

Correlazione e regressione lineare

Tecniche per analizzare la relazione tra 2 o più variabili continue

Correlazione: associazione lineare tra 2 variabili. La forza dell'associazione è data dal coefficiente di correlazione

Regressione: dipendenza di una variabile (dipendente) da un'altra variabile (indipendente)

Correlazione

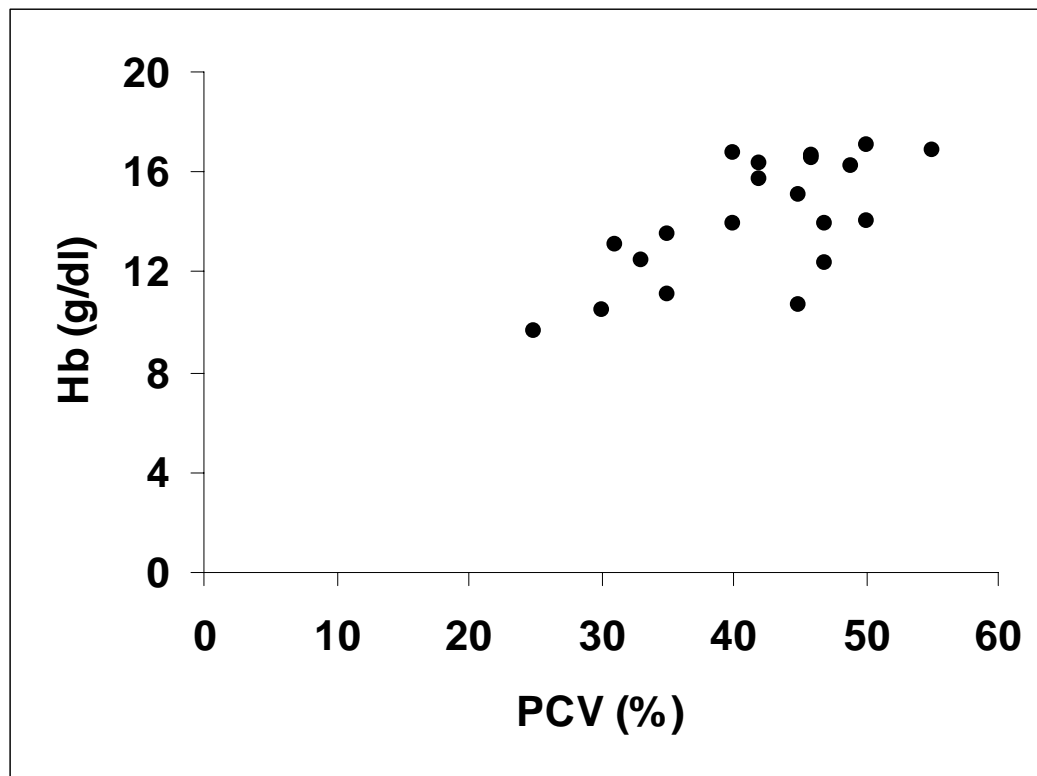
Risponde alla domanda:

esiste un'associazione lineare tra le variabili?

Esempio (Campbell et al. 1985): gruppo di donne di una determinata area geografica invitate a sottoporsi a un prelievo di sangue per la determinazione del livello di emoglobina (Hb) e dell'ematocrito (PCV). Si registra anche l'età e stato di menopausa (sì/no). Percentuale di adesione: circa il 90%.

Risultati da un campione randomizzato di 20 donne facenti parte del gruppo:

ID	Hb (g/dl)	PCV (%)	Età (anni)	Menopausa (0=no; 1=sì)
1	11,1	35	20	0
2	10,7	45	22	0
3	12,4	47	25	0
4	14,0	50	28	0
5	13,1	31	28	0
6	10,5	30	31	0
7	9,6	25	32	0
8	12,5	33	35	0
9	13,5	35	38	0
10	13,9	40	40	0
11	15,1	45	45	1
12	13,9	47	49	0
13	16,2	49	54	1
14	16,3	42	55	1
15	16,8	40	57	1
16	17,1	50	60	1
17	16,6	46	62	1
18	16,9	55	63	1
19	15,7	42	65	1
20	16,5	46	67	1



Vogliamo analizzare la relazione fra Hb e PCV.

Non ci chiediamo se Hb influenza PCV o PCV influenza Hb o se un alto valore di PVC causa un alto valore di Hb, ma se le due variabili sono associate.

Il coefficiente di correlazione del campione r ci permette di:

- riassumere la forza della relazione lineare fra le variabili
- verificare l'ipotesi che r sia zero, cioè se l'apparente associazione fra le variabili possa essere dovuta al caso

Coefficiente di correlazione di Pearson:

Dato un insieme di osservazioni appaiate $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$,

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}},$$

dove

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

sono le medie campionarie.

Ricordiamo che:

$$\sum_{i=1}^5 x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

ID	Hb (g/dl)
1	11,1
2	10,7
3	12,4
4	14,0
5	13,1

x: valori di Hb

$$\sum_{i=1}^5 x_i = 11,1 + 10,7 + 12,4 + 14,0 + 13,1 = 61,3$$

Esempio: Hb = x , PCV = y $\bar{x} = 14,12$ $\bar{y} = 41,65$

x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$
11,1	35	-3,02	-6,65	20,083	9,12	44,2225
10,7	45	-3,42	3,35	-11,457	11,70	11,2225
12,4	47	-1,72	5,35	-9,202	2,96	28,6225
14,0	50	-0,12	8,35	-1,002	0,01	69,7225
13,1	31	-1,02	-10,65	10,863	1,04	113,4225
10,5	30	-3,62	-11,65	42,173	13,10	135,7225
9,6	25	-4,52	-16,65	75,258	20,43	277,2225
12,5	33	-1,62	-8,65	14,013	2,62	74,8225
13,5	35	-0,62	-6,65	4,123	0,38	44,2225
13,9	40	-0,22	-1,65	0,363	0,05	2,7225
15,1	45	0,98	3,35	3,283	0,96	11,2225
13,9	47	-0,22	5,35	-1,177	0,05	28,6225
16,2	49	2,08	7,35	15,288	4,33	54,0225
16,3	42	2,18	0,35	0,763	4,75	0,1225
16,8	40	2,68	-1,65	-4,422	7,18	2,7225
17,1	50	2,98	8,35	24,883	8,88	69,7225
16,6	46	2,48	4,35	10,788	6,15	18,9225
16,9	55	2,78	13,35	37,113	7,73	178,2225
15,7	42	1,58	0,35	0,553	2,50	0,1225

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = 242,64 \quad \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2} = 360,33 \quad r = \frac{242,64}{360,33} = 0,673$$

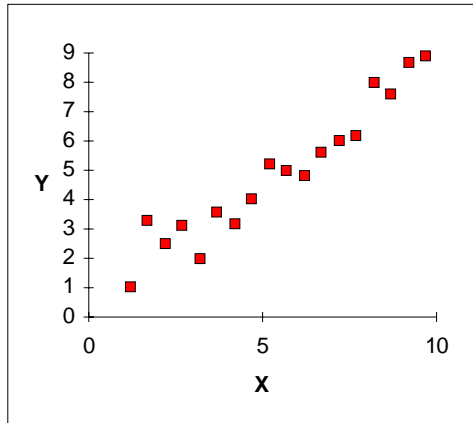
Il coefficiente di correlazione del campione r

- è una quantità a-dimensionale
 - varia da -1 a 1 ($r=1$ o $r=-1$: correlazione lineare esatta)
 - è positivo quando i valori delle variabili crescono insieme
 - è negativo quando i valori di una variabile crescono al decrescere dei valori dell'altra
 - non è influenzato dalle unità di misura
-
- r^2 : proporzione di variazione di una variabile che è "spiegata" dall'altra:

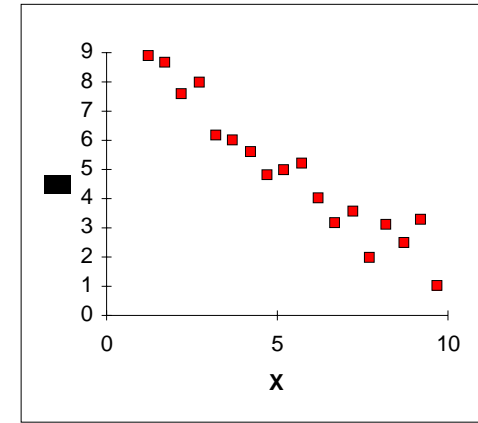
$r=0.9$ $r^2=0.81$ \Rightarrow circa l'80% della variazione di una variabile è spiegato dall'altra

Nell'esempio: $r=0.67$, $r^2=0.45$ \Rightarrow il 45% della variazione di Hb è spiegato dai valori di PCV

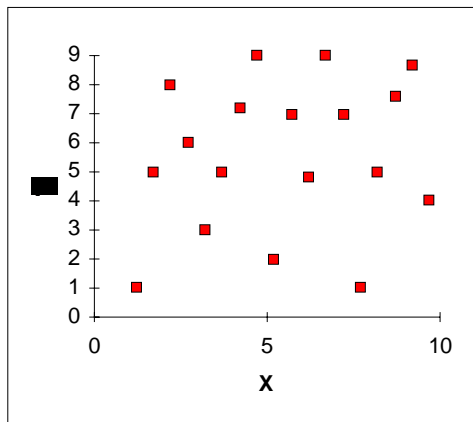
Esempi:



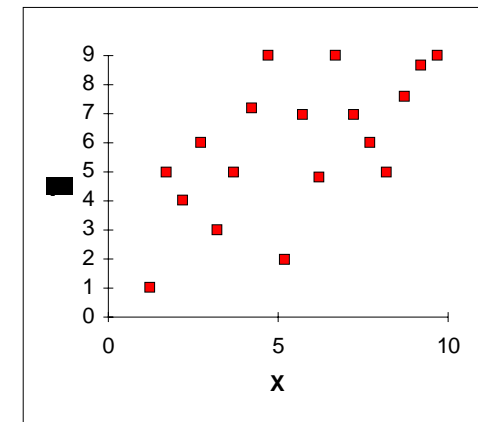
$r=0,96$



$r=-0,96$



$r=0,12$



$r=0,62$

$r=1$: punti perfettamente allineati su una retta con pendenza positiva

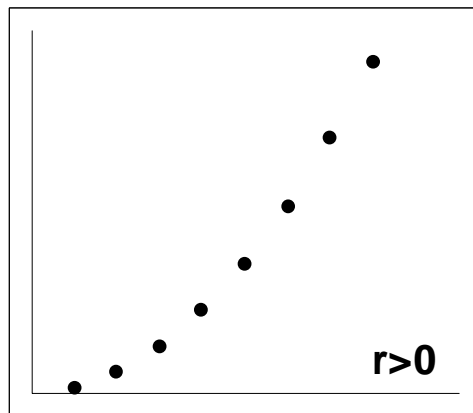
$r=-1$: punti perfettamente allineati su una retta con pendenza negativa

Il coefficiente di correlazione non è una buona misura di associazione fra due variabili quando:

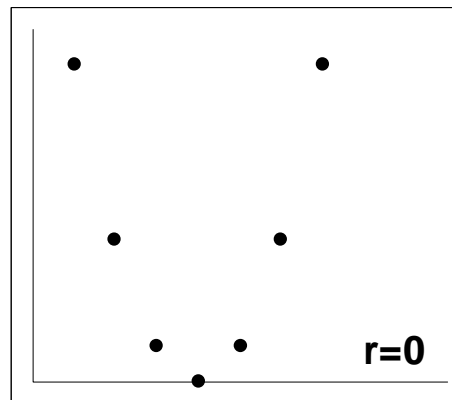
- la relazione fra le variabili è non-lineare
- in presenza di valori estremi

Esempi:

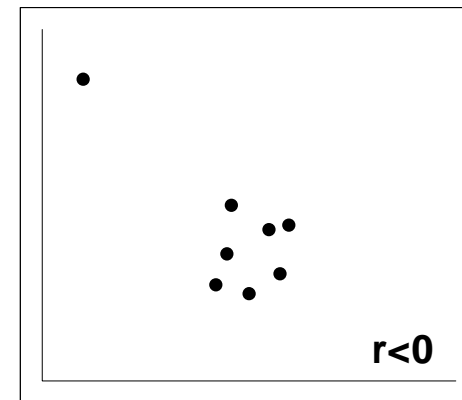
1.



2.



3.



1. relazione quadratica: $y = ax^2 + bx + c$
2. x e y sono fortemente associate, tuttavia $r=0$. E.g.: relazione fra mortalità globale della popolazione ed obesità
3. uno dei valori è molto distante dal gruppo principale degli altri valori e influenza fortemente il valore stimato di r: poiché è così estremo deriva probabilmente da una popolazione diversa. E.g.: studio sulla perdita di sangue e valori basali di Hb prima dell'inserimento di uno IUD. Valore estremo: donna affetta da malattia che causa importanti perdite ematiche e le anemizza

Il coefficiente di correlazione va utilizzato con cautela quando

- le variabili sono misurate da più di un gruppo distinto, e.g. pazienti affetti da una malattia e controlli sani: possono generare due gruppi di punti, ciascuno dei quali con $r=0$ ma $r \neq 0$ una volta combinati (effetto simile al caso che contiene un valore estremo)
- una delle due variabili è fissata a priori, e.g. quando si misura la risposta a dosi diverse di un farmaco. In questo caso la scelta di un particolare dosaggio può influenzare il coefficiente di correlazione, anche se la relazione dose-risposta è fissa

IMPORTANTE:

Un'elevata correlazione fra due variabili NON implica una relazione causa-effetto

I test di correlazione sono fra le procedure statistiche peggio utilizzate. Sono in grado di dimostrare se due variabili sono correlate, tuttavia NON sono in grado di dimostrare che due variabili NON sono correlate! Se una variabile dipende da un'altra, e se vi è una relazione causale, è sempre possibile trovare una qualche forma di correlazione fra le due. Ma se entrambe le variabili dipendono da una terza variabile, potremmo trovare correlazione fra le due variabili di partenza anche se fra di esse non vi fosse nessuna dipendenza causale.

Esempio: è stata trovata una correlazione fra il numero di ripetitori di telefoni cellulari e la diminuzione del numero dei passerotti. Domanda: sono i ripetitori a danneggiare i passerotti oppure entrambi gli effetti sono causati da qualcos'altro? Oppure sono osservazioni completamente indipendenti che per caso appaiono correlate?

Non lo sappiamo, i test di correlazione non rispondono a questa domanda e sono necessari altri studi.

Test di significatività (1)

Osservato il diagramma di dispersione dei valori delle due variabili e calcolato il coefficiente di dispersione si deve decidere se la correlazione osservata possa essere frutto del caso (spuria)

Cerchiamo la probabilità di ottenere un coefficiente di correlazione pari o più estremo del valore osservato r , posto che l'ipotesi nulla sia vera ($H_0: \rho=0$)

Calcoliamo $t = \frac{r - 0}{es(r)}$, dove l'errore standard stimato di r è dato da

$$es(r) = \sqrt{(1 - r^2)/(n - 2)}$$

Se le coppie di valori (x_i, y_i) sono state scelte casualmente e le due variabili x e y sono distribuite normalmente, t è distribuita come una variabile t di Student con $n-2$ gradi di libertà solo quando H_0 è vera

Test di significatività (2)

Nell'esempio: $n=20$, $r=0,67 \Rightarrow t=3,83$

Eseguiamo un test a due code dell'ipotesi nulla di assenza di associazione ad un livello di significatività $\alpha=0,05$

Per una distribuzione t di Student con 18 gradi di libertà, la probabilità di osservare un valore maggiore di 3,83 è 0,0006. Posto che il vero valore del coefficiente di correlazione ρ sia 0 (ipotesi nulla), la probabilità di osservare un valore di r tanto lontano da 0 quanto $r=3,83$ è $p=2 \times 0,0006=0,0012$

Rifiutiamo l'ipotesi nulla ad un livello di significatività pari a 0,05: in base a questo campione, c'è evidenza che la correlazione reale nella popolazione sia diversa da 0

Assunzione alla base del test di significatività: entrambe le variabili sono casuali e distribuite normalmente.

E.g.: nel caso in cui siano presenti valori estremi, la variabile non può essere distribuita normalmente e il test di significatività non è più valido.

Coefficiente di correlazione dei ranghi di Spearman r_s :

Si ordinano le due serie di valori delle variabili e si assegnano separatamente i ranghi (ad osservazioni uguali si assegnano ranghi medi), quindi si calcola il coefficiente di correlazione (di Pearson) dei ranghi.

Metodo equivalente:
$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n}$$
 dove d_i è la differenza

di ranghi per l'individuo i -esimo.

E' una misura di associazione più robusta e può essere utilizzata quando una o entrambe le variabili sono ordinali.