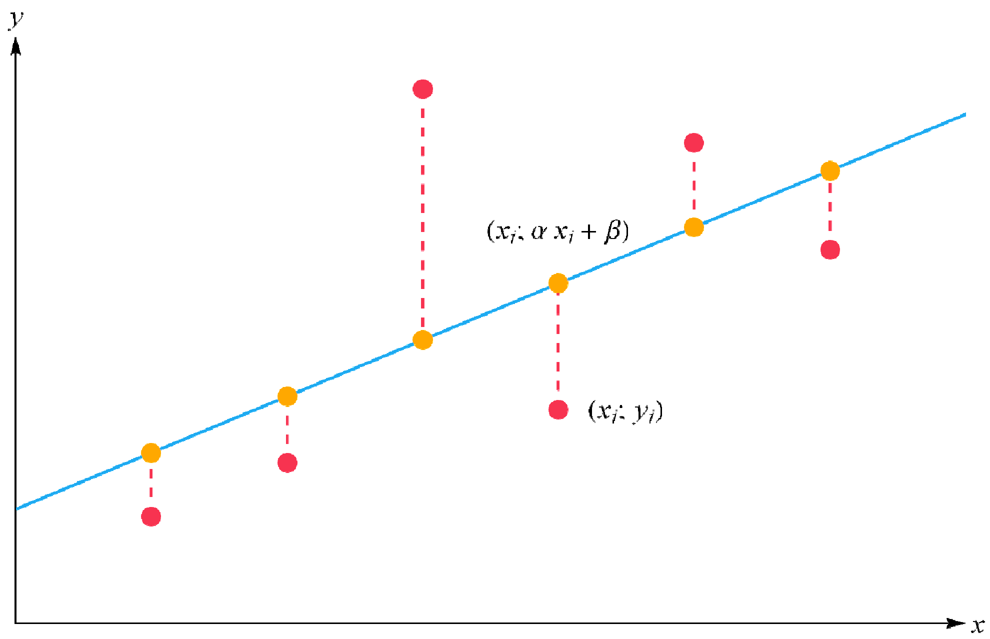
Un problema di minimo in statistica

Dati nel piano n punti $P_i = (x_i , y_i)$ con i valori x_i non tutti uguali, cerchiamo una retta $y = \alpha x + \beta$ che renda minima la funzione

$$E (\alpha , \beta) = \sum_{i=1}^{n} (y_i - \alpha x_i - \beta)^2$$

che rappresenta la somma dei quadrati delle distanze dei punti P_i dai corrispondenti punti $(x_i , \alpha x_i + \beta)$ sulla retta.



Poiché $E (\alpha , \beta) \to +\infty$ se $\| (\alpha , \beta) \| \to +\infty$, il minimo esiste.

In statistica a questa retta si dà il nome di retta di regressione.

Cerchiamo i punti stazionari della funzione. Essendo :

$$D_\alpha E = -2 \sum_{i=1}^{n} (y_i - \alpha x_i - \beta) x_i \qquad D_\beta E = -2 \sum_{i=1}^{n} (y_i - \alpha x_i - \beta)$$

dobbiamo studiare il sistema

$$\sum_{i=1}^{n} x_i y_i - \alpha \sum_{i=1}^{n} x_i^2 - \beta \sum_{i=1}^{n} x_i = 0 \qquad \sum_{i=1}^{n} y_i - \alpha \sum_{i=1}^{n} x_i - n\beta = 0 .$$

Dividiamo ambo i membri per n e poniamo:

$$\frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$ media degli x_i

$$\frac{\sum_{i=1}^{n} x_i^2}{n} = \bar{q}$$ media dei quadrati

$$\frac{\sum_{i=1}^{n} x_i\, y_i}{n} = \bar{p}$$ media dei prodotti.

Il sistema assume la forma

$$\begin{cases} \bar{q}\,\alpha + \bar{x}\,\beta = \bar{p} \\ \bar{x}\,\alpha + \beta = \bar{y} \end{cases}.$$

Poiché la matrice dei coefficienti ha determinante non nullo, in particolare positivo (vedere nota successiva), il sistema ha un'unica soluzione: il punto corrispondente è dunque quello di minimo. Si osservi che la matrice dei coefficienti è (a meno del fattore moltiplicativo 2 n) la matrice hessiana (costante) della funzione E: che il punto stazionario sia dunque di minimo (almeno localmente) è confermato dai risultati noti.

Ad esempio, per i punti (1 , 1) , (2 , 0) , (3 , 2) si ha : $\bar{x}=2$, $\bar{y}=1$, $\bar{p}=7/3$, $\bar{q}=14/3$ e dunque il sistema da risolvere diventa: 14/3 α + 2 β = 7/3 , 2 α + β = 1; la soluzione α = ½ , β = 0 fornisce la retta di regressione y = ½ x.

Nota

Dobbiamo provare che $\bar{q} > \bar{x}^2$.

Poiché abbiamo supposto che gli x_i non sono tutti uguali, le differenze x_i - $\bar{x}$ non sono tutte nulle; dunque:

$$\frac{\sum_{i=1}^{n} x_i}{n} + \frac{\sum_{i=1}^{n} \bar{x}^2}{n} - 2\,\bar{x}\,\frac{\sum_{i=1}^{n} x_i}{n} = \bar{q} + \bar{x}^2 - 2\,\bar{x}^2 = \bar{q} - \bar{x}^2 .$$