specifica del modello utilizzato, ma la regola generale consiste nel calcolare gli effetti confrontando i casi diversi (i diversi valori delle X originali)
- Sono possibili molte specificazioni non lineari, per cui è necessario riflettere:
  - Quali effetti non lineari si vogliono analizzare?
  - Quale ha senso nella particolare applicazione considerata?

Econometria I

Corso di laurea magistrale in Finanza e Assicurazioni

Slides per lezioni

secondo semestre 2020

Marco Ventura

Cap. 9 SW

Sommario

1. Validità interna ed esterna
2. Minacce alla validità interna
   a) Distorsione da variabili omesse
   b) Incorretta specificazione della forma funzionale
   c) Distorsione da errori nelle variabili
   d) Distorsione da dati mancanti e selezione campionaria
   e) Distorsione da causalità simultanea
3. Applicazione ai punteggi nei test

Validità interna ed esterna

- Facciamo un passo indietro e diamo uno sguardo più ampio alla regressione. Esiste un modo sistematico per valutare (criticare) gli studi di regressione? Sono noti i punti di forza della regressione multipla – ma quali sono le insidie?
  - Verranno elencate le ragioni più comuni per cui le stime di regressione multipla, fondate su dati basati sull'osservazione, possono produrre stime distorte **sull'effetto causale** di interesse.
  - Nell'applicazione sui punteggi nei test si cercherà di affrontare queste minacce nel modo migliore possibile – e di individuare i rischi ancora presenti. Dopo tutto questo lavoro, che cosa si sarà appreso sull'effetto sui punteggi nei test della riduzione delle dimensioni delle classi?

Quadro di riferimento per la valutazione di studi statistici: validità interna ed esterna (Paragrafo 9.1)

- ***Validità interna***: le inferenze statistiche sugli effetti causali sono valide per la popolazione studiata.

- ***Validità esterna***: le inferenze statistiche possono essere generalizzate dalla popolazione e dal contesto studiati ad altre popolazioni e altri contesti, dove il "contesto" fa riferimento all'ambiente legale, istituzionale e fisico e alle caratteristiche salienti. In una parola nel tempo e nello spazio.

Minacce alla validità esterna degli studi di regressione multipla

La valutazione delle minacce alla validità esterna richiede una conoscenza e un giudizio dettagliati e sostanziali caso per caso.

Fino a che punto è possibile generalizzare i risultati sulle dimensioni delle classi in California?

- Differenze nelle popolazioni
  - California nel 2011?
  - Massachusetts nel 2011?
  - Mexico nel 2011?
- Differenze di contesto
  - diversità di legislazione (per esempio le scuole speciali)
  - diversa gestione dell'educazione bilingue
- differenze nelle caratteristiche degli insegnanti

Minacce alla validità esterna degli studi di regressione multipla (continua)

Il caso specifico dell'effetto della dimensione della classe sull'apprendimento degli alunni sembra aver dato risultati sostanziali solo per Israele. ← classi da 40 bambini

Per altri paesi l'analisi ha fatto registrare guadagni molto contenuti. Es: in Italia i benefici sono assai modesti, la dimensione della classe non è un problema rilevante, piuttosto la falsificazione dei test invalida la possibilità di verificare la qualità dell'apprendimento e dell'offerta educativa.

Angrist Battistin e Vuri (2017)

In a Small Moment: Class Size and Mora Hazard in the Italian Mezzogiorno. AEJ:AE 9(4)

Minacce alla validità interna dell'analisi di regressione multipla (Paragrafo 9.2)

Validità interna: le inferenze statistiche sugli effetti causali sono valide per la popolazione studiata.

Cinque minacce alla validità interna degli studi di regressione:

- Distorsione da variabili omesse
- Forma funzionale incorretta
- Distorsione da errori nelle variabili
- Distorsione da selezione campionaria
- Distorsione da causalità simultanea

Tutte queste implicano che $E(u_i|X_{1i},\ldots,X_{ki}) \neq 0$ (o che non vale l'indipendenza in media condizionata) – nel qual caso lo stimatore OLS è distorto e inconsistente.

1. Distorsione da variabili omesse

La distorsione da variabili omesse nasce quando una variabile omesse è **sia:**

I. una determinante di *Y e*
II. correlata con almeno un regressore incluso.

- È stata esaminata in precedenza la distorsione da variabili omesse con una singola *X*. La distorsione nasce nelle regressioni multiple se la variabile omessa soddisfa le condizioni (i) e (ii) date in precedenza.
- Se la regressione multipla comprende variabili di controllo, occorre chiedersi se vi siano dei fattori omessi per i quali non esista un adeguato controllo, cioè se il termine di errore sia correlato con la variabile di interesse anche dopo che siano state inserite le variabili di controllo.

Soluzioni alla distorsione da variabili omesse

1. Se è possibile misurare la variabile causale omessa, inserirla come regressore aggiuntivo nella regressione multipla;
2. Se si possiedono dati su uno o più controlli e questi sono adeguati (nel senso del mantenimento della plausibilità dell'indipendenza in media condizionata), inserire le variabili di controllo;
3. Se possibile, usare *dati panel* nei quali ciascuna unità (individuo) venga osservata più di una volta; ← se variabil omesse time invariant
4. Se non è possibile misurare la variabile omessa, usare la *regressione con variabili strumentali*;
5. Condurre un esperimento controllato casualizzato.
   - *Perché funziona?* Si ricordi: se X viene assegnata casualmente, allora X sà necessariamente distribuita indipendentemente da u; perciò $E(u|X = x) = 0$.

2. Incorretta forma funzionale (incorretta specificazione della forma funzionale)

Nasce se la forma funzionale è incorretta – per esempio, un termine di interazione viene omesso in maniera incorretta; allora le inferenze sugli effetti causali saranno distorte.

Soluzioni alla incorretta specificazione della forma funzionale

1. Variabile dipendente continua: usare in X le specifiche non lineari "appropriate" (logaritmi, interazioni, ecc.)
2. Variabile dipendente discreta (*per esempio*: binaria): occorre un'estensione dei metodi di regressione multipla (l'analisi "probit" o "logit" per le variabili dipendenti binarie).

3. Distorsione da errori nelle variabili

Finora il presupposto è stato che *X* fosse misurata senza errori.

Nella realtà, i dati economici spesso presentano errori di misura

- Errori nell'inserimento dei dati amministrativi
- Errori di memoria nei sondaggi (quando ha iniziato a svolgere il suo lavoro attuale?)
- Ambiguità nelle domande (qual è stato il suo reddito dello scorso anno?)
- Problemi da risposte intenzionalmente errate ai sondaggi (Qual è la sua situazione finanziaria attuale? Quante volte si mette alla guida dopo avere bevuto?)

Distorsione da errori nelle variabili (continua)

In generale, un errore di misura in un regressore risulta in una **distorsione "da errori nelle variabili"**.

Qualche calcolo mostra come gli errori nelle variabili tipicamente portano alla correlazione tra la variabile misurate e l'errore di regressione. Si consideri il modello a regressore singolo:

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

e si supponga che $E(u_i|X_i) = 0$). Si ponga

X_i = valore reale non misurato di X

$\tilde{X}_i$ = versione misurata erroneamente di X (i dati osservati)

Allora

$Y_i = \beta_0 + \beta_1 X_i + u_i$ *sommando e sottraendo* $\beta_1 \tilde{X}_i$

$= \beta_0 + \beta_1 \tilde{X}_i + [\beta_1(X_i - \tilde{X}_i) + u_i]$

Per cui la regressione svolta è,

$Y_i = \beta_0 + \beta_1 \tilde{X}_i + v_i$, dove $v_i = \beta_1(X_i - \tilde{X}_i) + u_i$

Con l'errore di misura, tipicamente $\tilde{X}_i$ è correlata con v_i quindi $\hat{\beta}_1$ risulta distorta:

$\text{cov}(\tilde{X}_i, v_i) = \text{cov}(\tilde{X}_i, \beta_1(X_i - \tilde{X}_i) + u_i)$

$= \beta_1 \text{cov}(\tilde{X}_i, X_i - \tilde{X}_i) + \text{cov}(\tilde{X}_i, u_i)$

Spesso è plausibile che $\text{cov}(\tilde{X}_i, u_i) = 0$ (se $E(u_i|X_i) = 0$ allora $\text{cov}(\tilde{X}_i, u_i) = 0$ se l'errore di misurazione in $\tilde{X}_i$ è incorrelato con u_i). Ma tipicamente $\text{cov}(\tilde{X}_i, X_i - \tilde{X}_i) \neq 0$….

Distorsione da errori nelle variabili (continua)

$Y_i = \beta_0 + \beta_1 \tilde{X}_i + v_i$, dove $v_i = \beta_1(X_i - \tilde{X}_i) + u_i$

$$\begin{aligned}\text{cov}(\tilde{X}_i, v_i) &= \text{cov}(\tilde{X}_i, \beta_1(X_i - \tilde{X}_i) + u_i) \\ &= \beta_1 \text{cov}(\tilde{X}_i, X_i - \tilde{X}_i) + \text{cov}(,u_i) \\ &= \beta_1 \text{cov}(\tilde{X}_i, X_i - \tilde{X}_i) \quad \text{se cov}(\tilde{X}_i, u_i) = 0\end{aligned}$$

Per ottenere qualche intuizione per il problema si considerino due casi speciali:

A. Errore di misura classico
B. Errore di misura "migliore ipotesi"

A. Errore di misura classico

Il modello di errore di misura classico presume che

$$\tilde{X}_i = X_i + w_i,$$

dove w_i è rumore casuale a media zero con corr(X_i, w_i) = 0 e corr(u_i, w_i) = 0.

Con il modello di errore di misura classico, $\hat{\beta}_1$ è distorto verso zero. Questa è l'idea: si supponga di prendere la variabile vera e quindi aggiungere una grande quantità di rumore casuale – numeri casuali generati dal computer. Entro il limite del "solo rumore", $\tilde{X}_i$ sarà incorrelata a Y_i (e a qualsiasi altra cosa), quindi il coefficiente di regressione avrà valore atteso zero. Se $\tilde{X}_i$ contiene del rumore ma non è "solo rumore" allora la relazione tra $\tilde{X}_i$ e Y_i sarà attenuata, per cui $\hat{\beta}_1$ è distorto verso zero.

Errore di misura classico: i calcoli

$\tilde{X}_i = X_i + w_i$, dove corr(X_i, w_i) = 0 e corr(u_i, w_i) = 0.

L'errore di di misura è un errore casuale un WN

Quindi var($\tilde{X}_i$) = $\sigma^2_X + \sigma^2_w$

cov($\tilde{X}_i$, $X_i - \tilde{X}_i$) = cov(X_i + w_i, $-w_i$) = $-\ \sigma^2_w$ (*)

così

cov($\tilde{X}_i$, v_i) = cov($\tilde{X}_i$, $-\beta_1\ w_i + u_i$)

Ove $v_i = -\beta_1\ w_i + u$ per definizione vedi slide precedente

$= -\beta_1$ cov($\tilde{X}_i$, w_i) + cov($\tilde{X}_i$, u_i)

$= -\beta_1\ \sigma^2_w$ + cov(X+w, u)

Errore di misura classico: i calcoli (continua)

$$\text{cov}(\tilde{X}_i, v_i) = -\beta_1 \sigma_w^2 + \text{cov}(X+w, u)$$

Viene da (*)

Cov(xu)=0, cov(w,u)=0

Quindi $\text{cov}(\tilde{X}_i, v_i) = -\beta_1 \sigma_w^2$

Così dalla (6.1) abbiamo il plim

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1 \frac{\sigma_w^2}{\sigma_{\tilde{X}}^2} = \left(1 - \frac{\sigma_w^2}{\sigma_{\tilde{X}}^2}\right)\beta_1$$

$$= \left(\frac{\sigma_{\tilde{X}}^2 - \sigma_w^2}{\sigma_{\tilde{X}}^2}\right)\beta_1 = \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}\right)\beta_1$$

abbiamo sostituito $\text{Var}(\tilde{X}_i) = \sigma_X^2 + \sigma_v^2$

Errore di misura classico: i calcoli (riepilogo)

Quindi $\text{var}(\tilde{X}_i) = \sigma_X^2 + \sigma_v^2$

$\text{cov}(\tilde{X}_i, X_i - \tilde{X}_i) = \text{cov}(X_i + v_i, -v_i) = -\sigma_w^2$

così

$\text{cov}(\tilde{X}_i, \tilde{u}_i) = -\beta_1 \sigma_w^2$

così $\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1 \dfrac{\sigma_w^2}{\sigma_{\tilde{X}}^2} = \left(1 - \dfrac{\sigma_w^2}{\sigma_{\tilde{X}}^2}\right)\beta_1$

$= \left(\dfrac{\sigma_{\tilde{X}}^2 - \sigma_w^2}{\sigma_{\tilde{X}}^2}\right)\beta_1 = \left(\dfrac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}\right)\beta_1$

Errore di misura classico: i calcoli (continua)

$$\hat{\beta}_1 \xrightarrow{p} \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}\right)\beta_1 \qquad (9.2)$$

Il denominatore è sempre maggiore del numeratore, quindi $\hat{\beta}_1$ è distorto verso zero.

Il modello di errore di misura classico è speciale perché assume che $corr(X_i, w_i) = 0$.

Se l'errore di misura ha l'effetto di aggiungere un elemento casuale al valore effettivo di X, allora β_1 è inconsistente

Errore di misura classico: i calcoli (continua)

$$\hat{\beta}_1 \xrightarrow{p} \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}\right)\beta_1 \qquad (9.2)$$

Nel caso estremo in cui l'errore di misura è talmente grande che non rimane informazione in X_i, il rapporto tra le varianze in (9.2) è zero e $\hat{\beta}_1$ converge in probabilità a 0.

All'estremo opposto quando non c'è alcun errore di misura $\sigma_w^2 = 0$ e $\hat{\beta}_1 \xrightarrow{p} \beta_1$

B. Errore di misura "migliore ipotesi"

Si supponga che l'intervistato non ricordi X_i, ma faccia un'ipotesi del tipo $\tilde{X}_i = E(X_i|W_i)$, dove $E(u_i|W_i) = 0$.

ove W è il set informativo del rispondente,

Poiché $\tilde{X}_i$ è la migliore ipotesi, l'errore di misura $X_i - \tilde{X}_i$ non è correlato con con la riposta $\tilde{X}_i$

Se l'errore di misura fosse correlato con $\tilde{X}_i$ esisterebbe un'informazione utile per predire X, nel qual caso $\tilde{X}_i$ non sarebbe la migliore ipotesi di X.

Ovvero $E[(\tilde{X}_i - X), \tilde{X}_i] = 0$

B. Errore di misura "migliore ipotesi" (calcoli)

Formalmente possiamo scrivere che:

Dalla slide (13) sappiamo che $\beta_1(X_i - \tilde{X}_i) + u_i = v_i$

$$\text{cov}(\tilde{X}_i, v_i) = \text{cov}(\tilde{X}_i, \beta_1(X_i - \tilde{X}_i) + u_i)$$

$$= \beta_1 \text{cov}(\tilde{X}_i, X_i - \tilde{X}_i) + \text{cov}(\tilde{X}_i, u_i)$$

- $\text{cov}(\tilde{X}_i, X_i - \tilde{X}_i) = 0$ perché $\tilde{X}_i = E(X_i|W_i)$ (siccome $\tilde{X}_i$ è la migliore ipotesi, l'errore $X_i - \tilde{X}_i$ è incorrelato con $\tilde{X}_i$).
- $\text{Cov}(\tilde{X}_i, u_i) = 0$ perché $E(u_i|W_i) = 0$ ($\tilde{X}_i$ è funzione di W_i e $E(u_i|W_i) = 0$).
- Così $\text{cov}(\tilde{X}_i, v_i) = 0$, quindi $\hat{\beta}_1$ è non distorto.

Errore di misura "migliore ipotesi" (continua)

- Con il modello "migliore ipotesi", l'errore di misura è ancora presente – non si osserva il vero valore di X_i – ma questo errore di misura non introduce distorsione in $\hat{\beta}_1$!

- Il modello "migliore ipotesi" è estremo – non è sufficiente fare una buona ipotesi, è necessaria la "migliore" ipotesi $\tilde{X}_i = E(X_i|W_i)$, cioè il valore atteso condizionato di X data W, dove $E(u_i|W_i) = 0$.

Insegnamenti dai modelli classico e «migliore ipotesi»:

- Il livello di distorsione in $\hat{\beta}_1$ dipende dalla natura dell'errore di misura – questi due modelli sono casi speciali.
- Se a X_i viene aggiunto rumore puro, allora $\hat{\beta}_1$ è distorto verso zero.
- Il modello "migliore ipotesi" è estremo. In generale, se si pensa che vi sia un errore di misura, ci si dovrebbe preoccupare della distorsione da errore di misura.
- L'importanza potenziale della distorsione da errore di misura dipende dal modo in cui i dati vengono raccolti.
  - Spesso alcuni dati amministrativi (per esempio il numero di insegnanti in un distretto scolastico) sono molto accurati.
  - Spesso i sondaggi su argomenti sensibili (quanto guadagna?) presentano notevoli errori di misura.

Soluzioni alla distorsione da errori nelle variabili

1. Ottenere dati migliori (spesso più facile a dirsi che a farsi).
2. Sviluppare un modello specifico del processo degli errori di misura. Questo è possibile solo se si sa molto sulla natura dell'errore di misura – per esempio, un sottocampione dei dati viene sottoposto a controlli incrociati usando dati amministrativi e le discrepanze vengono analizzate e modellizzate. (Altamente specialistico; non ce ne occuperemo qui)
3. Regressione con variabili strumentali.

Esempio di errori di misura

In Angrist Battistin e Vuri (2017)

In a Small Moment: Class Size and Mora Hazard in the Italian Mezzogiorno. AEJ:AE 9(4)

Gli elaborati degli alunni devono essere trascritti dagli insegnanti, ma nelle classi grandi le insegnanti vengono coadiuvate nela correzione dei test ➔ punteggi affidabili

Nelle classi piccole non c'è bisogno di aiuto ➔ cheating

Usano variabili strumentali Z= regola di splitting della classe: 28 studenti.

4. Distorsione da dati mancanti e selezione campionaria

Spesso i dati mancano. A volte i dati mancanti introducono distorsione, ma a volte no. È utile considerare tre casi:

1.I dati sono mancanti a caso.

2.I dati sono mancanti in base al valore di una o più *X*

3.I dati sono mancanti in parte in base al valore di *Y* o *u*

I casi 1 e 2 non introducono distorsione: gli errori standard sono più grandi di come sarebbero se i dati non fossero mancanti, ma $\hat{\beta}_1$ è non distorto

Il caso 3 introduce la distorsione da "selezione campionaria".

Dati mancanti: Caso 1

1. I dati sono mancanti a caso

Si supponga di avere effettuato una semplice campionatura casuale di 100 lavoratori e avere registrato le risposte di ognuno su un foglio di carta – si supponga poi che il proprio cane abbia mangiato 20 dei fogli con le risposte (scelti a caso) prima che i dati potessero essere inseriti nel computer. Questo equivale ad avere effettuato una semplice campionatura casuale di 80 lavoratori (basta rifletterci), per cui il cane non ha introdotto alcuna distorsione.

Dati mancanti: Caso 2

2. I dati sono mancanti in base a un valore di una delle *X*

   Nell'applicazione su punteggi nei test/dimensioni delle classi, si supponga di restringere la propria analisi ai soli distretti scolastici con *STR* < 20. Considerando solo distretti con classi di piccole dimensioni non si sarà in grado di dire nulla sui distretti con classi di grandi dimensioni, la concentrazione sui distretti con classi di piccole dimensioni non introduce distorsione. Questo equivale ad avere dati mancanti, dove i dati mancano se *STR* > 20. Più in generale, se i dati sono mancanti in base unicamente a valori delle *X*, la loro mancanza non distorce lo stimatore OLS.

Dati mancanti: Caso 3

3. I dati sono mancanti in parte in base al valore di *Y* o *u*

In genere questo tipo di dati mancanti *introduce effettivamente* distorsione nello stimatore OLS. Questo tipo di distorsione è detta anche distorsione da selezione campionaria.

La distorsione da selezione campionaria nasce quando un processo di selezione:

(i) influenza la disponibilità dei dati e

(ii) è legato alla variabile dipendente.

Esempio #1: Statura degli studenti

Il prof di statistica chiede di stimare l'altezza media degli studenti maschi. Si raccolgono i dati (si ottiene il campione) stando in piedi fuori dallo spogliatoio della squadra di basket e registrando la statura degli studenti che vi entrano.

- Si tratta di un buon progetto – fornirà una stima non distorta della statura degli studenti?
- Formalmente, gli individui sono stati campionati in un modo legato alla Y (statura) risultante, il che si traduce in distorsione.

Esempio #2: I fondi comuni

- I fondi comuni gestiti attivamente superano quelli che seguono l'andamento del mercato?
- Strategia empirica:
  - Schema di campionatura: semplice campione causale dei fondi comuni disponibili al pubblico a una determinata data.
  - Dati: rendimenti dei 10 anni precedenti.
  - Stimatore: rendimento medio a dieci anni dei fondi comuni campione, meno rendimento a dieci anni dell'indice S&P500
  - Vi è distorsione da selezione campionaria?
    (o in modo equivalente, vi sono dati mancanti in base in parte al valore di Y o u?)
  - In che modo questo esempio è simile a quello dei giocatori di basket?

Esempio #2: I fondi comuni … continua

I fondi gestiti attivamente zono quei fondi il cui compito è quello di avere un rendimento maggiore del mercato.

I «dogs» - fondi con rendimento minore – vengono omessi dai dati perché usciti dal mercato o assorbiti da altri fondi.

Uno studio che usa dati sui fondi acquistabili oggi è soggetto a distorsione da variabili omesse: il campione è selezionato in base al valore della variabile dipendente, i rendimenti e quindi i fondi con rendimento basso vengono eliminati

La distorsione da selezione campionaria induce correlazione tra un regressore e l'errore.

Esempio dei fondi comuni:

$$rendimento_i = \beta_0 + \beta_1 fondo\_gestito_i + u_i$$

- Essere un fondo gestito nel campione ($fondo\_gestito_i = 1$) significa che il proprio rendimento è stato migliore di quello dei fondi gestiti estinti, che non si trovano nel campione – quindi corr($fondo\_gestito_i, u_i$) ≠ 0.
- I fondi comuni che sopravvivono sono i "giocatori di basket" dei fondi comuni.

Esempio #3: rendimento dello studio

- Quanto rende un anno aggiuntivo di studio?
- Strategia empirica:
  - Schema di campionatura: semplice campione casuale dei laureati con un impiego (se hanno un impiego, è possibile avere i dati sulle retribuzioni)
  - Dati: guadagni e anni di studio
  - Stimatore: regressione di ln(*guadagni*) su anni di studio
  - Ignorare i problemi da distorsione da variabili mancanti e da errori di misura – vi è distorsione da selezione campionaria?
  - Che rapporto c'è con l'esempio dei giocatori di basket?

Soluzioni alla distorsione da selezione campionaria

- Raccogliere il campione in un modo che eviti la selezione campionaria.
  - *Esempio dei giocatori di basket*: ottenere un vero campione casuale degli studenti, per esempio scegliendo gli studenti a caso dagli elenchi amministrativi degli iscritti.
  - *Esempio dei fondi comuni*: cambiare la popolazione del campione dai fondi disponibili alla *fine* del periodo di dieci anni, a quelli disponibili all'*inizio* del periodo (inclusi i fondi estinti)
  - *Esempio del rendimento dello studio*: campionare i laureati, non i lavoratori (comprendere i disoccupati)
- Esperimento casualizzato controllato.
- Costruire un modello del problema della selezione campionaria e farne una stima (non lo faremo in questa sede).

5. Distorsione da causalità simultanea

Finora si è ipotizzato che *X* causasse *Y*.

E se anche *Y* causa *X*?

Esempio: effetto delle dimensioni delle classi

- Un basso *STR* porta a migliori punteggi nei test
- Ma si supponga cha ai distretti con bassi risultati nei test vengano fornite risorse ulteriori.
  Con queste risorse aggiuntive vengono assunti nuovi insegnanti e STR si riduce come risultato di un processo politico anch'essi avranno un basso STR

5. Distorsione da causalità simultanea (continua)

- Che significato ha tutto ciò per una regressione di *TestScore* su *STR*?

Vediamone gli effetti in dettaglio.

Distorsione da causalità simultanea: in equazioni

(a) Effetto causale su Y di X: $Y_i = \beta_0 + \beta_1 X_i + u_i$

(b) Effetto causale su X di Y: $X_i = \gamma_0 + \gamma_1 Y_i + v_i$

- Un grande u_i significa un grande Y_i, *il che implica un* grande X_i (se $\gamma_1 > 0$)
- Quindi corr(X_i,u_i) ≠ 0
- Quindi $\hat{\beta}_1$ è distorto e inconsistente.
- *Esempio*: un distretto con risultati particolarmente negativi dato *STR* (u_i negativo) riceve risorse aggiuntive, che abbassano il suo *STR*; quindi STR_i e u_i sono correlati

Soluzioni alla distorsione da causalità simultanea

1. Eseguire un esperimento casualizzato controllato. Siccome X_i viene scelto a caso dallo sperimentatore, non vi è feedback dalla variabile risultante su Y_i (ipotizzando una perfetta corrispondenza).
2. Sviluppare e stimare un modello completo di entrambe le direzioni di causalità. È l'idea alla base di molti macromodelli (per esempio la Federal Reserve Bank-USA). *Questo nella pratica è estremamente difficile*.
3. Usare regressione a variabili strumentali per stimare l'effetto causale di interesse (effetto di X su Y, ignorando l'effetto di Y su X).

Piccola digressione su variabili strumentali

Lo stimatore a variabili strumentali ripristina la relazione causale tra X_1 e Y tramite una variabile esterna, detto strumento, Z, che ha due caratteristiche:

1) È correlate con X_1
2) Non ha un effetto causale su Y. La correlazione tra Y e Z passa solo tramite X_1

Es: stima dell'elasticità dell'offerta di latte al Prezzo. Strumento: la pioggia.

Piccola digressione su variabili strumentali (continua)

Es: stima dell'elasticità dell'offerta di latte al Prezzo. Strumento: la pioggia.

Regredendo X su Z e prendendo $\hat{X}$ è possibile isolare in X la componente che deriva da fluttuazioni casuali.

➔ Primo stadio

$\hat{X}$ verrà successivamente utilizzato nella regression al posto di X

➔ Secondo stadio

➔ Forma ridotta / primo stadio

Validità interna ed esterna quando la regressione è usata per le previsioni (Paragrafo 9.3)

- Previsione e stima degli effetti causali sono obbiettivi piuttosto diversi.
- Per la previsione,
  - $\bar{R}^2$ è importante (molto!)
  - La distorsione da variabili omesse non è un problema!
  - L'interpretazione dei coefficienti nei modelli di previsione non è importante – ciò che conta sono un buon adattamento e un modello che si possa ritenere "affidabile" per la propria applicazione
  - La validità esterna è fondamentale: il modello stimato con dati storici deve mantenersi valido per il futuro (immediato)
  - La previsione verrà trattata in seguito con i dati da serie storiche

quando la regressione è usata per le previsioni (continua)

- Finora abbiamo considerato la stima del *testscore* su STR per un policy maker che voglia ottimizzare le risorse pubbliche riducendo le classi. In questo caso serve una stima causale
- Si consideri un padre che cambia residenza con la famiglia e vuole scegliere il quartiere in base alla qualità delle scuole. Se nella regressione

$$Teststcore = 698.9 - 2.28\ STR$$

Vi è un errore di variabili omesse 2.28 non ha lettura causale, ma al genitore interessato alla previsione del punteggio

quando la regressione è usata per le previsioni (continua)

$$\textit{Teststcore} = 698.9 - 2.28\ \text{STR}$$

Vi è un errore di variabili omesse 2.28 non ha lettura causale.

Il genitore osserva STR nei diversi quartieri è interessato alla previsione del punteggio.

Ciò che rileva è che STR sia un predittore affidabile che spieghi il grosso della variazione dei punteggi tra distretti e che questa relazione sia stabile.

quando la regressione è usata per le previsioni (continua)

Sintesi: i modelli di regressione possono produrre previsioni affidabili anche se i loro coefficienti non hanno un'interpretazione causale

Applicare la validità interna ed esterna: punteggio nei test e dimensioni delle classi (Paragrafo 9.4)

- Obiettivo: valutare le minacce alla validità interna ed esterna dell'analisi empirica dei dati sui punteggi nei test della California.

- Validità esterna
  - Confrontare i risultati della California e del Massachusetts
  - Riflettere a lungo…
- Validità interna
  - Esaminare l'elenco delle cinque potenziali minacce per la validità interna e riflettere a lungo…

Verifica della validità esterna

Lo studio sulla California verrà confrontato a uno che usa i dati del Massachusetts

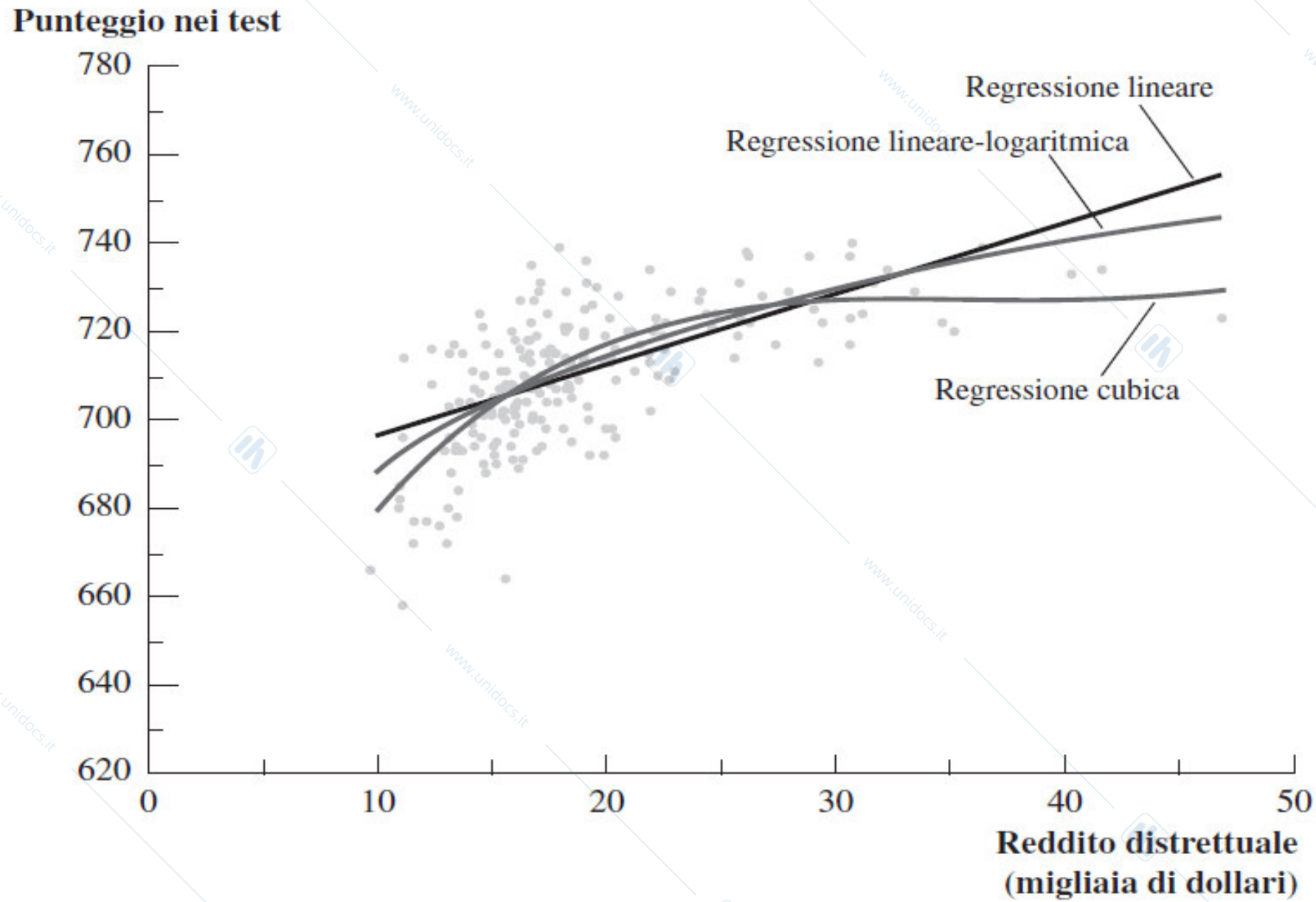
Il gruppo di dati del Massachusetts

- 220 distretti scolastici elementari
- Test: test MCAS del 1998 MCAS – totale quarto grado (Matematica + Inglese + Scienze)
- Variabili: *STR*, *TestScore*, *PctEL*, *LunchPct*, *Income*

I dati del Massachusetts: riepilogo delle statistiche

Tabella 9.1 Statistiche descrittive dei dati sul punteggio nei test in California e nel Massachusetts.

	California		Massachusetts	
	Media	Deviazione standard	Media	Deviazione standard
Punteggio nei test	654,1	19,1	709,8	15,1
Rapporto studenti/insegnanti	19,6	1,9	17,3	2,3
% studenti non di madrelingua	15,8%	18,3%	1,1%	2,9%
% aventi diritto al sussidio mensa	44,7%	27,1%	15,3%	15,1%
Reddito distrettuale medio ($)	15,317$	7226$	18,747$	5808$
Numero di osservazioni	420		220	
Anno	1999		1998	



Punteggi rispetto a reddito e rette di regressione:
dati del Massachusetts

Tabella 9.2 Stime di regressioni multiple del rapporto studenti/insegnanti e del punteggio nei test: dati del Massachusetts.

Variabile dipendente: media combinata dei punteggi ottenuti nei test di inglese, matematica e scienze nel distretto scolastico; 220 osservazioni

Regressore	(1)	(2)	(3)	(4)	(5)	(6)
STR	– 1,72** (0,50)	– 0,69* (0,27)	– 0,64* (0,27)	12,4 (14,0)	– 1,02** (0,37)	– 0,67* (0,27)
STR^2				– 0,680 (0,737)		
STR^3				0,011 (0,013)		
% studenti non madrelingua		– 0,411 (0,306)	– 0,437 (0,303)	– 0,434 (0,300)		
% studenti non madrelingua > mediana? (variabile binaria, *HiEL*)					– 12,6 (9,8)	
HiEL × *STR*					0,80 (0,56)	
% aventi diritto alla mensa gratuita		– 0,521** (0,077)	– 0,582** (0,097)	– 0,587** (0,104)	– 0,709** (0,091)	– 0,653** (0,72)
Reddito distrettuale (logaritmo)		16,53** (3,15)				
Reddito distrettuale			– 3,07 (2,35)	– 3,38 (2,49)	– 3,87* (2,49)	– 3,22 (2,31)
(Reddito distrettuale)^2			0,164 (0,085)	0,174 (0,089)	0,184* (0,090)	0,165 (0,085)
(Reddito distrettuale)^3			– 0,0022* (0,0010)	– 0,0023* (0,0010)	– 0,0023* (0,0010)	– 0,0022* (0,0010)
Intercetta	739,6** (8,6)	682,4** (11,5)	744,0** (21,3)	665,5** (81,3)	759,9** (23,2)	747,4** (20,2)

Statistiche F e valori-p per l'esclusione di gruppi di variabili						
Tutte le variabili *STR* e i termini d'interazione = 0				2,86 (0,038)	4,01 (0,020)	
STR^2 e $STR^3 = 0$				0,45 (0,641)		
$Income^2$, $Income^3$			7,74 (< 0,001)	7,75 (< 0,001)	5,85 (0,003)	6,55 (0,002)
HiEL, *HiEL* × *STR*					1,58 (0,208)	
SER	14,64	8,69	8,61	8,63	8,62	8,64
$\bar{R}^2$	0,063	0,670	0,676	0,675	0,675	0,674

Queste regressioni sono state stimate utilizzando i dati sui distretti scolastici elementari del Massachusetts, descritti nell'Appendice 9.1. Gli errori standard sono riportati tra parentesi sotto i coefficienti e i valori-*p* sono riportati tra parentesi sotto le statistiche *F*. I coefficienti sono statisticamente significativi al livello *5% o **1%.

Che somiglianza esiste tra i risultati di Massachusetts e California ?

- Funzione logaritmica rispetto a funzione cubica per *STR*?
- Evidenza di non linearità nella relazione *TestScore*-*STR*?
- Esiste una significativa interazione *HiEL* × *STR*?

Effetti previsti di una riduzione delle dimensioni delle classi di specificazione lineare 2 per il Massachusetts:

$$\widehat{TestScore} = \underset{(21,3)}{744,0} - \underset{(0,27)}{0,64}STR - \underset{(0,303)}{0,437}PctEL - \underset{(0,097)}{0,582}LunchPct$$

$$- \underset{(2,35)}{3,07}Income + \underset{(0,085}{0,164}Income^2 - \underset{(0,0010)}{0,0022}Income^3$$

- Effetto stimato = -0,64 × (-2) = 1,28
- Errore standard = 2 × 0,27 = 0,54

  ***NOTA*:** var(aY) = a^2var(Y); $SE(a\hat{\beta}_1) = |a|SE(\hat{\beta}_1)$
- 95% CI = 1,28 ± 1,96 × 0,54 = (0,22, 2,34)

Calcolo degli effetti previsti nei modelli non lineari

Si utilizzi il metodo "prima" e "dopo" :

$$\widehat{TestScore} = 655{,}5 + 12{,}4STR - 0{,}680STR^2 + 0{,}0115STR^3 - 0{,}434PctEL - 0{,}587LunchPct - 3{,}48Income + 0{,}174Income^2 - 0{,}0023Income^3$$

Riduzione stimata da 20 studenti a 18 :

$$\Delta\widehat{TestScore} = [12{,}4 \times 20 - 0{,}680 \times 20^2 + 0{,}0115 \times 20^3] - [12{,}4 \times 18 - 0{,}680 \times 18^2 + 0{,}0115 \times 18^3] = 1{,}98$$

- Si confronti con la stima data dal modello lineare di 1,28
- *Errori standard* di questo effetto stimato: si utilizzi il metodo "riordinamento della regressione" ("trasformazione dei regressori")

Riepilogo dei risultati per il Massachusetts

- Il coefficiente di *STR* si riduce da –1,72 a –0,69 quando vengono inserite le variabili di controllo per le caratteristiche di studenti e distretti – segno che la stima originaria presentava distorsione da variabili omesse.
- L'effetto delle dimensioni delle classi è statisticamente significativo al livello 1%, dopo il controllo delle caratteristiche di studenti e distretti
- Nessuna evidenza statistica di non linearità nella relazione *TestScore -STR*
- Nessuna evidenza statistica di interazione tra *STR* e *PctEL*

Confronto degli effetti stimati delle dimensioni delle classi tra California e Massachusetts

Tabella 9.3 Rapporto studenti/insegnanti e punteggio nei test: confronto tra le stime per la California e il Massachusetts.

			Stima degli effetti della riduzione di due studenti per insegnante, in unità di:	
	Stima OLS	**Deviazione standard del punteggio nei test nei distretti**	**Punti del test**	**Deviazione standard**
California				
Lineare: Tabella 8.3(2)	– 0,73 (0,26)	19,1	1,46 (0,52)	0,076 (0,027)
Cubica: Tabella 8.3(7) *STR* ridotto da 20 a 18	–	19,1	2,93 (0,70)	0,153 (0,037)
Cubica: Tabella 8.3(7) *STR* ridotto da 22 a 20	–	19,1	1,90 (0,69)	0,099 (0,036)
Massachusetts				
Lineare: Tabella 9.2(3)	– 0,64 (0,27)	15,1		0,085 (0,036)

Gli errori standard sono riportati tra parentesi.

Riepilogo: confronto tra le analisi di regressione di California e Massachusetts

- L'effetto delle dimensioni delle classi scende in entrambi i casi quando vengono aggiunte variabili di controllo per studenti e distretti.
- L'effetto delle dimensioni delle classi è statisticamente significativo in entrambi i casi.
- L'effetto stimato della riduzione di 2 studenti in *STR* è quantitativamente simile per California e Massachusetts.
- Nessuno dei gruppi di dati evidenzia interazione *STR* – *PctEL*.
- Esiste qualche evidenza di non linearità di *STR* nei dati della California ma non del Massachusetts.

Un passo indietro: che minacce per la validità interna rimangono nell'esempio punteggio nei test/dimensioni delle classi?

1. Distorsione da variabili omesse?

Quali fattori causali mancano?

- Caratteristiche degli studenti come le capacità innate
- Accesso a opportunità di apprendimento esterne
- Altre misure della qualità del distretto, come la qualità degli insegnanti

Le regressioni cercano di controllare questi fattori mancanti con variabili ci controllo che non sono necessariamente causali ma sono correlate con le variabili causali mancanti:

- Dati demografici dei distretti (reddito, % di diritto a sussidio mensa)
- Frazione di studenti non di madrelingua

Distorsione da variabili omesse (continua)

Le variabili di controllo sono efficaci? Cioè, dopo avere inserito le variabili di controllo l'errore è non correlato con *STR*?

- La risposta a queste domande richiede un ragionamento.
- Vi è qualche evidenza che le variabili di controllo stiano facendo il loro lavoro:
  - Il coefficiente di *STR* non cambia molto al cambiare della specificazione delle variabili
  - I risultati per California e Massachusetts sono simili – perciò se rimane della distorsione da variabili omesse, questa dovrebbe essere simile nei due gruppi di dati
- *Quali ulteriori variabili di controllo si potrebbero volere utilizzare – e cosa dovrebbero controllare?*

2. Forma funzionale incorretta?

- Si sono provate diverse forme funzionali, sia con i dati della California che del Massachusetts
- Gli effetti non lineari sono modesti
- Verosimilmente, non è una minaccia importante al momento.

3. Distorsione da errori nelle variabili?

- I dati sono amministrativi, per cui è probabilmente possibile escludere errori di registrazione o inserimento importanti.
- *STR* è una misura a livello di distretto, per cui gli studenti sottoposti ai test potrebbero non avere subito l'STR misurato per il distretto – un tipo di errore di misura complicato
- Idealmente si dovrebbero avere i dati sui singoli studenti per livello di grado.

4. Distorsione da selezione campionaria?

- Il campione è costituito da tutti i distretti scolastici elementari pubblici (in California e Massachusetts) – non ci sono dati mancanti
- Nessun motivo per pensare a un problema di selezione.

5. Distorsione da causalità simultanea?

- L'equiparazione del finanziamento in base ai punteggi nei test potrebbe provocare causalità simultanea.
- Questo non avveniva in California o Massachusetts durante i campionamenti, per cui la distorsione da causalità simultanea non appare verosimilmente importante.

Esempio ulteriore per una discussione in classe

Il fatto di apparire nello spettacolo televisivo *America's Most Wanted* aumenta le possibilità di essere catturati dalla polizia?

riferimento: Thomas Miles (2005), "Estimating the Effect of *America's Most Wanted*: A Duration Analysis of Wanted Fugitives," *Journal of Law and Economics*, 281-306.

- Unità di osservazione: criminali in fuga
- Schema di campionatura: 1200 fuggitivi maschi individuati sui siti web di FBI, NYCPD, LAPD, PhilaPD, USPS, e dei Federal Marshalls (*tutti i dati sono stati scaricati dal Web!*)
- Variabile dipendente: durata latitanza (anni prima della cattura)
- Regressori:
  - Apparizione su *America's Most Wanted* (175 dei 1200) (al tempo trasmesso su Fox, il sabato alle 21)
  - Tipo di crimine, caratteristiche personali

America's Most Wanted:

Minacce alla validità interna ed esterna

Validità esterna: in quale modo si vorrebbero approfondire i dati – durata maggiore della trasmissione? Realizzazione di una seconda trasmissione dello stesso tipo? E' richiesta precisione…

Validità interna: che importanza hanno queste minacce?

1. Distorsione da variabili omesse
2. Forma funzionale incorretta
3. Distorsione da errori nelle variabili
4. Distorsione da selezione campionaria
5. Distorsione da causalità simultanea

Altro?

Econometria I

Corso di laurea magistrale in Finanza e Assicurazioni

Slides per lezioni

secondo semestre 2019

Marco Ventura

Cap. 12 SW

Sommario

1. Regressione IV: cosa e perché; minimi quadrati in due stadi
2. Il modello generale di regressione IV
3. Verifica della validità degli strumenti
   a) Strumenti deboli e forti
   b) Esogeneità degli strumenti
4. Applicazione: domanda di sigarette
5. Esempi: dove trovare gli strumenti?

Regressione IV: perché?

Tre importanti minacce alla validità interna sono:

- Distorsione da variabili omesse per una variabile correlata con X ma inosservata (perciò non può essere inclusa nella regressione) e per cui vi sono variabili di controllo inadeguate;
- Distorsione da causalità simultanea (X causa Y, Y causa X);
- Distorsione da errori nelle variabili (X è misurata con errore)

Tutti e tre i problemi comportano $E(u|X) \neq 0$.

- La regressione con variabili strumentali può eliminare la distorsione quando $E(u|X) \neq 0$ – usando una *variabile strumentale* (IV), Z.

Lo stimatore IV con un singolo regressore e un singolo strumento (Paragrafo 12.1)

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

- La regressione IV divide X in due parti: una che potrebbe essere correlata con u, e una che non lo è. Isolando la parte che non è correlata con u, è possibile stimare β_1.
- Per fare questo si utilizza una ***variabile strumentale***, Z_i, che è correlata con X_i ma incorrelata con u_i.

Terminologia: endogeneità ed esogeneità

Una variabile ***endogena*** è una variabile correlata con *u*

Una variabile ***esogena*** è una variabile incorrelata con *u*

Nella regressione IV ci concentriamo sul caso in cui *X* è endogena ed esiste uno strumento, *Z*, esogeno.

Digressione sulla terminologia: "endogeno" significa letteralmente "determinato all'interno del sistema". Se *X* è congiuntamente determinata con *Y*, allora una regressione di *Y* su *X* è soggetta a distorsione da causalità simultanea. Ma questa definizione di endogeneità è troppo stretta perché sia possibile usare la regressione IV per risolvere i problemi di distorsione da variabili omesse e da errori nelle variabili, quindi usiamo la definizione più ampia fornita sopra.

Due condizioni per avere uno strumento valido

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

Perché una variabile strumentale (uno "***strumento***") Z sia valida, deve soddisfare due condizioni:

1. ***Rilevanza***: $\text{corr}(Z_i, X_i) \neq 0$
2. ***Esogeneità***: $\text{corr}(Z_i, u_i) = 0$

Supponiamo per ora di avere Z_i (vedremo più avanti come trovare variabili strumentali); come possiamo usarlo per stimare β_1?

Lo stimatore IV con una *X* e una *Z*

Spiegazione 1: minimi quadrati in due stadi (TSLS)

Ci sono due stadi – due regressioni:

(1) Si isola la parte di *X* che non è correlata con *u* mediante la regressione di *X* su *Z* usando gli OLS:

$$X_i = \pi_0 + \pi_1 Z_i + v_i \qquad (1)$$

- Poiché Z_i non è correlato con u_i, $\pi_0 + \pi_1 Z_i$ non è correlato con u_i. Non conosciamo π_0 o π_1 ma li abbiamo stimati, perciò…
- Si calcolano i valori predetti di X_i, , dove $\hat{X}_i = \widehat{\pi_0} + \widehat{\pi_1} Z_i$

Minimi quadrati in due stadi (continua)

(2) Si sostituisce X_i con $\hat{X}_i$ nella regressione di interesse:
si esegue la regressione di Y su $\hat{X}_i$ usando gli OLS:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i \qquad (2)$$

- **Poiché $\hat{X}_i$ è incorrelato con u_i, la prima assunzione dei minimi quadrati vale per la regressione (2).** (Ciò richiede che n sia grande in modo che π_0 e π_1 siano stimati con precisione)
- Quindi, in grandi campioni, β_1 può essere stimato con gli OLS usando la regressione (2)
- Lo stimatore risultante è detto *stimatore dei minimi quadrati in due stadi (TSLS)*, $\hat{\beta}_1^{TSLS}$.

Minimi quadrati in due stadi: riepilogo

Supponiamo che Z_i, soddisfi le due condizioni per uno strumento valido:

1. ***Rilevanza***: corr(Z_i,X_i) ≠ 0
2. ***Esogeneità***: corr(Z_i,u_i) = 0

<u>Minimi quadrati in due stadi:</u>

Stadio 1: Regressione di X_i su Z_i (inclusa intercetta), ottenendo i valori predetti $\hat{X}_i$

Stadio 2: Regressione di Y_i su $\hat{X}_i$(inclusa intercetta); il coefficiente di $\hat{X}_i$ è lo stimatore TSLS, $\hat{\beta}_1^{TSLS}$.

$\hat{\beta}_1^{TSLS}$ è uno stimatore consistente di β_1.

Lo stimatore IV, una X e una Z (continua)

Spiegazione 2: derivazione algebrica diretta

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

Allora

$$\begin{aligned}
\text{cov}(Y_i, Z_i) &= \text{cov}(\beta_0 + \beta_1 X_i + u_i, Z_i) \\
&= \text{cov}(\beta_0, Z_i) + \text{cov}(\beta_1 X_i, Z_i) + \text{cov}(u_i, Z_i) \\
&= \quad 0 \quad + \text{cov}(\beta_1 X_i, Z_i) + \quad 0 \\
&= \beta_1 \text{cov}(X_i, Z_i)
\end{aligned}$$

dove $\text{cov}(u_i, Z_i) = 0$ per l'esogeneità dello strumento; quindi

$$\beta_1 = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(X_i, Z_i)}$$

Lo stimatore IV, una X e una Z (continua)

$$\beta_1 = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(X_i, Z_i)}$$

Lo stimatore IV sostituisce queste covarianze della popolazione con covarianze campionarie:

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}},$$

s_{YZ} e s_{XZ} sono covarianze campionarie. Questo è lo stimatore TSLS – con una derivazione diversa!

Lo stimatore IV, una X e una Z (continua)

Spiegazione 3: derivazione dalla "forma ridotta"

La "forma ridotta" mette in relazione *Y* a *Z* e *X* a *Z*:

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$

$$Y_i = \gamma_0 + \gamma_1 Z_i + w_i$$

dove w_i è un termine d'errore. Poiché *Z* è esogena, è incorrelata con v_i e con w_i.

L'idea: una variazione unitaria in Z_i comporta una variazione in X_i di π_1 e una variazione in Y_i di γ_1. Poiché tale variazione in X_i nasce dalla variazione esogena in Z_i, tale variazione in X_i è esogena. Quindi una variazione esogena in X_i di π_1 unità è associata a una variazione in Y_i di γ_1 unità – perciò l'effetto su *Y* di una variazione esogena in *X* è $\boldsymbol{\beta_1 = \gamma_1/\pi_1}$ unità.

Lo stimatore IV dalla forma ridotta (continua)

I calcoli:

$X_i = \pi_0 + \pi_1 Z_i + v_i$

$Y_i = \gamma_0 + \gamma_1 Z_i + w_i$

Risolviamo l'equazione di *X* in *Z*:

$Z_i = -\pi_0/\pi_1 + (1/\pi_1)X_i - (1/\pi_1)v_i$

Sostituiamo nell'equazione di *Y* e raccogliamo i termini:

$$
\begin{aligned}
Y_i &= \gamma_0 + \gamma_1 Z_i + w_i \\
&= \gamma_0 + \gamma_1[-\pi_0/\pi_1 + (1/\pi_1)X_i - (1/\pi_1)v_i] + w_i \\
&= [\gamma_0 - \pi_0\gamma_1/\pi_1] + (\gamma_1/\pi_1)X_i + [w_i - (\gamma_1/\pi_1)v_i] \\
&= \beta_0 + \beta_1 X_i + u_i,
\end{aligned}
$$

dove $\beta_0 = \gamma_0 - \pi_0\gamma_1/\pi_1$, **$\beta_1 = \gamma_1/\pi_1$**, e $u_i = w_i - (\gamma_1/\pi_1)v_i$.

Lo stimatore IV dalla forma ridotta (continua)

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$

$$Y_i = \gamma_0 + \gamma_1 Z_i + w_i$$

quindi

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

dove

$$\beta_1 = \gamma_1 / \pi_1$$

Interpretazione: una variazione esogena in X_i di π_1 unità è associata a una variazione in Y_i di γ_1 unità – perciò l'effetto su Y di una variazione unitaria esogena in X è $\beta_1 = \gamma_1/\pi_1$.

Esempio 1: effetto dello studio sui voti

Qual è l'effetto sui voti di studiare un'ora in più al giorno?

Y = media voti

X = tempo di studio (ore al giorno)

Dati: voti e ore di studio di studenti del primo anno di college.

Vi aspettate che lo stimatore OLS di β_1 (l'effetto sulla media voti di studiare un'ora in più al giorno) sia non distorto? Perché, o perché no?

Effetto dello studio sui voti (continua)

Stinebrickner, Ralph and Stinebrickner, Todd R. (2008) "The Causal Effect of Studying on Academic Performance," *The B.E. Journal of Economic Analysis & Policy*: Vol. 8: Iss. 1 (Frontiers), Article 14.

- n = 210 studenti del primo anno al Berea College (Kentucky) nel 2001
- Y = media voti del primo semestre
- X = media ore di studi al giorno (sondaggio)
- I compagni di stanza sono stati assegnati a caso
- Z = 1 se il compagno di stanza ha portato un videogioco, = 0 altrimenti

Pensate che Z_i (indica se un compagno ha portato un videogioco) sia uno strumento valido?

1. È rilevante (correlato con X)?
2. È esogeno (incorrelato con u)?

Effetto dello studio sui voti (continua)

$X = \pi_0 + \pi_1 Z + v_i$

$Y = \gamma_0 + \gamma_1 Z + w_i$

*Y = media voti (*scala 4 punti)

X = tempo di studio (ore al giorno)

Z = 1 se il compagno ha portato un videogioco, = 0 altrimenti

Risultati di Stinebrinckner e Stinebrinckner

$\hat{\pi}_1 = -0{,}668$

$\hat{\gamma}_1 = -0{,}241$

$$\hat{\beta}_1^{IV} = \frac{\hat{\gamma}_1}{\hat{\pi}_1} = \frac{-0{,}241}{-0{,}668} = 0{,}360$$

Quali sono le unità? Queste stime hanno senso nel mondo reale?

(*Nota*: in realtà hanno eseguito le regressioni con regressori aggiuntivi – ci torneremo più avanti)

Consistenza dello stimatore TSLS

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$$

Le covarianze campionarie sono consistenti:

$s_{YZ} \xrightarrow{p} \text{cov}(Y,Z)$ e $s_{XZ} \xrightarrow{p} \text{cov}(X,Z)$. Quindi

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}} \xrightarrow{p} \frac{\text{cov}(Y,Z)}{\text{cov}(X,Z)} = \beta_1$$

- La condizione di rilevanza dello strumento, $\text{cov}(X,Z) \neq 0$, assicura che non si esegua una divisione per zero.

Esempio 2: offerta e domanda di burro

La regressione IV è stata sviluppata in origine per stimare l'elasticità della domanda per beni agricoli, per esempio il burro:

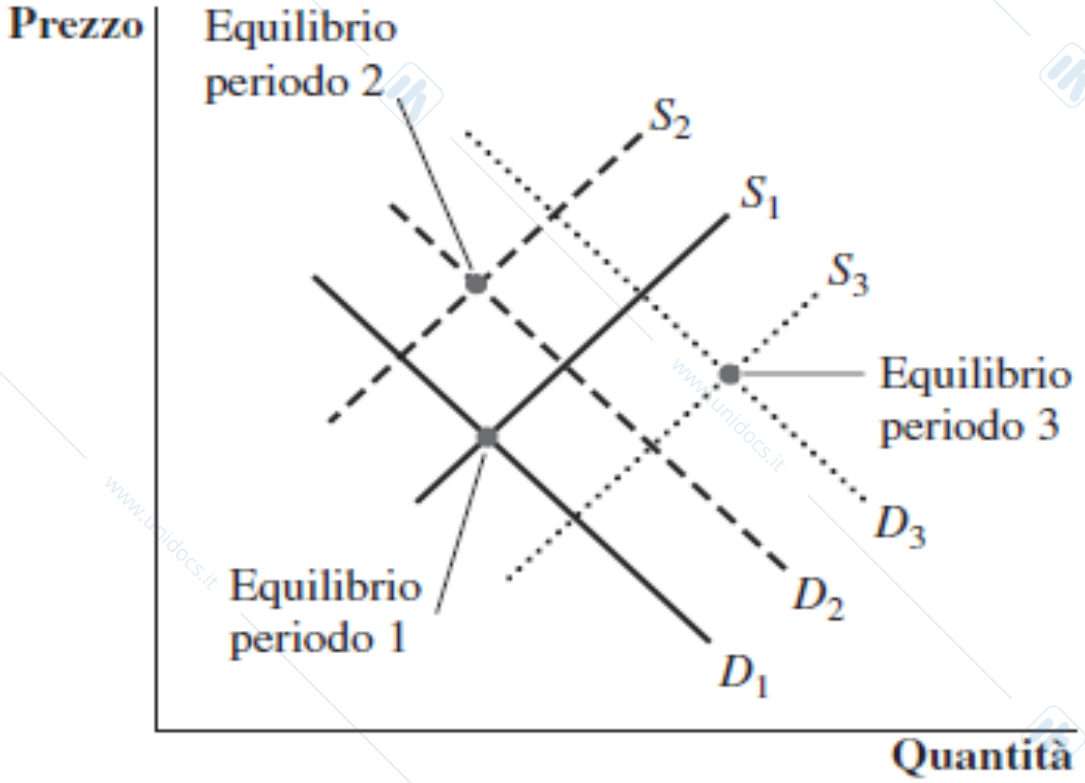
$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

- β_1 = elasticità del burro = variazione percentuale in quantità per una variazione dell'1% in prezzo (si ricordi la discussione sulla specifica log-log)

- Dati: osservazioni su prezzo e quantità di burro per diversi anni

- La regressione OLS di $\ln(Q_i^{butter})$ su $\ln(P_i^{butter})$ soffre di distorsione da causalità simultanea (*perché*?)

La distorsione da causalità simultanea nella regressione OLS di ln(Q_i^{butter}) su ln(P_i^{butter}) nasce perché prezzo e quantità sono determinati dall'interazione di domanda *e* offerta:



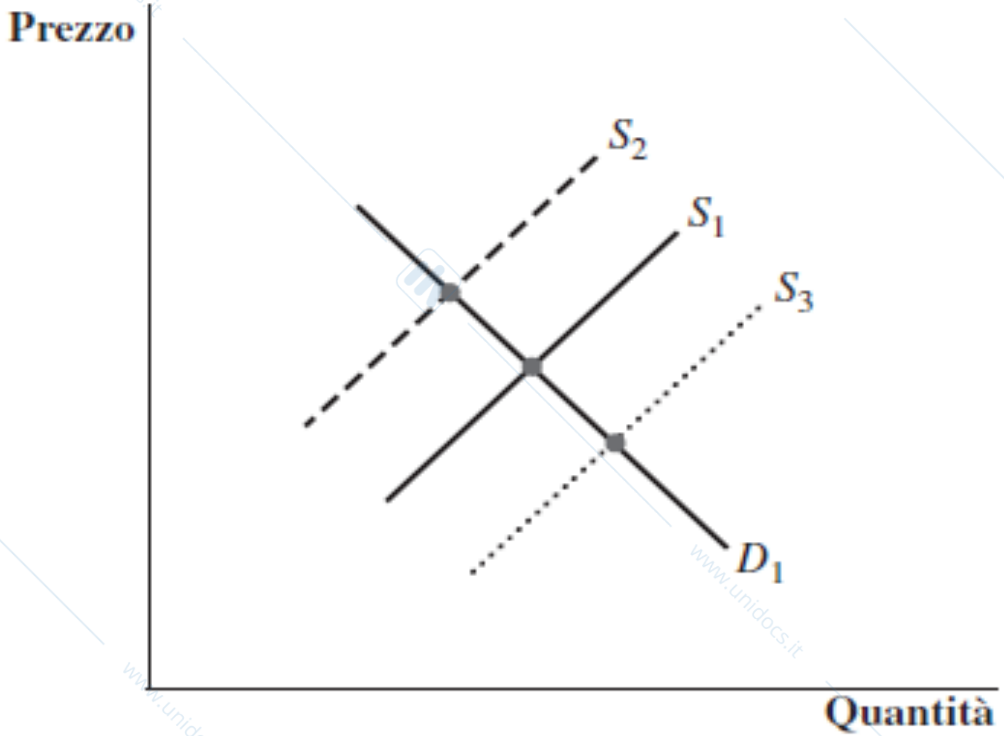
(a) Domanda e offerta in tre periodi

Questa interazione di domanda e offerta produce dati come…

(b) Prezzo e quantità di equilibrio per 11 periodi

Una regressione con questi dati produrrebbe la curva di domanda?

E se si spostasse solo l'offerta?



(c) Prezzo e quantità di equilibrio quando solo la curva di offerta si sposta

- TSLS stima la curva di domanda isolando gli spostamenti di prezzo e quantità conseguenti a spostamenti dell'offerta.
- *Z* è una variabile che sposta l'offerta ma non la domanda.

TSLS nell'esempio di domanda e offerta:

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

Sia Z = pioggia nelle aree di produzione lattiera.

Z è uno strumento valido?

(1) Rilevante? corr($rain_i$,ln(P_i^{butter})) ≠ 0?

Plausibilmente: pioggia insufficiente significa meno pascolo quindi meno burro e quindi prezzi più alti

(2) Esogeno? corr($rain_i$,u_i) = 0?

Plausibilmente: la pioggia nelle aree di produzione lattiera non dovrebbe influenzare la domanda di burro

TSLS nell'esempio di domanda e offerta (continua)

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

$Z_i = rain_i$ = pioggia nelle aree di produzione lattiera.

Passo 1: regressione di $\ln(P_i^{butter})$ su *rain*, dà $\ln(\widehat{P_i^{butter}})$

$\ln(\widehat{P_i^{butter}})$ isola le variazioni nel log del prezzo conseguenti all'offerta (o almeno a parte di essa)

Passo 2: regressione di $\ln(Q_i^{butter})$ su $\ln(\widehat{P_i^{butter}})$

Controparte dell'uso degli spostamenti della curva di offerta per tracciare la curva di domanda.

Esempio 3: punteggi nei test e dimensioni delle classi

- Le regressioni per punteggi nei test/dimensioni delle classi in California potrebbero avere distorsione da variabili omesse (per esempio per interessamento dei genitori).
- In linea di principio questa distorsione può essere eliminata dalla regressione IV (TSLS).
- La regressione IV richiede uno strumento valido, cioè che sia:

1. rilevante: $\text{corr}(Z_i, STR_i) \neq 0$
2. esogeno: $\text{corr}(Z_i, u_i) = 0$

Esempio 3: punteggi nei test e dimensioni delle classi (continua)

Ecco uno strumento ipotetico:

- alcuni distretti, colpiti casualmente da un terremoto, "raddoppiano" le classi:

  $Z_i = Quake_i$ = 1 se colpito da terremoto, = 0 altrimenti
- *Valgono le due condizioni per la validità dello strumento*?
- Il terremoto crea una situazione *come se* i distretti rientrassero in un esperimento con assegnazione casuale. Quindi, la variazione in *STR* conseguente al terremoto è esogena.
- Il primo stadio del TSLS prevede la regressione di *STR* su *Quake*, isolando così la parte esogena di *STR* (la parte "come se" fosse assegnata casualmente)

Inferenza con TSLS

- In grandi campioni, la distribuzione campionaria dello stimatore TSLS è normale
- L'inferenza (verifiche di ipotesi, intervalli di confidenza) procede nel modo consueto, ovvero $\pm$ 1,96*SE*
- Il concetto alla base della distribuzione normale in grandi campioni dello stimatore TSLS è che – come tutti gli altri stimatori che abbiamo considerato – comporta variabili casuali i.i.d. con media nulla, a cui possiamo applicare il TLC.
- Di seguito riportiamo i calcoli abbozzati (si veda l'Appendice 12.3 per i dettagli)...

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}$$

$$= \frac{\sum_{i=1}^{n} Y_i(Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i(Z_i - \bar{Z})}$$

Sostituiamo in $Y_i = \beta_0 + \beta_1 X_i + u_i$ e semplifichiamo:

$$\hat{\beta}_1^{TSLS} = \frac{\beta_1 \sum_{i=1}^{n} X_i(Z_i - \bar{Z}) + \sum_{i=1}^{n} u_i(Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i(Z_i - \bar{Z})}$$

quindi…

$$\hat{\beta}_1^{TSLS} = \beta_1 + \frac{\sum_{i=1}^{n} u_i (Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i (Z_i - \bar{Z})}$$

Quindi $$\hat{\beta}_1^{TSLS} - \beta_1 = \frac{\sum_{i=1}^{n} u_i (Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i (Z_i - \bar{Z})}$$

Moltiplicando entrambi i membri per: $\sqrt{n}$

$$\sqrt{n}(\hat{\beta}_1^{TSLS} - \beta_1) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i - \bar{Z}) u_i}{\frac{1}{n} \sum_{i=1}^{n} X_i (Z_i - \bar{Z})}$$

$$\sqrt{n}\,(\hat{\beta}_1^{TSLS} - \beta_1) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i - \bar{Z})u_i}{\frac{1}{n}\sum_{i=1}^{n}X_i(Z_i - \bar{Z})}$$

$$\frac{1}{n}\sum_{i=1}^{n}X_i(Z_i - \bar{Z}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z}) \xrightarrow{p} \text{cov}(X,Z) \neq 0$$

$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i - \bar{Z})u_i$ ha distribuzione $N(0,\text{var}[(Z-\mu_Z)u])$ (TLC)

quindi: $\hat{\beta}_1^{TSLS}$ ha distribuzione appr. $N(\beta_1, \sigma^2_{\hat{\beta}_1^{TSLS}})$,

dove $\sigma^2_{\hat{\beta}_1^{TSLS}} = \frac{1}{n}\frac{\text{var}[(Z_i - \mu_Z)u_i]}{[\text{cov}(Z_i, X_i)]^2}$

dove cov(*X*,*Z*) ≠ 0 perché lo strumento è rilevante

Inferenza con TSLS (continua)

$\hat{\beta}_1^{TSLS}$ ha distribuzione appr. $N(\beta_1, \sigma^2_{\hat{\beta}_1^{TSLS}})$,

- L'inferenza statistica procede nel modo consueto.
- La giustificazione è (come di consueto) basata su grandi campioni
- Tutto questo assume che gli strumenti siano validi – vedremo tra breve che cosa accade se non lo sono.
- ***Nota importante sugli errori standard***:
  - Gli errori standard OLS dalla regressione del secondo stadio non sono corretti – non tengono conto della stima al primo stadio ($\hat{X}_i$ è stimata).
  - Si utilizza invece un singolo comando apposito che calcola lo stimatore TSLS e gli errori standard corretti.
  - Come di consueto, si usano errori standard robusti all'eteroschedasticità

Esempio 4: domanda di sigarette

$$\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1 \ln(P_i^{cigarettes}) + u_i$$

Perché lo stimatore OLS di β_1 è probabilmente distorto?

- Data set: dati panel sul consumo annuo e i prezzi medi (comprese le imposte) delle sigarette, per stato, per i 48 stati contigui USA, 1985-1995.
- Variabile strumentale proposta:
- Z_i = imposta generale sulle vendite al pacchetto nello stato = $SalesTax_i$ una sorta di IVA nello stato i.
- Pensate che questo strumento sia valido?

1. Rilevante? corr($SalesTax_i$, ln($P_i^{cigarettes}$)) ≠ 0?
2. Esogeno? corr($SalesTax_i$,u_i) = 0?

Esempio 4: domanda di sigarette (continua)

Discussione:

Si noti che le aliquote negli USA cambiano di stato in stato!!!!

la condizione di rilevanza è certamente soddisfatta.

condizione di esogeneità: l'imposta sulle vendite deve essere incorrelata con u, i.e. incorrelata con $Q_i^{cigarettes}$ la domanda di sigarette. Le aliquote vengono definite sulla base di considerazioni politiche e di budget pubblico, non in base alla domanda.

Tuttavia, se le aliquote fossere fissate anche in funzione del reddito dello stato, allora è correlate con u ➔ strumento non valido. In questo caso IV è distorto e bisogna inserire il reddito nella equazione

Domanda di sigarette (continua)

Per ora usiamo solo i dati del 1995.

Primo stadio regressione OLS:

$$\widehat{\ln(P_i^{cigarettes})} = 4.63 + 0{,}031 SalesTax_i,\ n = 48$$

Secondo stadio regressione OLS: NB non forma ridotta!!!!

$$\widehat{\ln(Q_i^{cigarettes})} = 9{,}72 - 1{,}08 \widehat{\ln(P_i^{cigarettes})},\ n = 48$$

Regressione TSLS combinata con errori standard corretti, robusti all'eteroschedasticità:

$$\widehat{\ln(Q_i^{cigarettes})} = \underset{(1{,}53)}{9{,}72} - \underset{(0{,}32)}{1{,}08} \ln(P_i^{cigarettes}),\ n = 48$$

Esempio STATA: domanda di sigarette, primo stadio

Strumento = *Z* = *rtaxso* = imposta vendite (dollari reali/pacchetto)

```
                X         Z
. reg lravgprs rtaxso if year==1995, r

Regression with robust standard errors                Number of obs =      48
                                                      F(  1,    46) =   40.39
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4710
                                                      Root MSE      =  .09394

------------------------------------------------------------------------------
             |               Robust
    lravgprs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      rtaxso |   .0307289   .0048354     6.35   0.000     .0209956    .0404621
       _cons |   4.616546   .0289177   159.64   0.000     4.558338    4.674755
------------------------------------------------------------------------------

             X-hat
. predict lravphat   Ora abbiamo i valori predetti dal primo stadio
```

Secondo stadio

Y X-hat

```
. reg lpackpc lravphat if year==1995, r

Regression with robust standard errors                Number of obs =      48
                                                      F(  1,    46) =   10.54
                                                      Prob > F      =  0.0022
                                                      R-squared     =  0.1525
                                                      Root MSE      =  .22645

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravphat |  -1.083586   .3336949    -3.25   0.002    -1.755279   -.4118932
       _cons |   9.719875   1.597119     6.09   0.000     6.505042    12.93471
------------------------------------------------------------------------------
```

- Questi coefficienti sono le stime TSLS
- Gli errori standard sono sbagliati perché ignorano la stima al primo stadio

Tutto in un unico comando:

```
                    Y         X         Z
. ivregress 2sls lpackpc (lravgprs = rtaxso) if year==1995, vce(robust)

Instrumental variables (2SLS) regression          Number of obs =      48
                                                  Wald chi2(1)  =   12.05
                                                  Prob > chi2   =  0.0005
                                                  R-squared     =  0.4011
                                                  Root MSE      =  .18635

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.083587   .3122035    -3.47   0.001    -1.695494    -.471679
       _cons |   9.719876   1.496143     6.50   0.000     6.78749     12.65226
------------------------------------------------------------------------------
Instrumented:  lravgprs          This is the endogenous regressor
Instruments:   rtaxso            This is the instrumental varible
------------------------------------------------------------------------------
```

Equazione della domanda di sigarette stimata:

$$\widehat{\ln(Q_i^{cigarettes})} = \underset{(1,53)}{9.72} - \underset{(0,31)}{1,08}\ln(P_i^{cigarettes}) \qquad , n = 48$$

Riepilogo della regressione IV con singola *X* e *Z*

- Uno strumento valido *Z* deve soddisfare due condizioni:
  1. *rilevanza*: $\text{corr}(Z_i, X_i) \neq 0$
  2. *esogeneità*: $\text{corr}(Z_i, u_i) = 0$
- TSLS procede eseguendo prima la regressione di *X* su *Z* per ottenere $\hat{X}$, poi di *Y* su $\hat{X}$
- Il concetto chiave è che il primo stadio isola la parte della variazione in *X* che è incorrelata con *u*
- Se lo strumento è valido, allora la distribuzione in grandi campioni dello stimatore TSLS è normale, perciò l'inferenza procede come di consueto

Il modello generale di regressione IV (Paragrafo 12.2)

- Finora abbiamo considerato la regressione IV con un singolo regressore endogeno (X) e un singolo strumento (Z).
- Dobbiamo estenderla a:
  - più regressori endogeni ($X_1,\ldots,X_k$)
  - più variabili incluse esogene ($W_1,\ldots,W_r$) o variabili di controllo, che devono essere incluse per il consueto motivo delle variabili omesse
  - più variebili strumentali ($Z_1,\ldots,Z_m$). Più strumenti (rilevanti) possono produrre una minore varianza del TSLS: l'R^2 del primo stadio aumenta, perciò si ha maggiore variazione in $\hat{X}$.
- *Nuovi termini*: identificazione e sovraidentificazione

Identificazione

- In generale si dice che un parametro è ***identificato*** se diversi valori del parametro producono distribuzioni diverse dei dati.
- Nella regressione IV, il fatto che i coefficienti siano identificati dipende dalla relazione tra il numero di strumenti (m) e il numero di regressori endogeni (k)
- Intuitivamente, se ci sono meno strumenti che regressori endogeni, non possiamo stimare $\beta_1,\ldots,\beta_k$
  - Per esempio, supponiamo $k = 1$ ma $m = 0$ (nessuno strumento)!

Identificazione (continua)

I coefficienti $\beta_1,\ldots, \beta_k$ si dicono:

- ***esattamente identificati*** se $m = k$.

Ci sono esattamente gli strumenti sufficienti per stimare $\beta_1,\ldots,\beta_k$.

- ***sovraidentificati*** se $m > k$.

Ci sono più strumenti di quelli necessari per stimare $\beta_1,\ldots, \beta_k$. In questo caso si può verificare se gli strumenti sono validi (test delle "restrizioni sovraidentificanti") – torneremo sul tema in seguito

- ***sottoidentificati*** se $m < k$.

Ci sono troppo pochi strumenti per stimare $\beta_1,\ldots, \beta_k$. In questo caso occorre procurarsi più strumenti!

Il modello generale di regressione IV: riepilogo della terminologia

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$$

- Y_i è la ***variabile dipendente***
- $X_{1i},\ldots, X_{ki}$ sono i ***regressori endogeni*** (potenzialmente correlati con u_i)
- $W_{1i},\ldots,W_{ri}$ sono i ***regressori esogeni inclusi*** (incorrelati con u_i) o ***variabili di controllo*** (inclusi in modo che Z_i sia incorrelata con u_i, una volta inclusi i W)
- $\beta_0, \beta_1,\ldots, \beta_{k+r}$ sono i coefficienti di regressione ignoti
- $Z_{1i},\ldots,Z_{mi}$ sono le m ***variabili strumentali*** (***variabili esogene escluse***)
- I coefficienti sono ***sovraidentificati*** se $m > k$; ***esattamente identificati*** se $m = k$; ***sottoidentificati*** se $m < k$.

TSLS con un singolo regressore endogeno

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i$$

- m strumenti: $Z_{1i}, \ldots, Z_m$
- Primo stadio
  - Regressione di X_1 su *tutti* i regressori esogeni: regressione di X_1 su $W_1, \ldots, W_r$, $Z_1, \ldots, Z_m$, e un'intercetta, usando OLS
  - Calcolo dei valori predetti $\hat{X}_{1i}$, $i = 1, \ldots, n$
- Secondo stadio
  - Regressione di Y su $\hat{X}_{1i}$, $W_1, \ldots, W_r$, e un'intercetta, usando OLS
  - I coefficienti di questa regressione del secondo stadio sono gli stimatori TSLS, ma gli errori standard sono sbagliati
- Per ottenere errori standard corretti, occorre procedere in un singolo pasasggio con il software di regressione

Esempio 4: ancora la domanda di sigarette

Si supponga che il reddito sia esogeno (è plausibile – *perché*?), e di voler anche stimare l'elasticità:

$$\ln(\widehat{Q_i^{cigarettes}}) = \beta_0 + \beta_1 \ln(\widehat{P_i^{cigarettes}}) + \beta_2 \ln(Income_i) + u_i$$

Abbiamo due strumenti:

Z_{1i} = imposta generale sulle vendite, pensate all'IVA

Z_{2i} = imposta specifica sulle sigarette, accisa sul tabacco

- Variabile endogena: $\ln(\widehat{P_i^{cigarettes}})$ ("una sola *X*")
- Variabile esogena inclusa: $\ln(Income_i)$ ("una sola *W*")
- Strumenti (variabili escluse): imposta generale vendite, imposta specifica sulle sigarette ("due *Z*")
- *β_1 è sotto, sopra o esattamente identificata?*

Esempio: domanda di sigarette, un solo strumento

IV: rtaxso = real overall sales tax in state

Y *W* *X* *Z*

```
. ivreg lpackpc lperinc (lravgprs = rtaxso) if year==1995, r

IV (2SLS) regression with robust standard errors      Number of obs =      48
                                                      F(  2,    45) =    8.19
                                                      Prob > F      =  0.0009
                                                      R-squared     =  0.4189
                                                      Root MSE      =  .18957

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.143375   .3723025    -3.07   0.004    -1.893231   -.3935191
     lperinc |    .214515   .3117467     0.69   0.495     -.413375    .842405
       _cons |   9.430658   1.259392     7.49   0.000     6.894112    11.9672
------------------------------------------------------------------------------
Instrumented:  lravgprs
Instruments:   lperinc rtaxso
```

STATA lists ALL the exogenous regressors as instruments – slightly different terminology than we have been using

- Running IV as a single command yields the correct SEs
- Use , r for heteroskedasticity-robust *SE*s

Esempio: domanda di sigarette, due strumenti

Y *W* *X* Z_1 Z_2

```
. ivreg lpackpc lperinc (lravgprs = rtaxso rtax) if year==1995, r

IV (2SLS) regression with robust standard errors      Number of obs =      48
                                                      F(  2,    45) =   16.17
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4294
                                                      Root MSE      =  .18786

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.277424   .2496099    -5.12   0.000    -1.780164   -.7746837
     lperinc |   .2804045   .2538894     1.10   0.275     -.230955    .7917641
       _cons |   9.894955   .9592169    10.32   0.000     7.962993    11.82692
------------------------------------------------------------------------------
Instrumented:  lravgprs
Instruments:   lperinc rtaxso rtax    STATA lists ALL the exogenous regressors
                                      as "instruments" - slightly different
                                      terminology than we have been using
------------------------------------------------------------------------------
```

Stime TSLS, Z = imposta vendite (m = 1)

$$\ln(\widehat{Q_i^{cigarettes}}) = \underset{(1,26)}{9,43} - \underset{(0,37)}{1,14}\ln(P_i^{cigarettes}) + \underset{(0,31)}{0,21}\ln(Income_i)$$

Stime TSLS, Z = imposta vendite e imposta sig. (m = 2)

$$\ln(\widehat{Q_i^{cigarettes}}) = \underset{(0,96)}{9,89} - \underset{(0,25)}{1,28}\ln(P_i^{cigarettes}) + \underset{(0,25)}{0,28}\ln(Income_i)$$

- Errori standard minori *per m* = 2. Con 2 strumenti si hanno più informazioni, più "variazione come se casuale"
- Bassa elasticità al reddito (non è un bene di lusso); elasticità al reddito non significativamente diversa da zero a livello statistico
- Elasticità al prezzo sorprendentemente elevata

Assunzioni generali per la validità di uno strumento

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$$

(1) ***Esogeneità***: $\text{corr}(Z_{1i}, u_i) = 0, \ldots, \text{corr}(Z_{mi}, u_i) = 0$

(2) ***Rilevanza***: *caso generale, più X*

Supponiamo che la regressione del secondo stadio possa essere esguita usando i valori predetti dalla regressione del primo stadio. Allora non vi è perfetta collinearità in questa (impraticabile) regressione del secondo stadio.

- *Caso speciale di una sola X*: l'assunzione generale è equivalente a (a) almeno uno strumento deve entrare nella controparte della regressione del primo stadio e (b) i *W* non sono perfettamente collineari.

Assunzioni della regressione IV

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$$

1. $E(u_i | W_{1i},\ldots,W_{ri}) = 0$
- l'assunzione 1 dice "i regressori esogeni sono esogeni"

2. $(Y_i, X_{1i},\ldots,X_{ki}, W_{1i},\ldots,W_{ri}, Z_{1i},\ldots,Z_{mi})$ sono i.i.d.
- l'assunzione 2 non è nuova

3. X, W, Z e Y hanno momenti quarti finiti non nulli
- l'assunzione 3 non è nuova

4. Gli strumenti $(Z_{1i},\ldots,Z_{mi})$ sono validi.
   - Ne abbiamo parlato

- Sotto le assunzioni 1-4, il TSLS e la sua statistica t hanno distribuzione normale
- Il requisito fondamentale è che gli strumenti siano validi

W come variabili di controllo

- In molti casi le *W* sono incluse allo scopo di controllare per fattori omessi, cosicché, una volta incluse le *W*, *Z* è incorrelata con *u*. In questo caso le *W* non devono essere esogene, ma devono essere variabili di controllo effettive nel senso discusso nel Capitolo 7 – ora però focalizzandosi sulla produzione di uno strumento esogeno.

- Tecnicamente, la condizione perché le *W* siano variabili di controllo effettive è che la media condizionata degli u_i non dipenda da Z_i, date W_i:

$$E(u_i|W_i, Z_i) = E(u_i|W_i)$$

W come variabili di controllo (continua)

- Quindi un'alternativa alla prima assunzione della regressione IV è che valga l'indipendenza in media condizionata:

$$E(u_i|W_i, Z_i) = E(u_i|W_i)$$

  Questa è la versione IV dell'assunzione dell'indipendenza in media condizionata del Capitolo 7.
- *Ecco il punto chiave*: in molte applicazioni occorre includere variabili di controllo (*W*) affinché *Z* sia verosimilmente esogena (incorrelata con *u*).
- Per i dettagli si veda l'Appendice 12.6

Esempio 1: effetto dello studio sui voti (continua)

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

Y = media voti del primo semestre

X = media di ore di studio al giorno

Z = 1 se il compagno di stanza ha portato un videogioco, = 0 altrimenti

I compagni di stanza sono stati assegnati a caso

Conoscete un motivo per cui Z potrebbe essere correlata con u – anche se è assegnata casualmente? Che cos'altro entra nel termine d'errore, quali sono altri determinanti dei voti, oltre al tempo speso studiando?

Esempio 1: effetto dello studio sui voti (continua)

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

Perché Z potrebbe essere correlata u?

- Ecco una *ipotetica* possibilità: il genere. Supponiamo:
  - le donne ottengono voti migliori degli uomini, a parità di ore di studio
  - Gli uomini hanno più probabilità di portare un videogioco, rispetto alle donne
  - Allora corr(Z_i, u_i) < 0 (i maschi hanno più probabilità di avere un compagno di stanza [maschio] che porti un videogioco, ma i maschi tendono anche ad avere voti inferiori, a parità di tempo di studio).
- È solo un altro caso di distorsione da variabili omesse. La soluzione sta nel controllare per (o includere) la variabile omessa, in questo caso il genere.

Esempio 1: effetto dello studio sui voti (continua)

- Questa logica porta a includere *W* = genere come variabile di controllo nella regressione IV:

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

- La stima TSLS qui riportata proviene da una regressione che ha incluso il genere come variabile *W* – tra altre variabili come la materia di specializzazione.

Verifica della validità degli strumenti (Paragrafo 12.3)

Ricordiamo i due requisiti per strumenti validi:

1. *Rilevanza* (caso speciale di una sola X)
   Almeno uno strumento deve entrare nella controparte di popolazione della regressione del primo stadio.
2. *Esogeneità*

Tutti gli strumenti devono essere incorrelati con il termine d'errore:
corr(Z_{1i},u_i) = 0,…, corr(Z_{mi},u_i) = 0

Che cosa accade se uno di questi requisiti non è soddisfatto? Come si può verificare? Che cosa occorre fare?

Se si hanno più strumenti, quale si deve usare?

Verifica dell'assunzione 1: rilevanza dello strumento

Ci concentreremo su un singolo regressore incluso:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i$$

Regressione del primo stadio:

$$X_i = \pi_0 + \pi_1 Z_{1i} + \ldots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \ldots + \pi_{m+k} W_{ki} + u_i$$

- Gli strumenti sono rilevanti se almeno uno dei $\pi_1, \ldots, \pi_m$ è diverso da zero.
- Gli strumenti si dicono ***deboli*** se tutti i $\pi_1, \ldots, \pi_m$ sono uguali o vicini a zero.
- ***Gli strumenti deboli*** dicono molto poco sulla variazione in X, oltre a ciò che dicono le W

Quali sono le conseguenze di strumenti deboli?

Se gli strumenti sono deboli, la distribuzione campionaria del TSLS e della sua statistica *t* non è normale, anche con *n* grande.

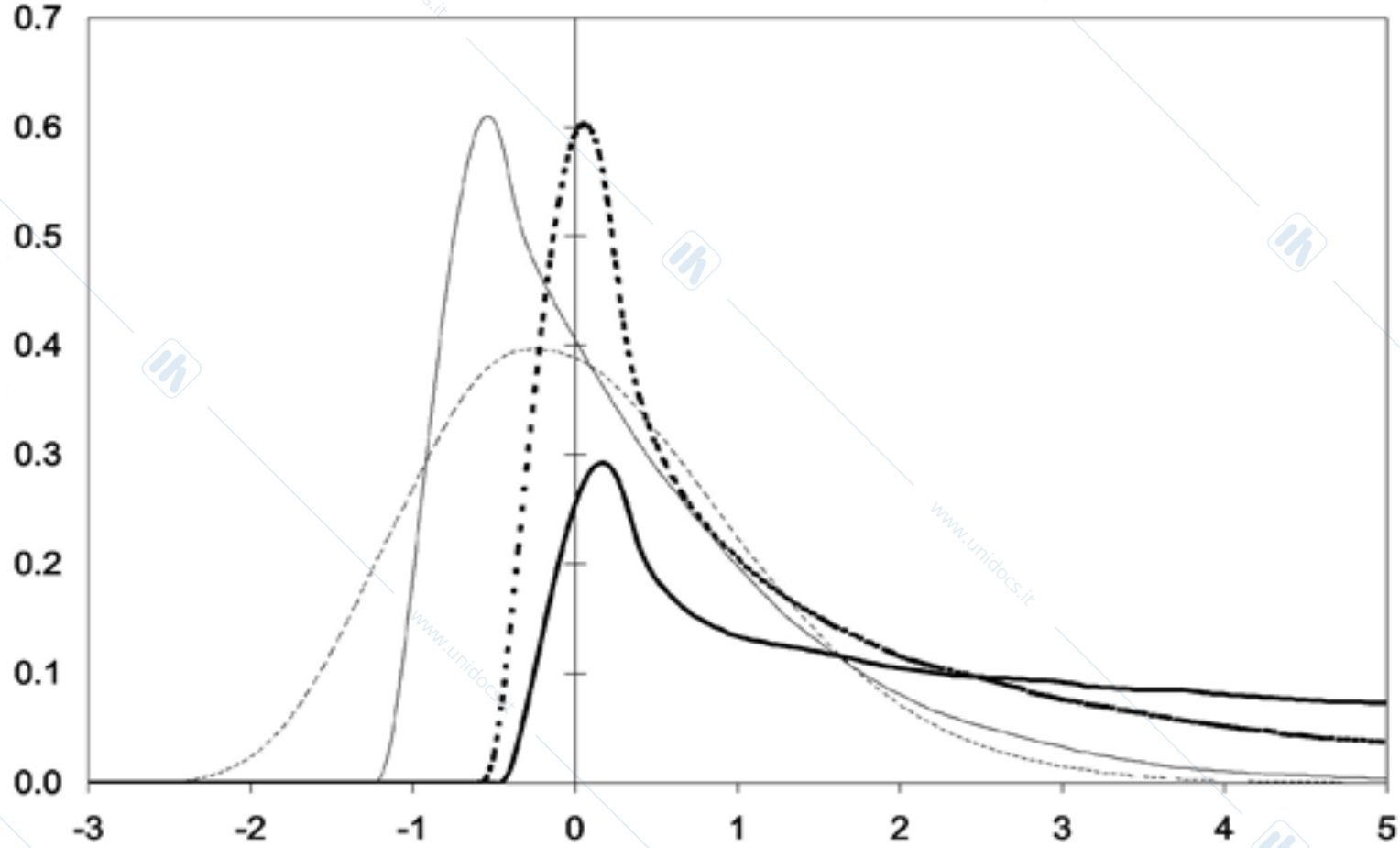
Consideriamo il caso più semplice:

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

$$X_i = \pi_0 + \pi_1 Z_i + u_i$$

- Lo stimatore IV è $\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$
- Se cov(*X*,*Z*) è zero o minore, allora s_{XZ} sarà piccolo: con strumenti deboli, il denominatore è quasi zero.
- In questo caso, la distribuzione campionaria di $\hat{\beta}_1^{TSLS}$ (e la sua statistica *t*) non è ben approssimata dall'approssimazione normale per *n* grande…

Esempio: la distribuzione campionaria della statistica *t* del TSLS con strumenti deboli



Linea scura = strumenti non rilevanti

Linea chiara tratteggiata = strumenti forti

Perché la nostra approssimazione normale ci tradisce?

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$$

- Se cov(X,Z) è piccola, piccole variazioni in s_{XZ} (da un campione al successivo) può indurre grandi variazioni in $\hat{\beta}_1^{TSLS}$
- Supponiamo di calcolare in un campione s_{XZ} = 0,00001...
- Allora l'approssimazione normale per n grande non è una buona approssimazione della distribuzione campionaria di $\hat{\beta}_1^{TSLS}$
- Un'approssimazione migliore è quella di $\hat{\beta}_1^{TSLS}$ come il *rapporto* di due variabili casuali normali correlate (si veda l'Appendice 12.4)
- Se gli strumenti sono deboli, i consueti metodi di inferenza sono inaffidabili – potenzialmente molto inaffidabili.

Misurazione della forza degli strumenti in pratica: la statistica *F* del primo stadio

- La regressione del primo stadio (una sola X):
- Regressione di X su $Z_1,..,Z_m,W_1,\ldots,W_k$.
- Strumenti totalmente irrilevanti ←→ *tutti* i coefficienti di $Z_1,\ldots,Z_m$ sono zero.
- La ***statistica F del primo stadio*** verifica l'ipotesi che $Z_1,\ldots,Z_m$ non entrino nella regressione del primo stadio.
- Strumenti deboli implicano un valore basso della statistica *F* del primo stadio.

Verifica di strumenti deboli con una singola *X*

- Si calcola la statistica *F* del primo stadio.

Regola empirica: se la statistica F del primo stadio è minore di 10, allora l'insieme di strumenti è debole.

- In questo caso, lo stimatore TSLS sarà distorto, e le inferenze statistiche (errori standard, verifiche di ipotesi, intervalli di confidenza) possono essere fuorvianti.

Verifica di strumenti deboli con una singola *X* (continua)

- Perché confrontare la statistica *F* del primo stadio con 10?
- Non è sufficiente respingere l'ipotesi nulla che i coefficienti delle *Z* siano zero – serve un contenuto predittivo sostanziale per una buona approssimazione normale.
- Il confronto della statistica *F* del primo stadio con 10 verifica se la distorsione del TSLS, rispetto all'OLS, è minore del 10%. Se la *F* è minore di 10, la distorsione relativa è superiore al 10%, cioè il TSLS può avere una distorsione sostanziale (si veda l'Appendice 12.5).

Che cosa fare se si hanno strumenti deboli

- Procurarsi strumenti migliori (più facile a dirsi che a farsi!)
- Se si hanno molti strumenti, alcuni sono probabilmente più deboli di altri ed è una buona idea scartare i più deboli (scartando uno strumento irrilevante si aumenta la statistica *F* del primo stadio)
- Se si hanno pochi strumenti, e sono tutti deboli, allora occorre eseguire un'analisi IV al di là del TSLS…
  - Separare il problema della stima di β_1 e della costruzione di intervalli di confidenza
  - Sembra strano, ma se il TSLS non ha distribuzione normale, ha senso (davvero?)

Intervalli di confidenza con strumenti deboli

- Con strumenti deboli, gli intervalli di confidenza TSLS non sono validi, ma altri intervalli di confidenza lo sono. Riportiamo due modi per calcolare intervalli di confidenza validi in grandi campioni anche se gli strumenti sono deboli:

1. L'intervallo di confidenza di Anderson-Rubin

- L'intervallo di confidenza di Anderson-Rubin si basa sulla statistica test di Anderson-Rubin che verifica $\beta_1 = \beta_{1,0}$:
  - Si calcola = $Y_i - \beta_{1,0}X_i$
  - Si esegue la regressione su $W_{1i},\ldots, W_{ri}, Z_{1i},\ldots, Z_{mi}$
  - Il test AR è la statistica F su $Z_{1i},\ldots, Z_{mi}$
- Ora si inverte il test: l'intervallo di confidenza AR al 95% è l'insieme di β_1 non rifiutati al livello del 5% dal test AR.
- Calcolo: si usa software specialistico.

Intervalli di confidenza con strumenti deboli (continua)

2. Intervallo di confidenza del rapporto di verosimiglianza condizionato di Moreira

- L'intervallo di confidenza del rapporto di verosimiglianza condizionato è basato sull'inversinoe del test del repporto di verosimiglianza condizionato di Moreira. Per calcolare questo test, il suo valore critico e l'intervallo di confidenza del rapporto di verosimiglianza condizionato, è richiesto un software specialistico.
- Questo intervallo di confidenza tende a essere più ristretto di quello di Anderson-Rubin, soprattutto quando vi sono molti strumenti.
- Se si dispone di un software che produce questo intervallo, è il caso di usarlo.

Stima con strumenti deboli

Non ci sono stimatori non distorti se gli strumenti sono deboli o irrilevanti. Tuttavia, alcuni stimatori hanno una distribuzione più centrata su β_1 del TSLS.

- Uno di questi stimatori è quello di massima verosimiglianza con informazione limitata (LIML)
- Lo stimatore LIML
  - può essere derivato come stimatore di massima verosimiglianza
  - è il valore di β_1 che minimizza il valore-p del test AR (!)
- Per approfondire stimatori, verifiche e intervalli di confidenza nel caso di strumenti deboli, si veda l'Appendice 12.5

Verifica dell'assunzione 2: esogeneità dello strumento

- Esogeneità dello strumento: ***Tutti*** gli strumenti sono correlati con il termine d'errore:
  $\text{corr}(Z_{1i}, u_i) = 0, \ldots, \text{corr}(Z_{mi}, u_i) = 0$
- Se gli strumenti sono correlati con il termine d'errore, il primo stadio del TSLS non può isolare una componente di X incorrelata con il termine d'errore, perciò $\hat{X}$ è correlata con u e il TSLS è inconsistente.
- Se ci sono più strumenti che regressori endogeni, è possibile verificare – *parzialmente* – l'esogeneità dello strumento.

Verifica di restrizioni di sovraidentificazione

Consideriamo il caso più semplice:

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

- Supponiamo che vi siano due strumenti validi: Z_{1i}, Z_{2i}
- Allora potremmo calcolare due stime TSLS separate.
- Intuitivamente, se queste due stime TSLS sono molto diverse tra loro, ci dev'essere qualcosa di sbagliato: uno strumento o l'altro (o entrambi) devono essere non validi.
- Il test *J* di restrizioni sovraidentificanti esegue questo confronto in un modo statisticamente preciso.
- Si può fare soltanto se il numero di *Z* è maggiore del numero di *X* (sovraidentificazione).

Il test *J* di restrizioni di sovraidentificazione

Supponiamo che il numero di strumenti = m > numero di $X = k$ (sovraidentificazione)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$$

Il test *J* è il test di Anderson-Rubin, usando lo stimatore TSLS al posto del valore ipotizzato $\beta_{1,0}$. Procedura:

1. Prima si stima l'equazione di interesse usando TSLS e tutti gli m strumenti; si calcolano i valori predetti $\hat{Y}_i$, usando le X effettive (non le $\hat{X}$ usate per stimare il secondo stadio)
2. Si calcolano i residui $\hat{u}_i = Y_i - \hat{Y}_i$
3. Si esegue la regressione rispetto a $Z_{1i},\ldots,Z_{mi}$, $W_{1i},\ldots,W_{ri}$
4. Si calcola la statistica F che verifica l'ipotesi che i coefficienti di $Z_{1i},\ldots,Z_{mi}$ siano tutti zero;
5. La ***statistica J*** è $J = mF$

Il test *J* (continua)

$J = mF$, dove F = la statistica F che verifica i coefficienti di $Z_{1i},\ldots,Z_{mi}$ in una regressione dei residui TSLS rispetto a $Z_{1i},\ldots,Z_{mi}$, $W_{1i},\ldots,W_{ri}$.

Distribuzione della statistica *J*

- Sotto l'ipotesi nulla che tutti gli strumenti siano esogeni, J ha una distribuzione chi-quadro con $m–k$ gradi di libertà
- Se $m = k$, $J = 0$ (*ha senso?*)
- Se alcuni strumenti sono esogeni e altri sono endogeni, la statistica J sarà grande, e l'ipotesi nulla che tutti gli strumenti sono esogeni sarà respinta.

Verifica della validità degli strumenti: riepilogo

Questo riepilogo considera il caso di una singola *X*. I due requisiti per la validità degli strumenti sono:

1. *Rilevanza*

- Almeno uno strumento deve entrare nella controparte della regressione del primo stadio.
- Se gli strumenti sono deboli, allora lo stimatore TSLS è distorto e la statistica *t* ha una distribuzione non normale
- Per verificare strumenti deboli con un singolo regressore endogeno incluso, si verifica la statistica *F* del primo stadio
  - Se $F>10$, gli strumenti sono forti – si usa il TSLS
  - Se $F<10$, gli strumenti sono deboli – si fa qualcosa.

2. *Esogeneità*

- ***Tutti*** gli strumenti devono essere incorrelati con il termine d'errore: corr(Z_{1i},u_i) = 0,…, corr(Z_{mi},u_i) = 0
- Possiamo eseguire una verifica parziale di esogeneità: se $m>1$, possiamo verificare l'ipotesi nulla che tutti gli strumenti siano esogeni contro l'alternativa che almeno $m-1$ siano endogeni (correlati con u)
- Si usa il test J, realizzato usando i residui TSLS.
- Se il J respinge l'ipotesi, allora almeno alcuni degli strumenti sono endogeni, perciò occorre prendere una decisione difficile e scartare alcuni (o tutti) gli strumenti.

Applicazione alla domanda di sigarette (Paragrafo 12.4)

Perché siamo interessati a conoscere l'elasticità della domanda di sigarette?

- Teoria della tassazione ottimale. L'aliquota d'imposta ottimale è inversamente proporzionale all'elasticità al prezzo: maggiore è l'elasticità, minore la quantità influenzata da una data percentuale d'imposta, perciò minore è la variazione di consumo e perdita secca.
- Esternalità del fumo – ruolo dell'intervento pubblico per scoraggiare il fumo
  - effetti di salute del fumo passivo? (non monetari)
  - esternalità monetarie

Dati panel

- Consumo annuo di sigarette, prezzi medi pagati dal consumatore finale (tasse incluse), reddito personale e percentuali d'imposta (specifiche per le sigarette e generali sulle vendite nello stato)
- 48 stati continentali USA, 1985-1995

Strategia di stima

- Dobbiamo usare metodi di stima IV per genstire la distorsione da causalità simultanea che nasce dall'interazione di offerta e domanda.
- Indicatori binari di stato = variabili *W* (variabili di controllo) che controllano per caratteristiche inosservate a livello di stato che influiscono sulla domanda si sigarette e la percentuale d'imposta, purché tali caratteristiche non varino nel tempo.

Modello con effetti fissi della domanda di sigarette

$$\ln(Q_{it}^{cigarettes}) = \alpha_i + \beta_1 \ln(P_{it}^{cigarettes}) + \beta_2 \ln(Income_{it}) + u_{it}$$

- $i = 1,\ldots,48 t = 1985, 1986,\ldots,1995$
- corr($\ln(P_{it}^{cigarettes})$,u_{it}) è verosimilmente diverso da sero a causa di interazioni offerta-domanda
- α_i riflette fattori omessi inosservati che variano tra stati ma non nel tempo, per esempio l'atteggiamento verso il fumo
- Strategia di stima:
  - Usiamo metodi di regressione con dati panel per eliminare α_i
  - Usiamo TSLS per gestire la distorsione da causalità simultanea
  - Usiamo $T = 2$ con variazioni 1985 – 1995 (metodo "prima e dopo") – osserviamo la risposta a lungo termine, non la dinamica di breve termine (elasticità a breve v. lungo termine)

Il metodo "prima e dopo" (quando T=2)

- Un modo per modellare gli effetti a lungo termine è quello di considerare variazioni su 10 anni, tra il 1985 e il 1995
- Riscriviamo la regressione in forma "prima e dopo" :

$$\ln(Q_{i1995}^{cigarettes}) - \ln(Q_{i1985}^{cigarettes})$$

$$= \beta_1[\ln(P_{i1995}^{cigarettes}) - \ln(P_{i1985}^{cigarettes})]$$

$$+\beta_2[\ln(Income_{i1995}) - \ln(Income_{i1985})]$$

$$+ (u_{i1995} - u_{i1985})$$

- Creaiamo variabili di "variazione a 10 anni", per esempio:
- Variazione a 10 anni nel log del prezzo = $\ln(P_{i1995}) - \ln(P_{i1985})$
- Poi stimiamo l'elasticità della domanda mediante TSLS usando variazioni a 10 anni nelle variabili strumentali
- Questo è equivalente a usare i dati originali e includere gli indicatori binari di stato (variabili "*W*") nella regressione

STATA: domanda di sigarette

Prima si creano variabili di "variazione a 10 anni"

10-year change in log price

$$= \ln(P_{it}) - \ln(P_{it-10}) = \ln(P_{it}/P_{it-10})$$

```
. gen dlpackpc = log(packpc/packpc[_n-10]);          _n-10 è il valore a 10 anni
. gen dlavgprs = log(avgprs/avgprs[_n-10]);
. gen dlperinc = log(perinc/perinc[_n-10]);
. gen drtaxs   = rtaxs-rtaxs[_n-10];
. gen drtax    = rtax-rtax[_n-10];
. gen drtaxso = rtaxso-rtaxso[_n-10];
```

Usiamo TSLS per stimare l'elasticità della domanda con la specifica "variazione a 10 anni"

```
                    Y       W          X         Z
. ivregress 2sls dlpackpc dlperinc (dlavgprs = drtaxso) , r;

IV (2SLS) regression with robust standard errors      Number of obs =      48
                                                      F(  2,    45) =   12.31
                                                      Prob > F      =  0.0001
                                                      R-squared     =  0.5499
                                                      Root MSE      =  .09092

------------------------------------------------------------------------------
             |               Robust
    dlpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    dlavgprs |  -.9380143   .2075022    -4.52   0.000    -1.355945   -.5200834
    dlperinc |   .5259693   .3394942     1.55   0.128    -.1578071    1.209746
       _cons |   .2085492   .1302294     1.60   0.116    -.0537463    .4708446
------------------------------------------------------------------------------
Instrumented:  dlavgprs
Instruments:   dlperinc drtaxso
------------------------------------------------------------------------------
```

NOTE:

- *Tutte le variabili – Y, X, W e Z – sono in variazioni a 10 anni*
- *Elasticità stimata = –0,94 (SE = 0,21) – sorprendentemente elastica!*
- *Elasticità del reddito piccola, non statisticamente diversa da zero*
- *Occorre verificare se lo strumento è rilevante…*

Verifica della rilevanza dello strumento: si calcola la statistica *F* del primo stadio

```
.  reg dlavgprs drtaxso dlperinc;

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  2,    45) =   23.86
       Model |  .191437213     2  .095718606           Prob > F      =  0.0000
    Residual |  .180549989    45  .004012222           R-squared     =  0.5146
-------------+------------------------------           Adj R-squared =  0.4931
       Total |  .371987202    47  .007914621           Root MSE      =  .06334
------------------------------------------------------------------------------
    dlavgprs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     drtaxso |   .0254611   .0037374     6.81   0.000     .0179337    .0329885
    dlperinc |  -.2241037   .2119405    -1.06   0.296    -.6509738    .2027664
       _cons |   .5321948    .031249    17.03   0.000     .4692561    .5951334
------------------------------------------------------------------------------

.  test drtaxso;
 ( 1)  drtaxso = 0

       F(  1,    45) =   46.41
            Prob > F =    0.0000
```

Non serviva eseguire "test" qui!
Con strumento m=1, la stat F è il quadrato della stat t:
*6,81*6,81 = 46,41*

F del primo stadio = 46,5 > 10 perciò lo strumento non è debole

Possiamo verificare l'esogeneità dello strumento? ***No****: m = k*

Domanda di sigarette, variazionie a 10 anni – 2 IV

```
                    Y          W         X         Z1     Z2
. ivregress 2sls dlpackpc dlperinc (dlavgprs = drtaxso drtax) , vce(r);

Instrumental variables (2SLS) regression          Number of obs =      48
                                                  Wald chi2(2)  =   45.44
                                                  Prob > chi2   =  0.0000
                                                  R-squared     =  0.5466
                                                  Root MSE      =  .08836

------------------------------------------------------------------------------
             |               Robust
    dlpackpc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    dlavgprs |  -1.202403   .1906896    -6.31   0.000    -1.576148   -.8286588
    dlperinc |   .4620299   .2995177     1.54   0.123    -.1250139    1.049074
       _cons |   .3665388   .1180414     3.11   0.002     .1351819    .5978957
------------------------------------------------------------------------------
Instrumented:  dlavgprs
Instruments:   dlperinc drtaxso drtax
------------------------------------------------------------------------------
```

drtaxso = solo imposta generale sull vendite

drtax = solo imposta specifica sulle sigarette

Elasticità stimata = -1,2, anche più elastica rispetto all'uso della sola imposta generale sulle vendite!

Statistica *F* del primo stadio – entrambi gli strumenti

X Z1 Z2 W

```
. reg dlavgprs drtaxso drtax dlperinc ;

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  3,    44) =   51.36
       Model |  .289359873     3  .096453291            Prob > F      =  0.0000
    Residual |  .082627329    44  .001877894            R-squared     =  0.7779
-------------+------------------------------           Adj R-squared =  0.7627
       Total |  .371987202    47  .007914621            Root MSE      =  .04333

------------------------------------------------------------------------------
    dlavgprs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     drtaxso |    .013457   .0030498     4.41   0.000     .0073106    .0196033
       drtax |   .0075734   .0010488     7.22   0.000     .0054597     .009687
    dlperinc |  -.0289943   .1474923    -0.20   0.845    -.3262455    .2682568
       _cons |   .4919733   .0220923    22.27   0.000     .4474492    .5364973
------------------------------------------------------------------------------

. test drtaxso drtax;

 ( 1)  drtaxso = 0
 ( 2)  drtax = 0
       F(  2,    44) =   75.65        75.65 > 10 perciò gli strumenti non sono deboli
            Prob > F =    0.0000
```

Con m>k, possiamo verificare le restrizioni di sovraidentificazione…

Verifica delle restrizioni di sovraidentificazione

```
. predict e, resid;          Calcola valori predetti per l'ultima regressione
                             (la precedente regressione TSLS)
. reg e drtaxso drtax dlperinc;              Regress e on Z's and W's

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  3,    44) =    1.64
       Model |  .037769176     3  .012589725            Prob > F      =  0.1929
    Residual |  .336952289    44  .007658007            R-squared     =  0.1008
-------------+------------------------------           Adj R-squared =  0.0395
       Total |  .374721465    47  .007972797            Root MSE      =  .08751

------------------------------------------------------------------------------
           e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     drtaxso |   .0127669   .0061587     2.07   0.044      .000355    .0251789
       drtax |  -.0038077   .0021179    -1.80   0.079     -.008076    .0004607
    dlperinc |  -.0934062   .2978459    -0.31   0.755    -.6936752    .5068627
       _cons |    .002939   .0446131     0.07   0.948    -.0869728    .0928509
------------------------------------------------------------------------------
. test drtaxso drtax;
 ( 1)  drtaxso = 0                  Calcola la statistica J, che è m*F,
 ( 2)  drtax = 0                    dove F verifica se i coefficienti
                                    degli strumenti sono zero
       F(  2,    44) =    2.47      perciò J = 2 × 2.47 = 4.93
            Prob > F =    0.0966              ** ATTENZIONE – usa la f.d. sbagliata **
```

I gradi di libertà corretti per la statistica J sono $m-k$:

- $J = mF$, dove F = la statistica F che verifica i coefficienti di $Z_{1i},\ldots,Z_{mi}$ in una regressione dei residui TSLS rispetto a $Z_{1i},\ldots,Z_{mi}$, $W_{1i},\ldots,W_{mi}$.
- Sotto l'ipotesi nulla che tutti gli strumenti siano esogeni, J ha una distribuzione chi-quadro con $m-k$ gradi di libertà
- Qui J = 4,93, distribuzione chi-quadro con f.d. = 1; il valore critico al 5% è 3,84, perciò respinge al livello di significatività del 5%.
- In STATA:

```
.  dis "J-stat = " r(df)*r(F) "  p-value = "  chiprob(r(df)-1,r(df)*r(F));
J-stat = 4.9319853  p-value = .02636401
```

J = 2 × 2.47 = 4.93 *valore-p* da distribuzione chi-quadro(1)

E ora???

Riepilogo dei risultati in forma di tabella

Tabella 12.1 Stime TSLS del consumo di sigarette su dati panel per 48 stati USA.

Variabile dipendente: $\ln(Q_{i,1995}^{cigarettes}) - \ln(Q_{i,1985}^{cigarettes})$

Regressore	(1)	(2)	(3)
$\ln(P_{i,1995}^{cigarettes}) - \ln(P_{i,1985}^{cigarettes})$	– 0,94** (0,21)	– 1,34** (0,23)	– 1,20** (0,20)
$\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})$	0,53 (0,34)	0,43 (0,30)	0,46 (0,31)
Intercetta	– 0,12 (0,07)	– 0,02 (0,07)	– 0,05 (0,06)
Variabili strumentali	Sales tax	Sigarette-specific tax	Both sales tax and sigarette-specific tax
Statistica *F* first stage	33,70	107,20	88,60
Restrizioni di sovraidentificazione test *J* e *p*-value			4,93 (0,026)

Queste regressioni sono state stimate sui dati per 48 stati USA (48 osservazioni su differenze decennali). I dati sono descritti nell'Appendice 12.1. Il test-*J* per le restrizioni di sovraidentificazione è descritto nel Concetto chiave 12.6 (il suo valore-*p* è riportato tra parentesi) e la statistica *F* per il primo stadio è descritta nel Concetto chiave 12.5. I coefficienti sono statisticamente significativi al livello *5% o **1%.

Come dobbiamo interpretare il rifiuto del test *J*?

- Il test *J* rifiuta l'ipotesi nulla che entrambi gli strumenti siano esogeni
- Questo significa che o *rtaxso* è endogeno, o *rtax* è endogeno, o entrambi!
- Il test *J* non ci dice quale! *Occorre ragionare…*
- Perché *rtax* (imposta sulle sigarette) potrebbe essere endogeno?
  - Forze politiche: tradizione del fumare o molti fumatori -> pressione politica per basse imposte sulle sigarette
  - In questo caso, l'imposta sulle sole sigarette è endogena
- Questo ragionamento non vale per l'imposta generale sulle vendite
- → usiamo solo uno strumento, l'imposta generale sulle vendite

La domanda di sigarette: riepilogo di risultati empirici

- Usiamo l'elasticità stimata in base al TSLS con l'imposta generale sulle vendite come unico strumento:

Elasticità = -0,94, *SE* = 0,21

- Questa elasticità è sorprendentemente elevata (non anelastica) – un incremento dell'1% nei prezzi riduce le vendite di sigarette di quasi l'1%. È un'elasticità molto maggiore di quanto si pensi comunemente nella letteratura sull'economia e la salute.
- È un'elasticità di lungo periodo (variazione a 10 anni). *Che cosa vi aspettate riguardo l'elasticità di breve periodo (variazione a un anno), sarà maggiore o minore?*

Valutazione della validità dello studio

Altre minacce alla validità interna?

1. Distorsione da variabili omesse?
   - *Lo stimatore con effetti fissi controlla per fattori inosservati che variano tra stati ma non nel tempo*
2. Errata specificazione delal forma funzionale? (*si potrebbe verificare*)
3. Altra distorsione da causalità simultanea?
   - *Non se l'imposta generale sulle vendite è uno strumento valido, una volta inclusi gli effetti fissi di stato!*
4. Distorsione da errori nelle variabili?
5. Distorsione da selezione campionaria? (*no, abbiamo tutti gli stati*)
6. Altre minacce alla validità interna degli studi di regressione IV riguardano il dubbio che lo strumento sia (1) rilevante e (2) esogeno. *Quanto sono significative queste minacce nell'applicazione dell'elasticità della domanda di sigarette?*

Valutazione della validità dello studio (continua)

Validità esterna?

- Abbiamo stimato un'elasticità nel lungo periodo, possiamo generalizzarla al breve periodo? Perché, o perché no?
- Supponiamo di voler usare l'elasticità stimata di -0,94 per orientare la politica odierna. Riportiamo due variazioni a partire dal periodo coeprto dai dati (1985-95) – queste variazioni pongono una minaccia alla validità esterna (generalizzazione dal periodo 1985-95 a oggi)?
  - Oggi si fuma meno che nel 1985-1995
  - L'atteggiamento culturale verso il fumo è cambiato in senso negativo dal 1985-95.

Dove trovare strumenti validi? (Paragrafo 12.5)

Note generali

Nell'analisi IV il difficile è trovare strumenti validi

- Metodo 1: "variabili in un'altra equazione" (per es. fattori di spostamento dell'offerta che non hanno effetto sulla domanda)
- Metodo 2: cercare una variazione esogena (Z) che sia "come se" assegnata casualmente (non influisce direttamente su Y) ma influisca su X.
- Questi sono due modi diversi di pensare agli stessi problemi – vedere il collegamento…
  - La pioggia sposta la curva di offerta del burro ma non la curva di domanda. È "come se" assegnata casualmente
  - L'imposta sulle vendite sposta la curva di offerta per le sigarette ma non la curva di domanda; le imposte sulle vendite sono "come se" assegnate casualmente

Esempio: cateterizzazione cardiaca

McClellan, Mark, Barbara J. McNeil, and Joseph P. Newhouse (1994), "Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality?" *Journal of the American Medical Association*, vol. 272, no. 11, pp. 859 – 866.

La cateterizzazione cardiaca migliora la longevità dei pazienti affetti da attacco cardiaco?

Y_i = sopravvivenza (in giorni) di pazienti colpiti da attacco cardiaco

X_i = 1 se il paziente riceve la cateterizzazione cardiaca,

= 0 altrimenti

- Trial clinici mostrano che *CardCath* influisce su *SurvivalDays*.
- Ma il trattamento è efficace "sul campo"?

Cateterizzazione cardiaca (continua)

$$SurvivalDays_i = \beta_0 + \beta_1 CardCath_i + u_i$$

- L'OLS è non distorto? La decisione di trattare un paziente con cateterizzazione cardiaca è endogena – è (*stata*) presa sul campo dal soccorritore e dipende da u_i (caratteristiche di salute del paziente inosservate)
- Se i pazienti più sani sono cateterizzati, allora l'OLS è affetto da distorsione da causalità simultanea e sovrastiva l'effetto del trattamento
- Strumento proposto: distanza del più vicino ospedale dotato di cateterizzazione cardiaca meno distanza del più vicino ospedale "normale"

Cateterizzazione cardiaca (continua)

- Z = distanza differenziale dall'ospedale dotato di CC
  - Rilevante? Se un ospedale dotato di CC è lontano, il paziente non vi sarà portato e non sarà trattato con CC
  - Esogena? Se la distanza dell'ospedale dotato di CC non influisce sulla sopravvivenza, se non per l'effetto su $CardCath_i$, allora corr(distance,u_i) = 0 perciò è esogena
  - Se la posizione del paziente è casuale, allora la distanza differenziale è "come se" fosse assegnata casualmente.
  - *Il primo stadio è un modello di probabilità lineare: la distanza influisce sulla probabilità di ricevere il trattamento*
- Risultati:
  - OLS stima un significativo e ampio effetto della CC
  - TSLS stima un effetti piccolo, spesso insignificante

Esempio: crowding out (spiazzamento) della spesa privata in beneficenza

Gruber, Jonathan and Daniel M. Hungerman (2005), "Faith-Based Charity and Crowd Out During the Great Depression," NBER Working Paper 11332.

La spesa sociale pubblica spiazza la spesa di beneficenza privata (chiesa, croce rossa, ecc.)?

Y = spesa di beneficenza privata (chiese)

X = spesa pubblica

Qual è il motivo per usare variabili strumentali?

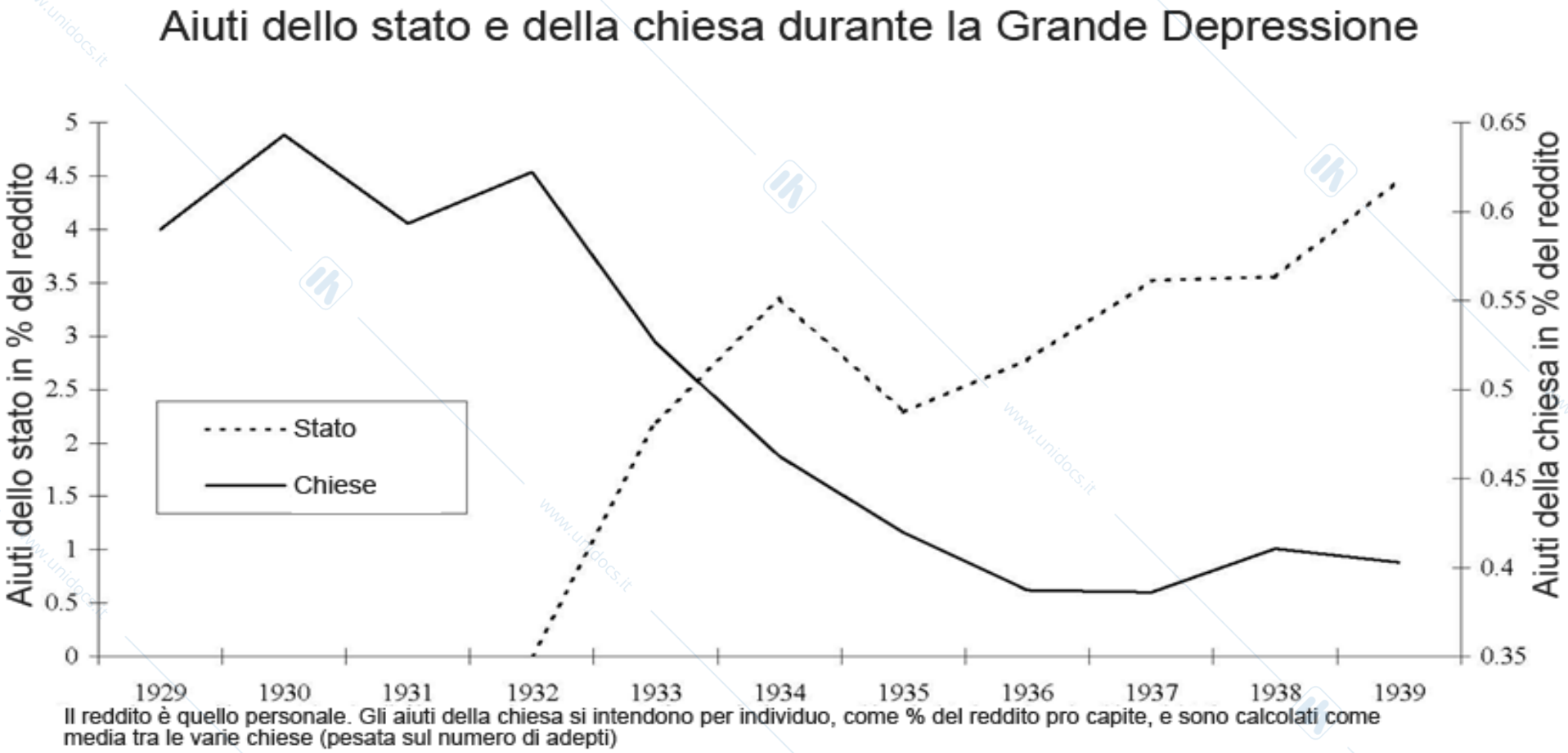
Strumento proposto:

Z = forza della delega congressuale

Spesa di beneficenza privata (continua). I dati – alcuni dettagli

- Dati panel, annui, per stato, 1929-1939, U.S.A.
- *Y* = totale spesa beneficenza da parte di sei chiese (CCC, Lutheran, Northern Baptist, Presbyterian (2), Southern Baptist); beneficenza = ¼ della spesa totale delle chiese.
- *X* = spesa sociale federale sotto il New Deal (General Relief, Work Relief, Civil Works Administration, Aid to Dependent Children,…)
- *Z* = durata mandato dei rappresentanti dello stato al House & Senate Appropriations Committees, in mesi
- *W* = insiemi di effetti fissi

Spesa di beneficenza privata (continua)



Il reddito è quello personale. Gli aiuti della chiesa si intendono per individuo, come % del reddito pro capite, e sono calcolati come media tra le varie chiese (pesata sul numero di adepti)

Spesa di beneficenza privata (continua)

***Valutazione di validità*:**

- Validità strumento:
  - Rilevanza?
  - Esogeneità?
- Altre minacce alla validità interna:

1. Distorsione da variabili omesse
2. Forma funzionale
3. Errore di misura
4. Selezione del campione
5. Causalità simultanea

- Validità esterna a oggi negli USA? Per aiuti ai paesi in via di sviluppo?

Esempio: concorrenza tra scuole

Hoxby, Caroline M. (2000), "Does Competition Among Public Schools Benefit Students and Taxpayers?" *American Economic Review* 90, 1209-1238

Qual è l'effetto della concorrenza tra scuole pubbliche sulla performance degli studenti?

Y = punteggi nei test al 12-esimo livello

X = misura di scelta tra distretti scolastici (funzione del numero di distretti nell'area)

Qual è la motivazione per usare variabili strumentali?

Strumento proposto:

Z = numero di piccoli corsi d'acqua nell'area

Concorrenza tra scuole (continua)

Dati – alcuni dettagli

- dati sezionali, USA, area metropolitana, fine anni Novanta (n = 316),
- Y = voto al 12-esimo livello scolastico (anche altre misure)
- X = indice preso da una revisione della letteratura di settore misurando il livello di concorrenza (“indice di Gini”) – in base al numero di “imprese” e alla “quota di mercato”
- Z = misura del numero di piccoli corsi d’acqua – che hanno formato confini geografici naturali.
- W = insiemi di variabili di controllo

Concorrenza tra scuole (continua)

***Valutazione di validità*:**

- Validità strumento:
  - Rilevanza?
  - Esogeneità?
- Altre minacce alla validità interna:

1. Distorsione da variabili omesse
2. Forma funzionale
3. Errore di misura
4. Selezione del campione
5. Causalità simultanea

- Validità esterna a oggi negli USA?

Conclusioni (Paragrafo 12.6)

- Uno strumento valido ci consente di isolare una parte di *X* che è incorrelata con *u*, e quella parte può essere usata per stimare l'effetto su *Y* di una variazione in *X*
- La regressione IV richiede strumenti validi:

1. *Rilevanza*: verifica tramite statistica *F* del primo stadio
2. *Esogeneità*: verifica di restrizioni di sovraidentificazione tramite la statistica *J*

- Uno strumento valido isola la variazione in *X* che è "come se" assegnata casualmente.
- Il requisito fondamentale di almeno *m* strumenti validi non può essere verificato – *occorre usare la testa.*

Domande e risposte sulla regressione IV

1. Quando usare la regressione IV?

Ogni volta che *X* è correlata con *u* e si ha uno strumento valido. I motivi principali per la correlazione tra *X* e *u* potrebbero essere:

- Variabili omesse che portano a distorsione
  - Esempio: distorsione da talento nel rendimento dell'istruzione
- Errore di misura
  - Esempio: errore di misura negli anni di istruzione
- Distorsione da selezione del campione
  - I pazienti scelgono il trattamento
- Distorsione da causalità simultanea
  - Esempio: offerta e domanda di burro, sigarette

2. Quali sono le minacce alla validità interna di una regressione IV?

- La minaccia principale alla validità interna di una regressione IV è la non validità dell'assunzione di strumenti validi. Dato un insieme di variebili di controllo *W*, gli strumenti sono validi se sono relativi ed esogeni.
  - La rilevanza può essere valutata verificando se gli strumenti sono deboli o forti: la statistica *F* del primo stadio è > 10?
  - L'esogeneità può essere verificata usando la statistica *J* – purché si abbiano *m* strumenti esogeni con cui partire! In generale, l'esogeneità deve essere valutata basandosi su una conoscenza approfondita dell'applicazione considerata.