

Analisi Multivariata

Corso di laurea in Statistica. A.A. 2009/2010

Carla Rampichini

1 Covarianza

La covarianza è una misura di dipendenza tra due variabili casuali. Date due variabili casuali X e Y la covarianza teorica tra X e Y è definita:

$$\sigma_{XY} = Cov(X, Y) = E(XY) - E(X)E(Y) \quad (1)$$

dove $E(X) = \int xf(x)dx$, $E(XY) = \int \int xyf(xy)dxdy$

Se X e Y sono indipendenti, la covarianza $Cov(X, Y)$ è uguale a zero, in generale il viceversa non è vero.

La covarianza di X con se stessa è la varianza di X :

$$\sigma_{XX} = Cov(X, X) = Var(X)$$

Se X è una variabile multivariata a p dimensioni, cioè se $\mathbf{X} = [X_1 \dots X_p]'$, allora le covarianze teoriche tra gli elementi di $\mathbf{X}$ possono essere organizzate in una matrice detta matrice di covarianza

$$\Sigma = \begin{pmatrix} \sigma_{11} & \dots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{1p} & \dots & \sigma_{pp} \end{pmatrix}$$

Vedremo più avanti le proprietà della matrice di covarianza. La versione empirica della varianza e della covarianza è data da:

$$s_{XY} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (2)$$

$$s_{XX} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3)$$

Per n piccolo, diciamo $n \leq 20$, si sostituisce al fattore $1/n$ che compare nella (2) e nella (3) il fattore $1/(n-1)$ per correggere la *distorsione* (o *bias*). Per

una variabile casuale $\mathbf{X}$ p -dimensionale la matrice di covarianza empirica è data da:

$$\mathbf{S} = \begin{pmatrix} s_{11} & \dots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{1p} & \dots & s_{pp} \end{pmatrix}$$

In uno scatterplot bi-dimensionale la covarianza fornisce una misura di quanto lo scatterplot sia simile a una retta, deve essere quindi chiaro che la covarianza misura solo la dipendenza lineare tra due variabili.

1.1 Esempio

Il manager di un grande negozio di abbigliamento sta analizzando le vendite di pullover classici di colore blu in 10 differenti periodi dell'anno. Il manager dispone dei dati relativi al numero di pullover venduti X_1 , alla variazione del prezzo X_2 , in euro, al costo della pubblicità apparsa su quotidiani a diffusione locale X_3 (in euro) e alla presenza di commesso nel reparto X_4 (in ore per periodo). I dati raccolti sono riportati nella matrice dei dati seguente.

$$\mathbf{X} = \begin{pmatrix} 230 & 125 & 200 & 109 \\ 181 & 99 & 55 & 107 \\ 165 & 97 & 105 & 98 \\ 150 & 115 & 85 & 71 \\ 97 & 120 & 0 & 82 \\ 192 & 100 & 150 & 103 \\ 181 & 80 & 85 & 111 \\ 189 & 90 & 120 & 93 \\ 172 & 95 & 110 & 86 \\ 170 & 125 & 130 & 78 \end{pmatrix}$$

Il manager è convinto che il prezzo influenzi il numero di pullover venduti. Per valutare questa ipotesi, disegna lo scatterplot di X_2 vs X_1 . A una prima occhiata sembra che la nuvola di punti sia inclinata negativamente. Il calcolo della covarianza empirica fornisce il risultato seguente:

$$s_{12} = \frac{1}{9} \sum_{i=1}^{10} (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) = -80.02$$

un valore negativo, come atteso in base al grafico. Si noti che il valore della covarianza dipende dall'unità di misura. Così, se i prezzi questo esempio fossero in Yen (JPY), avremmo ottenuto un valore della covarianza differente.