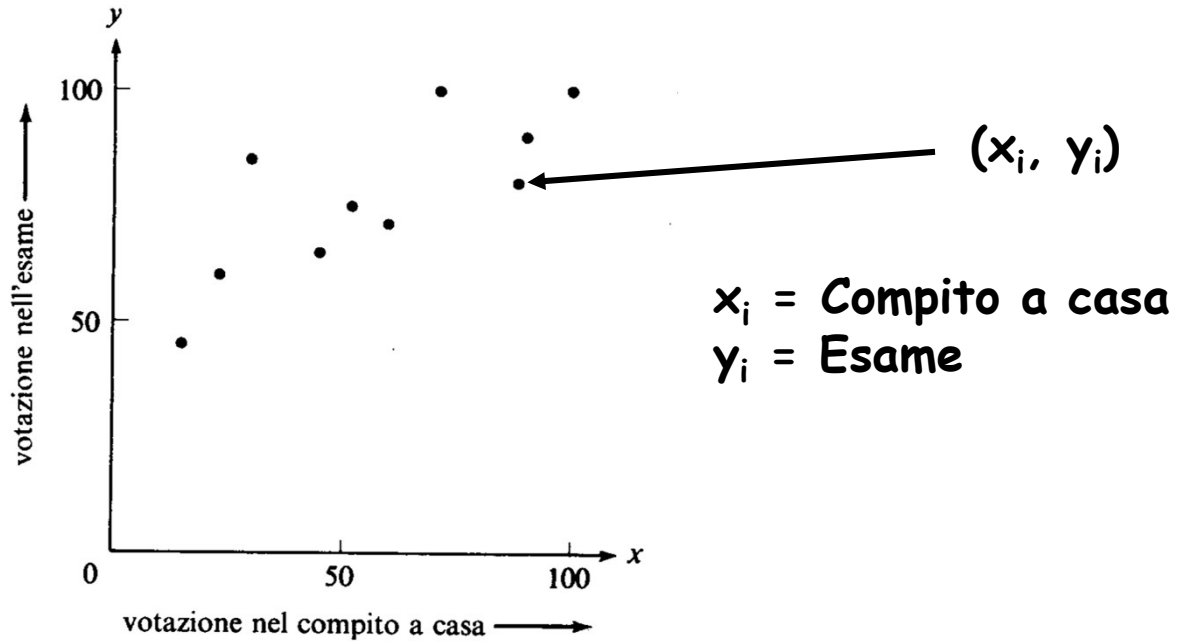
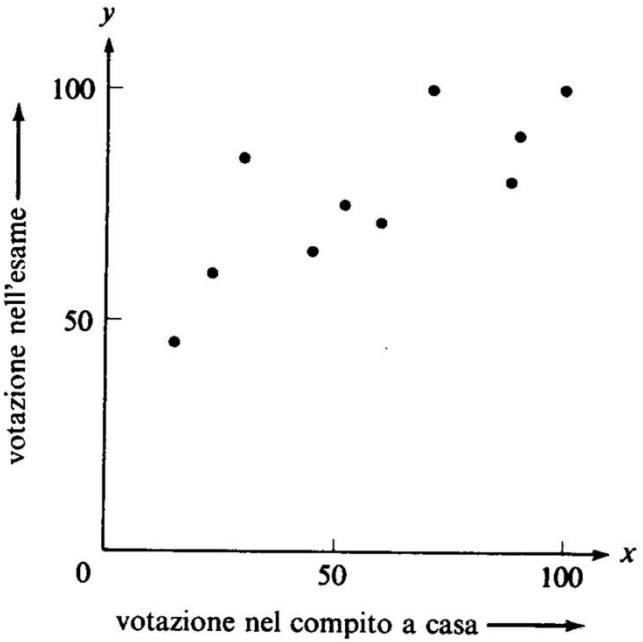
Associazione tra due variabili quantitative

Esempio (1)

Supponiamo che un professore voglia dimostrare che esercitarsi a casa aiuti gli studenti nel superamento dell'esame. A tal fine registra la votazione dei compiti a casa e degli esami e li rappresenta graficamente su un diagramma cartesiano (votazioni espresse in 100-esimi).



Esempio (2)



I risultati confermano l'opinione del docente secondo cui chi fa bene gli esercizi è candidato a superare a pieni voti l'esame?

- Gli studenti con alte votazioni nel compito a casa tendono ad avere alte votazioni all'esame.
- L'andamento fra le due variabili sembra essere lineare.

Associazione

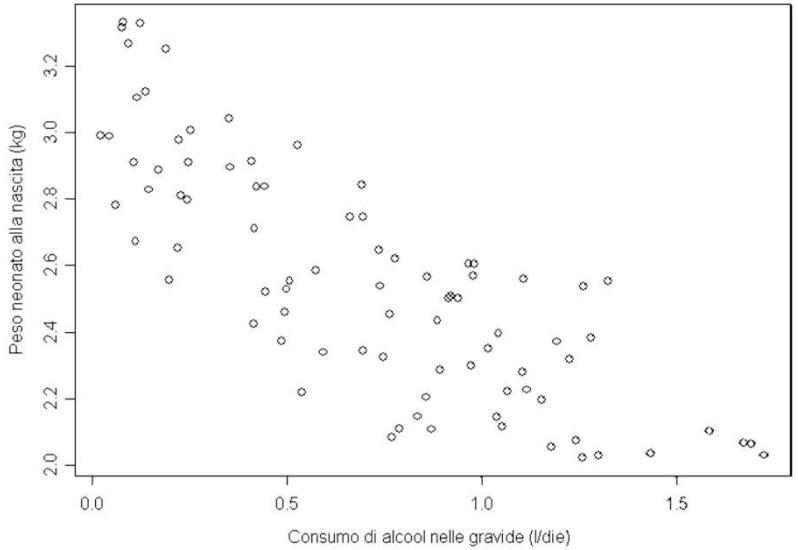
Si parla di **associazione** quando si studia la relazione esistente tra due variabili casuali.

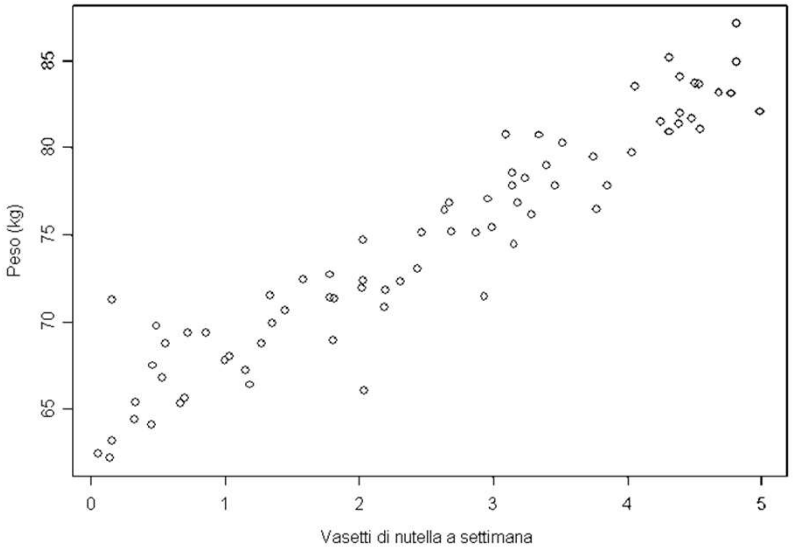
L'obiettivo è quello di valutare come si comporta una variabile al variare dell'altra.

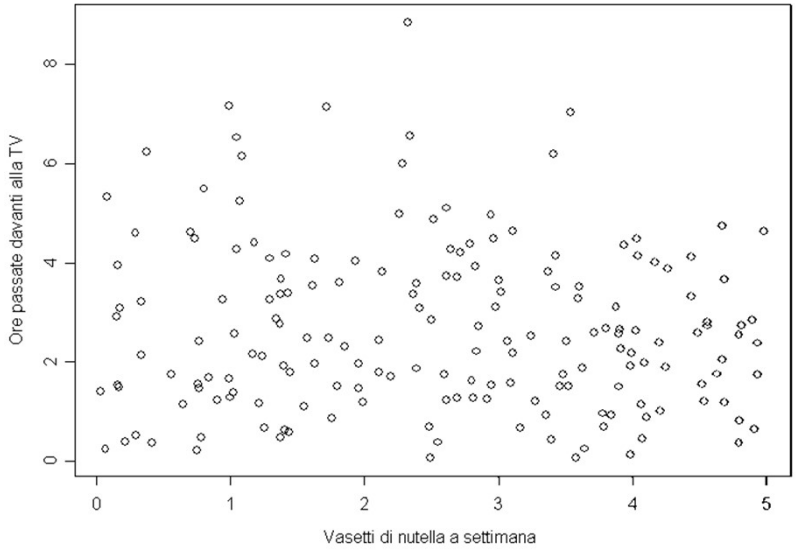
Ad esempio:

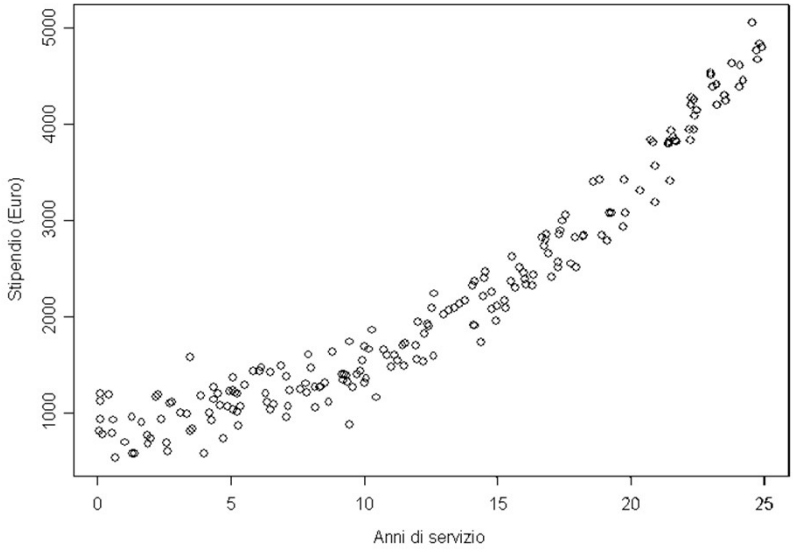
1) "Peso del neonato" vs "Consumo alcolici della madre"
2) "Consumo di nutella" vs "Peso"
3) .....

Associazione - esempi

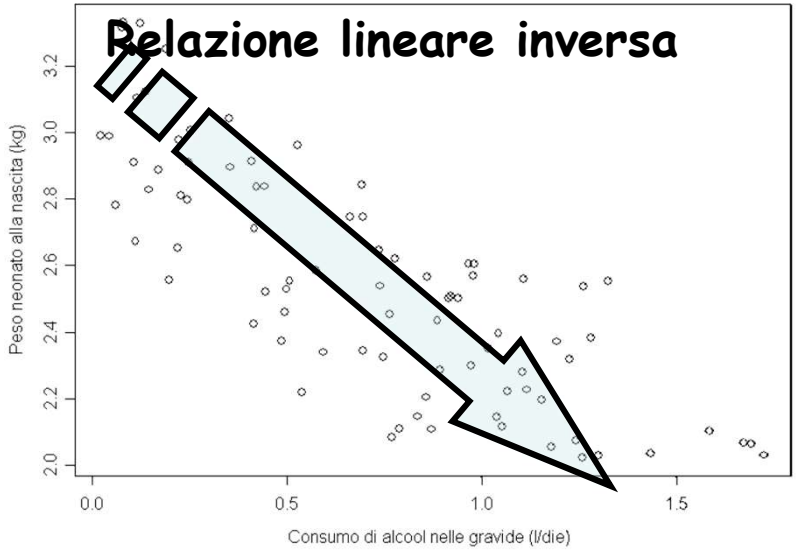


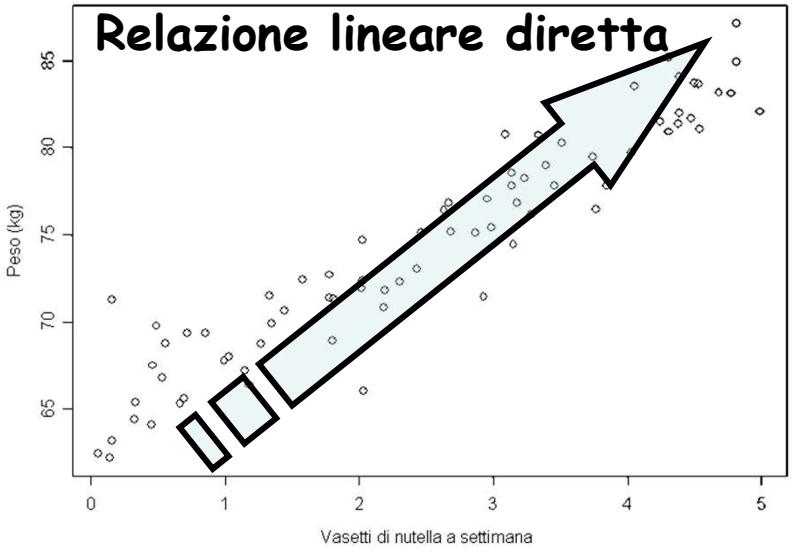


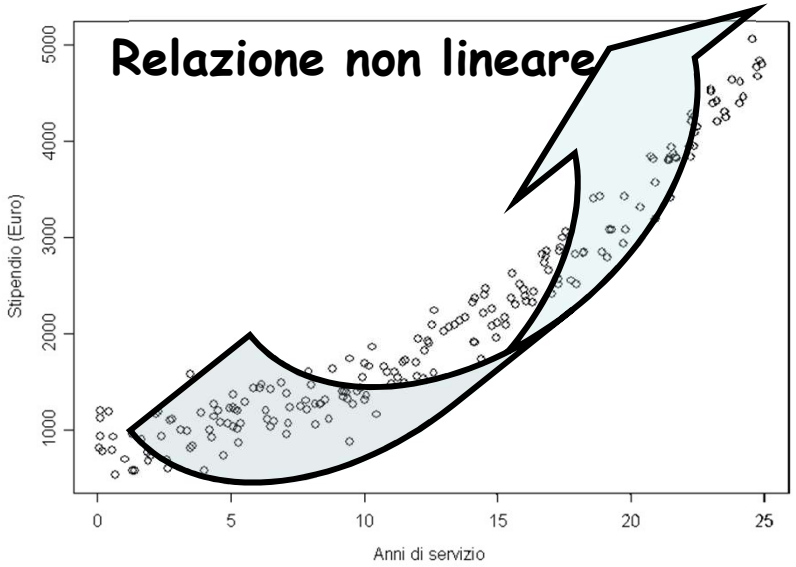




Associazione - esempi







Tipi di Associazione

Tra due variabili quantitative può esistere una relazione:

- ✓ lineare
  - proporzionalità inversa
  - proporzionalità diretta
- ✓ non lineare
- ✓ di indipendenza

Distribuzione campionaria congiunta di due variabili

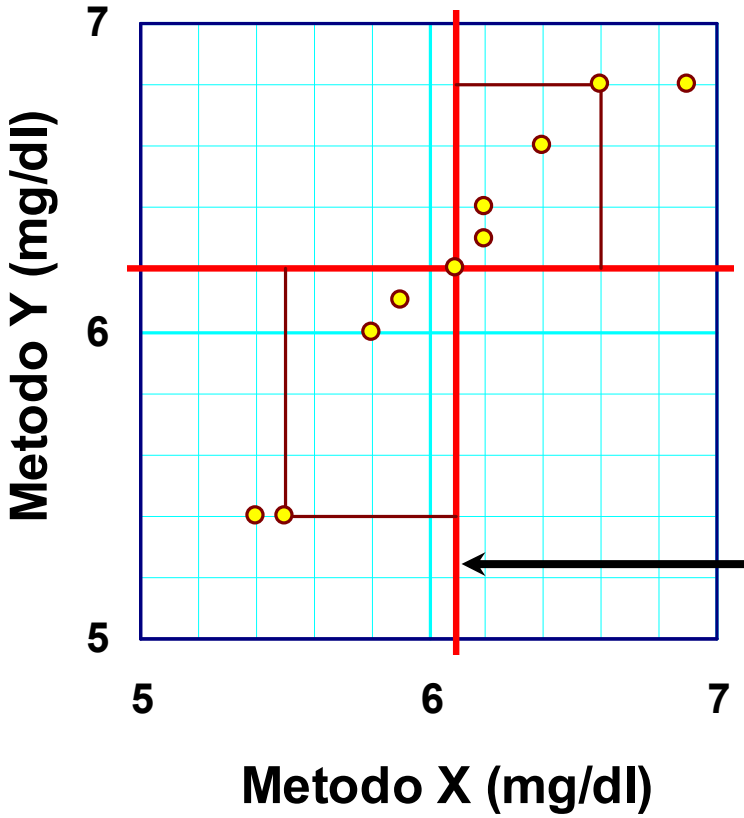
Si consideri un insieme di coppie (x_i, y_i) di valori di uricemia, misurati con due metodi (X ed Y) in un gruppo di 10 uomini anziani.

Ciascun prelievo di sangue (uno per soggetto) è stato ripartito in due aliquote, una analizzata con il metodo X e l'altra con il metodo Y.

	Soggetti									
Metodo	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X	5.8	6.2	6.9	6.1	5.4	6.2	5.9	5.5	6.6	6.4
Y	6.0	6.3	6.8	6.2	5.4	6.4	6.1	5.4	6.8	6.6

Grafico di dispersione

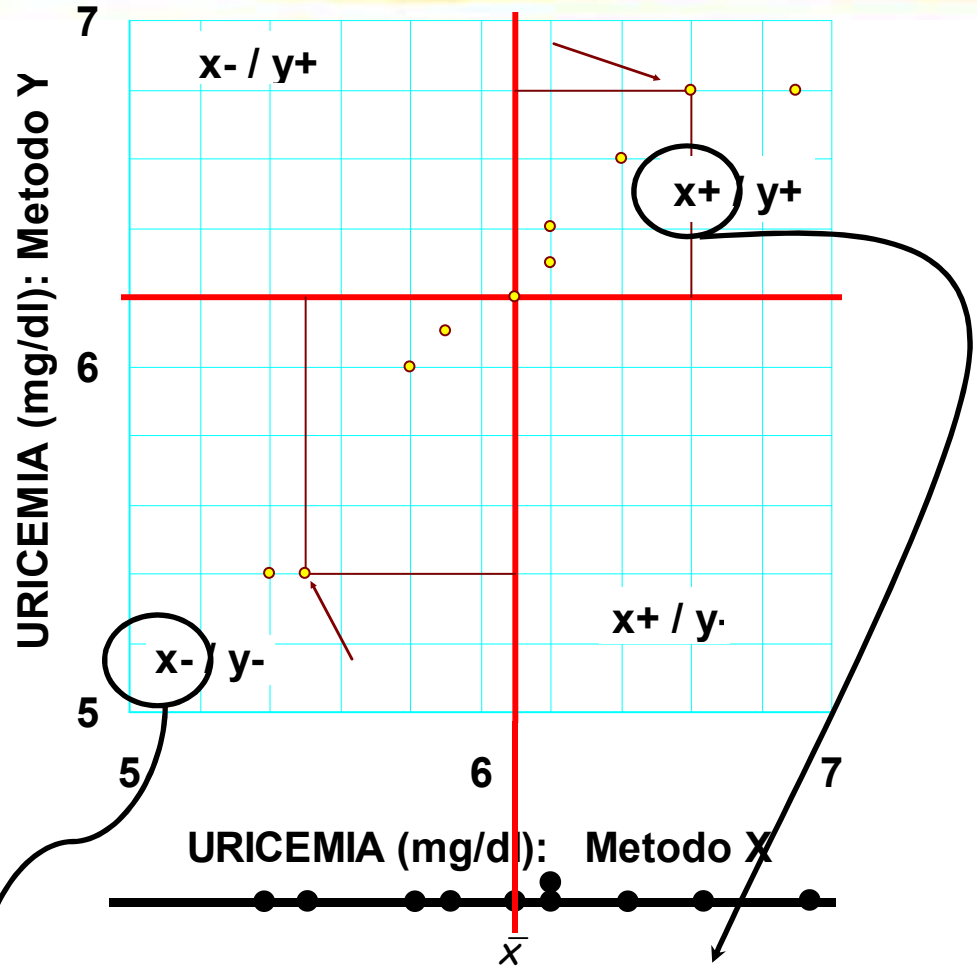
I dati possono essere rappresentati in un diagramma cartesiano: ogni punto è rappresenta una coppia (x_i, y_i).



Media delle misurazioni ottenute con il metodo Y ($\bar{y}$)

Media delle misurazioni ottenute con il metodo X ($\bar{x}$)

… continua



$X^- = (x - \bar{x}) < 0$
(scarti dalla media negativi)

$X^+ = (x - \bar{x}) > 0$
(scarti dalla media positivi)

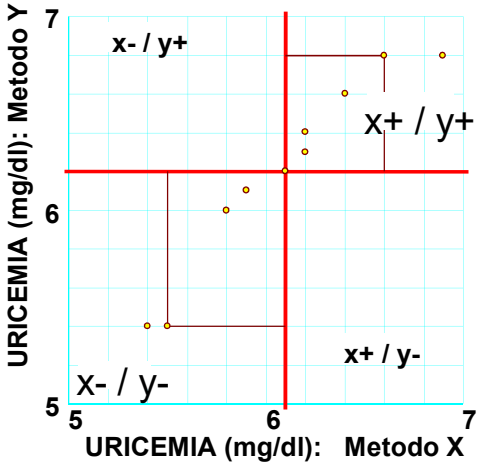
Osservazione

In presenza di una *relazione lineare diretta*, ci si attende che se una misura x_i è maggiore della media, anche la corrispondente misura y_i sia maggiore della media.

e che:

$x^+ \Rightarrow y^+$

$x^- \Rightarrow y^-$



I punti (x_i,y_i) sono addensati nel primo e nel terzo quadrante.

Le due variabili X e Y possono essere descritte da una legge del tipo: $Y = a+bX$.

Indici di Covariazione - Codevianza

La codevianza è la somma dei prodotti degli scarti:

$$C_{XY} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- Se $C_{XY}>0$ le coppie di scarti concordi (+/+ o -/-) prevalgono su quelle di scarti discordi (+/- o -/+);
- Se $C_{XY}<0$ le coppie di scarti discordi (+/- o -/+) prevalgono su quelle di scarti concordi (+/+ o -/-) ;
- Se $C_{XY}=0$ le coppie concordi e discordi si equivalgono.

Indici di Covariazione - Covarianza

La covarianza è la media dei prodotti degli scarti:

$$s_{XY} = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{(n-1)} C_{XY}$$

- ✓ La covarianza, in analogia con quanto visto per la varianza campionaria, è definita come il rapporto tra codevianza e n-1.
- ✓ La covarianza segue lo stesso andamento della codevianza.

Indice di correlazione lineare (1)

Il coefficiente di correlazione lineare è pari al rapporto tra la covarianza e il prodotto delle deviazioni standard (s_x e s_y) delle variabili x e y:

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{C_{XY}}{\sqrt{D_X D_Y}}$$

E' una misura:

1) <u>della direzione</u>
2) <u>della forza</u>

del **legame lineare** fra due variabili X e Y.

Indice di correlazione lineare (2)

Il coefficiente di correlazione lineare r:

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{C_{XY}}{\sqrt{D_X D_Y}}$$

1) è un numero puro (adimensionale)

2) è simmetrico (rimane invariato se si scambiano le due variabili)

3) è invariante rispetto a trasformazioni lineari (come il cambio di origine o il cambio di scala)

Indice di correlazione lineare (3)

4) Il coefficiente di correlazione lineare può assumere valori compresi tra -1 e +1.

$$-1 \leq r \leq 1$$

(il segno di r dipende solo dal numeratore)

r=-1	legame perfettamente lineare (relazione inversa)
-1<r<0	legame tendenzialmente lineare (relazione inversa)
r=0	assenza di correlazione lineare
0<r<1	legame tendenzialmente lineare (relazione diretta)
r=1	legame perfettamente lineare (relazione diretta)

Calcoli dell'esempio – uricemia in 10 uomini

	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	5.8	6.0	-0.3	-0.2	0.09	0.04	+0.06
2	6.2	6.3	+0.1	+0.1	0.01	0.01	+0.01
3	6.9	6.8	+0.8	+0.6	0.64	0.36	+0.48
4	6.1	6.2	0	0.0	0.0	0.0	0.0
5	5.4	5.4	-0.7	-0.8	0.49	0.64	+0.56
6	6.2	6.4	+0.1	+0.2	0.01	0.04	+0.02
7	5.9	6.1	-0.2	-0.1	0.04	0.01	+0.02
8	5.5	5.4	-0.6	-0.8	0.36	0.64	+0.48
9	6.6	6.8	+0.5	+0.6	0.25	0.36	+0.30
10	6.4	6.6	+0.3	+0.4	0.09	0.16	+0.12
Tot					1.98	2.26	2.05

… quindi

$\bar{x} = 6.1 \quad D_X = 1.98 \; s_X = \sqrt{\frac{1.98}{10-1}} = 0.469$ mg/dl

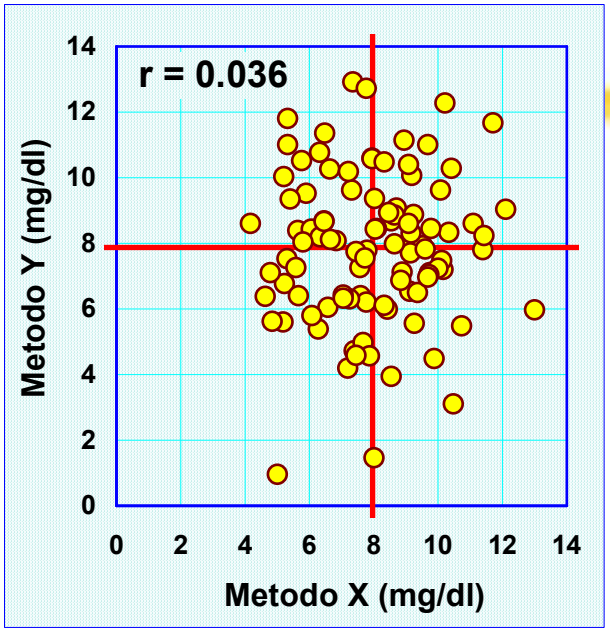
$\bar{y} = 6.2 \quad D_Y = 2.26 \; s_Y = \sqrt{\frac{2.26}{10-1}} = 0.501$ mg/dl

Codevianza $C_{X,Y} = \sum (x - \bar{x})(x - \bar{y}) = 2.05$ $(\text{mg/dl})^2$

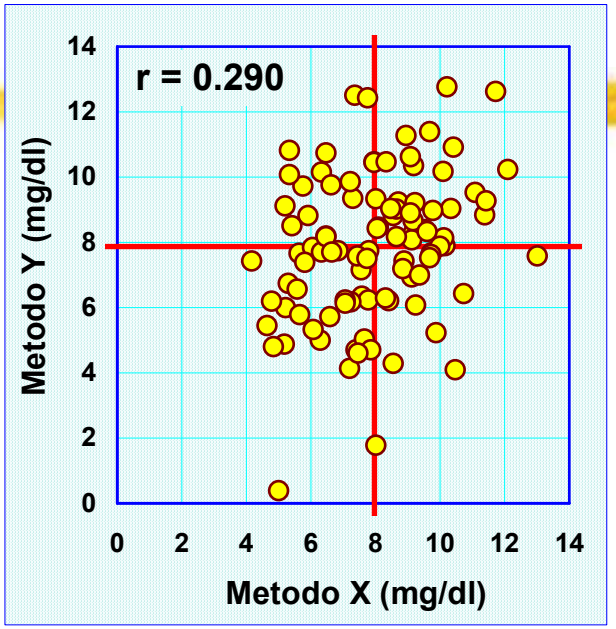
Covarianza $s_{X,Y} = \frac{2.05}{(10-1)} = 0.228 (\text{mg/dl})^2$

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{C_{XY}}{\sqrt{D_X D_Y}} = 0.97$$

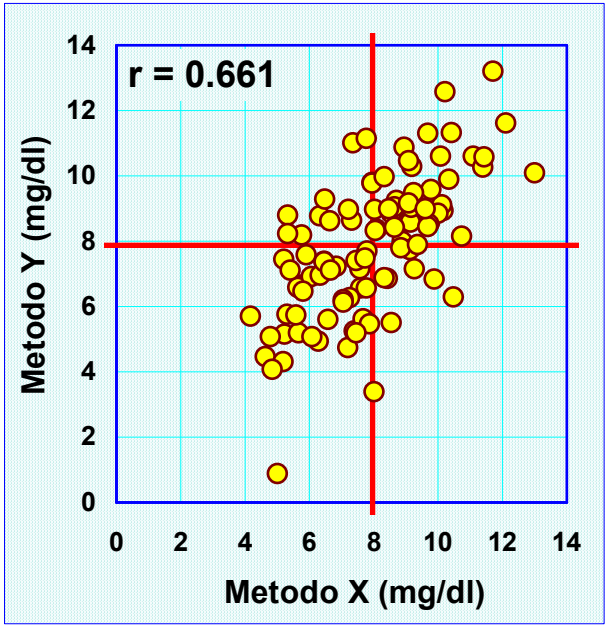
uno studente alla prima lezione



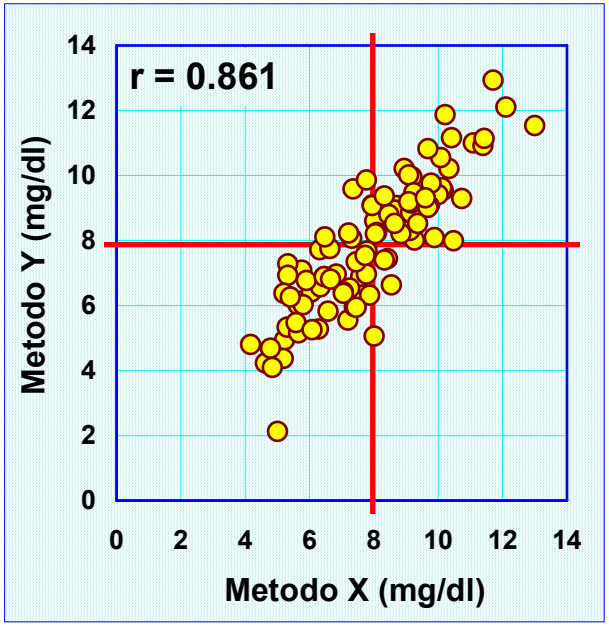
uno studente alla seconda lezione



uno studente all'ultima lezione



un analista esperto



Gli esempi qui riportati si riferiscono alla correlazione tra i valori di uricemia rilevati, *in differenti condizioni*, con due metodi di misura (X e Y) su un campione di 100 soggetti anziani.

Linee guida per l'interpretazione del valore di r

Valore assoluto di r	Giudizio sulla correlazione
0.0-0.2	Bassissima
0.2-0.4	Bassa
0.4-0.6	Media
0.6-0.8	Alta
0.8-1.0	Molto alta

Relazione non lineare

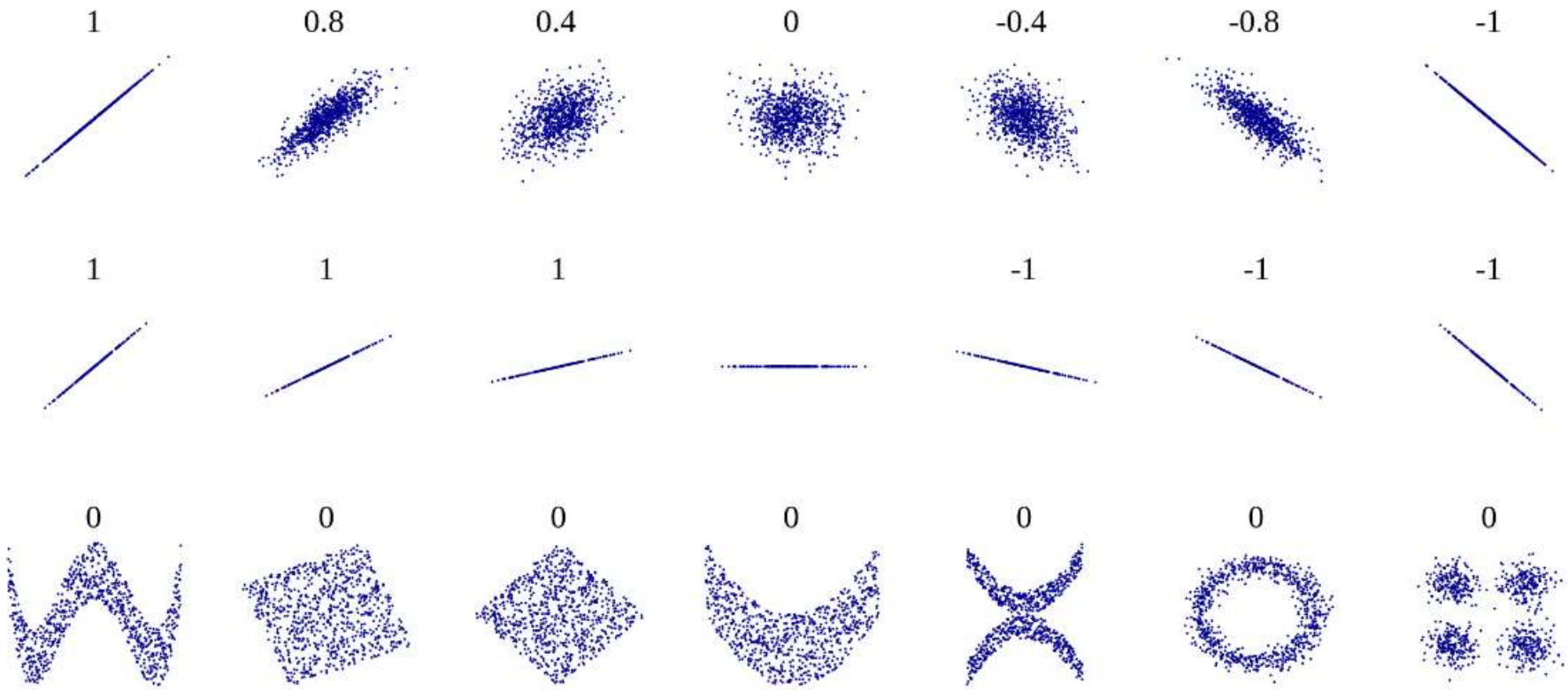
r=0.15

Un valore basso di r non indica assenza di relazione, ma significa mancanza di relazione LINEARE!

$Y=a+bX \longrightarrow Y=a+bX+cX^2$

Non interpretare mai il valore di r da solo, è sempre meglio costruire un grafico di dispersione dei dati.

Esempi di correlazione



fonte Wikipedia

Indovina l'indice di correlazione

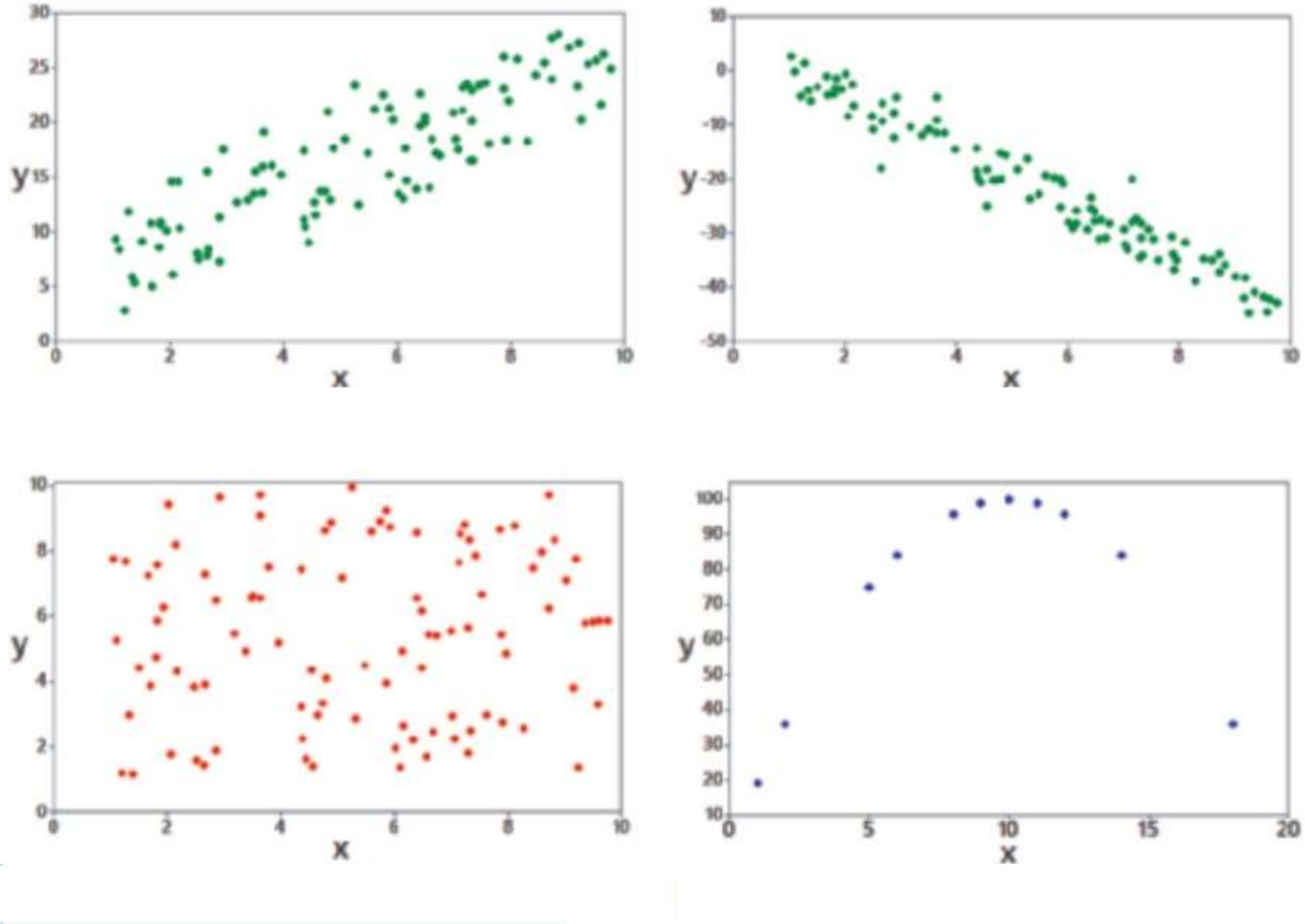


FIGURE 10-2 Scatterplots

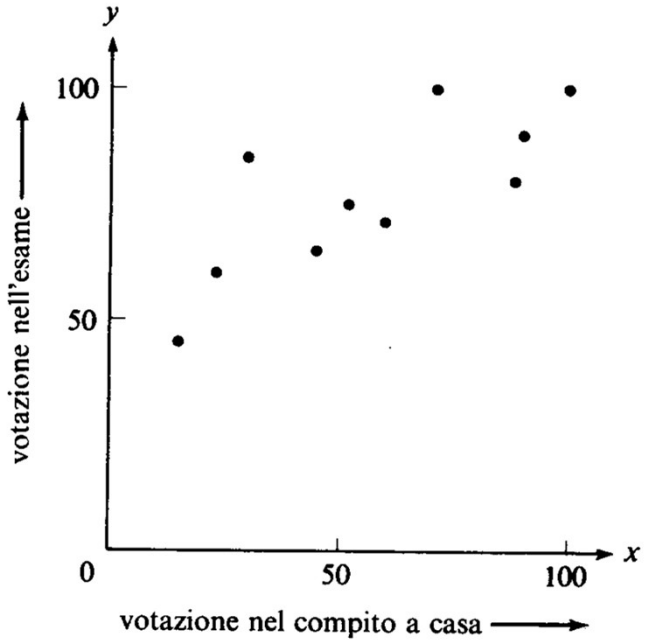
http://guessthecorrelation.com/
http://www.wilderdom.com/301/int/cor-guess.html

Esercizio

Studenti	1	2	3	4	5	6	7	8	9	10
X_i (voti casa)	90	60	45	100	15	23	52	30	71	88
Y_i (voti esame)	90	71	65	100	45	60	75	85	100	80

$$s_{XY} = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{(n-1)} C_{XY}$$

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{C_{XY}}{\sqrt{D_X D_Y}}$$



1) Calcolare r

r=408.29/(29.67x17.63)=0.78

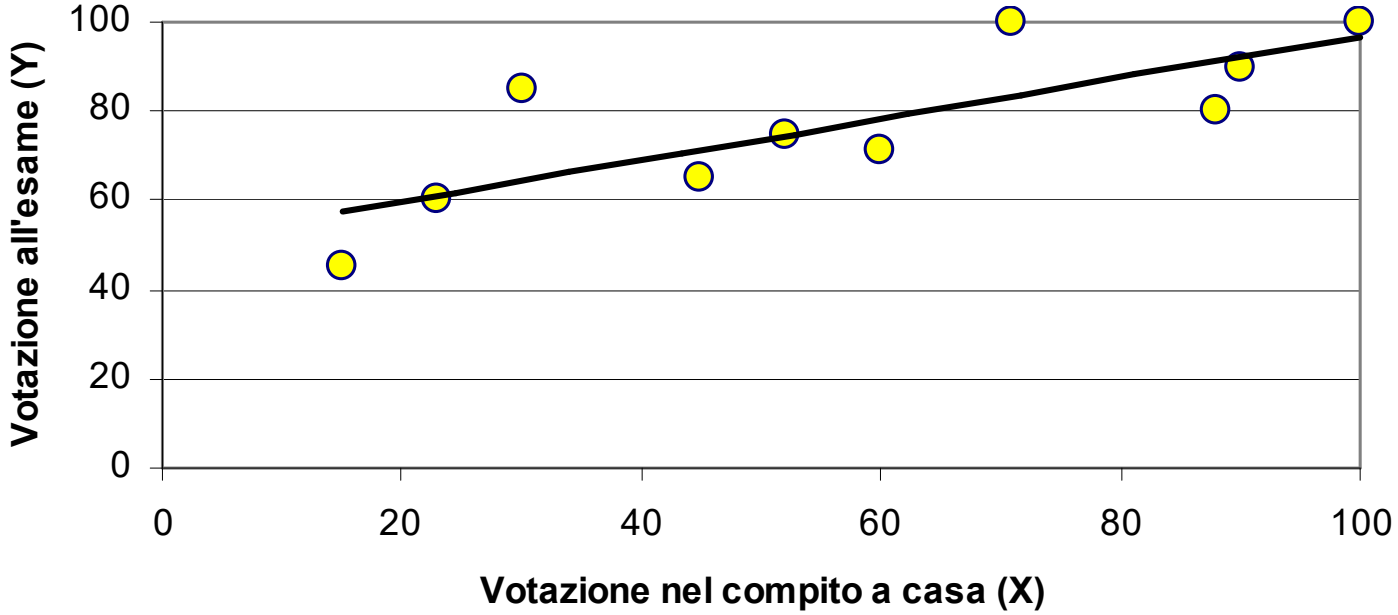
2) ….e se il professore avesse trovato un valore di r prossimo a 0?

3) ….e se il professore avesse trovato un valore di r prossimo a -1?

Retta di regressione

Se esiste una relazione lineare tra X e Y …

… posso interpolare i punti con una retta



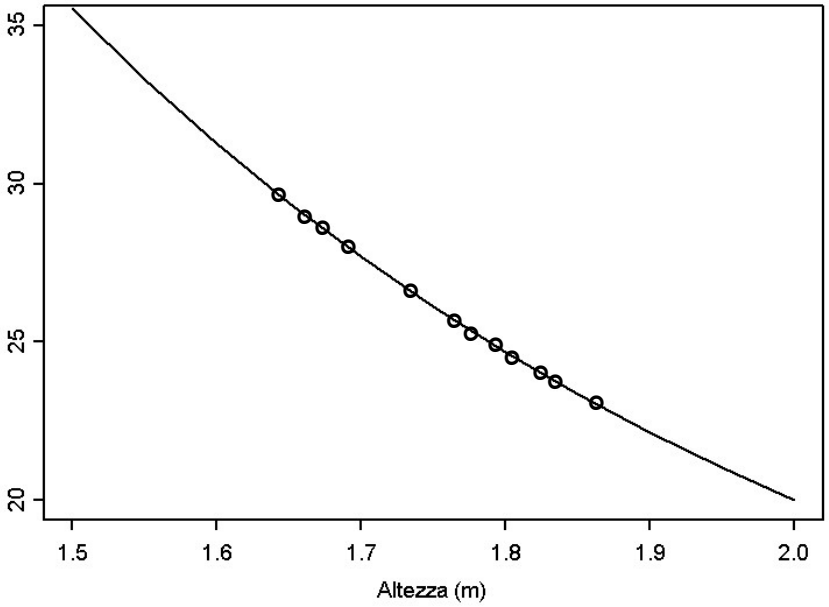
Potrei così cercare di 'predire' il voto dell'esame (variabile dipendente) in funzione del voto nel compito a casa (variabile indipendente) !

Relazione deterministica vs relazione statistica

BMI=(peso-kg)/(altezza-m)^2

Valutiamo il BMI per un insieme di valori di altezza che abbiamo osservato su un insieme di soggetti con un particolare peso:

- peso = 80 kg
- altezza = da 1.6 a 1.9 m



Relazione deterministica

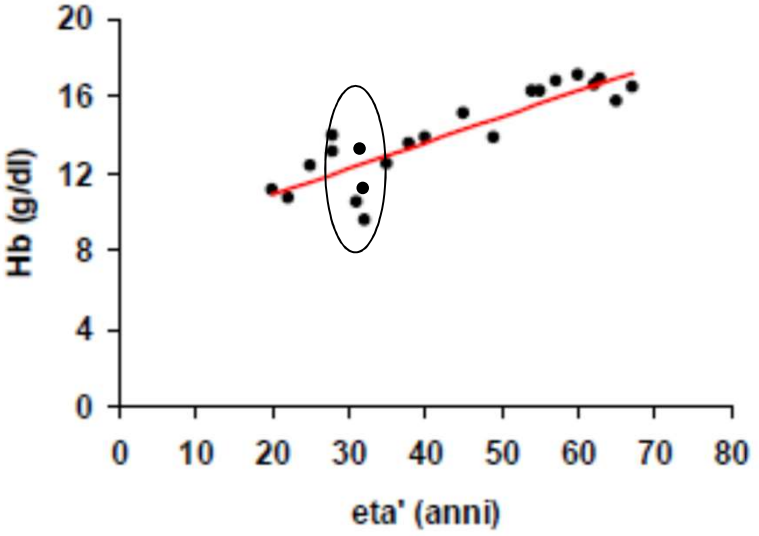
Relazione deterministica vs relazione statistica

$$\mathbf{Hb} = \beta_0 + \beta_1 * \text{età} + \varepsilon$$

Valutiamo i valori di emoglobina (g/dl) di un insieme di donne con età variabile.

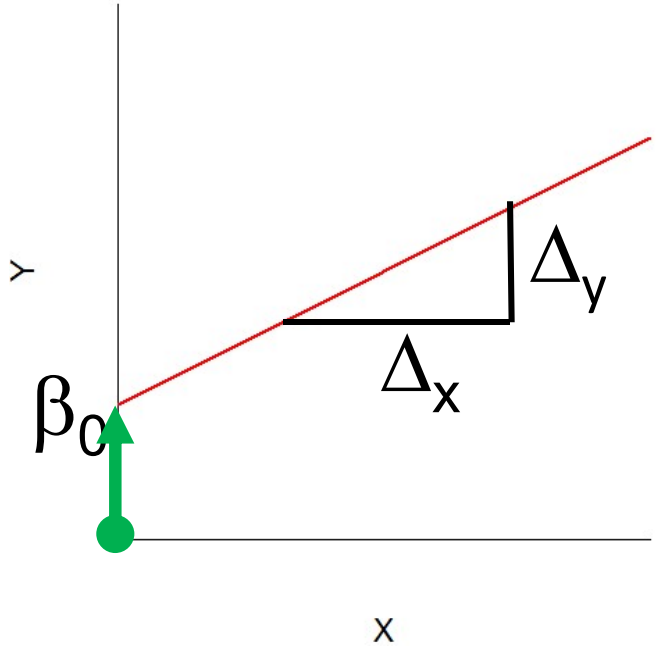
- Hb = da 10 a 16.5 g/dl
- età da 20 a 70 anni

I valori di emoglobina variano in donne con la stessa età



Relazione statistica

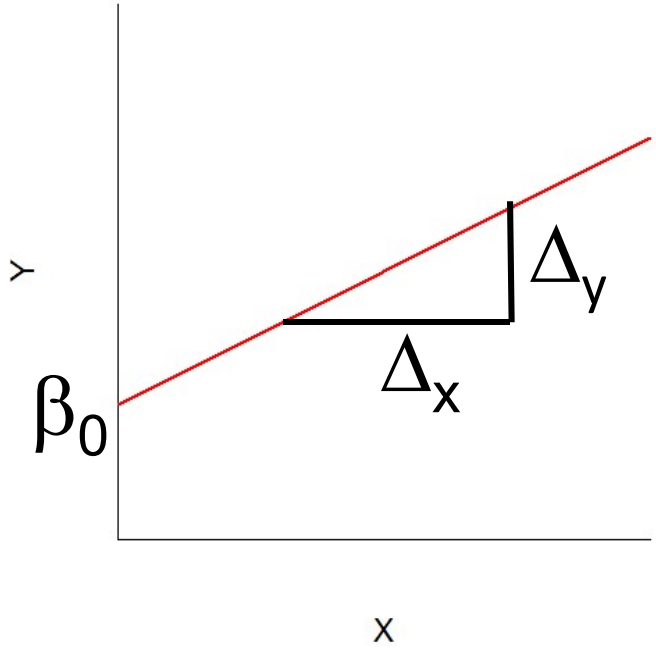
Equazione di una retta $Y = \beta_0 + \beta_1 X$



$\beta_1 = \Delta_y / \Delta_x$

β_0 = intercetta, cioè il punto in cui la retta interseca l'asse delle Y (valore di Y quando X=0)

Equazione di una retta $Y = \beta_0 + \beta_1 X$



$\beta_1 = \Delta_y / \Delta_x$

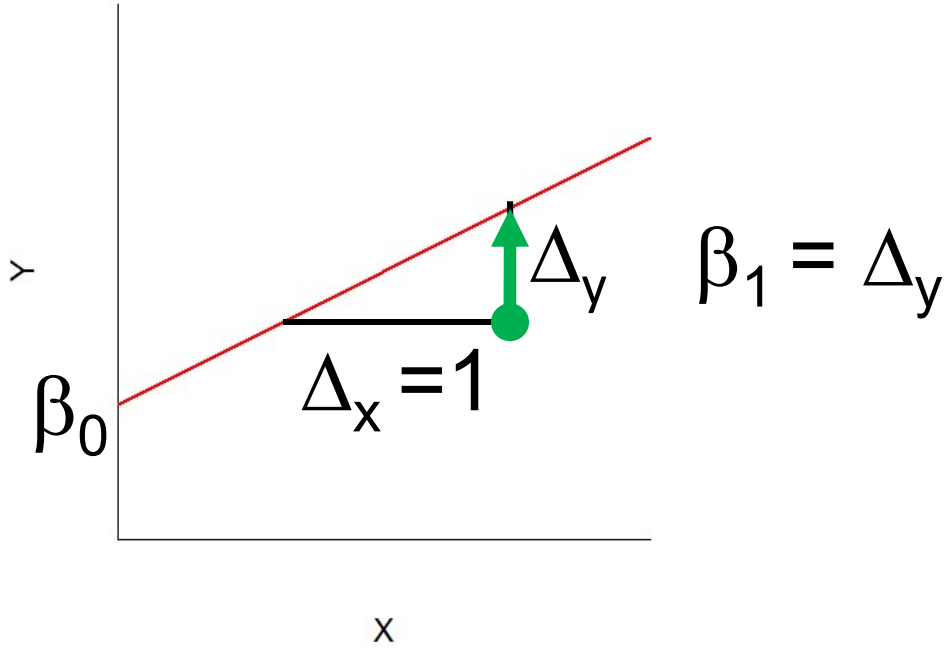
β_0 = intercetta, cioè il punto in cui la retta interseca l'asse delle Y (valore di Y quando X=0)

β_1 = coefficiente angolare o pendenza della retta (variazione di Y quando X aumenta di una unità)

$\beta_1 < 0$ -> proporzionalità inversa

$\beta_1 > 0$ -> proporzionalità diretta

Equazione di una retta $Y = \beta_0 + \beta_1 X$



β_0 = intercetta, cioè il punto in cui la retta interseca l'asse delle Y (valore di Y quando X=0)

β_1 = coefficiente angolare o pendenza della retta (variazione di Y quando X aumenta di una unità)

$\beta_1 < 0$ -> proporzionalità inversa

$\beta_1 > 0$ -> proporzionalità diretta

Modello di regressione lineare semplice

La **regressione lineare semplice** è un modello che studia come una variabile di risposta Y (*var. dipendente*) dipenda da una variabile esplicativa X (*var. indipendente*).

La relazione fra Y e X è riassunta dall'equazione di una retta:

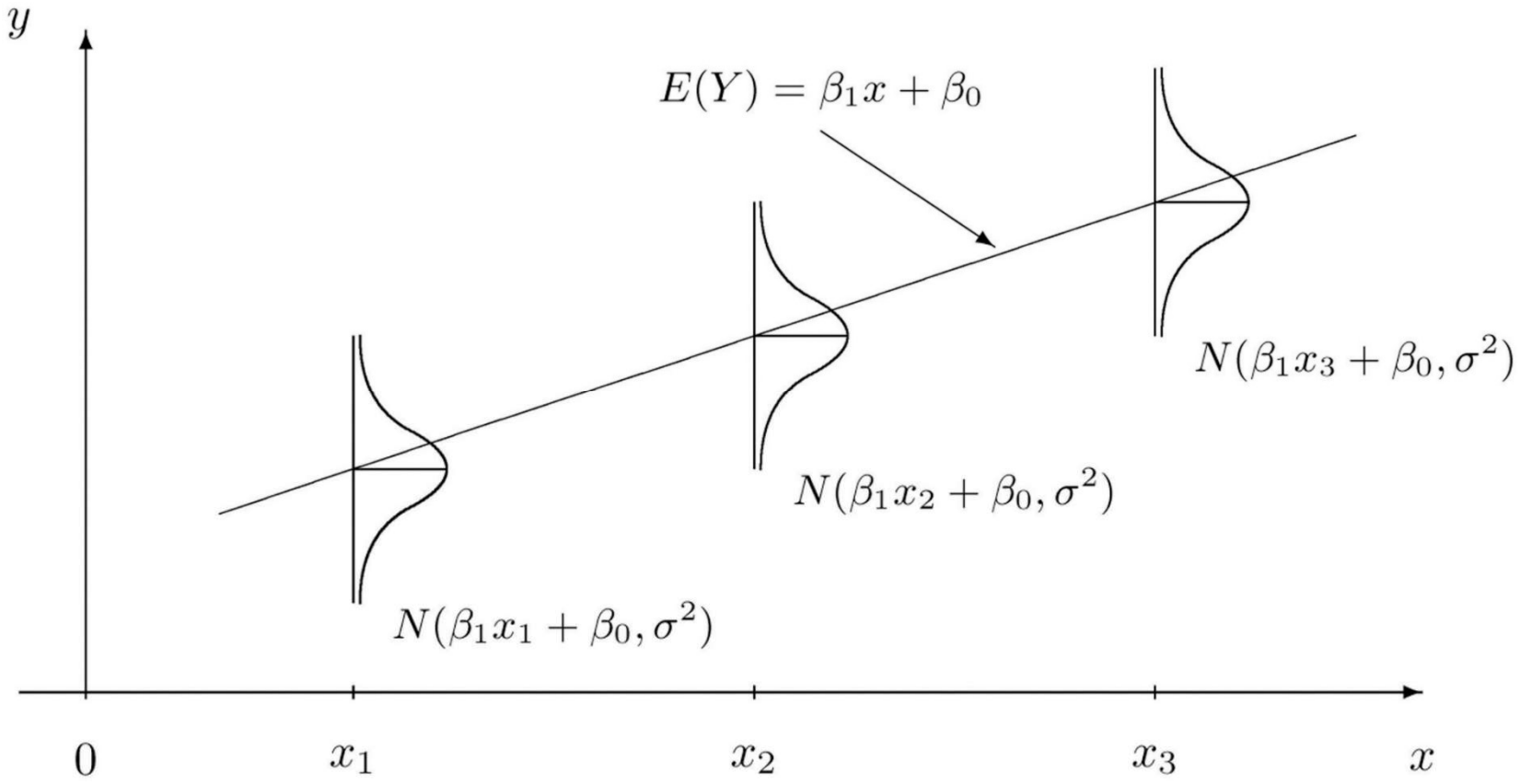
$$Y = \underbrace{\beta_0 + \beta_1 X}_{\text{parte sistematica (segnale)}} + \underbrace{\varepsilon}_{\text{parte casuale (rumore)}}$$

Questo modello ipotizza che la risposta y sia il risultato della somma di:

- una parte sistematica (che è funzione lineare di x)
- una parte casuale (che essendo puramente accidentale non dipende da x)

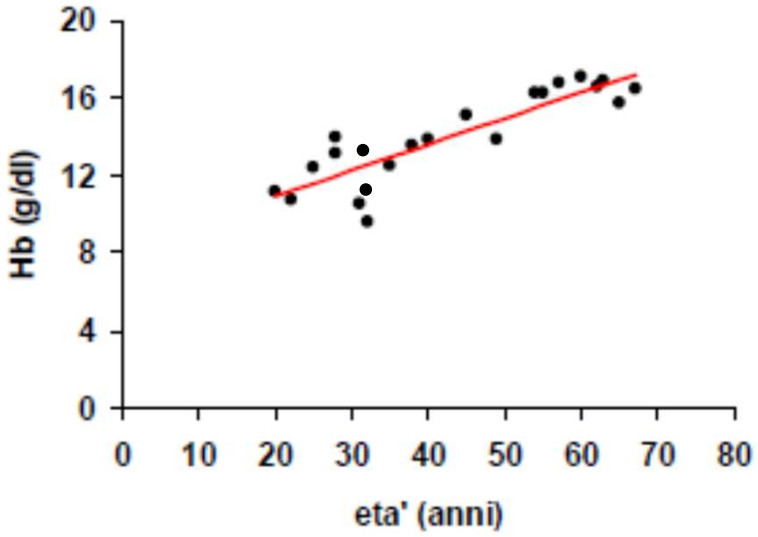
Modello di regressione lineare semplice

L'idea di base del modello di regressione e' che **le medie di Y al variare di x** stiano su una linea retta.



Dal modello di regressione (teorico) alla linea di regressione (stima)

La retta di regressione di popolazione è un modello con parametri β_0 e β_1 che vengono stimati da b_0 e b_1 a partire dal campione di dati osservati.



Modello di regressione

$$y_x = \beta_0 + \beta_1 * x + \varepsilon$$

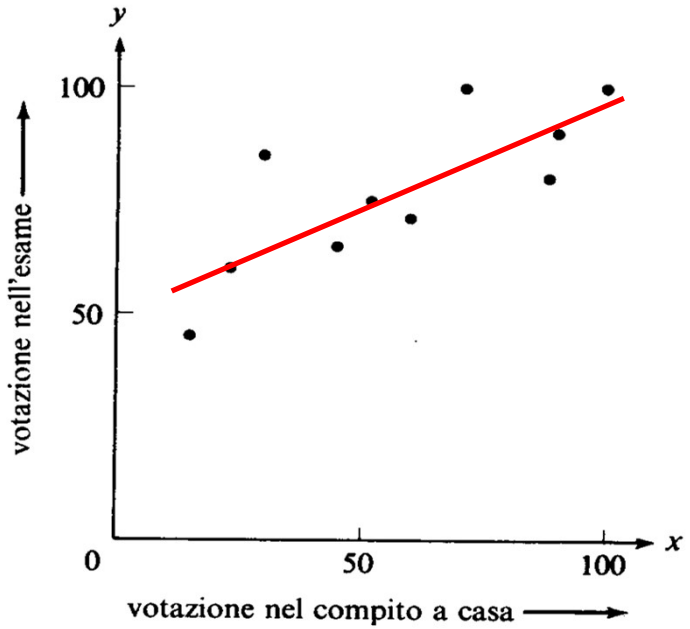
Linea di regressione

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

Da notare che nell'equazione della retta non compare più la componente casuale ε, dato che si assume che in media essa sia nulla e non dipenda da X.

Stima della retta di regressione

Poiché le rette del piano sono infinite, dobbiamo definire un criterio che permetta di stimare la "migliore retta" di regressione per i punti osservati.
Una scelta ragionevole è quella della retta che passa il più vicino possibile all'insieme di questi punti.



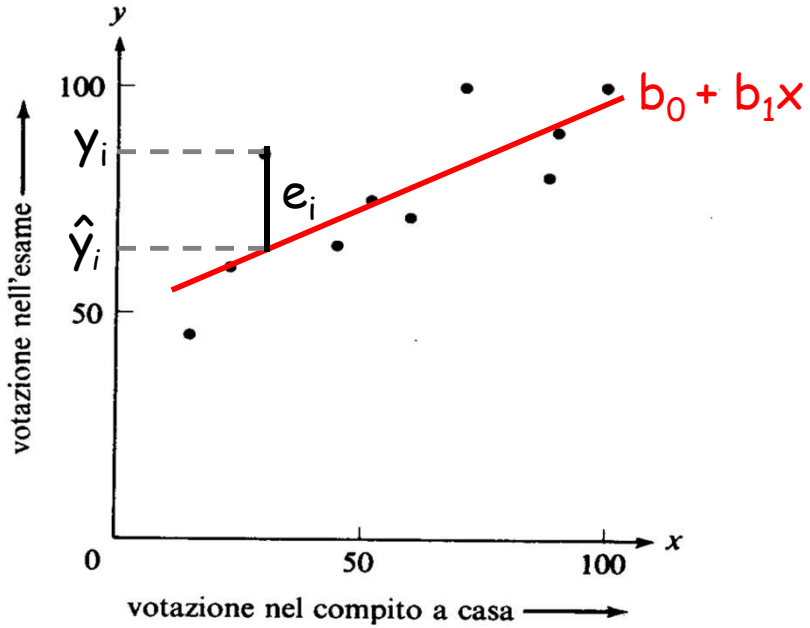
Stima della retta di regressione

Nello stabilire una misura di distanza tra la retta e l'insieme dei punti, occorre definire:

y_i = valore *osservato* di Y per l'unità i

$\hat{y}_i = b_0 + b_1 x_i$ = valore *previsto* di Y per l'unità i

$e_i = y_i - \hat{y}_i$ = errore



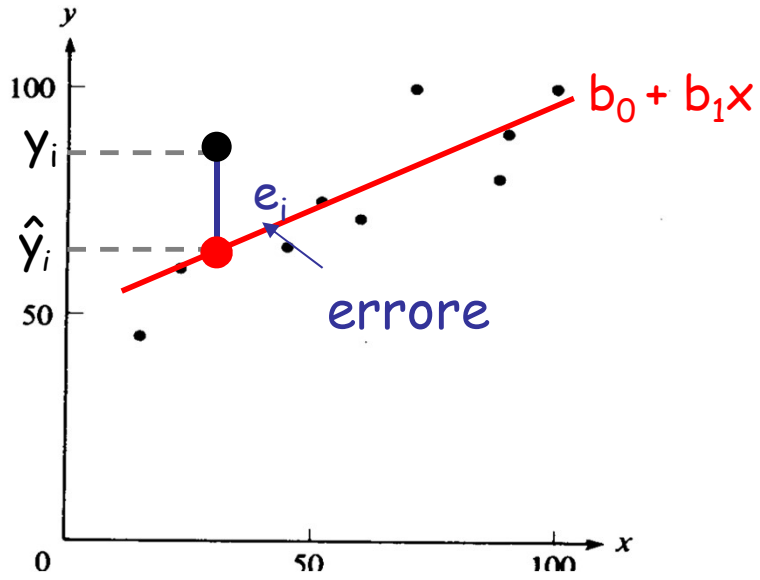
Stima della retta di regressione

Si sceglie la retta che minimizza la somma dei quadrati degli errori, ovvero degli scarti fra i valori osservati di y e quelli predetti sulla base della retta

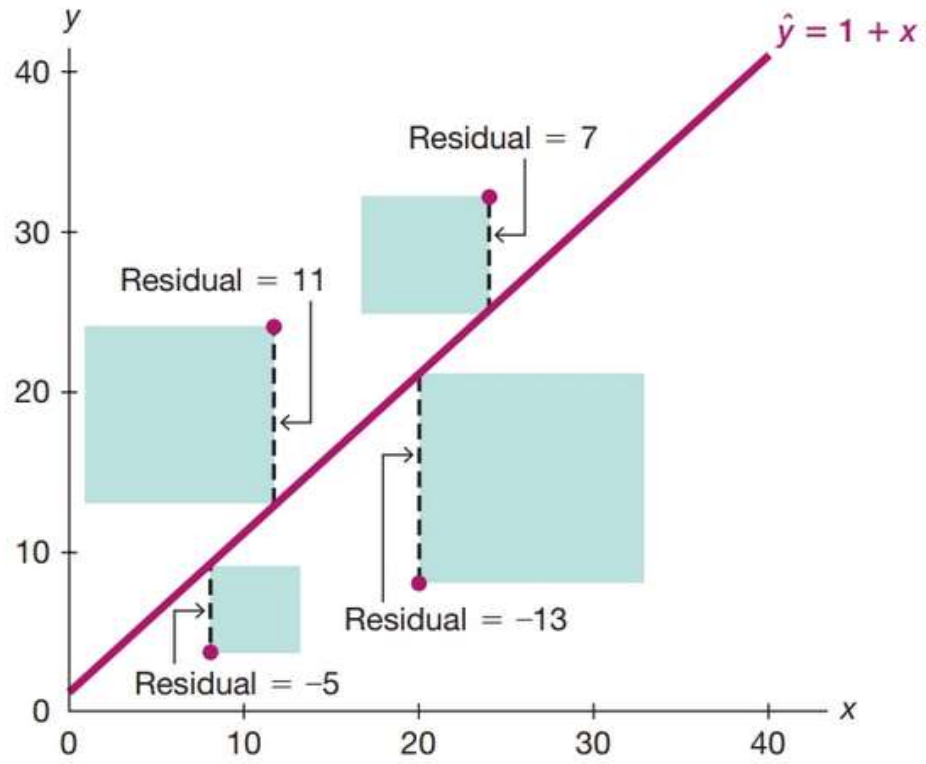
$$SS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

valore osservato

valore previsto



Esempio



i: (x=8,y=4)

$$y = 4$$

$$\hat{y} = 1 + 8 = 9$$

Stima della retta di regressione

Le stime b_1 e b_0 si ottengono minimizzando SS:

$$b_1 = \frac{C_{XY}}{D_X} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s^2_x}$$

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

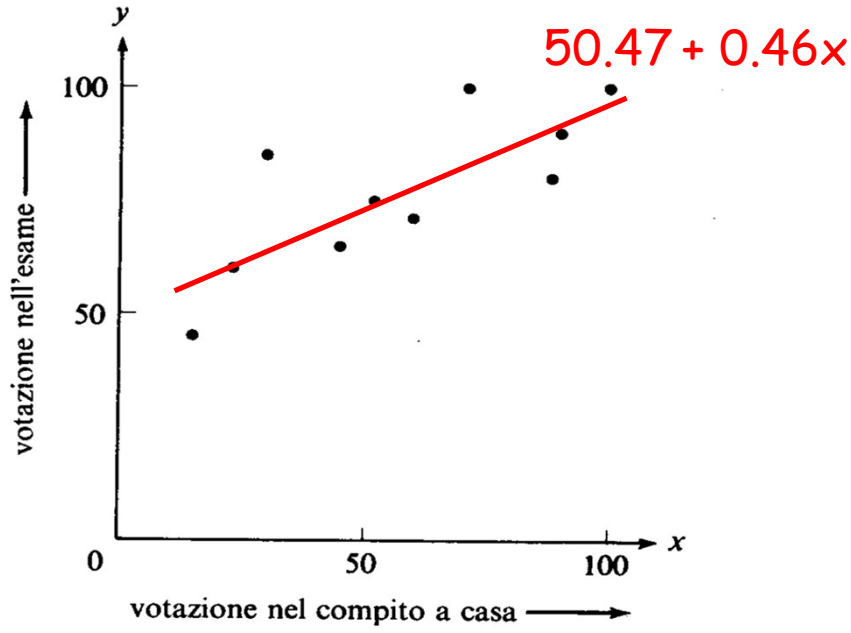
Nell'esempio:

$\bar{y} = 77.1 \quad \bar{x} = 57.4$

$s_{xy} = 408.3 \quad s_x = 29.7$

$b_1 = 408.3/(29.7)^2 = 0.46$

$b_0 = 77.1 - 0.46 \cdot 57.4 = 50.47$



Interpretazione della retta di regressione

$$\hat{y} = 50.47 + 0.46 \cdot x$$

Interpretazione dell'intercetta (b_0=50.47) ?

Voto d'esame di uno studente con votazione 0 nel compito a casa.

Interpretazione della pendenza (b_1=0.46)?

Se il voto del compito a casa migliora di un punto, il voto all'esame cresce di circa mezzo punto.

Stima della retta di regressione

Proprietà:

1) La retta stimata passa per il baricentro delle osservazioni.

$$\text{per} \quad x = \bar{x} \quad \Rightarrow \quad \hat{y} = \bar{y}$$

2) La somma dei residui è nulla.

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - \hat{y}_i = 0$$

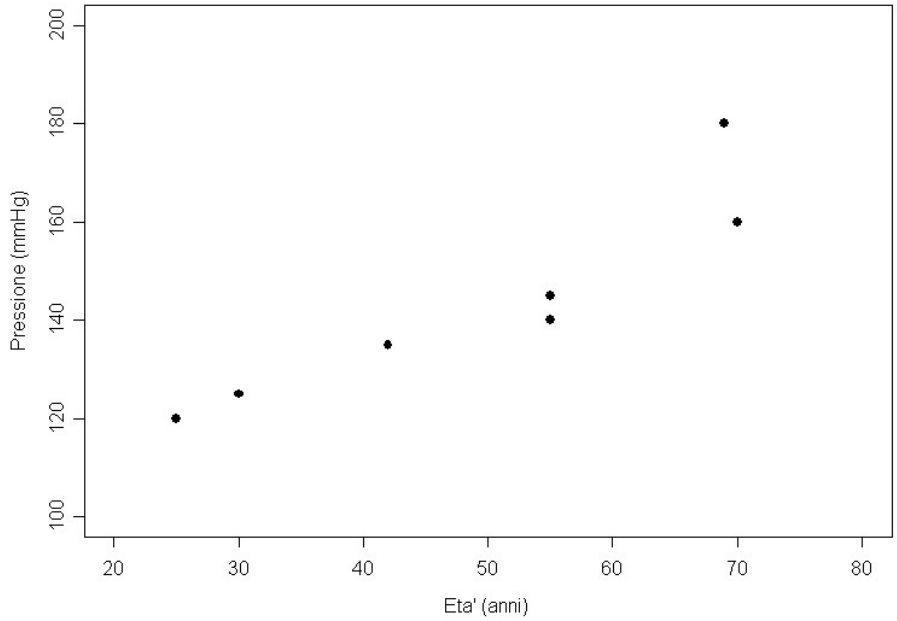
3) La somma dei valori stimati è uguale alla somma dei valori osservati.

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$$

Esempio

Dato un campione di sette individui, si vuole valutare la relazione tra la pressione arteriosa e l'età.

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



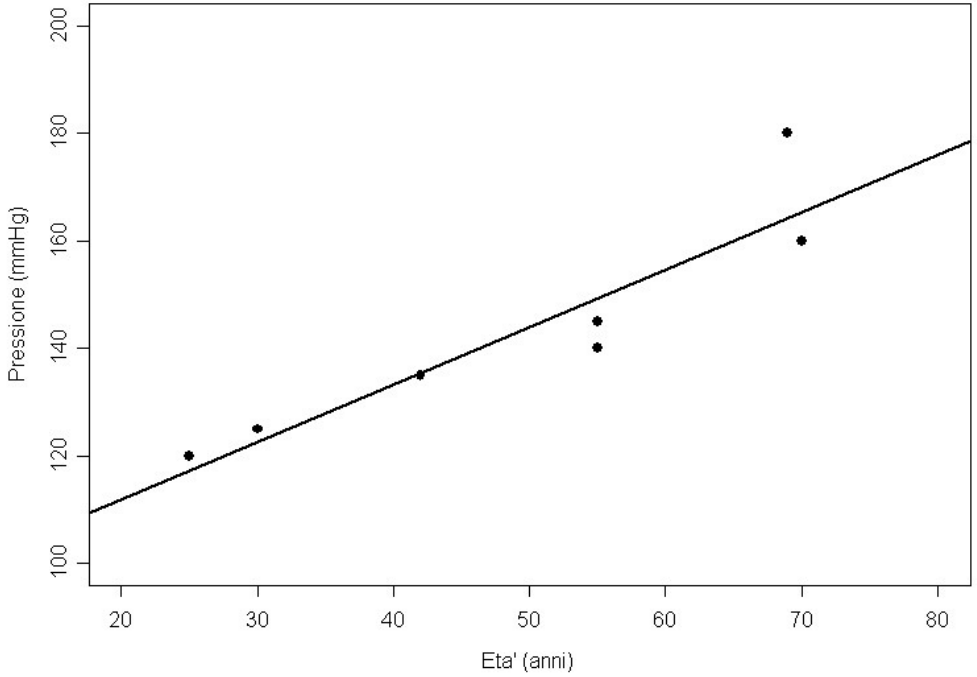
Come varia la pressione in funzione dell'età?

Esempio - soluzione

Ingredienti per la stima della retta di regressione:

$\bar{y} = 143.57$ $\bar{x} = 49.43$ $s^2_x = 271.10$ $s_{xy} = 291.33$

$b_1 = 291.33/271.10 = 1.07$ $b_0 = 143.57 - 1.07 \cdot 49.43 = 90.46$

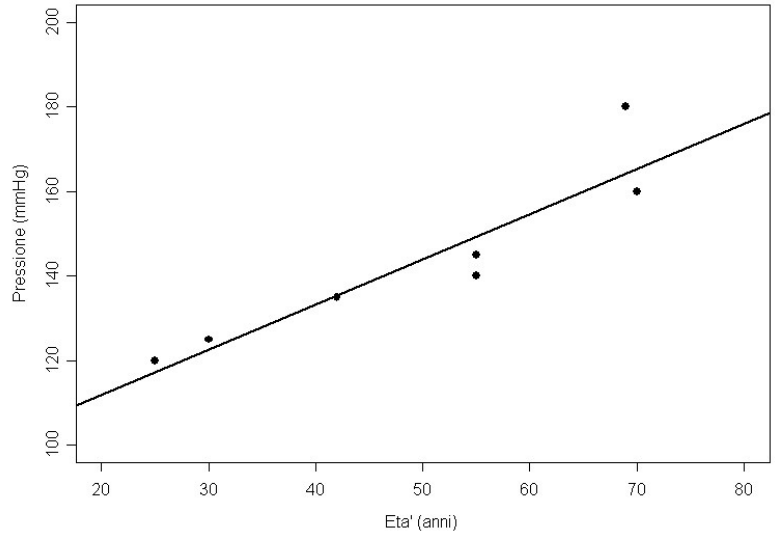


La pressione aumenta di 1 mmHg per ogni anno di età.

Esempio - quesiti

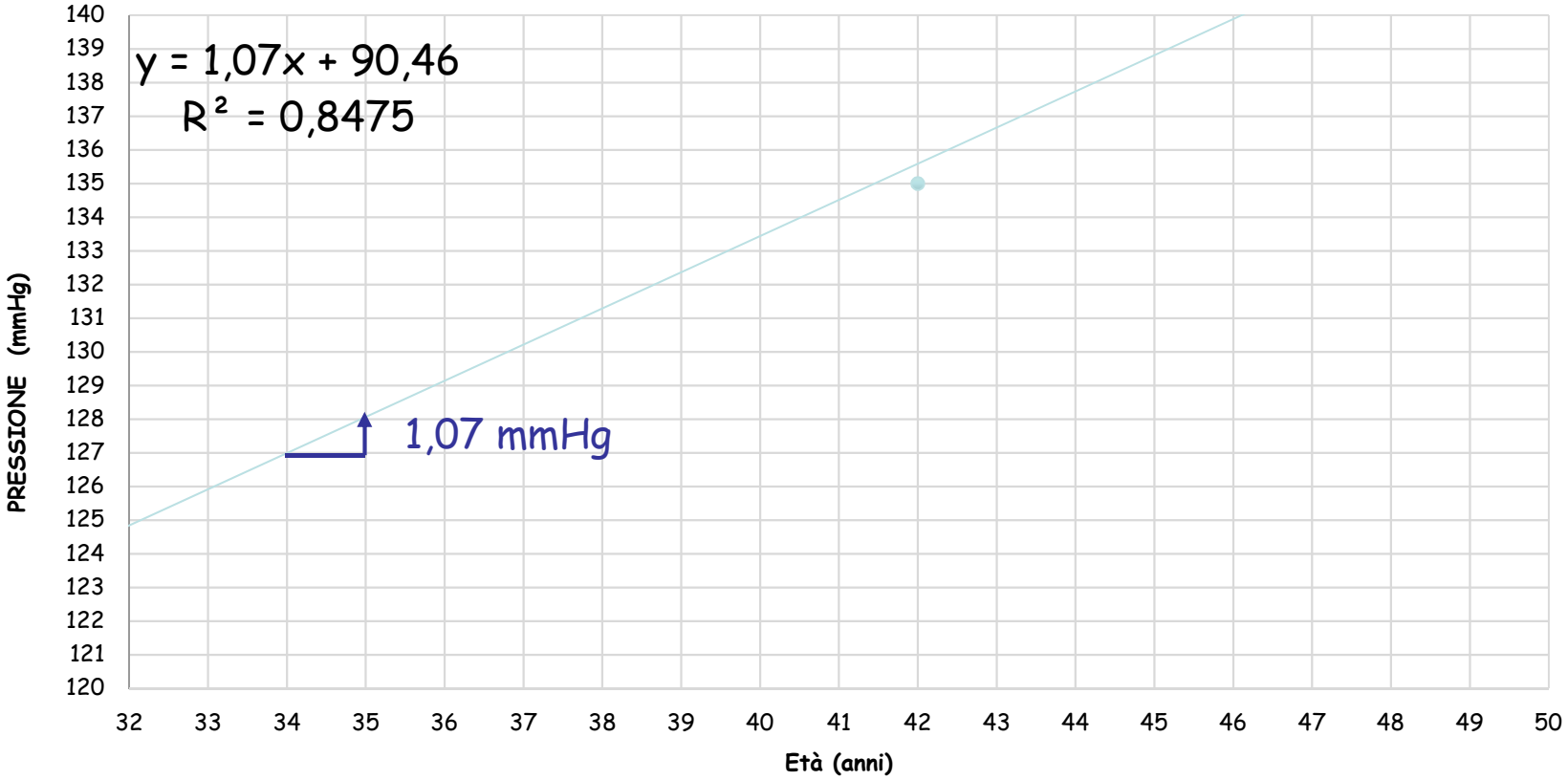
pressione = 90.46 + 1.07età

- Interpretare l'intercetta ed il coefficiente angolare.
- Di quanto varia la pressione
  - all'aumentare dell'età di 2 anni?
  - all'aumentare dell'età di 10 anni?
  - al diminuire dell'età di 1 anno?



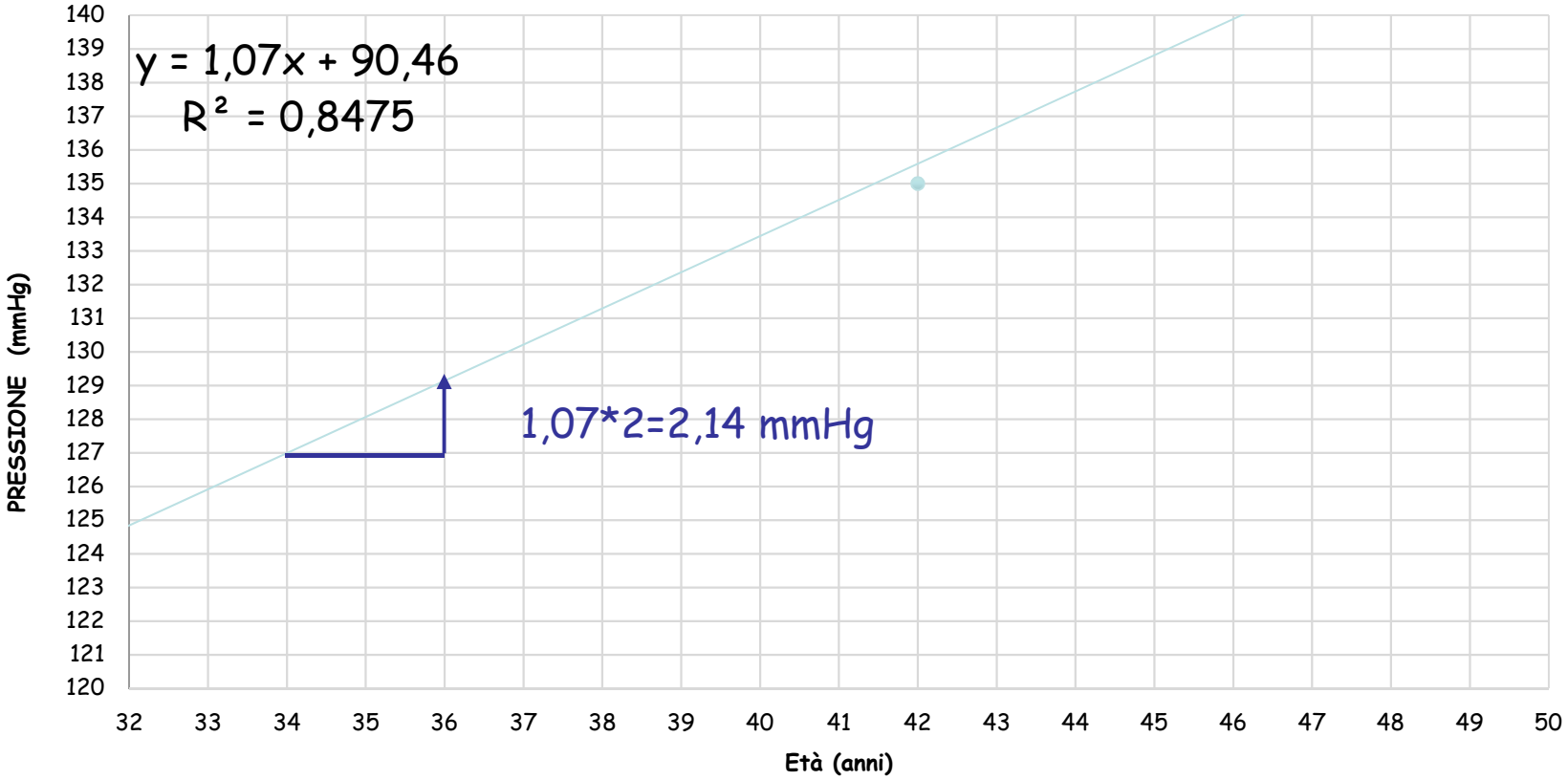
Esempio -

Di quanto varia la pressione all'aumentare dell'età di 1 anno?



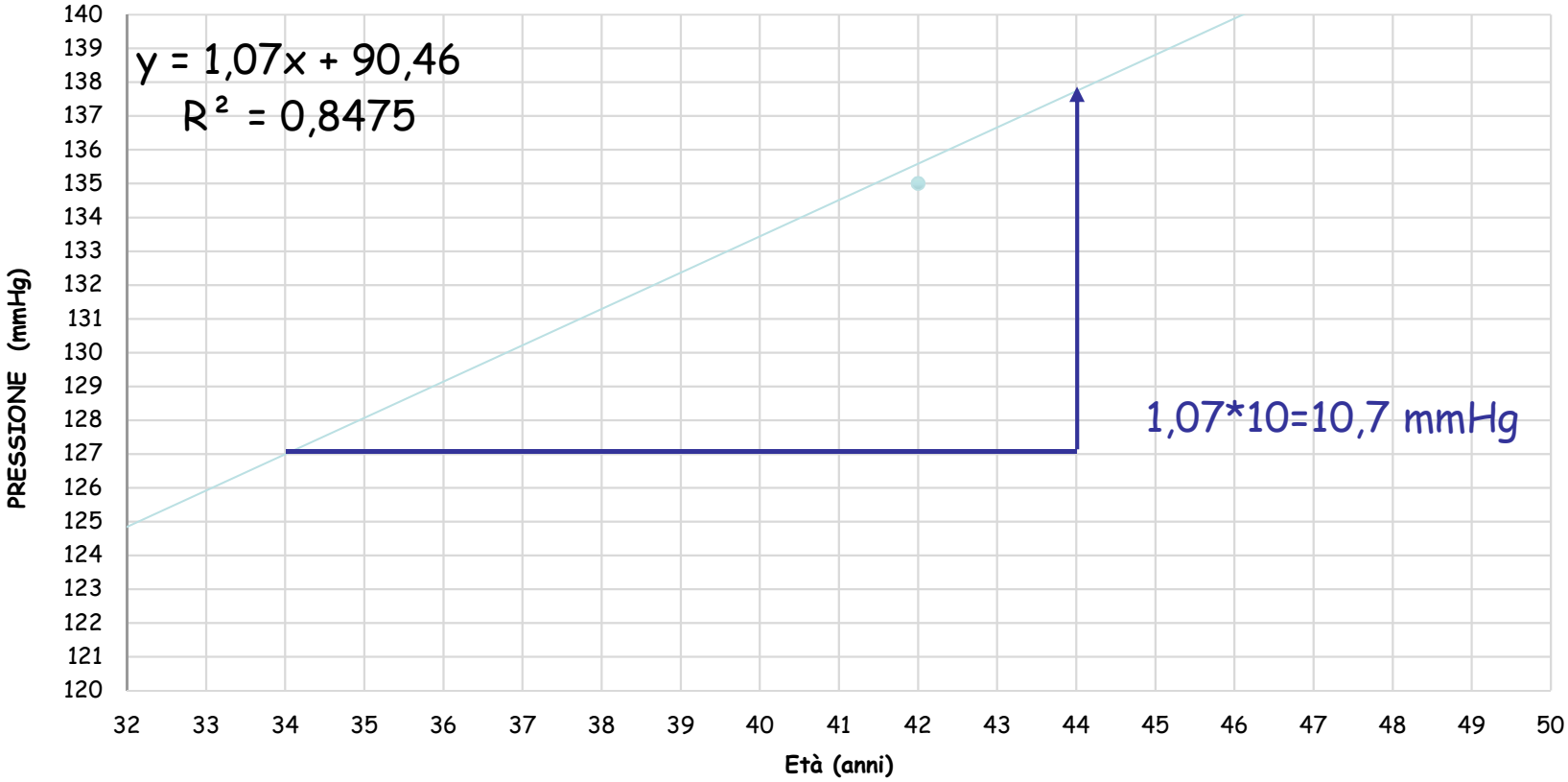
Esempio -

Di quanto varia la pressione all'aumentare dell'età di 2 anni?



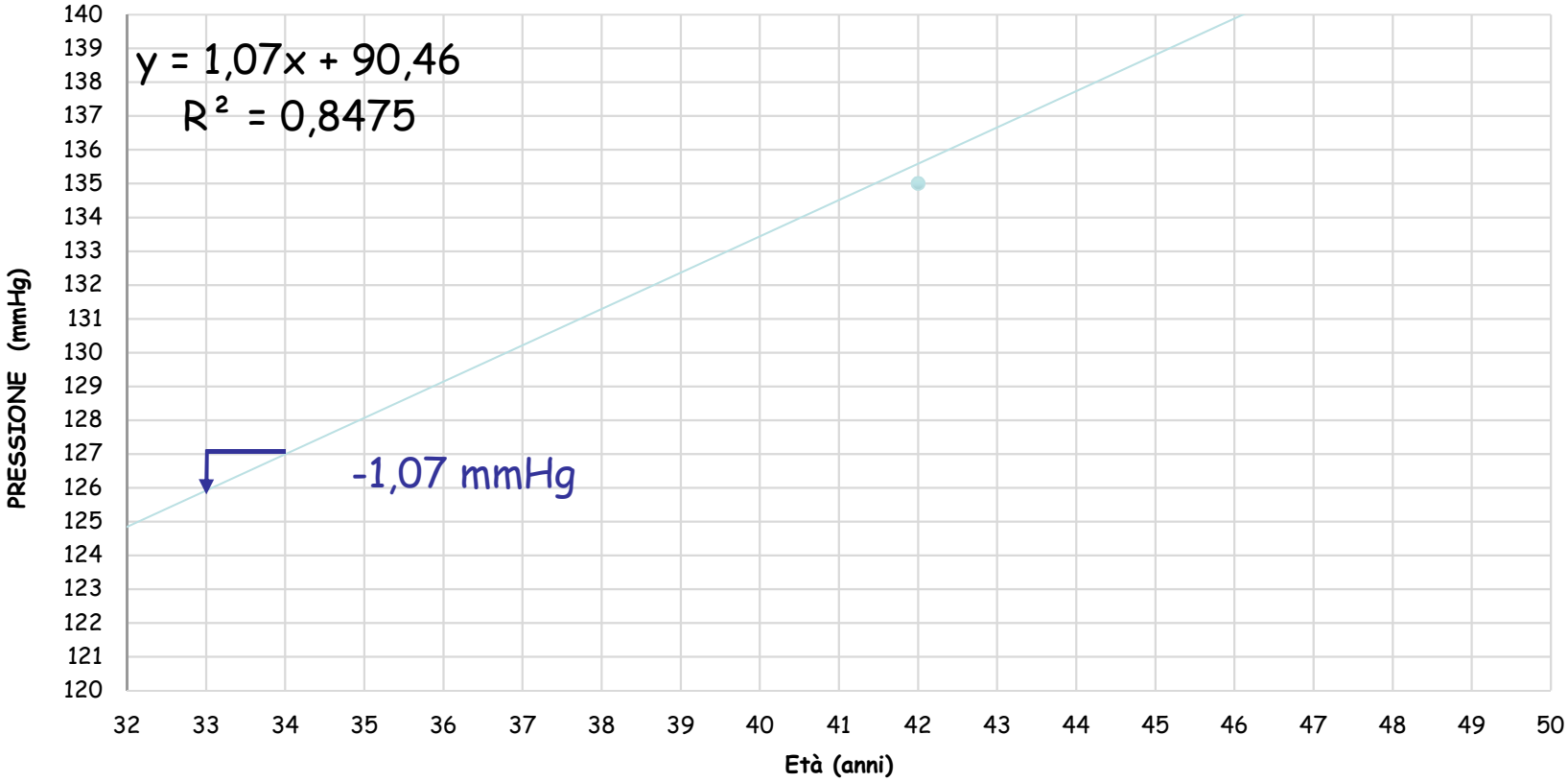
Esempio -

Di quanto varia la pressione all'aumentare dell'età di 10 anni?



Esempio -

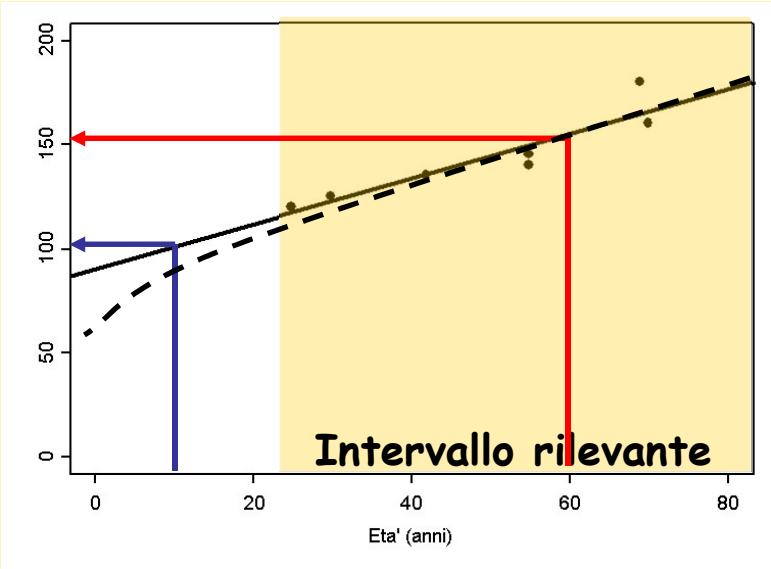
Di quanto varia la pressione al diminuire dell'età di 1 anno?



Interpolazione vs estrapolazione

Sulla base della relazione lineare fra X e Y, potrei tentare delle predizioni (ad es. pressione in funzione dell'età).

⚠ Attenzione a fare previsioni dove non ci sono dati!

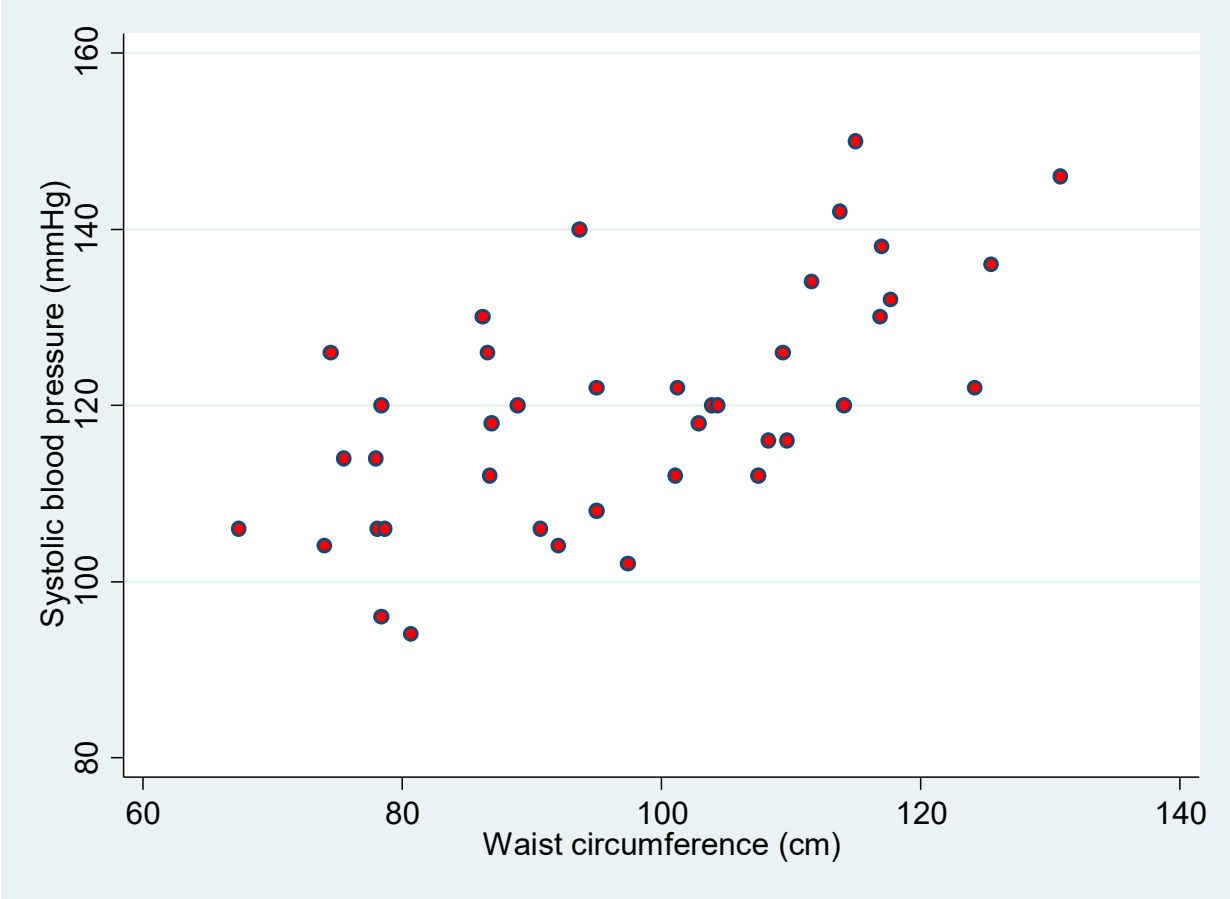


ESTRAPOLAZIONE: prevedere y in corrispondenza di un valore di X esterno dell'intervallo rilevante (mi aspetto che un soggetto di 10 anni abbia una pressione intorno a 100 mmHg).

INTERPOLAZIONE: prevedere y in corrispondenza di un valore di X interno dell'intervallo rilevante (mi aspetto che un soggetto di 60 anni abbia una pressione intorno a 150 mmHg).

Esempio (studio osservazionale «body data») -

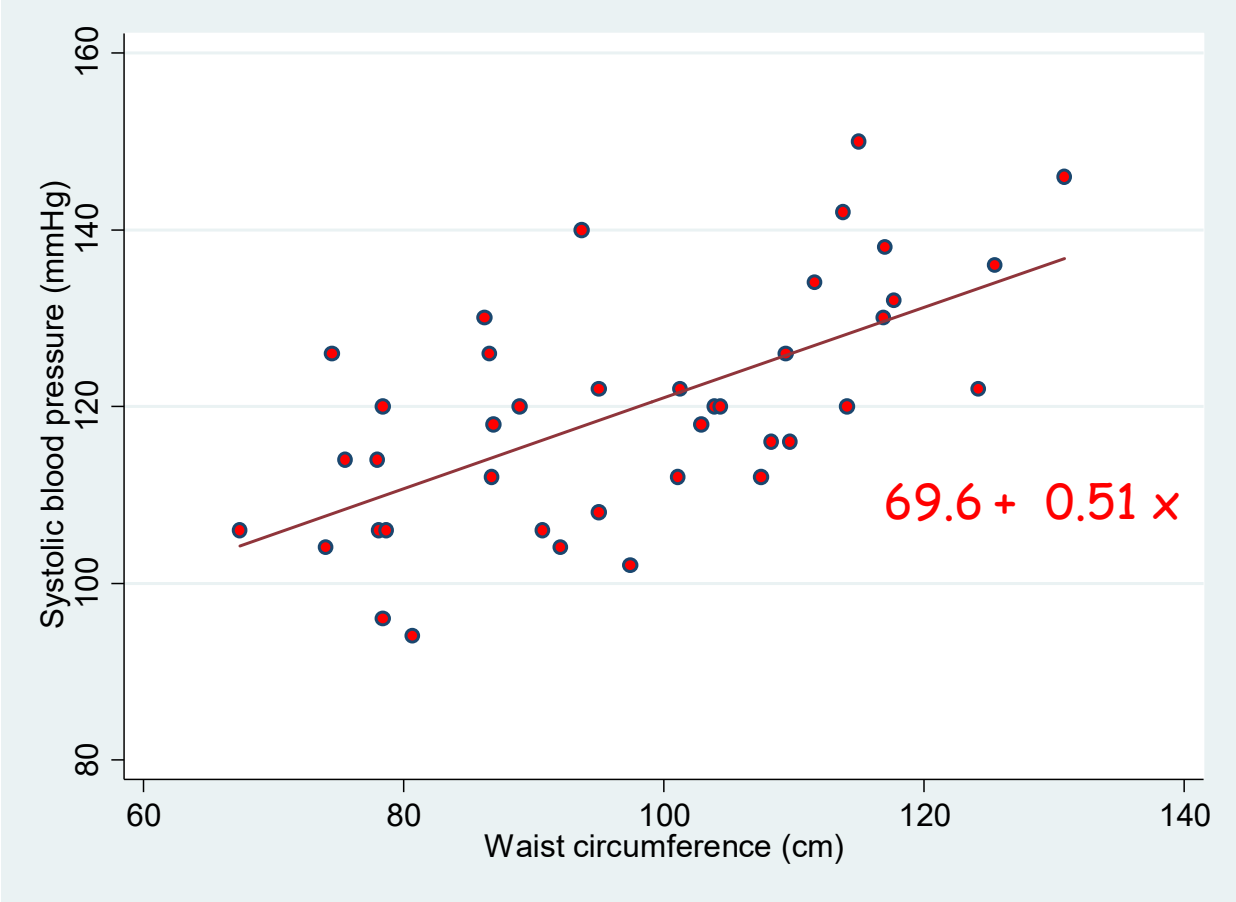
Esiste un'associazione tra circonferenza vita e pressione sanguigna sistolica?
Campione di 40 uomini.



Relazione lineare diretta.

Esempio (studio osservazionale «body data») -

Esiste un'associazione tra circonferenza vita e pressione sanguigna sistolica?
Campione di 40 uomini.



Relazione lineare diretta.

Quando forte?

$r = 0.6347$

Come varia mediamente la pressione in funzione della circonferenza vita?

Errore standard del coefficiente di regressione

$$y = 69.6 + 0.51 \cdot x$$

Se la circonferenza vita aumenta di un cm, la pressione sistolica aumenta di 0.51 mmHg.

Quanto è affidabile questa stima?

Stima del coefficiente: $b_1 = 0.51$

Errore standard: $SE(b_1)=0.101$

Intervallo di Confidenza (IC) 95%

$0.51 \pm 1.96 * 0.101 = [0.312; 0.708]$

Assunti:

- Osservazioni indipendenti tra di loro
- Scostamenti dalla linea di regressione Normali con varianza costante (omoschedasticità)

Errore standard del coefficiente di regressione

$$y = 69.6 + 0.51 \cdot x$$

Se la circonferenza vita aumenta di un cm, la pressione sistolica aumenta di 0.51 mmHg.

Quanto è affidabile questa stima?

Stima del coefficiente: $b_1 = 0.51$

Errore standard: $SE(b_1)=0.101$

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

$$\text{t-test}=\frac{b_1-0}{SE(b_1)} = \frac{0.51}{0.101} = 5.06 \quad \text{p-value}<0.0001$$

Assunti:

- Osservazioni indipendenti tra di loro
- Scostamenti dalla linea di regressione Normali con varianza costante (omoschedasticità)

Serum Uric Acid and Pulse Wave Velocity Among Healthy Adults: Baseline Data From the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil)

Cristina Pellegrino Baena,[1,2] Paulo Andrade Lotufo,[2] José Geraldo Mill,[3] Roberto de Sa Cunha,[3] and Isabela J Benseñor[2]

BACKGROUND
We aimed to evaluate a possible association between serum uric acid (SUA) levels and carotid-to-femoral pulse wave velocity (cf-PWV) among healthy participants of the ELSA-Brasil.

METHODS
We excluded subjects using antihypertensive medication, diuretics, allopurinol, binge drinkers, body mass index (BMI) >35 kg/m², and those with history of cardiovascular diseases (CVD). In a cross-sectional and sex-specific analysis, linear regression models were built having cf-PWV as dependent variable and SUA as independent variable. Multiple adjustments were subsequently made for age, heart rate and blood pressure, BMI, and fasting glucose levels as covariates. Product interaction terms were built to test interaction between SUA and other covariates.

RESULTS
We analyzed 1,875 men and 1,713 women (mean ages, 48.9±8.4 and 50.2±8.7 years, respectively). SUA was linearly associated with cf-PWV in men ($P = 0.01$) and in women ($P = 0.01$). After full adjustment, the association remained significant for men ($P = 0.01$) and no longer significant for women ($P = 0.10$). Fully adjusted linear coefficients β (95% CI) were 0.06 (0.015; 0.112) and 0.04 (−0.01; 0.12) in men and women, respectively. Significant interaction between SUA and age ($P = 0.02$) fasting glucose ($P < 0.01$) and BMI ($P = 0.02$) was found only for women.

CONCLUSION
In an apparently healthy population, SUA was significantly associated to cf-PWV in men but not in women.

Keywords: aortic stiffness; blood pressure; gender; hypertension; pulse wave velocity; uric acid.

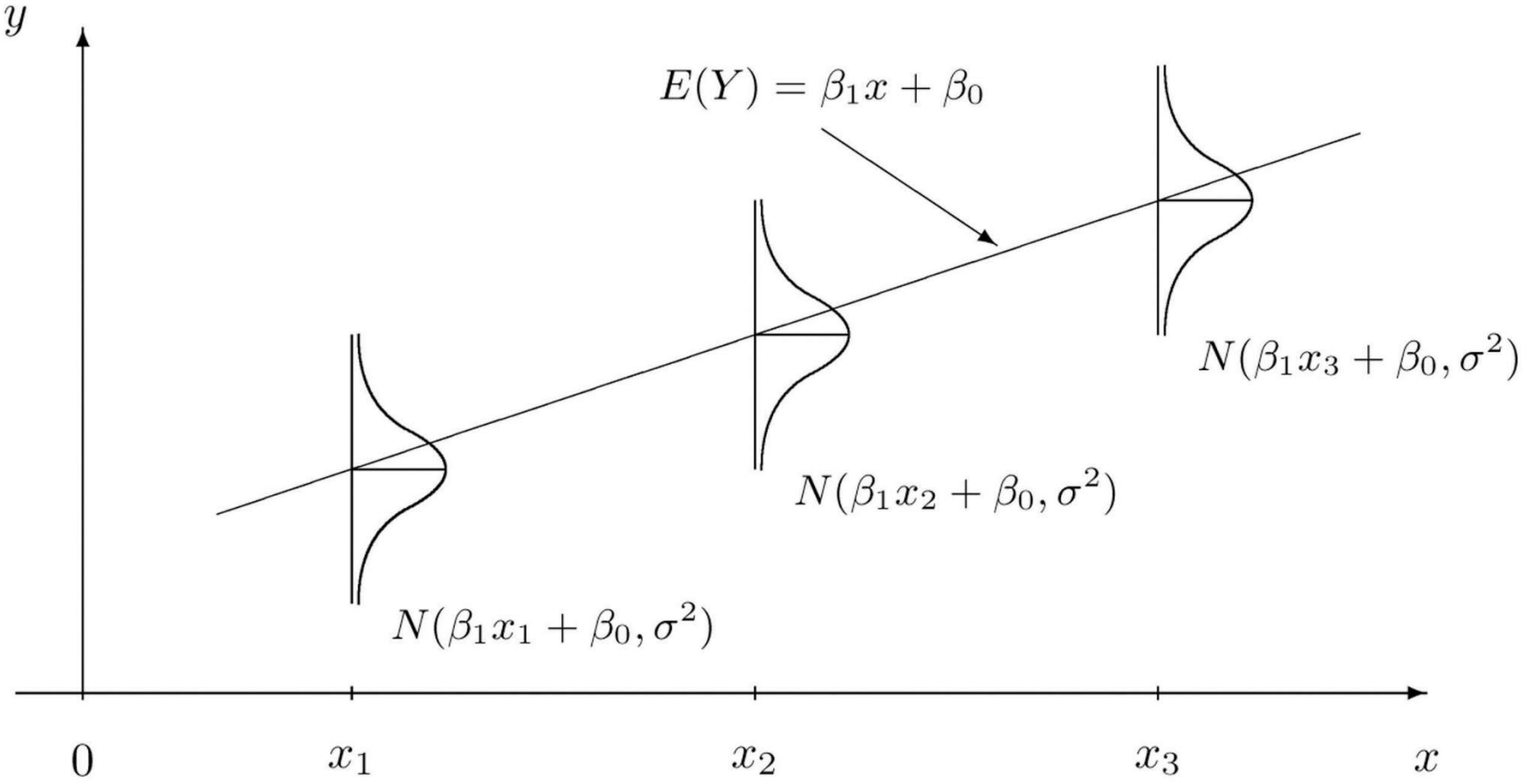
doi:10.1093/ajh/hpu298

Table 1. Linear association between serum uric acid with pulse wave velocity by genders

	Men (n = 1,865)			Women (n = 1,713)		
	β	CI (95%)	P	β	CI (95%)	P
Univariate	0.07	(0.03, 0.13)	<0.01	0.14	(0.13, 0.26)	<0.01
Adjusted for age	0.36	(0.05, 0.67)	<0.01	0.38	(0.05, 0.07)	<0.01
Plus blood pressure and heart rate	0.04	(−0.003, 0.089)	0.06	0.07	(0.013, 0.125)	0.01
Fully adjusted	0.06	(0.015, 0.112)	0.01	0.04	(−0.01, 0.12)	0.10

Fully adjusted model: age, blood pressure, and heart rate measured before pulse wave velocity measurement, body mass index, and fasting glucose. Results are presented as difference in cf-PWV unit by each serum uric acid unit (mg/dl).

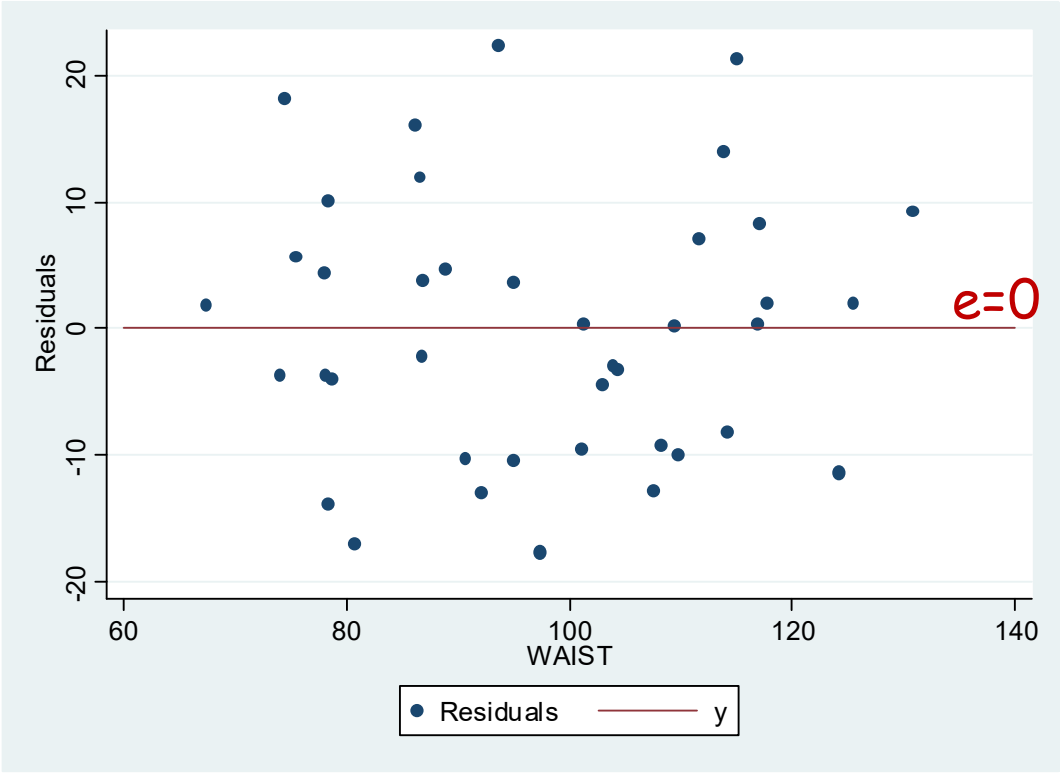
Assunti regressione lineare



Per ogni valore X, la corrispondente distribuzione di Y nella popolazione è Gaussiana con media situata sulla linea di regressione «vera» e varianza costante.

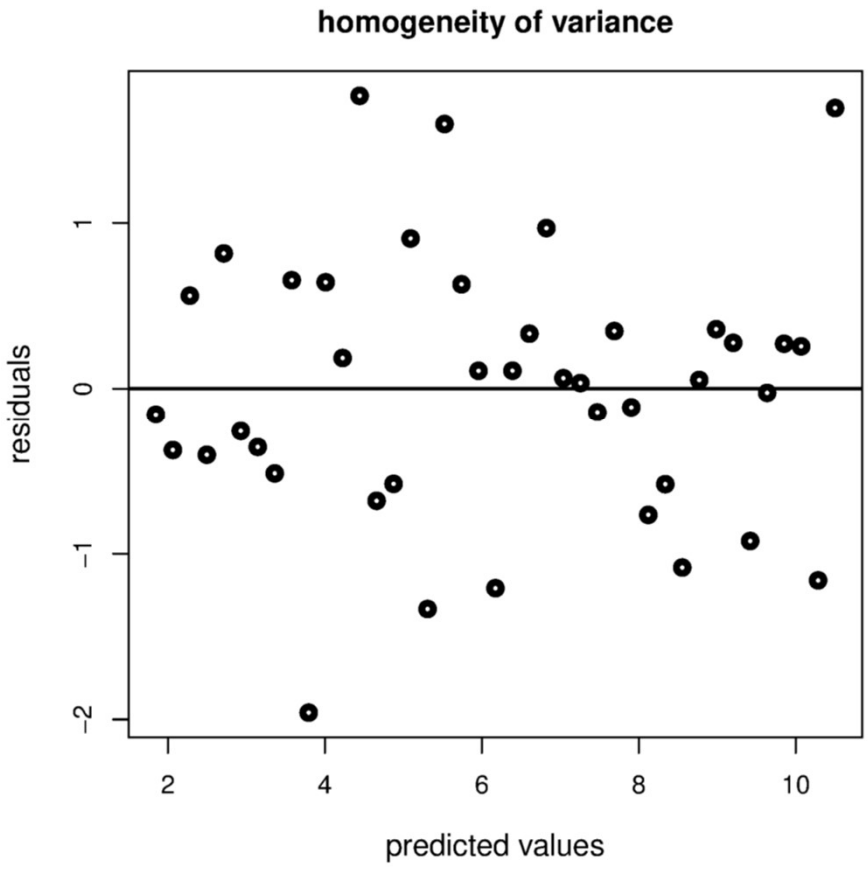
Grafico dei residui

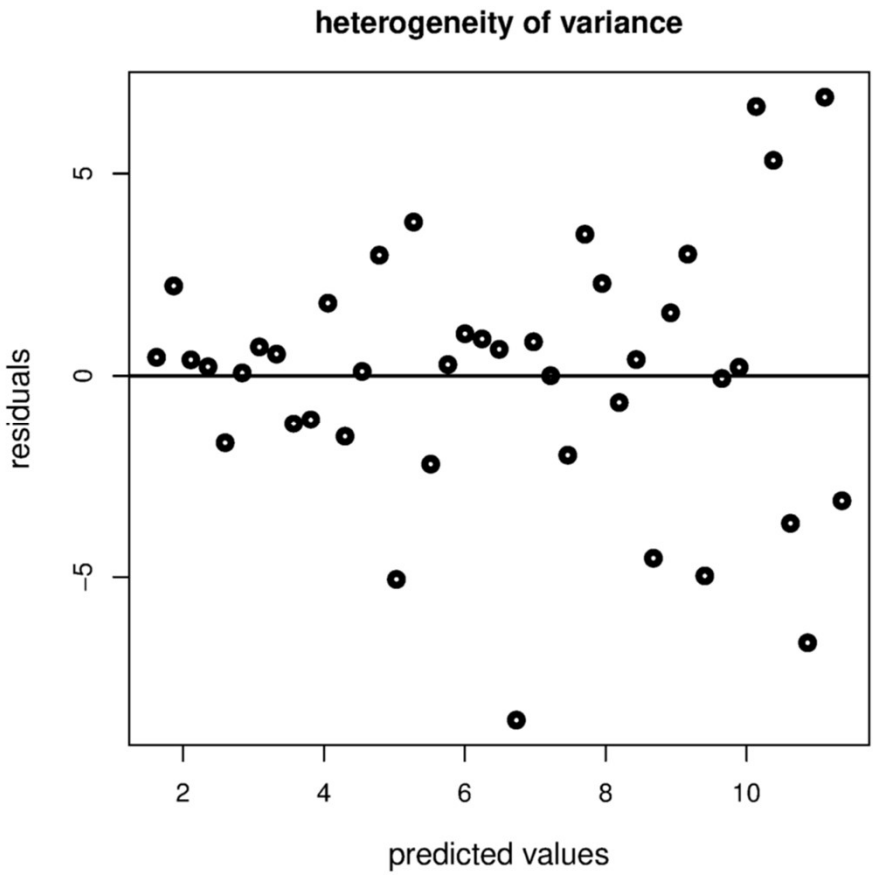
Gli assunti di normalità e omoschedasticità (varianza costante) possono esser valutati tramite il grafico dei residui $e_i = y_i - \hat{y}_i$



Quando gli assunti sono soddisfatti il grafico dovrebbe apparire così:

1) Una nuvola di punti senza tendenze
2) Simmetrico rispetto a e=0
3) Più punti vicino alla linea e=0
4) Stessa variabilità sopra e sotto la linea e=0 per ogni X

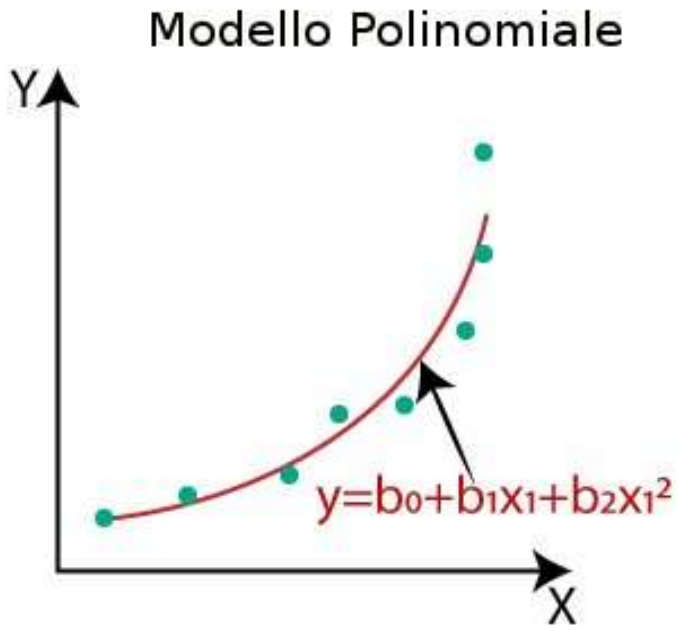

homogeneity of variance
residuals
predicted values
−2
−1
0
1
2
4
6
8
10


heterogeneity of variance
residuals
predicted values
−5
0
5
2
4
6
8
10

Regressione polinomiale

Abbiamo assunto che la relazione tra X e Y sia lineare ma non è necessariamente così.
L'ipotesi di linearità può esser valutata tramite polinomi

$$Y=\beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_j X_1^j + \ldots + \varepsilon$$



L’uso della retta di regressione per la previsione

Se gli assunti sono rispettati potrei cercare di «predire» la pressione sistolica in base al valore di circonferenza vita.

Nell’esempio:

Che valore di pressione sistolica mi attendo mediamente per un uomo con una circonferenza vita di 100 cm?

$\hat{y} = 69.6 + 0.51\,x$

$\hat{y} = 69.9 + 0.51 \cdot 100 = 121\ mmHg$

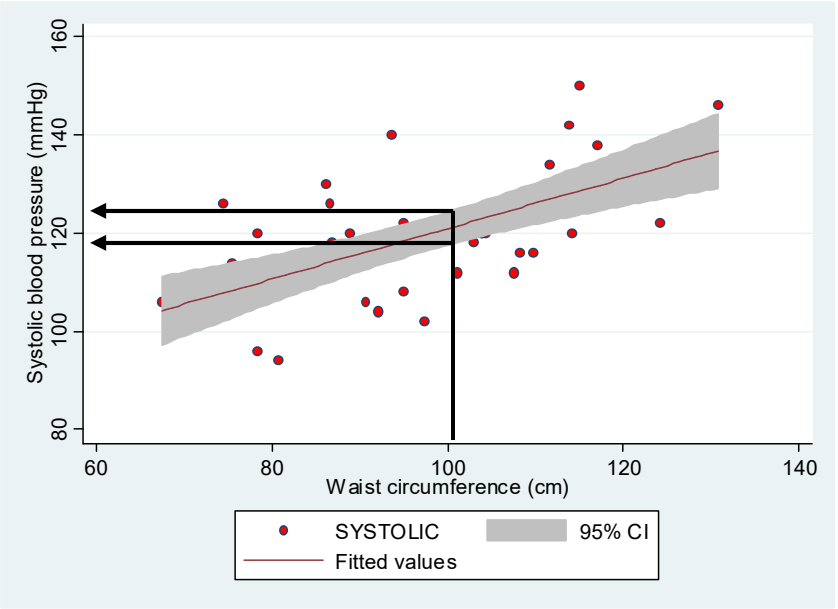
L’uso della retta di regressione per la previsione

Se gli assunti sono rispettati potrei cercare di «predire» la pressione sistolica in base al valore di circonferenza vita.

Nell’esempio:

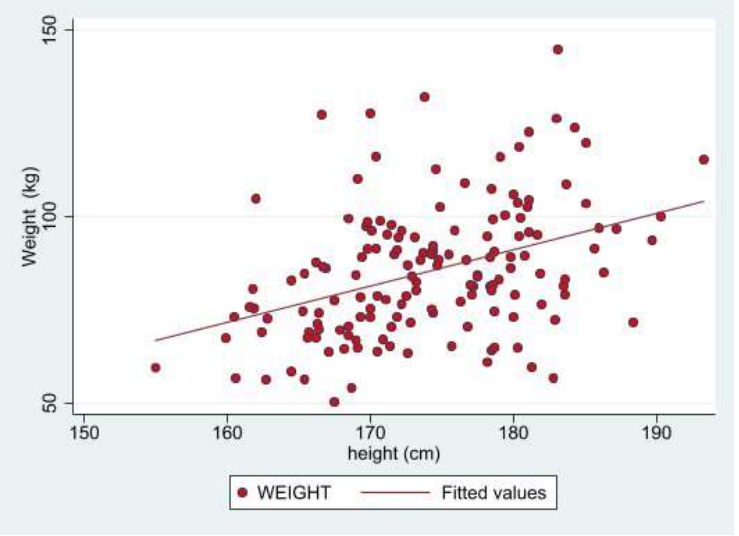
Che valore di pressione sistolica mi attendo per un uomo con una circonferenza vita di 100 cm?

$$\hat{y} = 69.9 + 0.51 \cdot 100 = 121\ mmHg$$



Esercizio

In un campione casule di 153 uomini si sono misurate le altezze (cm) e i pesi (kg). Le altezze variano da 155 a 193 cm con media uguale a 174.1 cm, e i pesi variano da 50 a 145 kg, con media uguale a 85.6 kg.
Il coefficiente di correlazione tra pesi e altezze è risultato 0.394 e i coefficienti della retta di regressione tra peso (variabile dipendente) e altezza (variabile indipendente) sono riportati in tabella:



	Coefficienti (IC95%)	Errore std	t	P-value
Intercetta	-85.1 (-149.1;-21.1)	32.4	-2.63	0.010
Altezza (cm)	0.98 (0.61;1.35)	0.19	5.27	<0.001

Trovare il miglior valore previsto di peso per un uomo alto 186 cm.

$$\hat{y} = -85.1 + 0.98 \cdot 186 = 97.2\ kg$$

Valutazione della bontà di adattamento

La valutazione della bontà di adattamento della curva si misura tramite il coefficiente di determinazione:

$SS_{tot} = \sum_i^n (y_i - \bar{y})^2$ devianza totale di Y

$SS_{tot} = \sum_i^n (y_i - \bar{y})^2 = \sum_i^n (y_i - \hat{y}_i)^2 + \sum_i^n (\hat{y}_i - \bar{y})^2$

SS_{errore} + $SS_{regressione}$

Varaibilità casuale — Variabilità di Y spiegata da X

$r^2 = \frac{\sum_i^n (\hat{y}_i - \bar{y})^2}{\sum_i^n (y_i - \bar{y})^2} = \frac{SS_{regressione}}{SS_{tot}}$ = Proporzione di varianza di Y spiegata da X

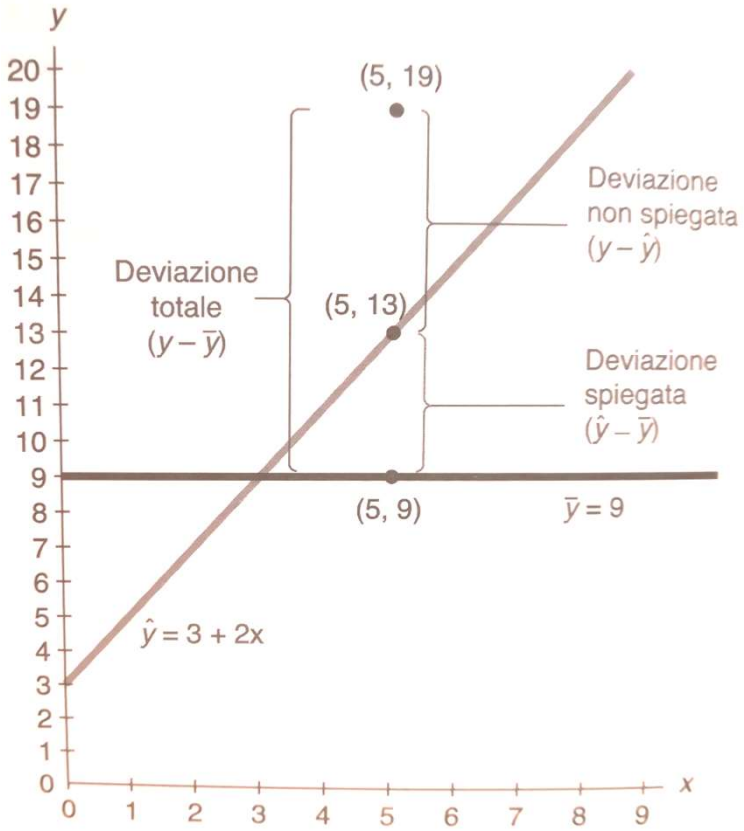
$0 \leq r^2 \leq 1$

$r^2 = \frac{variazione\ spiegata}{variazione\ totale}$

Valutazione della bontà di adattamento

La valutazione della bontà di adattamento della curva si misura tramite il coefficiente di determinazione:

$$r^2 = \frac{variazione\ spiegata}{variazione\ totale} \qquad 0 \leq r^2 \leq 1$$



Nota: nella regressione lineare semplice il valore di r^2 è semplicemente il quadrato del coefficiente di correlazione lineare (di Pearson).

Variabilità spiegata: r^2

Un valore di Y predetto non sarà necessariamente uguale al valore di Y osservato, perché in aggiunta a X, vi sono altri fattori che influiscono su Y, come fluttuazione aleatorie ed altre variabili non incluse nello studio.

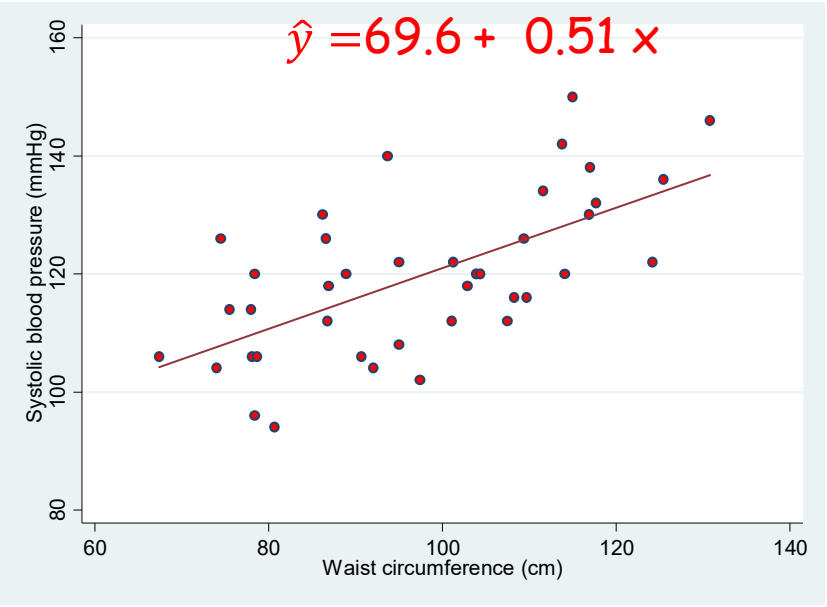
Il valore di r^2 rappresenta la proporzione di variabilità di Y spiegata dalla relazione lineare tra X e Y

Esempio: $r=0.78$ $r^2=0.6084$

Circa il 61% della variabilità del voto di esame è spiegata dalla sua relazione con il voto del compito a casa

Esempio

Esiste un'associazione tra circonferenza vita e pressione sanguigna sistolica?



	coefficient	SE	95%CI	T	P-value	r^2
WAIST	0.51	0.101	[0.312-0.708]	5.06	<0.001	0.4028

Il 40% della variabilità della pressione sistolica è spiegato dalla circonferenza vita, e il resto?

Esercizio

In un campione casuale di 153 uomini si sono misurate le altezze (cm) e i pesi (kg). Le altezze variano da 155 a 193 cm con media uguale a 174.1 cm, e i pesi variano da 50 a 145 kg, con media uguale a 85.6 kg.
Il coefficiente di correlazione tra pesi e altezze è risultato 0.394 e i coefficienti della retta di regressione tra peso (variabile dipendente)
e altezza (variabile indipendente) sono riportati in tabella:

	Coefficienti (IC95%)	**Errore std**	**t**	**P-value**
intercetta	-85.1 (-149.1;-21.1)	32.4	-2.63	0.010
Altezza (cm)	0.98 (0.61;1.35)	0.19	5.27	<0.001

1) Trovare il coefficiente di determinazione.
2) Quali informazioni pratiche fornisce il coefficiente di determinazione?

1) $r^2 = 0.155$
2) 15.5 % della variazione del peso è spiegata dalla sua relazione (lineare) con l'altezza (84.5% è da imputare ad altri fattori).