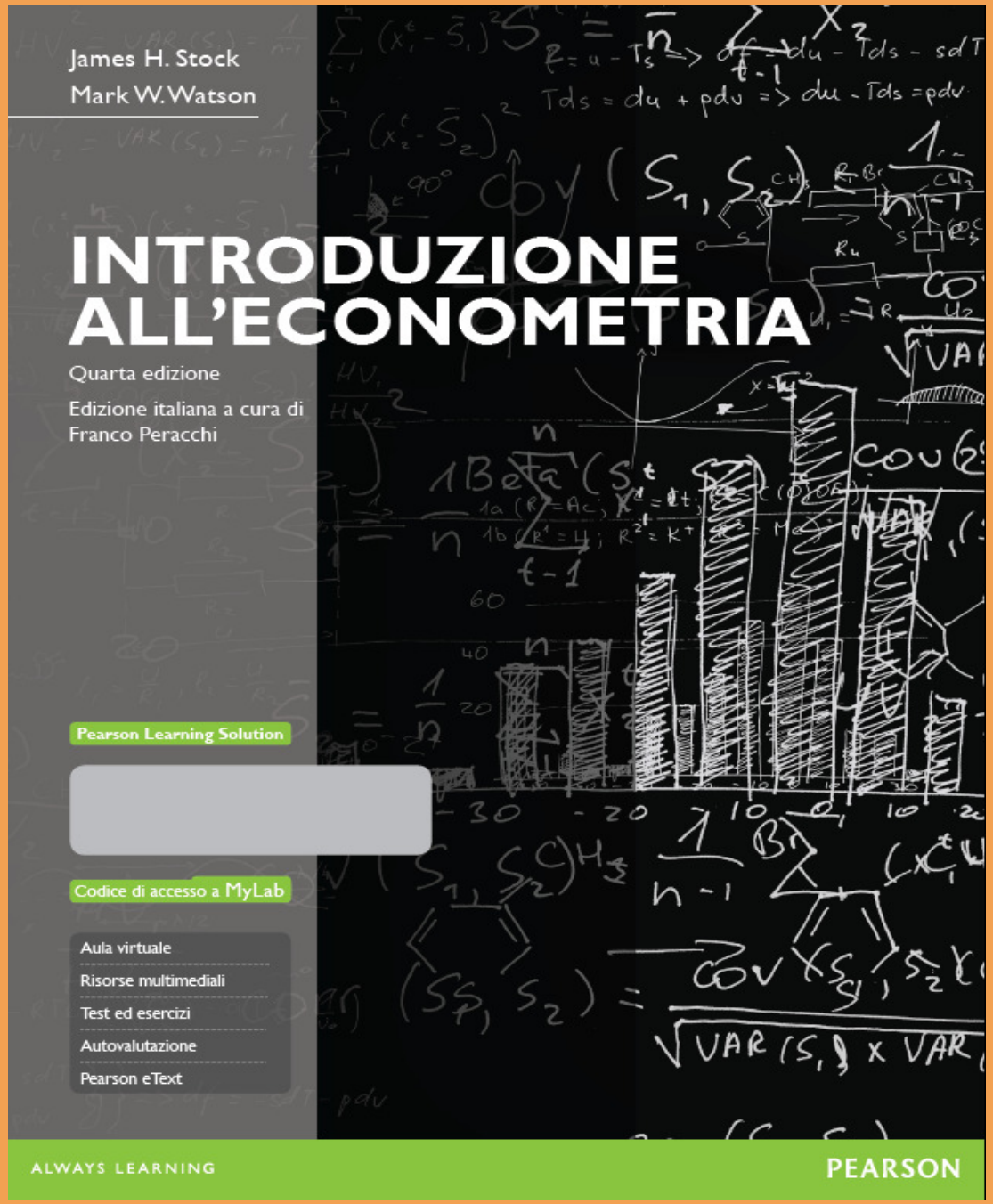
Capitolo 4

Regressione lineare con un singolo regressore



© 2016 Pearson Italia – Milano, Torino

Sommario

1. Il modello di regressione lineare
2. Stima dei coefficienti del modello di regressione lineare
3. Misure di bontà dell'adattamento
4. Le assunzioni dei minimi quadrati
5. Distribuzione campionaria degli stimatori OLS
6. Conclusioni

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*: ossia di quanto varia **in media Y** se X varia di +1.(effetto causale su Y di un var. 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*.

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* nella popolazione

$E[TestScore|STR] = \beta_0 + \beta_1 STR$

β_1 = pendenza della retta di regressione

$$= \frac{\Delta E[Testscore|STR]}{\Delta STR}$$

= variazione nella media del punteggio nei test per una variazione unitaria in *STR*

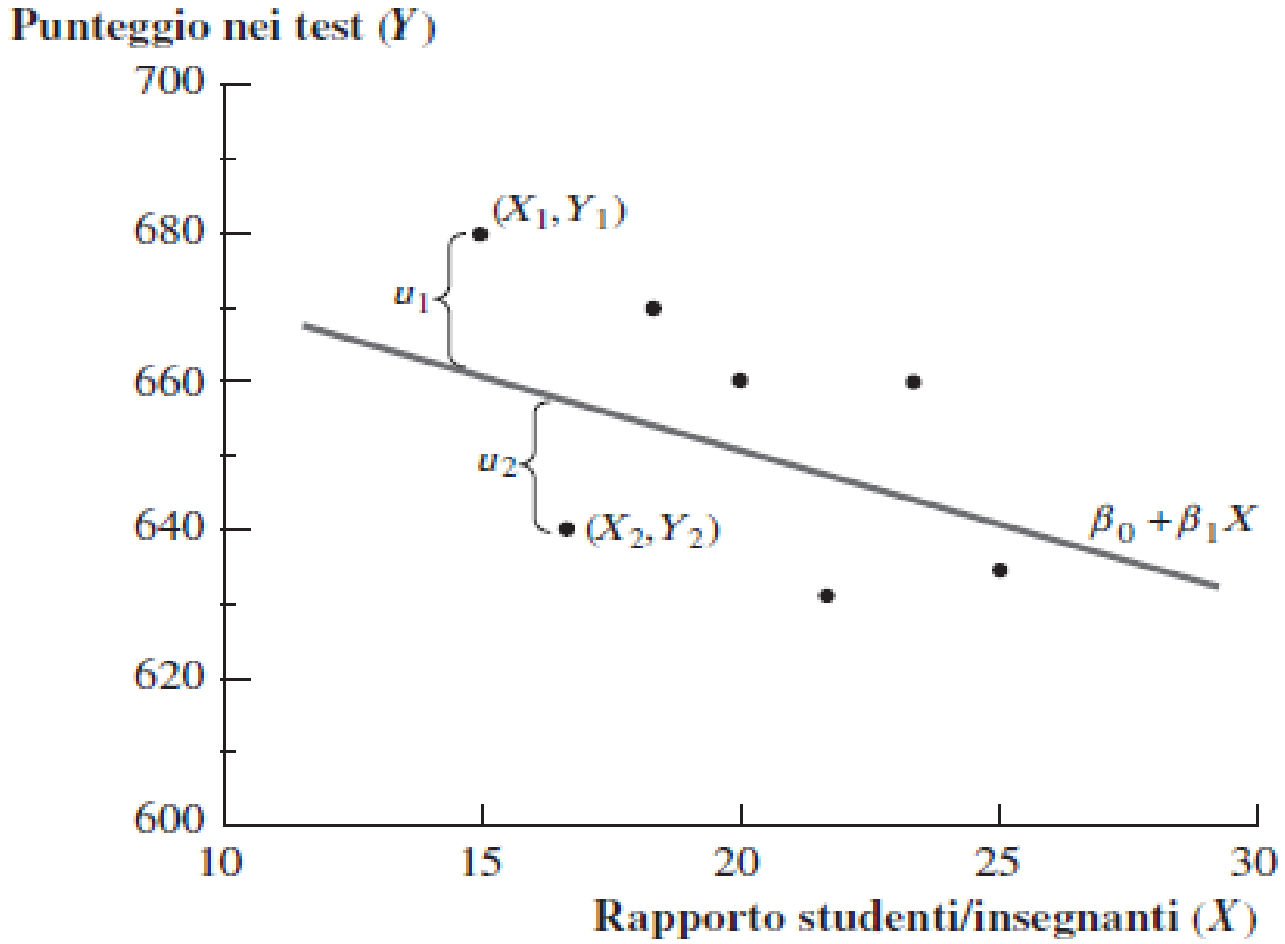
- *Perché β_0 e β_1 sono parametri di "popolazione"*?
- Vorremmo conoscere il valore di β_1.
- Non conosciamo β_1, perciò dobbiamo stimarlo utilizzando i dati.

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 dagli effetti dei fattori omessi su Y. 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 dati?

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

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

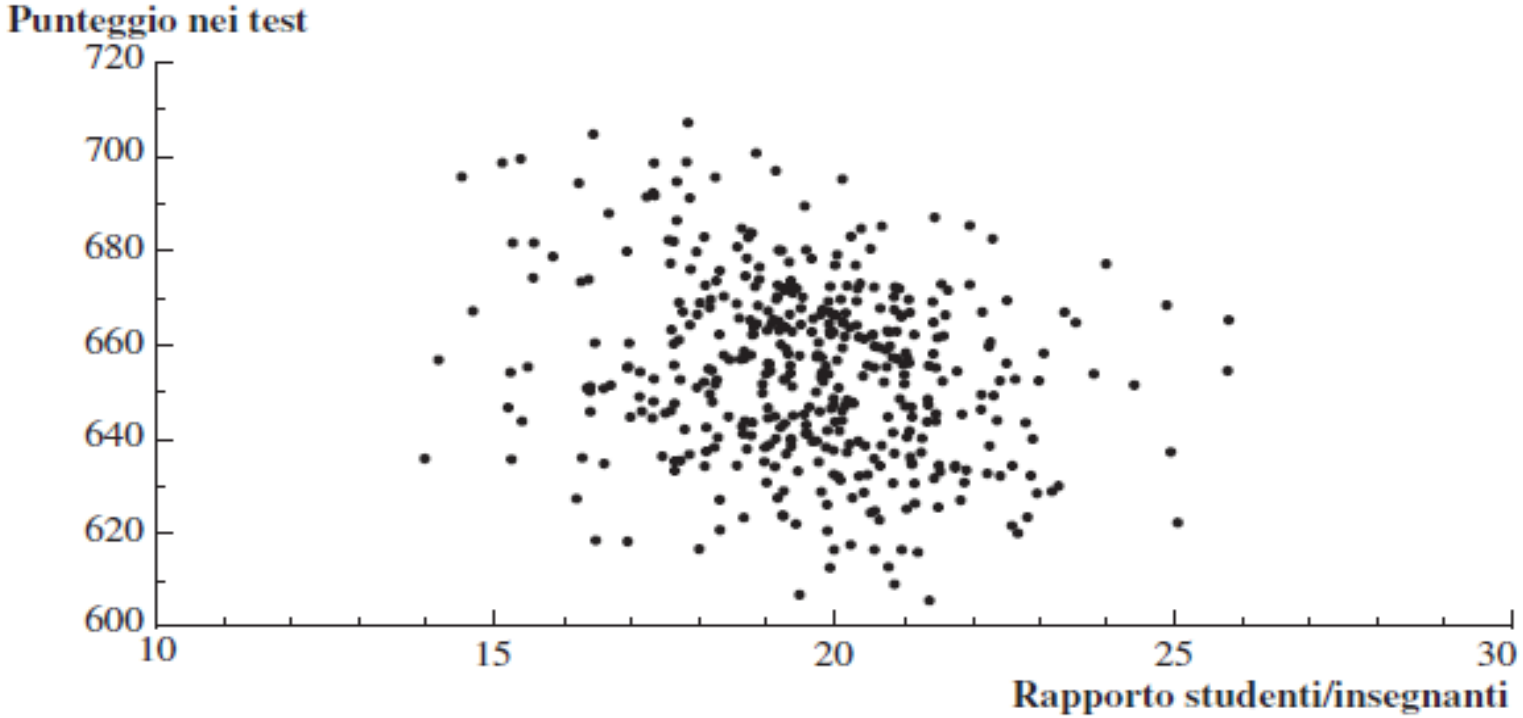
Per analogia, **ci concentreremo sullo stimatore dei minimi quadrati (OLS, "*ordinary least squares*") dei parametri ignoti β_0 e β_1.** Lo stimatore OLS è dato da

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

Meccanismo dei minimi quadrati ordinari

La retta di regressione: E[*TestScore*|*STR*] = $\beta_0 + \beta_1 STR$

$$\beta_1 = \frac{\Delta E[Testscore \mid STR]}{\Delta STR} = ??$$



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:

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

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

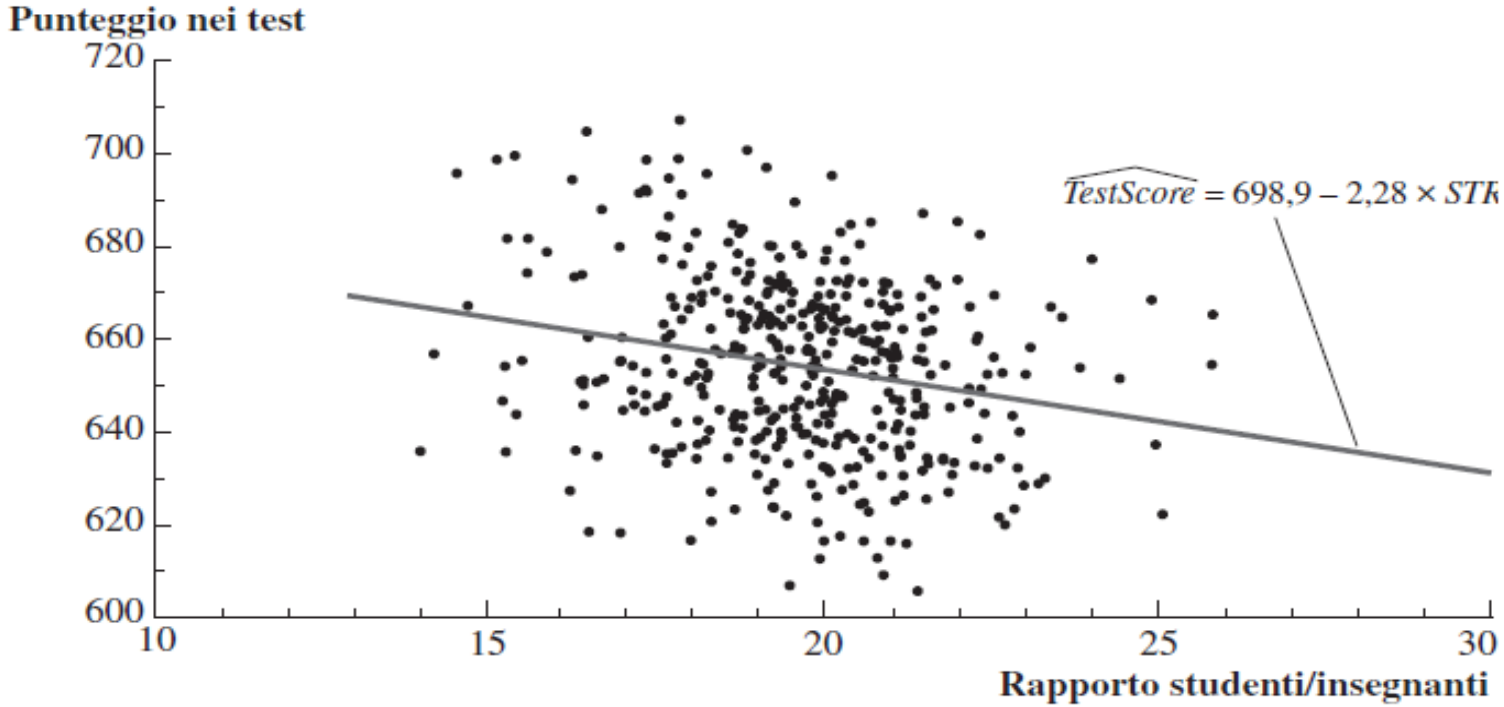
I valori predetti $\hat{Y}_i$ e i residui $\hat{u}_i$ sono:

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

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

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*

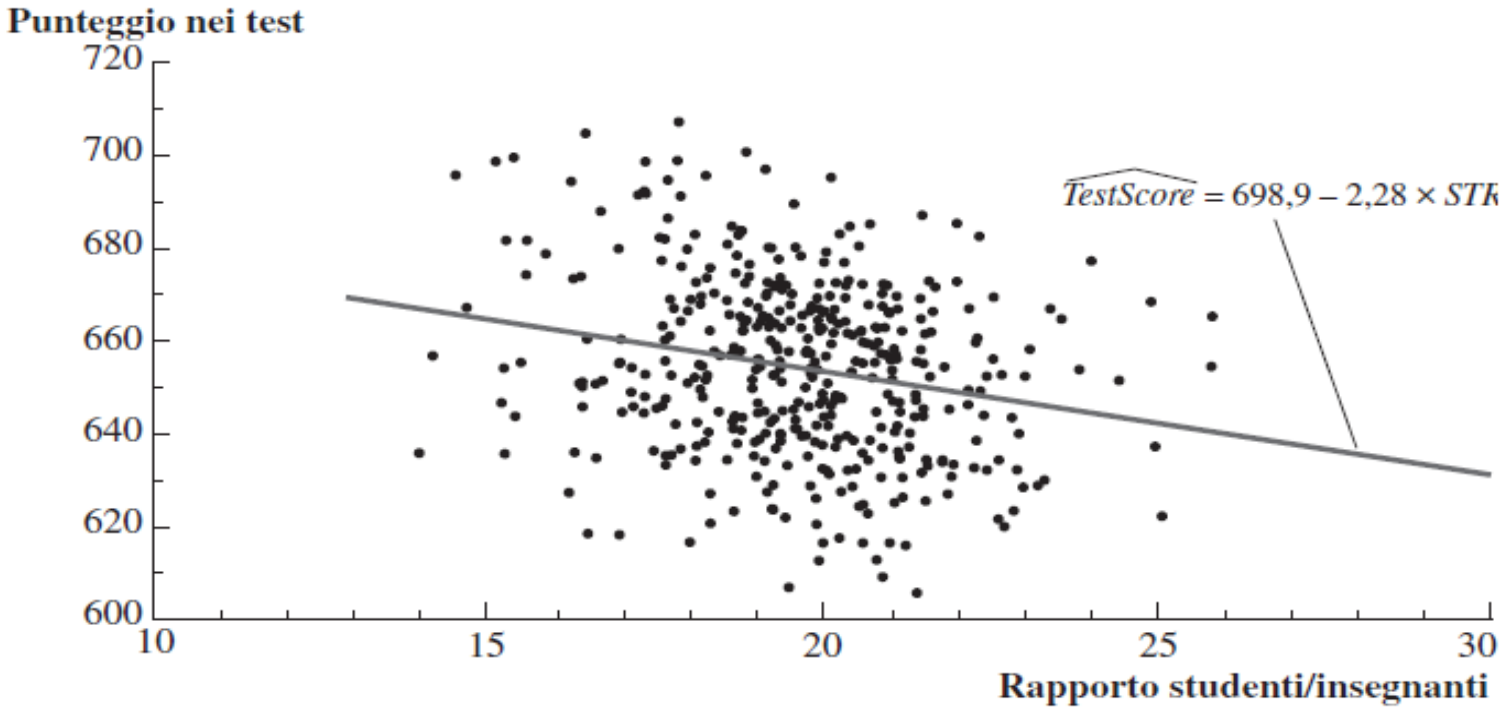


- Pendenza stimata = $\hat{\beta}_1$ = – 2,28
- Intercetta stimata = $\hat{\beta}_0$ = 698,9
- Retta di regressione stimata: $\widehat{TestScore} = 698{,}9 - 2{,}28 \times STR$

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 **la stima** di $\frac{\Delta E[Testscore \mid STR]}{\Delta STR} = -2{,}28$
- 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 bontà dell'adattamento della regressione ai dati:

- L'***R*² della regressione** misura la frazione della varianza (devianza) di *Y* spiegata da *X*; è priva di unità e può variare tra zero (nessun adattamento) e uno (adattamento perfetto)

- L'***errore standard della regressione* (*SER*)** misura la dimensione di un tipico residuo di regressione nelle unità (di misura) 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(SS) Totali = SS "spiegata" + 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 $ESS = 0$
- $R^2 = 1$ 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 dei residui. È (quasi) la deviazione standard campionaria dei residui OLS:

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

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

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

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

Il *SER*:

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

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

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

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

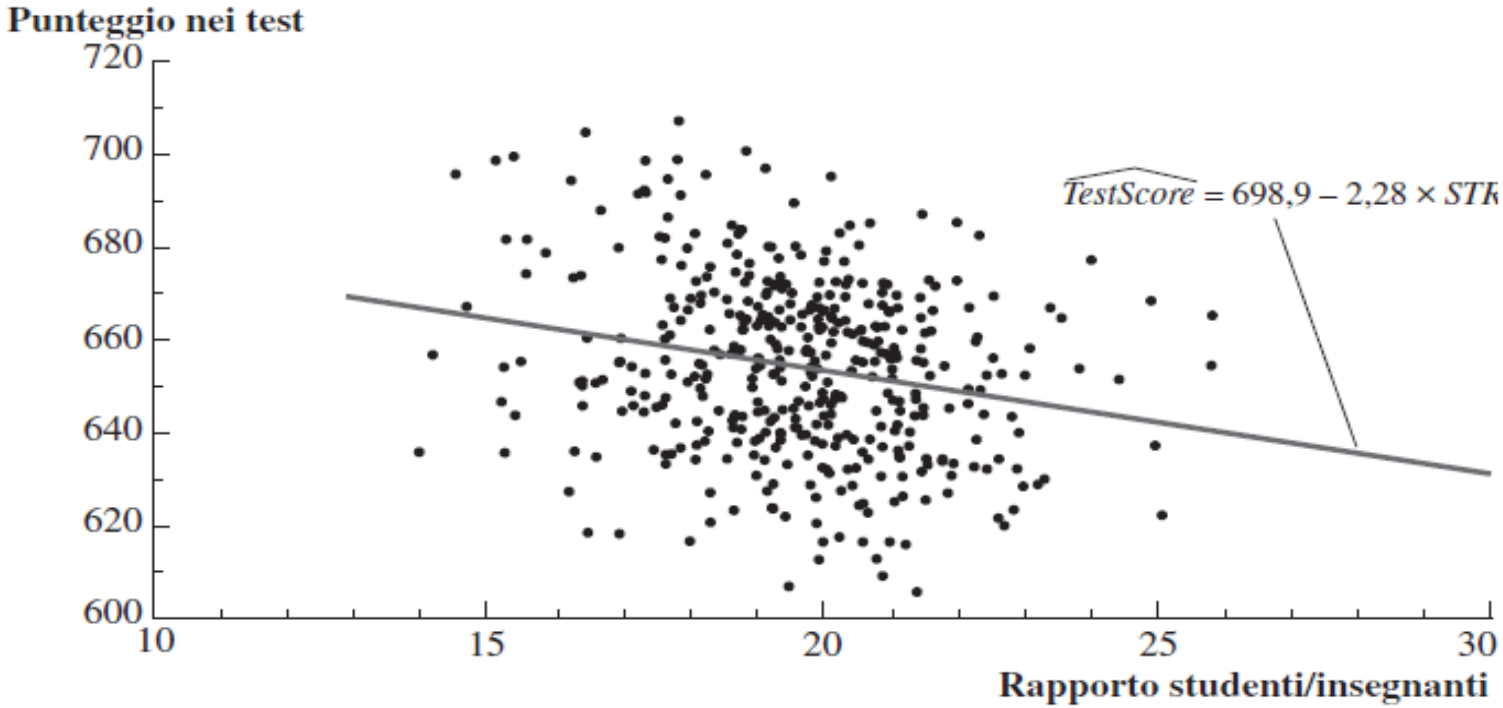
Misura la stessa cosa del *SER* – la differenza sta nel fattore 1/*n* anziché 1/(*n*–2).

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 $\overline{Y}$).
- 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 ottenuti i dati (lo schema di campionamento)

- Queste assunzioni – sono tre – sono note come assunzioni dei minimi quadrati (o meglio, del **modello di regressione lineare semplice**)

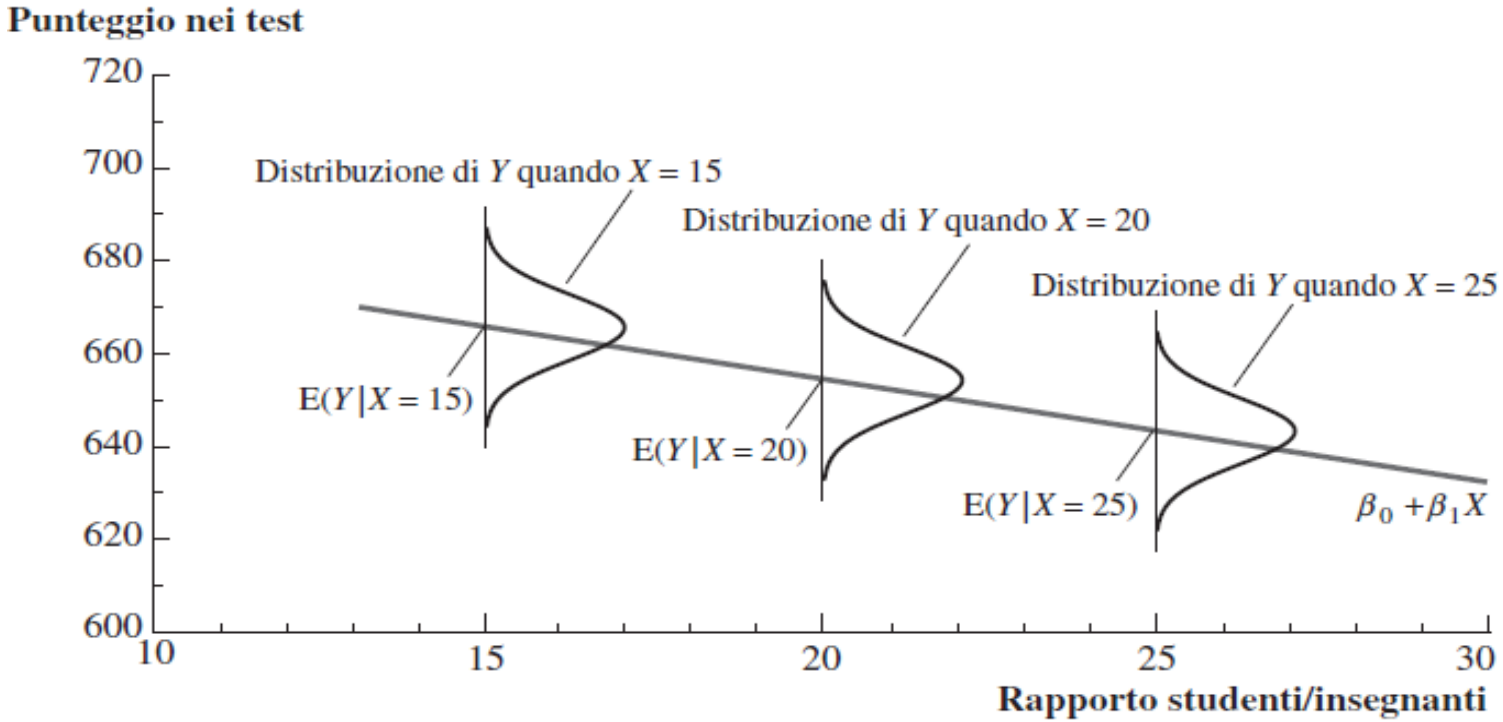
Le assunzioni dei minimi quadrati

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

1. La distribuzione di u condizionata a X ha media nulla, cioè $E(u|X = x) = 0$ per ogni x.
   - *Questo implica, insieme all'ass. 2, 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 semplice*
   - *Questo fornisce la distribuzione camp. di* $\hat{\beta}_0$ *e* $\hat{\beta}_1$ *in grandi campioni*
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?

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:

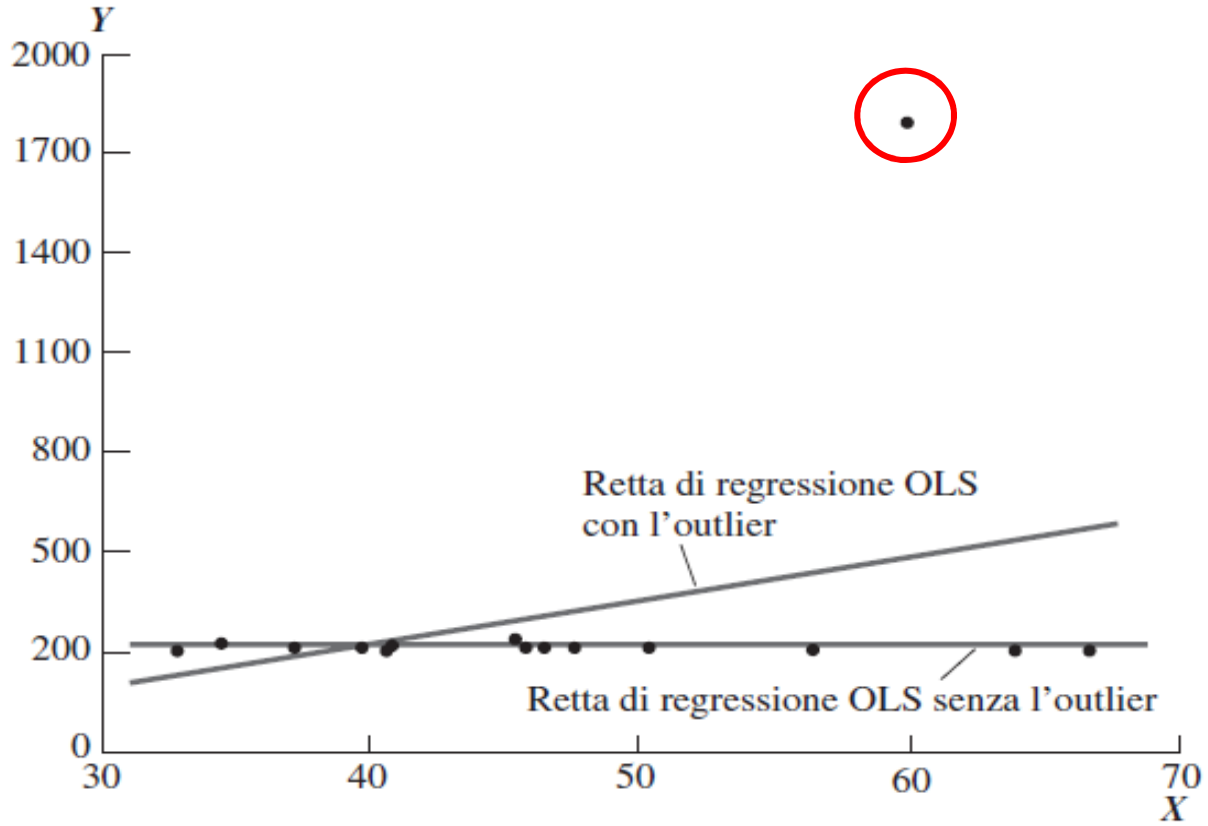
- Le unità sono scelte dalla stessa popolazione, perciò (X_i, Y_i) sono *identicamente distribuite* per ogni $i = 1,\ldots, n$.
- Le unità sono scelte a caso, perciò i valori di (X, Y) per unità diverse sono *indipendentemente distribuite*.

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: $0<E(X^4)<\infty$ e $0<E(Y^4)<\infty$

- 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 H0: $\beta_1 = 0$
- 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(Y|X)=g(X) è una f. lineare

•$E(u|X) = 0$ (prima assunzione dei minimi quadrati)

•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 assunzione)

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:

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

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

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

Quindi

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

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

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

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($\hat{\beta}_1$):

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

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

= 0 **poiché $E(u_i|X_i=x) = 0$ per l'assunzione OLS 1 (e grazie all'ass. 2)**

- Quindi l'assunzione 1 e 2 implicano insieme 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($\hat{\beta}_1$):

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

dove $v_i = (X_i - \bar{X})u_i$. Se n è grande, $s_X^2 \approx \sigma_X^2$, $\frac{n-1}{n} \approx 1$, e $\bar{X} \approx \mu_X$, perciò

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

dove $v_i = (X_i - \mu_X)u_i$ (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 var($\hat{\beta}_1 - \beta_1$) = var($\hat{\beta}_1$)

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

dove l'uguaglianza finale usa l'assunzione 2. Quindi

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

Riepilogo

1. $\hat{\beta}_1$ è non distorto: $E(\hat{\beta}_1) = \beta_1$ – proprio come $\bar{Y}$!
2. 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 $\text{var}(\hat{\beta}_1) \propto 1/n$ 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 $\text{var}(v) = \sigma^2$ finita. 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 $\hat{\beta}_1$:

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

- Quando *n* è grande, $v_i = (X_i - \overline{X})u_i \approx (X_i - \mu_X)u_i$, 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

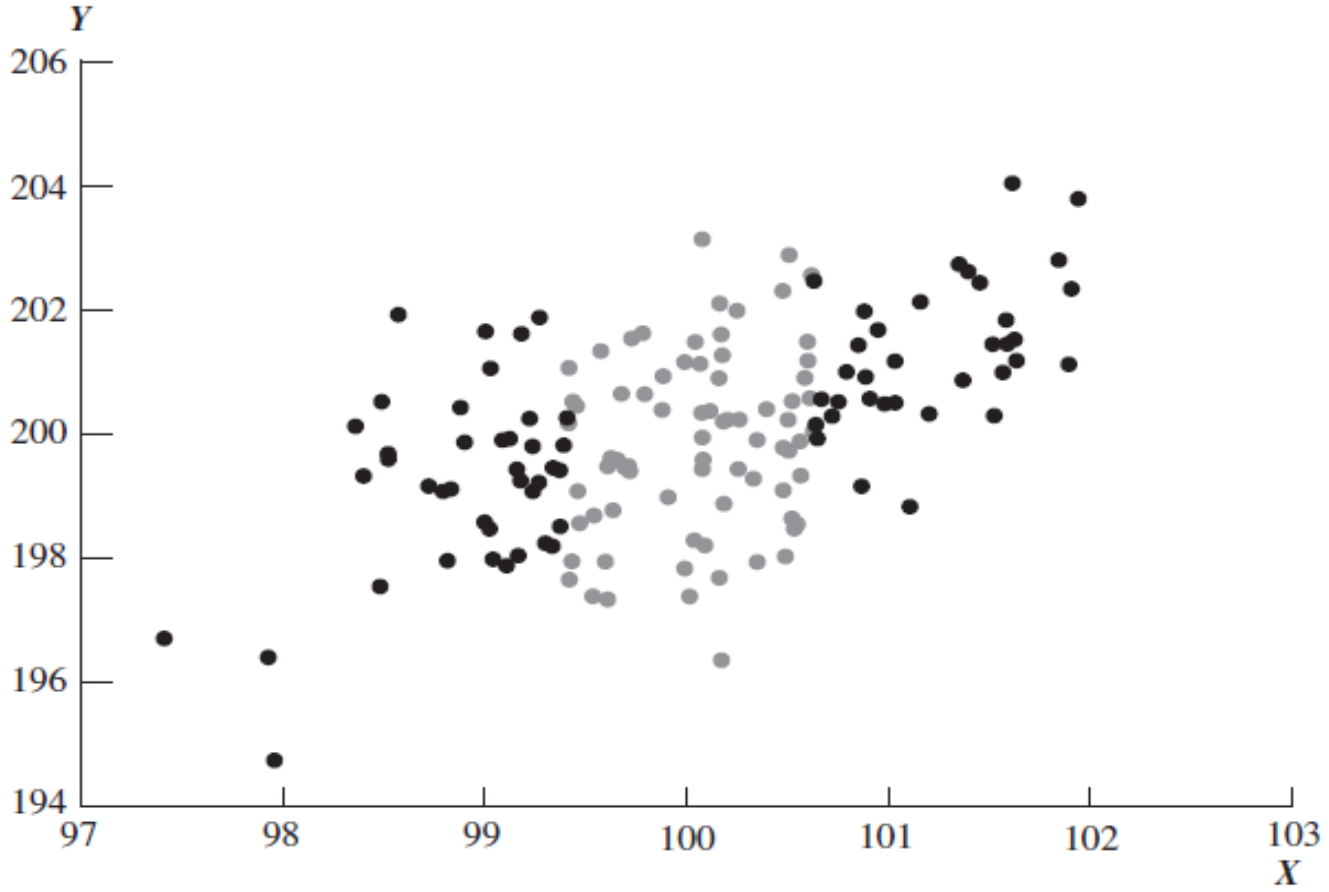
$$\text{var}(\hat{\beta}_1 - \beta_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{(\sigma_X^2)^2}$$

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…

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)
  - var($\hat{\beta}_1$) = $\frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{\sigma_X^4} \propto \frac{1}{n}$.
- 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

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

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

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

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