Metodi statistici per la ricerca sociale

Capitolo 14. Regressione multipla: la scelta dei modelli

Alessandra Mattei

Dipartimento di Statistica, Informatica, Applicazioni (DiSIA)
Università degli Studi di Firenze
mattei@disia.unifi.it

LM–88 SOCIOLOGIA E RICERCA SOCIALE

Diagnostiche di regressione

- Le proprietà degli stimatori dei parametri del modello di regressione valgono se sono soddisfatte le assunzioni che sono alla base del modello
- Se una o più assunzioni sono violate si dice che il modello è **mal specificato**
- Obiettivo: Diagnostiche che permettono di stabilire se
  - ✓ Le assunzioni del modello di regressione sono violate (errata specificazione del modello)
  - ✓ Esistono osservazioni che influenzano pesantemente l'adattamento del modello e le inferenze: valori di leva (*leverage*), osservazioni *influenti*, *valori anomali/outliers*

Il modello di regressione lineare: Assunzioni

Le osservazioni $y_1, \ldots, y_i, \ldots, y_n$ sono realizzazioni di variabili aleatorie $Y_1, \ldots, Y_i, \ldots, Y_n$ Normali indipendenti aventi media che è funzione lineare delle variabili esplicative $X_1, \ldots, X_K$, e varianza costante indipendentemente dal valore delle variabili esplicative $X_1, \ldots, X_K$:

Per $i = 1, \ldots, n$

$$Y_i | X_{i1} = x_{i1}, \ldots X_{iK} = x_{iK} \sim N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK}, \sigma^2) \quad \text{indipendenti}$$

Quindi, per ogni $i = 1, \ldots, n$

$$\mathbb{E}(Y_i | X_{i1} = x_{i1}, \ldots X_{iK} = x_{iK}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK}$$

e

$$Var(Y_i | X_i = x_i) = \sigma^2 \qquad \textbf{(ipotesi di omoschedasticità)}$$

Il modello di regressione lineare multipla: Assunzioni

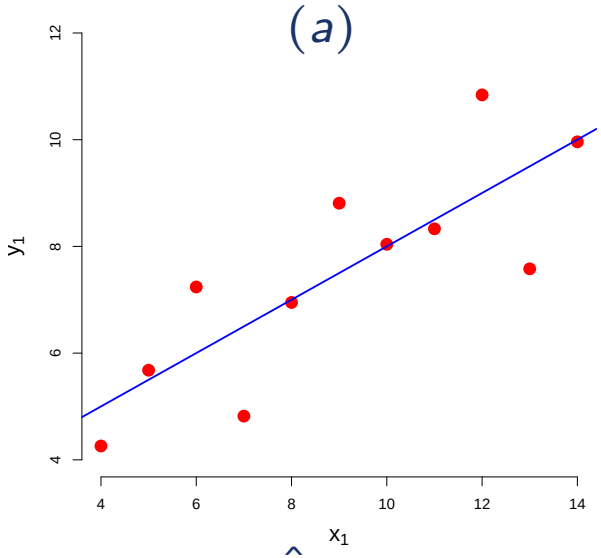
Assunzione 1. Per ogni unità $i = 1, \ldots, n$

$$Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \cdots + \beta_k \cdot x_{ik} + \epsilon_i$$

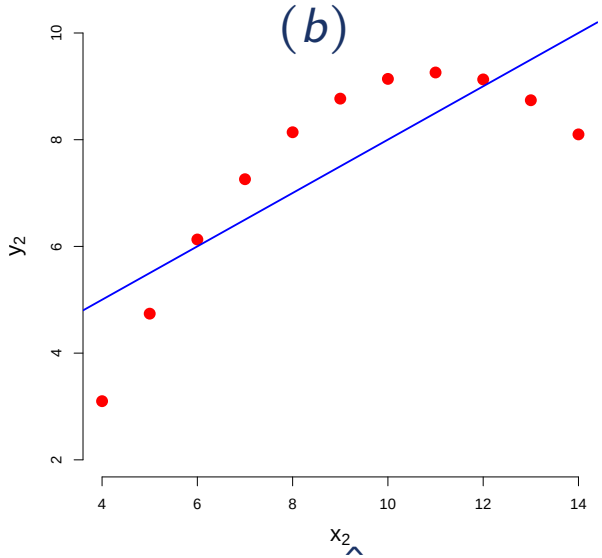
Assunzione 2. Gli errori ϵ_i, $i = 1, \ldots, n$, sono variabili aleatorie indipendenti aventi media nulla e varianza costante indipendentemente dal valore delle variabili esplicative $X_1, \ldots, X_K$:

$$\text{Per} \quad i = 1, \ldots, n \qquad \epsilon_i \sim N(0, \sigma^2) \qquad \text{indipendenti}$$

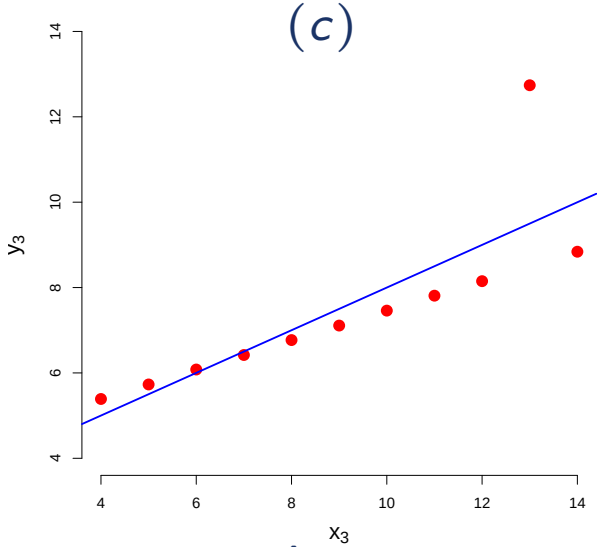
Quattro data set: Anscombe (1973)



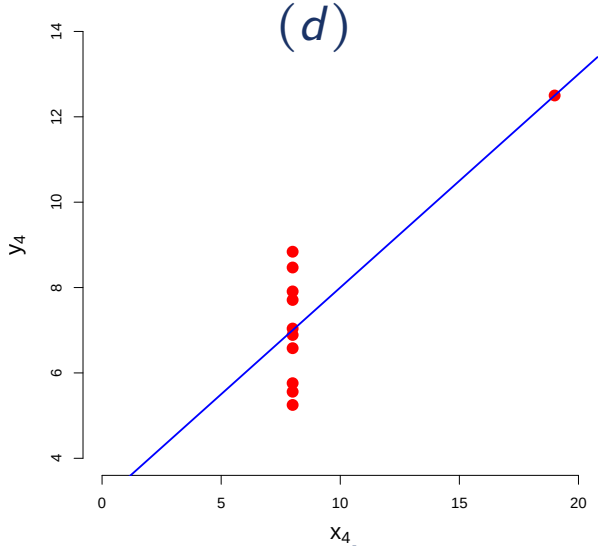
$r_{X_1,Y_1} = 0.816 \quad \hat{Y}_{i1} = 3.001 + 0.5 \cdot x_{i1}$



$r_{X_2,Y_2} = 0.816 \quad \hat{Y}_{i2} = 3.001 + 0.5 \cdot x_{i2}$



$r_{X_3,Y_3} = 0.816 \quad \hat{Y}_{i3} = 3.001 + 0.5 \cdot x_{i3}$



$r_{X_4,Y_4} = 0.816 \quad \hat{Y}_{i4} = 3.001 + 0.5 \cdot x_{i4}$

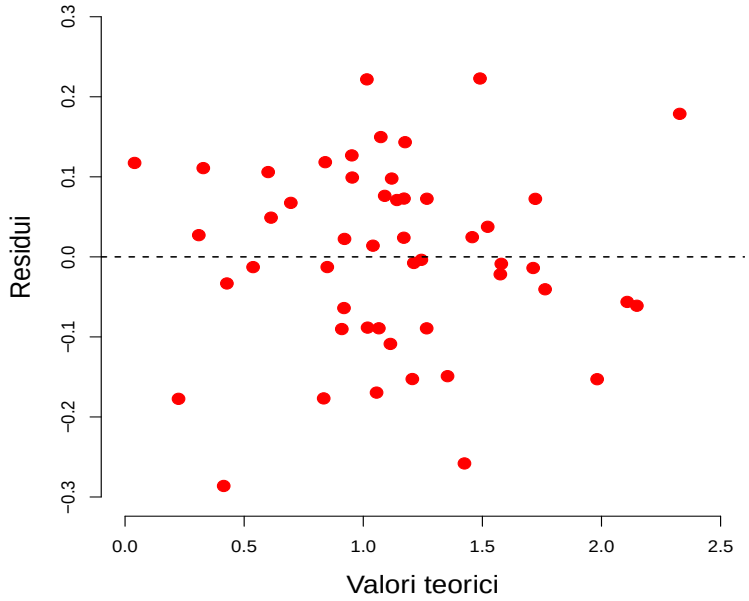
Quattro data set: Anscombe (1973)

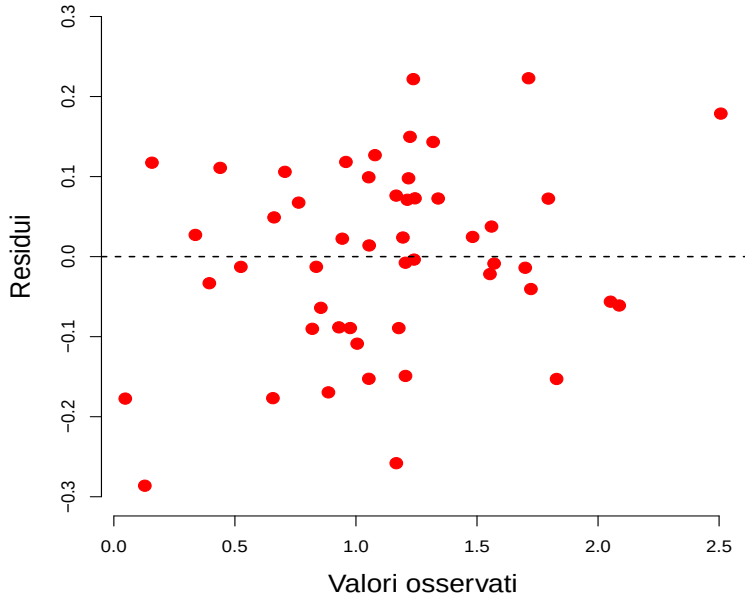
- La figura (a) suggerisce che il modello lineare può essere ragionevole
- La figura (b) suggerisce una relazione nonlineare
- La figura (c) mostra che i dati seguono bene il modello lineare a eccezione di un punto che si trova distante dalla retta di regressione stimata. Tale osservazione può rappresentare un valore anomalo che deve essere esaminato con attenzione
- La figura (d) suggerisce problemi nella scelta del campione o nel disegno dell'esperimento.

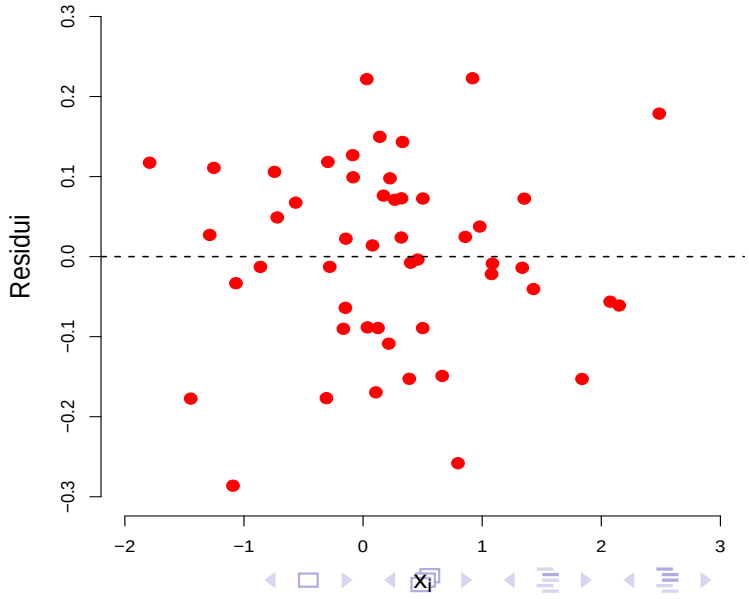
Analisi dei residui

- Una tecnica che consente di investigare sulle cause di errata specificazione del modello si basa sull'analisi dei residui
- Un metodo grafico: grafico dei residui
- Il grafico dei residui è un grafico di dispersione in cui l'asse delle ordinate è riferito ai residui e l'asse delle ascisse è riferito ai valori teorici, o ai valori osservati della variabile risposta o ai valori di ciascuna variabile esplicativa
- Se il modello è ben specificato, i residui tenderanno a distribuirsi in modo casuale attorno alla retta $\widehat{e} = 0$ senza mostrare valori anomali né tendenze di fondo o comportamenti sistematici

Il grafico dei residui in caso di corretta specificazione del modello







Tipi di residui

- Residui di regressione

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \left[\widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_{i1} + \ldots + \widehat{\beta}_k \cdot x_{ik}\right]$$

- Residui standardizzati

$$\hat{e}_i^* = \frac{\hat{e}_i}{s} \qquad \text{dove } s = \sqrt{\frac{\sum_i \hat{e}_i^2}{n-k-1}}$$

- Residui studentizzati

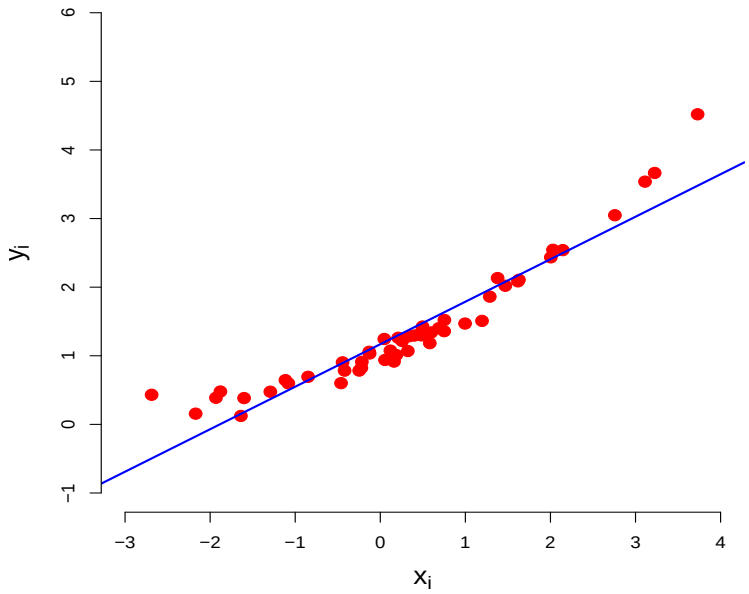
$$\tilde{\hat{e}}_i = \frac{\hat{e}_i}{\widehat{e.s.}(\hat{e}_i)}$$

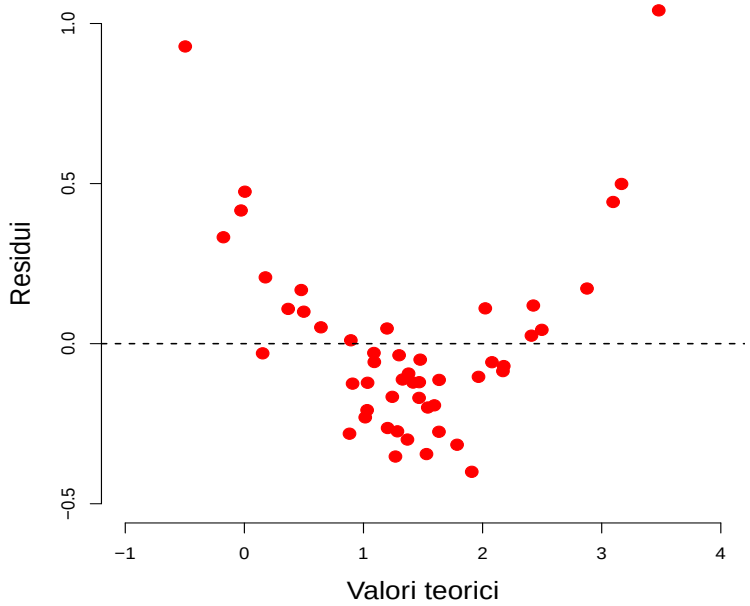
Linearità

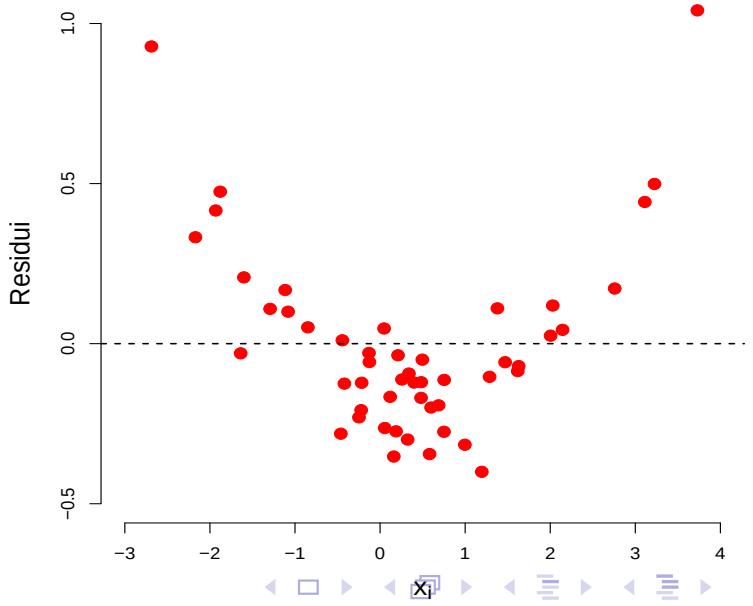
In caso di relazione non lineare:

- I parametri del modello di regressione perdono di significato
- Le stime del valor medio per un dato valore della variabile esplicativa potrebbero risultare fortemente distorte

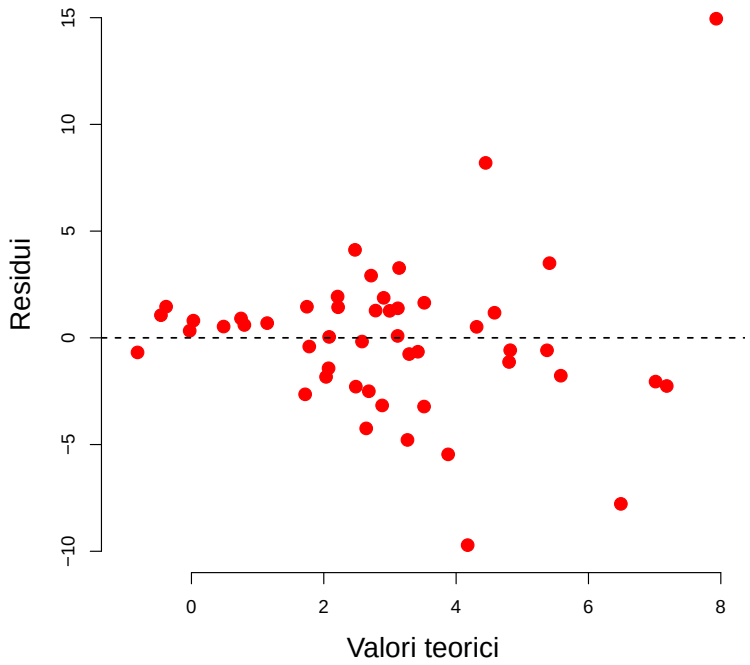
Linearità

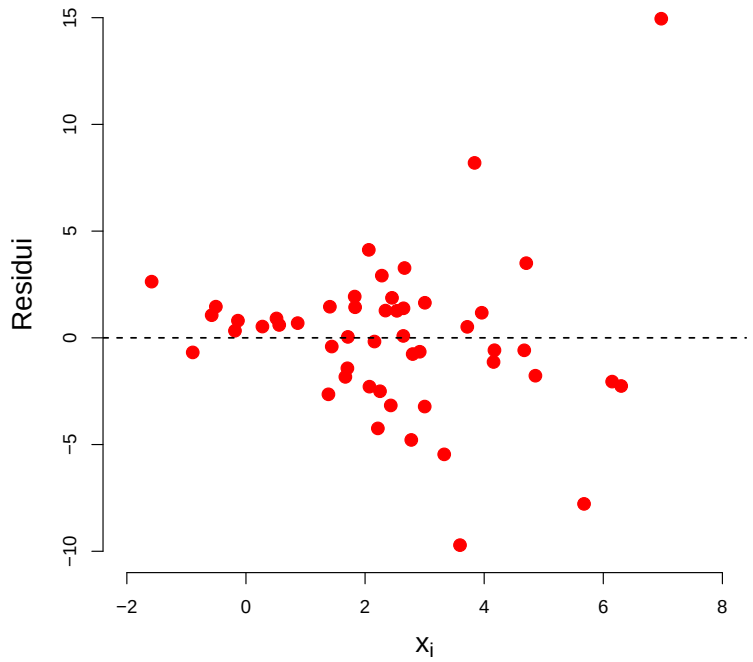






Eteroschedasticità





Eteroschedasticità

In presenza di eteroschedasticità

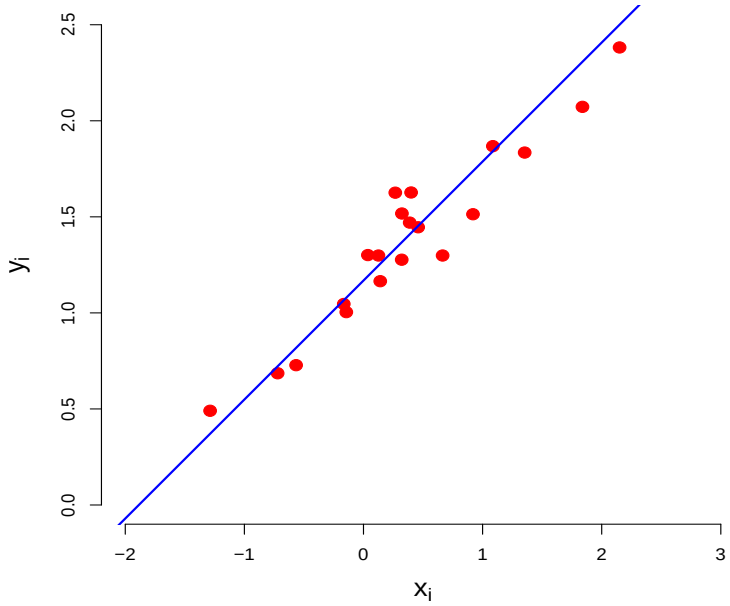
- Gli stimatori dei coefficienti di regressione e della risposta media rimangono corretti, ma perdono di efficienza;
- Le formule della varianza e dell'errore standard degli stimatori non sono più validi e il loro uso conduce ad intervalli di confidenza errati ed a verifiche di ipotesi non affidabili

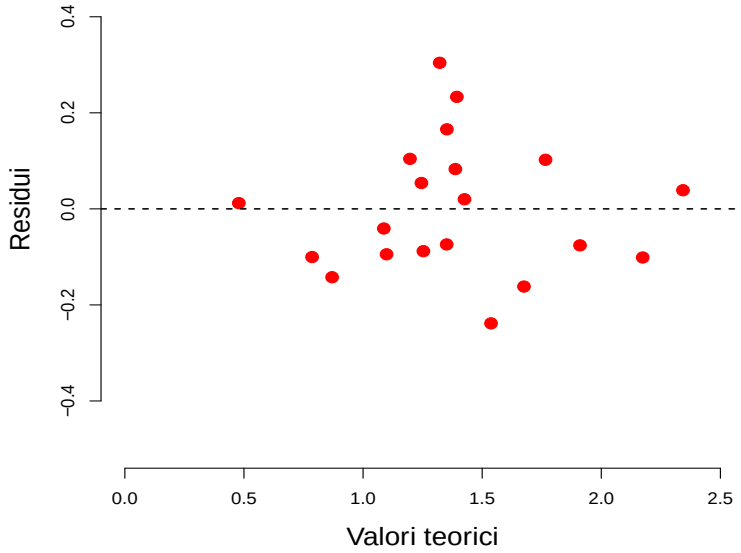
Per ovviare a tali incovenienti è necessario utilizzare stimatori dei parametri di regressione diversi dagli stimatori dei minimi quadrati standard

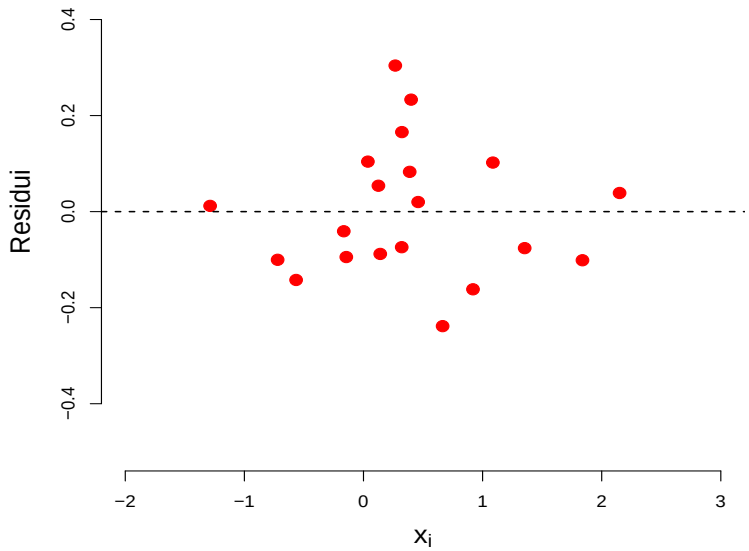
Assunzione di indipedenza

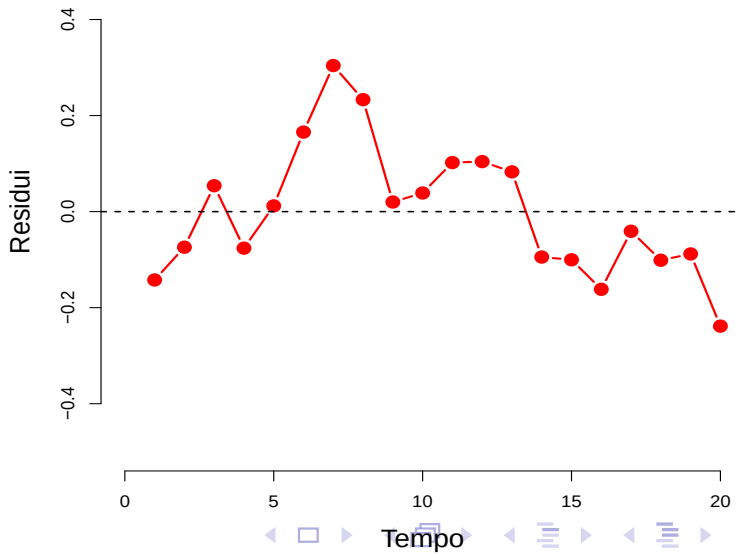
- Se le osservazioni sono in una sequenza temporale, in genere gli errori non sono indipendenti
- Grafico dei residui rispetto al tempo
- I residui contigui tendono ad assumere stesso segno: **autocorrelazione positiva**
  - ✓ Sul grafico i residui mostrano comportamenti ciclici intorno allo zero
- I residui contigui tendono ad assumere segno opposto: **autocorrelazione negativa**
  - ✓ Sul grafico i residui tendono sistematicamente a cambiare segno
- In presenza di (auto-)correlazione (negativa o positiva) tra i termini di errore
  - ✓ Gli stimatori dei minimi quadrati dei coefficienti di regressione rimangono corretti ma perdono di efficienza
  - ✓ Le procedure inferenziali che ignorano l'autocorrelazione possono portare a risultati fuorvianti e errati

Assunzione di indipedenza









Assunzione di Normalità

Retta di regressione

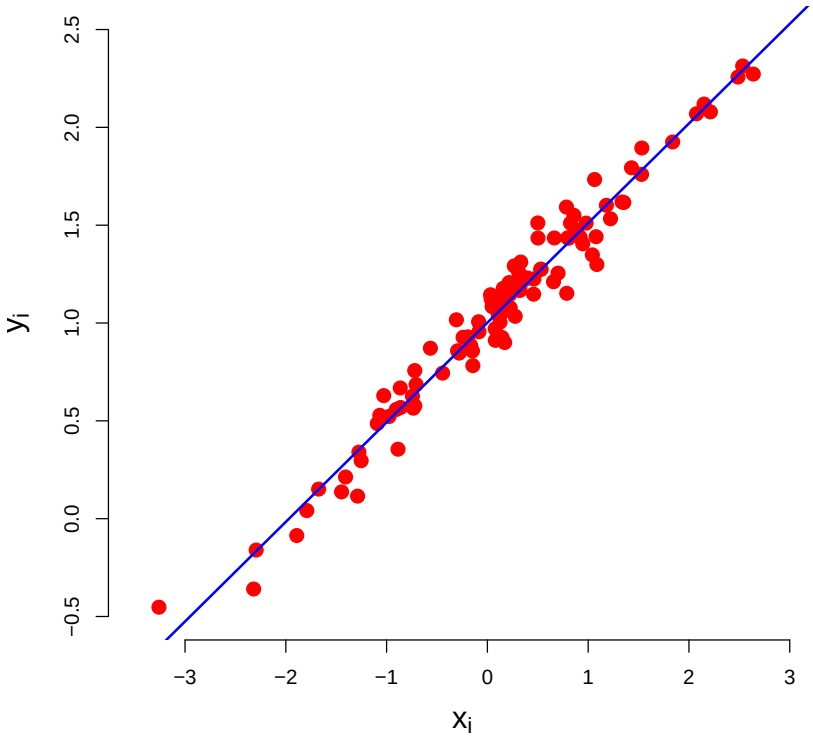
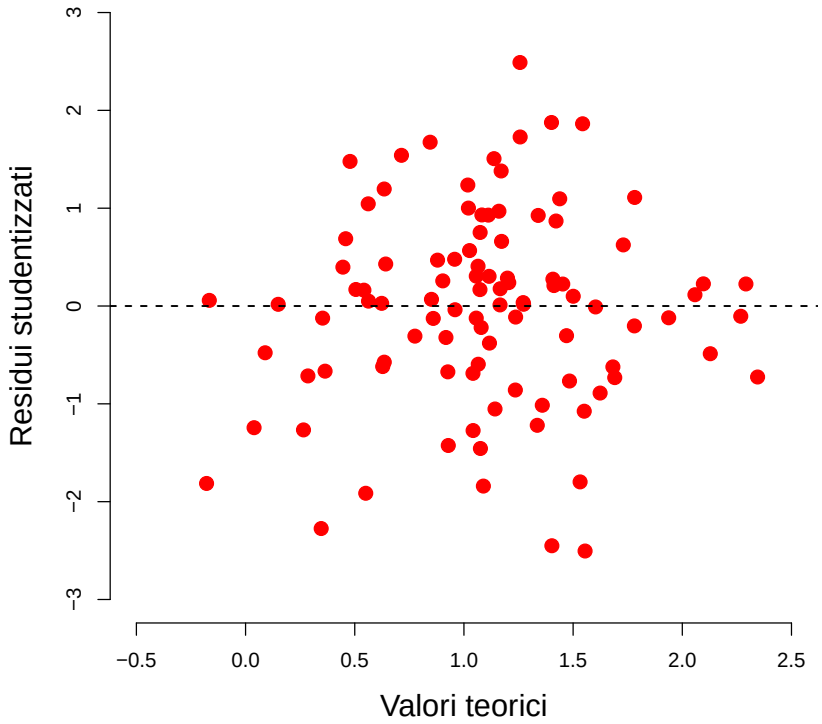
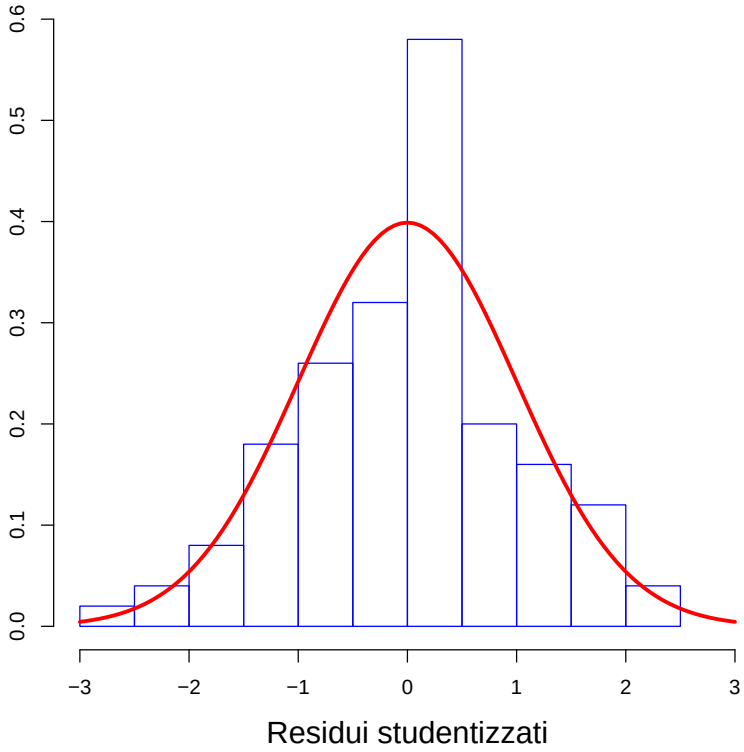


Grafico dei residui studentizzati

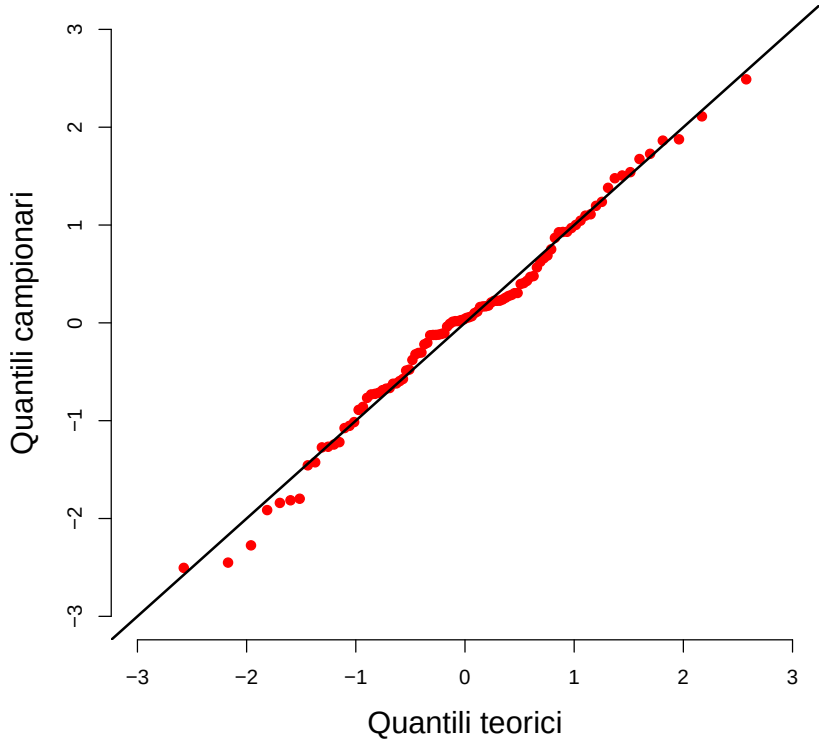


Assunzione di Normalità

Istogramma dei residui studentizzati



Q-Q plot dei residui studentizzati



Osservazioni particolari: valori anomali

- Un valore anomalo è un'osservazione che presenta caratteristiche diverse dal resto dei dati
- Data una generica osservazione (x_i, y_i), questa rappresenta un valore anomalo se
  - ✓ tra le unità che presentano lo stesso valore della variabile esplicativa, l'unità in esame presenta un valore della variabile risposta molto diverso
  - ✓ il valore della variabile esplicativa, x_i è *isolato* dal resto dei valori della variabile esplicativa osservati

Osservazioni particolari

Valori anomali (outlier) nella variabile risposta

- Osservazioni con residui standardizzati o studentizzati elevati (valori anomali della variabile risposta, Y)

Valori di leva (leverage)

- Osservazioni con valori delle variabili esplicative distanti dalle rispettive medie (Valori anomali nelle variabile esplicative)
- Le osservazioni di leva non necessariamente presentano residui elevati

Osservazioni Influenti

- Osservazioni con comportamento anomalo che influenzano notevolmente i risultati: Un'osservazione è influente se la sua rimozione comporta un cambiamento notevole nelle stime dei parametri e/o in R^2
- Non tutti i valori anomali nella variabile risposta e i valori di leva sono necessariamente osservazioni influenti

Valori anomali: Esempio - Abilità dei neonati

- Per 21 bambini è rilevata l'età in mesi in cui è stata pronunciata la prima parola (X) e il punteggio ad un test di abilità (Y)
- Statistiche descrittive

$$\bar{x} = 14.381 \qquad \bar{y} = 93.667$$

$$s_x^2 = 63.148 \qquad s_y^2 = 195.633 \qquad s_{x,y} = -71.167$$

- Modello di regressione lineare

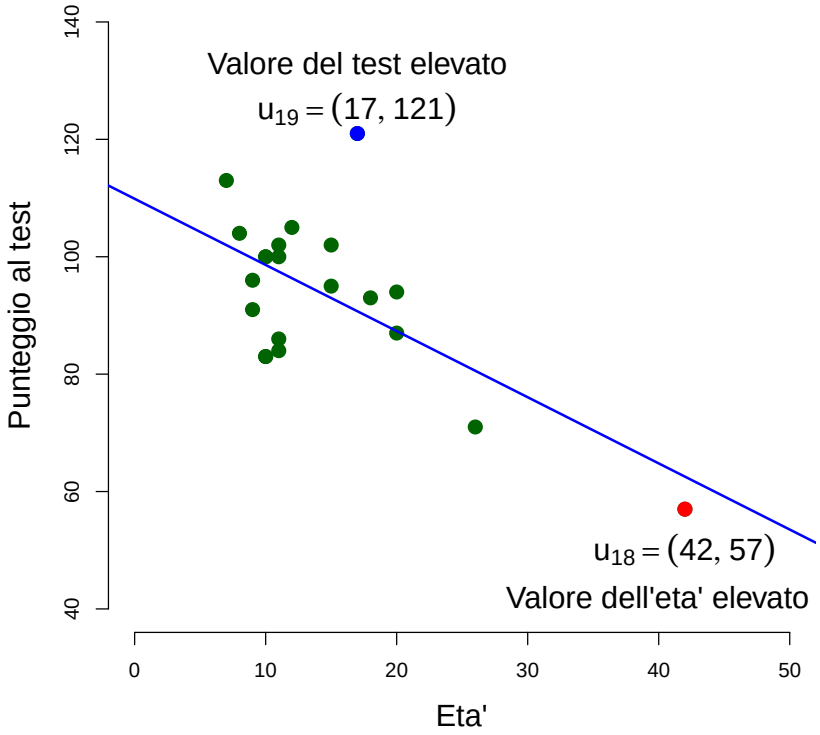
$$Test_i = \beta_0 + \beta_1 \cdot Età_i + \epsilon_i$$

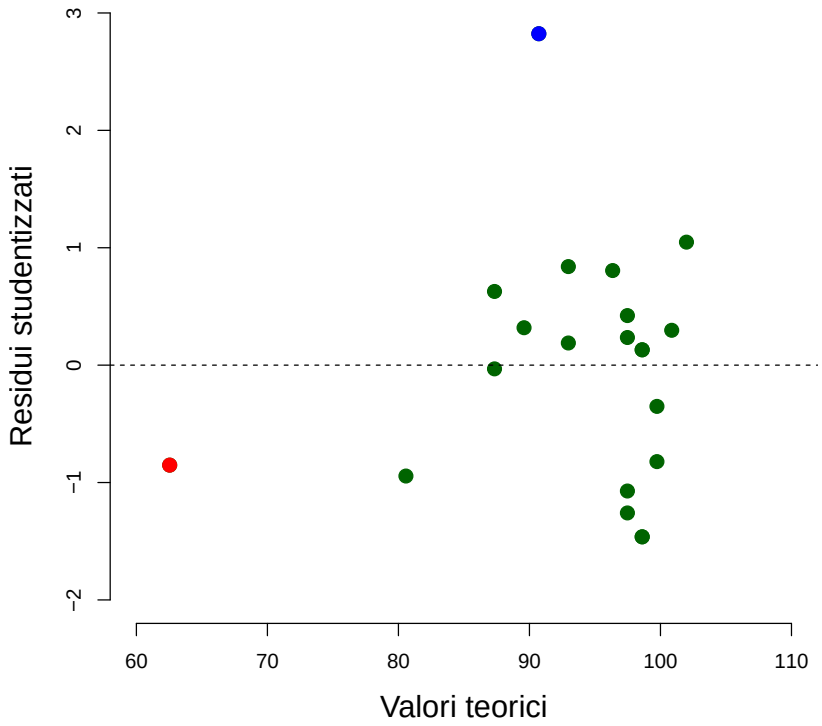
$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_i = 109.874 - 1.127 \cdot x_i \qquad s = 11.02 \qquad \text{e} \qquad R^2 = 0.41$$

u_i	Età$_i$	Test$_i$	$\widehat{y}_i$	$\widehat{e}_i$	$\widehat{e}_i^*$	$\widetilde{\widehat{e}}_i$
1	15	95	92.97	2.03	0.18	0.19
2	26	71	80.57	−9.57	−0.87	−0.94
3	10	83	98.60	−15.60	−1.42	−1.46
4	9	91	99.73	−8.73	−0.79	−0.82
5	15	102	92.97	9.03	0.82	0.84
6	20	87	87.33	−0.33	−0.03	−0.03
7	18	93	89.59	3.41	0.31	0.32
8	11	100	97.48	2.52	0.23	0.24
9	8	104	100.86	3.14	0.29	0.30
10	20	94	87.33	6.67	0.60	0.63
11	7	113	101.98	11.02	1.00	1.05
12	9	96	99.73	−3.73	−0.34	−0.35
13	10	83	98.60	−15.60	−1.42	−1.46
14	11	84	97.48	−13.48	−1.22	−1.26
15	11	102	97.48	4.52	0.41	0.42
16	10	100	98.60	1.40	0.13	0.13
17	12	105	96.35	8.65	0.78	0.81
18	42	57	62.54	−5.54	−0.50	−0.85
19	17	121	90.72	30.28	2.75	2.82
20	11	86	97.48	−11.48	−1.04	−1.07
21	10	100	98.60	1.40	0.13	0.13

Valori anomali: Esempio - Abilità dei neonati

$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} \cdot x_i = 109.874 - 1.127 \cdot x_i \quad s = 11.02$ con $gdl = 19$ e $R^2 = 0.41$





Punti di leva

- Misura di leveraggio nella regressione lineare semplice

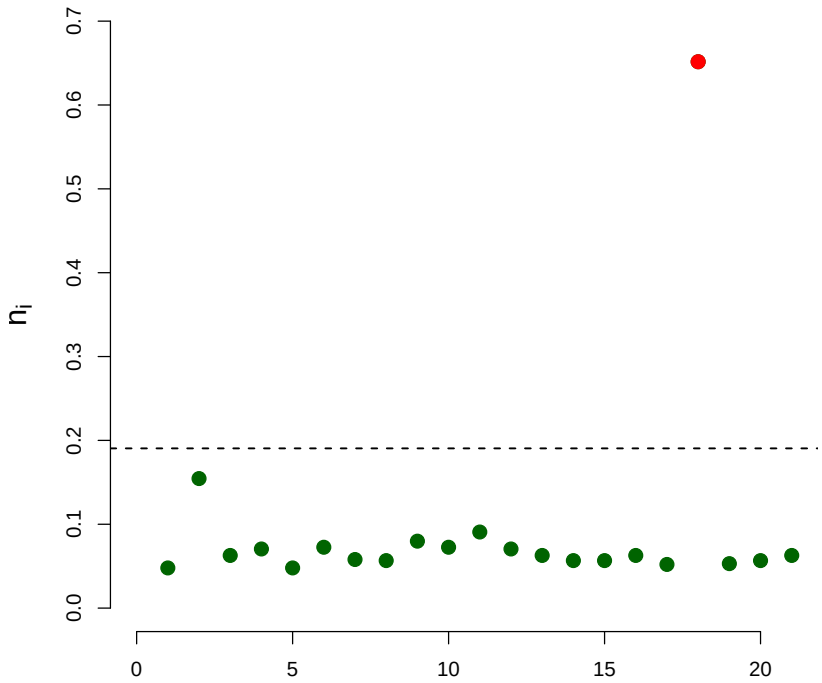
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Nella regressione lineare multipla il cacolo dei valori h_i è più complesso
- Media dei valori h_i: $(k+1)/n$
  - ✓ Nella regressione lineare semplice: $(k+1)/n = 2/n$
- Valori di h_i maggiori di $2 \cdot (k+1)/n$ indicano valori di leva molto elevati

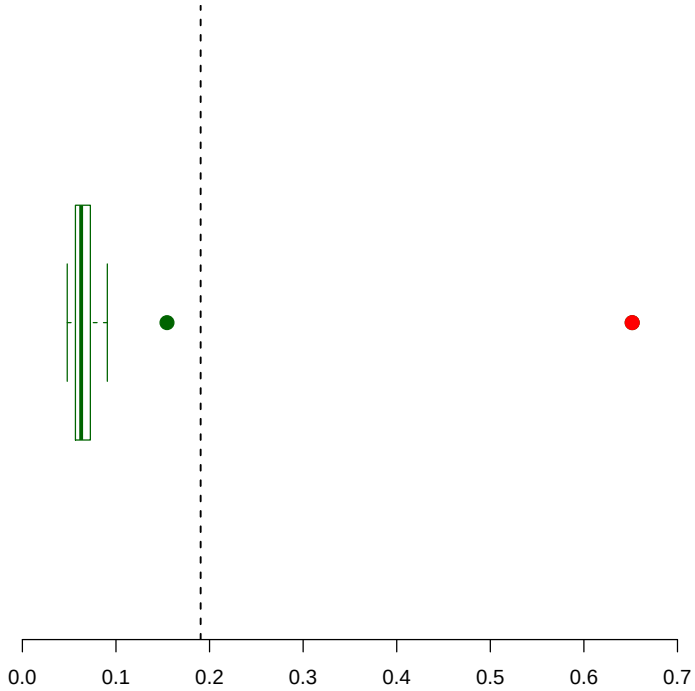
Punti di leva: Esempio - Abilità dei neonati

$$\text{Media dei valori } h_i : \frac{2}{21} = 0.095 \Longrightarrow 2 \cdot \text{Media dei valori } h_i = 0.19$$

Dot-plot dei valori h_i



Box-plot dei valori h_i



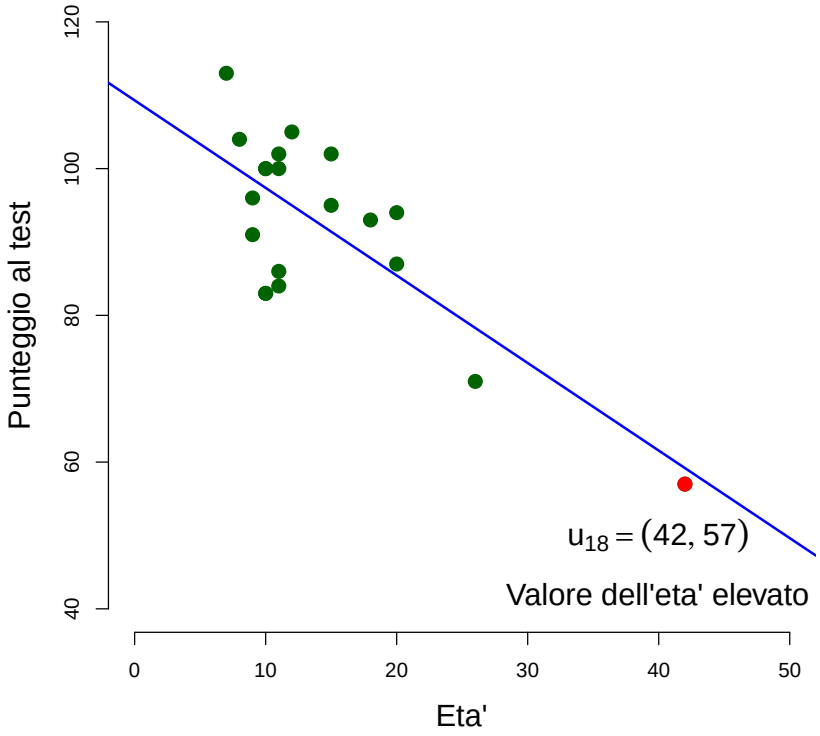
Cosa fare in presenza di valori anomali

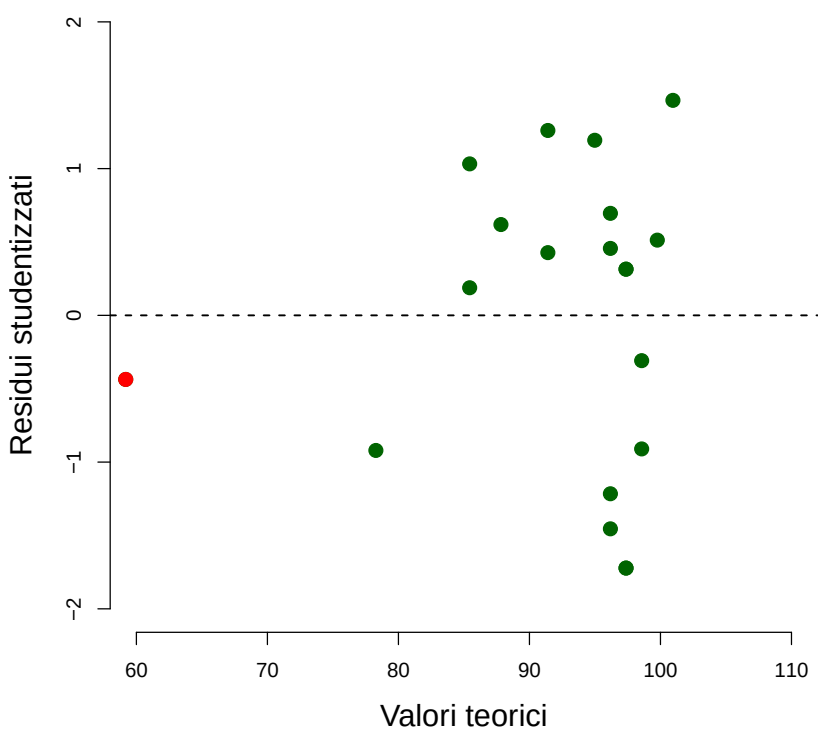
- La presenza di valori anomali può avere degli effetti rilevanti sulla regressione alterando i risultati dell'analisi.
- Gli effetti della presenza di valori anomali possono essere analizzati confrontando il modello stimato su tutti i dati con il modello stimato eliminando preliminarmente i dati che si sospetta essere anomali
- L'effetto di disturbo sulla regressione dovuto alla presenta di valori anomali potrebbe essere neutralizzato eliminando i valori anomali.
- Prima di eliminare i valori anomali è importante comprendere la ragione della loro anomalia
- In assenza di errori di rilevazione, la presenta di valori anomali indica che il modello non è in grado di prevedere adeguatamente tutti i dati

Valori anomali: Esempio - Abilità dei neonati

Eliminando l'osserzione $u_{19} = (17, 121)\ldots$

$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} \cdot x_i = 109.305 - 1.193 \cdot x_i \quad s = 8.628$ con $gdl = 18 \quad$ e $\quad R^2 = 0.5716$

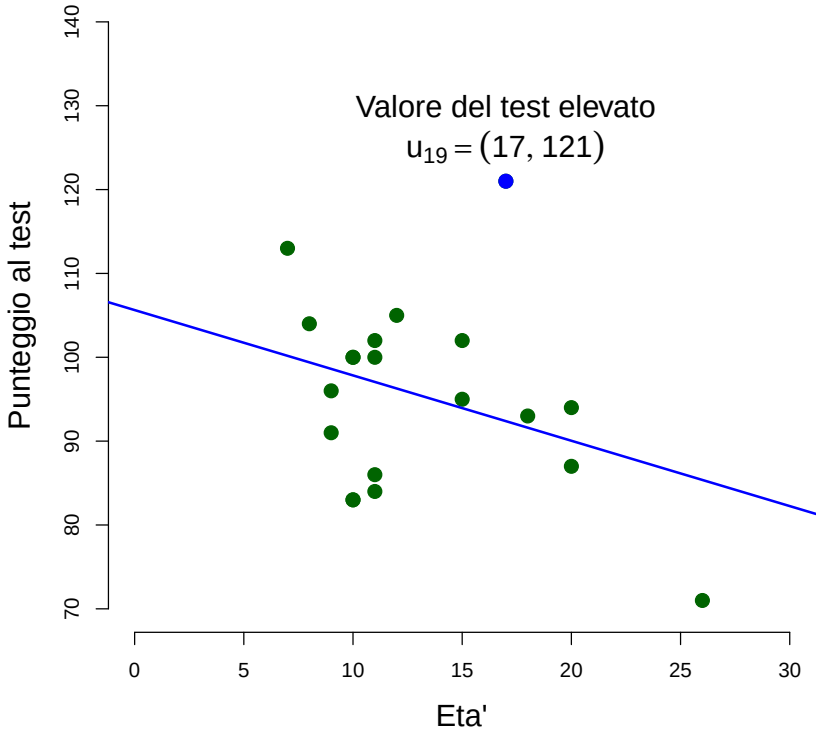


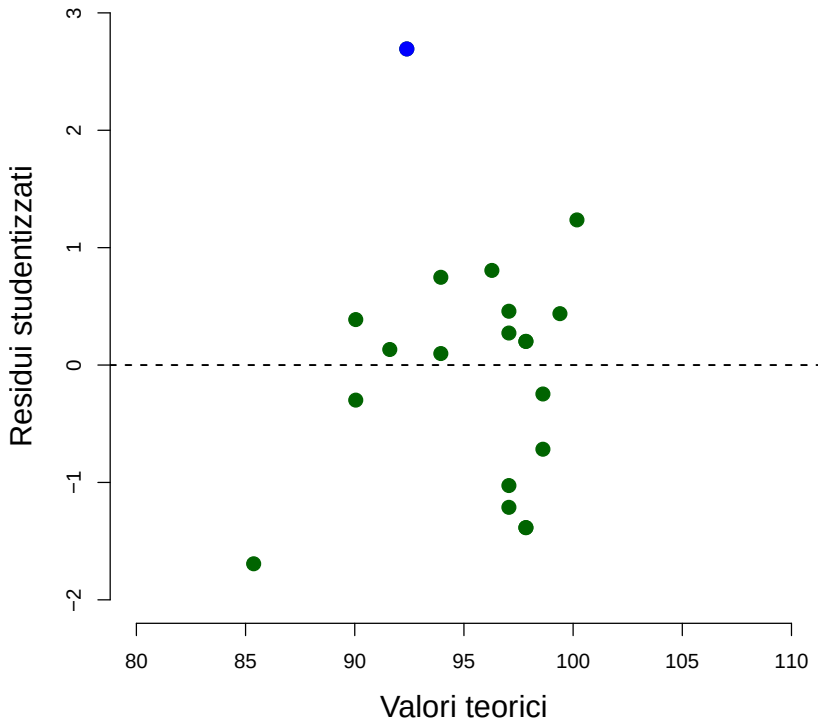


Valori anomali: Esempio - Abilità dei neonati

Eliminando l'osserzione $u_{18} = (42, 57) \ldots$

$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} \cdot x_i = 105.630 - 0.779 \cdot x_i \quad s = 11.11$ con $gdl = 18 \quad$ e $\quad R^2 = 0.1122$

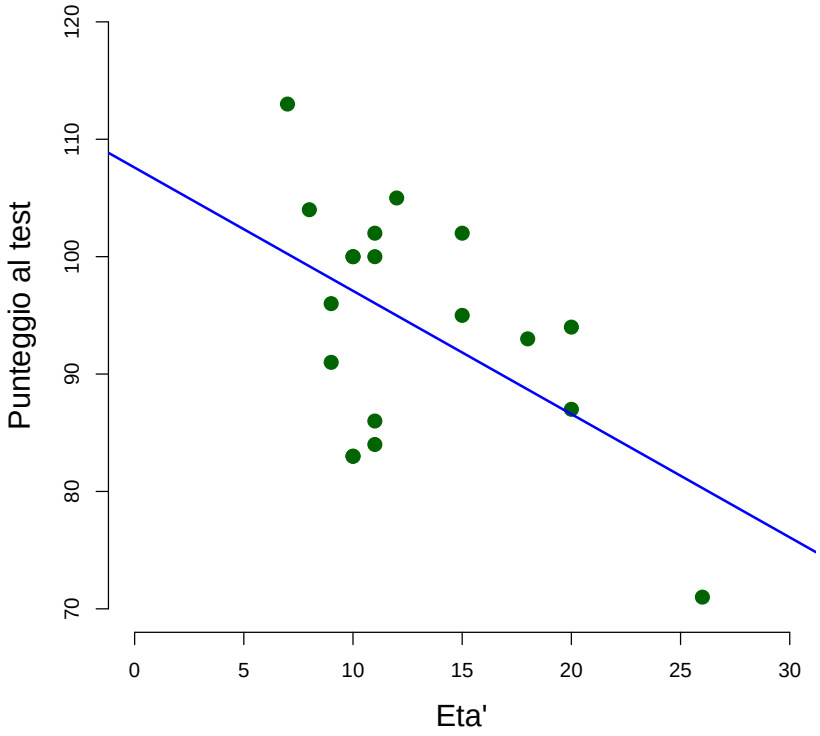


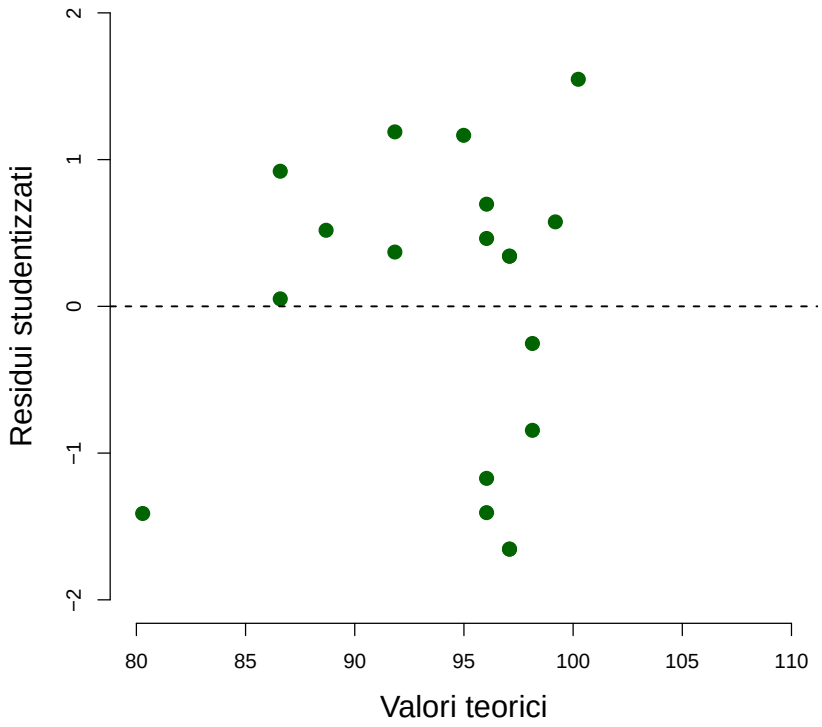


Valori anomali: Esempio - Abilità dei neonati

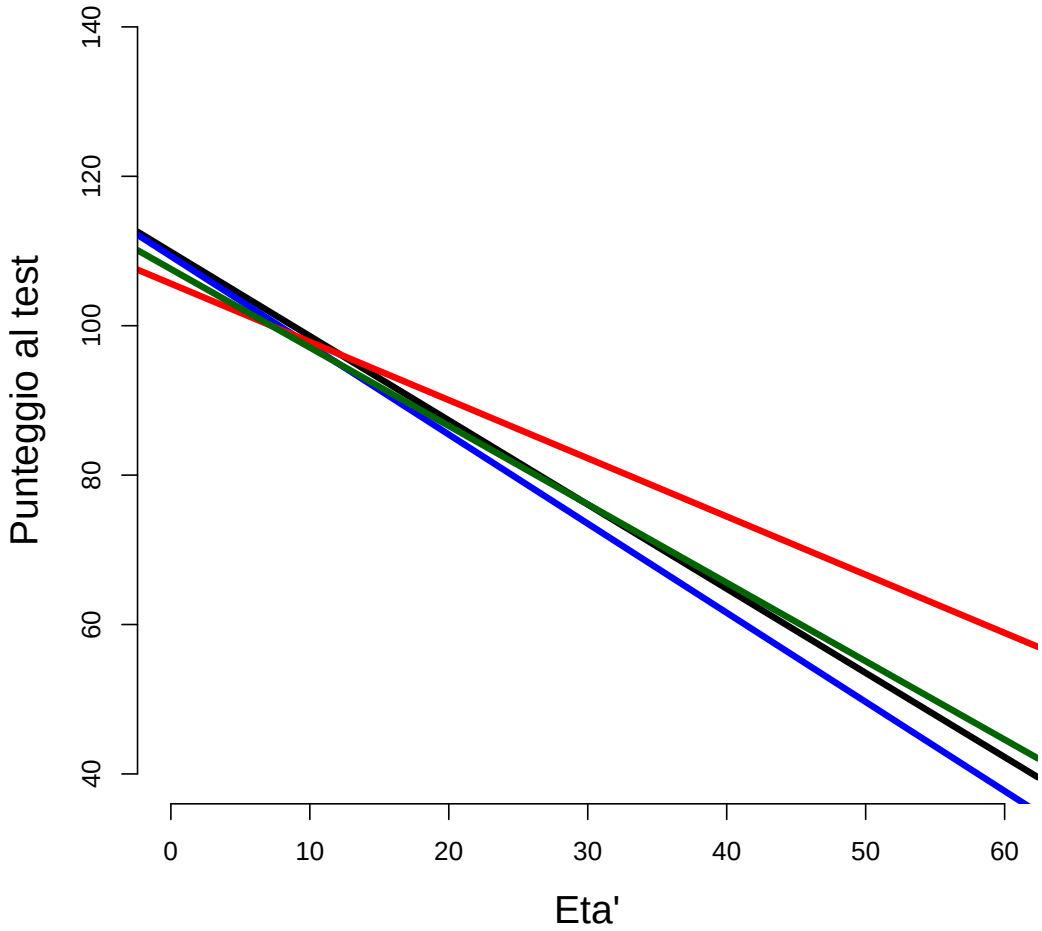
Eliminando le osserzioni $u_{18} = (42, 57)$ e $u_{19} = (17, 121)\ldots$

$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_i = 107.586 - 1.05 \cdot x_i \quad s = 8.831$ con $gdl = 17 \quad$ e $\quad R^2 = 0.2701$





Valori anomali: Esempio - Abilità dei neonati



- $n = 21$: Tutte le osservazioni
- $n = 20$: Eliminata l'osservazione $u_{18} = (42, 57)$
- $n = 20$: Eliminata l'osservazione $u_{19} = (17, 121)$
- $n = 20$: Eliminata le osservazioni $u_{18} = (42, 57)$ e $u_{19} = (17, 121)$

Multicollinearità

- Esiste **multicollinearità** in presenza di variabili esplicative tra loro molto correlate
  - ✓ Una variabile è collineare con le altre se l'indice R^2 di un modello di regressione che pone tale variabile in funzione di tutte le altre variabili esplicative è elevato
- L'effetto principale dovuto alla multicollinearità è quello di aumentare considerevolmente la varianza degli stimatori dei minimi quadrati dei coefficienti di regressione.
- L'aumento della varianza dovuto alla multicollinearità ha degli effetti negativi sull'inferenza dei coefficienti di regressione.
  - ✓ L'incremento dell'errore standard degli stimatori dei coefficienti porta all'aumento dell'ampiezza dell?intervallo di confidenza
  - ✓ L'aumento dell'errore standard fa diminuire il valore assoluto della statistica test portando più facilmente a non rifiutare l'ipotesi nulla anche se questa non è vera, ossia a commettere con maggiore probabilità un errore del secondo tipo.

VIF: Variance Inflaction Factor

- Variance inflaction factor (fattore di inflazione della varianza): indice utilizzato per misurare il livello di multicollinearità di una variabile esplicativa X_j con le altre

$$VIF_j = \frac{1}{1 - R_j^2}$$

dove R_j^2 è il coefficiente di determinazione lineare del modello di regressione nel quale la variabile X_j dipende dalle altre variabili esplicative

  - ✓ Il valore minimo del VIF_j è 1 e indica che la variabile X_j è incorrelata dalle altre
  - ✓ In genere, un valore superiore a 2 indica un livello sufficientemente alto di multicollinearità

Multicollinearità: Esempio - Numero di crimini

- Dati relativi a 141 aree metropolitane degli Stati Uniti nel 1977

  Y = Numero di crimini
  X_1 = Popolazione (in migliaia)
  X_2 = % di popolazione con 65 anni e più
  X_3 = % di popolazione con diploma superiore
  X_4 = Forza lavoro (in migliaia)

- Modello di regressione lineare

$$Y_i = \beta_0 + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \beta_3 \cdot X_{i3} + \beta_4 \cdot X_{i4} + \epsilon_i$$

Variabile	Coefficiente	E.S
Costante	−29059.26	13902.84
Popolazione (migliaia)	76.44	17.61
% popolazione di 65 anni e più	168.18	208.30
% con diploma superiore	388.49	653.44
Forza lavoro (migliaia)	−22.20	38.40

Multicollinearità: Esempio - Numero di crimini

- VIF

Variabile	R_j^2	VIF
Popolazione (migliaia)	0.995	205.956
% popolazione di 65 anni e più	0.08	1.087
% con diploma superiore	0.081	1.089
Forza lavoro (migliaia)	0.995	206.150

- la variabile Popolazione e Forza Lavoro presentano un valore molto elevato del VIF indicando presenza di multicollinearità

Multicollinearità: Esempio - Numero di crimini

- Se una variabile esplicativa è inserita nel modello soprattutto per ragioni di controllo (ma non è di diretto interesse il suo effetto sulla variabile risposta) non è importante considerare problemi di multicollinearità
- Fattori che possono suggerire la presenza di problemi potenzialmente dovuti a multicollinearità:
  - ✓ Radicali cambiamenti nella stima di un coefficiente quando un'altra variabile esplicativa è inclusa nel modello
  - ✓ Test F per il confronto del modello con il modello nullo suggerisce di rifiutare l'ipotesi nulla ma nessun coefficiente parziale è statisticamente significativo singolarmente preso
- L'interpretazione del coefficiente perde di senso in presenza di multicollinearità

Multicollinearità: Osservazioni

- Alcuni approcci per rimediare alla multicollinearità
  - ✓ Scegliere un sottoinsieme di variabili esplicative, rimuovendo le variabili che spiegano solo una piccola parte della variabilità residua
  - ✓ Se diverse variabili esplicative sono altamente correlate e si riferiscono tutte a una stessa caratteristica sottostante, è possibile costruire un indice riassuntivo combinando i valori di tali variabili
- La multicollinearità non influisce sull'analisi degli effetti congiunti delle variabili esplicative: la multicollinearità non altera la bontà di adattamento del modello e la sua capacità previsiva rispetto alla variabile risposta.