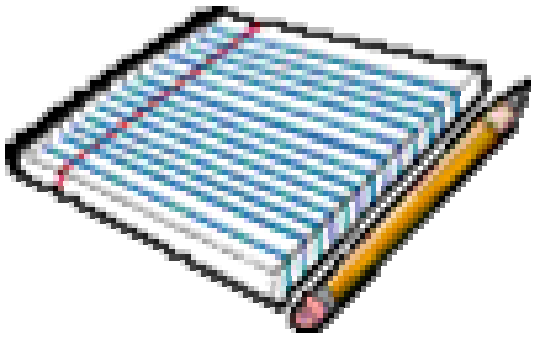
Lezione 9

Correlazione e Regressione lineare



La correlazione

Una tecnica per individuare una *relazione fra due variabili*

Il concetto di correlazione si basa sullo studio del "variare insieme" di almeno due variabili

Il "variare insieme" si definisce COVARIANZA

ESEMPIO

Supponiamo di voler studiare la relazione tra età e peso

Ci aspettiamo che, almeno fino ad una certa età, il peso vari insieme all'età:

più età ⇨ più peso

CORRELAZIONE

- La misura del grado di associazione di due variabili si ottiene sulla base della

COVARIANZA

$$Cov_{XY} = \frac{\sum (X - M_X)(Y - M_Y)}{N}$$

- ⇨ può assumere valori positivi e negativi (direzione)
- ⇨ quando è 0 X e Y sono indipendenti
- ⇨ aumenta al crescere del grado di dipendenza tra X e Y

CORRELAZIONE

Il coefficiente di correlazione r varia

$$-1 \leq r \leq +1$$

misura contemporaneamente

la forza della relazione ⇨ il *valore*

la direzione della relazione ⇨ il *segno*

la forma della relazione ⇨ lineare

CORRELAZIONE

- se $r = \pm 1$ ⇨ relazione lineare *perfetta*
- se $r = 0$ ⇨ *assenza* di relazione lineare
- se $-1 < r < -.50$ ⇨ relazione negativa da *forte→debole*
- se $.50 < r < 1$ ⇨ relazione positiva da *debole→forte*

REGRESSIONE

Uno degli interessi preminenti dello studio del comportamento è quello di poterlo predire sulla base della relazione tra quel comportamento ed un altro

La REGRESSIONE è un insieme di procedure statistiche che consentono di usare le informazioni che si hanno su una variabile per predirne un'altra

REGRESSIONE

- Studiare la relazione tra due variabili significa descrivere in che modo una variabile "dipenda" da un'altra

⇩

- Come una variazione nella variabile **X** (variabile INDIPENDENTE = "causa") influenza il variare di una variabile **Y** (variabile DIPENDENTE= "effetto")

REGRESSIONE

- **Ha due obiettivi:**
  - **misurare il *grado* e il *verso* dell'influenza della variabile indipendente X sulla variabile dipendente Y**
  - **ottenere un'equazione che permetta di *prevedere* il valore della variabile dipendente Y, conoscendo solo quello della indipendente X**

$$Y_i = \alpha + \beta X_i + \varepsilon$$

REGRESSIONE

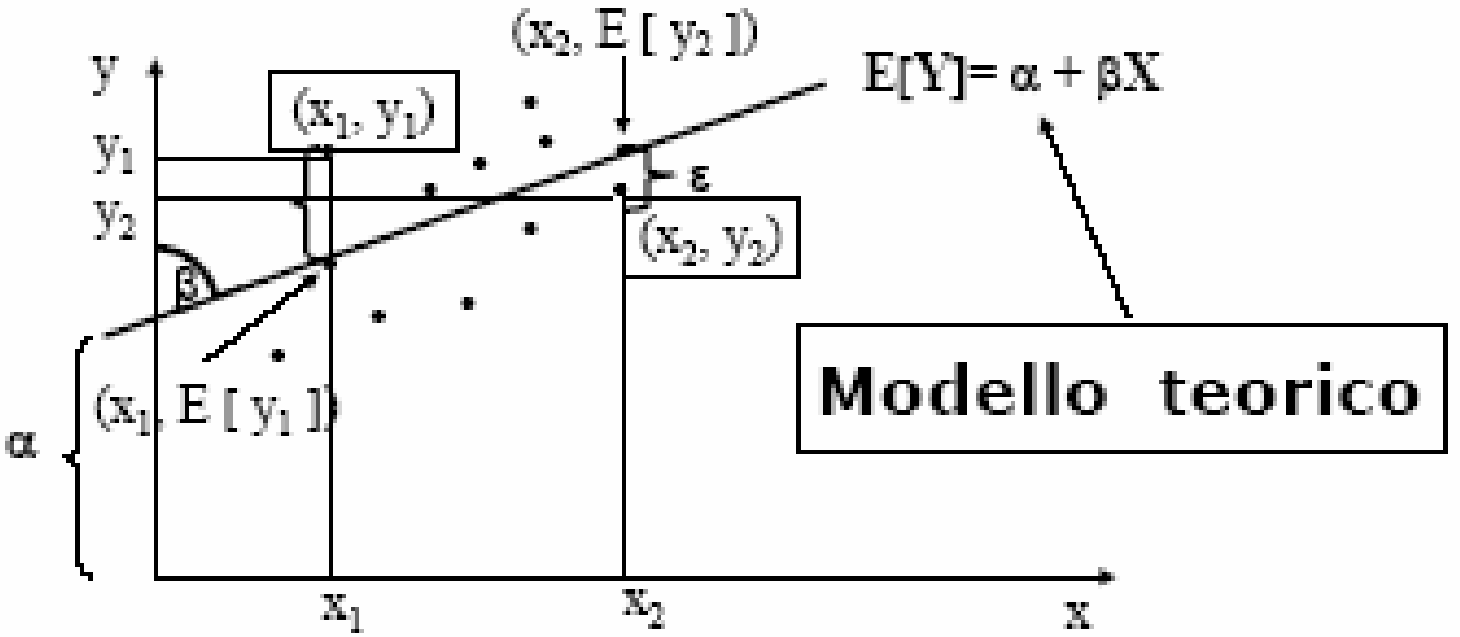
$$Y = \alpha + \beta X + \varepsilon$$

dove: α = *intercetta*, punto in cui la retta incontra l'asse delle *Y*, rappresenta il *valore predetto* di Y in corrispondenza di X uguale zero

β = *coefficiente di regressione*, inclinazione della retta, parametro della popolazione, rappresenta *l'incremento predetto* di Y per un incremento unitario di X

ε = *errore*

REGRESSIONE



REGRESSIONE

- **Relazione forte: Se X è legata ad Y, sapendo i valori di X possiamo prevedere quelli di Y, e osserveremo che la Y ricavata dalla X coincide con la Y osservata**

$$Y_i = \hat{Y}_i = a + bx_i$$

- **Relazione debole: Per legami deboli o inesistenti fra X e Y invece la Y prevista sarà distante dalla Y osservata**

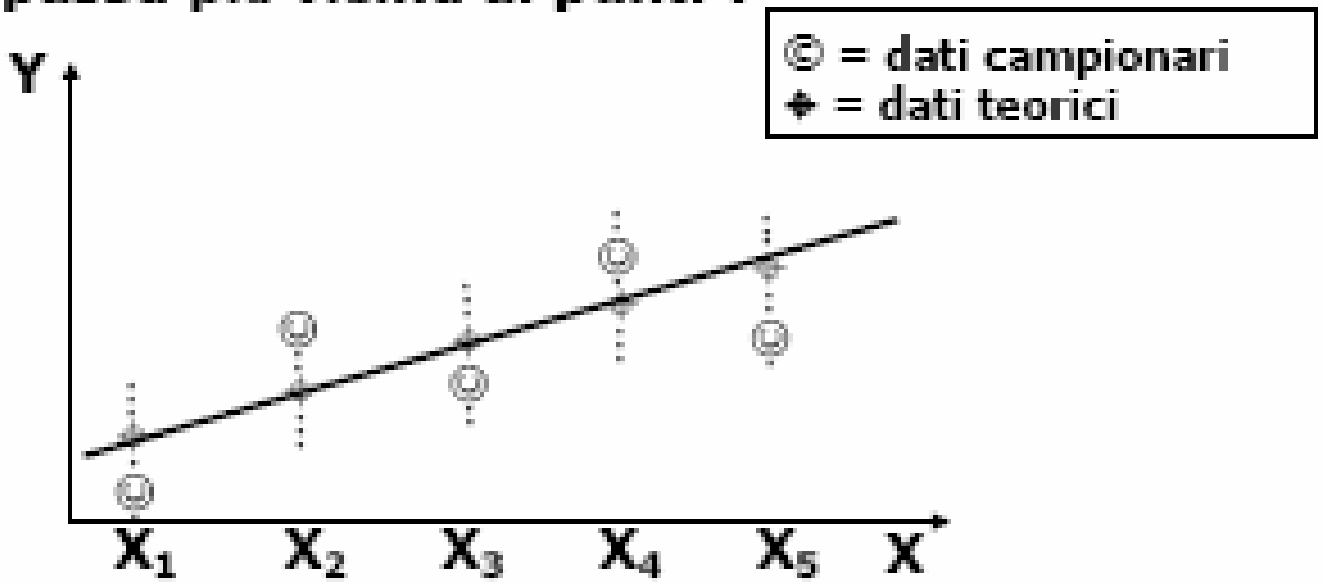
$$Y_i \neq \hat{Y}_i = a + bX_i$$

$$\hat{Y}_i - Y_i = e_i$$

→ Tanto più è debole la relazione e tanto più il valore di e_i sarà elevato

REGRESSIONE

- Come si fa a stabilire qual è la retta che passa più vicina ai punti ?



⇨ *Stima* della retta di regressione:

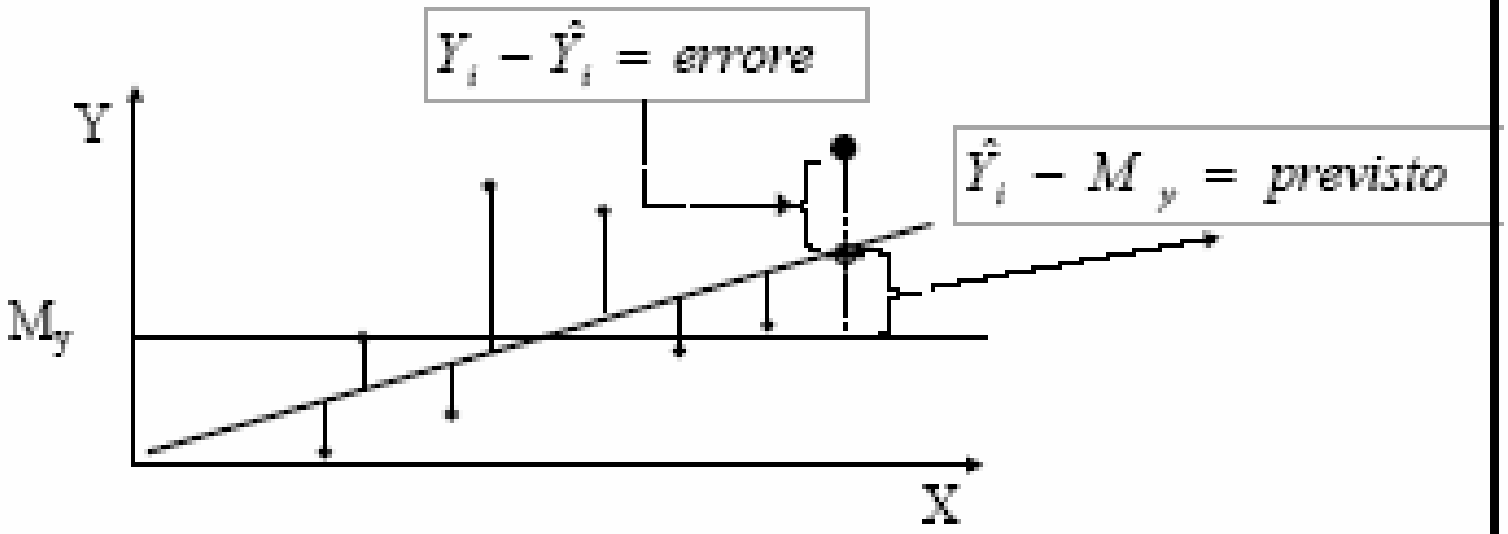
$$Y = a + bX$$

dove: a = stima di α

b = stima di β

⇩

METODO DEI MINIMI QUADRATI: per trovare la retta che rende *minima* la somma degli scarti tra *valori stimati* o *teorici* (⌖) e *valori empirici* o *osservati* (•)



Obiettivo: individuare la retta che rende minimo l'errore e massima la parte prevista

- Attraverso il METODO DEI MINIMI QUADRATI si ricavano *a* e *b*:

$$b = \frac{\sum_{i=1}^{N} (X_i - M_X)(Y_i - M_Y)}{\sum_{i=1}^{N} (X_i - M_X)^2}$$

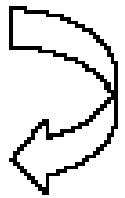
CODEVIANZA: Misura di come *x* e *y* variano insieme. Più forte è la relazione maggiore, in valore assoluto, è la codevianza che può essere negativa. Dividendo per *n* si ottiene la **COVARIANZA**

DEVIANZA di *x* (somma di tutti gli scarti dalla media). Dividendo per *n* si ottiene la **VARIANZA**

$$a = M_Y - bM_x$$

- Il *coefficiente di regressione* può anche essere espresso come:

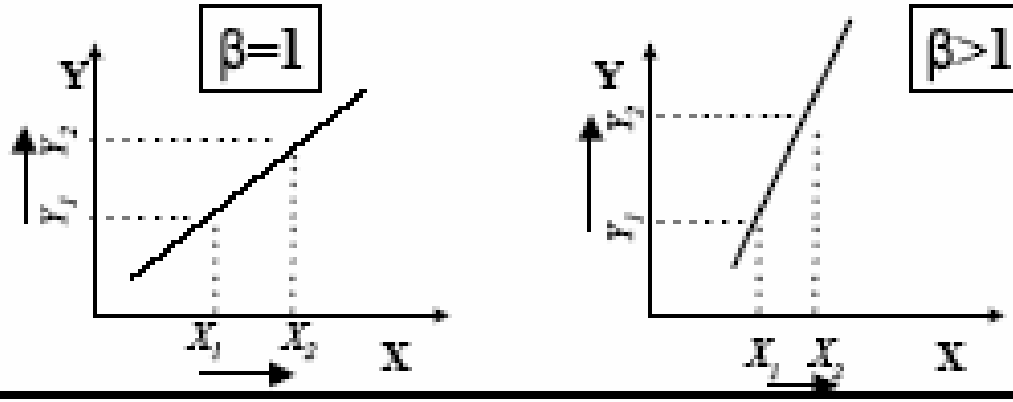
$$b = \frac{\text{Covarianza}_{xy}}{\text{Varianza}_{x}}$$



■ $\beta > 0$ = all'*aumentare* di X *aumenta* Y

Ad ogni variazione di X corrisponde un'*uguale* variazione in Y

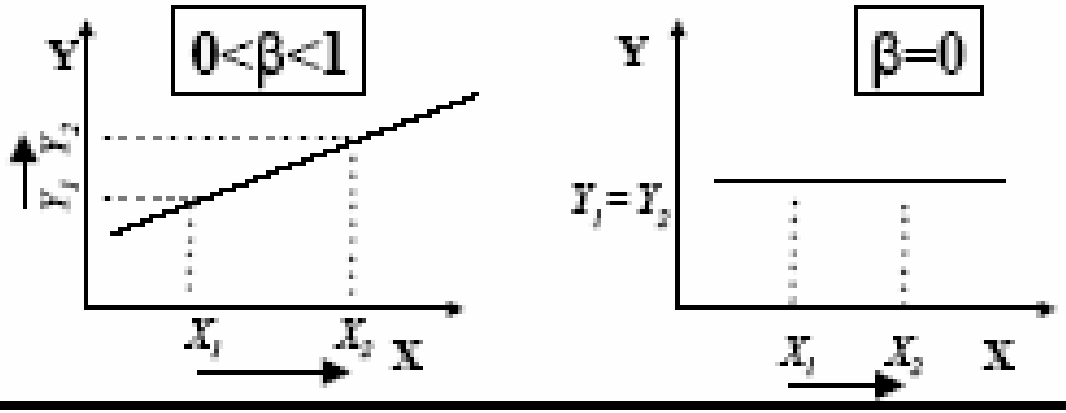
Ad ogni variazione di X corrisponde una variazione *maggiore* in Y



- $\beta > 0$ = all'*aumentare* di X *aumenta* Y

Ad ogni variazione di X corrisponde una variazione *minore* in Y

Ad ogni variazione di X *non* corrisponde *alcuna* variazione in Y ⇨ ASSENZA DI RELAZIONE

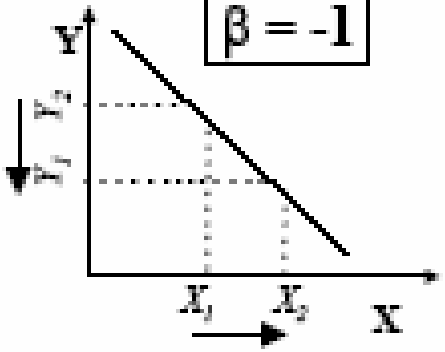


■ $\beta < 0$ = all'*aumentare* di X *diminuisce* Y

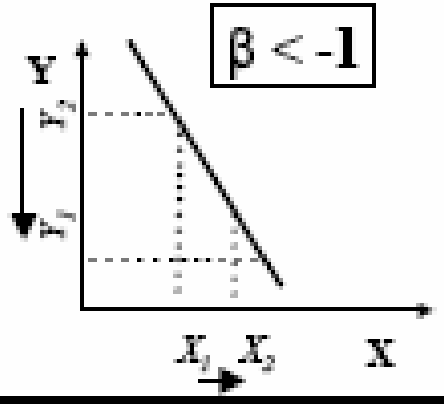
Ad ogni variazione di X corrisponde un'*uguale* variazione in Y

Ad ogni variazione di X corrisponde una variazione *maggiore* in Y

$\beta = -1$



$\beta < -1$



COEFFICIENTE di DETERMINAZIONE

R^2 R square

Misura 'quanto' della variabile dipendente Y sia spiegato dalla variabile indipendente X

$$0 \leq R^2 \leq 1$$

Formula per il calcolo di R^2:

$$R^2 = \frac{\text{cov}^2(x, y)}{\text{var}(x) * \text{var}(y)}$$

... coefficiente di determinazione:

$$R^2 = \frac{\mathrm{cov}^2(x, y)}{\mathrm{var}(x)\,\mathrm{var}(y)}$$

... coefficiente di correlazione:

$$r = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

... allora: