

Regressione lineare semplice

- Prof. Giuseppe Verlato
- Sezione di Epidemiologia e Statistica Medica, Università di Verona

Statistica con due variabili

var. nominale, var.
nominale: gruppo
sanguigno - cancro
gastrico

chi-quadrato

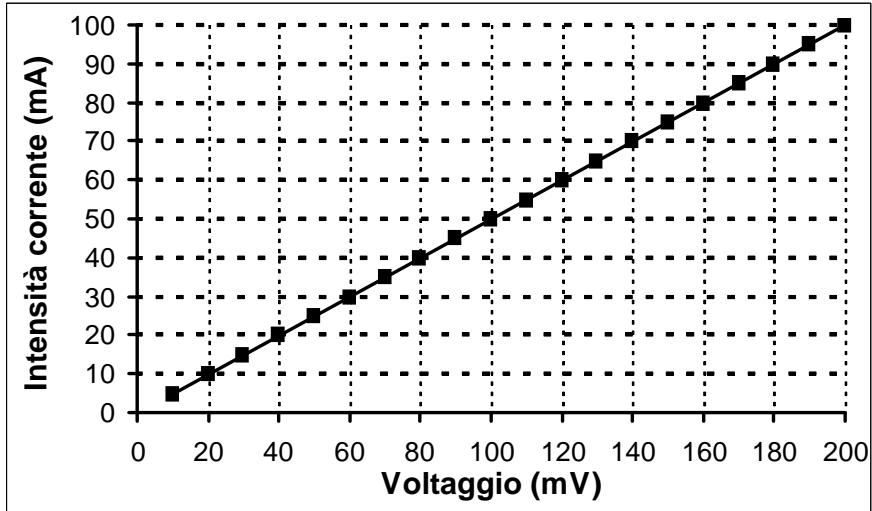
var. nominale,
var. quantitativa:
sesso - pressione
sistolica

t di Student,
ANOVA

var. quantitativa,
var. quantitativa:
peso - glicemia

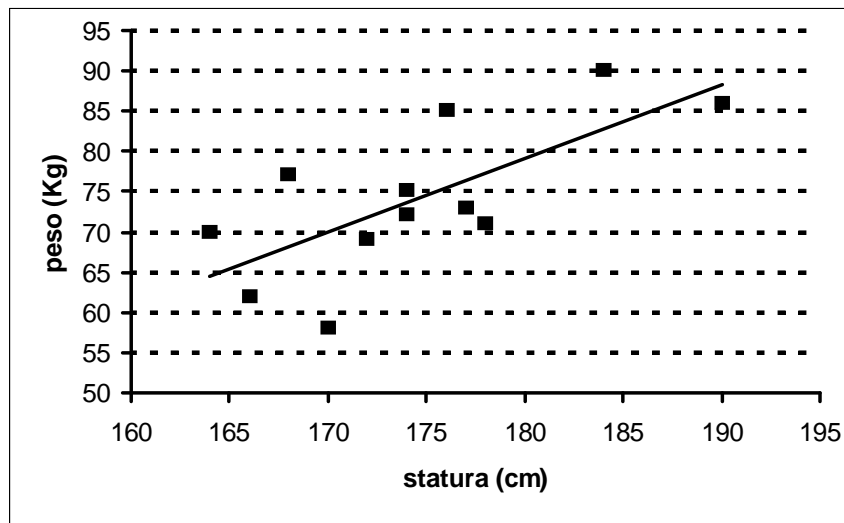
correlazione,
regressione

Correlazione perfetta tra variabile X (Differenza di potenziale V) e variabile Y (intensità della corrente: I legge di Ohm)



$$\text{Conduttanza} = \Delta I / \Delta V$$

In medicina non esistono relazioni perfette, in quanto una variabile Y è influenzata non da una sola variabile X, ma da molte altre variabili, perlopiù ignote (variabilità biologica)

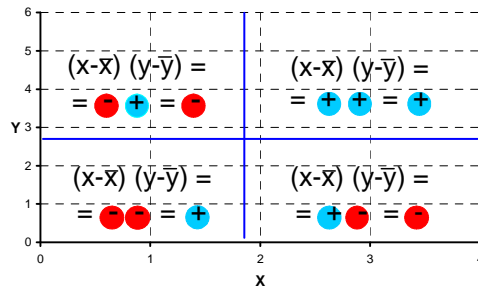


Misure di dispersione

Statistica univariata \longrightarrow devianza

Statistica bivariata \longrightarrow codevianza

	Formula euristica	Formula empirica	
Devianza	$\Sigma(x-\bar{x})^2$	$\Sigma x^2 - (\Sigma x)^2/n$	sempre ≥ 0
Codevianza	$\Sigma(x-\bar{x})(y-\bar{y})$	$\Sigma xy - (\Sigma x * \Sigma y)/n$	$< 0, = 0, > 0$



Misure di dispersione

Statistica univariata \longrightarrow varianza = devianza / (n-1)

Statistica bivariata \longrightarrow covarianza = codevianza / (n-2)

$$\hat{COV}[X_1, X_2] = \frac{\sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{n - 2}$$

2 campioni casuali di dimensione n

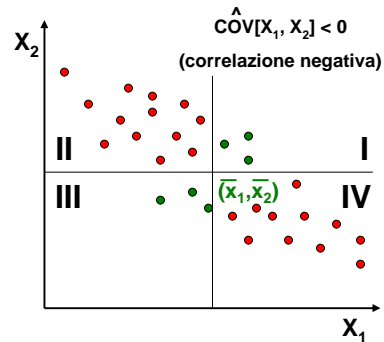
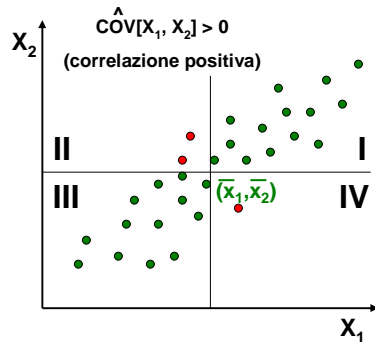
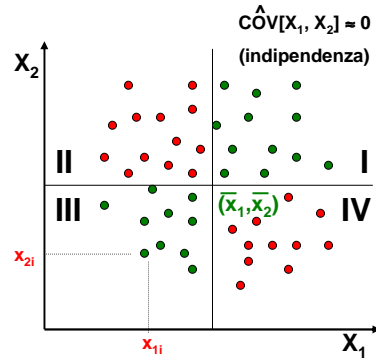
$\Rightarrow n$ valori: $(x_{1i} - \bar{x}_1) (x_{2i} - \bar{x}_2)$

quadrante I - valori positivi

quadrante II - valori negativi

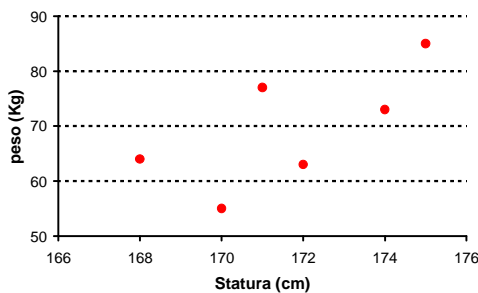
quadrante III - valori positivi

quadrante IV - valori negativi



	Statura (cm)	Peso (Kg)	xy	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
	172	63	10836	0,3	-6,5	-2,17
	174	73	12702	2,3	3,5	8,17
	171	77	13167	-0,7	7,5	-5,00
	175	85	14875	3,3	15,5	51,67
	168	64	10752	-3,7	-5,5	20,17
	170	55	9350	-1,7	-14,5	24,17
Totale	1030	417	71682			97,00

MEDIA 171,7 69,5



codevarianza =

$$\sum (x-\bar{x})(y-\bar{y}) = 97$$

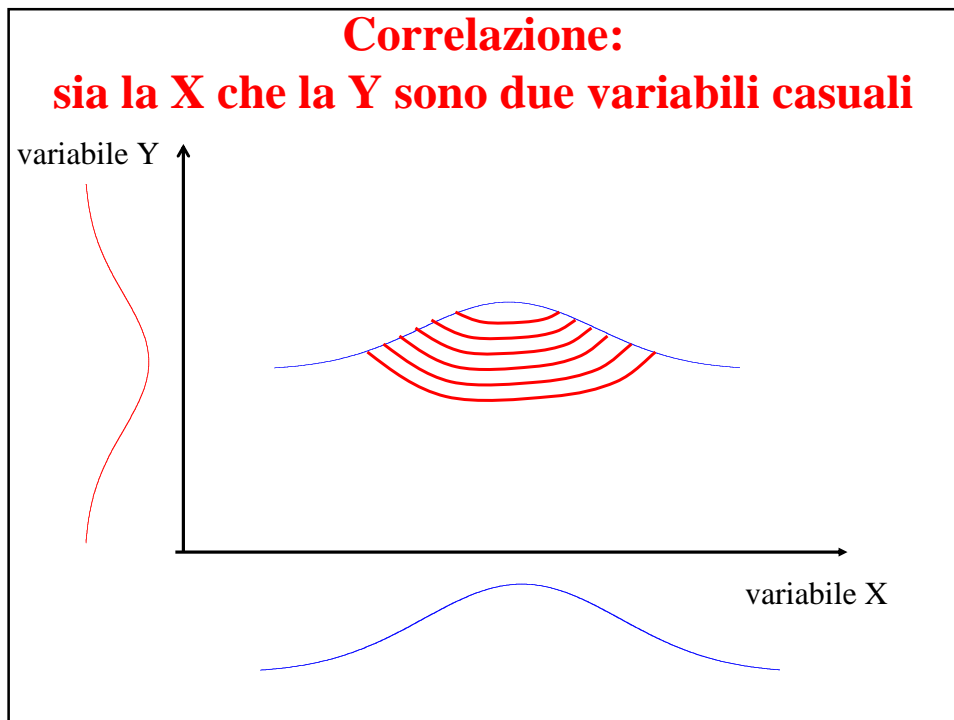
$$\sum xy - (\sum x \sum y) / n =$$

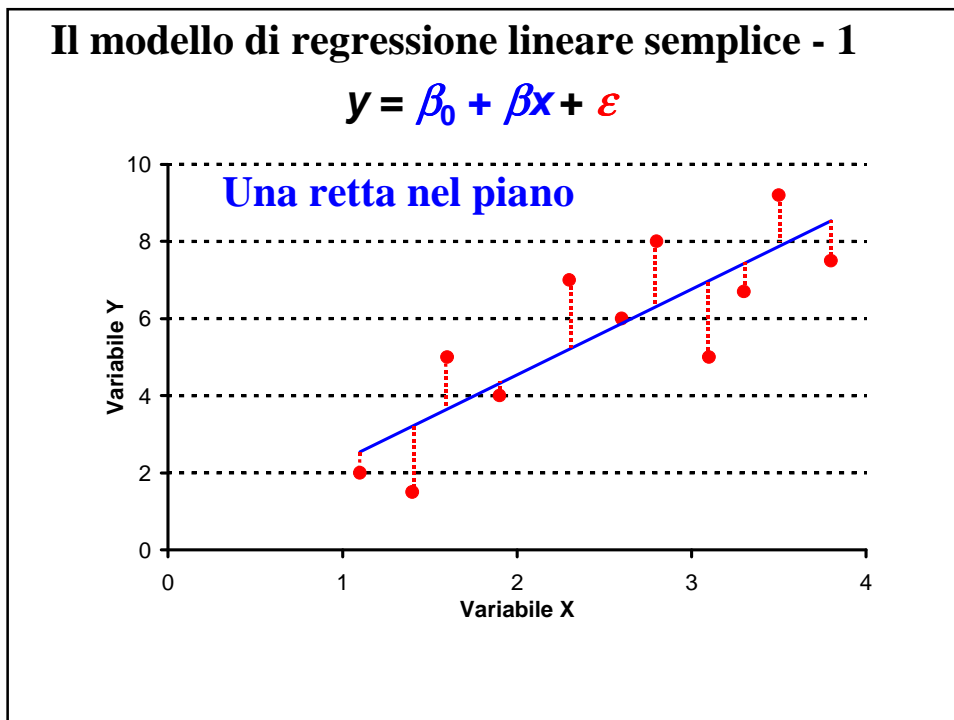
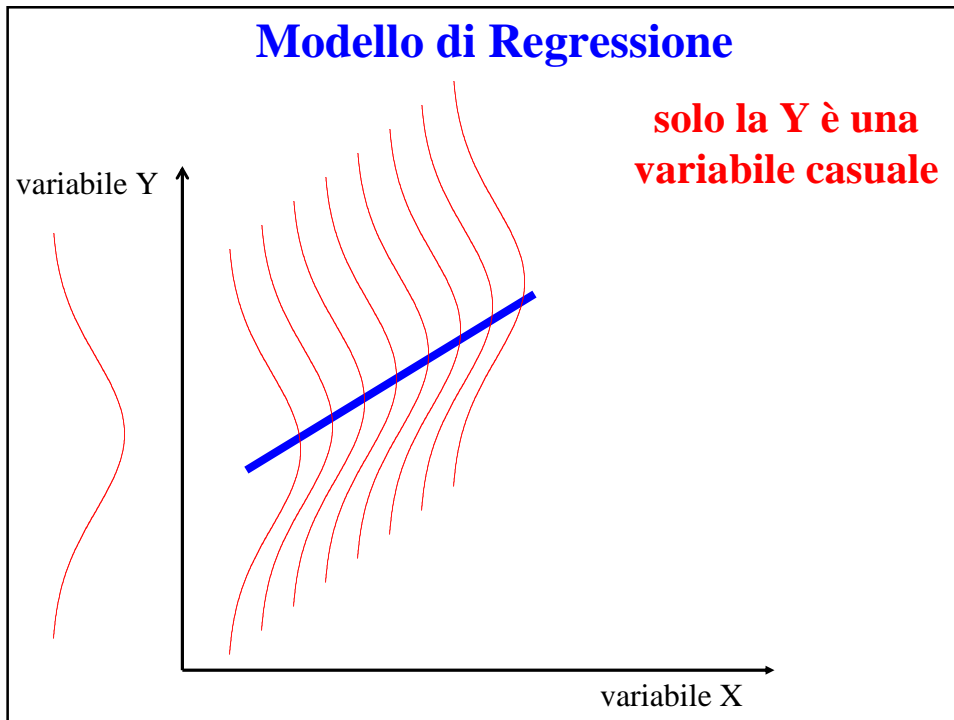
$$71682 - 1030 * 417 / 6 =$$

$$71682 - 71585 = 97$$

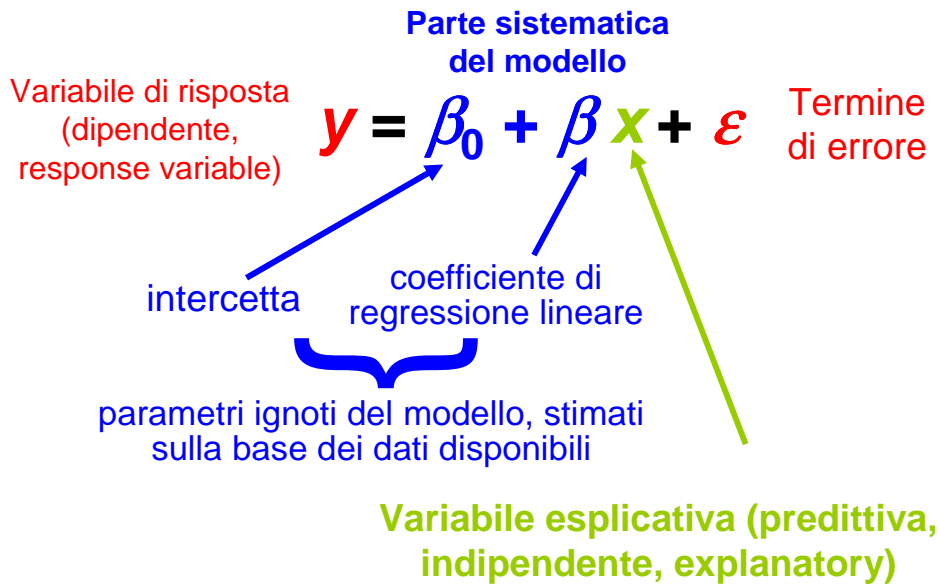
Regressione = relazione di tipo asimmetrico:
una variabile casuale (Y) dipende da una
variabile fissa (X)

Correlazione = relazione di tipo simmetrico:
le due variabili sono sullo stesso piano





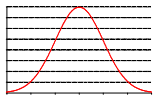
Il modello di regressione lineare semplice - 2



Il modello di regressione lineare semplice - 3

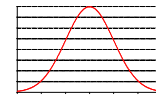
$$y = \beta_0 + \beta x + \varepsilon$$

Variabile di risposta (dipendente)



Predittore lineare, parte deterministica del modello, senza variabilità casuale

Termine di errore, parte probabilistica



L'errore, e quindi la variabile di risposta, si distribuisce NORMALMENTE

Il modello di regressione lineare semplice - 4

Il peso (Y) dipende dalla statura (X_1)

$$E(y) = \beta_0 + \beta_1 x_1$$

$E(y)$ = valore atteso (media) del peso degli individui che hanno quella determinata statura

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

y = peso di un determinato individuo, che dipende dalla statura, (parte sistematica del modello), ma anche da altre caratteristiche individuali (ε , parte probabilistica)

Il modello di regressione lineare semplice - 5

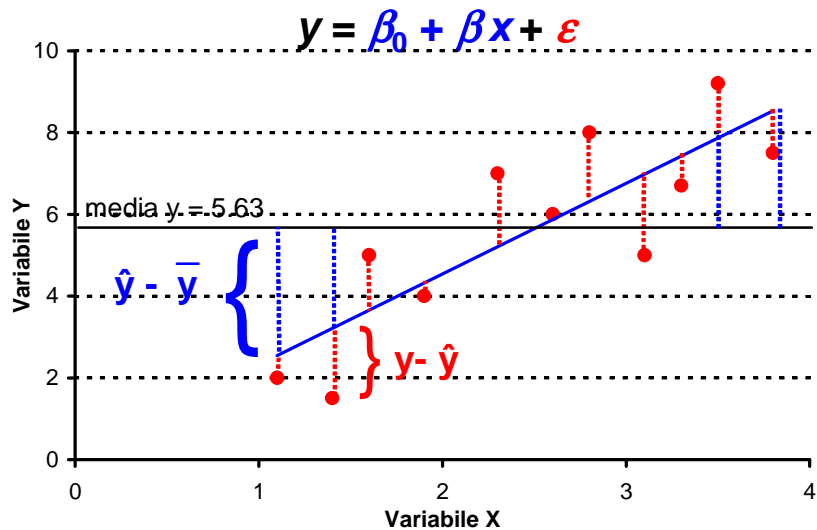
- **Modello teorico (ignoto)**

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- **Regressione Lineare stimata**

$$\hat{y} = b_0 + b_1 x$$

SCOMPOSIZIONE DELLA DEVIANZA nella Regressione lineare semplice - 1



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

SCOMPOSIZIONE DELLA DEVIANZA nella Regressione lineare semplice - 2

Variabilità
totale
 $(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$
Variabilità
spiegata dalla
regressione

Si può dimostrare che:

Devianza
totale, SST
 $\Sigma(y - \bar{y})^2 = \Sigma(\hat{y} - \bar{y})^2 + \Sigma(y - \hat{y})^2$
Devianza
spiegata dalla
regressione, SSR

Correlazione

Il coefficiente di correlazione (r) è un numero adimensionale, che varia tra -1 e +1

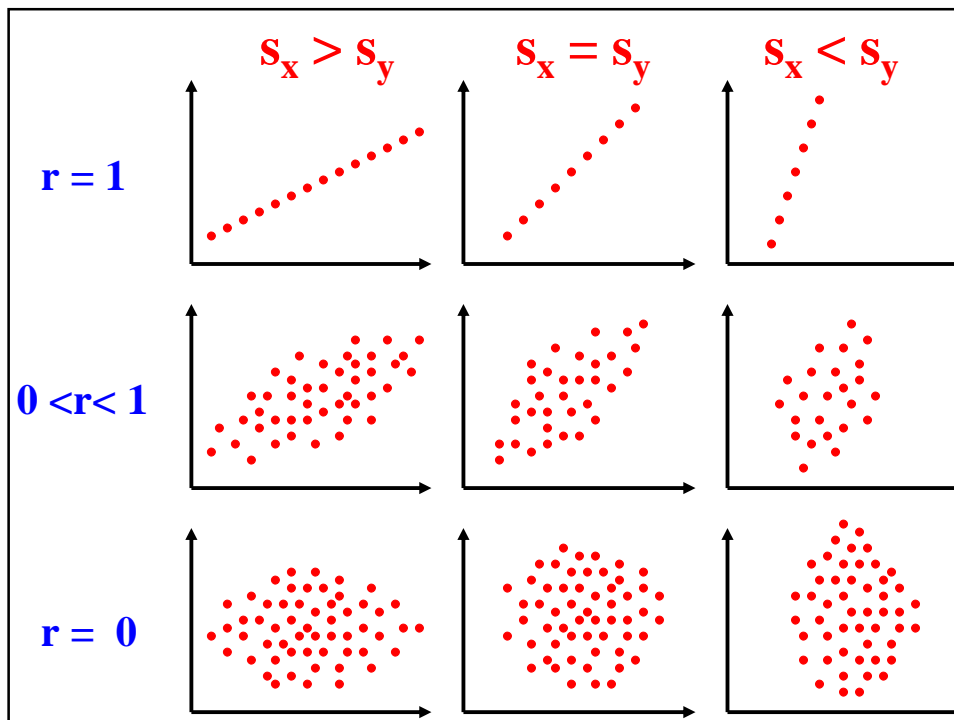
$r = -1$ i punti si allineano lungo una retta discendente

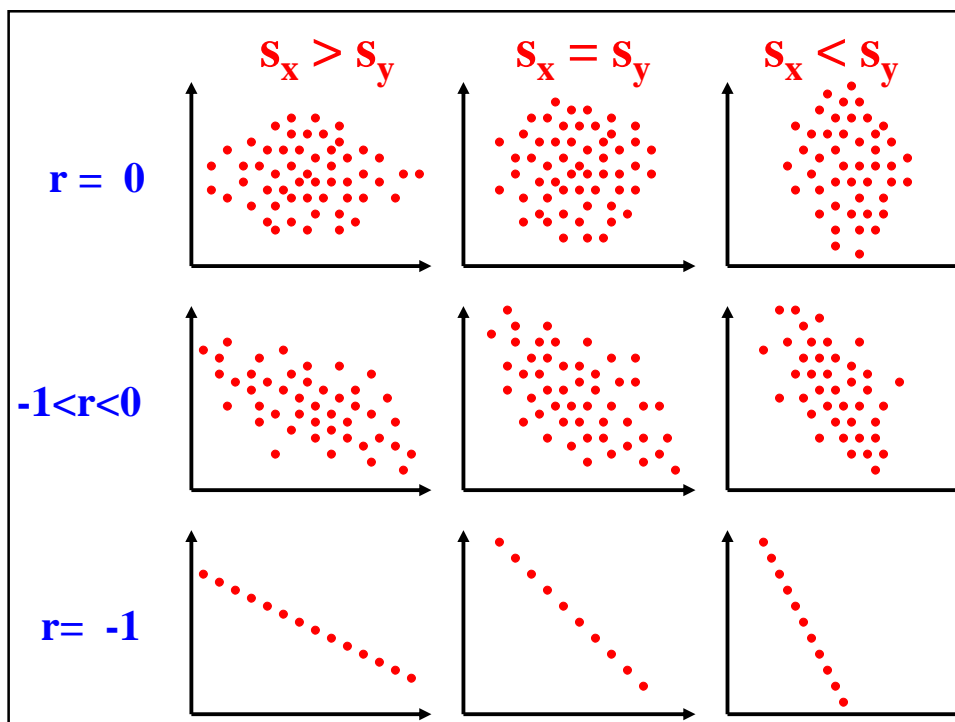
$r = 0$ i punti si dispongono a caso, senza mostrare un andamento crescente o decrescente

$r = +1$ i punti si allineano lungo una retta ascendente

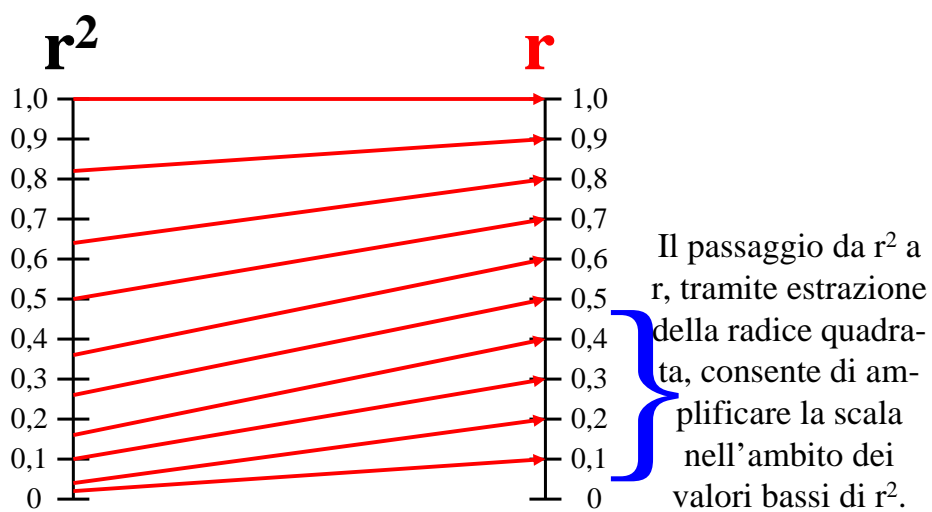
$$r = \frac{\text{codevarianza}_{xy}}{\sqrt{\text{devianza}_x * \text{devianza}_y}}$$

$$r^2 = \frac{\text{Devianza spiegata dalla regressione, SSR}}{\text{Devianza totale, SST}} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$





Nella realtà biologica la maggior parte delle correlazioni tra variabili è piuttosto debole: l' r^2 oscilla tra 0 e 0,5.



Regressione lineare semplice

Si cerca di trovare la retta che meglio interpola, che meglio si adatta alla nuvola di punti.

METODO DEI MINIMI QUADRATI

Si sceglie la retta che **riduce al minimo la devianza residua, SSE, $\Sigma(y - \hat{y})^2$**

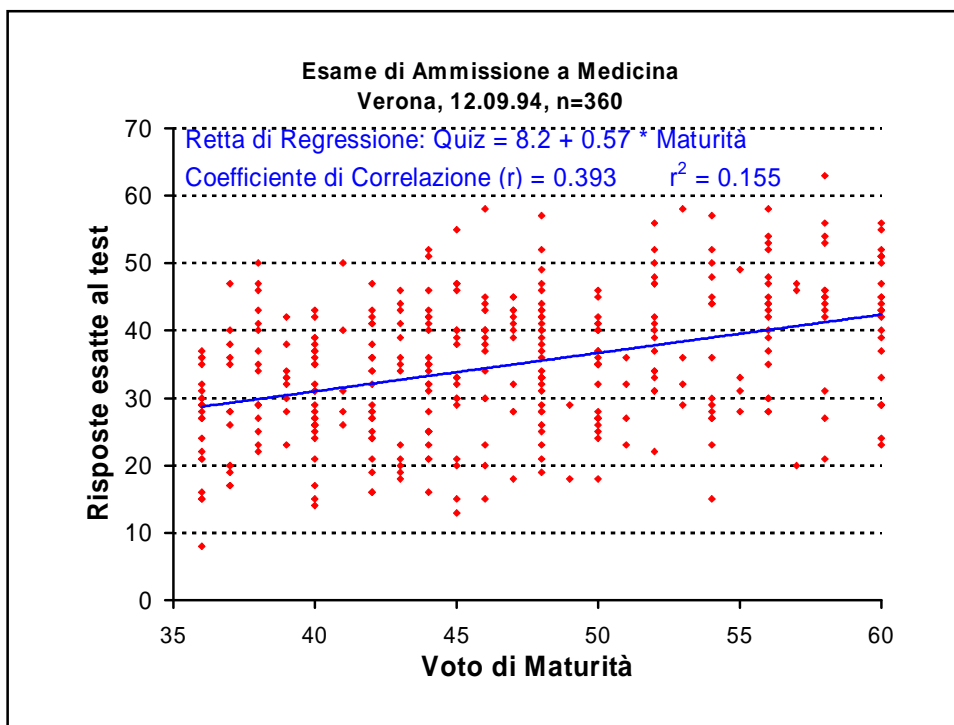
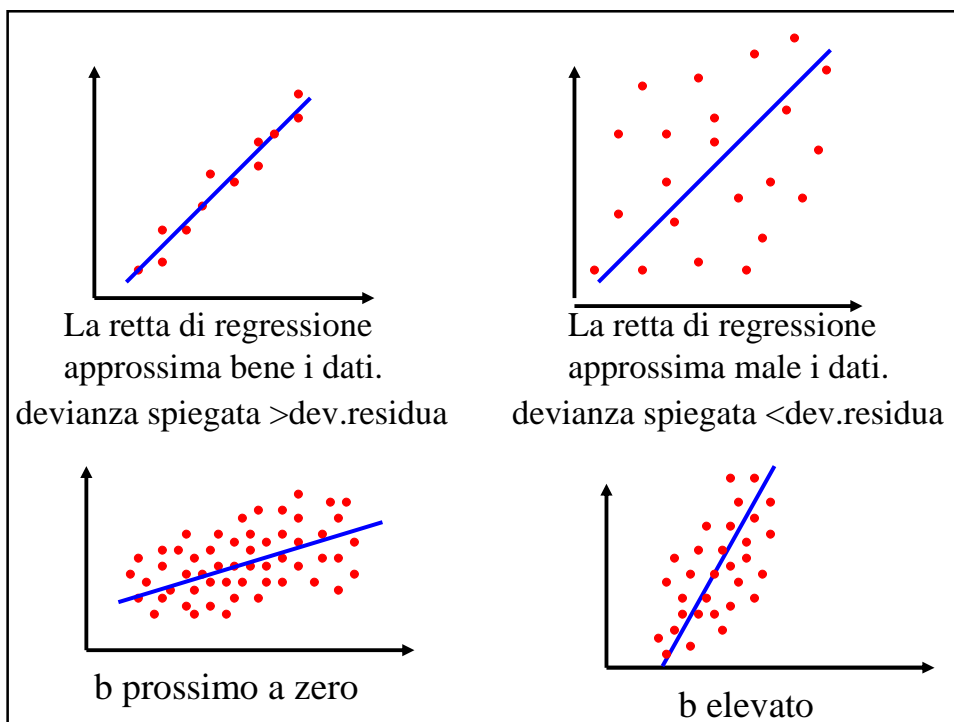
Regressione lineare semplice

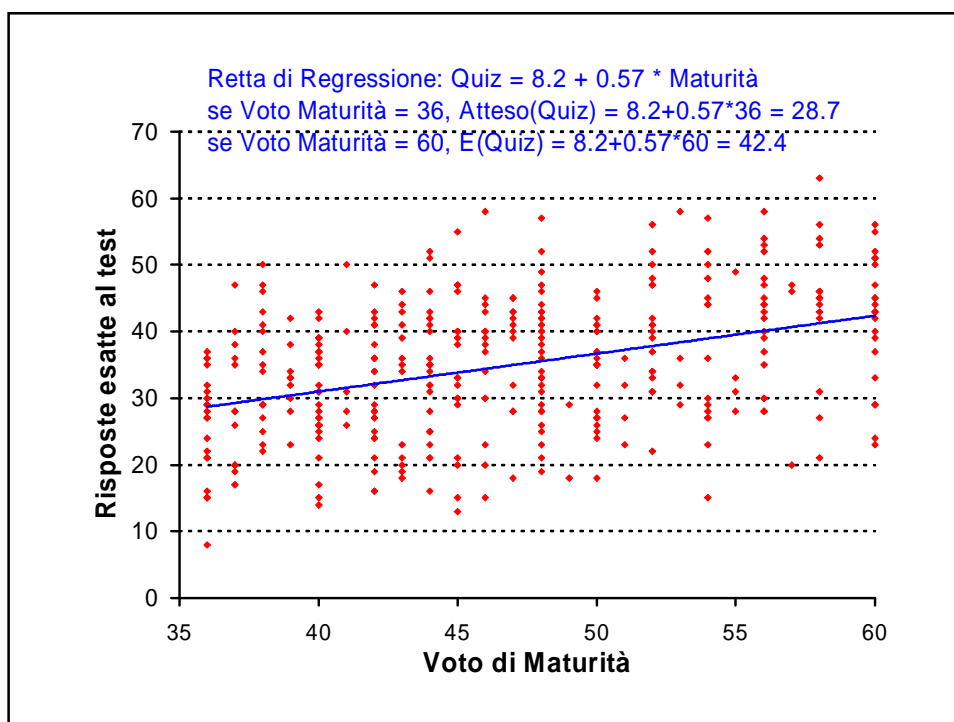
b_1 = coefficiente di regressione lineare, pendenza, slope $b_1 = \frac{\text{codevianza}_{xy}}{\text{devianza}_x}$

b_1 varia tra $-\infty$ e $+\infty$, ha come unità di misura il rapporto tra l'unità di misura della variabile Y e l'unità di misura della variabile X

il valore assoluto di b_1 dipende dalla unità di misura utilizzate

b_0 = intercetta $b_0 = \bar{y} - b_1 \bar{x}$





Correlazione

Coefficiente di Correlazione (r) = 0.393

$$r^2 = 0.155$$

$$r^2 = \frac{\text{Devianza spiegata dalla regressione, SSR}}{\text{Devianza totale, SST}} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$

Il 15,5% della variabilità (devianza) del punteggio al test è spiegato dalla variabilità del voto di maturità.

ESEMPIO:

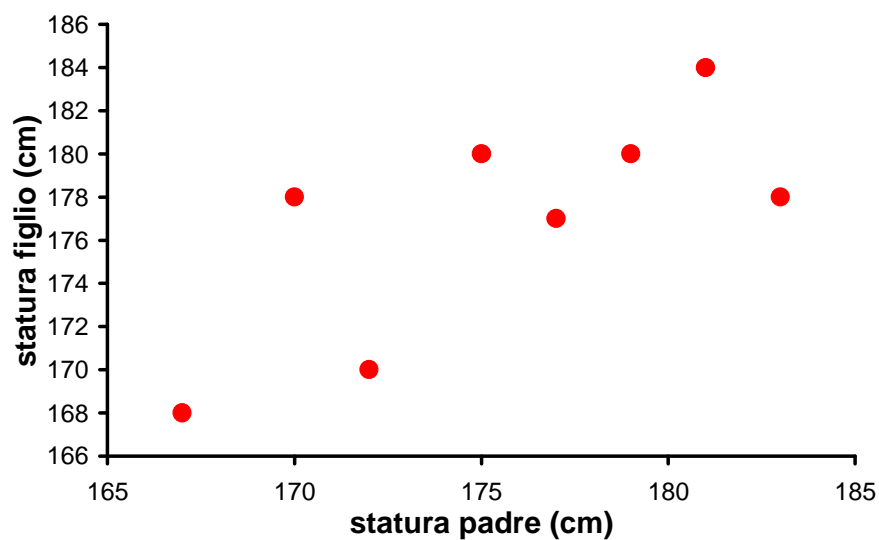
Esiste una relazione tra altezza dei padri e altezza dei figli maschi?

(A padri più bassi corrispondono figli più bassi?)

(A padri più alti corrispondono figli più alti?)

Padre	Figlio
167 cm	168 cm
175 cm	180 cm
183 cm	178 cm
170 cm	178 cm
181 cm	184 cm
172 cm	170 cm
177 cm	177 cm
179 cm	180 cm

I passo: rappresentazione grafica mediante diagramma di dispersione (scatterplot)



II passo: si ipotizza un modello statistico, che possa essere utile ad interpretare i dati

Ipotizziamo un modello lineare del tipo: $y = \beta_0 + \beta x + \varepsilon$

$$(\text{altezza figli}) = \beta_0 + \beta (\text{altezza padri}) + \varepsilon$$

(i figli di uno stesso padre hanno statura abbastanza simile, ma non necessariamente uguale, anche se ομοιομετρός, cioè figli della stessa madre)

III passo: Statistica descrittiva uni- e bi-variata

	Σx	Σx^2	n	Σxy	media
Statura padri	1404	246618	8	248483	175,5
Statura figli	1415	250477	8		176,875
		devianza	varianza	dev.standard	
Statura padri	216	30,86	5,55		
Statura figli	198,875	28,41	5,33		

$$\text{codevarianza} = \Sigma xy - \Sigma x \Sigma y / n = 248483 - 1404 * 1415 / 8 = 150,5$$

IV passo: Stima dei parametri del modello con il metodo dei minimi quadrati

$$b_1 = \frac{\text{codevianza}_{xy}}{\text{devianza}_x} = 150,5 / 216 = 0,697 \text{ cm/cm}$$

$$b_0 = \bar{y} - b_1 \bar{x} = 176,875 - 0,697 * 175,5 = 54,59 \text{ cm}$$

Retta di regressione:

$$\text{altezza figlio (cm)} = 54,6 \text{ cm} + 0,697 \text{ cm/cm} * \text{altezza padre (cm)}$$

Quando la statura del padre cresce di 1 cm, la statura del figlio cresce in media di 7 mm.

V passo: Calcolo del coefficiente di correlazione

$$r = \frac{\text{codevianza}_{xy}}{\sqrt{\text{devianza}_x * \text{devianza}_y}} = \frac{150,5}{\sqrt{216 * 198,9}} = 0,726$$

$$r^2 = 0,7261^2 = 0,527$$

Il 52,7% della variabilità nella statura dei figli è spiegata dalla variabilità nell'altezza dei padri

**VI passo: Inferenza sui parametri:
i dati “supportano” il modello proposto?**

Test t di Student, basato su b_1 (stima di β_1)

$$\begin{cases} H_0: \beta_1 = 0 & \text{Livello di significatività} = 5\% \\ H_1: \beta_1 \neq 0 & \text{Gradi di libertà} = n - 2 = 8 - 2 = 6 \end{cases}$$

test a due code Soglia critica = $t_{6, 0,025} = 2,447$

$$t = \frac{b-0}{ES_b} = \frac{b}{\sqrt{\text{var}_{\text{res}}/\text{dev}_x}} = \frac{0,697}{\sqrt{15,67 / 216}} = 2,588$$

$$\text{dev}_{\text{res}} = \text{dev}_y - \text{codev}_{xy}^2/\text{dev}_x = 198,9 - 150,5^2/216 = 94,01$$

$$\text{var}_{\text{res}} = \text{dev}_{\text{res}} / (n-2) = 94,01 / 6 = 15,67$$

**VI passo: Inferenza sui parametri:
i dati “supportano” il modello proposto?**

Test t di Student, basato su r (stima di ρ)

$$\begin{cases} H_0: \rho = 0 & \text{Livello di significatività} = 5\% \\ H_1: \rho \neq 0 & \text{Gradi di libertà} = n - 2 = 8 - 2 = 6 \end{cases}$$

test a due code Soglia critica = $t_{6, 0,025} = 2,447$

$$t = \frac{r-0}{ES_r} = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = \frac{0,726}{\sqrt{(1-0,527) / 6}} = 2,587$$

$$| t \text{ osservato} | > t \text{ tabulato}$$

$$2,588 > 2,447$$



Rifiuto H_0

La relazione tra statura dei figli e statura dei padri non è dovuta al caso, ma è un fatto reale ($P=0,041$).

VII passo: Previsione sui valori della variabile Y

Per $x = 185$ cm, qual è il valore atteso di Y?

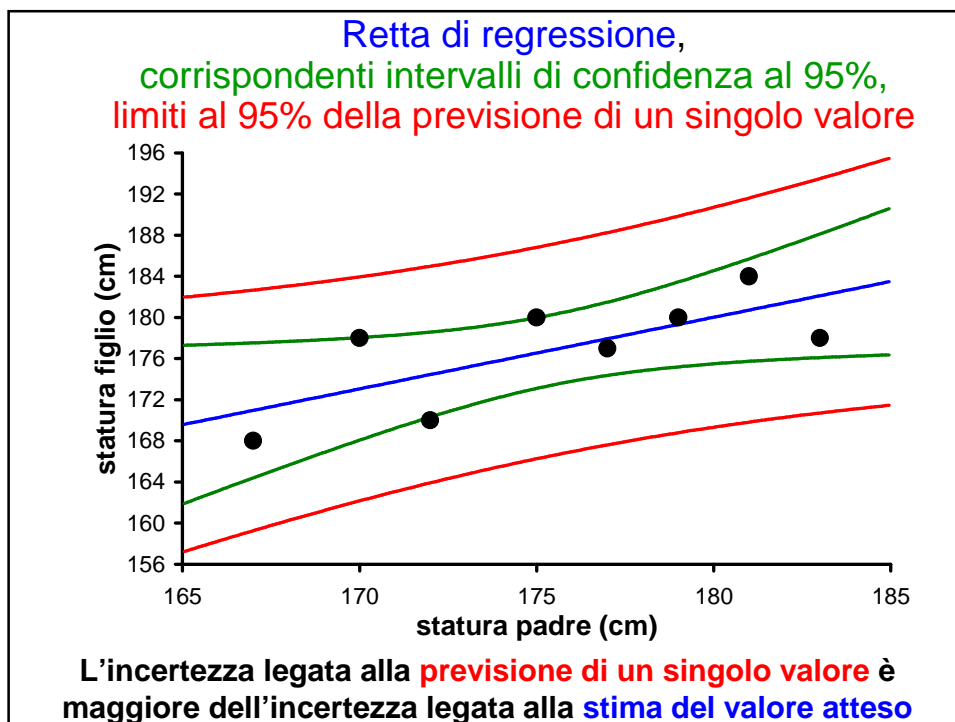
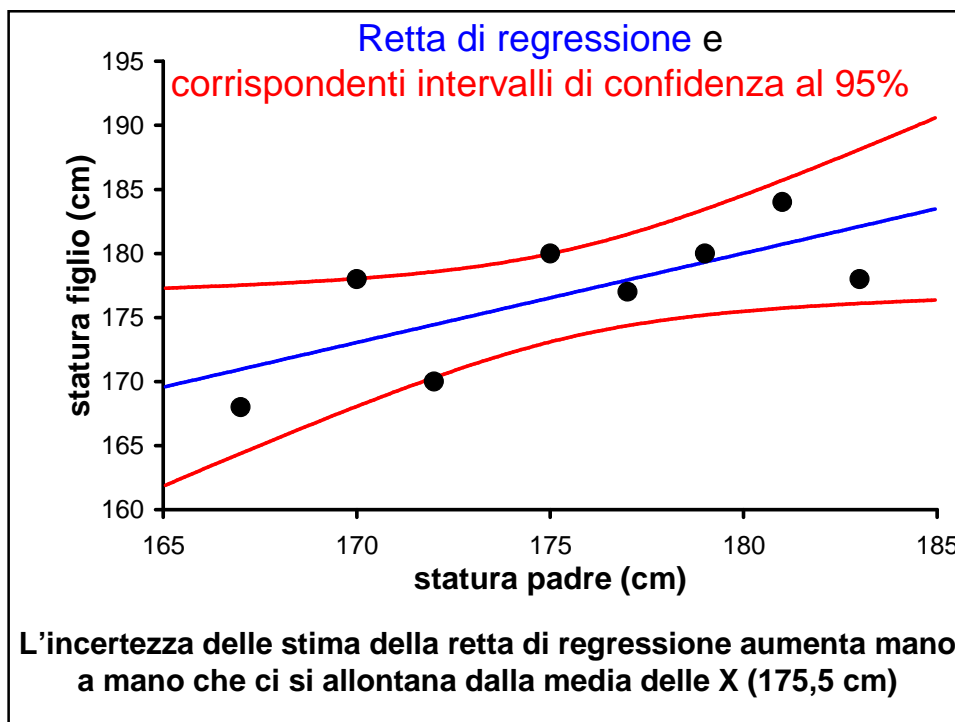
Retta di regressione:

$$\hat{y} = 54,6 \text{ cm} + 0,697 \text{ cm/cm} * 185 \text{ cm} = 183,49 \text{ cm}$$

$$ES_{\hat{y}} = \sqrt{\text{var}_{\text{res}} [1/n + (x-\bar{x})^2 / \text{dev}_x]} =$$

$$= \sqrt{15,67 [1/8 + (185-175,5)^2 / 216]} = 2,916$$

$$IC_{95\%} = \hat{y} \pm t_{v,\alpha/2} * ES_{\hat{y}} = 183,49 \pm 2,447 * 2,916 = \begin{bmatrix} 190,63 \\ 176,36 \end{bmatrix}$$

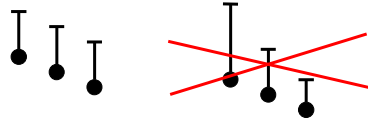


Regressione lineare semplice

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

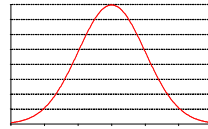
ASSUNZIONI

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



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

- 4) **Distribuzione NORMALE** degli errori



Regressione lineare semplice ASSUNZIONI

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

