§20 Regressione lineare

Date due variabili statistiche A e B, di tipo numerico, vogliamo quantificare in che misura i valori assunti da una siano correlati a quelli assunti dall'altra.

Esempio
È stato chiesto a 59 studenti di ripetere due volte una prova di abilità, il cui risultato è un punteggio fra 0 e 15. I punteggi delle due prove, A e B, sono riassunti in questo diagramma.

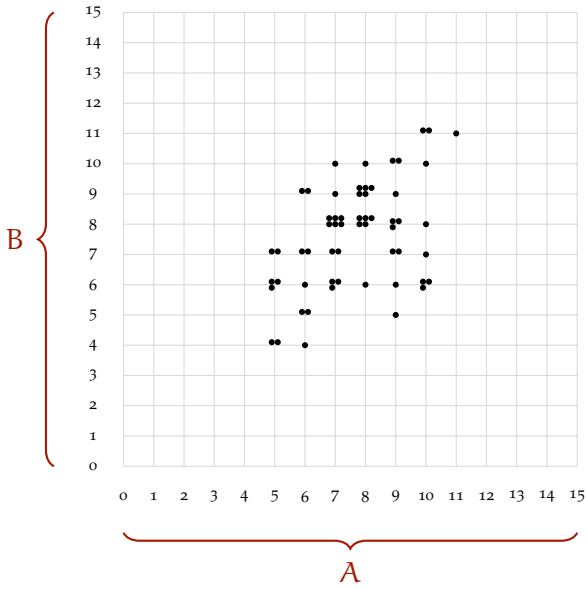
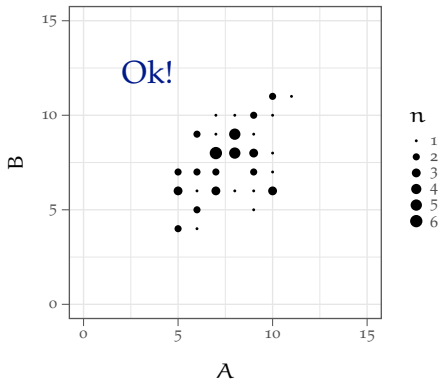


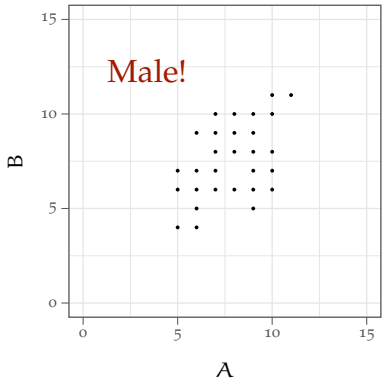
Diagramma di dispersione (in inglese *scatter plot*).

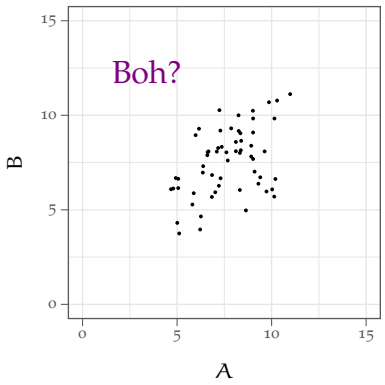
Ciascun punto rappresenta un individuo, con la sua coppia di valori (a, b).



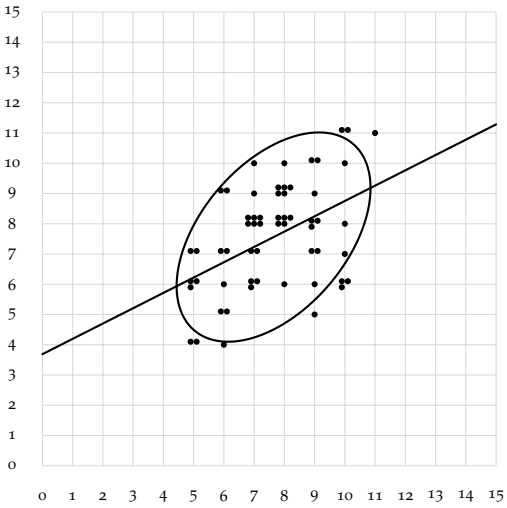
Un'altra forma di diagramma di dispersione: qui il numero di individui è rappresentato dalla dimensione dei cerchi.

Rappresentando ogni coppia (a, b) con un punto si avrebbero delle sovrapposizioni che oscurano l'accumulo dei dati al centro della nuvola.





In questa versione i punti sono stati mossi di poco, in modo casuale, per rimuovere le sovrapposizioni.

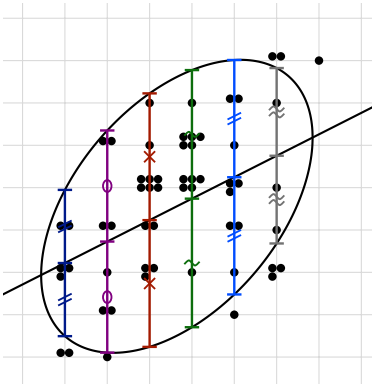


I punti — le coppie (a, b) — formano una nuvola grossolanamente ellittica. Questo avviene perché la distribuzione delle variabili A e B è approssimativamente congiuntamente gaussiana. L'ellisse è inclinata: questo indica una correlazione fra A e B e, in particolare, una correlazione positiva, perché, al crescere di A, anche B tende a crescere; la situazione opposta si chiamerebbe correlazione negativa.

Il diagramma mostra altresì la retta di regressione di B rispetto ad A. Questa retta divide a metà ogni sezione verticale della nuvola, vale a dire, per ogni fissato valore $A = a_0$, il corrispondente

$$b_0 = \text{pendenza} \cdot a_0 + \text{intercetta}$$

è una stima della media delle coordinate B dei punti che hanno $A = a_0$.

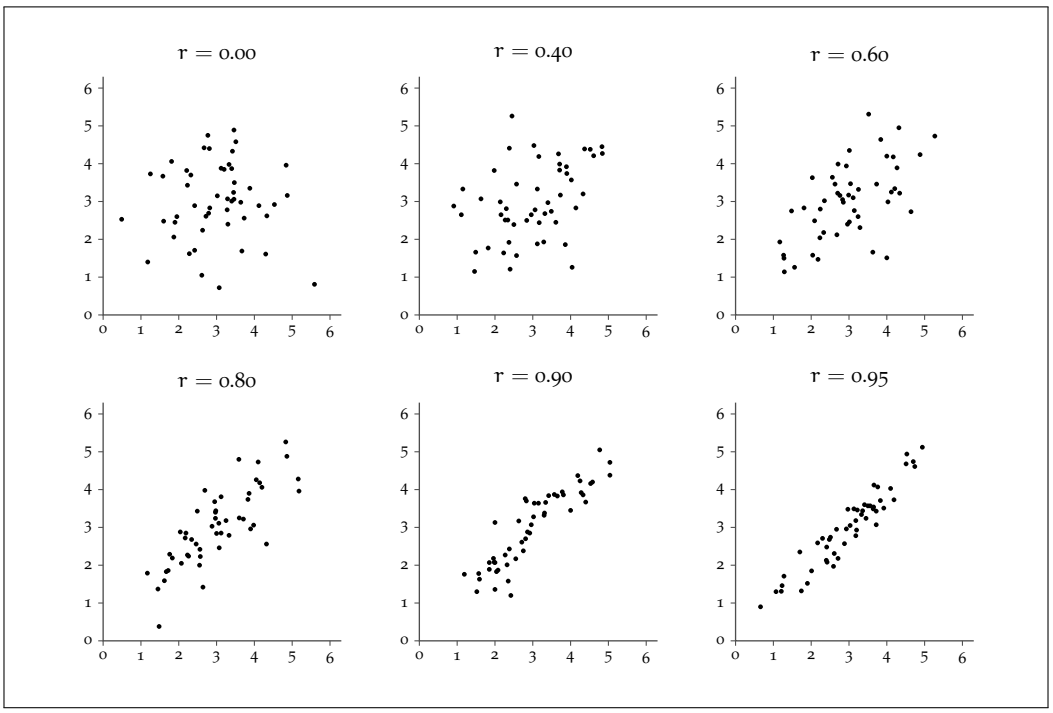


Vedremo come calcolare le seguenti quantità:

- i coefficienti della retta di regressione;
- il coefficiente di correlazione di Pearson r, che misura la forza della correlazione;
- un p-value, che dice se i dati bastano per asserire che le variabili sono realmente correlate.

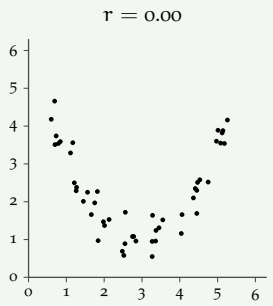
Abbiamo già descritto il significato della retta di regressione. Per quanto riguarda r, si tratta di un numero nell'intervallo $-1 \leqslant r \leqslant 1$, il cui segno indica il segno della correlazione. Il valore assoluto $|r|$ dice quanto la correlazione è forte, ossia in che misura è possibile prevedere B conoscendo A per mezzo della retta di regressione. A $|r| = 0$ corrisponde l'assenza completa di correlazione lineare (in questo caso la retta di regressione è orizzontale: una costante pari alla media di B). A $|r| = 1$ corrisponde il caso in cui B è esattamente determinata da $B = mA + q$ per opportuni coefficienti m (la pendenza) e q (l'intercetta).

Grossolanamente, r descrive la forma della nuvola di punti. Se $r = 0$ gli assi dell'ellisse sono paralleli agli assi coordinati, per cui, a scegliere le giuste unità di misura, la nuvola appare circolare. Se $r > 0$ (risp. $r < 0$) l'asse maggiore è disposto diagonalmente così: ／ (risp. così: ＼). Più $|r|$ si avvicina a 1, più l'ellisse è schiacciata, e, per $|r| = 1$, questa ha la forma di un segmento di retta.



Osservazione

La regressione lineare cerca una dipendenza del tipo $B = mA + q$. Questo metodo è insensibile — anzi confuso — da leggi di dipendenza più complicate. Per esempio le variabili del diagramma qui a fianco risultano linearmente scorrelate ($r = 0$), anche se c'è una correlazione evidente.



Discuteremo in dettaglio il p-value più avanti. Molto sinteticamente, l'idea è che, in un campione casuale, potremmo osservare una correlazione che, in realtà, nella popolazione non c'è, per via di uno scherzo della fortuna. Il p-value misura quanto è verosimile che la correlazione apparentemente presente nei dati sia un mero frutto del caso. Quindi minore è il p-value, più è verosimile che la correlazione sia reale, ossia maggiore è la significatività statistica della correlazione. Un p-value grande, ossia una scarsa significatività statistica, indica che i dati non bastano per concludere che A e B sono linearmente correlate (perché il campione — il numero di punti (a, b) rilevati — è troppo piccolo, o perché la correlazione non c'è). Soglie tipiche con cui si confrontano i p-value (si chiamano livelli di significatività) nelle scienze della vita sono < 0.05 per ritenere un risultato significativo e < 0.01 perché sia fortemente significativo. Una triste macumba circonda queste soglie, per cui è bene non farne un feticcio: si tratta di semplici numeri scelti per convenzione.

§20.1 Formule e calcoli

Chiamiamo $(a_1, b_1), (a_2, b_2), \ldots, (a_n, b_n)$ i punti del nostro campione. Si comincia col calcolare media e deviazione standard di A e di B (sul campione, con n al denominatore): μ_A, σ_A, μ_B, σ_B. Poi si convertono i punti del campione in coordinate standard (le stesse del capitolo 18, dove le abbiamo chiamate z; qui le variabili sono due, e le distinguiamo con un asterisco)

$$\boxed{A^* = \frac{A - \mu_A}{\sigma_A} \qquad B^* = \frac{B - \mu_B}{\sigma_B}}$$

ottenendo $(a_1^*, b_1^*), (a_2^*, b_2^*), \ldots, (a_n^*, b_n^*)$.

Ora, il coefficiente di correlazione di Pearson r è la media dei prodotti $a_i^* b_i^*$:

$$\boxed{r = \frac{a_1^* b_1^* + a_2^* b_2^* + \ldots + a_n^* b_n^*}{n}}$$

E la retta di regressione, in coordinate standard, è semplicemente

$$B^* = r\, A^*$$

Da cui, sostituendo A^* e B^* con le rispettive definizioni,

$$\frac{B - \mu_B}{\sigma_B} = r\,\frac{A - \mu_A}{\sigma_A} \qquad \Longrightarrow \qquad B = \overbrace{r\,\frac{\sigma_B}{\sigma_A}}^{\text{pendenza}} A + \overbrace{\mu_B - r\,\frac{\sigma_B}{\sigma_A}\,\mu_A}^{\text{intercetta}}$$

Non conviene memorizzare quest'ultima formula, ma la molto più semplice

Regola (retta di regressione in coordinate standard)

$$B^* = r\,A^*$$

da cui è facile ricavare la retta di regressione che lega A e B come appena visto.

Esercizio
Calcola il coefficiente di correlazione r e la retta di regressione di B rispetto ad A per i dati seguenti.

a_i	4	5	1	9	4	7	5	4	5	6
b_i	5	6	8	2	4	1	5	3	2	4

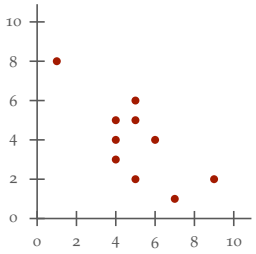
Soluzione.

Otteniamo $\mu_A = 5.00$ $\sigma_A = 2.00$

$\mu_B = 4.00$ $\sigma_B = 2.00$



I dati sono finti...



Da cui

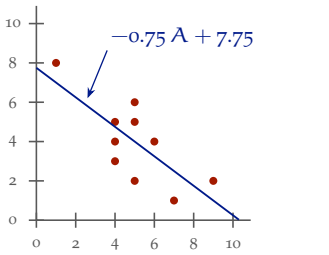
a_i^*	-0.5	0.0	-2.0	2.0	-0.5	1.0	0.0	-0.5	0.0	0.5
b_i^*	0.5	1.0	2.0	-1.0	0.0	-1.5	0.5	-0.5	-1.0	0.0
$a_i^* b_i^*$	-0.25	0	-4	-2	0	-1.5	0	0.25	0	0

$$r = \text{media di questi} = \frac{-7.5}{10} = -0.75$$

Per cui la retta di regressione è $B^* = -0.75\,A^*$, che possiamo scrivere nella forma

$$\frac{B-4}{2} = -0.75\,\frac{A-5}{2}$$

$$B = -0.75\,A + 7.75$$



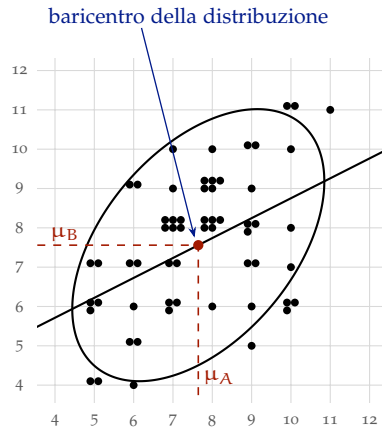
□

Esempio

I risultati che abbiamo discusso di una prova di abilità hanno

$$\mu_A = 7.644 \qquad \sigma_A = 1.603$$
$$\mu_B = 7.559 \qquad \sigma_B = 1.730 \qquad r = 0.469$$

quindi la retta di regressione è



$$\frac{B - 7.559}{1.730} = 0.469\,\frac{A - 7.644}{1.603} \qquad \Longrightarrow \qquad B = 0.506\,A + 3.69$$

Non stupisce che ci sia correlazione fra due ripetizioni della medesima prova: questo avviene, in generale, con le prove di abilità.

Osservazione
La retta di regressione passa sempre per il baricentro (μ_A, μ_B) dei punti del campione, quindi la si può disegnare tracciando la retta per questo punto che ha pendenza $r\,\frac{\sigma_B}{\sigma_A}$.

Nota (regressione verso la media)
Nell'esempio, la retta di regressione in coordinate standard è $B^* = 0.469\,A^*$: siccome $|r| < 1$, il valore previsto per B è, in coordinate standard, più vicino alla media di quanto lo sia A. I dati lo confermano: gli 8 studenti che hanno ottenuto 10 alla prima prova hanno avuto, alla seconda, 8.1 in media; i 7 che hanno ottenuto 5 hanno avuto 5.7 in media. Chi è andato molto bene tende a peggiorare un po', chi è andato male a migliorare. Questo fenomeno si chiama regressione verso la media — è da qui che viene il nome "retta di regressione".

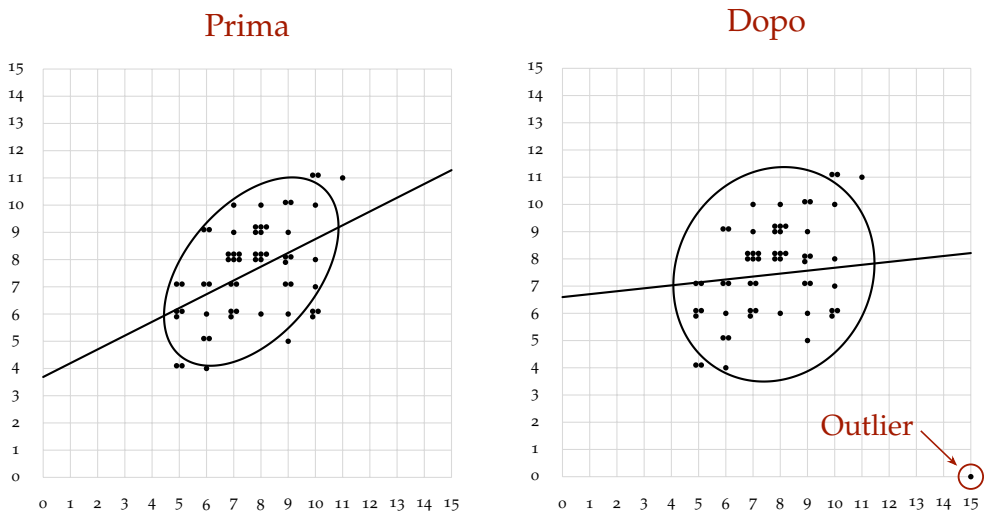
Non serve cercare spiegazioni: i primi non si sono montati la testa, gli ultimi non si sono impegnati di più, e, contrariamente al motto evangelico, regrediscono verso la *media*, non verso il massimo della distribuzione. Questo fenomeno è una conseguenza meccanica del fatto che la correlazione non è perfetta. Il punteggio dipende in parte dall'abilità e in parte dalla fortuna: fra chi ha ottenuto un punteggio molto alto ci sono più fortunati che sfortunati, e la fortuna non si ripete. Infatti il fenomeno si presenta anche a ritroso: chi è andato molto bene alla seconda prova, alla prima era andato, in media, un po' peggio.

Attenzione se si valuta una terapia. Di solito un paziente comincia a curarsi quando sta peggio: alla visita successiva starà, in media, meglio, qualunque cosa abbia preso, per cui, senza un gruppo di controllo, è facile scambiare questo

miglioramento per un effetto del trattamento. È una delle ragioni per cui servono gli studi controllati.

§20.2 Regressione lineare e outlier

In statistica un outlier è un dato che pare anomalo se confrontato con la distribuzione degli altri. Questa definizione è imprecisa perché non c'è una definizione precisa. Possiamo introdurre artificialmente un outlier fra i dati della nostra prova di abilità aggiungendo il singolo punto $(15, 0)$ ai 59 presenti.



$r = 0.469$ — p-value $= 0.00018$

$r = 0.101$ — p-value $= 0.44$

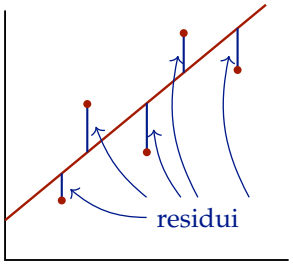
Come si vede, un singolo outlier confonde completamente il calcolo: il coefficiente di correlazione e la retta di regressione non sono misure statistiche robuste. Allo stesso modo la media non è robusta,

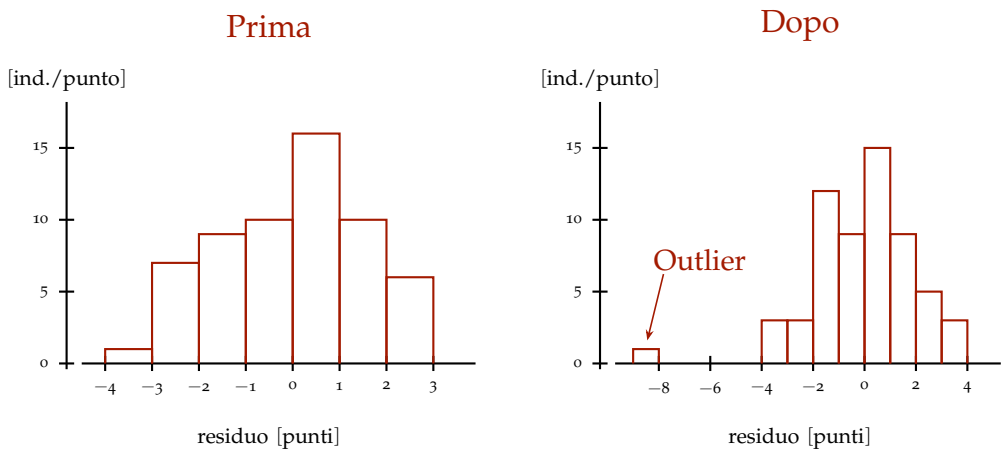
perché un singolo outlier può avere un peso arbitrariamente grande sulla media, mentre, per esempio, questo non accade per la mediana.

Perché il metodo che abbiamo visto sia giustificato, conviene che i residui

$$\text{i-esimo residuo} = b_i - \overbrace{(m\,a_i + q)}^{\text{retta di regressione}}$$

siano distribuiti in maniera (approssimativamente) gaussiana. Questa condizione si può verificare a posteriori, per esempio calcolando i residui e tracciandone un istogramma.





In conclusione, è importante, prima di applicare la regressione lineare, controllare su un diagramma di dispersione che non ci siano evidenti outlier. È buona norma verificare altresì la distribuzione dei residui.

§20.3 Il p-value

I p-value compaiono, in generale, nel contesto dei test di significatività. Nel caso della regressione lineare, noi vogliamo saggiare l'ipotesi seguente — che, per qualche ragione, chiamiamo ipotesi alternativa.

Ipotesi alternativa: A e B sono correlate linearmente.

Per farlo, raccogliamo un campione di n individui, calcoliamo il coefficiente r_{oss} e la retta di regressione, e ci convinciamo che i residui sono distribuiti in modo compatibile con una gaussiana. Ora possiamo chiederci: se A e B fossero scorrelate, quanto sarebbe probabile che un campione di n individui abbia $|r| \geqslant |r_{oss}|$? Detto in altri termini: se A e B fossero scorrelate, quanti dei campioni di n individui, in proporzione a tutti quelli possibili, sarebbero correlati al pari del nostro, o più? Questo è il p-value e misura il grado di significatività statistica con cui possiamo rifiutare l'ipotesi nulla:

Ipotesi nulla: A e B non sono correlate.

Più basso è il p-value, meno è verosimile che la correlazione osservata sia uno scherzo del caso, più, quindi, è ragionevole che sia reale.

Calcolare questo p-value può essere complicato, per via della necessità di caratterizzare l'insieme di tutti i campioni di n individui possibili. Infatti, a seconda di quali ipotesi si fanno sulla popolazione, si ottengono diversi insiemi di campioni possibili, quindi p-value diversi. Il fatto seguente è basato sull'ipotesi che i residui siano indipendenti fra loro e distribuiti in maniera gaussiana.

Fatto
Se valesse l'ipotesi nulla, allora

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

sarebbe distribuito secondo una t di Student con $n-2$ gradi di libertà.

La correlazione è significativa quando possiamo rifiutare l'ipotesi nulla — l'ipotesi di non correlazione. Vale quindi la regola seguente.

Regola (significatività della correlazione)
La correlazione è significativa al livello α se

$$|t| > t_{\alpha/2,\,n-2}$$

Dove, ricordiamo, $t_{\alpha/2,\,n-2}$ è il valore della tavola della t di Student nella colonna $\frac{\alpha}{2}$ alla riga $n-2$ (per $\alpha = 0.05$ la colonna 0.025, per $\alpha = 0.01$ la colonna 0.005).

Esercizio
Considera la situazione dell'esercizio a pagina 149:

a_i	4	5	1	9	4	7	5	4	5	6
b_i	5	6	8	2	4	1	5	3	2	4

$n = 10$

La correlazione è significativa al livello 0.05? E al livello 0.01?

Soluzione. Abbiamo già calcolato $r = -0.75$, da cui

$$t = -0.75\sqrt{\frac{10-2}{1-0.75^2}} \approx -3.21$$

Guardiamo la tavola della t di Student (appendice C).

				$\alpha = 0.05$		$\alpha = 0.01$	
gradi di libertà	0.25	0.10	0.05	0.025	0.01	0.005	
...	...	...	...	...	...	...	
6	0.72	1.44	1.94	2.45	3.14	3.71	
7	0.71	1.41	1.89	2.36	3.00	3.50	
8	0.71	1.40	1.86	2.31	2.90	3.36	$2.31 < 3.21 < 3.36$
9	0.70	1.38	1.83	2.26	2.82	3.25	
10	0.70	1.37	1.81	2.23	2.76	3.17	

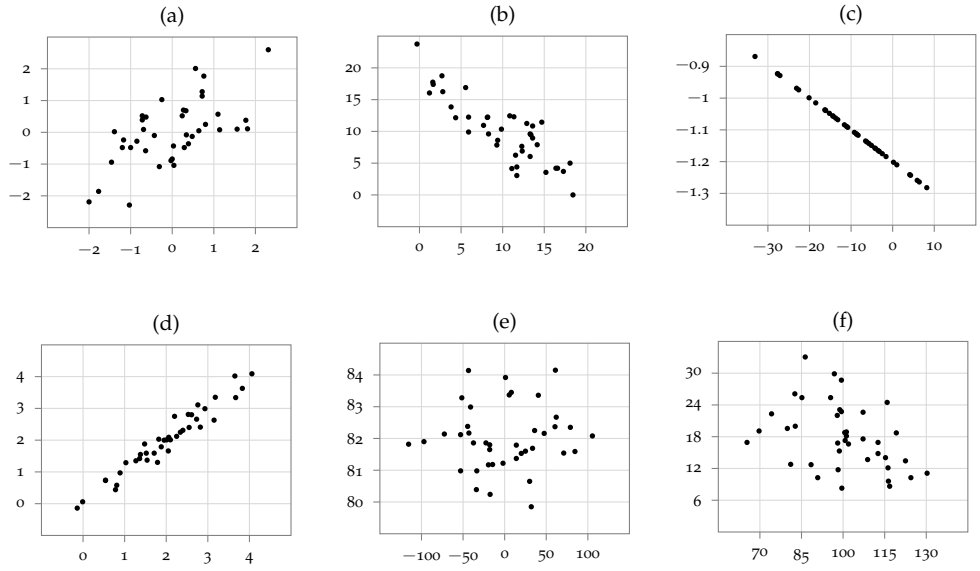
Applicando la regola: alla riga $8 = 10 - 2$ le soglie sono $t_{0.025,8} = 2.31$ e $t_{0.005,8} = 3.36$. Siccome $2.31 < |t| = 3.21 < 3.36$, la correlazione è significativa al livello 0.05, ma non al livello 0.01. □

§20.4 Esercizi

Esercizio 110 (soluzione a p. 318)
I coefficienti di correlazione corrispondenti ai seguenti diagrammi di dispersione sono

$$-0.85 \qquad -0.38 \qquad -1.00 \qquad 0.06 \qquad 0.97 \qquad 0.62$$

Associa ad ogni diagramma il suo coefficiente.



I valori indicati sugli assi sono spaziati di una deviazione standard delle rispettive variabili, con la media posta al centro. Per ciascuno dei diagrammi, calcola l'equazione della retta di regressione e traccia questa retta sul diagramma stesso.

Esercizio 111 (soluzione a p. 320)
Considera le seguenti coppie di valori assunti dalle variabili statistiche X e Y su un campione di sei individui.

x_i	4.9	13.6	0.3	5.8	4.0	5.6
y_i	5.4	2.9	11.7	1.8	6.1	8.1

1. Rappresenta questi punti su un diagramma di dispersione.
2. Calcola la retta di regressione di Y (variabile dipendente) rispetto a X (variabile indipendente).
3. È possibile affermare, al livello di significatività $\alpha = 0.05$, che X e Y sono linearmente correlate?

Esercizio 112 (soluzione a p. 321)
Considera le seguenti coppie di valori assunti dalle variabili statistiche X e Y su un campione di sei individui.

x_i	0.7	9.4	4.3	3.2	6.9	5.3
y_i	4.1	9.9	6.8	6.6	7.2	9.8

1. Rappresenta questi punti su un diagramma di dispersione.
2. Calcola la retta di regressione di Y (variabile dipendente) rispetto a X (variabile indipendente).
3. È possibile affermare, al livello di significatività $\alpha = 0.05$, che X e Y sono linearmente correlate?

Esercizio 113 (soluzione a p. 323)
Considera le seguenti coppie di valori assunti dalle variabili statistiche X e Y su un campione di sei individui.

x_i	4.4	2.1	8.0	4.3	4.7	2.7
y_i	9.4	5.4	9.1	5.1	7.7	3.9

1. Rappresenta questi punti su un diagramma di dispersione.
2. Calcola la retta di regressione di Y (variabile dipendente) rispetto a X (variabile indipendente).
3. È possibile affermare, al livello di significatività $\alpha = 0.05$, che X e Y sono linearmente correlate?

Esercizio 114 (soluzione a p. 324)
Le variabili statistiche X e Y hanno rispettivamente media 3.60 e 84.1, e deviazione standard 0.45 e 8.6. Sapendo che la pendenza della retta di regressione di Y (variabile dipendente) rispetto a X (variabile indipendente) è 14.6, determina l'espressione della retta e il coefficiente di correlazione di Pearson. Supponendo che i dati precedenti siano stati ricavati da un campione di dieci individui, è possibile affermare, al livello di significatività $\alpha = 0.05$, che X e Y sono correlate? Qual è l'espressione della retta di regressione di X (variabile dipendente) rispetto a Y (variabile indipendente), ossia la retta di regressione che si ottiene scambiando le variabili?

1. Con $\hat{p} = 0.40$,

$$\widehat{\text{SEM}} = \sqrt{\frac{0.40 \cdot 0.60}{1000}} \approx 0.0155 \qquad 1.96 \cdot 0.0155 \approx 0.030$$

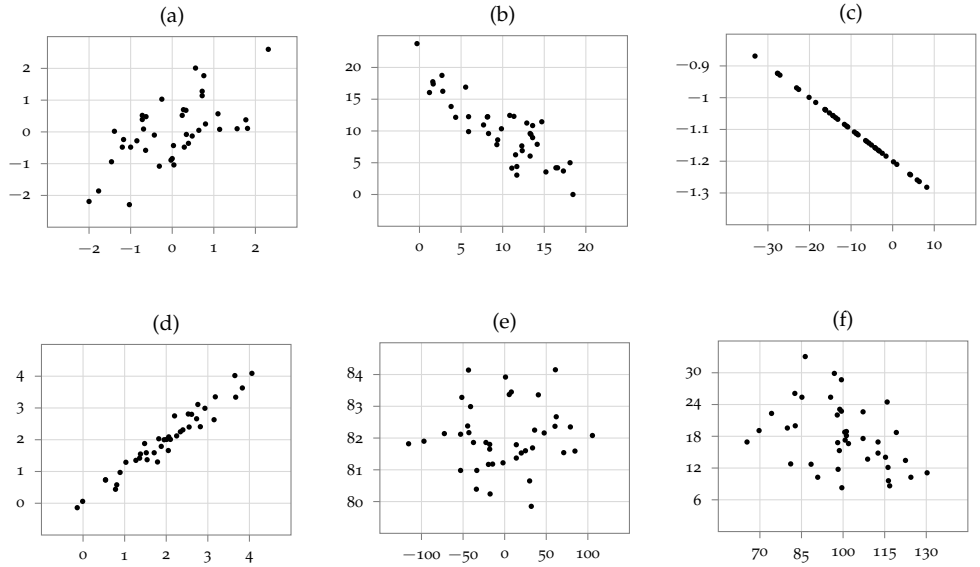
   cioè $\pm 3\%$.

2. Con quattro volte gli intervistati l'errore standard si dimezza: $\pm 1.5\%$.

Esercizio 110 (p. 154)
I coefficienti di correlazione corrispondenti ai seguenti diagrammi di dispersione sono

$$-0.85 \qquad -0.38 \qquad -1.00 \qquad 0.06 \qquad 0.97 \qquad 0.62$$

Associa ad ogni diagramma il suo coefficiente.



I valori indicati sugli assi sono spaziati di una deviazione standard delle rispettive variabili, con la media posta al centro. Per ciascuno dei diagrammi, calcola l'equazione della retta di regressione e traccia questa retta sul diagramma stesso.

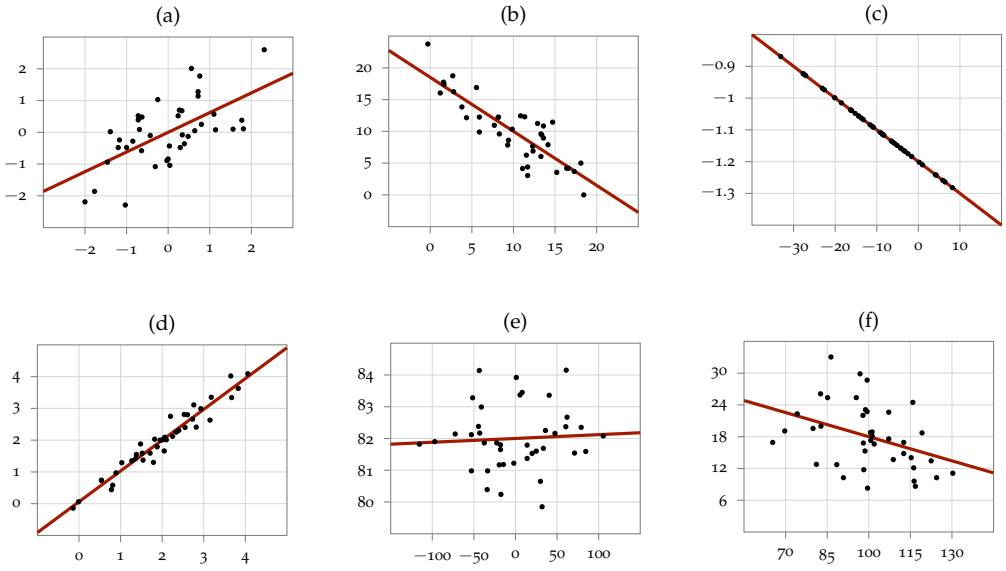
Soluzione. Il valore -1.00 va al diagramma (c), i cui punti stanno esattamente su una retta decrescente. Fra gli altri, il segno si legge dall'inclinazione della nuvola, il valore assoluto da quanto la nuvola è schiacciata:

(a) 0.62 (b) -0.85 (c) -1.00 (d) 0.97 (e) 0.06 (f) -0.38

Per le rette usiamo $B = \mu_B + r\,\frac{\sigma_B}{\sigma_A}\,(A - \mu_A)$, leggendo medie e deviazioni standard sugli assi:

	μ_A	σ_A	μ_B	σ_B	retta
(a)	0	1	0	1	$B = 0.62\,A$
(b)	10	5	10	5	$B = 18.5 - 0.85\,A$
(c)	-10	10	-1.1	0.1	$B = -1.2 - 0.01\,A$
(d)	2	1	2	1	$B = 0.06 + 0.97\,A$
(e)	0	50	82	1	$B = 82 + 0.0012\,A$
(f)	100	15	18	6	$B = 33.2 - 0.152\,A$

Per tracciarle conviene ricordare che la retta di regressione passa per il baricentro (μ_A, μ_B), al centro del diagramma, e che spostandosi a destra di una deviazione standard σ_A la retta sale di r deviazioni standard σ_B (o scende, se $r < 0$). Nei diagrammi, dove ogni quadretto è una deviazione standard, la retta passa per il centro e sale di r quadretti per ogni quadretto verso destra.



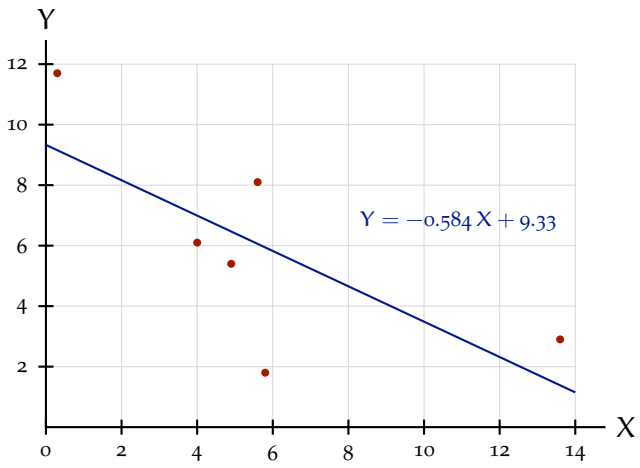
Esercizio 111 (p. 154)
Considera le seguenti coppie di valori assunti dalle variabili statistiche X e Y su un campione di sei individui.

x_i	4.9	13.6	0.3	5.8	4.0	5.6
y_i	5.4	2.9	11.7	1.8	6.1	8.1

1. Rappresenta questi punti su un diagramma di dispersione.
2. Calcola la retta di regressione di Y (variabile dipendente) rispetto a X (variabile indipendente).
3. È possibile affermare, al livello di significatività $\alpha = 0.05$, che X e Y sono linearmente correlate?

Soluzione.

Soluzione della domanda 1. Riportiamo i sei punti sul diagramma di dispersione, insieme alla retta che calcoleremo nella seconda domanda.



□

Soluzione della domanda 2. Calcoliamo media e deviazione standard di X e di Y (con $n = 6$ al denominatore):

$$\mu_X = \frac{34.2}{6} = 5.7000 \qquad \sigma_X = \sqrt{\frac{95.12}{6}} = 3.9816$$

$$\mu_Y = \frac{36.0}{6} = 6.0000 \qquad \sigma_Y = \sqrt{\frac{64.52}{6}} = 3.2792$$

Poi convertiamo i punti in coordinate standard, $x_i^* = (x_i - \mu_X)/\sigma_X$ e $y_i^* = (y_i - \mu_Y)/\sigma_Y$, e moltiplichiamo fra loro le due coordinate di ogni punto:

x_i^*	−0.2009	1.9841	−1.3562	0.0251	−0.4270	−0.0251
y_i^*	−0.1830	−0.9453	1.7382	−1.2808	0.0305	0.6404
$x_i^* y_i^*$	0.0368	−1.8757	−2.3574	−0.0322	−0.0130	−0.0161

Il coefficiente di correlazione è la media di questi prodotti,

$$r = \frac{-4.2576}{6} = -0.7096$$

quindi la retta di regressione, scritta in coordinate standard, è $Y^* = -0.7096\,X^*$. Sostituendo X^* e Y^* con le loro definizioni troviamo la retta nelle unità originali:

$$\frac{Y - 6.0000}{3.2792} = -0.7096\,\frac{X - 5.7000}{3.9816} \qquad \Longrightarrow \qquad Y = -0.584\,X + 9.33$$

Riportata sul diagramma, la retta taglia a metà la nuvola dei punti nella maniera giusta — scende insieme alla nuvola, e lascia punti sopra e sotto da un capo all'altro. □

Soluzione della domanda 3. Per saggiare la correlazione calcoliamo, con $n = 6$,

$$t = r\sqrt{\frac{n-2}{1-r^2}} = -0.7096 \cdot \sqrt{\frac{4}{1-0.7096^2}} = -0.7096 \cdot 2.838 \approx -2.01$$

Se fra X e Y non ci fosse correlazione, t seguirebbe una t di Student con $n - 2 = 4$ gradi di libertà, che supera la soglia $t_{0.025,\,4} = 2.78$ in valore assoluto solo nel 5% dei casi (tavola in appendice, riga 4, colonna 0.025). Siccome $2.01 < 2.78$, la correlazione non è significativa al livello 0.05, e la risposta è no — con sei punti soli, anche un coefficiente $r \approx -0.710$ potrebbe essere dovuto al caso. □

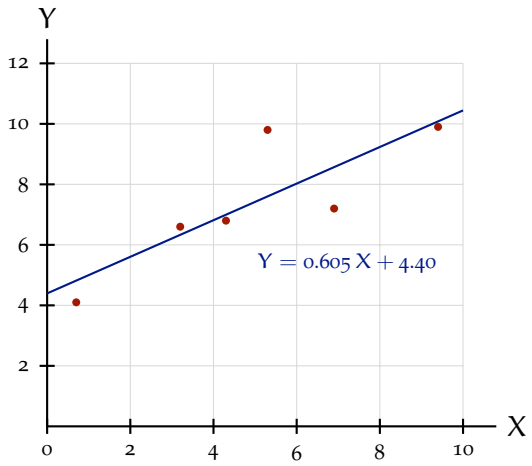
Esercizio 112 (p. 155)
Considera le seguenti coppie di valori assunti dalle variabili statistiche X e Y su un campione di sei individui.

x_i	0.7	9.4	4.3	3.2	6.9	5.3
y_i	4.1	9.9	6.8	6.6	7.2	9.8

1. Rappresenta questi punti su un diagramma di dispersione.
2. Calcola la retta di regressione di Y (variabile dipendente) rispetto a X (variabile indipendente).
3. È possibile affermare, al livello di significatività $\alpha = 0.05$, che X e Y sono linearmente correlate?

Soluzione.

Soluzione della domanda 1. Oltre ai sei punti, sul diagramma di dispersione abbiamo tracciato la retta di regressione, ricavata più avanti.



□

Soluzione della domanda 2. Ci servono per prime le medie e le deviazioni standard, con $n = 6$ al denominatore:

$$\mu_X = \frac{29.8}{6} = 4.9667 \qquad \sigma_X = \sqrt{\frac{45.27}{6}} = 2.7469$$

$$\mu_Y = \frac{44.4}{6} = 7.4000 \qquad \sigma_Y = \sqrt{\frac{23.94}{6}} = 1.9975$$

Con queste passiamo alle coordinate standard $x_i^* = (x_i - \mu_X)/\sigma_X$ e $y_i^* = (y_i - \mu_Y)/\sigma_Y$, e nell'ultima riga scriviamo i loro prodotti.

x_i^*	−1.5533	1.6139	−0.2427	−0.6431	0.7038	0.1213
y_i^*	−1.6521	1.2516	−0.3004	−0.4005	−0.1001	1.2015
$x_i^* y_i^*$	2.5661	2.0199	0.0729	0.2576	−0.0705	0.1458

Facendo la media dei prodotti otteniamo il coefficiente di correlazione

$$r = \frac{4.9918}{6} = 0.8320$$

Abbiamo quindi $Y^* = 0.8320\, X^*$ in coordinate standard, e tornando a X e Y

$$\frac{Y - 7.4000}{1.9975} = 0.8320\, \frac{X - 4.9667}{2.7469} \qquad \Longrightarrow \qquad Y = 0.605\, X + 4.40$$

(controprova sul diagramma: la retta sale con la nuvola dei punti e la taglia a metà nella maniera giusta, con punti da una parte e dall'altra lungo tutto il tratto). □

Soluzione della domanda 3. Il valore di t, per $n = 6$, è

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.8320 \cdot \sqrt{\frac{4}{1 - 0.8320^2}} = 0.8320 \cdot 3.605 \approx 3.00$$

e va confrontato con la soglia $t_{0.025,\,4} = 2.78$ (tavola in appendice, riga $n - 2 = 4$, colonna 0.025) — in mancanza di correlazione, t sarebbe distribuito secondo una t di Student con questi gradi di libertà, che oltrepassa 2.78, in valore assoluto, nel 5% dei casi. Abbiamo $3.00 > 2.78$, quindi sì, la correlazione è significativa al livello 0.05. □

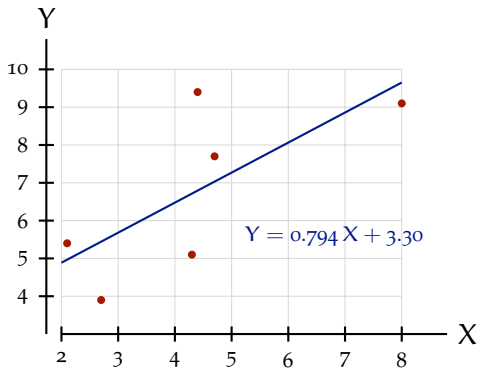
Esercizio 113 (p. 155)
Considera le seguenti coppie di valori assunti dalle variabili statistiche X e Y su un campione di sei individui.

x_i	4.4	2.1	8.0	4.3	4.7	2.7
y_i	9.4	5.4	9.1	5.1	7.7	3.9

1. Rappresenta questi punti su un diagramma di dispersione.
2. Calcola la retta di regressione di Y (variabile dipendente) rispetto a X (variabile indipendente).
3. È possibile affermare, al livello di significatività $\alpha = 0.05$, che X e Y sono linearmente correlate?

Soluzione.

Soluzione della domanda 1. Nel diagramma di dispersione anticipiamo anche la retta di regressione, che troveremo fra poco.



□

Soluzione della domanda 2. Le due medie e le due deviazioni standard (queste ultime calcolate con $n = 6$ al denominatore) sono

$$\mu_X = \frac{26.2}{6} = 4.3667 \qquad \sigma_X = \sqrt{\frac{21.23}{6}} = 1.8812$$

$$\mu_Y = \frac{40.6}{6} = 6.7667 \qquad \sigma_Y = \sqrt{\frac{26.11}{6}} = 2.0862$$

e da queste ricaviamo le coordinate standard dei sei punti, $x_i^* = (x_i - \mu_X)/\sigma_X$ e $y_i^* = (y_i - \mu_Y)/\sigma_Y$, insieme ai loro prodotti.

x_i^*	0.0177	−1.2049	1.9314	−0.0354	0.1772	−0.8860
y_i^*	1.2623	−0.6551	1.1185	−0.7989	0.4474	−1.3741
$x_i^* y_i^*$	0.0224	0.7893	2.1602	0.0283	0.0793	1.2174

La media dell'ultima riga è il coefficiente di correlazione,

$$r = \frac{4.2969}{6} = 0.7161$$

e la retta di regressione, che in coordinate standard è $Y^* = 0.7161\,X^*$, nelle unità originali diventa

$$\frac{Y - 6.7667}{2.0862} = 0.7161\,\frac{X - 4.3667}{1.8812} \qquad \Longrightarrow \qquad Y = 0.794\,X + 3.30$$

Sul diagramma vediamo che la retta segue l'inclinazione della nuvola dei punti e la divide a metà come deve, lasciandone alcuni sopra e alcuni sotto a entrambe le estremità. □

Soluzione della domanda 3. Con $n = 6$ abbiamo

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.7161 \cdot \sqrt{\frac{4}{1 - 0.7161^2}} = 0.7161 \cdot 2.865 \approx 2.05$$

Per decidere ci serve la soglia della t di Student a $n - 2 = 4$ gradi di libertà, $t_{0.025,\,4} = 2.78$ (tavola in appendice, riga 4, colonna 0.025). Se le due variabili non fossero correlate, t seguirebbe proprio questa distribuzione, e andrebbe oltre 2.78 in valore assoluto nel 5% dei casi. Il nostro valore resta sotto, $2.05 < 2.78$, quindi la risposta è no: la correlazione non è significativa al livello 0.05. Sei punti sono pochi — non bastano a escludere che un coefficiente $r \approx 0.716$ sia frutto del caso. □

> **Esercizio 114** (p. 155)
> Le variabili statistiche X e Y hanno rispettivamente media 3.60 e 84.1, e deviazione standard 0.45 e 8.6. Sapendo che la pendenza della retta di regressione di Y (variabile dipendente) rispetto a X (variabile indipendente) è 14.6, determina l'espressione della retta e il coefficiente di correlazione di Pearson. Supponendo che i dati precedenti siano stati ricavati da un campione di dieci individui, è possibile affermare, al livello di significatività $\alpha = 0.05$, che X e Y sono correlate? Qual è l'espressione della retta di regressione di X (variabile dipendente) rispetto a Y (variabile indipendente), ossia la retta di regressione che si ottiene scambiando le variabili?

Soluzione. Ricordiamo che in coordinate standard la retta di regressione è $Y^* = r\,X^*$, e che tornando alle unità originali la sua pendenza diventa $r\,\frac{\sigma_Y}{\sigma_X}$. Conoscendo la pendenza, ricaviamo quindi il coefficiente di correlazione

$$r \cdot \frac{8.6}{0.45} = 14.6 \qquad \Longrightarrow \qquad r = 14.6 \cdot \frac{0.45}{8.6} \approx 0.764$$

Per scrivere la retta basta ora sapere che passa per il baricentro $(\mu_X, \mu_Y) = (3.60,\, 84.1)$, da cui

$$Y - 84.1 = 14.6\,(X - 3.60) \qquad \Longrightarrow \qquad Y = 14.6\,X + 31.5$$

Passiamo alla significatività. Il campione ha $n = 10$ individui, quindi

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.764 \cdot \sqrt{\frac{8}{1-0.764^2}} = 0.764 \cdot 4.383 \approx 3.35$$

In assenza di correlazione questo t seguirebbe una t di Student con $n - 2 = 8$ gradi di libertà, che in valore assoluto va oltre la soglia $t_{0.025,8} = 2.31$ nel 5% dei casi (tavola in appendice, riga 8, colonna 0.025). Dato che $3.35 > 2.31$, possiamo rispondere sì: la correlazione è significativa al livello 0.05.

Infine scambiamo le variabili. Il coefficiente di correlazione resta lo stesso, e in coordinate standard la nuova retta di regressione è $X^* = r\,Y^*$, ossia

$$\frac{X - 3.60}{0.45} = 0.764\,\frac{Y - 84.1}{8.6} \qquad \Longrightarrow \qquad X = 0.0400\,Y + 0.238$$

Nota che questa non è la retta di prima risolta rispetto a X, che avrebbe pendenza $\frac{1}{14.6} \approx 0.0685$. Le due rette di regressione passano entrambe per il baricentro, ma coincidono solo se $r = \pm 1$, perché la regressione verso la media agisce in tutti e due i versi.

Esercizio 115 (p. 168)
Supponi di avere due farmaci A e B contro la medesima malattia. L'efficacia del farmaco A è corroborata da uno studio clinico al livello di significatività 0.05 (p-value < 0.05), l'efficacia di B è corroborata da uno studio al livello di significatività 0.01 (p-value < 0.01). Per ciascuna delle seguenti affermazioni indica se sia giustificata oppure no, spiegando la ragione della tua scelta.

1. Il farmaco A è efficace almeno nel 95% dei casi di malattia, mentre il farmaco B almeno nel 99%.
2. C'è maggiore evidenza statistica a favore dell'efficacia del farmaco B rispetto al farmaco A.
3. Anche se supponiamo che entrambi i farmaci siano effettivamente efficaci, non è detto che il farmaco B produca un effetto più forte del farmaco A.
4. Anche se supponiamo che entrambi i farmaci siano effettivamente efficaci, non è detto che il farmaco B curi un maggior numero di casi del farmaco A.
5. Se il farmaco A non fosse efficace, una replica dello studio relativo avrebbe il 95% di probabilità di rifiutare l'ipotesi nulla.
6. Se il farmaco A non fosse efficace, una replica dello studio relativo avrebbe il 95% di probabilità di non riprodurre il risultato positivo di quest'ultimo.
7. Se il farmaco A fosse efficace, una replica dello studio relativo avrebbe il 5% di probabilità di non rifiutare l'ipotesi nulla.
8. La probabilità che il farmaco A non sia efficace è minore del 5%.