Regressione e correlazione

TEMI PROPOSTI:

- **Concetto di associazione**
- **Rappresentazione grafica**
- **Modello lineare in due variabili**
  - ***Stima dei coefficienti di regressione***
- **Correlazione**
  - ***Coefficiente di correlazione***

Concetto di associazione

Se si è interessati all'associazione tra due insiemi di dati quantitativi occorre trovare una misura appropriata per determinare il grado di associazione o *correlazione*, come viene chiamato, tra i due insiemi di variabili.

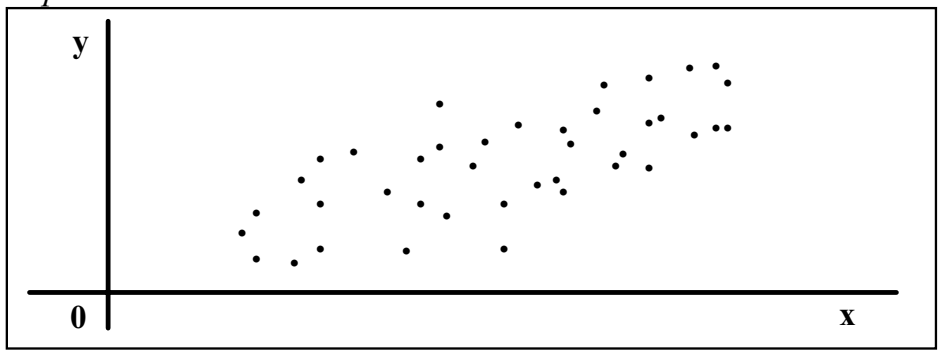
Esempi

Alcuni esempi di variabili correlate possono essere i seguenti:

- temperatura massima giornaliera in due città;
- quantità di azoto somministrata ad un appezzamento e resa di grano ottenuta da questo;
- altezza di una certa pianta ed età della stessa;
- quantità di pioggia caduta e produzione di grano;

Rappresentazione grafica della associazione

Nella pratica quotidiana, si nota l'esistenza di casi in cui tra due variabili pur *i valori dell'una tendono a disporsi in un modo che risente del modo in cui si dispongono i valori dell'altra variabile*. Ad ogni coppia di valori corrisponde un punto nel piano (x, y). Il grafico formato da tutti questi punti è chiamato *diagramma di dispersione* .



Modello lineare in due variabili

Di solito, una delle variabili, che è chiamata variabile di base o *dipendente*, mentre la variabile *indipendente* aiuta a fornire informazioni concernenti la variabile basilare.

$$Y = \alpha + \beta . X$$

Con ***n coppie*** di osservazioni **(x1, y1), (x2, y2), ..., (xn, yn)**.

A causa della presenza di errori di varia natura, la relazione effettiva tra le osservazioni non è esatta. La presenza di dispersione nei dati osservati impone di sostituire alla relazione teorica precedente la forma stocastica:

$$y_i = \alpha + \beta . x_i + u_i \qquad \textbf{(con i = 1, 2, ..., n)}$$

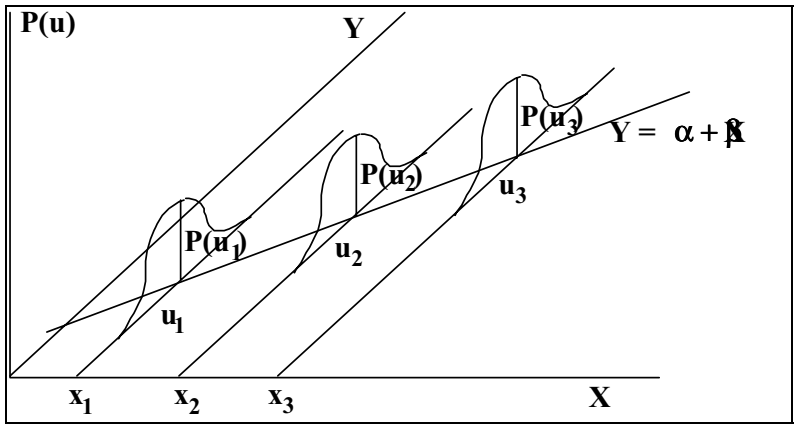
La ipotesi più semplice che si può avanzare riguardo la distribuzione degli errori è la seguente:

- $\mu(u) = 0$;

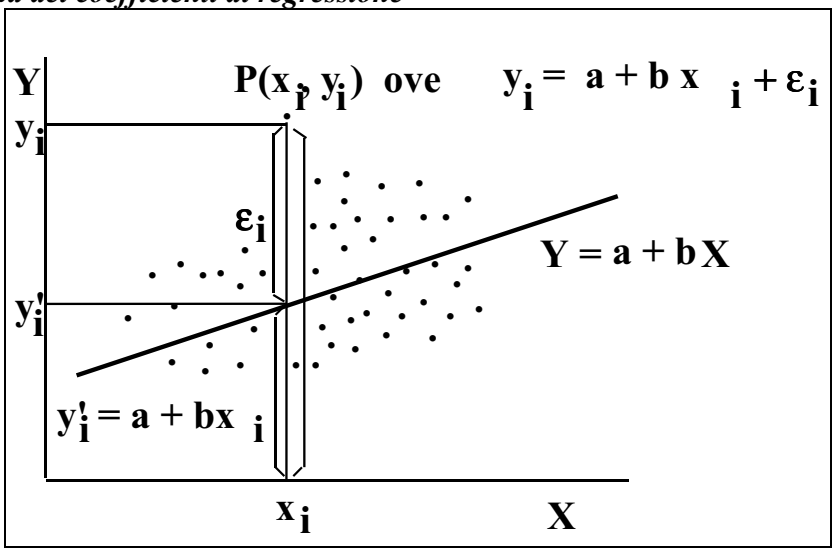
- $\sigma^2(u) = K$ ed indipendente da X;

- $\sigma_{ij} = 0$ per $i \neq j$.

Graficamente si ha:



Stima dei coefficienti di regressione



In realtà, non si conosce la retta $Y = \alpha + \beta . X$!!

Per contro, si dispone di un insieme di ipotesi che affermano quanto segue:

- $y_i = \alpha + \beta . x_i + u_i$ **(con i = 1, 2, ..., n)**
- **$\mu(u) = 0$;**
- **$\sigma^2(u) = K$ ed indipendente da X;**
- **$\sigma_{ij} = 0$ per $i \neq j$.**

α, β e $\sigma^2(u)$ sono parametri incogniti !

L'approccio classico è quello di usare il metodo dei minimi quadrati, secondo il quale, la condizione di minimo per il modello considerato diviene:

$$\sum_{i=1}^{i=n} \varepsilon_i^2 = \sum_{i=1}^{i=n} \left(y_i - y_i'\right)^2 = \sum_{i=1}^{i=n} \left(y_i - a - b . x_i\right)^2 = \text{minimo}$$

Le derivate prime rispetto ai parametri incogniti a e b, annullate e riordinate nelle incognite, forniscono il sistema di equazioni normali:

$$\begin{cases} \sum_{i=1}^{i=n} y_i = a . n + b . \sum_{i=1}^{i=n} x_i \\ \sum_{i=1}^{i=n} y_i . x_i = a . \sum_{i=1}^{i=n} x_i + b . \sum_{i=1}^{i=n} x_i^2 \end{cases}$$

La prima equazione di tale sistema può essere anche scritta

$$\frac{\sum_{i=1}^{i=n} y_i}{n} = a + b . \frac{\sum_{i=1}^{i=n} x_i}{n}$$

La retta interpolatrice passa per il punto medio della variabile statistica (x,y)
Risolvendo si ottiene:

$$b = \frac{n.\sum y_i.x_i - (\sum y_i).(\sum x_i)}{n.\sum x_i^2 - (\sum x_i)^2} = \frac{cov(y,x)}{var(x)} = \frac{s_{y,x}}{s_x^2}$$

$$a = M(y) - b.M(x)$$

Le stime dei minimi quadrati godono di quattro requisiti importanti. Esse sono: *corrette, consistenti, efficienti e sufficienti.*

$\sigma^2(u)$ non è conosciuto.

Si *stima la varianza degli errori* sulla base dei quadrati degli scarti dei valori osservati dalla funzione di regressione.

una stima *corretta* della varianza dell'universo, che non si conosce, si ottiene con la relazione:

$$\overline{\sigma}_u^2 = \frac{1}{(n-2)}.\sum[\varepsilon_i - M(\varepsilon_i)]^2$$

Infatti si ha:

$$E\left[\frac{\sum(\varepsilon_i - \bar{\varepsilon})^2}{n-2}\right] = E\left[\frac{\sum(\varepsilon_i)^2}{n-2}\right] = \sigma^2(\varepsilon)$$

Nel caso del modello lineare,in due variabili, il parametro b indica l'inclinazione della retta di regressione e in statistica è chiamato anche *coefficiente di regressione*.

Esso è una *misura della dipendenza tra variabili*, quando si ipotizza che la variabile esplicativa o indipendente sia l'antecedente della variabile da spiegare o dipendente e, inoltre, si ammette che le *misure della X non siano affette da errore*.

Quando, però, si ha ragione di ritenere che la Y sia la variabile antecedente ed X la risultante, si può ottenere una seconda misura della dipendenza, questa volta di X da Y, cioè un secondo coefficiente di regressione della X alla Y.

Il termine

$$S(x,y) = \frac{\sum(x-\overline{X})(y-\overline{Y})}{n}$$

è chiamato anche *covarianza* tra y ed x.

Siano inoltre :

$$S^2(x) = \frac{\sum(x-\overline{X})^2}{n}$$

$$S^2(y) = \frac{\sum(y-\overline{Y})^2}{n}$$

Quello che finora si è indicato come coefficiente di regressione di Y ad X può rappresentarsi come segue:

$$b_{y/x} = \frac{n.\sum y_i.x_i - (\sum y_i).(\sum x_i)}{n.\sum x_i^2 - (\sum x_i)^2}, = \frac{S(x,y)}{S^2(x)} = \frac{\sum (x-\overline{X})(y-\overline{Y})}{\sum (x-\overline{X})^2}$$

quando si suppone che sia vero il contrario, si può stimare il coefficiente di regressione di X ad Y con la relazione

$$b_{x/y} = \frac{n.\sum y_i.x_i - (\sum y_i).(\sum x_i)}{n.\sum y_i^2 - (\sum y_i)^2}, = \frac{S(x,y)}{S^2(y)} = \frac{\sum (x-\overline{X})(y-\overline{Y})}{\sum (y-\overline{Y})^2}$$

Se si ha ragione di ritenere che tra le due variabili sussista più che la dipendenza di una dall'altra una effettiva *interdipendenza* tra le due, si può assumere come misura di tale interdipendenza la *media geometrica dei due coefficienti di regressione* di Y ad X e di X ad Y. Si perviene così a quello che in letteratura è correntemente denominato *coefficiente di correlazione*, che risulta pari a

$$r_{y,x} = \sqrt{b_{y/x} * b_{x/y}}$$

Il coefficiente di correlazione può essere determinato anche in modo diverso.
Si ha:

$$\boxed{r_{y,x} = \sqrt{b_{y/x} * b_{x/y}} = \frac{S(x,y)}{S(x)S(y)}}$$

Inoltre sia la somma dei quadrati degli scarti :

$$\sum_{i=1}^{i=n} \varepsilon_i^2 = \sum_{i=1}^{i=n} [y_i - (a + bx_i)]^2$$

Si ha:

$$\frac{1}{n} \sum_{i=1}^{i=n} \varepsilon_i^2 = \sigma_y^2 \{1 - r_{y,x}^2\}$$

risulta che è sempre:

$$|r_{y,x}| < 1$$

Se $r_{y,x}$ é uguale ad 1, allora la somma dei quadrati degli scarti sarà uguale a zero e ciò sarà possibile *se e solo se* tutti i punti rappresentativi dei dati osservati giaceranno su una retta.

Se $r_{y,x} = 0$, allora il numeratore del parametro b, che coincide con il numeratore di $r_{y,x}$, sarà uguale a zero e l'equazione che descrive la relazione tra le due variabili si riduce ad y = a.

Ossia Y è indipendente da X e conseguentemente non esiste alcuna relazione tra le due variabili e vi è assenza di correlazione tra la X e la Y.

Il coefficiente di correlazione è un indicatore soddisfacente del grado di interrelazione esistente tra due insiemi di dati quantitativi.

- ***La correlazione*, quindi, *misura l'interdipendenza tra le variabili X ed Y.***

- **Per contro, *la regressione misura la dipendenza di una delle due variabili dall'altra.***