

Analisi di regressione

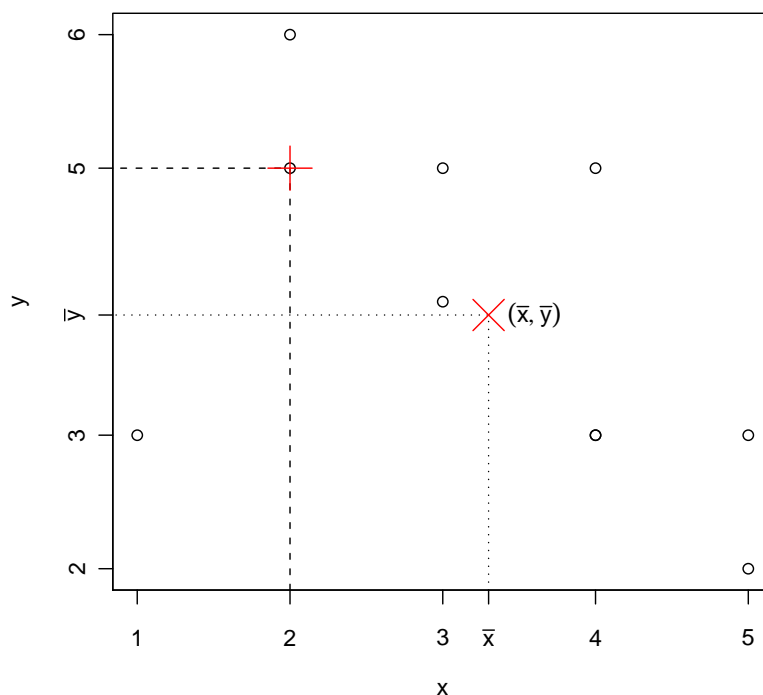
Analisi congiunta di due fenomeni di tipo quantitativo (meglio se continui)

Supponiamo di aver raccolto i seguenti dati

x_i	2	3	4	2	5	4	5	3	4	1
y_i	5	4	3	6	2	5	3	5	3	3

La tabella di contingenza sarebbe una $n \times n$ con solo 0 e 1

Si rappresentano i dati (x_i, y_i) su un grafico *a dispersione*



Si vuole capire se i punti (ovvero i fenomeni statistici) si disperdono attorno ad un particolare valore detto **baricentro** della distribuzione (\bar{x}_n, \bar{y}_n)

Nell'esempio $\bar{x}_n = 3.3$ e $\bar{y}_n = 3.9$

L'indice che misura la dispersione delle coppie di punti dal baricentro è la *covarianza*

La covarianza, al contrario della varianza, si occupa anche di misurare l'eventuale direzione della variabilità

Covarianza tra X ed Y

$$\sigma_{xy} = \text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)$$

ovvero

Covarianza tra X ed Y (formula alternativa)

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i \cdot y_i) - (\bar{x}_n \cdot \bar{y}_n)$$

Calcoliamo la covarianza tra X e Y

x_i	2	3	4	2	5	4	5	3	4	1	
y_i	5	4	3	6	2	5	3	5	3	3	
$x_i \cdot y_i$	10	12	12	12	10	20	15	15	12	3	121

$$\sigma_{xy} = \frac{121}{10} - 3.3 \cdot 3.9 = -0.77$$

Indice di correlazione

Vale la seguente relazione

$$-\sigma_x \cdot \sigma_y \leq \sigma_{xy} \leq \sigma_x \cdot \sigma_y$$

e quindi possiamo definire l'indice di *correlazione*

Indice di correlazione X ed Y

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \quad -1 \leq \rho_{xy} \leq 1$$

- $\rho_{xy}=0$ solo se X ed Y sono **incorrelate**
- $\rho_{xy}=1$ solo se X ed Y sono in relazione lineare **diretta**
- $\rho_{xy}=-1$ solo se X ed Y sono in relazione lineare **inversa**

L'assenza di relazione lineare non implica che non siano presenti altri tipi di relazione

x_i	-2	-1	0	1	2	0
y_i	4	1	0	1	4	0
$x_i \cdot y_i$	-8	-1	0	1	8	0

$$\rho_{xy} = \frac{1}{5} \sum_{i=1}^5 x_i y_i - \bar{x}_n \bar{y}_n = \frac{0}{5} - 0 \cdot 0 = 0$$

ovvero c'è assenza di relazione lineare tra X ed Y

Costruiamo la tabella di contingenza

	Y	0	1	4	
X					
-2		0	0	1	1
-1		0	1	0	1
0		1	0	0	1
1		0	1	0	1
2		0	0	1	1
		1	2	2	5

Calcoliamo l'indice $\tilde{\chi}^2$

$$\tilde{\chi}^2 = \frac{\sum_{i=1}^5 \sum_{j=1}^3 \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1}{\min(3 - 1, 5 - 1)} = \frac{\frac{1}{2} + \frac{1}{2} + 1 + \frac{1}{2} + \frac{1}{2} - 1}{2} = 1$$

Siamo in presenza di massima connessione!

Retta di regressione

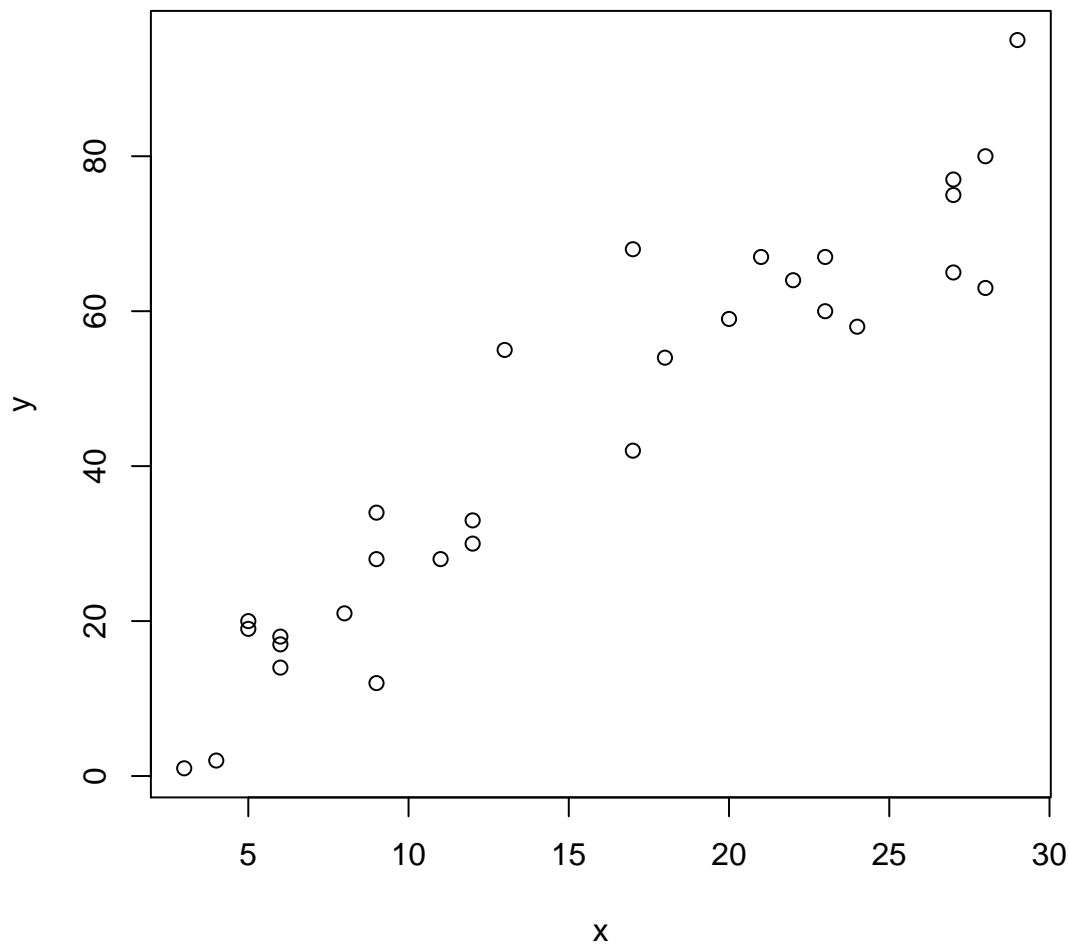
Analisi congiunta di due fenomeni di tipo quantitativo (meglio se continui)

Vogliamo studiare un tipo particolare di dipendenza tra due variabili: quella lineare

Abbiamo rilevato X (età) ed Y (peso) su n individui. Abbiamo $n = 30$ coppie di numeri (x_i, y_i) , $i = 1, \dots, n$

(x_i, y_i)	(x_i, y_i)	(x_i, y_i)
11 28	6 14	29 95
8 21	24 58	9 12
28 63	18 54	3 1
17 42	21 67	12 30
9 28	6 18	9 34
4 2	22 64	23 67
28 80	27 65	5 20
5 19	17 68	27 75
12 33	27 77	20 59
23 60	6 17	13 55

Prima cosa da fare sempre: rappresentare i punti su un grafico



Dal grafico a dispersione ci aspettiamo un valore positivo per la covarianza e quindi per ρ

Ci chiediamo però se esiste una relazione funzionale tra la variabile X e la variabile Y del tipo $Y = f(X)$

Guardando il grafico si può ipotizzare che sia del tipo lineare

$$Y = f(X) = a + b \cdot X$$

Se ipotizziamo il modello $Y = a + b \cdot X$ in corrispondenza di x_i , osservato sulla variabile X , (indipendente) dovremmo osservare il valore

$$y_i^* = a + b \cdot x_i$$

per la variabile Y (dipendente)

I valori y_i^* sono detti valori **teorici** o **previsti** della variabile Y

Cerchiamo “la retta migliore” passante per i punti (x_i, y_i)

Cerchiamo la retta che minimizza la distanza *quadratica*

$$(y_i - y_i^*)^2$$

Si tratta di trovare i valori di a e b che rendono minima

$$\sum_{i=1}^n (y_i - y_i^*)^2 = \sum_{i=1}^n (y_i - (a + b \cdot x_i))^2 = g(a, b)$$

Per trovare tali valori occorre risolvere il sistema

$$\begin{cases} \frac{\partial g(a,b)}{\partial a} = 0 \\ \frac{\partial g(a,b)}{\partial b} = 0 \end{cases}$$

Il sistema ha come soluzione

$$\begin{cases} b = \frac{\sigma_{xy}}{\sigma_x^2} \\ a = \bar{y} - b \cdot \bar{x} \end{cases}$$

Quindi ricapitolando:

Retta di regressione di Y in funzione di X

Se abbiamo n coppie di punti (x_i, y_i) la miglior retta passante “vicino” ad essi, detta retta di regressione, è quella di equazione:

$$Y = a + b \cdot X$$

dove

$$b = \frac{\sigma_{xy}}{\sigma_x^2} \quad a = \bar{y} - b \cdot \bar{x}$$

Prima di calcolare i coefficienti della retta è bene calcolare ρ_{xy}

$$\bar{x} = \frac{469}{30} = 15.6\bar{3} \quad \bar{y} = \frac{1326}{30} = 44.2$$

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = 205.04$$

$$\sigma_x = \sqrt{\frac{2192.967}{30}} = 8.55 \quad \sigma_y = \sqrt{\frac{19284.8}{30}} = 25.35$$

$$\rho_{xy} = \frac{205.04}{8.55 \cdot 25.35} = 0.95$$

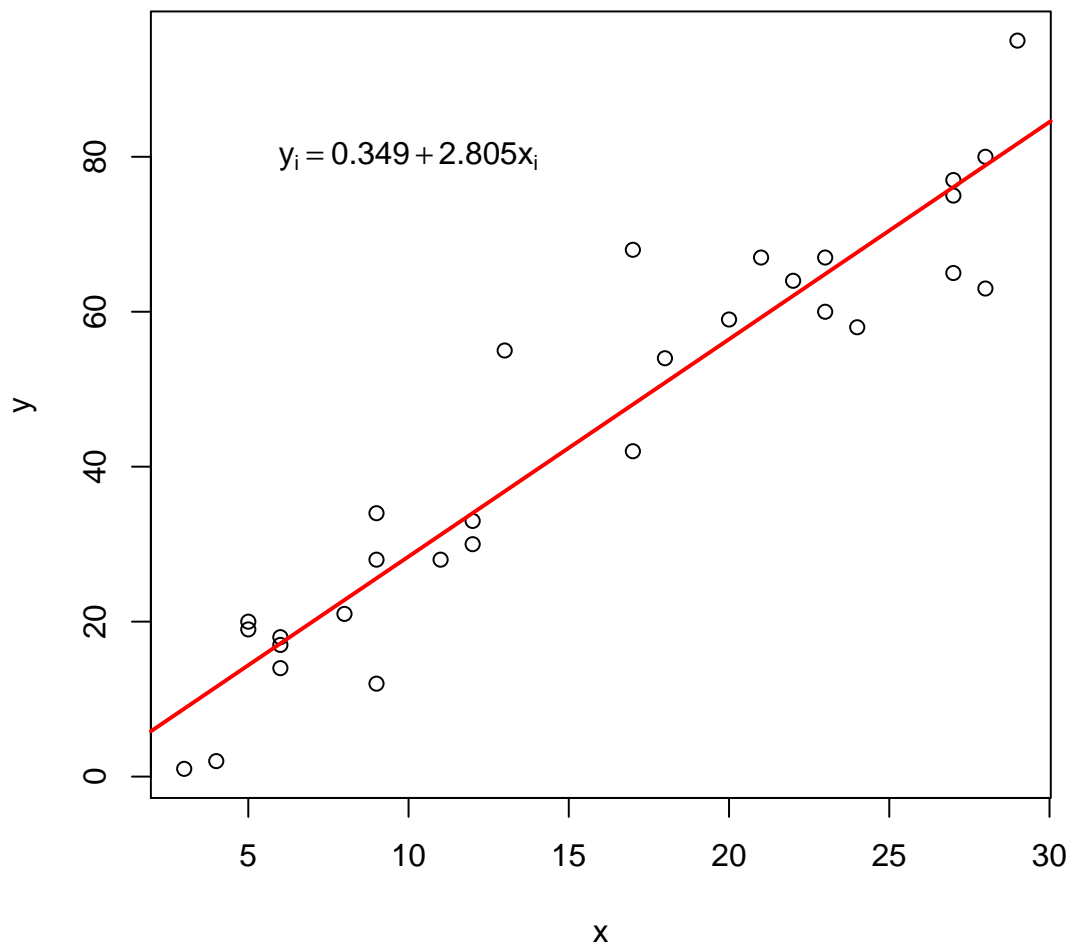
Calcoliamo i coefficienti della retta:

$$b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{205.04}{73.1} = 2.805$$

$$a = \bar{y} - b\bar{x} = 44.2 - 2.805 \cdot 15.6\bar{3} = 0.349$$

La miglior retta (nel senso della distanza quadratica) che passa per i punti (x_i, y_i) è data da

$$y = 0.349 + 2.805 \cdot x$$



La retta è il modello interpretativo del fenomeno

Cosa ne facciamo?

Possiamo fare delle previsioni

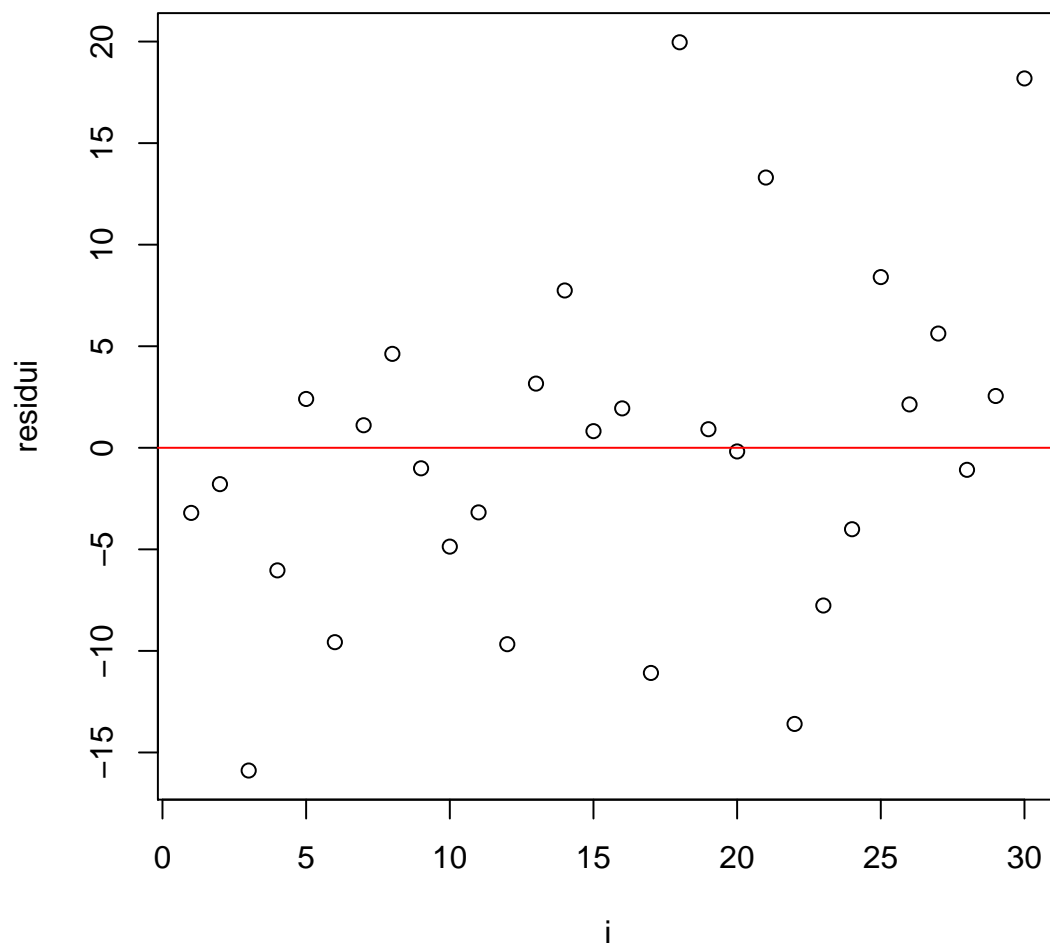
Bontà di adattamento

Si basa sull'analisi dei residui

$$e_i = y_i - y_i^*$$

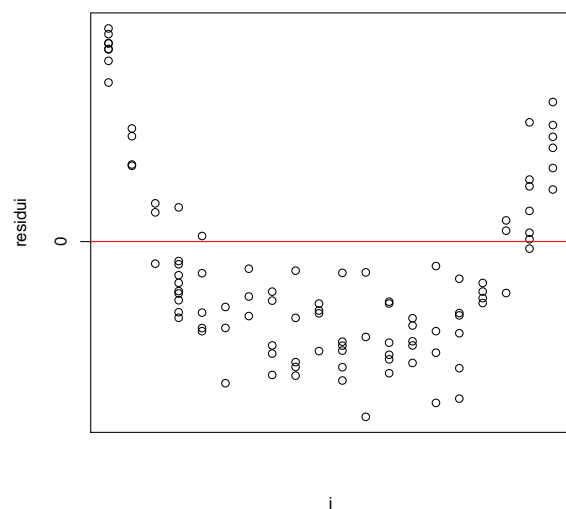
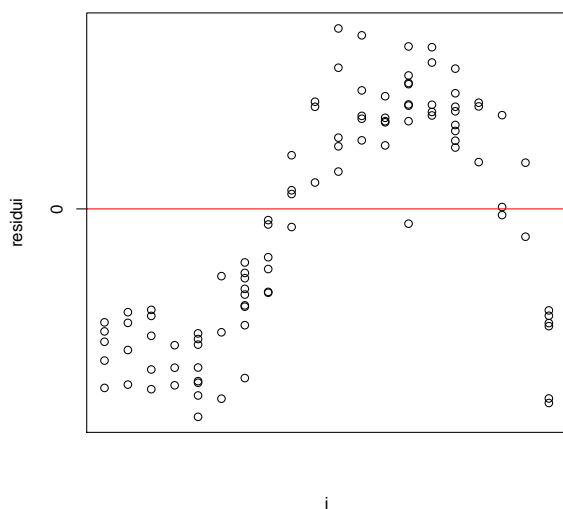
ottenuti come differenza tra i valori osservati (y_i) e i valori teorici previsti (y_i^*)

Prima di tutto si rappresentano graficamente



Ci si deve aspettare che i residui siano bassi (in termini dell'unità di misura di Y) e che ve ne siano un po' positivi ed un po' negativi ma senza troppa regolarità

Andamenti come questi indicano che il modello lineare non è adatto a spiegare il legame tra le variabili



L'indice di determinazione o R^2

Partiamo dalla varianza di Y ottenuta sommando i termini $(y_i - \bar{y})^2$

Aggiungendo e sottraendo i valori y_i^* otteniamo

$$(y_i - \bar{y} \pm y_i^*)^2 = ((y_i^* - \bar{y}) + (y_i - y_i^*))^2$$

da cui

Scomposizione della varianza di Y

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i^* - \bar{y})^2 + \frac{1}{n} \sum_{i=1}^n (y_i - y_i^*)^2 = \bar{\sigma}_y^2 + \sigma_e^2$$

- *varianza dovuta alla regressione* ($\bar{\sigma}_y^2$)
- *varianza dei residui* (σ_e^2).

L'indice che misura la bontà di adattamento (della retta di regressione ai dati) è

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2}$$

Tanto più è alto, tanto più ci riterremo soddisfatti del nostro modello

Calcoliamo tutti i residui $e_i = y_i - y_i^*$

La varianza dei residui è pari a

$$\sigma_e^2 = \frac{1}{n} \sum_{i=1}^n (y_i - y_i^*)^2 = 67.7$$

La varianza totale (varianza di Y) è

$$\sigma_y^2 = 19284.8/30 = 664.83$$

Otteniamo dunque

$$R^2 = 1 - \frac{67.7}{664.83} = 0.8947 = \rho^2$$

La regola empirica vuole che valori di $\rho^2 > 0.7$ (ovvero di R^2) siano indice di un buon adattamento del modello ai dati

Effetto degli outlier sulla retta di regressione

Abbiamo i seguenti dati

x_i	y_i
1	4
1	3
2	3
2	2

Il coefficiente di correlazione vale $\rho = -0.71$ La retta di regressione è (calcolare a e b per esercizio)

$$y = 4.5 - 1 \cdot x$$

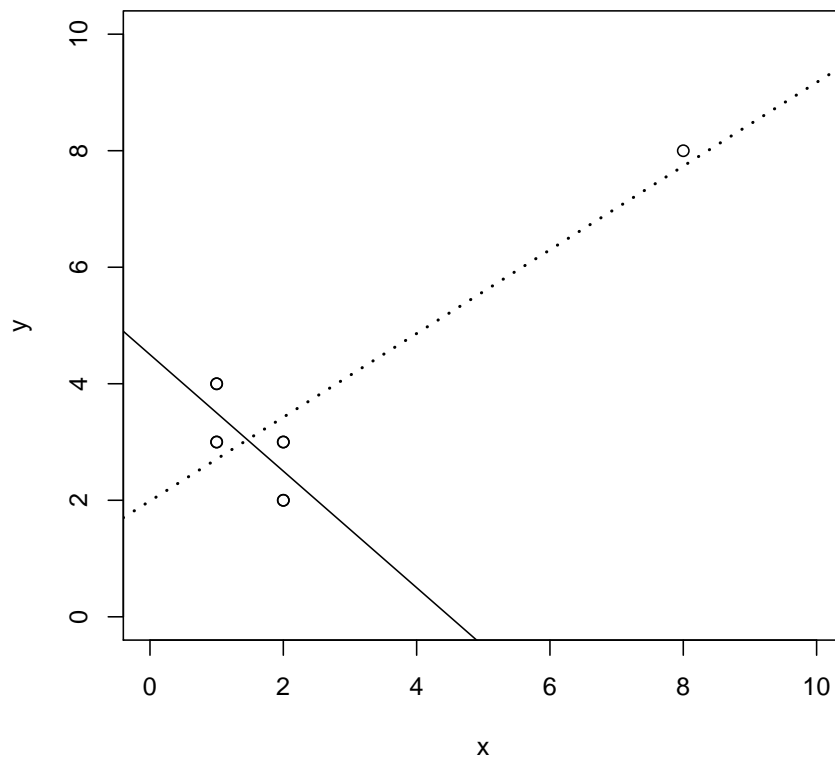
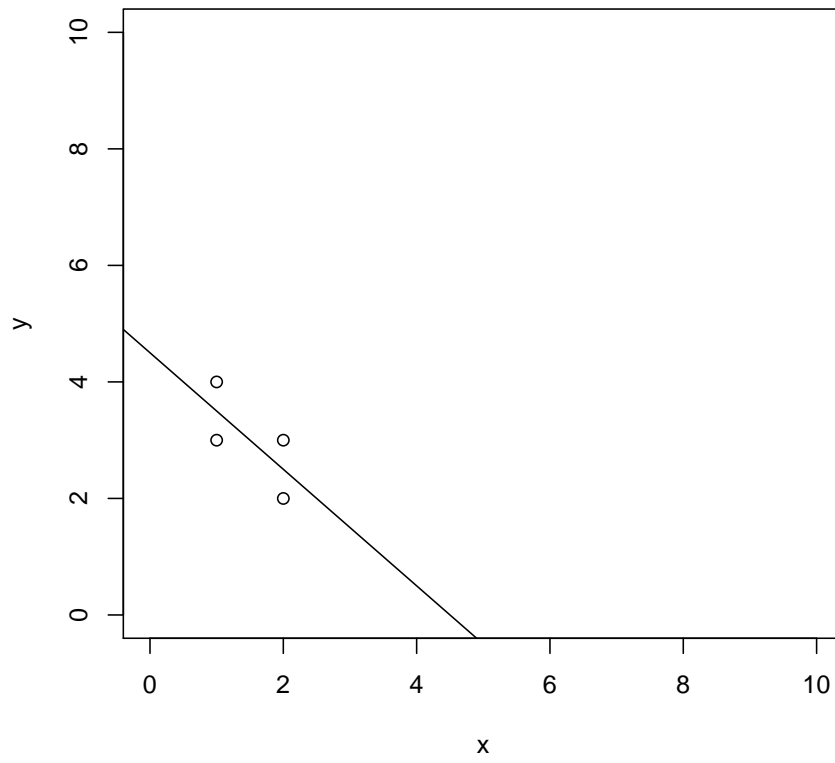
Modifichiamo la tabella aggiungendo un punto

x_i	y_i
1	4
1	3
2	3
2	2
8	8

Il coefficiente di correlazione vale $\rho = 0.9$ La retta di regressione è (calcolarle a e b per esercizio)

$$y = 1.98 + -0.72x$$

Cosa è accaduto?

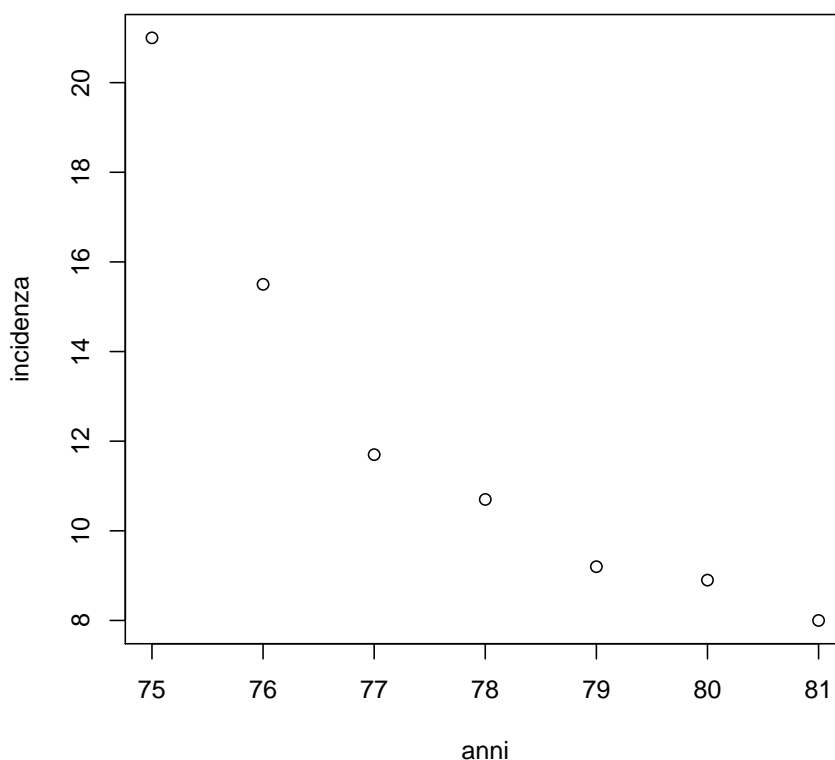


Cambiamenti di scala

Consideriamo i dati ottenuti da un'indagine epidemiologica condotta a seguito della somministrazione di un nuovo tipo di vaccino ritenuto efficace nella cura del contagio da febbre tifoidea

<i>Anno</i>	1975	1976	1977	1978	1979	1980	1981
<i>Casi</i>	21	15.5	11.7	10.7	9.2	8.9	8

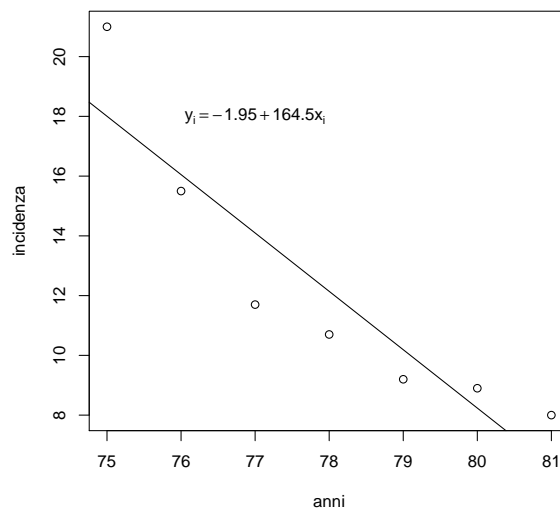
Il grafico di dispersione indica che l'andamento difficilmente possa considerarsi lineare



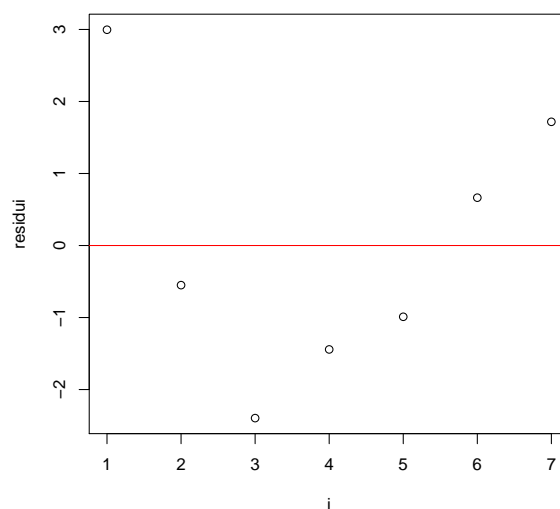
Calcoliamo il coefficiente di correlazione $\rho = -0.91$:
indica forte correlazione negativa

Calcoliamo la retta di regressione e otteniamo

$$y = 164.5 - 1.95x$$



Il grafico dei residui ci segnala un'anomalia



L'andamento che lega Y ad X sembra più prossimo ad un andamento di tipo esponenziale negativo, cioè del tipo $Y = e^{-X}$

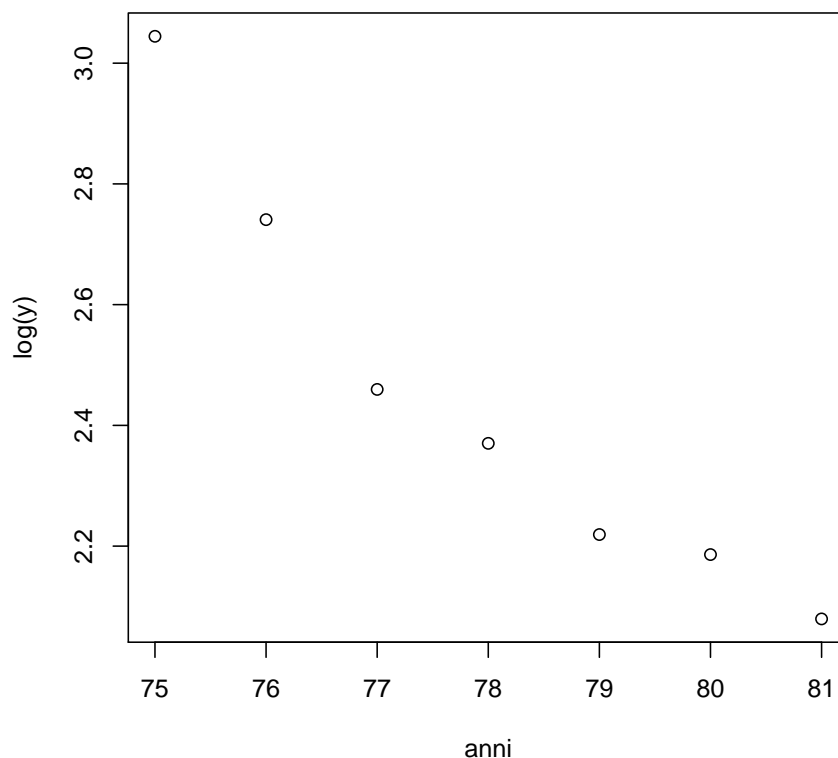
Passando ai logaritmi

$$\log(Y) = \log(e^{-X}) = -X$$

Il modello di regressione lineare più appropriato sembra essere del tipo

$$\log(Y) = a + b X$$

Rappresentiamo i punti $(x_i, \log y_i)$



Abbiamo effettuato un *cambiamento di scala* sulla variabile Y ottenendo $\log(Y)$

Calcoliamo la correlazione tra $\log(Y)$ ed X

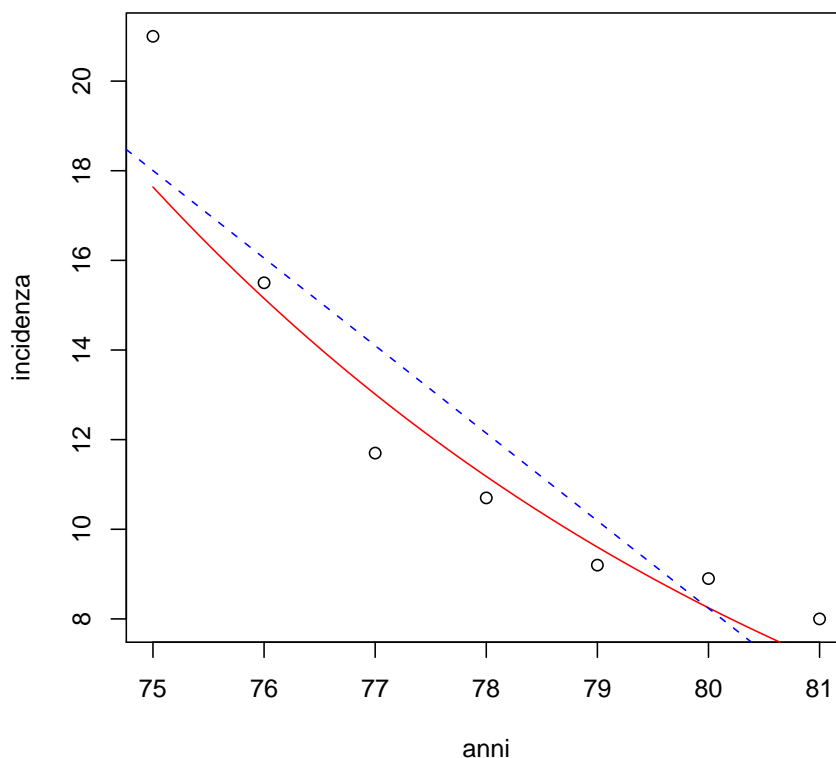
$$\rho = -0.96$$

Calcoliamo la retta di regressione usando $\log(y_i)$ al posto di y_i ed otteniamo come soluzione

$$\log(Y) = -0.152X + 14.27$$

Ovvero passando all'esponenziale

$$Y = e^{\log(Y)} = e^{-0.152X + 14.27}$$



Il modello con il cambiamento di scala dimostra la sua efficacia nel caso in cui vogliamo ottenere delle previsioni

Vogliamo prevedere il numero medio di casi di tifo per il 1985

Otteniamo per i due modelli:

$$y = -1.95x + 164.5 = -1.95 \cdot 85 + 164.5 = -1.25$$

$$y = e^{-0.152x+14.27} = e^{-0.152 \cdot 85 + 14.27} = 4.6$$

Il primo fallisce clamorosamente

Il secondo è attendibile