STATISTICA

Regressione-2

Fare sempre il grafico!

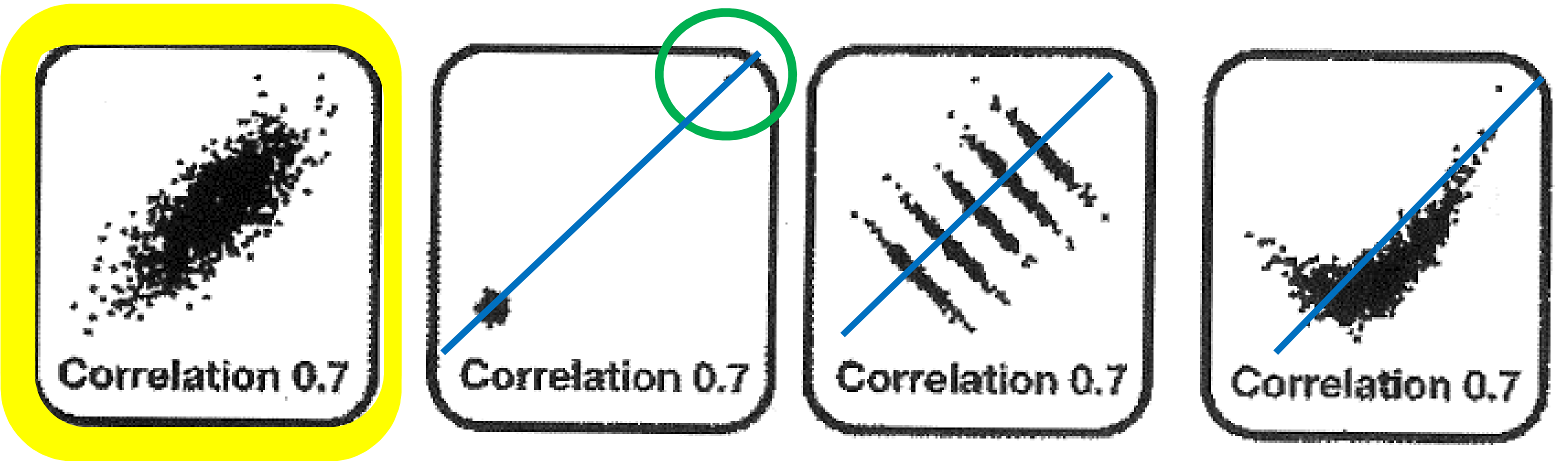
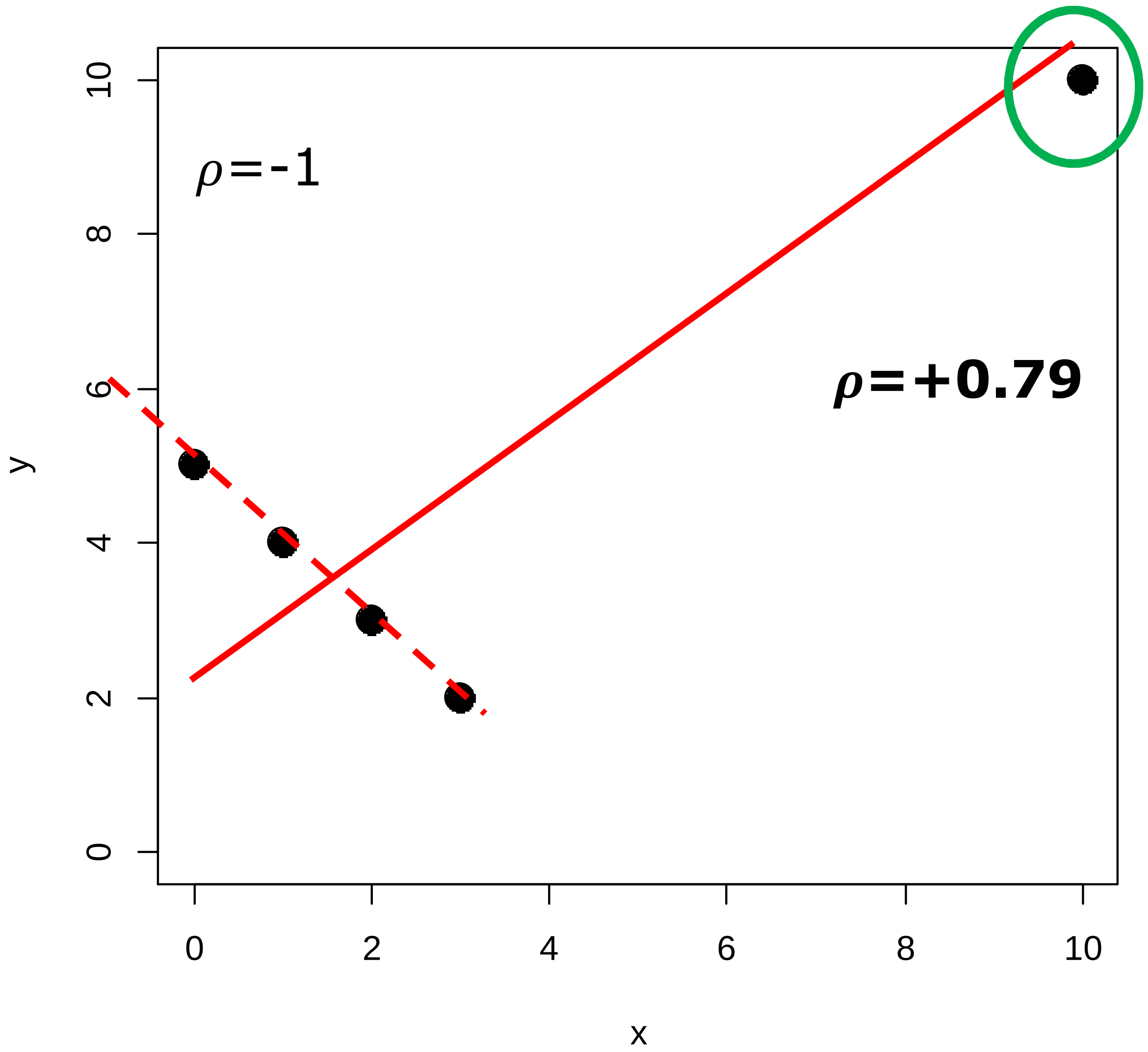


Figura 5.12 - Coefficiente di correlazione lineare e bontà di adattamento: ecco quattro esempi in cui il coefficiente di correlazione risulta essere pari a 0.7! (Fonte immagine: section on Statistical Graphics, American Statistical Association).

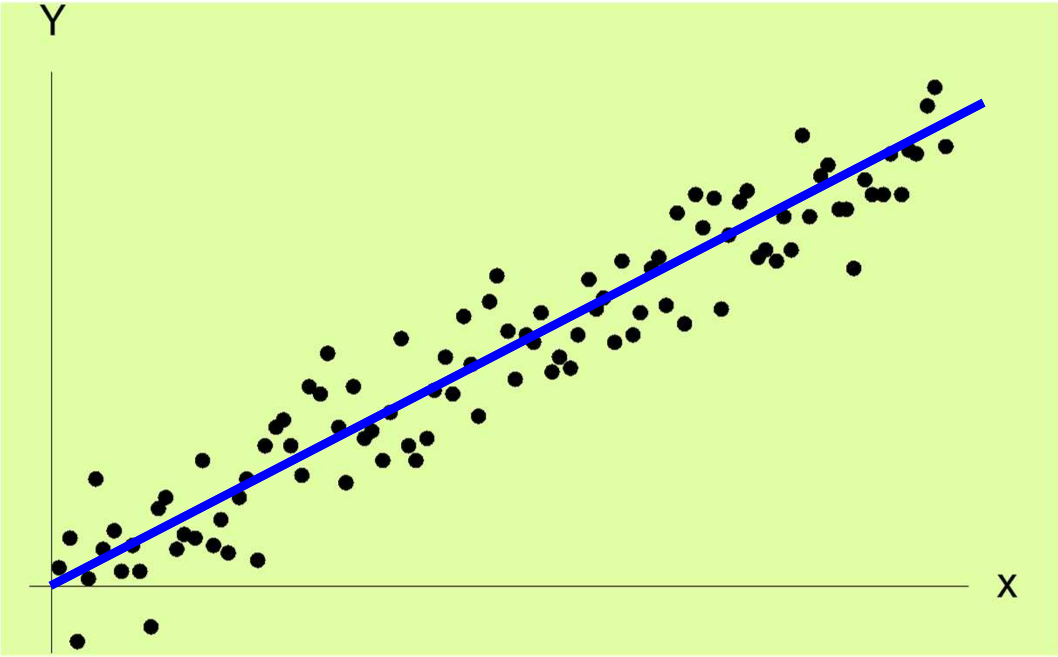
Fig. 5.12 pg 178, “Statistica” di S. Iacus

Fare sempre il grafico!



outlier
o
dato influente

Inferenza



Il modello della **regressione lineare semplice**:

$$f(x) = a + bx$$

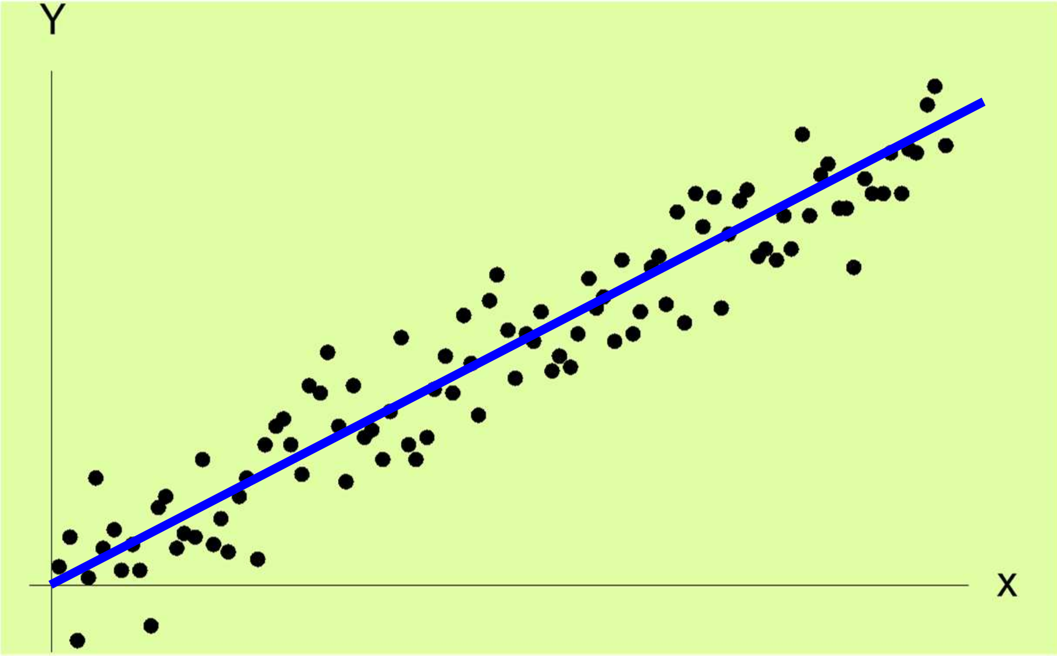
$$\approx \Leftrightarrow \ \varepsilon_i \sim N(0, \sigma^2)$$

ε_i **indipendenti**

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

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

Inferenza



Il modello della **regressione lineare semplice**:

$$f(x) = a + bx$$

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

$$\varepsilon_i \text{ indipendenti}$$

⇩

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

⇩

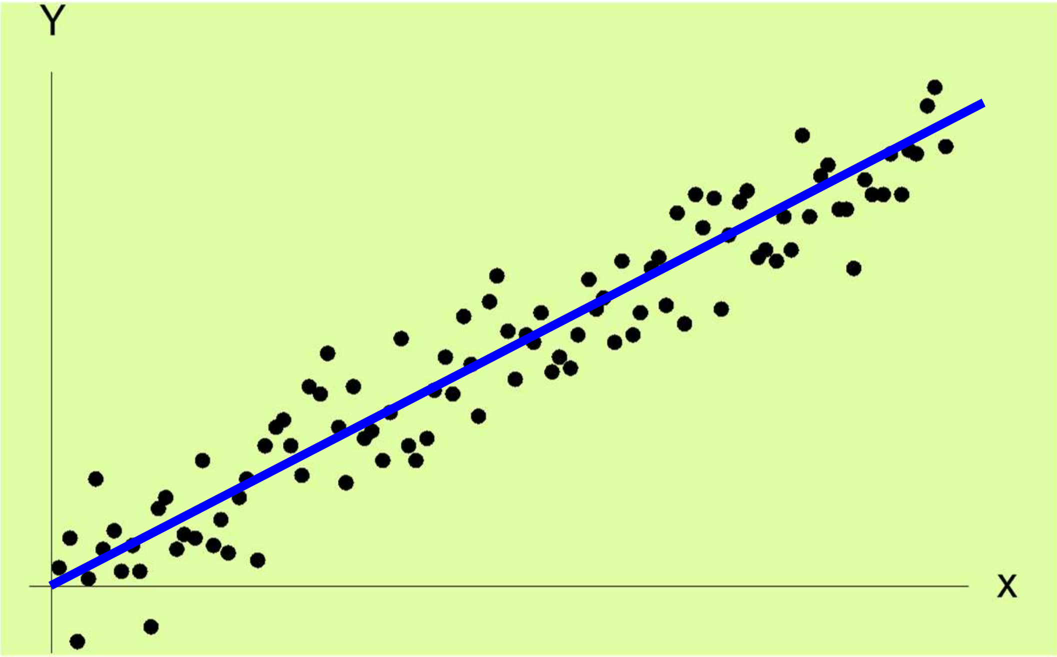
$$Y_i \sim N(a + bx_i, \sigma^2)$$

Il valore medio di Y_i in corrispondenza a tutte le unità statistiche per cui $X = x_i$ è

$$a + bx_i$$

$$E(Y_i) = a + bx_i$$

Inferenza



Il modello della **regressione lineare semplice**:

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

$$\boldsymbol{Y_i = a + bx_i + \varepsilon_i}$$

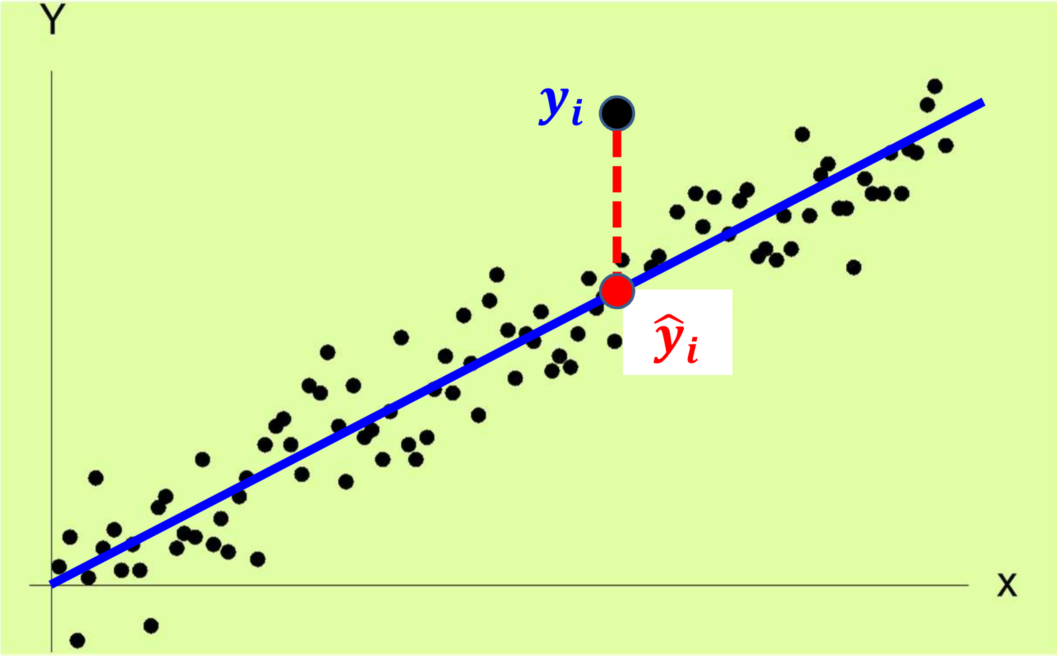
Il modello ha tre parametri incogniti: a, b, σ^2

1. Stimare a, b e σ^2

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

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

Inferenza



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

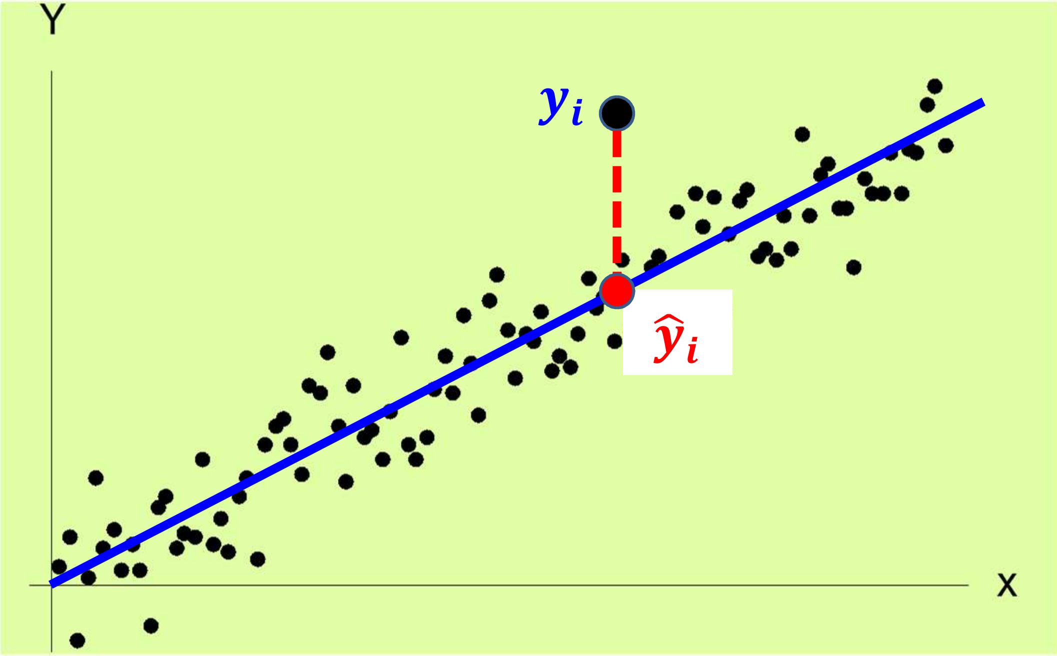
$$Y_i = a + bx_i + \varepsilon_i$$

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

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

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

Inferenza



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

$$\boldsymbol{Y_i = a + bx_i + \varepsilon_i}$$

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

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

varianza degli errori

stima di σ^2

errori ≈ **residui**

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

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

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

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

$\boldsymbol{H_0 : b = b_0}$ $\boldsymbol{H_1 : b \neq b_0}$

rifiutiamo H_0 se:

$$\frac{\left|\hat{b} - b_0\right|}{\sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} > t(\boldsymbol{n-2})_{\frac{\alpha}{2}}$$

Inferenza

dalle stime agli stimatori:

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

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

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

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

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

A_n e B_n v.c. gaussiane

Intervallo di confidenza di livello $1 - \alpha$ per b :

$$\left(\hat{b} - t(\boldsymbol{n-2})_{\frac{\alpha}{2}} \times \sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \hat{b} + t(\boldsymbol{n-2})_{\frac{\alpha}{2}} \times \sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \right)$$

Inferenza

dalle stime agli stimatori:

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

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

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

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

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

...aussiane

E SE CONTIENE LO 0?

Tipo: (-1.23, 2.17)

Intervallo di confidenza di livello $1 - \alpha$ per b :

$$\left(\hat{b} - t(\boldsymbol{n-2})_{\frac{\alpha}{2}} \times \sqrt{\frac{s^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}, \hat{b} + t(\boldsymbol{n-2})_{\frac{\alpha}{2}} \times \sqrt{\frac{s^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \right)$$

Inferenza

$$H_0 : a = a_0 \quad H_1 : a \neq a_0$$

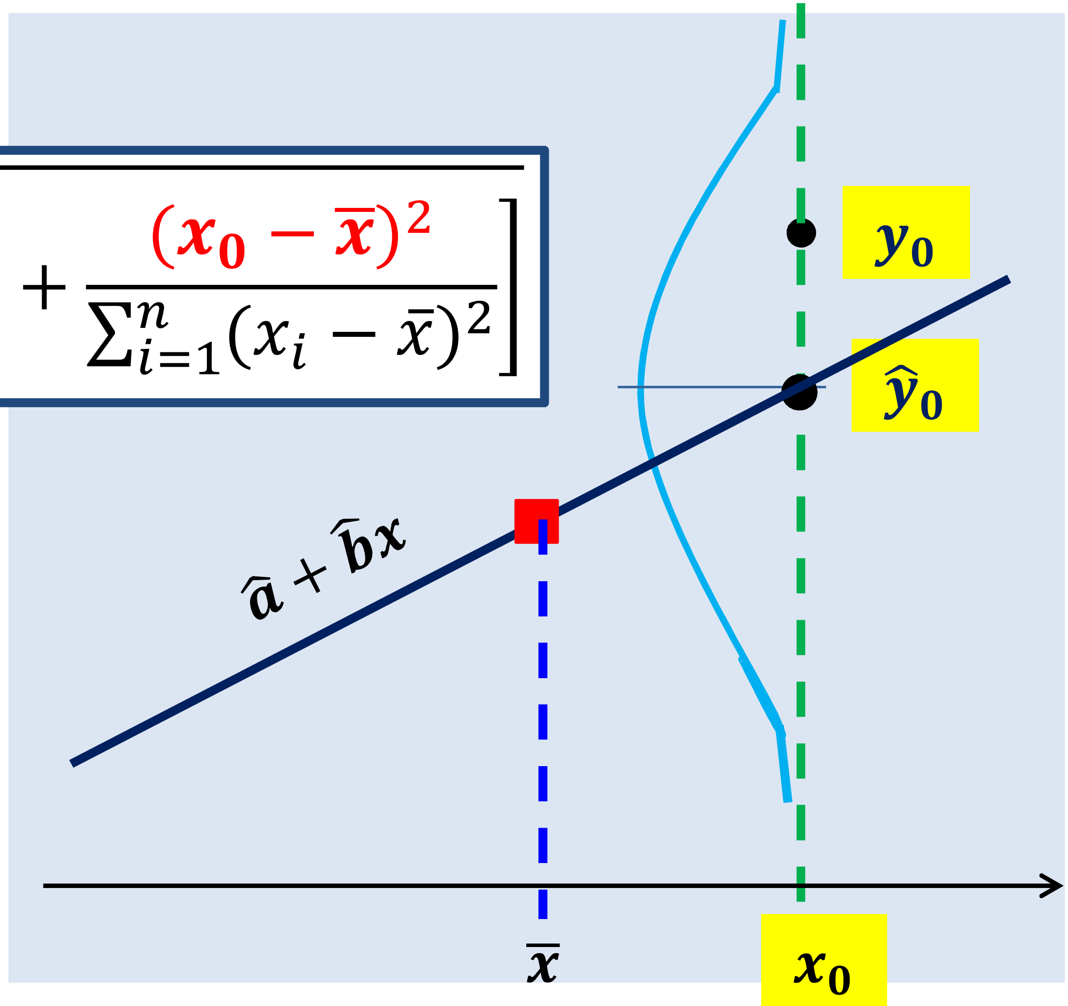
$$\frac{|\hat{a} - a_0|}{\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}} > t(n-2)_{\frac{\alpha}{2}}$$

Intervallo di confidenza di livello $1 - \alpha$ per a :

$$\left(\hat{a} - t(\boldsymbol{n}-\mathbf{2})_{\frac{\alpha}{2}} \times \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \hat{a} + t(\boldsymbol{n}-\mathbf{2})_{\frac{\alpha}{2}} \times \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right)$$

Inferenza per la previsione

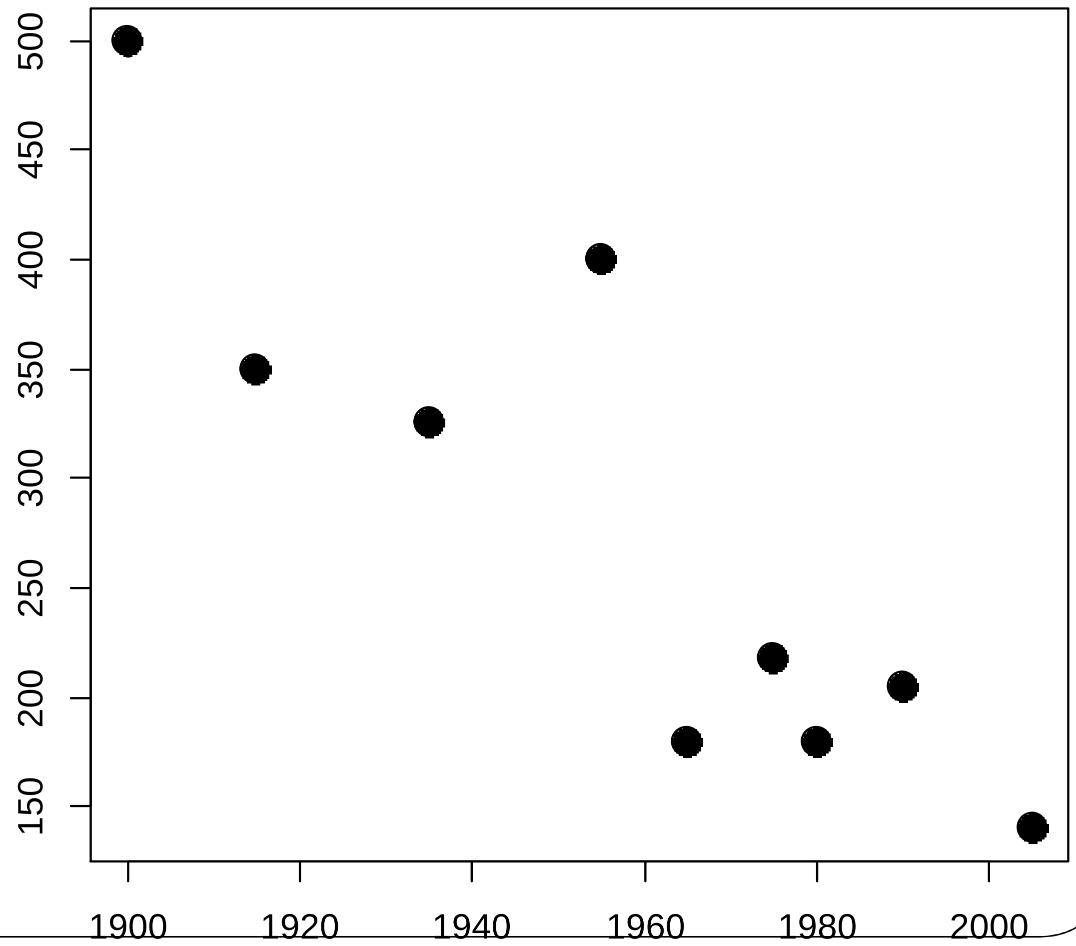
$$\hat{y}_0 \mp t(n-2)_{\frac{\alpha}{2}} \times \sqrt{s^2\left[1 + n^{-1} + \frac{(\boldsymbol{x_0} - \overline{\boldsymbol{x}})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$



Esercizio 2

X anno, Y consumo medio annuo procapite di pane

X	1900	1915	1935	1955	1965	1975	1980	1990	2005
Y (kg)	500	350	325	400	180	218	180	205	140



Esercizio 2

X anno, Y consumo medio annuo procapite di pane

X	1900	1915	1935	1955	1965	1975	1980	1990	2005
Y (kg)	500	350	325	400	180	218	180	205	140

$\bar{x} = 1957.78$

$\bar{y} = 277.65$

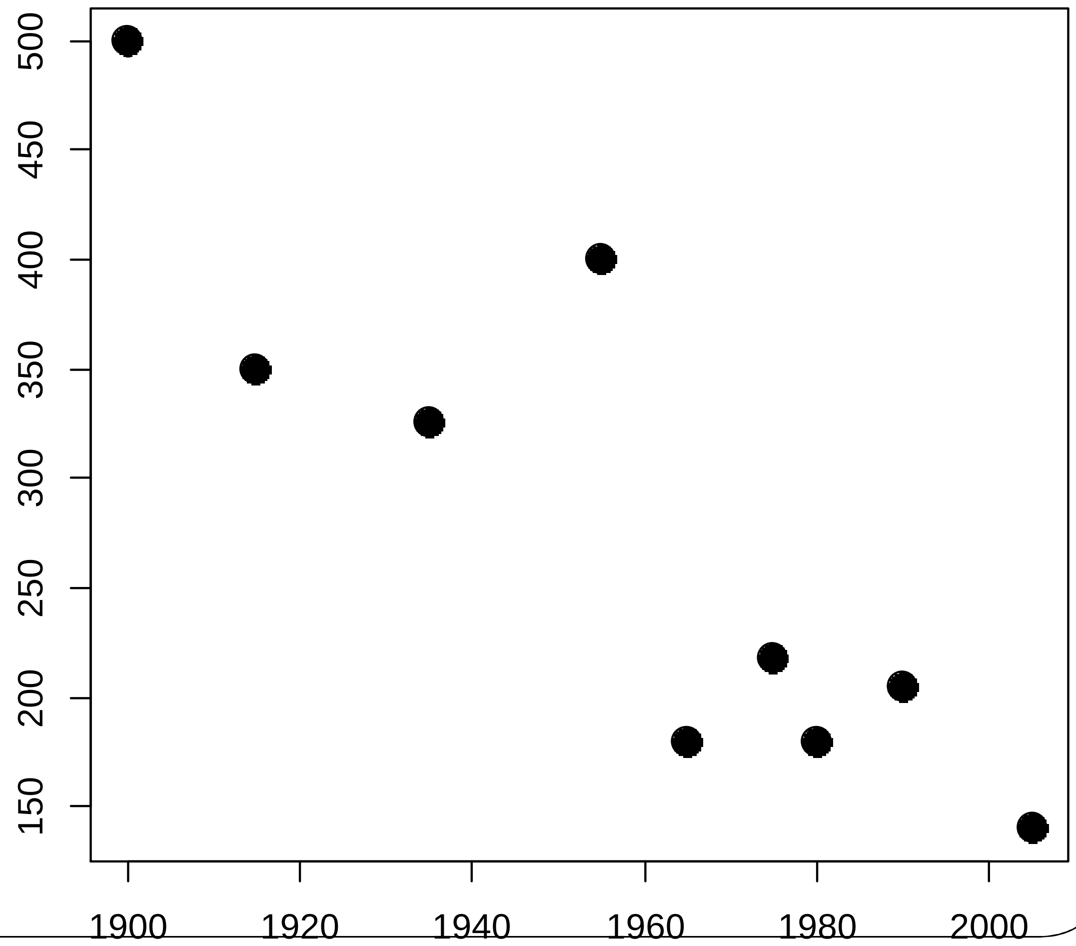
$\sigma_x{}^2 = 1089.51$

$\sigma_y{}^2 = 13193.36$

$cov(x,y) = -3344.877$

$$\rho_{xy} = \frac{-3344.877}{\sqrt{1089.51 \times 13193.36}} = \mathbf{-0.88}$$

$R^2 = (-0.88)^2 = \mathbf{0.77}$



Esercizio 2

X anno, Y consumo medio annuo procapite di pane

X	1900	1915	1935	1955	1965	1975	1980	1990	2005
Y (kg)	500	350	325	400	180	218	180	205	140

$\bar{x} = 1957.78$

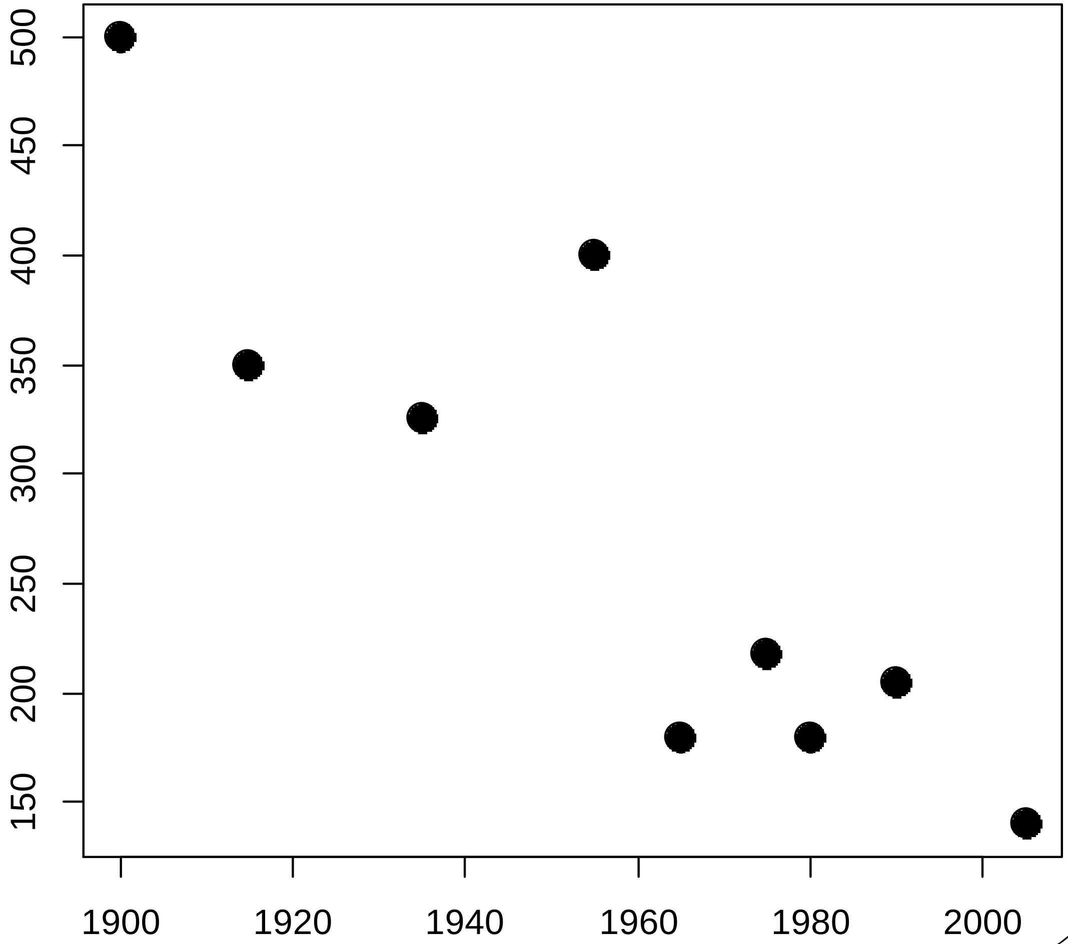
$\bar{y} = 277.65$

${\sigma_x}^2 = 1089.51$

$cov(x, y) = -3344.877$

$$\hat{b} = \frac{-3344.877}{1089.51} = \mathbf{-3.07}$$

$$\hat{a} = 277.65 + 3.07 \times 1957.78 = \mathbf{6288.0}$$



Esercizio 2

X anno, Y consumo medio annuo procapite di pane

X	1900	1915	1935	1955	1965	1975	1980	1990	2005
Y (kg)	500	350	325	400	180	218	180	205	140

$\bar{x} = 1957.78$

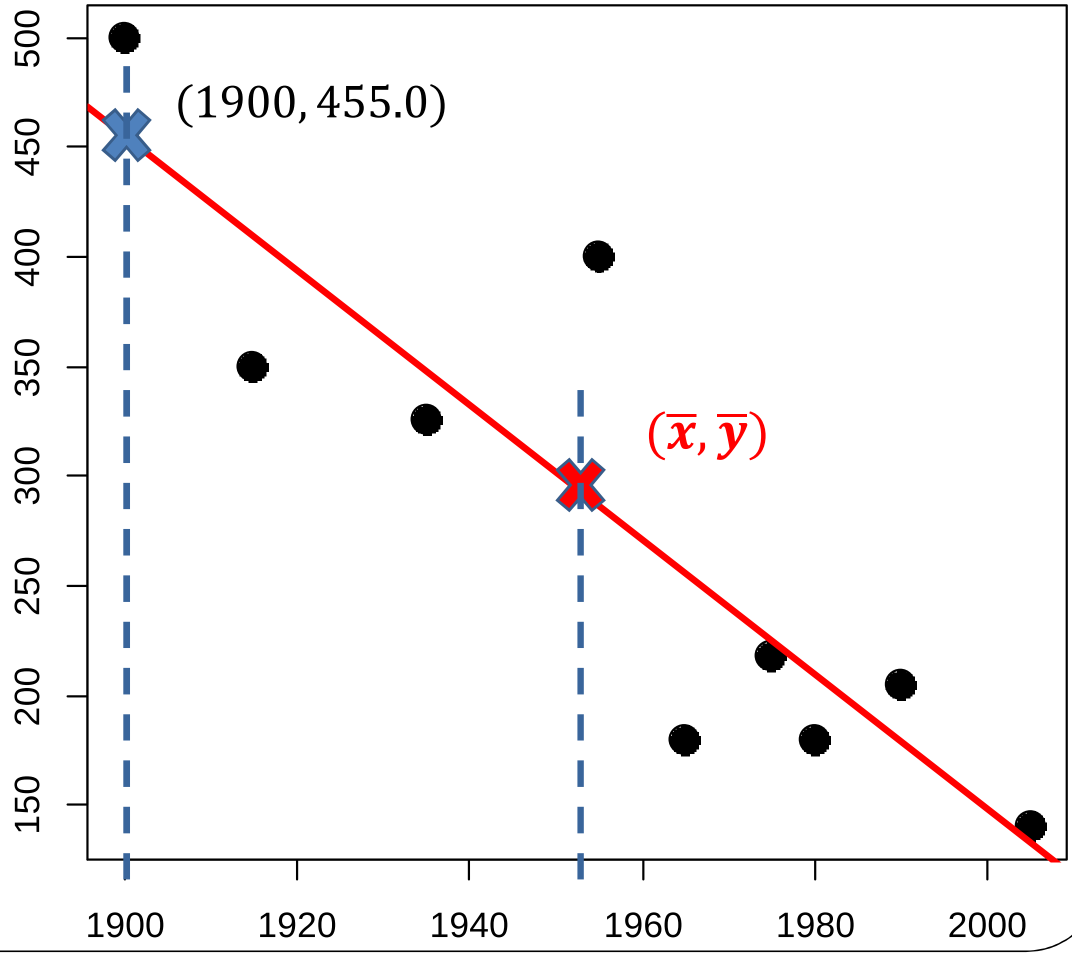
$\bar{y} = 277.65$

${\sigma_x}^2 = 1089.51$

$cov(x, y) = -3344.877$

$$\hat{b} = \frac{-3344.877}{1089.51} = \mathbf{-3.07}$$

$$\hat{a} = 277.65 + 3.07 \times 1957.78 = \mathbf{6288.0}$$



Esercizio 2

X anno, Y consumo medio annuo procapite di pane

X	1900	1915	1935	1955	1965	1975	1980	1990	2005
Y (kg)	500	350	325	400	180	218	180	205	140
$\hat{y}$	455	408.95	347.55	286.15	255.45	224.75	209.4	178.7	132.65

$\bar{x} = 1957.78$

$\bar{y} = 277.65$

${\sigma_x}^2 = 1089.51$

$cov(x, y) = -3344.877$

$\hat{a} = 6288.0$

$$s^2 = \frac{1}{7}\sum_{i=1}^{9}(y_i - \hat{y}_i)^2 = 3759.85$$

$\Rightarrow s = 61.318$

$\hat{b} = \mathbf{-3.07}$

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

$$\frac{|\hat{b}|}{\sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} = \frac{3.07}{\sqrt{\frac{3759.85}{9 \times 1089.51}}} =$$

$$= 4.958 \; ?>? \; t(7)_{\frac{0.05}{2}} = 2.3646$$

rifiutiamo l'ipotesi che $b = 0$!

Esercizio 2

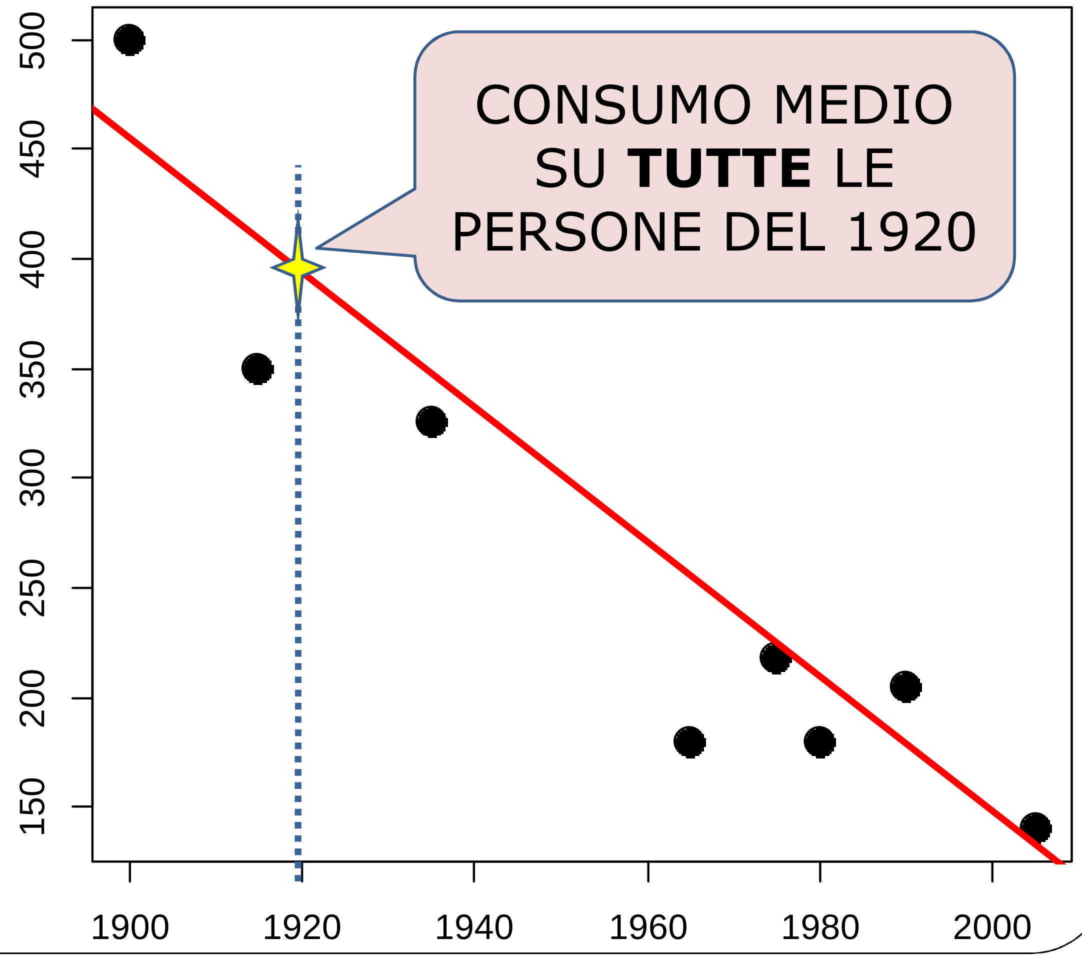
X anno, Y consumo medio annuo procapite di pane

X	1900	1915	1935	1955	1965	1975	1980	1990	2005
Y (kg)	500	350	325	400	180	218	180	205	140

previsione per $x = 1920$

$\hat{y} = 6288.0 - 3.07 \times 1920 = 393.6$ kg

in media



Esercizio 2

X anno, Y consumo medio annuo procapite di pane

X	1900	1915	1935	1955	1965	1975	1980	1990	2005
Y (kg)	500	350	325	400	180	218	180	205	140

previsione per $x = 1920$

$\hat{y} = 6288.0 - 3.07 \times 1920 = 393.6$ kg

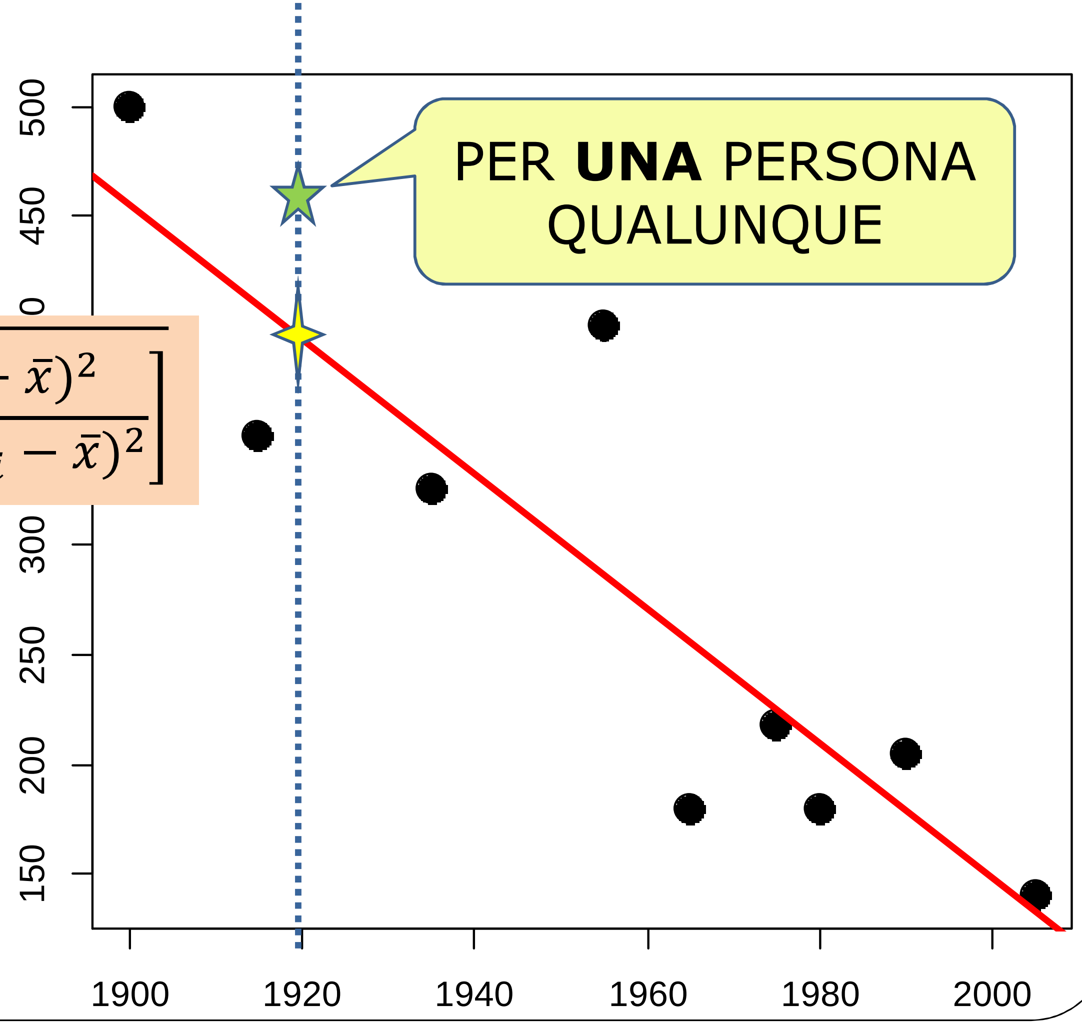
$$\hat{y}_0 \mp t(n-2)_{\frac{\alpha}{2}} \times \sqrt{s^2 \left[1 + n^{-1} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$

$s^2 = 3759.83$

$\sum_{i=1}^{n}(x_i - \bar{x})^2 = n\sigma_x{}^2 = 7626.57$

$\alpha = 0.05;\ t(n-2)_{\frac{\alpha}{2}} = 2.3646$

$393.6 \pm \mathbf{165.2} : (228.4, 558.8)$



Esercizio 2

X anno, Y consumo medio annuo procapite di pane

X	1900	1915	1935	1955	1965	1975	1980	1990	2005
Y (kg)	500	350	325	400	180	218	180	205	140

previsione per $x = 2020$

$\hat{y} = 6288.0 - 3.07 \times 2020 = 86.6$ kg

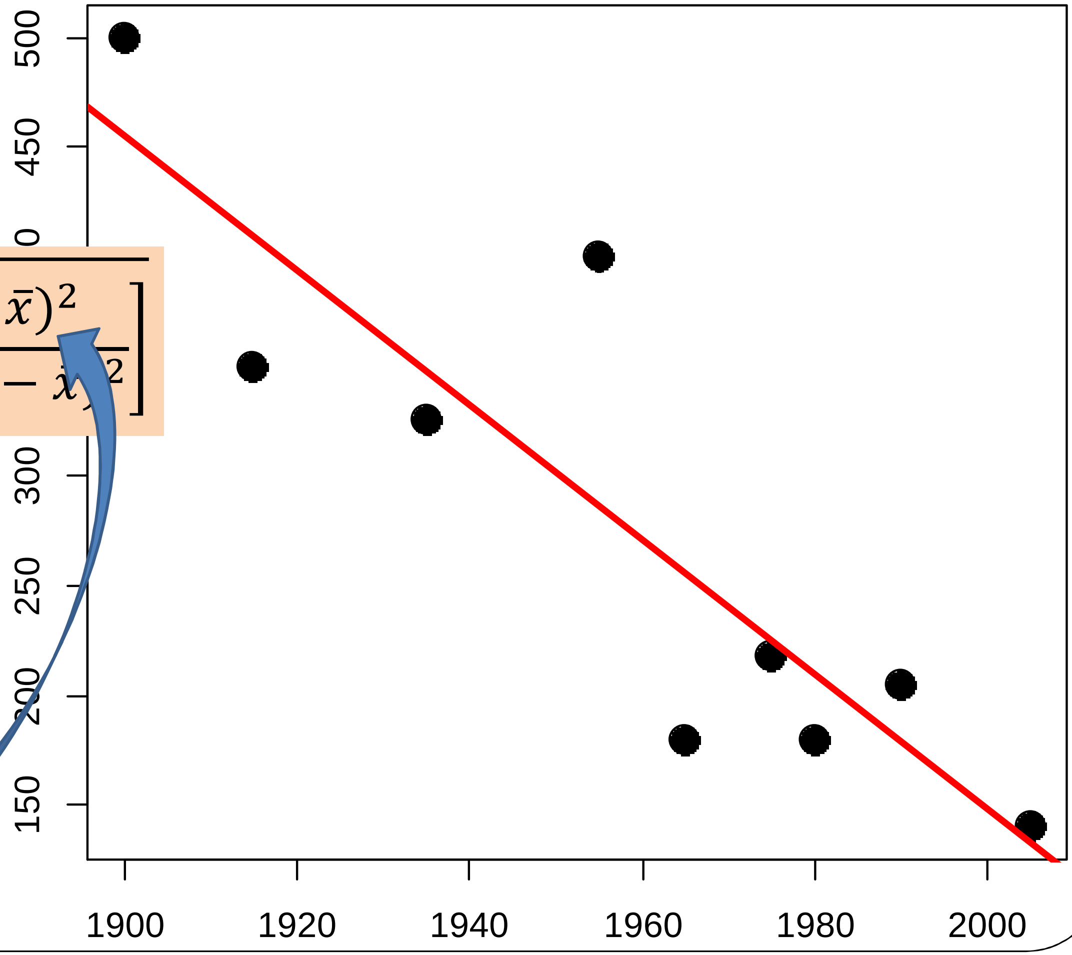
$$\hat{y}_0 \mp t(n-2)_{\frac{\alpha}{2}} \times \sqrt{s^2 \left[1 + n^{-1} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$

$s^2 = 3759.83$

$\sum_{i=1}^{n}(x_i - \bar{x})^2 = n\sigma_x{}^2 = 7626.57$

$\alpha = 0.05;\ t(n-2)_{\frac{\alpha}{2}} = 2.3646$

$86.6 \mp \mathbf{184.5}$



Esercizio 2

X anno, Y consumo medio annuo procapite di pane

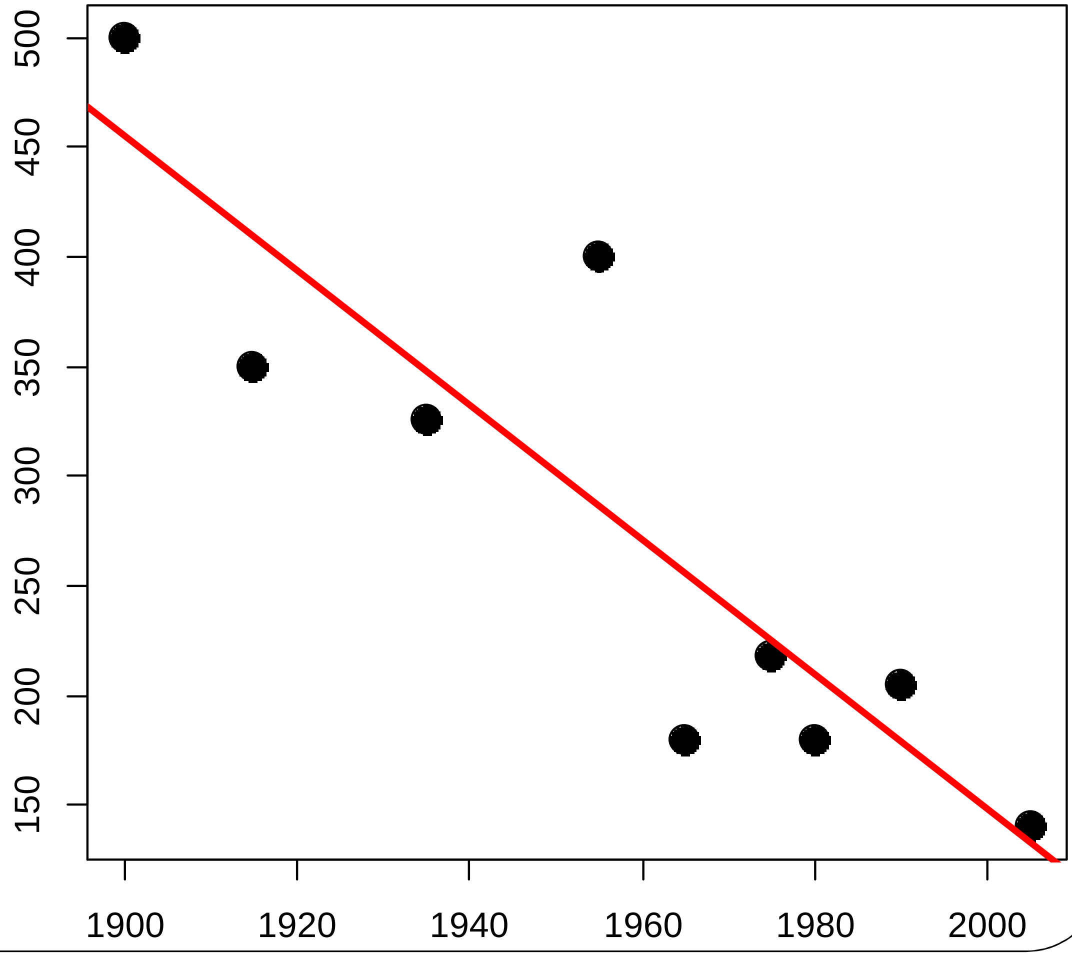
X	1900	1915	1935	1955	1965	1975	1980	1990	2005
Y (kg)	500	350	325	400	180	218	180	205	140

previsione per $x = 2050$

$y = 6288.0 - 3.07 \times 2050 =$ **-5.5 kg**

$$\hat{b} = \frac{-3344.877}{1089.51} = -3.07$$

$$\hat{a} = 277.65 + 3.07 \times 1957.78 = 6288.0$$



Esercizio 2

X anno, Y consumo medio annuo procapite di pane

X	1900	1915	1935	1955	1965	1975	1980	1990	2005
Y (kg)	500	350	325	400	180	218	180	205	140

previsione per $x = 2050$

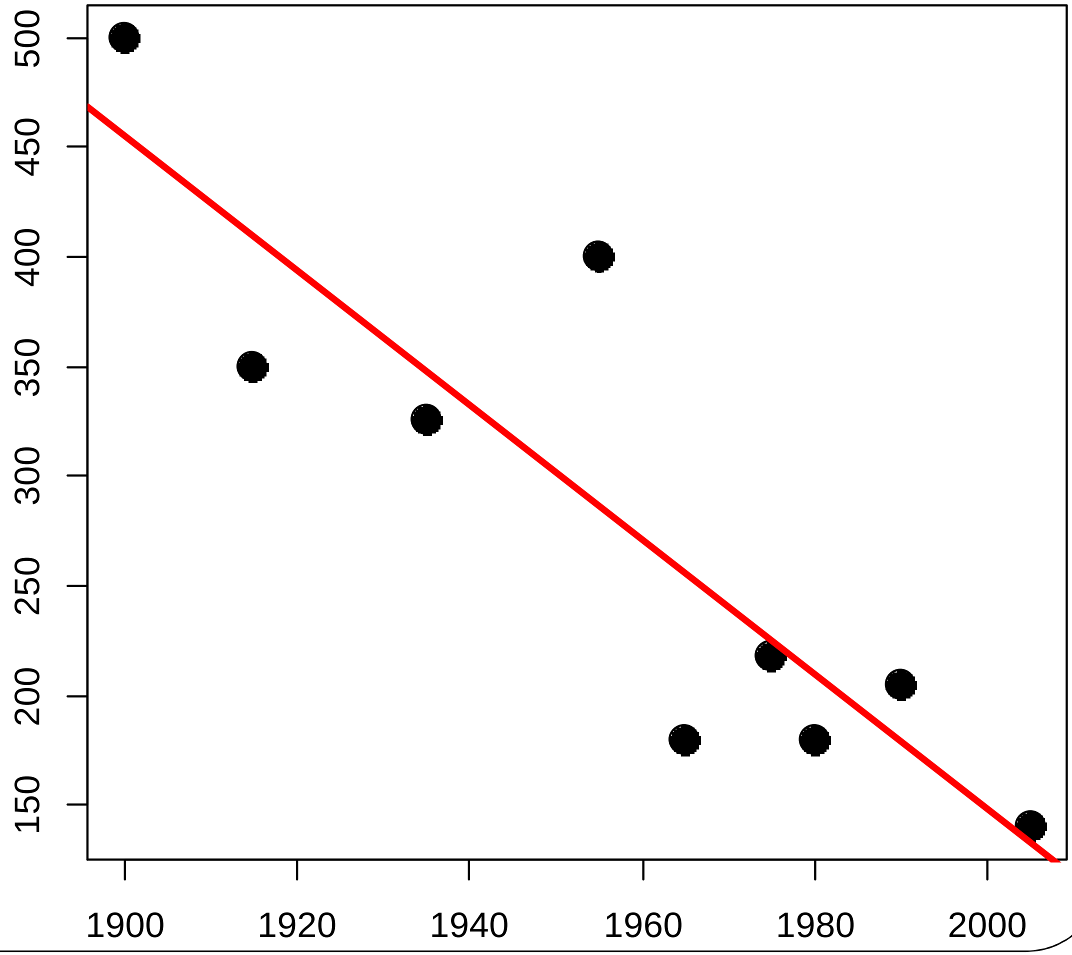
$y = 6288.0 - 3.07 \times 2050 = \mathbf{-5.5}$ **kg**

"**previsioni di lungo periodo**"

sono *fuori* dal range dei dati!

Meglio non farle! Ma se proprio...

attenzione al senso!



Esercizio 2

X anno, Y consumo medio annuo procapite di pane

X	1900	1915	1935	1955	1965	1975	1980	1990	2005
Y (kg)	500	350	325	400	180	218	180	205	140

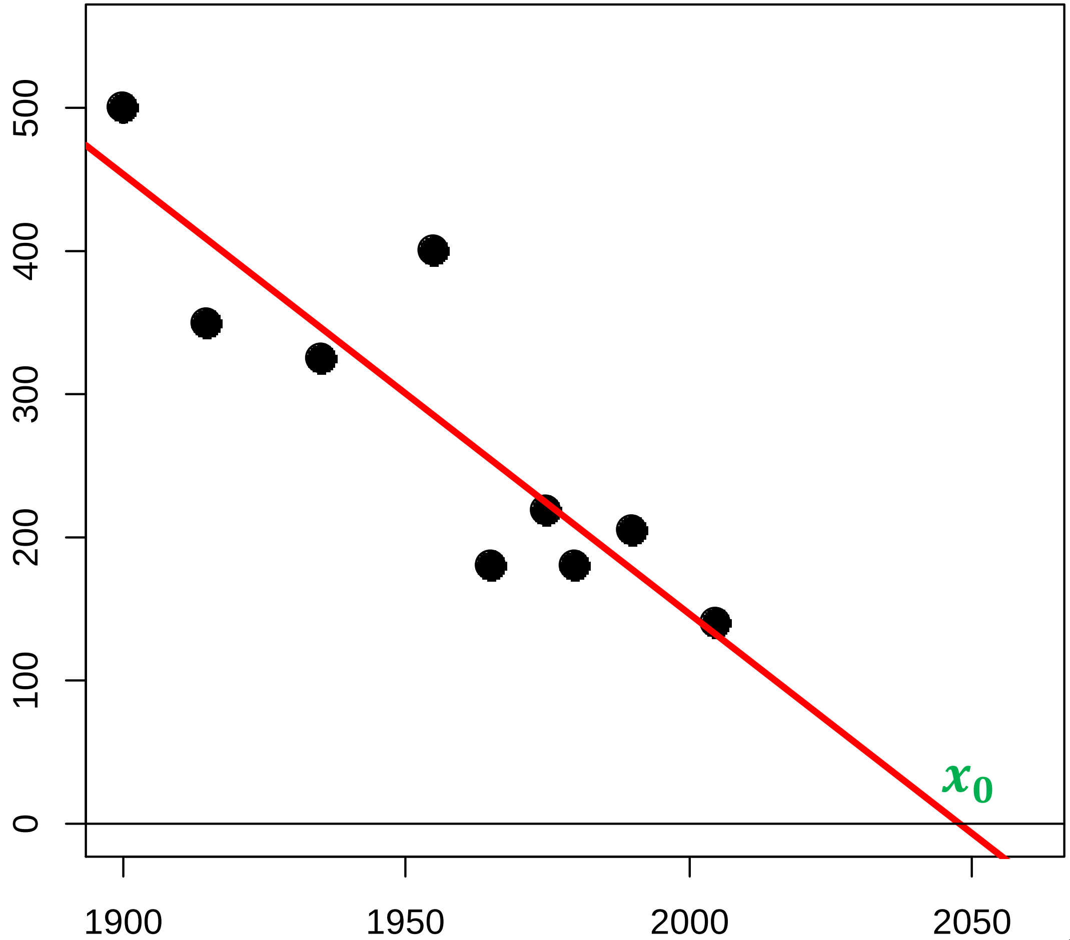
previsione per $x = 2050$

$y = 6288.0 - 3.07 \times 2050 = \mathbf{-5.5}$ kg

$$6288.0 - 3.07 \times \boldsymbol{x_0} = 0 \iff$$

$$x_0 = \frac{6288.0}{3.07} = 2048.21$$

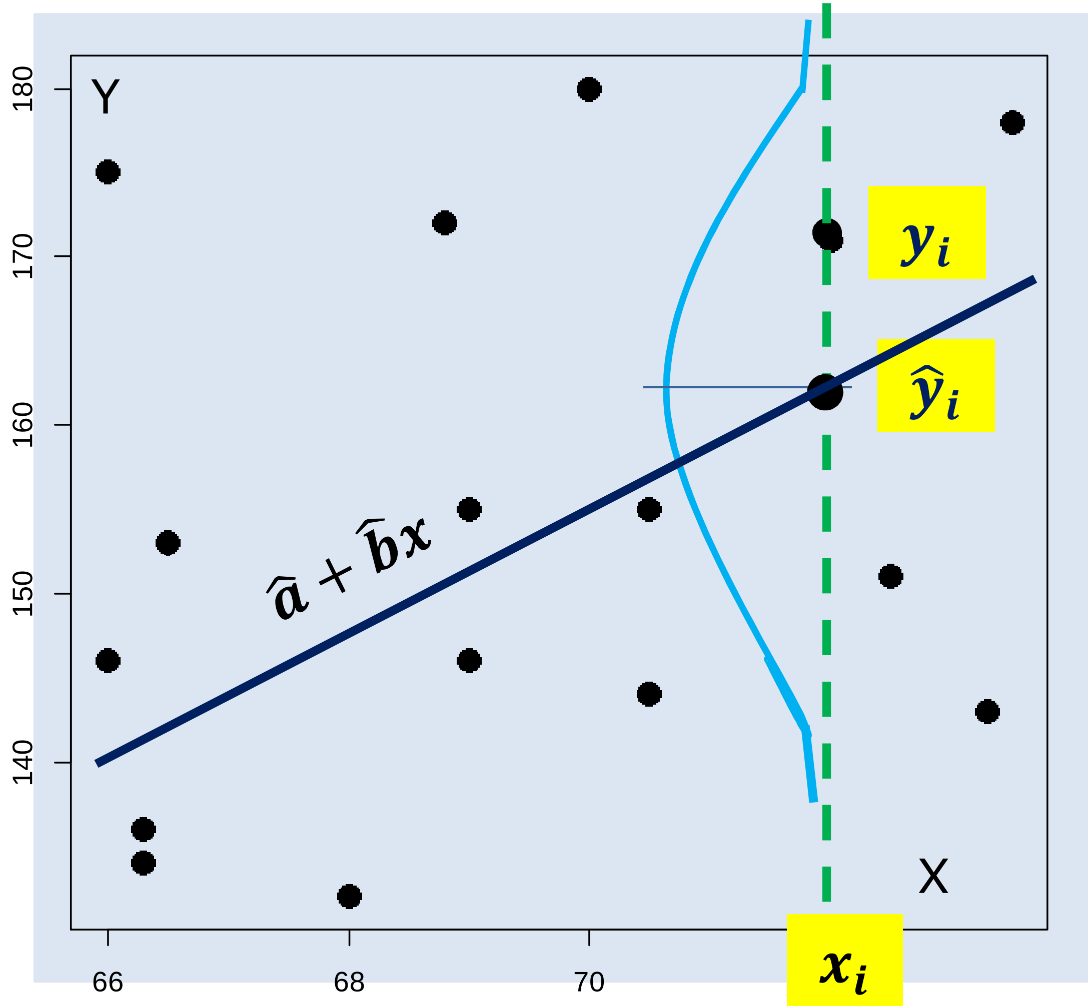
previsione a lungo termine
solo fino al 2048…



Il **modello** di regressione lineare

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

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.



Il **modello** di regressione lineare

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

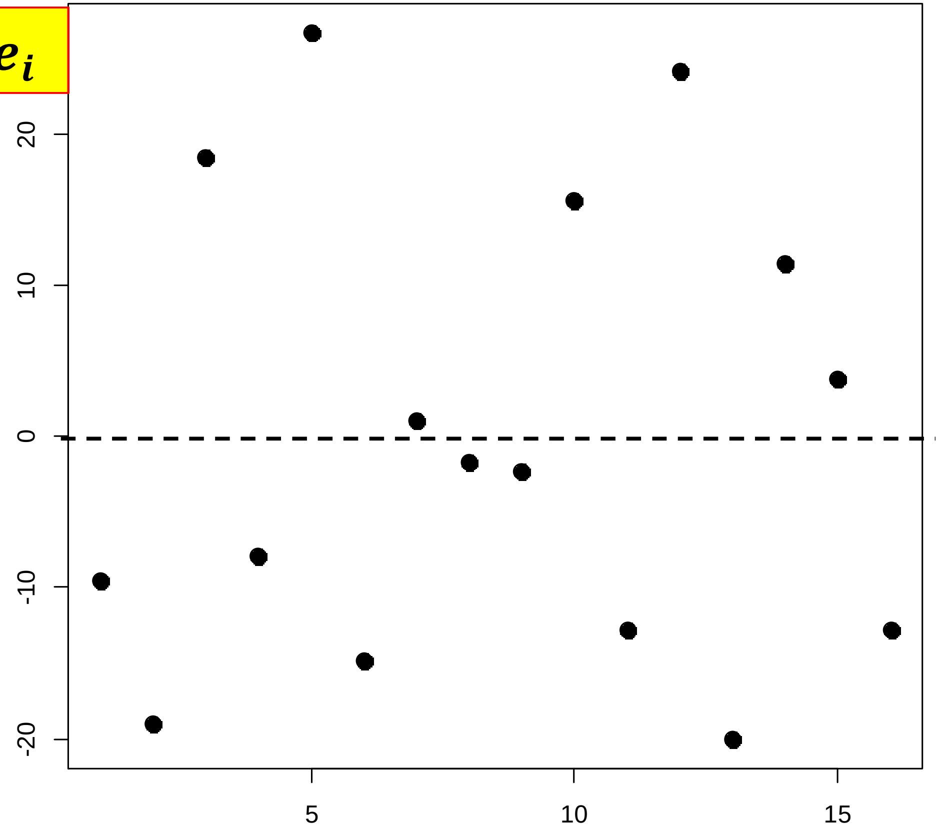
$\Updownarrow$

$y_i - \hat{y}_i$

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

GRAFICO DEI RESIDUI

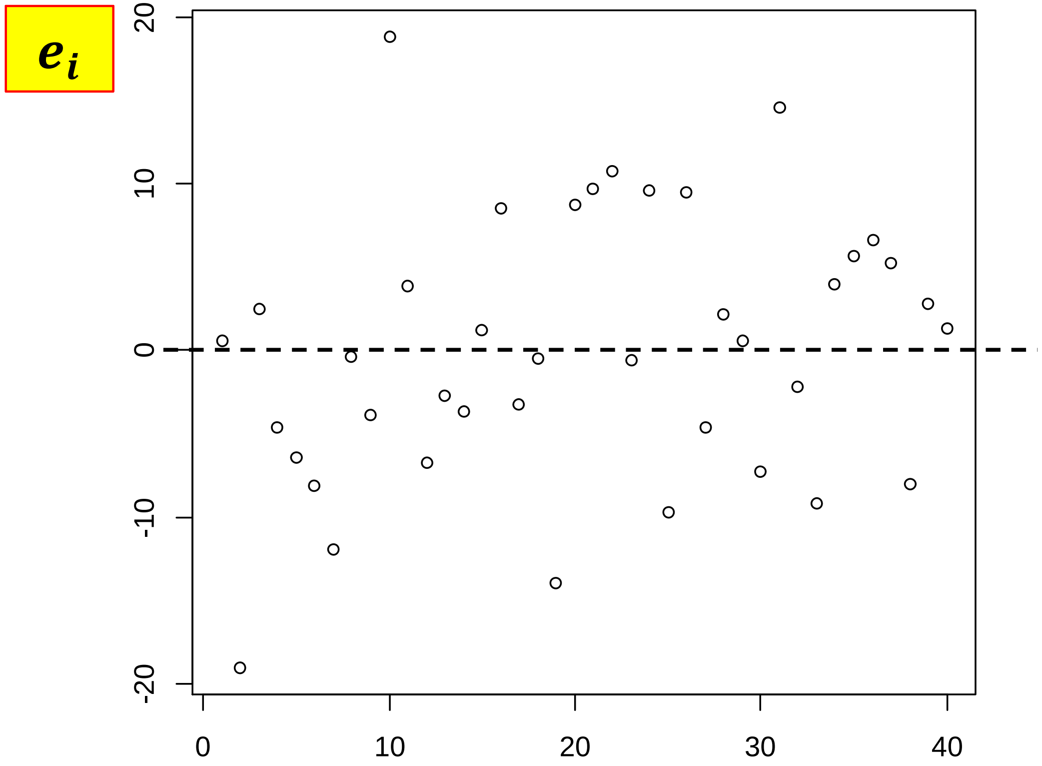
$\boldsymbol{e_i}$



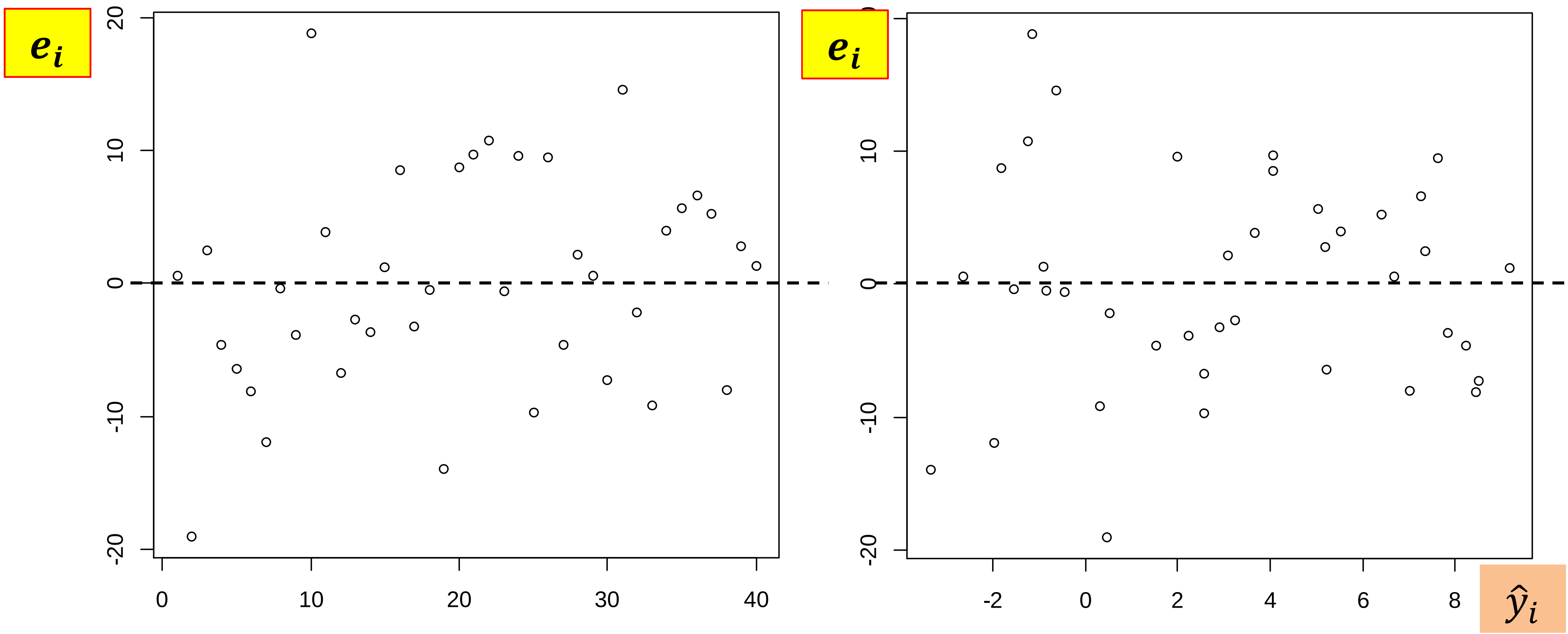
nell'ordine dei dati

- non sono «troppo grandi»: $(-3s^2, +3s^2)$;
- sono in parte positivi e in parte negativi;
- il loro grafico è "sparpagliato".

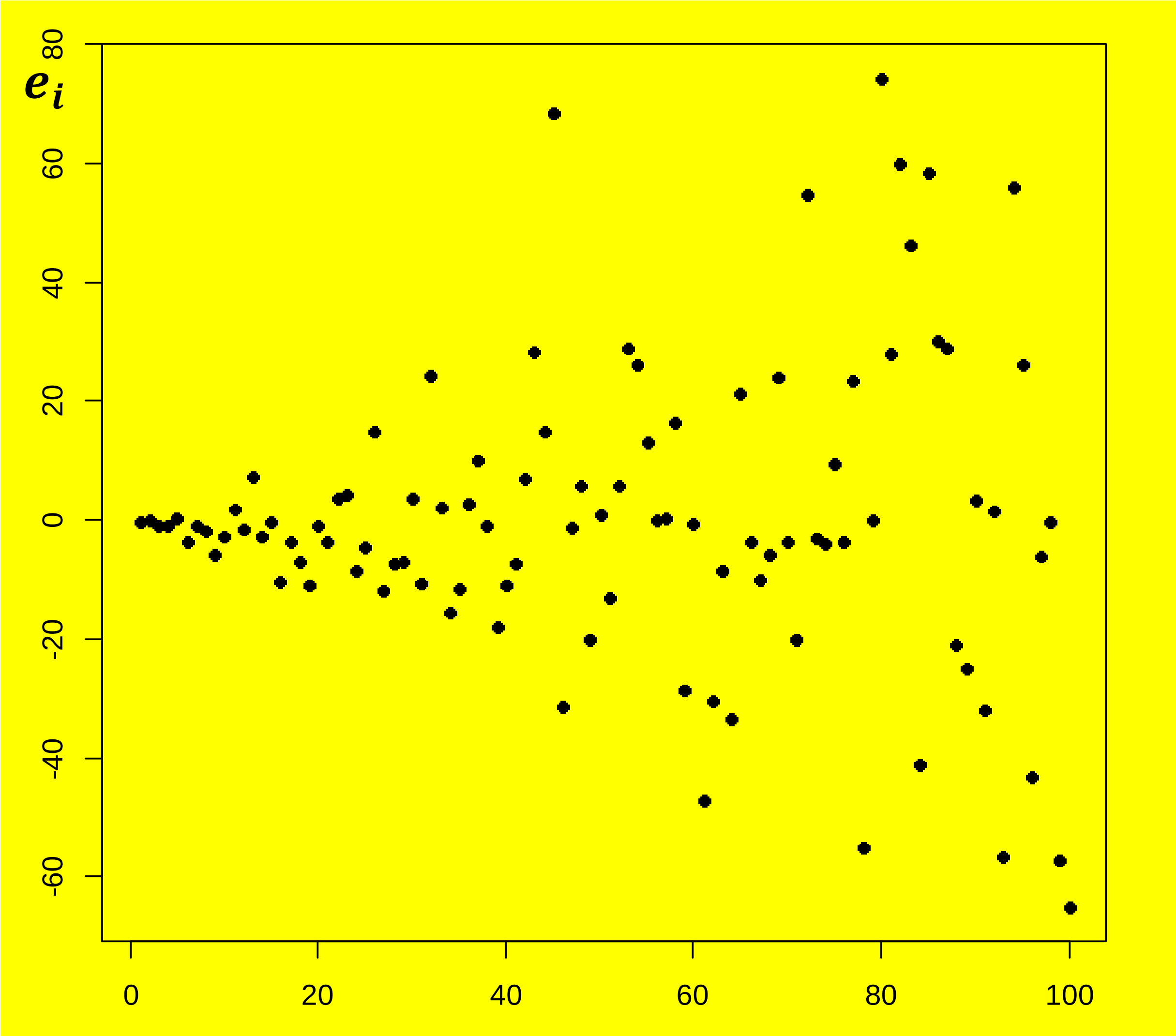
Verifica della Gaussianità



Verifica della Gaussianità

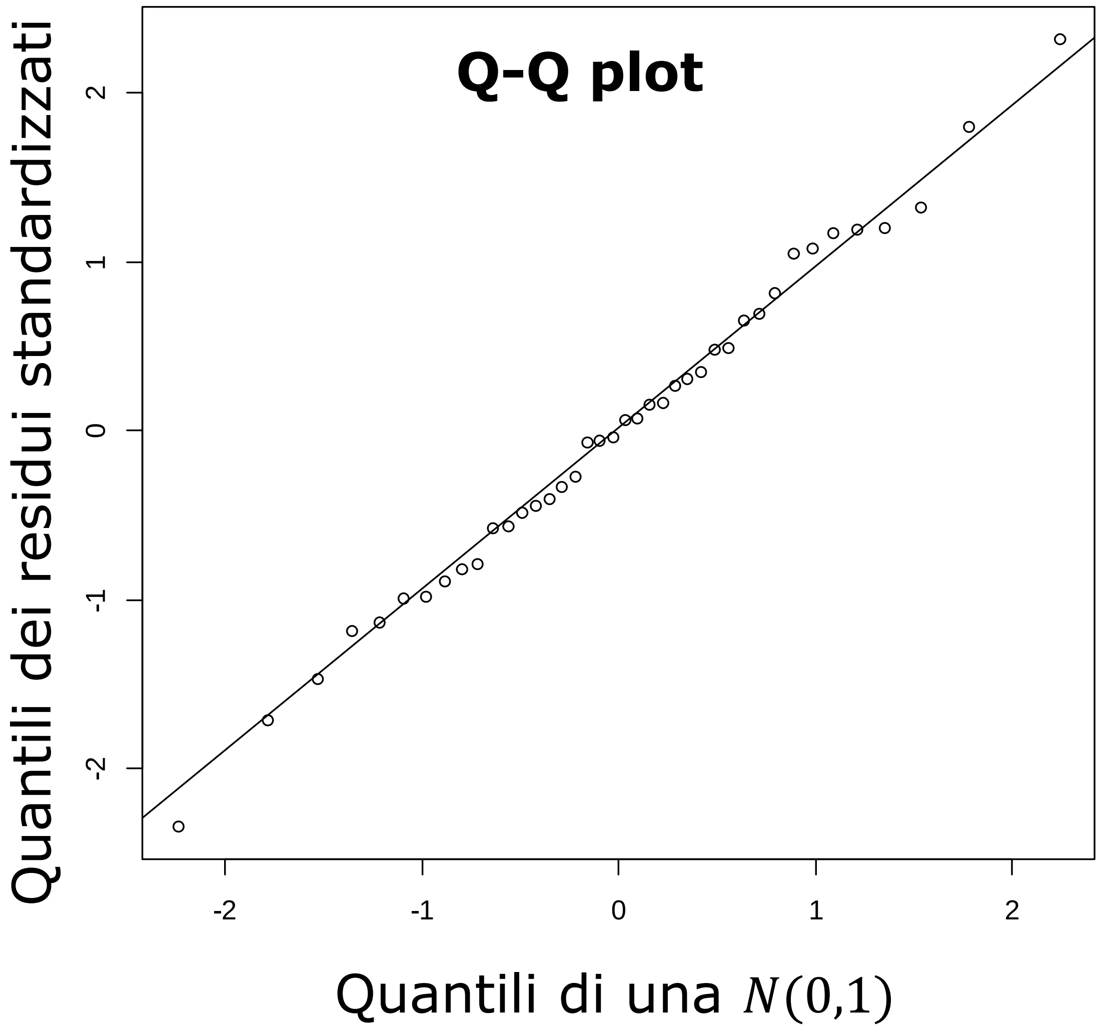


Verifica della Gaussianità

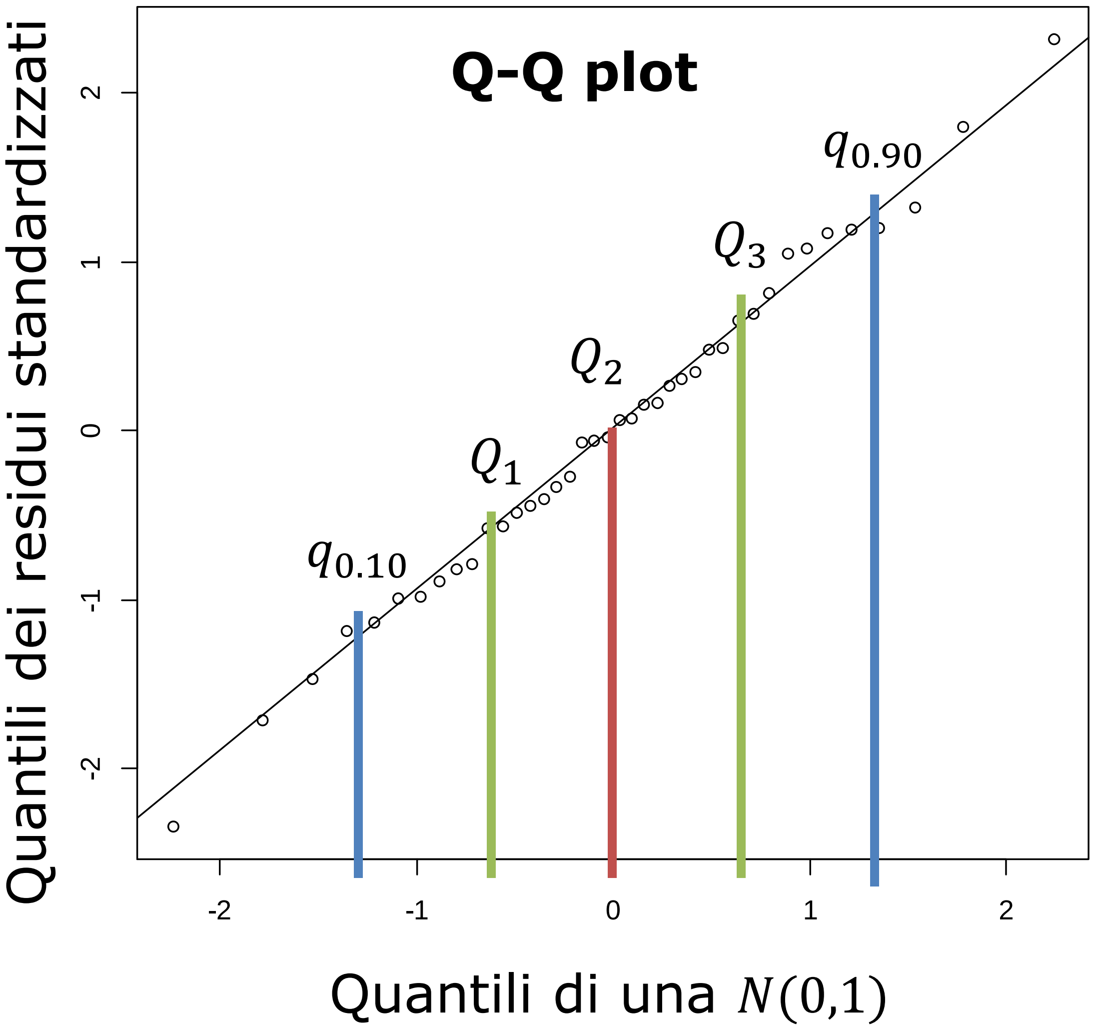


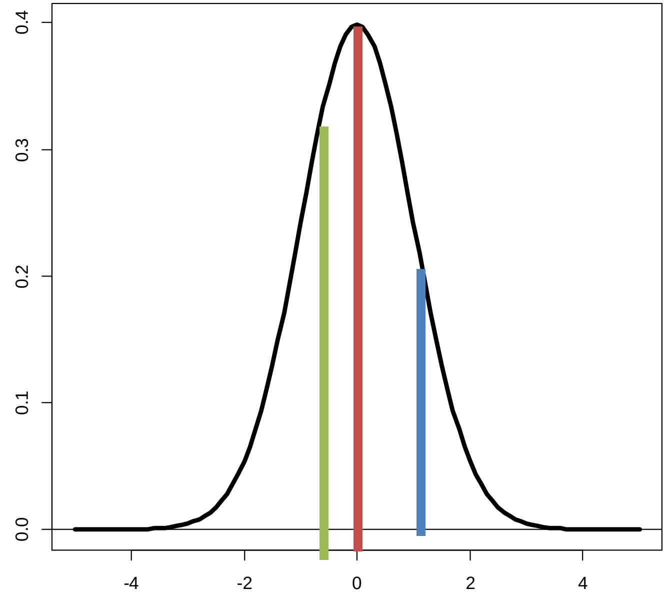
La varianza non
è costante

Verifica della Gaussianità



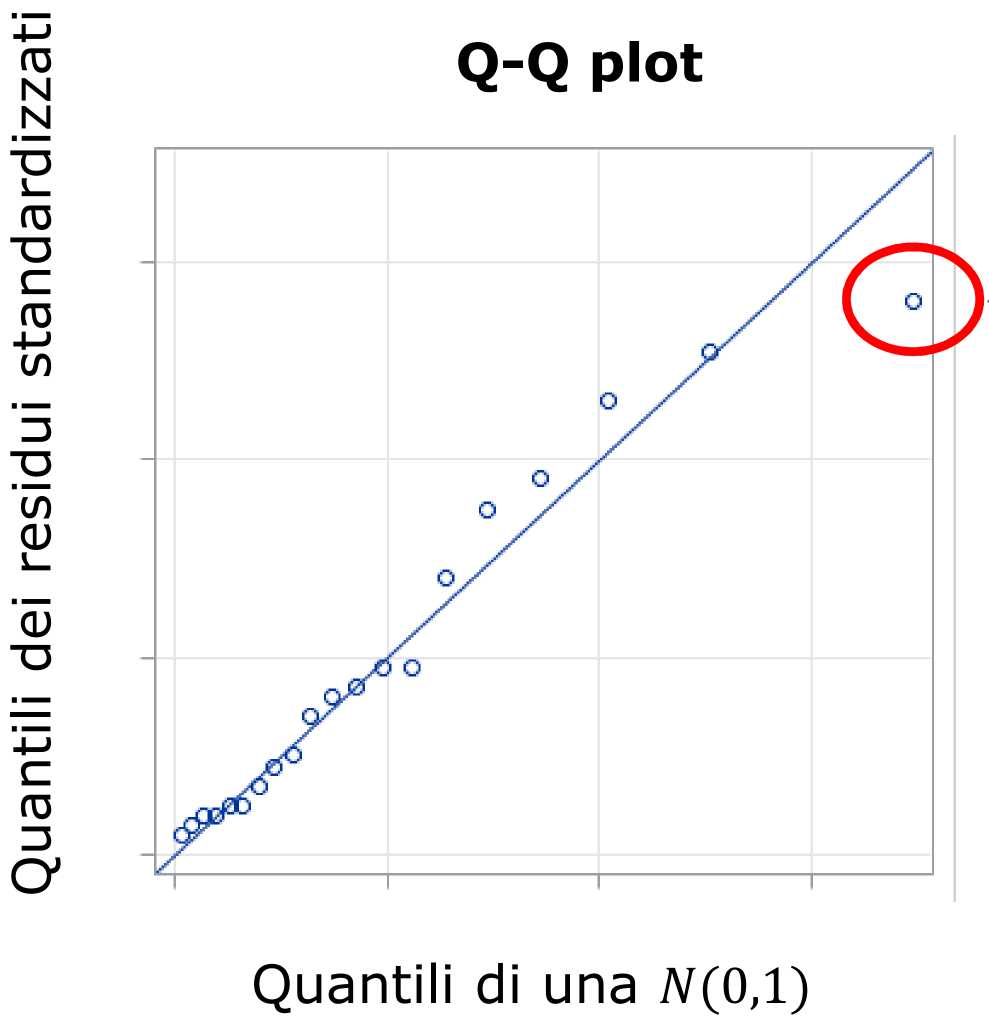
Verifica della Gaussianità





$Q_1 = -0.57$ $q_{0.90} = 1.38$

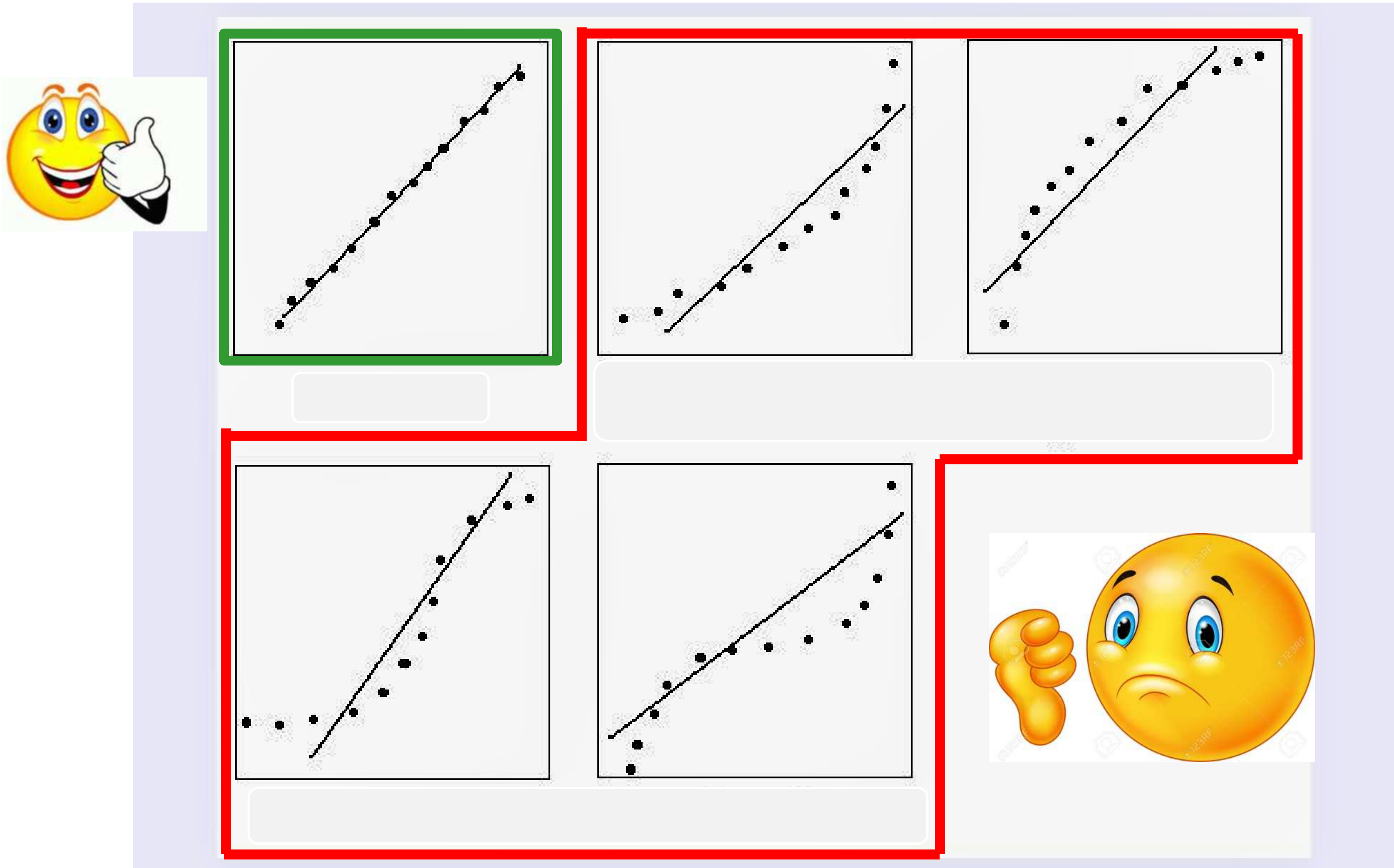
Verifica della Gaussianità



dato anomalo/ outlier

Ci sono tecniche di diagnostica *ad hoc*

Verifica della Gaussianità

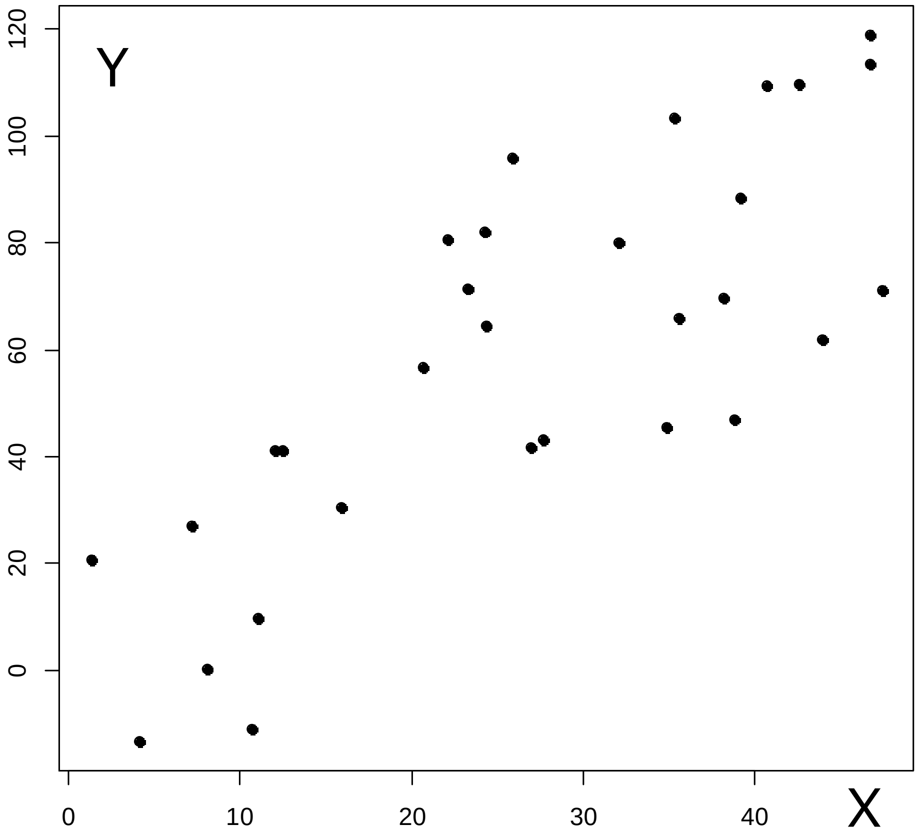


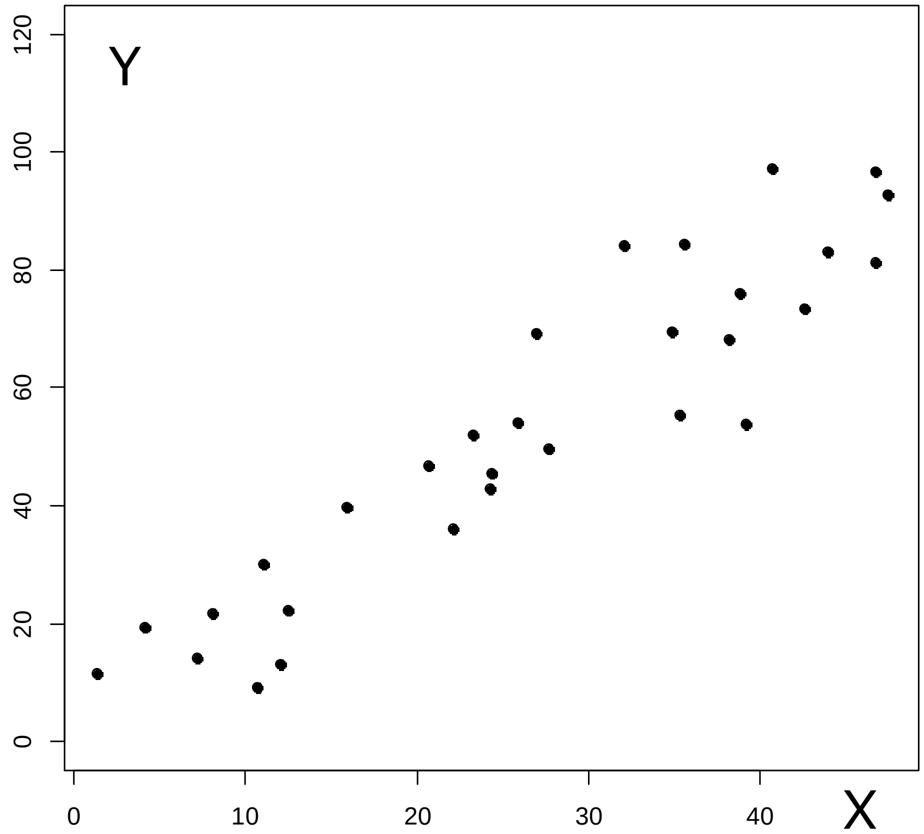
http://condor.depaul.edu/sjost/it223/documents/normal-plot.htm

Esercizio 3

Variabile	Coeff.	Dev. std.	Statistica *t*	*p-value*
Intercetta	3.8199	9.0891	0.420	0.677
X	2.0642	0.3029	6.816	0

$R^2 = 0.624$





Esercizio 3

Variabile	Coeff.	Dev. std.	Statistica *t*	*p-value*
Intercetta	3.8199	9.0891	0.420	0.677
X	2.0642	0.3029	6.816	0

$R^2 = 0.624$

$$y_i = 3.8199 + 2.0642x_i + \varepsilon_i$$

Esercizio 3

Variabile	Coeff.	Dev. std.	Statistica *t*	*p-value*
Intercetta	3.8199	9.0891	0.420	0.677
X	2.0642	0.3029	6.816	0

$R^2 = 0.624$

$$y_i = 3.8199 + 2.0642 x_i + \varepsilon_i$$

valori della statistica per i due test d'ipotesi $H_0 : a = 0$ e $H_0 : b = 0$:

$$\frac{\hat{a}}{\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}} \quad \text{e} \quad \frac{\hat{b}}{\sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$$

Esercizio 3

Variabile	Coeff.	Dev. std.	Statistica *t*	*p-value*
Intercetta	3.8199	9.0891	0.420	0.677
X	2.0642	0.3029	6.816	0

$R^2 = 0.624$

$$y_i = 3.8199 + 2.0642x_i + \varepsilon_i$$

valori del denominatore nella statistica per i due test d'ipotesi

$H_0 : a = 0$ e $H_0 : b = 0$:

$$\frac{\hat{a}}{\sqrt{\boldsymbol{s}^2\left(\frac{\boldsymbol{1}}{\boldsymbol{n}} + \frac{\overline{\boldsymbol{x}}^2}{\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^2}\right)}} \quad \text{e} \quad \frac{\hat{b}}{\sqrt{\frac{\boldsymbol{s}^2}{\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^2}}}$$

Esercizio 3

Variabile	Coeff.	Dev. std.	Statistica *t*	*p-value*
Intercetta	3.8199	9.0891	0.420	0.677
X	2.0642	0.3029	6.816	0

$R^2 = 0.624$

$$\frac{\textbf{Coeff}}{\textbf{Dev. Std}} = \textbf{Statistica } \boldsymbol{t}$$

$$y_i = 3.8199 + 2.0642x_i + \varepsilon_i$$

valori del denominatore nella statistica per i due test d'ipotesi

$H_0 : a = 0$ e $H_0 : b = 0$:

$$\frac{\hat{a}}{\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}} \quad \text{e} \quad \frac{\hat{b}}{\sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$$

Esercizio 3

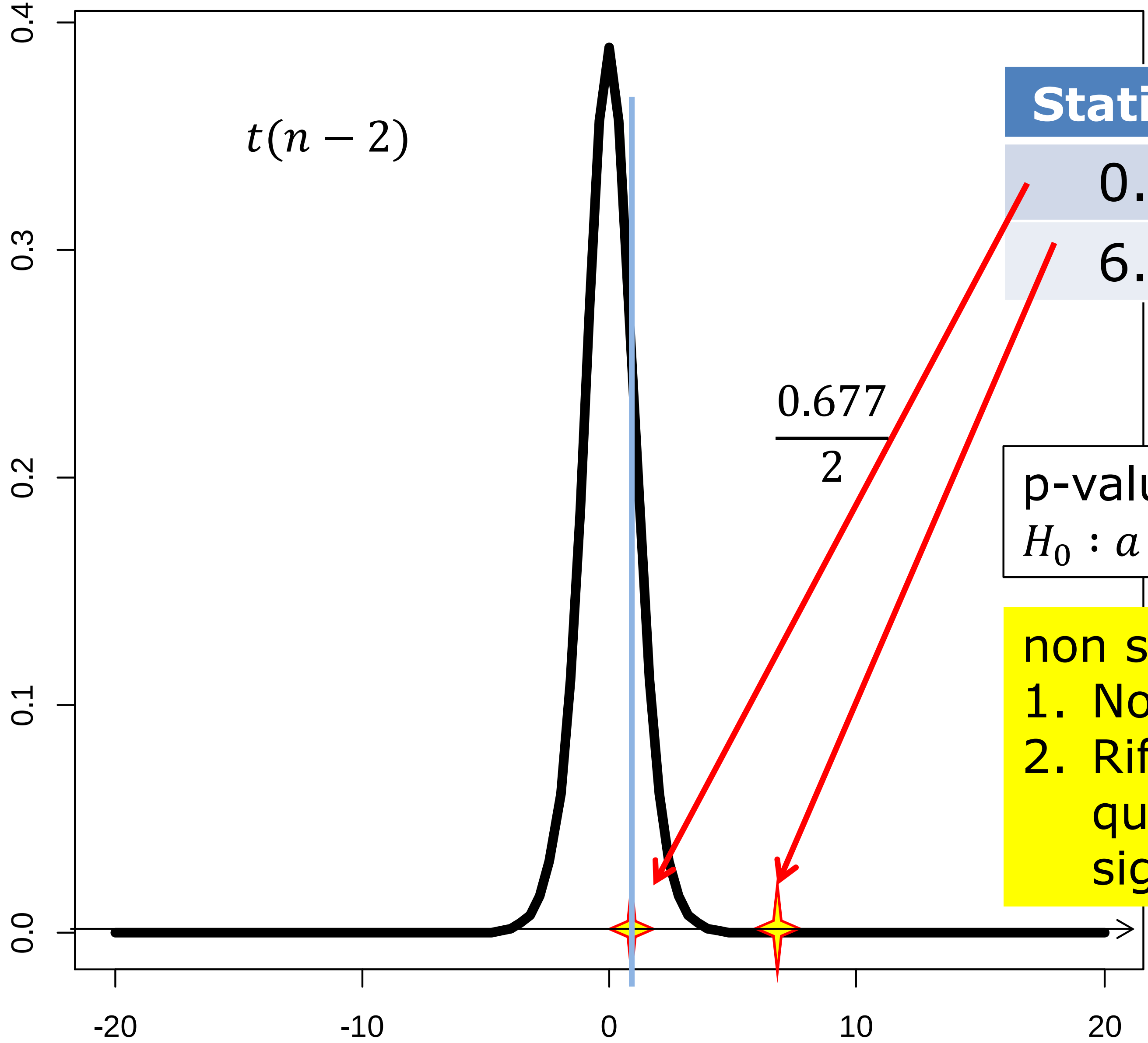
Variabile	Coeff.	Dev. std.	Statistica *t*	*p-value*
Intercetta	3.8199	9.0891	0.420	0.677
X	2.0642	0.3029	6.816	0

$R^2 = 0.624$

p-value per i due test d'ipotesi
$H_0 : a = 0$ e $H_0 : b = 0$

non sappiamo n, però

Esercizio 3



Statistica *t*	*p-value*
0.420	0.677
6.816	0

p-value per i due test d'ipotesi $H_0: a = 0$ e $H_0: b = 0$

non sappiamo n, però:

1. Non rifiutiamo $H_0: \ a = 0$
2. Rifiutiamo $H_0: b = 0$ a qualunque livello di significatività