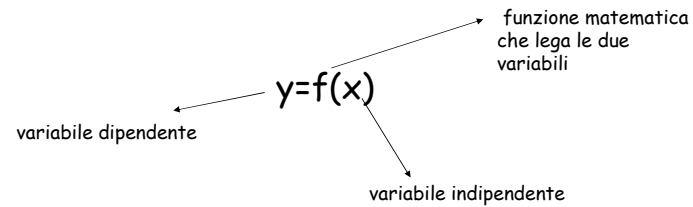


## REGRESSIONE

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



SCOPO

- ⇒ descrittivo
- ⇒ predittivo

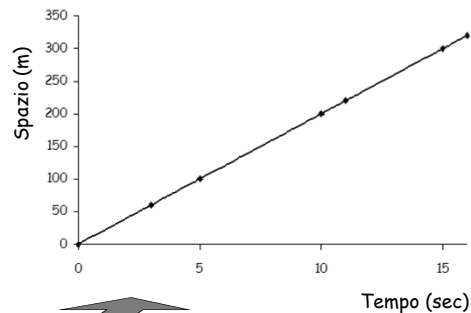
### Esempio di relazione lineare

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

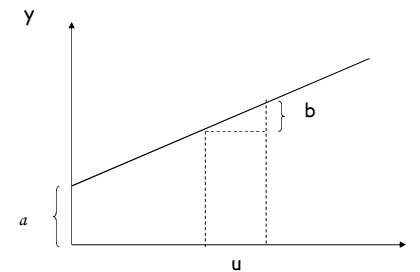


$$y=20x$$

## ◆ Regressione lineare

Ampiamente usata in ambito biomedico

intercetta →  $y=a+b(x)$  ← Coefficiente angolare



◆ 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

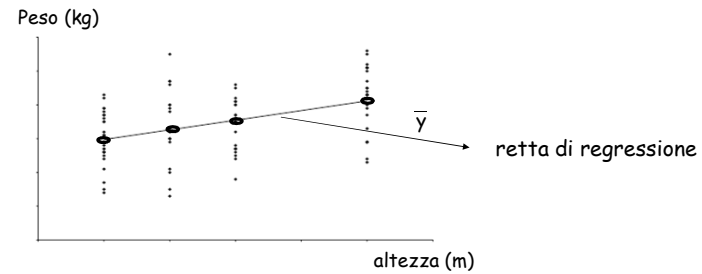
MA → Il modello lineare permette di approssimare la descrizione del fenomeno

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

se curva=retta

RETTE DI REGRESSIONE DI  $y$  SU  $x$

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

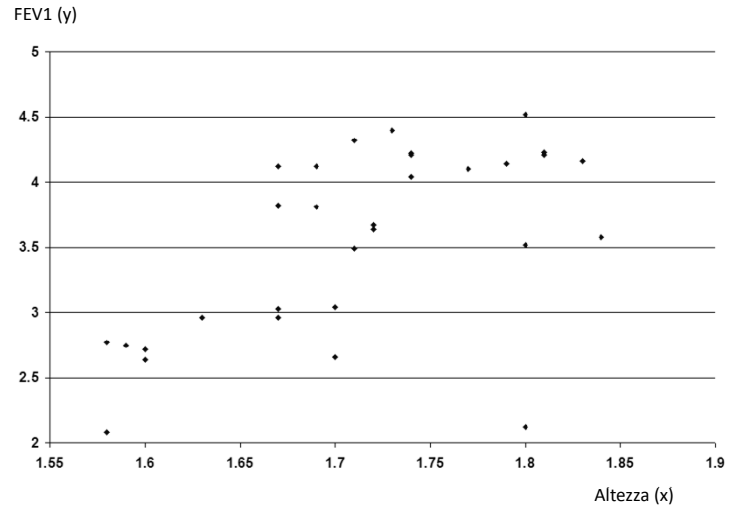
Esercizio

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:

$$y_i = \alpha + \beta(x_i) + \varepsilon_i$$

COMPONENTE FISSA o PREDITTORE LINEARE

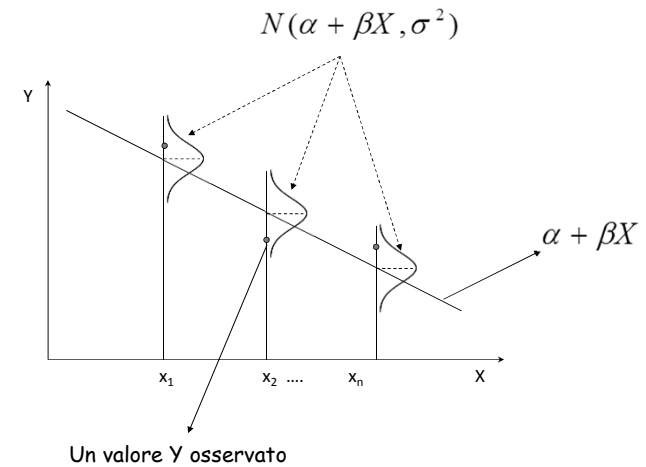
ERRORE CASUALE associato ad ogni osservazione

Dove:

- Y è la variabile di risposta (o dipendente)
- $\alpha + \beta$  sono parametri ignoti da stimare sulla base dei dati disponibili
- X è la variabile esplicativa (indipendente)
- $\varepsilon_i$  (errore casuale)  $\sim N(0, \sigma^2)$



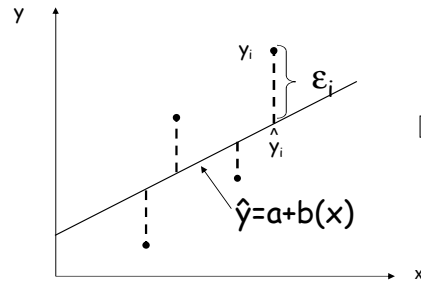
• Y, cioè il FEV1, dipende dall'altezza dell'individuo (X, parte sistematica) e da altre caratteristiche individuali ( $\varepsilon_i$ , parte probabilistica)



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

➡ Come stimiamo  $\alpha$  e  $\beta$ ?

### ➤ STIMA DEI PARAMETRI CON IL METODO DEI MINIMI QUADRATI



➡ Cerchiamo la retta che rende minima la distanza tra  $y$  e  $\hat{y}$ , per ogni  $i$

➡ Cerchiamo  $a$  e  $b$  (stime di  $\alpha$  e  $\beta$ ) in modo da minimizzare la seguente quantità:

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



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

Quindi nell'esempio:

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

$$S_{xx} = 0.1748, \quad S_{yy} = 15.1098, \quad S_{xy} = 0.9562$$

$$b = S_{xy} / S_{xx} = 0.9562 / 0.1748 = 5.47$$

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

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

$$b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

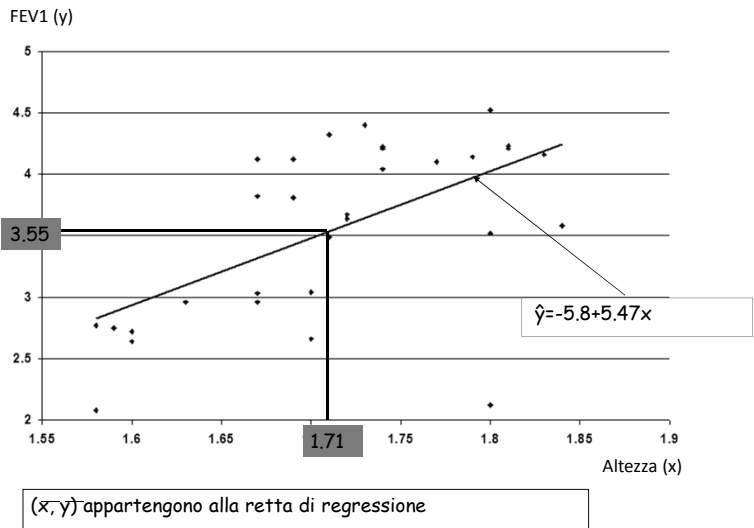
$$a = \bar{y} - b\bar{x}$$

Stima dei parametri della retta di regressione

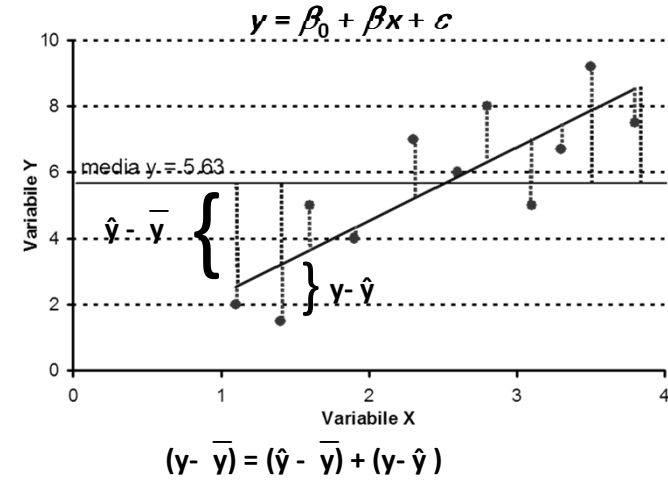
Si noti che il punto di coordinate  $(\bar{x}, \bar{y})$  appartiene alla retta di regressione. Infatti:

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

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



### SCOMPOSIZIONE DELLA DEVIANZA nella Regressione lineare semplice - 1



### SCOMPOSIZIONE DELLA DEVIANZA nella Regressione lineare semplice - 2

Variabilità totale

Variabilità residua

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

Variabilità spiegata dalla regressione

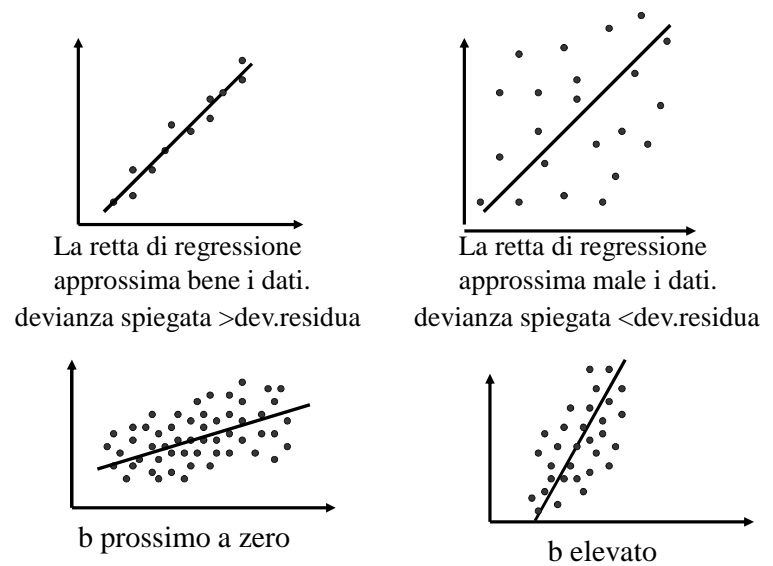
Si può dimostrare che:

Devianza totale, SST

Devianza residua, SSE

$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2$$

Devianza spiegata dalla regressione, SSR



**COEFFICIENTE DI DETERMINAZIONE**

$$r^2 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$

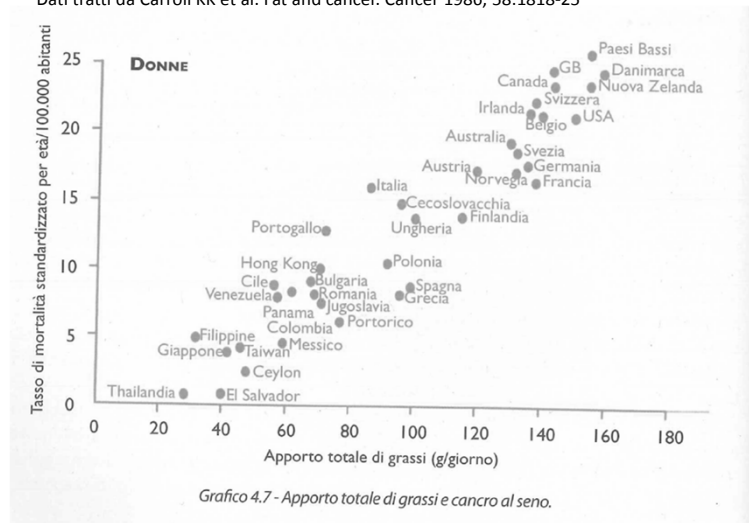
$$r^2 = 1 - \text{SSE} / \text{SST} = \text{SSR} / \text{SST}$$

SST = Devianza (o Somma dei quadrati) totale

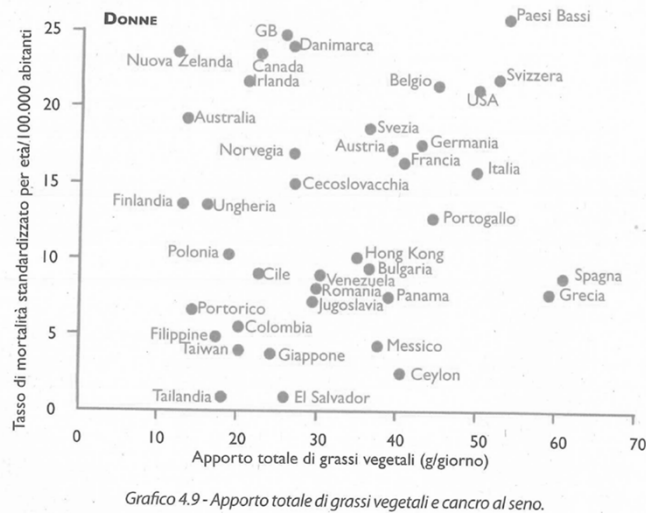
SSR = Devianza (o Somma dei quadrati) spiegata dalla Regressione

Proporzione di variazione totale della variabile dipendente Y che è spiegata dalla variabile indipendente X

Rielaborato da TC Campbell TM Campbell – The China Study  
 Dati tratti da Carroll KK et al. Fat and cancer. Cancer 1986, 58:1818-25



Rielaborato da TC Campbell TM Campbell – The China Study  
 Dati tratti da Carroll KK et al. Fat and cancer. Cancer 1986, 58:1818-25



5. Stima della varianza residua

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

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

g.l. = n° osservazioni - n° parametri

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

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

Si può dimostrare che: 
$$ES(b) = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{s_e}{\sqrt{S_{xx}}}$$

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

$$\begin{cases} H_0: \beta_0 = \beta = 0 \\ H_1: \beta \neq 0 \end{cases} \quad t = \frac{b - \beta_0}{ES(b)} \sim t_{n-2}$$

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

intervallo di confidenza (95%)

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

$$5.47 \pm 3.30$$

$$(2.17; 8.77)$$

Nell'esempio:

$$s_e^2 = \frac{\sum_i (y_i - \hat{y})^2}{n-2} = 0.34 \quad \text{g.l.} = 31-2=29$$

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

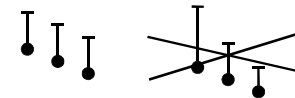
$$t = \frac{b - \beta_0}{ES(b)} = \frac{5.47 - 0}{1.396} = 3.92$$

## Regressione lineare semplice

$$y = \beta_0 + \beta x + \varepsilon$$

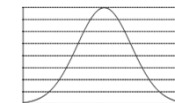
## ASSUNZIONI

- Il valore atteso degli errori  $E(\varepsilon)$  deve essere pari a ZERO
- OMOSCEDASTICITA' (La varianza degli errori rimane costante)



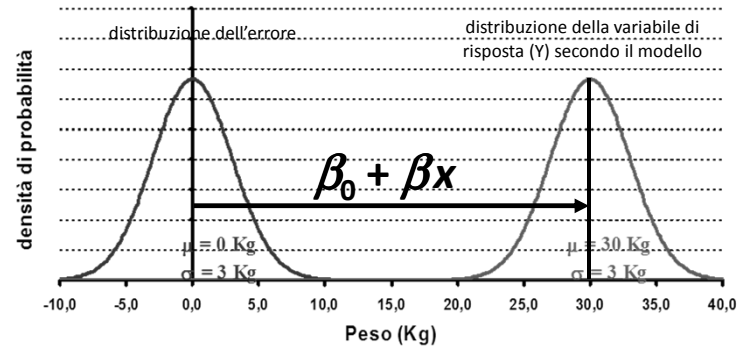
- INDIPENDENZA degli errori  
se le provette tra un esame e l'altro non vengono lavate adeguatamente, una determinazione risente della determinazione precedente

- Distribuzione NORMALE degli errori



## Regressione lineare semplice ASSUNZIONI

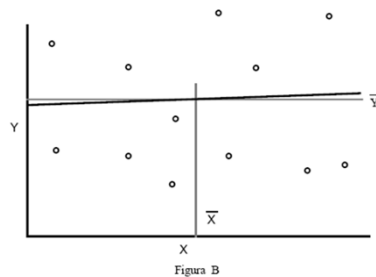
- 1) Il valore atteso degli errori  $E(\epsilon)$  deve essere pari a ZERO
- 4) Gli errori si distribuiscono normalmente



## Inferenza sui parametri: i dati “supportano” il modello proposto?

La retta calcolata non rappresenta un miglioramento effettivo della distribuzione dei punti, rispetto alla loro media.

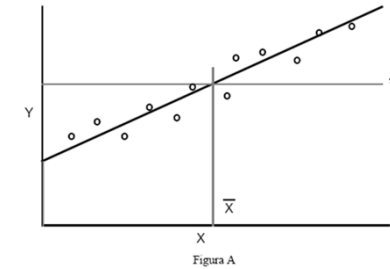
In questo caso la retta calcolata ha una pendenza positiva. Ma sarebbe sufficiente raccogliere un'altra osservazione per ottenere una retta differente, che potrebbe addirittura avere un coefficiente angolare  $b$  di segno opposto.



La retta calcolata non è rappresentativa di una relazione reale tra X ed Y.

## Inferenza sui parametri: i dati “supportano” il modello proposto?

I punti hanno distanze dalla retta di regressione che sono sensibilmente minori di quelle dalla media.



Il valore stimato mediante la retta di regressione si avvicina molto all'osservazione reale.

Raccogliendo altri punti campionari, la retta calcolata resterebbe probabilmente immutata. La retta di regressione esprime la **relazione reale** che esiste tra X ed Y.

### SCHEMA RIASSUNTIVO

#### UTILIZZO DEL MODELLO DI REGRESSIONE

Date  $n$  unità sperimentali in cui vengono misurate 2 variabili, X e Y, lo studio della relazione tra X e Y segue i seguenti punti:

- Plot descrittivo dei dati per valutare l'esistenza di una relazione lineare
- Si ipotizza che nell'universo esista una legge funzionale del tipo  $y = a + \beta x$ , con  $a + \beta$  parametri incogniti



- ↵ Si stimano con il metodo dei minimi quadrati i parametri ignoti
- ↵ Si procede all'inferenza sui parametri per valutare se i dati "supportano" il modello proposto
- ↵ Inferenza e previsione sui valori della variabile  $Y$
- ↵ Analisi dei residui per ulteriore conferma del modello