

Capitolo 4. Regressione e Correlazione.

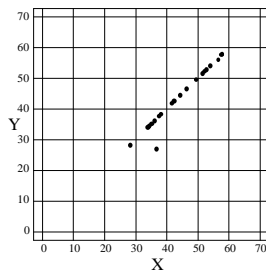
Regressione

Il termine regressione ha un'origine antica ed un significato molto particolare. L'inventore è un certo F.Galton, genetista, che nel 1889 pubblicò un articolo in cui dimostrava come “ogni caratteristica di un individuo è ereditata dalla prole, ma in media ad un livello minore”. Ad esempio, i figli di un genitore di statura alta sono anch'essi alti, ma in media sono meno alti del genitore. Tale fenomeno, descritto anche graficamente, fu chiamato regressione e da allora tale termine è rimasto per definire quelle tecniche statistiche che analizzano la relazione tra due o più variabili.

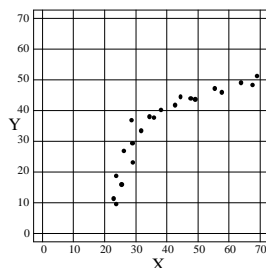
Per analizzare la relazione tra due variabili occorre:

- (1) assumere un modello di relazione (lineare o anche non lineare)
 - (2) valutare i parametri del modello e la loro variabilità
 - (3) in base ai dati ottenuti, verificare la validità del modello assunto inizialmente
- Ciò consente di fare stime e previsioni di notevole interesse scientifico.

L'analisi della regressione si appoggia molto alla rappresentazione grafica. Le due variabili sono rappresentate dagli assi di un sistema cartesiano. Le osservazioni sono rappresentate dai punti:



Il modello di relazione che per ora consideriamo è quello lineare. Vale a dire che la variazione tra la variabile X e la variabile Y è rappresentabile da una retta. Ovviamente, situazioni come la seguente non sono rappresentabili da una retta e per queste occorrerà trovare in seguito delle soluzioni alternative.

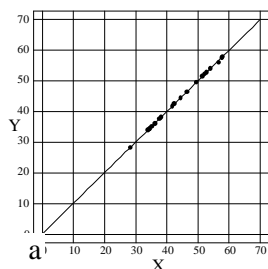


Poiché la retta è definita dall'equazione:

$$y = a + bx$$

occorre innanzitutto trovare i valori dei parametri **a** e **b** che meglio adattano la retta al modello (talvolta l'equazione è scritta con altri simboli, come per es.: $y = b_0 + b_1x$).

Il parametro **a** è detto **intercetta** (sull'asse Y) mentre il parametro **b** è detto **pendenza** o fattore angolare o, internazionalmente, slope.



Quando, come nel grafico, l'intercetta è zero, la retta passa per l'origine e l'equazione si semplifica:

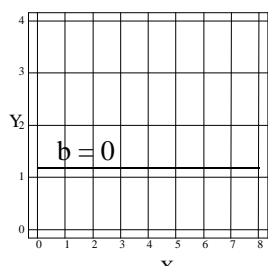
$$y = bx$$

La relazione esprime in tal caso (ma solo in tal caso) un rapporto di proporzionalità. Si può infatti scrivere:

$$\frac{y}{x} = b$$

Attenzione: c'è proporzionalità se e solo se il parametro **a** è pari a zero. In altre parole, se la retta non passa per l'origine, la relazione può essere lineare ma non può esservi proporzionalità. Ad es., non esiste proporzionalità tra gradi centigradi e gradi Fahrenheit. Per questo motivo non è lecito, in ogni caso ed a qualsiasi titolo, elaborare rapporti, formulare indici, ecc. tra variabili se prima non si dimostra che esiste proporzionalità, cioè che la retta di regressione passa per l'origine ($a=0$).

Il parametro **b** invece deve essere comunque diverso da zero. Se **b** è zero non sussiste relazione in quanto $y = a$. Graficamente $b = 0$ corrisponde ad una retta orizzontale.



Per calcolare i parametri dell'equazione bisogna introdurre un nuovo parametro statistico: la **codevianza** ($S_{x,y}$) in grado di stimare la variabilità appaiata delle due variabili. La formula è analoga a quella della devianza, in quanto consiste nella sommatoria dei prodotti degli scarti tra i valori x e y e le rispettive medie. [Per ragioni che appariranno evidenti in seguito, d'ora in avanti in questo capitolo indicheremo le medie con la notazione x_{medio} e y_{medio} anziché m_x e m_y].

$$S_{x,y} = \sum (x - x_{\text{medio}})(y - y_{\text{medio}})$$

Ad esempio:

X	Y	$(x - x_{\text{medio}})(y - y_{\text{medio}})$
1	2	$(1-2)(2-4)=-2$
2	3	$(2-2)(3-4)=0$
3	7	$(3-2)(7-4)=3$
		$S_{x,y}=1$
$x_{\text{medio}}=2$	$y_{\text{medio}}=4$	
$S_x=2$	$S_y=14$	

Il parametro **b** si calcola come:

$$b = \frac{S_{x,y}}{S_x}$$

Nell'esempio di sopra, $S_x = 2$, per cui $b = 1/2 = 0.5$.

Per calcolare l'intercetta, basta sapere che la retta deve necessariamente passare per il punto di intersezione delle due medie x_{medio} e y_{medio} . Pertanto, in base all'equazione della retta, possiamo scrivere:

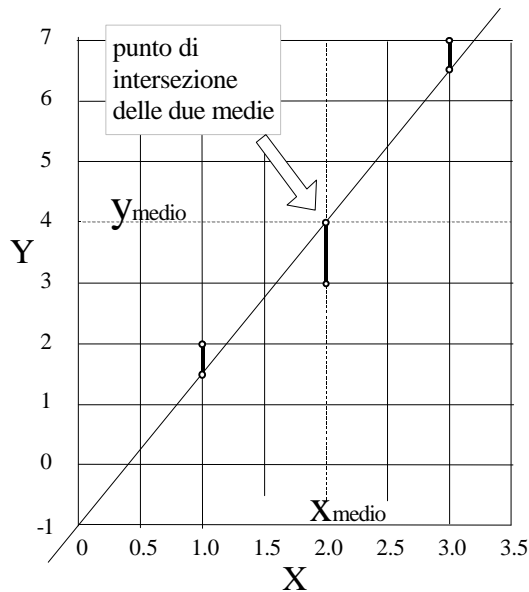
$$y_{\text{medio}} = a + b x_{\text{medio}}$$

da cui si ricava l'intercetta:

$$a = y_{\text{medio}} - b x_{\text{medio}}$$

Nell'esempio riportato sopra, $a = 4 - (0.5 \cdot 2) = 3.5$

L'equazione della retta di regressione sarà quindi: $y = 3.5 + 0.5x$



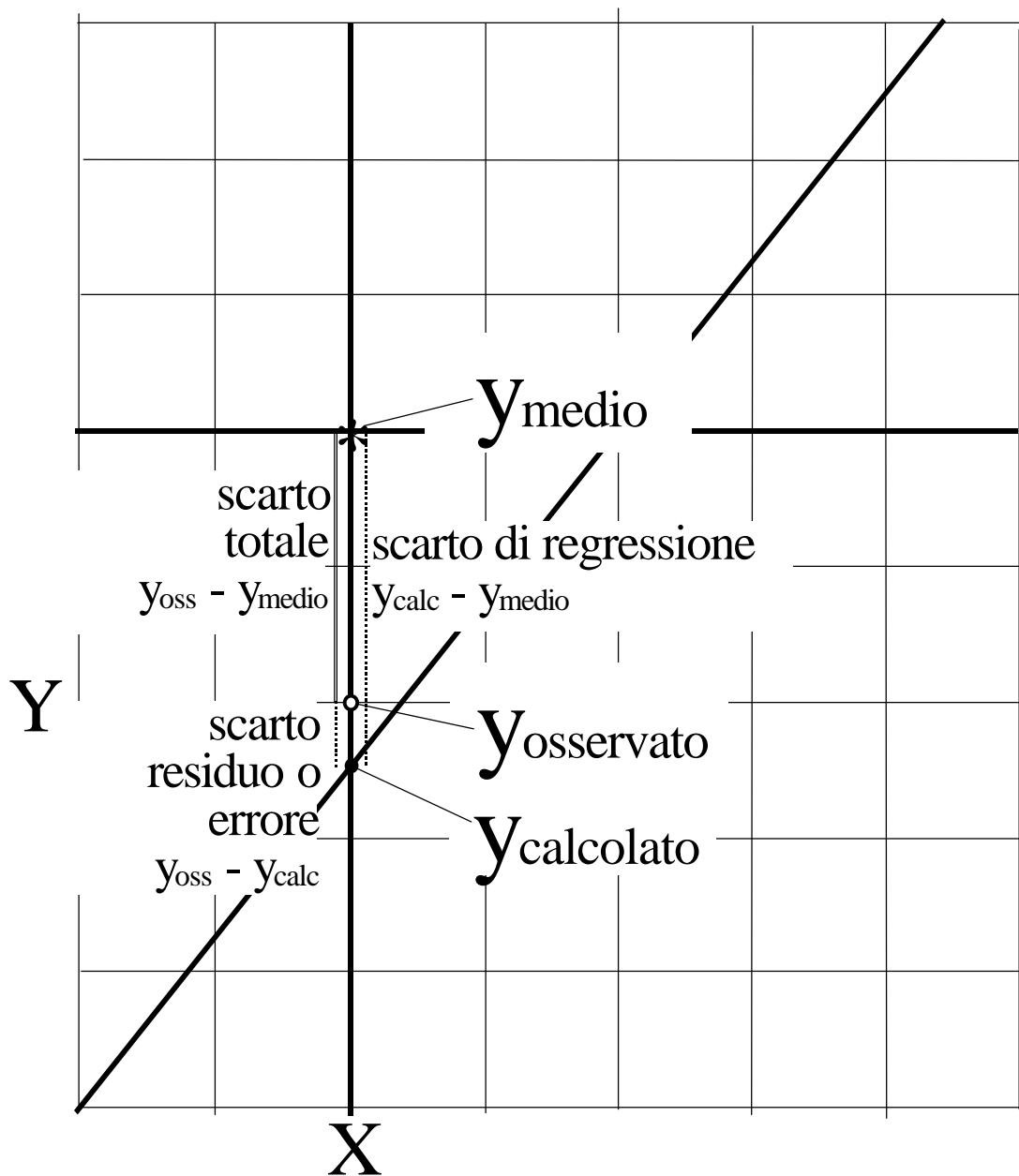
Prima di andare avanti, bisogna fare una importante osservazione. Le variabili X e Y , poste rispettivamente in ascissa ed in ordinata, non sono intercambiabili. X è infatti la cosiddetta **variabile indipendente**, mentre Y è la cosiddetta **variabile dipendente**. Non si tratta di una dipendenza causa-effetto (su questo punto anche molti autorevoli testi sono ambigui se non fuorvianti). Molto spesso può anche esservi tra X e Y una vera dipendenza causa-effetto (come nel caso X =dose del farmaco, Y =effetto del farmaco). Ma per noi il termine indipendente si riferisce solamente al tipo di variabilità nel senso di 'variato a priori, anche arbitrariamente'. Mentre dipendente significa 'libero di variare statisticamente senza condizionamenti imposti dal campionamento'. Per esempio: se decidiamo di valutare la crescita di ragazzi possiamo scegliere di prendere ragazzi di età diversa (10, 11, 12, ecc. anni) e poi di misurarne l'altezza (quella che risulterà). Quindi noi interveniamo rendendo uniforme la distribuzione della variabile X (es, prendendo 10 soggetti per ogni classe di età), mentre non interveniamo affatto sulla variabile Y che varierà spontaneamente, sperando che lo faccia seguendo la distribuzione normale ed avendo una certa relazione lineare con i valori di X . Questo è il vero significato di variabile indipendente e variabile dipendente. Un caso un po' paradossale, ma utile per chiarire questo aspetto, è quello riportato in una nostra indagine sulla maturazione dei denti dei ragazzi in cui è stato necessario e opportuno mettere come variabile indipendente lo stadio di sviluppo del dente e come variabile dipendente l'età dei ragazzi: l'opposto esatto della relazione di dipendenza biologica che vuole che i denti maturino in funzione dell'età. La statistica non entra nel merito di questi fatti. Nel nostro caso serviva assumere lo stadio di sviluppo del dente come variabile indipendente in modo da poter prevedere l'età fisiologica del ragazzo (più importante dell'età anagrafica). Inoltre lo stadio di sviluppo dentale non era distribuito normalmente, e quindi, anche volendo, non poteva essere assunto come variabile dipendente mentre l'età rispondeva a questo requisito (lo studio fu accettato senza obiezioni).

Nel paragrafo precedente, abbiamo anticipato che i valori della variabile dipendente devono essere distribuiti normalmente per ogni valore x in ascissa. Non solo, ma dovrebbero anche avere media linearmente variante al variare di x e deviazione standard costante. E' ovvio che quando si hanno pochi dati è impossibile verificare tali requisiti. Comunque, se dovessero verificarsi evidenti asimmetrie dei valori y attorno alla loro media (sospetta non-normalità) oppure distribuzioni bivariate a forma di punta di freccia (sospetto trend della deviazione standard) sarebbe bene approfondire il problema e/o consultare un esperto.

Analizziamo ora la variabilità della variabile dipendente Y . A fianco dei valori osservati mettiamo i valori calcolati in base all'equazione di regressione. Vediamo quindi che esistono tre tipi di scarti o differenze:

- differenze tra valori di y osservati (y_{oss}) e y_{medio}
→ **devianza totale o semplicemente devianza di y** , indicata con S_y
- differenze tra valori di y calcolati (y_{calc}) e y_{medio}
→ **devianza di regressione**, indicata con S_{reg}
- differenze tra valori di y osservati (y_{oss}) e valori calcolati (y_{calc})
→ **devianza di errore**, indicata con S_{res}

				variabilità totale	variabilità dovuta alla regressione	variabilità residua o di errore
				$\Sigma(y_{oss} - y_{medio})^2$	$\Sigma (y_{calc} - y_{medio})^2$	$\Sigma (y_{oss} - y_{calc})^2$



Come si noterà nel grafico, tenendo conto del segno delle operazioni, per ogni valore osservato di y , lo scarto totale corrisponde alla somma dello scarto di regressione più lo scarto di errore residuo. Questo vale anche per la somma dei quadrati degli scarti: la devianza totale di y corrisponde alla somma della devianza dovuta alla regressione più la devianza residua:

$$\begin{aligned} S_y &= S_{\text{reg}} + S_{\text{res}} \\ 14 &= 12.5 + 1.5 \end{aligned}$$

Lo stesso vale per i gradi di libertà:

$$\begin{aligned} \text{GDL}_y &= \text{GDL}_{\text{reg}} + \text{GDL}_{\text{res}} \\ n-1 &= 1 + n-2 \\ 2 &= 1 + 1 \end{aligned}$$

Si tratta quindi di una decomposizione della variabilità totale di y (dispersione dei valori osservati attorno alla media) in una variabilità dovuta alla regressione (dispersione dei valori dell'equazione attorno alla media) ed una variabilità residua o di errore (dispersione dei valori osservati attorno ai valori calcolati dall'equazione). E' come rendere giustizia dicendo che la variabilità imputabile ad Y è solo quella residua, in quanto l'altra, quella dovuta alla regressione, è unicamente imputabile alla relazione tra Y ed X .

Le varianze ottenute dalla decomposizione della devianza totale di Y sono importanti per testare l'ipotesi nulla secondo cui la varianza di regressione e quella residua siano uguali, che significa che non v'è relazione. Il test si esegue con il rapporto:

$$F = \frac{s_{\text{reg}}^2}{s_{\text{res}}^2}$$

In questo caso si tratta quindi di una **analisi della varianza applicata alla regressione**. Per l'ipotesi nulla, $F=1$. Come in precedenza, valori di F maggiori di 1 indeboliscono via via la probabilità a favore dell'ipotesi nulla. La tabella ci dirà se il valore di F ottenuto dal test è più grande di quello corrispondente ad $\alpha = 0.05$. Ovviamente, se F non risulterà significativo si dovrà accettare l'ipotesi nulla e concludere che non vi è relazione significativa tra Y ed X .

Finora abbiamo stimato solo i parametri dell'equazione. Bisogna adesso valutare la deviazione standard di questi. Nota bene: poiché si tratta di parametri globali di campione - come nel caso delle medie - è indifferente parlare di deviazione standard o errore standard. Quindi la notazione s_a si definisce indifferentemente deviazione standard o errore standard dell'intercetta. Idem per la pendenza e per altri parametri che vedremo in seguito.

Attenzione. Non confondiamoci. In diversi testi la devianza dei residui (seguita dalla varianza, deviazione standard ed errore standard dei residui) ha $y.x$ o y,x a pedice. Quindi:

$S_{x,y}$ o S_{xy} indicano la covarianza
 $S_{y,x}$ o $S_{y.x}$ indicano la devianza dei residui di Y, scritta anche come anche S_{res}

L'errore standard della pendenza è dato dalla formula:

$$s_b = \sqrt{\frac{S_{res}^2}{S_x}}$$

L'errore standard per i valori di y della retta è:

$$s_y = \sqrt{S_{res}^2 \left(\frac{1}{n} + \frac{(x - x_{medio})^2}{S_x} \right)}$$

Notare che il termine $(x - x_{medio})^2$ è tanto maggiore quanto più x si allontana dalla media. Per cui l'errore standard sarà minimo in corrispondenza del valore x medio, e andrà crescendo a destra e a sinistra. Quindi tracciando tutti gli errori standard dei valori y si ottiene una cintura attorno alla retta di regressione che è stretta al centro e si allarga ai lati. Questa cintura è detta cintura di confidenza (*confidence belt*, vedi grafico successivo).

L'errore standard dell'intercetta è semplicemente l'errore standard del valore y corrispondente a $x=0$. Pertanto sarà:

$$s_a = \sqrt{S_{res}^2 \left(\frac{1}{n} + \frac{x_{medio}^2}{S_x} \right)}$$

Finalmente possiamo saggiare l'ipotesi che l'intercetta sia uguale a zero ($H_0: a=0$) mediante il test t :

$$t = \frac{a}{s_a}, \quad \text{con } n-2 \text{ gradi di libertà, gli stessi di } s_{res}.$$

Analogamente, l'ipotesi che la pendenza sia uguale a zero ($H_0: b=0$) si può verificare mediante il test:

$$t = \frac{b}{s_b}, \quad \text{sempre con } n-2 \text{ gradi di libertà.}$$

Abbiamo già visto cosa succede se $a=0$ o $b=0$. Quindi i risultati di questi due test dovranno farci trarre delle importanti conclusioni.

Una volta calcolata l'equazione della retta, le previsioni di y per un nuovo valore di x (oltre a quelli usati per la valutazione del modello e dei parametri) sono consentite solo nell'ambito dell'intervallo sperimentale

L'errore standard delle previsioni, cioè di nuovi valori di y è:

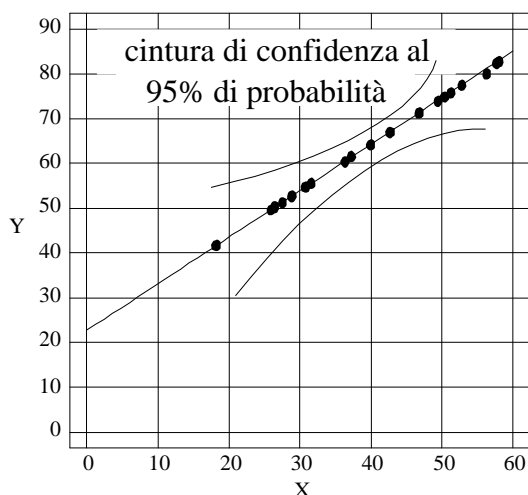
$$s_{y \text{ da nuovo } x} = \sqrt{s_{\text{res}}^2 \left(1 + \frac{1}{n} + \frac{(x - x_{\text{medio}})^2}{S_x} \right)}$$

Pertanto un valore y ottenuto in base all'equazione della retta per un nuovo valore x avrà un intervallo fiduciale:

$$LF = y \pm t s_y$$

dove il valore di t è come al solito riferito al livello di probabilità adottato ed ai gradi di libertà della varianza dei residui: $n-2$.

Se poi si calcolano i limiti fiduciali per ogni previsione y al livello di probabilità del 95% o del 99% si ottengono le cosiddette fasce o **cinture di confidenza** o **confidence belts** (una più ristretta, per il 95%, ed una più dilatata, per il 99%, dato il maggiore valore di t) entro cui si spera (matematicamente) che si sia compreso il valore vero oggetto delle previsioni.



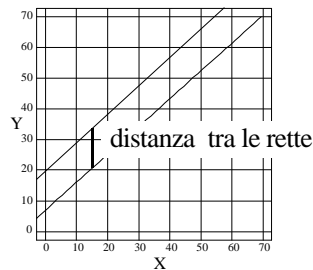
Si noterà che le fasce di confidenza si restringono ad arco verso il centro. Ciò è dovuto al termine $(x - x_{\text{medio}})^2$ presente nella formula dell'errore standard. Quanto più ci allontaniamo dal valore medio di x , tanto più aumenta l'errore standard della previsione. In altre parole, la previsione più forte di y è quella corrispondente al

valore medio di x , per la quale tutto il termine $(x - x_{\text{medio}})^2$ dell'errore standard si azzera.

Avendo calcolato 2 rette di regressione ci si può domandare se i 2 coefficienti angolari b_a e b_b siano significativamente diversi ($b_a \neq b_b$), in alternativa all'ipotesi nulla che le 2 rette siano parallele ($b_a = b_b$). Il test che saggia la probabilità a favore di tale ipotesi è detto **test di parallelismo**:

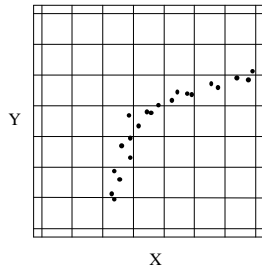
$$t = \frac{b_a - b_b}{\sqrt{\frac{S_{\text{res}_a} + S_{\text{res}_b}}{n_a + n_b - 4} \left(\frac{1}{S_{x_a}} + \frac{1}{S_{x_b}} \right)}}$$

Ad es., può essere interessante verificare se due differenti tassi di crescita siano significativamente diversi o assumibili come equivalenti. Se poi i tassi di crescita risultassero equivalenti (pendenza comune, come da ipotesi nulla), sarebbe interessante calcolare la precocità di una popolazione nei confronti dell'altra, risultante nella distanza che separa verticalmente le due rette.

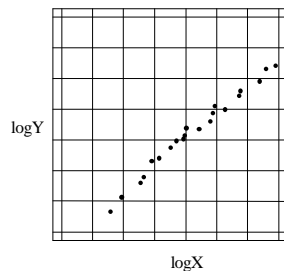


Situazioni non-lineari

Se X e Y sono legate da una **relazione non-lineare** si può ricorrere a trasformazioni dei dati tali da restituire un andamento lineare. Classica è la trasformazione log-log che linearizza la relazione tra dimensione dello step (X) e misura del perimetro (Y) dei contorni frattali e quindi in genere delle forme naturali. In pratica una relazione della forma:



dopo trasformazione log-log appare abbastanza linearizzata:



Un altro classico caso di rapporto non-lineare è quello dose (X) - effetto (Y) che normalmente si linearizza valutando il logaritmo della concentrazione della dose. In teoria, si può introdurre qualsiasi trasformazione (logaritmica, esponenziale, ecc.) utile a restituire linearità ai dati. Ovviamente i migliori risultati si ottengono con un po' di studio ed un po' di esperienza. Esistono comunque tecniche che suggeriscono il tipo trasformazione, e giudicano anche tra diverse trasformazioni quella che meglio linearizza i dati.

Correlazione

Per correlazione si intende un semplice rapporto di associazione bidirezionale tra due variabili, che non implica dipendenza né mira a fare previsioni. Tanto che in alcuni testi si evita di chiamare le due variabili X e Y, definendole semplicemente X_1 e X_2 . Noi, per mantenere una connessione con alcuni aspetti della regressione, manterremo X e Y. Il coefficiente di correlazione r è un indice di adattamento dei punti alla retta (o alla curva, nel caso di modelli non-lineari).

- r è un numero puro, adimensionale compreso tra -1 e +1.
- r è positivo quando b è positivo (x e y crescono insieme).
- r è negativo quando b è negativo (per x che cresce, y decresce e viceversa).
- r è esattamente +1 o -1 quando i punti coincidono perfettamente con la retta.
- r è esattamente zero quando i punti formano una nuvola omogenea e circolare.

r si calcola come:

$$r = \frac{S_{x,y}}{\sqrt{S_x S_y}} \quad (1^\circ \text{ equazione})$$

Poichè la codevarianza (al numeratore) non è mai maggiore della radice del prodotto delle due devianze (al denominatore) r, in valore assoluto, non potrà mai essere maggiore di 1.

E' evidente la simmetria della formula. Significa che r non varia invertendo gli assi.

r si può ottenere anche come:

$$r = \sqrt{\frac{S_{\text{reg}}}{S_y}} \quad (2^\circ \text{ equazione})$$

Questa formula è importante perché se eleviamo tutto al quadrato otteniamo:

$$r^2 = \frac{S_{\text{reg}}}{S_y}$$

Poiché la devianza di regressione (al numeratore) non è che una frazione della devianza totale di Y (al denominatore), r^2 esprime la frazione della devianza totale di Y dovuta alla regressione. Per questo r^2 è anche detto **coefficiente di determinazione**. Per il bravo statistico, il valore di r^2 è ancora più importante del valore r. Generalmente r^2 è espresso in percentuale (basta moltiplicare r^2 per 100).

L'errore standard di r è dato da:

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

Se sostituiamo r^2 mediante la seconda equazione otteniamo:

$$s_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-\frac{S_{\text{reg}}}{S_y}}{n-2}} = \sqrt{\frac{\frac{S_{\text{res}}}{S_y}}{n-2}} = \sqrt{\frac{S_{\text{res}}}{n-2} \frac{1}{S_y}} = \sqrt{\frac{S_{\text{res}}^2}{S_y}}$$

notare la somiglianza con l'errore standard della pendenza: $s_b = \sqrt{\frac{S_{\text{res}}^2}{S_x}}$

Possiamo quindi saggiare la significatività di r ($H_0: r = 0$) con il solito test t :

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

Applicazione. Riprendiamo i dati da cui abbiamo calcolato i parametri della regressione.

X	Y	$(x-x_{\text{medio}})(y-y_{\text{medio}})$
1	2	$(1-2)(2-4)=2$
2	3	$(2-2)(3-4)=0$
3	7	$(3-2)(7-4)=3$
		$S_{x,y}=5$
$x_{\text{medio}}=2$	$y_{\text{medio}}=4$	
$S_x=2$	$S_y=14$	

$S_y = 14$	$GDL_y = 2$
$S_{\text{reg}} = 12.5$	$GDL_{\text{reg}} = 1$
$S_{\text{res}} = 1.5$	$GDL_{\text{res}} = 1$

Calcoliamo r .

In base alla prima equazione

$$r = \frac{S_{x,y}}{\sqrt{S_x S_y}} = \frac{5}{\sqrt{2 \times 14}} = \frac{5}{\sqrt{2 \times 14}} = 0.945$$

In base alla seconda equazione

$$r = \sqrt{\frac{S_{\text{reg}}}{S_y}} = \sqrt{\frac{12.5}{14}} = 0.945$$

Calcoliamo l'errore standard di r .

In base alla prima equazione

$$s_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0.945^2}{n-2}} = \sqrt{\frac{0.107}{1}} = 0.327$$

In base alla seconda equazione

$$s_r = \sqrt{\frac{s_{\text{res}}^2}{S_y}} = \sqrt{\frac{1.5}{14}} = 0.327$$

Per la significatività di r calcoliamo il rapporto t

$$t = \frac{r}{s_r} = \frac{0.945}{0.327} = 2.890$$

Il valore di t sembra buono, ma purtroppo non è significativo. Con un solo grado di libertà ($n - 2 = 1$), la significatività si raggiunge solo con un t pari o maggiore a 12.706 (vedi la tabella nel Cap. 2). Concludiamo pertanto che, nonostante l'alto valore (0.945), il coefficiente di correlazione non è significativamente diverso da zero. In altre parole, non siamo autorizzati a ritenere che esista una correlazione tra le due variabili.

Alcuni grafici con valori indicativi di r :

