Esercitazione 8

Statistica

Alfonso Iodice D'Enza
iodicede@unicas.it

Università degli studi di Cassino

Outline

Misura del legame

Nel caso di variabili quantitative preferibile utilizzare una misura del legame che coinvolga, oltre le frequenze, anche le modalità (numeriche) delle variabili.
Le componenti della variabile doppia X e Y possono essere caratterizzate da diversa posizione e variabilità, risulta in genere che

$$\mu_x \neq \mu_y \text{ e } \sigma_x \neq \sigma_y$$

Volendo misurare le variazioni congiunte delle modalità di X ed Y, si fa riferimento alla versione standardizzata delle variabili, data da

$$Z_x = \frac{X - \mu_x}{\sigma_x} \text{ e } Z_y = \frac{Y - \mu_y}{\sigma_y}$$

questo per escludere dalla misura del legame gli effetti della differente media e varianza (essendo $\mu_x \neq \mu_y$ e $\sigma_x \neq \sigma_y$)

Il coefficiente di correlazione lineare di Pearson ρ

L'indice corrispondente alla media aritmetica del prodotto delle modalità standardizzate delle variabili si definisce coefficiente di correlazione lineare di Pearson ρ ed dato da

$$\rho_{xy} = \frac{1}{n}\sum_{i=1}^{n}(z_{x,i}z_{y,i}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu_x}{\sigma_x} \times \frac{y_i - \mu_y}{\sigma_y}\right)$$

Con piccole trasformazioni si ottiene la presente formalizzazione

$$\rho_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sigma_x\sigma_y} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

La quantità al numeratore si definisce covarianza: essa corrisponde alla media del prodotto degli scarti delle modalità di X e Y dalle rispettive medie. La covarianza misura la contenporanea variazione di X e Y con riferimento alle loro medie.

Proprietà del coefficiente di correlazione

- se X e Y sono indipendenti, allora $\rho_{xy} = 0$ (NON vale il contrario)
- se $\rho_{xy} = 1$, allora $Y = \alpha + \beta X$ (ovvero Y una trasformazione lineare di X)
- se $\rho_{xy} = -1$, allora $Y = \alpha - \beta X$ (ovvero Y una trasformazione lineare di X)
- $\rho_{xy} = \rho_{yx}$
- $\rho_{xx} = 0$

Il coefficiente di correlazione lineare di Pearson ρ

Esercizio

Si considerino i voti riportati da $n = 8$ studenti negli esami di $matematica$ e $statistica$.

	$matematica(x_i)$	$statistica(y_i)$
1	24	23
2	27	28
3	30	30
4	26	27
5	29	30
6	18	20
7	21	20
8	22	25

- Si misuri il legame lineare che caratterizza le due variabili

Svolgimento

É necessario calcolare le medie aritmetiche μ e gli scarti quadratici medi σ

- Il voto medio ottenuto dagli studenti all'esame di matematica è
  $\mu_m = \frac{\sum_{i=1}^{8} x_i}{n} = \frac{197}{8} = 24.625$
- Il voto medio ottenuto dagli studenti all'esame di statistica è $\mu_s = \frac{\sum_{i=1}^{8} y_i}{n} = \frac{203}{8} = 25.375$

	x_i	y_i	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x)^2$	$(y_i - \mu_y)^2$
1	24	23	-0.62	-2.38	0.39	5.64
2	27	28	2.38	2.62	5.64	6.89
3	30	30	5.38	4.62	28.89	21.39
4	26	27	1.38	1.62	1.89	2.64
5	29	30	4.38	4.62	19.14	21.39
6	18	20	-6.62	-5.38	43.89	28.89
7	21	20	-3.62	-5.38	13.14	28.89
8	22	25	-2.62	-0.38	6.89	0.14
Tot	197	203			119.875	115.875

scarti quadratici medi: $\sigma_m = \sqrt{\frac{\sum_{i=1}^{8} (x_i - \mu_m)^2}{n}} = \sqrt{\frac{119.875}{8}} = 3.87$

$\sigma_s = \sqrt{\frac{\sum_{i=1}^{8} (y_i - \mu_s)^2}{n}} = \sqrt{\frac{115.875}{8}} = 3.805$

Svolgimento

Per calcolare il coefficiente di correlazione lineare resta da calcolare la covarianza, ovvero la media aritmetica del prodotto degli scarti dalla media.

	x_i	y_i	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x) \times (y_i - \mu_y)$
1	24.00	23.00	-0.62	-2.38	1.48
2	27.00	28.00	2.38	2.62	6.23
3	30.00	30.00	5.38	4.62	24.86
4	26.00	27.00	1.38	1.62	2.23
5	29.00	30.00	4.38	4.62	20.23
6	18.00	20.00	-6.62	-5.38	35.61
7	21.00	20.00	-3.62	-5.38	19.48
8	22.00	25.00	-2.62	-0.38	0.98
Tot	197	203			111.125

La covarianza è

$$\sigma_{ms} = \frac{\sum_{i=1}^{8} (x_i - \mu_m)(y_i - \mu_s)}{n} = \frac{111.125}{8} = 13.89$$

É ora possibile calcolare il coefficiente di correlazione dato da

$$\rho_{ms} = \frac{\sigma_{ms}}{\sigma_m \sigma_s} = \frac{13.89}{3.87 \times 3.805} = 0.943$$

Metodo alternativo per il calcolo di ρ

Da un punto di vista computazionale risulta conveniente l'utilizzo della seguente formulazione alternativa del coefficiente di correlazione lineare ρ basata sulle somme delle modalità delle componenti $(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i)$, sulle somme dei quadrati delle modalità delle componenti $(\sum_{i=1}^{n} (x_i)^2, \sum_{i=1}^{n} (y_i)^2)$, sulla somma dei prodotti tra le modalità $(\sum_{i=1}^{n} x_i y_i)$

$$\rho = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{(n \sum_{i=1}^{n} (x_i)^2 - \left[\sum_{i=1}^{n} x_i\right]^2)(n \sum_{i=1}^{n} (y_i)^2 - \left[\sum_{i=1}^{n} y_i\right]^2)}}$$

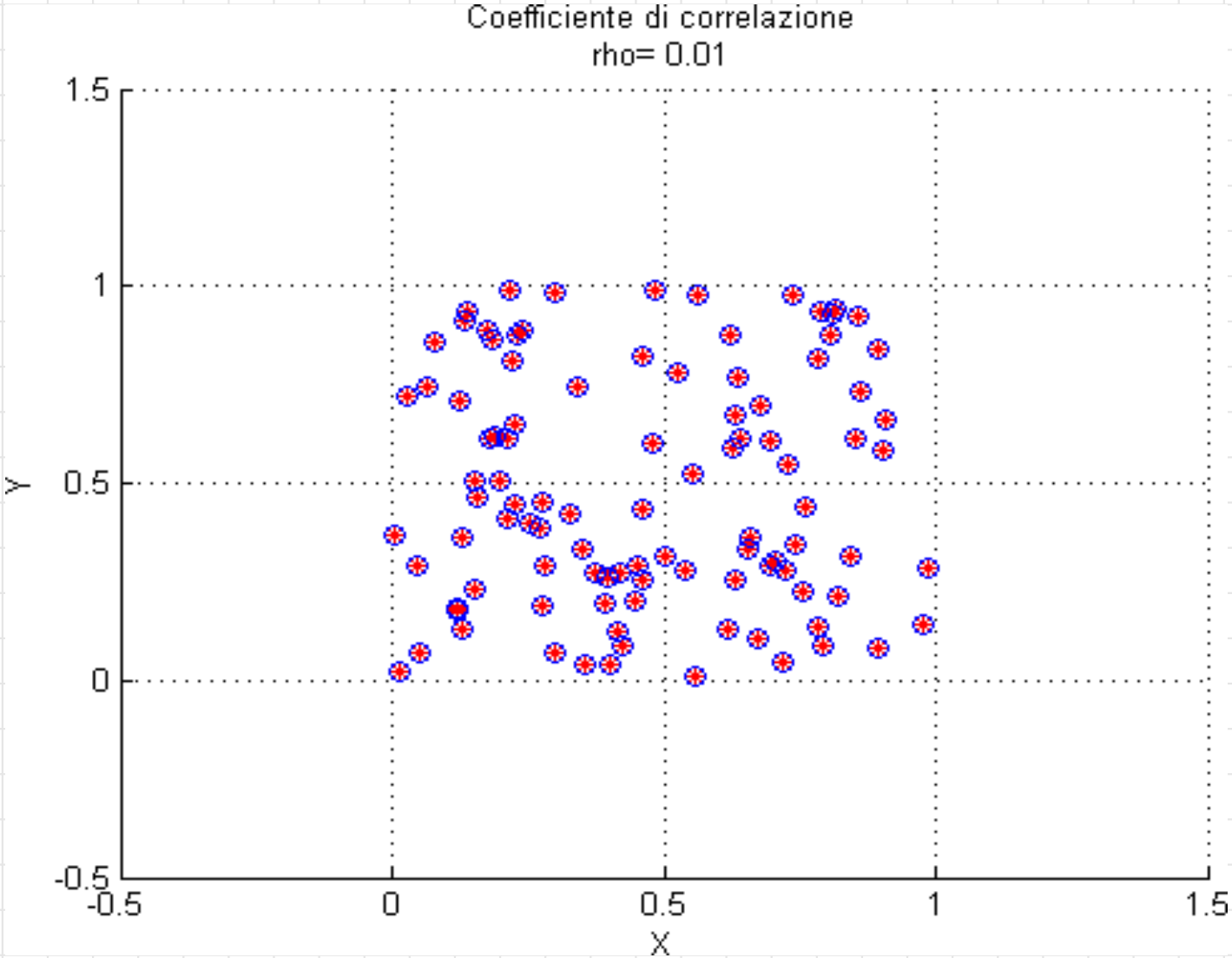
Metodo alternativo per il calcolo di ρ

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	24	23	576	529	552
2	27	28	729	784	756
3	30	30	900	900	900
4	26	27	676	729	702
5	29	30	841	900	870
6	18	20	324	400	360
7	21	20	441	400	420
8	22	25	484	625	550
	$\sum x = 197$	$\sum y = 203$	$\sum x^2 = 4971$	$\sum y^2 = 5267$	$\sum xy = 5110$

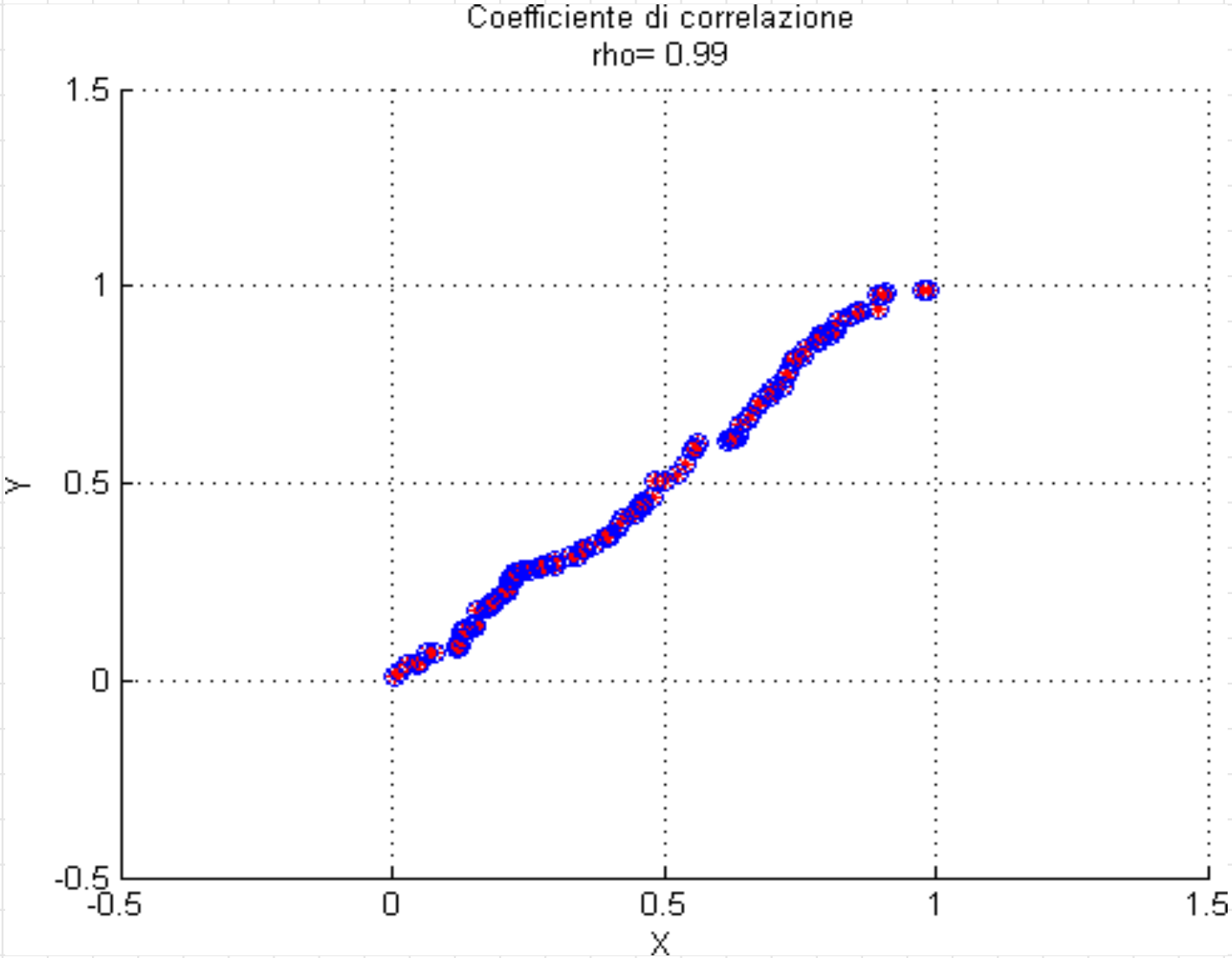
$$\rho = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{(n\sum_{i=1}^{n}(x_i)^2 - \left[\sum_{i=1}^{n} x_i\right]^2)(n\sum_{i=1}^{n}(y_i)^2 - \left[\sum_{i=1}^{n} y_i\right]^2)}} =$$

$$= \frac{8 \times 5110 - (197 \times 203)}{\sqrt{(8 \times 4971 - (197)^2) \times (8 \times 5267 - (203)^2)}} = 0.943$$

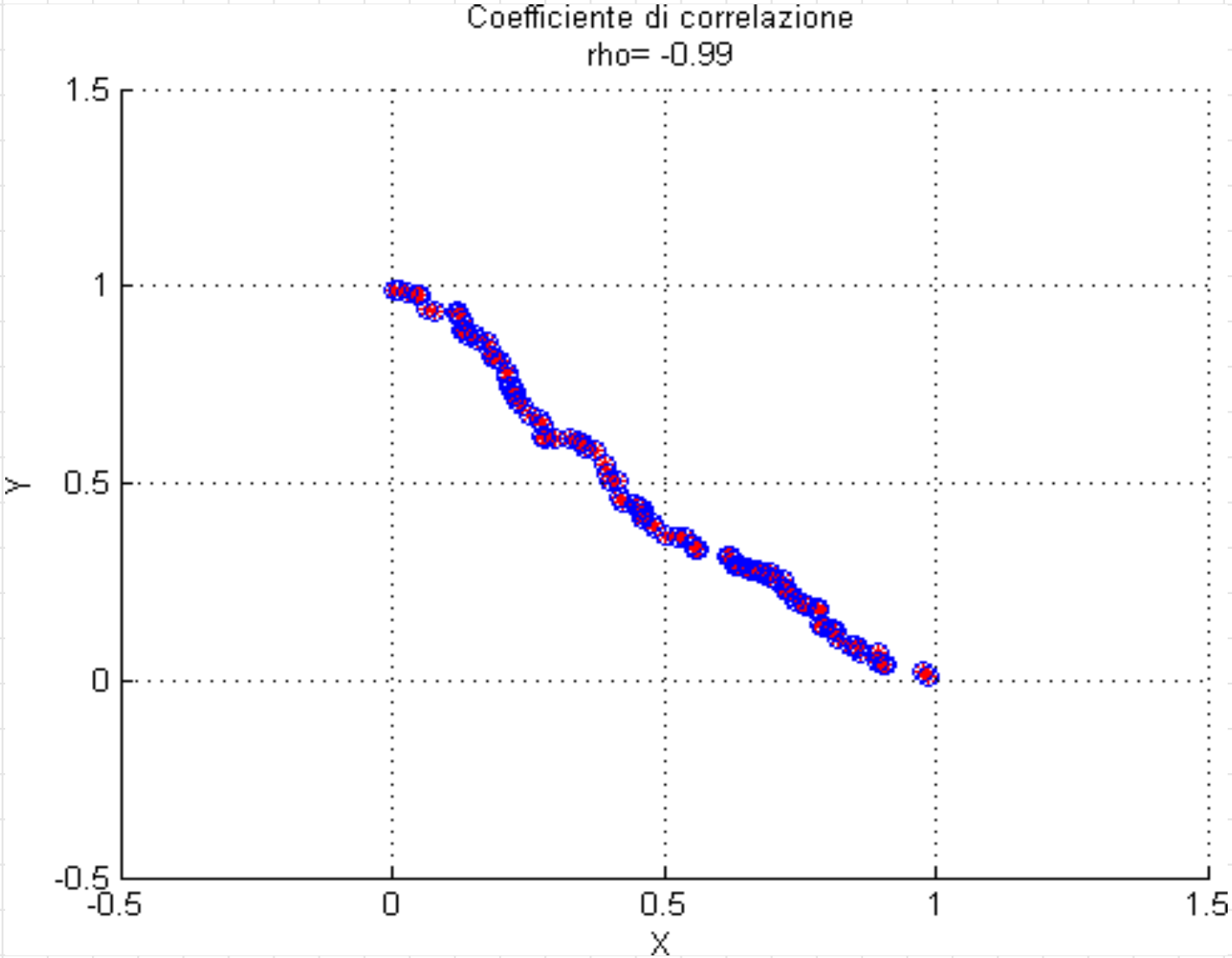
Coefficiente di correlazione: esempi di casi limite



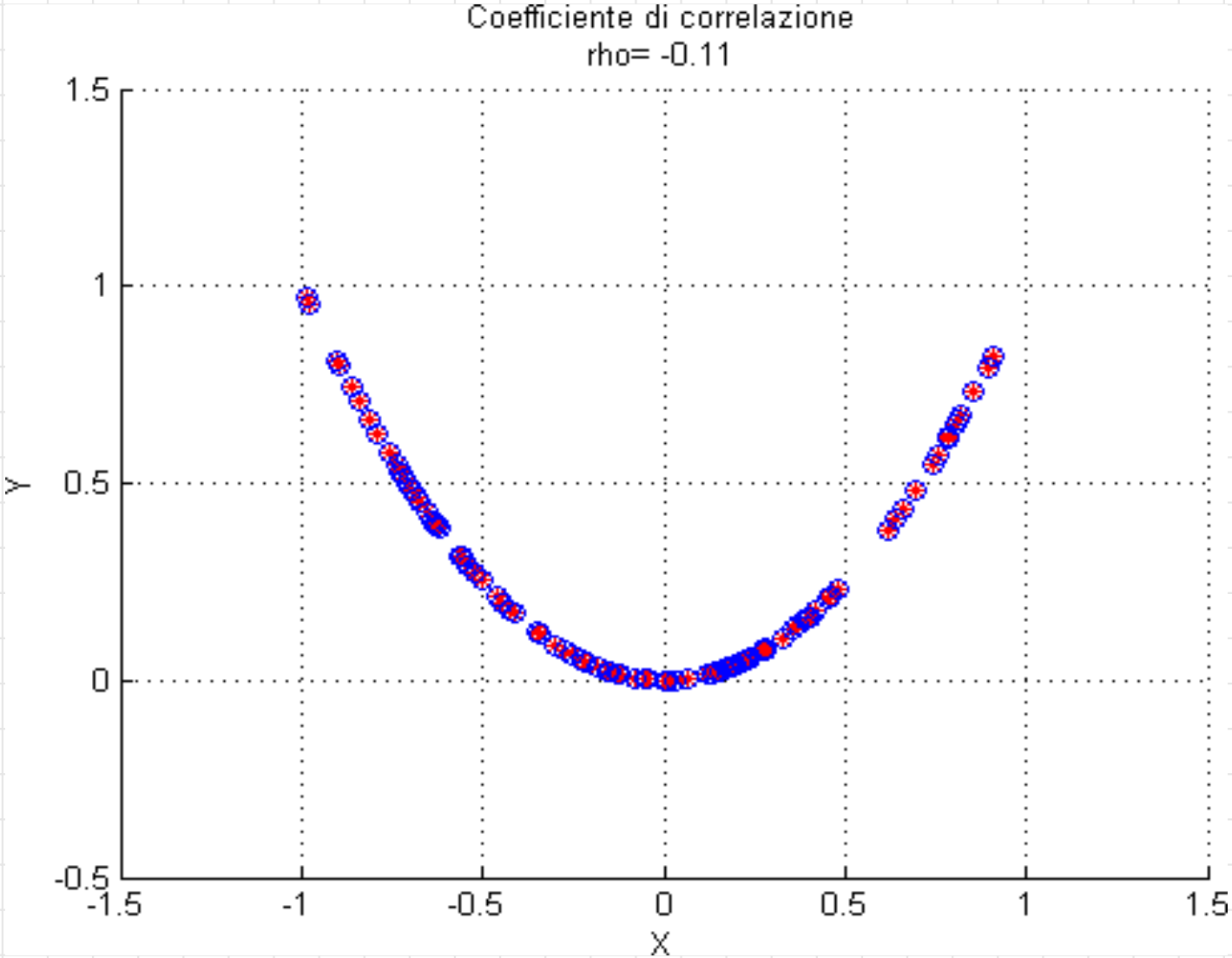
Coefficiente di correlazione: esempi di casi limite



Coefficiente di correlazione: esempi di casi limite



Coefficiente di correlazione: esempi di casi limite

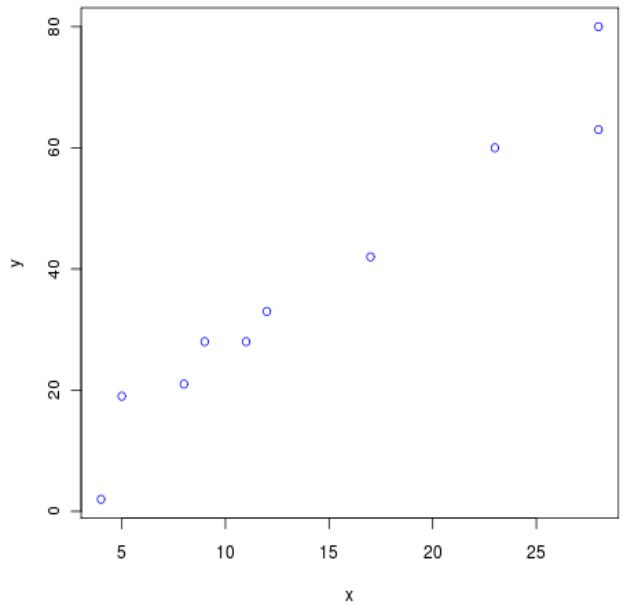


Dipendenza lineare

Lo studio della relazione tra caratteri statistici è, nel caso della interdipendenza, di tipo simmetrico: due caratteri quantitativi X e Y hanno lo stesso ruolo e si vuole studiare se essi siano indipendenti o meno. A questo scopo sono stati introdotti gli indici di covarianza σ_{xy} e di correlazione lineare ρ. Si consideri di aver osservato due caratteri quantitativi X ed Y. Si riportano i valori e il grafico di dispersione:

	Y	X
1	28	11
2	21	8
3	63	28
4	42	17
5	28	9
6	2	4
7	80	28
8	19	5
9	33	12
10	60	23
	376	145

Il diagramma di dispersione (scatter plot)



Dipendenza lineare

covarianza e coefficiente di correlazione

- $\mu_x = \frac{\sum_{i=1}^{10} x_i}{10} = 14.5$
- $\mu_y = \frac{\sum_{i=1}^{10} y_i}{10} = 37.6$
- $\sigma_x = \sqrt{\frac{\sum_{i=1}^{10} (x_i - \mu_x)^2}{10}} = 8.57$
- $\sigma_y = \sqrt{\frac{\sum_{i=1}^{10} (y_i - \mu_y)^2}{10}} = 22.49$
- $\sigma_{xy} = \frac{\sum_{i=1}^{10} (x_i - \mu_x)(y_i - \mu_y)}{10} = 187.3$
- $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = 0.97$

Dipendenza funzionale lineare

Essendo il valore del coefficiente di correlazione lineare prossimo ad 1 esiste una forte relazione lineare tra X ed Y. Come confermato dal grafico di dispersione, i dati sono approssimativamente allineati lungo una retta crescente. Ci si può dunque aspettare che sussista una relazione funzionale tra i dati del tipo

$$Y = f(X) = b_0 + b_1 X$$

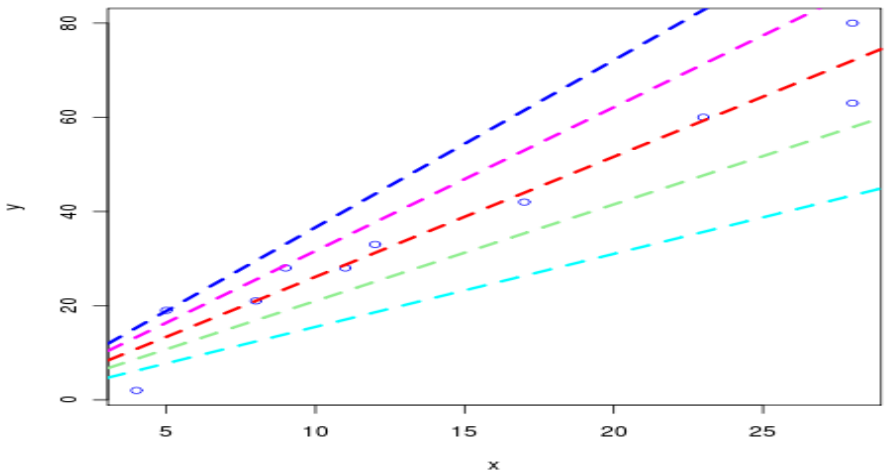
che rappresenta l'equazione di una retta passante attraverso la nube di punti di coordinate (x_i, y_i).

La retta di regressione

La retta di regressione

La retta di regressione fornisce una approssimazione della dipendenza dei valori di Y dai valori di X. La relazione di dipendenza non è esattamente riprodotta dalla retta; i valori $y_i^* = b_0 + b_1 x_i$ sono dunque i valori teorici, ovvero i valori che la variabile Y assume, secondo il modello $Y = b_0 + b_1 X$, in corrispondenza dei valori x_i osservati.

rette passanti per la nube di punti



Determinazione della retta di regressione

L'identificazione della retta avviene attraverso la determinazione dei valori di b_0, l'intercetta, e b_1, il coefficiente angolare o pendenza. La retta 'migliore' è quella che passa più 'vicina' ai punti osservati. In altre parole, si vuole trovare la retta per la quale le differenze tra i valori teorici y_i^* e i valori osservati y_i siano minime.

La retta di regressione

I residui

le differenze tra i valori teorici y_i^* e i valori osservati y_i vengono definite residui. La retta di regressione è tale che la somma dei residui al quadrato sia minima. Formalmente

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y_i^*)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Il problema consiste dunque nel ricercare b_0 e b_1 che minimizzano la precedente espressione. Da un punto di vista operativo bisogna risolvere il seguente sistema di equazioni

$$\frac{\partial}{\partial b_0} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = 0$$

$$\frac{\partial}{\partial b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = 0$$

Ricerca dei parametri della retta di regressione: (b_0)

$$-2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) =$$

$$\sum_{i=1}^{n} y_i - n * b_0 - b_1 \sum_{i=1}^{n} x_i = 0$$

$$b_0 = \mu_y - b_1 \mu_x$$

La retta di regressione

I residui

le differenze tra i valori teorici y_i^* e i valori osservati y_i vengono definite residui. La retta di regressione è tale che la somma dei residui al quadrato sia minima. Formalmente

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

$$= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Il problema consiste dunque nel ricercare b_0 e b_1 che minimizzano la precedente espressione. Da un punto di vista operativo bisogna risolvere il seguente sistema di equazioni

$$\frac{\partial}{\partial b_0} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = 0$$

$$\frac{\partial}{\partial b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = 0$$

Ricerca dei parametri della retta di regressione:(b_1)

$$-2 \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \left(\frac{\sum_{i=1}^{n} y_i}{n} - b_1 \frac{\sum_{i=1}^{n} x_i}{n} \right)$$

$$b_1 \left(n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2 \right) = n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\sigma_{xy}}{\sigma_x^2}$$

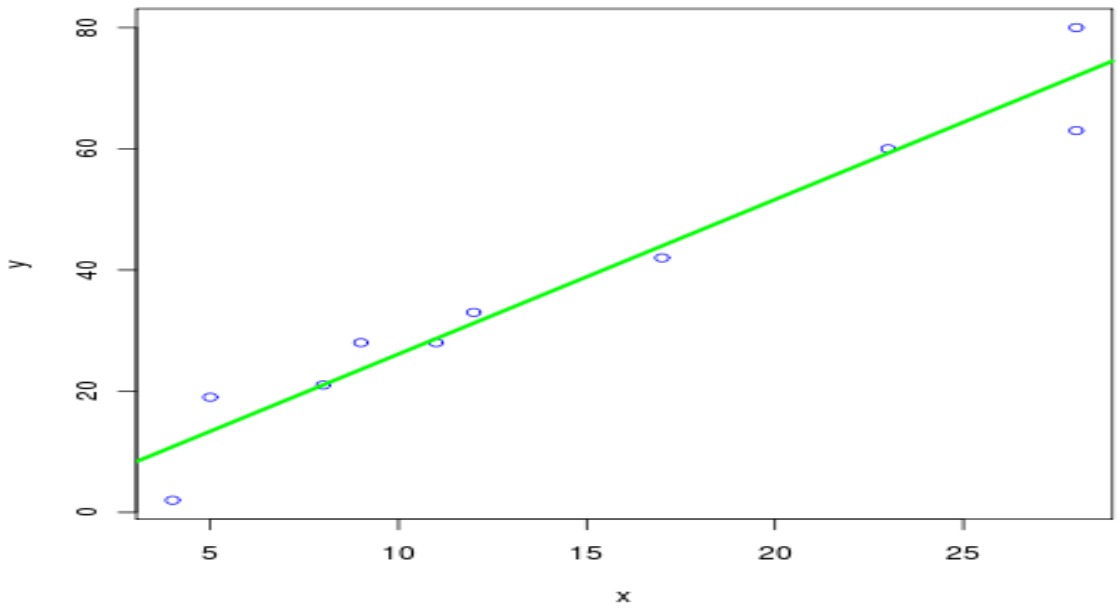
Determinazione della retta di regressione

Calcolo dei coefficienti

Richiamando le quantità calcolate in precedenza e le formule per il calcolo dei parametri si ha

- $b_1 = \frac{\sigma_{xy}}{\sigma_x^2} = 2.55$
- $b_0 = \mu_y - b_1 \mu_x = 37.6 - (2.55 * 14.5) = 0.62$

La retta 'migliore'



Interpretazione dei valori dei coefficienti di regressione

- b_0 rappresenta l'intercetta della retta di regressione ed indica il valore della variabile di risposta Y quando il predittore X assume valore 0.
- b_1 rappresenta l'inclinazione della retta di regressione, ovvero la variazione della variabile di risposta Y in conseguenza di un aumento unitario del predittore X.

Bontà di adattamento

Esistono diversi strumenti grafici ed analitici per valutare la bontà dell'adattamento della retta di regressione ai dati

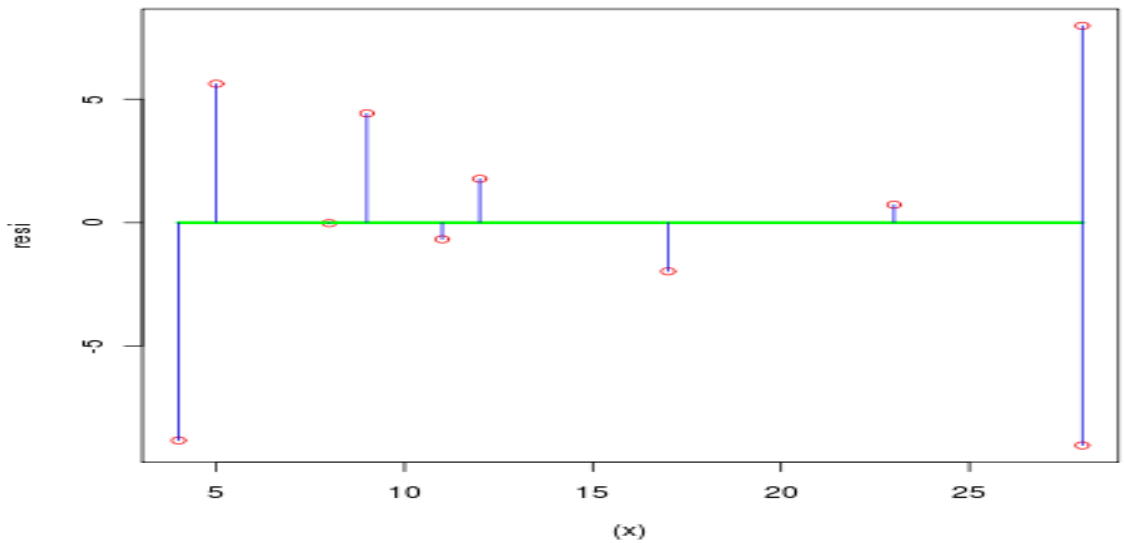
- Strumenti grafici: plot dei residui
- Strumenti analitici:coefficiente di determinazione lineare R^2

Plot dei residui

Perchè la retta possa essere considerata una buona approssimazione della relazione che intercorre tra Y ed X è necessario che i residui abbiano un andamento casuale rispetto ai valori della X. Se, ad esempio, all'aumentare dei valori della X aumentassero sistematicamente anche i residui, allora la relazione potrebbe non essere non lineare: la retta di regressione ne sarebbe dunque una cattiva approssimazione.

Plot dei residui

Per verificare che l'andamento dei residui sia effettivamente casuale rispetto ad X, è possibile utilizzare un diagramma di dispesione tra i valori x_i ed i corrispondenti residui $e_i (i = 1, \ldots, n)$



coefficiente di determinazione lineare R^2

Ricordando che la devianza il numeratore della varianza...

$$Dev_y = \sum_{i=1}^{n} (y_i - \mu_y)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \mu_y)^2 =$$
$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \mu_y)^2 + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \mu_y)$$
$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \mu_y)^2 + 2(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i)(\sum_{i=1}^{n} \hat{y}_i - n\mu_y)$$

Il metodo dei minimi quadrati assicura che $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$, quindi

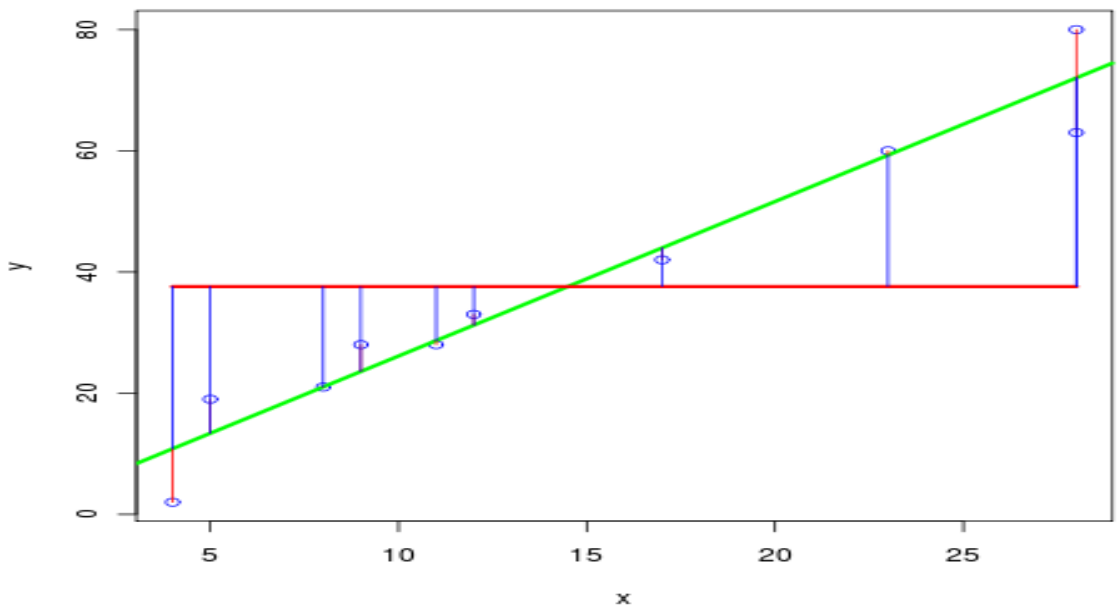
$$Dev(y) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \mu_y)^2 + 2 * 0 * (\sum_{i=1}^{n} \hat{y}_i - n\mu_y)$$
$$= \sum_{i=1}^{n} (\hat{y}_i - \mu_y)^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = Dev_r + Dev_e$$

Decomposizione della devianza

La devianza può essere decomposta dunque nelle seguenti quantità $Dev_y = Dev_r + Dev_e$

- $Dev_y = \sum_{i=1}^{n} (y_i - \mu_y)^2$ devianza totale
- $Dev_r = \sum_{i=1}^{n} (\hat{y}_i - \mu_y)^2$ devianza di regressione
- $Dev_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ devianza dei residui

Interpretazione grafica



Bontà dell'adattamento

Intituitivamente, l'adattamento della retta è migliore quanto maggiore sarà proporzione di variabilità totale che la retta di regressione riesce a spiegare; ovvero, l'adattamento della retta è migliore quanto minore sarà la variabilità residua. Una misura di come il modello approssima i dati osservati è data dal coefficiente di determinazione lineare R^2, dato da

$$R^2 = \frac{Dev_r}{Dev_y} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \mu_y)^2}{\sum_{i=1}^{n} (y_i - \mu_y)^2}$$

ovvero

$$R^2 = 1 - \frac{Dev_e}{Dev_y} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \mu_y)^2}$$

esempio di calcolo R^2

- $Dev_y = \sum_{i=1}^{n} (y_i - \mu_y)^2 = 5058.4$
- $Dev_r = \sum_{i=1}^{n} (\hat{y}_i - \mu_y)^2 = 4776.214$
- $Dev_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 282.1862$

$$R^2 = \frac{Dev_r}{Dev_y} = \frac{4776.214}{5058.4} = 0.94$$

ovvero

$$R^2 = 1 - \frac{Dev_e}{Dev_y} = 1 - \frac{282.1862}{5058.4} = 1 - 10.53 = 0.94$$

Influenza di un outlier sulla soluzione

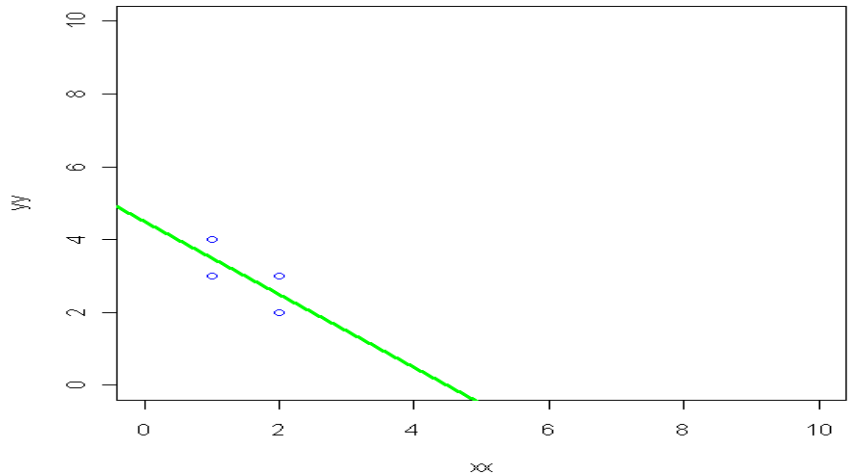
Un piccolo esempio

Si considerino le seguenti osservazioni

```
xx  yy
 1   4
 1   3
 2   3
 2   2
```

Retta di regressione

La soluzione induce a concludere che vi sia una relazione di proporzionalità inversa: poichè la retta è decrescente si deduce che all'aumentare di X, la variabile dipendente Y diminuisce.



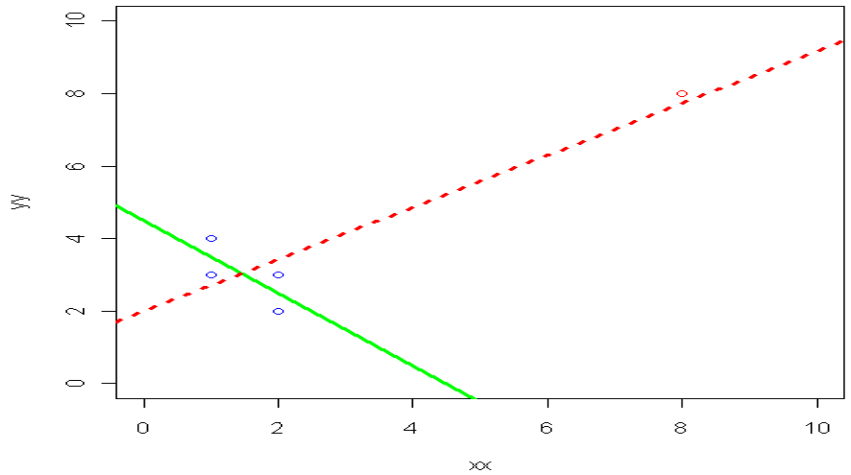
Influenza di un outlier sulla soluzione

Un (altro) piccolo esempio

Si considerino le osservazioni precedenti a cui è aggiunta un unica coppia di valori $(8, 8)$. I dati sono

xx2	yy2
1	4
1	3
2	3
2	2
8	8

Retta di regressione



In questo caso, la sola presenza della nuova osservazione conduce all'identificazione di una retta di regressione diversa dalla prima: l'inclinazione positiva della retta indica una relazione di diretta proporzionalità. Tuttavia tale soluzione è unicamente dovuta dalla presenza dell' osservazione $(8, 8)$ che pertanto induce a valutare la relazione di dipendenza tra Y ed X in maniera errata. L'osservazione $(8, 8)$ si definisce pertanto un outlier. L'identificazione e la conseguente eliminazione degli eventuali outlier è un elemento molto importante nello studio della dipendenza tra fenomeni.

Esercizio regressione: distribuzione doppia di frequenze

Si consideri di aver osservato su 10 rivenditori di componenti informatiche le variabili *numero di punti vendita* e *Fatturato settimanale complessivo*. Si studi la dipendenza del fatturato dal numero di punti vendita.

	fino a 2	tra 2 e 4	tra 4 e 6
fino a 5000	3	2	0
tra 5000 e 1000	1	2	2

- Si stimino i coefficienti della retta di regressione.
- Si valuti la bontà di adattamento della retta ai dati.

Esercizio regressione: distribuzione doppia di frequenze

Essendo le modalità delle variabili qualitative espresse in intervalli di valori, è necessario fare riferimento ai centri di ciascun intervallo. La tabella è dunque data da

Y/X	1	3	5	Tot
2500	3	2	0	5
7500	1	2	2	5
Tot	4	4	2	10

- Le medie aritmetiche si ottengono a partire dalle distribuzioni marginali di frequenze:

$$\mu_x = \frac{1}{n}\sum_{j=1}^{k} x_j n_{.j} = \frac{1}{10} \times (1 \times 4) + (3 \times 4) + (5 \times 2) = \frac{4 + 12 + 10}{10} = 2.6$$

$$\mu_y = \frac{1}{n}\sum_{i=1}^{h} y_i n_{i.} = \frac{1}{10} \times (2500 \times 5) + (7500 \times 5) = \frac{12500 + 37500}{10} = 5000$$

dove h rappresenta numero di righe della tabella, k il numero di colonne della tabella.

Esercizio regressione: distribuzione doppia di frequenze

Per calcolare le varianze si fa riferimento agli scarti dalla media al quadrato

Y/X	$(1-2.6)^2$	$(3-2.6)^2$	$(5-2.6)^2$	Tot
$(2500-5000)^2$	3	2	0	5
$(7500-5000)^2$	1	2	2	5
Tot	4	4	2	10

- Le varianze si ottengono a partire dalle distribuzioni marginali di frequenze:

$$\sigma_x^2 = \frac{1}{n}\sum_{j=1}^{k}(x_j-\mu_x)^2 n_{.j} = \frac{1}{10}\times((1-2.6)^2\times 4)+((3-2.6)^2\times 4)+$$

$$+((5-2.6)^2\times 2) = \frac{10.24+0.64+11.52}{10} = 2.24$$

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{h}(y_i-\mu_y)^2 n_{i.} = \frac{1}{10}\times(2500\times 5)^2+(7500\times 5)^2$$

$$= \frac{31250000+31250000}{10} = 6250000$$

dove h rappresenta numero di righe della tabella, k il numero di colonne della tabella.

Esercizio regressione: distribuzione doppia di frequenze

Per calcolare la covarianza si deve fare riferimento alle distribuzioni condizionate di frequenza.

Y/X	$(1-2.6)$	$(3-2.6)$	$(5-2.6)$	Tot
$(2500-5000)$	3	2	0	5
$(7500-5000)$	1	2	2	5
Tot	4	4	2	10

y_i	x_i	$y_i-\mu_y$	$x_i-\mu_x$
2500	1	(2500-5000)	(1-2.6)
2500	1	(2500-5000)	(1-2.6)
2500	1	(2500-5000)	(1-2.6)
2500	3	(2500-5000)	(3-2.6)
2500	3	(2500-5000)	(3-2.6)
7500	1	(7500-5000)	(1-2.6)
7500	3	(7500-5000)	(3-2.6)
7500	3	(7500-5000)	(3-2.6)
7500	5	(7500-5000)	(5-2.6)
7500	5	(7500-5000)	(5-2.6)

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{h}\sum_{j=1}^{k}(y_i-\mu_y)\times(x_j-\mu_x)\times n_{ij} =$$

$$= \frac{1}{10}((2500-5000)(1-2.6)\times 3 + (2500-5000)(3-2.6)\times 2 +$$

$$+ (7500-5000)(1-2.6)\times 1 + (7500-5000)(3-2.6)\times 2 +$$

$$+ (7500-5000)(5-2.6)\times 2) = \frac{12000-2000-4000+2000+12000}{10} = 2000$$

Esercizio regressione: distribuzione doppia di frequenze

Avendo calcolato le quantità $\mu_x = 2.6$, $\mu_y = 5000$, $\sigma_x^2 = 2.24$ e $\sigma_{xy} = 2000$, è possibile calcolare i coefficienti della retta di regressione

Calcolo dei coefficienti

- $b_1 = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{2000}{2.24} = 892.571$
- $b_0 = \mu_y - b_1 \mu_x = 5000 - (892.571 * 2.6) = 2679.315$

quindi l'equazione della retta di regressione è

$$y = b_0 + b_1 x = 2679.315 + 892.571x$$

Dunque, il valore stimato $\hat{y}_i$ corrispondente ad un valore x_i assegnato è $\hat{y}_i = b_0 + b_1 x$.

Valutazione della bontà di adattamento

Ricordando che

$$R^2 = \frac{Dev_r}{Dev_y} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \mu_y)^2}{\sum_{i=1}^{n} (y_i - \mu_y)^2}$$

ovvero

$$R^2 = 1 - \frac{Dev_e}{Dev_y} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \mu_y)^2}$$

con $Dev_y = Dev_r + Dev_e$

- $Dev_y = \sum_{i=1}^{n} (y_i - \mu_y)^2$ devianza totale
- $Dev_r = \sum_{i=1}^{n} (\hat{y}_i - \mu_y)^2$ devianza di regressione
- $Dev_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ devianza dei residui

Per ottenere R^2, misura della bontà di adattamento, si deve calcolare solo la devianza dei residui, avendo già calcolato σ_y^2.

Calcolo della devianza dei residui

- $Dev_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ devianza dei residui

in base alla retta di regressione stimata, i valori $\hat{y}_i$ stimati in funzione dei valori x_i sono

- $\hat{y}_1 = b_0 + b_1 x_1 = 2679.315 + 892.571 \times 1 = 3571.886$
- $\hat{y}_2 = b_0 + b_1 x_2 = 2679.315 + 892.571 \times 3 = 5357.028$
- $\hat{y}_3 = b_0 + b_1 x_3 = 2679.315 + 892.571 \times 5 = 7142.17$

Calcolo della devianza dei residui

Per calcolare i residui $y_i - \hat{y}_i$ nel caso di tabella a doppia entrata si procede come segue

$y_i / \hat{y}_j$	$\hat{y}_1 = 3571.886$	$\hat{y}_2 = 5357.028$	$\hat{y}_3 = 7142.17$	Tot
$y_1 = 2500$	3	2	0	5
$y_2 = 7500$	1	2	2	5
Tot	4	4	2	10

- $Dev_e = \sum_{i=1}^{h} \sum_{i=1}^{k} ((y_i - \hat{y}_j)^2) \times n_{ij}$ devianza dei residui per tabella doppia

calcolo della devianza dei residui

$$Dev_e = \sum_{i=1}^{h} \sum_{j=1}^{k} ((y_i - \hat{y}_j)^2) \times n_{ij} = ((2500 - 3571.886)^2) \times 3 + ((2500 - 5357.028)^2) \times 2 +$$
$$+ ((7500 - 3571.886)^2) \times 1 + ((7500 - 5357.028)^2) \times 2 + ((7500 - 7142.17)^2) \times 2 =$$
$$= 44642859$$

$$dev_y = \sum_{i=1}^{n} (y_i - \mu_y)^2 = \sigma_y^2 \times n = 6250000 \times 10 = 62500000$$

$$R^2 = 1 - \frac{dev_e}{dev_y} = 1 - 0.71 = 0.29$$