


CDL in MEDICINA & CHIRURGIA

Statistica Medica

gbarbati@units.it

A.A. 2024-25



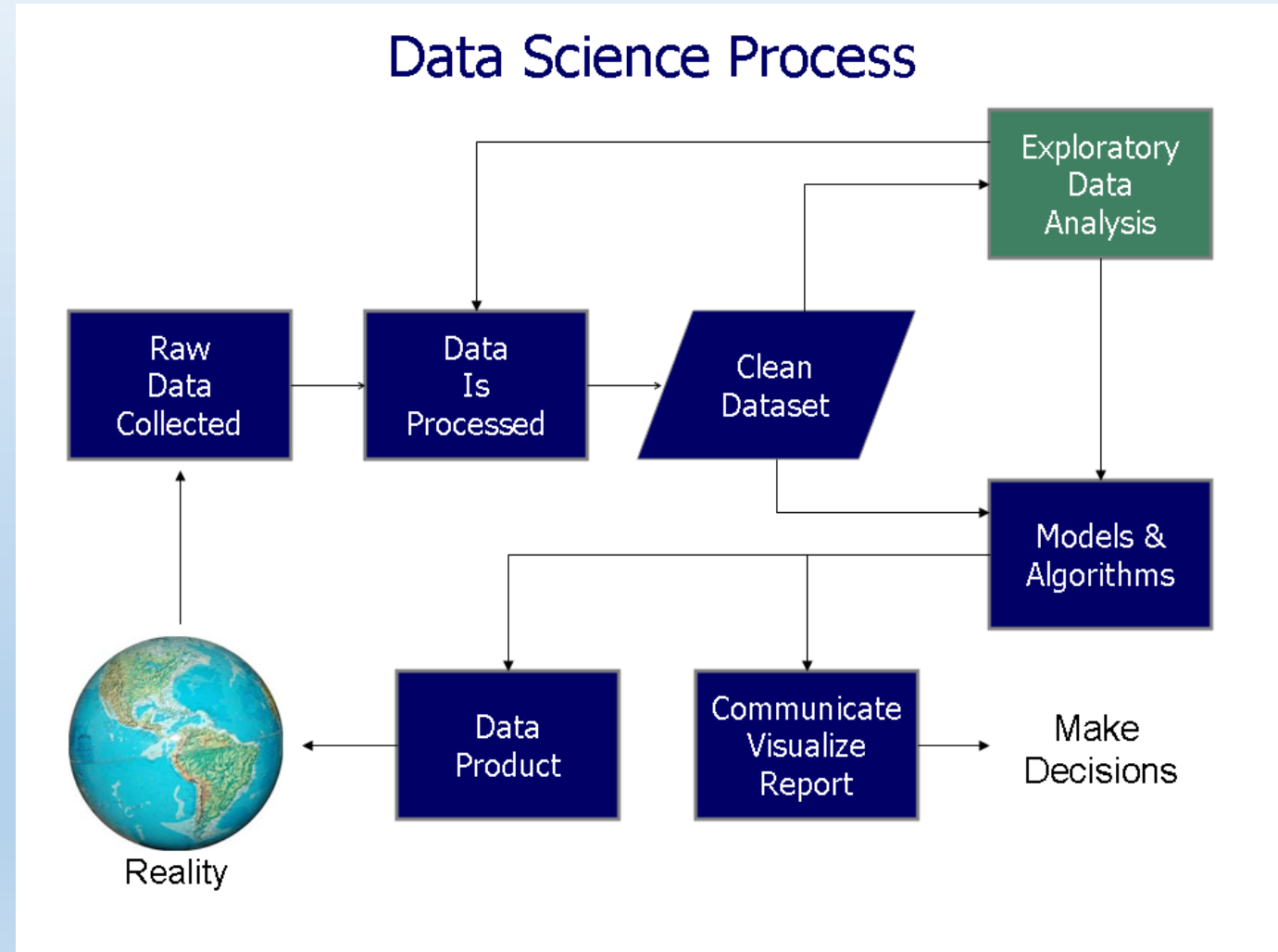
UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute

Sommario:

- Correlazione
- Regressione

There are lies, damned lies and statistics.

Mark Twain

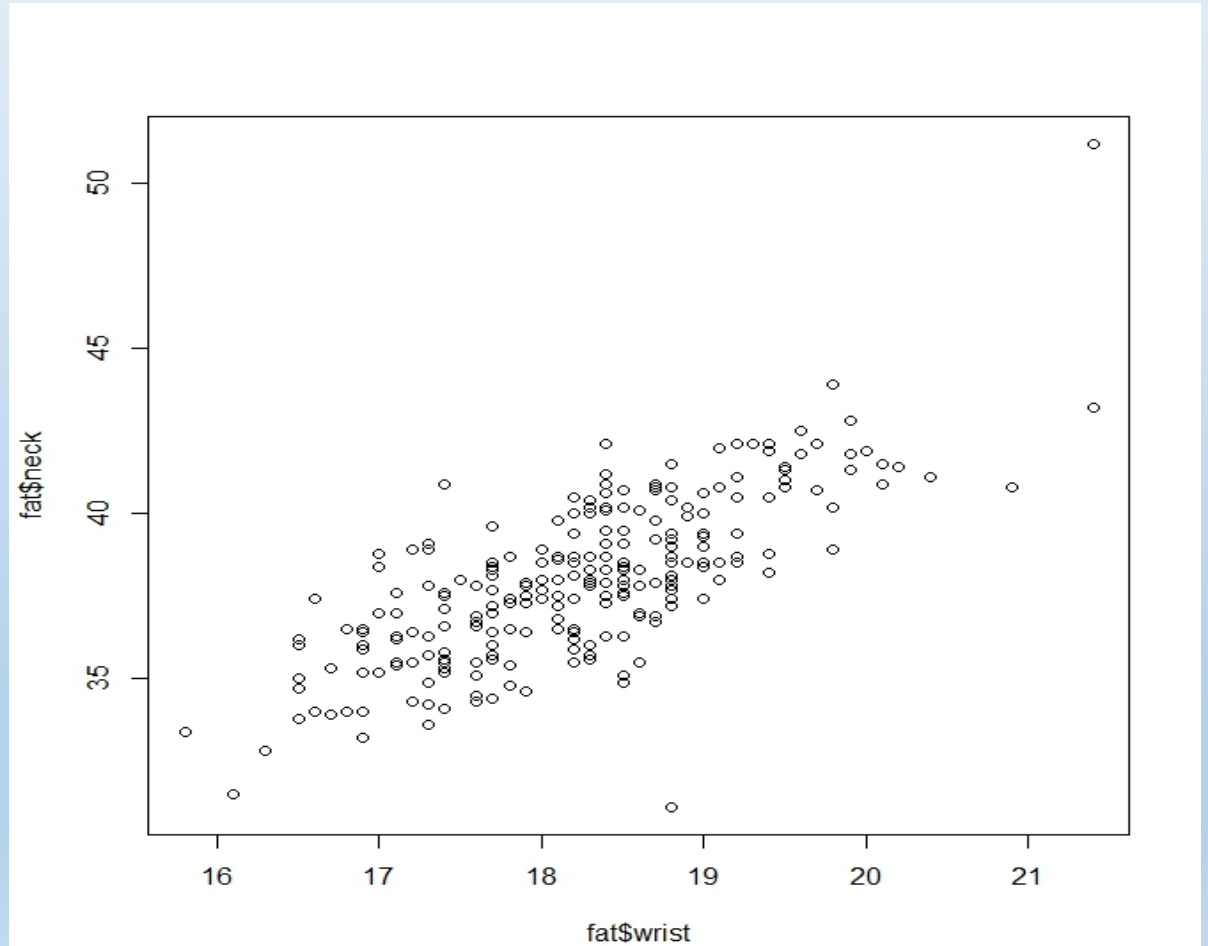


Indice di Correlazione (variabili su scala di misura numerica)

Se si rilevano **due** caratteri in una popolazione su scala numerica e si è interessati a studiare l'associazione tra i due caratteri, la rappresentazione grafica usuale è lo **scatter plot (diagramma cartesiano)**:

Distribuzione su un campione di 252 maschi delle dimensioni del polso e del collo: si ipotizza che l'ampiezza del collo sia circa due volte quella del polso.

Vogliamo definire un indice che quantifichi numericamente la associazione lineare tra due variabili su scala quantitativa.



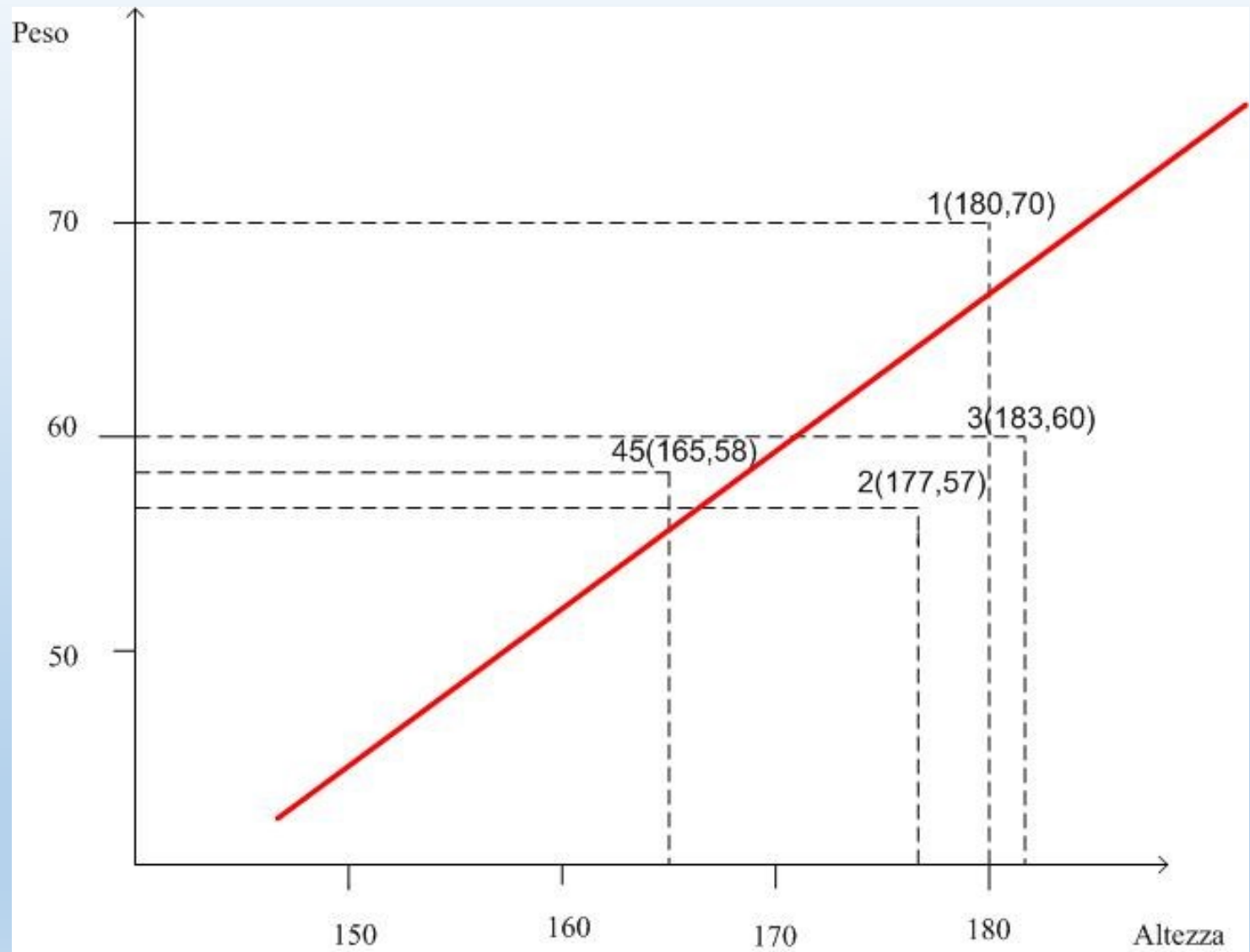
Per descrivere la relazione che intercorre tra due variabili **quantitative**, gli strumenti statistici di base sono:

- (1) il coefficiente di *covarianza*;
- (2) il coefficiente di *correlazione*;
- (3) la *regressione* lineare.

Ex: abbiamo un campione di N osservazioni (N=45, studenti) di cui abbiamo rilevato l'altezza ed il peso. Ogni studente è individuato da una *coppia* di valori: (x_i, y_i) , $i=1\dots 45$, dove x_i = valore dell'altezza nello studente i , y_i = valore del peso dello studente i .

| Studente | Altezza (cm) | Peso (kg) |
|----------|--------------|-----------|
| 1 | 180 | 70 |
| 2 | 177 | 57 |
| 3 | 183 | 60 |
| ... | ... | ... |
| 45 | 165 | 58 |

Vogliamo studiare l'associazione tra peso e altezza in questo campione



Il grafico cartesiano illustra "*visivamente*" se c'è una tendenza di *associazione* tra le variabili rilevate.

In questo caso la tendenza di associazione si può rappresentare tramite una retta che "*interpola*" (=passa attraverso) i punti rilevati.

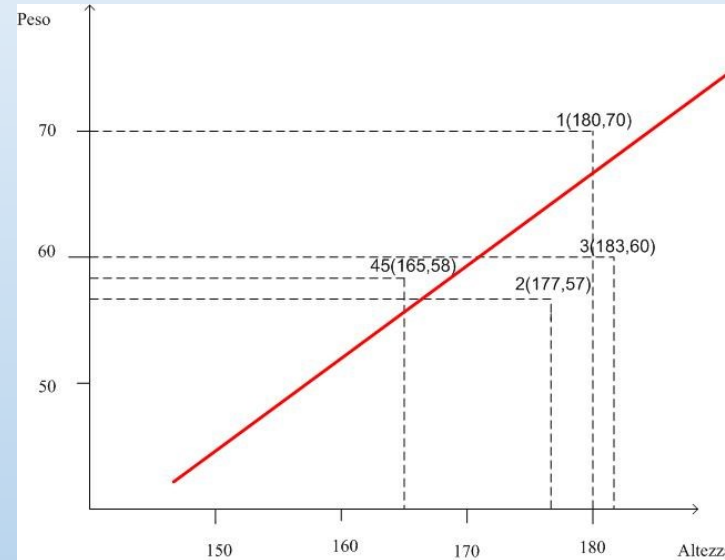
Per quantificare pero' numericamente la relazione tra due variabili dobbiamo usare dei coefficienti che si basano sulla *media* e sulla *varianza* dei dati.

I valori che occorre calcolare per ottenere un coefficiente numerico di associazione tra X ed Y sono i seguenti:

Media di X:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

| Unità | Variabile X | Variabile Y |
|-------|-------------|-------------|
| 1 | x_1 | y_1 |
| 2 | x_2 | y_2 |
| 3 | x_3 | y_3 |
| | ... | ... |
| n | x_n | y_n |



Media di Y:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

Varianza di X:

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

Varianza di Y:

$$s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$$

Deviazione Standard X:

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

Deviazione standard Y:

$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}$$

A cosa serve calcolare questi indici ?

COEFFICIENTE DI COVARIANZA:

Sintesi numerica dell'associazione tra X ed Y è il *coefficiente di covarianza*, definito come la sommatoria dei prodotti degli scarti di X e di Y dalle rispettive medie:

$$Cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Media di X:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = 6$$

| x | y | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|----|-----|-----------------|-----------------|----------------------------------|
| 3 | 30 | -3 | -80 | 240 |
| 4 | 70 | -2 | -40 | 80 |
| 5 | 130 | -1 | 20 | -20 |
| 12 | 210 | 6 | 100 | 600 |
| | | | | 900/3=300 |

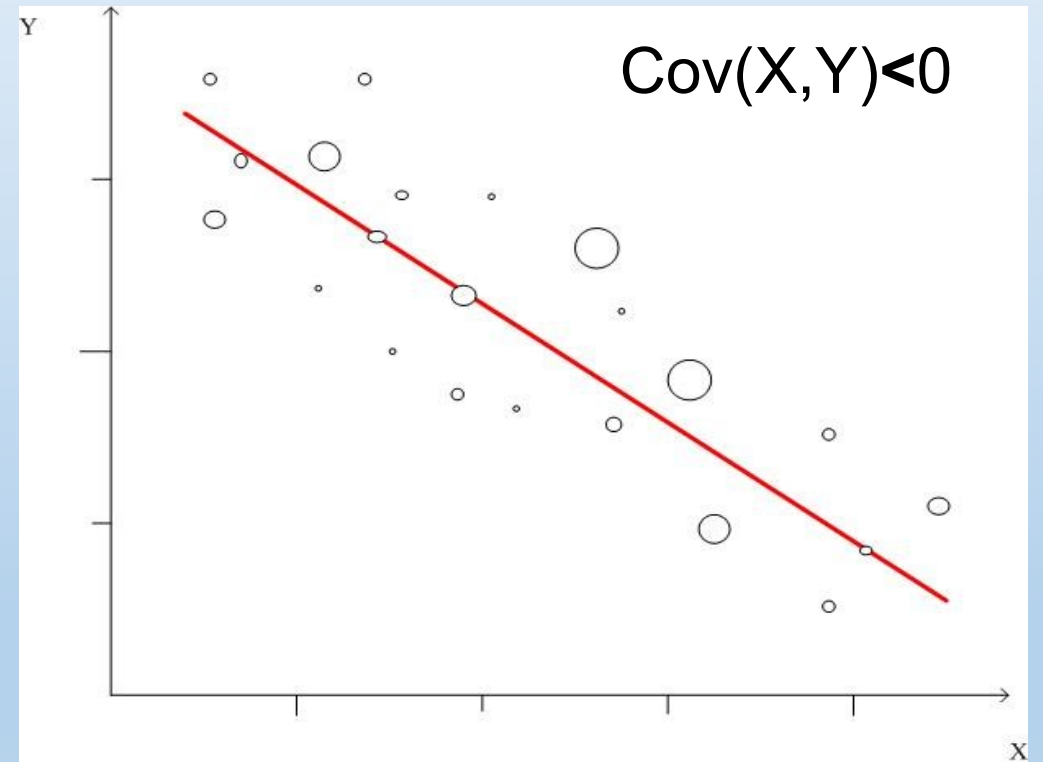
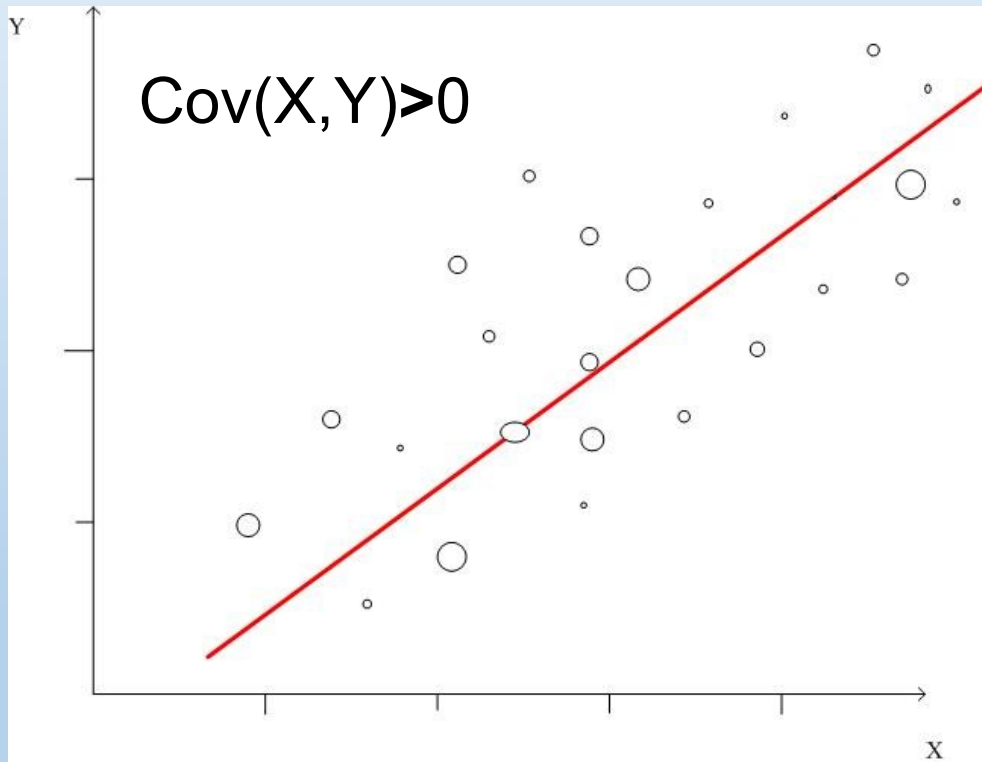
Media di Y:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = 110$$

$$Cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = 300$$

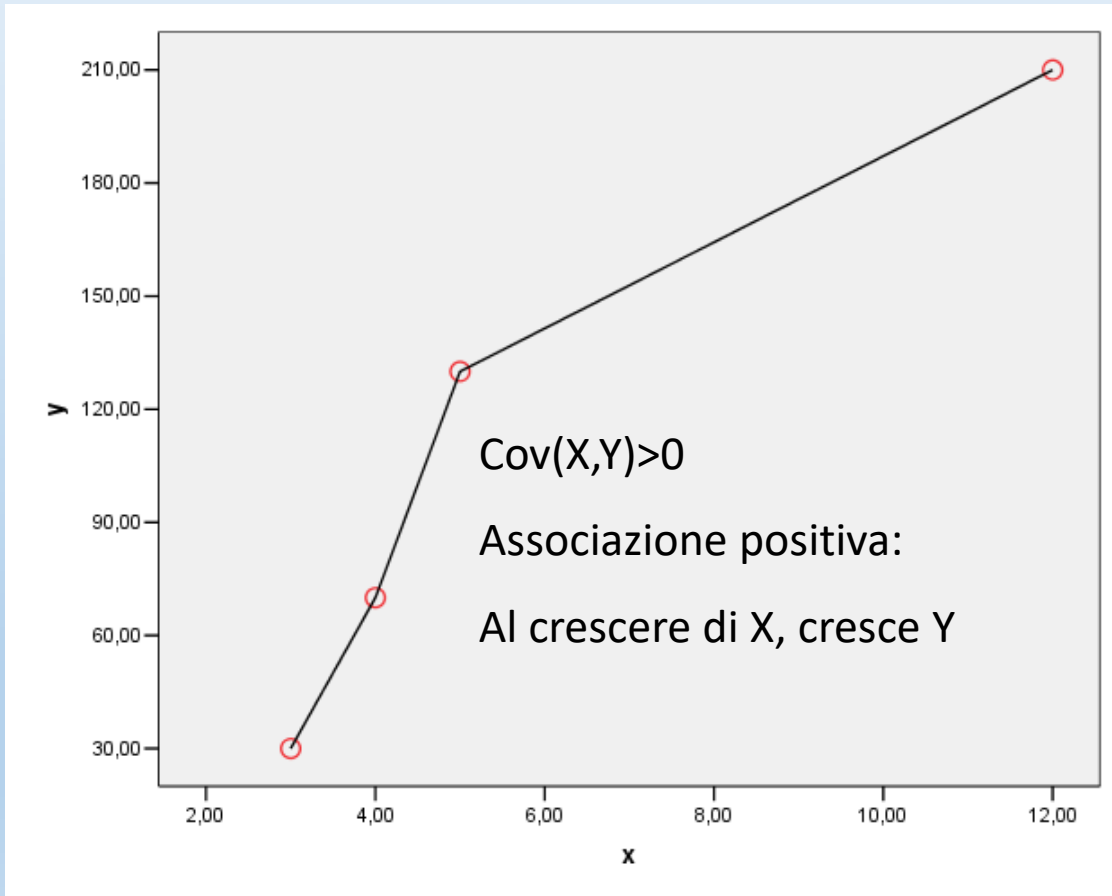
Quanto più il coefficiente di covarianza è **positivo** (cioè **maggiore di zero**) tanto più le due variabili sono associate positivamente, cioè *vanno nella stessa direzione*.

Quanto più il coefficiente di covarianza è **negativo** (ossia **minore di zero**) tanto più le due variabili sono associate negativamente, cioè *vanno in direzioni opposte*.



$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = 300$$

| x | y | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|----|-----|-----------------|-----------------|----------------------------------|
| 3 | 30 | -3 | -80 | 240 |
| 4 | 70 | -2 | -40 | 80 |
| 5 | 130 | -1 | 20 | -20 |
| 12 | 210 | 6 | 100 | 600 |
| | | | | 900/3=300 |



Regola aritmetica dei prodotti
con il segno:

+ * + = +

+ * - = -

- * - = +

Il coefficiente di covarianza *dipende* dall'*unità di misura* dei dati: nel caso di altezza e peso assume valori diversi se si passa da cm/kg a metri/grammi. Inoltre non ha un intervallo definito di valori.

Si introduce quindi un coefficiente che *non dipende* dalla scala di misura, e che assume valori in un *range* (intervallo) prestabilito:

(2) COEFFICIENTE DI CORRELAZIONE:

Il coefficiente di correlazione tra X e Y si definisce come il rapporto tra la covarianza di X e Y divisa per il prodotto tra le rispettive deviazioni standard:

$$r(X, Y) = \frac{Cov(x, y)}{S_x S_y}$$

$-1 < r < 1$; non dipende dalla scala di misura di X e Y.

r -> **1** : le due variabili sono associate *positivamente*, vanno nella stessa direzione;

r -> **-1** le due variabili sono associate *negativamente*, vanno in direzioni opposte;

r -> **0** le due variabili *non sono associate*.

$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = 16.67$$

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = 4.08$$

| x | y | x _i - \bar{x} | y _i - \bar{y} | (x _i - \bar{x})(y _i - \bar{y}) |
|----|-----|----------------------------|----------------------------|------------------------------------------------------------|
| 3 | 30 | -3 | -80 | 240 |
| 4 | 70 | -2 | -40 | 80 |
| 5 | 130 | -1 | 20 | -20 |
| 12 | 210 | 6 | 100 | 600 |
| | | | | 900/4=225 |

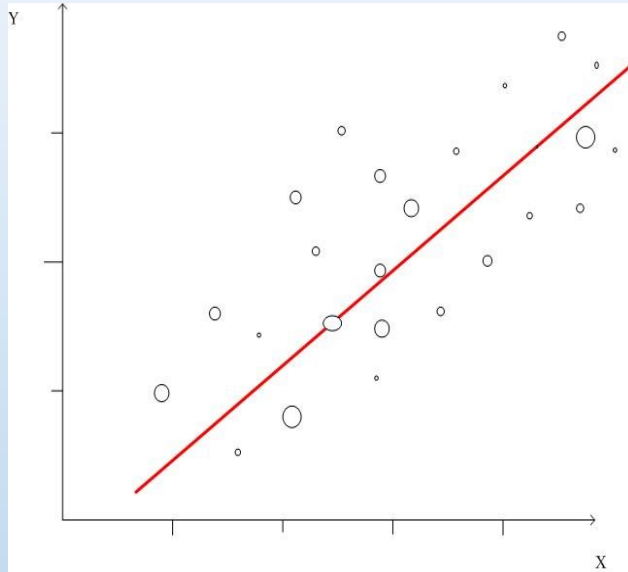
$$s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = 6133$$

$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}} = 78.32$$

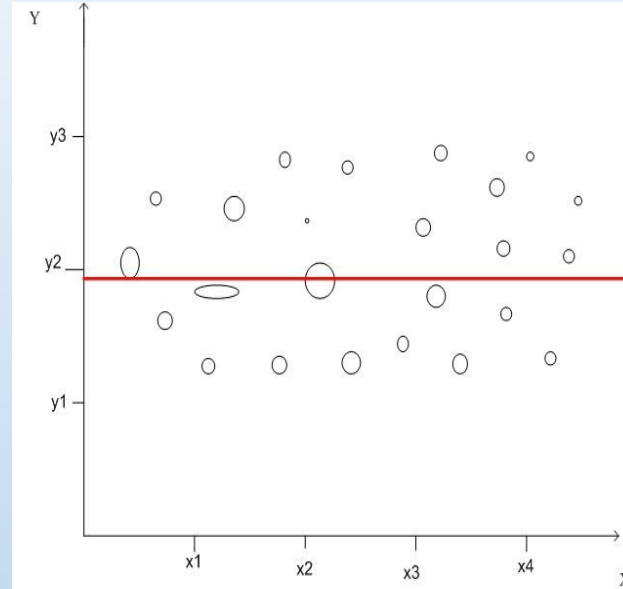
$$r(X, Y) = \frac{Cov(x, y)}{s_x s_y} = \frac{300}{4.08 * 78.32} = \frac{300}{319.77} = 0.94$$

$r(X, Y)$ è molto vicino ad **1** ed indica quindi che X e Y sono correlate positivamente (al crescere di X aumenta Y e viceversa).

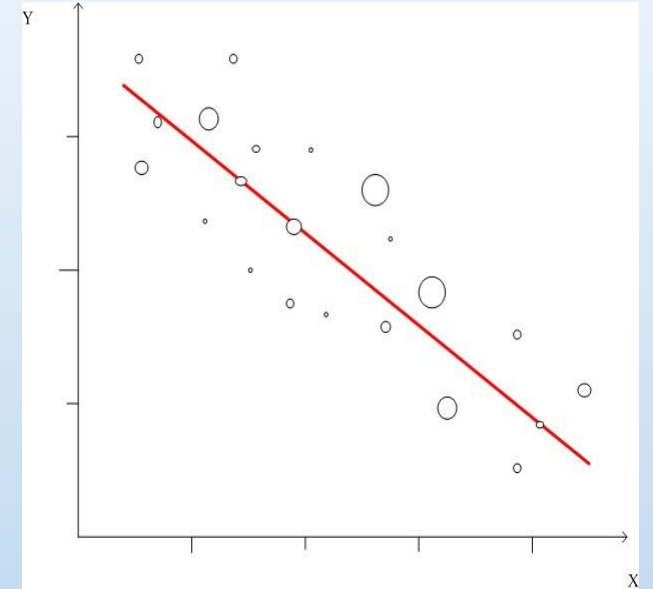
(a) $r(X,Y)$ vicino a **+1**



(b) $r(X,Y)$ vicino a **0**



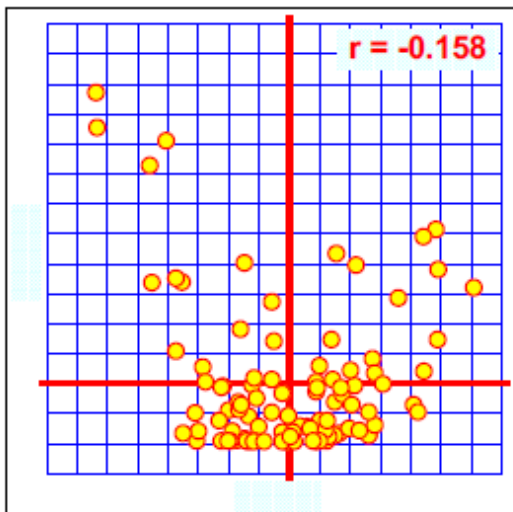
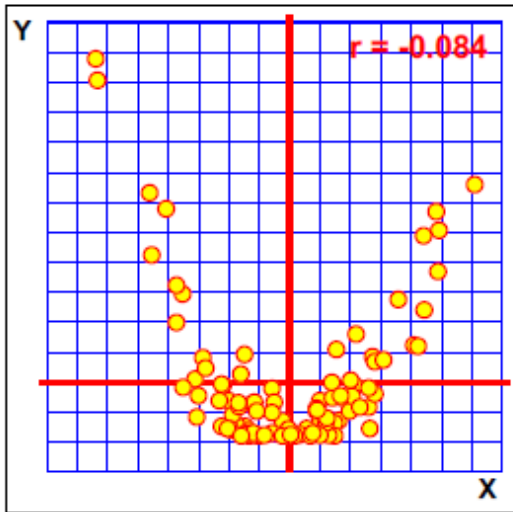
(c) $r(X,Y)$ vicino a **-1**



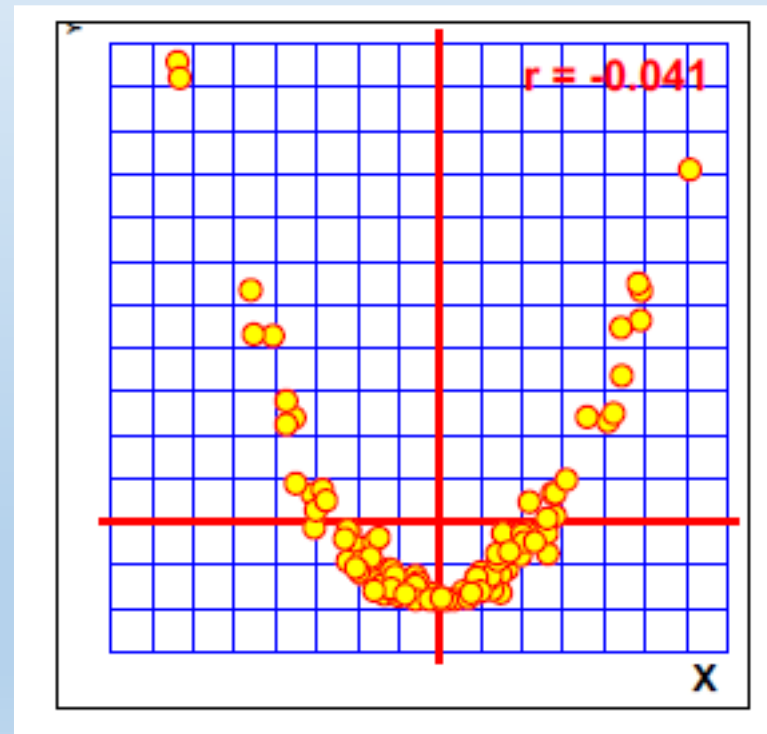
- (a) X ed Y sono fortemente associati in senso positivo (crescono insieme o decrescono insieme);
(b) X ed Y non sono associati (al variare di X, Y non cambia);
(c) X ed Y sono fortemente associati in senso negativo: al crescere di X diminuisce Y e viceversa

Vedremo in che modo è possibile effettuare un test di ipotesi sul coefficiente di correlazione per stabilire se il valore numerico trovato indica una associazione “significativa” tra i due fenomeni oppure no.

- r è un numero puro adimensionale (senza unità di misura)
- può assumere valori **compresi tra -1 (correlazione negativa) e +1 (correlazione positiva)**. **0 corrisponde ad assenza di correlazione lineare**
- non risente dello scambio delle variabili: $r(x,y)=r(y,x)$
- non risente dell'aggiunta di una stessa quantità costante a tutti i valori di una variabile
- non risente della moltiplicazione per un numero positivo costante di tutti i valori di una variabile
- **NON** misura la associazione in generale ma solo quella **LINEARE: Dispersione dei punti intorno ad una retta**
- **NON** definisce una relazione **causa-effetto**



Il coefficiente di correlazione lineare **è indice di quanto i punti si allineano su di una retta**: vi possono essere associazioni anche forti, ma di tipo non lineare per le quali il coefficiente di correlazione è prossimo a 0.



E' sempre quindi necessario **VISUALIZZARE** i dati tramite lo scatter plot per capire se è opportuno calcolare un coefficiente di correlazione lineare !!

Interpretare il coefficiente di correlazione :

L'interpretazione del coefficiente di correlazione dipende fundamentalmente dalle caratteristiche della ricerca. Come regola di pratica utilità possono essere utili i seguenti suggerimenti (che valgono anche per il segno negativo):

| ρ | <i>grado di associazione</i> |
|---------|------------------------------|
| 0.8-1.0 | forte |
| 0.5-0.8 | buona |
| 0.2-0.5 | debole |
| 0.0-0.2 | trascurabile |

Esistono 2 versioni principali del coefficiente di correlazione: **Pearson** e **Spearman**.

Il coefficiente di Spearman è da utilizzare quando una o entrambe le variabili sono **asimmetriche** oppure hanno **una scala di misura discreta con pochi valori possibili** (ad esempio dei punteggi ad un test) ed è più "robusto" quando vi siano dei **valori anomali/estremi** nella distribuzione delle variabili da correlare.

Problema: come decidere se la correlazione è *significativamente diversa* da zero?

Soluzione: Test statistico di ipotesi ! Ipotesi nulla: la coppia di variabili **non è correlata** ($r_{XY} = 0$) ; è sufficientemente screditata dai dati ??

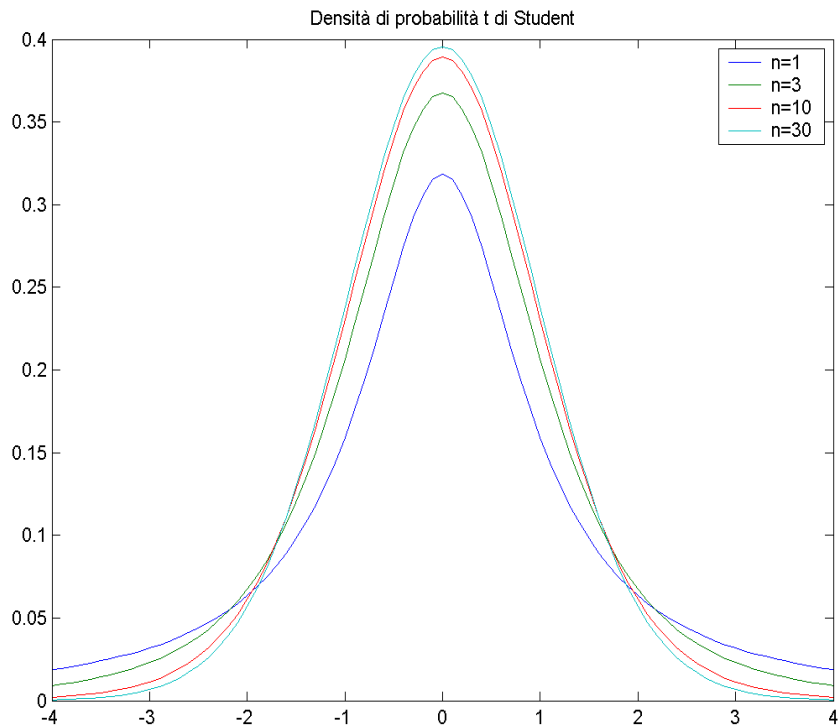


Per verificare se l'ipotesi nulla è screditata, si ha bisogno di una “distribuzione di riferimento” : la distribuzione di probabilità della statistica di test *se fosse vera* H_0 .

$$t = \frac{R_{xy}}{\sqrt{1 - R_{xy}^2}} \sqrt{n - 2} = t_{n-2}$$

(t_{n-2} : *t di Student a n-2 gradi di libertà*)

La statistica di test che viene calcolata sui dati è una V.A. che si comporta (se fosse vera l'ipotesi nulla) come una *t di Student* con «n-2» gradi di libertà.



La distribuzione ***t di Student*** è molto simile alla gaussiana. Ma cambia di *forma* in relazione alla numerosità *n* del campione: tende ad avvicinarsi alla **distribuzione normale standard** $N(0,1)$, al crescere di *n*.

Per **$n > 30$** le due distribuzioni sono indistinguibili.

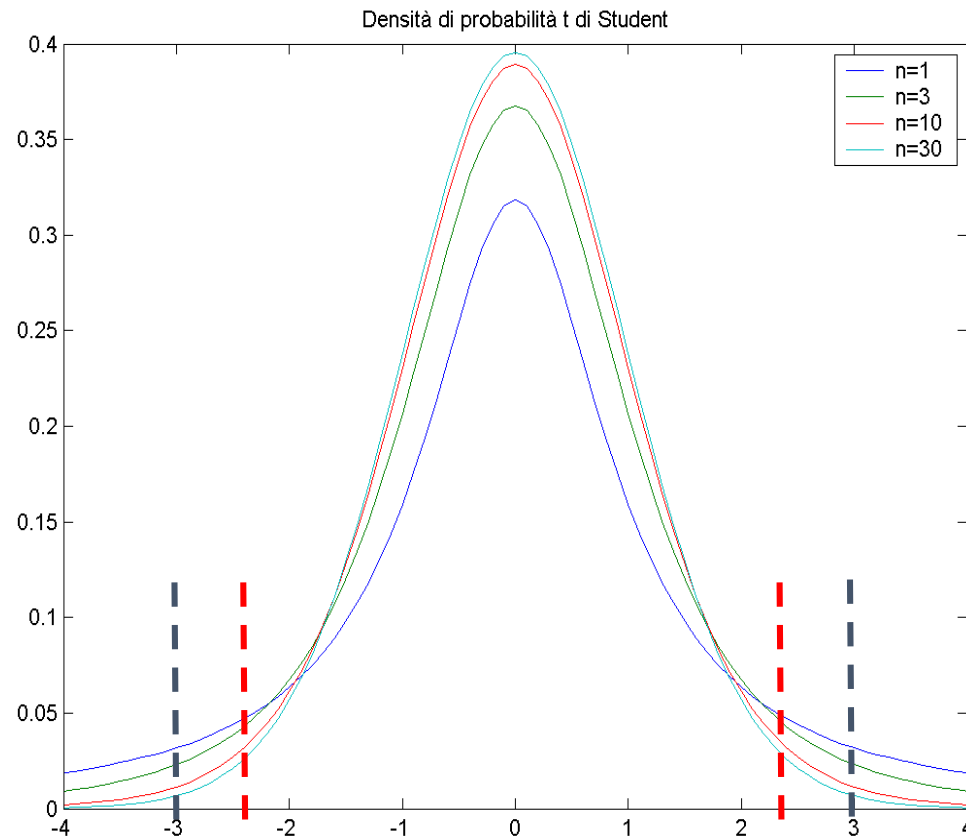
Per «gradi di libertà» della distribuzione si intende appunto questa dipendenza della forma della distribuzione dalla dimensione campionaria.

Distribuzione di probabilità se fosse vera H_0 : non c'è correlazione fra le due variabili

Test di correlazione (significatività α): Avendo calcolato t

- $|t| > t_{\alpha/2} \Rightarrow$ respingo l'ipotesi nulla $\Rightarrow R_{xy}$ è significativamente $\neq 0$

$|t| \leq t_{\alpha/2} \Rightarrow$ non respingo l'ipotesi nulla $\Rightarrow R_{xy}$ non è significativamente $\neq 0$



Sulla base della dimensione campionaria, ci saranno delle costanti t che delimiteranno le regioni di accettazione e di rifiuto della ipotesi nulla.

Il software ci darà in output il valore della statistica di test ed il corrispondente «**p-value**»:

Si «rigetta» H_0 se il valore osservato di t è troppo *poco probabile*

(se H_0 fosse vera): $p \leq \alpha$

Introduzione alla Retta di Regressione

Quando si studia una relazione causale tra due variabili quantitative, occorre definire:

- una variabile **esplicativa** o indipendente o **CAUSA**
- una variabile **dipendente** o risposta cioè un **EFFETTO**

Consideriamo le seguenti coppie di variabili:

- X= Peso; Y= Tasso di colesterolo;
- X= Età (nei bambini tra 0 e 12 anni); Y= Statura (nei bambini tra 0 e 12 anni)
- X= Dose di un farmaco; Y= "livello" di malattia/ tempo di guarigione.

Quali sono le **cause** e quali gli **effetti** ? Si cerca di *stimare* una relazione tra X e Y tramite una *funzione matematica*: **$Y=f(X)$**



In matematica, una **funzione** è una relazione tra due insiemi, che ad ogni elemento del primo insieme fa corrispondere uno e un solo elemento del secondo insieme.

Il termine 'regressione' è stato introdotto da Sir Francis Galton, antropologo inglese, nell'articolo "*Regression towards mediocrity in hereditary stature*" pubblicato nel Journal of the Anthropological Institute nel 1885.

'Regressione' si riferiva alla tendenza dei figli ad avere altezze più prossime alla media rispetto ai genitori.

Galton voleva verificare se la statura dei figli potesse essere prevista sulla base di quella dei genitori. Ed esprimere questa corrispondenza in una legge matematica.

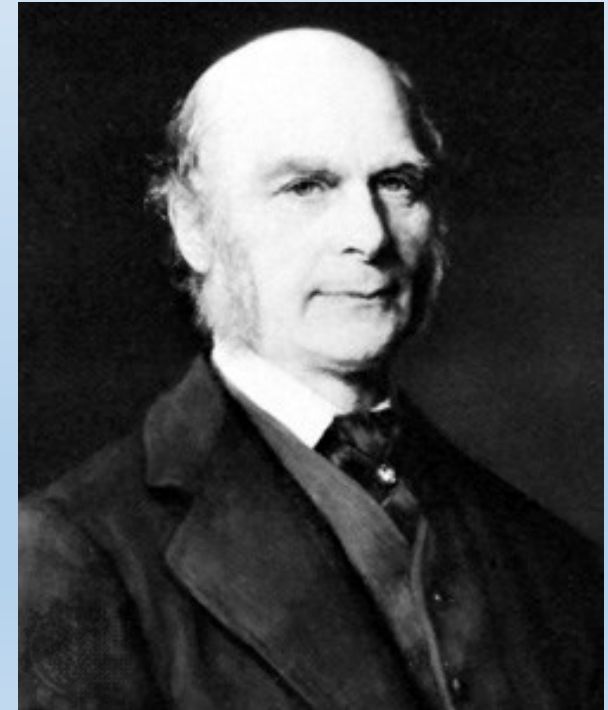
Raccolse dei dati sull'altezza dei genitori (=X) e dei loro figli (=Y), e notò che padri alti avevano figli alti, ma **più bassi** dei rispettivi genitori, mentre padri bassi avevano figli bassi, ma tendenzialmente **più alti** dei loro padri.

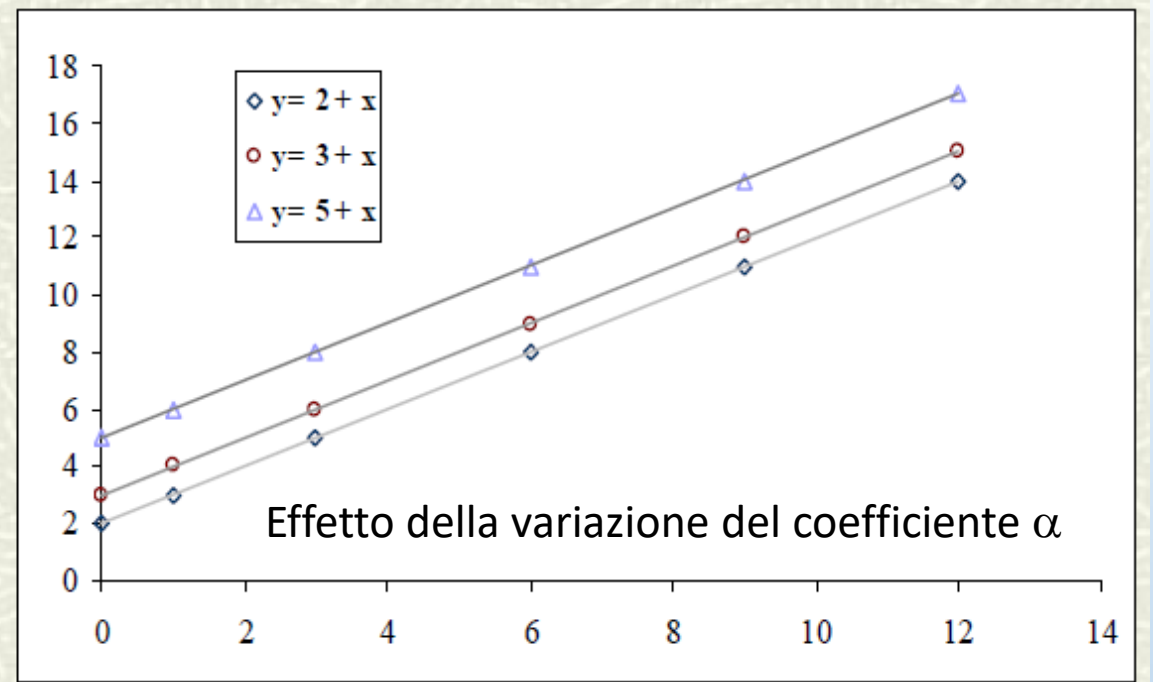
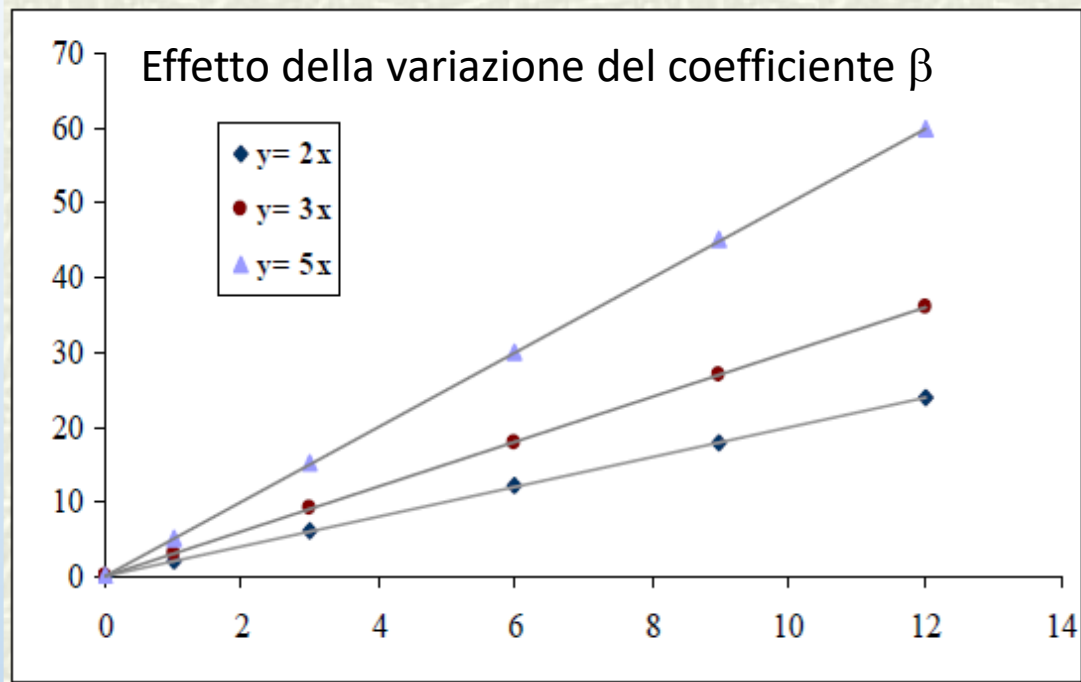
Galton chiamò questo fenomeno «**regressione alla mediocrità**», e il termine regressione venne poi applicato a tutte le analisi di questo tipo.

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

α = intercetta della retta

β = pendenza angolare





Il problema statistico è quello di determinare (stimare) i valori di α e β , detti **'coefficienti di regressione'**, a partire dai valori di X e di Y osservati su un campione.

Il metodo più comunemente usato per stimare i coefficienti di regressione è definito: *metodo dei minimi quadrati (least squares)*.

Il metodo consiste nel *minimizzare* rispetto alle incognite α e β gli scarti al quadrato tra i valori osservati di Y e i valori 'teorici' di Y , cioè quelli che ci aspetteremmo di ottenere calcolandoli dai valori assunti da X , tramite l'equazione di regressione.

A differenza però del concetto matematico di funzione, dove la relazione tra X e Y è *deterministica* cioè priva di errore casuale, in statistica tale relazione viene *stimata* usando dei dati campionari, quindi *aleatori o stocastici o casuali*. Dunque si aggiunge una componente di «errore»:

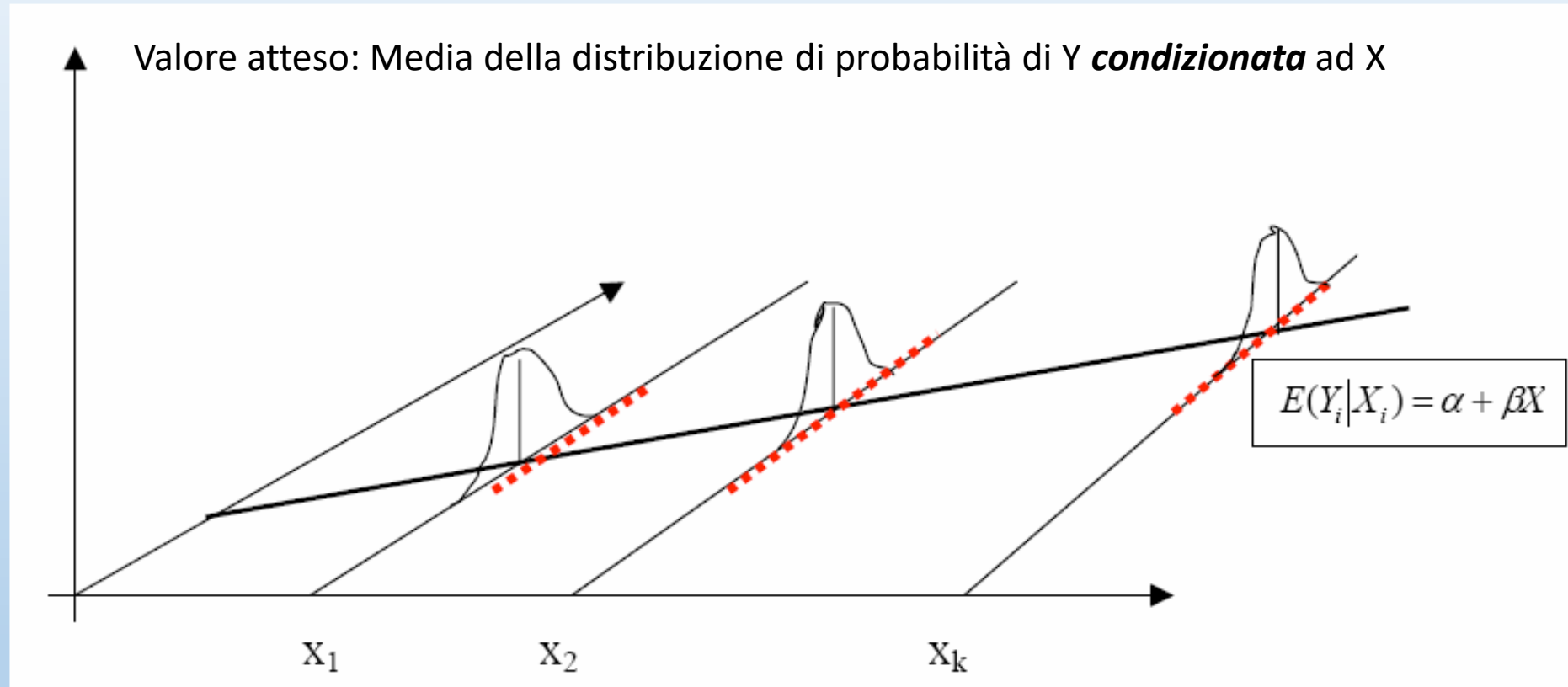
$$\begin{array}{ccc}
 \text{deterministica} & & \text{stocastica} \\
 \downarrow & & \downarrow \\
 \underbrace{\alpha + \beta X_i}_{\text{deterministica}} + u_i & & \\
 Y_i = \alpha + \beta X_i + u_i & & i=1, \dots, n
 \end{array}$$

per ogni valore assunto dalla variabile osservata X esiste un'intera distribuzione di probabilità di valori della Y e ciò significa che per ogni valore di X non sarà possibile conoscere con certezza il valore di Y ...

...quindi **Y è una variabile casuale** la cui distribuzione di probabilità è determinata dai valori della X e dalla distribuzione di probabilità del termine di errore, rappresentato da u ...

...la completa specificazione del modello di regressione include, oltre l'equazione della regressione, anche la specificazione della distribuzione di probabilità della **componente casuale** u ...[es: *gaussiana a media zero*...]

In altri termini: la retta di regressione è un'equazione lineare che associa ad ogni valore di X (*variabile esplicativa*) un **valore atteso** di Y (*variabile dipendente*):



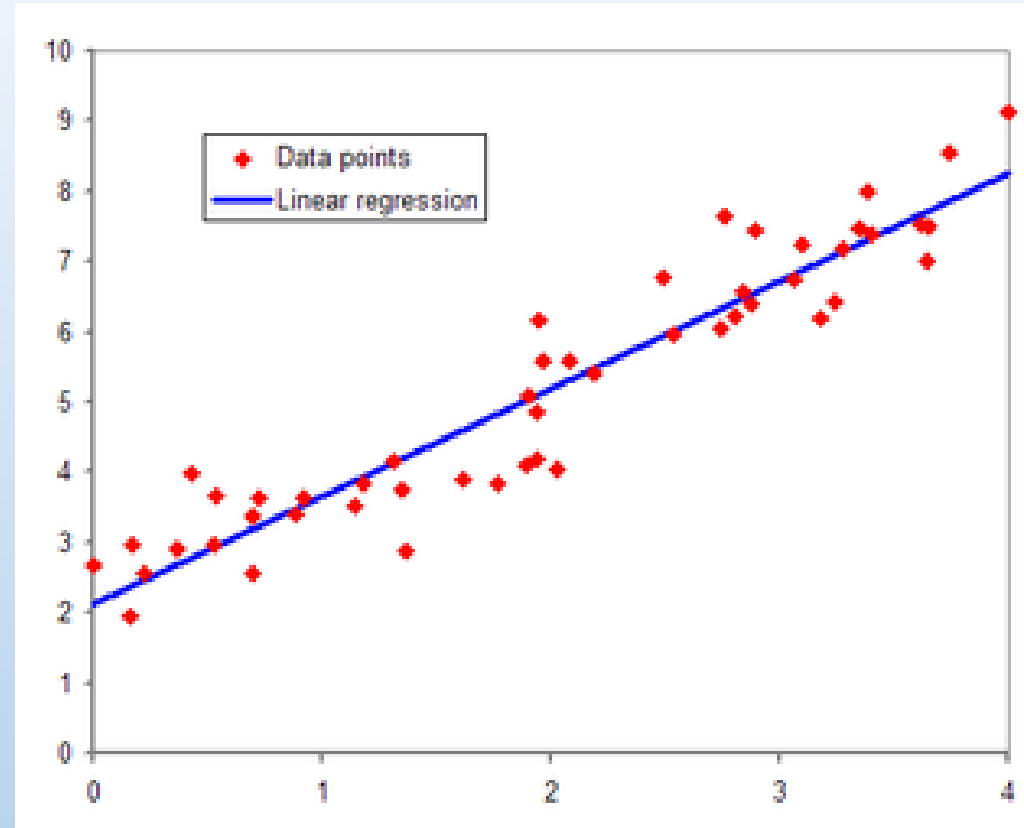
α = "intercetta" della retta

β = "pendenza" o "coefficiente angolare" della retta di regressione di Y su X .

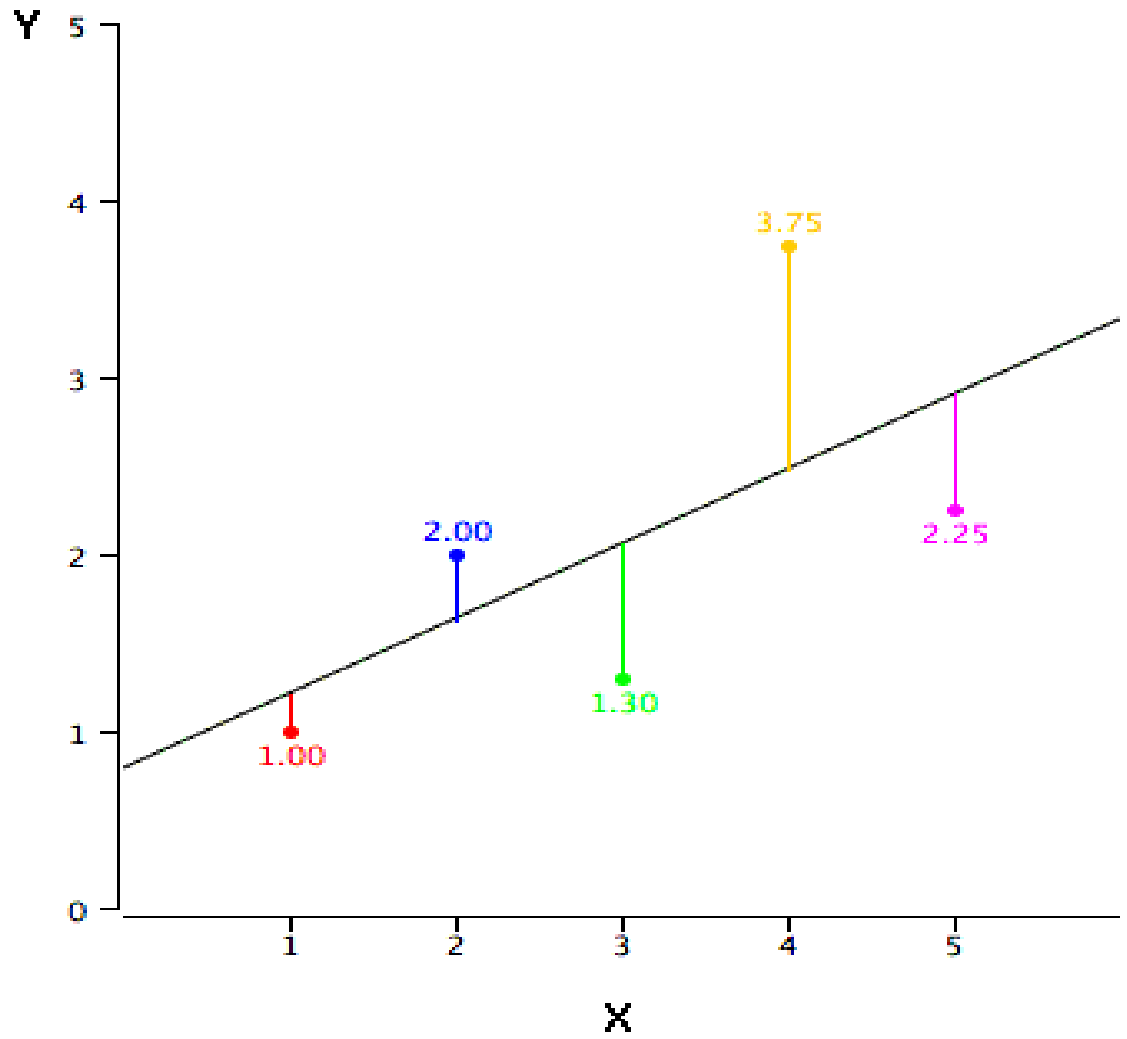
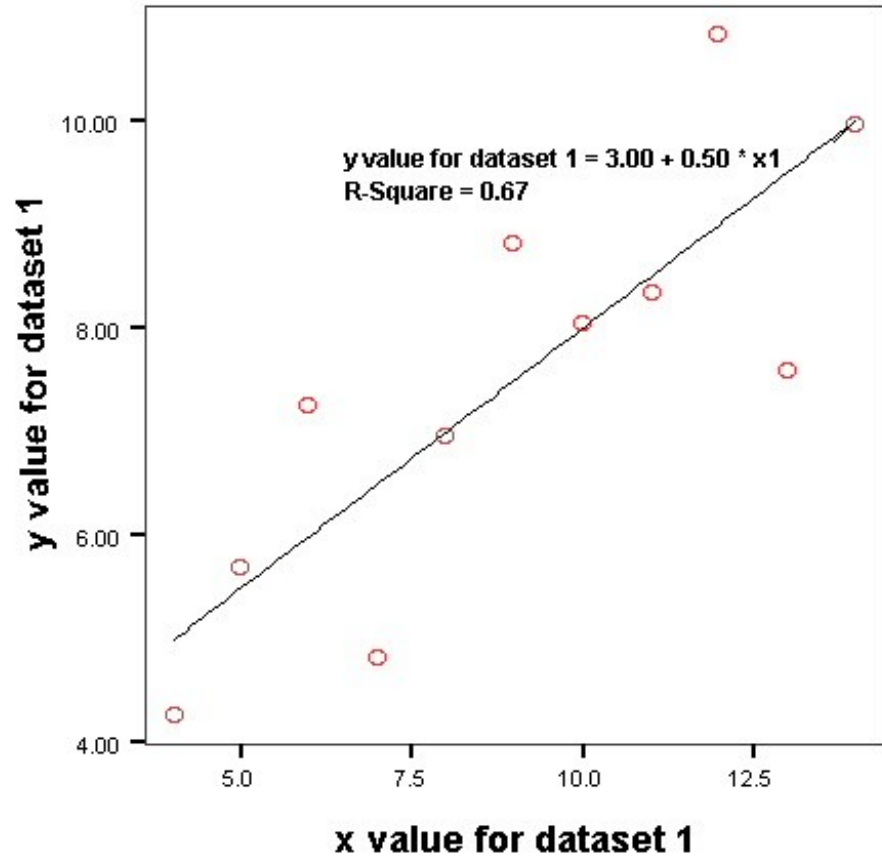
| unità | carattere X | carattere Y |
|-------|-------------|-------------|
| 1 | X_1 | Y_1 |
| 2 | X_2 | Y_2 |
| 3 | X_3 | Y_3 |
| ... | ... | ... |
| n | X_n | Y_n |

$$\min_{a,b} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = \alpha + \beta x_i$$



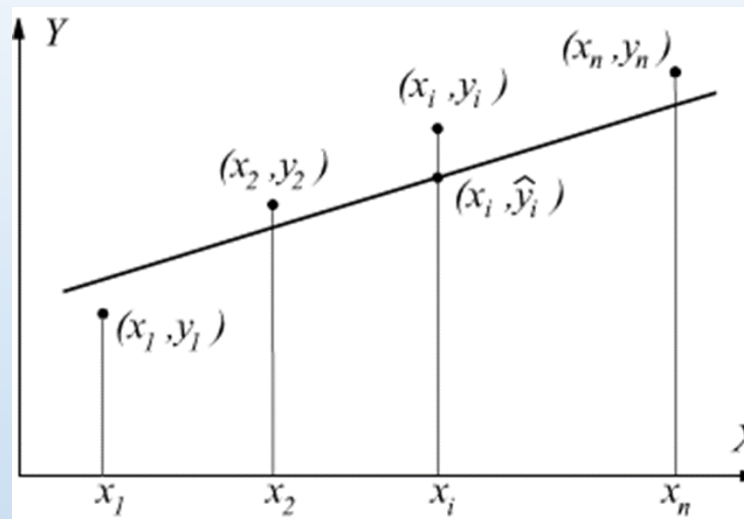
Si ottiene così l'equazione di una retta che "interpola" (=passa attraverso) la nuvola di punti osservati, in modo tale che la *distanza media* dei punti da questa retta sia *minima*.



Dunque: i coefficienti della retta di regressione si stimano dal campione di coppie di valori osservati (x_i, y_i) , per $i=1, \dots, n$, mediante il metodo dei minimi quadrati:

$$\min_{\alpha, \beta} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$$

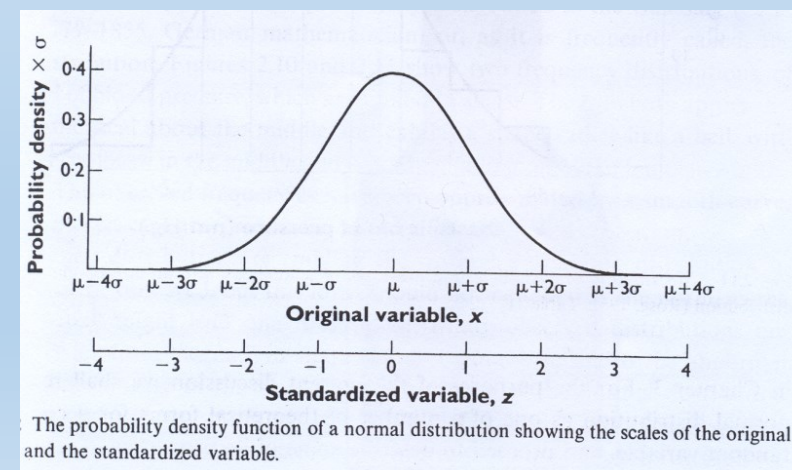


Stimando i parametri della retta sulla base del campione si commette un errore ε : la stima campionaria della parte casuale u , detta anche **residui** della regressione, che dovrebbe seguire una distribuzione gaussiana:

$$\varepsilon = y - \hat{y}$$

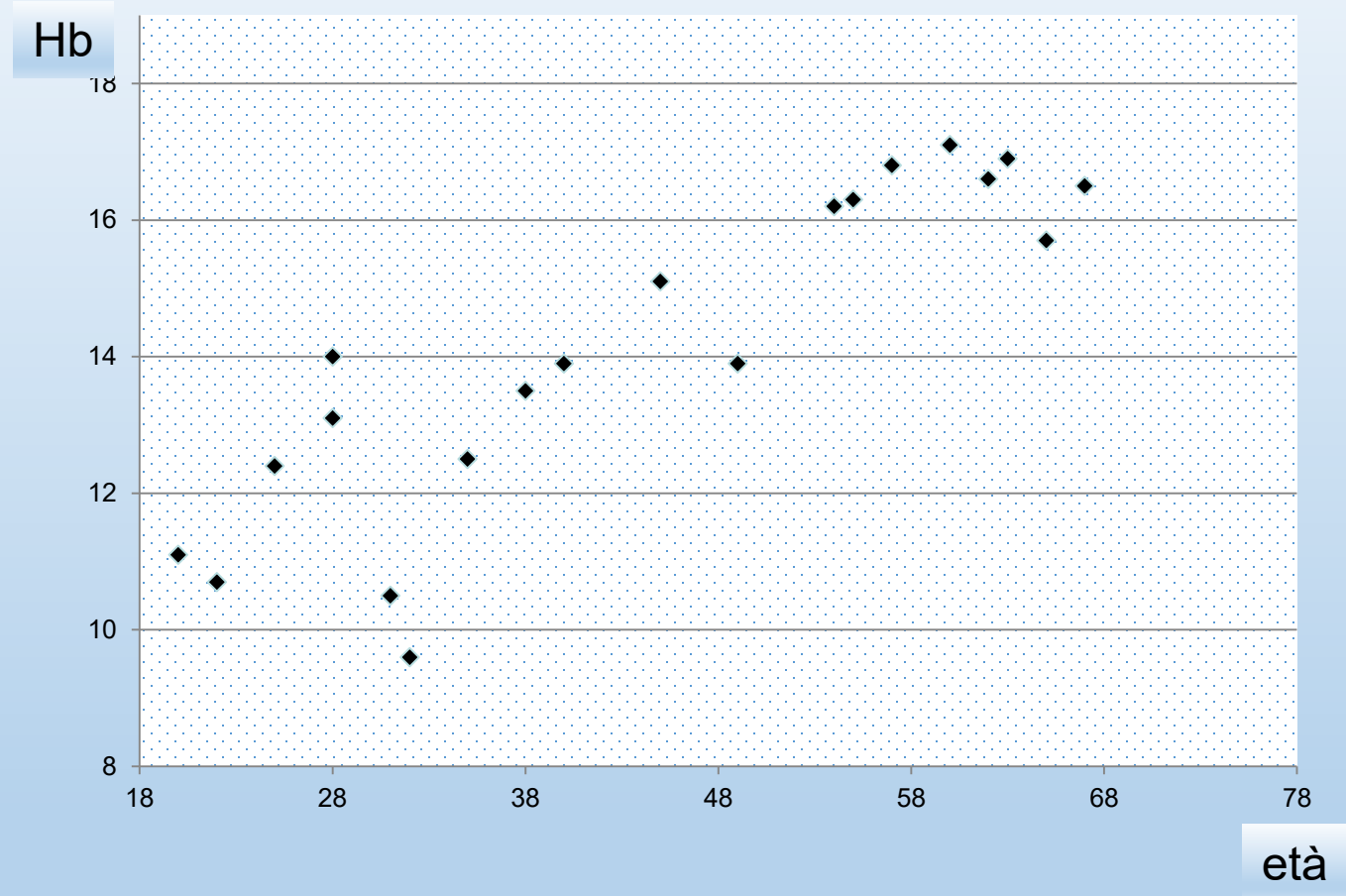
$$\varepsilon \approx N(0, \sigma)$$

la variabilità residua «non spiegata» dal modello lineare dovrebbe provenire da una V.A. gaussiana a media zero...



Esempio: valori di emoglobina ed età in 20 donne:

| Soggetto | Hb (g/dl) | età |
|----------|-----------|-----|
| 1 | 11.1 | 20 |
| 2 | 10.7 | 22 |
| 3 | 12.4 | 25 |
| 4 | 14 | 28 |
| 5 | 13.1 | 28 |
| 6 | 10.5 | 31 |
| 7 | 9.6 | 32 |
| 8 | 12.5 | 35 |
| 9 | 13.5 | 38 |
| 10 | 13.9 | 40 |
| 11 | 15.1 | 45 |
| 12 | 13.9 | 49 |
| 13 | 16.2 | 54 |
| 14 | 16.3 | 55 |
| 15 | 16.8 | 57 |
| 16 | 17.1 | 60 |
| 17 | 16.6 | 62 |
| 18 | 16.9 | 63 |
| 19 | 15.7 | 65 |
| 20 | 16.5 | 67 |



Vogliamo prevedere i valori di emoglobina al crescere della età.



Retta di regressione:

$$E(Hb) = \alpha + \beta * Et\grave{a} = 8.24 + 0.13 * Et\grave{a}$$

Output di R:

```
Call:
lm(formula = Hb..g.dl. ~ et\grave{a}, data = DATI_HB)

Residuals:
    Min       1Q   Median       3Q      Max
-2.9358 -0.5536  0.1888  0.8042  2.0012

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.23979    0.79426  10.374 5.06e-09 ***
et\grave{a}      0.13425    0.01711   7.844 3.24e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

GLOSSARIO:

1. «E»= valore atteso della emoglobina al crescere dell'età;
2. quale è il significato dell'errore standard che viene riportato sui coefficienti ?
3. Test di ipotesi associato ai coefficienti di regressione...?



Retta di regressione:

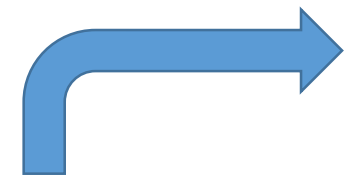
$$E(Hb) = \alpha + \beta * Et\grave{a} = 8.24 + 0.13 * Et\grave{a}$$

Output di R:

```
Call:
lm(formula = Hb..g.dl. ~ et\grave{a}, data = DATI_HB)

Residuals:
    Min       1Q   Median       3Q      Max
-2.9358 -0.5536  0.1888  0.8042  2.0012

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.23979    0.79426  10.374 5.06e-09 ***
et\grave{a}      0.13425    0.01711   7.844 3.24e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Il **test di ipotesi** viene effettuato su ogni coefficiente stimato per vedere **se \u00e8 significativamente diverso da zero**:

H0: beta=0
H1: beta\u22600

Regressione lineare multipla

- Relazione tra una variabile continua e **un insieme** di variabili continue o categoriche:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i$$

- coefficienti di *regressione parziale* β_i
 - valore del cambiamento di y **in media** quando x_i cambia di una unità e tutte le altre x_j , per $j \neq i$, rimangono costanti
 - misura l'associazione tra x_i ed y **corretta** per tutte le altre x_j

Esempio:

- PAS (pressione arteriosa sistolica, **outcome**) *verso* : età, sesso, PAD, etc...

Terminologia della Regressione lineare multipla

$$\underline{E(y)} = \alpha + \underline{\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i}$$

Variabile dipendente

Outcome

Predetta

Variabile Risposta

Variabile Esito...

Variabili indipendenti

Variabili predittive

Variabili esplicative

Covariate...

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i + \boldsymbol{\varepsilon} \longrightarrow \text{componente stocastica di errore}$$

[Si assume sempre una distribuzione gaussiana per l'errore]

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

Su un campione di n pazienti si misurano k variabili

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Stessa equazione di prima, ma scritta utilizzando vettori/matrici



Regressione lineare multipla con R

MATRICE DEI DATI

```
RegModel.1 <- lm(pas~eta.arr+pad+sezzo, data=dati)
summary(RegModel.1)
```

```
Call:
lm(formula = pas ~ eta.arr + pad + sesso, data = dati)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-29.157  -8.191  -0.437   7.643  32.074
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.1135     9.5881    5.122 1.33e-06 ***
eta.arr       0.2191     0.0962    2.277  0.0253 *
pad           0.8697     0.1355    6.418 7.24e-09 ***
sezzo        1.7413     2.8042    0.621  0.5363
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

| ID | PAS | ETA' | PAD | SESSO |
|-----|-----|------|-----|-------|
| 1 | 100 | 57 | 90 | 0 |
| 2 | 110 | 81 | 78 | 0 |
| 3 | 120 | 35 | 60 | 1 |
| 4 | 100 | 48 | 87 | 1 |
| ... | ... | ... | ... | ... |

Ho: per **ogni** coefficiente di regressione: $\beta_i=0$

$$E(PAS) = \alpha + \beta_1 * SESSO + \beta_2 * ETA + \beta_3 * PAD$$

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | 49.1135 | 9.5881 | 5.122 | 1.83e-06 | *** |
| eta.arr | 0.2191 | 0.0962 | 2.277 | 0.0253 | * |
| pad | 0.8697 | 0.1355 | 6.418 | 7.24e-09 | *** |
| sezzo | 1.7413 | 2.8042 | 0.621 | 0.5363 | |

$$E(PAS) = 49.1 + 0.22 * ETA' + 0.87 * PAD$$

SESSO : 0=F; 1=M (non significativo)

La PAS aumenta in media di 0.22 all'aumentare di un anno di età e di 0.87 all'aumentare di una unità di PAD (pressione arteriosa diastolica).

A parità di PAD, la PAS è mediamente più alta di 0.22 all'aumentare di una unità di ETA (-> aumentare di 1 anno di età).

A parità di età, la PAS è mediamente più alta di 0.87 all'aumentare di una unità di PAD.

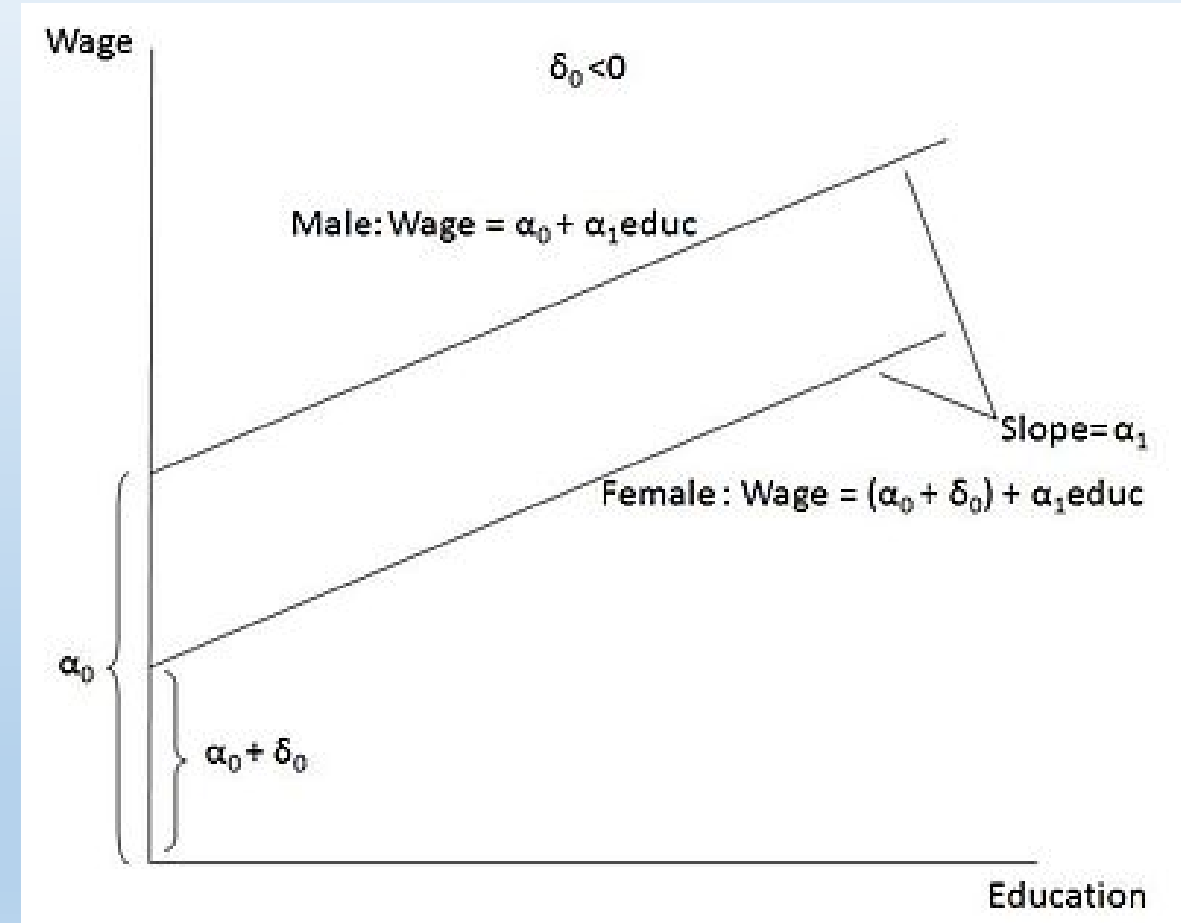
Variabile indipendente binaria

Esempio: Si vuole mettere in relazione il reddito al sesso (e gli anni di scolarità)

$$\text{Wage} = \alpha_0 - \delta_0 * \text{Sesso} + \alpha_1 * \text{education}$$

Dobbiamo aver stabilito un ***livello di riferimento*** per la variabile binaria: in questo caso è essere Maschio (livello 0).

Sesso= 0 (Maschio) ; Sesso=1 (Femmina)



Variabile indipendente categoriale: le *dummy variables*

Quando la variabile indipendente è di tipo qualitativo (categorico a più livelli o ordinale) dobbiamo trasformarla in tante variabili dicotomiche quante sono il numero delle categorie della variabile qualitativa - 1

N-1 perché una categoria è utilizzata come **riferimento**

Un'unica variabile sarà quindi trasformata in N-1 variabili e avremo N-1 coefficienti di regressione da stimare.

Esempio:

- Peso (variabile dipendente)
- Mangiare frutta : variabile indipendente, categorica a 3 livelli:
 - mai
 - saltuariamente
 - tutti i giorni



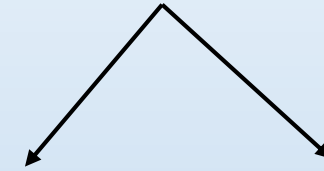
Mangiare frutta:

Mai : livello di riferimento (non ha coefficiente)

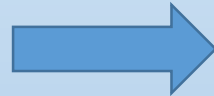
Saltuariamente : se si 1 altrimenti 0

Tutti i giorni : se si 1 altrimenti 0

Dummy variables



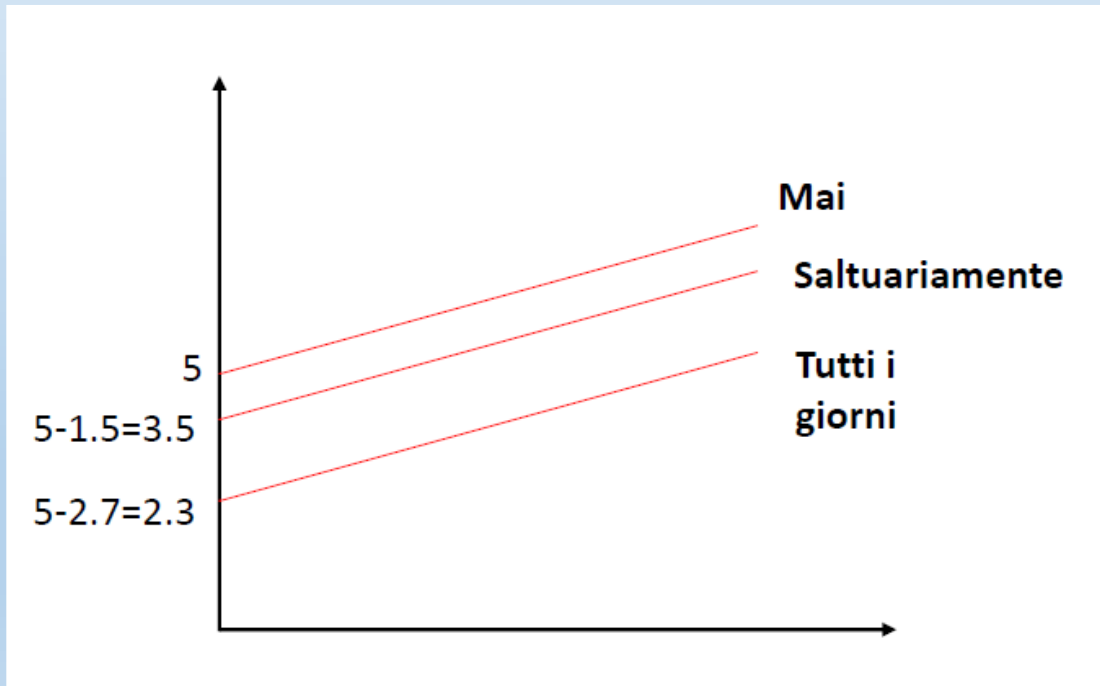
| SOGGETTO | PESO (kg) | FRUTTA |
|----------|-----------|----------------|
| 1 | 56 | mai |
| 2 | 81 | saltuariamente |
| 3 | 90 | tutti i giorni |
| 4 | 47 | tutti i giorni |
| 5 | 65 | tutti i giorni |
| 6 | 67 | mai |
| 7 | 72 | mai |
| 8 | 65 | mai |
| 9 | 54 | saltuariamente |
| 10 | 89 | saltuariamente |
| 11 | 60 | tutti i giorni |
| 12 | 55 | tutti i giorni |



| SOGGETTO | PESO (kg) | FRUTTA_saltuaria | FRUTTA_tutti |
|----------|-----------|------------------|--------------|
| 1 | 56 | 0 | 0 |
| 2 | 81 | 1 | 0 |
| 3 | 90 | 0 | 1 |
| 4 | 47 | 0 | 1 |
| 5 | 65 | 0 | 1 |
| 6 | 67 | 0 | 0 |
| 7 | 72 | 0 | 0 |
| 8 | 65 | 0 | 0 |
| 9 | 54 | 1 | 0 |
| 10 | 89 | 1 | 0 |
| 11 | 60 | 0 | 1 |
| 12 | 55 | 0 | 1 |

$$\text{peso} = 5 - 1.5 * \text{saltuariamente} - 2.7 * \text{tutti_i_giorni}$$

- Se un soggetto mangia frutta *saltuariamente* peserà in media 1.5 kg in meno **rispetto a chi non la mangia mai**
- Se un soggetto mangia frutta *tutti i giorni* peserà in media 2.7 kg in meno **rispetto a chi non la mangia mai**



Le variabili categoriche traslano la retta verso l'alto o verso il basso [intercetta]*

I coefficienti vanno infatti a sommarsi/sottrarsi all'intercetta

*a meno che non si inseriscano esplicitamente delle **interazioni**

Regressione logistica

- Modella la relazione tra un set di variabili continue o categoriche x_i
 - dicotomiche (sesso: maschio/femmina)
 - categoriche (classe sociale, titolo di studio...)
 - continue (età, peso, altezza...)



e

• **Variabile di outcome dicotomica Y** $y_i = \begin{cases} 1 \\ 0 \end{cases}$

- esito dicotomico (binario) situazione molto comune in biologia e epidemiologia:
evento sì/no (patologia, morte, infarto,)

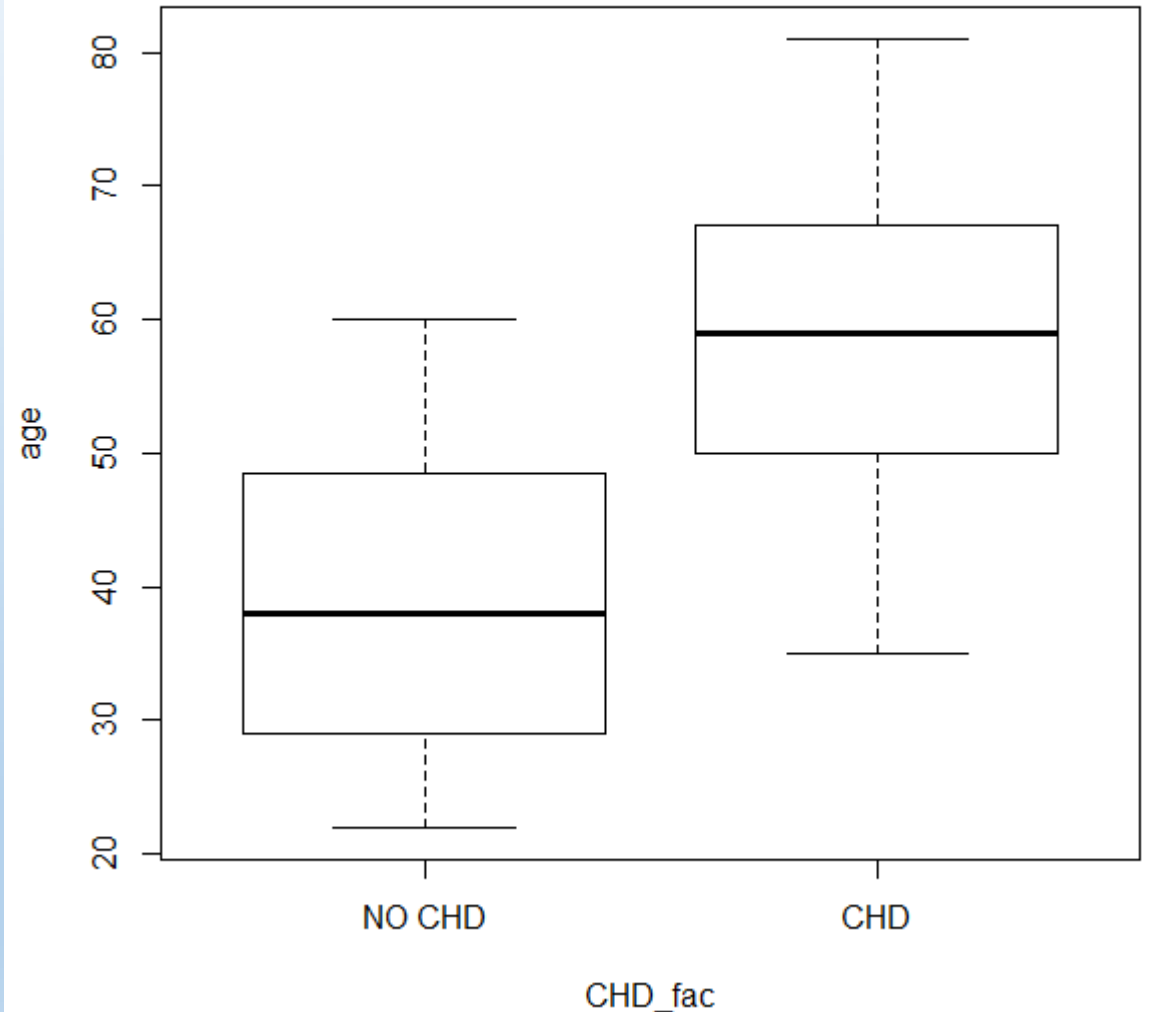
Esempio: età e sintomi di malattia coronarica (CHD:0=NO; 1=SI)

| Età | CHD | Età | CHD | Età | CHD |
|-----|-----|-----|-----|-----|-----|
| 22 | 0 | 40 | 0 | 54 | 0 |
| 23 | 0 | 41 | 1 | 55 | 1 |
| 24 | 0 | 46 | 0 | 58 | 1 |
| 27 | 0 | 47 | 0 | 60 | 1 |
| 28 | 0 | 48 | 0 | 60 | 0 |
| 30 | 0 | 49 | 1 | 62 | 1 |
| 30 | 0 | 49 | 0 | 65 | 1 |
| 32 | 0 | 50 | 1 | 67 | 1 |
| 33 | 0 | 51 | 0 | 71 | 1 |
| 35 | 1 | 51 | 1 | 77 | 1 |
| 38 | 0 | 52 | 0 | 81 | 1 |

Su un campione di 33 soggetti è stata registrata l'età e la presenza o assenza di sintomi di malattia coronarica. L'obiettivo è studiare se esiste una *relazione di tipo causa-effetto* tra l'età e la presenza di tali sintomi.

Come possiamo rispondere al quesito ?

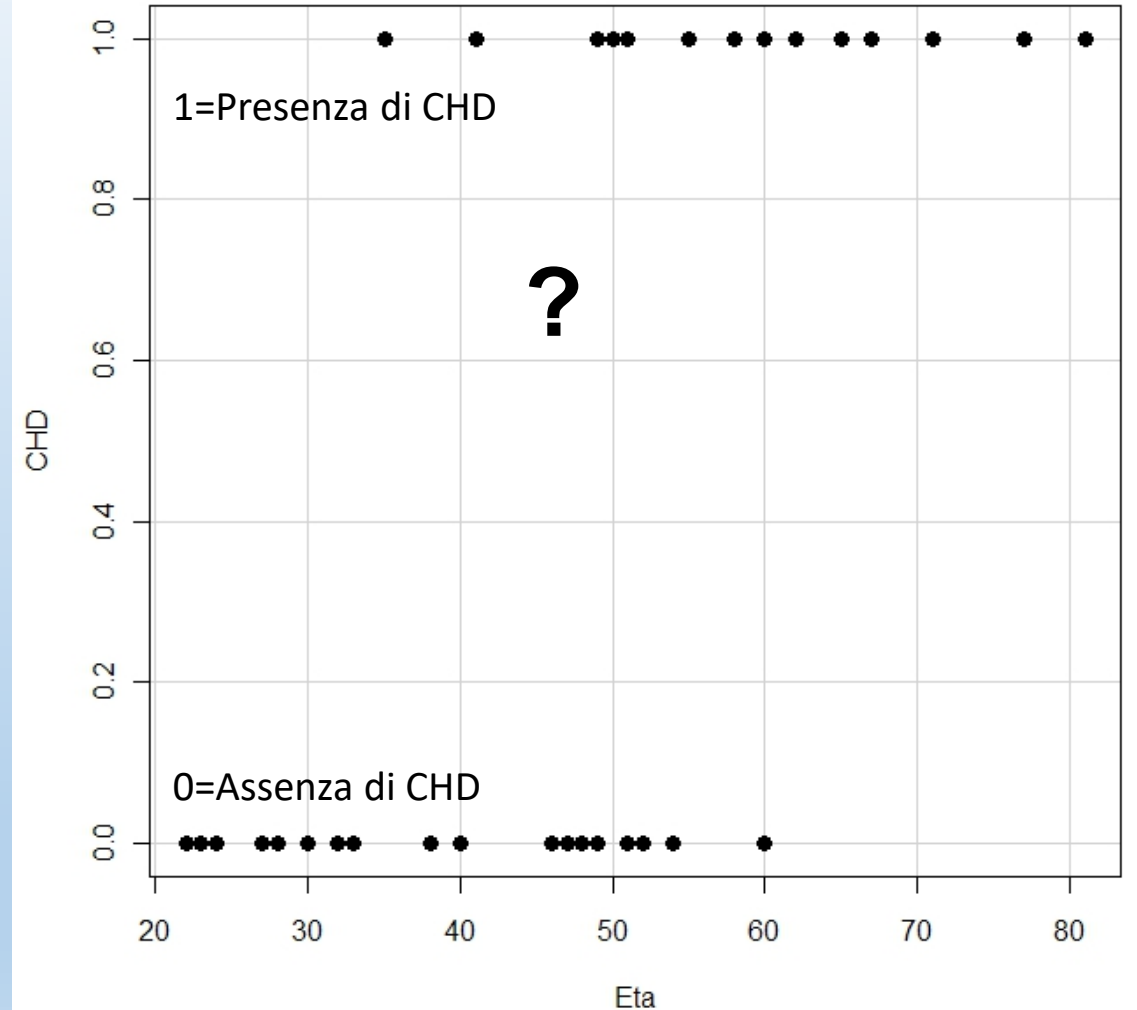
- Confronto della età media di soggetti malati e non-malati (t-test):
 - Non-Malati: 38.6 anni
 - Malati: 58.7 anni
($p=0.0001$)
- Ma la media da sola non è sufficiente a rispondere al quesito: vogliamo infatti valutare il rapporto *causa-effetto* tra età e sintomi...
- ...Regressione Lineare?



Scatter plot: età vs sintomi

Lo scatter plot non ha senso, perché la variabile risposta CHD è su una scala dicotomica; ci offre una indicazione che le persone più giovani di età sono prevalentemente nel gruppo **senza** malattia coronarica e le persone più anziane sono prevalentemente nel gruppo **con** malattia coronarica....

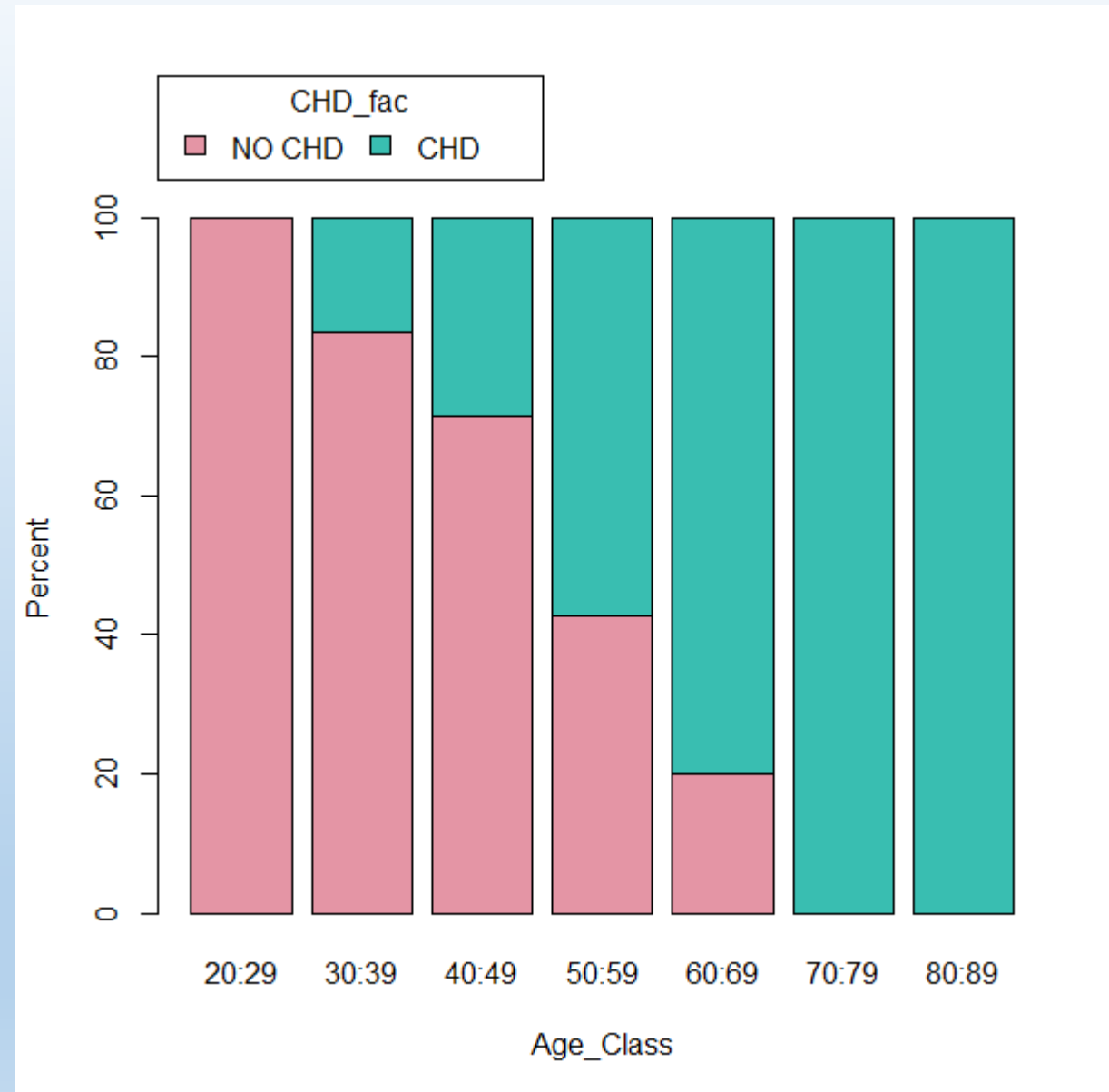
Ma non abbiamo ancora una relazione che ci dica numericamente **di quanto cresce** il **rischio** di malattia coronarica al crescere della età.



Prevalenza (%) dei segni di CHD in base al gruppo di età: Tabella di Contingenza ?

| Classe età | # in gruppo | Malati | |
|------------|-------------|--------|-----|
| | | # | % |
| 20 - 29 | 5 | 0 | 0 |
| 30 - 39 | 6 | 1 | 17 |
| 40 - 49 | 7 | 2 | 29 |
| 50 - 59 | 7 | 4 | 57 |
| 60 - 69 | 5 | 4 | 80 |
| 70 - 79 | 2 | 2 | 100 |
| 80 - 89 | 1 | 1 | 100 |

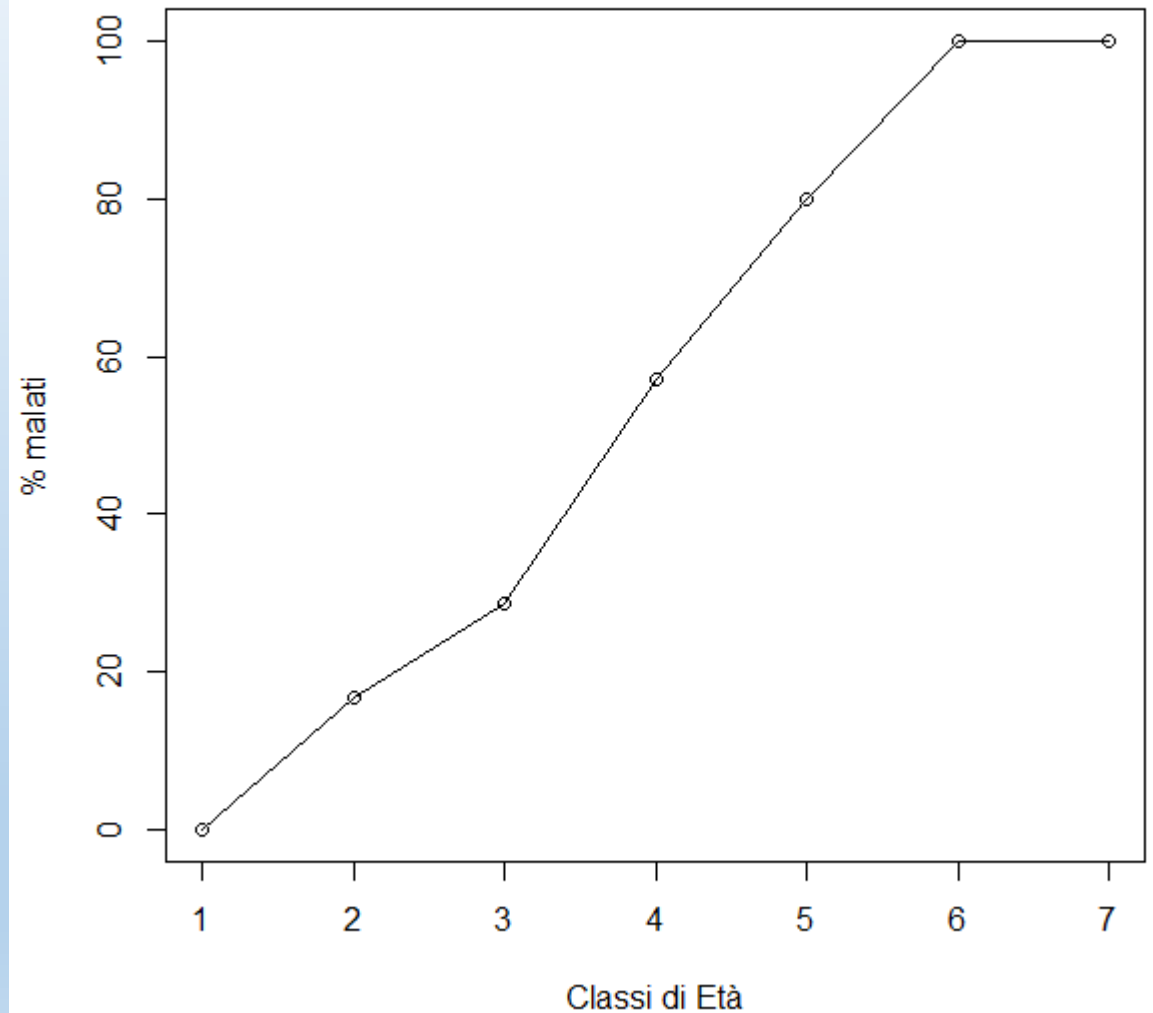
Ci stiamo avvicinando...



Scatter plot: classi di età vs % di malati

Questo scatter plot ha già più senso: abbiamo messo sull'asse delle ordinate una frequenza percentuale, che potrebbe essere interpretata come una probabilità condizionata: $P(\text{Malattia} | \text{Classe di Età})$. Vediamo che tale probabilità cresce al crescere della classe di età di appartenenza. Dobbiamo quindi trovare una funzione matematica (F) che ci faccia passare da una variabile dipendente dicotomica ad una su una scala numerica compresa tra 0 e 1 (come quella della probabilità).

$$P(Y|X) = F(\alpha + \beta X)$$



La funzione logistica:

Si *linearizza* la relazione tra X e la probabilità dell'evento tramite il «**logit**» e così si possono stimare i coefficienti di regressione α e β in modo «semplice»

$$\ln \left[\frac{P(y | x)}{1 - P(y | x)} \right] = \alpha + \beta x$$

Questa funzione ha le proprietà matematiche *opportune* per essere utilizzata in un modello di regressione.

Inoltre, produce delle stime che hanno una *interpretazione biologicamente ragionevole* in termini di rischio di evento.

La funzione logistica e il modello di regressione logistica:

$$\ln \left[\frac{P(y|x)}{1 - P(y|x)} \right] = \alpha + \beta x \quad \longleftrightarrow \quad P(y|x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

logit di $P(y|x)$

Variabile Aleatoria di Bernoulli (singola unità):

$$\Pr\{Y_i = y_i\} = \pi_i^{y_i} (1 - \pi_i)^{1 - y_i}$$

$$y_i = 0, 1$$

$$E(Y_i) = \mu_i = \pi_i$$

$$\text{var}(Y_i) = \sigma_i^2 = \pi_i(1 - \pi_i)$$

Variabile Aleatoria Binomiale (sul campione)*:

$$Y_i \sim B(n_i, \pi_i)$$

$$E(Y_i) = \mu_i = n_i \pi_i$$

$$\text{var}(Y_i) = \sigma_i^2 = n_i \pi_i (1 - \pi_i)$$

$$y_i = \begin{cases} 1 \\ 0 \end{cases}$$

*La parte **stocastica** del modello di regressione logistica è descritta dalla distribuzione di probabilità della Y

• Vantaggi del logit:

- Crea una relazione lineare con X
- è continua (logit può variare tra -∞ e ∞)
- c'è un diretto legame con la nozione di malattia...

$$\text{ODDS RATIO} = \frac{\text{odds di esposizione nei casi}}{\text{odds di esposizione nei controlli}}$$

$$\text{ODDS RATIO} = \frac{a/c}{b/d} = \frac{a}{b} \cdot \frac{d}{c} = \frac{a \cdot d}{b \cdot c}$$

dove

| | | | |
|-------------|----|----------|----------|
| | | malattia | |
| | | SI | NO |
| esposizione | SI | a | b |
| | NO | c | d |

Esposizione aumentata

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta x$$

$$\frac{P}{1-P} = e^{\alpha + \beta x}$$

| | | | Totale |
|-------------|--------|------|--------|
| | MALATI | SANI | |
| ESPOSTI | a | b | a+b |
| NON ESPOSTI | c | d | c+d |
| Totale | a+c | b+d | N |

Interpretazione di β in termini di ODDS:
(se la covariata X è dicotomica)

| | Evento | Non Evento |
|-----------------------|------------|--------------|
| Esposizione (X=1) | $P(Y X=1)$ | $1-P(Y X=1)$ |
| Non Esposizione (X=0) | $P(Y X=0)$ | $1-P(Y X=0)$ |



| | Evento | Non Evento |
|-----------------------|---------------------------------------------------------|--------------------------------------|
| Esposizione (X=1) | $\frac{\exp(\alpha + \beta)}{1 + \exp(\alpha + \beta)}$ | $\frac{1}{1 + \exp(\alpha + \beta)}$ |
| Non Esposizione (X=0) | $\frac{\exp(\alpha)}{1 + \exp(\alpha)}$ | $\frac{1}{1 + \exp(\alpha)}$ |

Calcolo dell'Odds Ratio:

$$\frac{\frac{\exp(\alpha + \beta)}{1 + \exp(\alpha + \beta)} * \frac{1}{1 + \exp(\alpha)}}{\frac{\exp(\alpha)}{1 + \exp(\alpha)} * \frac{1}{1 + \exp(\alpha + \beta)}} = \frac{\exp(\alpha + \beta)}{\exp(\alpha)} = \exp(\beta)$$

L'Odds Ratio ci dice quanto è probabile che ci sia l'evento tra coloro che hanno X=1 rispetto a coloro che hanno X=0.

Esempio: X=Fumo (Si, No);
Evento=cancro al polmone.
OR=2 significa che il cancro al polmone è diagnosticato il doppio delle volte tra i fumatori rispetto ai non fumatori.

Interpretazione di β :
 (se la covariata X è continua)

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta x$$

Se X è una variabile continua (per es. età in anni) β esprime la variazione in *log odds* della probabilità di evento per una variazione unitaria di X , sulla sua scala di misura.

Esempio: effetto dell'età sulla presenza di malattia coronarica (CHD):

$$\begin{aligned} \ln\left(\frac{P}{1-P}\right) &= \alpha + \beta X \\ &= -6.71 + 0.13 * Age \end{aligned}$$

$$\exp(0.13) = 1.14$$

Per ogni anno di età in più il rischio di CHD aumenta di 1.14 volte.

```
Call:
glm(formula = chd ~ age, family = binomial(logit), data = dati)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.70846     2.35397  -2.850  0.00437 **
age           0.13150     0.04634   2.838  0.00454 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Inferenza per β :

```
Confint(GLM.1, level=0.95, type="Wald")
```

| | Estimate | 2.5 % | 97.5 % | exp(Estimate) | 2.5 % | 97.5 % |
|-------------|----------|--------|--------|---------------|----------|--------|
| (Intercept) | -6.70 | -11.32 | -2.09 | 0.001 | 0.000012 | 0.12 |
| age | 0.13 | 0.04 | 0.22 | 1.14 | 1.04 | 1.24 |

$$W = \frac{\hat{\beta}}{SE(\beta)} \approx N(0,1)$$

Intervallo di confidenza:

$$95\%CI = e^{(\beta \pm SE_{\beta})}$$

- β = incremento del *log-odds* per incremento unitario di x
- Test d'ipotesi $H_0 : \beta=0$
(test di Wald)



Esempio con R

Vogliamo vedere se il sesso è un fattore di rischio per la mortalità o il trapianto in un gruppo di malati cardiopatici.

P= probabilità di evento morte o trapianto

SESSO: 1=F; 0=M



$$\ln \left(\frac{P}{1-P} \right) = \alpha + \beta * \mathbf{SESSO}$$

$$95\%CI = e^{(\beta \pm SE_{\beta})}$$



Valutiamo tramite β il rischio di evento nei maschi rispetto alle femmine

Coefficients:

| | Estimate | Std. Error | z value | Pr(> z) | |
|-------------|-----------------|------------|---------|----------|-----|
| (Intercept) | -0.63599 | 0.07526 | -8.450 | < 2e-16 | *** |
| sex | -0.54857 | 0.14969 | -3.665 | 0.000248 | *** |



 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Exponentiated coefficients ("odds ratios")

| | sex |
|-------------|------------------|
| (Intercept) | 0.5294118 |
| | 0.5777778 |

Fatto pari a «100%» il rischio di evento (se si fosse uomini), essere donna espone ad un rischio del 58% di morte o trapianto [rispetto ad essere uomo]. Quindi le donne sono maggiormente «protette» dal rischio di evento (OR < 1).

Confint (GLM.1, level=0.95, type="Wald")

| | Est | 2.5% | 97.5% | exp(Est) | 2.5% | 97.5% |
|-------------|-------|-------|-------|-------------|-------------|-------------|
| (Intercept) | -0.63 | -0.78 | -0.48 | 0.52 | 0.45 | 0.61 |
| sex | -0.54 | -0.84 | -0.25 | 0.57 | 0.43 | 0.77 |

L'intervallo di confidenza stimato intorno a OR:
 [0.43 ; 0.77] **non contiene** 1.

Regressione logistica multipla

- Più' di una variabile indipendente
 - dicotomiche , ordinali, nominali, continue ...



$$\ln \left(\frac{P}{1-P} \right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i$$

Interpretazione di β_i :

- Incremento del log-odds per un incremento unitario di x_i con tutte le altre x_j costanti;
- misure di associazione tra x_i e log-odds *corretta* per tutte le altre x_j



Esempio con R

Vogliamo vedere se il sesso, la NYHA (*New York Advanced Heart Failure*, un punteggio di gravità da 1 a 4) e la frazione di eiezione del ventricolo sinistro (FEVSIN, variabile continua) sono fattori di rischio per la mortalità o il trapianto in un gruppo di malati cardiopatici.

P= probabilità di evento morte o trapianto

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

SESSO: 1=F; 0=M

$$= \alpha + \beta_1 * Sesso + \beta_2 * FEVSIN + \beta_3 * NYHA$$

NYHA: 1, 2, 3, 4

FEVSIN: parametro continuo [da 5% a 100%]



Coefficients:

| | Estimate | Std. Error | z value | Pr(> z) | |
|-------------|-----------|------------|---------|--------------|-----|
| (Intercept) | -0.364770 | 0.369227 | -0.988 | 0.323187 | |
| sex | -0.550281 | 0.163794 | -3.360 | 0.000781 | *** |
| NYHA | 0.499550 | 0.092551 | 5.398 | 0.0000000675 | *** |
| FEVSIN | -0.040125 | 0.007863 | -5.103 | 0.0000003347 | *** |



Exponentiated coefficients ("odds ratios")

| (Intercept) | sex | NYHA | FEVSIN |
|-------------|-----------|-----------|-----------|
| 0.6943563 | 0.5767880 | 1.6479796 | 0.9606691 |

Confint (GLM.2, level=0.95, type="Wald")

| | Est | 2.5% | 97.5% | exp (Est) | 2.5% | 97.5% |
|--------|-------|-------|-------|-----------|------|-------|
| sex | -0.55 | -0.87 | -0.22 | 0.57 | 0.41 | 0.79 |
| NYHA | 0.49 | 0.31 | 0.68 | 1.64 | 1.37 | 1.97 |
| FEVSIN | -0.04 | -0.05 | -0.02 | 0.96 | 0.94 | 0.97 |

Le donne hanno 57% il rischio degli uomini;

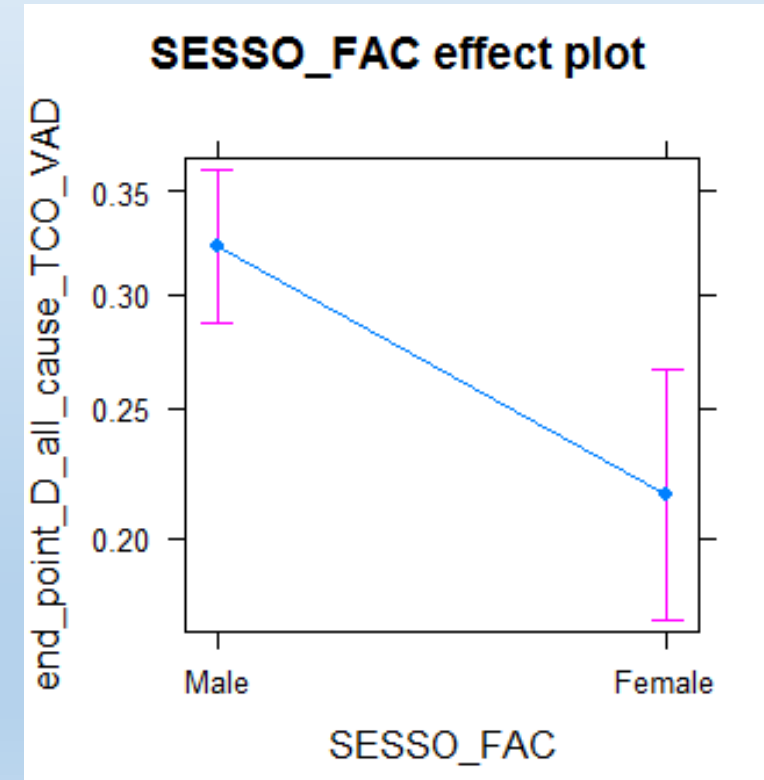
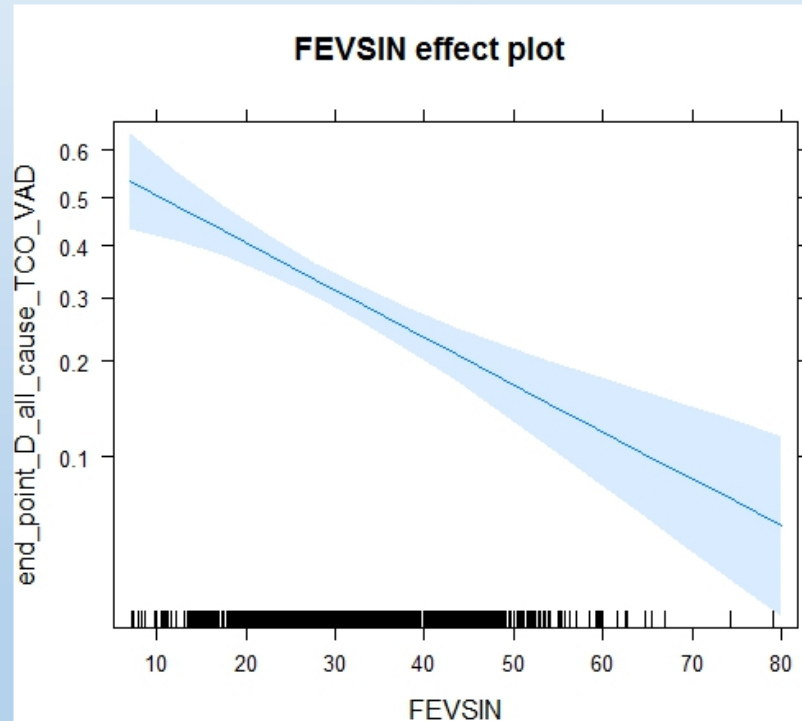
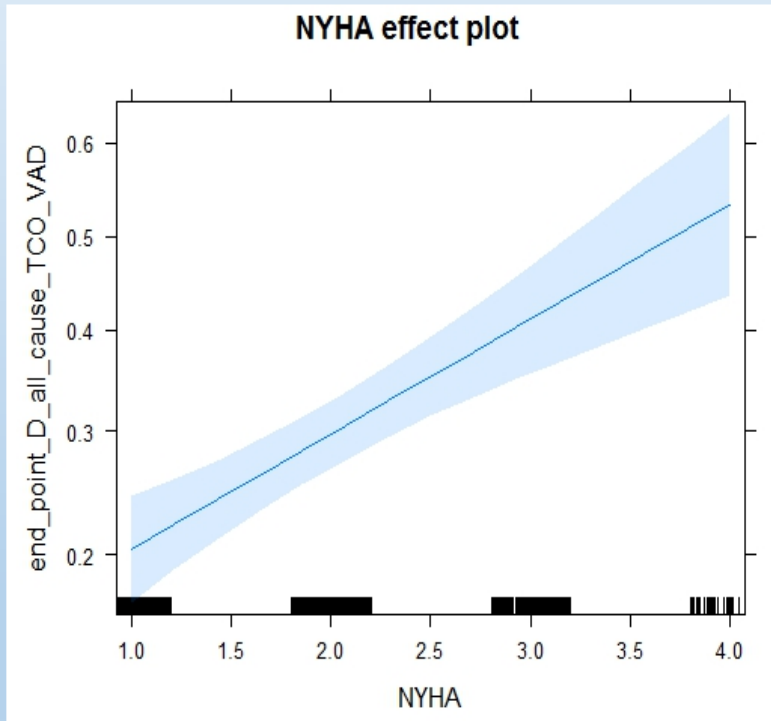
All'aumentare di una unità di FEVSIN il rischio diminuisce di 0.96;

All'aumentare di una classe NYHA il rischio aumenta di 1.64 volte

A parità di sesso e di classe NYHA, un aumento unitario di FEVS diminuisce il rischio di 0.96.

A parità di FEVS e classe NYHA, le donne hanno 57% il rischio degli uomini;

A parità di sesso e di FEVS, un aumento di classe NYHA incrementa il rischio di 1.64 volte.



Regola «**del pollice**» per la dimensione campionaria in studi con modelli di regressione multivariati:

| | |
|---------------------|-------------------------------------------------------------|
| Linear regression | # patients (samples) = 15 (10-20) x # independent variables |
| Logistic regression | Min(# events, # non-events) = 10 x # independent variables |

E' cruciale quindi avere una idea del **numero di eventi attesi** (se studio prospettico) per poter stimare con potenza adeguata i coefficienti β dei modelli di regressione multivariata !

REGRESSION MODELING STRATEGIES

Frank E Harrell Jr
 Department of Biostatistics
 Vanderbilt University School of Medicine
 Nashville TN 37232 USA
 f.harrell@vanderbilt.edu
 biostat.mc.vanderbilt.edu/rms