Spesso ad una fissata popolazione vengono associate coppie di dati: età e peso, età e reddito, reddito e scelte politiche, ecc.). In questo caso l'insieme dei dati è dato da coppie numeriche

$$(x_1, y_1), \dots, (x_N, y_N).$$

In genere, non ci sarà una legge funzionale precisa che leghi le y_i alle x_i. Ma in certi casi potrà succedere che a valori $x_i < \bar{x}$ corrispondano valori $y_i < \bar{y}$, e che viceversa quando $x_i > \bar{x}$ si abbia anche $y_i > \bar{y}$. Oppure potrà succedere l'opposto, cioè che per $x_i < \bar{x}$ si abbia $y_i > \bar{y}$ e per $x_i > \bar{x}$ si abbia $y_i < \bar{y}$. O magari non succederà nulla di tutto ciò.

Una misura numerica del modo in cui le x_i si associano alle y_i è data dalla <u>covarianza</u> di X e Y,

$$\mathrm{Cov}(X,Y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}).$$

Nei due casi prima prospettati, il 1° corrisponde a una covarianza positiva, il 2° a una covarianza negativa.

Se $\mathrm{Cov}(X,Y) = 0$ si dice che X e Y sono <u>non correlate</u>.

Elenchiamo le proprietà della covarianza:

$$\mathrm{Cov}(X,Y) = \frac{1}{N} \sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y} = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

$$\mathrm{Cov}(X,X) = \sigma_X^2 = \mathrm{Var}[X],$$

$$\mathrm{Cov}(aX+b, cY+d) = ac\ \mathrm{Cov}(X,Y);$$

Inoltre

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{Cov}(X,Y).$$

Il numero

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y}$$

si chiama coefficiente di correlazione di X e Y.

Si ha $\rho(aX+b, cY+d) = \rho(X,Y)$ se $ac>0$. Inoltre

$$-1 \le \rho(X,Y) \le 1:$$

infatti per la disuguaglianza di Cauchy-Schwarz

$$|\mathrm{Cov}(X,Y)| = \frac{1}{N}\left|\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})\right| \le$$

$$\le \frac{1}{N}\left[\sum_{i=1}^{N}(x_i-\bar{x})^2\right]^{1/2}\left[\sum_{i=1}^{N}(y_i-\bar{y})^2\right]^{1/2} =$$

$$= \sigma_X\,\sigma_Y.$$

È facile provare, inoltre, che

$$|\rho(X,Y)| = 1 \iff \exists\, a\neq 0,\ b\in\mathbb{R}:\ Y = aX+b.$$

Infatti, se $Y = aX+b$, dalle proprietà della covarianza segue

$$\mathrm{Cov}(X,Y) = a\cdot\mathrm{Cov}(X,X) = a\,\mathrm{Var}[X];$$

d'altronde

$$\mathrm{Var}[Y] = a^2\,\mathrm{Var}[X],$$

cosicché

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y} = \frac{a}{|a|} = \begin{cases} +1 & \text{se } a>0 \\ -1 & \text{se } a<0. \end{cases}$$

Viceversa, se $\rho = 1$ allora

$$\mathrm{Var}\left[\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right] = \frac{\mathrm{Var}[X]}{\sigma_X^2} + \frac{\mathrm{Var}[Y]}{\sigma_Y^2} - 2\frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = 2(1-\rho) = 0,$$

per cui

$$\rightarrow \quad \frac{x_i}{\sigma_X} = \frac{y_i}{\sigma_X}, \quad i = 1, \dots, N$$

ossia esiste $c \in \mathbb{R}$ tale che

$$\frac{x_i}{\sigma_X} - \frac{y_i}{\sigma_Y} = c, \quad i = 1, \dots, N$$

Dunque tutte le coppie (x_i, y_i) appartengono alla retta

$$\frac{x}{\sigma_X} - \frac{y}{\sigma_Y} = c,$$

ossia

$$y = \frac{\sigma_Y}{\sigma_X} x - c\sigma_Y,$$

vale a dire

$$y = aX + b \quad \text{con } a = \frac{\sigma_Y}{\sigma_X}, \; b = -c\sigma_Y.$$

Similmente, se $\rho = -1$ si trova

$$\mathrm{Var}\left[\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right] = 2(1+\rho) = 0,$$

e analogamente si arriva a

$$y = aX + b \quad \text{con } a = \frac{\sigma_Y}{\sigma_X}, \; b = c\sigma_Y.$$

Il coefficiente di correlazione esprime una misura della dipendenza lineare fra X e Y. L'insieme delle coppie (x_i, y_i) è il <u>grafico di dispersione</u>: se l'insieme di questi punti è

molto addensato intorno a una retta, allora $\rho(X,Y)$ sarà vicino a $+1$ o a -1, a seconda che il coefficiente angolare di tale retta sia positivo o negativo. Se invece l'insieme dei punti (x_i, y_i) è abbastanza "rotondo", allora $\rho(X,Y)$ sarà vicino a 0.

Esempio L'insieme dei dati

$$(-2,4),\ (-1,1),\ (0,0),\ (1,1),\ (2,4)$$

soddisfa la relazione $Y = X^2$, e si ha

$$\mathbb{E}[X]=0,\quad \mathbb{E}[Y]=2,\quad \mathrm{Var}[X]=2,\quad \mathrm{Var}[Y]=\frac{11}{5},\quad \mathrm{Cov}(X,Y)=0,$$

e dunque $\rho(X,Y)=0$.

In molte applicazioni, le variabili X, Y possono essere legate da una relazione lineare $Y = aX+b$, ma, a causa di errori di misura, non è possibile determinare a, b; in altri casi, tale legame non è precisamente lineare. Tuttavia, se si ritiene che il legame statistico fra X e Y sia approssimabile con un'opportuna funzione lineare $Y = aX+b$, si deve cercare un metodo per selezionare la retta "migliore" fra le infinite possibili.

La logica è la seguente: se i punti (x_i, y_i) appartenessero

tutti alla stessa retta $y=ax+b$, avremmo

$$(y_i - ax_i - b)^2 = 0 \quad , \; i=1,\dots,N.$$

Se una tale retta non esiste, ci si accontenta di determinare la retta che rende minima la funzione

$$f(a,b) = \sum_{i=1}^{N} (y_i - ax_i - b)^2.$$

Tale retta si chiama <u>retta di regressione di Y su X</u> e il metodo è quello <u>dei minimi quadrati</u>. Per trovare i coefficienti di a e b basterà annullare il gradiente di f:

$$\nabla f(a,b) = 0 \iff \begin{cases} -2 \sum_{i=1}^{N} x_i (y_i - ax_i - b) = 0 \\ -2 \sum_{i=1}^{N} (y_i - ax_i - b) = 0 \, , \end{cases}$$

ossia

$$\begin{cases} \sum_{i=1}^{N} x_i y_i - a \sum_{i=1}^{N} x_i^2 - b \sum_{i=1}^{N} x_i = 0 \\ \sum_{i=1}^{N} y_i - a \sum_{i=1}^{N} x_i - Nb = 0 \, . \end{cases}$$

Ciò equivale a

$$\begin{cases} N\,\mathbb{E}[XY] - Na\,\mathbb{E}[X^2] - Nb\,\mathbb{E}[X] = 0 \\ N\,\mathbb{E}[Y] - Na\,\mathbb{E}[X] - Nb = 0 \, , \end{cases}$$

da cui, semplificando N e ricavando b dalla 2^a equazione, si ottiene

$$\begin{cases} a = \dfrac{\mathrm{Cov}(X,Y)}{\sigma_X^2} = \rho(X,Y)\,\dfrac{\sigma_Y}{\sigma_X} \\ b = \mathbb{E}[Y] - \mathbb{E}[X]\,\rho(X,Y)\,\dfrac{\sigma_Y}{\sigma_X} \, . \end{cases}$$

La retta di regressione è dunque

$$y - \mathbb{E}[Y] = \rho \frac{\sigma_Y}{\sigma_X} (x - \mathbb{E}[X]),$$

ovvero, in forma più simmetrica,

$$\frac{y - \mathbb{E}[Y]}{\sigma_Y} = \rho \frac{x - \mathbb{E}[X]}{\sigma_X}.$$

In modo completamente analogo, la retta di regressione di X su Y, che ha forma $x = ay + b$, è data da

$$x - \mathbb{E}[X] = \rho \frac{\sigma_X}{\sigma_Y} (y - \mathbb{E}[Y])$$

ovvero

$$\frac{x - \mathbb{E}[X]}{\sigma_X} = \rho \frac{y - \mathbb{E}[Y]}{\sigma_Y}.$$

Si noti che il punto $(\mathbb{E}[X], \mathbb{E}[Y])$ appartiene sempre alla retta di regressione.

Si noti infine che

$$\rho(X,Y) = \pm 1 \iff \mathrm{Var}\left[\frac{Y - \mathbb{E}[Y]}{\sigma_Y} - \rho \frac{X - \mathbb{E}[X]}{\sigma_X}\right] = 0.$$