La retta dei minimi quadrati

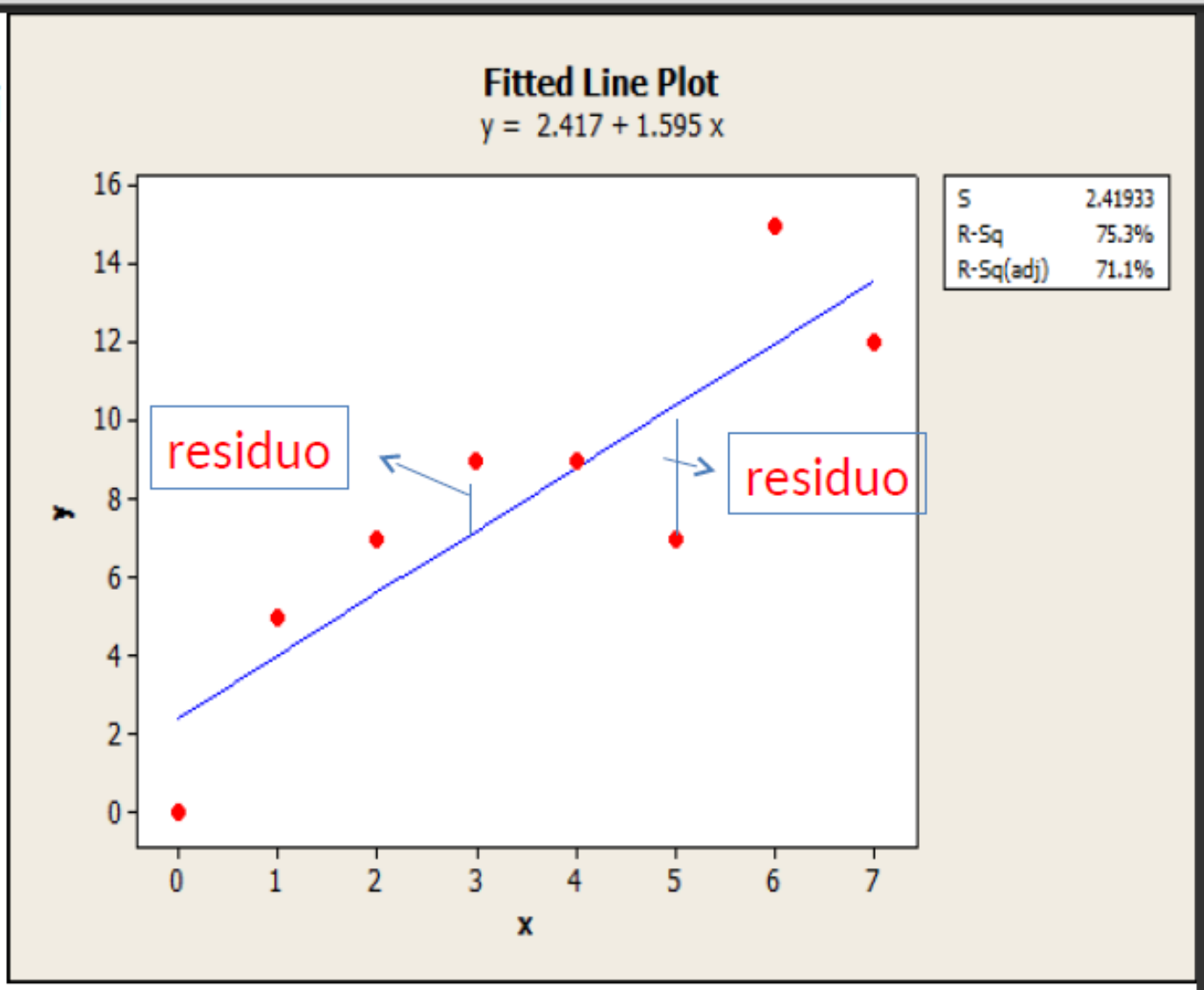
- Poiché la relazione espressa dal modello di regressione lineare semplice consiste nell'equazione di una retta, si tratterà di trovare la retta che **meglio approssima** i punti osservati.
- Secondo il metodo dei minimi quadrati si sceglie la "migliore" retta di regressione minimizzando la somma dei quadrati delle **distanze verticali** tra i punti osservati e la retta stessa, ovvero la somma dei quadrati dei

residui o errori di previsione.

I residui sono anche chiamati errori stimati.

Solo 2° gruppo di
Topi trattati

x	y
0	0
1	5
2	7
3	9
4	9
5	7
6	15
7	12



Qual è la risposta osservata? Qual è la risposta prevista?

Residui

Il **residuo** è la differenza fra un valore osservato della variabile di risposta e il valore previsto dalla stima della retta di regressione Y=bX+a

residuo = y osservato- y previsto

Valore di ingresso x_i : risposta osservata y_i ,

risposta prevista $bx_i + a$.

Residuo: $y_i - bx_i - a$.

Un residuo può essere positivo o negativo. Per determinare la retta dei minimi quadrati si minimizza la somma dei quadrati dei residui

Retta di regressione dei minimi quadrati: formule

Supponiamo di avere dei dati su una variabile esplicativa *x* e su una variabile di risposta *y* per *n* unità. In base ai dati, ricaviamo le mediee le deviazioni standard delle due variabili e la loro correlazione *r*. La retta di regressione dei minimi quadrati è la retta

$$\widehat{y} = a + b\,x$$

$$b = r\,\frac{s_y}{s_x}$$

$$a = \bar{y} - b\,\bar{x}$$

I residui

- Perché ci sia un ***buon adattamento del modello*** ai dati vogliamo che:
- i residui non individuino alcun andamento ulteriormente interpolabile con termini di ordine superiore.
- il segno dei residui sia "casuale", sia cioè, in qualche modo, ripartito equamente tra + e-, per escludere errori sistematici
- rappresentando su di un grafico i punti $(\mathbf{x_i},\mathbf{r_i})$ se la relazione tra **X** e **Y** è lineare dovremmo osservare dei punti che oscillano casualmente sopra e sotto lo 0.

Esempio: età e altezza. Grafico dei residui.

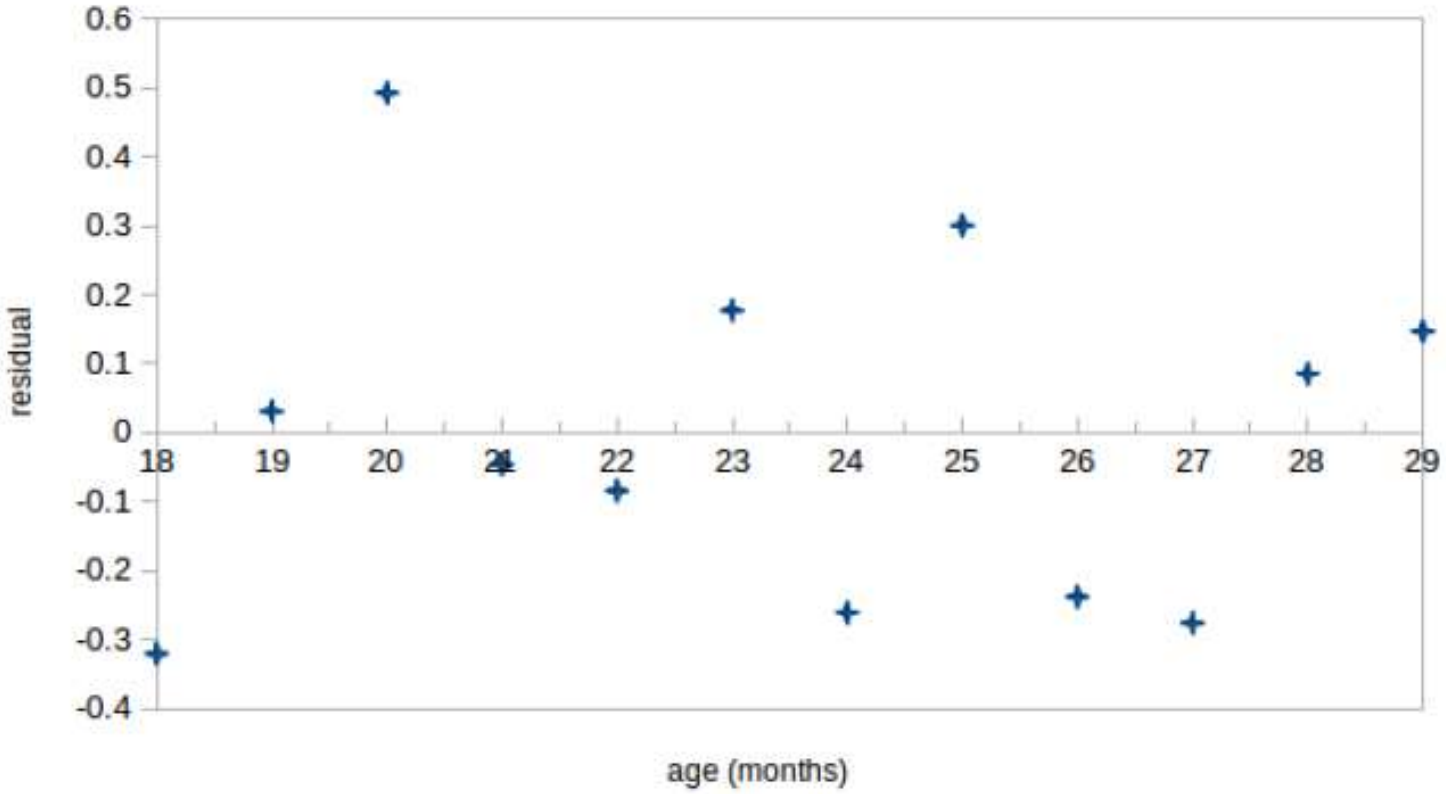
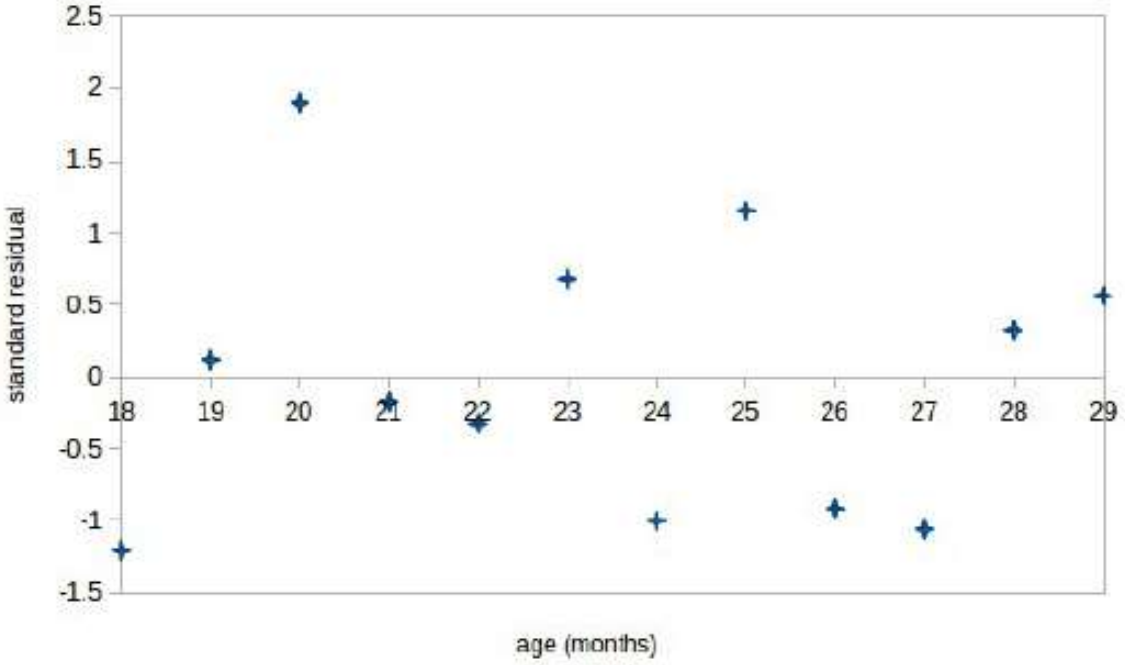


Grafico dei residui standardizzati



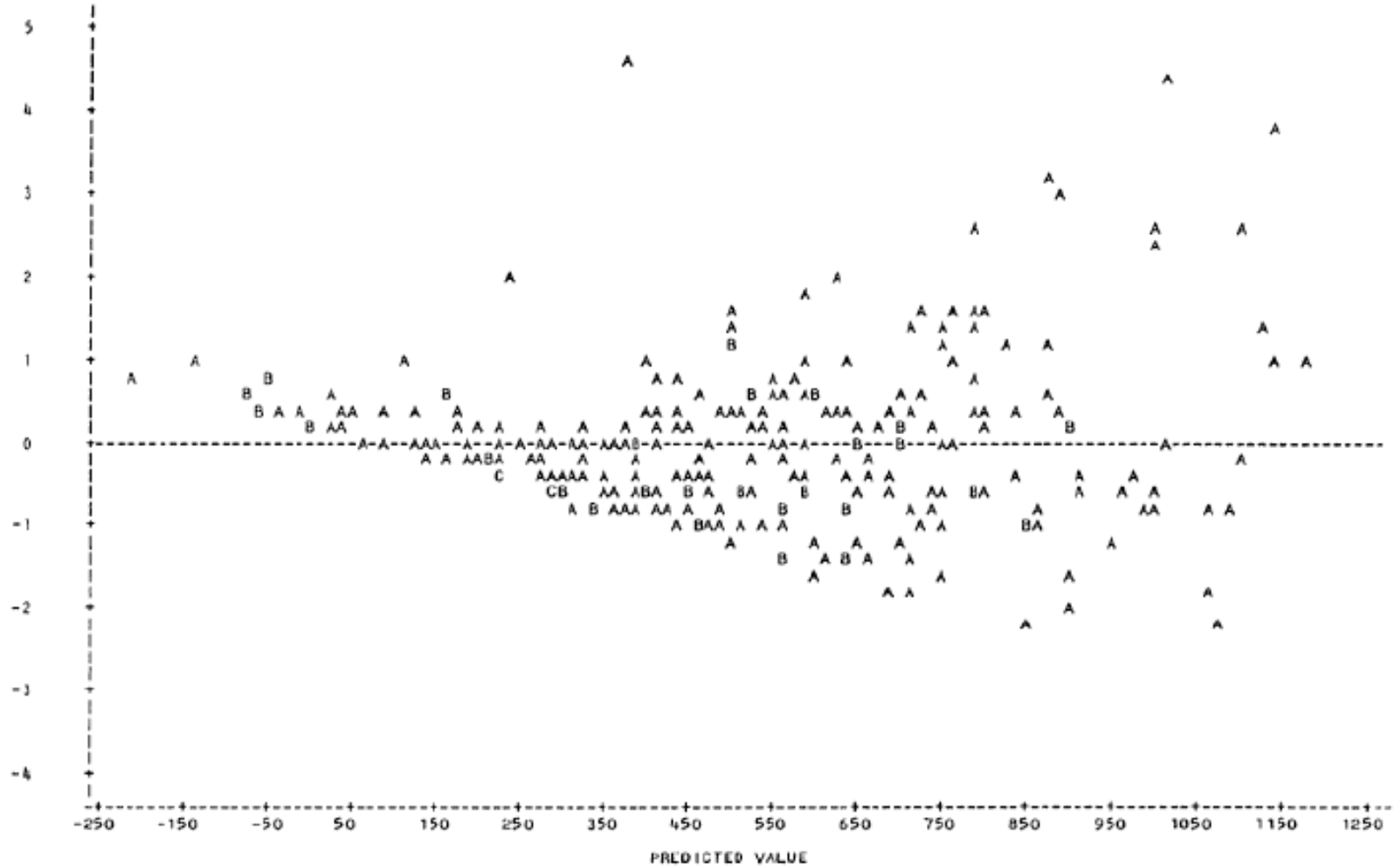
Residui standardizzati

$$(Y_i - a - b\,x_i)/\sqrt{(SS_{res}/(n-2))}$$

Se il modello di regressione è corretto, i residui standardizzati sono variabili aleatorie appross. normali standard, quindi

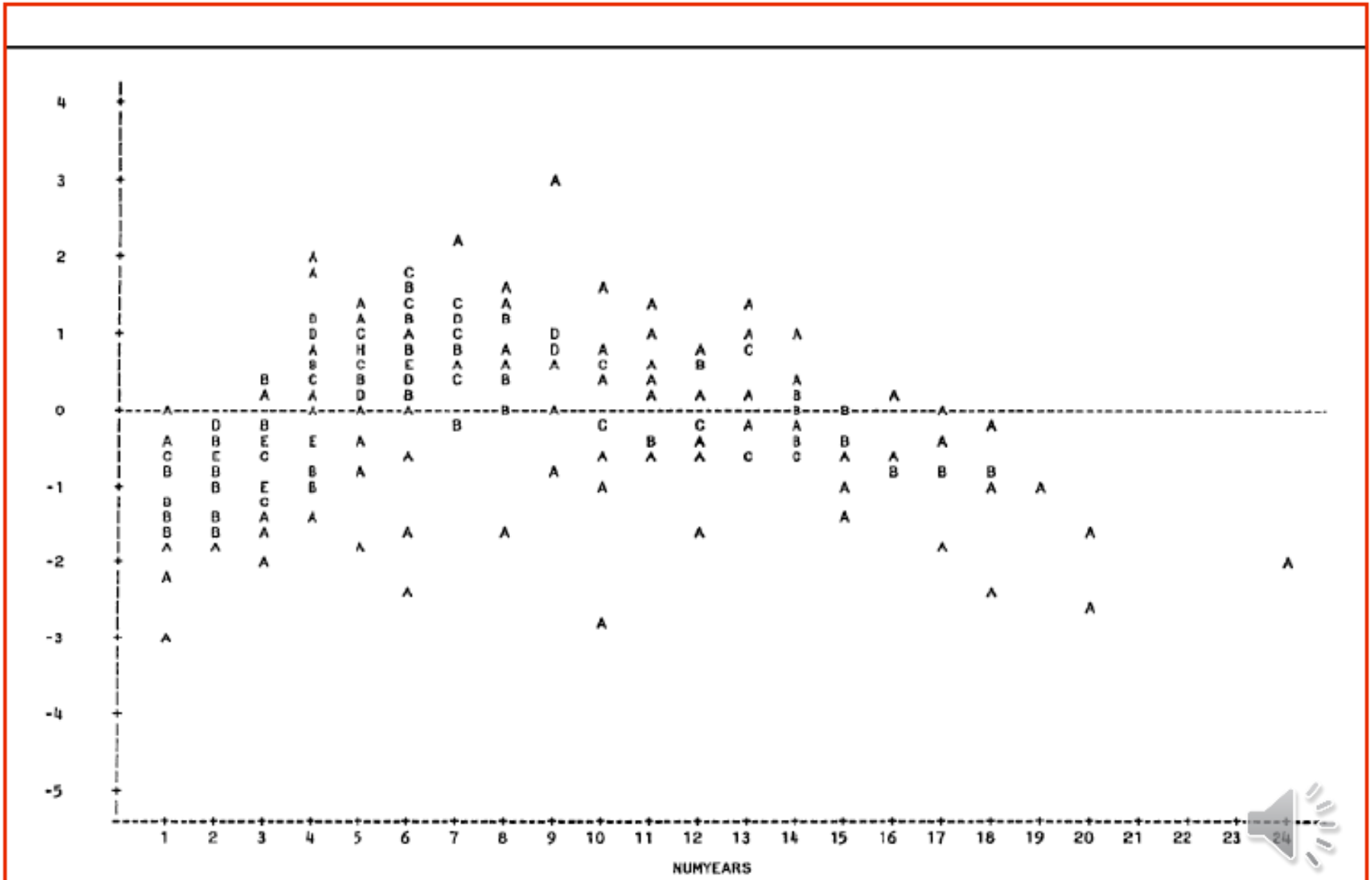
- Dovrebbero essere distribuite intorno allo 0
- Circa il 95 % dei valori dovrebbe essere compreso tra -2 e 2.

La variabilità dei residui non è costante, ma aumenta all'aumentare di x.



(a)

I residui si dispongono lungo una curva: la dipendenza non è lineare.



Regressione lineare $r^2=0.92$

r^2 ha un valore superiore rispetto al modello precedente

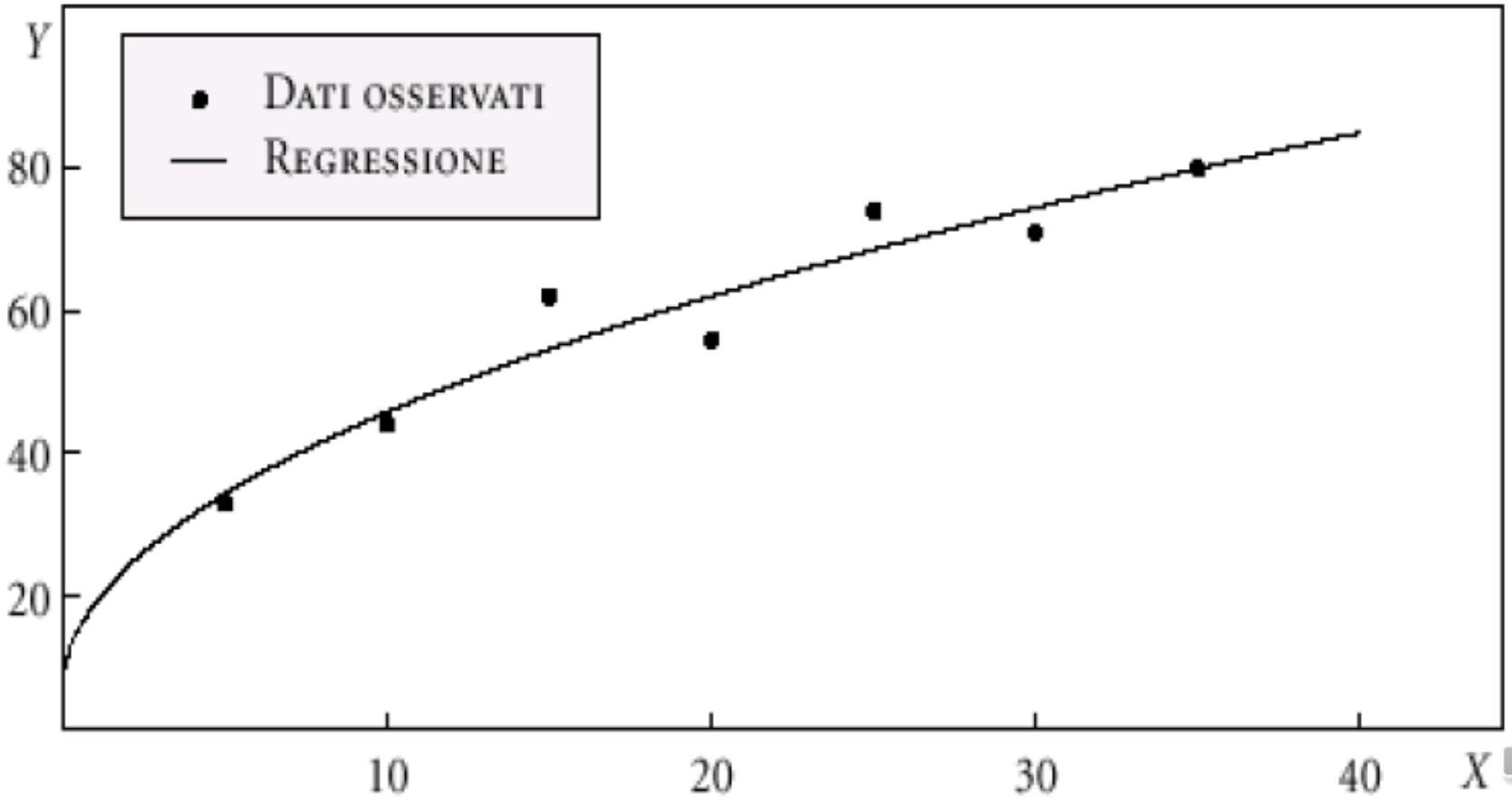


Grafico dei residui dell'es. precedente

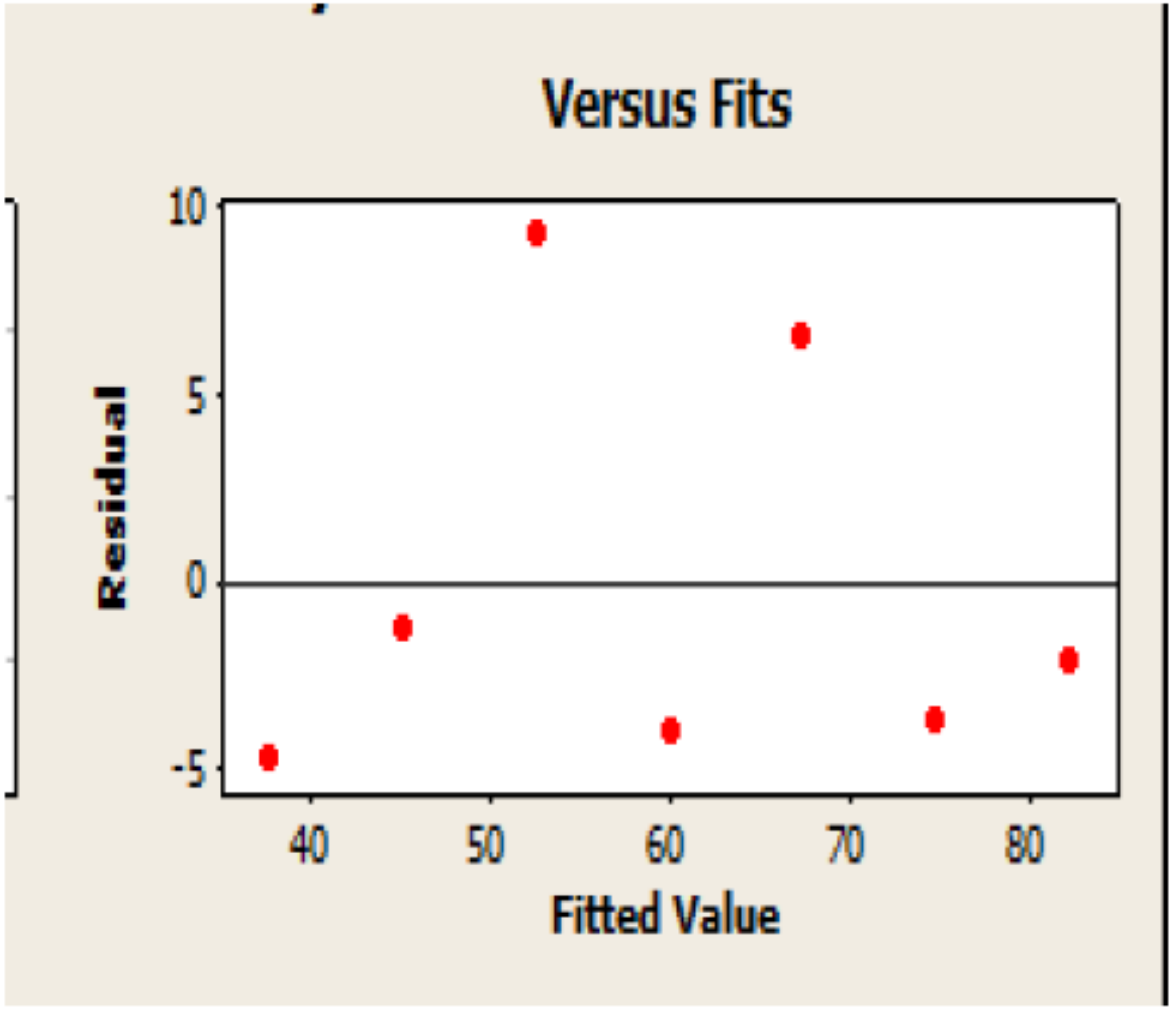
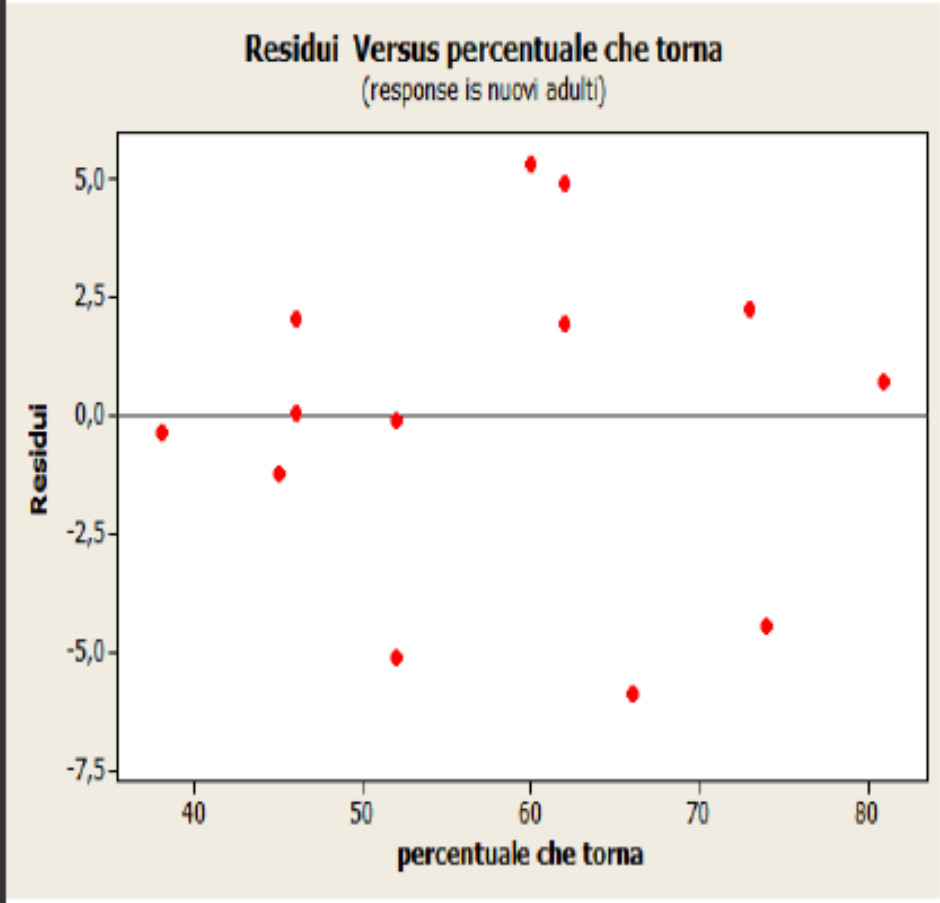
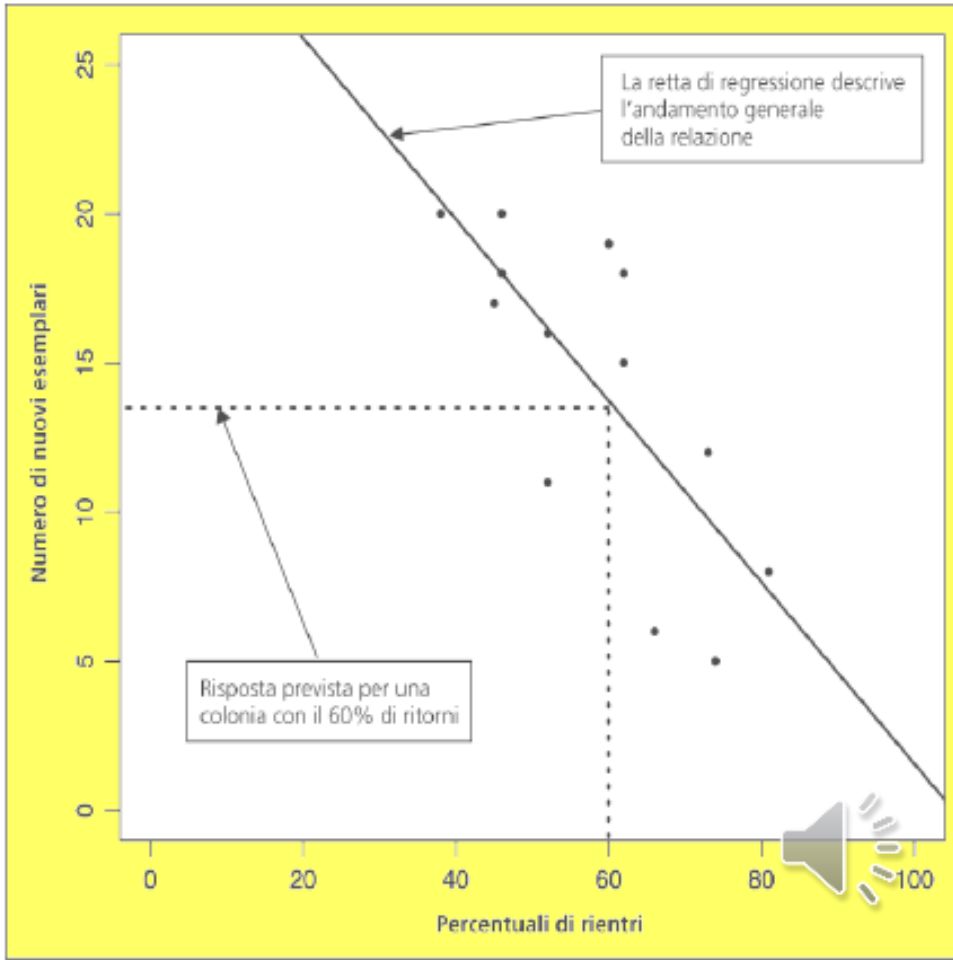


Grafico dei residui

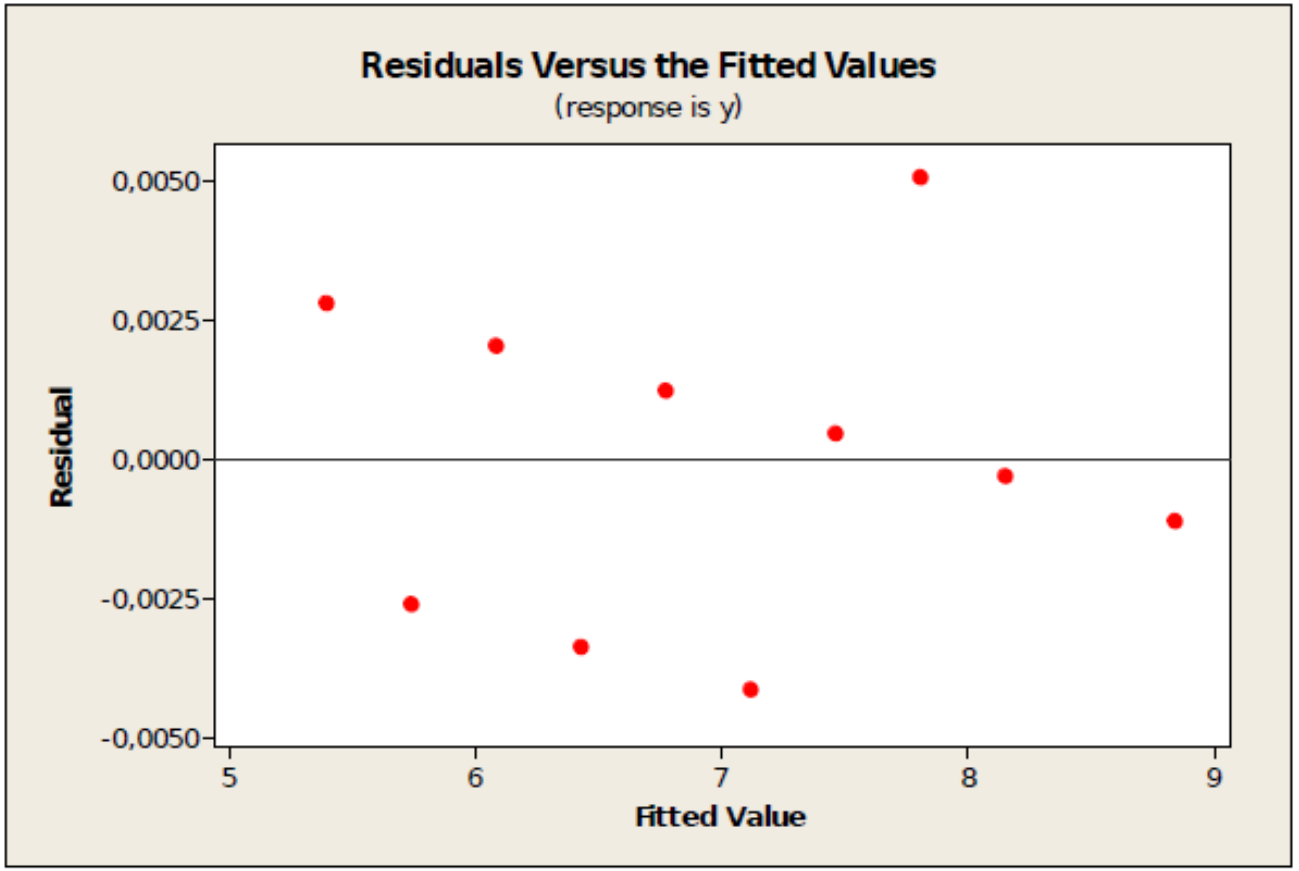
Se il grafico dei residui non mostra alcuna configurazione particolare (ossia presenta una configurazione casuale) l'equazione di regressione è un buon modello per rappresentare l'associazione tra le 2 variabili





Analisi dei residui e coefficiente di determinazione

Il grafico dei residui per l'es precedente mostra una **configurazione casuale**



L'equazione di regressione è un buon modello per rappresentare l'associazione tra le 2 variabili

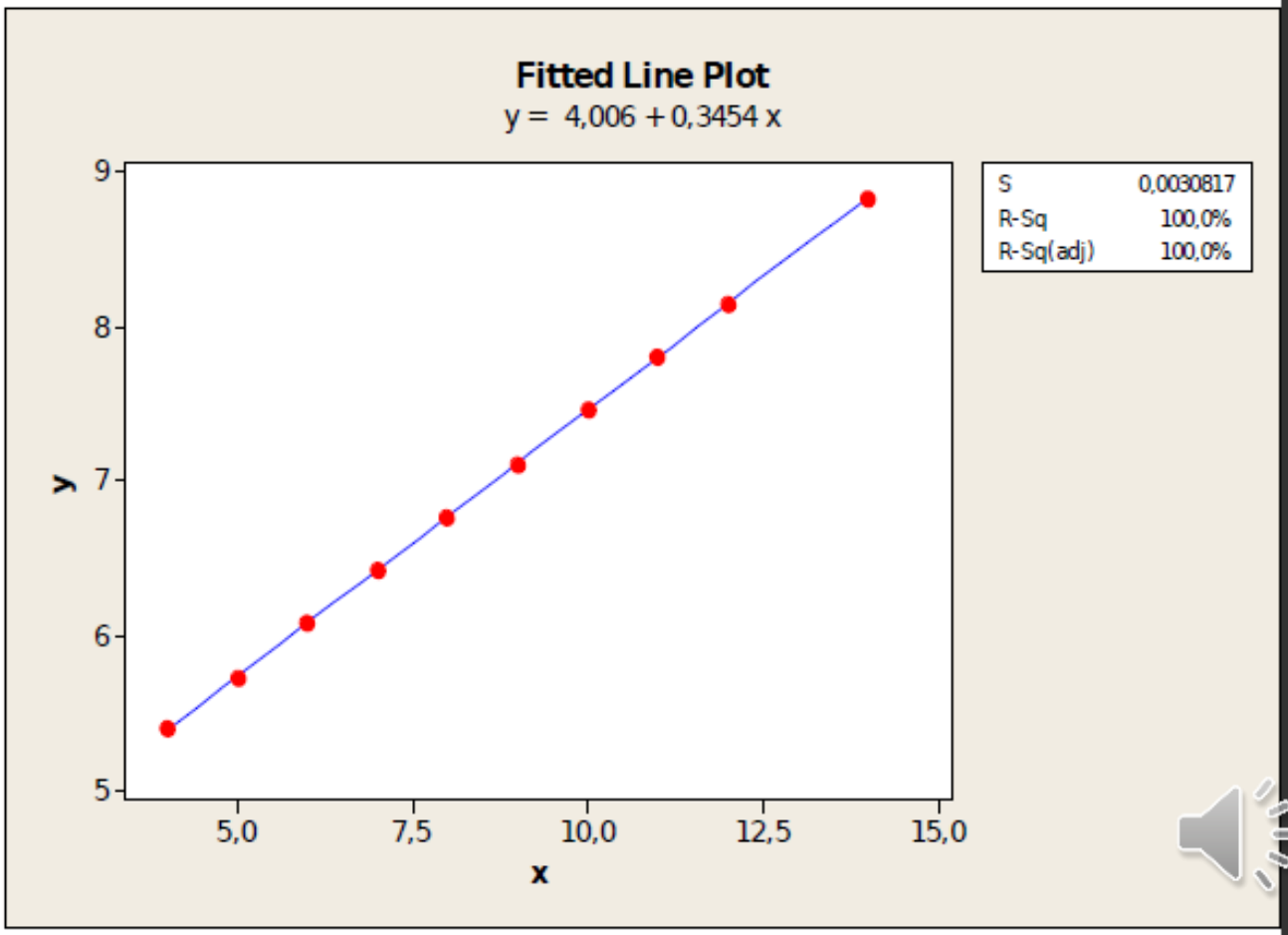
Retta di regressione

r = 1

r^2 = 100%

x	y
10	7,46
8	6,77
9	7,11
11	7,81
14	8,84
6	6,08
4	5,39
12	8,15
7	6,42
5	5,73

Retta di regressione per i dati osservati



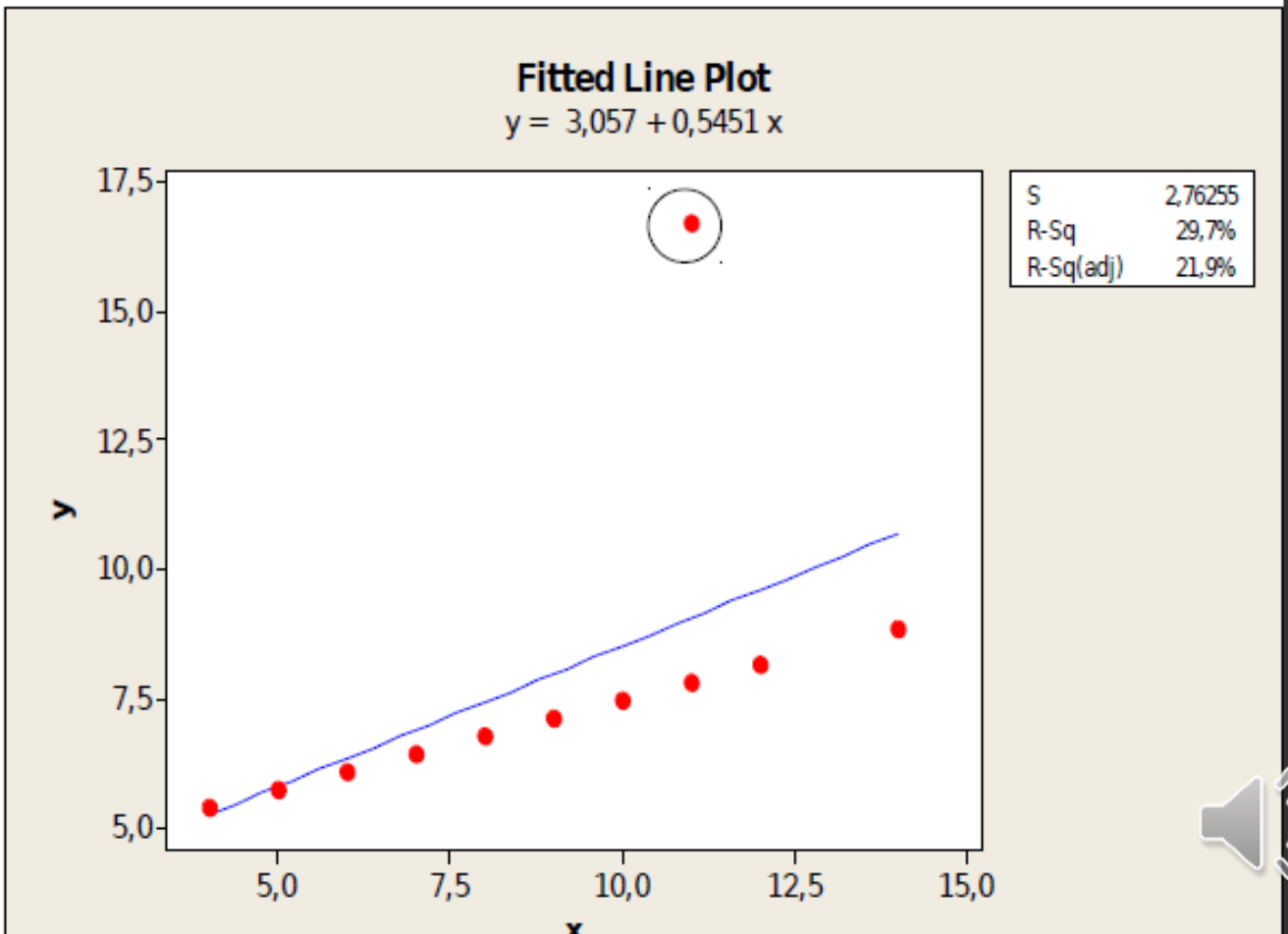
Analisi dei residui e coefficiente di determinazione

Al campione precedente è stata **aggiunta un'osservazione (outlier)**. Il modello lineare non è un buon modello (r^2 = 29.7%)

r = 0.545

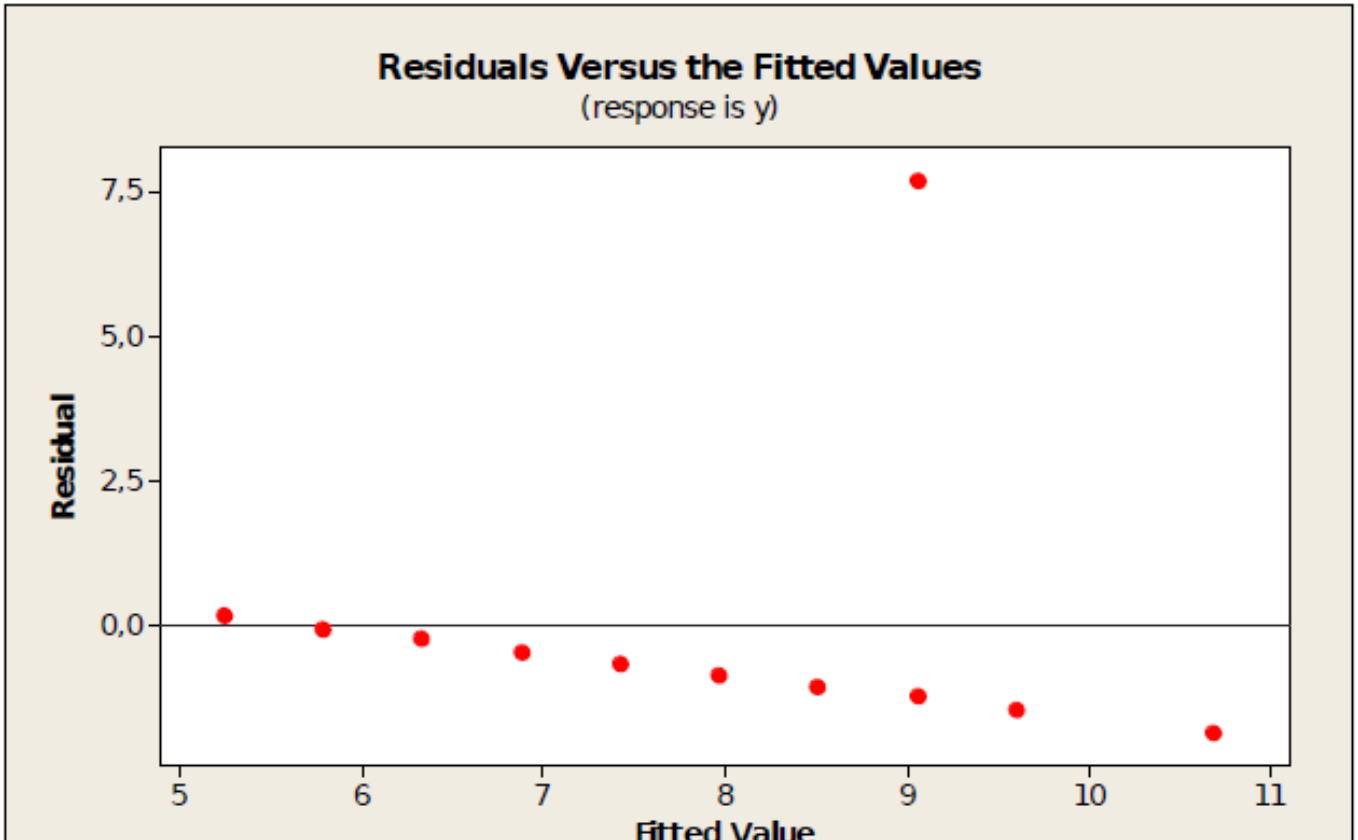
r^2 = 29.7%

x	y
10	7,46
8	6,77
11	16,74
9	7,11
11	7,81
14	8,84
6	6,08
4	5,39
12	8,15
7	6,42
5	5,73



Regressione:analisi dei residui e coefficiente di determinazione

Grafico dei residui (es. precedente, campione con outlier) **non** mostra una configurazione casuale. Il modello lineare **non** è un buon modello

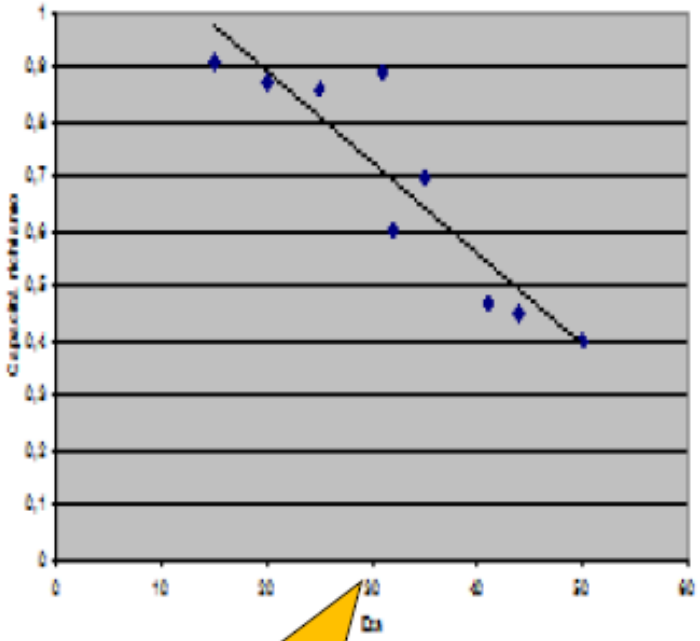


Osservazioni influenti

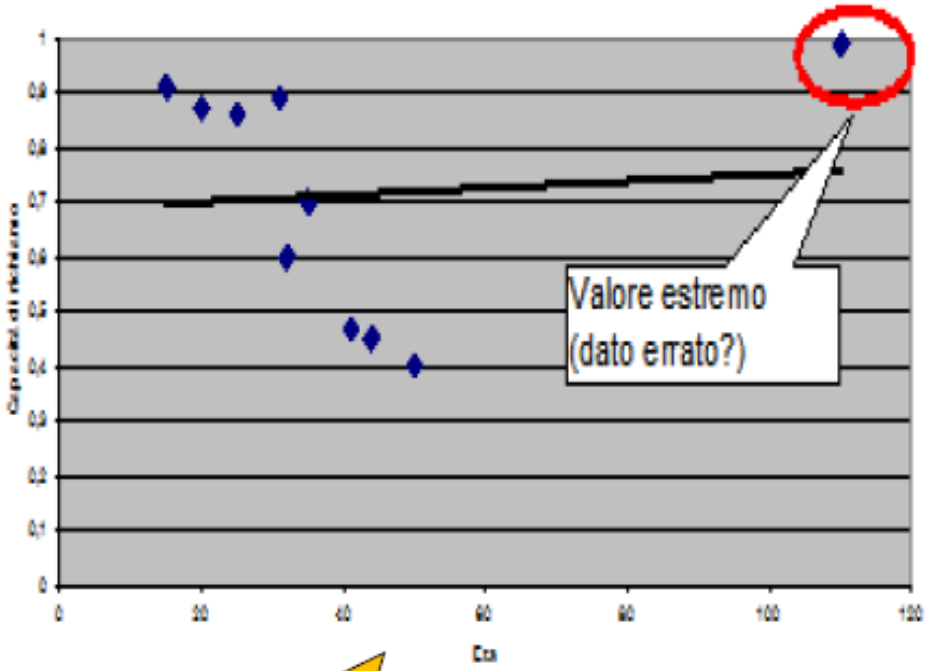
- **Outlier e osservazioni influenti nella regressione**
- Un **outlier** è un'osservazione che non segue il modello generale assunto dalla maggior parte delle osservazioni. I punti che, guardando un diagramma di dispersione, possiamo **considerare** outlier in direzione di *y*, hanno residui elevati.

- Un'osservazione è **influente** se, eliminandola, cambierebbe profondamente il risultato. I punti che, in un diagramma di dispersione, possiamo considerare **outlier in direzione della *x*** sono spesso punti influenti nella determinazione della retta di regressione dei minimi quadrati.

Osservazioni influenti



retta di regressione con un buon adattamento ai dati



ai dati originali è stata aggiunta una osservazione influente

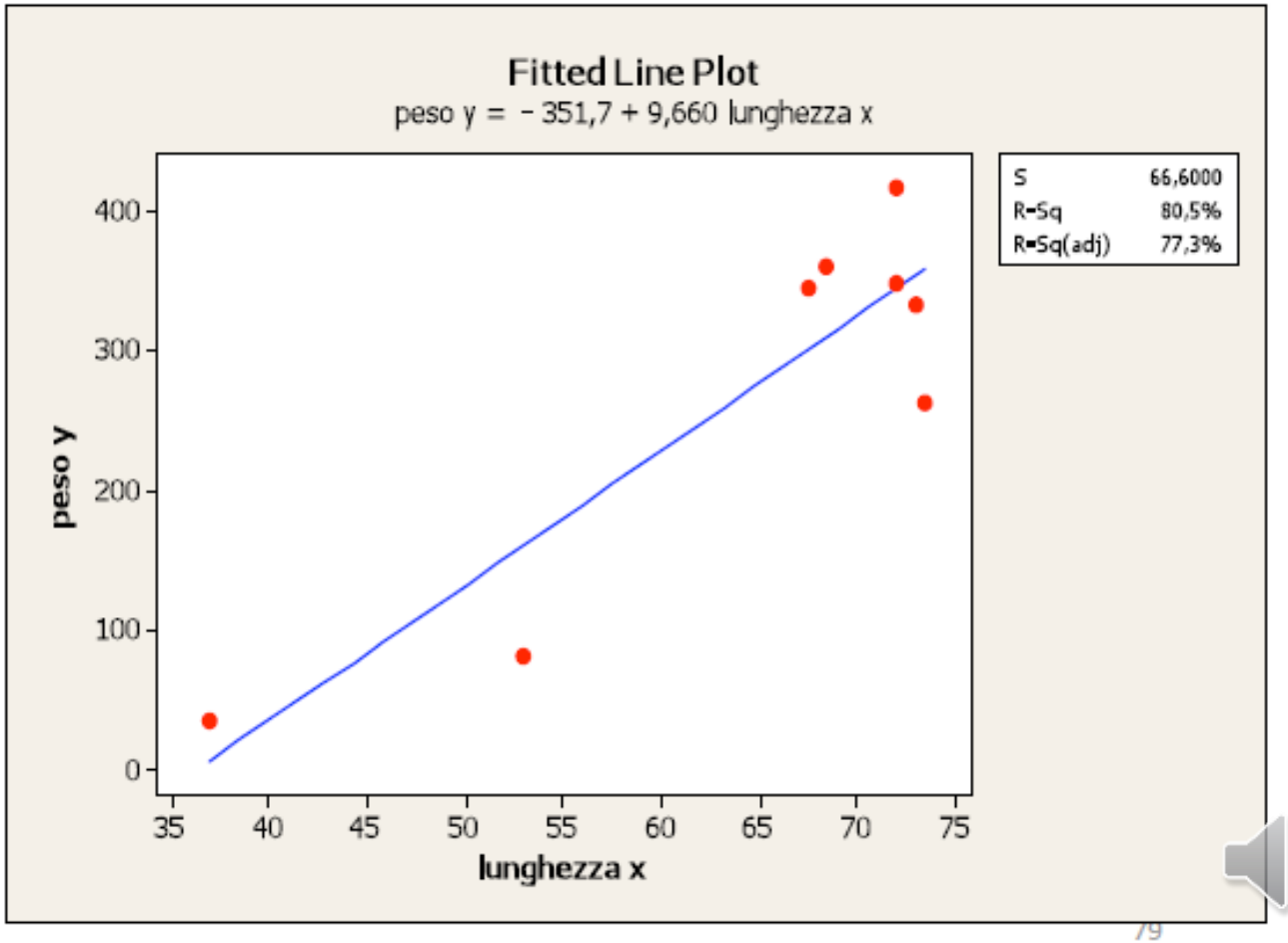
Regressione: Lunghezza e peso di un campione di 8 orsi maschi

la retta di regressione interpola bene i dati

$r = 0.897$

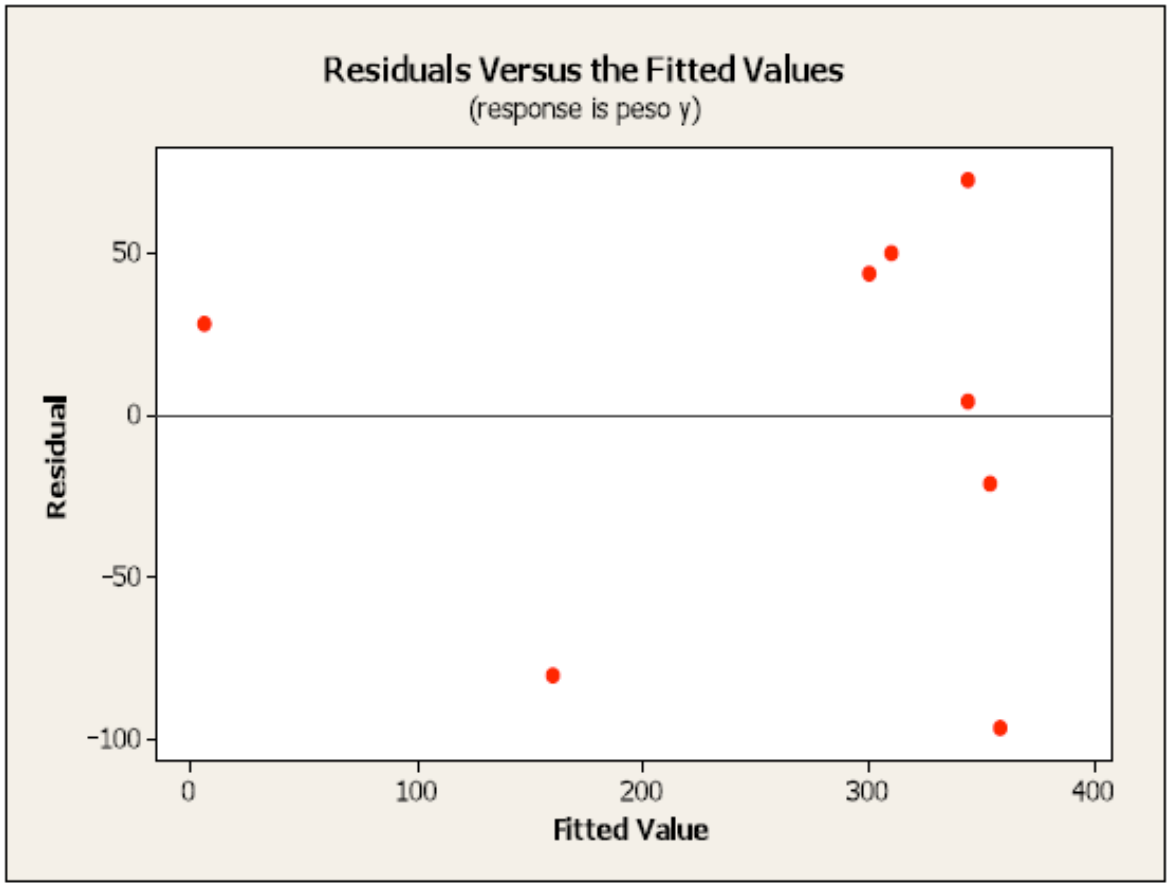
$r^2 = 80.5\%$

lunghezza (pollici)	peso (libbre)
53,0	80
67,5	344
72,0	416
72,0	348
73,5	262
68,5	360
73,0	332
37,0	34



Regressione: continuazione esempio

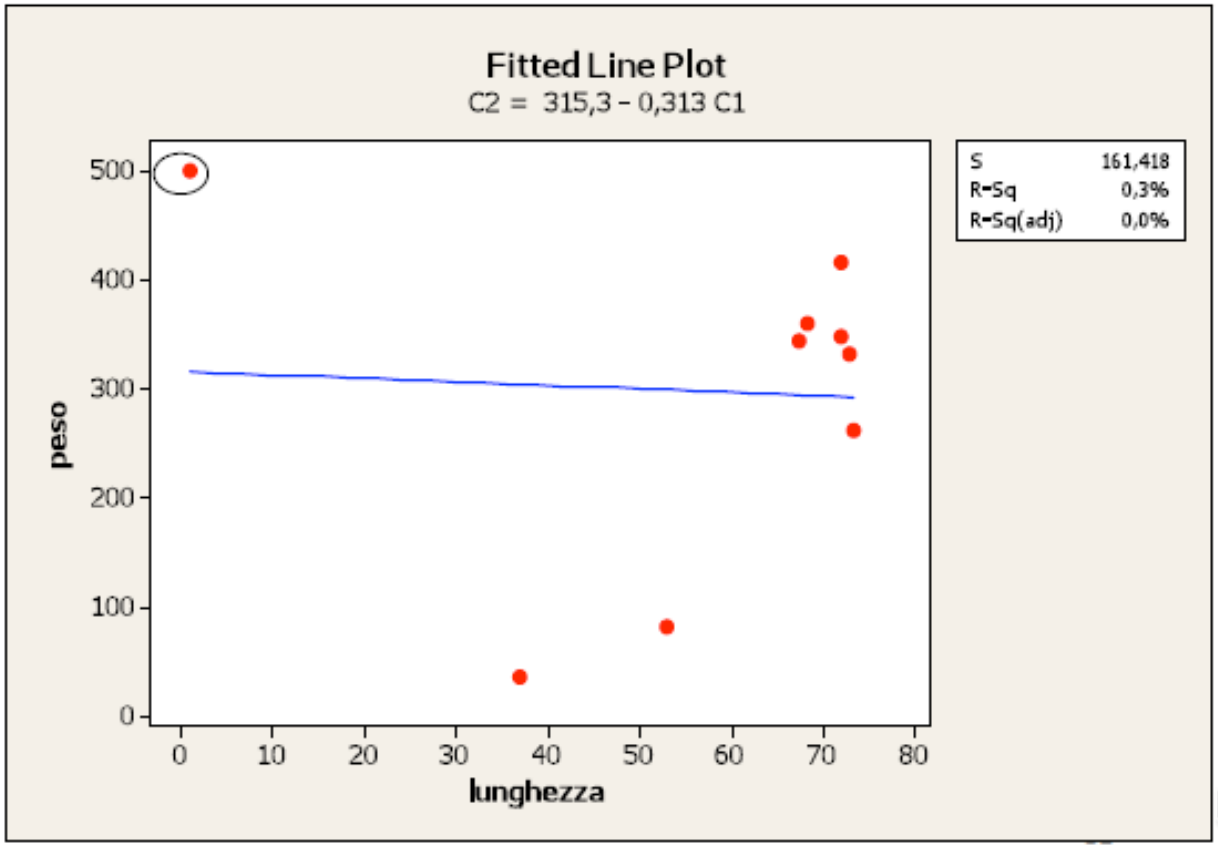
Il grafico dei residui presenta una configurazione casuale



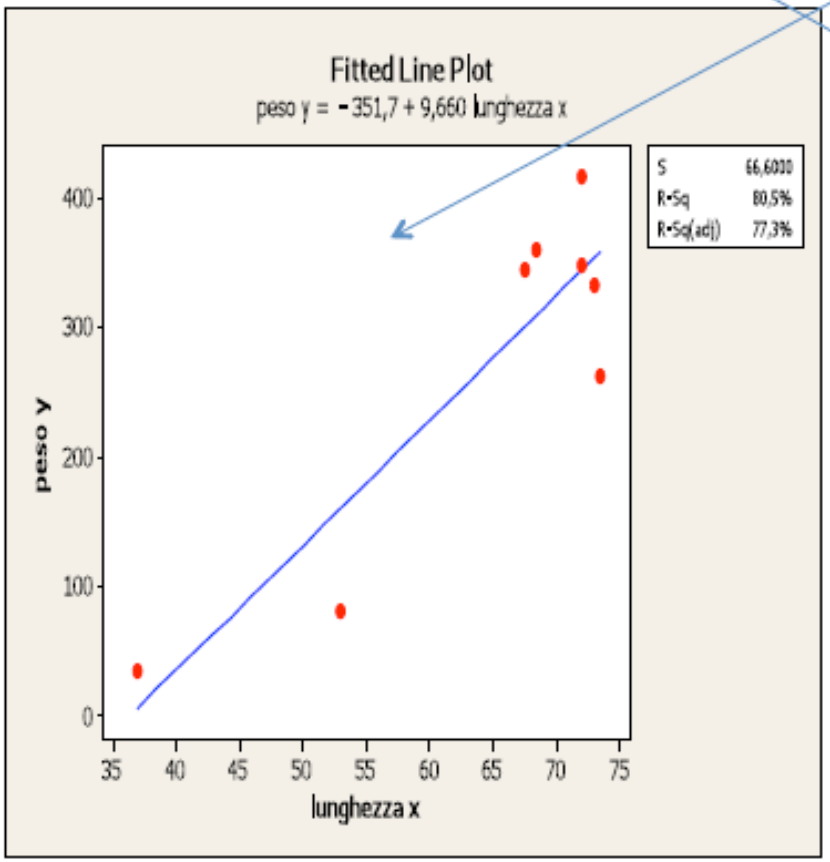
Regressione: continuazione esempio

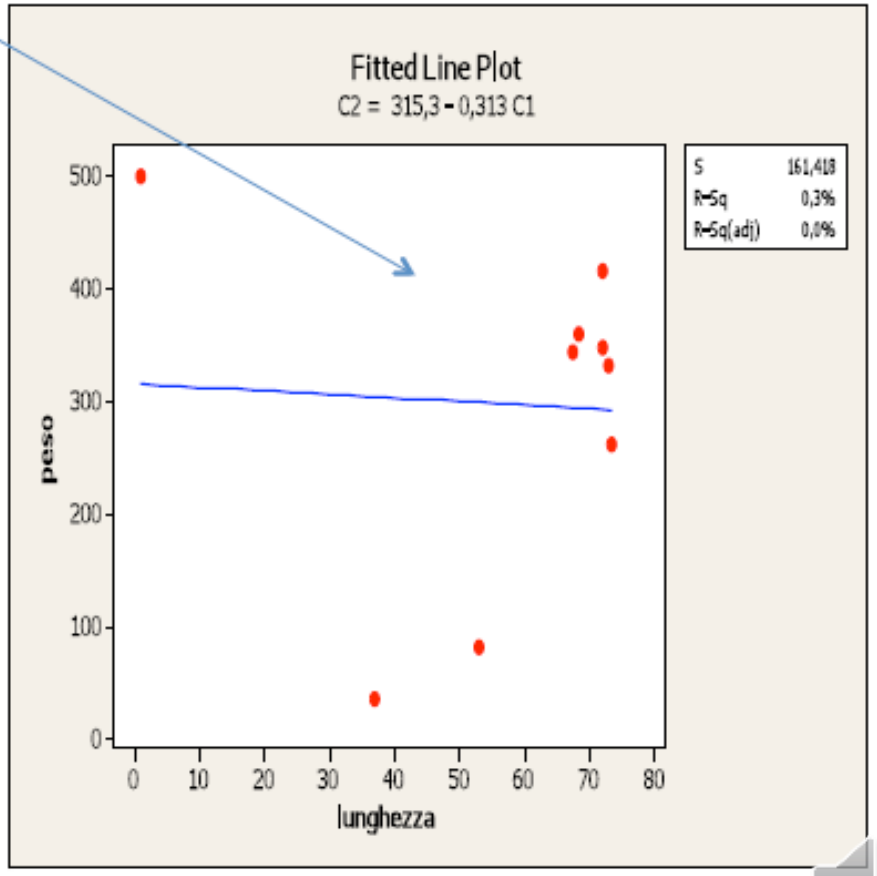
Se aggiungessimo al campione un nuovo orso lungo 1 pollice e di peso pari a 500 libbre!!! avremmo un punto influente. Infatti si avrebbe un effetto notevole sulla retta di regressione (cambia l'equazione).

x	y
53,0	80
67,5	344
72,0	416
72,0	348
73,5	262
68,5	360
73,0	332
37,0	34
1,0	**500**



Le due rette a confronto (senza osservazione influente: $r^2 = 80,5\%$ e con osservazione influente: $r^2 = 0,3\%$)

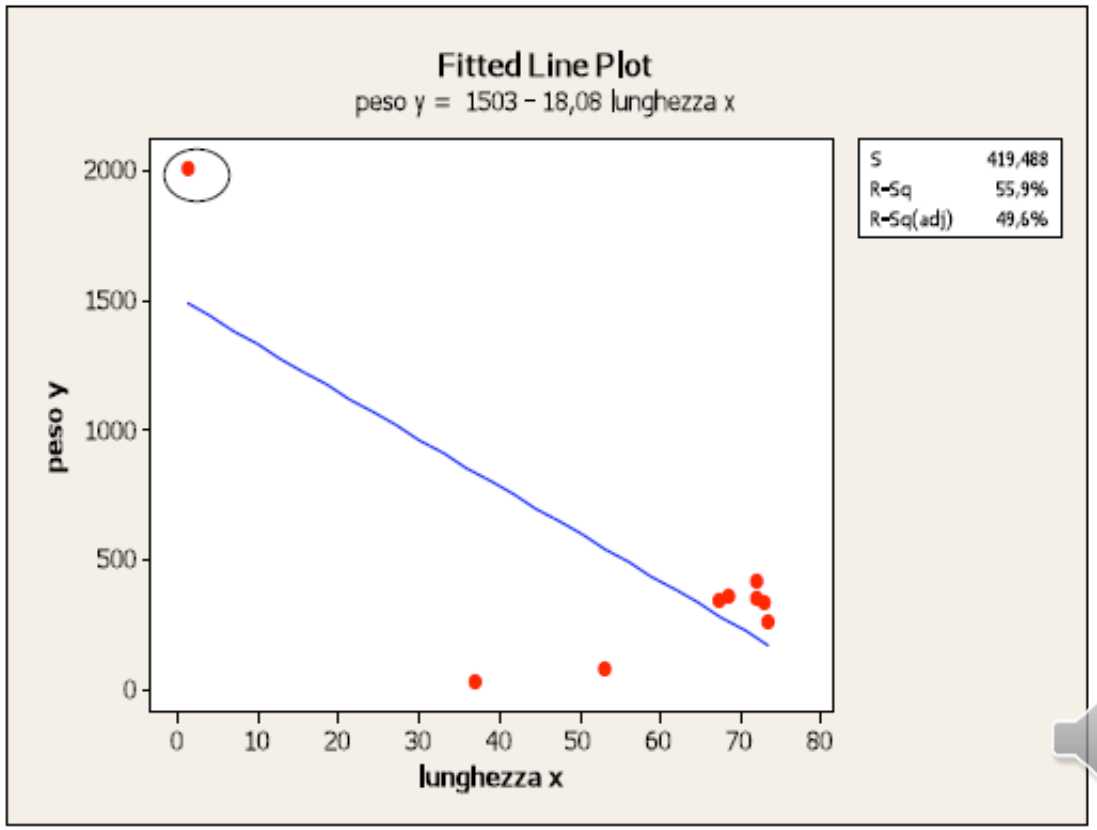


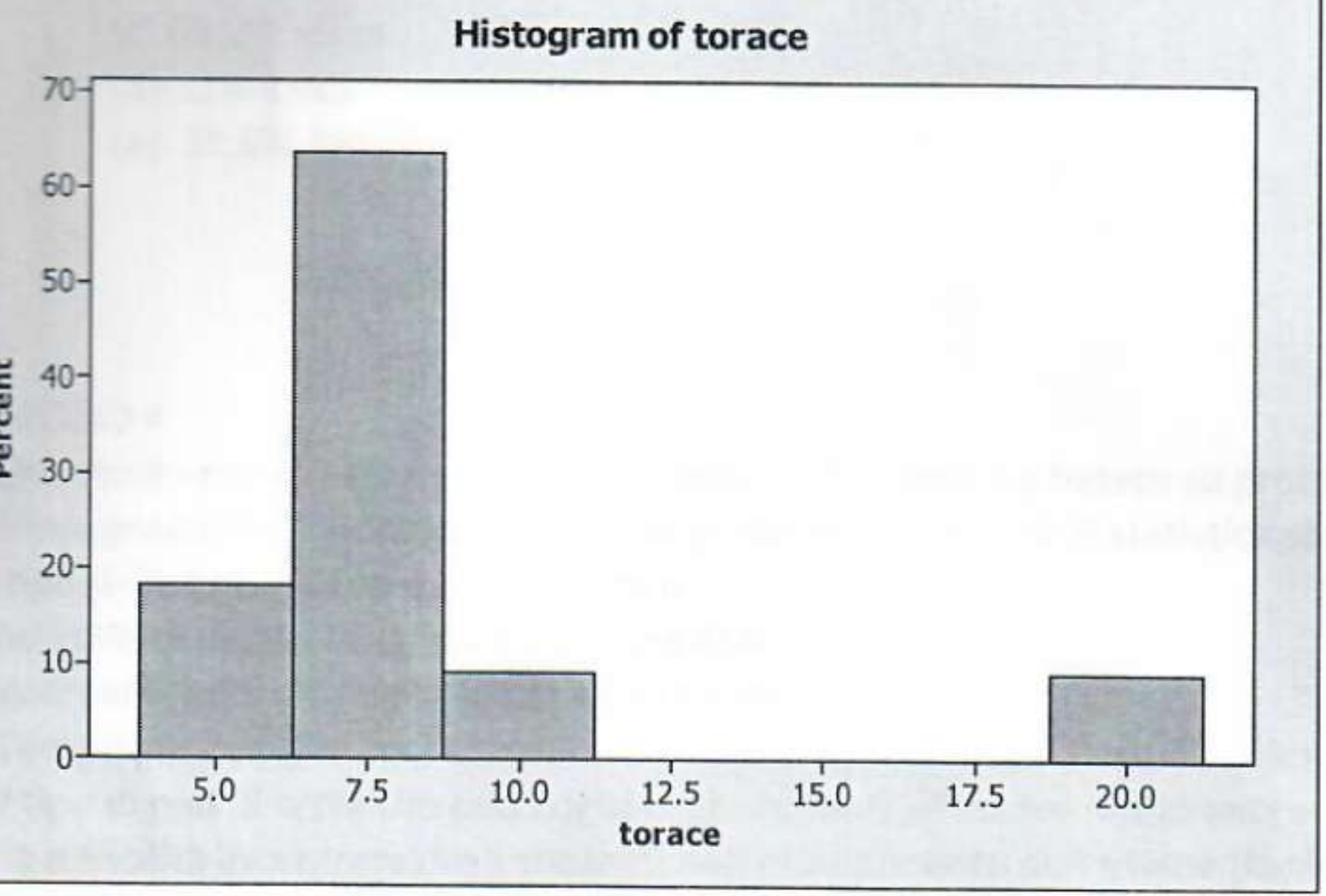


Regressione: continuazione esempio

Se aggiungessimo al campione un nuovo orso lungo 1 pollice e di peso pari a 2000 libbre!!! avremmo di nuovo un punto influente (e anche outlier nella direzione della y). Infatti si avrebbe un effetto drammatico sulla retta di regressione.

x	y
53,0	80
67,5	344
72,0	416
72,0	348
73,5	262
68,5	360
73,0	332
37,0	34
1,0	2000





Cosa si può dire sulla forma del grafico?