Il metodo dei minimi quadrati e la retta di regressione

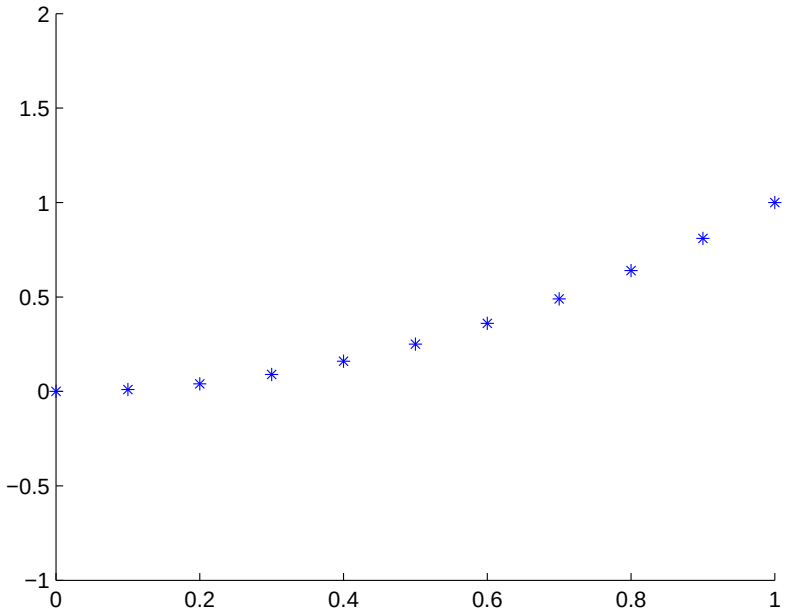
Francesco Dell'Accio

Dipartimento di Matematica e Informatica
Università della Calabria, 87036 Rende (CS), Italia

Nuovo Progetto Lauree Scientifiche 13/04/2016

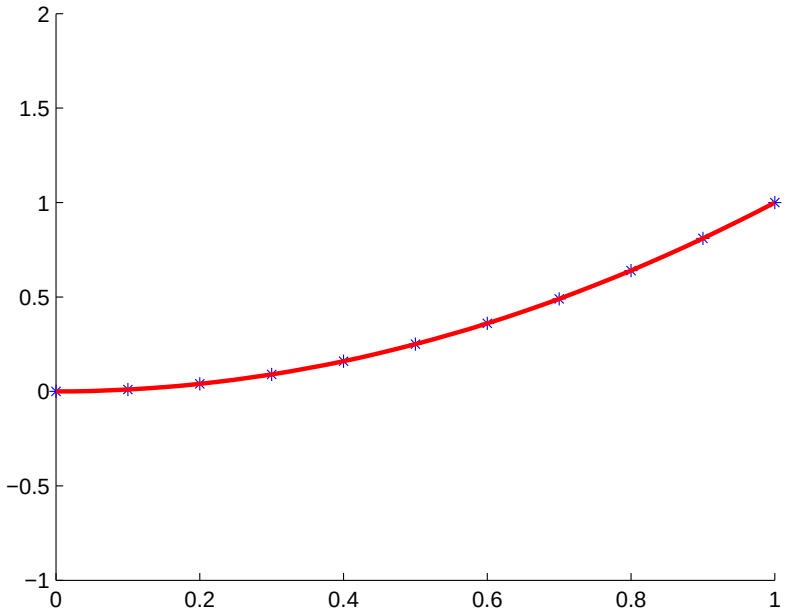
Spesso, in diversi campi scientifici, si deve affrontare il problema di interpretare, valutare e gestire dati ricavati da rilevazioni statistiche o da esperimenti riguardanti un fenomeno. Il problema è piuttosto complesso, mentre noi vogliamo soltanto dare un'idea di come si può affrontarlo e ci limitiamo a casi semplici. Studiamo problemi che riguardano relazioni fra due sole variabili x e y, delle quali conosciamo alcune coppie di valori (x_i, y_i), rilevati da un'indagine statistica e che vogliamo interpretare tramite una funzione $y = f(x)$. Consideriamo quindi le coppie ordinate di valori (x_i, y_i) e rappresentiamole in un piano cartesiano tramite punti, ottenendo quello che chiamiamo diagramma a dispersione o nuvola di punti

Vogliamo determinare una funzione matematica, che chiameremo funzione interpolante, in grado di rappresentare il fenomeno studiato.



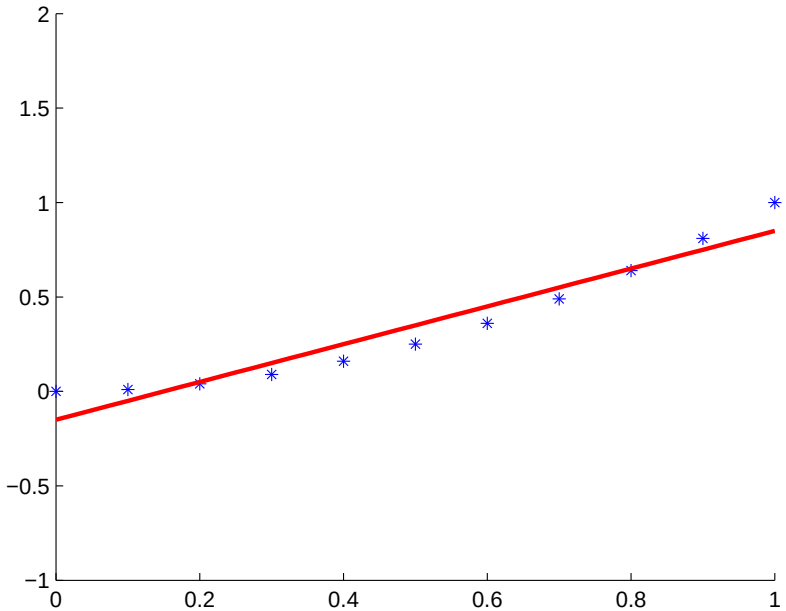
Vogliamo determinare una funzione matematica, che chiameremo funzione interpolante, in grado di rappresentare il fenomeno studiato.

1. Se la funzione assume esattamente i valori rilevati, e quindi il suo grafico passa per tutti i punti del diagramma a dispersione, parliamo di interpolazione per punti noti o interpolazione matematica;



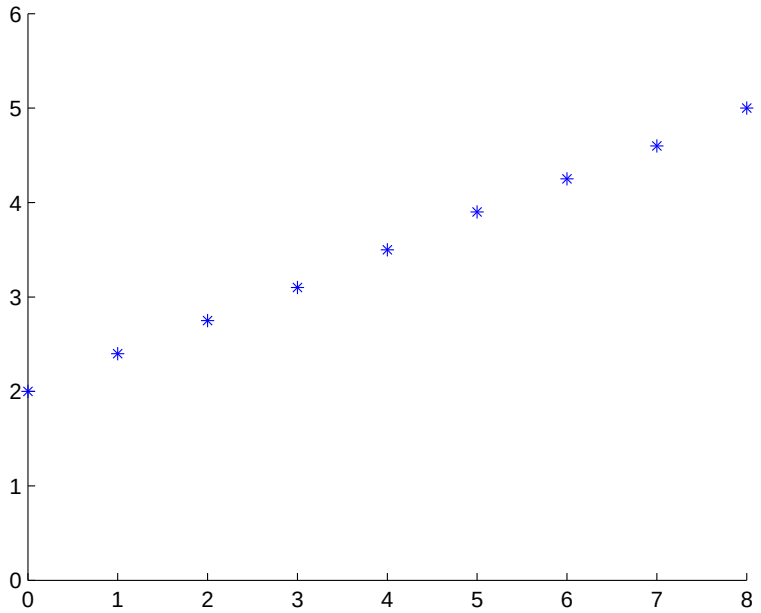
Vogliamo determinare una funzione matematica, che chiameremo funzione interpolante, in grado di rappresentare il fenomeno studiato.

1. Se la funzione assume esattamente i valori rilevati, e quindi il suo grafico passa per tutti i punti del diagramma a dispersione, parliamo di interpolazione per punti noti o interpolazione matematica;
2. Se la funzione assume valori *vicini* ai valori rilevati e quindi il suo grafico passa fra i punti del diagramma a dispersione, parliamo di interpolazione fra punti noti o interpolazione statistica



Supponiamo di avere un insieme di valori della quantità y corrispondenti a valori della quantità x, ad esempio, la seguente sequenza di valori x_i, y_i che plottiamo in un grafico.

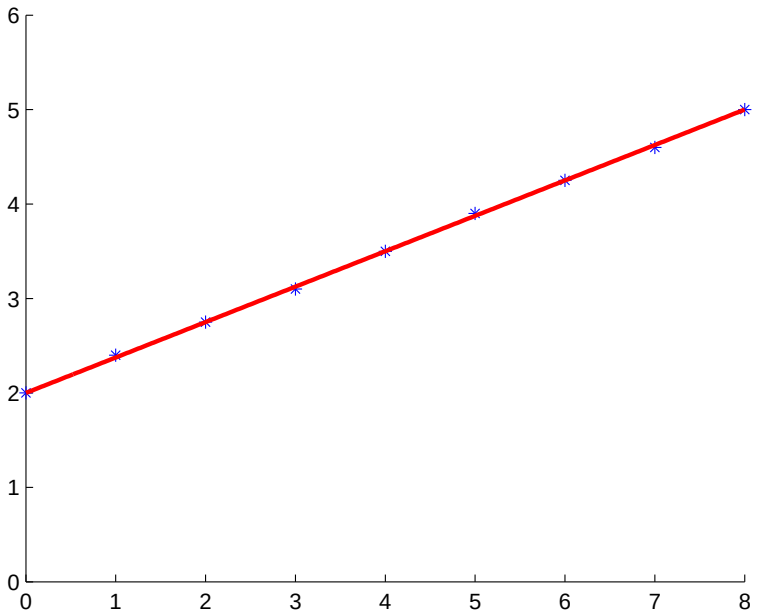
x_i	0	1	2	3	4	5	6	7	8
y_i	2.0	2.4	2.75	3.1	3.5	3.9	4.25	4.6	5.0



Dal grafico, sospettiamo che esista una relazione lineare tra x e y:

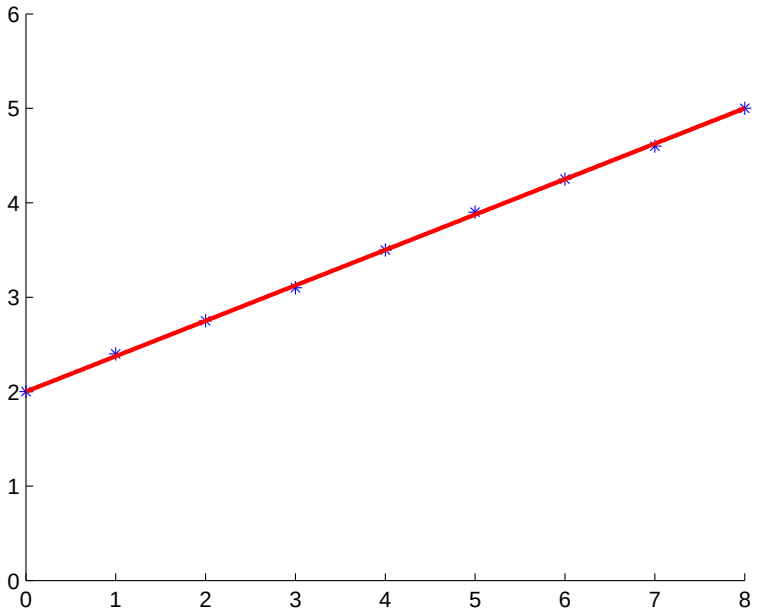
$$y = mx + q.$$

Effettivamente i punti possono essere *congiunti* abbastanza facilmente con una retta. La pendenza e l'intercetta possono essere misurate e, in particolare, usando il primo e l'ultimo dei punti plottati, si trova che la pendenza della retta è $\frac{3}{8}$, mentre l'intercetta è 2. L'equazione della retta è dunque $y = \frac{3}{8}x + 2$



Da un'analisi più attenta del grafico si intuisce che alcuni punti (x_i, y_i) non appartengono alla retta. Verifichiamo ciò calcolando i residui o errori di accostamento $r_i = y_i - mx_i - q$

x_i	0	1	2	3	4	5	6	7	8
r_i	0	0.0250	0	-0.0250	0	0.0250	0	$-0,0250$	0

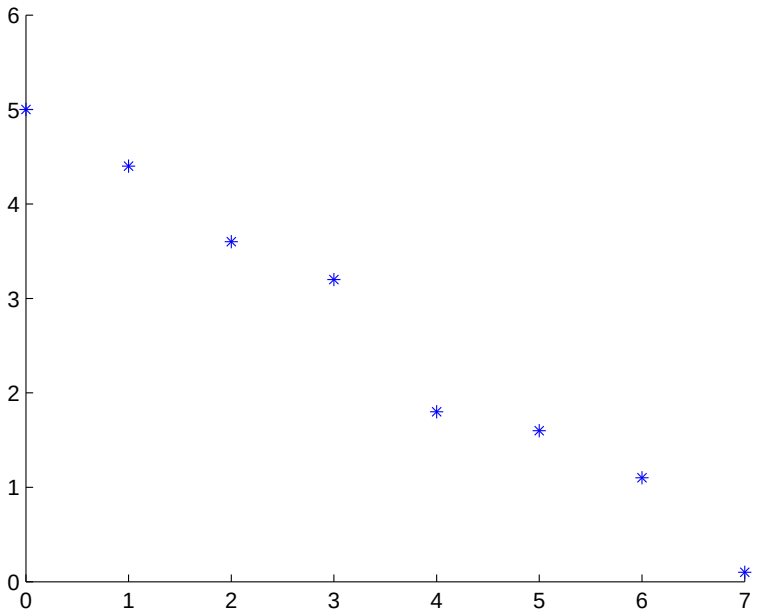


Il grafico tracciato appare però soddisfacente, poichè tutti i punti risultano *quasi* sulla retta.

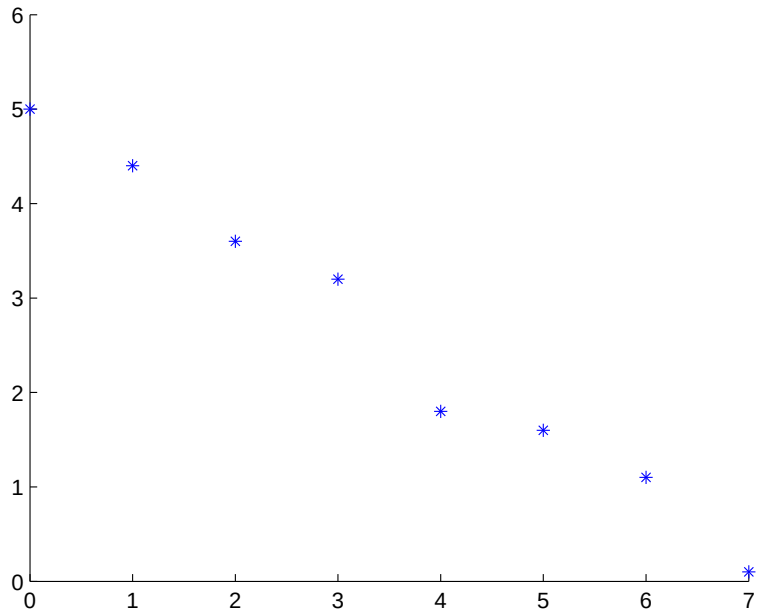
Più spesso, tuttavia, specialmente quando si ha a che fare con dati sperimentali, dall'analisi dei grafici si ha una forte impressione che una relazione lineare debba esistere, ma è veramente difficile determinare *ad occhio* la posizione della retta. Consideriamo ad esempio il seguente set di punti

x_i	0	1	2	3	4	5	6	7
y_i	5	4.4	3.6	3.2	1.8	1.6	1.1	0.1

che plottiamo nel grafico seguente

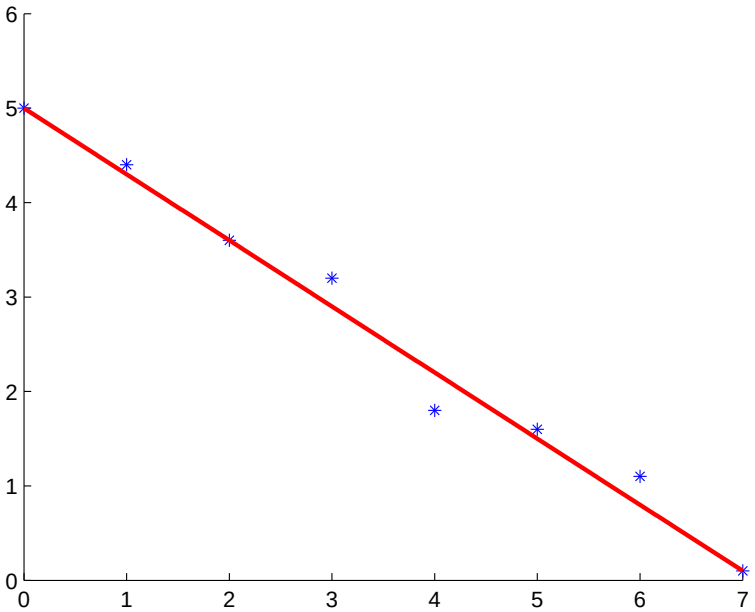


L'impressione è che i dati si dispongano intorno ad una retta, ma la posizione esatta della retta diventa una questione di gusto personale



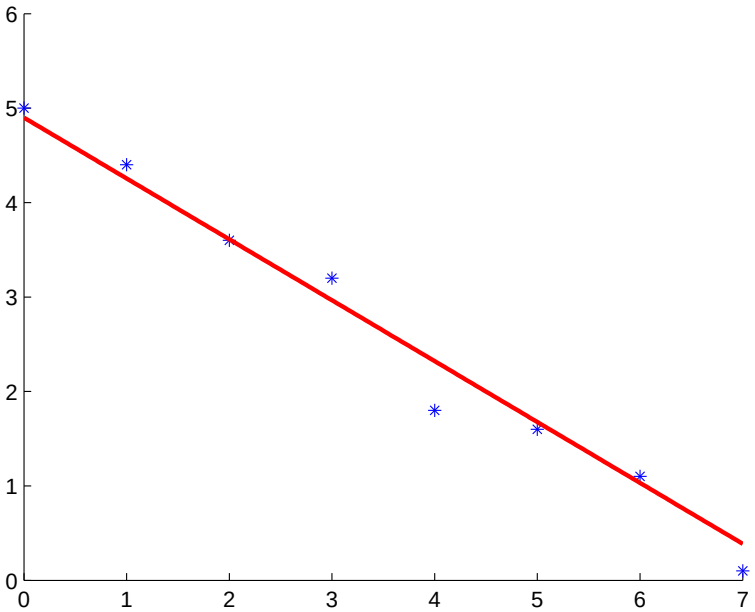
L'impressione è che i dati si dispongano intorno ad una retta, ma la posizione esatta della retta diventa una questione di gusto personale

1. Come prima, infatti, possiamo tracciare la retta passante per il primo e l'ultimo dei punti, di equazione $y = -\frac{4.9}{7}x + 5$ e giudicarla adeguata



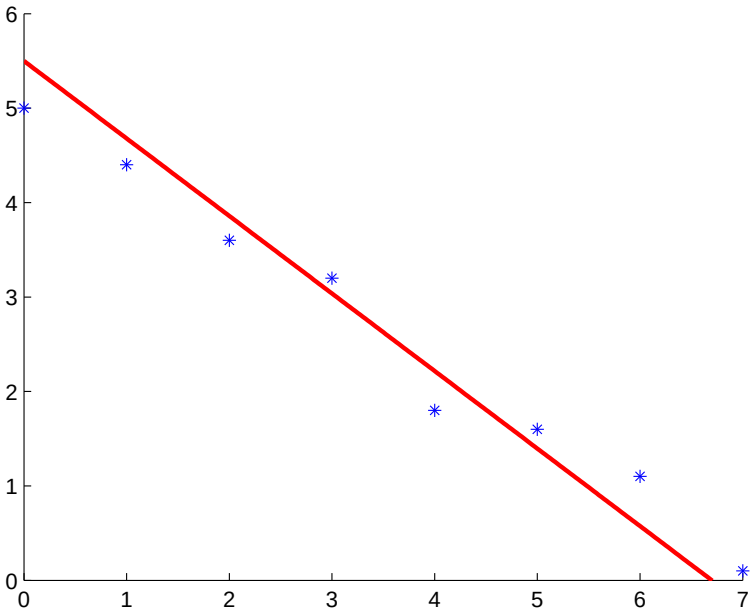
L'impressione è che i dati si dispongano intorno ad una retta, ma la posizione esatta della retta diventa una questione di gusto personale

1. Come prima, infatti, possiamo tracciare la retta passante per il primo e l'ultimo dei punti, di equazione $y = -\frac{4.9}{7}x + 5$ e giudicarla adeguata
2. Qualcuno potrebbe giudicare la retta di equazione $y = -\frac{4.9}{7.6}x + 4.9$ che lascia sopra e sotto di se un numero uguale di punti più adeguata



L'impressione è che i dati si dispongano intorno ad una retta, ma la posizione esatta della retta diventa una questione di gusto personale

1. Come prima, infatti, possiamo tracciare la retta passante per il primo e l'ultimo dei punti, di equazione $y = -\frac{4.9}{7}x + 5$ e giudicarla adeguata
2. Qualcuno potrebbe giudicare la retta di equazione $y = -\frac{4.9}{7.6}x + 4.9$ che lascia sopra e sotto di se un numero uguale di punti più adeguata
3. Qualcun'altro potrebbe decidere che i punti più al centro sono più importanti e pertanto giudicare migliore la rappresentazione fornita dalla retta di equazione $y = -\frac{5.5}{6.7}x + 5.5$



Come determinare la retta $y = mx + q$? Agire sui residui!

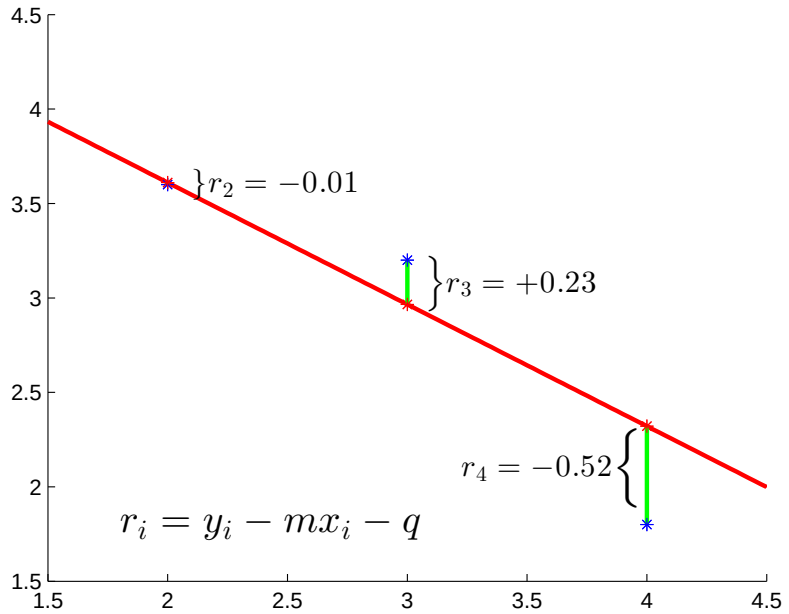
Se il punto (x_i, y_i) giace sulla retta $y = mx + q$, allora risulta

$$r_i = y_i - mx_i - q = 0;$$

d'altro canto, se il punto (x_i, y_i) è esterno alla retta $y = mx + q$, allora risulta

$$r_i = y_i - mx_i - q \neq 0.$$

Il residuo r_i rappresenta la distanza (con segno) tra il dato (x_i, y_i) e il punto sulla retta $(x_i, mx_i + q)$ corrispondente al valore x_i della variabile. Il residuo r_i ha valore positivo o negativo dipendentemente dal fatto che il punto plottato giace sopra o sotto la retta.



Non potendo annullare tutti i residui (ciò capita solo se i punti sono allineati, e $y = mx + q$ è esattamente la retta passante per essi) possiamo pensare di annullare la loro somma:

$$\sum_{i=1}^{n} (y_i - mx_i - q) = 0.$$

La precedente equazione può anche essere scritta nel seguente modo

$$\sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - nq = 0. \tag{1}$$

Il numero

$$r = \sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - nq$$

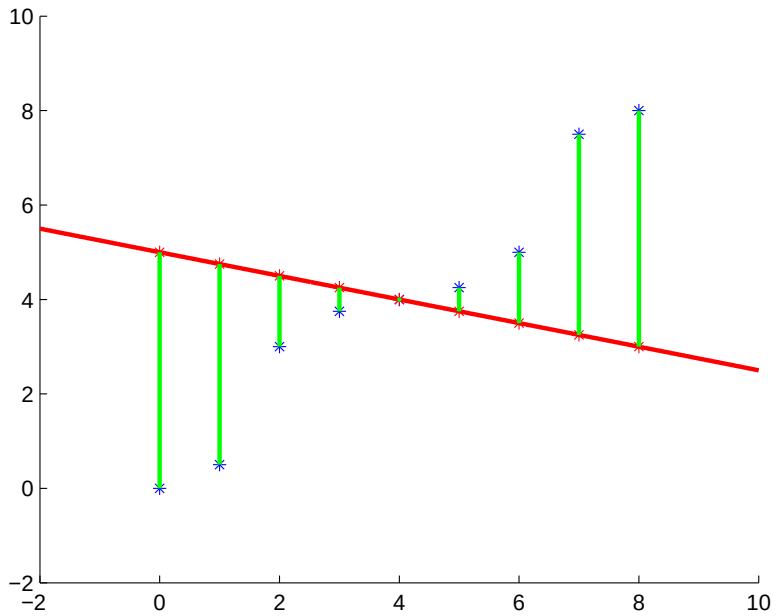
è detto residuo totale.

La sola equazione (1) nelle incognite m e q non è sufficiente a determinare in modo univoco una retta $y = mx + q$. Inoltre il residuo totale può essere nullo anche nel caso in cui la retta $y = mx + q$ non si adatta affatto ai dati!

Consideriamo il seguente set di dati

x_i	0	1	2	3	4	5	6	7	8
y_i	0	0.5	3	3.75	4	4.25	5	7.5	8

non è difficile verificare che ogni retta di equazione $y - 4 = m(x - 4)$ è tale $\sum_{i=1}^{n} r_i = 0$ con residui che tendono ad infinito per m tendente ad infinito.



Naturalmente non riusciamo a determinare i valori di m e q dall'equazione (1) poichè essa contiene due incognite. Però possiamo dividere i punti dati in due gruppi e costituire due equazioni separate della forma (1). Detti quindi I_1, I_2 due insiemi non vuoti di indici tali che $I_1 \cup I_2 = \{0, 1, \ldots, n\}$ e $I_1 \cap I_2 = \emptyset$ consideriamo il sistema di due equazioni in due incognite

$$\begin{aligned} \sum_{i \in I_1} y_i - m \sum_{i \in I_1} x_i - n_1 q = 0 \\ \sum_{i \in I_2} y_i - m \sum_{i \in I_2} x_i - n_2 q = 0 \end{aligned}$$

dove con n_1 e n_2 abbiamo denotato rispettivamente le cardinalità (numero di elementi di) di I_1 e I_2 . Le due equazioni ci consentiranno di calcolare i valori di m e di q. Come si vede il metodo delle medie non presenta alcuna difficoltà di calcolo; è inoltre possibile mostrare che questo metodo può produrre soluzioni sorprendentemente buone, a patto che il raggruppamento delle equazioni è realizzato in modo appropriato.[1]

[1] G. Dahlquist, B. Sjoberg and P. Svensson, Comparison of the Method of Averages with the Method of Least Squares, Mathematics of Computation, Vol. 22, No. 104 (Oct., 1968), pp. 833-845

Applichiamo la procedura al secondo set di dati presentati.

Gruppo 1		Gruppo 2	
x	y	x	y
0	5.0	4	1.8
1	4.4	5	1.6
2	3.6	6	1.1
3	3.2	7	0.1
$\sum_{I_1} x$	$\sum_{I_1} y$	$\sum_{I_2} x$	$\sum_{I_2} y$
6	16.2	22	4.6

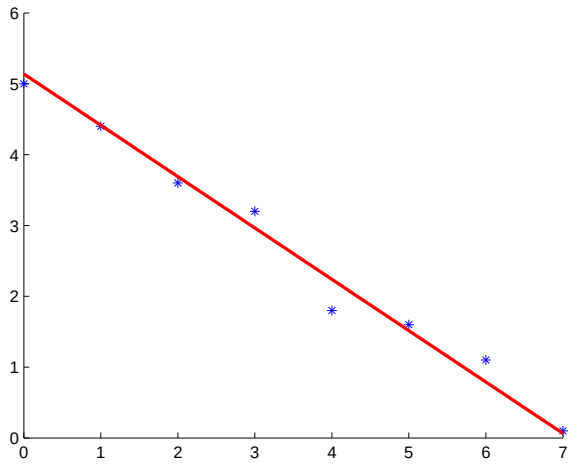
Poniamo le due equazioni a sistema

$$\begin{cases} 16.2 - 6m - 4q = 0 \\ 4.6 - 22m - 4q = 0 \end{cases}$$

e risolvendo si ha

$$\begin{cases} m = -\frac{11.6}{16} = 0.725 \\ q = 5.14 \end{cases}$$

per cui la retta richiesta ha equazione $y = -0.725x + 5.14$.



Come visto in precedenza, la sola equazione del residuo totale nullo non è sufficiente a garantire che i residui siano effettivamente piccoli, poichè residui di segno opposto tendono ad annullarsi nella somma. D'altro canto dalla disuguaglianza

$$0 \leq |r_i| \leq \sqrt{\sum_{j=1}^{n} r_j^2} \text{ per ogni } i = 1, \ldots, n,$$

deduciamo che se la radice quadrata della somma dei quadrati dei residui è piccola, allora ogni residuo r_i è vicino a zero. Equivalentemente *se la somma dei quadrati dei residui è minima, allora ogni residuo r_i è vicino a zero* per la nota proprietà dei numeri reali positivi[2]
Ci poniamo quindi il problema della determinazione della retta $y = mx + q$ per cui risulta minima (leggi *più piccola possibile*) la quantità

$$S(m, q) = \sum_{i=1}^{n} (y_i - mx_i - q)^2 .$$

Notiamo subito che $S(m, q)$ è un polinomio di grado due in m e q.

[2] *Per ogni* $a, b \geq 0$, $a \leq b \iff a^2 \leq b^2$.

Fissiamo inizialmente il numero $q \in \mathbb{R}$ che assumiamo come parametro. Per ogni $i = 1, \ldots, n$ sviluppiamo i quadrati

$$(y_i - mx_i - q)^2 = m^2 x_i^2 - 2my_i x_i + 2mqx_i + \ldots$$

quindi sommiamo su tutti gli indici e riarrangiamo la somma come segue

$$S(m, q) = \sum_{i=1}^{n} x_i^2 m^2 - 2\left(\sum_{i=1}^{n} y_i x_i - q\sum_{i=1}^{n} x_i\right) m + \ldots$$

La relazione precedente lega, per q fissato, la variabile S alla variabile m e definisce una parabola di equazione

$$y = ax^2+bx+c \ \textit{con}\ y = S, x = m, a = \sum_{i=1}^{n} x_i^2 > 0 \ \textit{e}\ b = -2\left(\sum_{i=1}^{n} y_i x_i - q\sum_{i=1}^{n} x_i\right)$$

La parabola volge la concavità verso l'alto, per cui il più piccolo dei valori y al variare di x si ha in corrispondenza del vertice, di ascissa $x = -\dfrac{b}{2a}$. Riscriviamo la relazione precedente nelle variabili m e q per ottenere

$$m = \frac{\sum_{i=1}^{n} y_i x_i - q\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} \tag{2}$$

In modo del tutto analogo, fissiamo il numero $m \in \mathbb{R}$ che ora assumiamo come parametro. Nello sviluppo dei quadrati

$$(y_i - mx_i - q)^2 = q^2 - 2y_i q + 2mx_i q + \dots$$

scriviamo prima le potenze di q in ordine decrescente, quindi sommiamo su tutti gli indici e riarrangiamo la somma come segue

$$S(m,q) = \sum_{i=1}^{n} q^2 - 2\left(\sum_{i=1}^{n} y_i - m\sum_{i=1}^{n} x_i\right) q + \dots$$

La relazione precedente lega, per m fissato, la variabile S alla variabile q e definisce una parabola di equazione

$$y = ax^2+bx+c \ \textit{con}\ y = S, x = q, a = \sum_{i=1}^{n} 1 = n > 0 \ \textit{e}\ b = -2\left(\sum_{i=1}^{n} y_i - m\sum_{i=1}^{n} x_i\right)$$

La parabola volge la concavità verso l'alto, per cui il più piccolo dei valori y al variare di x si ha in corrispondenza del vertice, di ascissa $x = -\frac{b}{2a}$. Riscriviamo la relazione precedente nelle variabili m e q per ottenere

$$q = \frac{\sum_{i=1}^{n} y_i - m\sum_{i=1}^{n} x_i}{n} \tag{3}$$

Il problema delle determinazione dei valori m e q in corrispondenza dei quali è minima la quantità $S(m,q)$ si riduce quindi alla determinazione delle soluzioni del seguente sistema di due equazioni lineari in due incognite

$$\begin{cases} \sum_{i=1}^{n} y_i - m\sum_{i=1}^{n} x_i - nq = 0 \\ \sum_{i=1}^{n} x_i y_i - m\sum_{i=1}^{n} x_i^2 - q\sum_{i=1}^{n} x_i = 0 \end{cases}$$

che riscriviamo in forma coincisa come segue

$$\begin{cases} \sum y - m\sum x - nq = 0 \\ \sum xy - m\sum x^2 - q\sum x = 0 \end{cases} \tag{4}$$

Il sistema precedente, nelle incognite m e q, ammette una e una sola soluzione. Ciò risulta dalla disuguaglianza

$$\left(\sum x\right)^2 - n\sum x^2 < 0^3$$

e dal Teorema di Cramer. Risolviamo il sistema mediante la regola di Cramer:

$$\begin{cases} m = \dfrac{n\sum xy - \sum x\sum y}{n\sum x^2 - \left(\sum x\right)^2} \\ q = \dfrac{\sum y\sum x^2 - \sum x\sum xy}{n\sum x^2 - \left(\sum x\right)^2} \end{cases} \tag{5}$$

[3] Si applica la disuguaglianza di Cauchy-Schwarz: $\left(\sum_{i=1}^{n} x_i y_i\right)^2 \leq \left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right)$

Come esempio, consideriamo ancora una volta l'insieme di dati della sezione precedente. Prima di tutto compiliamo la seguente tabella di dati

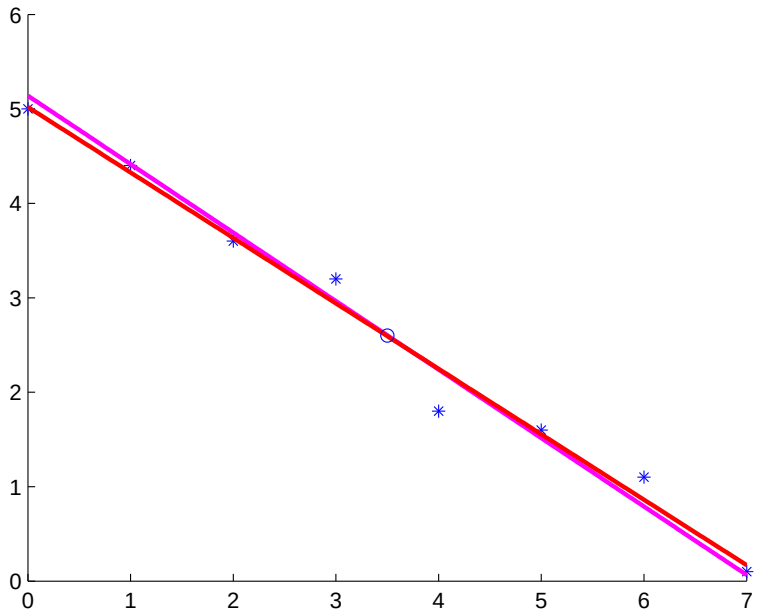
x	y	xy	x^2
0	5.0	0	0
1	4.4	4.4	1
2	3.6	7.2	4
3	3.2	9.6	9
4	1.8	7.2	16
5	1.6	8.0	25
6	1.1	6.6	36
7	0.1	0.7	49
$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$
28	20.8	43.7	140

Il numero di punti dati è 8 per cui inserendo i valori delle quantità in (5) otteniamo dopo qualche calcolo

$$\begin{cases} m = -\dfrac{29.1}{42} = -0.693 \\ q = 5.02 \end{cases}$$

per cui la retta richiesta ha equazione $y = -0.693x + 5.02$

Metodo dei minimi quadrati vs metodo delle medie



Nel grafico precedente confrontiamo la retta ottenuta con il metodo delle medie (in magenta) con la retta ottenuta con il metodo dei minimi quadrati (in rosso). In generale, il metodo dei minimi quadrati è più accurato del metodo delle medie. La retta ottenuta con il metodo dei minimi quadrati passa per il *baricentro* dei punti dati

$$\left(\frac{1}{n} \sum_{i=1}^{n} x_i, \frac{1}{n} \sum_{i=1}^{n} y_i \right)$$

La verifica di questa proprietà è immediata.

Otteniamo la retta che ben si adatta ai punti dati mediante il metodo dei minimi quadrati risolvendo il sistema (4) o il sistema equivalente

$$\begin{cases} \frac{1}{n}\sum y - m\frac{1}{n}\sum x - q = 0 \\ \frac{1}{n}\sum xy - m\frac{1}{n}\sum x^2 - q\frac{1}{n}\sum x = 0 \end{cases}$$

ottenuto dividendo entrambe le equazioni per n. Poniamo

$$\bar{x} = \frac{1}{n}\sum x, \quad \bar{y} = \frac{1}{n}\sum y$$

e riscriviamo il sistema precedente nella forma più compatta

$$\begin{cases} \bar{y} - m\bar{x} - q = 0 \\ \frac{1}{n}\sum xy - \frac{m}{n}\sum x^2 - q\bar{x} = 0 \end{cases} \tag{6}$$

Sottraiamo dalla prima equazione moltiplicata per $\bar{x}$ la seconda equazione:

$$\bar{x}\bar{y} - m\bar{x}^2 - \frac{1}{n}\sum xy + \frac{m}{n}\sum x^2 = 0$$

per ricavare

$$m = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} \tag{7}$$

e sostituendo il valore di m nella prima delle equazioni (6) otteniamo

$$q = \frac{\bar{y}\sum x^2 - \bar{x}\sum xy}{\sum x^2 - n\bar{x}^2}. \tag{8}$$

La retta ottenuta con il metodo dei minimi quadrati, scritta nella forma

$$y = mx + q$$

con m e q calcolabili mediante le formule

$$m = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} \qquad q = \frac{\bar{y}\sum x^2 - \bar{x}\sum xy}{\sum x^2 - n\bar{x}^2}$$

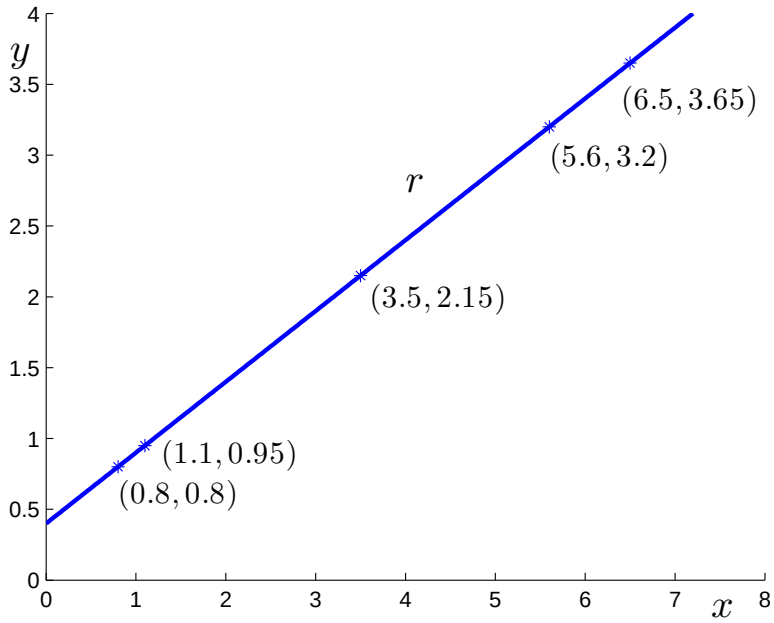
è nota col nome di retta di regressione. Questa retta, in particolare, stima i valori di y dai valori di x e pertanto è nota col nome di retta di regressione di y su x. In modo del tutto analogo possiamo stimare i valori di x dai valori di y. In questo caso calcoliamo la retta di regressione di x su y. Tale retta ha equazione

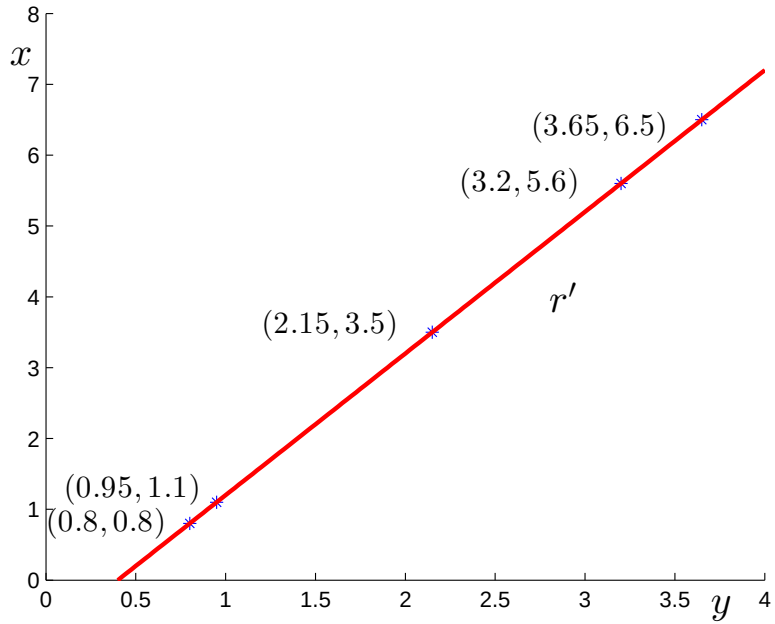
$$x = m'y + q'$$

dove

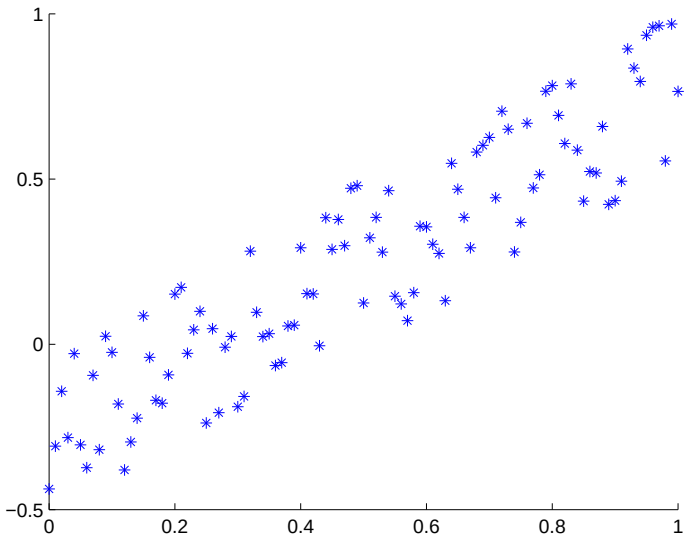
$$m' = \frac{\sum xy - n\bar{x}\bar{y}}{\sum y^2 - n\bar{y}^2} \qquad q' = \frac{\bar{x}\sum y^2 - \bar{y}\sum xy}{\sum y^2 - n\bar{y}^2}$$

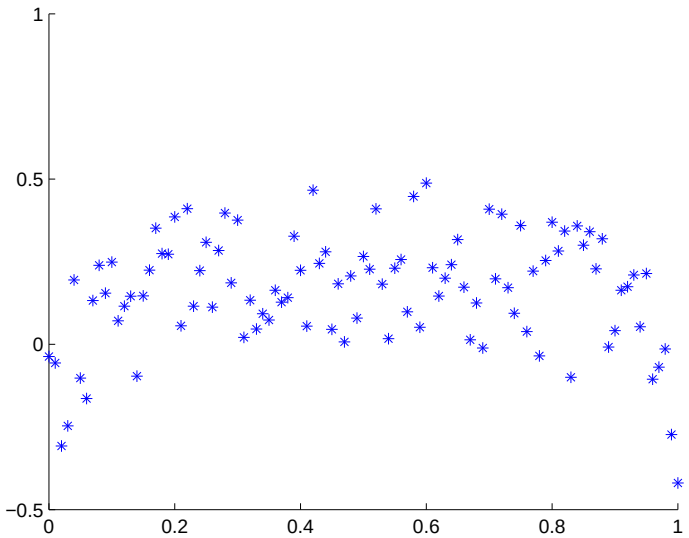
Se tutti i punti dati giacciono esattamente su una retta r le due rette di regressione coincideranno con la retta r: la ragione di ciò giace nel fatto che, nel calcolare la retta di regressione di y su x, le deviazioni verticali di questi punti dalla retta sono minimizzate, così come sono minimizzate le deviazioni verticali dei punti dalla retta di regressione di x su y.





Nonostante le formule derivate siano intese a produrre una retta per un insieme di punti dati, esse possono essere applicate a qualsiasi insieme di punti del piano, ad esempio ai punti della figure sottostanti





Sorge quindi il problema della misurazione del grado di linearità dei punti dati, cioè di quanto bene la retta prodotta si adatta ai dati, ovvero della misurazione della probabilità che i punti dati provengano da un fenomeno avente una legge lineare.

Partiamo dal fatto che se i punti dati sono allineati (e quindi y dipende linearmente da x) la retta di regressione di y su x di equazione $y = mx + q$ e la retta di regressione di x su y di equazione $x = m'y + q'$ sono la stessa retta obliqua del piano. In questo caso i coefficienti angolari delle due rette

$$m = \frac{\Delta y}{\Delta x}, \quad m' = \frac{\Delta x}{\Delta y}$$

sono uno il reciproco dell'altro per cui

$$m \cdot m' = 1.$$

Nel caso limite in cui y è indipendente da x la retta $y = mx + q$ è orizzontale e

$$m = 0.$$

Analogamente, se x è indipendente da y allora (nel piano yOx)

$$m' = 0.$$

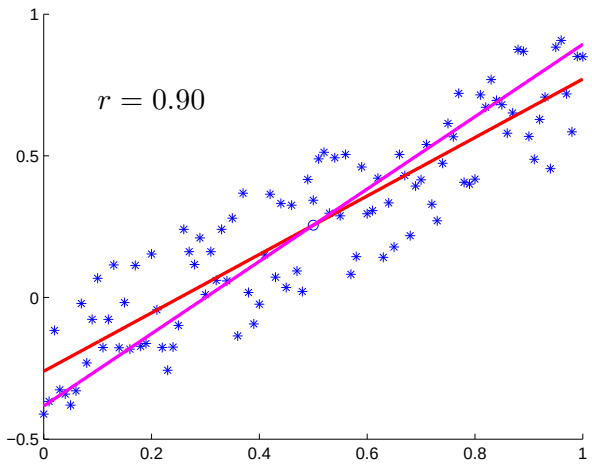
Misuriamo quindi il grado di linearità tra x e y attraverso il *coefficiente di correlazione* tra le due variabili x e y definito mediante

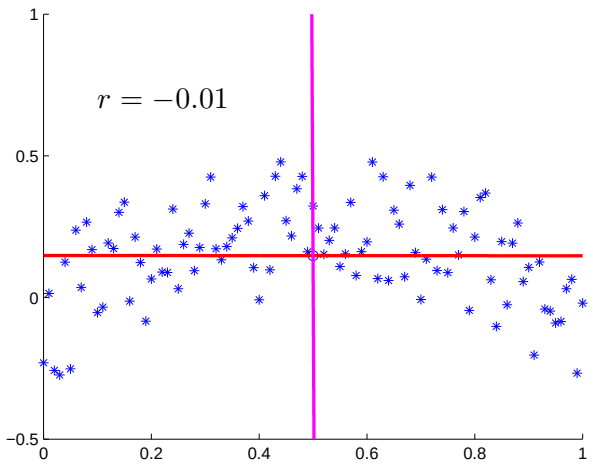
$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\left(\sum x^2 - n\bar{x}^2\right)^{1/2}\left(\sum y^2 - n\bar{y}^2\right)^{1/2}} = \pm\sqrt{m \cdot m'}$$

Si dimostra che

- r assume valori nell'intervallo $[-1, 1]$;
- $r > 0$ se, e solo se la relazione tra x e y è diretta, (a valori più grandi di x corrispondono valori più grandi di y);
- $r < 0$ se, e solo se relazione tra y e x è inversa (a valori più grandi di x corrispondono valori più piccoli di y);
- se $r = \pm 1$ allora la regressione è perfetta e in questo caso i punti sono allineati;
- se $r = 0$ non c'è dipendenza lineare tra le due variabili.

Il coefficiente di correlazione è legato al coseno dell'angolo formato dalle due rette di regressione, come evidenziato dalle seguenti figure





Allo scopo di rendere più esplicito questo legame, introduciamo alcune grandezze statistiche che entrano in gioco nella teoria della correlazione. Date n coppie (x_i, y_i) di una rilevazione statistica su due variabili X e Y, calcolate le medie

$$\bar{x} = \frac{\sum x_i}{n} \text{ e } \bar{y} = \frac{\sum y_i}{n}$$

ricaviamo tutti gli scarti $x_i' = x_i - \bar{x}$ e $y_i' = y_i - \bar{y}$ dai valori medi $\bar{x}$ e $\bar{y}$.
La covarianza di X e di Y è la media dei prodotti degli scarti:

$$\sigma_{XY} = \frac{\sum x_i' y_i'}{n}$$

I numeri

$$\sigma_X = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \text{ e } \sigma_Y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}}$$

sono le deviazioni standard di X e Y.
Si dimostra facilmente che

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_X}.$$

Introduciamo quindi i vettori degli scarti

$$X' = (x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}) \text{ e } Y' = (y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_n - \bar{y})$$

Notiamo che

- σ_{XY} è uguale al prodotto scalare $X' \cdot Y' = \sum x_i' y_i'$ dei vettori degli scarti X', Y' diviso n:
$$\sigma_{XY} = \frac{X' \cdot Y'}{n}.$$
- le deviazioni standard di X e Y sono uguali rispettivamente alle lunghezze dei vettori degli scarti $\|X'\| = \sqrt{\sum (x_i - \bar{x})^2}$, $\|Y'\| = \sqrt{\sum (y_i - \bar{y})^2}$ diviso $\sqrt{n}$

$$\sigma_X = \frac{\|X'\|}{\sqrt{n}}, \qquad \sigma_Y = \frac{\|Y'\|}{\sqrt{n}}$$

Quindi il coefficiente di correlazione r è uguale al coseno dell'angolo θ tra i due vettori X', Y':

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_X} = \frac{X' \cdot Y'}{\|X'\| \|Y'\|} = \cos\theta$$

Esempi di nuvole di dati e correlazioni relative