Principi di Econometria

lezione 9

AA 2016-2017

Paolo Brunori

Dove siamo arrivati?

- la regressione lineare multipla ci permette di stimare l'effetto della variabile X sulla Y tenendo ferme tutte le altre variabili osservabili che hanno un impatto su Y
- le stime si ottengono in modo analogo alla regressione univariata, minimizzando la somma degli errori compiuti nell'interpolare i dati
- in questo caso non usiamo una retta ma un (iper)piano di regressione di k dimensioni, tante quanti sono le variabili di controllo X
- ogni volta che aggiungiamo un regressore spieghiamo un po' della variabilità di Y
- per questo usiamo $R^2 - corretto$ invece che l'R^2

assunzioni necessarie per OLS

- le assunzioni sono le stesse già viste + 1
  1. $E(u_i|X_i) = 0 \; \forall i = 1, ..., k$
  2. $(X_1, ..., X_n)$ sono i.i.d
  3. gli outlier sono improbabili

$\Rightarrow$ assenza di collinearità perfetta

collninearità perfetta

- esempio 1: X_1= età, X_2=anno di nascita ($X_1 = 2016 - X_2$)
- esempio 2: inclusione di tutte le variabili che indicano categorie:
- immaginiamo di voler stimare il modello:

$$Y = \beta_0 + \beta_1 DONNA + \beta_2 UOMO + u$$

- qui $DONNA = 1 - UOMO$
- a una variazione di un regressore si associa sempre una variazione lineare dell'altro: non posso distinguere come varia Y al variare di uno tenendo fermo l'altro

nel caso del consumo di tabacco in Turchia

- immaginate di introdurre una variabile PIL ottenuta moltiplicando il reddito per 70milioni
- le due quantità PIL pro capite e PIL sono una funzione lineare dell'altra
- $PIL = 70.000.000 \times PIL_{PRO\ CAPITE}$
- non sarà quindi possibile calcolare i coefficienti
- ma in generale i software semplicemente elimineranno la variabile collineare ed effettueranno i conti come se non fosse parte del modello
- R ad esempio restituisce "NA" al posto del coefficiente

collinearità non perfetta

- se il legame fra una o più variabili è vicino ad un legame lineare la regressione OLS può essere stimata
- immaginate ad esempio di inserire una variabile 'stipendio medio' insieme alla variabile 'PIL pro capite'
- in questo caso la stima dei coefficienti è corretta
- ma per almeno uno di questi è imprecisa
- in effetti per stimare l'effetto della variazione di una variabile tenendo ferma l'altra si può sfruttare solo quella piccola parte della variabilità non identica per i due regressori

Regressione multipla del consumo di tabacco

	coefficiente	errore standard	t	$valore - p$
β_0	1.6572	0.1237	13.394	0.0000
β_Y	0.0003	0.0000	6.518	0.0000
β_P	-0 .4231	0.096	-4.3662	0.0001

Regressione multipla del consumo di tabacco: PIL e reddito pro capite fortemente correlati

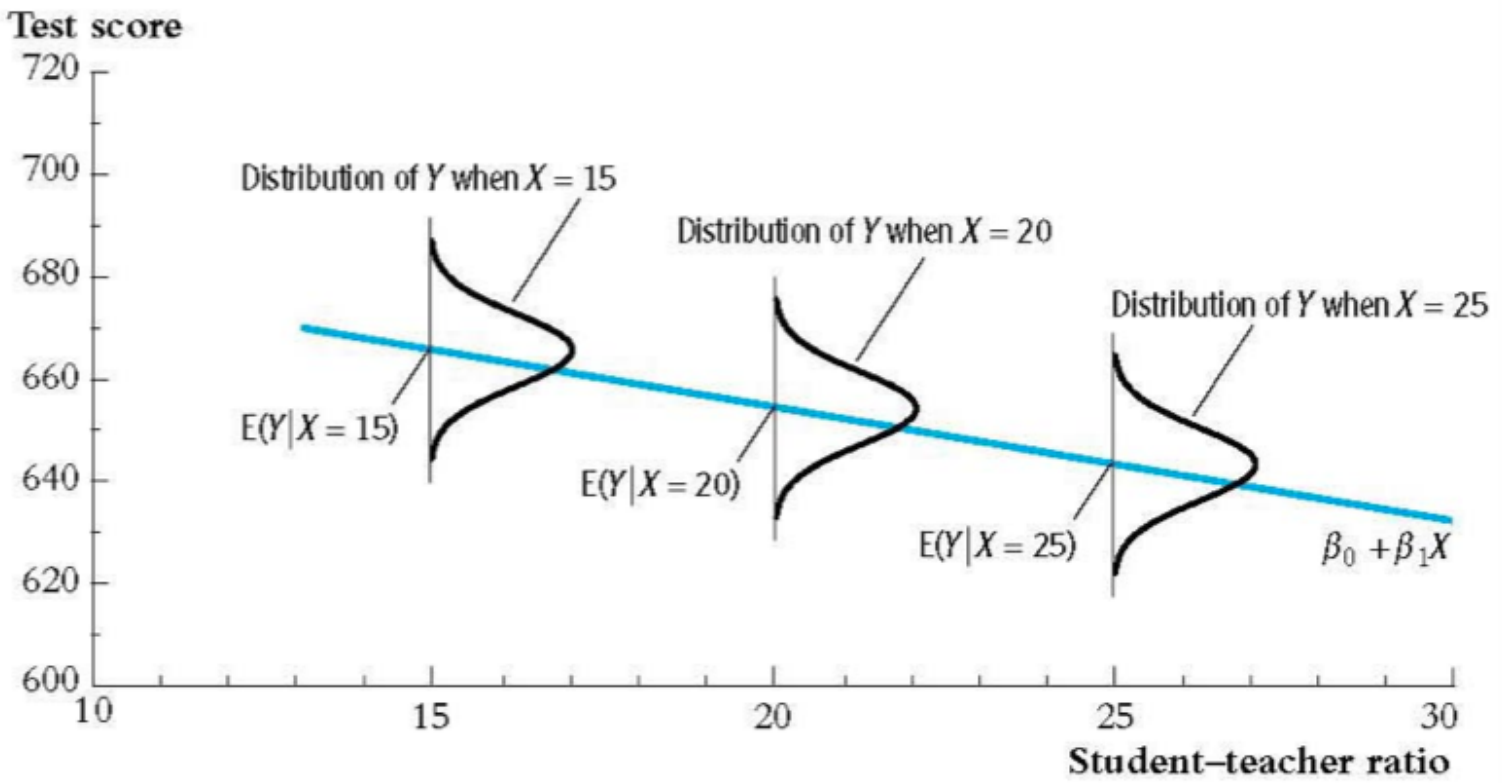
	coefficiente	errore standard	t	$valore - p$
β_0	1.5440	0.1339	11.53	0.0000
β_P	-0.4299	0.0930	-4.621	0.0001
β_Y	0.0003	0.0000	-1.043	0.3069
β_{PIL}	0.0008	0.0004	1.809	0.0825

- come posso interpretare questi coefficienti?

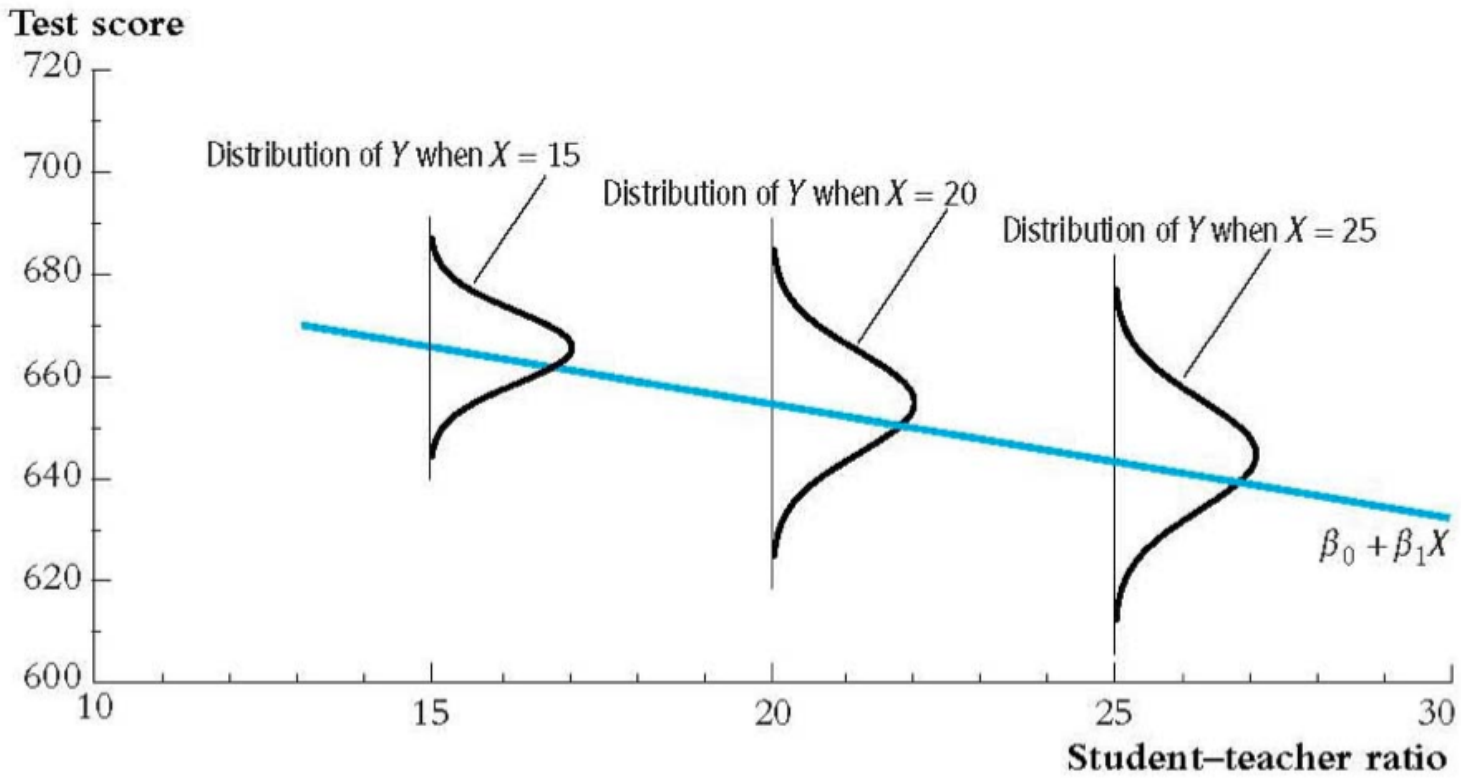
eteroschedasticità

- la prima assunzione per la validità degli OLS è che $E(u|X) = 0 \ \forall i = 1, .., n$
- se oltre a questo assumiamo che la varianza di u condizionata a X è costante allora si dice che gli errori sono omoschedastici
- la violazione di questa condizione non rende le stime distorte, le rende solo meno precise.

errori omoschedastici



errori eteroschedastici



efficienza dello stimatore OLS con omoschedasticità

- anche in caso di eteroschedasticità OLS è non distorto, consistente ed asintoticamente normale
- l'omoschedasticità aggiunge una proprietà: gli stimatori OLS sono stimatori lineari efficienti
- la prova è nel teorema di Gauss-Markov (appendice 5.2 del libro Stock & Watson)

Teorema di Gauss-Markov

- "lo stimatore OLS è il miglior stimatore lineare condizionatamente non distorto" (BLUE)
- uno stimatore lineare si può scrivere come:

$$\tilde{\beta}_1 = \sum_{i=1}^{n} a_i \, Y_i m$$

dove i pesi a_i possono dipendere da X_i ma non da Y_i

- la non distorsione condizionata implica:

$$E(\tilde{\beta}|X_1, ..., X_n) = \beta_1$$

- sotto le assunzioni dell'OLS (1-3) + l'omoschedacità + $u_i \sim N(0, \sigma_u^2)$: il teorema è valido

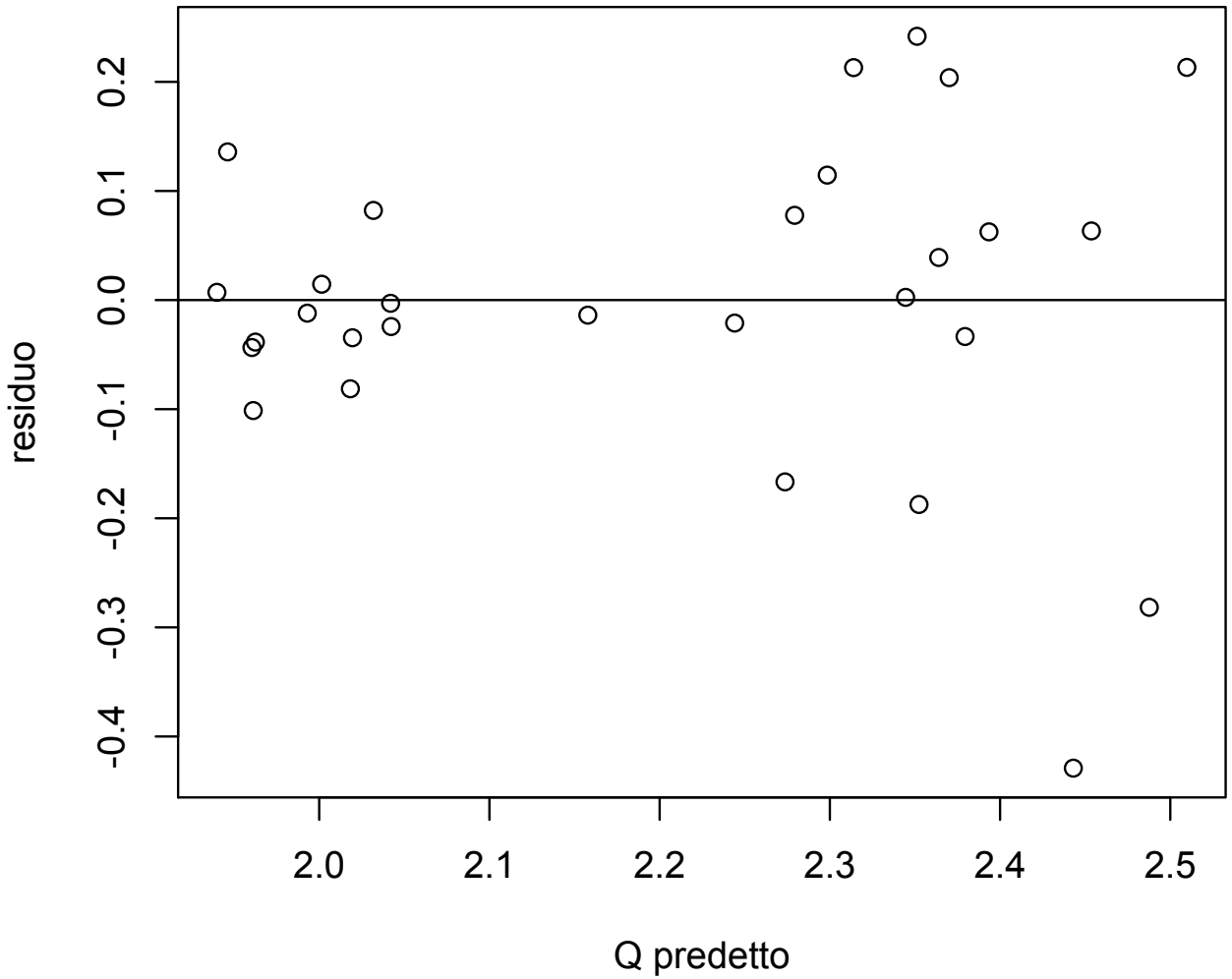
stimatore dei minimi quadrati ponderati

- se conosciamo l'eteroschedasticità (varianza di u condizionata a x)
- allora si può costruire uno stimatore più efficiente dell'OLS
- lo stimatore *dei minimi quadrati ponderati* utilizza un peso per ogni osservazione i:

$$w_i = \frac{1}{\sqrt{\sigma^2_{u|X_i}}}$$

- il peso impone che le osservazioni che si trovano in zone dove l'errore è più disperso attorno all'iperpiano di regressione contino meno nel definire i coefficienti

consumo di tabacco in Turchia: errori eteroschedastici



Regressione multipla del consumo di tabacco

	coefficiente	errore standard	t	$valore - p$
β_0	1.6572	0.1237	13.394	0.0000
β_Y	0.0003	0.0000	6.518	0.0000
β_P	-0 .4231	0.096	-4.3662	0.0001

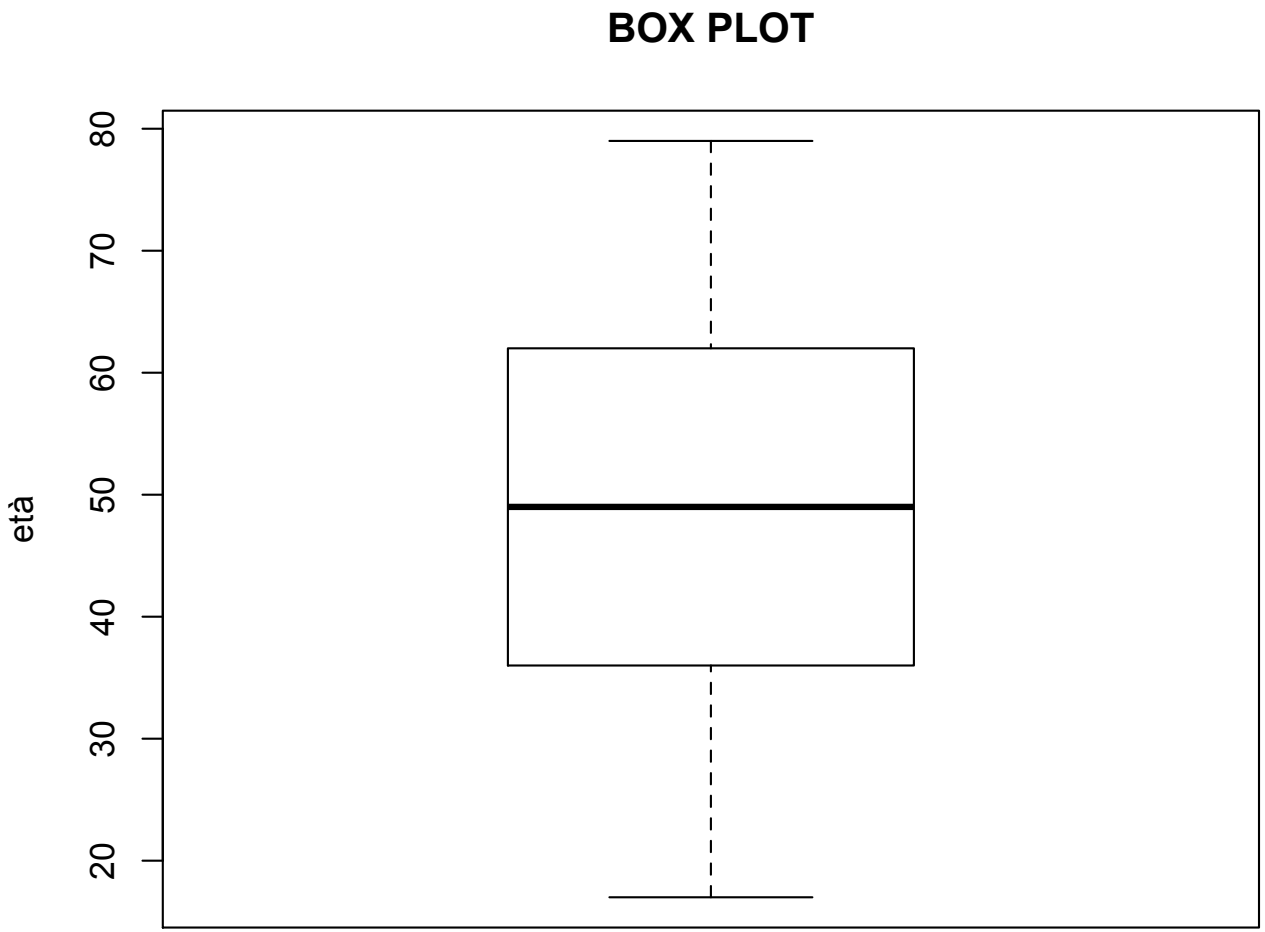
- stimando lo stesso modello specificando a R che siamo in presenza di eteroschedasticità ci costringe a tener conto del fatto che la nostra stima è meno efficiente

Regressione multipla robusta per eteroschedasticità

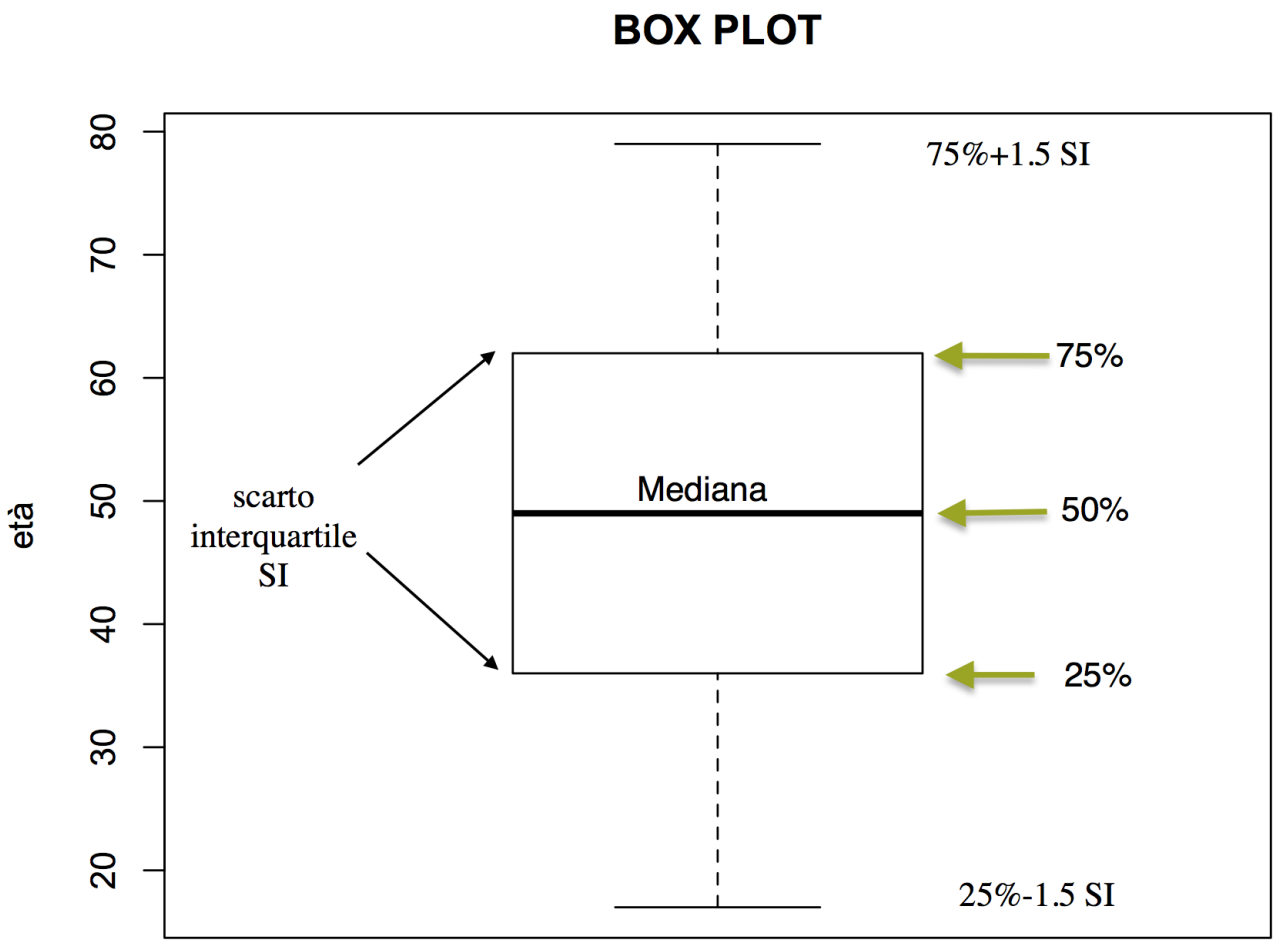
	coefficiente	errore standard	t	$valore - p$
β_0	1.6572	0.1746	9.4859	0.0000
β_Y	0.0003	0.0000	3.599	0.0014
β_P	-0 .4231	-0.24	-1.7494	0.09202

- questo ha effetto sugli errori standard e quindi sulla significatività

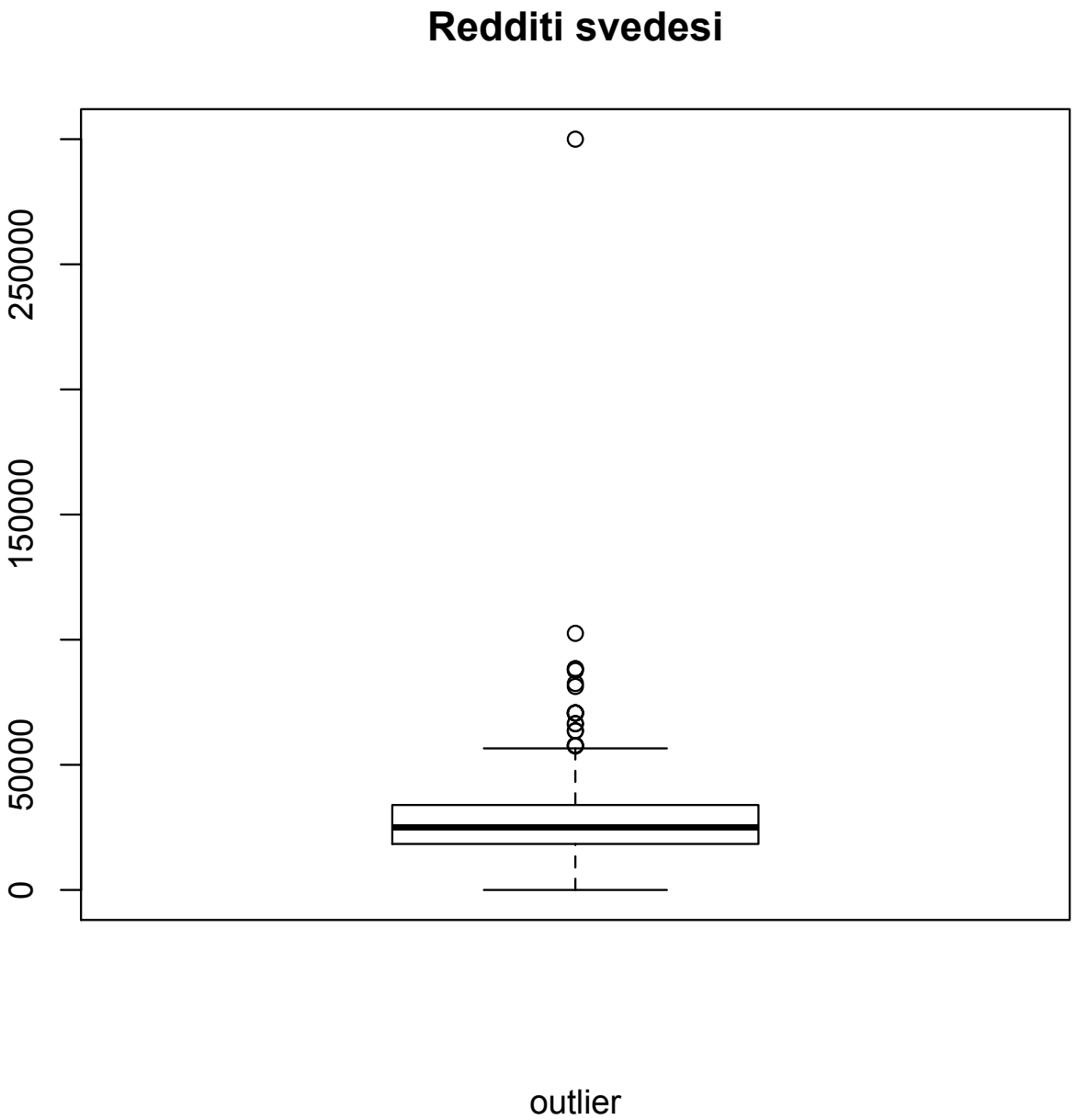
Come si individuano gli outlier?



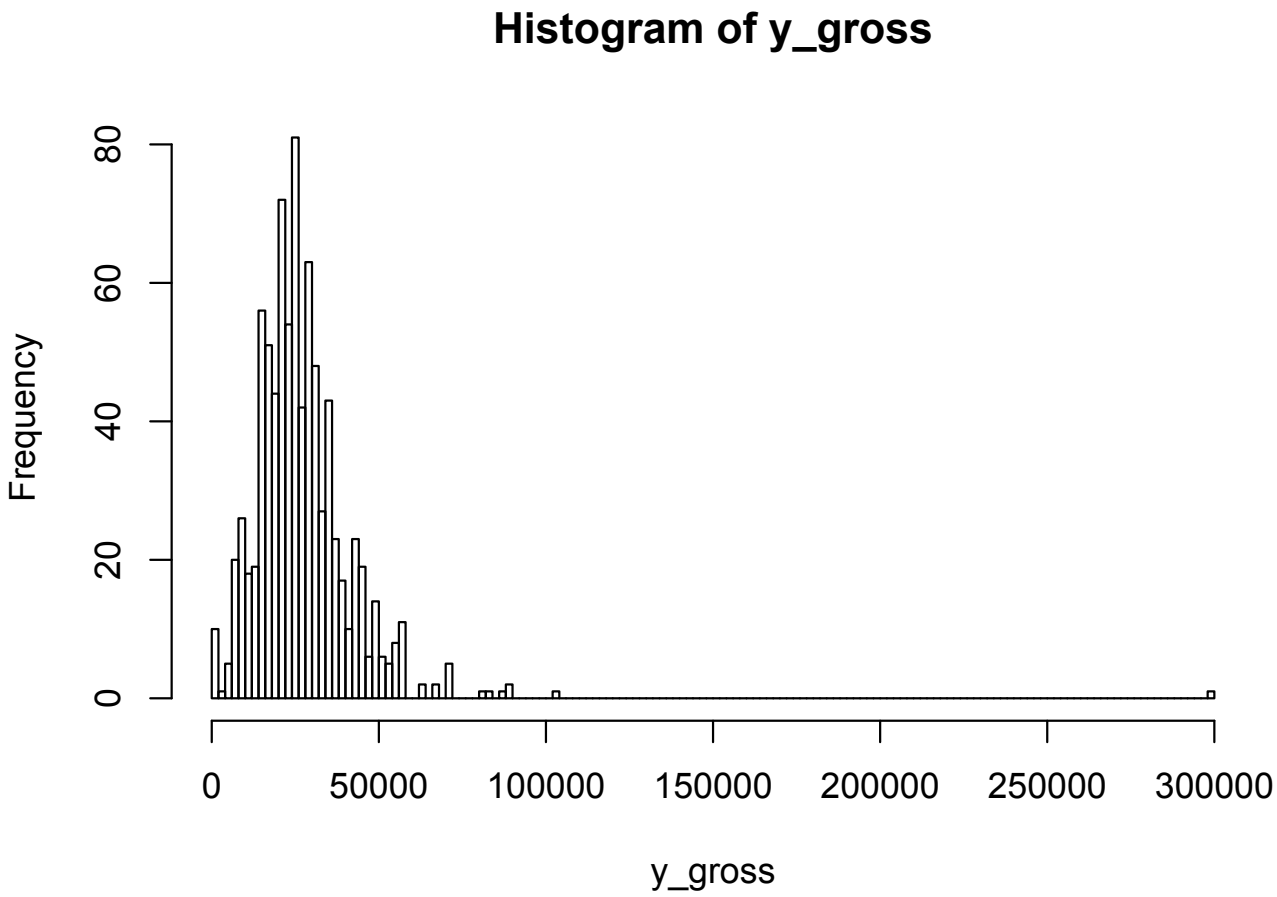
Blox plot


BOX PLOT
età
80
70
60
50
40
30
20
75%+1.5 SI
75%
Mediana
50%
scarto
interquartile
SI
25%
25%-1.5 SI

Redditi svedesi: outlier

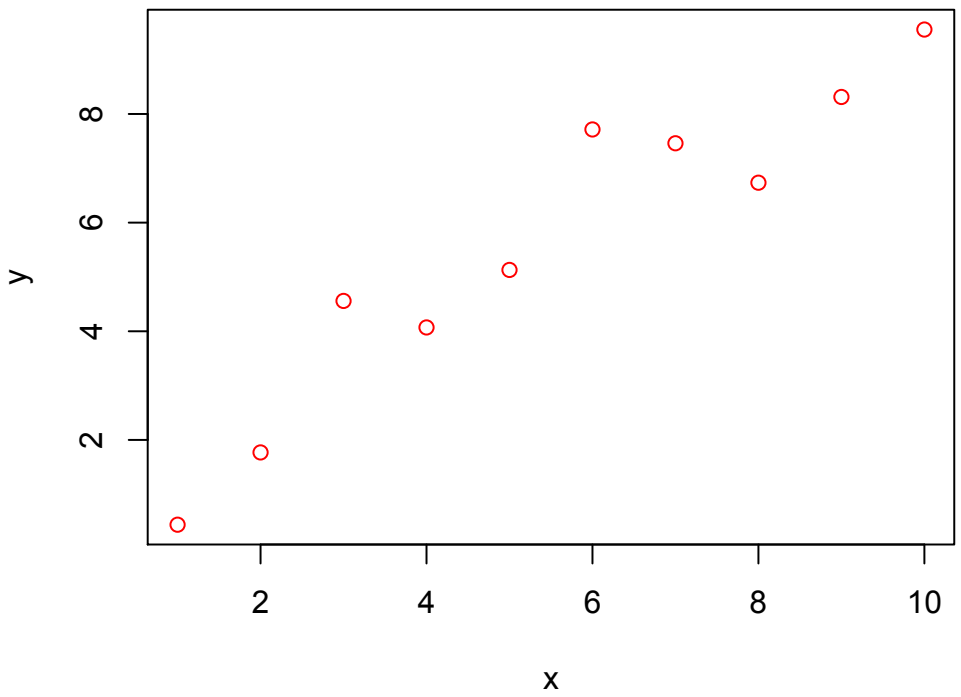


la distribuzione del reddito: istogramma

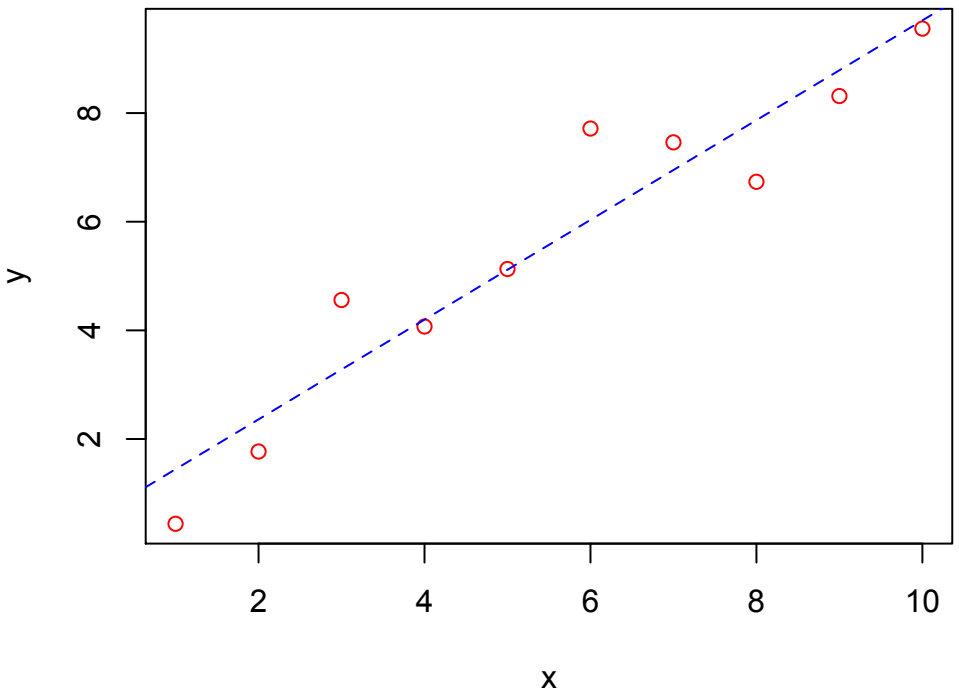


L'effetto di un outlier sulla retta di regressione

Immaginate due variabili legate liearmente



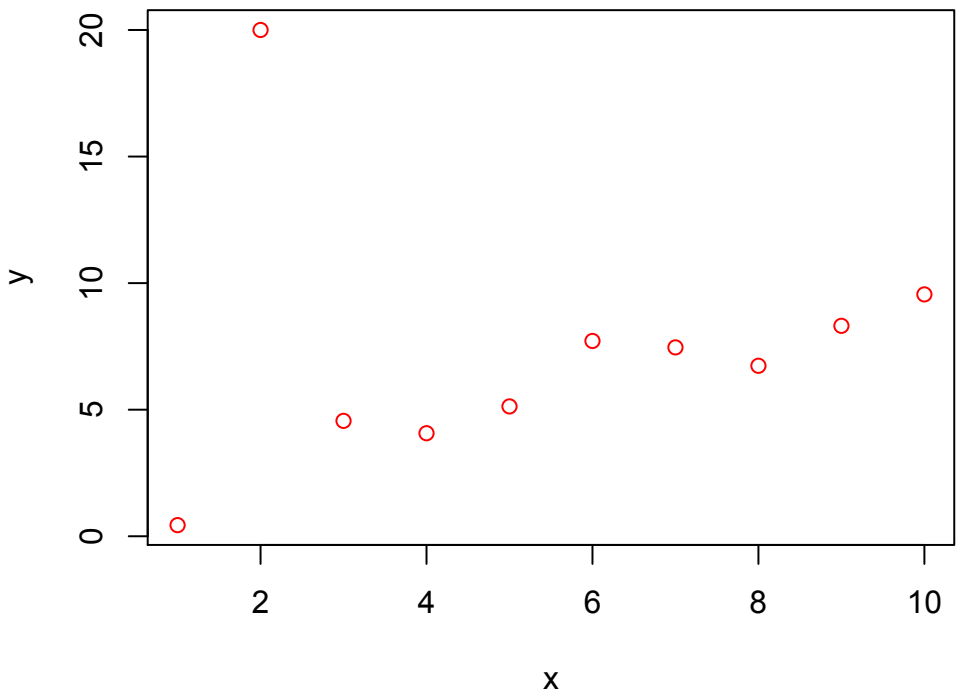
L'effetto di un outlier sulla retta di regressione



La retta di regressione interpola bene i dati

L'effetto di un outlier sulla retta di regressione

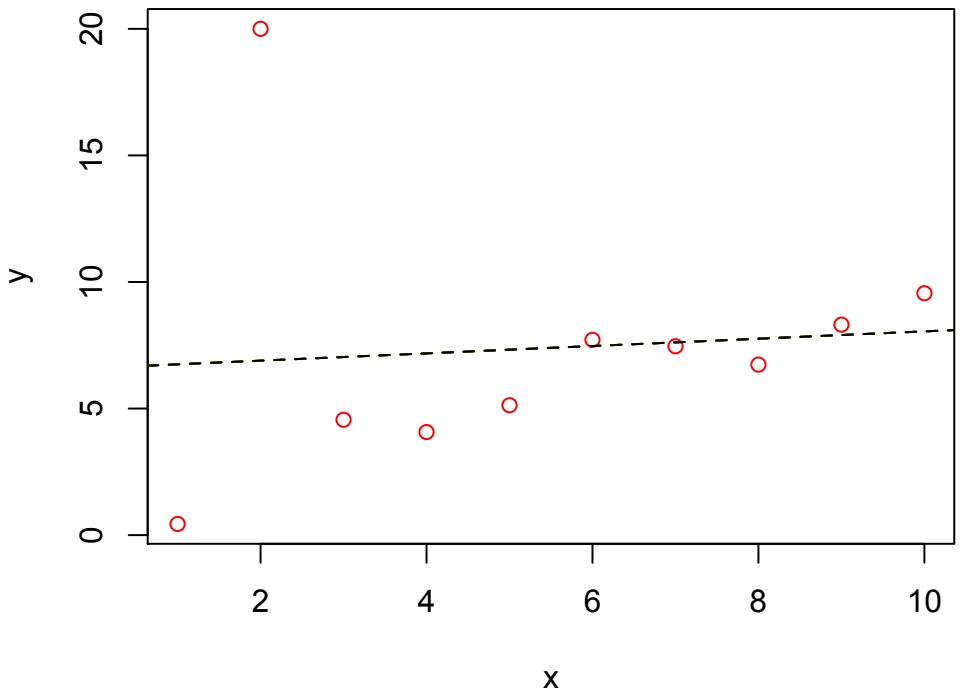
manipoliamo i dati in modo che il valore dell'osservazione che ha valore x=2 abbia valore y=20



Si tratta chiaramente di un outlier

L'effetto di un outlier sulla retta di regressione

la retta stimata è estremamente sensibile alla presenza dell'outlier



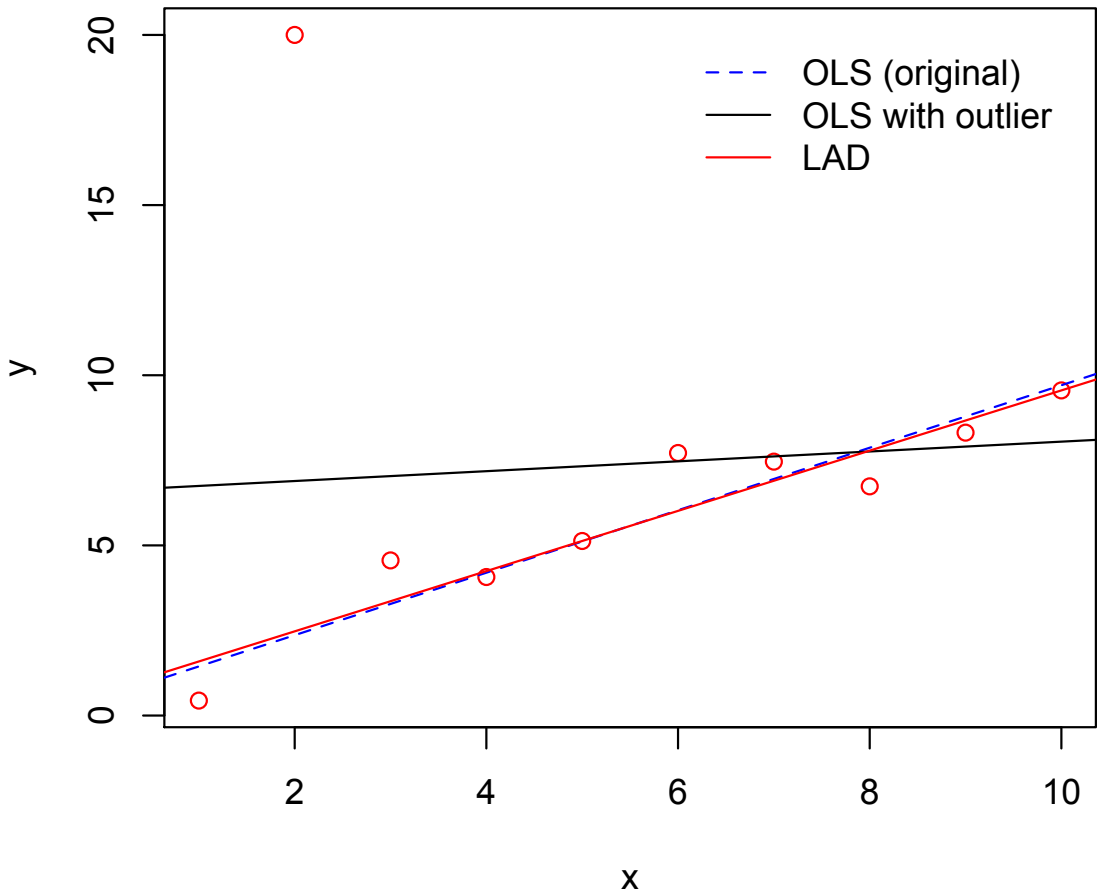
stimatore delle minime deviazioni assolute

- se gli outlier non sono rari
- è più efficiente uno stimatore meno sensibile ai valori u_i elevati
- lo stimatore delle minime deviazioni assolute (Least Absolute Deviations - LAD) ottenuto derivando b_0, b_1 che minimizzano:

$$\sum_{i=1}^{n} | Y_i - b_0 - b_1 X_i|$$

L'effetto di un outlier sulla retta di regressione

Se invece si utilizza la stima LAD si ottiene una retta praticamente identica a quella stimata sui dati originari

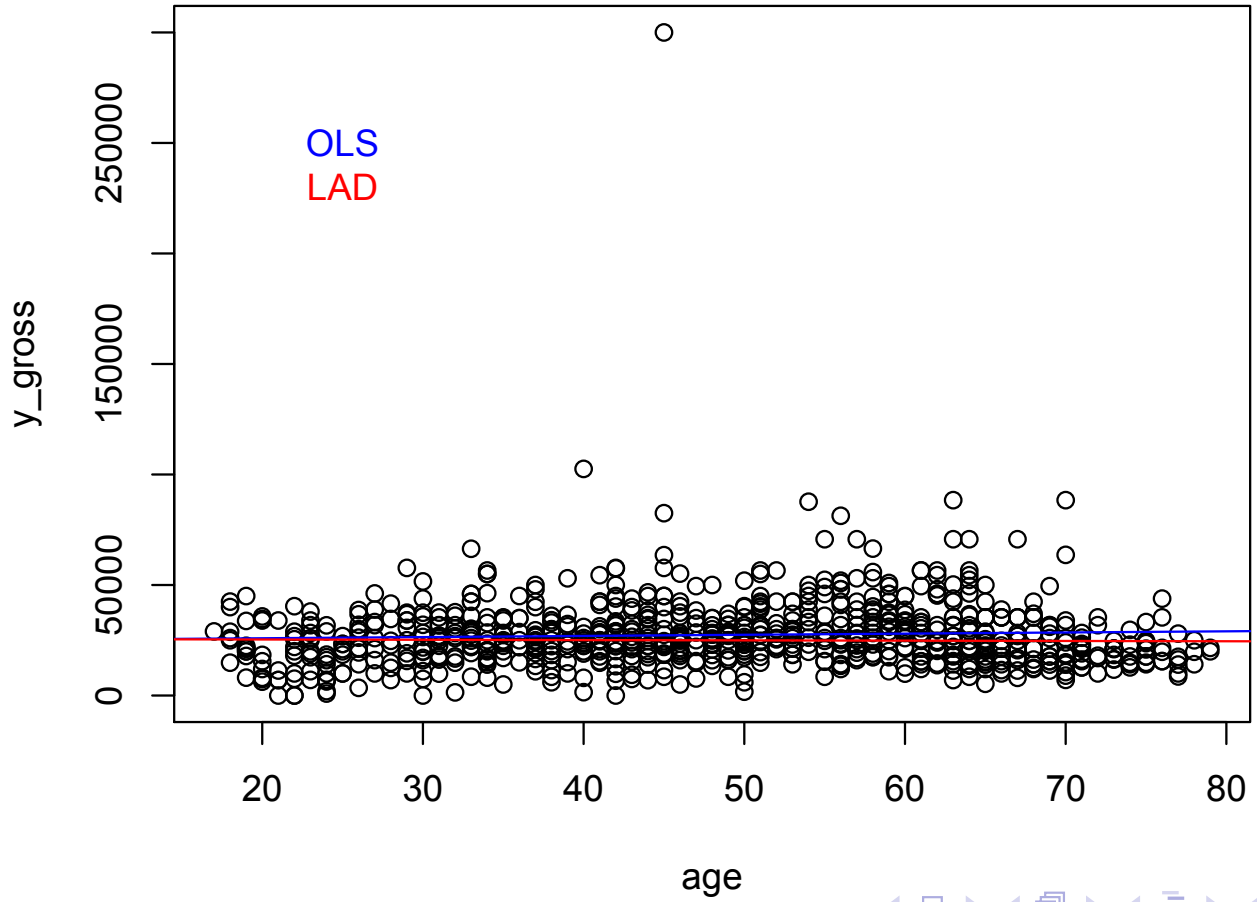


Redditi svedesi

- il dataset contiene i dati di reddito di 838 individui
- il dataset contiene le variabili:
  - sex = sesso
  - age = età
  - edu = anni di istruzione
  - y_gross = reddito lordo

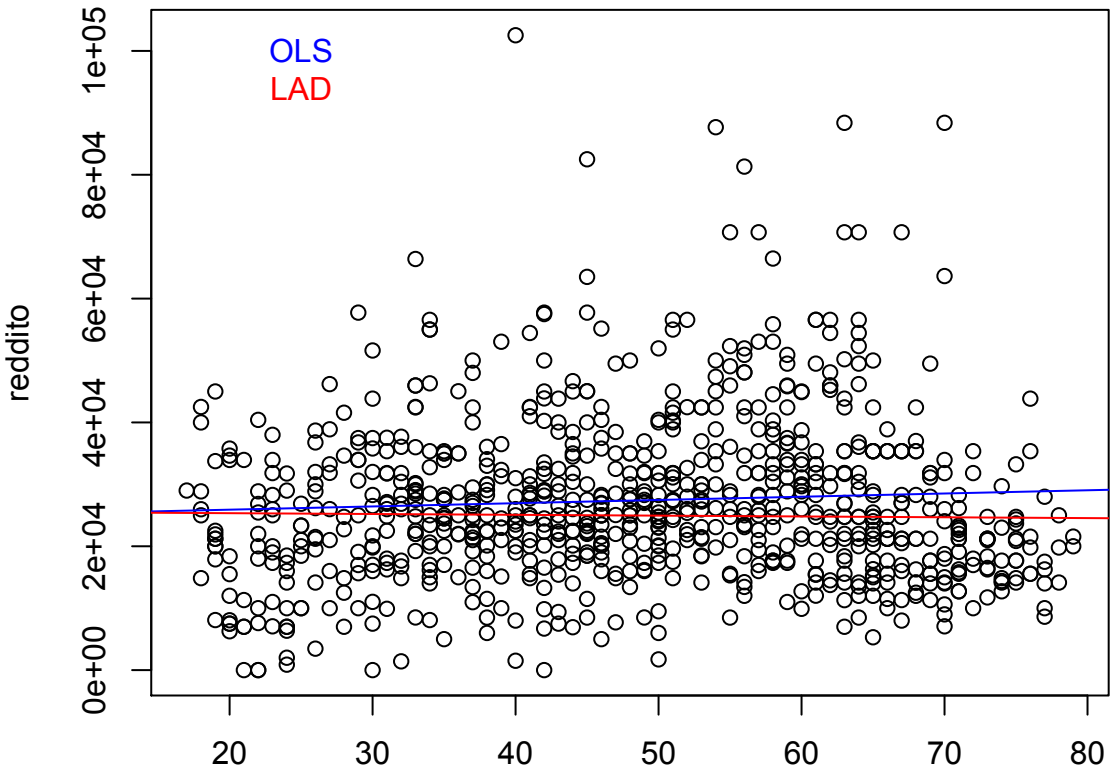
Redditi svedesi: outlier

Il caso dei dati di reddito svedesi è simile qui la distorsione dei coefficienti è molto meno marcato perché l'outlier è solo uno su oltre 800 osservazioni.



Redditi svedesi: outlier

la differenza è meglio apprezzabile se si mostra nel grafico la nuvola di punti escludendo l'outlier



attenzione: un outlier è stato rimosso dal grafico