Regressione e correlazione

Regressione e correlazione

In molti casi si osservano grandezze che tendono a covariare, ma…

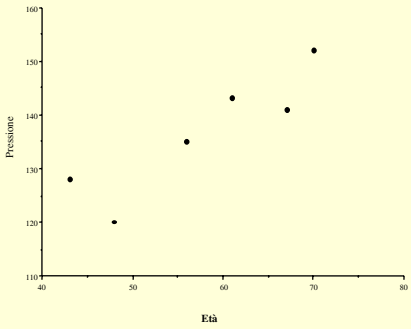
(1) Se c'è una relazione di dipendenza fra due variabili, ovvero se il valore di una variabile (dipendente) si può determinare come funzione di una seconda variabile (indipendente), allora si può usare una **regressione**.

Esempio: la pressione arteriosa dipende dall'età del soggetto

(2) Se non c'è una relazione di dipendenza fra le variabili, ovvero se nessuna delle due è causa delle variazioni dell'altra, la tendenza a covariare si misura in termini di **correlazione**.

Esempio: lunghezza e peso di un organismo

Età	Pressione
43	128
48	120
56	135
61	143
67	141
70	152

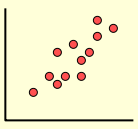


Per misurare l'intensità di una relazione (lineare) si usa il coefficiente di correlazione di Bravais-Pearson.

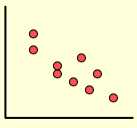
Per un campione: r

r e ρ variano fra +1 e -1

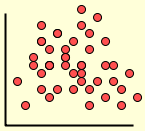
Per una popolazione: ρ (rho)



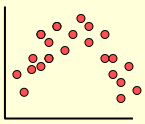
Proporzionalità diretta: r tende a +1



Proporzionalità inversa: r tende a -1



Nessuna relazione: r tende a 0



Nessuna relazione <u>lineare</u>: r tende a 0

$$r = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{[n(\Sigma X^2) - (\Sigma X)^2][n(\Sigma Y^2) - (\Sigma Y)^2]}}$$

Per l'esempio sulla pressione arteriosa:

Soggetto	Età(X)	PA(Y)	XY	X^2	Y^2
A	43	128	5504	1849	16384
B	48	120	…	…	…
C	56	135	…	…	…
D	61	143	…	…	…
E	67	141	…	…	…
F	70	152	…	…	…
	$\Sigma X=345$	$\Sigma Y=819$	$\Sigma XY=47634$	$\Sigma X^2=20399$	$\Sigma Y^2=112443$

r = .897

Cioè: forte relazione positiva

Se r = .897 indica una forte relazione positiva, si può affermare che questa relazione non è frutto del caso ed è quindi significativa?

Ipotesi da testare per la significatività di una correlazione:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

$$t = r\sqrt{\frac{N - 2}{1 - r_2}} = 4.059$$

$$t_{crit(.05,\ df=1)} = 2.776$$

Poichè t=4.059>2.776, si rigetta H_0 e si conclude che esiste una correlazione positiva e significativa fra età e pressione arteriosa.

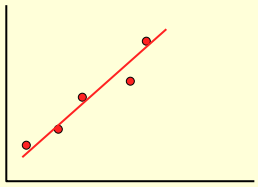
Attenzione!

Una correlazione positiva e significativa **non** implica un rapporto causale.

Regressione lineare

Analizza la natura e l'intensità di una relazione lineare fra due variabili, di cui una dipende dall'altra (o almeno una è misurata senza errore).

Interpoliamo una retta…

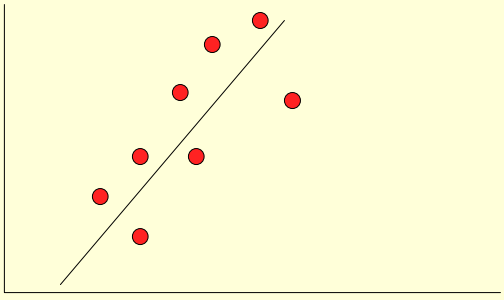


Una retta qualsiasi è descritta dall'equazione:

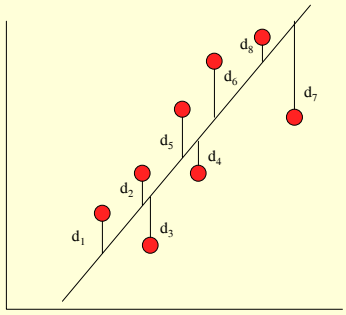
$Y = a + bX$ (per un campione)

$Y = \alpha + \beta X$ (per una popolazione)

Per determinare la retta che meglio si adatta ai dati, si usa il metodo dei ***minimi quadrati***.



Per determinare la retta che meglio si adatta ai dati, si usa il metodo dei ***minimi quadrati***.



Si calcola la distanza di ogni punto dalla retta nello spazio della variabile dipendente (Y)

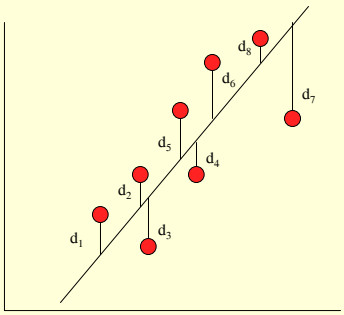
La somma

$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 \ldots . d_8^2$$

ovvero

$$\Sigma[Y - f(\overline{X})]^2$$

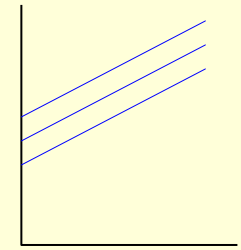
deve essere minimizzata



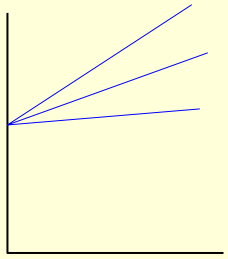
(N.B. Questa somma è una componente della somma dei quadrati – e quindi della varianza – della variabile Y)

Nell'equazione $Y = a + bX$,

a è l'intercetta sull'asse Y
b è la pendenza della retta o coefficiente di regressione



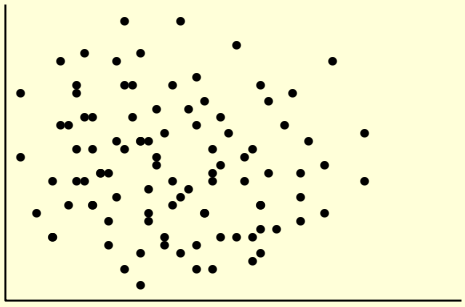
stessa b - differente a



stessa a - differente b

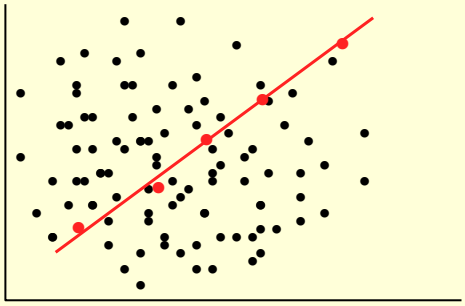
Esiste una retta di regressione per qualsiasi insieme di dati.

Immaginiamo una popolazione di dati per cui $\beta = 0$ …



Esiste una retta di regressione per qualsiasi insieme di dati.

Immaginiamo una popolazione di dati per cui $\beta = 0$ …



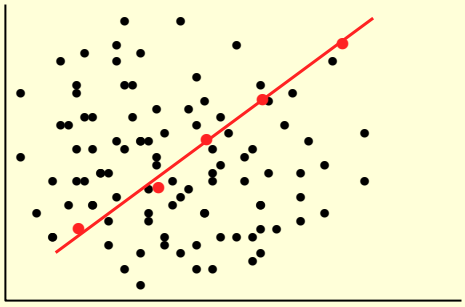
Se un campione casuale comprendesse i punti (•),

la retta $Y = a + bX$ che si interpolerebbe avrebbe $b \neq 0$

Qual'è la probabilità che l'insieme di punti in rosso sia stato estratto dalla popolazione studiata e che esso descriva accuratamente la relazione fra X e Y?

Definiamo l'ipotesi nulla e l'ipotesi alternativa:

H_0: $\beta = 0$
H_1: $\beta \neq 0$



Quindi usiamo un'ANOVA

1) Si calcola la somma dei quadrati ovvero la variabilità complessiva di Y

$$SS_T = \Sigma(Y_i - \overline{Y})^2$$

2) Si calcola la somma dei quadrati per la regressione (cioè per il modello usato)

$$SS_R = \frac{\left[\Sigma X_i Y_i - \frac{\Sigma X_i \Sigma Y_i}{n}\right]^2}{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}}$$

3) Si calcola la somma dei quadrati per i residui (scarti dalla regressione)

$$SS_D = SS_T - SS_R$$

4) Si calcolano i quadrati medi per la regressione e per i residui

$MS_x = SS_x/df_x$ dove $df_T = n-1$, $df_R = 1$, $df_D = df_T - df_R$

5) Si determina F: $F = MS_R/MS_D$

6) Si determina il valore di p corrispondente

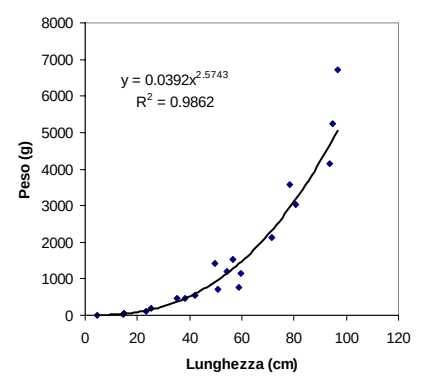
7) Il coefficiente di determinazione $r^2 = SS_R/SS_D$ è la proporzione di varianza totale spiegata dalla regressione

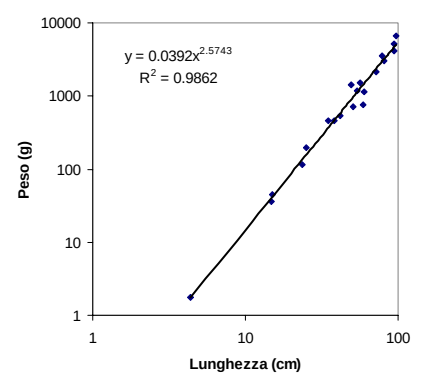
Relazioni non lineari

- Se una retta non descrive la relazione fra due variabili, si deve usare una funzione non lineare
- Spesso a questo fine si usano delle trasformazioni non lineari dei dati, per esempio in logaritmo
- Un caso tipico è quello di una relazione lineare fra i logaritmi delle due variabili, tale che la curva che si interpola è:

  $Y = a X^b$ [cioè $\log(Y)=a+b \log(X)$]

- Esempio: relazioni peso-lunghezza in pesci





Una relazione peso-lunghezza si descrive con la regressione lineare log-log, ovvero con una funzione di potenza $Y=aX^b$

Domanda #1: il peso dipende dalla lunghezza?

Domanda #2: se accettiamo di usare la lunghezza come variabile indipendente (è più facile da misurare), possiamo affermare che l'errore di misura della lunghezza è nullo?

Domanda #3: possiamo affermare che l'errore di misura della lunghezza è << di quello del peso?

Il peso non dipende dalla lunghezza (e viceversa).

Cosa sappiamo:

- sono grandezze che covariano
- quindi i valori dell'una possono essere utili per stimare i valori dell'altra
- entrambe le misure sono affette da errore
- l'ordine di grandezza dell'errore nella stima della lunghezza (assunta come variabile indipendente) può variare in funzione del metodo di misura e degli organismi da misurare

Il peso non dipende dalla lunghezza (e viceversa).

Quindi, la regressione lineare **non** è un metodo adatto a descrivere questa relazione, a meno che l'errore di misura della lunghezza non sia << di quello del peso.

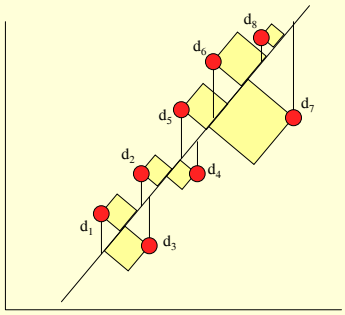
Asse Maggiore e Asse Maggiore Ridotto

- Regola empirica: se la varianza delle X è >1/3 di quella delle Y, non si dovrebbe usare la regressione lineare
- L'Asse Maggiore considera sia l'errore della X che quello della Y: è la bisettrice dell'angolo formato della retta di regressione della X sulla Y con quella di regressione della Y sulla X.
- L'Asse Maggiore Ridotto è quasi concidente con l'Asse Maggiore, ma è più semplice da ottenere.

Asse Maggiore

- Si minimizza la somma dei quadrati delle proiezioni dei punti sull'Asse Maggiore
- Il calcolo implica:
  - Estrazione di autovalori ed autovettori dalla matrice di covarianza

    oppure
  - Calcolo delle regressioni Y su X e X su Y e della bisettrice delle due rette



Asse maggiore

Asse Maggiore Ridotto

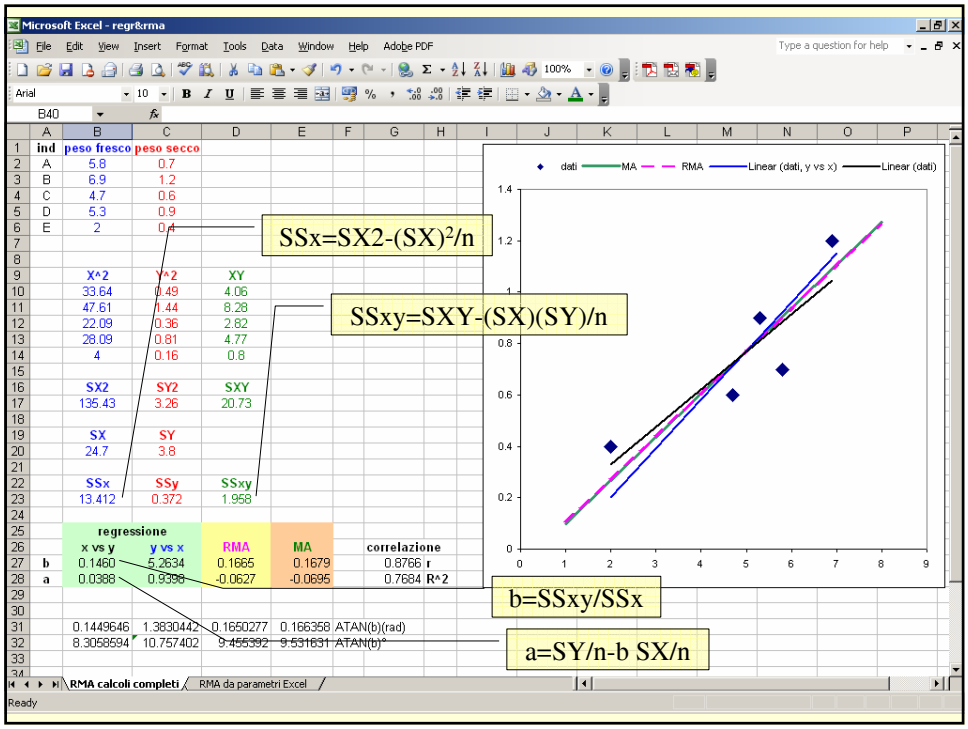
- In pratica, quasi coincide con l'Asse Maggiore
- Il calcolo implica:
  - Calcolo delle regressioni Y su X e X su Y e quindi

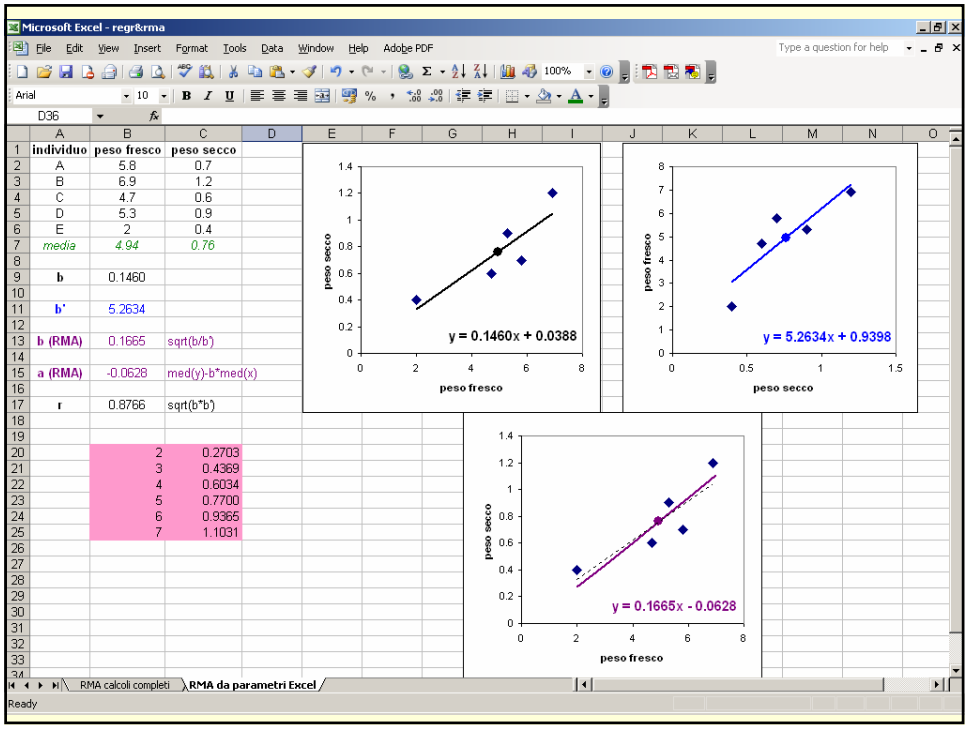
$$b_{RMA} = \sqrt{b_{Y=f(X)} / b_{X=f(Y)}}$$

  - Calcolo delle somme dei quadrati SS_X e SS_Y o delle varianze

$$b_{RMA} = \sqrt{SS_Y / SS_X} = \sqrt{s_Y^2 / s_X^2}$$

  - In ogni caso: $a_{RMA} = \overline{Y} - b_{RMA}\overline{X}$





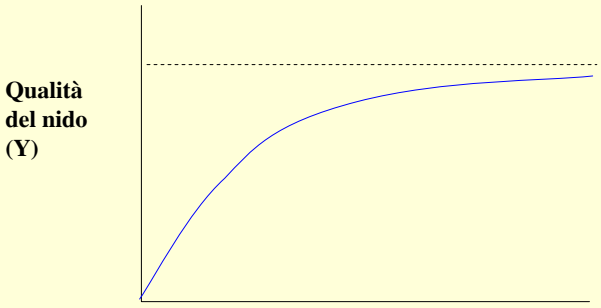
	A	B	C
1	individuo	peso fresco	peso secco
2	A	5.8	0.7
3	B	6.9	1.2
4	C	4.7	0.6
5	D	5.3	0.9
6	E	2	0.4
7	*media*	*4.94*	*0.76*
8			
9	b	0.1460	
10			
11	b'	5.2634	
12			
13	b (RMA)	0.1665	sqrt(b/b')
14			
15	a (RMA)	-0.0628	med(y)-b*med(x)
16			
17	r	0.8766	sqrt(b*b')
18			
19			
20		2	0.2703
21		3	0.4369
22		4	0.6034
23		5	0.7700
24		6	0.9365
25		7	1.1031

Dati ordinali e relazioni monotoniche: la correlazione di rango di Spearman

Esperimento:
valutare la relazione fra qualità dei nidi costruiti e tempo di apprendimento

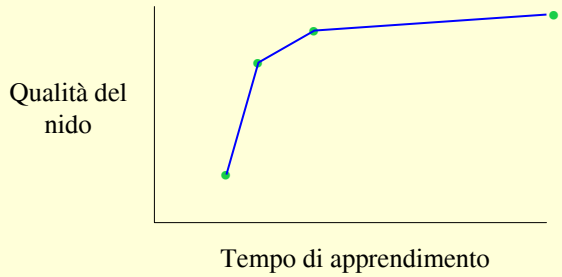
Cosa dobbiamo attenderci?

- una relazione non lineare (l'apprendimento non consente di migliorare all'infinito)
- una relazione monotonica (con l'apprendimento la qualità dei nidi non può peggiorare)



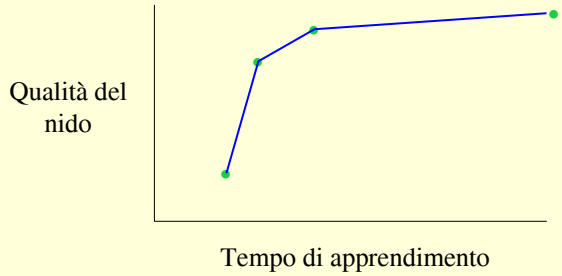
Dati (fittizi) :

Uccello	Tempo di apprendimento	Qualità del nido
A	4	9
B	2	2
C	10	10
D	3	8



Assegnamo dei ranghi ai dati :

Uccello	Tempo di apprendimento	Qualità del nido
A	4 → 3	9 → 3
B	2 → 1	2 → 1
C	10 → 4	10 → 4
D	3 → 2	8 → 2



Calcolo della correlazione di Spearman (metodo di base)

1) Si assegnano i ranghi ai valori di X e Y

2) Si calcola il coefficiente di Bravais-Pearsono sui dati trasformati

Uccello	Tempo di apprendimento	Qualità del nido	XY
A	3	3	9
B	1	1	1
C	4	4	16
D	2	2	4

$\Sigma X = 10 \quad \Sigma X^2 = 31 \quad \Sigma XY = 31$

$SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 6$

Analogamente, $SS_Y = 6$ e $SP = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n} = 6$

Quindi $r_s = \frac{SP}{\sqrt{(SS_X)(SS_Y)}} = 1.0$

Calcolo di r_S dai ranghi

Se non ci sono ranghi assegnati ex-aequo, il calcolo può essere semplificato, essendo:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

dove *d* è la differenza fra il rango della *i*-ma osservazione per il descrittore *j* e quello per il descrittore *k*.

Calcolo di r_S dai ranghi

Se ci sono (molti) ranghi assegnati ex-aequo, il calcolo deve essere corretto come segue:

$$r_S = \frac{2n^3 - 2n - \sum_{h=1}^{m}(q_{hj}^3 - q_{hj}) - \sum_{h=1}^{m}(q_{hk}^3 - q_{hk}) - 12\sum_{i=1}^{n} d_i^2}{2 \cdot \sqrt{\cdot n^3 - n - \sum_{h=1}^{m}(q_{hj}^3 - q_{hj})} \cdot \sqrt{n^3 - n - \sum_{h=1}^{m}(q_{hk}^3 - q_{hk})}}$$

dove *m* è il numero di ranghi e q_{hj} e q_{hk} sono il numero di osservazioni di rango *h* per il descrittore *j* e per quello *k*