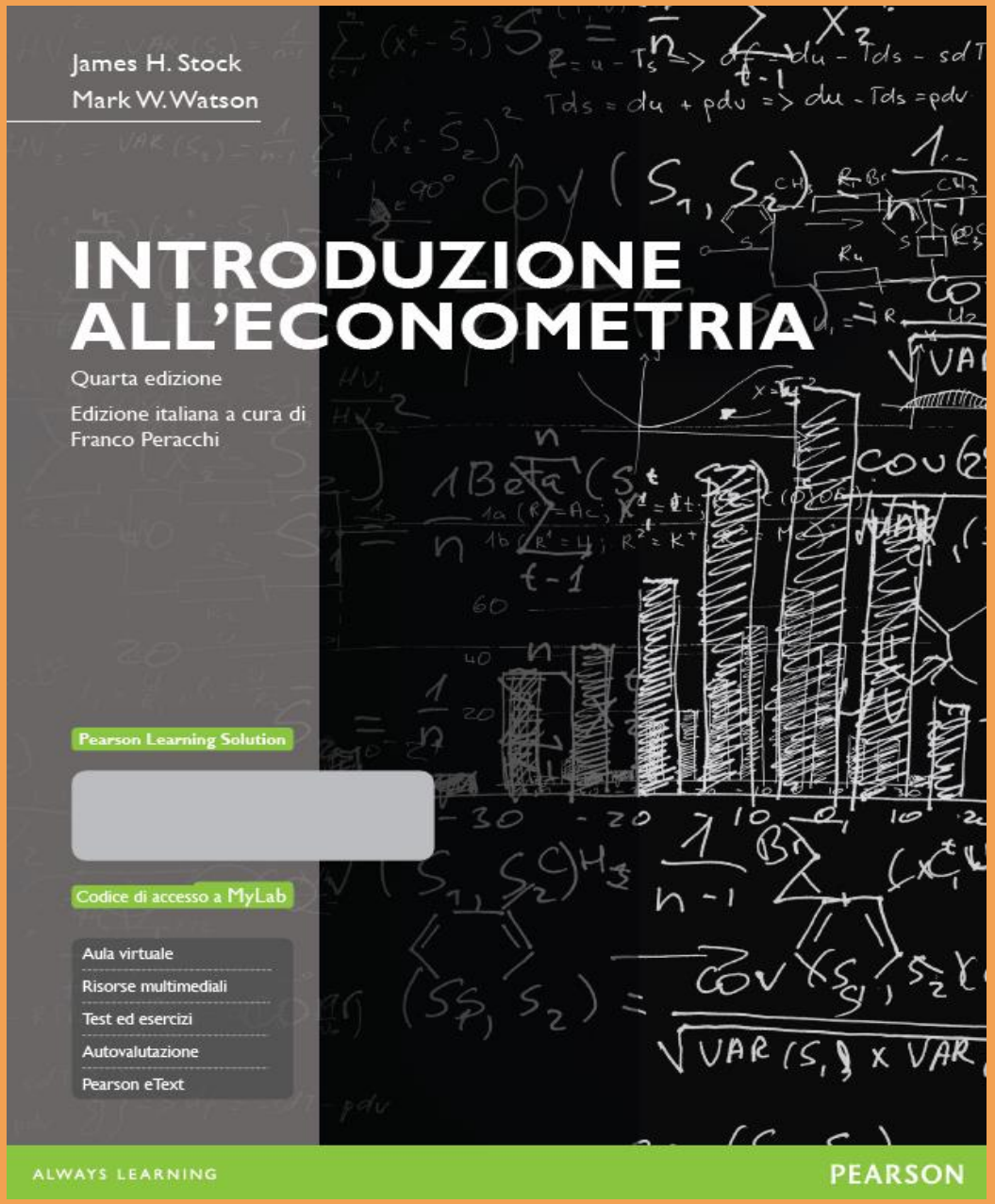
Capitolo 9

Valutazione di studi basati sulla regressione multipla



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.

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 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 a 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;
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

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

$$= \beta_0 + \beta_1 X_i + [\beta_1(X_i - X_i) + u_i]$$

Per cui la regressione svolta è,

$$Y_i = \beta_0 + \beta_1 X_i + u_i \text{ dove } u_i = \beta_1(X_i - X_i) + u_i$$

Con l'errore di misura, tipicamente X_i è correlata con u_i quindi risulta $\hat{\beta}_1$ distorta:

$$\text{cov}(X_i, u_i) = \text{cov}(X_i, \beta_1(X_i - X_i) + u_i)$$

$$= \beta_1 \text{cov}(X_i, X_i - X_i) + \text{cov}(X_i, u_i)$$

Spesso è plausibile che cov(X_i, u_i) = 0 (se $E(u_i|X_i) = 0$ allora cov(X_i, u_i) = 0 se l'errore di misurazione in X_i è incorrelato con u_i).

Ma tipicamente cov(X_i, $X_i - X_i$) ≠ 0.

Distorsione da errori nelle variabili (continua)

$Y_i = \beta_0 + \beta_1 X_i + u_i$ dove $u_i = \beta_1(X_i - X_i) + u_i$

$$\begin{aligned}\text{cov}(X_i, u_i) &= \text{cov}(X_i, \beta_1(X_i - X_i) + u_i) \\ &= \beta_1 \text{cov}(X_i, X_i - X_i) + \text{cov}(X_i, u_i) \\ &= \beta_1 \text{cov}(X_i, X_i - X_i) \text{ poiché } \text{cov}(X_i, u_i) = 0\end{aligned}$$

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

$$\tilde{X}_i = X_i + v_i,$$

dove v_i è rumore casuale a media zero con corr(X_i, v_i) = 0 e corr(u_i, v_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

$\tilde{X}_i = X_i + v_i$, dove corr(X_i, v_i) = 0 e corr(u_i, v_i) = 0.

Quindi var($\tilde{X}_i$) = $\sigma^2_X + \sigma^2_v$

cov($\tilde{X}_i, X_i$ – $\tilde{X}_i$) = cov($X_i + v_i$, $-v_i$) = – σ^2_v

così

cov($\tilde{X}_i$, u_i) = $-\beta_1 \sigma^2_v$

così $\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1 \frac{\sigma^2_v}{\sigma^2_{\tilde{X}}} = \left(1 - \frac{\sigma^2_v}{\sigma^2_{\tilde{X}}}\right)\beta_1$

$= \left(\frac{\sigma^2_{\tilde{X}} - \sigma^2_v}{\sigma^2_{\tilde{X}}}\right)\beta_1 \neq \beta_1 \quad \left(\frac{\sigma^2_X}{\sigma^2_X + \sigma^2_v}\right)\beta_1$

Quindi $\hat{\beta}_1$ è distorto verso zero.

Il modello di errore di misura classico è speciale perché assume che corr(X_i, v_i) = 0.

B. Errore di misura "migliore ipotesi"

Si supponga che l'intervistato non ricordi X_i, ma faccia un'ipotesi del tipo $\hat{X}_i = E(X_i|W_i)$, dove $E(u_i|W_i) = 0$. Allora,

$$\text{cov}(\hat{X}_i, u_i) = \text{cov}(\hat{X}_i, \beta_1(X_i - \hat{X}_i) + u_i)$$
$$= \beta_1\text{cov}(\hat{X}_i, X_i - \hat{X}_i) + \text{cov}(\hat{X}_i, u_i)$$

- $\text{cov}(\hat{X}_i, X_i - \hat{X}_i) = 0$ perché $\hat{X}_i = E(X_i|W_i)$ (siccome $\hat{X}_i$ è la migliore ipotesi, l'errore $X_i - \hat{X}_i$ è incorrelato con $\hat{X}_i$).
- $\text{Cov}(\hat{X}_i, u_i) = 0$ perché $E(u_i|W_i) = 0$ ($\hat{X}_i$ è funzione di W_i e $E(u_i|W_i) = 0$).
- Così $\text{cov}(\hat{X}_i, u_i) = 0$, 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 qui 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 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.

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 è non distorto $\hat{\beta}_1$

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 nell' 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?

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: come risultato di un processo politico anch'essi avranno un basso *STR*
- Che significato ha tutto ciò per una regressione di *TestScore* su *STR*?

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*).

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

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

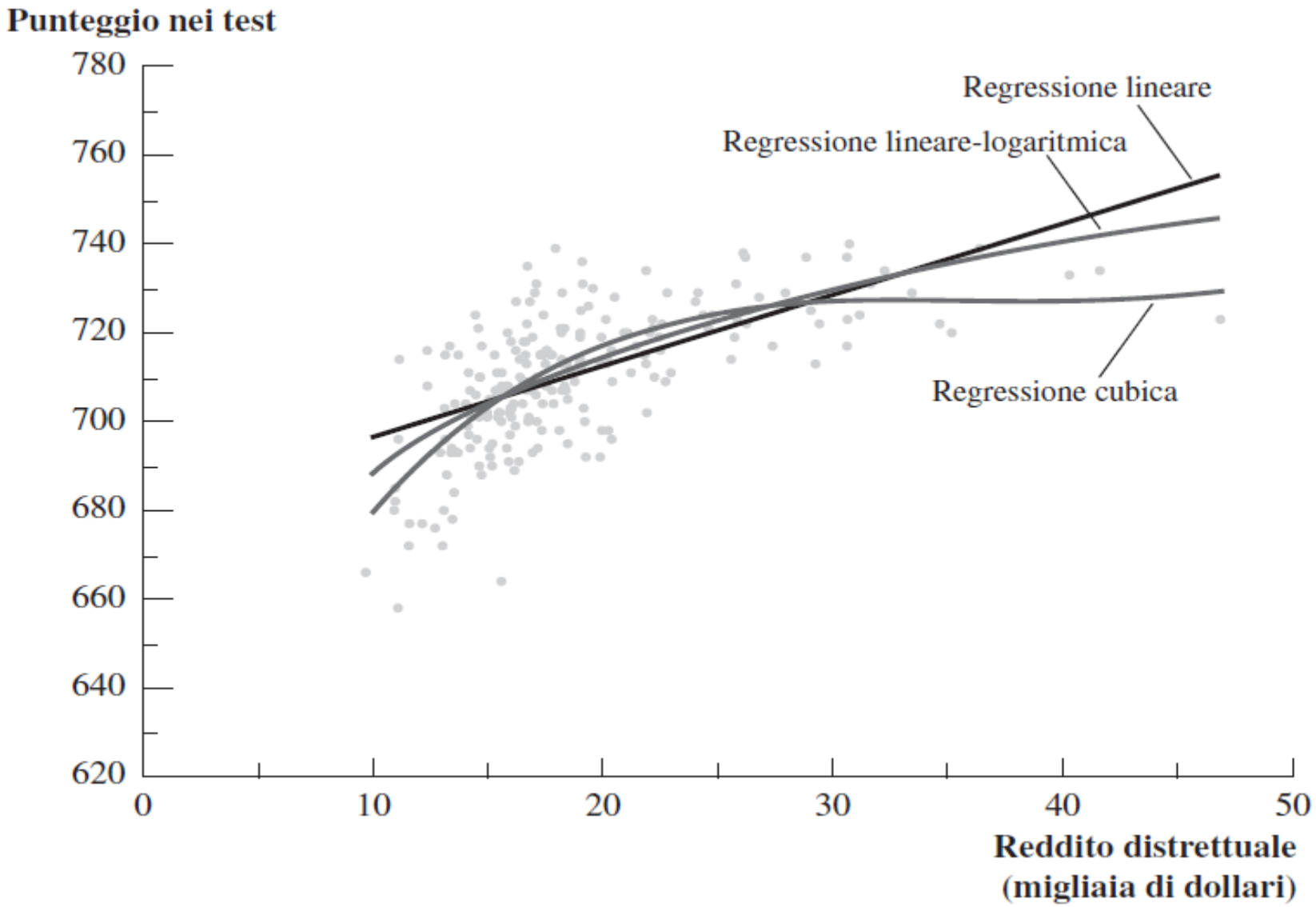
Il gruppo di dati del Massachusetts

- 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:

$$\widehat{TestScore} = 744{,}0 - 0{,}64STR - 0{,}437PctEL - 0{,}582LunchPct$$
$$(21{,}3)\ \ (0{,}27)\qquad (0{,}303)\qquad (0{,}097)$$

$$- 3{,}07Income + 0{,}164Income^2 - 0{,}0022Income^3$$
$$(2{,}35)\qquad (0{,}085\qquad (0{,}0010)$$

- 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?