ANNO ACCADEMICO
2018/2019

CORRELAZIONE
E
REGRESSIONE

CORRELAZIONE E REGRESSIONE

Relazioni esistenti tra due o più variabili = che tipo di relazione esiste e quanto è forte?

CORRELAZIONE = variabili che dipendono da cause comuni e tendono quindi a variare congiuntamente (**covariare**)

→ **DIPENDENZA FUNZIONALE**

REGRESSIONE = rapporto di dipendenza tra due o più variabili

→ **DIPENDENZA CAUSALE**

CORRELAZIONE

La correlazione indica la tendenza che hanno due variabili (X e Y) a ***variare insieme***, ovvero, a ***covariare***.

Non implica rapporti di causa-effetto ma si ha una descrizione quantitativa dell'associazione tra le due variabili.

1. Le variabili variano in modo associato?

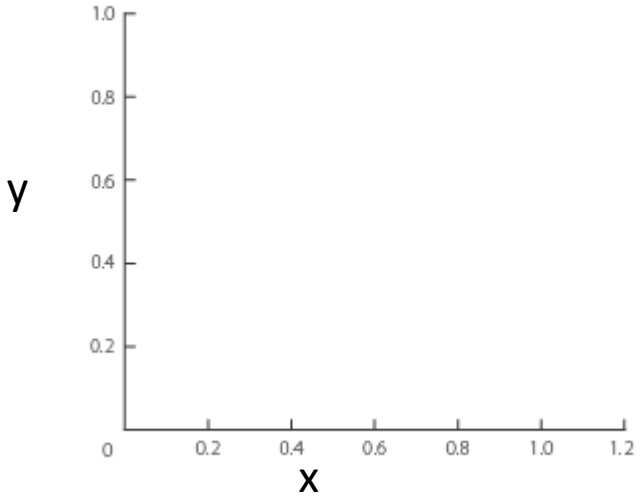
2. Quanto è grande questo grado di associazione?

CORRELAZIONE

Rappresentazione grafica

Il primo modo per verificare l'esistenza di una correlazione lineare tra due caratteri quantitativi X e Y, è quello di rappresentare la distribuzione doppia (X,Y) attraverso un **grafico a dispersione** (o *scatterplot*).

Uno *scatterplot* è un grafico in cui ogni osservazione della variabile doppia (x_I, y_I) viene rappresentata come un punto sugli assi cartesiani in cui:
- all'asse delle ascisse sono associati i valori della variabile X
- all'asse delle ordinate sono invece associati i valori della variabile Y



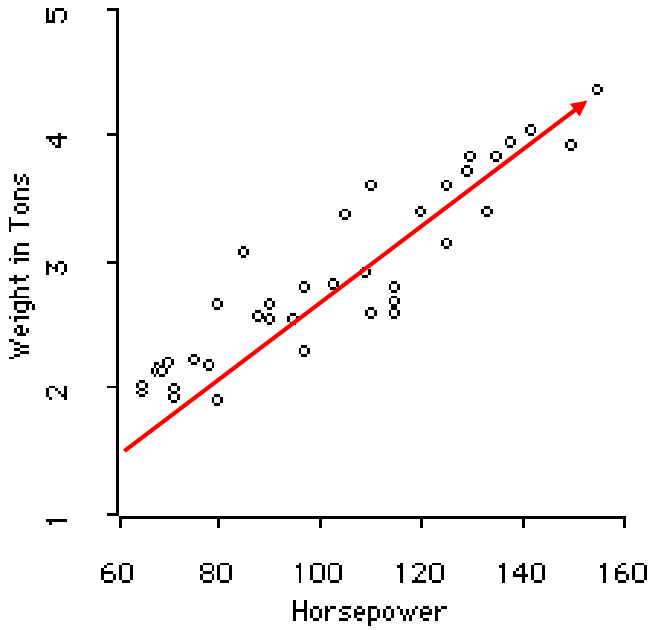
CORRELAZIONE

Quando si parla di correlazione bisogna prendere in considerazione due aspetti: *il* ***tipo di relazione esistente*** tra due variabili e ***la forma della relazione (direzione ed entità)***

TIPO DI RELAZIONE

LINEARE:

All'aumentare o al diminuire di X aumenta o diminuisce Y. Rappresentata su assi cartesiani assume la forma di una nube allungata

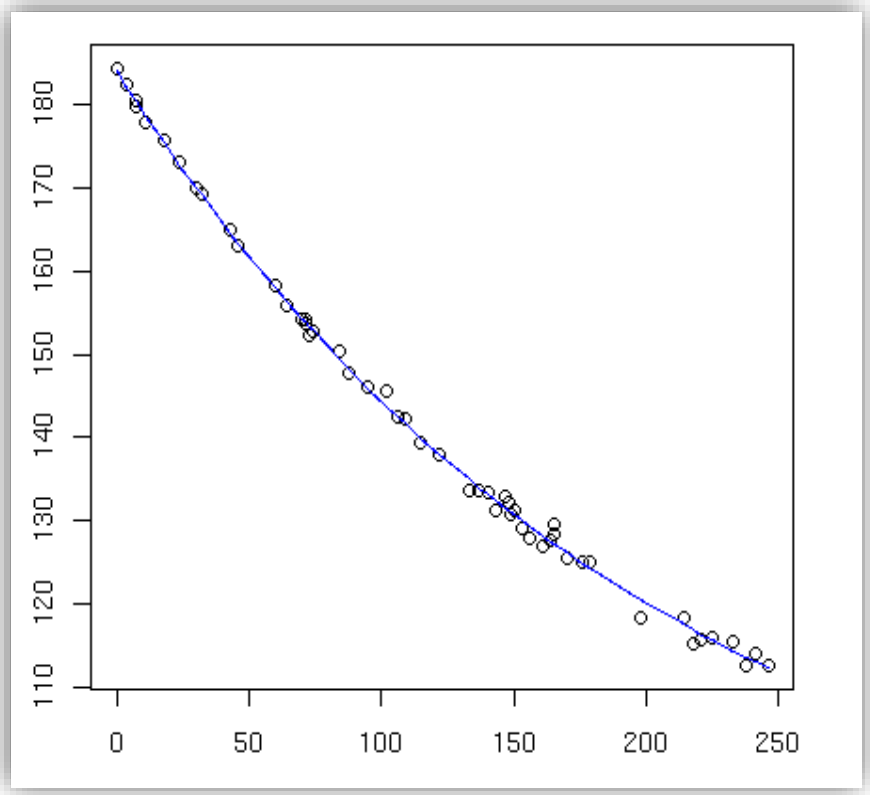


CORRELAZIONE

TIPO DI RELAZIONE

NON LINEARE O CURVILINEA

In questo caso a livelli bassi e alti di X corrispondono livelli bassi di Y, mentre a livelli intermedi di X corrispondono livelli intermedi di Y. Se rappresentata su assi cartesiani ha un andamento curvilineo (a parabola)

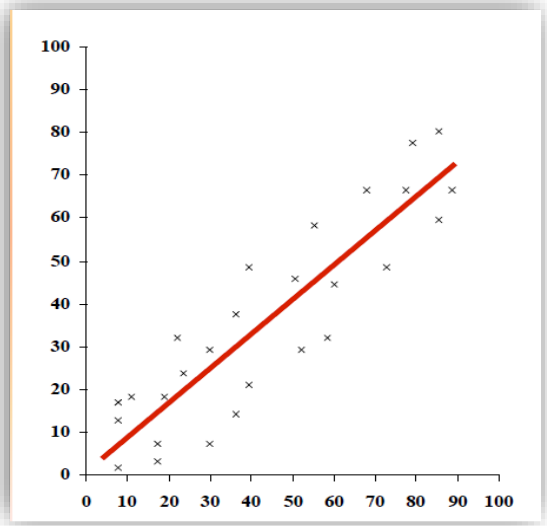


CORRELAZIONE

DIREZIONE DELLA RELAZIONE

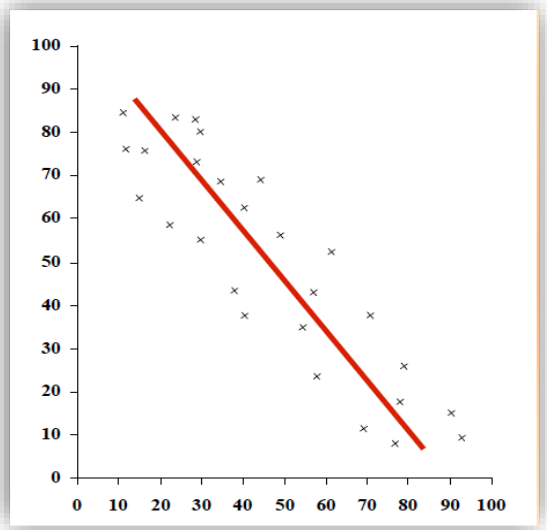
POSITIVA

All’aumentare della variabile X aumenta anche la variabile Y



NEGATIVA

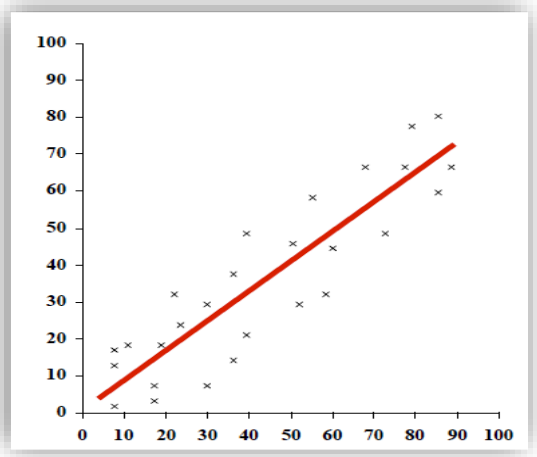
All’aumentare della variabile X diminuisce la variabile Y



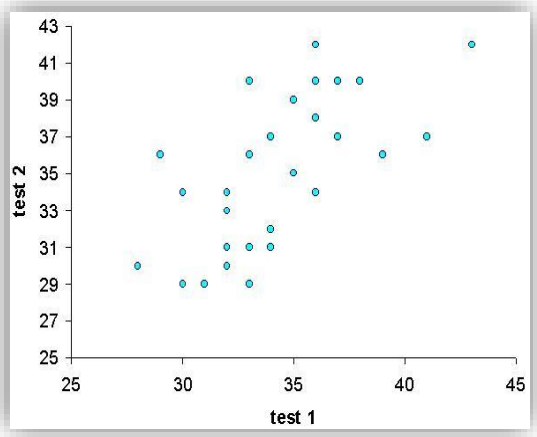
CORRELAZIONE

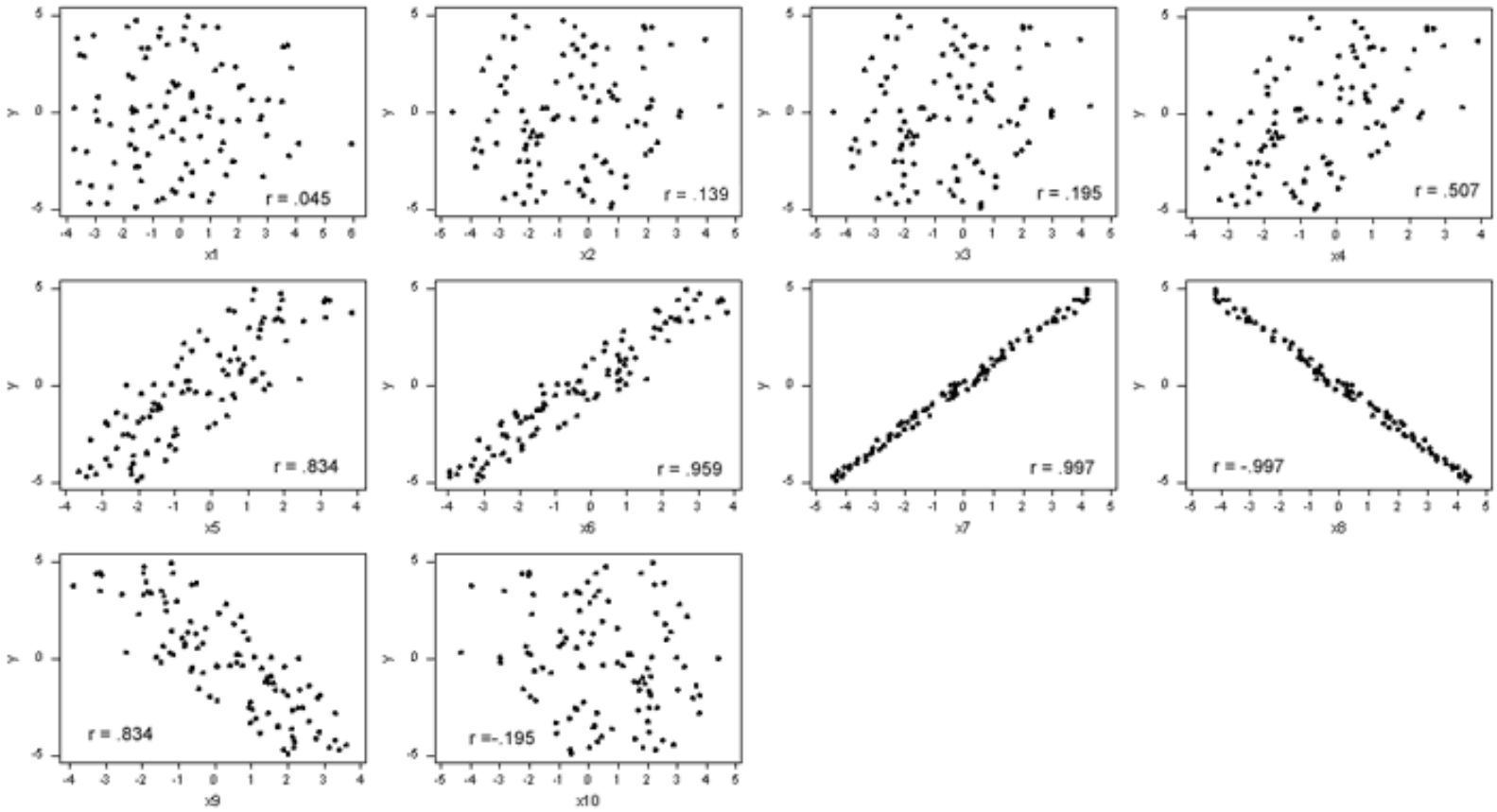
ENTITÀ DELLA RELAZIONE
Forza della relazione esistente tra due variabili

Quanto più la «nube» è allungata tanto più alta sarà la correlazione



Tanto più invece valori sono dispersi in maniera uniforme minore sarà la correlazione tra i due




r = .045
x1
r = .139
x2
r = .195
x3
r = .507
x4
r = .834
x5
r = .959
x6
r = .997
x7
r = -.997
x8
r = .834
x9
r =-.195
x10
y

CORRELAZIONE lineare

Per rispondere alle domande riguardanti il tipo (positiva o negativa) e l'entità della correlazione esistente (o meno) tra due variabili è necessario calcolare il **COEFFICIENTE DI CORRELAZIONE (r) di Pearson**

Il COEFFICIENTE DI CORRELAZIONE:

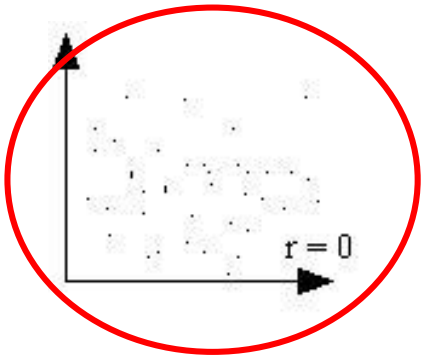
- misura l'intensità della relazione (lineare) tra due variabili X e Y;
- può assumere valori che vanno dal – 1.00 al + 1.00;
- quando ha valore +1 significa perfetta correlazione positiva: i valori della Y si dispongono esattamente su una retta con pendenza positiva;
- quando ha valore -1 significa perfetta correlazione negativa: i valori della Y si dispongono esattamente su una retta con pendenza negativa;
- Un valore pari a 0 indica che non c'è nessuna correlazione LINEARE tra le due variabili.

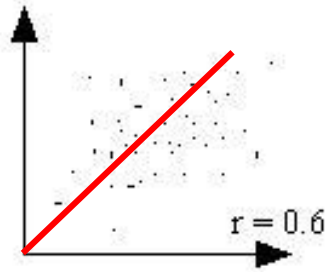
CORRELAZIONE

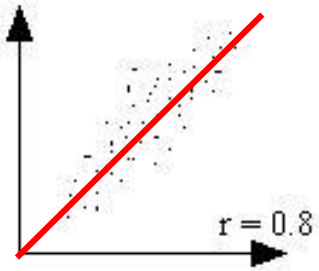
- **r = 1, r = -1 :** associazione completa positiva e negativa tra le due variabili. Dato il valore assunto da una variabile possiamo prevedere senza errori quello dell'altra.

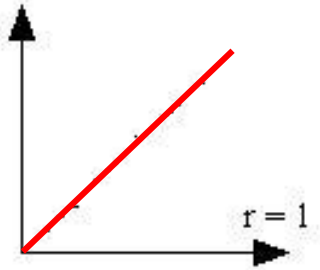
- **r = 0.6, r = 0.8 :** l'associazione è incompleta, cioè a una delle due variabili corrisponde una popolazione di valori dell'altra. È possibile prevedere la media.

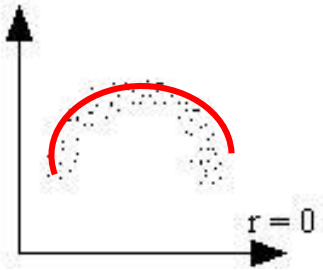
- **r = 0 :** le due variabili non sono correlate o non sono correlate in maniera lineare.











CORRELAZIONE

COEFFICIENTE DI CORRELAZIONE (COEFFICIENTE DI PEARSON)

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$$

$$r = \frac{cov_{xy}}{\sqrt{varianza_x \; x \; varianza_y}} \quad \Rightarrow \quad r = \frac{cov_{xy}}{\sqrt{ds_x^2 \; x \; ds_y^2}}$$

Dove:

cov_{xy} : è la covarianza tra X e Y, cioè la somma dei prodotti delle deviazioni delle misurazioni associate dalle loro rispettive medie

ds_x : deviazione standard di X

ds_y : deviazione standard di Y

CORRELAZIONE

COVARIANZA

Esprime l'intensità con cui due variabili variano insieme

$$cov_{x,y} = \frac{\sum(x - \bar{x})(y - \bar{y})}{N}$$

in cui:
$\bar{X}$ è la media di X;
$\bar{Y}$ è la media di Y;
N è la numerosità del campione

CODEVIANZA

$$\text{COV}_{XY} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{N}$$

in cui:
$\sum XY$ è la somma dei prodotti di XY;
$\sum x$ è la somma dei valori di X;
$\sum y$ è la somma dei valori di Y

SOMMA PRODOTTI

FATTORE DI CORREZIONE

CORRELAZIONE: esempio

ESEMPIO

Supponiamo di avere misurato la statura di 10 bambini di età compresa tra 6 e 12 anni e di riportare i dati su una tabella:

soggetto	età (anni) X	statura (centimetri) Y
1	6	115
2	6	120
3	7	122
4	8	130
5	8	128
6	9	134
7	10	136
8	10	140
9	11	147
10	12	151

CORRELAZIONE

Vogliamo calcolare il coefficiente di correlazione (r) e ci dobbiamo perciò ricavare la covarianza e la deviazione standard

$$\mathrm{cov}_{XY} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{N}$$

$\sum XY = 11723$

$\sum x = 87$

$\sum y = 1323$

Soggetto	Età X	Statura Y	XY
1	6	115	690
2	6	120	720
3	7	122	854
4	8	130	1040
5	8	128	1024
6	9	134	1206
7	10	136	1360
8	10	140	1400
9	11	147	1617
10	12	151	1812
Σ	**87**	**1323**	**11723**

$$\mathrm{cov}_{XY} = \frac{11723 - \frac{87 \cdot 1323}{10}}{10} = 21.29$$

COVARIANZA

CORRELAZIONE

Calcoliamo adesso la deviazione standard di X e di Y

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

$$DS_x = 2{,}057$$

$$DS_y = 11{,}653$$

CORRELAZIONE

Possiamo ora calcolare il coefficiente di correlazione

$$r = \frac{cov_{xy}}{\sqrt{ds_x^2 \; x \; ds_y^2}}$$

cov_{xy} = 21.29

$DS_x = 2.06$

$DS_y = 11.65$

$$r = \frac{21.29}{\sqrt{2.05^2 \; x \; 11.65^2}}$$

→ 0.987 →

ALTA CORRELAZIONE POSITIVA

CORRELAZIONE

SIGNIFICATIVITÀ

Nel nostro caso abbiamo **10** soggetti e un r di **0.986**.

Il nostro valore **è più alto** di quello riportato in tabella per il 5% a 8 gradi di libertà (n-2)

LA CORRELAZIONE RISULTA SIGNIFICATIVA

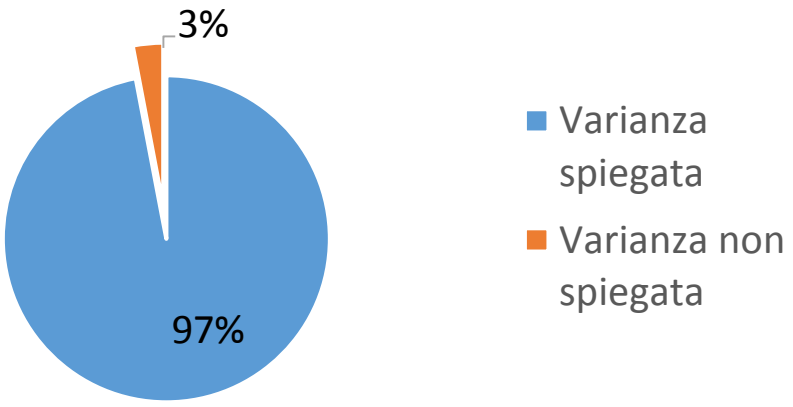
Numero di coppie in esame meno due	r 5 %	r 1 %
1	0,996 92	0,999 877
2	0,950 00	0,990 000
3	0,878 3	0,958 73
4	0,811 4	0,917 20
5	0,754 5	0,874 5
6	0,706 7	0,834 3
7	0,666 4	0,797 7
8	0,631 9	0,764 6
9	0,602 1	0,734 8
10	0,576 0	0,707 9
11	0,552 9	0,683 5
12	0,532 4	0,661 4
13	0,513 9	0,641 1
14	0,497 3	0,622 6
15	0,482 1	0,605 5
16	0,468 3	0,589 7
17	0,455 5	0,575 1
18	0,443 8	0,561 4
19	0,432 9	0,548 7
20	0,422 7	0,536 8
25	0,380 9	0,486 9
30	0,349 4	0,448 7
35	0,324 6	0,418 2
40	0,304 4	0,393 2
45	0,287 5	0,372 1
50	0,273 2	0,354 1
60	0,250 0	0,324 8
70	0,231 9	0,301 7
80	0,217 2	0,283 0
90	0,205 0	0,267 3
100	0,194 6	0,254 0

CORRELAZIONE

IL COEFFICIENTE DI DETERMINAZIONE r^2

Il coefficiente di determinazione misura l'ammontare di variabilità di una variabile spiegato dalla sua relazione con un'altra variabile. Nel caso specifico della correlazione **il coefficiente r^2 indica la percentuale di varianza che hanno in comune due variabili.**

Nell'esempio precedente, abbiamo trovato un r pari a **0.986**, da cui ricaviamo $r^2 = 0.986^2 =$ **0.972**. Questo sta ad indicare che il **97% delle variazioni della statura sono dovute all'età.**



REGRESSIONE

Rapporto di causa-effetto

Relazione quantitativa e funzionale tra una variabile **dipendente** Y e variabile **indipendente** X

Le variazioni della variabile dipendente Y sono considerate la risposta alle variazioni della variabile indipendente X

VARIABILE INDIPENDENTE = CAUSA
(sull'asse delle ascisse)

VARIABILE DIPENDENTE = EFFETTO
(sull'asse delle ordinate)

REGRESSIONE

1. Misurazione **quantitativa della** relazione funzionale esistente tra due o più variabili;
2. **Misura della variabilità della Y** non associata alla variazione della X
3. **Previsione** del valore di Y avendo noto il valore di X;

A seconda del numero di variabili indipendenti si distingue tra:

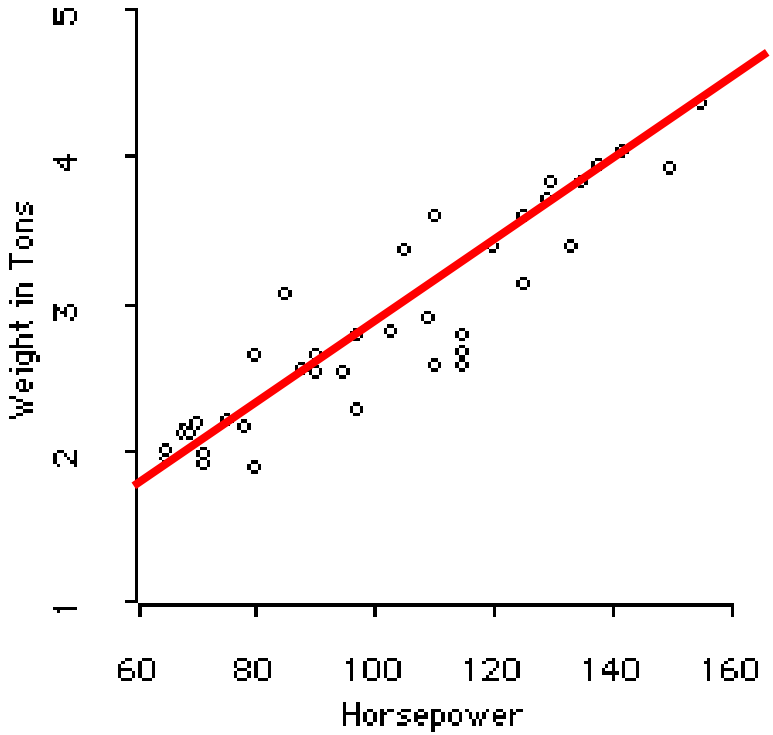
- Regressione semplice
- Regressione multipla

A seconda dell'andamento assunto dalla variabile indipendente in relazione ai valori assunti dalla variabile indipendente si distingue:

- Regressione lineare
- Regressione non lineare

REGRESSIONE: rappresentazione grafica

Retta di regressione



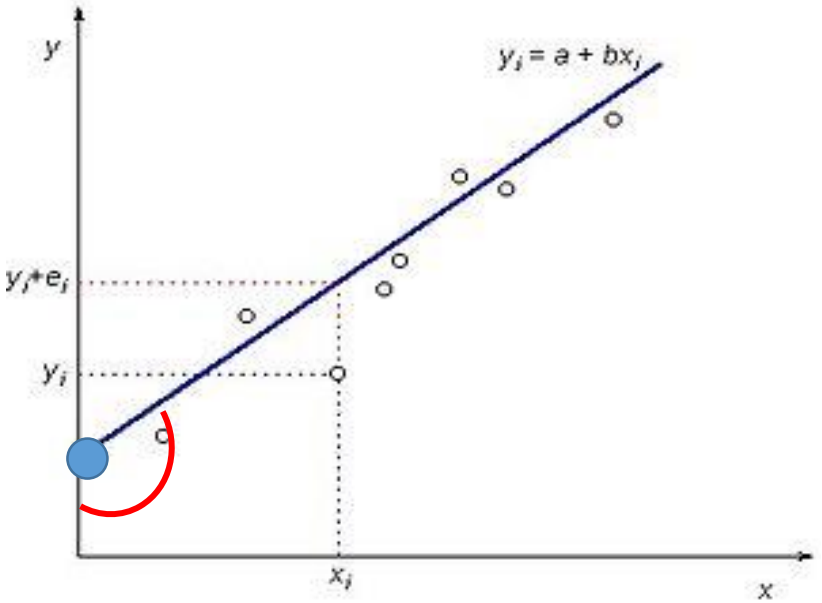
REGRESSIONE LINEARE SEMPLICE

Dobbiamo determinare una funzione matematica che descriva la relazione osservata.
L'equazione che rappresenta la retta del modello di regressione lineare semplice è:

$$Y = a + bX$$

Coefficiente di regressione che misura la variazione di Y quando X cambia di una sola unità.
GRADO DI INCLINAZIONE E VERSO DELLA RETTA

Intercetta della retta sull'asse delle Y, cioè il valore che Y assume quando X = 0
POSIZIONE DELLA RETTA



REGRESSIONE LINEARE SEMPLICE

$$Y = a + bX$$

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

$$b_{y,x} = \frac{covarianza\ (x,y)}{varianza\ (x)}$$

Una volta calcolati l'intercetta e il coefficiente di regressione otteniamo l'equazione rappresentativa della retta, con cui possiamo **PREVEDERE** il valore di Y (o una sua popolazione di valori) conoscendo solo il valore di X (indipendente)

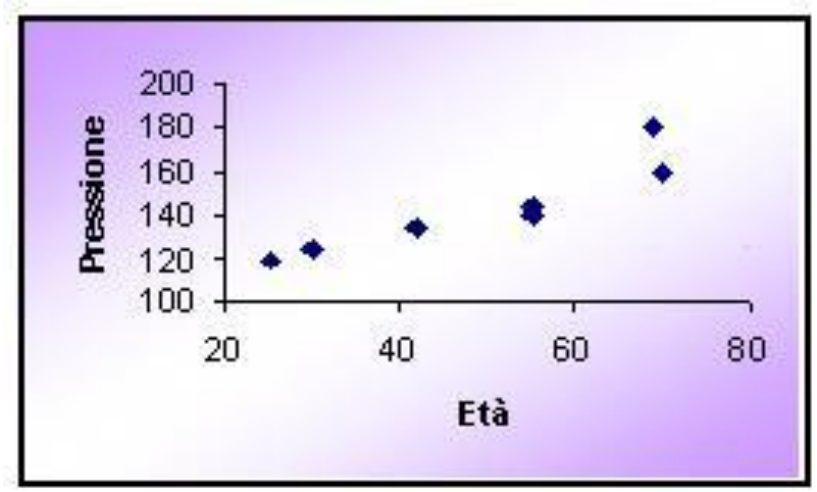
La retta di regressione esprime una **tendenza,** alla quale viene associata una misura di variabilità (errore standard della regressione); questo vuol dire che mediamente al variare della *X* la *Y* assumerà certi valori, non un valore preciso.

REGRESSIONE LINEARE SEMPLICE

ESEMPIO

Data una popolazione di individui si vuole stimare la relazione tra pressione arteriosa ed età e stimare i coefficienti della retta

ETA'	PRESSIONE
25	120
30	125
42	135
55	140
55	145
69	180
70	160



$Y = a + bX$

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

$$b_{y,x} = \frac{covarianza\ (x,y)}{varianza\ (x)}$$

ETA' (x)	PRESSIONE (Y)
25	120
30	125
42	135
55	140
55	145
69	180
70	160

	Età	**Pressione**
media	49,43	143,57
varianza	316	430,95
covarianza	340	

$$\boldsymbol{Y = 90,43 + 1,075(X)}$$

$$b = \frac{340}{316} = 1{,}075$$

$$a = 143{,}57 - 1{,}075(49{,}43) = 90{,}43$$