Statistica 1

Dati bivariati. II - Correlazione e regressione

Alessandra Salvan e Laura Ventura

Dipartimento di Scienze Statistiche
Università di Padova
http://www.stat.unipd.it/

salvan@stat.unipd.it
ventura@stat.unipd.it

copyright@2013-2014

Dati bivariati

- In molte situazioni interessa studiare se esiste una relazione tra due variabili misurate sulle stesse unità. Esempi:
  – *"Le misurazioni del peso prima della terapia sono in relazione con le misurazioni dopo la terapia?"*
  – *"il voto di maturità è in relazione con la performance universitaria?"*
- Oppure si desidera prevedere il valore di una variabile conoscendo il valore di un'altra. Esempi:
  – *"conoscendo l'altezza del padre, è possibile prevedere l'altezza di un figlio?"*
  – *"conoscendo la durata della gravidanza, si può stimare il peso alla nascita?"*
- La statistica permette di rispondere a questo tipo di domande, con strumenti adatti alla natura delle variabili in esame. A tale scopo, per variabili quantitative, si tratteranno:
  – La CORRELAZIONE, che misura la dipendenza lineare tra due variabili;
  – La REGRESSIONE, che valuta la relazione lineare tra due variabili.

Correlazione

- La correlazione misura l'associazione tra due variabili quantitative. È lo strumento che si utilizza quando si hanno a disposizione coppie di valori di variabili ⇒ permette di valutare come variano i valori di una variabile al variare dell'altra e viceversa.
- Esempi:
  – Numero di sigarette fumate in gravidanza e tasso di crescita del feto ⇒ all'aumentare del numero di sigarette fumate diminuisce il tasso di crescita (correlazione negativa).
  – Livello di colesterolo e BMI (Body Mass Index = peso (kg)/altezza^2 (m^2)) ⇒ tanto è maggiore il livello di colesterolo quanto è maggiore il BMI (correlazione positiva).
  – Il valor medio della temperatura (ambiente) e il BMI ⇒ non c'è motivo di pensare che la temperatura influenzi il BMI delle persone (assenza di correlazione).
- La relazione può essere valutata tramite:
  – Un grafico (grafico di dispersione)
  – Un indice che quantifica il grado di correlazione (coefficiente di correlazione)

Diagramma di dispersione

- Nello studio dell'associazione tra due variabili quantitative misurate sulle stesse unità statistiche, indicate con X e Y, è molto utile disegnare un grafico, il diagramma di dispersione, prima di procedere con altre analisi formali.

Nel grafico di dispersione le coppie

$$(x_1, y_1)\ (x_2, y_2)\ \ldots\ (x_n, y_n)$$

di valori di due variabili quantitative misurate sulle n unità sono rappresentati come punti di un piano cartesiano, i cui assi corrispondono alle due variabili.

Medie e varianze di X e Y

La media aritmetica e la varianza di X sono

$$m_x = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \,,$$

e

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - m_x^2 \,.$$

Analogamente, si indicano con m_y e S_y^2 media e varianza di Y.

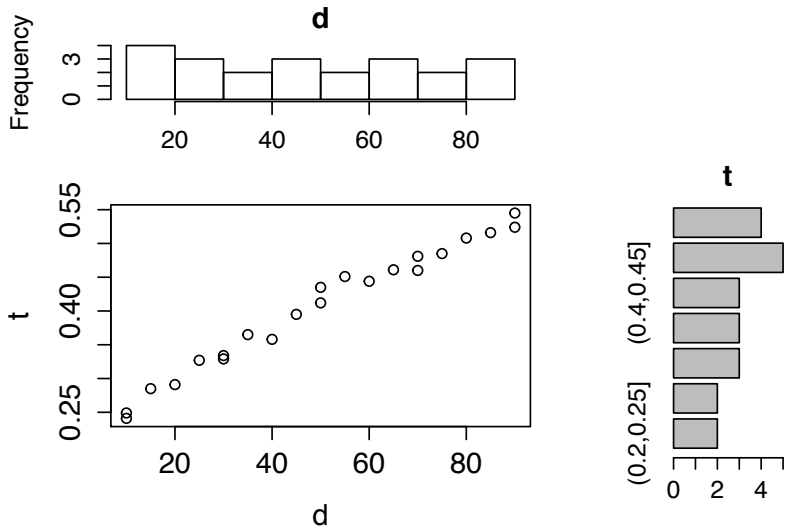
DIAGRAMMA DI DISPERSIONE

– Ogni punto del grafico rappresenta una unità.

– Permette di verificare visivamente se le coppie di punti presentano una qualche forma di regolarità e per vedere come i punti si dispersdono intorno a un particolare punto di riferimento: il baricentro della nuvola dei punti, ossia il punto di coordinate (m_x, m_y).

– La nuvola di punti ha una forma allungata verso l'alto $\Rightarrow$ a modalità crescenti della X corrispondono più frequentemente modalità crescenti della Y.

– Si possono considerare convenzioni grafiche per punti ripetuti.



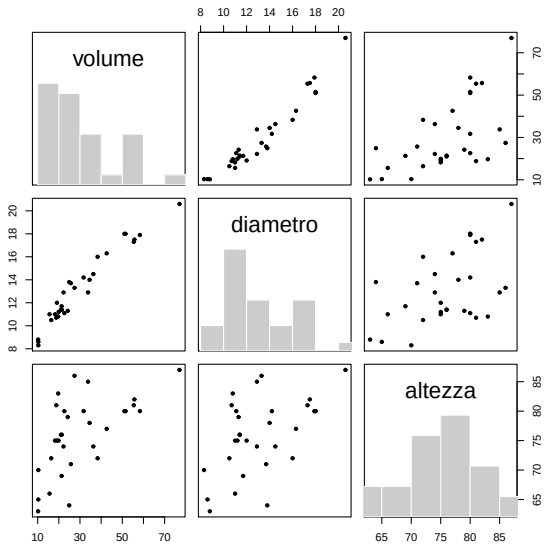
Esempio: Ciliegi neri

diametro tronco (in *pollici*)	altezza (in *piedi*)	volume legno (in $piedi^3$)
8.3	70	10.3
8.6	65	10.3
8.8	63	10.2
10.5	72	16.4
10.7	81	18.8
10.8	83	19.7
11.0	66	15.6
11.0	75	18.2
11.1	80	22.6
11.2	75	19.9
11.3	79	24.2
11.4	76	21.0
11.4	76	21.4
11.7	69	21.3
12.0	75	19.1
12.9	74	22.2
12.9	85	33.8
13.3	86	27.4
13.7	71	25.7
13.8	64	24.9
14.0	78	34.5
14.2	80	31.7
14.5	74	36.3
16.0	72	38.3
16.3	77	42.6
17.3	81	55.4
17.5	82	55.7
17.9	80	58.3
18.0	80	51.5
18.0	80	51.0
20.6	87	77.0

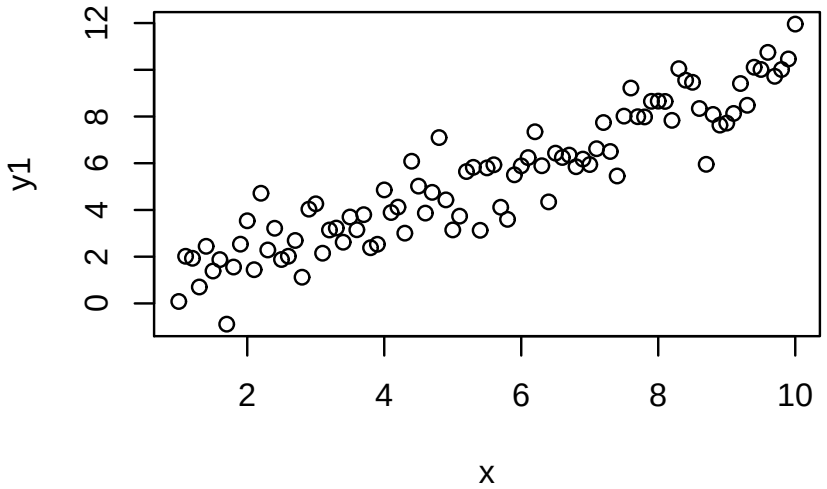
Il ruolo delle variabili X e Y è simmetrico?

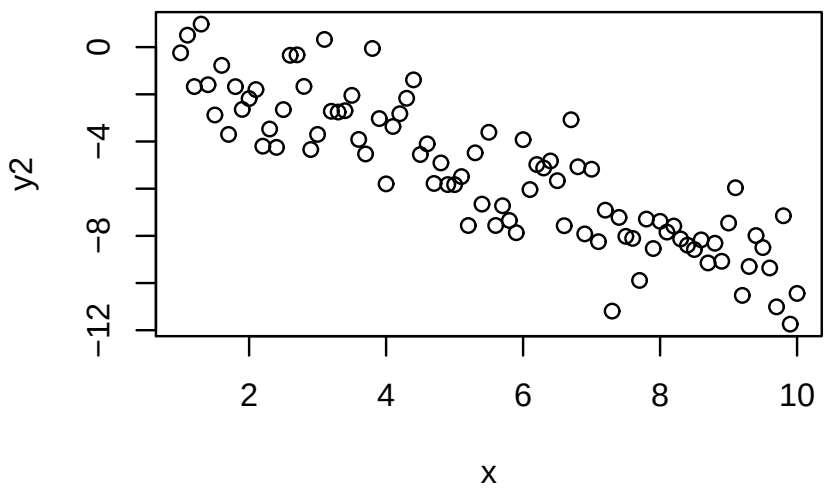
- A volte può essere importante spiegare una delle due variabili in funzione dell'altra. Si avrà quindi una VARIABILE ESPLICATIVA X e una VARIABILE RISPOSTA Y.
- Ma a volte non ha importanza quale sia l'una e quale sia l'altra.

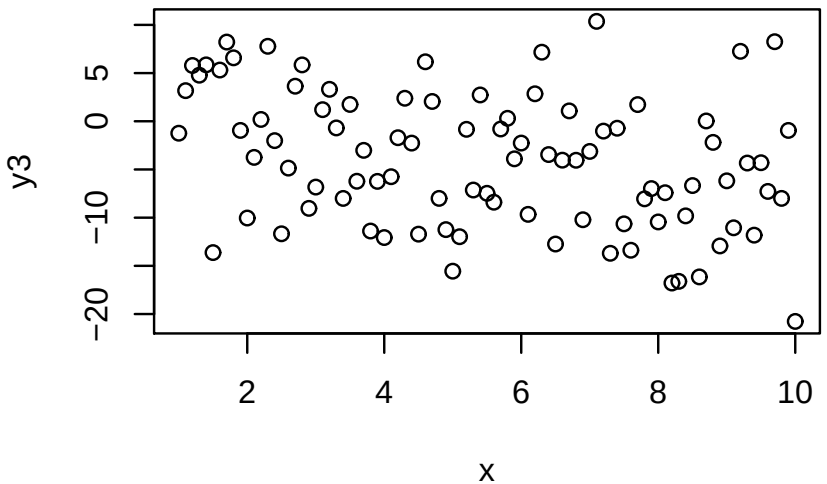
Nell'ESEMPIO dei ciliegi è ragionevole voler esprimere il volume del legno (Y), noto solo dopo che l'albero è stato abbattuto, a partire dal diametro (X), misurabile anche senza abbattere l'albero. Dal grafico di dispersione si vede che, in generale, negli alberi con diametro grande anche il volume del legno è elevato $\Rightarrow$ correlazione positiva.

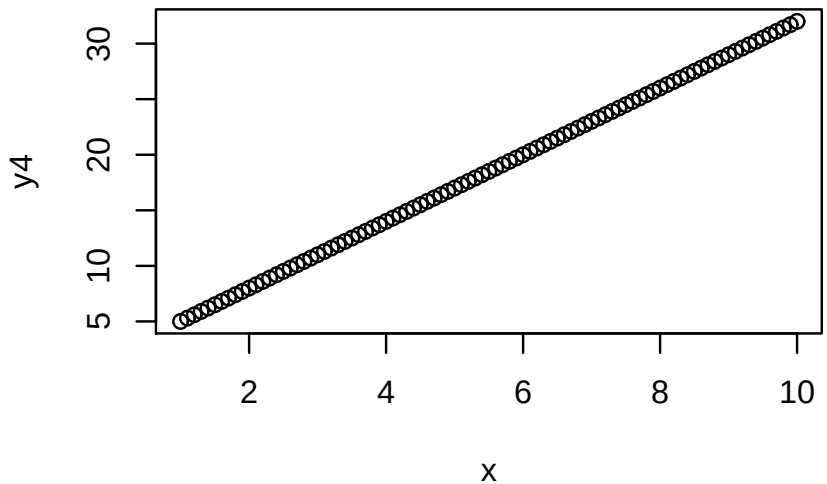


. . . qualche situazione tipo . . .



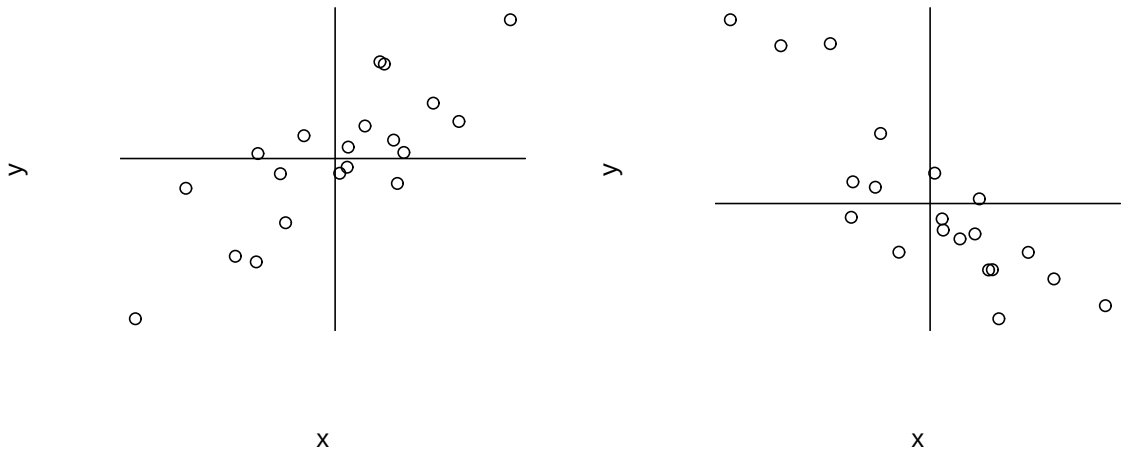






La covarianza

- Per avere una valutazione analitica del grado di associazione tra due variabili quantitative, esiste un indice che misura la dispersione nel piano dei punti dal proprio centro: la COVARIANZA.
- Il nome lascia intuire che si tratta di un'estensione al caso di due variabili della varianza. La covarianza si basa infatti sugli scarti delle x_i dalla propria media, $(x_i - m_x)$, e delle y_i dalla propria media, $(y_i - m_y)$.
- La covarianza, a differenza della varianza che è sempre positiva, misura l'eventuale direzione del legame, ovvero se le due variabili si muovono nella stessa direzione o in direzioni opposte. Il segno della covarianza riflette il senso crescente o decrescente dell'allineamento tendenziale.



La covarianza

- La covarianza segnala una concordanza (sia che X e Y decrescono o crescono) con un segno $+$ e una discordanza (quando X cresce e Y decresce, o viceversa) con il segno -. Formalmente, l'indice è

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y) \ .$$

- Una formula alternativa per il calcolo della covarianza è

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - m_x m_y$$

- Si noti che $S_{xx} = S_x^2$, ossia la covarianza tra X e X coincide con la varianza di X.

Campo di variazione della covarianza

La covarianza può assumere valori sia positivi sia negativi. In particolare, vale

$$-S_x S_y \leq S_{xy} \leq S_x S_y$$

Dimostrazione.
La varianza della combinazione $aX - bY$ (Appendice), per a e b costanti, è $a^2 S_x^2 + b^2 S_y^2 - 2ab S_{xy}$.

Si consideri ora la variabile T definita come $T = S_y^2 X - S_{xy} Y$. Allora, la variabile T ha varianza

$$\begin{aligned} S_T^2 &= S_y^4 S_x^2 + S_{xy}^2 S_y^2 - 2 S_y^2 S_{xy} S_{xy} \\ &= S_y^4 S_x^2 - S_{xy}^2 S_y^2 \end{aligned}$$

Ma poiché vale $S_T^2 \geq 0$, deve valere la diseguaglianza

$$S_y^4 S_x^2 - S_{xy}^2 S_y^2 \geq 0$$

ossia, dividendo per S_y^2,

$$S_{xy}^2 \leq S_y^2 S_x^2$$

da cui segue la tesi.

La correlazione

Il coefficiente di correlazione

- Dalla proprietà $-S_x S_y \leq S_{xy} \leq S_x S_y$, può essere costruito un indice relativo semplicemente dividendo S_{xy} per il prodotto degli scarti quadratici medi di X e Y. L'indice così ottenuto prende valori in [-1,1] e viene detto coefficiente di correlazione:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} \qquad -1 \leq r_{xy} \leq 1$$

- La formula del coefficiente di correlazione non è poi così terribile come appare!! Può solo essere noioso calcolarla a mano. In genere si usa un software opportuno.
- Un modo di procedere può essere il seguente:

  – Per le due variabili si calcolano le medie $m_x = \frac{1}{n}\sum x_i$ e $m_y = \frac{1}{n}\sum y_i$
  – Si calcola la media dei prodotti $\frac{1}{n}\sum x_i y_i$
  – Si calcolano le medie dei quadrati $\frac{1}{n}\sum x_i^2$ e $\frac{1}{n}\sum y_i^2$
  – Si calcola la covarianza $S_{xy} = \frac{1}{n}\sum x_i y_i - m_x m_y$
  – Si calcolano $S_x = [\frac{1}{n}\sum x_i^2 - m_x^2]^{1/2}$ e $S_y = [\frac{1}{n}\sum y_i^2 - m_y^2]^{1/2}$
  – Questi sono i numeri che servono per calcolare r_{xy}

- In sintesi: come si interpreta il valore trovato di r_{xy}?

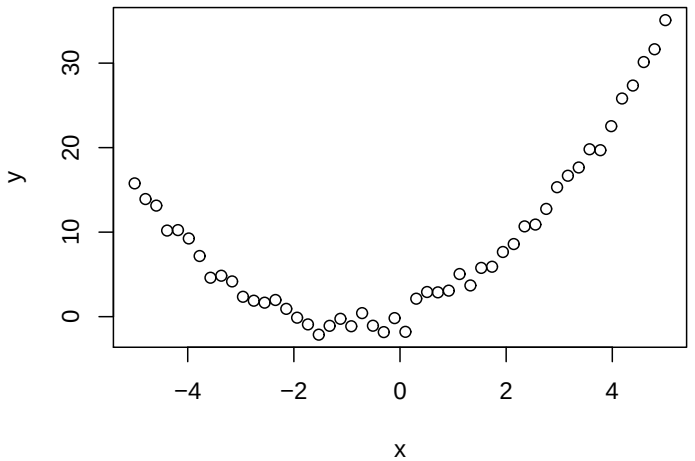
Guida all'interpretazione di r_{xy}

- $-1 \leq r_{xy} \leq 1$
- $r_{xy} = +1$: correlazione positiva perfetta (tutti i punti su una retta: concordi)
- $r_{xy} = -1$: correlazione negativa perfetta (tutti i punti su una retta: discordi)
- $r_{xy} > 0$: correlazione positiva
- $r_{xy} < 0$: correlazione negativa
- $r_{xy} \cong 0$: assenza di relazione lineare

Se $r_{xy} = \pm 1$ le variabili sono legate da una relazione lineare perfetta (diretta o inversa, rispettivamente). Si parla di relazione lineare in quanto r_{xy} misura se le coppie di valori (x_i, y_i) sono allineate lungo una retta del tipo $y = a + bx$.

Quando tra X e Y non vi è una relazione lineare o essa è estremamente debole, il valore dell'indice r_{xy} è zero o circa zero, e le variabili sono dette incorrelate.

ATTENZIONE: Il coefficiente di correlazione misura una associazione lineare. Il valore $r_{xy} = 0$ non indica tuttavia un'assenza di relazione tra le due variabili. Può esserci una relazione curvilinea.



Esempio: r_{xy} per i ciliegi

- Siano $Y =$ volume di legno (piedi^3) e $X =$ diametro del tronco (pollici).
- Si ha

  $m_x = 3.24$
  $m_y = 30.17$
  $\sum(x_i - m_x)^2 = 295.44$
  $\sum(y_i - m_y)^2 = 8106.08$
  $\sum(x_i - m_x)(y_i - m_y) = 1496.644$
- Allora:

$$r_{xy} = \frac{1496.644}{\sqrt{295.4 \times 8106.08}} = 0.967$$

- Il valore 0.967 indica una correlazione positiva molto forte tra il diametro del tronco e il volume del legno (come ci si aspettava dal grafico di dispersione).
- Con una relazione così forte, il volume del legno potrebbe essere previsto in modo accurato conoscendo il diametro del tronco.

La regressione

La regressione

- Quando dall'analisi di un diagramma di dispersione emerge un particolare andamento della nuvola di punti di X e Y, è naturale chiedersi se esiste una qualche relazione statistica $Y = f(X)+$errore tra X e Y.
- Il problema è lo stesso di prima: si vuole studiare una relazione tra le variabili. La relazione non è più simmetrica!! Perchè si vuole comprendere come la variabile risposta Y sia influenzata dalla variabile esplicativa X.
- Se la relazione che emerge è di tipo lineare, si può esprimere la relazione statistica tra X e Y usando un modello molto semplice: l'equazione della retta.

Il modello è del tipo:

$$Y = a + bX + \text{errore}$$

con

$a =$ intercetta
$b =$ coefficiente angolare
errore $=$ la deviazione dalla retta dei punti osservati

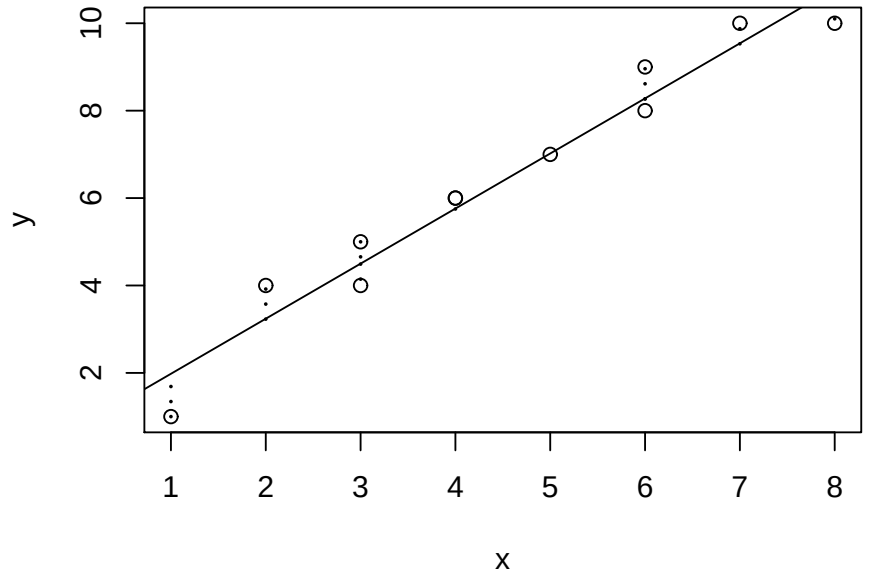
La regressione

- Se si calcolano "opportunamente" i valori di a e b, l'equazione può essere usata per prevedere il valore della Y a partire da un qualunque valore della X.
- PROBLEMA: come trovare la retta che si adatta nel modo migliore ai dati?
- Si devono determinare i valori di a e b che rendono la retta la più "vicina" possibile alle coppie osservate (x_i, y_i): la retta interpolante, cioè quella che passa tra i punti lasciando da essa scarti complessivamente minimi.
- I punti che stanno sulla retta sono le coppie di punti $(x_i, \hat{y}_i) = (x_i, a + bx_i)$, con $\hat{y}_i$ valori teorici o previsti, cioè i valori che la variabile Y dovrebbe assumere per $X = x_i$ se la relazione tra X e Y fosse esattamente quella ipotizzata $Y = a + bX$.
- r_{xy} misura quanto bene i dati sono allineati lungo tale retta. Come regola empirica, valori da 0.80/0.85 a 1 (o da -1 a -0.85/0.80) rivelano una accettabile relazione lineare di tipo diretto (o inverso). Ricordiamo che quando $r_{xy} = 0$ non è escluso che X e Y possono essere legate da altre relazioni, come $Y = \cos(X) + \exp(X^3)$, o altre "mostruosità" del genere.

Minimi quadrati

- Come cerchiamo la retta interpolante? Si noti che le quantità $e_i = y_i - \hat{y}_i$ misurano la distanza o scarto tra i valori di Y osservati e quelli teorici. In particolare, prendiamo la distanza quadratica, data da $(y_i - \hat{y}_i)^2$. Ne consegue che la distanza totale tra i valori osservati e teorici è

$$d(a,b) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \ .$$



La retta dei minimi quadrati

- La somma dei quadrati $d(a,b) = \sum_{i=1}^{n}(y_i - a - bx_i)^2$ dipende dalle incognite a e b, mentre y_i e x_i sono numeri osservati.
- La retta interpolante è quella i cui valori di a e di b che rendono minima $d(a,b)$, che viene detta retta dei minimi quadrati.

Si mostra che i valori a e b che minimizzano $d(a,b)$ sono dati da

$$\hat{b} = \frac{S_{xy}}{S_x^2} \qquad \hat{a} = m_y - \hat{b}\, m_x$$

- I calcoli richiesti sono gli stessi che servono per determinare il coefficiente di correlazione . . . non serve molto lavoro in più.
- Sia r_{xy} sia $\hat{b}$ dipendono al numeratore dalla covarianza S_{xy}. Essendo le quantità al denominatore sempre positive, è evidente che i segni di r_{xy} e di $\hat{b}$ sono concordi con il segno di S_{xy}.

Dimostrazione

Posto $y_i^* = y_i - bx_i$, $i = 1, \ldots, n$, la somma dei quadrati $d(a, b)$ può essere riscritta come $\sum_{i=1}^n (y_i^* - a)^2$. Quindi, per la proprietà dei minimi quadrati della media aritmetica, la quantità $\sum_{i=1}^n (y_i^* - a)^2$ è minima per

$$\hat{a} = \frac{1}{n}\sum_{i=1}^n y_i^* = \frac{1}{n}\sum_{i=1}^n (y_i - bx_i) = \frac{1}{n}\sum_{i=1}^n y_i - b\,\frac{1}{n}\sum_{i=1}^n x_i = m_y - b\,m_x \ .$$

Sostituendo tale valore in $d(a, b)$ si ottiene

$$\begin{aligned}
\sum_{i=1}^n (y_i - m_y - bx_i + bm_x)^2 &= \sum_{i=1}^n \left[(y_i - m_y) - b(x_i - m_x)\right]^2 \\
&= \sum_{i=1}^n (y_i - m_y)^2 + b^2 \sum_{i=1}^n (x_i - m_x)^2 - 2b \sum_{i=1}^n (y_i - m_y)(x_i - m_y) \\
&= nb^2 S_x^2 - 2nbS_{xy} + nS_y^2
\end{aligned}$$

Come funzione di b, si tratta di una funzione quadratica, il cui grafico è una parabola con concavità rivolta verso l'alto. Il minimo si ha in corrispondenza del vertice, ossia per

$$\hat{b} = \frac{-(-2nS_{xy})}{2nS_x^2} = \frac{S_{xy}}{S_x^2}$$

Esempio: alberi di ciliegio

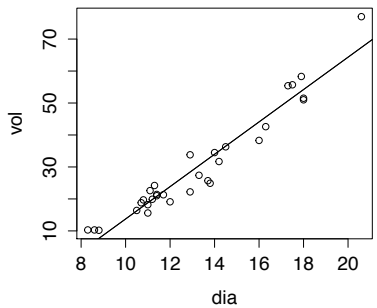
Nell'esempio del volume di legno (Y) e diametro del tronco (X) si trovano i seguenti valori di $\hat{a}$ e $\hat{b}$:

$\hat{b} = 1496.644/295.44 = 5.06$piedi^3/pollici $\qquad$ $\hat{a} = 30.17 + 5.06 \times 13.25 = -36.87$piedi^3

La retta di regressione per questi dati è:

$$\hat{Y} = -36.87 + 5.06\, X = -36.87 + 5.06\, \text{diametro}$$

Abbiamo il risultato: ma come interpretarlo e usarlo?? La retta è UTILE per fare previsioni sulla variabile risposta. Ad esempio per $X = 15$, si trova $Y = -36.87 + 5.06 \times 15 = 39.03$. Per $X = 5$ si ha $Y = -36.87 + 5.06 \times 5 = -11.57$. Ma ATTENZIONE perchè $X = 5$ è "poco realistico".



Bontà dell'adattamento della retta ai dati

- Come possiamo valutare se la retta si adatta bene ai dati? Abbiamo bisogno di un indice capace di riassumere l'adattamento globale e la capacità esplicativa complessiva del modello in rapporto ai dati osservati.
- Si può utilizzare ancora il coefficiente di correlazione r_{xy}. E poiché non ha importanza se la correlazione è positiva o negativa, si eleva r_{xy} al quadrato $\Rightarrow$ COEFFICIENTE DI DETERMINAZIONE:

$$R^2 = r_{xy}^2$$

NOTA:

Se $R^2 = 1$: adattamento perfetto (tutti i punti sulla retta)
Se $R^2 = 0$: la retta non ha nulla da vedere con i dati
Se $R^2 = 0.8$: "buon livello" di adattamento

- ESEMPIO: $r_{xy} = 0.967 \Rightarrow R^2 = 0.935$, ossia la retta di regressione si adatta molto bene ai dati.

Interpretazione di R^2 come proporzione di varianza spiegata

- Siano $\hat{y}_i = \hat{a} + \hat{b}x_i$, $i = 1, \ldots, n$, i valori calcolati sulla retta dei minimi quadrati.
- La somma dei residui $y_i - \hat{y}_i$ vale zero.
  Infatti, $\sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n (y_i - \hat{a} - \hat{b}x_i) = \sum_{i=1}^n (y_i - m_y + \hat{b}m_x - \hat{b}x_i) =$
  $\sum_{i=1}^n (y_i - m_y) - \hat{b}\sum_{i=1}^n (x_i - m_x) = 0$ (proprietà di baricentro).
- Inoltre, $\sum_{i=1}^n (y_i - \hat{y}_i)x_i = \sum_{i=1}^n (y_i - \hat{y}_i)(x_i - m_x) =$
  $\sum_{i=1}^n (y_i - m_y + \hat{b}m_x - \hat{b}x_i)(x_i - m_x) = nS_{xy} - \hat{b}nS_x^2 = 0\,.$
- Allora, dall'identità $\sum_{i=1}^n (y_i - m_y)^2 = \sum_{i=1}^n (y_i \pm \hat{y}_i - m_y)^2 =$
  $\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - m_y)^2 + 2\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - m_y)$, usando le due relazioni precedenti, si vede facilmente che l'ultima sommatoria vale zero.
  Dunque $\frac{1}{n}\sum_{i=1}^n (y_i - m_y)^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \frac{1}{n}\sum_{i=1}^n (\hat{y}_i - m_y)^2$
  ossia

  VARIANZA TOTALE=VARIANZA RESIDUA+VARIANZA SPIEGATA
- Si vede infine che R^2 =VARIANZA SPIEGATA/VARIANZA TOTALE.

  Infatti, $\sum_{i=1}^n (\hat{y}_i - m_y)^2 = \sum_{i=1}^n (m_y - \hat{b}m_x + \hat{b}x_i - m_y)^2 = n\hat{b}^2 S_x^2 = nS_{xy}^2/S_x^2$. E quindi

$$\frac{\sum_{i=1}^n (\hat{y}_i - m_y)^2}{\sum_{i=1}^n (y_i - m_y)^2} = \frac{nS_{xy}^2}{S_x^2 nS_y^2} = R^2\,.$$

Esempio: Tensione, corrente e resistenza

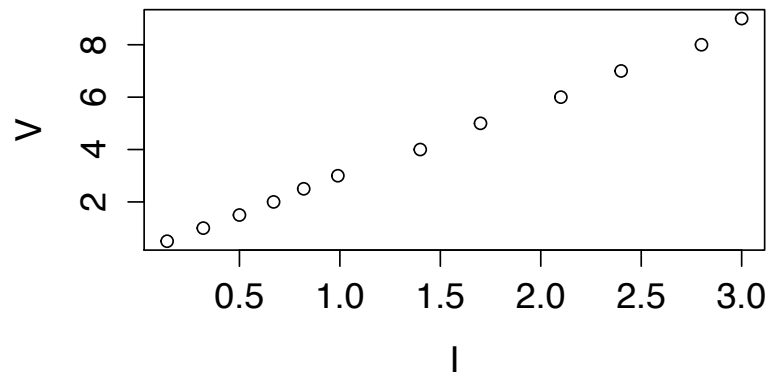
I seguenti dati riportano $n = 12$ misurazioni della tensione (`V`) e della corrente (`I`):

`V = (0.5,1,1.5,2,2.5,3,4,5,6,7,8,9)` in *volt*
`I = (0.14,0.32,0.50,0.67,0.82,0.99,1.4,1.7,2.1,2.4,2.8,3)` in *ampere*

La relazione lineare tra le due variabili è esprimibile come

$$V = a + bI + \text{errore}$$

e ci si attende dal modello teorico $a \doteq 0$ *volt*, $b = Res$ *volt/ampere*, dove Res è una costante di proporzionalità che misura la resistenza, e un valore di R^2 estremamente elevato.



Posto $X = I$ e $Y = V$, si ha:

$m_x = 1.403$ e $m_y = 4.125$

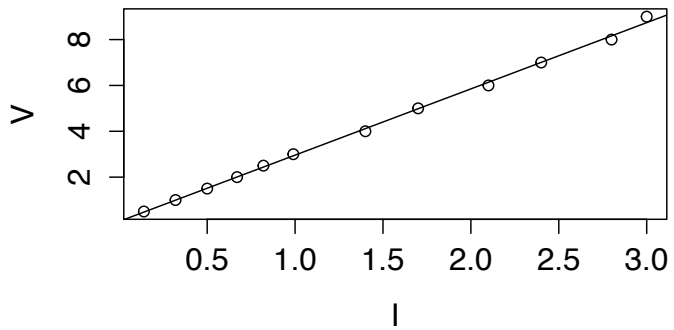
$S_x^2 = 0.892$, $S_y^2 = 7.463$ e $S_{xy} = 2.578$

$\rightarrow \hat{b} = 2.578/0.892 = 2.89$ *volt/ampere* e $\hat{a} = 4.125 - 2.89 \times 1.403 = 0.07$ *volt*.

La retta di regressione per questi dati è:

$$\hat{Y} = 0.07 + 2.89\, X$$

Con correlazione $r_{xy} = 0.999$ $(R^2 = 0.9985)$, tale modello evidenza una relazione lineare tra le due variabili. Inoltre, $a \doteq 0$ *volt* come ci si aspettava dal modello teorico, mentre $Res = 2.89$ *volt/ampere*.



Esempio: Intensità luminosa e inverso del quadrato della distanza

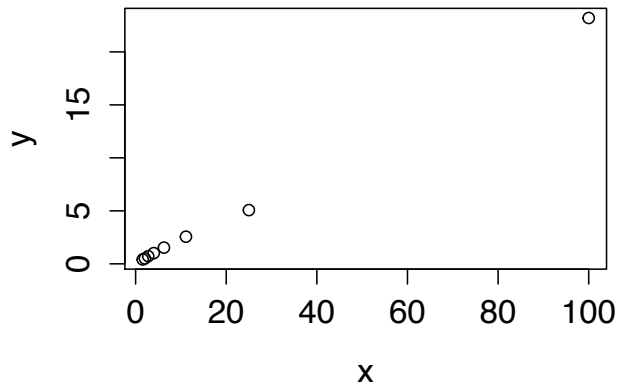
I seguenti dati riportano $n = 8$ misurazioni dell'intensità luminosa della luce di una lampadina (Y) raccolta da un sensore a distanza d e la grandezza $X = 1/d^2$:

```
x = (100,25,11.11,6.25,4,2.778,2.041,1.563)
y = (23.2,5.07,2.56,1.53,1.01,0.72,0.51,0.41)
```

La relazione lineare tra le due variabili è esprimibile come:

$$Y = a + bX + \text{errore}$$

e ci si attende dal modello teorico $a \doteq 0$, $b = k$, dove k è una costante di proporzionalità tale che $Y = kX$, e un valore di R^2 estremamente elevato.



Si ha:

$m_x = 19.09$ e $m_y = 4.38$

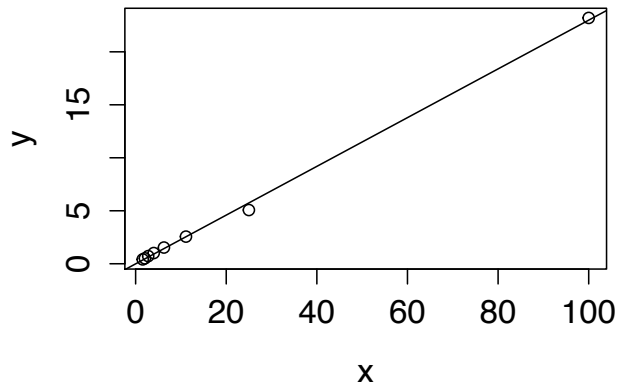
$S_x^2 = 987.69$, $S_y^2 = 52.69$ e $S_{xy} = 228.004$

$\rightarrow \hat{b} = 228.004/987.69 = 0.23$ e $\hat{a} = 4.38 - 0.23 \times 19.09 = -0.01$.

La retta di regressione per questi dati è:

$$\hat{Y} = -0.01 + 0.23\, X$$

Con correlazione $r_{xy} = 0.999$ ($R^2 = 0.9988$), tale modello evidenza una relazione lineare tra le due variabili. Inoltre, $a \doteq 0$ come ci si aspettava dal modello teorico, mentre $k = 0.23$.



Esempio: peso alla nascita e durata della gravidanza

Peso alla nascita (in grammi) per $n = 32$ neonati, durata della gravidanza (in settimane), madre fumatrice (S/N).

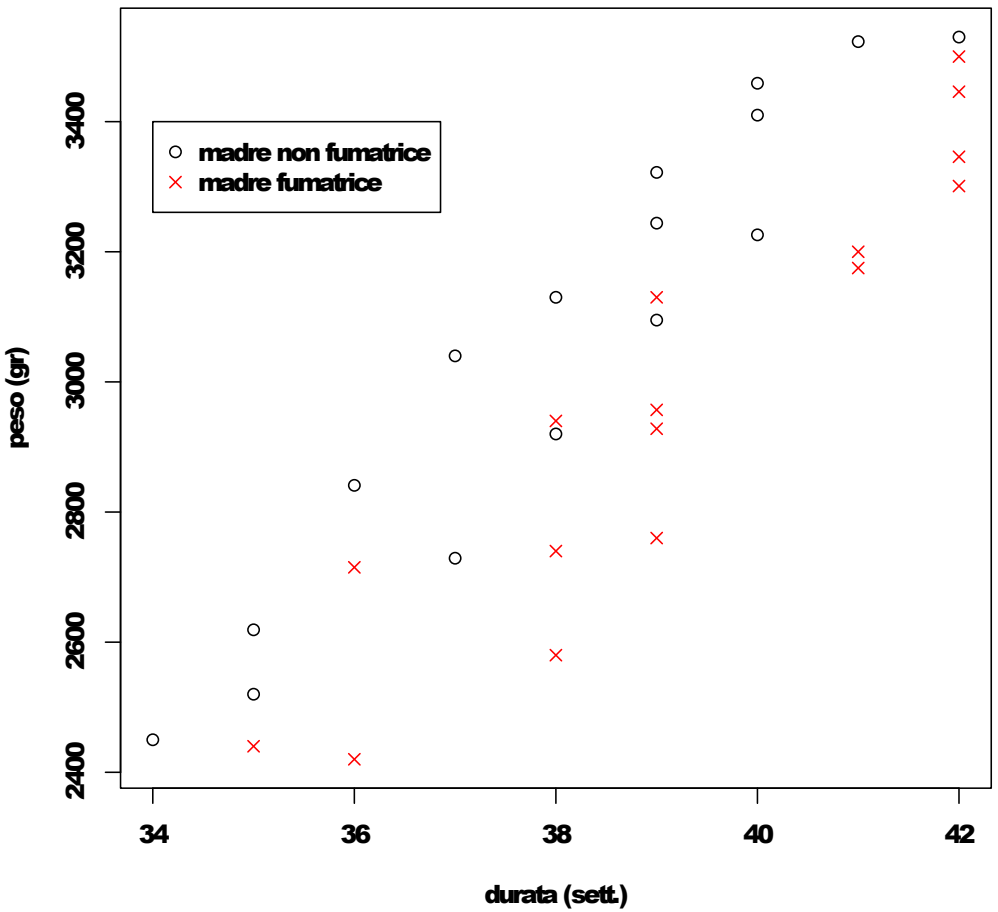
Madre fumatrice

Peso	2940	2420	2760	2440	3301	2715	3130	2928
	3446	2957	2580	3500	3200	3346	3175	2740
Durata	38	36	39	35	42	36	39	39
	42	39	38	42	41	42	41	38

Madre non fumatrice

Peso	3130	2450	3226	2729	3410	3095	3244	2520
	3523	2920	3530	3040	3322	3459	2619	2841
Durata	38	34	40	37	40	39	39	35
	41	38	42	37	39	40	35	36

Esempio: peso alla nascita e durata della gravidanza



- Posto Y_{F} = peso da madre fumatrice e X_{F} = durata per madre fumatrice, si ha:
  $m_{x_F} = 39.1875$ e $m_{y_F} = 2973.625$
  $S^2_{x_F} = 5.027$, $S^2_{y_F} = 111192.9$ e $S_{xy_F} = 698.9453$
  La retta di regressione per questi dati è:

$$\hat{Y}_{F} = -2474.6 + 139.0\, X_{F}$$

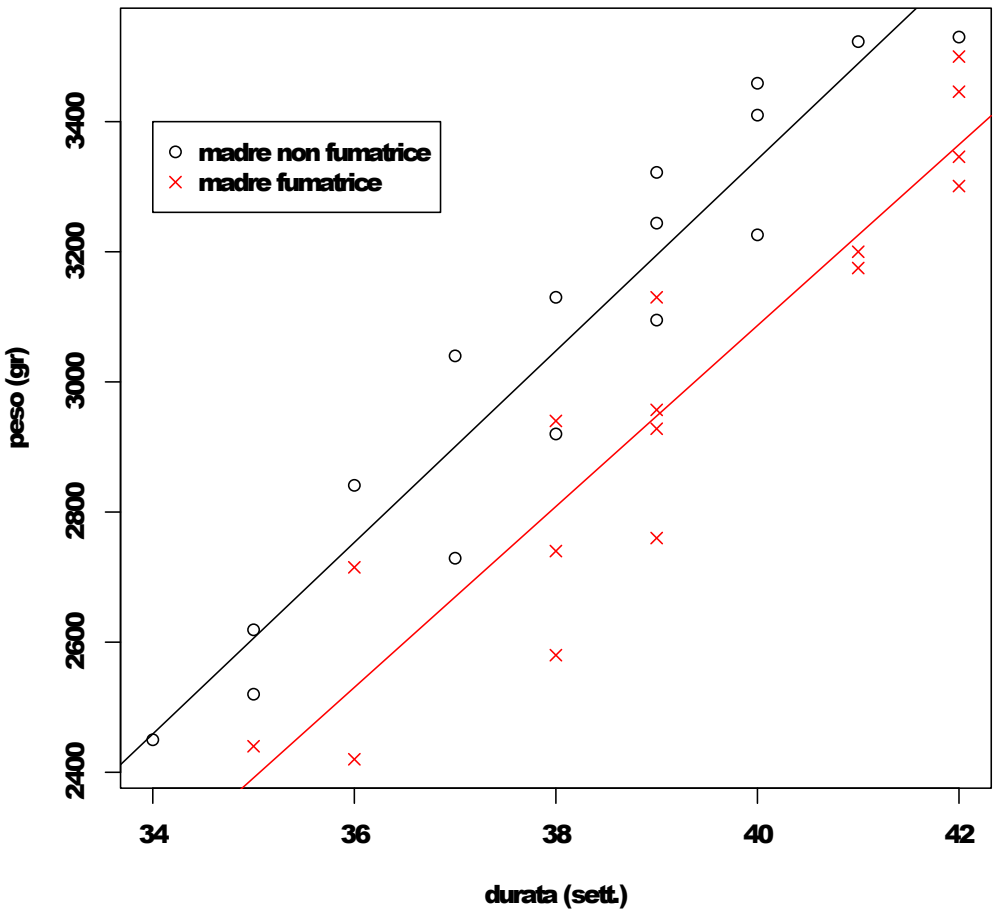
  La correlazione è $r_{xy_F} \doteq 0.93$.

- Posto Y_{NF} = peso da madre non fumatrice e X_{NF} = durata per madre non fumatrice, si ha:
  $m_{x_{NF}} = 38.125$ e $m_{y_{NF}} = 3066.125$
  $S^2_{x_{NF}} = 4.9844$, $S^2_{y_{NF}} = 118015.9$ e $S_{xy_{NF}} = 733.73$
  La retta di regressione per questi dati è:

$$\hat{Y}_{NF} = -2546.1 + 147.2\, X_{NF}$$

  La correlazione è $r_{xy_{NF}} \doteq 0.96$.

Esempio: peso alla nascita e durata della gravidanza



Appendice: proprietà della media e della varianza

Media

- Linearità: $m_{a+bx} = a + bm_x$, con $a, b \in \mathbb{R}$
- Combinazione lineare: $m_{ax+by} = am_x + bm_y$, con $a, b \in \mathbb{R}$

Varianza

- Invarianza rispetto a traslazioni: $S^2_{a+x} = S^2_x$, con $a \in \mathbb{R}$
- Omogeneità (di secondo grado): $S^2_{bx} = b^2 S^2_x$, con $b \in \mathbb{R}$

  $\rightarrow S^2_{a+bx} = b^2 S^2_x$, con $a, b \in \mathbb{R}$
- Combinazione lineare: $S^2_{ax+by} = a^2 S^2_x + b^2 m^2_y + 2abS_{xy}$, con $a, b \in \mathbb{R}$ e $S^2_{ax-by} = a^2 S^2_x + b^2 m^2_y - 2abS_{xy}$, con $a, b \in \mathbb{R}$

Esercizi

(1) La gascromotografia è una tecnica per analizzare miscele di gas. I dati che seguono mostrano la quantità di una certa sostanza (Y) e la corrispondente misura ottenuta da un gascromatografo (X):

quantità	0.25	0.25	0.25	1	1	1	5	5	5	20	20	20
misura	6.55	7.98	6.54	29.7	30	30.1	211	204	212	929	905	922

1) Disegnare il diagramma di dispersione dei dati
2) Calcolare la quantità media di sostanza
3) Calcolare la retta di regressione che permette di prevedere la quantità di sostanza come funzione della misura ottenuta dal gascromatografo
4) Calcolare l'indice di correlazione
5) Per una quantità di sostanza pari a 2, il gascromatografo ha fornito una misura pari a?

(2) La seguente tabella mostra per vari anni il numero di incidenti stradali in una certa regione:

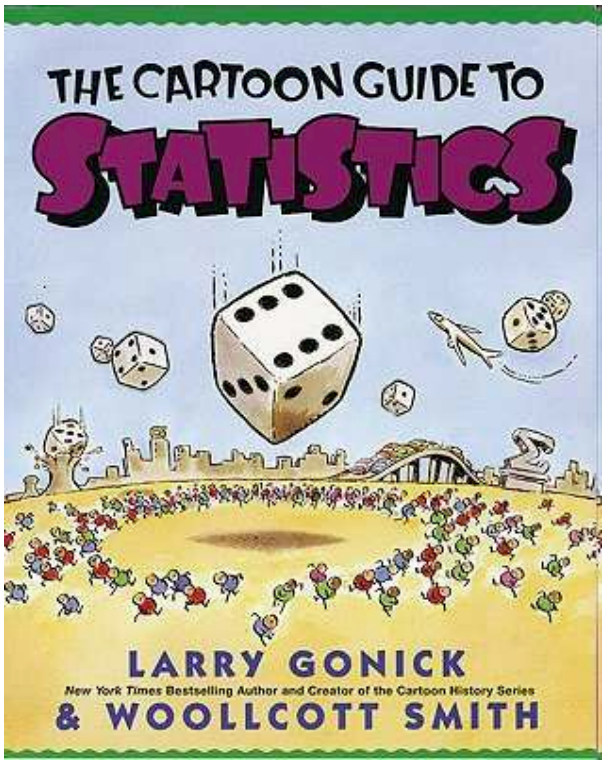
Anno	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Incidenti	5413	6122	6705	6824	7790	7698	8571	8688	9422	9904

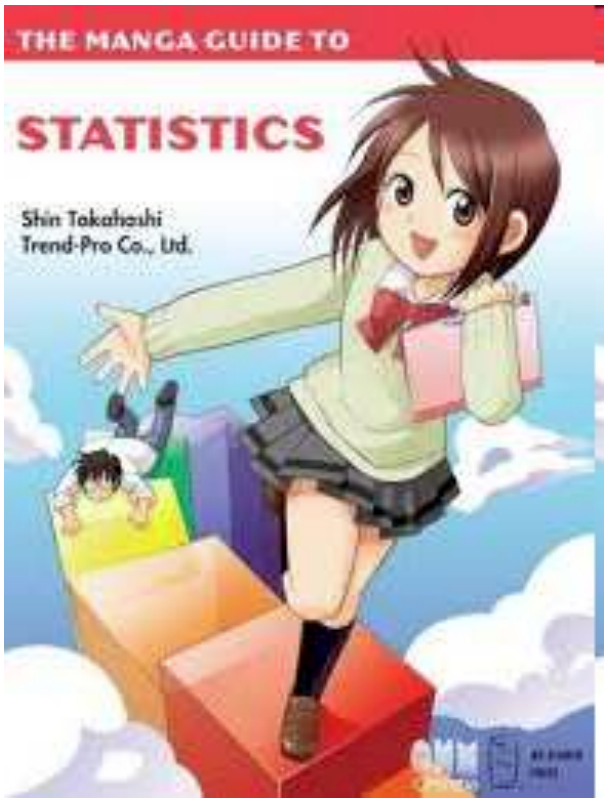
1) Si calcoli il numero medio di incidenti in un anno.
2) Si fornire una rappresentazione grafica dei dati opportuna.
3) Si calcoli la retta di regressione che permette di prevedere il numero di incidenti come funzione dell'anno.
4) Si calcoli il coefficiente di correlazione.
5) Si fornisca una previsione per il numero di incidenti per il 2001.

Alcuni riferimenti bibliografici

- Agresti, A., Finlay, B. (2009). *Statistica per le scienze sociali*. Pearson.
- Agresti, A., Franklin, C. (2013). *Statistics. The Art and Science of Learning from Data.* Pearson.
- Bernstein, S., Bernstein, R. (2003). *Statistica Descrittiva*. McGraw-Hill.
- Bradstreet, T.E. (1996). Teaching introductory statistics courses so nonstatisticians experience statistical reasoning. *The American Statistician*, Vol. 50, 69 – 78.
- Diamond, I., Jefferies, J. (2001). *Introduzione alla statistica per le scienze sociali*. McGraw-Hill.
- Pace, L., Salvan, A. (1996). *Introduzione alla Statistica. I Statistica Descrittiva*. Cedam.
- Rosenthal, J.S. (2005). *Le Regole del Caso: Istruzioni per l'Uso*. Longanesi.

Oppure...


THE CARTOON GUIDE TO
STATISTICS
LARRY GONICK
New York Times Bestselling Author and Creator of the Cartoon History Series
& WOOLLCOTT SMITH


THE MANGA GUIDE TO
STATISTICS
Shin Takahashi
Trend-Pro Co., Ltd.