Momento del secondo ordine

$$E[XY] = \int_{-\infty}^{+\infty} xy f_{XY}(x,y)dxdy$$

Se non si conosce f_{XY} è possibile stimare il momento congiunto del secondo ordine dai dati. Date *n* osservazioni delle variabili X e Y, $x = (x_1, x_2, \cdots, x_n)$ e $y = (y_1, y_2, \cdots, y_n)$ possiamo scrivere

$$E[XY] = \frac{1}{n}\sum_{k=1}^{n} x_k y_k \qquad (1)$$

Matlab

```
x=[x1 x2 x3 x4 … xn];
y=[y1 y2 y3 y4 … yn];

E[xy]=x*y'/n;
```

Covarianza

$$C_{X,Y} = E[(X - E[X])(Y - E[Y])]$$

con $E[X] = \eta_X$ e $E[Y] = \eta_Y$ valori medi di *X* e *Y*.

Se non si conoscono le distribuzioni marginali e congiunte si possono stimare i valori medi e la covarianza dai dati.

$$E[X] = \frac{1}{n}\sum_{k=1}^{n} x_k \quad \text{e} \quad E[Y] = \frac{1}{n}\sum_{k=1}^{n} y_k$$

Il momento congiunto centrale delle variabili rispetto al valor medio è dato da

$$C_{X,Y} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \eta_X)(y_k - \eta_Y) \qquad (2)$$

Si dimostra che la stima non polarizzata della covarianza delle variabili *X* e *Y* è data da

$$C_{X,Y} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \eta_X)(y_k - \eta_Y) \qquad (3)$$

Matlab

```
x=[x1 x2 x3 x4 … xn];
y=[y1 y2 y3 y4 … yn];
```

Il momento congiunto centrale in (2) si stima nel modo seguente

```
c_xy=cov(x,y,1);
```

La stima della covarianza in (3) si stima nel modo seguente

c_xy=cov(x,y,0) o c_xy=cov(x,y)

N.B. Nel seguito faremo riferimento alla notazione in (2)

Coefficiente di Correlazione

$$\rho_{XY} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[XY] - E[X]E[Y]}{\sqrt{E\left[(X-\eta_X)^2\right]E[(Y-\eta_Y)2]}}$$

dove la deviazione standard σ_X è la radice quadrata della varianza

$$\sigma_X^2 = E\left[(X-\eta_X)^2\right]$$

Se non si conoscono le distribuzioni marginali e congiunte si possono le deviazioni standard dai dati

Il momento centrale del secondo ordine è dato da

$$\sigma_X^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \eta_X)^2 \qquad (4)$$

Si dimostra che la stima non polarizzata della varianza è data da

$$\sigma_X^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \eta_X)^2 \qquad (5)$$

N.B. Nel seguito faremo riferimento alla notazione in (4)

Matlab

x=[x1 x2 x3 x4 … xn];
y=[y1 y2 y3 y4 … yn];

Tramite il comando C=corrcoef(x,y) viene stimata una matrice C di dimensione 2x2 così formata

$$C = \begin{pmatrix} \rho_{XX} & \rho_{YX} \\ \rho_{XY} & \rho_{YY} \end{pmatrix}$$

N.B. L'elemento (i,j) è calcolato come $\frac{\mathrm{cov}(Variabile\_i, Variabile\_j, 1)}{\mathrm{std}(Variabile\_i,1)*\mathrm{std}(Variabile\_j,1)}$ dove $\begin{matrix} Variabile\_i = X \\ Variabile\_j = Y \end{matrix}$

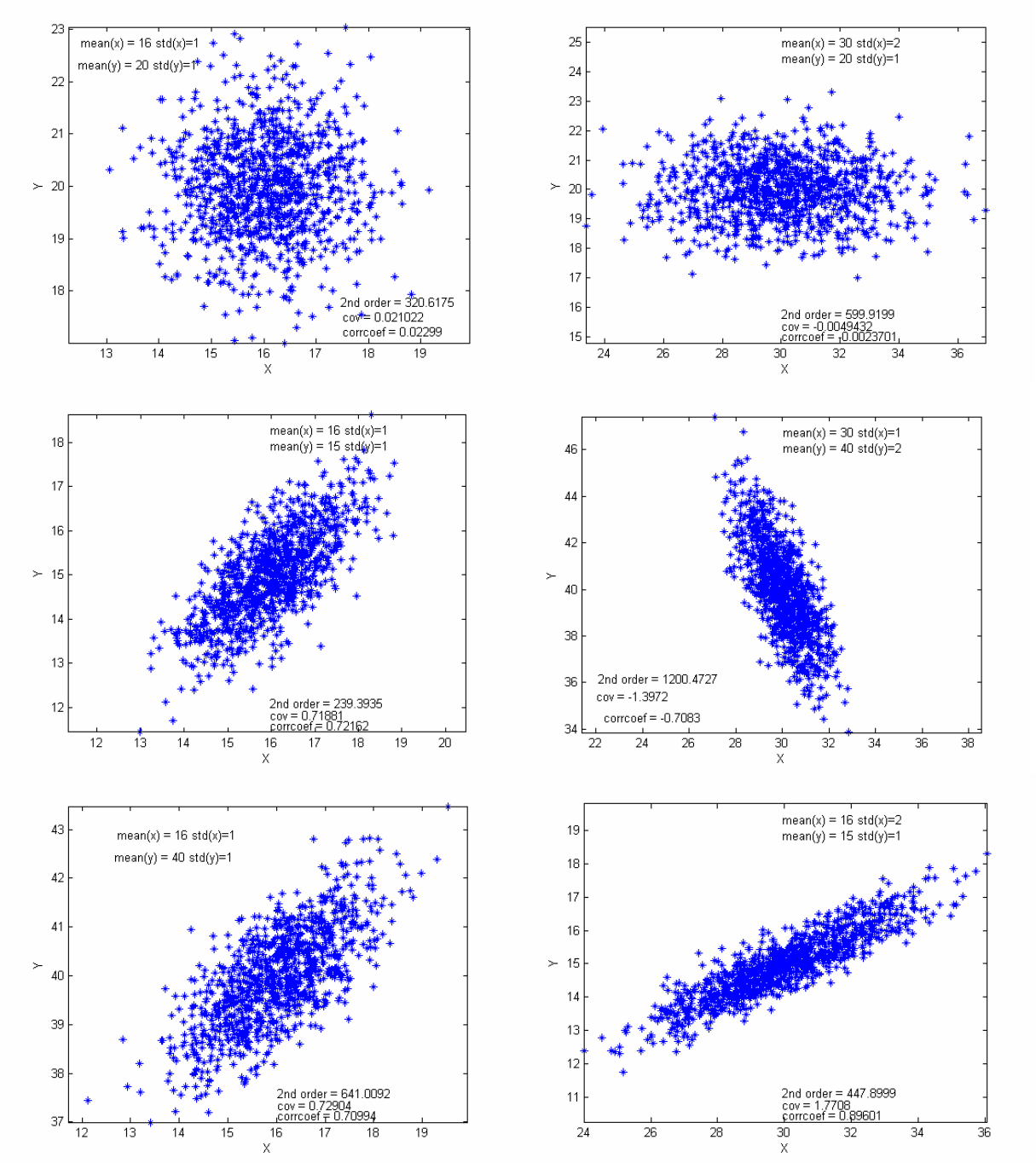
o equivalentemente $\frac{\mathrm{cov}(Variabile\_i, Variabile\_j)}{\mathrm{std}(Variabile\_i)*\mathrm{std}(Variabile\_j)}$

Alcuni esempi

Nei grafici seguenti sono mostrati gli scatter plot relativi a variabili aleatorie X e Y a diverso valore medio, deviazione standard e coefficiente di correlazione.
Una volta generati i due vettori

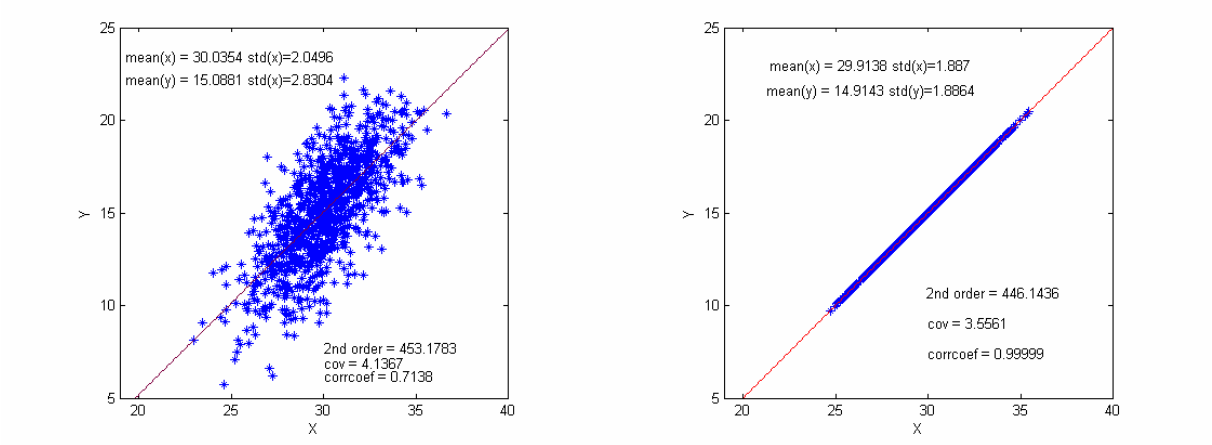
x=[x1 x2 x3 x4 … xn];
y=[y1 y2 y3 y4 … yn];

ogni punto del piano ha coordinate (xi,yi). In Matlab un grafico di questo tipo si può ottenere disegnando un simbolo, ad esempio un '*' in tali punti, eseguendo il comando plot(x,y,'*')



Si deve notare come i valori medi influenzino la posizione dei dati, mentre il coefficiente di correlazione insieme alle deviazioni standard siano legate al tipo di distribuzione spaziale dei dati.

Si deve precisare che il coefficiente di correlazione non è la pendenza della retta sulla quale si distribuiscono i dati. Si vedano a tale proposito i seguenti grafici. La retta può essere stimata attraverso la regressioni lineare. Vedremo in seguito la relazione che li lega.



Regressione

Nel modello di regressione lineare si assume una relazione di tipo lineare tra il valore medio della variabile dipendente Y da quello indipendente per cui

$$E(Y \mid X) = \eta_{Y|X} = a + b * x$$

Il modello si scrive come

$$y = a + b * x + \varepsilon$$

Dove ε è l'errore gaussiano con varianza σ^2 e valore medio nullo.

Dati due vettori delle osservazioni x e y, per gli elemento *i*-esimi dei vettori si può scrivere:

yi=a+b*xi+εi

per cui dati

x=[x1 x2 x3 x4 … xn];
y=[y1 y2 y3 y4 … yn];

si hanno n relazioni

y1=a+b*x1+ε1
y2=a+b*x2+ε2
.
.
.
yn=a+b*xn+εn

Usando la notazione matriciale si può scrivere:

y=a+bx+ε

$\begin{bmatrix} y1 \\ \vdots \\ yn \end{bmatrix} = \begin{bmatrix} 1 & x1 \\ \vdots & \vdots \\ 1 & xn \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} \varepsilon 1 \\ \vdots \\ \varepsilon n \end{bmatrix}$ con $X = \begin{bmatrix} 1 & x1 \\ \vdots & \vdots \\ 1 & xn \end{bmatrix}$ la matrice nota del modello

e $\beta = [a \quad b]$' la matrice incognita dei coefficienti

In Matlab tale matrice può essere stimata come

beta=X\y

gli elementi di beta sono tali da minimizzare l'errore quadratico medio $\sum_{k=1}^{n} (yk - (a + b * xk))^2$

Relazione tra coefficiente di correlazione e pendenza della retta di regressione

Date la variabile dipendente y e la variabile indipendente x il coefficiente di correlazione ρ_{xy} è legato alla pendenza della retta dalla seguente relazione

$$b = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)} = \rho_{xy} \frac{std(y)}{std(x)}$$