STATISTICA MULTIVARIATA

Corso di laurea in matematica | Anno accademico 2004/05 | Numero 3

Il prodotto scalare

Situazione 10.1. Siano $x, y \in \mathbb{R}^n$ con $x = (x_1, \dots, x_n)^t$, $y = (y_1, \dots, y_n)^t$. Supponiamo di nuovo che x ed y non siano diagonali. Possiamo allora formare il coefficiente di correlazione

$$r := r_{xy} = \frac{\|x^{CE}, y^{CE}\|}{|x^{CE}||y^{CE}|} = \|x^{NG}, y^{NG}\|$$

già introdotto nella definizione 8.4.

$\varphi, \lambda, \tau, p$ sono definiti come a pagina 8.

Nota 10.2. L'equazione $\|x, 1^n\| = n\overline{x}$ dell'osservazione 6.10, benché immediata nella dimostrazione, stabilisce un importante legame tra un concetto statistico, la media $\overline{x}$, e un concetto geometrico, il prodotto scalare.

Il coefficiente di correlazione è definito mediante un prodotto scalare. Il prodotto scalare di due vettori $u, v \in \mathbb{R}^n$ è a sua volta profondamente legato alla lunghezza $|u+v|$ della somma di due vettori oppure anche alla lunghezza $|u-v|$ della differenza. Abbiamo infatti

$$\begin{aligned} |u+v|^2 &= \sum_{k=1}^n (u_k + v_k)^2 \\ &= \sum_{k=1}^n u_k^2 + \sum_{k=1}^n v_k^2 + 2\sum_{k=1}^n u_k v_k \\ &= |u|^2 + |v|^2 + 2\|u, v\| \end{aligned}$$

e similmente

$$|u-v|^2 = |u|^2 + |v|^2 - 2\|u, v\|$$

I due punti u e v formano insieme all'origine 0 un triangolo (eventualmente degenerato) i cui lati hanno le lunghezze $|u|$, $|v|$ e $|u-v|$. Assumiamo che il triangolo non sia degenerato e sia α l'angolo opposto al lato di lunghezza $|u-v|$. Per il teorema del coseno abbiamo

$$|u-v|^2 = |u|^2 + |v|^2 - 2|u||v|\cos\alpha$$

da cui

$$\|u, v\| = |u||v|\cos\alpha$$

come abbiamo già osservato a pagina 8.

Il coefficiente di correlazione di x ed y, nonostante il nome prometta molto di più, è essenzialmente un parametro che lega x^{NG} ed y^{NG} con $x^{NG} + y^{NG}$ ed $x^{NG} - y^{NG}$.

Corollario 10.3. *Siano $u, v \in \mathbb{R}^n$ ed $\alpha\beta \in \mathbb{R}$. Allora*

$$|\alpha u + \beta v|^2 = \alpha^2|u|^2 + \beta^2|v|^2 + 2\alpha\beta\|u, v\|$$

Dimostrazione. Per la nota 10.2 e la bilinearità del prodotto scalare abbiamo

$$\begin{aligned} |\alpha u + \beta v|^2 &= |\alpha u|^2 + |\beta v|^2 + 2\|\alpha u, \beta v\| \\ &= \alpha^2|u|^2 + \beta^2|v|^2 + 2\alpha\beta\|u, v\| \end{aligned}$$

Osservazione 10.4. *Siano $u, v \in \mathbb{R}^n$ ed $\alpha, \beta \in \mathbb{R}$. Allora*

$$(\alpha u + \beta v)^{CE} = \alpha u^{CE} + \beta v^{CE}$$

Dimostrazione. Per la linearità della media abbiamo

$$\begin{aligned} (\alpha u + \beta v)^{CE} &= \alpha u + \beta v - \overline{\alpha u + \beta v}1^n \\ &= \alpha u + \beta v - (\alpha\overline{u} + \beta\overline{v})1^n \\ &= \alpha(u - \overline{u}1^n) + \beta(v - \overline{v}1^n) \\ &= \alpha u^{CE} + \beta v^{CE} \end{aligned}$$

Lemma 10.5. *Siano $u, v \in \mathbb{R}^n$ vettori di lunghezza 1, cioè $|u| = |v| = 1$. Allora*

$$\|u, v\| = 1 - \frac{1}{2}|u-v|^2$$

Dimostrazione. Per la nota 10.2 abbiamo

$$\begin{aligned} |u-v|^2 &= |u|^2 + |v|^2 - 2\|u, v\| \\ &= 2 - 2\|u, v\| \end{aligned}$$

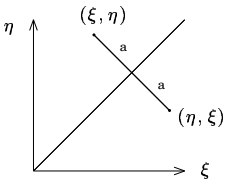
per cui

$$2\|u, v\| = 2 - |u-v|^2$$

Ciò implica l'enunciato.

Nota 10.6. Consideriamo un punto $z = (\xi, \eta)$ nel piano e il punto $z' = (\eta, \xi)$ che si ottiene riflettendo z alla retta $\eta = \xi$.

a sia la distanza di z da questa retta.



Allora $z - z' = (\xi - \eta, \eta - \xi)$, per cui

$$(2a)^2 = |z - z'|^2 = 2(\xi - \eta)^2$$

cosicché $a^2 = \frac{1}{2}(\xi - \eta)^2$.

$\frac{1}{2}(\xi - \eta)^2$ è quindi il quadrato della distanza del punto (ξ, η) dalla retta $\eta = \xi$.

Nota 10.7. Siano $u, v \in \mathbb{R}^n$ vettori di lunghezza 1. Per $i = 1, \dots, n$ sia a_i la distanza del punto (u_i, v_i) dalla retta $\eta = \xi$ in $\mathbb{R}_2$. Allora

$$\|u, v\| = 1 - \sum_{i=1}^n a_i^2$$

Dimostrazione. Dalla nota 10.6 sappiamo che $a_i^2 = \frac{1}{2}(u_i - v_i)^2$. Per il lemma 10.5

$$\begin{aligned} \|u, v\| &= 1 - \frac{1}{2}|u-v|^2 \\ &= 1 - \sum_{i=1}^n \frac{1}{2}(u_i - v_i)^2 = 1 - \sum_{i=1}^n a_i^2 \end{aligned}$$

In questo numero

Algebra della varianza

Proposizione 10.8. *Siano $u, v \in \mathbb{R}^n$ ed $\alpha, \beta \in \mathbb{R}$. Allora*

$$s^2_{\alpha u + \beta v} = \alpha^2 s_u^2 + \beta^2 s_v^2 + 2\alpha\beta s_{uv}$$

Dimostrazione. Usando il corollario 10.3 e l'osservazione 10.4 abbiamo

$$\begin{aligned} s^2_{\alpha u + \beta v} &= \frac{|(\alpha u + \beta v)^{CE}|^2}{n-1} \\ &= \frac{|\alpha u^{CE} + \beta v^{CE}|^2}{n-1} \\ &= \alpha^2 \frac{|u^{CE}|^2}{n-1} + \beta^2 \frac{|v^{CE}|^2}{n-1} \\ &\quad + 2\alpha\beta \frac{\|u^{CE}, v^{CE}\|}{n-1} \\ &= \alpha^2 s_u^2 + \beta^2 s_v^2 + 2\alpha\beta s_{uv} \end{aligned}$$

Corollario 10.9. *Siano $u, v \in \mathbb{R}^n$. Allora*

$$s^2_{u+v} = s_u^2 + s_v^2 + 2s_{uv}$$

Osservazione 10.10. *Siano $u, v \in \mathbb{R}^n$ ed $\alpha, \beta \in \mathbb{R}$. Allora*

$$s_{\alpha u, \beta v} = \alpha\beta s_{uv}$$

Dimostrazione. Usando l'osservazione 10.4 abbiamo

$$\begin{aligned} s_{\alpha u, \beta v} &= \frac{\|(\alpha u)^{CE}, (\beta v)^{CE}\|}{n-1} \\ &= \frac{\|\alpha u^{CE}, \beta v^{CE}\|}{n-1} \\ &= \alpha\beta \frac{\|u^{CE}, v^{CE}\|}{n-1} = \alpha\beta s_{uv} \end{aligned}$$

Osservazione 10.11. *Siano $u, v \in \mathbb{R}^n$ ed $\alpha, \beta \in \mathbb{R}$. Allora $s_{u+\alpha 1^n, v+\beta 1^n} = s_{uv}$.*

In particolare $s_{u+\alpha 1^n} = s_u$.

Dimostrazione. Abbiamo

$$(u + \alpha 1^n)^{CE} = u^{CE} + \alpha(1^n)^{CE} = u^{CE}$$

perché $(1^n)^{CE} = 0$; per la stessa ragione $(v + \beta 1^n)^{CE} = v^{CE}$.

Ciò implica l'enunciato.

Il coefficiente di correlazione

„Il falsificatore astuto è più abile. Applica metodi formalmente inattaccabili a dati non adatti a questi metodi ...“ (trad. da Fassl, 3)

Corollario 11.1. $r = 1 - \sum_{i=1}^{n} a_i^2$

dove a_i^2 è il quadrato della distanza di (x_i^{NG}, y_i^{NG}) dalla retta $\eta = \xi$ nel piano $\mathbb{R}_2$.

Questa è una della più importanti interpretazioni del coefficiente di correlazione.

Dimostrazione. Siccome

$$r = \|x^{NG}, y^{NG}\|$$

l'enunciato segue dalla nota 10.7.

Proposizione 11.2. $r = \dfrac{s_{xy}}{s_x s_y}$

Dimostrazione. Abbiamo

$$s_{xy} = \frac{\|x^{CE}, y^{CE}\|}{n-1}$$

$$r = \frac{\|x^{CE}, y^{CE}\|}{|x^{CE}||y^{CE}|} = \frac{\|x^{CE}, y^{CE}\|}{(n-1)s_x s_y} = \frac{s_{xy}}{s_x s_y}$$

Corollario 11.3. $r = 0 \iff s_{xy} = 0$.

Corollario 11.4.

$r = 0 \iff s^2_{x+y} = s^2_x + s^2_y$.

Dimostrazione. Ciò segue dai corollari 11.3 e 10.9.

Corollario 11.5. *Siano $\alpha, \beta \in \mathbb{R} \setminus 0$. Allora*

$$r_{\alpha x, \beta y} = (\operatorname{sgn} \alpha\beta) \cdot r_{xy}$$

Dimostrazione. Usando l'osservazione 10.10 e la proposizione 10.8 dalla proposizione 11.2 abbiamo

$$r_{\alpha x, \beta y} = \frac{s_{\alpha x, \beta y}}{s_{\alpha x} s_{\beta y}} = \frac{\alpha\beta}{|\alpha||\beta|} \frac{s_{xy}}{s_x s_y}$$

Corollario 11.6. *Siano $\alpha, \beta \in \mathbb{R}$. Allora*

$$r_{x+\alpha 1^n, y+\beta 1^n} = r_{xy}$$

Dimostrazione. Ciò segue dalla proposizione 11.2 e dall'osservazione 10.11, oppure in modo geometrico dalla figura nella definizione 8.4.

Nota 11.7. Abbiamo visto nella nota 8.5 che la retta di regressione di y rispetto ad x può essere scritta nella forma

$$\eta - \overline{y} = \lambda(\xi - \overline{x}) \quad \text{con } \lambda = r\frac{|y^{CE}|}{|x^{CE}|}.$$

Sostituituiamo adesso x ed y con x^{NG} ed y^{NG}. Il coefficiente di correlazione non cambia e le medie sono uguali a 0. Inoltre

$$|(x^{NG})^{CE}| = |x^{NG}| = 1$$
$$|(y^{NG})^{CE}| = |y^{NG}| = 1$$

per cui l'equazione della retta di regressione di y^{NG} rispetto ad x^{NG} è semplicemente

$$\eta = r\xi$$

Il coefficiente di correlazione è quindi la pendenza della retta di regressione di y^{NG} rispetto ad x^{NG}.

---

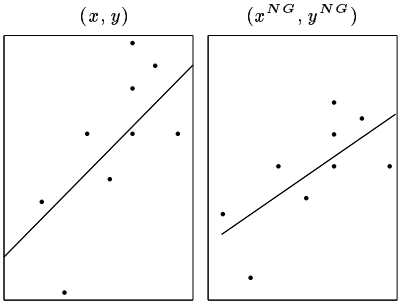
x ed y siano i dati relativi ai minerali di ematite della prima tabella a pagina 9. Calcoliamo le normalizzazioni geometriche con la nostra funzione `Sg.ng` (pagina 7):

```
x=c(28,29,30,31,32,32,32,33,34)
y=c(27,23,30,28,30,32,34,33,30)

xng=Sg.ng(x); yng=Sg.ng(y)
print(round(xng,2))
print(round(yng,2))
```

Otteniamo così la tabella

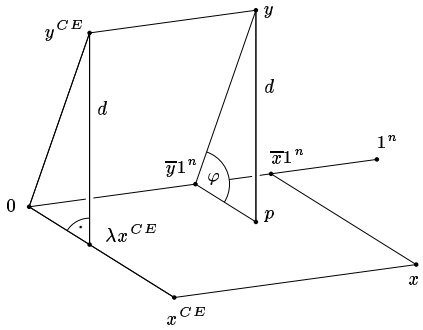
x	y	x^{NG}	y^{NG}
28	27	−0.59	−0.28
29	23	−0.41	−0.70
30	30	−0.22	0.04
31	28	−0.04	−0.18
32	30	0.14	0.04
32	32	0.14	0.25
32	34	0.14	0.46
33	33	0.33	0.35
34	30	0.51	0.04



Nota 11.8. Ricordiamo dalla nota 8.3 che

$$p = \lambda x + \tau 1^n$$

Il vettore dei residui $d := y - p$ è stato introdotto a pagina 9.



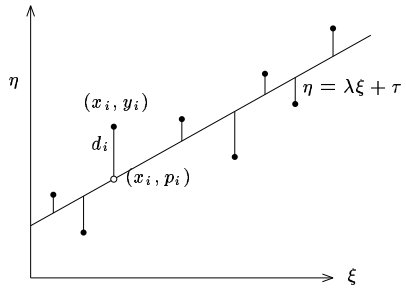
Le coordinate di $p = \lambda x + \tau 1^n$ sono naturalmente

$$p_i = \lambda x_i + \tau$$

i punti (x_i, p_i) sono quindi esattamente i punti sulla retta di regressione con ascisse uguale ad x_i.

A pagina 9 abbiamo definito i residui

$$d_i = y_i - (\lambda x_i + \tau) = y_i - p_i$$



Proposizione 11.9.

$$|d|^2 = (1 - r^2) \cdot |y^{CE}|^2$$

Dimostrazione. Nella prima figura della nota 11.8 vediamo che

$$\frac{d}{|y^{CE}|} = |\sin\varphi| = \sqrt{1 - r^2}$$

e ciò implica il risultato.

Osservazione 11.10. $-1 \le r \le 1$.

Dimostrazione. Sappiamo che $r = \cos\varphi$.

Corollario 11.11. *Sono equivalenti:*

(1) $d = 0$.
(2) *I punti (x_i, y_i) si trovano tutti sulla retta di regressione di y rispetto ad x.*
(3) $r^2 = 1$.
(4) $r = \pm 1$.

Dimostrazione. (1) $\iff$ (2): Chiaro.

(1) $\iff$ (3): Siccome $|y^{CE}| \neq 0$, dalla proposizione 11.9 segue che

$$d = 0 \iff 1 - r^2 = 0.$$

(3) $\iff$ (4): Chiaro.

Osservazione 11.12.
(1) $r = 1 \iff x^{NG} = y^{NG}$.
(2) $r = -1 \iff x^{NG} = -y^{NG}$.
(3) $r = 0 \iff x^{NG} \perp y^{NG}$
$\iff x^{CE} \perp y^{CE}$.

Dimostrazione. $r = \cos\varphi$ e abbiamo già osservato che φ è anche l'angolo tra le normalizzazioni geometriche x^{NG} ed y^{NG}.

Osservazione 11.13. *Sia $\overline{x} = 0$. Allora*

$$\|x^{CE}, y^{CE}\| = \|x, y\|$$

Dimostrazione. Per il lemma 7.2 abbiamo

$$\|x^{CE}, y^{CE}\| = \|x^{CE}, y\| = \|x, y\|$$

perché $\overline{x} = 0$ implica $x^{CE} = x$.

Corollario 11.14. *Sia $\overline{x} = 0$. Allora*

$$r = 0 \iff x \perp y$$

Nel linguaggio comune il termine *correlazione* significa un rapporto stretto tra due elementi e questo significato viene spesso meccanicamente applicato al coefficiente di correlazione che invece deve essere compreso solo come un parametro numerico che non individua una precisa configurazione statistico-causale tra due variabili.

Decomposizione della varianza

Osservazione 12.1. *Sia $u \in \mathbb{R}^n$. Allora*

$$s_{u^{CE}} = s_u$$

Dimostrazione. Ciò segue dall'osservazione 10.11.

Osservazione 12.2. $\overline{p} = \overline{y}$.

Dimostrazione. Dalla prima figura della nota 11.8 è chiaro che $\overline{y}1^n$ non è solo la proiezione ortogonale di y sulla retta $\mathbb{R}1^n$ (teorema 6.13), ma anche la proiezione ortogonale di p sulla stessa retta e ciò implica (ancora per il teorema 6.13) che $\overline{y}1^n = \overline{p}1^n$ e quindi $\overline{y} = \overline{p}$.

La dimostrazione analitica è altrettanto facile: Dalla nota 8.3 sappiamo che

$$p = \overline{y}1^n + \lambda x^{CE}$$

Però $\overline{x^{CE}} = 0$, per cui $\overline{p} = \overline{\overline{y}1^n} = \overline{y}$.

Corollario 12.3. $\overline{d} = 0$ *e quindi* $d = d^{CE}$.

Dimostrazione. Ciò segue dall'osservazione 12.2 ed è evidente anche dalla prima figura nella nota 11.8, da cui si vede che d è ortogonale a 1^n e possiede quindi media 0 per il corollario 6.11.

Corollario 12.4. $p^{CE} = \lambda x^{CE}$.

Dimostrazione. Infatti

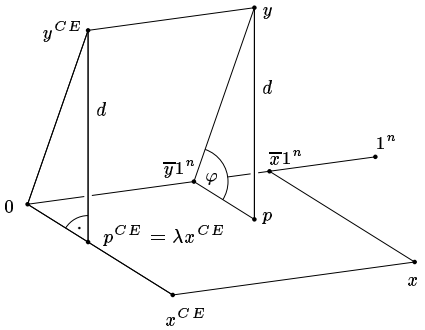
$$\lambda x^{CE} = p - \overline{y}1^n = p - \overline{p}1^n = p^{CE}$$

per l'osservazione 12.2.

Corollario 12.5.

$$\begin{aligned} |y^{CE}|^2 &= |y - p|^2 + |p^{CE}|^2 \\ &= |d|^2 + |p^{CE}|^2 \end{aligned}$$

Usando il corollario 12.4 l'enunciato segue dalla figura - non è altro che il teorema di Pitagora applicato al triangolo a sinistra.



Proposizione 12.6. $|p^{CE}|^2 = r^2 |y^{CE}|^2$.

Dimostrazione. Per il corollario 12.5 e la proposizione 11.9 abbiamo

$$\begin{aligned} |p^{CE}|^2 &= |y^{CE}|^2 - d^2 \\ &= |y^{CE}|^2 - (1 - r^2)|y^{CE}|^2 \\ &= r^2 |y^{CE}|^2 \end{aligned}$$

Proposizione 12.7. $s_p^2 = \lambda^2 s_x^2$.

Dimostrazione. Dal corollario 12.4 abbiamo $p^{CE} = \lambda x^{CE}$. L'enunciato segue dall'osservazione 12.1 e dalla proposizione 10.8.

Teorema 12.8. $s_y^2 = s_p^2 + s_d^2 = \lambda^2 s_x^2 + s_d^2$.

Dimostrazione. Ciò segue dal corollario 12.5, perché dal corollario 12.3 sappiamo che $d = d^{CE}$, per cui abbiamo

$$\begin{aligned} |y^{CE}|^2 &= (n-1)s_y^2 \\ |p^{CE}|^2 &= (n-1)s_p^2 \\ |d|^2 = |d^{CE}|^2 &= (n-1)s_d^2 \end{aligned}$$

Nota 12.9. Il teorema 12.8 è molto importante in statistica e costituisce una *decomposizione della varianza* di y nella somma tra la varianza di p, cioè la parte di s_y che deriva direttamente dalla regressione di y rispetto ad x, e la varianza di d, cioè la varianza del vettore dei residui.

s_d^2 perciò si chiama anche la *varianza residua* (di y rispetto ad x). La varianza di y è quindi uguale alla varianza dovuta alla regressione più la varianza residua.

Il coefficiente

$$\frac{s_p^2}{s_y^2} = \lambda^2 \frac{s_x^2}{s_y^2}$$

dà una misura di quanto la regressione da sola determina la varianza di y e si chiama per questa ragione il *coefficiente di determinazione* (di y rispetto ad x).

Proposizione 12.10. *Il coefficiente di determinazione è uguale al quadrato del coefficiente di correlazione:* $\frac{s_p^2}{s_y^2} = r^2$.

Dimostrazione. Ciò segue direttamente dalla proposizione 12.7 e dall'equazione

$$\lambda = r \frac{s_y}{s_x}$$

che abbiamo visto a pagina 9.

Nota 12.11. Per il corollario 11.11 il coefficiente di determinazione è uguale a 1 se e solo se i punti (x_i, y_i) si trovano tutti sulla retta di regressione di y rispetto ad x. Dalla proposizione 12.10 segue inoltre che il coefficiente di determinazione non cambia se scambiamo x ed y; infatti per definizione $r_{xy} = r_{yx}$.

Nota 12.12. Nelle ipotesi che abbiamo fatto nelle osservazioni che seguono la nota 9.1 le variabili x_i ed y_i hanno ruoli diversi. In situazioni in cui nessuna delle due variabili può essere considerata indipendente si può disegnare anche la retta di regressione

$$\xi = \lambda' \eta + \tau'$$

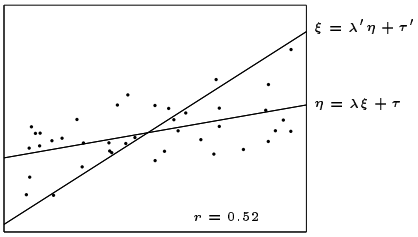
degli x_i rispetto agli y_i. Allora, siccome $r_{xy} = r_{yx} = r$, abbiamo

$$\lambda = r \frac{|y^{CE}|}{|x^{CE}|} \qquad \lambda' = r \frac{|x^{CE}|}{|y^{CE}|}$$

Da ciò segue

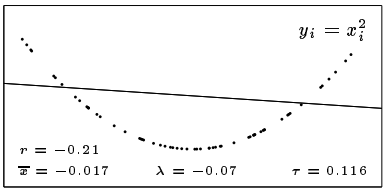
$$r^2 = \lambda \lambda'$$

$|r| = \sqrt{\lambda \lambda'}$ è quindi la *media geometrica* delle pendenze delle due rette di regressione.



Le critiche

Nota 12.13. Abbiamo visto finora le più importanti interpretazioni del coefficiente di correlazione. Esse mostrano che si tratta di un concetto essenzialmente geometrico che dovrebbe essere quindi utilizzato solo in quei casi in cui i legami geometrici hanno un significato statistico per il problema che si studia (cfr. però quanto detto nella nota 10.2). In particolare si dovrebbero evitare interpretazioni *causali*, anche in casi di correlazioni vicina a 1. Una correlazione uguale o vicina a 0 a sua volta non implica che non ci sono legami statistici o causali tra le variabili. Se ad esempio $\overline{x} = 0$ e con ogni punto (x_i, y_i) anche $(-x_i, y_i)$ appartiene ai dati (con la stessa molteplicità se presente più volte), per il corollario 11.14 il coefficiente di correlazione si annulla, anche quando sussiste un semplice legame funzionale tra le variabili, ad esempio ogni volta che $y_i = f(x_i)$, dove f è una funzione simmetrica, cioè tale che $f(\xi) = f(-\xi)$.



In questo caso la retta di regressione è data da $\eta = \overline{y}$, come segue dalla relazione $\tau = \overline{y} - \lambda \overline{x}$.

Un coefficiente di correlazione nullo non significa quindi una mancanza di legami causali tra x ed y, ma esprime piuttosto una forma di *simmetria*.

Il coefficiente di correlazione e i coefficienti della retta di regressione sono molto sensibili alla presenza anche di pochi valori eccezionali (in inglese *outliers*). Talvolta valori estremi possono essere semplicemente eliminati, ma ciò è permesso solo quando si può assumere che questi valori derivino da errori nelle misurazioni; in medicina valori estremi, quando non dovuti ad errori, hanno spesso significati diagnostici, per cui bisogna ricorrere ad un altro modello.

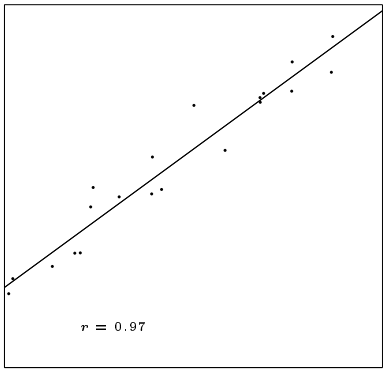
L'uso indiscriminato del coefficiente di correlazione viene spesso e giustamente criticato. J. Carroll chiama il coefficiente di correlazione

„one of the most frequently used tools of psychometricians ... and perhaps also one of the most frequently misused“

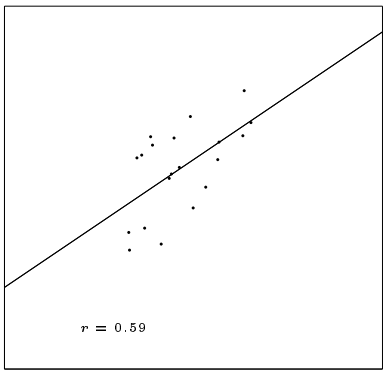
(citato in Rodgers/Nicewander, 61), e Arak Mathai, un famoso esperto di probabilità geometrica, è dell'opinione che il nome *coefficiente di correlazione* non dovrebbe essere più utilizzato, come risulta dalla recensione di uno dei suoi lavori sullo *Zentralblatt*:

„One of the most widely used concepts in statistical literature is the concept of correlation. In applied areas this correlation is interpreted as measuring relationship between variables. This article examines the structure of the expression defining correlation and shows that this concept cannot be meaningfully used to measure relationship or lack of it, or linearity or nonlinearity or independence or association or any such thing, and recommends that this misnomer correlation be replaced with something else in statistical literature.“

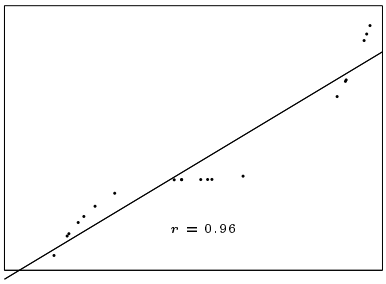
Esempi commentati



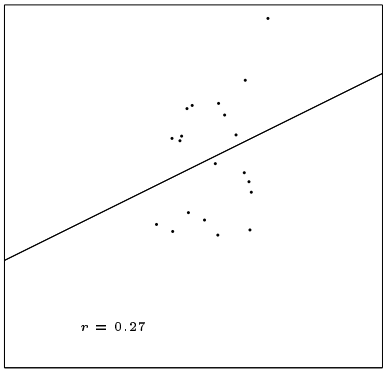
In questo caso y sembra veramente dipendere in modo lineare da x; la retta di regressione può essere utilizzata correttamente come legge che lega le due variabili.

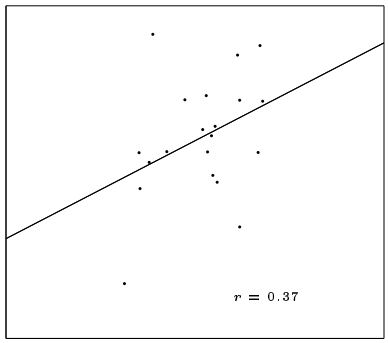


Questo caso è simile al precedente con il coefficiente di correlazione che esprime correttamente il più debole legame rispetto al caso precedente.

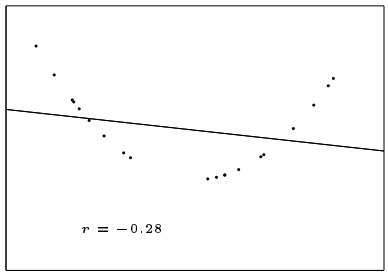


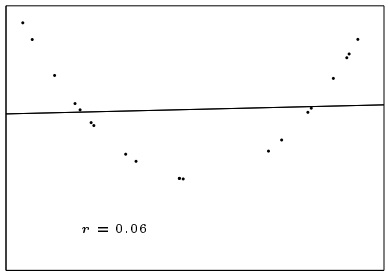
Nonostante il coefficiente di correlazione sia uguale a 0.96, il legame sembra sinusoidale piuttosto che lineare e quindi è più appropriato un modello nonlineare.



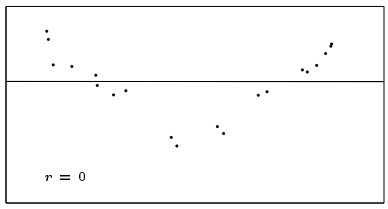


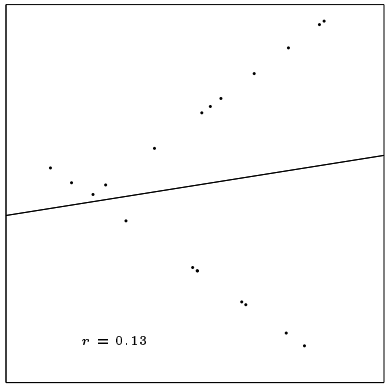
In questi due esempi il legame lineare è molto debole e nella seconda figura si ha l'impressione che la correlazione maggiore sia dovuta più a una certa simmetria e concentrazione al centro che a una dipendenza di y da x.



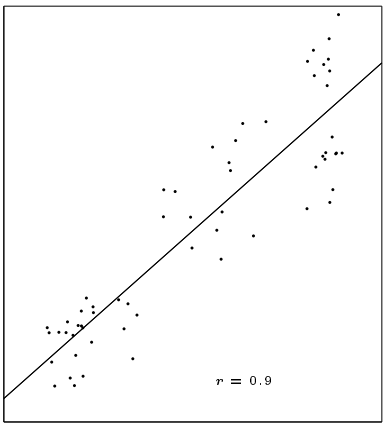


La dipendenza funzionale di tipo quadratico è evidente; il coefficiente di correlazione è vicino a 0; cfr. pagina 12. Infatti il coefficiente di correlazione misura solo la dipendenza *lineare* tra le due variabili.

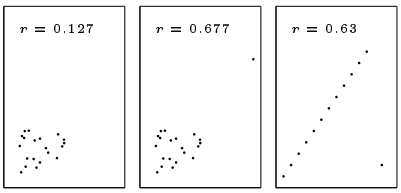




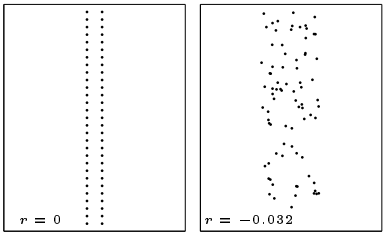
Nonostante che il coefficiente di correlazione sia molto vicino a zero, si notano in ciascuna delle ultime due figure due gruppi che esprimono una dipendenza lineare piuttosto spiccata di y da x. Questa situazione è tipica per dati non omogenei.



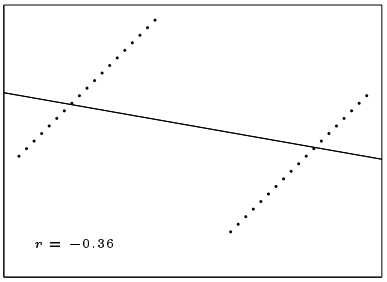
Anche questo è un caso di disomogeneità dei dati in cui però i tre gruppi distinti producono insieme un coefficiente di correlazione alto, benché all'interno di ogni gruppo la dipendenza lineare è piuttosto debole.



Si vede il forte effetto di un singolo valore eccezionale sul coefficiente di correlazione; persino nella seconda figura il coefficiente di correlazione è maggiore di quello nella terza!



Queste configurazioni illustrano un'altra volta quanto detto nella nota 12.13 riguardo al caso in cui i punti sono (almeno approssimativamente) simmetrici rispetto a una retta parallela all'asse delle y.



La correlazione totale è negativa, benché ogni gruppo presenti al suo interno una forte correlazione positiva.

Il quartetto di Anscombe

Esempi particolarmente impressivi sono stati costruiti da Francis Anscombe (citato in Bahrenberg/, 199-200). Consideriamo le seguenti serie di dati, noti nella letteratura come *quartetto di Anscombe*:

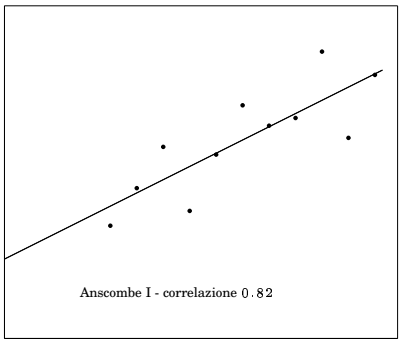
x_{I-III}	y_I	y_{II}	y_{III}
10.0	8.04	9.14	7.46
8.0	6.95	8.14	6.77
13.0	7.58	8.74	12.74
9.0	8.81	8.77	7.11
11.0	8.33	9.26	7.81
14.0	9.96	8.10	8.84
6.0	7.24	6.13	6.08
4.0	4.26	3.10	5.39
12.0	10.84	9.13	8.15
7.0	4.82	7.26	6.42
5.0	5.68	4.74	5.73

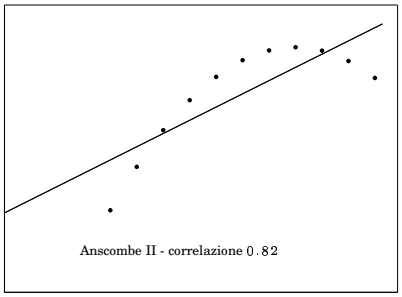
x_{IV}	y_{IV}
8.0	6.58
8.0	5.76
8.0	7.71
8.0	8.84
8.0	8.47
8.0	7.04
8.0	5.25
19.0	12.50
8.0	5.56
8.0	7.91
8.0	6.89

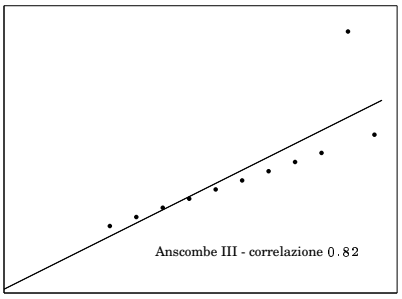
Questi dati hanno in comune le seguenti caratteristiche:

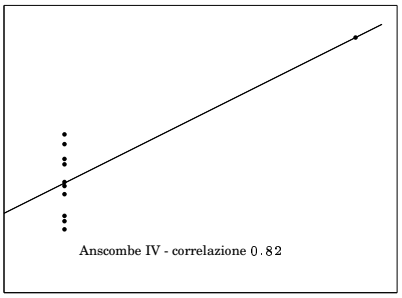
$n = 11$;
$\overline{x} = 9, \overline{y} = 7.5$;
retta di regressione $\eta = 0.5\xi + 0.3$;
coefficiente di correlazione $r = 0.82$.

Nonostante ciò le figure mostrano relazioni di dipendenza completamente diverse.









Solo nel primo caso l'analisi regressionale lineare può essere applicata. Gli esempi fanno vedere chiaramente che i valori numerici dei parametri statistici non sono sufficienti per una corretta interpretazione statistica che deve affiancata dalla rappresentazione grafica e uno studio il più dettagliato possibile dei meccanismi interni da cui i dati derivano.

Che nonostante le critiche, con un uso ragionato del coefficiente di correlazione si possono ottenere anche rappresentazioni molto convincenti di legami statistici, lo mostrano i grafici alle pagine 188-189 del libro di Bahrenberg/, in cui sono illustrate le correlazioni tra le diverse zone climatiche della Germania.

Correlazione parziale

Nota 14.1. Talvolta una correlazione tra x ed y è riconducibile alla correlazione di entrambe le variabili con una terza variabile; per studiare questi influssi si introduce la *correlazione parziale*. D'altra parte, anche l'interpretazione della correlazione parziale pone problemi e si basa su ipotesi che spesso non sono facilmente identificabili. Si dovrebbe quindi fare un uso molto cauto del coefficiente di correlazione parziale; in particolare si dovrebbe sempre poter assumere che x ed y dipendono in modo lineare dalla terza variabile. Cfr. Linder/, 38-43.

Definizione 14.2. Sia $u \in \mathbb{R}^n$ un terzo vettore non diagonale. Denotiamo con p_x la proiezione ortogonale di x sul piano P_u generato da 1^n ed u e con p_y la proiezione ortogonale di y su P_u. Siano $d_x := x - p_x$ e $d_y := y - p_y$. Assumiamo inoltre che $r^2_{xu} \neq 1$ ed $r^2_{yu} \neq 1$. Per il corollario 11.11 allora d_x e d_y sono $\neq 0$; questi vettori, essendo ortogonali a 1^n, sono allora anche non diagonali. Definiamo in queste ipotesi la *correlazione parziale* $r_{xy|u}$ di x ed y rispetto ad u tramite

$$r_{xy|u} := r_{d_x d_y}$$

Nota 14.3. Nelle ipotesi della definizione 14.2, i vettori d_x e d_y sono entrambi ortogonali al piano P_u. Per $n \leq 3$ ciò implica che sono paralleli, ma ciò non è più vero per $n \geq 4$; infatti il complemento ortogonale di un piano in $\mathbb{R}^n$ ha dimensione $n - 2$.

Proposizione 14.4. *Nelle ipotesi della definizione 14.2 si ha*

$$r_{xy|u} = \frac{r_{xy} - r_{xu} r_{yu}}{\sqrt{1 - r^2_{xu}}\sqrt{1 - r^2_{yu}}}$$

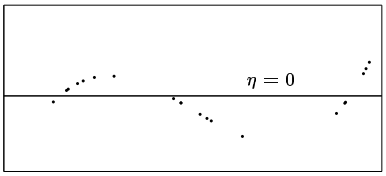
Dimostrazione. Paset, 32-35, dove si trova anche un'interpretazione di questa formula, che talvolta viene posta come definizione della correlazione parziale, nell'ambito della trigonometria sferica.

La funzione Sreg.residui

Definiamo una funzione per il calcolo del vettore dei residui d:

```
Sreg.residui = function (x,y)
{u=Sreg.retta(x,y)
y-(u[1]*x+u[2])}
```

Applichiamo questa funzione al terzo esempio a pagina 13:



Bibliografia

F. Anscombe: Graphs in statistical analysis. Am. Statistician 27 (1973), 17-21.

15142 **G. Bahrenberg/E. Giese/J. Nipper:** Statistische Methoden in der Geographie I. Teubner 1999.

J. Carroll: The nature of the data, or how to choose a correlation coefficient. Am. Statistician 38 (1984), 58-60.

14030 **H. Fassl:** Einführung in die medizinische Statistik. Barth 1999.

5221 **A. Linder/W. Berchtold:** Statistische Methoden III. Multivariate Verfahren. Birkhäuser 1982.

A. Mathai: The concept of correlation and misinterpretations. Int. J. Math. Stat. Sci. 7/2 (1998), ...

A. Mathai: On Pearson's statistic for goodness of fit. Int. J. Math. Stat. Sci. 7/2 (1998), ...

16226 **F. Paset:** Regressione, correlazione e analisi delle componenti principali. Tesi Univ. Ferrara 2003.

15908 **J. Rodgers/W. Nicewander:** Thirteen ways to look at the correlation coefficient. Am. Statistician 42/1 (1988), 59-66.