STATISTICA

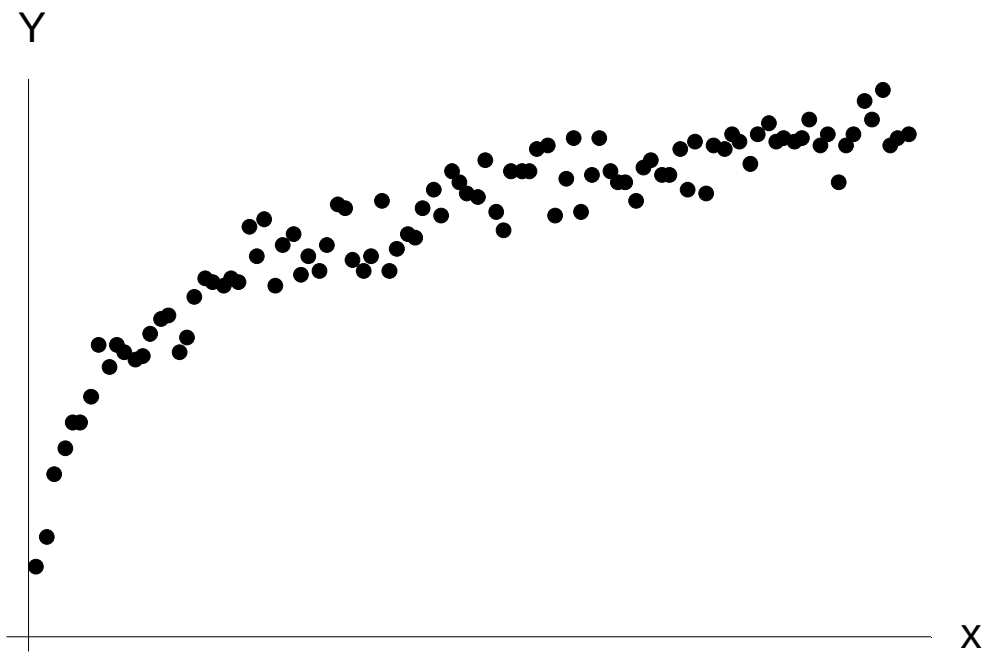
Regressione-1

Il problema

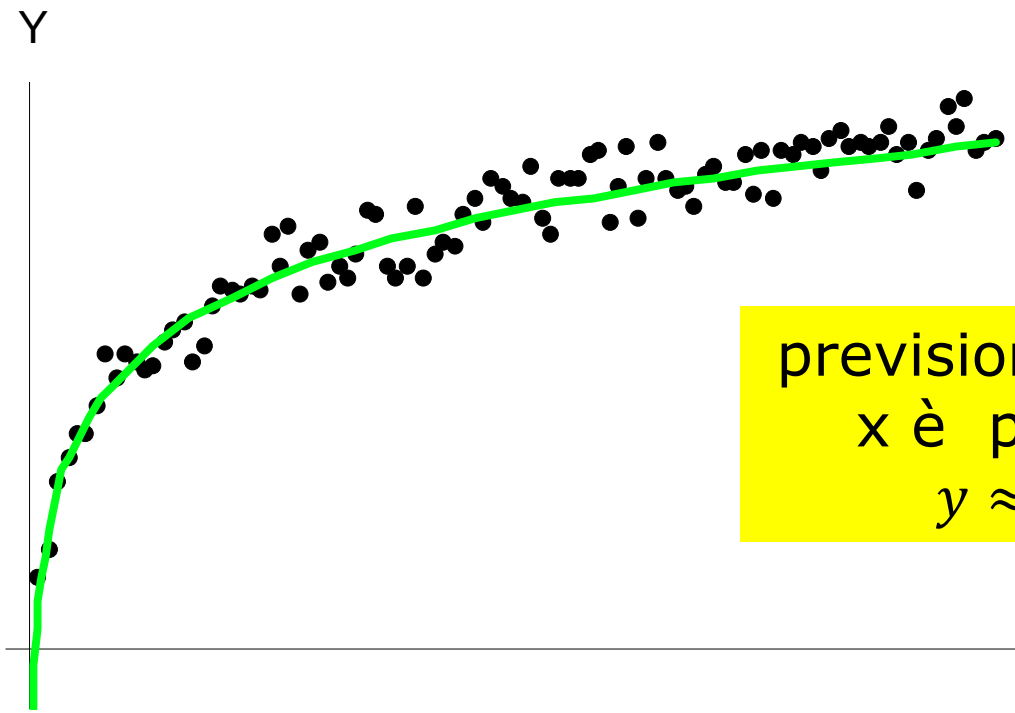
	Peso	Altezza
1	118	64.5
2	151	72.5
3	143	73.3
4	172	68.8
5	147	65.0
6	146	69.0
⋮	⋮	⋮
38	139	64.5
39	148	74.0
40	179	75.5

- I due fenomeni sono collegati?
- Se aumenta l'altezza, aumenta il peso?
- Chiedersi il viceversa ha senso?

Obiettivo generale



Obiettivo generale

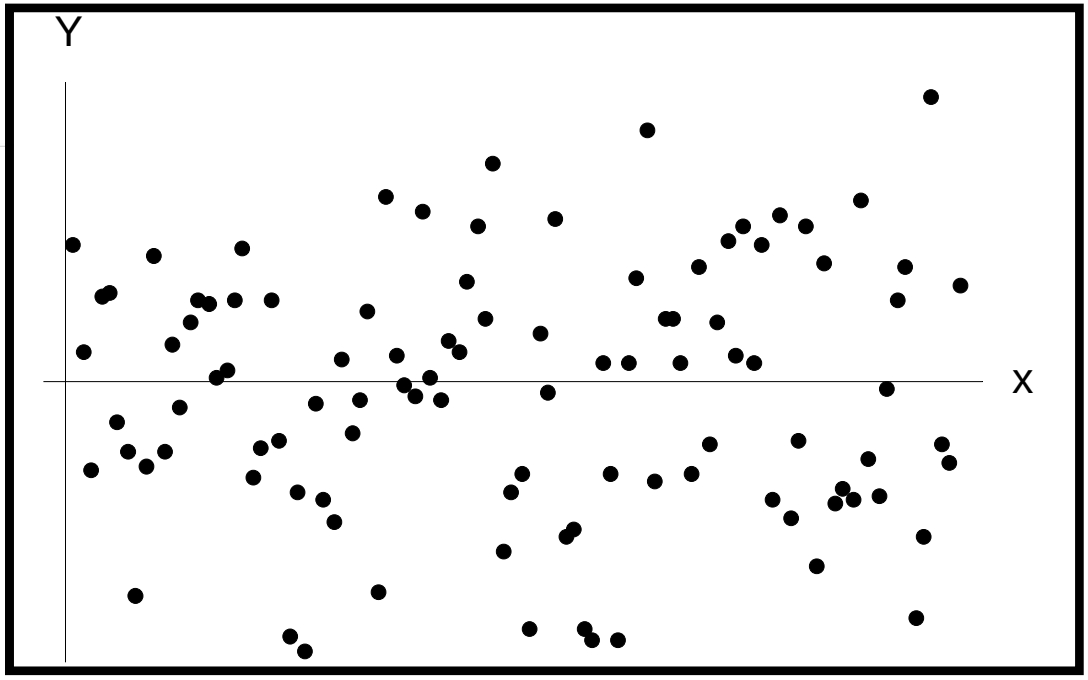


previsione di y da x è possibile

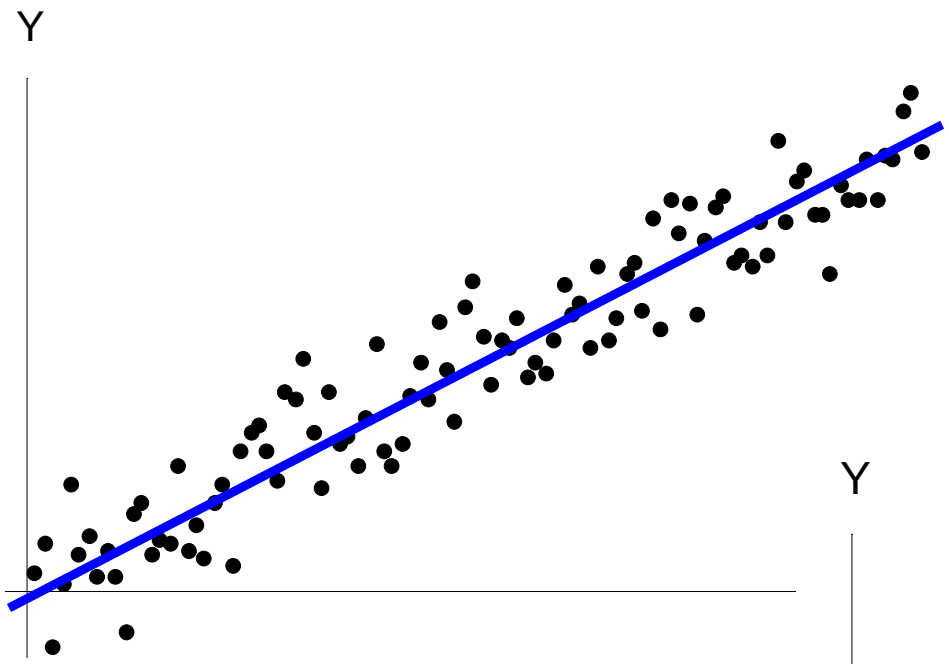
$y \approx f(x)$

Obiettivo generale

previsione di y da x
non è possibile

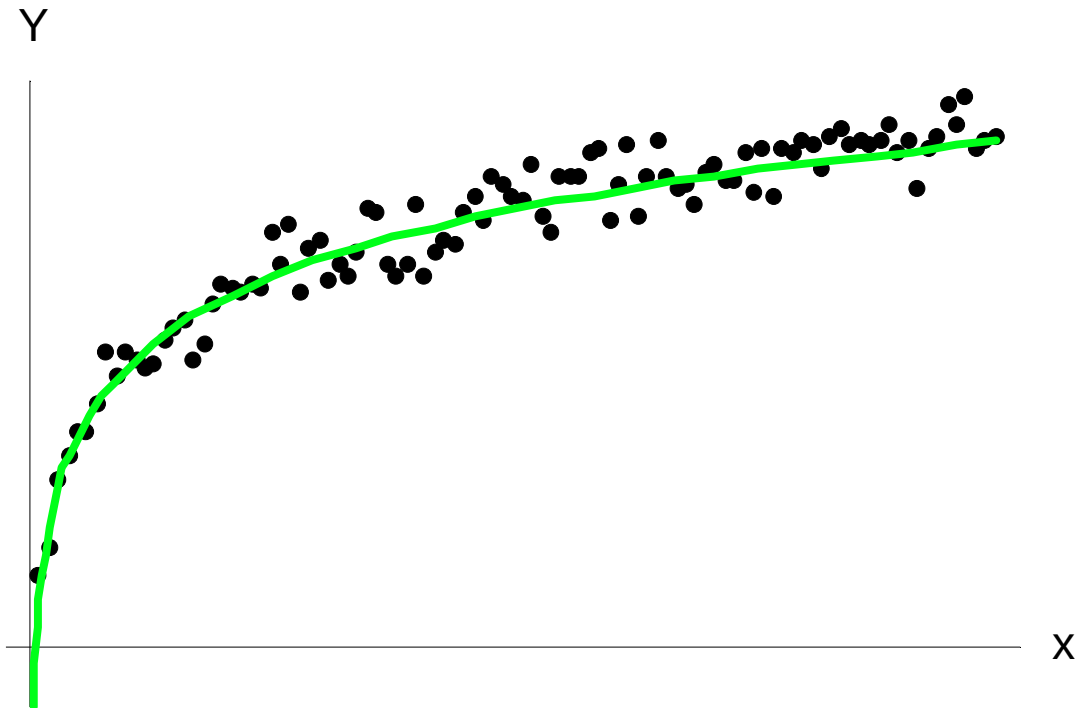


Obiettivo generale

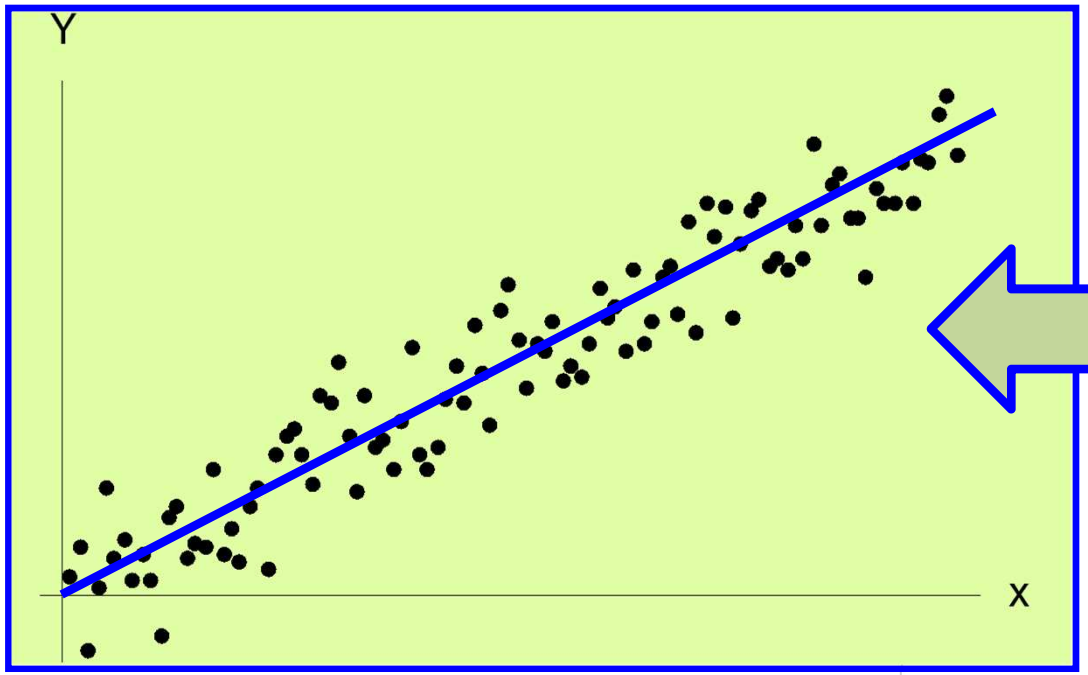


Un **modello** per la previsione di una variabile Y che dipende da un'altra variabile, X:

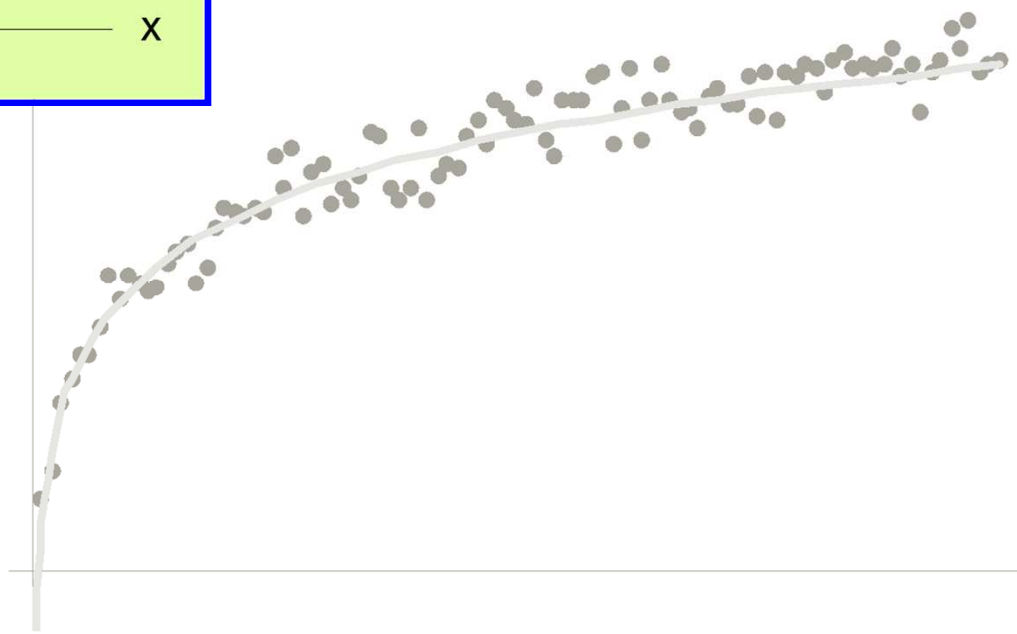
$$f(x) \quad \& \quad y \approx f(x)$$



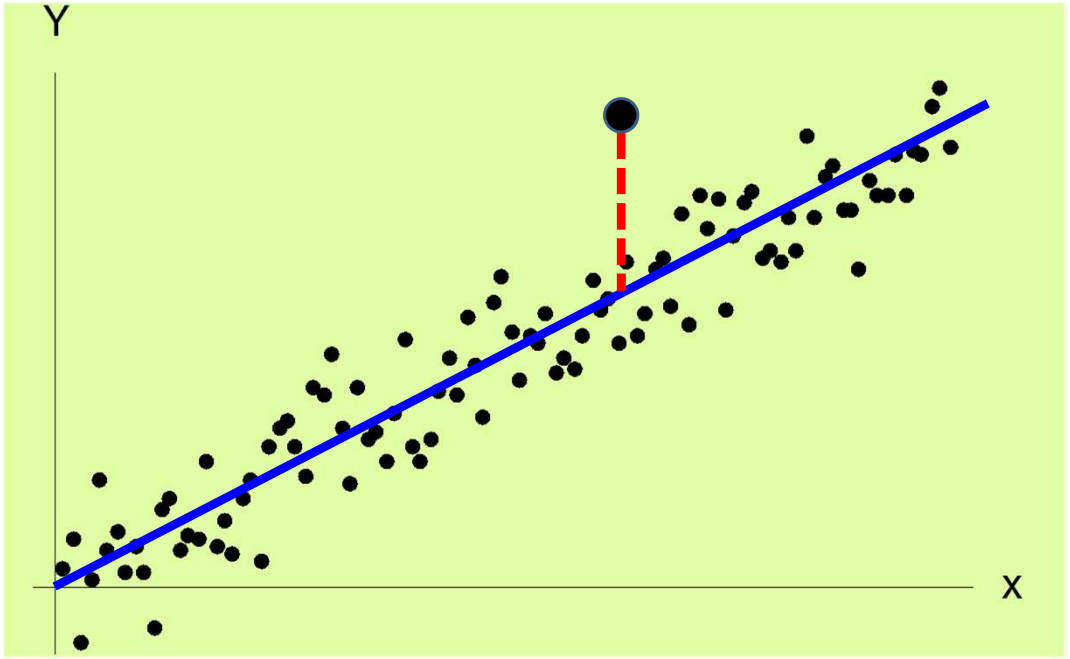
Obiettivo specifico



Il modello della **regressione lineare semplice**:



Obiettivo specifico



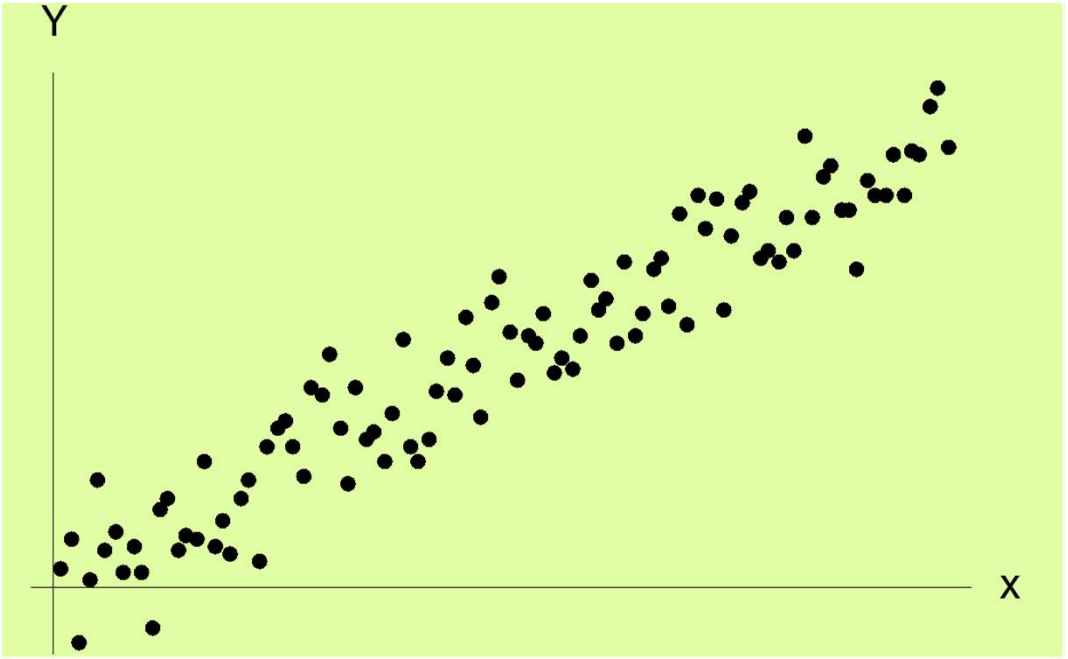
Il modello della **regressione lineare semplice**:

$$f(x) = a + bx$$
$$\approx \Leftrightarrow \ \varepsilon_i \sim N(0, \sigma^2)$$
$$\varepsilon_i \text{ indipendenti}$$

$$Y_i = a + bx_i + \varepsilon_i$$

x_i punti del disegno
Y_i variabile casuale

Regressione lineare



1. GRAFICO di dispersione

2. OPPORTUNI INDICI STATISTICI

La covarianza

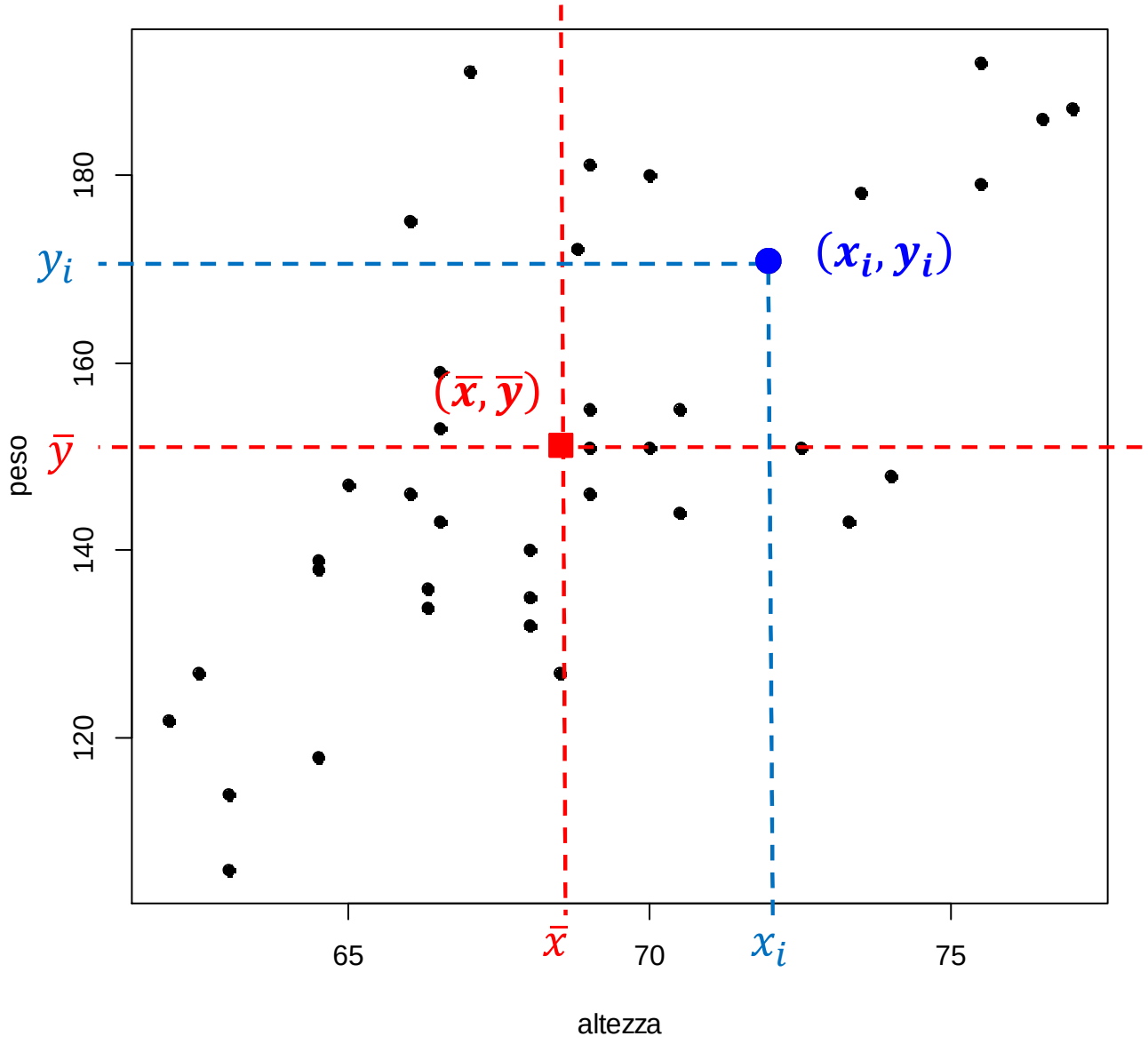
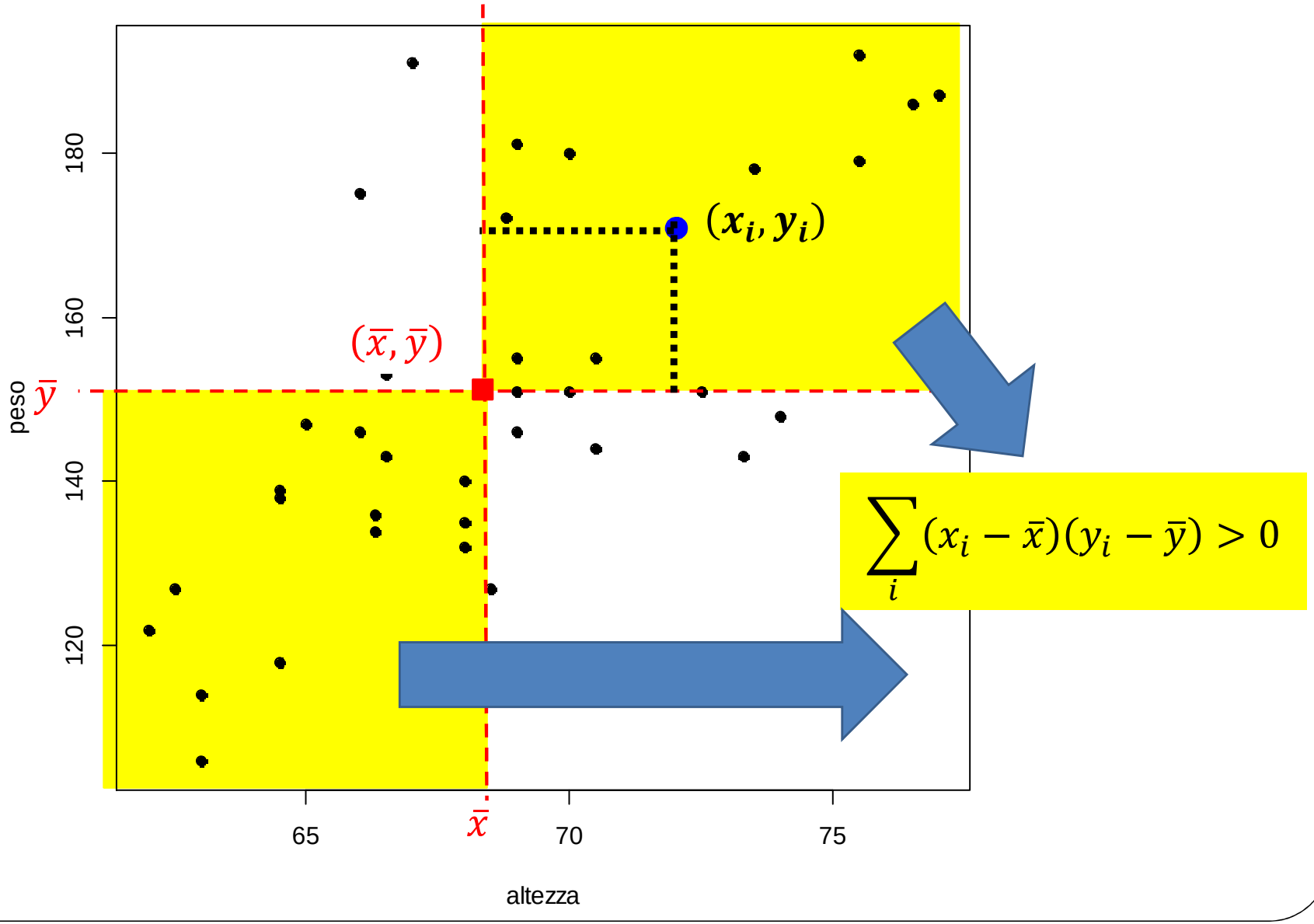
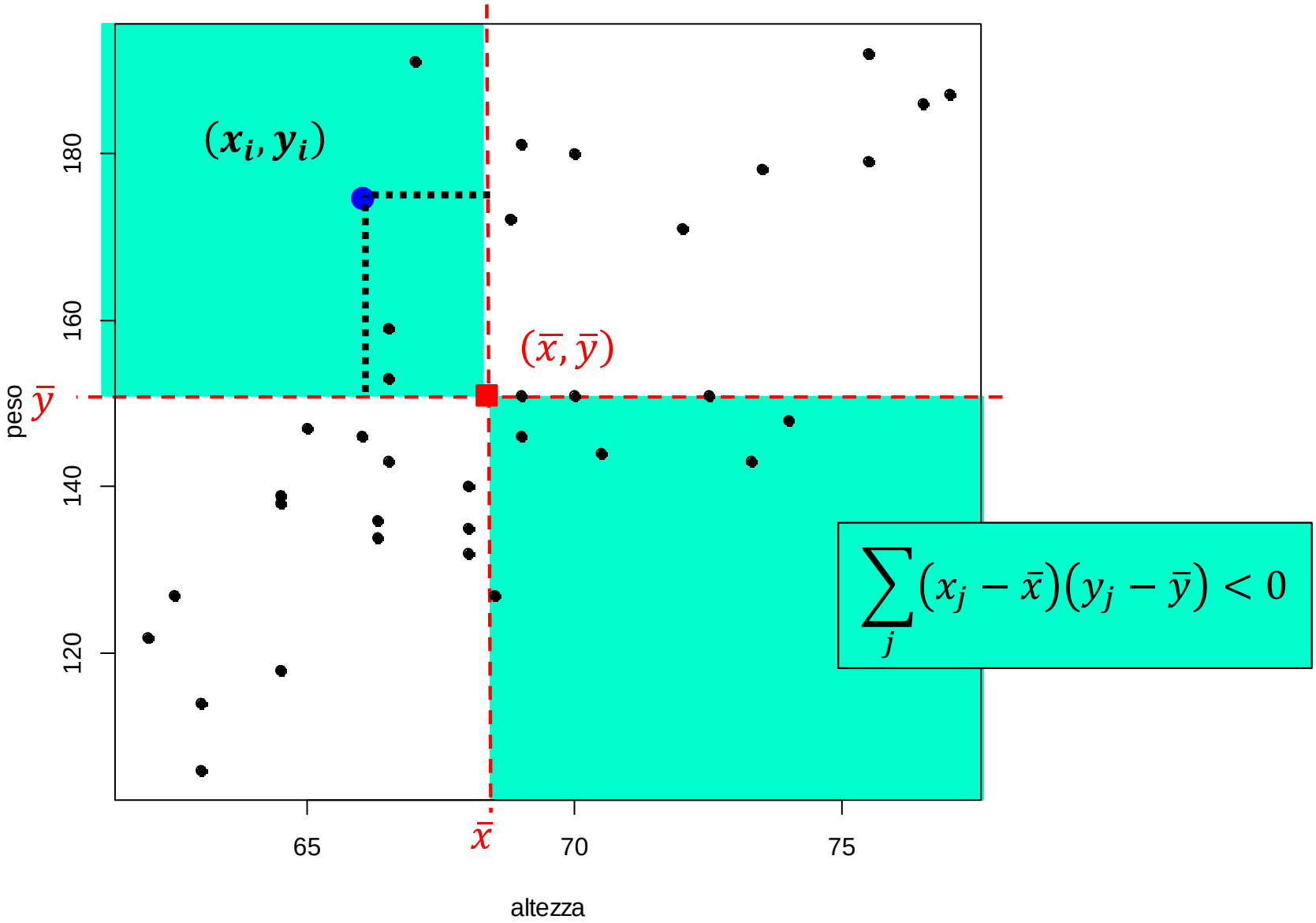


grafico di dispersione attorno al baricentro (centroide)

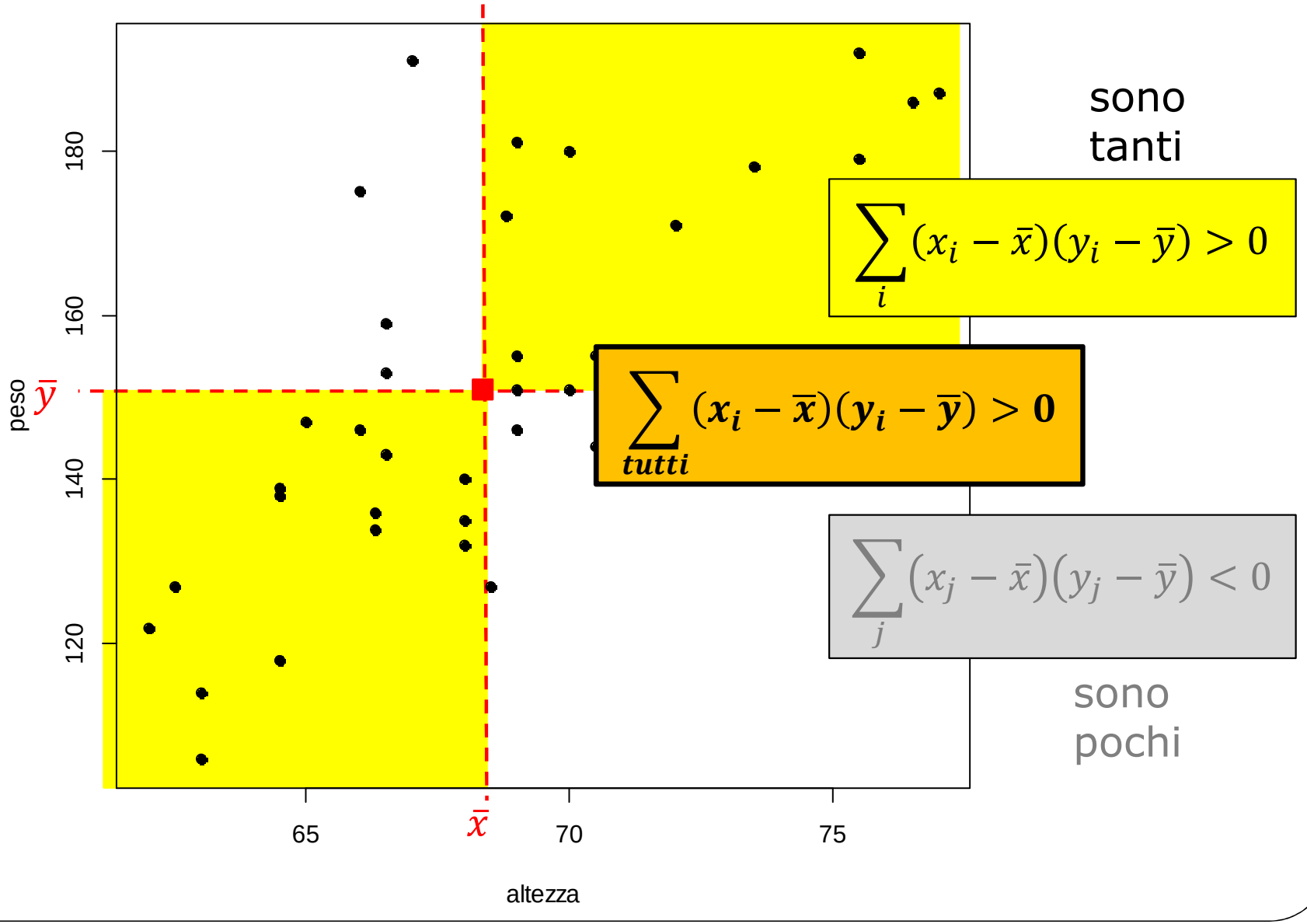
La covarianza



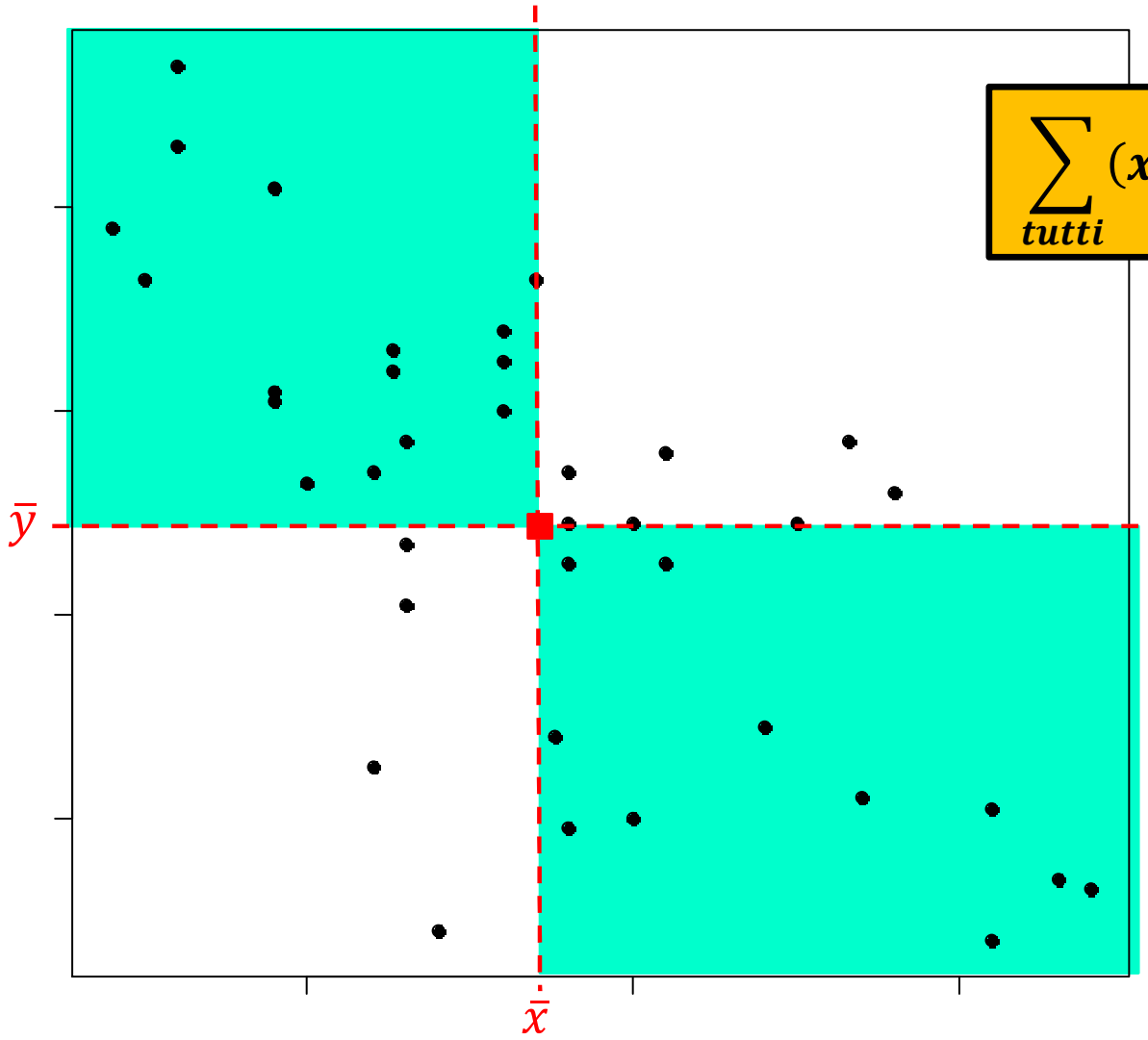
La covarianza



La covarianza

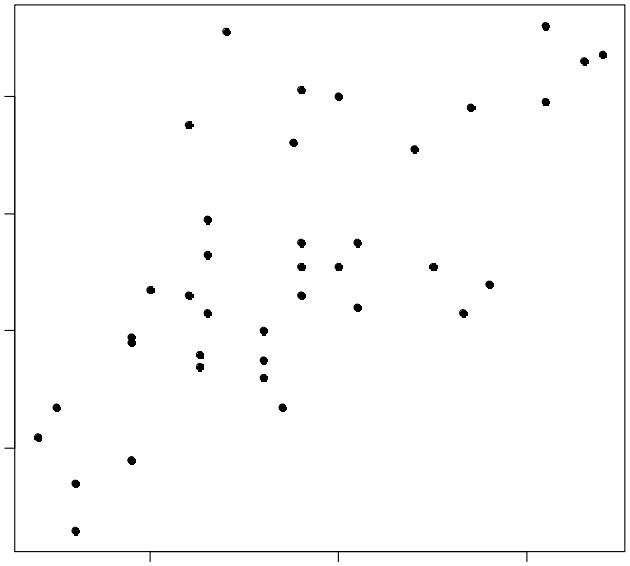


La covarianza



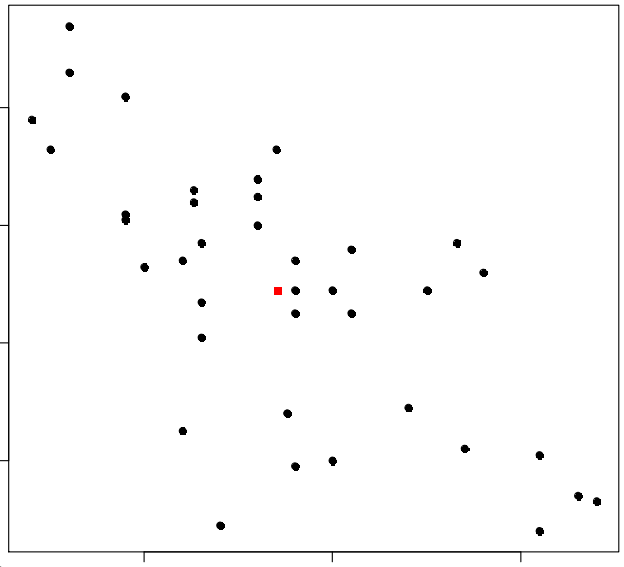
$$\sum_{tutti} (x_i - \bar{x})(y_i - \bar{y}) < 0$$

La covarianza



$$\boldsymbol{\sigma_{xy} = cov(x,y) = \frac{1}{n}\sum_i (x_i - \overline{x})(y_i - \overline{y})}$$

$cov(x,y) > 0 \iff$ quando x tende a crescere y fa lo stesso

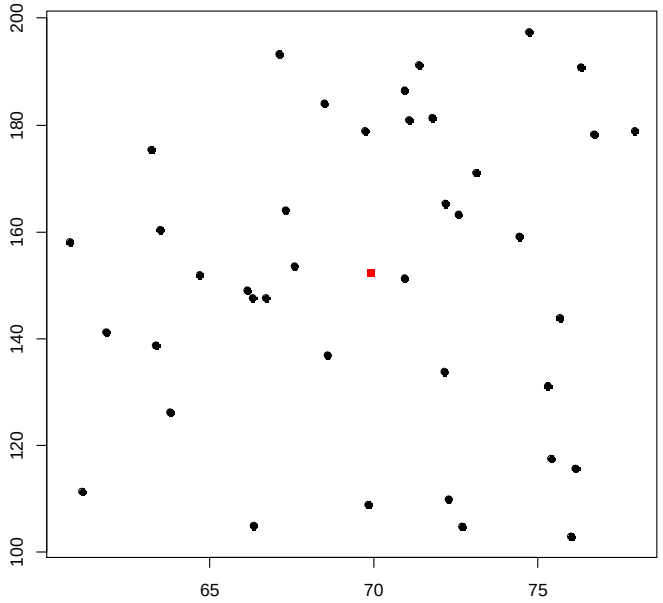


$cov(x,y) < 0 \iff$ quando x tende a crescere y tende a decrescere

$$\boldsymbol{\sigma_{xy} = cov(x,y) = \left[\frac{1}{n}\sum_i x_i y_i\right] - \overline{x} \times \overline{y}}$$

La covarianza

$$\sigma_{xy} = cov(x,y) = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$



$$cov(x,y) = 0 \quad \Leftrightarrow \quad \left[\frac{1}{n}\sum_i \boldsymbol{x_i y_i}\right] = \bar{\boldsymbol{x}} \times \bar{\boldsymbol{y}}$$

$$\sigma_{xy} = cov(x,y) = \left[\frac{1}{n}\sum_i x_i y_i\right] - \bar{x} \times \bar{y}$$

La correlazione lineare

$$\rho_{xy} = r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$\boldsymbol{\sigma_{xy} = cov(x,y) = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})}$$

$$\boldsymbol{-1 \leq \rho_{xy} \leq 1}$$

$\rho_{xy} > 0$ $\quad cov(x,y) > 0 \iff$ quando x tende a crescere y fa lo stesso

$\rho_{xy} < 0$ $\quad cov(x,y) < 0 \iff$ quando x tende a crescere y tende a decrescere

$\rho_{xy} = 0$ $\quad cov(x,y) = 0$

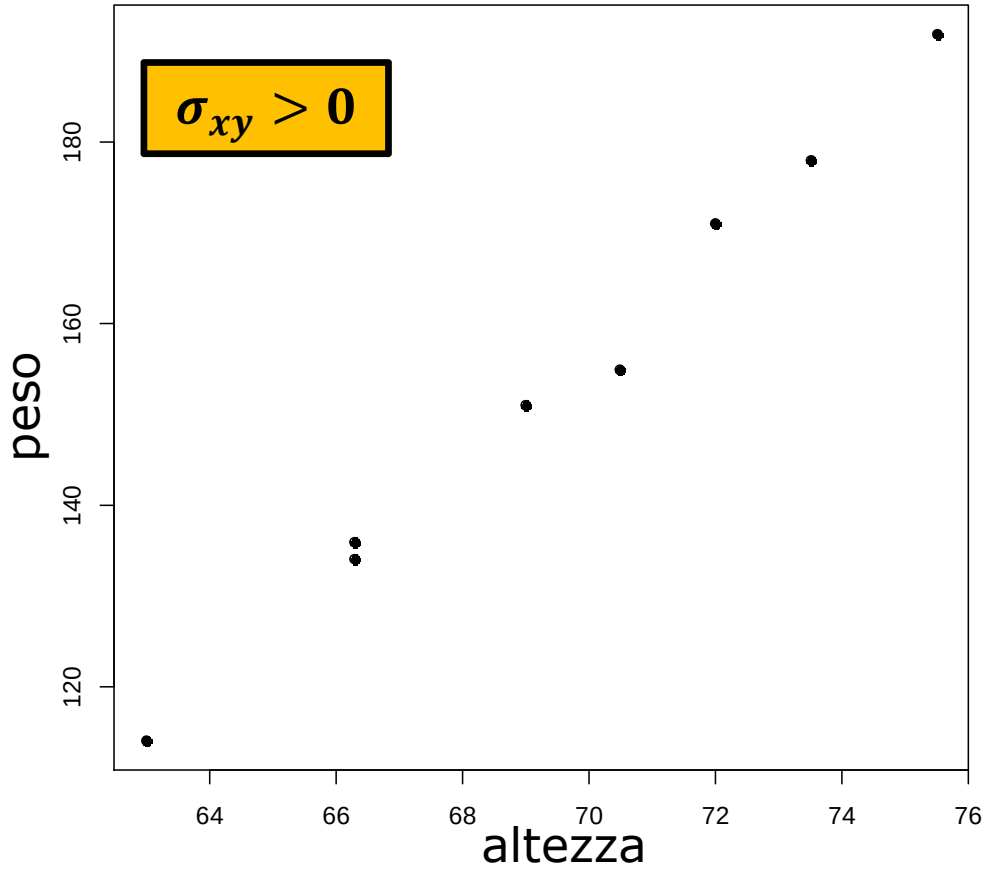
coeff. di correlazione lineare

$$\sigma_{xy} = cov(x,y) = \left[\frac{1}{n}\sum_i x_i y_i\right] - \bar{x} \times \bar{y}$$

$$-\sigma_x \sigma_y \leq \sigma_{xy} \leq \sigma_x \sigma_y$$

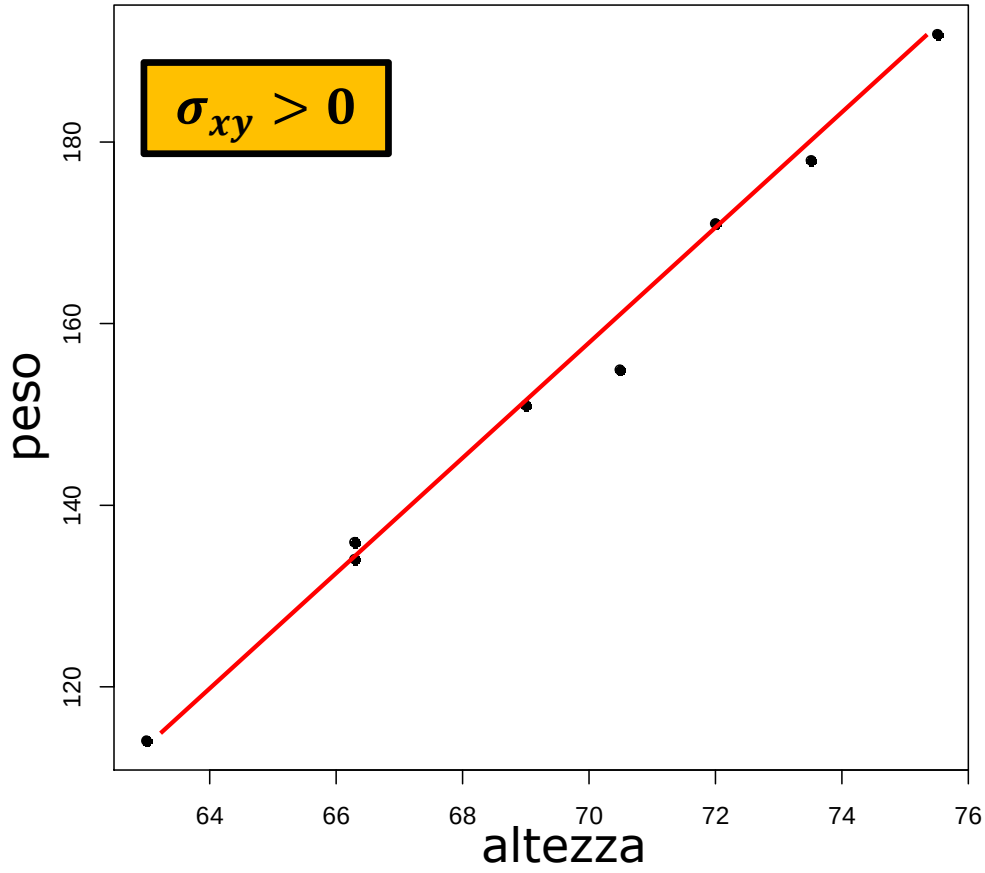
Esempio 1

Alt. (x)	66.3	70.5	73.5	66.3	69.0	63.0	72.0	75.5
Peso (y)	134	155	178	136	151	114	171	192



Esempio 1

Alt. (x)	66.3	70.5	73.5	66.3	69.0	63.0	72.0	75.5
Peso (y)	134	155	178	136	151	114	171	192



Esempio 1

Alt. (x)	66.3	70.5	73.5	66.3	69.0	63.0	72.0	75.5
Peso (y)	134	155	178	136	151	114	171	192
$x_i y_i$	8884.2	10927.5	13083.0	9016.8	10419.0	7182.0	12312.0	14496.0

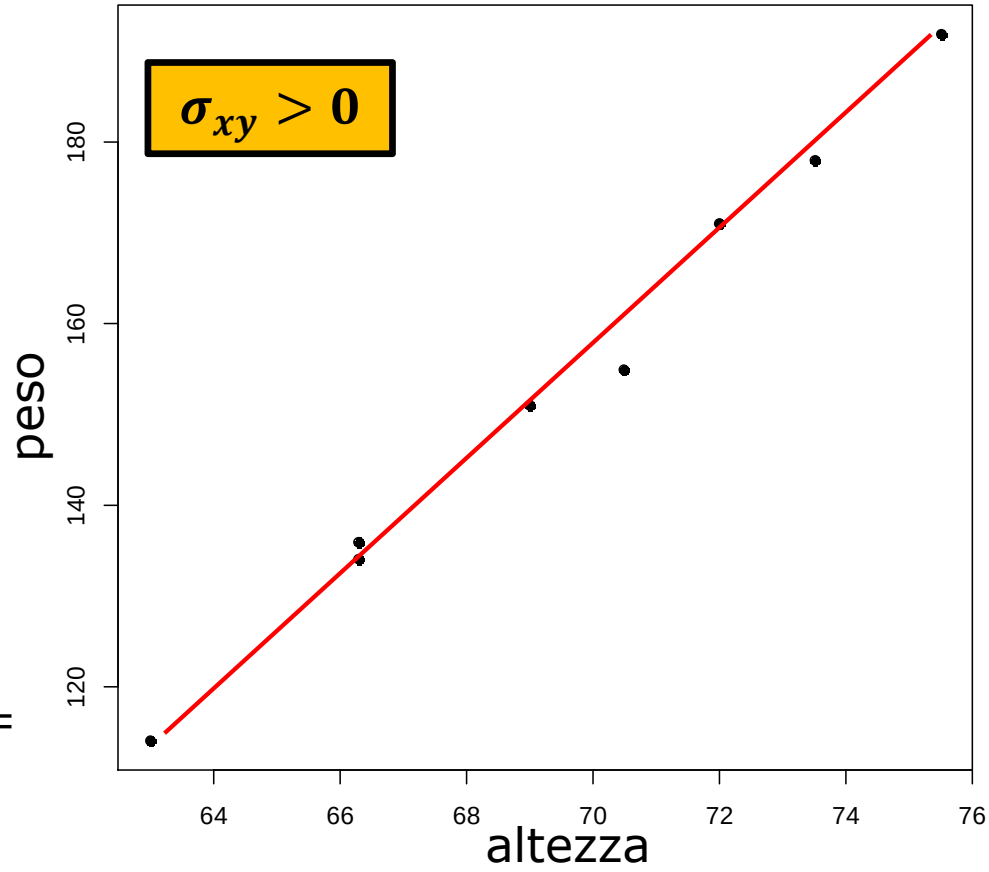
$$n = 8$$

$$\bar{x} = 69.51 \qquad \sigma_x = 3.908784$$
$$\bar{y} = 153.875 \qquad \sigma_y = 24.09065$$

$$\frac{1}{8}\sum_{i=1}^{8} x_i y_i = \frac{86320.5}{8} = 10790.06$$

$$\sigma_{xy} = 10790.06 - 69.51 \times 153.875 =$$
$$= 94.21125$$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{94.21125}{3.908784 \times 24.09065} =$$
$$= 1!$$



(*Contro*)Esempio

X	-2	-1	0	1	2
Y	4	1	0	1	4
XY	-8	-1	0	1	8

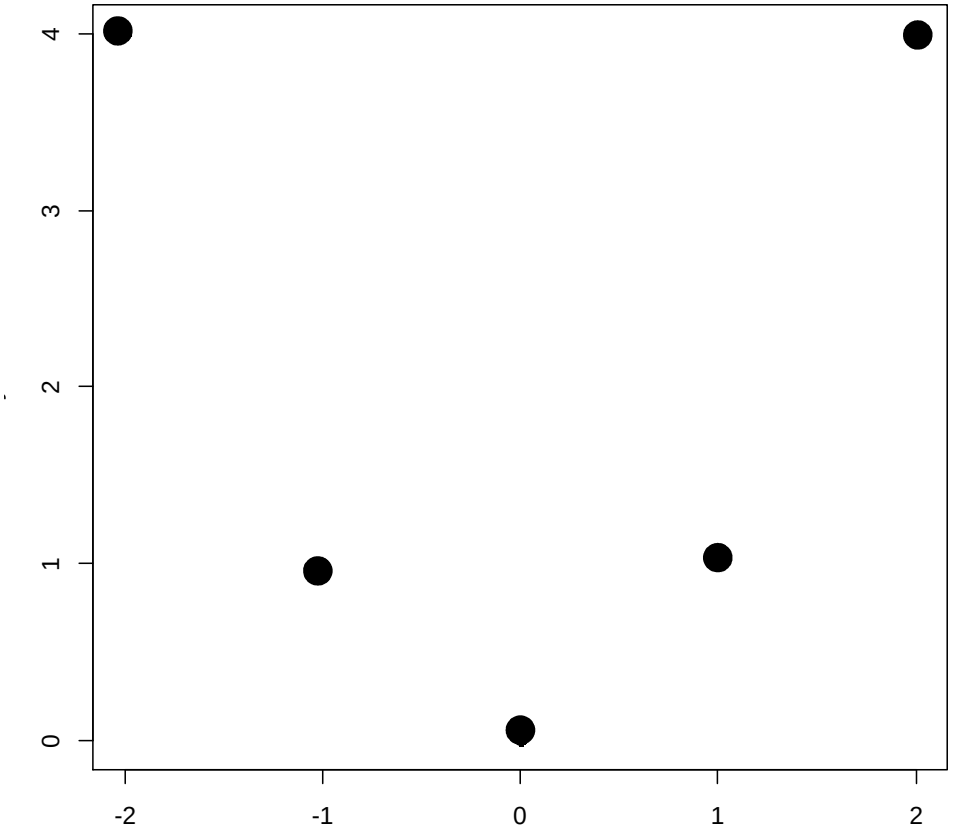
$$n = 5$$

$$\bar{x} = 0$$

$$\bar{y} = 2$$

$$\frac{1}{5}\sum_i x_i y_i = 0!$$

$$\begin{cases} \boldsymbol{\sigma_{xy} = 0} \\ \boldsymbol{\rho_{xy} = 0} \end{cases}$$



(*Contro*)Esempio

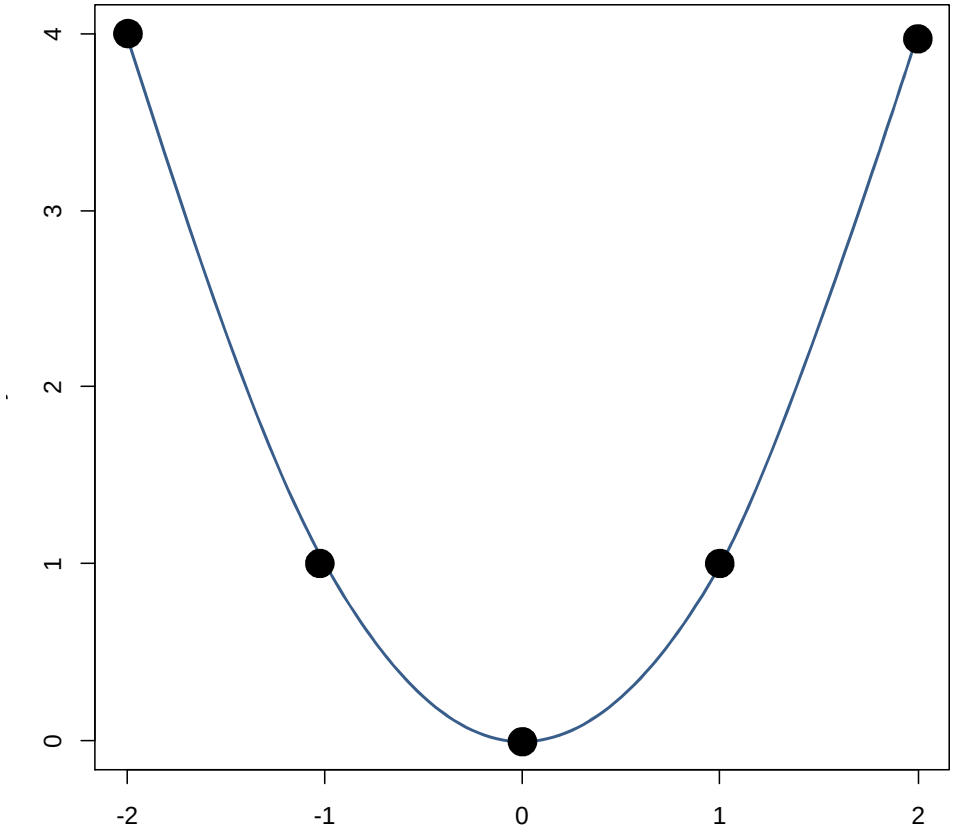
X	-2	-1	0	1	2
Y	4	1	0	1	4

$\sigma_{xy} = 0$

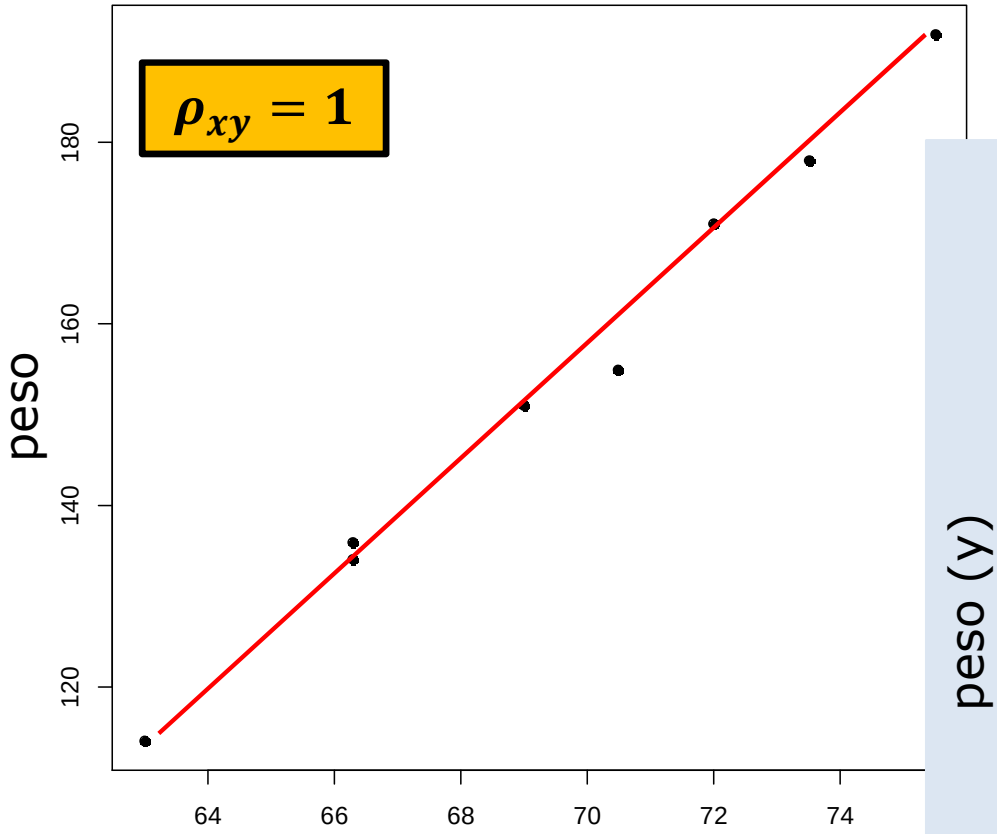
$\rho_{xy} = 0$

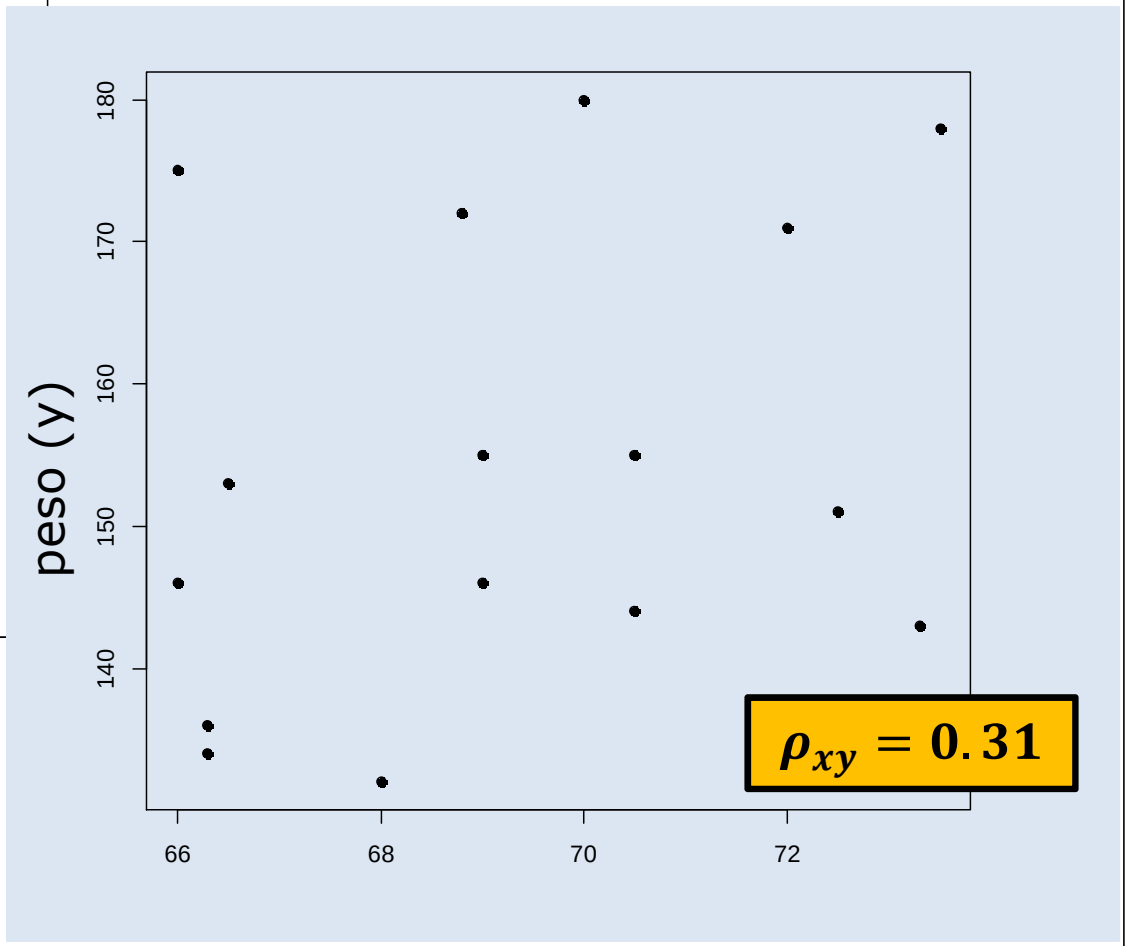
assenza di correlazione **lineare**…

ma chiara correlazione quadratica!



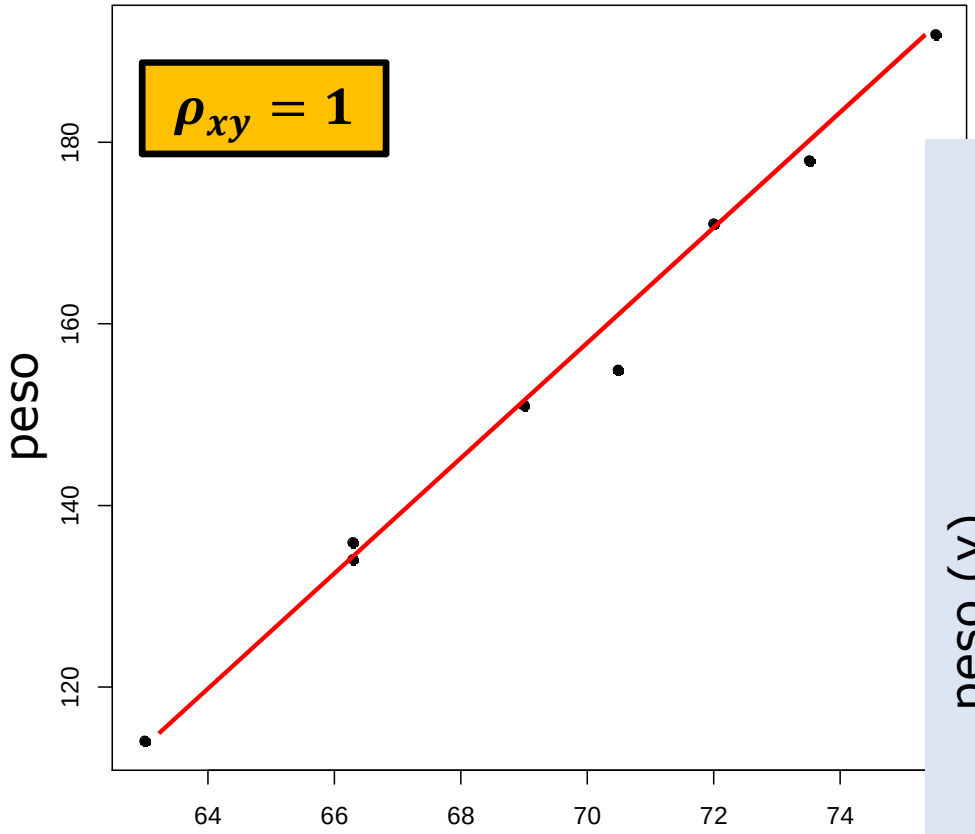
La retta di regressione



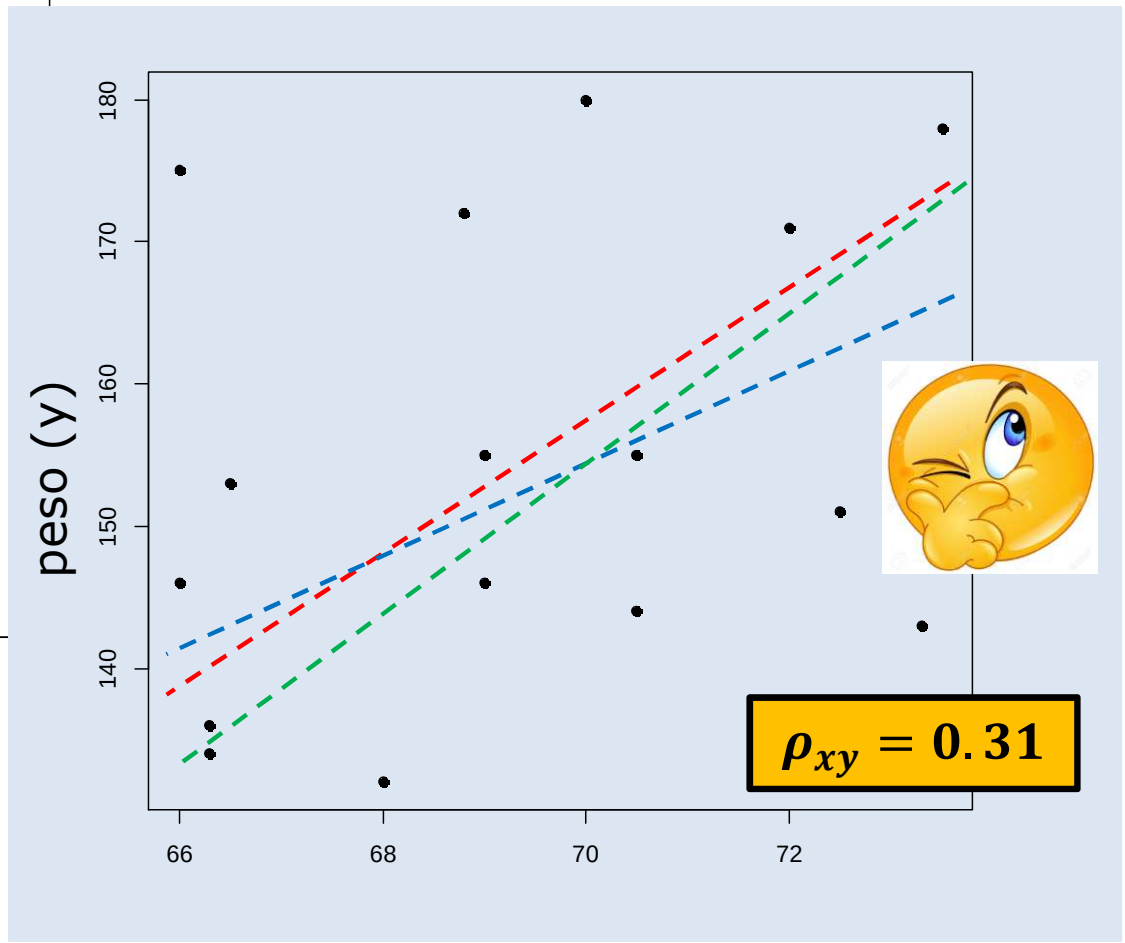


altezza (x)

La retta di regressione



Qual è la retta che passa più vicino a tutti i punti, più o meno?



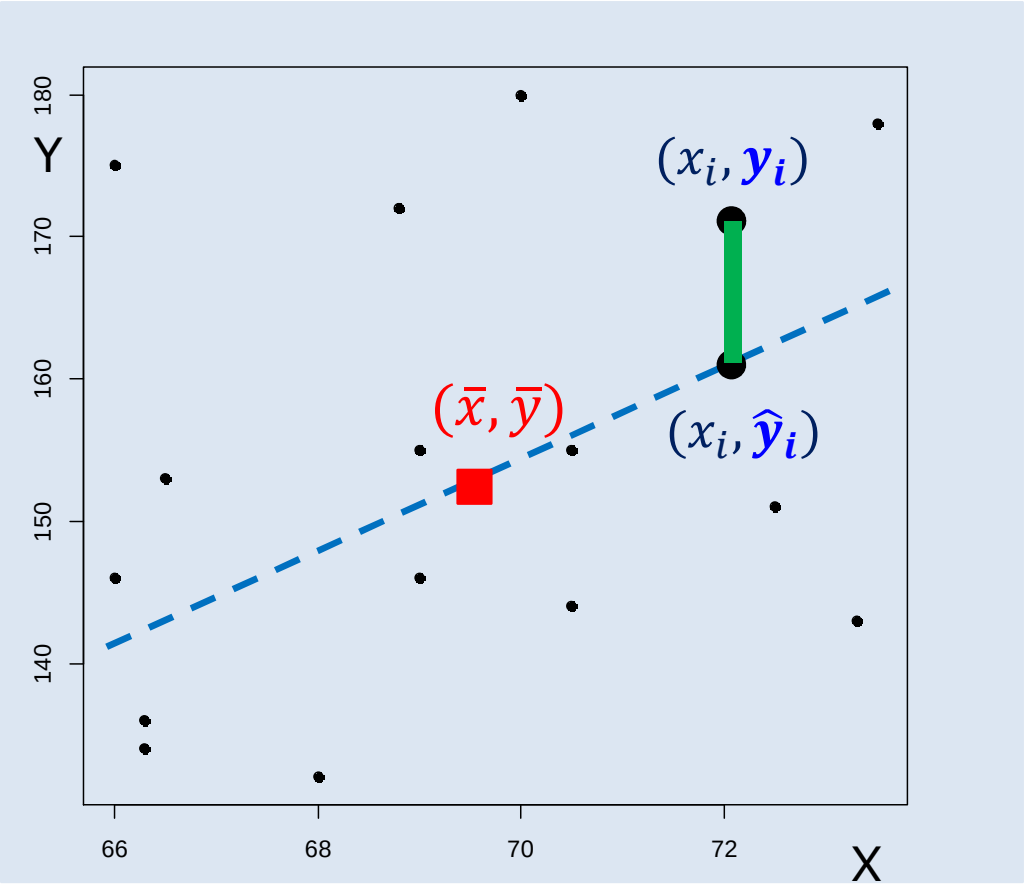
La retta di regressione

Di tutte le rette $\boldsymbol{y = a + bx}$ quella che passa più vicino a tutti i punti, nel senso dei *minimi quadrati,* è quella con coeff. $\hat{a}$ e $\hat{b}$ che rendono minima la quantità:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}[y_i - (a + bx_i)]^2$$

$$\hat{b} = \frac{\sigma_{xy}}{\sigma^2{}_x} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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



Esempio 1, cont.

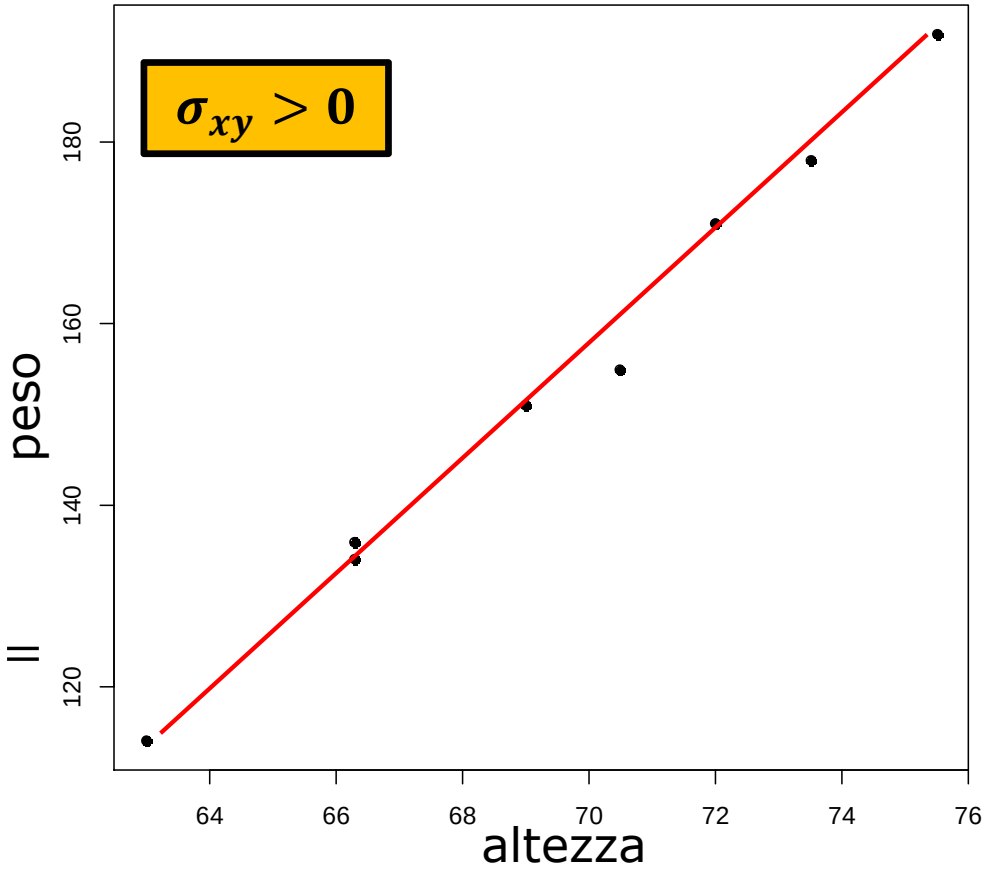
Alt. (x)	66.3	70.5	73.5	66.3	69.0	63.0	72.0	75.5
Peso (y)	134	155	178	136	151	114	171	192
$x_i y_i$	8884.2	10927.5	13083.0	9016.8	10419.0	7182.0	12312.0	14496.0

$$n = 8$$

$$\bar{x} = 69.51 \qquad \sigma_x = 3.908784$$
$$\bar{y} = 153.875 \qquad \sigma_y = 24.09065$$
$$\sigma_{xy} = 94.21125$$
$$\rho_{xy} = 1$$

$$\hat{b} = \frac{\sigma_{xy}}{{\sigma_x}^2} = \frac{94.21125}{15.27859} = 6.166$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 153.875 - 6.166 \times 69.51 = \\ = -274.724$$



Esempio 2

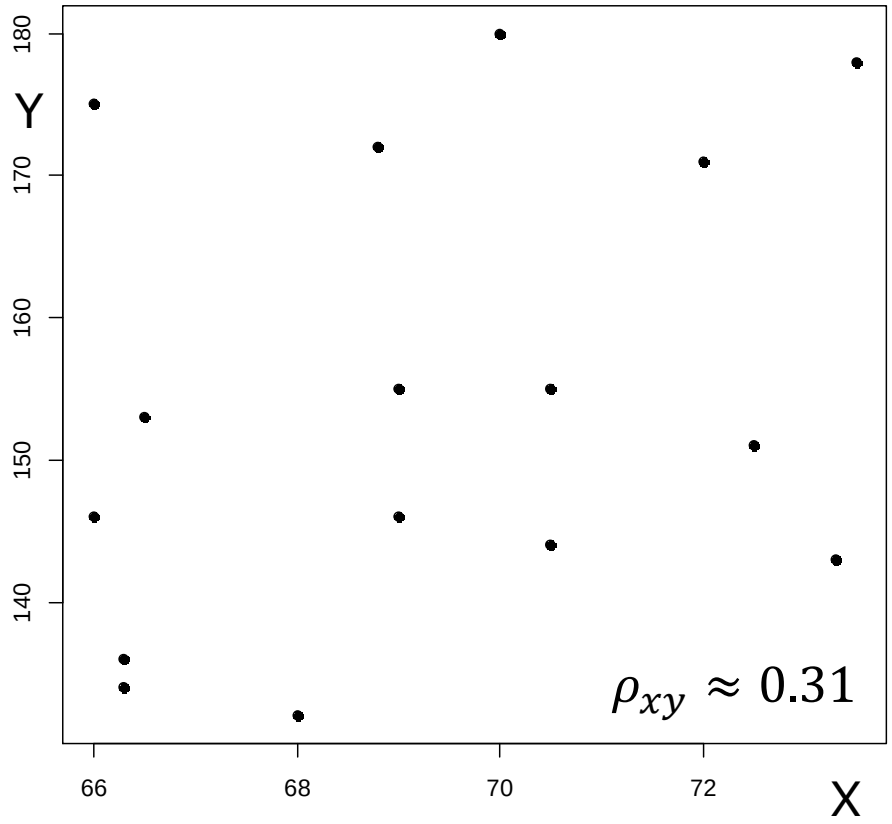
X	72.5	73.3	68.8	69	66	66.3	69	70.5	66	73.5	66.3	70	68	72	66.5	70.5
Y	151	143	172	146	175	134	155	155	146	178	136	180	132	171	153	144

$n = 16$

$\bar{x} = 69.2625$
$\bar{y} = 154.4375$

$\sigma_x = 2.552664$
$\sigma_y = 15.55622$

$\sigma_{xy} = 12.27266$



Esempio 2

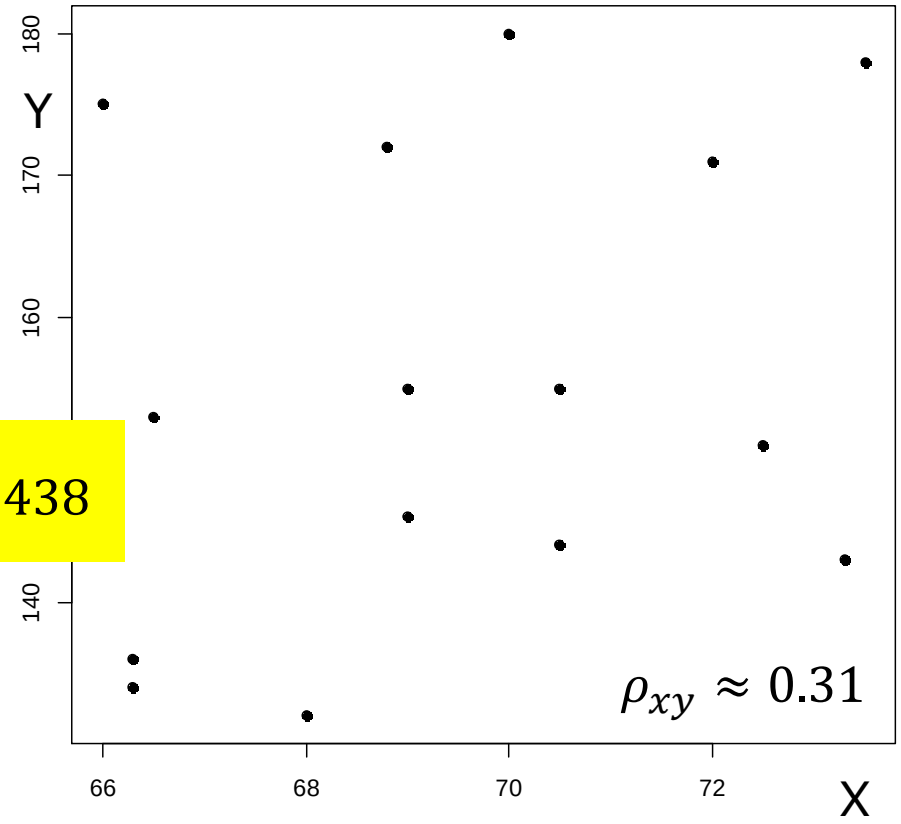
X	72.5	73.3	68.8	69	66	66.3	69	70.5	66	73.5	66.3	70	68	72	66.5	70.5
Y	151	143	172	146	175	134	155	155	146	178	136	180	132	171	153	144

$n = 16$

$\bar{x} = 69.2625$
$\bar{y} = 154.4375$

$\sigma_x = 2.552664$
$\sigma_y = 15.55622$

$\sigma_{xy} = 12.27266$ ⇒ $\hat{b} = \frac{12.27266}{2.552664^2} = 1.883438$



Esempio 2

X	72.5	73.3	68.8	69	66	66.3	69	70.5	66	73.5	66.3	70	68	72	66.5	70.5
Y	151	143	172	146	175	134	155	155	146	178	136	180	132	171	153	144

$n = 16$

$\bar{x} = 69.2625$

$\bar{y} = 154.4375$

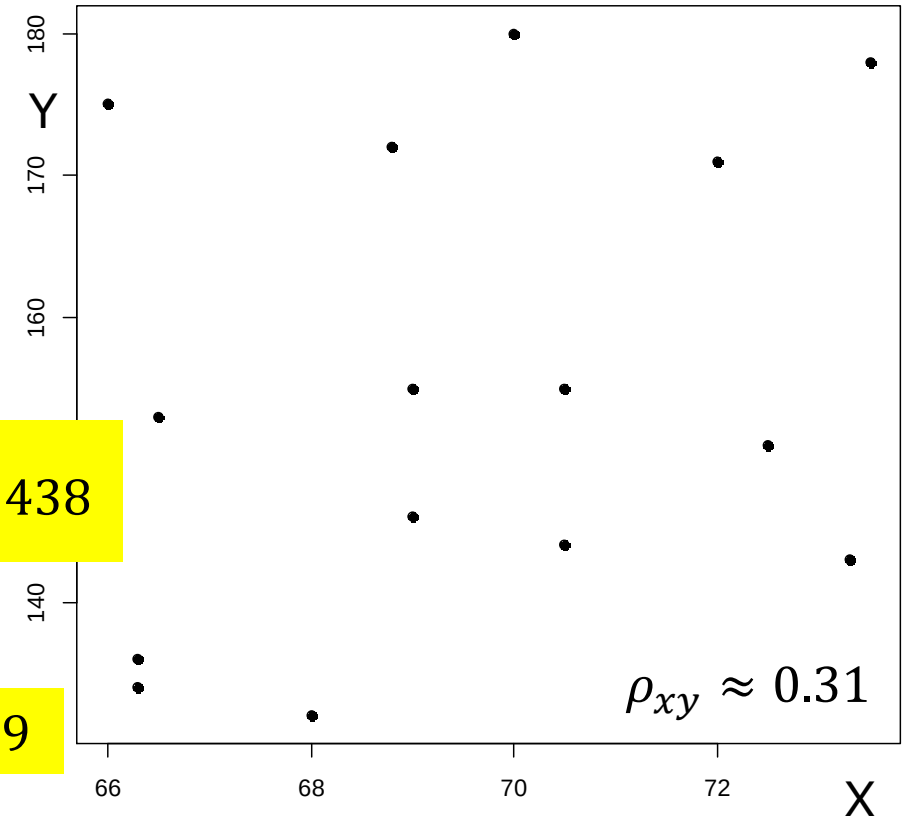
$\sigma_x = 2.552664$

$\sigma_y = 15.55622$

$\sigma_{xy} = 12.27266$ $\Rightarrow$ $\hat{b} = \frac{12.27266}{2.552664^2} = 1.883438$

$\Downarrow$

$\hat{a} = 154.4375 - 1.883438 \times 69.2625 = 23.9859$



Esempio 2

X	72.5	73.3	68.8	69	66	66.3	69	70.5	66	73.5	66.3	70	68	72	66.5	70.5
Y	151	143	172	146	175	134	155	155	146	178	136	180	132	171	153	144

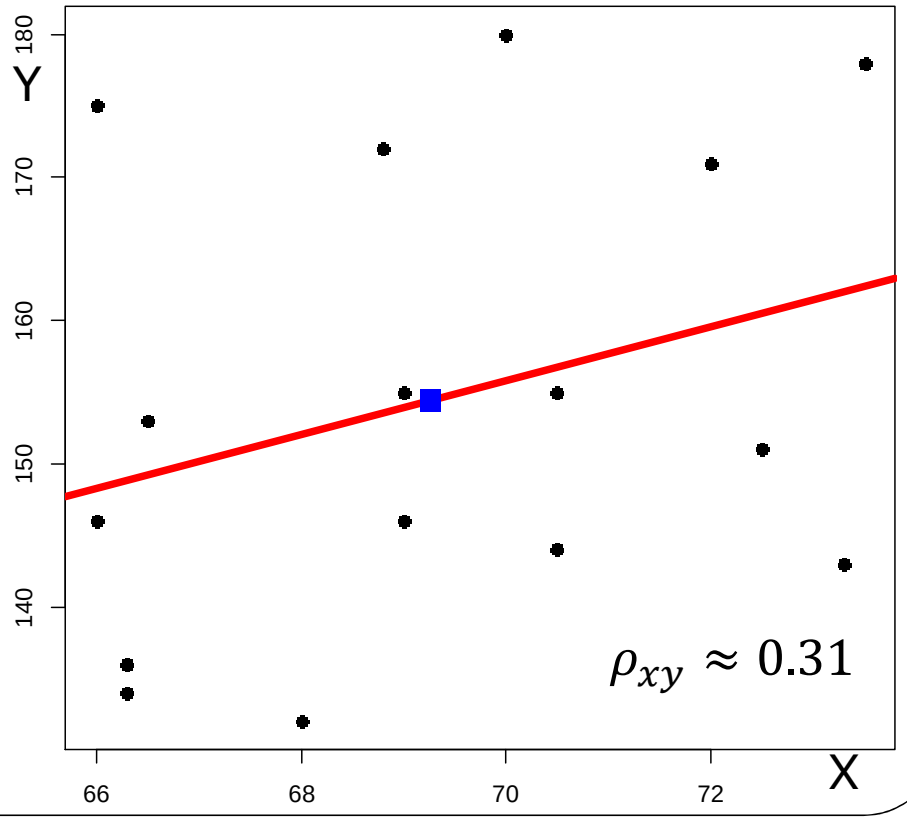
$n = 16$

$\bar{x} = 69.2625$
$\bar{y} = 154.4375$

$\sigma_x = 2.552664$
$\sigma_y = 15.55622$

$\sigma_{xy} = 12.27266$

$Y = 23.9859 + 1.883438X$

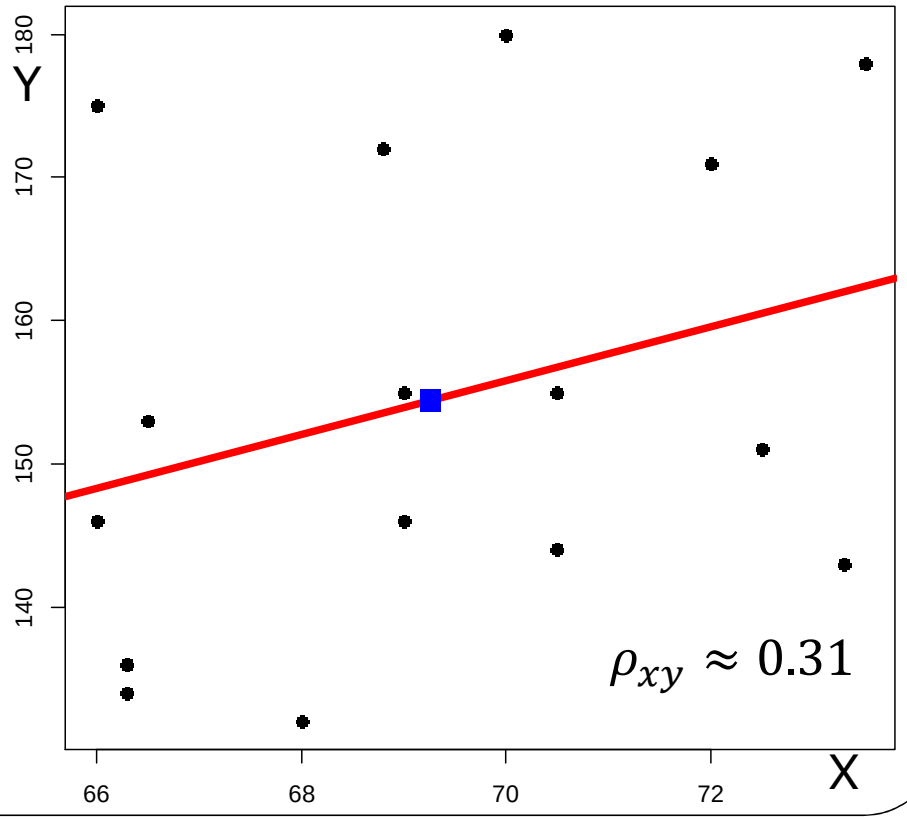


Esempio 2

X	72.5	73.3	68.8	69	66	66.3	69	70.5	66	73.5	66.3	70	68	72	66.5	70.5
Y	151	143	172	146	175	134	155	155	146	178	136	180	132	171	153	144

$$Y = 23.9859 + 1.883438\,x$$

Per l'aumento di 1 piede nell'altezza in media il peso aumenta di 1.88… libbre



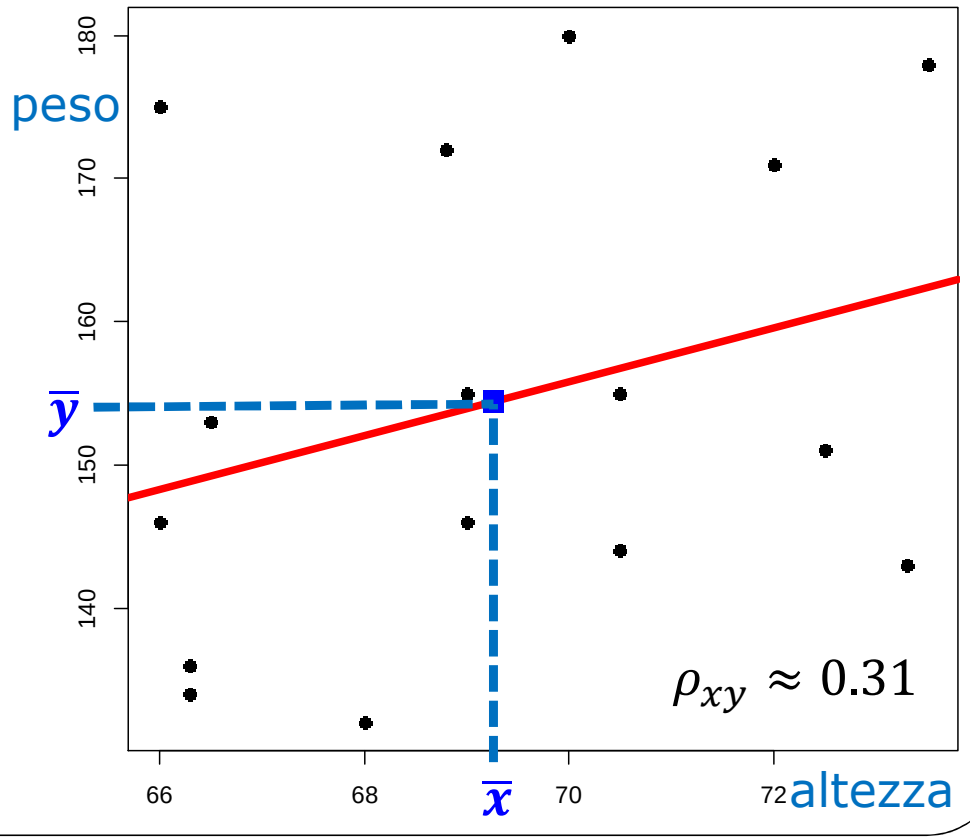
Esempio 2

X	72.5	73.3	68.8	69	66	66.3	69	70.5	66	73.5	66.3	70	68	72	66.5	70.5
Y	151	143	172	146	175	134	155	155	146	178	136	180	132	171	153	144

$$Y = 23.9859 + 1.883438X$$

previsione per un valore non osservato di X:

$$X = \bar{x} = 69.2625 \Longrightarrow \hat{y} = \bar{y} = 154.4375$$



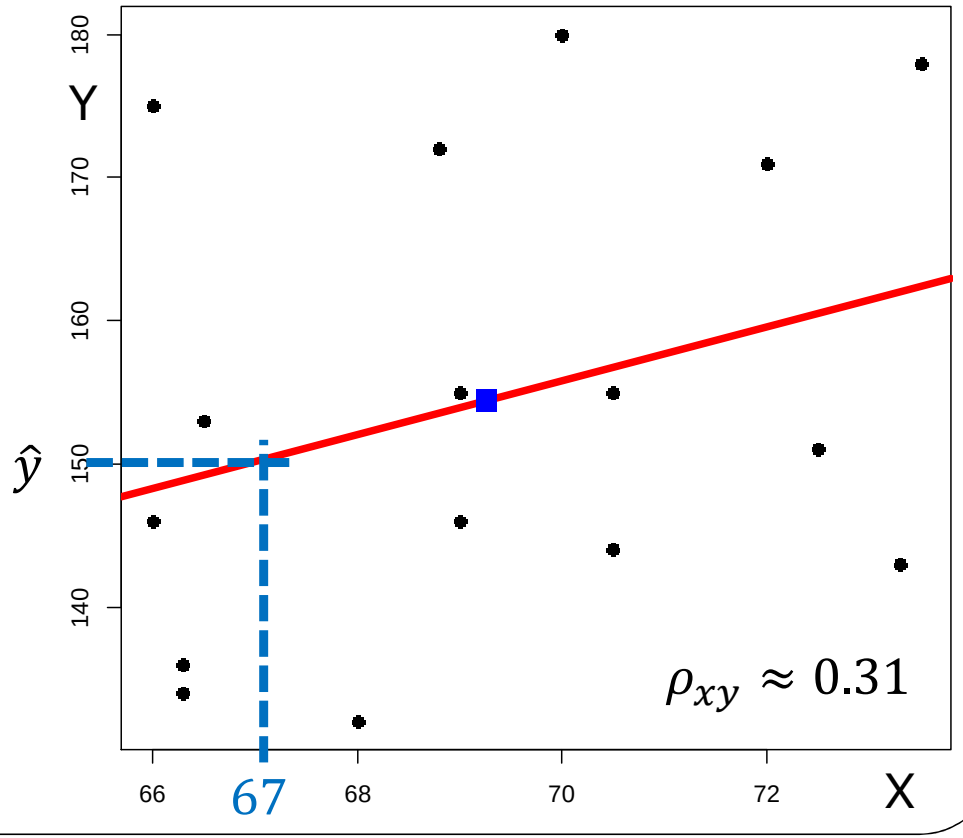
Esempio 2

X	72.5	73.3	68.8	69	66	66.3	69	70.5	66	73.5	66.3	70	68	72	66.5	70.5
Y	151	143	172	146	175	134	155	155	146	178	136	180	132	171	153	144

$$Y = 23.9859 + 1.883438x$$

previsione per un valore non osservato di X:

$$X = 67 \Longrightarrow \hat{y} = \hat{a} + \hat{b}x =$$
$$= 23.9859 + 1.883438 \times 67$$
$$= 150.1762$$



Esempio 2

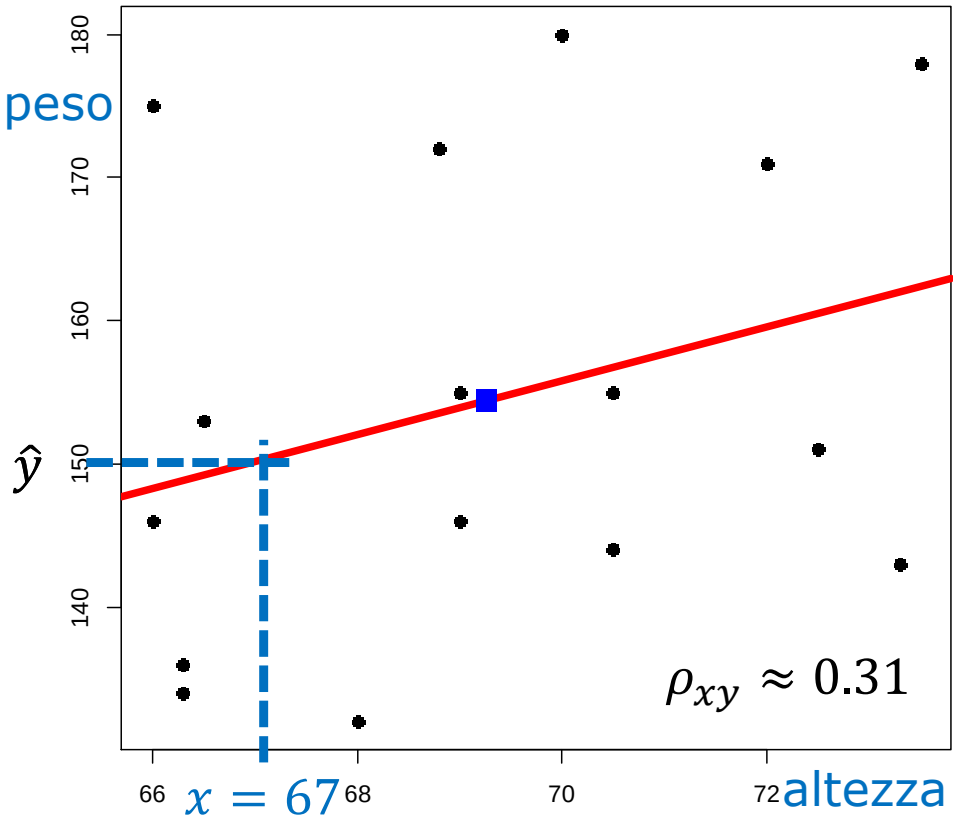
X	72.5	73.3	68.8	69	66	66.3	69	70.5	66	73.5	66.3	70	68	72	66.5	70.5
Y	151	143	172	146	175	134	155	155	146	178	136	180	132	171	153	144

$$Y = 23.9859 + 1.883438X$$

previsione per un valore non osservato di X:

$$X = 67 \Longrightarrow \hat{y} = \hat{a} + \hat{b}x = 150.1762$$

Sulla base di questo modello, una persona alta 67 pollici (170.18 cm) dovrebbe pesare circa 150 libbre (68 kg).



Esempio 2

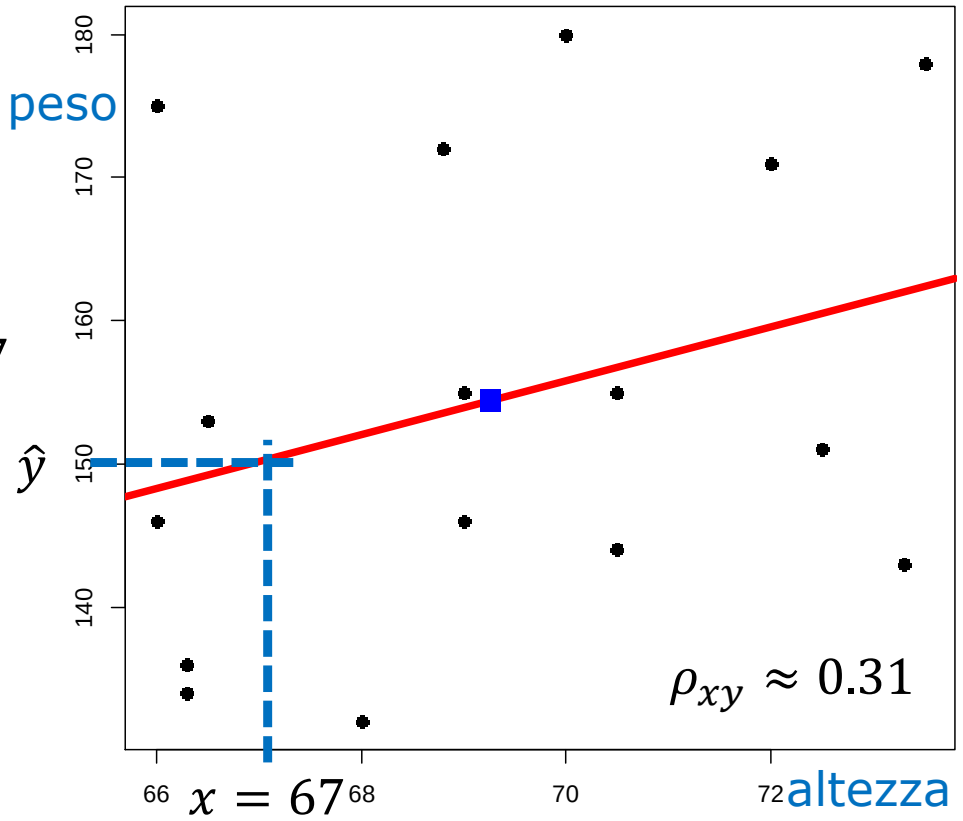
X	72.5	73.3	68.8	69	66	66.3	69	70.5	66	73.5	66.3	70	68	72	66.5	70.5
Y	151	143	172	146	175	134	155	155	146	178	136	180	132	171	153	144

$$Y = 23.9859 + 1.883438X$$

previsione per un valore non osservato di X:

$$X = 67 \Longrightarrow \hat{y} = 23.9859 + 1.883438 \times 67 = 150.1762$$

La previsione è attendibile?



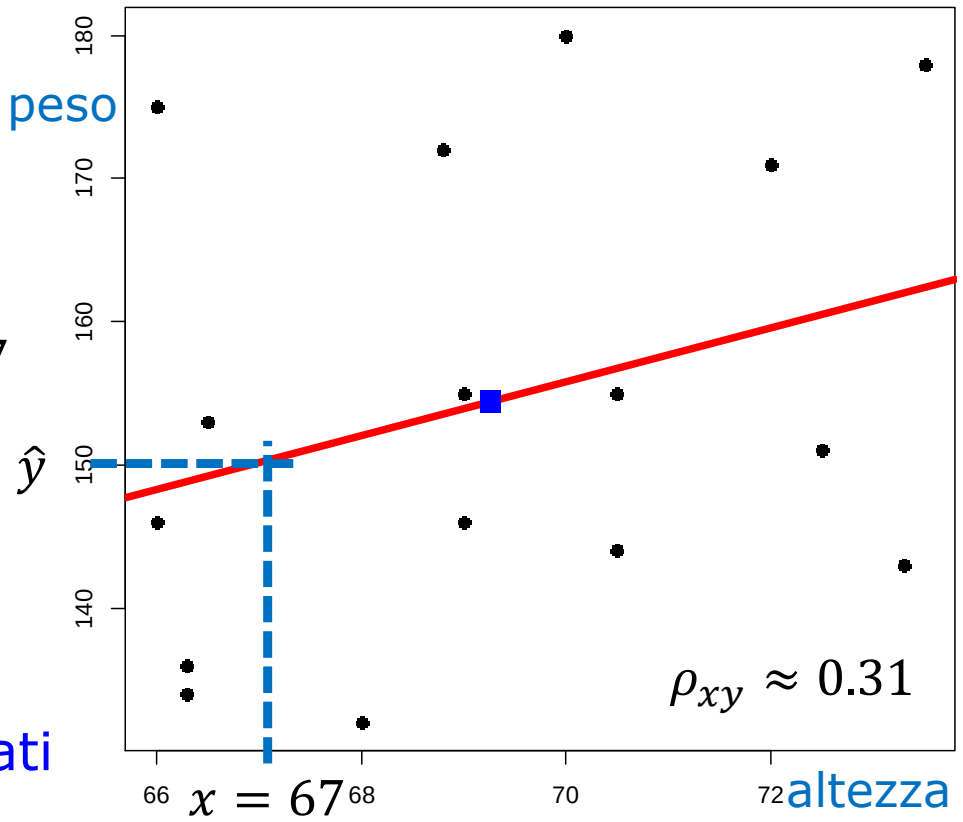
Esempio 2

X	72.5	73.3	68.8	69	66	66.3	69	70.5	66	73.5	66.3	70	68	72	66.5	70.5
Y	151	143	172	146	175	134	155	155	146	178	136	180	132	171	153	144

$$Y = 23.9859 + 1.883438X$$

previsione per un valore non osservato di X:

$$X = 67 \Longrightarrow \hat{y} = 23.9859 + 1.883438 \times 67$$
$$= 150.1762$$

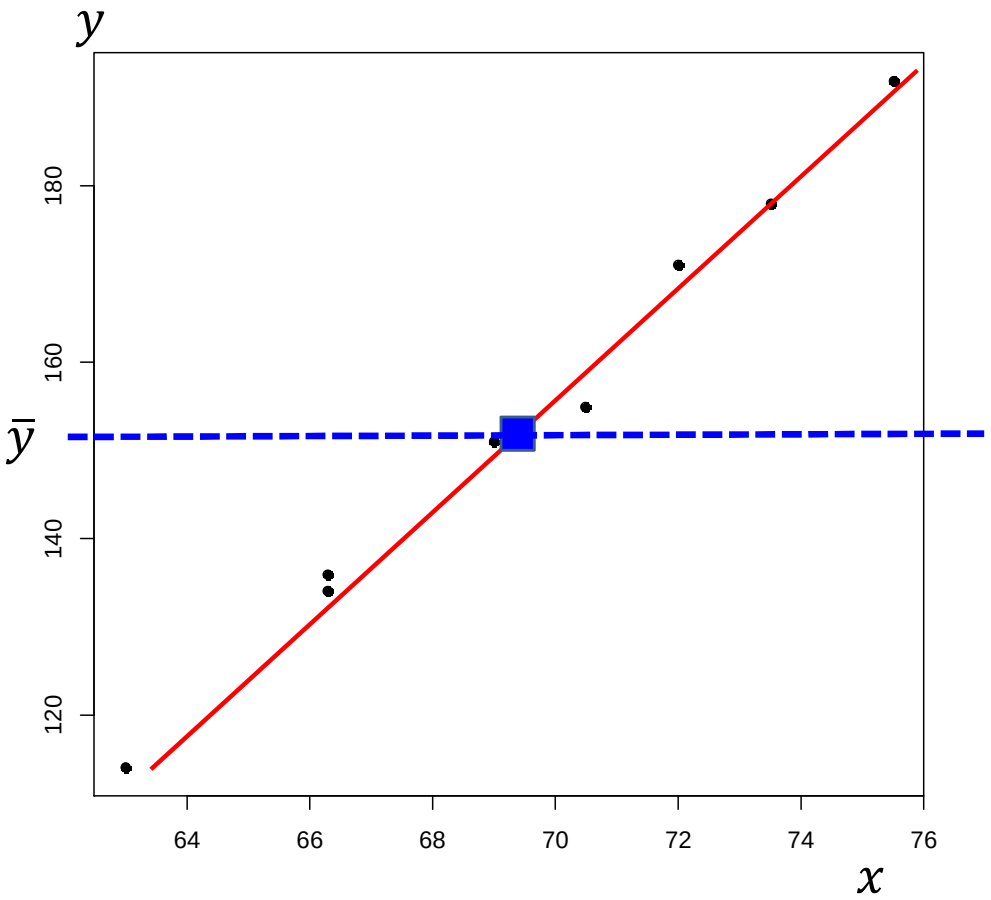


La previsione è attendibile?

solo se il modello si adatta bene ai dati

La varianza spiegata dalla retta

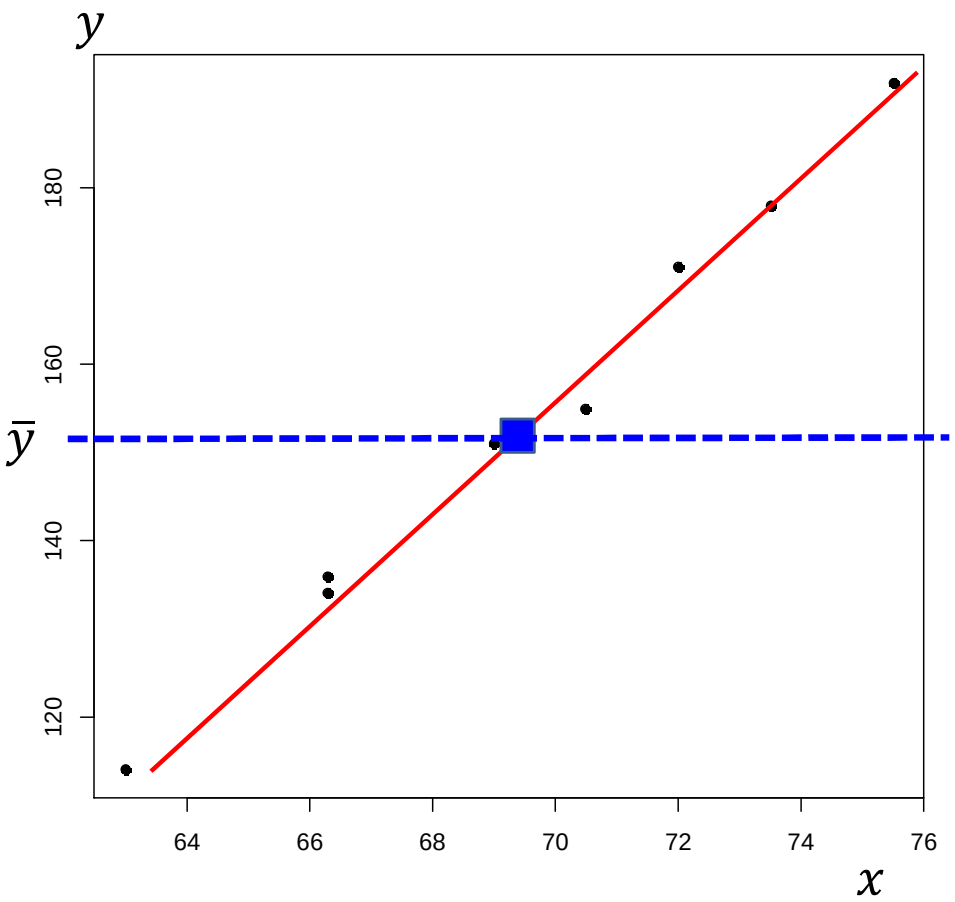
$$\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \bar{y}$$

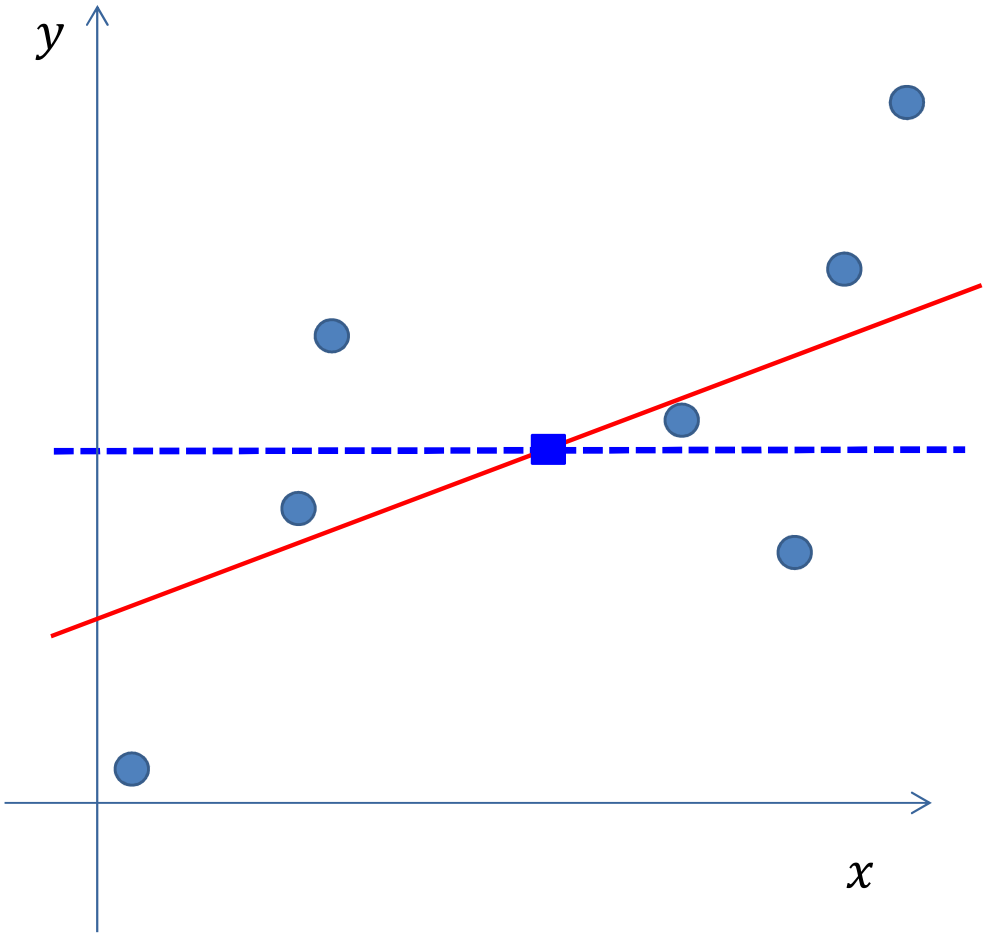


$$\sigma_y{}^2 = \frac{1}{n}\sum(y_i - \bar{y})^2 \approx \sigma_{\hat{y}}{}^2 = \frac{1}{n}\sum(\hat{y}_i - \bar{y})^2$$

La varianza spiegata dalla retta

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





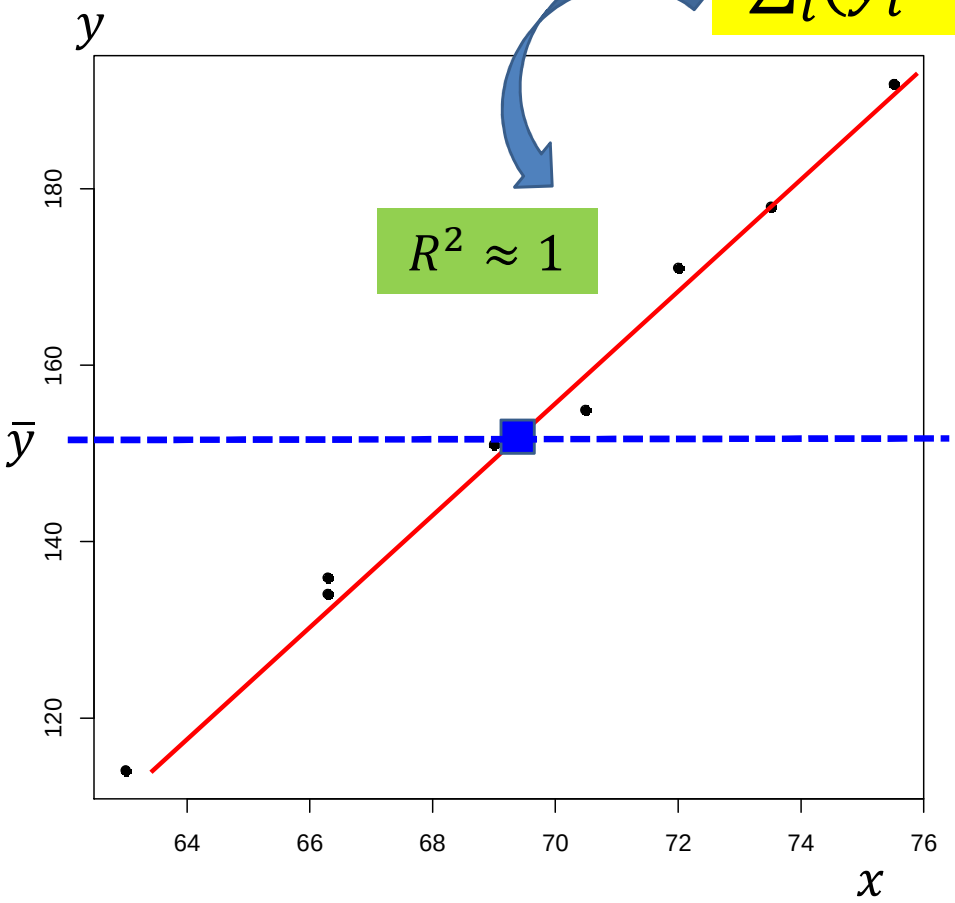
$$\sigma_y{}^2 = \frac{1}{n}\sum(y_i - \bar{y})^2 \approx \sigma_{\hat{y}}{}^2 = \frac{1}{n}\sum(\hat{y}_i - \bar{y})^2$$

e qui???

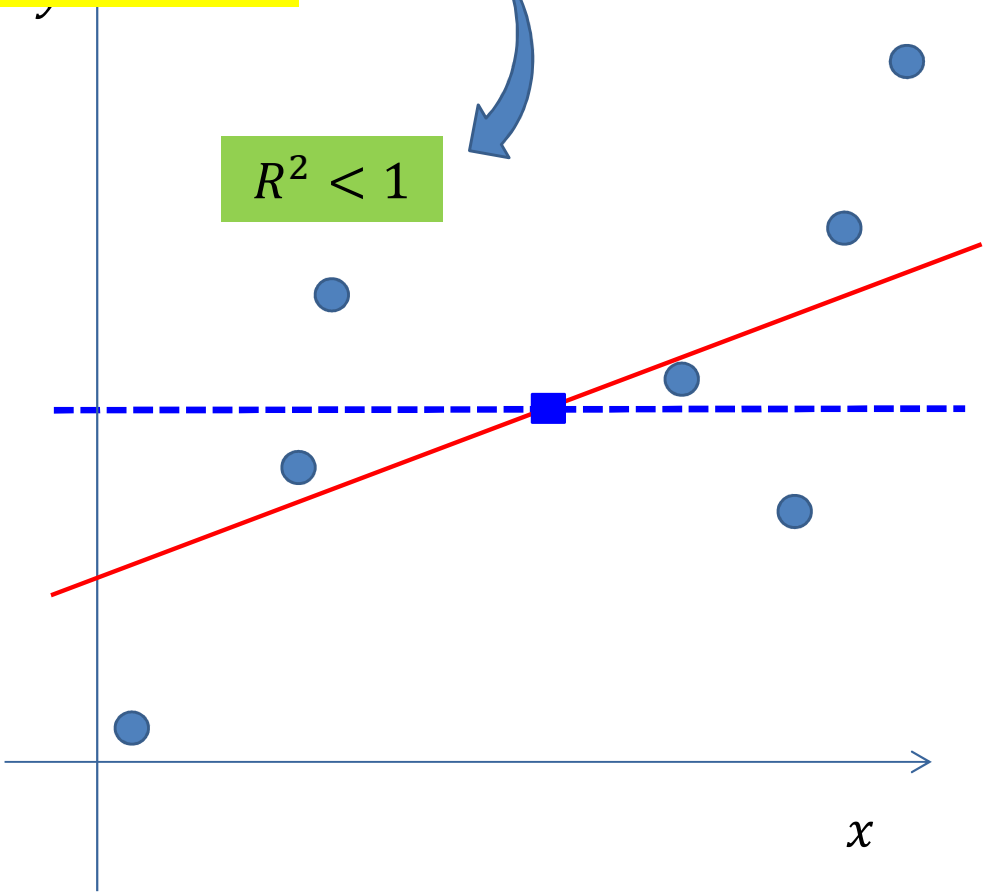
La varianza spiegata dalla retta

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

$$\frac{\sum_i(\hat{y}_i - \bar{y})^2}{\sum_i(y_i - \bar{y})^2} = R^2$$



$R^2 \approx 1$



$R^2 < 1$

$$\sigma_y{}^2 = \frac{1}{n}\sum(y_i - \bar{y})^2 \approx \sigma_{\hat{y}}{}^2 = \frac{1}{n}\sum(\hat{y}_i - \bar{y})^2$$

e qui???

La bontà della regressione

$$\frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = R^2$$

coeff. di determinazione

$$0 \leq R^2 \leq 1$$

$$R^2 = \rho_{xy}{}^2$$

$$\boldsymbol{R^2 > 0.7 \Leftrightarrow \rho_{xy} > 0.837} \text{ o } \boldsymbol{\rho_{xy} < -0.837}$$

(per tendenze crescenti)

(per tendenze decrescenti)

Analisi della Varianza

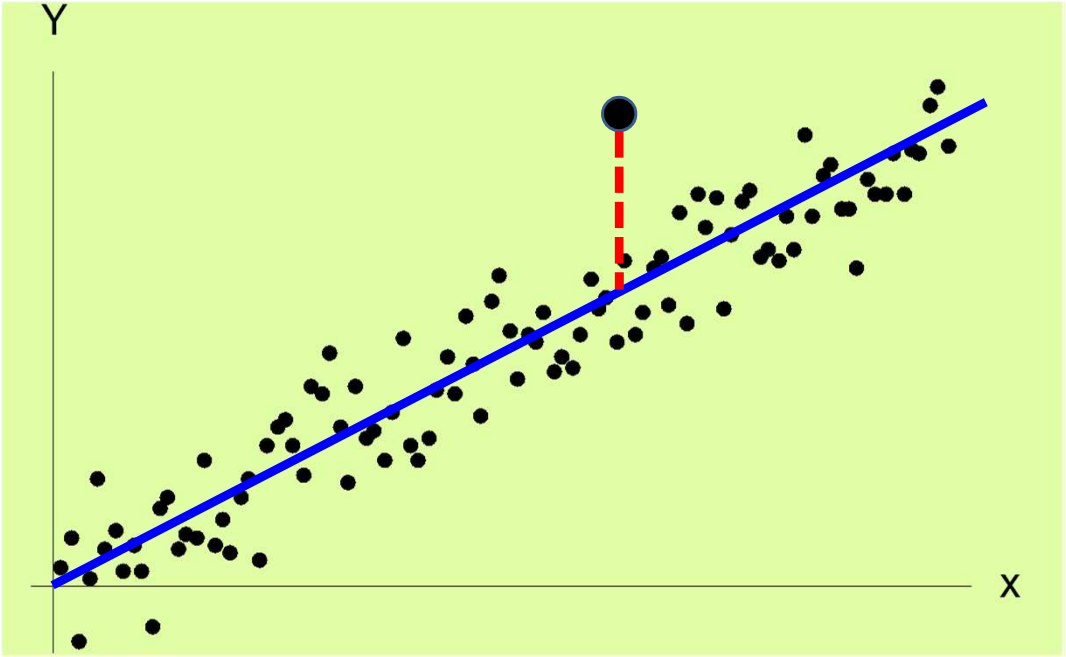
Fonte di variabilità	Gradi di libertà (gl)	SS (Sum of Squares)	Mean Square (SS/gl)
Retta di regressione	1	$\sum_i (\hat{y}_i - \bar{y})^2$	
Attorno alla retta	$n-2$	$\sum_i (y_i - \hat{y}_i)^2$	$\frac{1}{n-2}\sum_{i=1}^{n} {e_i}^2$
Totale	$n-1$	$\sum_i (y_i - \bar{y})^2$	

varianza ***spiegata***

varianza ***totale***

num. di parametri stimati (a e b)

Inferenza



Il modello della **regressione lineare semplice**:

$$f(x) = a + bx$$
$$\approx \Leftrightarrow \ \varepsilon_i \sim N(0, \sigma^2)$$
$$\varepsilon_i \text{ indipendenti}$$

$$Y_i = a + bx_i + \varepsilon_i$$

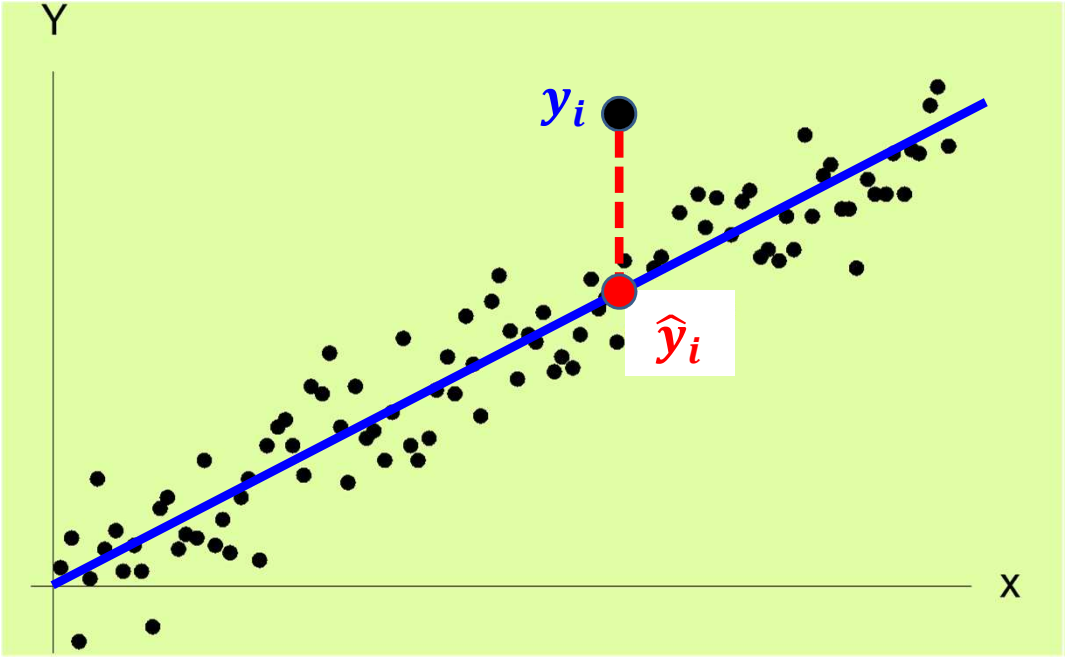
Il modello ha tre parametri incogniti: a, b, σ^2

1. Stimare σ^2

2. Verificare se il vero valore della pendenza nella popolazione è davvero diverso da zero ($\Leftrightarrow$ previsione) oppure no:

$$H_0 : b = 0, \quad H_1 : b \neq 0$$

Inferenza



$$f(x) = a + bx$$
$$\approx \Leftrightarrow \ \varepsilon_i \sim N(0, \sigma^2)$$
$$\varepsilon_i \text{ indipendenti}$$

$$Y_i = a + bx_i + \varepsilon_i$$

$$\hat{y}_i = \hat{a} + \hat{b}y_i$$

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

$$\sum_{i=1}^{n} e_i = 0$$

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} {e_i}^2$$

stima di σ^2

varianza degli errori

Inferenza

dalle stime agli stimatori:

$$B_n = \frac{\sum(Y_i - \bar{Y}_n)(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$$

$$A_n = \bar{Y}_n - B_n\bar{x}$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$Y_i \sim N(a + bx_i, \sigma^2)$$

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} {e_i}^2$$

A_n e B_n v.c. gaussiane

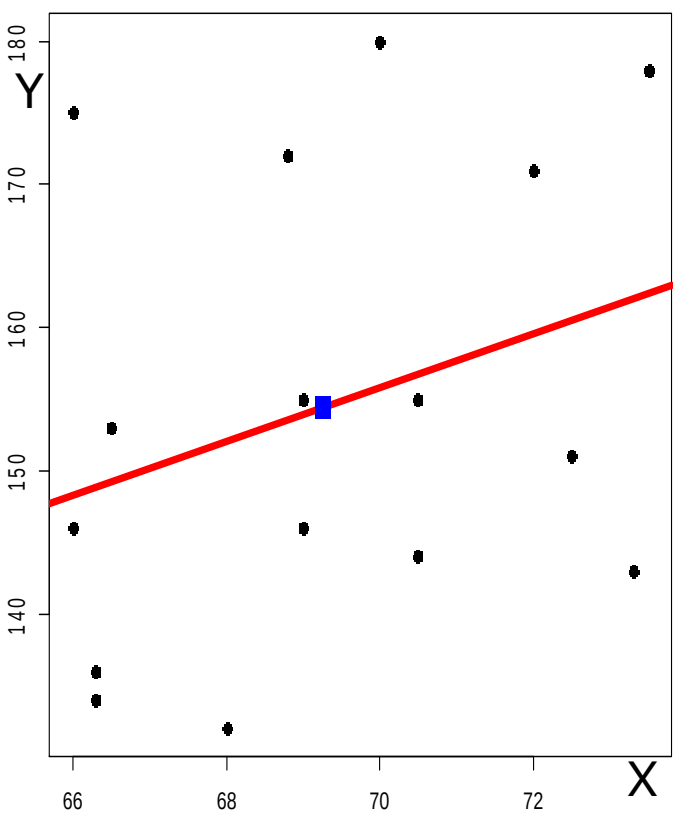
$$H_0 : b = 0 \qquad H_1 : b \neq 0$$

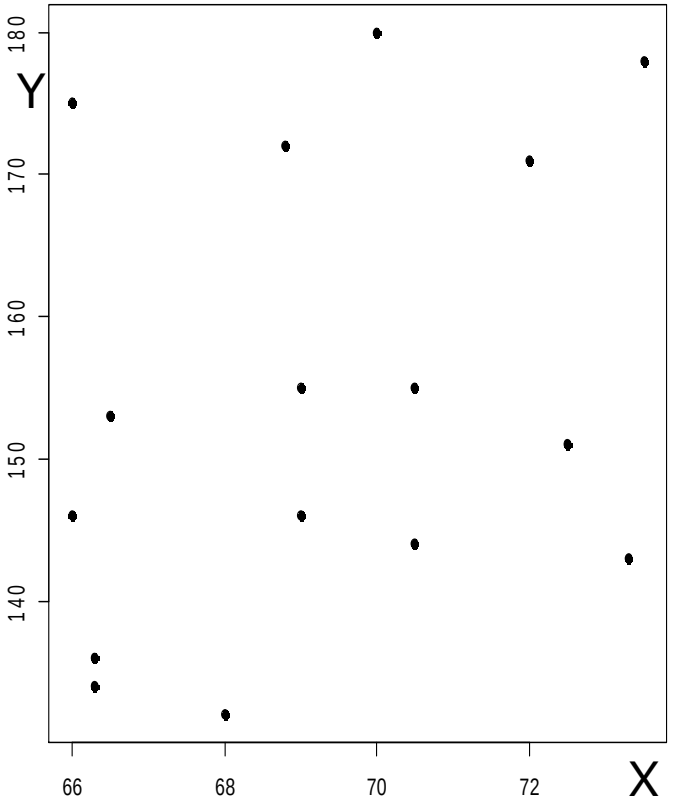
rifiutiamo H_0 se:

$$\frac{|\hat{b}|}{\sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} > t(\boldsymbol{n-2})_{\frac{\alpha}{2}}$$

Inferenza

La relazione lineare stimata sui dati vale in generale nella popolazione?





Inferenza

$$H_0 : a = 0 \quad H_1 : a \neq 0$$

$$\frac{|\hat{a}|}{\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}} > t(n-2)_{1-\frac{\alpha}{2}}$$

Il **modello** di regressione lineare

$$Y_i = a + bx_i + \varepsilon_i \quad , \qquad \varepsilon_i \sim N(0, \sigma^2)$$

In questo modello, **mi aspetto** di osservare il valore $\hat{y}_i$ **(sulla retta),** ma **l'incertezza** del fenomeno può produrre **un'osservazione** y_i **che non sta sulla retta**. Questo *errore*, $e_i = y_i - \hat{y}_i$, è supposto **gaussiano**, quindi non può essere troppo grande ("$-3\sigma, 3\sigma$"), e deve essere simmetrico, nel senso che l'istogramma degli e_i deve dare una «campana» simmetrica.