REGRESSIONE

Sezione di Epidemiologia & Statistica Medica
Università degli Studi di Verona

REGRESSIONE

- permette di esprimere la relazione tra due variabili con un modello funzionale

$y=f(x)$

funzione matematica che lega le due variabili

variabile dipendente

variabile indipendente

SCOPO
- descrittivo
- predittivo

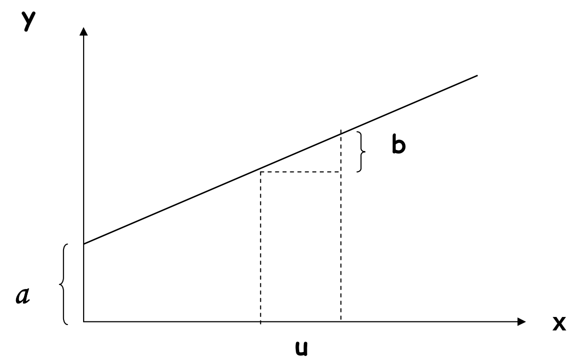
Regressione lineare

Ampiamente usata in ambito biomedico

$y=a+b(x)$

intercetta

Coefficiente angolare



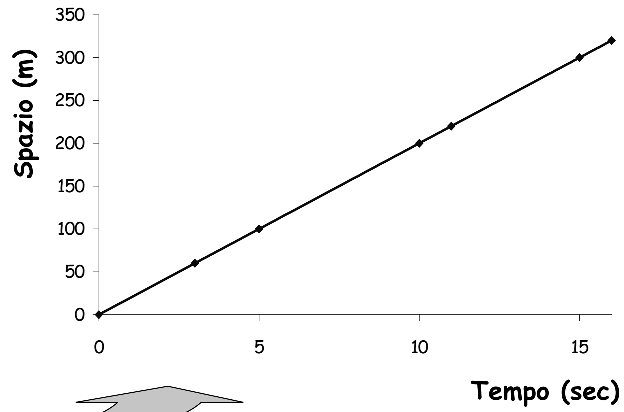
Esempio di relazione lineare

Si abbiano due variabili X e Y.

X è il tempo (in secondi) a cui viene osservato un corpo.

Y è lo spazio (in metri) che il corpo ha percorso da un certo punto.

X	Y
0	0
3	60
5	100
10	200
11	220
15	300
16	320



$y=20x$

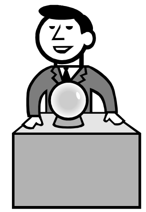
- La variabilità di Y è completamente spiegata dalla retta

- La retta descrive perfettamente i dati e individua la "legge" che li ha prodotti (*legge del moto uniforme*)

- Il coefficiente angolare (b=20) rappresenta l'incremento nello spazio per incremento unitario nel tempo (la velocità) ed è misurato come metri al secondo

- Tale modello è completamente deterministico:

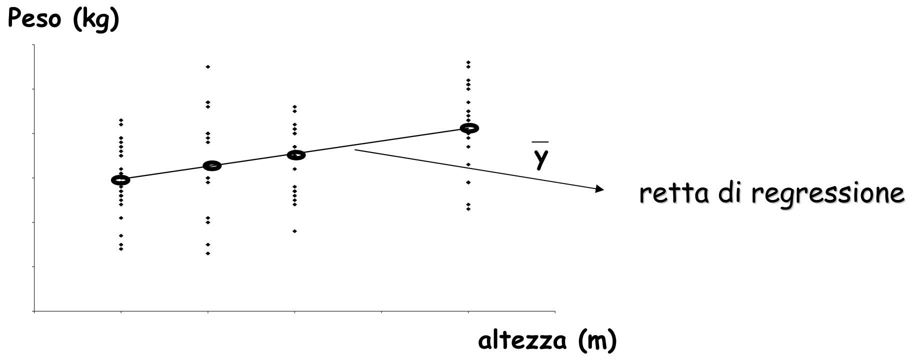
  noti i valori di x, si possono predire esattamente i valori di Y



- In biologia e medicina, la relazione tra variabili non è sempre perfettamente lineare

MA → Il modello lineare permette di approssimare la descrizione del fenomeno

Esempio: relazione tra peso e altezza



- Per ogni altezza esiste un range di pesi → Variabilità biologica + Errore di misura

- In media il peso cresce linearmente con l'altezza

- Il luogo geometrico delle medie di Y per dati valori di X è detto CURVA DI REGRESSIONE DI Y SU X

se curva=retta

RETTA DI REGRESSIONE DI y SU x

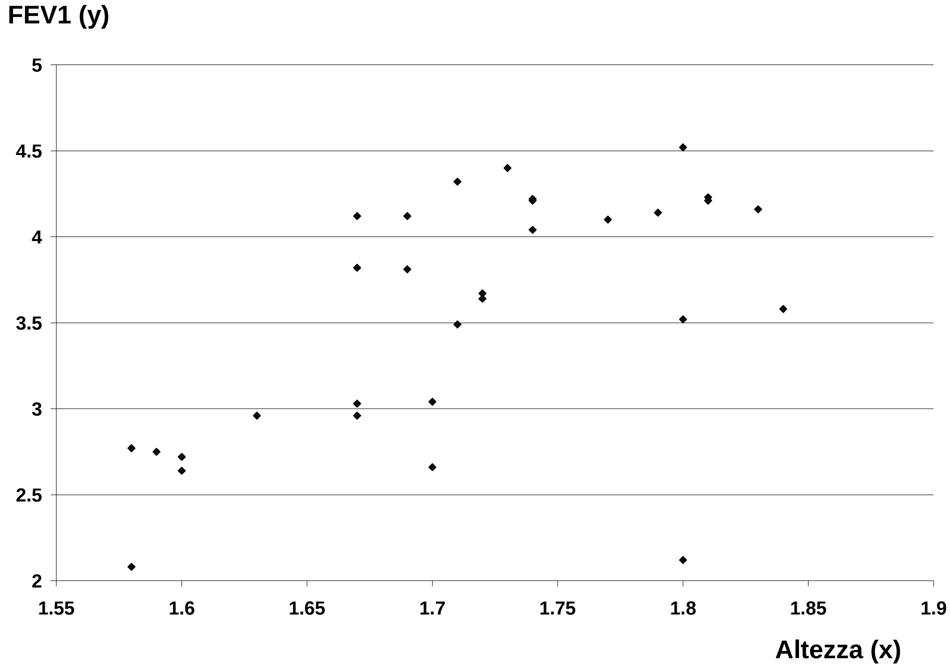
Esercizio

Nella tabella seguente sono riportati i dati relativi ad altezza e FEV1 (*forced expiratory volume in 1 second*) per 30 soggetti (dati ECRHS).

altezza (m)	FEV1 (l)
X	Y
1.79	4.14
1.8	4.52
1.72	3.64
1.69	4.12
1.72	3.67
1.84	3.58
1.6	2.72
1.7	3.04
1.83	4.16
1.58	2.08
1.74	4.04
1.74	4.22
1.67	3.82
1.71	3.49
1.67	2.96
1.58	2.77

altezza (m)	FEV1 (l)
X	Y
1.71	4.32
1.67	3.03
1.67	4.12
1.73	4.4
1.81	4.21
1.81	4.23
1.8	3.52
1.69	3.81
1.7	2.66
1.74	4.21
1.77	4.1
1.8	2.12
1.6	2.64
1.63	2.96
1.59	2.75

1. Rappresentiamo i dati in un diagramma a dispersione di punti



2. Assumeremo che nella popolazione il legame tra altezza (X) e FEV1 (Y) possa essere espressa da:

$$E(y)=\alpha+\beta(x)$$

L'osservazione di Y nell'i-mo individuo avrà quindi la seguente struttura:

COMPONENTE FISSA o PREDITTORE LINEARE

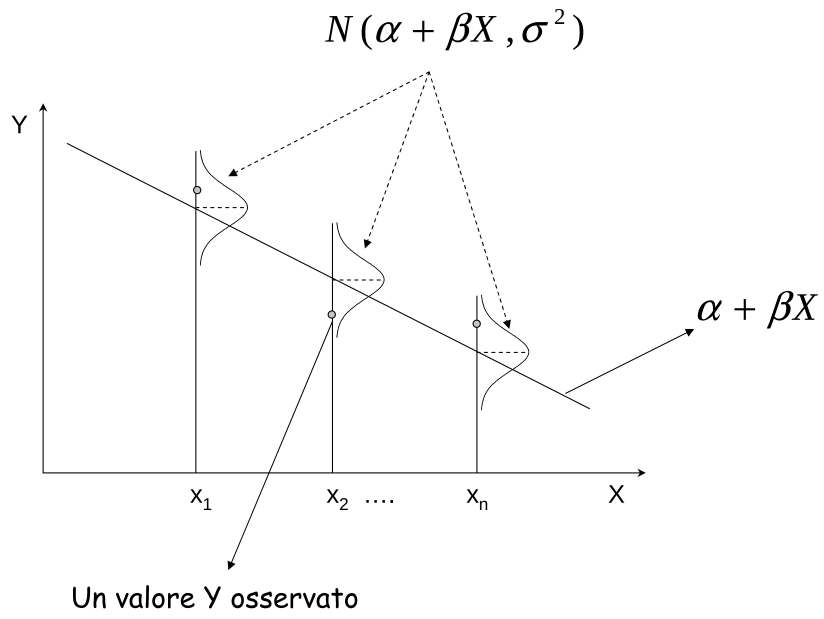
$$y_i = \boxed{\alpha+\beta(x_i)} + \boxed{\varepsilon_i}$$

ERRORE CASUALE associato ad ogni osservazione

Dove:

- **Y** è la variabile di risposta (o <u>dipendente</u>)
- $\alpha+\beta$ sono parametri ignoti da <u>stimare</u> sulla base dei dati disponibili
- **X** è la variabili esplicativa (<u>indipendente</u>)
- ε_ι (errore casuale) ~**N(0,σ²)**

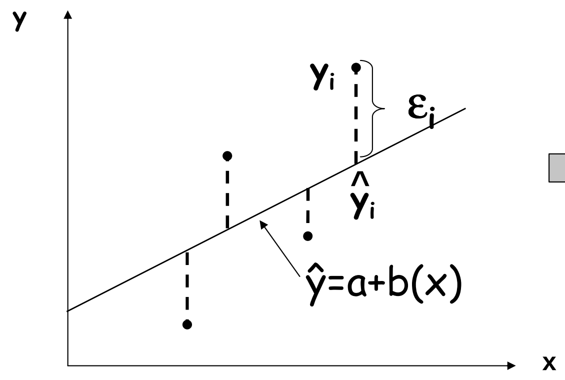
•Y, cioè il FEV1, dipende dall'altezza dell'individuo (X, parte sistematica) e da altre caratteristiche individuali (ε_ι, parte probabilistica)



3. A questo punto, come scegliamo la retta che meglio si adatta ai nostri dati?

Come stimiamo α e β?

STIMA DEI PARAMETRI CON IL METODO DEI MINIMI QUADRATI



Cerchiamo la retta che rende minima la distanza tra y e ŷ, per ogni i

Cerchiamo a e b (stime di α e β) in modo da minimizzare la seguente quantità:

$$\sum_i \varepsilon_i^2 = \sum_i (y_i - \alpha - \beta x_i)^2 = \sum_i (y_i - \hat{y}_i)^2$$

$$b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i x_i y_i - \left[\left(\sum_i x_i\right)\right]\left(\sum_i y_i\right)/n}{\sum_i x_i^2 - \left(\sum_i x_i\right)^2 / n} = \frac{codev.}{dev} = \frac{Sxy}{Sxx}$$

$$b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{Sxy}{Sxx}$$

$$a = \bar{y} - b\bar{x}$$

Stima dei parametri della retta di regressione

Si noti che il punto di coordinate $(\bar{x}, \bar{y})$ appartiene alla retta di regressione. Infatti:

$$\hat{y} = \bar{y} - b\bar{x} + bx \Rightarrow \hat{y} = \bar{y} + b(x - \bar{x})$$

E per $x = \bar{x} \Rightarrow \hat{y} = \bar{y}$

Quindi nell'esempio:

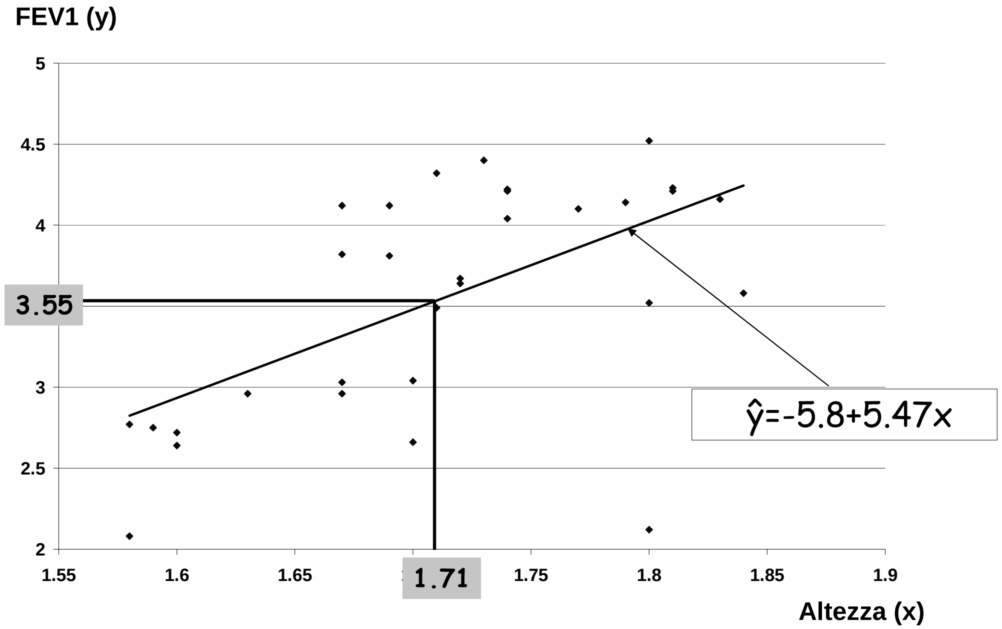
$\bar{x}=1.71$, $\bar{y}=3.55$

$Sxx=0.1748$, $Syy=15.1098$, $Sxy=0.9562$

$b=Sxy/Sxx=0.9562/0.1748=5.47$

$a=\bar{y}-b\bar{x}=3.55-5.47\cdot 1.71=-5.8$

$$\hat{y}=-5.8+5.47x$$



$(\bar{x}, \bar{y})$ appartengono alla retta di regressione

5. Stima della varianza residua

Varianza delle osservazioni, Y, intorno al modello di regressione (varianza d'errore)

$$s_e^2 = \frac{\sum_i (y_i - \hat{y})^2}{n-2} =$$

g.l=n°osservazioni-n° parametri

$$s_e^2 = (S_{yy} - \frac{S_{xy}^2}{S_{xx}})/(n-2)$$

Dimostrazione:

$$\sum_i (y_i - \hat{y})^2 = \sum_i (y_i - a - bx_i)^2$$

$$= \sum_i (y_i - \bar{y} + b\bar{x} - bx_i)^2 = \sum_i \{(y_i - \bar{y}) - b(x_i - \bar{x})\}^2 =$$

$$= \sum_i (y_i - \bar{y})^2 + b^2 \sum (x_i - \bar{x})^2 - 2b \sum (y_i - \bar{y})(x_i - \bar{x})$$

•($b=S_{XY}/S_{XX}$)

$$= S_{yy} + \frac{S_{xy}^2}{S_{xx}^2} S_{xx} - 2\frac{S_{xy}}{S_{xx}} S_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

6. Errore standard di b e test per il modello di regressione

- Si può dimostrare che: $ES(b) = \dfrac{s_e}{\sqrt{\sum\left(x_i - \bar{x}\right)^2}} = \dfrac{s_e}{\sqrt{Sxx}}$

- La validità del modello viene valutata mediante il seguente sistema d'ipotesi

$$\begin{cases} H_0: \beta_0 = \beta = 0 \\ H_1: \beta \neq 0 \end{cases}$$

$$\boxed{t = \frac{b - \beta_0}{ES(b)}} \sim t_{n-2}$$

- intervallo di confidenza (95%) $\boxed{b \pm t_{n-2,\,\alpha/2} \cdot ES(b)}$

Nell'esempio;

$$s_e^2 = \frac{\sum_i (y_i - \hat{y})^2}{n-2} = 0.34$$

g.l.=31-2=29

$$ES(b) = \frac{s_e}{\sqrt{\sum\left(x_i - \bar{x}\right)^2}} = \frac{\sqrt{0.34}}{\sqrt{0.17}} = 1.396$$

$$t = \frac{b - \beta_0}{ES(b)} = \frac{5.47 - 0}{1.396} = 3.92$$

intervallo di confidenza (95%)

$$b \pm t_{0.025,29} \cdot ES(b) \Rightarrow 5.47 \pm 2.364 \cdot 1.396$$

$$5.47 \pm 3.30$$

$$\left(2.17; 8.77\right)$$