STATISTICA

Regressione-3

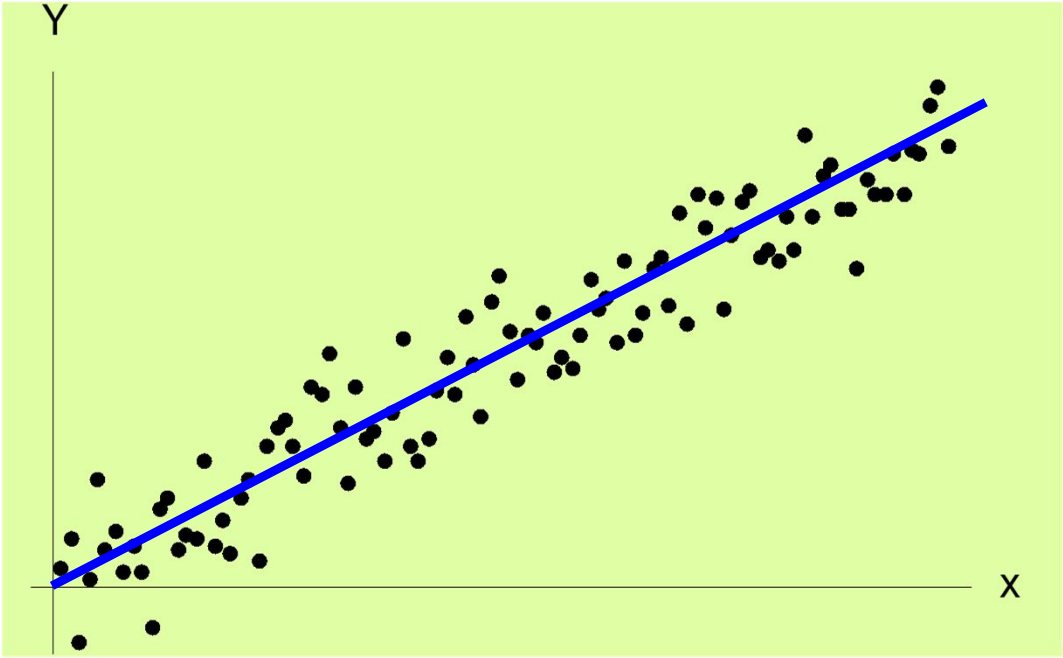
L'inferenza per il modello lineare semplice

La bontà della regressione

Per fare un buon modello lineare serve:

- ✓ una **correlazione alta**, che dice che i dati stanno vicini alla retta
- ✓ alcune **ipotesi** che dicano che il meccanismo che genera i dati è (ragionevolmente) lineare

Inferenza



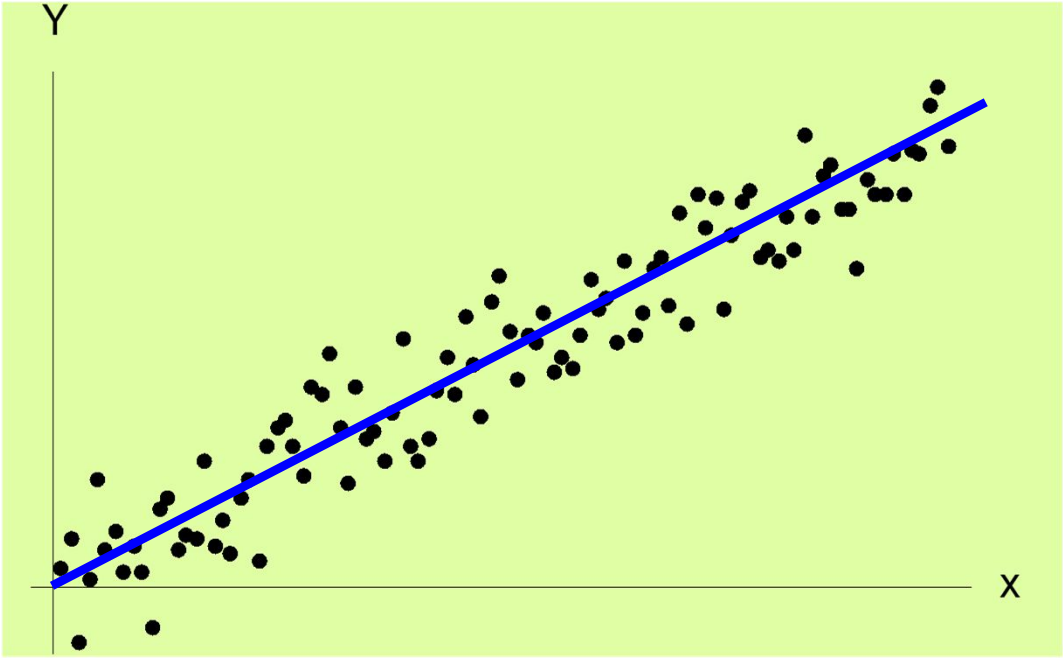
Il modello della **regressione lineare semplice**:

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

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

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

Inferenza



Il modello della **regressione lineare semplice**:

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

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

$$\varepsilon_i \text{ indipendenti}$$

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

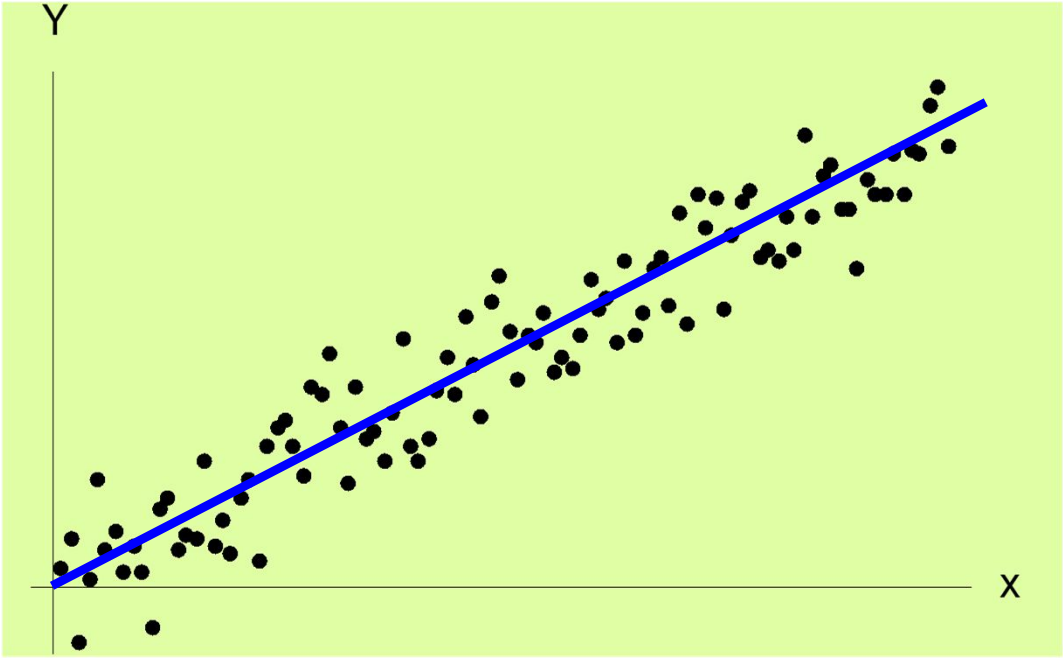
$$Y_i \sim N(a + bx_i, \sigma^2)$$

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**:

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

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

$$\varepsilon_i \text{ indipendenti}$$

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

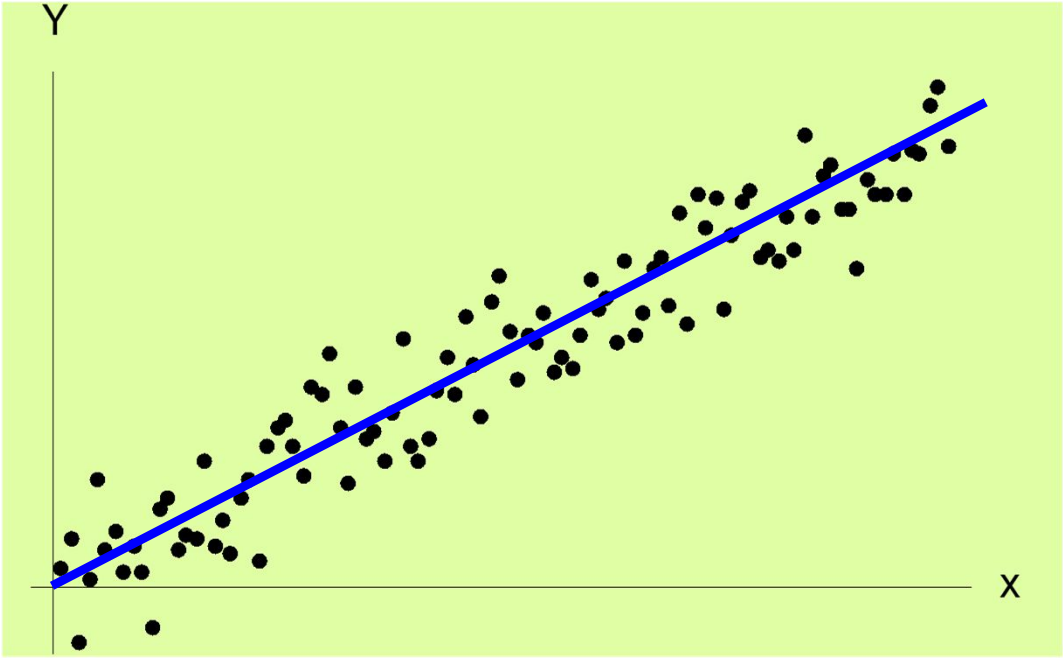
$$a + bx_i$$

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

$$Var(Y_i) = \sigma^2$$

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

Inferenza



Il modello della **regressione lineare semplice**:

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

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

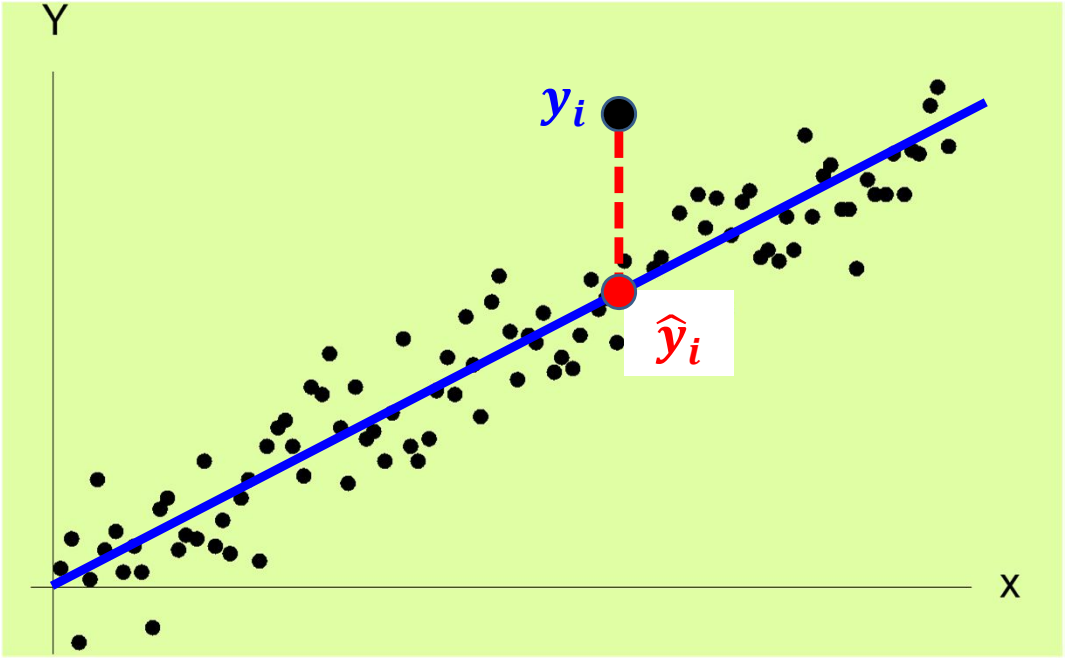
Il modello ha tre parametri incogniti: a, b, σ^2

1. Stimare a, b e 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



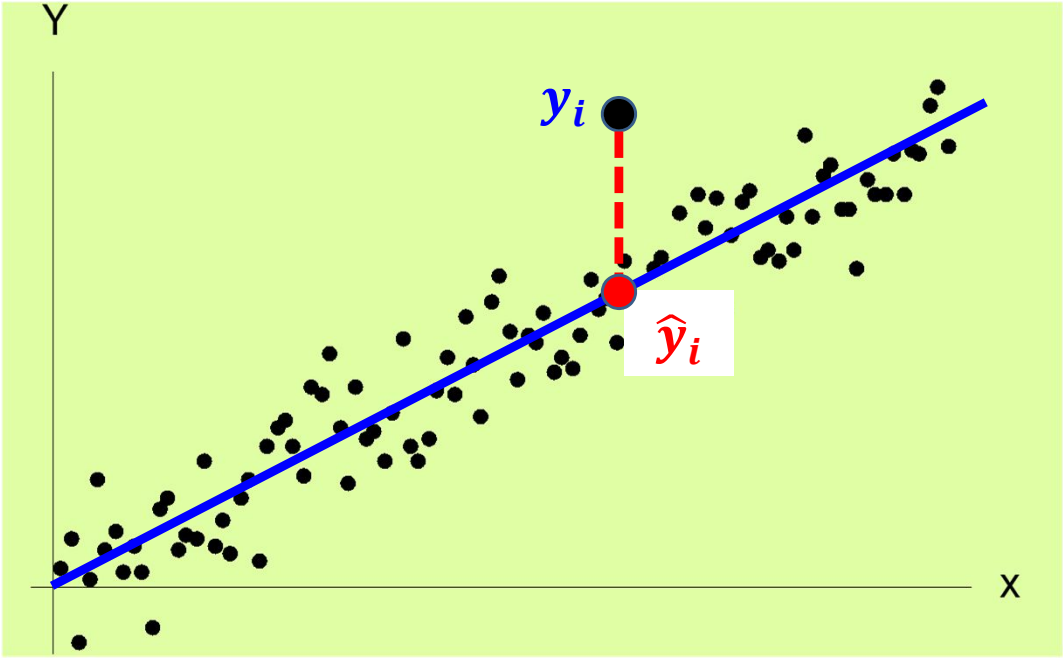
$$f(x) = a + bx$$
$$\approx \Leftrightarrow \ \varepsilon_i \sim N(0, \sigma^2)$$
$$\varepsilon_i \text{ indipendenti}$$
$$Y_i = a + bx_i + \varepsilon_i$$

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

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

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

Inferenza



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

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

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

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

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

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

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

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

rifiutiamo H_0 se:

(rifiutiamo la casualità di una pendenza $\neq 0$)

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

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:

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

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

$\boldsymbol{H_0 : a = 0}$ $\boldsymbol{H_1 : a \neq 0}$

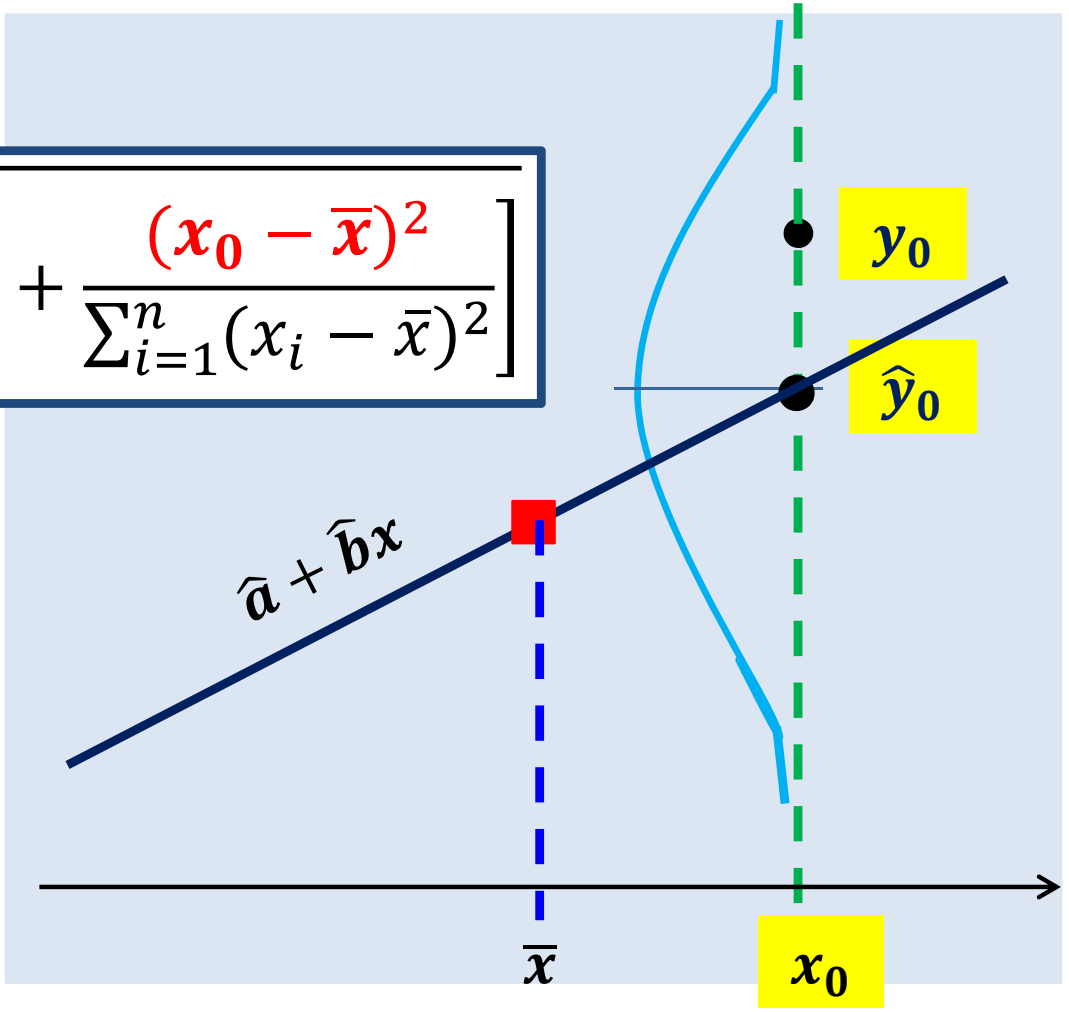
rifiutiamo H_0 se:

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

Inferenza per la previsione

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

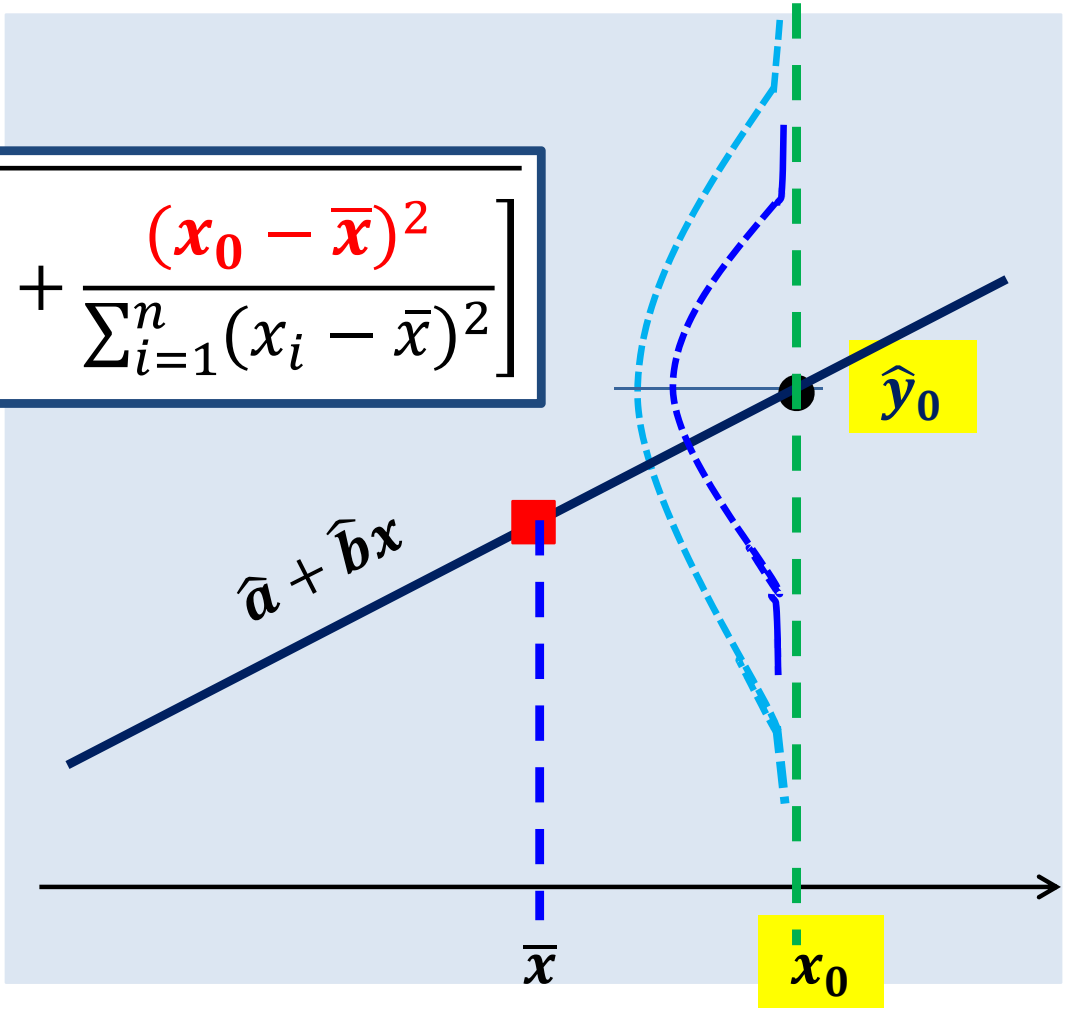
IC della **risposta** di un nuovo "individuo" con covariata pari a x_0



Inferenza per la previsione

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

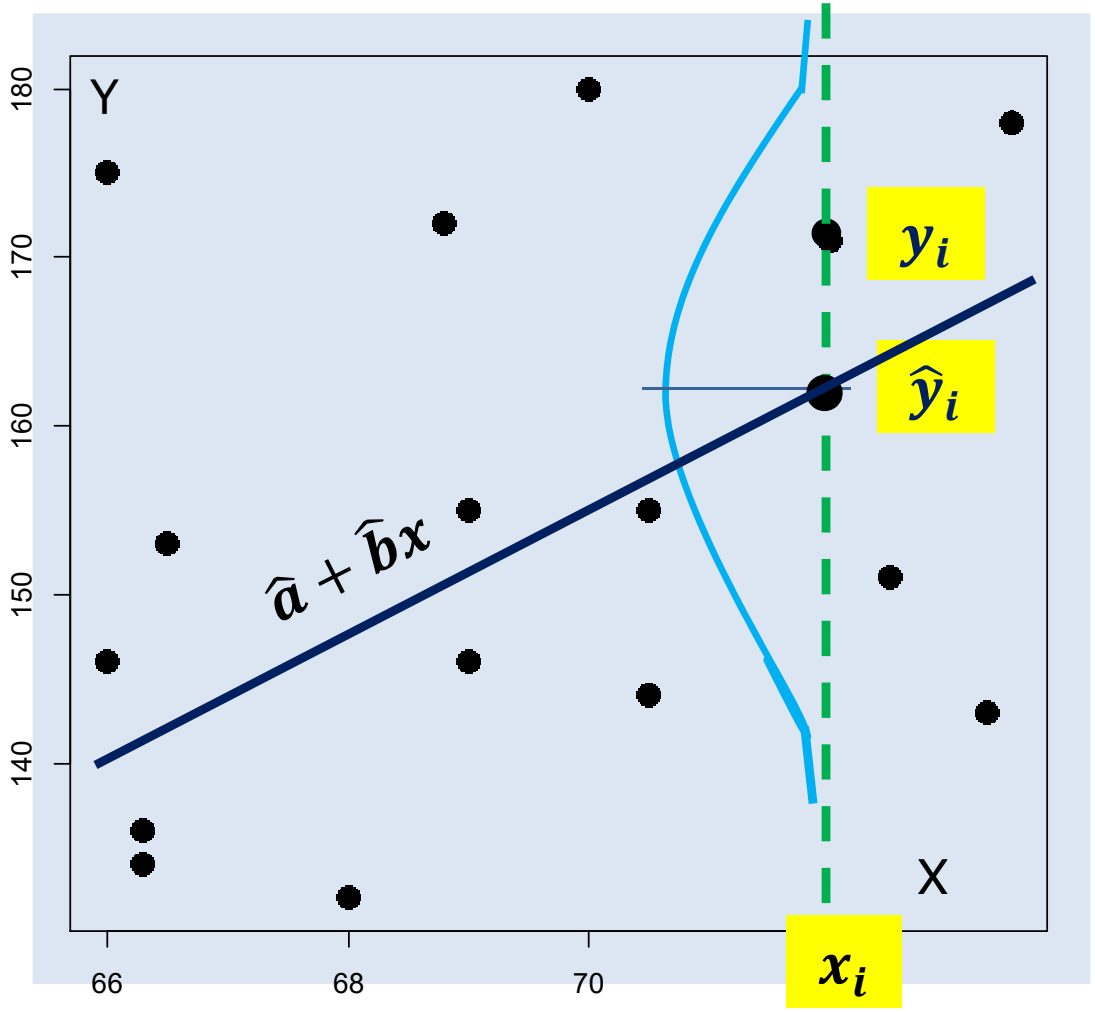
IC della **risposta media** di tutti gli "individui" con covariata pari a x_0



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.



Il **modello** di regressione lineare

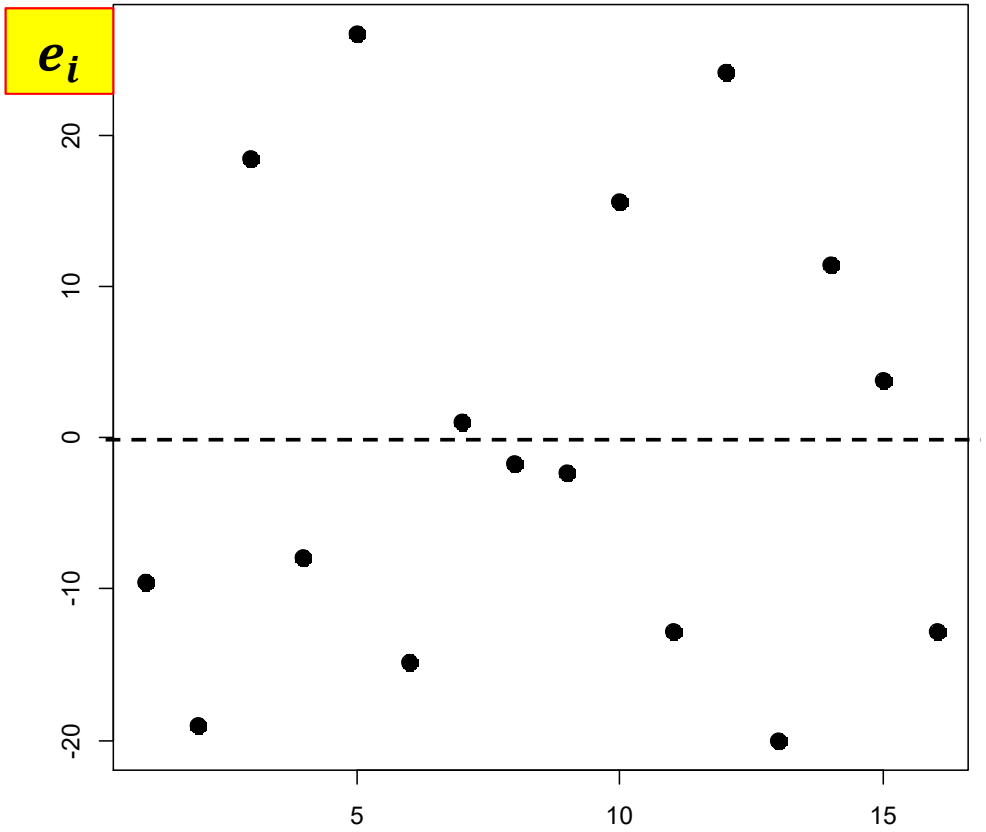
$$Y_i = a + bx_i + \varepsilon_i \quad ,$$

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

$$\boldsymbol{y_i - \hat{y}_i}$$

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

GRAFICO DEI RESIDUI

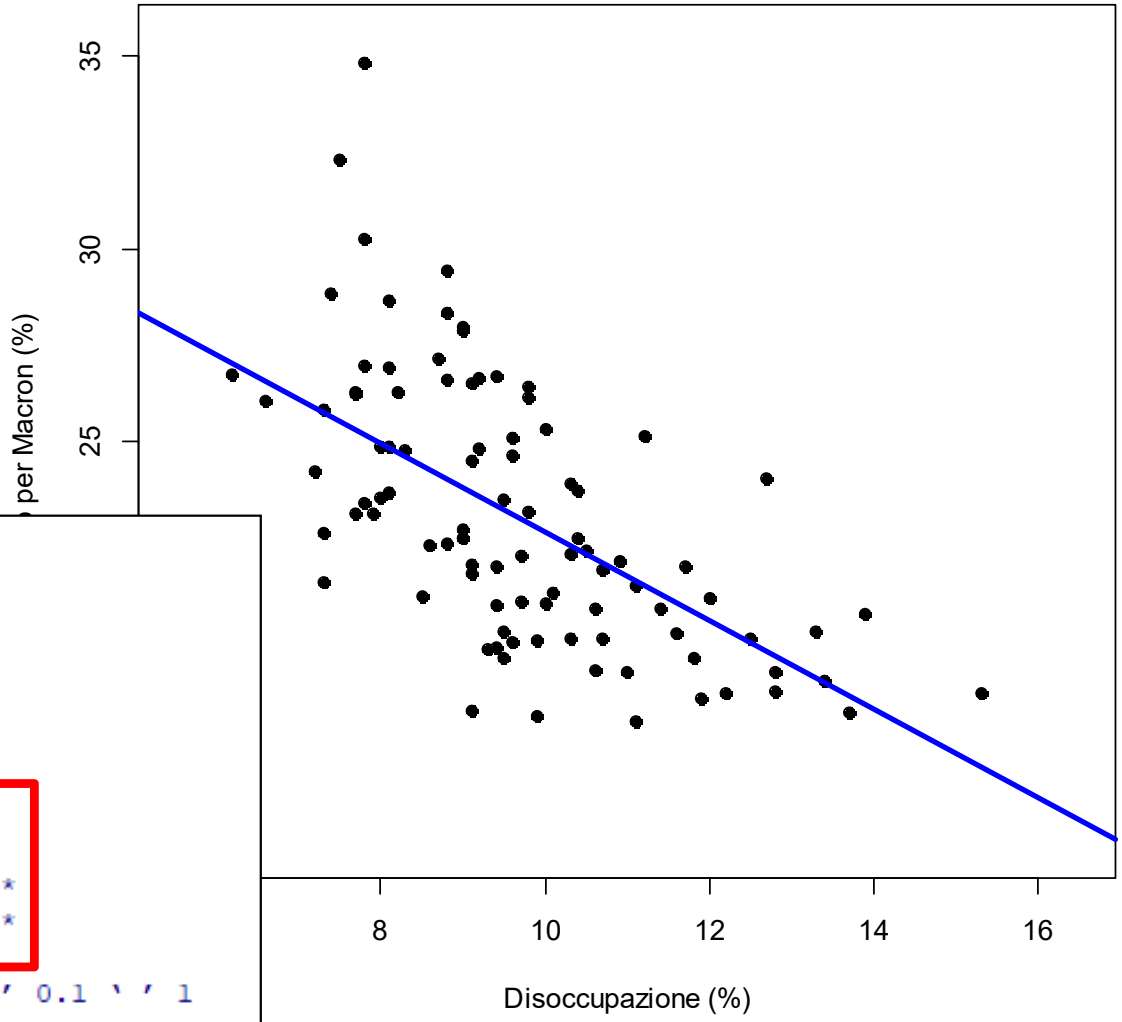


nell'ordine dei dati

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

Facciamo un salto in

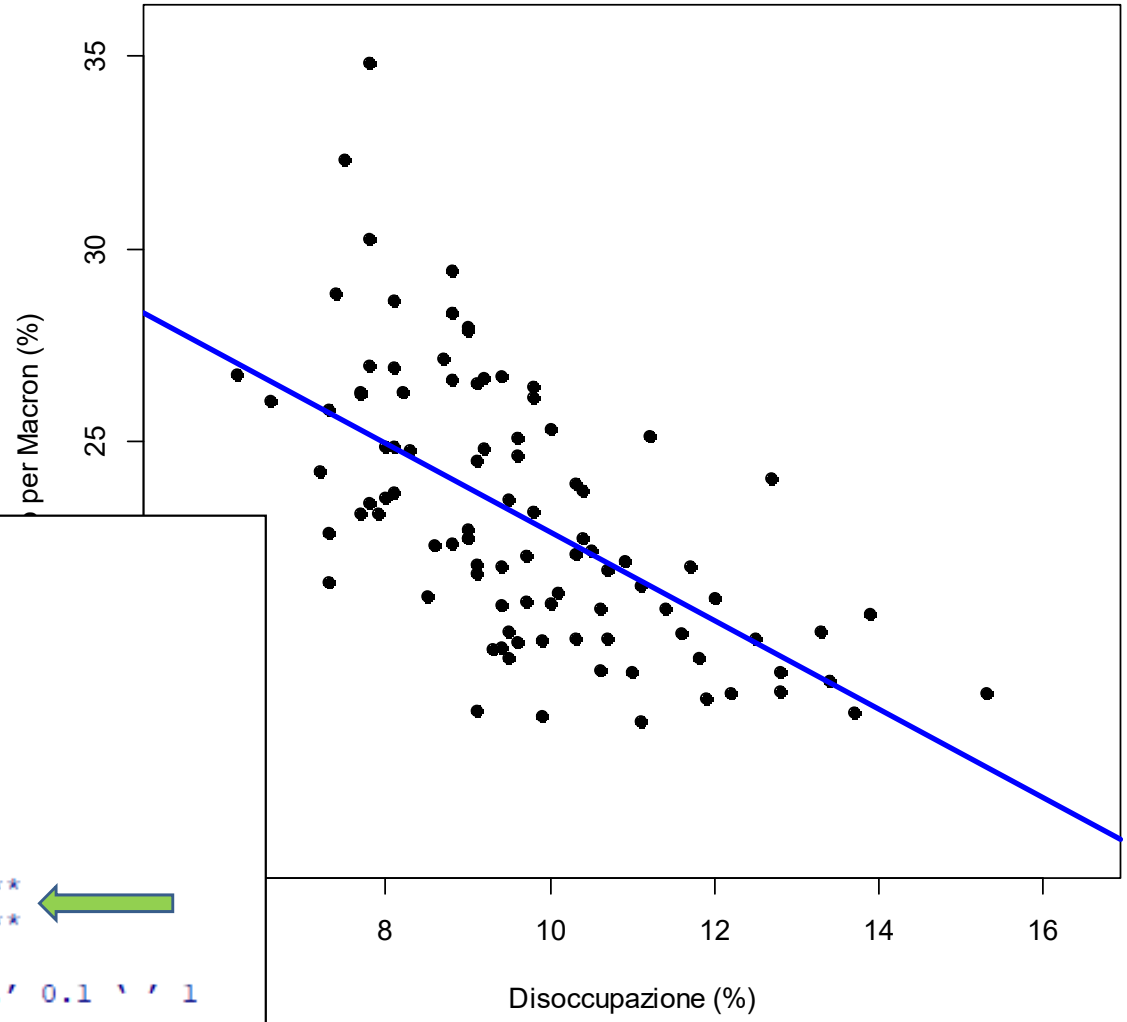
e in Francia!




```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6817 -1.9000 -0.2081  1.6560  9.6499

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.1702     1.5616   21.88  < 2e-16 ***
X            -1.1526     0.1592   -7.24 1.21e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.779 on 94 degrees of freedom
Multiple R-squared:  0.358,     Adjusted R-squared:  0.3512
F-statistic: 52.42 on 1 and 94 DF,  p-value: 1.212e-10
```

Facciamo un salto in

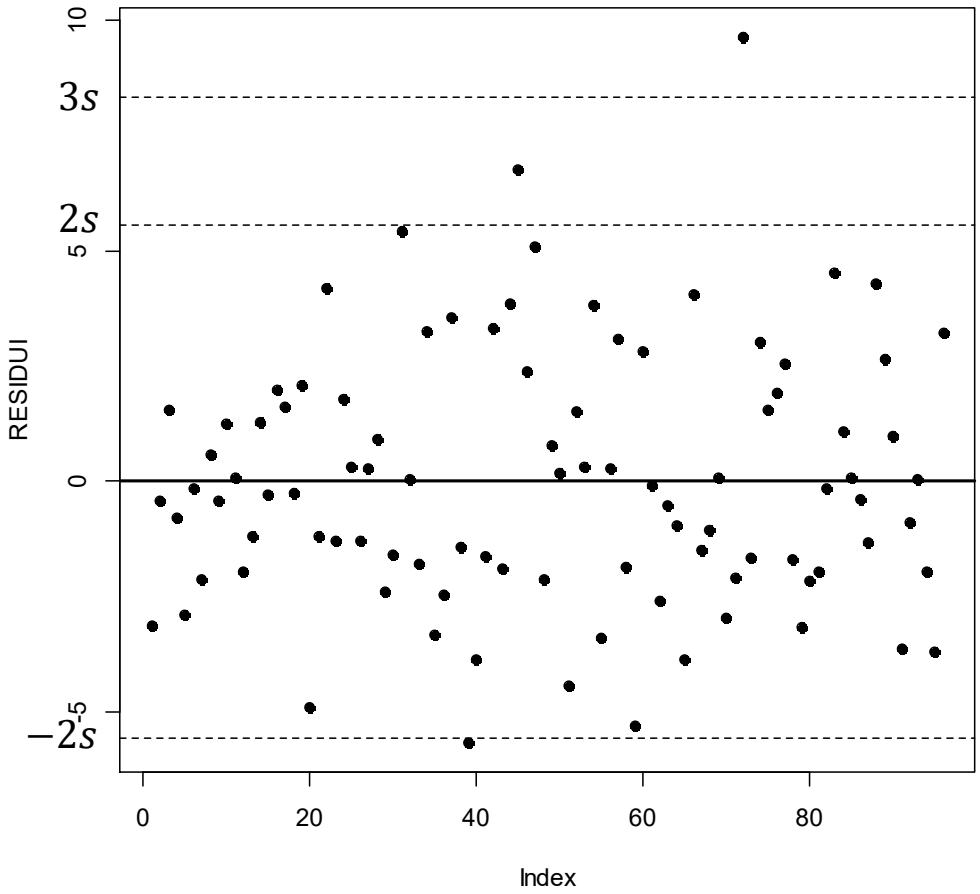
e in Francia!




```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6817 -1.9000 -0.2081  1.6560  9.6499

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.1702     1.5616   21.88  < 2e-16 ***
X            -1.1526     0.1592   -7.24 1.21e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.779 on 94 degrees of freedom
Multiple R-squared:  0.358,     Adjusted R-squared:  0.3512
F-statistic: 52.42 on 1 and 94 DF,  p-value: 1.212e-10
```

Facciamo un salto in

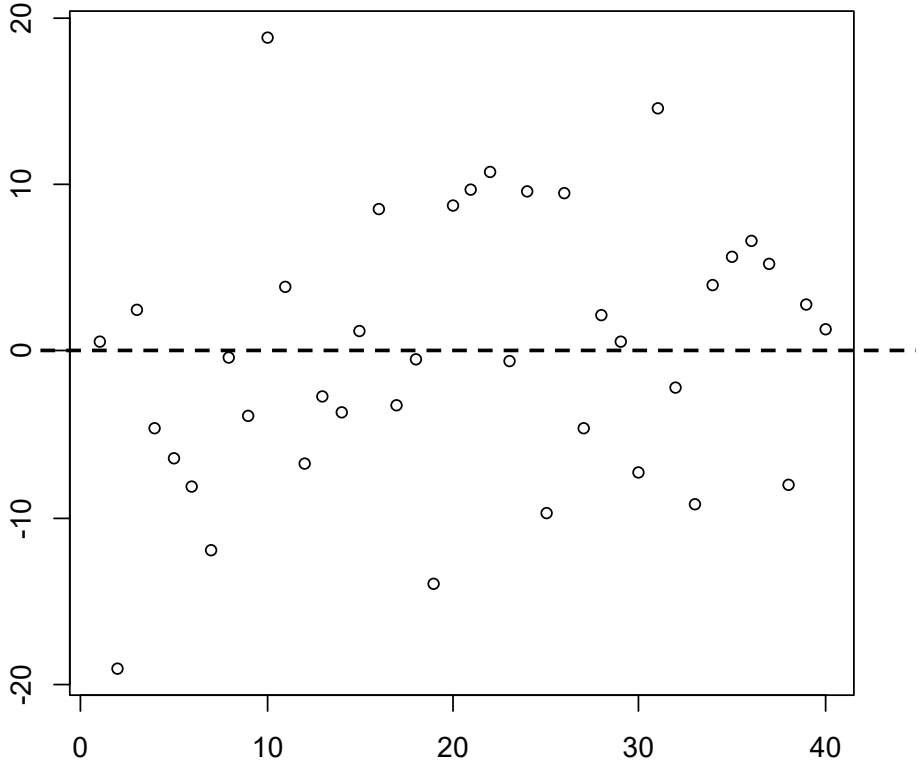
e in Francia!



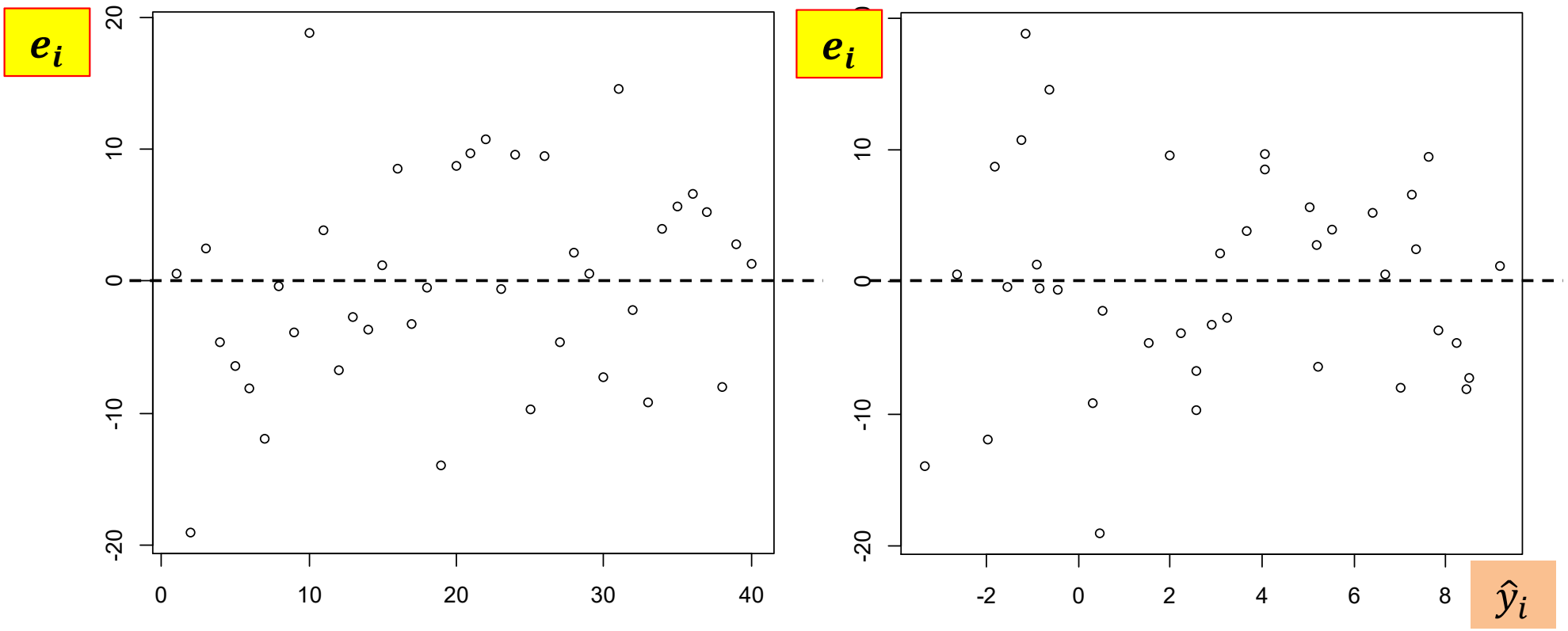

```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6817 -1.9000 -0.2081  1.6560  9.6499

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.1702     1.5616   21.88  < 2e-16 ***
X            -1.1526     0.1592   -7.24 1.21e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.779 on 94 degrees of freedom
Multiple R-squared:  0.358,     Adjusted R-squared:  0.3512
F-statistic: 52.42 on 1 and 94 DF,  p-value: 1.212e-10
```

Verifica della Gaussianità

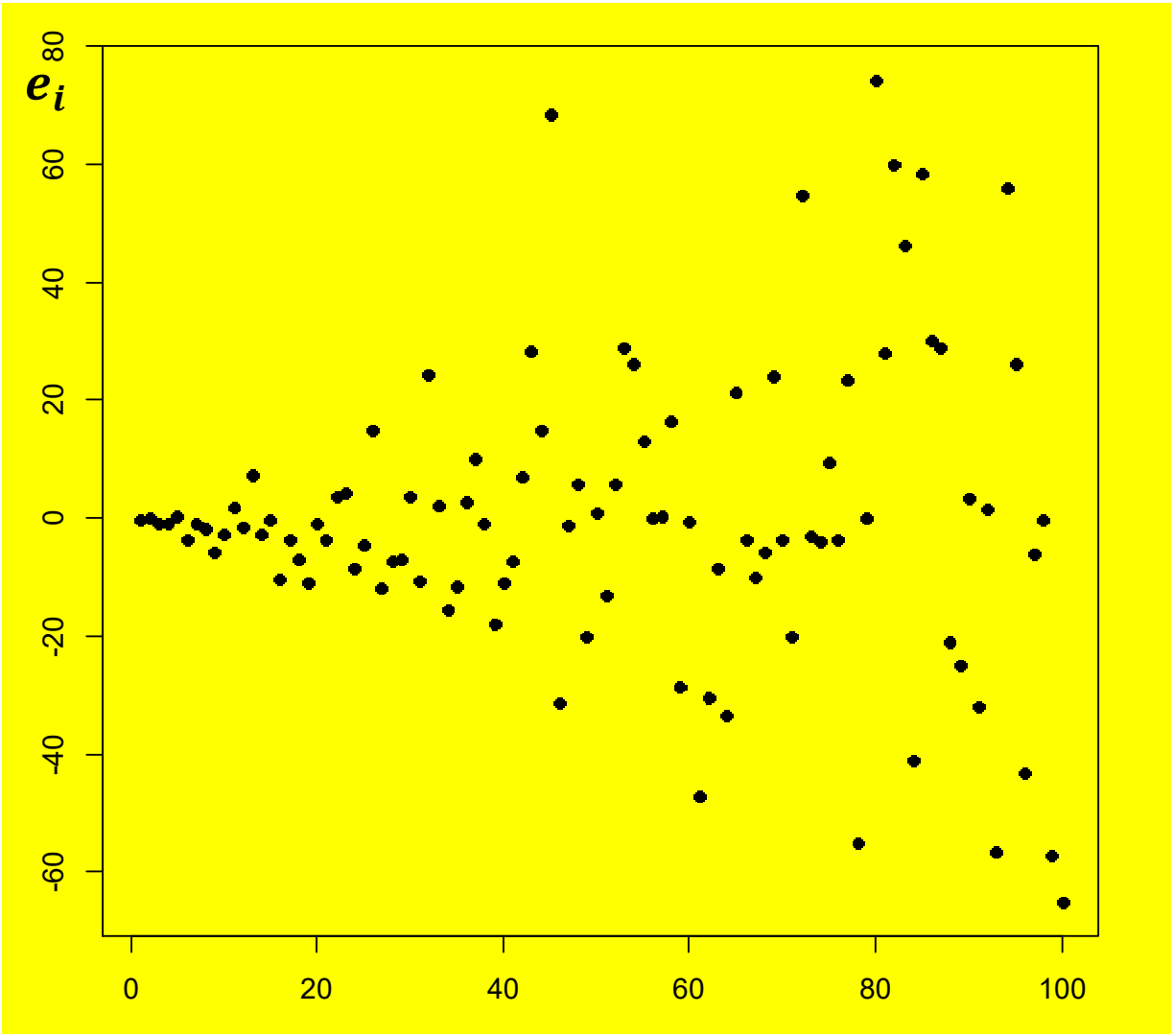
e_i



Verifica della Gaussianità

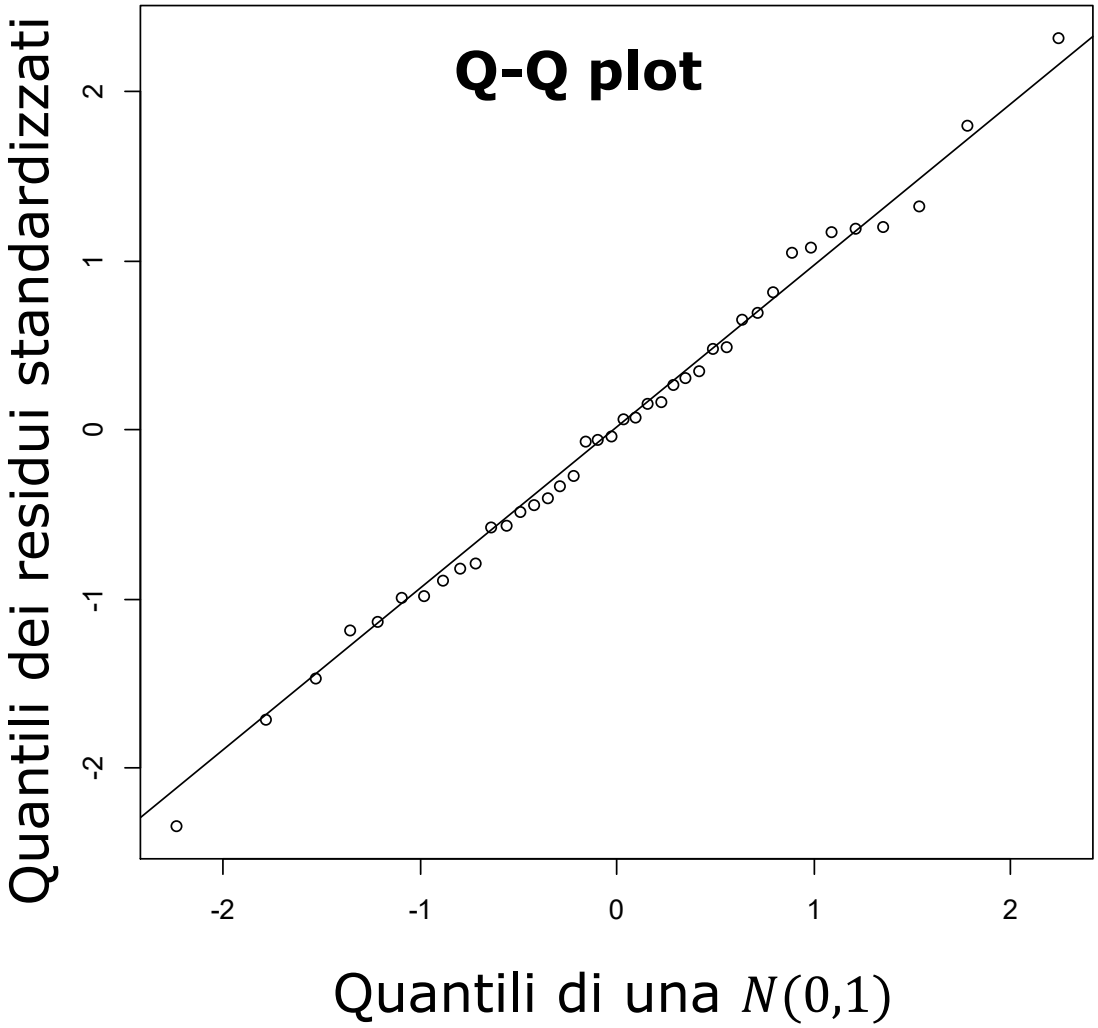


Verifica della Gaussianità

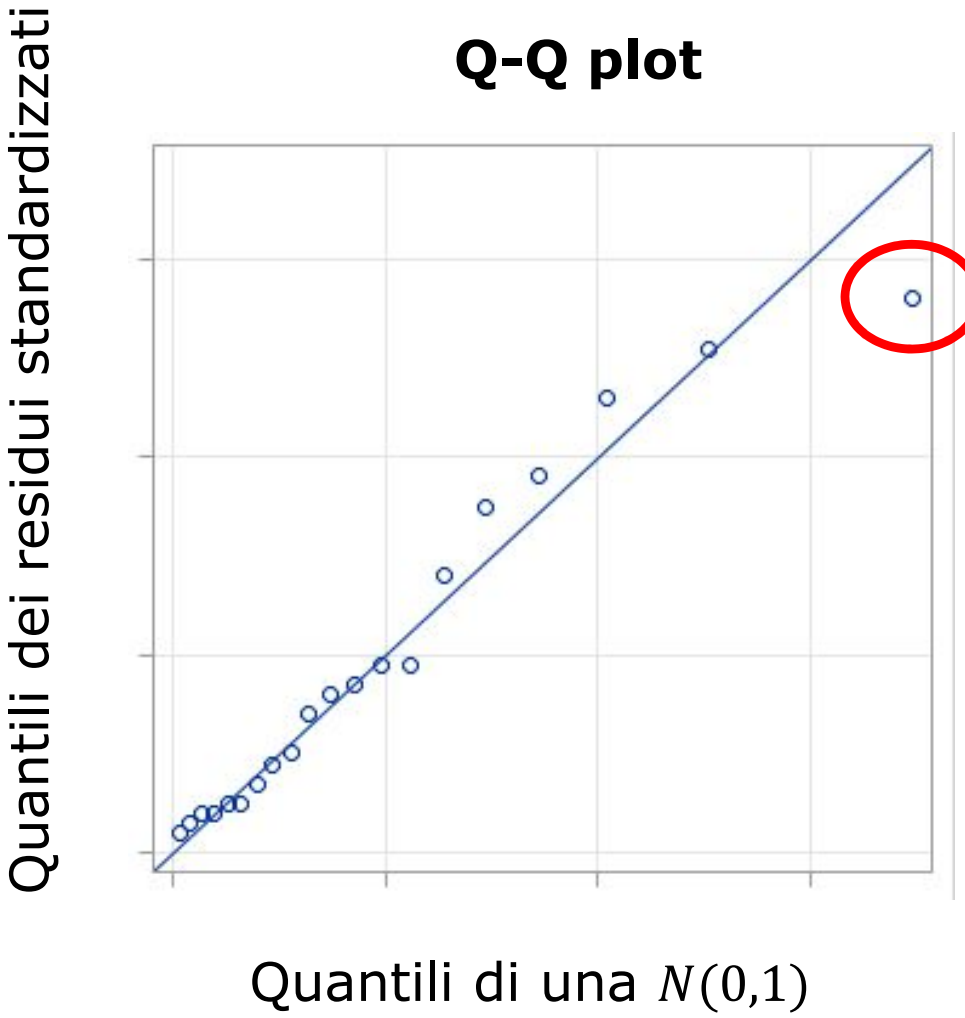


La varianza non è costante

Verifica della Gaussianità



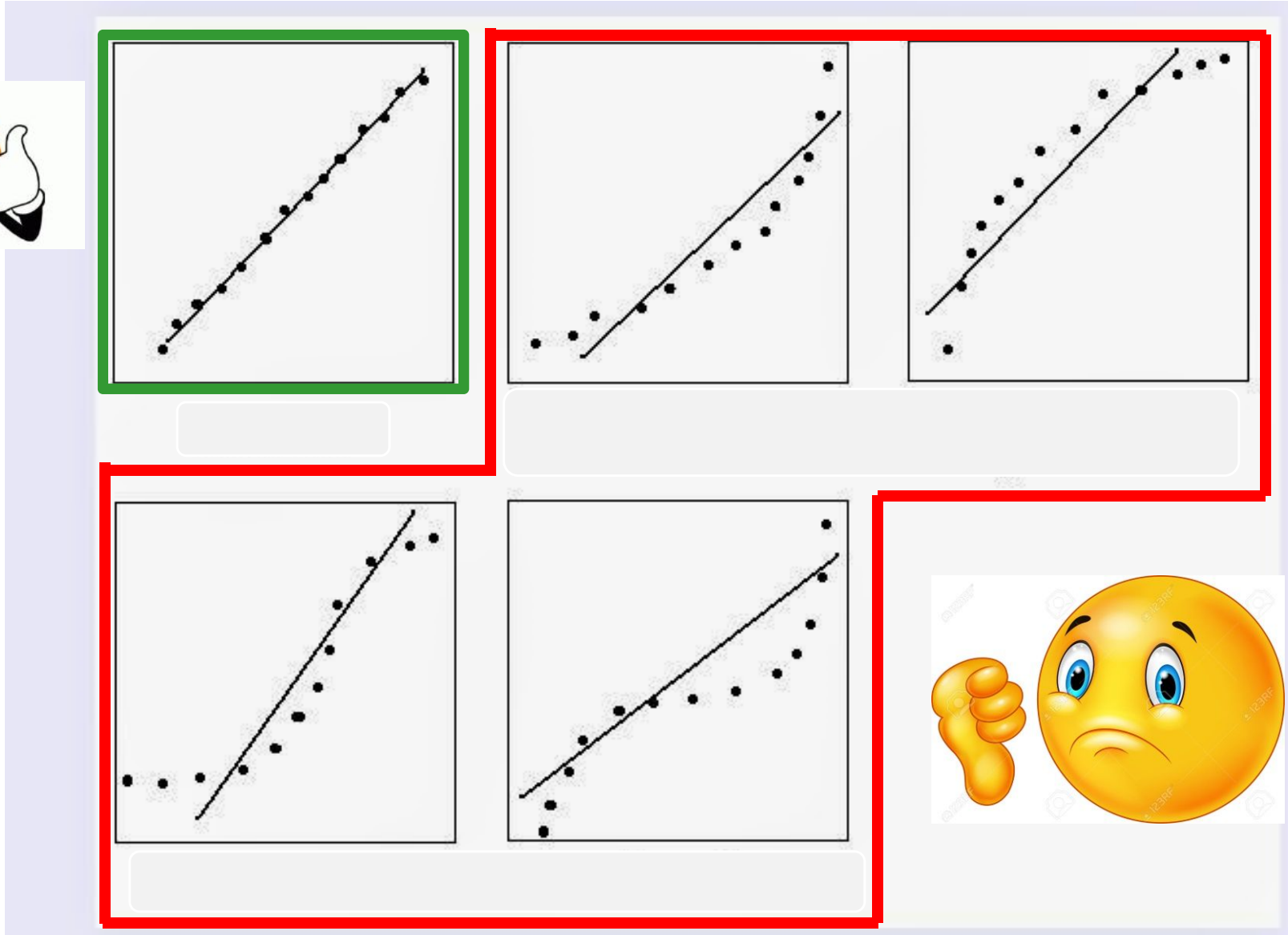
Verifica della Gaussianità



dato anomalo/ outlier

Ci sono tecniche di diagnostica *ad hoc*

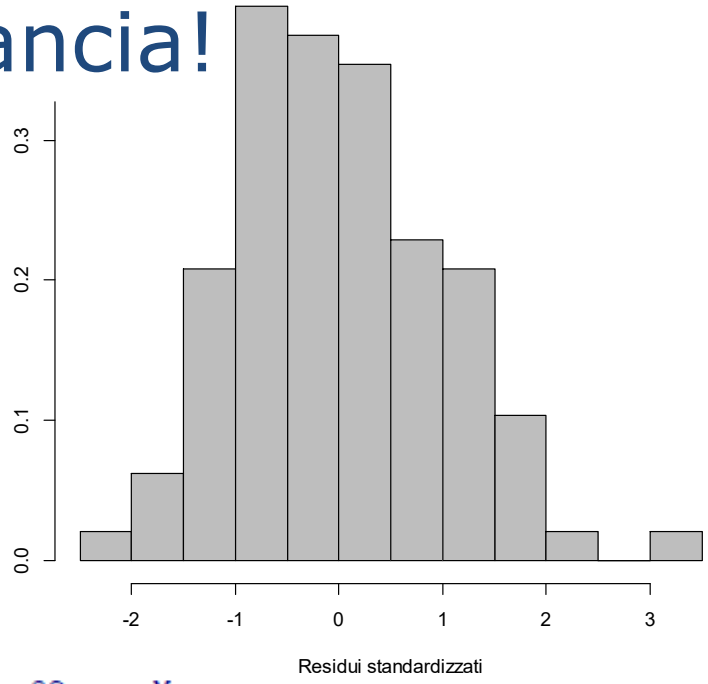
Verifica della Gaussianità



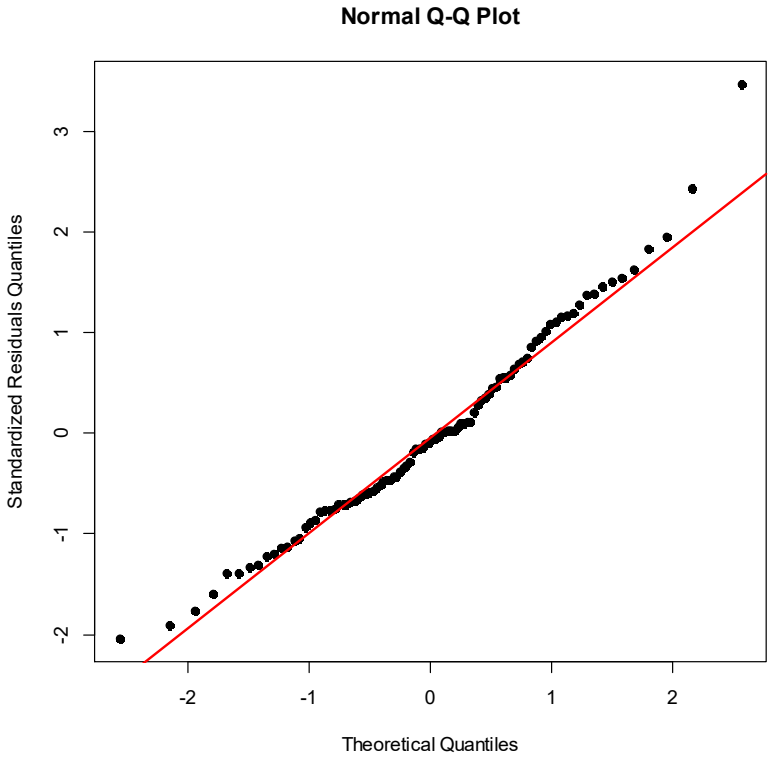
http://condor.depaul.edu/sjost/it223/documents/normal-plot.htm

Facciamo un salto in

e in Francia!

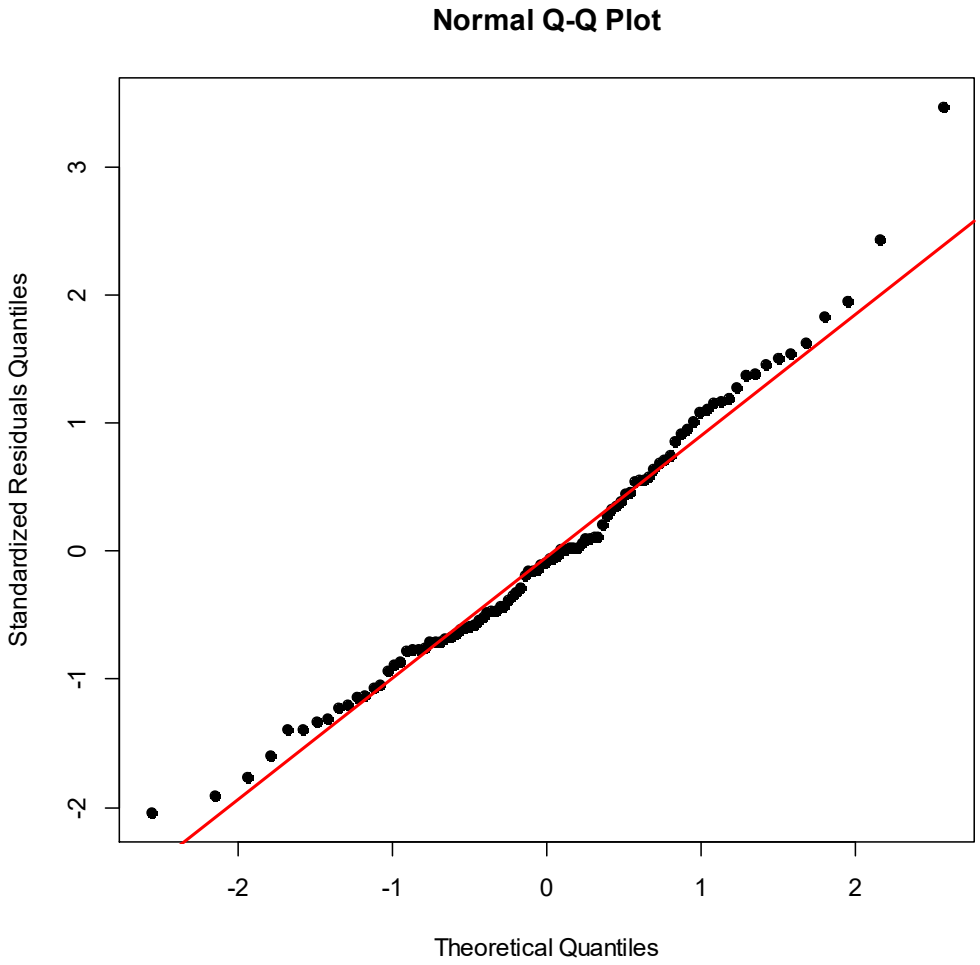



```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6817 -1.9000 -0.2081  1.6560  9.6499

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.1702     1.5616   21.88  < 2e-16 ***
X            -1.1526     0.1592   -7.24 1.21e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.779 on 94 degrees of freedom
Multiple R-squared:  0.358,     Adjusted R-squared:  0.3512
F-statistic: 52.42 on 1 and 94 DF,  p-value: 1.212e-10
```



Facciamo un salto in

e in Francia!



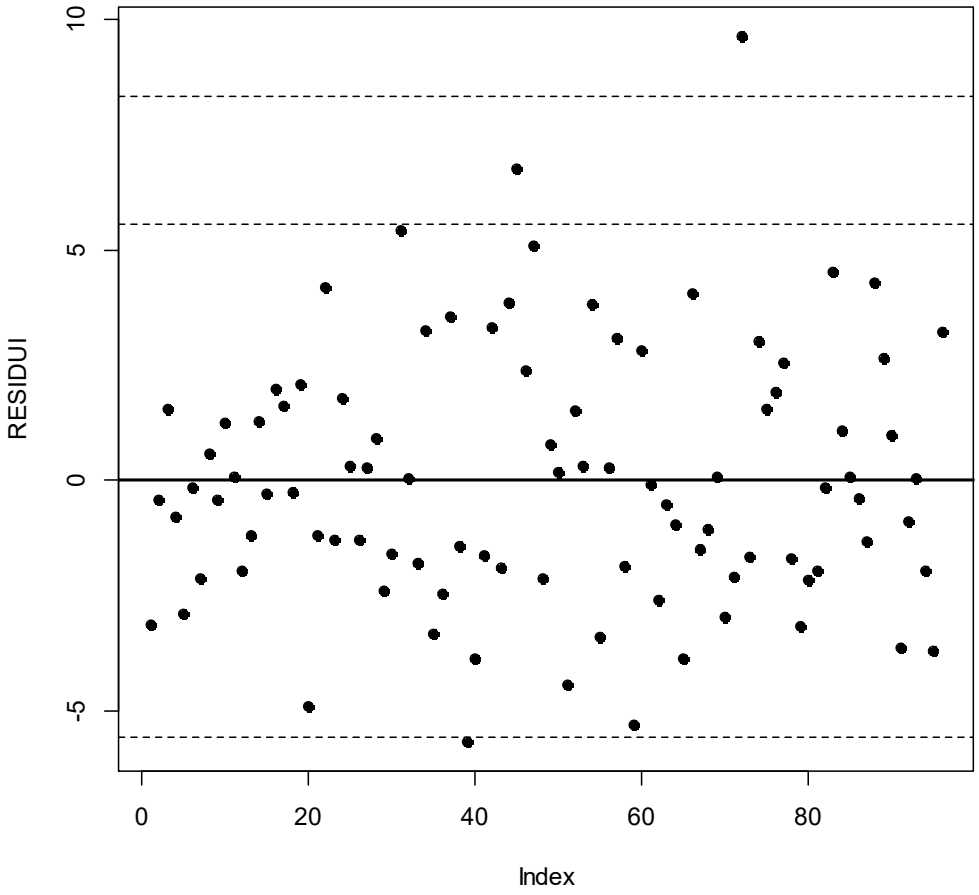
Shapiro-Wilks (Madansky, p. 20)

```
> shapiro.test(«residui»)

        Shapiro-Wilk normality test
```

(test di regressione, cfr. di varianze)

Facciamo un salto in

e in Francia!



Breusch-Pagan test (Madansky, p. 81)

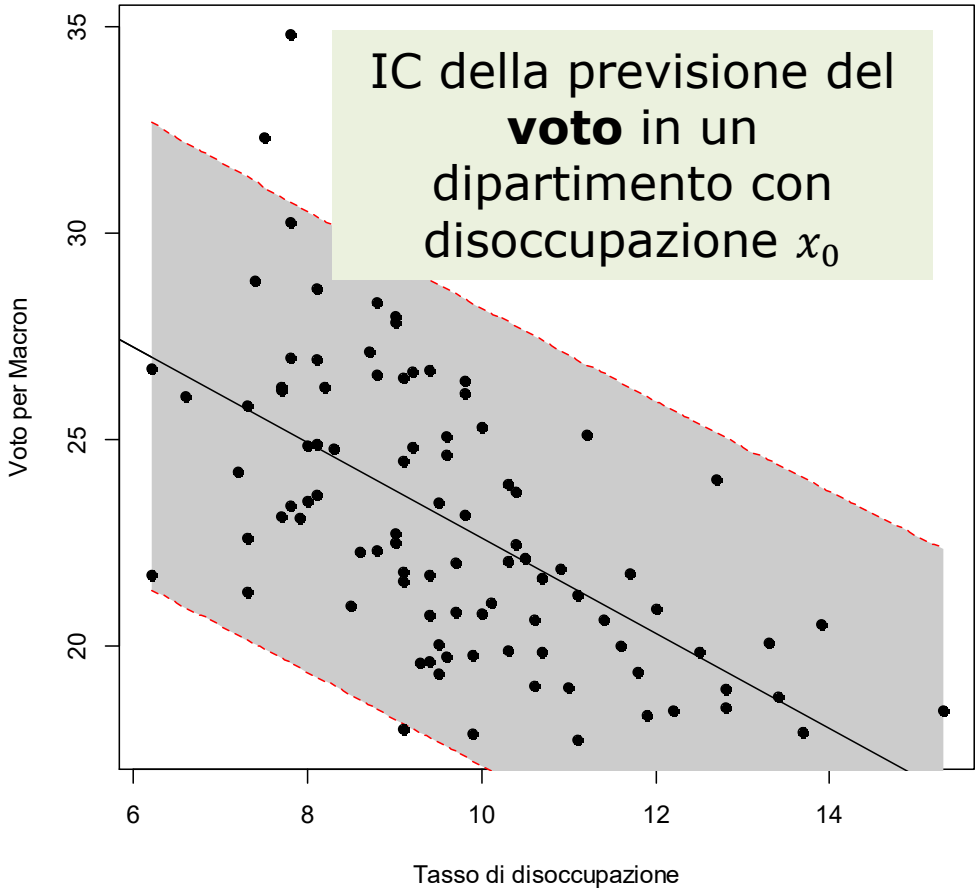
```
> library(car)
> ncvTest(«lm»)

        Non-constant Variance Score Test
```

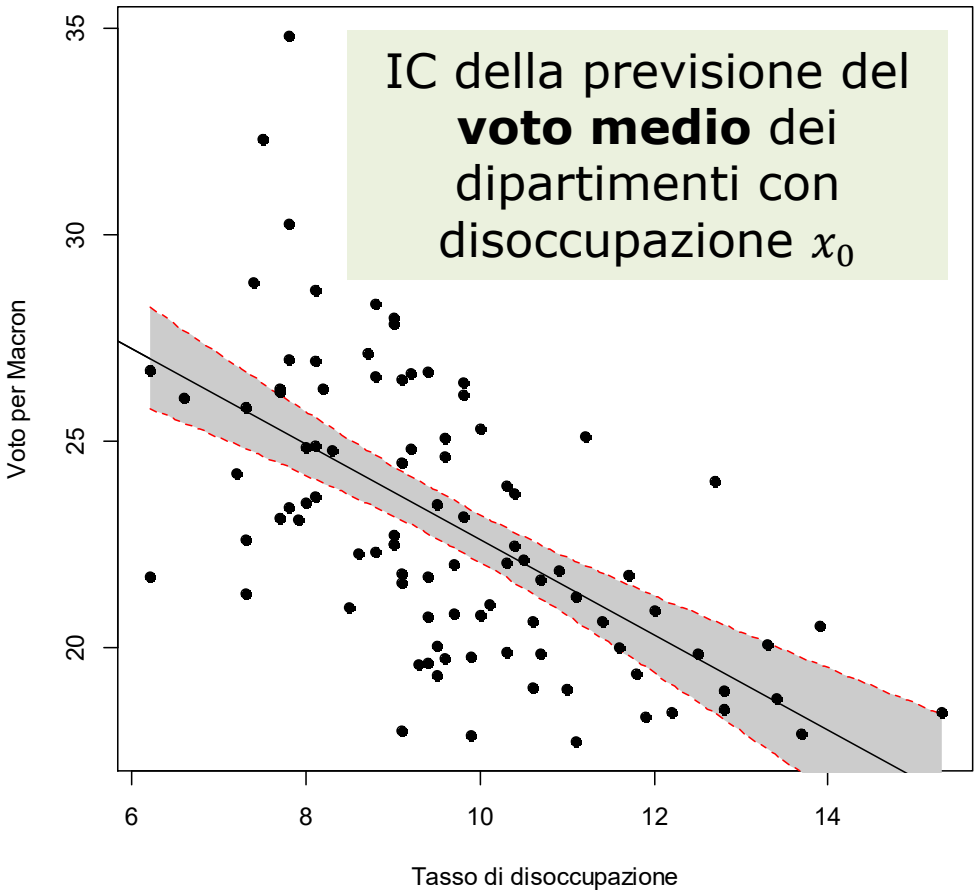
Durbin-Watson test (Madansky, p. 93)

```
> library(car)
> durbinWatsonTest(«lm»)

Alternative hypothesis: rho != 0
```

Facciamo un salto in R

e in Francia!



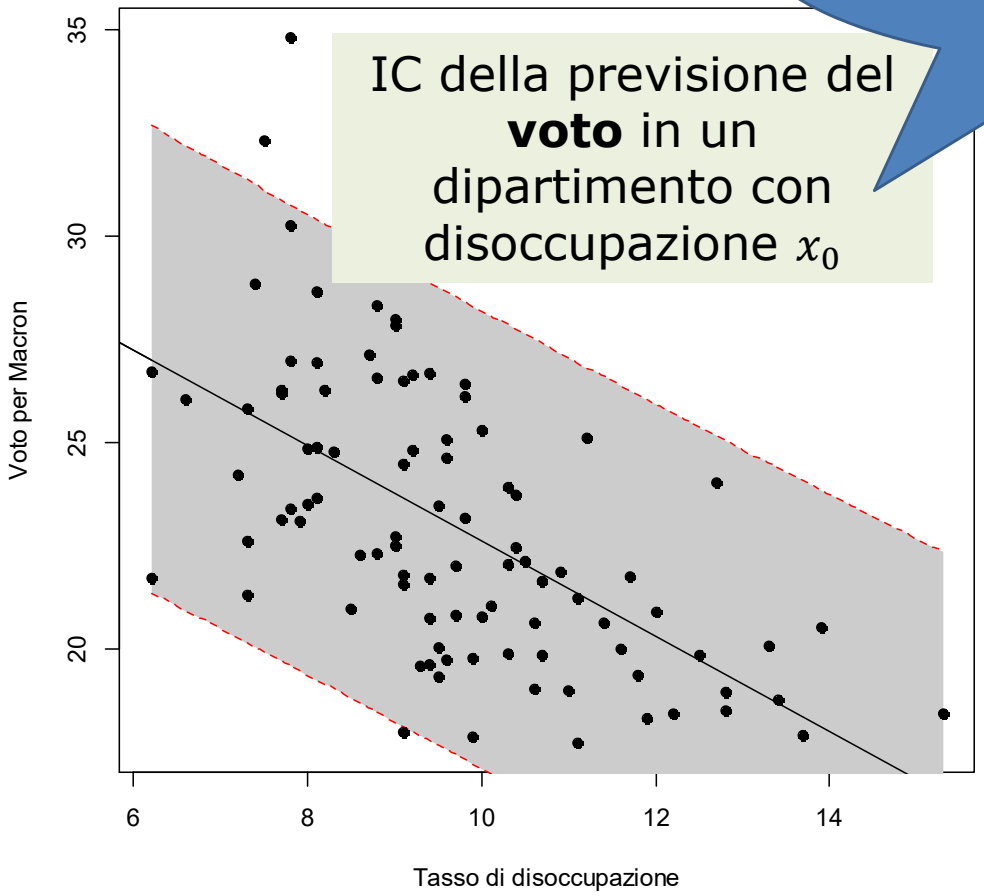


Facciamo un salto in R

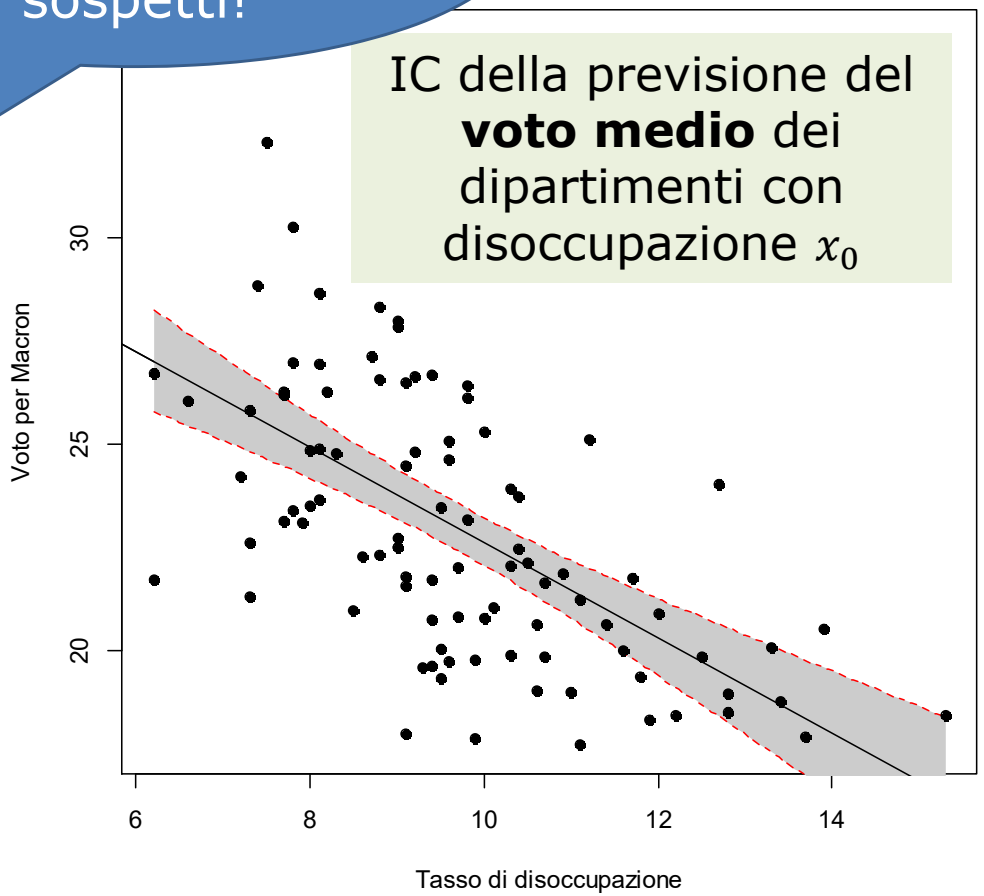
e in Francia!

I dati che cadano fuori dall'IC sono sospetti!

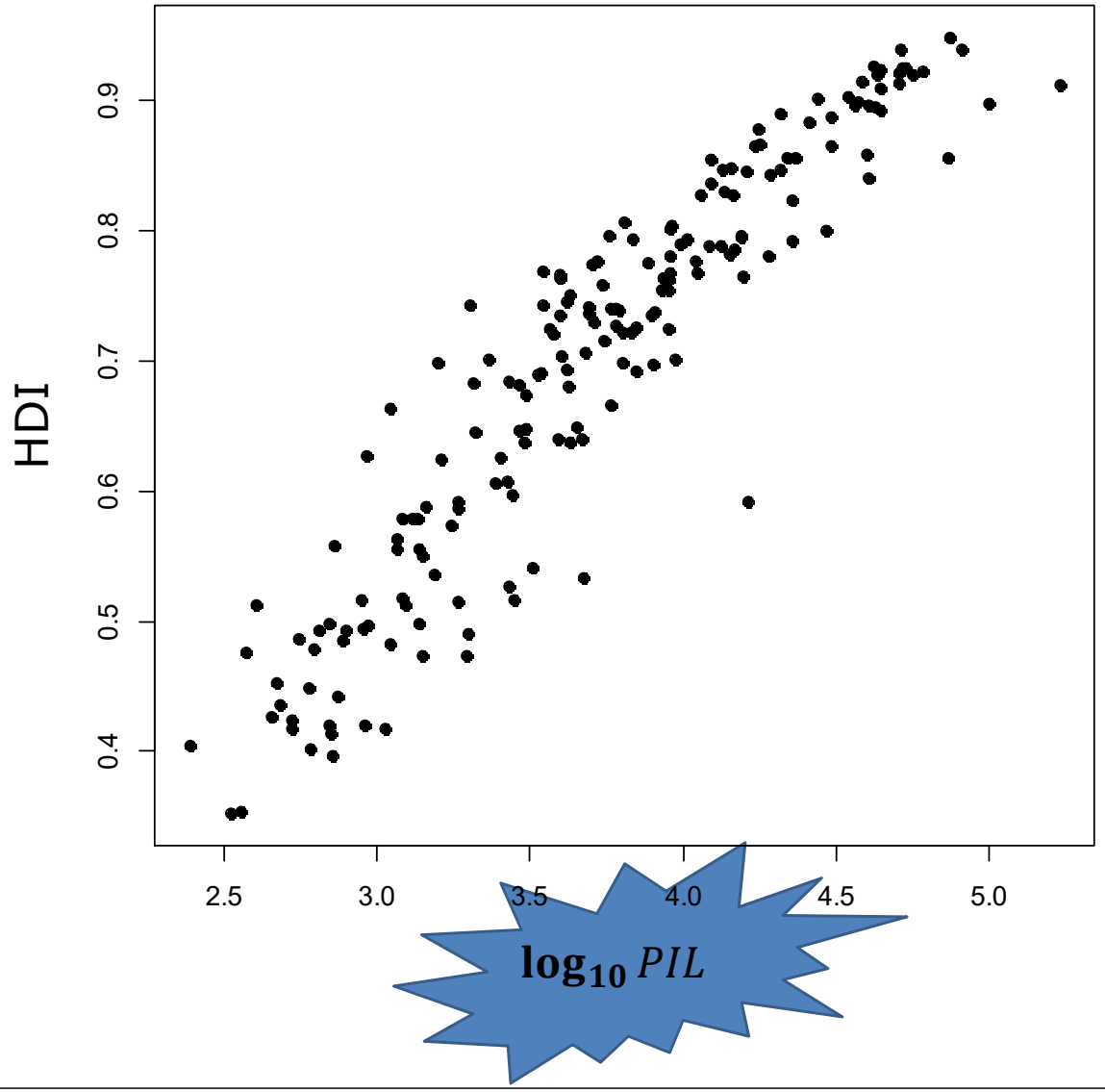
IC della previsione del **voto** in un dipartimento con disoccupazione x_0



IC della previsione del **voto medio** dei dipartimenti con disoccupazione x_0



Esercizio di compito

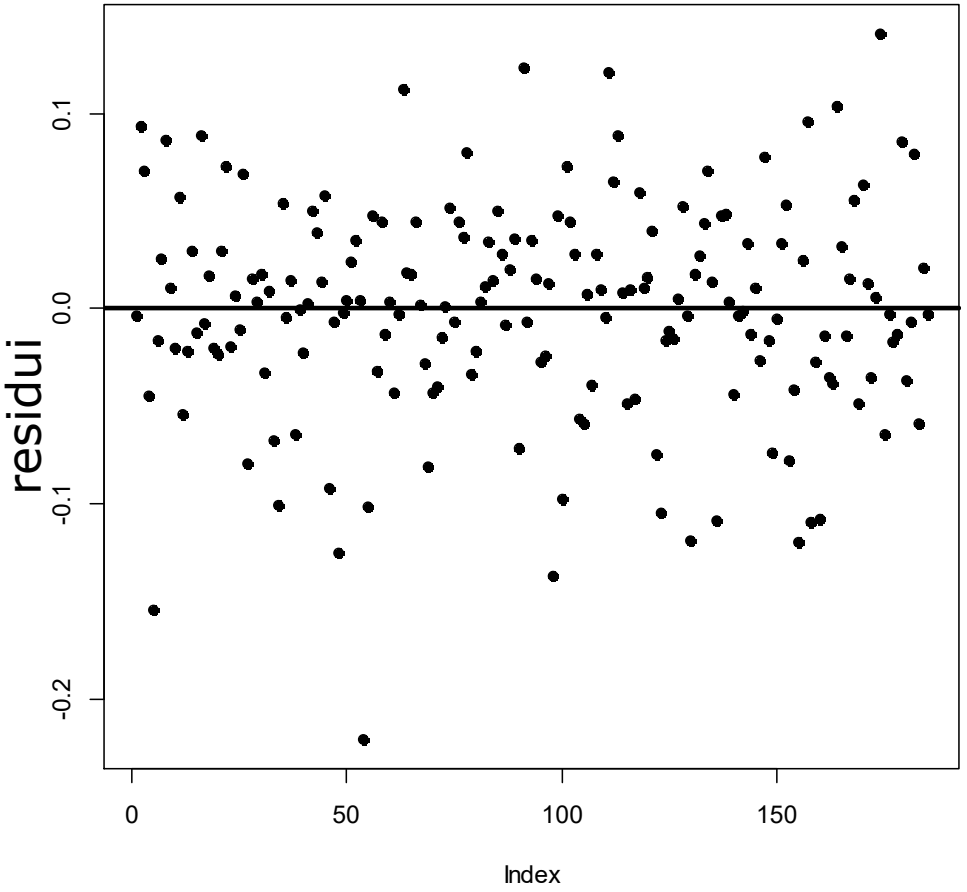


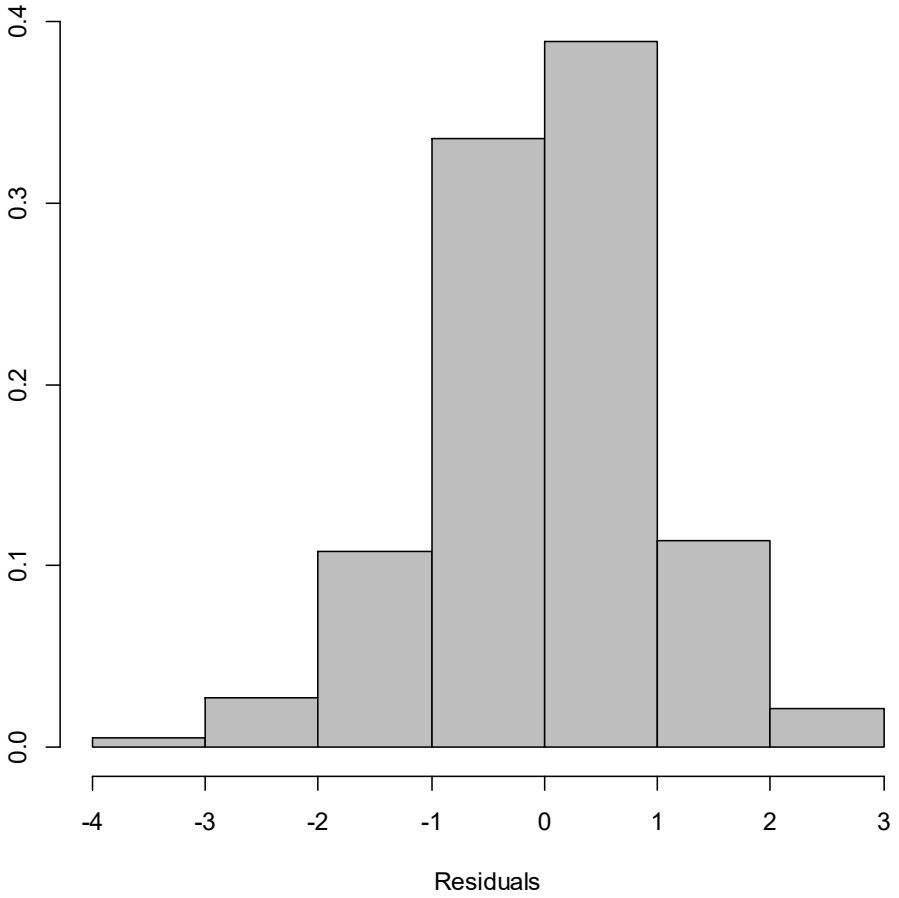
L'indice di sviluppo umano (ISU) (in inglese: **HDI**-**Human Development Index**) è un indice comparativo dello sviluppo dei vari paesi calcolato tenendo conto dei diversi tassi di aspettativa di vita, istruzione e reddito nazionale lordo procapite.

dati:
gdp-hdi-2105.txt

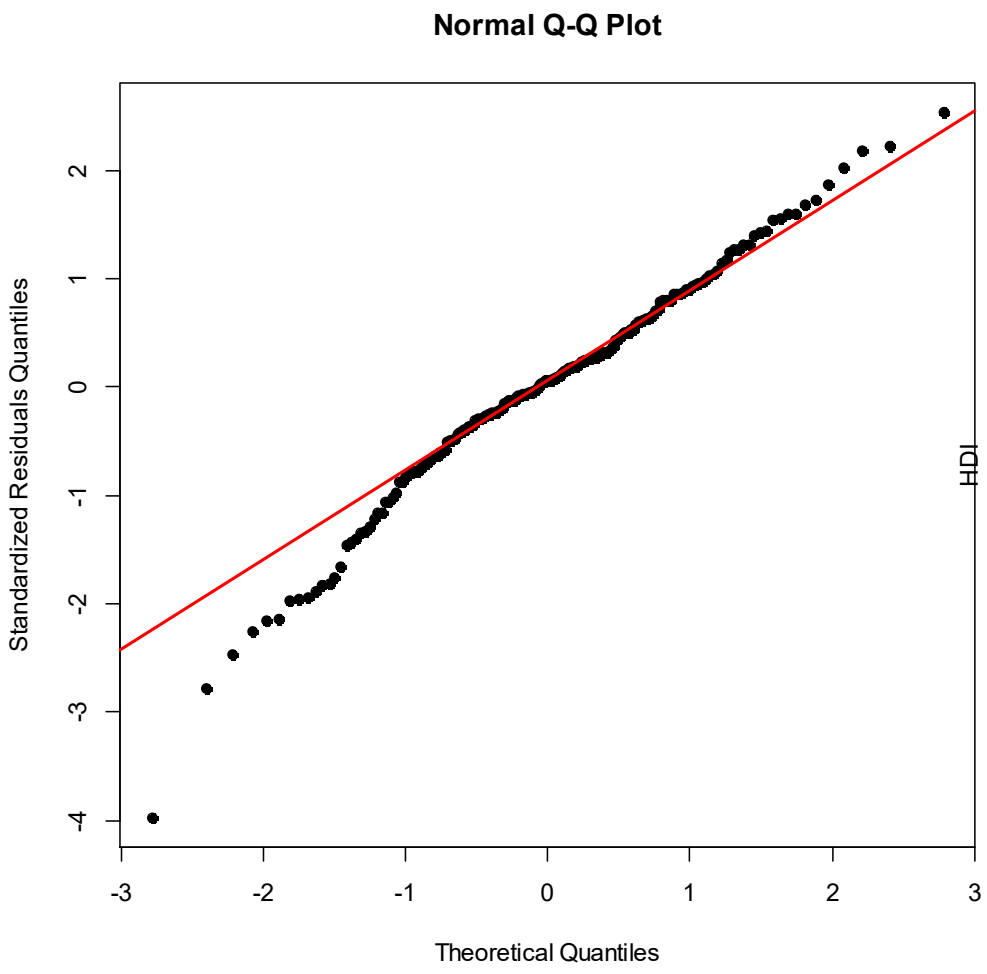
(si ringrazia la studentessa M. Lintner)

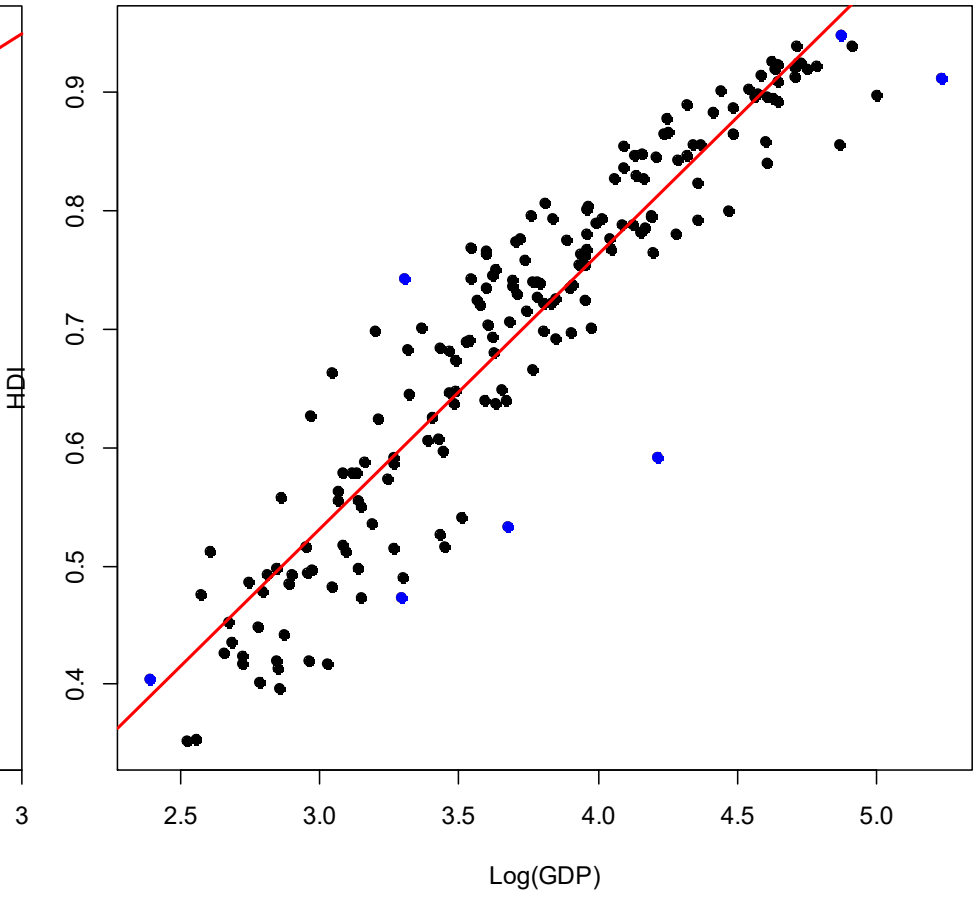
Esercizio


residui
0.1
0.0
-0.1
-0.2
0
50
100
150
Index


0.4
0.3
0.2
0.1
0.0
-4
-3
-2
-1
0
1
2
3
Residuals

Esercizio

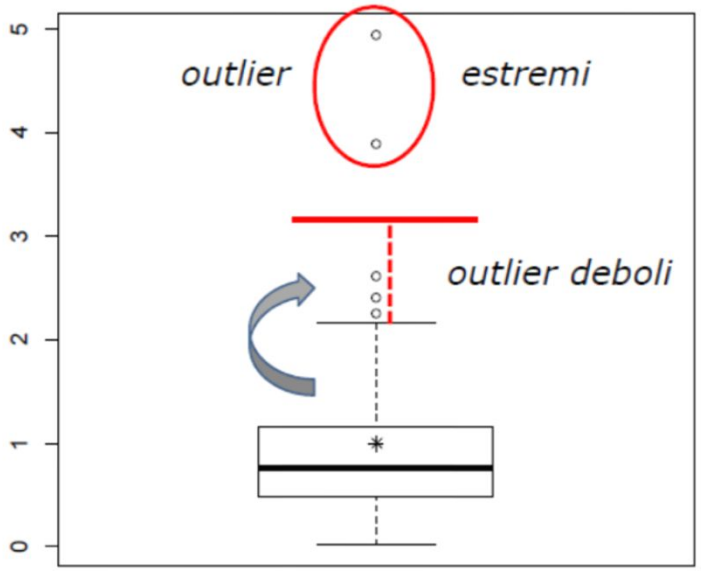




morale della favola: R^2 alto non basta!!

Outlier e dati influenti

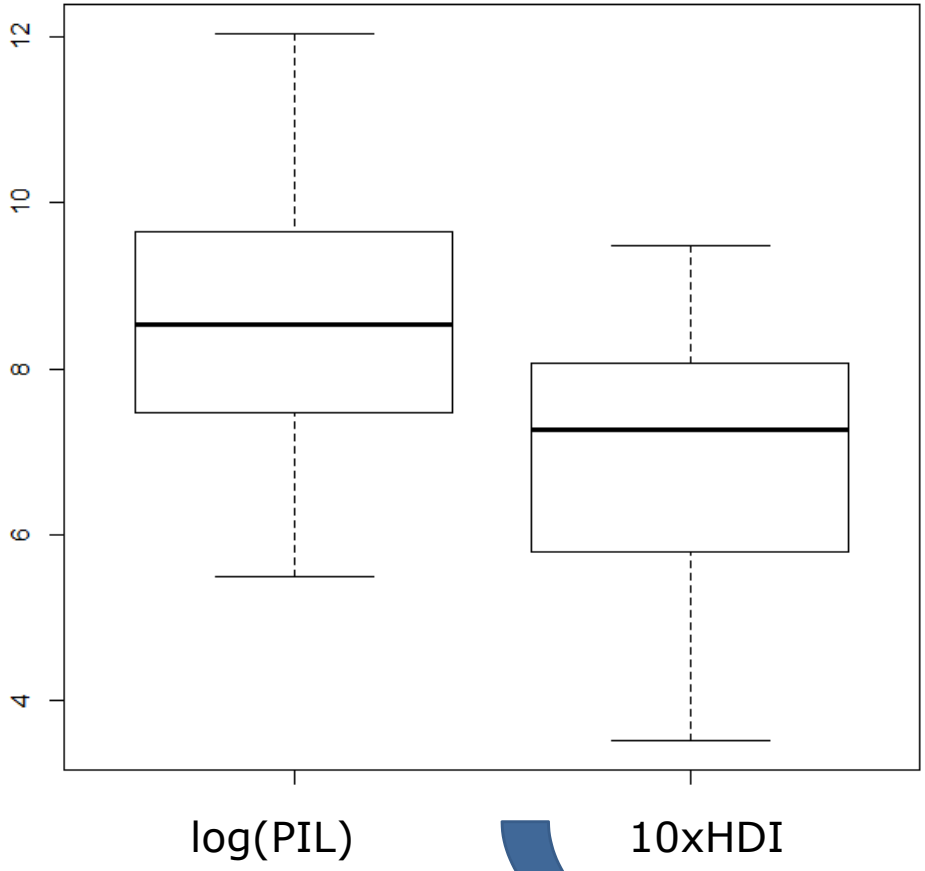
Boxplot

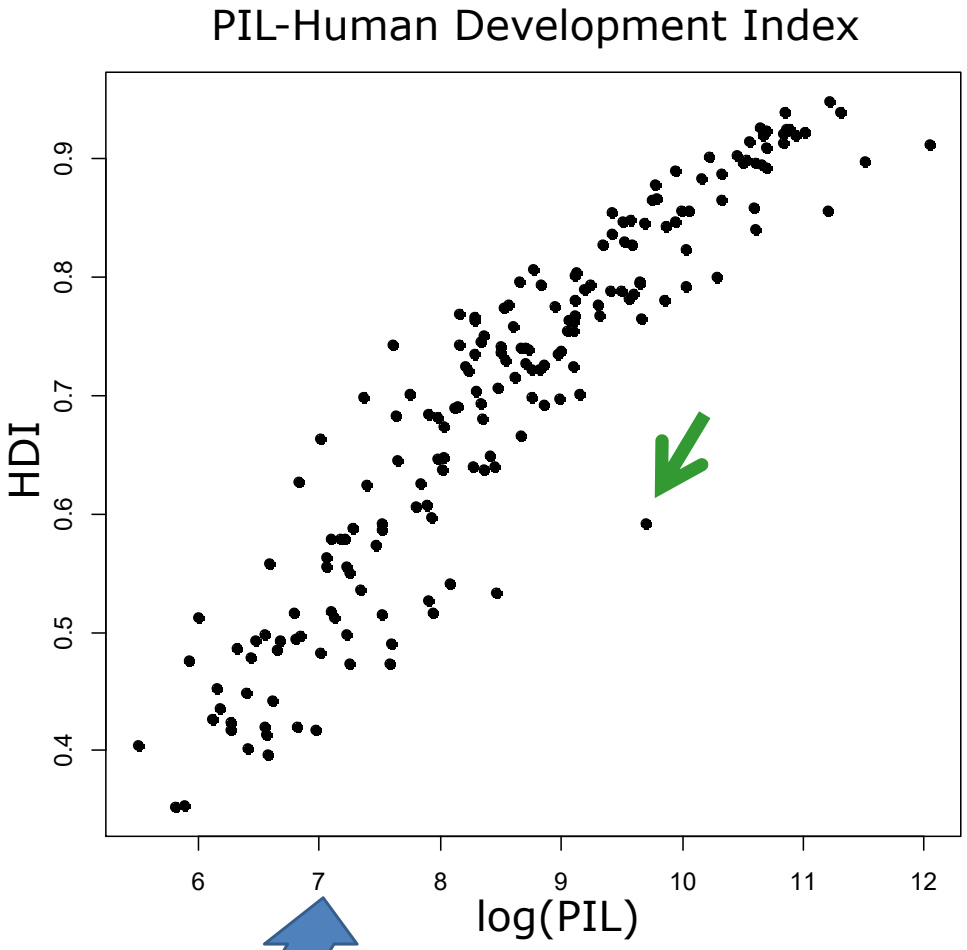


Dati fuori dal baffo (superiore): *outlier*

Baffo più corto della lunghezza massima

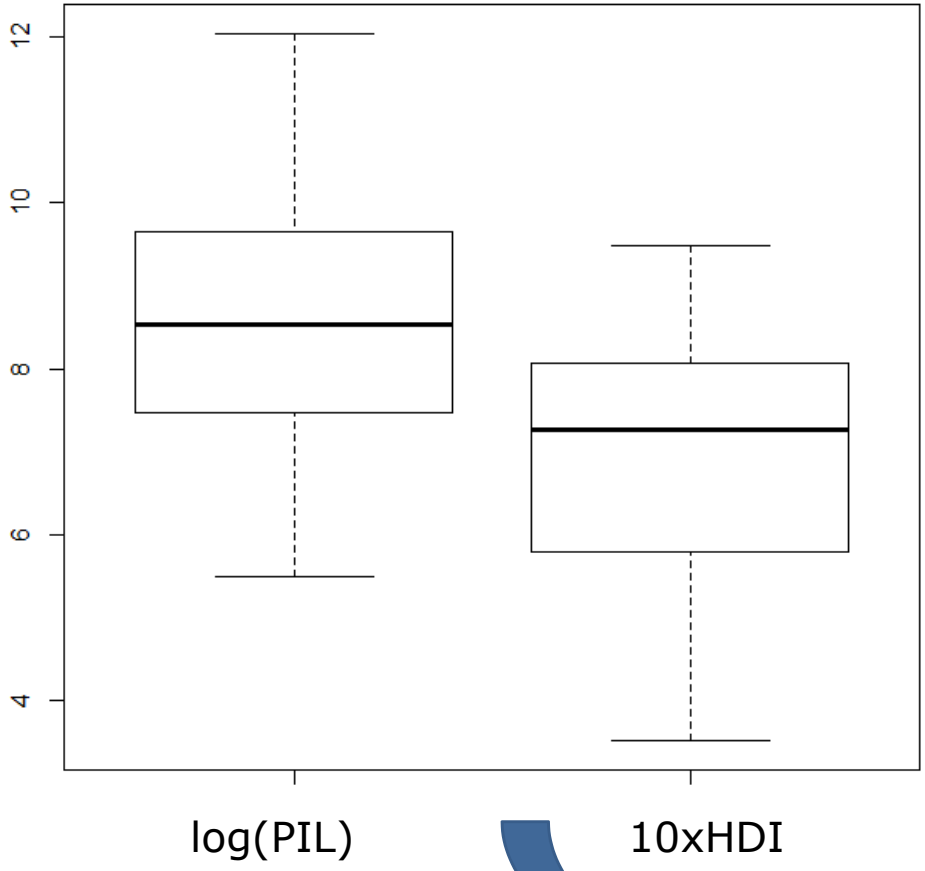
Outlier e dati influenti

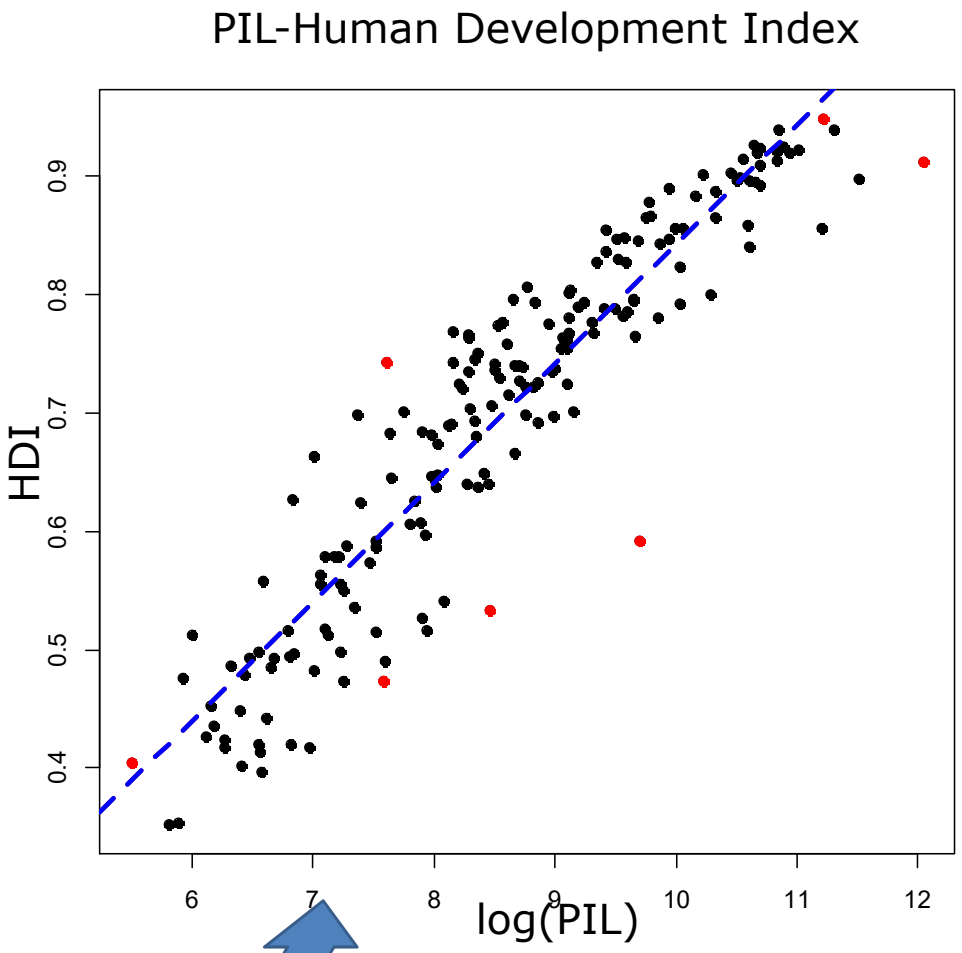




dati: gdp-hdi-2105.txt

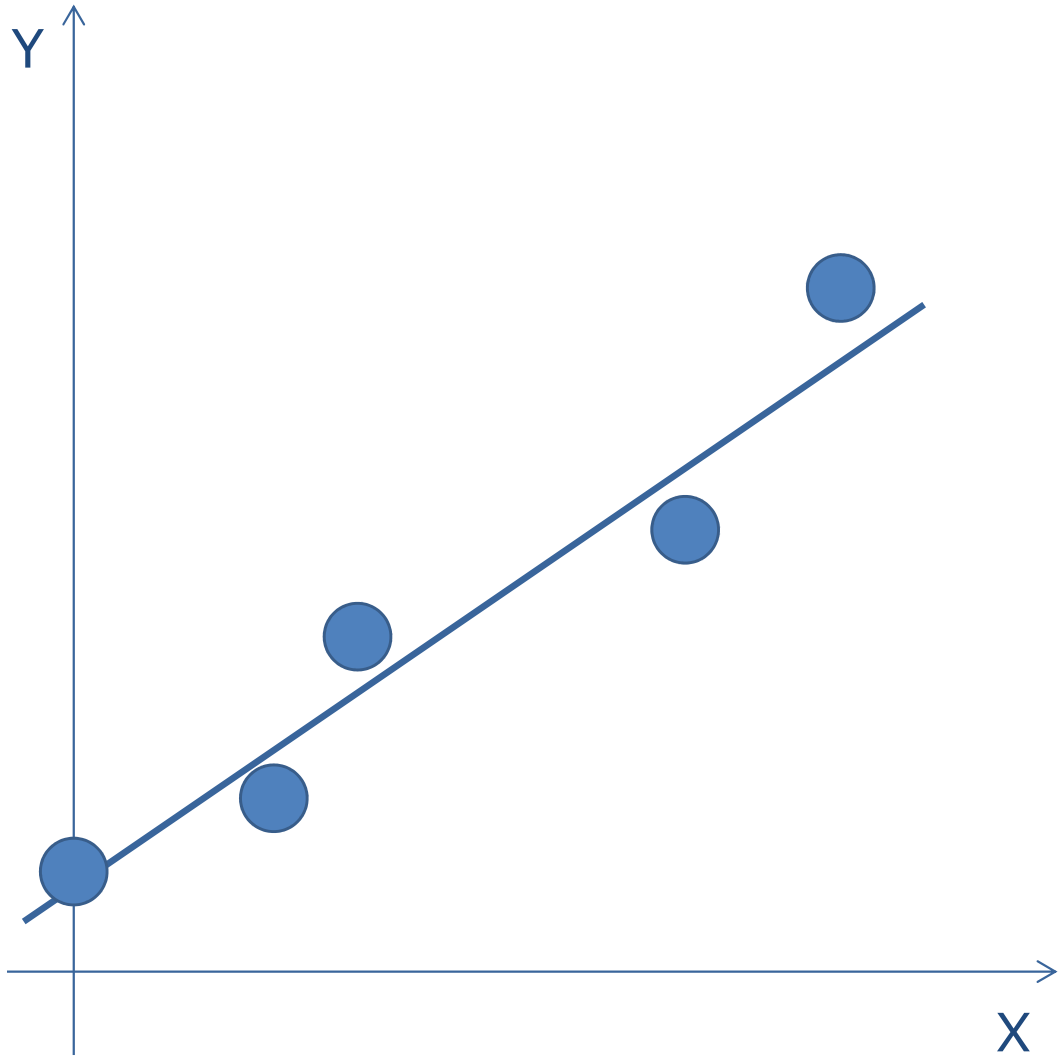
Outlier e dati influenti



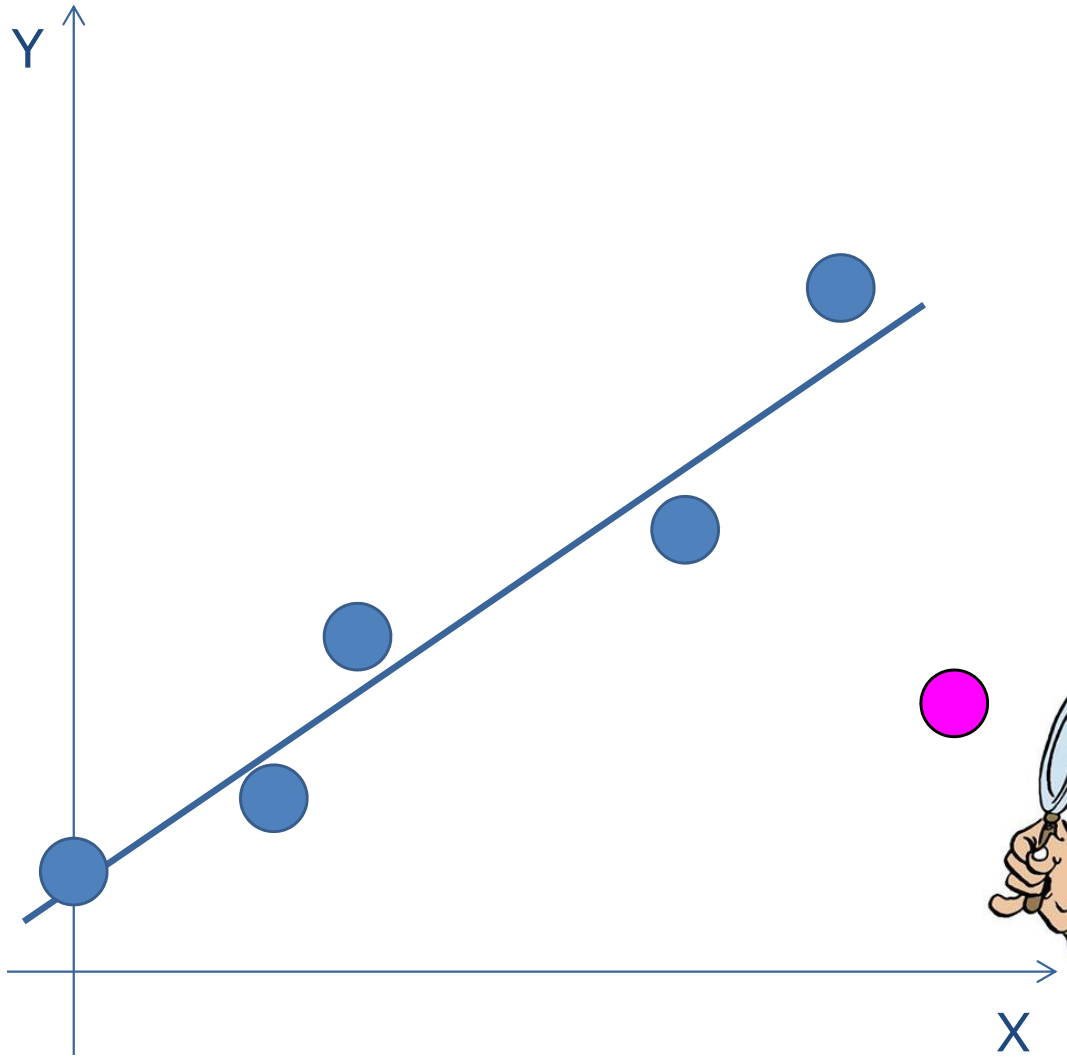


dati: gdp-hdi-2105.txt

Outlier e dati influenti



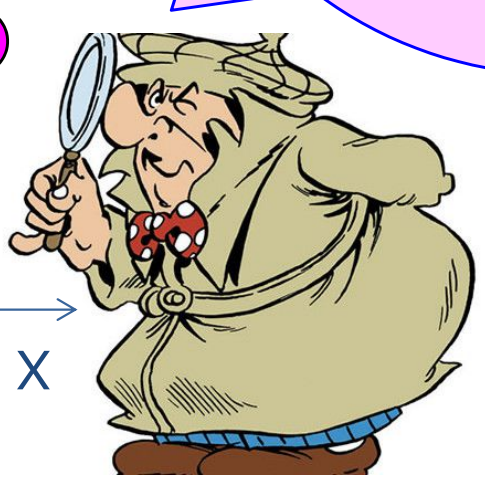
Outlier e dati influenti



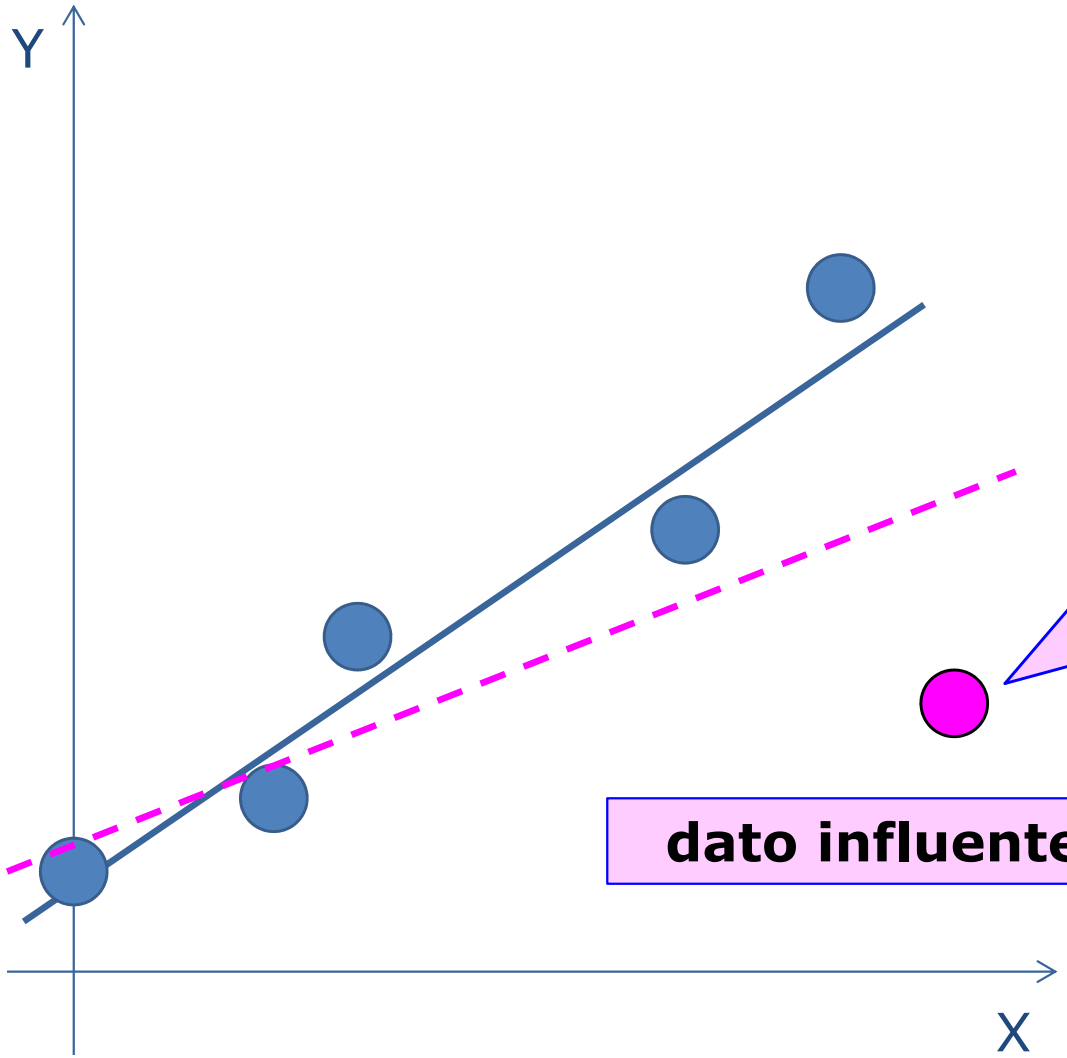
E’ un outlier rispetto a X?

Lo è rispetto a Y?

Che effetto ha sulla retta dei minimi quadrati?



Outlier e dati influenti



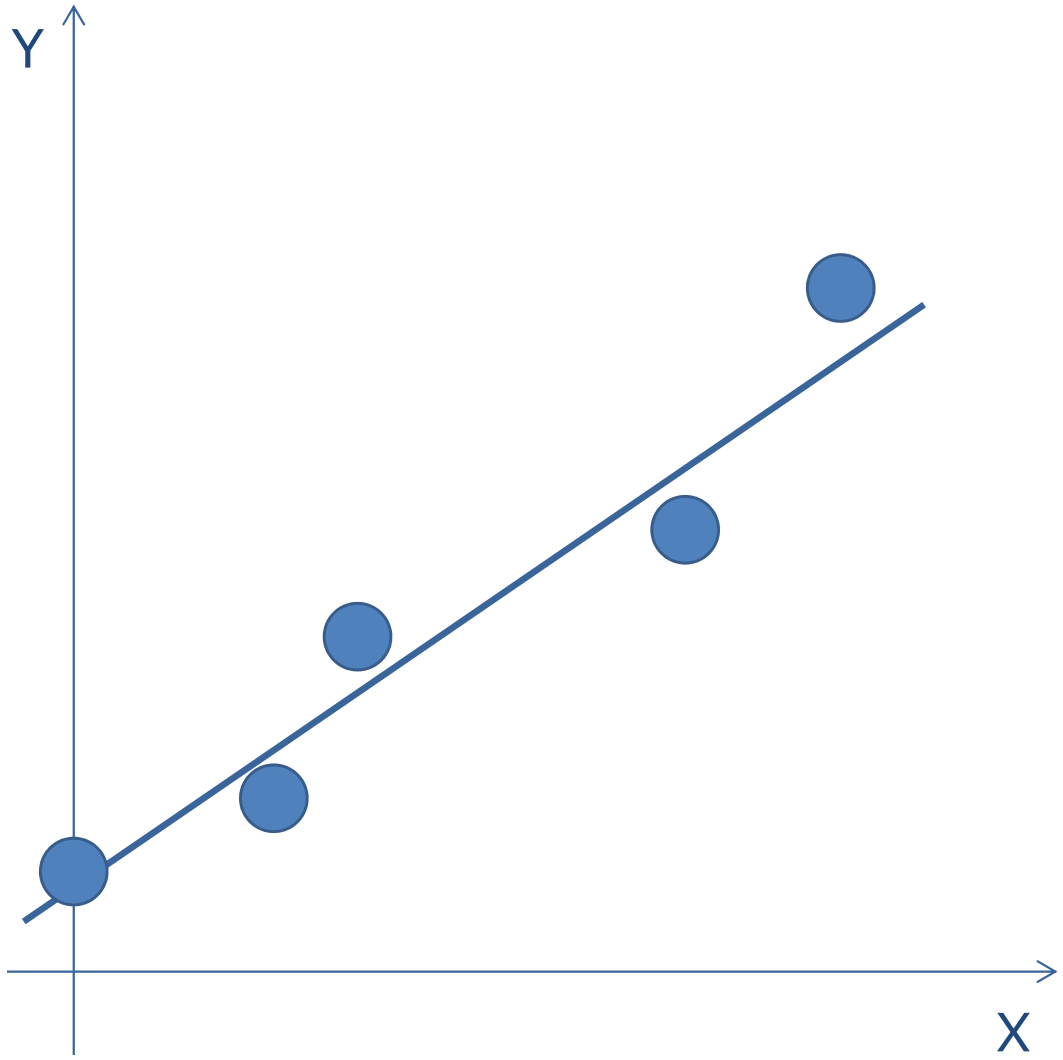
E’ un outlier rispetto a X?

Lo è rispetto a Y?

Che effetto ha sulla retta dei minimi quadrati?

dato influente (*leverage*)

Outlier e dati influenti



E’ un outlier rispetto a X?

Lo è rispetto a Y?

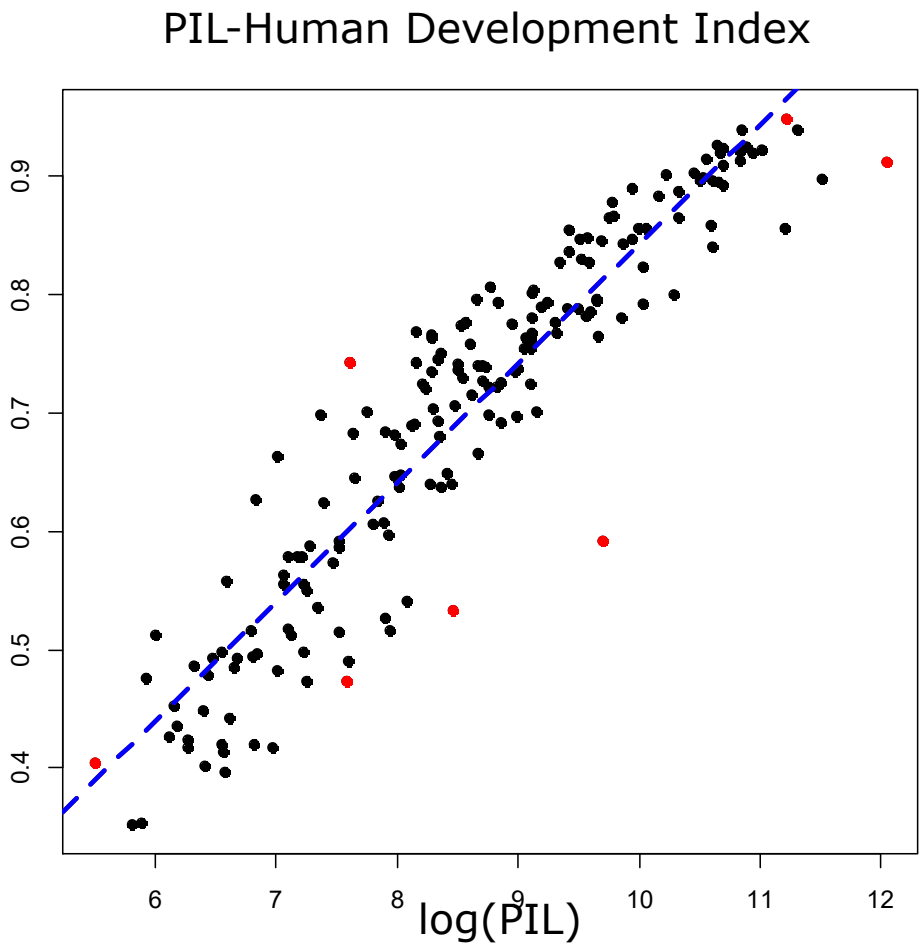
Che effetto ha sulla retta dei minimi quadrati?

Outlier e dati influenti

```
> influence.measures
```

Regression Deletion Diagnostics Description

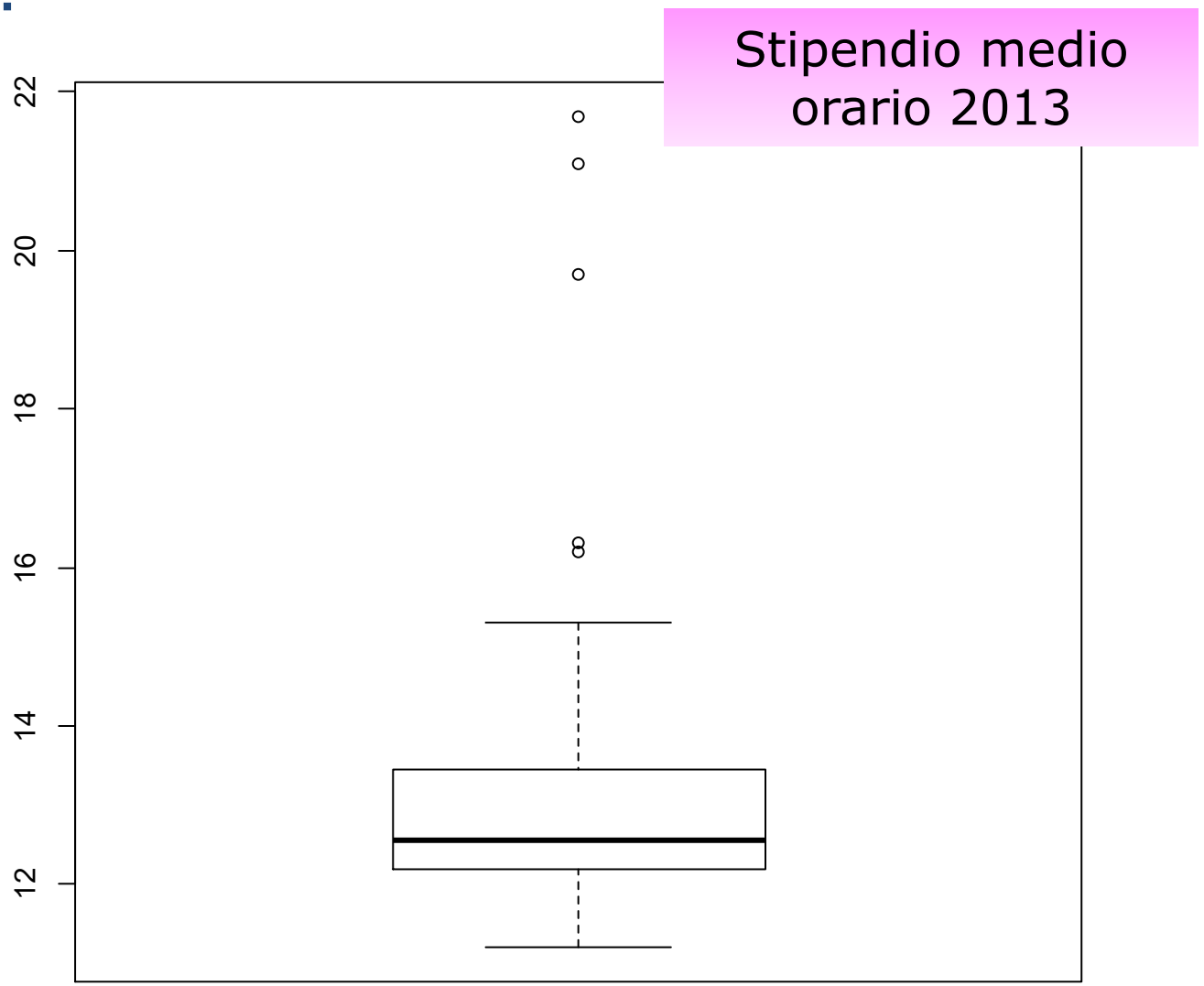
This suite of functions can be used to compute some of the regression (leave-one-out deletion) diagnostics for linear and generalized linear models discussed in Belsley, Kuh and Welsch (1980), Cook and Weisberg (1982), etc.



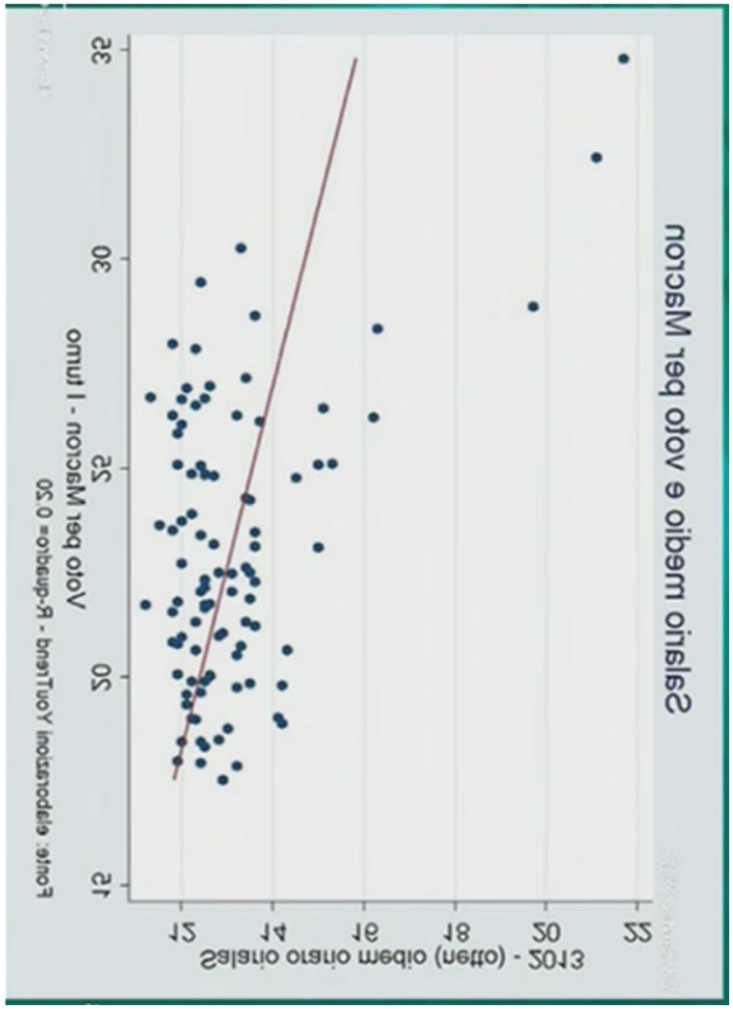
dati: gdp-hdi-2105.txt

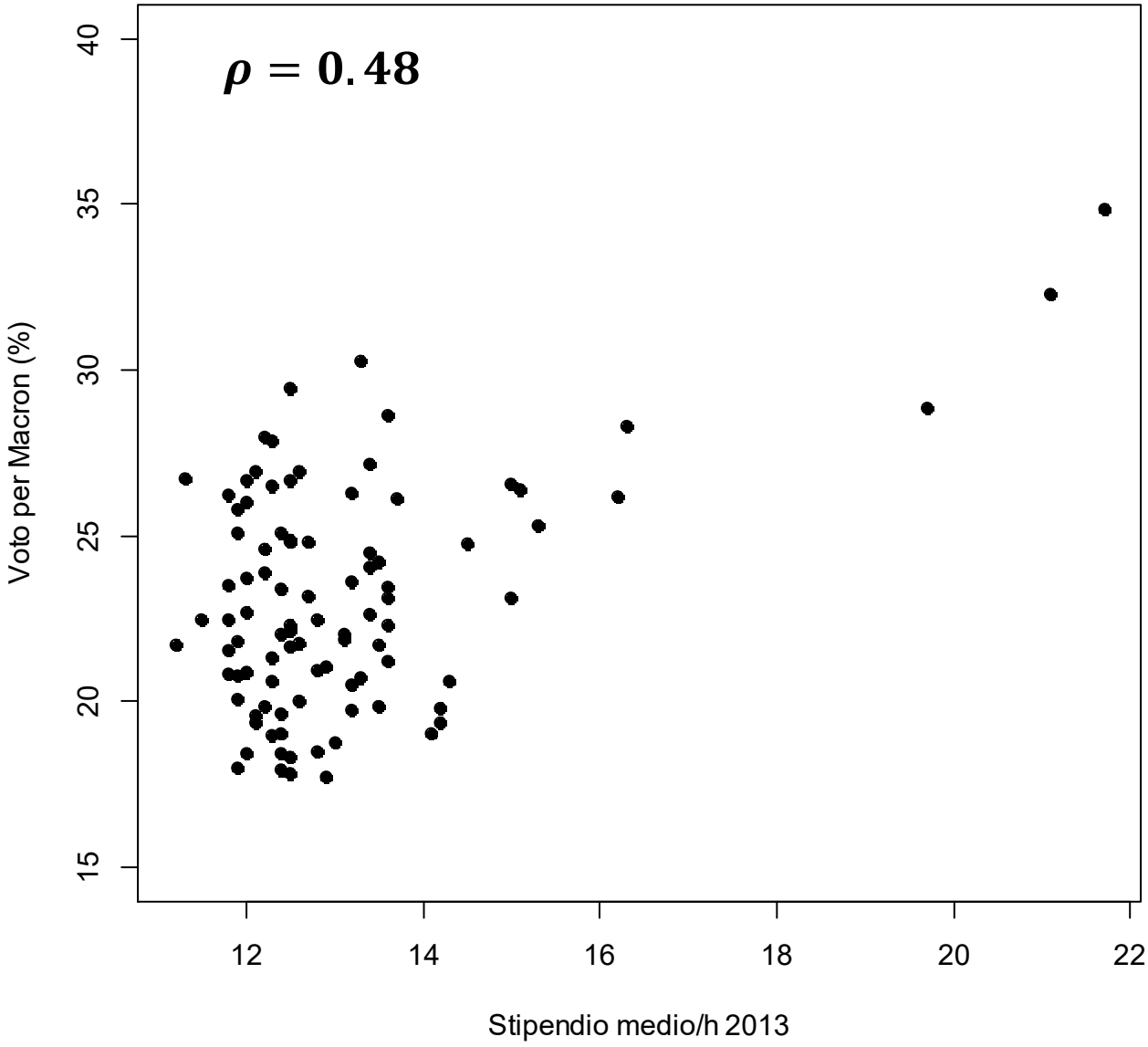
Facciamo un salto in

e in Francia!

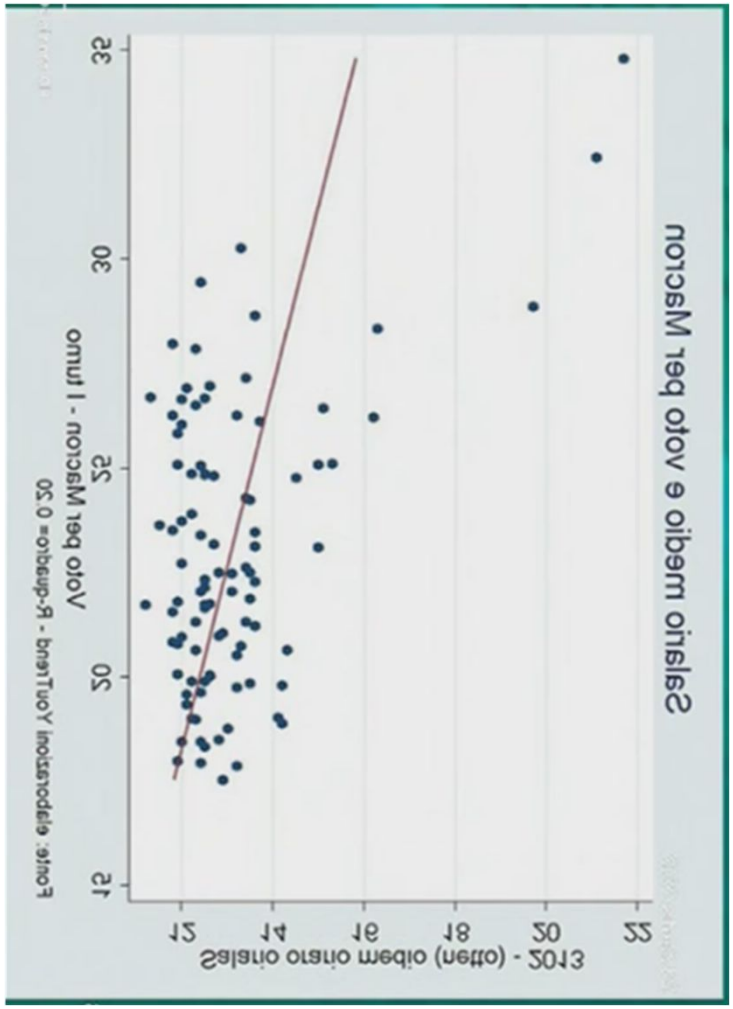


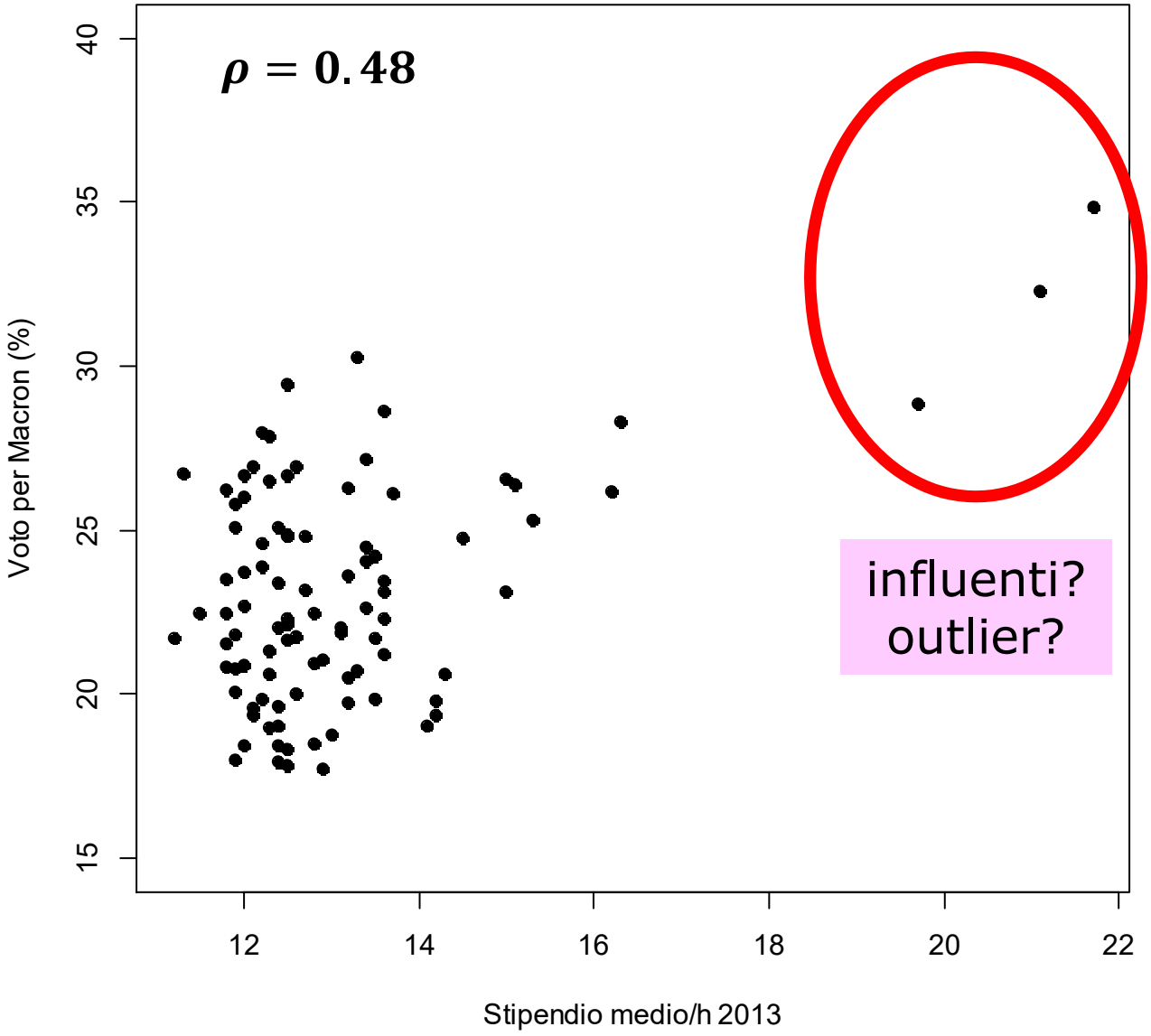
Facciamo un salto in



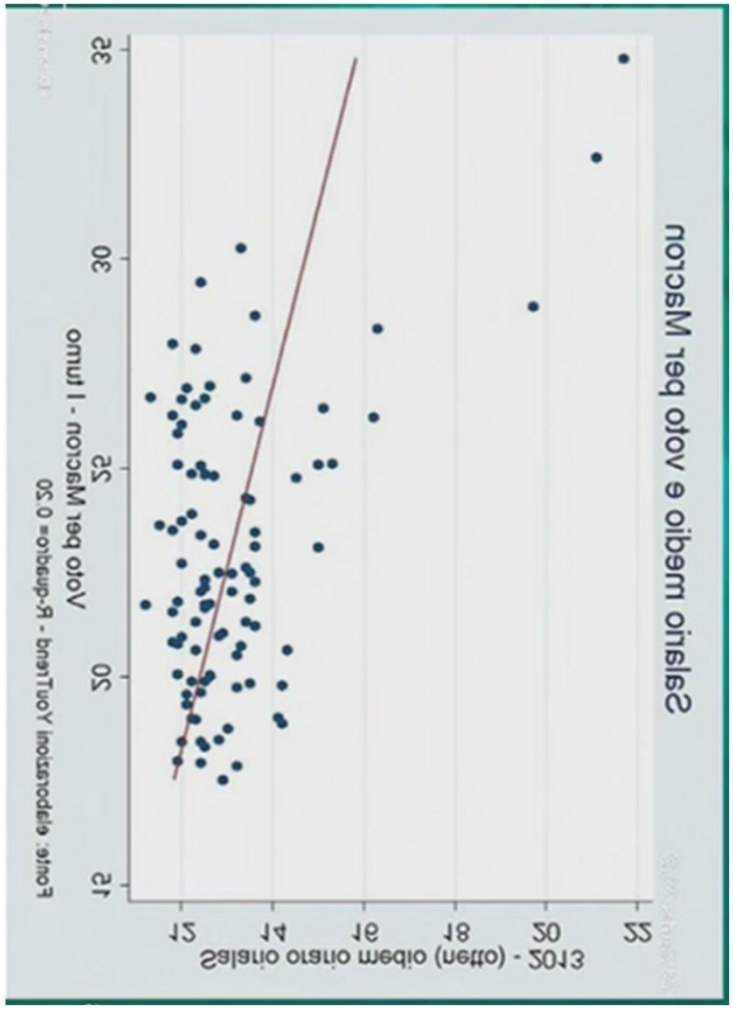


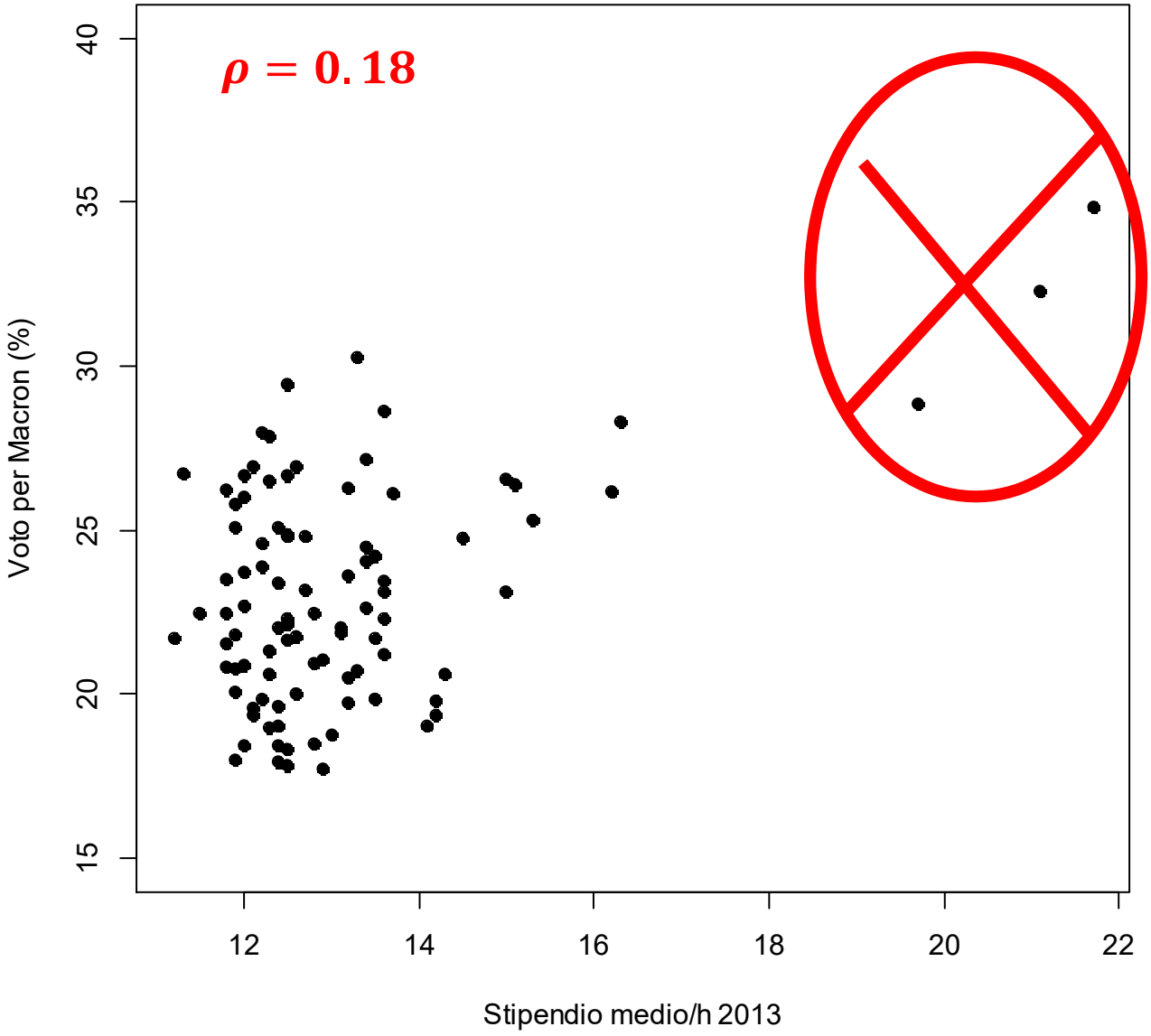
Facciamo un salto in R





Facciamo un salto in R





CI VOGLIAMO **PROPRIO** FAR PASSARE
IN MEZZO UNA RETTA?
UNA PARABOLA, UN'IPERBOLE, UNA
FUNZIONE SINUSOIDALE…?!