Econometria | 2023/2024

Lezione 3: Modello Lineare Univariato I

Lezione 4: Modello Lineare Univariato II

Giuseppe Ragusa

https://gragusa.org

Roma, 26 febbraio 2024

Sommario

- Il modello di regressione lineare
- Lo stimatore dei minimi quadrati ordinari OLS
- Misure di bontà della regressione campionaria
- Le assunzioni dei minimi quadrati
- La distribuzione campionaria dello stimatore OLS

Regressione: motivazione

- Obiettivo stimare l'effetto di un'aumento (o di una diminuzione) della dimensione della classe sulla qualità della didattica misurata dal punteggio nei test standardizzati
- Approccio: stimare

$$\Delta = E(testscr|STR < 20) - E(testscr|STR \geqslant 20)$$

usando la differenza delle medie campionarie

```r
testscr <- Caschool$testscr
str <- Caschool$str
Delta_hat <- mean(testscr[str<20]) -
              mean(testscr[str>=20])
cat("Differenza delle medie", Delta_hat)
```

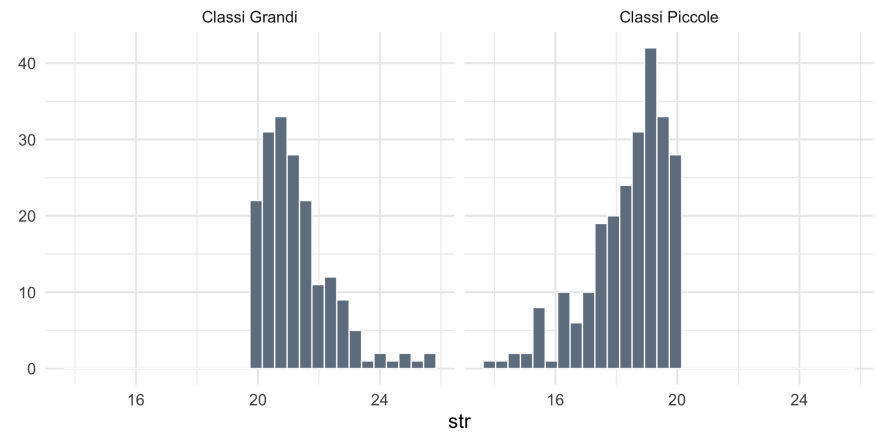
```
Differenza delle medie 7.372
```

- $\hat{\Delta} = 7.3724$: scuole con classi grandi ($str \geqslant 20$) hanno un punteggio più alto di circa 7,3 punti rispetto a scuole con classo piccole ($str < 20$)

Problemi

Ci sono (almeno) due problemi con l'approccio appena descritto:

1. difficile quantificare un aumento puntuale della dimensione delle classi
2. difficile tenere conto di altre variabili che potrebbe influenzare il puntegio nei test e la dimensione delle classi in ciascuna scuola



Problemi con Δ

- Il problema 1. potrebbe essere risolto considerando gruppi piu' piccoli, e.g.,

$$\Delta = E(testscr|STR = 20) - E(testscr|STR = 22)$$

- Il problema con questo approccio è la dimensione del campione:

```r
n_20 = sum(str==20)
n_22 = sum(str==22)
cat("distretti con str=20: ", n_20,
    "; distretti con str=22:", n_22)
```

```
distretti con str=20:  5 ; distretti con
str=22: 2
```

```r
gr1 <- testscr[str==20]
gr2 <- testscr[str==22]
Delta_hat <- mean(gr1) - mean(gr2)
s1 <- var(gr1); n1 <- length(gr1)
s2 <- var(gr2); n2 <- length(gr2)
## Errore standard
stderr = sqrt(s2/n2 + s1/n1)
## Intervallo di confidenza
c(Delta_hat -1.96*stderr,
  Delta_hat + 1.96*stderr)
```

```
[1] -47.32 -11.58
```

stime molto imprecise e intervallo di confidenza molto ampio

Alternativa: considerare piccoli intervalli, e.g.,

$$\Delta = E(testscr|19 \leqslant str \leqslant 21) - E(testscr|21 < STR \leqslant 23).$$

```
gr1 <- testscr[str>=19 & str<=21]
gr2 <- testscr[str>21 & str<=23]
Delta_hat <- mean(gr1) - mean(gr2)
s2_20 <- var(gr1); n_20 <- length(gr1)
s2_22 <- var(gr2); n_22 <- length(gr2)
## Errore standard
stderr = sqrt(s2_20/n_20 + s2_22/n_22)
## Intervallo di confidenza
c(Delta_hat -1.96*stderr,
  Delta_hat + 1.96*stderr)
```

```
[1]  1.747 10.746
```

Anche in questo caso poche osservazioni e intervallo di confidenza molto grande

La stima cambia anche drammaticamente per piccole variazioni dell'intervallo; e.g., se considerassimo

$$\Delta = E(testscr|19.5 \leqslant str \leqslant 20.5) - E(TestScore|21.5 \leqslant str \leqslant 22.5),$$

otterremmo un intervallo di confidenza $(-3.894, 9.591)$

Un framework per l'analisi degli effetti causali

Problema 2. tenere conto di altre variabile che possono contemporaneamente influenzare sia i punteggi sia la dimensione delle classi

In altre parole, abbiamo bisogno di un framework che ci consenta di valutare di volta in volta se le nostre stime hanno carattere causale oppure no…

Il modello lineare univariato

Il modello di regressione lineare è

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

Il valore atteso condizionale di Y_i dato X_i è una funzione lineare di X.

I due parametri hanno una semplice interpretazione:

- β_0 (intercetta) è il valore atteso di Y_i associato a $X = 0$

$$\beta_0 = E(Y_i|X_i = 0)$$

- β_1 (pendenza) la variazione di $E(Y|X)$ associata ad una variazione unitaria di X

$$\beta_1 = E(Y_i|X_i = x + 1) - E(Y_i|X_i = x).$$

Il modello lineare univariato

Due modi equivalenti di descrivere il modello lineare:

1. Modello per $E(Y_i|X_i)$

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

2. Modello per Y_i con errore

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad E(u_i|X_i) = 0$$

I due modelli sono equivalenti.

Il modello lineare univariato

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

n osservazioni su (X_i, Y_i)

- X_i è la variabile indipendente o regressore
- Y_i è la variabile dipendente
- β_0 è l'intercetta
- β_1 è la pendenza
- u_i è l'errore di regressione

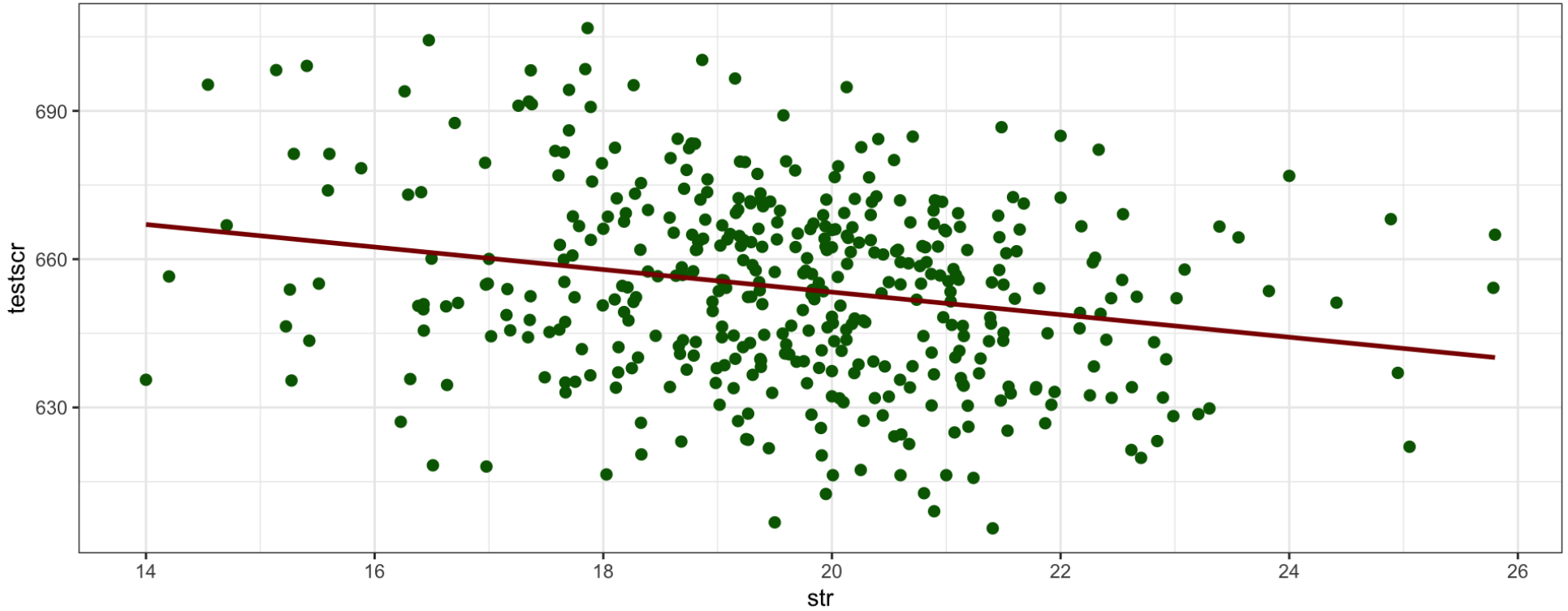
Motivi per includere il termine di errore

1. Anche con più regressori, alcune determinanti di Y saranno in pratica omesse sempre dal modello.
    - Ad esempio perché i fattori che inflenzano Y sono troppi per essere inclusi in un modello
    - oppure perché alcuni di questi fattori sono non osservabili o non misurabili
2. Ci possono essere errori nel modo in cui Y è misurata che non possono essere modellati.
3. Ci possono essere fattori che per loro natura non possono essere modellati come, ad esempio, un attacco terroristico o uno shock non prevedibile (COVID).

Roadmap

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

Come tracciare una retta attraverso i dati?



Il modello di regressione lineare (Par 4.1)

La retta di regressione per il nostro modello della relazione lineare tra STR e TestScore è dato da:

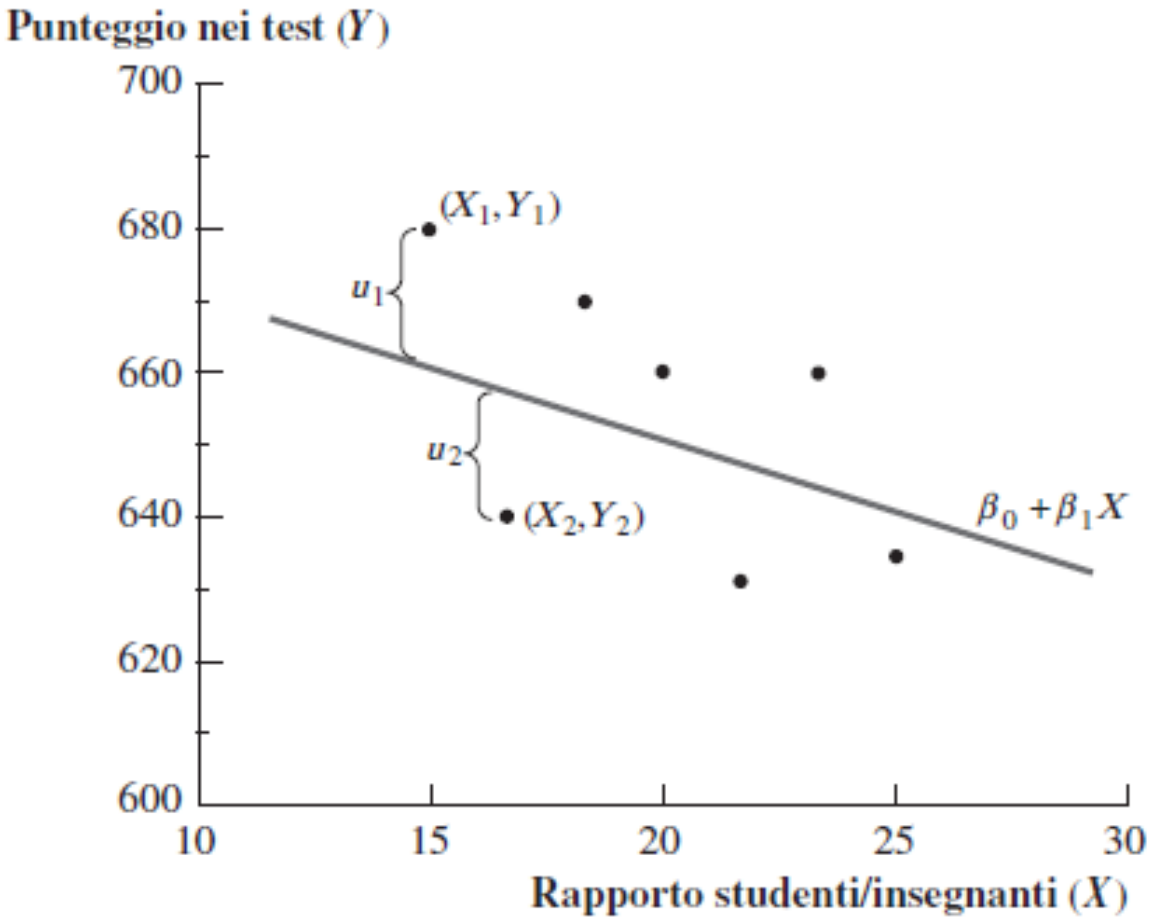
$$testscr_i = \beta_0 + \beta_1 str_i + u_i \quad \forall i = 1, \ldots, n$$

$$\begin{aligned} \beta_1 &= \text{pendenza della retta di regressione} \\ &= \text{variazione nel punteggio nei test per una variazione unitaria in } str \end{aligned}$$

Non conosciamo β_1 perciò dobbiamo stimarlo utilizzando i dati.

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 (Par. 4.2)

Come possiamo stimare β_0 e β_1 dai dati?

Si ricordi che lo stimatore OLS di μ, $\bar{Y}$ 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.
- Lo stimatore OLS di β_0 e β_1 è la soluzione al seguente problema di ottimizzazione:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \hat{u}_i^2 = \min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

- La soluzione puo’ essere ottenuta mediante il calcolo differenziale

Lo stimatore OLS

Lo stimatore OLS minimizza la differenza quadratica media tra i valori reali di Y_i e la previsione $\hat{Y}_i$ (“valori predetti”) basata sulla retta stimata.

Gli stimatori OLS (cioe’ le soluzioni al problema di ottimizzazione) sono dati da

$$\hat{\beta}_0 = \bar{Y} - \bar{X}\hat{\beta}_1,$$

e

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

Applicazione ai dati della California

```
Call:
lm(formula = testscr ~ str, data = Caschool)

Coefficients:
(Intercept)          str
     698.93        -2.28
```

Pendenza stimata: $\hat{\beta}_1 = -2.28$

Intercetta stimata: $\hat{\beta}_0 = 698.9$

Retta di regressione stimata: $\widehat{testscr}_i = 698.9 - 2.28 \times str$

Interpretazione delle stime di pendenza e intercetta

$$\widehat{testscr} = 698.9 - 2.28 \times str$$

Scuole con uno studente in più per insegnante in media ottengono punteggi nei test inferiori di 2,28 punti.

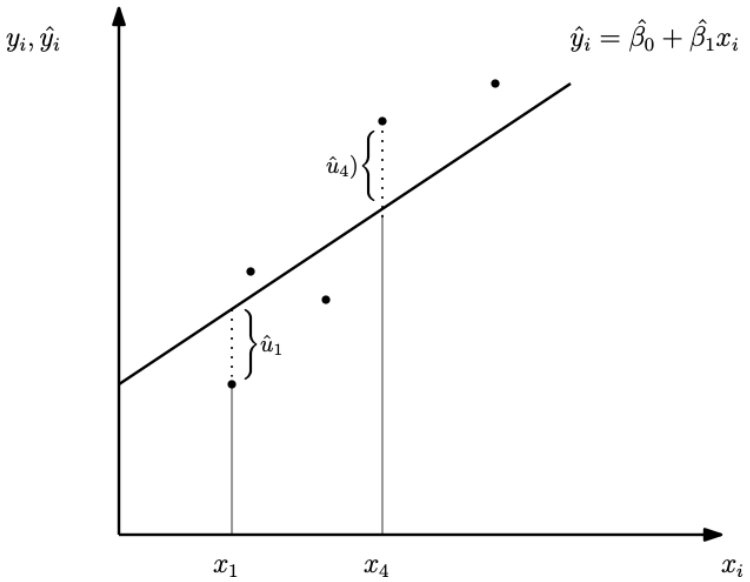
L'intercetta puo' essere (letteralmente) interpretata come una stima del valore atteso dei punteggi in distretti con zero studenti per insegnante $(\hat{\beta}_0 = 698.9)$.

Questa interpretazione non ha molto 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 Elementary che si trova nella contea di Tehama, CA

	district	county	testscr	str
242	Antelope Elementary	Tehama	657.8	19.3



Valore predetto $\hat{y}_i$:

$$\widehat{testscr} = 698.933 – 2.28 \times 19.3 = 654.857$$

Residuo $\hat{u}_i$:

$$\hat{u}_i = 657.75 – 654.8567 = 2.8933$$

Valori predetti e residui

	district	county	testscr	str
242	Antelope Elementary	Tehama	657.8	19.3

```r
lm1 <- lm(testscr~str, data=Caschool)
predict(lm1, newdata = list(str=19.33))
```

```
    1
654.9
```

```r
uhat <- residuals(lm1)
uhat[242]
```

```
  242
2.893
```

Proprietà algebriche delle stime OLS

Sono proprietà derivanti dal metodo OLS a prescindere da come i dati sono generati.

1. La stimatore OLS è espresso solo in termini di osservabili (X, Y).
2. Lo stimatore OLS è uno stimatore puntuale.
3. La retta OLS passa sempre dalla media del campione $(\bar{X}, \bar{Y})$.
4. La somma dei residuai è uguale a zero: $\sum_{i=1}^{n} \hat{u}_i = 0$.
5. La media delle osservazioni è uguale alla media dei valori predetti: $\bar{Y} = \bar{\hat{Y}}$.
6. Il prodotto incrociato tra X_i ed i residui $\hat{u}_i$ è zero: $\sum_{i=1}^{n} X_i \hat{u}_i = 0$.
7. Il prodotto incrociato tra $\hat{Y}_i$ ed i residui $\hat{u}_i$ è zero: $\sum_{i=1}^{n} \hat{Y}_i \hat{u}_i = 0$.

Misure di bontà dell'adattamento (Par 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 di Y spiegata da X
  - R^2 è privo 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.

R^2 della regressione

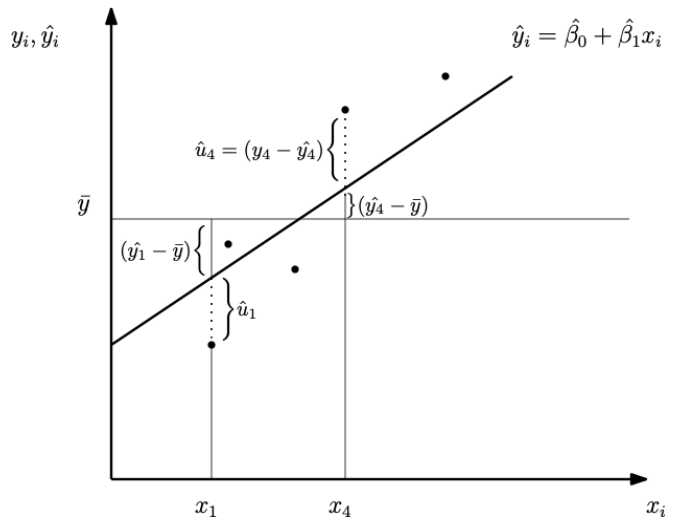
La distanza di Y_i dalla media $\bar{Y}$ può essere suddivisa in due parti:

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

Prendendo i quadrati e sommando sul campione otteniamo la variazione totale e, quindi:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

$$= \underbrace{\sum_{i=1}^{n}\hat{u}_i^2}_{RSS} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}_{ESS}$$



R^2 della regressione è la frazione della varianza campionaria "spiegata"

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \underbrace{\sum_{i=1}^{n}\hat{u}_i^2}_{RSS} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}_{ESS}$$

- Dividendo questa espressione per $n-1$, otteniamo che la varianza campionaria di Y_i è uguale alla varianza campionaria di $\hat{Y}_i$ piu' la varianza campionaria di $\hat{u}_i$ (questo segue dal fatto che la covarianza fra X_i e $\hat{u}_i$ è uguale a zero)

Definizione di R^2:

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

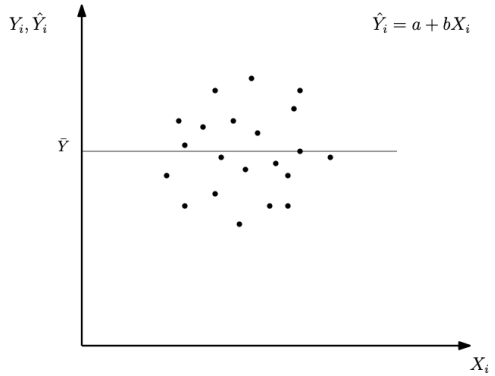
R^2 della regressione è la frazione della varianza campionaria "spiegata"

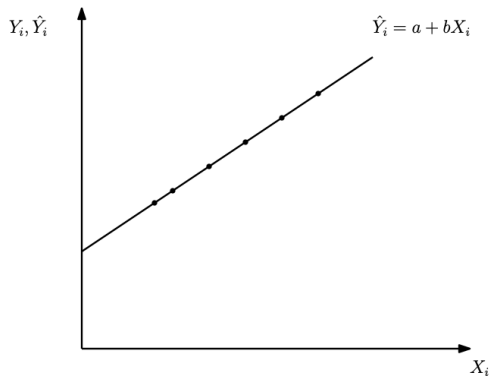
$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

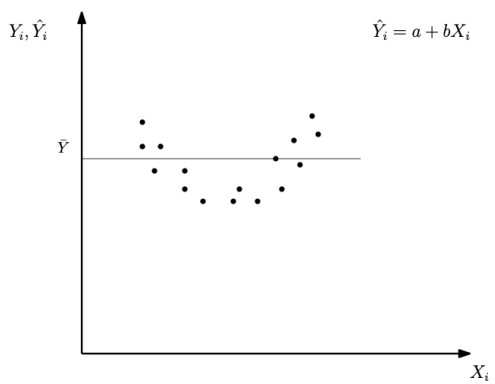
- $R^2 = 0 \implies ESS = 0$
- $R^2 = 1 \implies ESS = TSS$
- $0 \leqslant R^2 \leqslant 1$

Important

Per la regressione con una singola X, l' R^2 è il quadrato del coefficiente di correlazione tra X e Y


$Y_i, \hat{Y}_i$
$\hat{Y}_i = a + bX_i$
$\bar{Y}$
X_i


$Y_i, \hat{Y}_i$
$\hat{Y}_i = a + bX_i$
X_i


$Y_i, \hat{Y}_i$
$\hat{Y}_i = a + bX_i$
$\bar{Y}$
X_i

L’errore standard della regressione (SER)

Il SER misura la dispersione della distribuzione di $\hat{u}_i$ attorno alla retta di regressione. 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

$$\bar{\hat{u}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0$$

Il SER:

- ha le unità di u_i, che sono le unità di Y_i
- misura la “dimensione” media del residuo OLS (l’“errore” medio della retta di regressione OLS)

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, mentre in s_Y^2 ne è stato stimato solo uno μ.
- Quando n è grande la differenze fra $1/n$, $1/(n-1)$ e $1/(n-2)$ sono molto piccole e quindi la scelta del denominatore ha poca importanza pratica
- Per i dettagli, cfr. Paragrafo 17.4

Esempio

```r
lm1 <- lm(testscr~str, data=Caschool)
summary(lm1)
```

```
Call:
lm(formula = testscr ~ str, data = Caschool)

Residuals:
   Min     1Q Median     3Q    Max
-47.73 -14.25   0.48  12.82  48.54

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   698.93       9.47   73.82  < 2e-16 ***
str            -2.28       0.48   -4.75  2.8e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.6 on 418 degrees of freedom
Multiple R-squared:  0.0512,    Adjusted R-squared:  0.049
F-statistic: 22.6 on 1 and 418 DF,  p-value: 2.78e-06
```

```r
## Calcolo R2
uhat <- resid(lm1)
TSS <- var(Caschool$testscr)
ESS <- var(predict(lm1))
R2 <- ESS/TSS
## Calcolo SER
SER <- sqrt(sum(uhat^2)/(420-2))
cat("R2: ", R2, " SER: ", SER)
```

```
R2:  0.05124  SER:  18.58
```

Le assunzioni dei minimi quadrati (Par. 4.4)

- Quali sono 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)

Le assunzioni dei minimi quadrati

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

Assunzione 1:

La distribuzione di u condizionata a X ha media nulla, cioè

$$E(u_i|X_i) = 0;$$

Assunzione 2:

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

Assunzione 3:

Gli outlier in X e/o Y sono rari.

Assunzione dei minimi quadrati n. 1:

$$E(u_i|X_i) = 0$$

Per ogni dato valore di X_i, la media di u_i è zero:

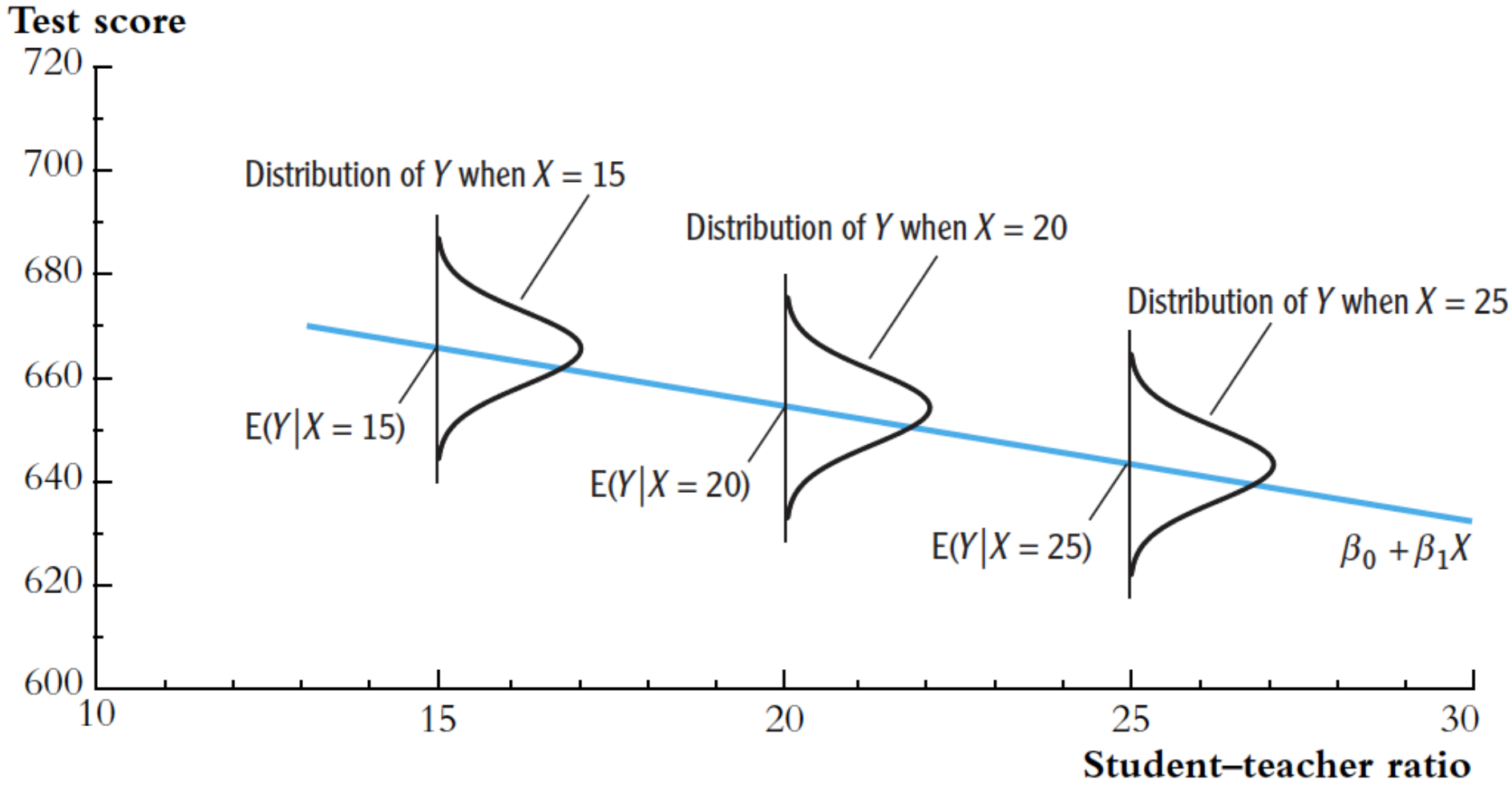
- se $E(u_i|X_i) = 0$, allora $corr(u_i, X_i) = 0$
  - quindi stiamo assumendo che i fattori omessi dalla regressione (cioè le altre variabili importanti nella determinazione di Y_i contenute in u_i) non siano correlati con X_i

Esempio:

$$testscr_i = \beta_0 + \beta_1 str_i + u_i$$

- Quali sono alcuni di questi "altri fattori"?
- E' plausibile che questi fattori non siano correlati con X_i?

Assunzione dei minimi quadrati n. 1 (continua)



Assunzione dei minimi quadrati n. 2:

$$(Y_i, X_i),\ i = 1, \ldots, n \ \text{ i.i.d.}$$

- Questo si verifica automaticamente se l'unità (individuo, distretto) è campionata mediante campionamento casuale semplice:
  - Le unità sono scelte dalla stessa popolazione, perciò (Y_i, X_i) sono identicamente distribuite per ogni $i = 1 \ldots, n$.
  - Le unità sono scelte a caso, perciò i valori di (Y_i, X_i) per unità diverse (diversi i) sono indipendentemente distribuiti.
- 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.

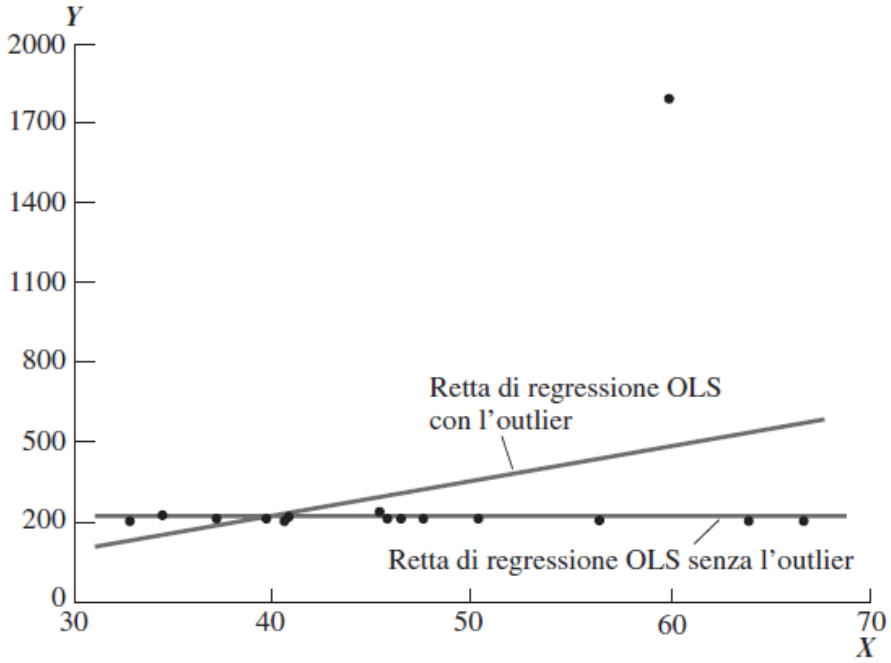
Assunzione dei minimi quadrati n. 3:

$$E(X^4) < \infty, \quad E(Y^4) < \infty$$

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?
- Gli outlier indicano problemi nei dei dati (problemi nella codifica o nella registrazione).
- Controllare sempre i dati e assicurarsi che non ci siano errori

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 $\hat{\beta}_1$.

Dobbiamo

- quantificare l'incertezza campionaria associata a $\hat{\beta}_1$
- usare $\hat{\beta}_1$ per verificare ipotesi quali $H_0 : \beta_1 = 0$
- costruire un intervallo di confidenza per β_1

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

1. Qual è $E(\hat{\beta}_1)$? • Se $E(\hat{\beta}_1) = \beta_1$, allora lo stimatore OLS non è distorto

2. Qual $Var(\hat{\beta}_1)$ (misura di incertezza campionaria)

- Dobbiamo derivare una formula per poter calcolare l'errore standard di $\hat{\beta}_1$

3. Qual è la distribuzione di $\hat{\beta}_1$ in piccoli campioni? • È molto complessa, in generale

4. Qual è la distribuzione di $\hat{\beta}_1$ in grandi campioni?

- In grandi campioni, $\hat{\beta}_1$ ha distribuzione normale.

Media e varianza di $\hat{\beta}_1$

Un po' di algebra preliminare:

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

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

Sostituendo questa espressione nella formula di $\hat{\beta}_1$ otteniamo

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

Si noti che

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\bar{X})(u_i-\bar{u}) &= \sum_{i=1}^{n}(X_i-\bar{X})u_i - \bar{u}\sum_{i=1}^{n}(X_i-\bar{X}) \\ &= \sum_{i=1}^{n}(X_i-\bar{X})u_i - \bar{u}\left(\sum_{i=1}^{n}X_i - n\bar{X}\right) \\ &= \sum_{i=1}^{n}(X_i-\bar{X})u_i\end{aligned}$$

pertanto

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

Ora possiamo calcolare $E(\hat{\beta}_1)\ldots$

$$
\begin{aligned}
E(\hat{\beta}_1) &= \beta_1 + E\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) \\
&= \beta_1 + \underbrace{E\left[\underbrace{E\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\middle| X_1, \ldots, X_n\right)}\right]}_{\text{per la legge delle aspettative iterate}} \\
&= \beta_1 + E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})\underbrace{E(u_i|X_1, \ldots, X_n)}_{=0, \text{ per Assunzione n. 1}}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] \\
&= \beta_1
\end{aligned}
$$

Quindi l'assunzione 1 implica che

$$E(\hat{\beta}_1) = \beta_1$$

Ora calcoliamo la varianza di $\hat{\beta}_1$

Scriviamo

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

dove $\nu_i = (X_i - \bar{X})u_i$. Se n è grande, $s_X^2 \approx \sigma_X^2$ e $\frac{n-1}{n} \approx 1$. Possiamo pertanto scrivere

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

Quindi, usando Assunzione n. 2, possiamo scrivere:

$$var(\hat{\beta}_1) \approx \frac{var(\nu_i)/n}{(\sigma_X^2)^2}.$$

Riepilogo

- $\hat{\beta}_1$ è non distorto:

$$E(\hat{\beta}_1) = \beta_1$$

- La varianza dello stimatore OLS

$$var(\hat{\beta}_1) \approx \frac{1}{n} \frac{var((X_i - \bar{X})u_i)}{(\sigma_X^2)^2} \propto \frac{1}{n}$$

è inversamente proporzionale a n – proprio come $var(\bar{Y})$

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

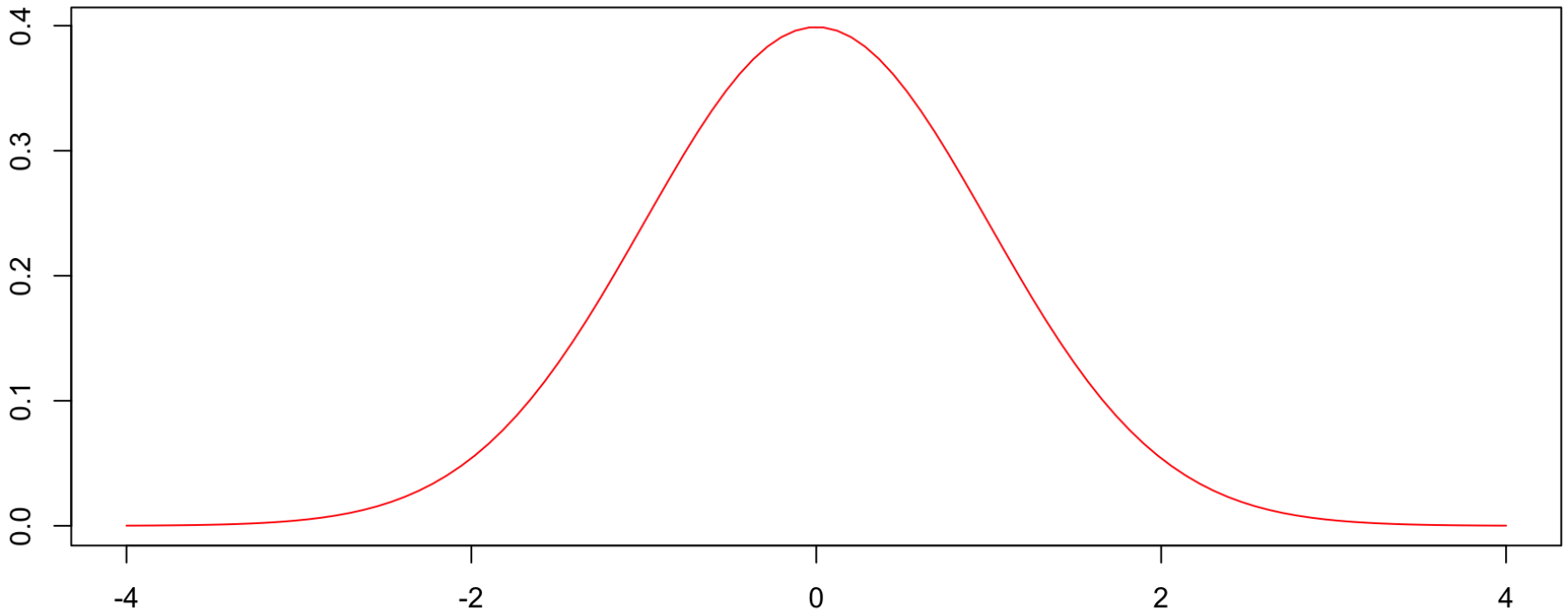
- Determinare la distribuzione campionaria esatta è complicato – dipende dalla distribuzione di (Y_i, X_i)
- Quando n è grande otteniamo alcune buone e semplici approssimazioni:
  - Quando n è grande, la distribuzione campionaria di $\hat{\beta}_1$ è ben approssimata da una distribuzione normale (TLC)

Theorema del Limite Centrale

Se $Z_i, i = 1, \ldots, n$ i.i.d con $E(Z_i) = 0$ e $var(Z_i) = \sigma^2$, allora quando n e' grande, la distribuzione di

$$\bar{Z} = \frac{1}{n} \sum_{i=1}^{n} Z_i$$

e' approssimata da $N(0, \sigma^2)$


0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4

Approssimazione della distribuzione di $\hat{\beta}_1$

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

- Quando n è grande, $(X_i - \bar{X})u_i \approx (X_i - \mu_X)u_i$, che e' i.i.d e $var((X_i - \mu_X)u_i) < \infty$ (per via Assunzione n 3)
- Quindi, per il TLC, la distribuzione di $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i$ e' approssimata da $N(0, var((X_i - \mu_x)u_i)/n)$
- Per n grande, segue che

$$\hat{\beta}_1 \xrightarrow{d} N(\beta_1, \sigma^2_{\hat{\beta}_1}),\ \sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{var((X_i - \bar{X})u_i)}{(\sigma^2_X)^2}$$

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

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

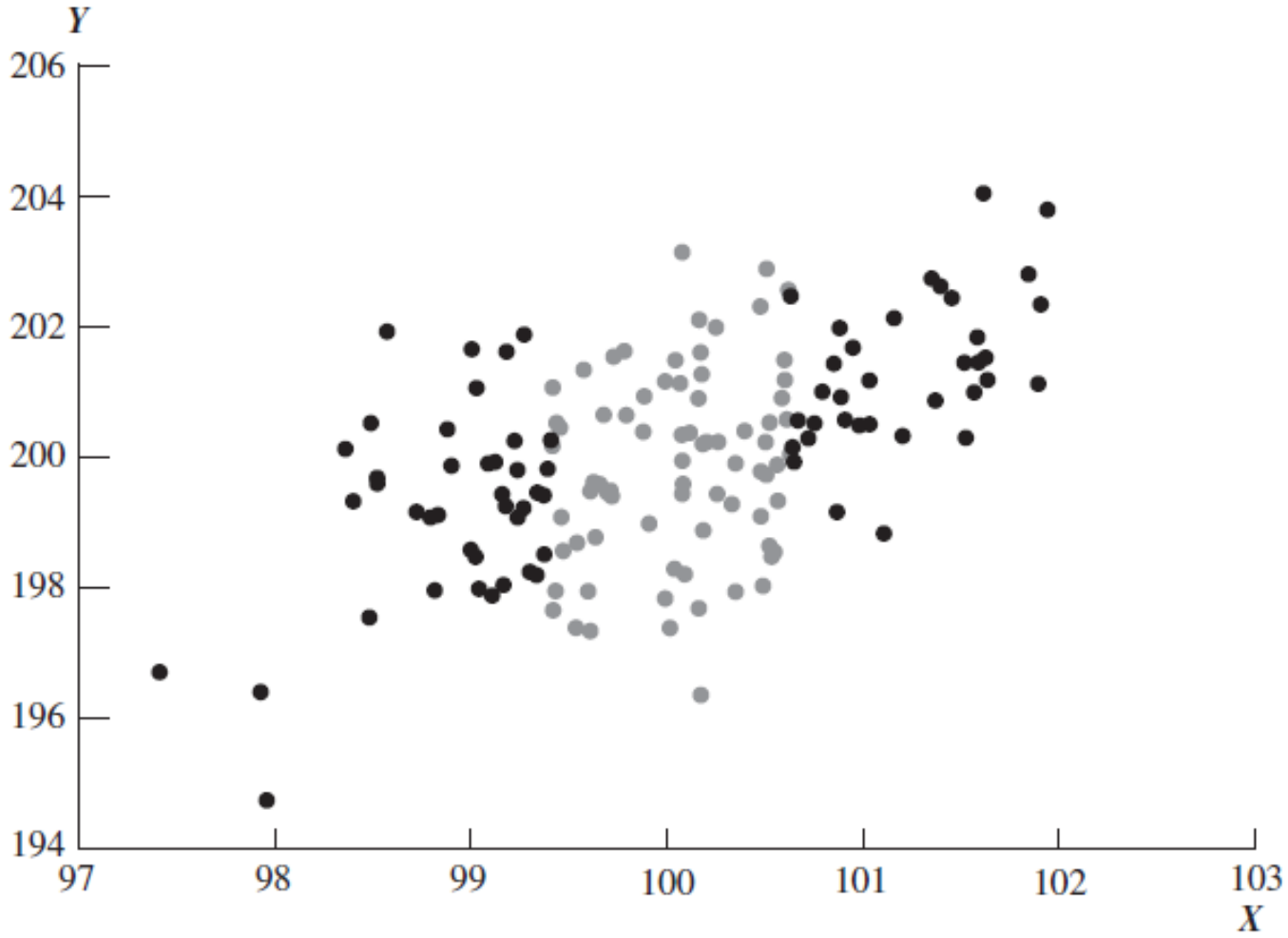
La distribuzione in grandi campioni di $\hat{\beta}_0$ è $N(\beta_0, \sigma^2_{\hat{\beta}_0})$, dove

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

Maggiore è la varianza di X, minore è la varianza di

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{var((X_i - \bar{X})u_i)}{(\sigma^2_X)^2}$$

- La varianza di X, σ^2_X, appare al quadrato al denominatore – perciò aumentando la dispersione di X diminuisce la varianza di $\hat{\beta}_1$
- Se vi è più variazione in X, allora vi sono più informazioni nei dati per l'adattamento della retta di regressione.



Il numero di punti neri e grigi è lo stesso.

Quali consentono di ottenere una retta di regressione più accurata?

Maggiore è la varianza di u, maggiore è la varianza di

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{var((X_i - \bar{X})u_i)}{(\sigma^2_X)^2}$$

- La varianza di u appare al numeratore – perciò aumentando la dispersione di u aumenta la varianza di $\hat{\beta}_1$
- Se gli errore sono più piccoli (mantenendo fissi gli X) i dati si concentrerebbero maggiormente attorno alla retta di regressione della popolazione -> la pendenza sarebbe quindi stimata meglio

Errore Standard di $\hat{\beta}_1$

L'errore standard è una stima della variabilita' di $\hat{\beta}_1$, cioe' una stima di

$$\sigma_{\hat{\beta}_1} = \sqrt{\sigma^2_{\hat{\beta}_1}}$$

Questa stima e' calcolata da R (e da altri software statistici), ma vedremo che bisogna "istruire" R per ottenere una stima appropriata e coerente con le assunzioni.

```r
library(fixest)
feols(testscr~str, data=Caschool, vcov="hetero")
```

```
OLS estimation, Dep. Var.: testscr
Observations: 420
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error t value   Pr(>|t|)
(Intercept)   698.93    10.3644  67.436  < 2.2e-16 ***
str            -2.28     0.5195  -4.389 1.4467e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 18.5   Adj. R2: 0.04897
```