

METODI STATISTICI PER LA BIOINGEGNERIA (B)

# PARTE 9: UN CASO DI STUDIO SULLA REGRESSIONE LINEARE

A.A. 2025-2026

Prof. Martina Vettoretti

# IL MODELLO DI REGRESSIONE LINEARE MULTIPLA (RIPASSO)

- Modello di regressione lineare multipla:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \varepsilon_i, \qquad i = 1, \dots, n$$

$$\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- Assunzioni:
  - Relazione lineare tra $X_j, j = 1, \dots, m$ e $Y$.
  - $\varepsilon_i$ normali e tra loro indipendenti e $\varepsilon_i \sim N(0, \sigma_i^2)$

- Dati necessari per l'identificazione del modello:
  - $(x_{i1}, x_{i2}, \dots, x_{im}, y_i)$, i=1,…,n

# IDENTIFICAZIONE DEL MODELLO

- Stima dei coefficienti di regressione con il metodo dei minimi quadrati lineari
  - Assunzione: $\sigma_i^2 = \sigma^2 \; \forall i = 1, \dots, n$
  - Applicazione dello stimatore:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

- Stima a posteriori del valore di $\sigma^2$

$$\hat{\sigma}^2 = \frac{SSE}{n - (m+1)}, \qquad SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

# VALUTAZIONE DELLA BONTA' DEL MODELLO

- Confronto tra uscita misurata e uscita predetta

$$y_i \quad \text{vs.} \quad \hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_{ij}$$

$$\boldsymbol{y} \quad \text{vs.} \quad \hat{\boldsymbol{y}} = \boldsymbol{X} \cdot \hat{\boldsymbol{\beta}}$$

$$MSE = \frac{SSE}{n}, \qquad RMSE = \sqrt{MSE}$$

- Coefficiente di determinazione $R^2$

$$R^2 = 1 - \frac{SSE}{SST} \qquad SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Test F
  - $H_0$: $\beta_1 = \beta_2 = \cdots = \beta_m = 0$
  - $H_1$: almeno un coefficiente $\beta_j \neq 0, \; j \neq 0$

# ANALISI DEI RESIDUI

- Calcolo dei residui:

$$r_i = y_i - \hat{y}_i, \qquad i = 1, \dots, n$$

- Check distribuzione normale
  - Istogramma, test di normalità, q-q plot, indici di forma campionari
- Check media nulla
  - Calcolo media campionaria + test di verifica ipotesi
- Check campioni scorrelati (bianchezza)
  - Plot $r_i$ vs. $\hat{y}_i$ + funzione di autocorrelazione
- Check varianza omogenea, no trend, no outlier
  - Plot $r_i$ vs. $\hat{y}_i$

# VALUTAZIONE DEI PARAMETRI STIMATI

- Calcolo dello standard error:
  - $SE_j$ è radice quadrata dell'elemento in posizione $j$ su diagonale di $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$
- Calcolo coefficiente di variazione delle stime:

$$CV_j = \frac{SE_j}{|\hat{\beta}_j|} \cdot 100\ \%$$

- Valutazione valori stimati ed intervallo di confidenza al 95%

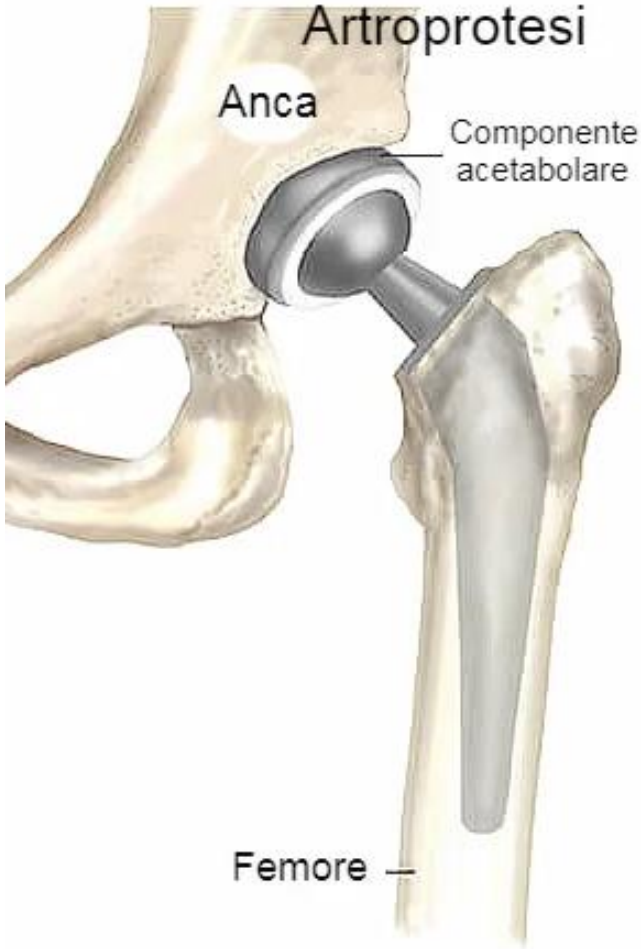
$$\hat{\beta}_j \pm 1.96 \cdot SE_j$$

  - Segno di $\hat{\beta}_j$
  - Valore assoluto di $\hat{\beta}_j$
- Test statistico sulle stime dei parametri (t test):
  - $H_0$: $\beta_j = 0$
  - $H_1$: $\beta_j \neq 0$

# CASO DI STUDIO

- Problema: predizione diametro della componente acetabolare di una protesi all'anca utilizzando variabili antropometriche.

- Zou et al. «Development and validation of multiple linear regression models for predicting total hip arthroplasty acetabular prosthesis», Journal of Orthopaedic Surgery and Research, 2024.



# DATASET

Dati raccolti su 500 pazienti di età compresa tra 65 e 85 anni.

- Variabile dipendente Y: diametro della componente acetabolare [mm]
- Variabili indipendenti:
  - $X_1$: altezza [cm]
  - $X_2$: peso [kg]
  - $X_3$: girovita [cm]
  - $X_4$: lunghezza del piede [cm]
  - $X_5$: età [anni]

Esercizio svolto in Matlab. Di seguito i risultati principali.

# IDENTIFICAZIONE DEL MODELLO DI REGRESSIONE LINEARE MULTIPLA

- Equazione del modello

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

- Stima dei coefficienti del modello con il metodo dei minimi quadrati lineari, assumendo varianza d'errore costante.

Matrice 500x6

$$\boldsymbol{X} = [\mathbf{1}\ \boldsymbol{x_1}\ \boldsymbol{x_2}\ \boldsymbol{x_3}\ \boldsymbol{x_4}\ \boldsymbol{x_5}] \Rightarrow \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \Rightarrow \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \\ \hat{\beta}_5 \end{bmatrix} = \begin{bmatrix} 21.0441 \\ 0.0884 \\ 0.0735 \\ 0.0069 \\ 0.5155 \\ -0.0111 \end{bmatrix}$$

Vettore colonna 500x1 di valori 1

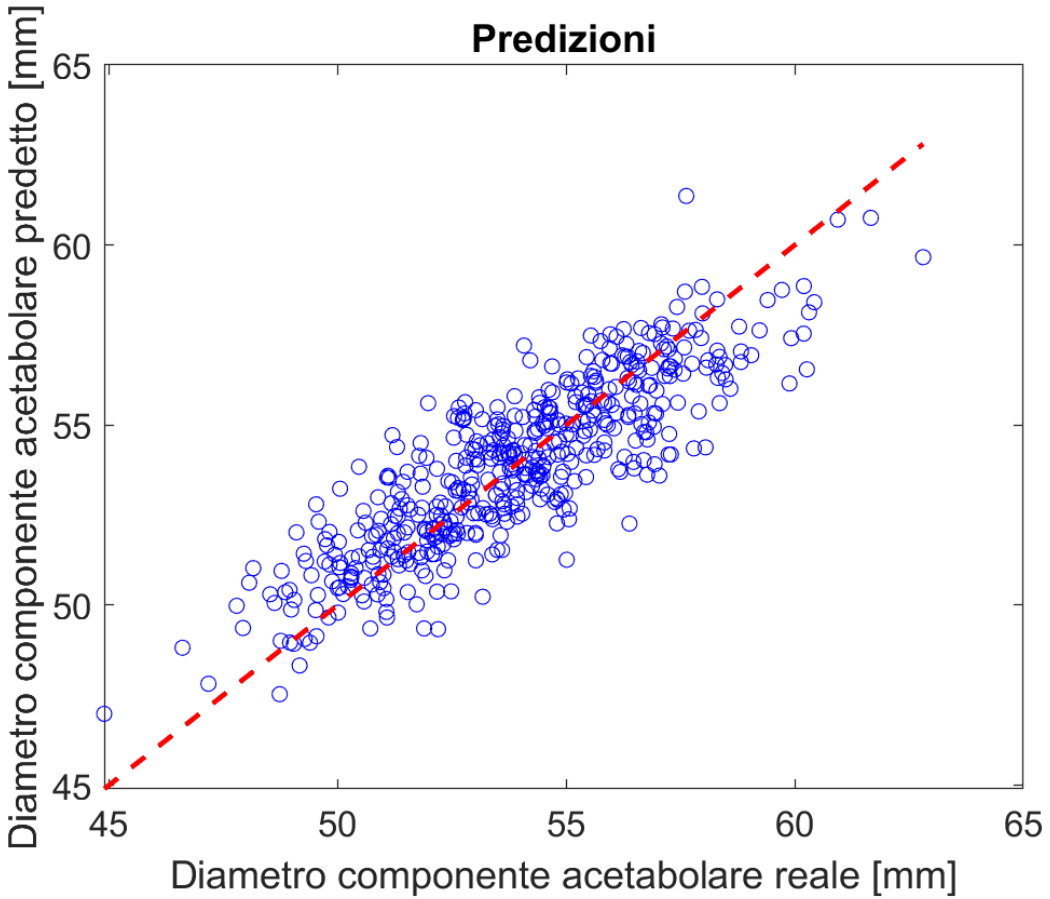
Vettori colonna 500x1 con i valori delle variabili indipendenti

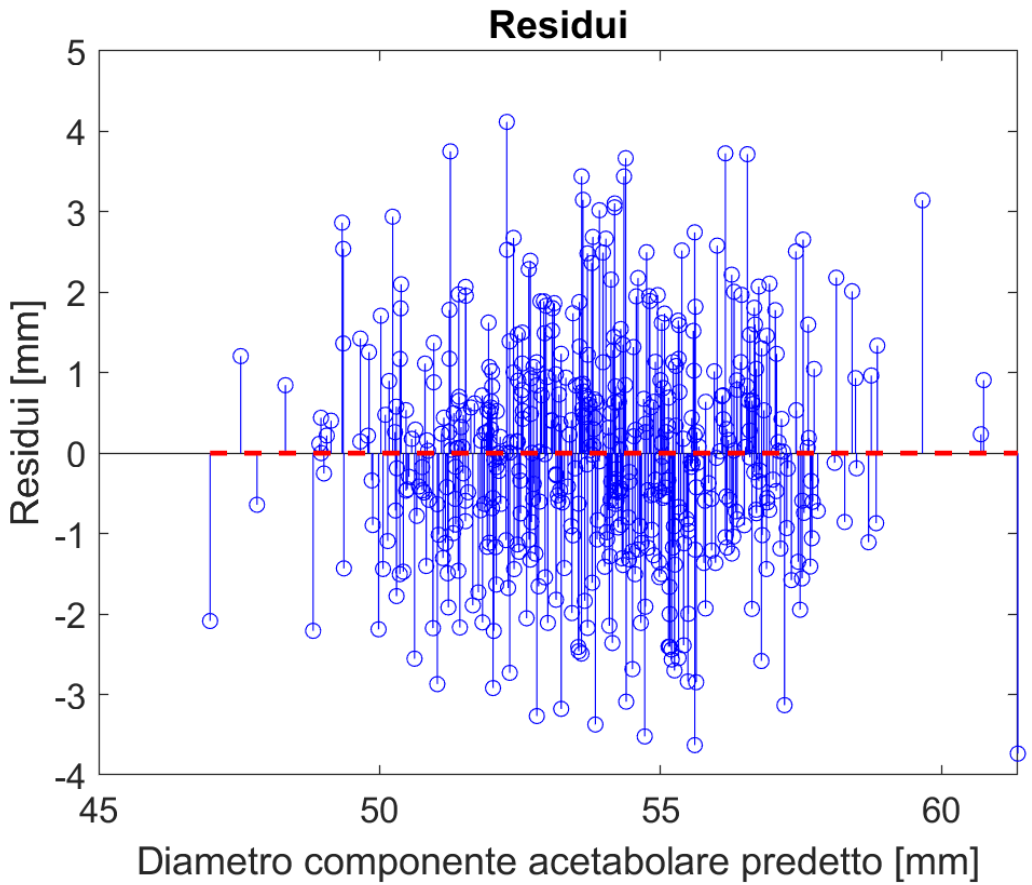
Vettore colonna 500x1 con i valori di Y

# PREDIZIONI E RESIDUI

Predizioni: $\hat{y} = X \cdot \hat{\beta}$

Residui: $y - \hat{y}$



# STIMA DELLA VARIANZA DELL'ERRORE

- Calcolo di SSE

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 980.2855\ [mm^2]$$

- Varianza dell'errore stimata a posteriori: $\hat{\sigma}^2 = \frac{SSE}{n-(m+1)} = 1.9844\ [mm^2]$

# VALUTAZIONE DEL MODELLO

- Calcolo di MSE ed RMSE

$$MSE = \frac{SSE}{n} = 1.96\ mm^2,$$
$$RMSE = \sqrt{MSE} = 1.40\ mm$$

- Calcolo di $R^2$

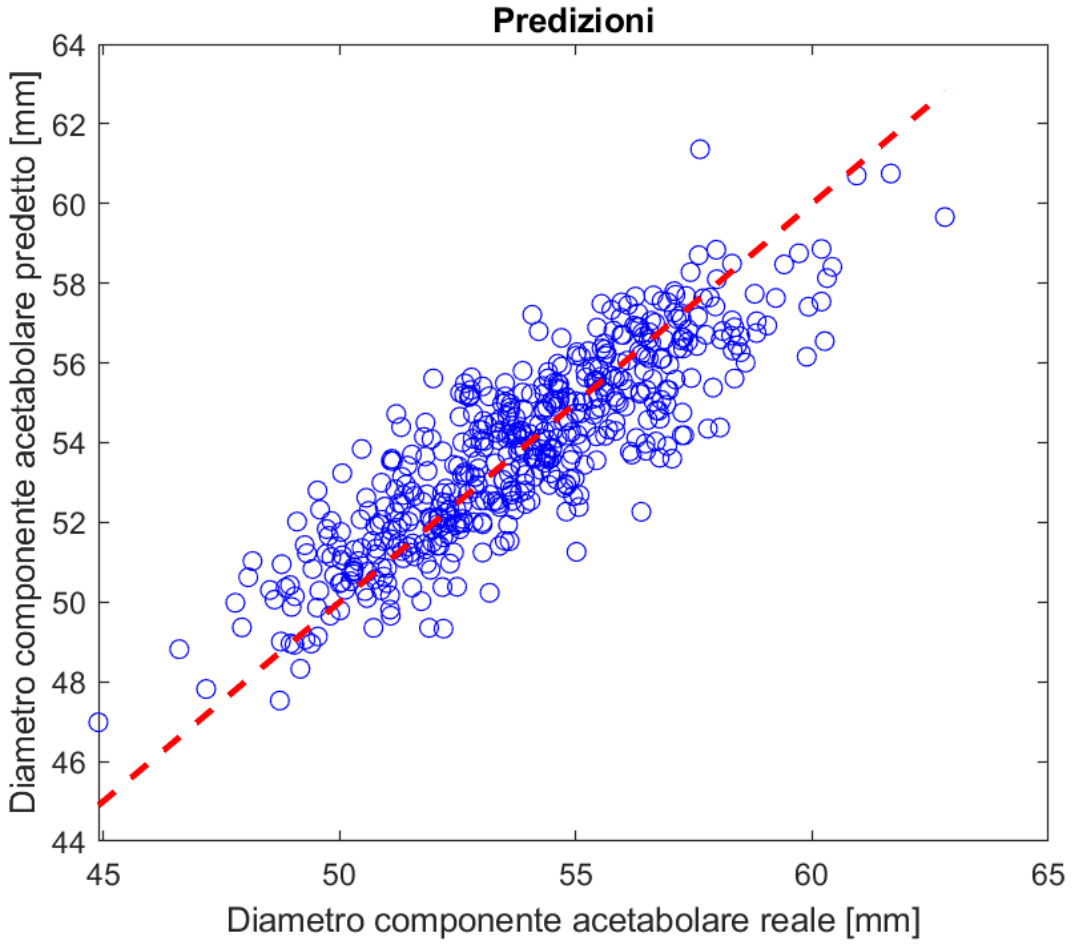
$$R^2 = 1 - \frac{SSE}{SST} = 0.7374$$

- Test F

$$F = \frac{(SST - SSE)/m}{SSE/(n - m - 1)} = 277.4024$$

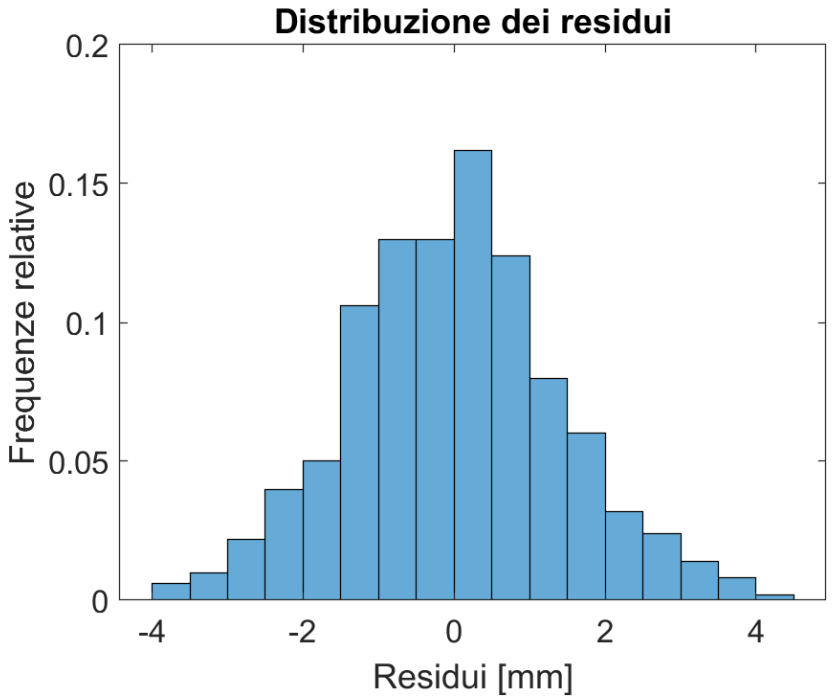
$$\alpha = 0.05 \rightarrow F_{\alpha,m,n-m-1} = 2.23$$

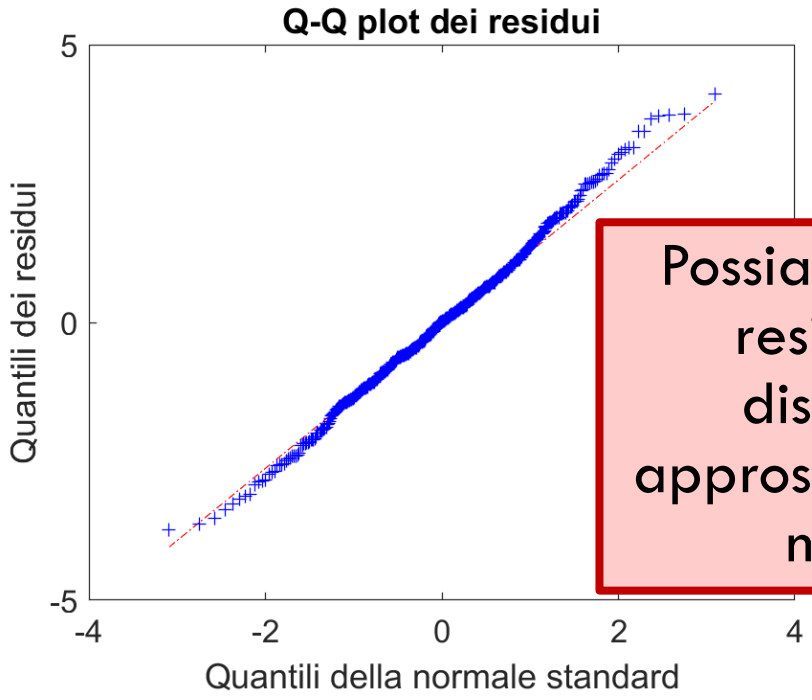
Commenti?



# ANALISI DEI RESIDUI: NORMALITA' E MEDIA NULLA

- Check distribuzione normale e media nulla



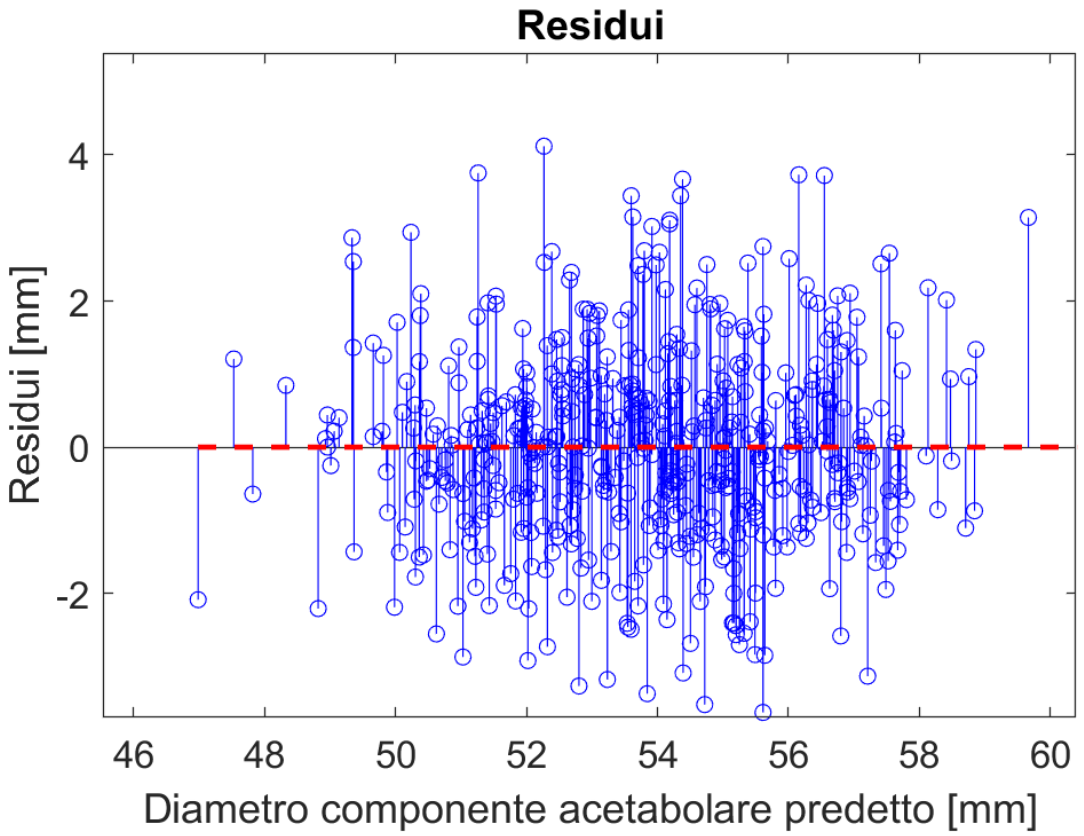



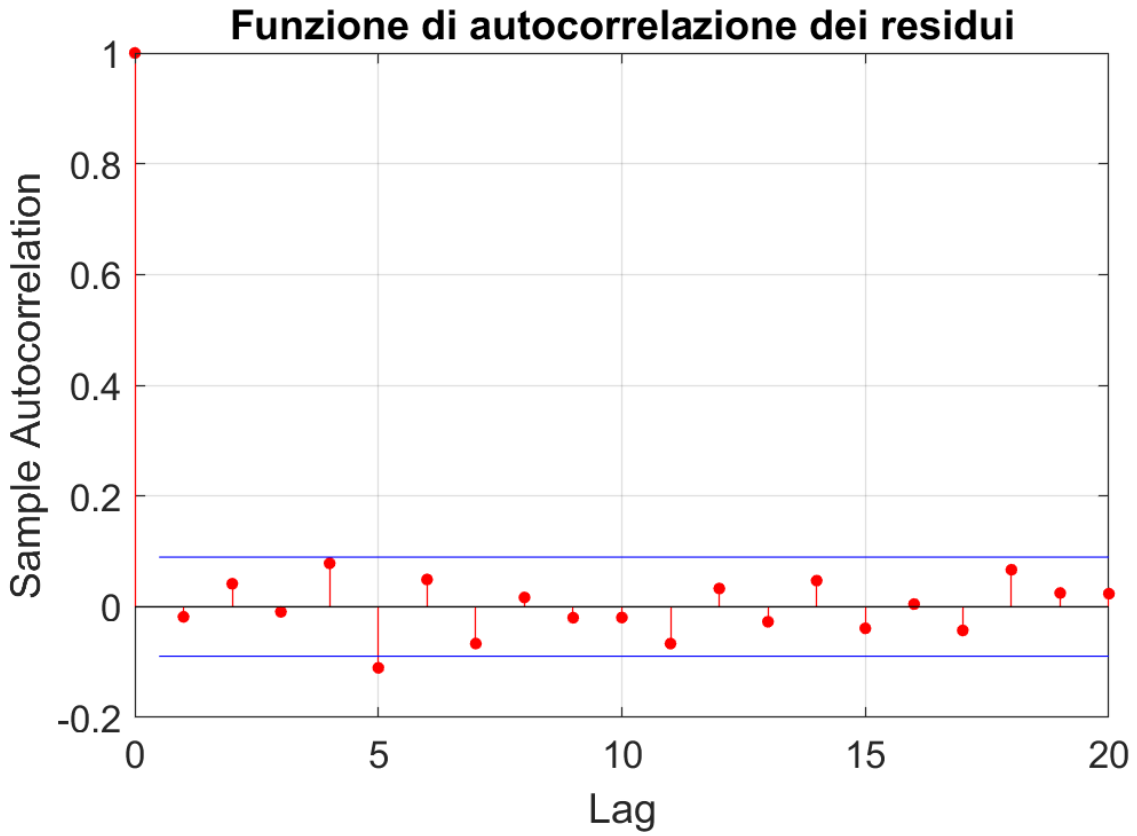

Possiamo dire che i residui hanno distribuzione approssimativamente normale?

- Skewness campionaria = 0.1138
- Curtosi campionaria = 3.0341
- Media campionaria = 4 x $10^{-13}$

- Lilliefors test: p-value=0.3664
- T test ($H_0$: $\mu = 0$): p-value = 1.00

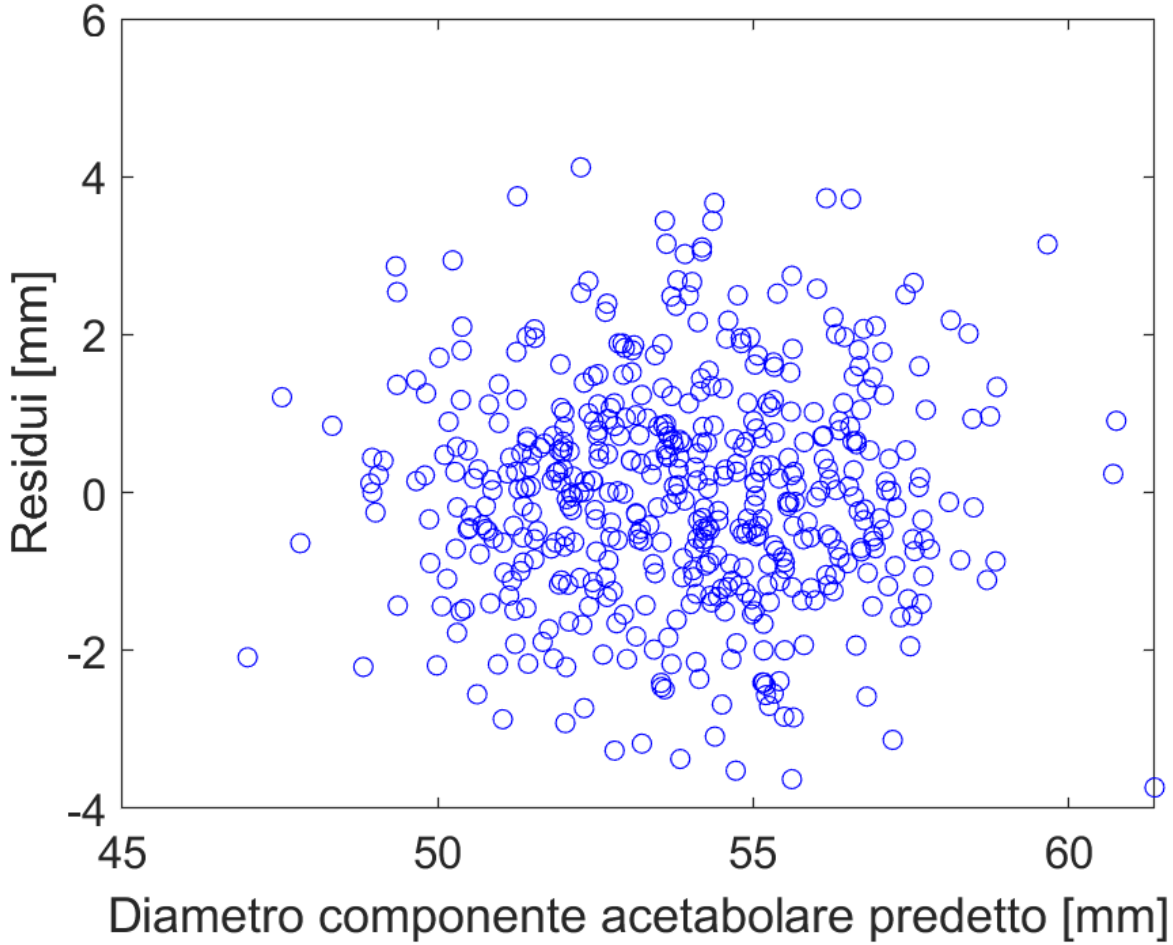
# ANALISI DEI RESIDUI: BIANCHEZZA




Possiamo dire che i residui sono a campioni scorrelati (bianchi)?

# ANALISI DEI RESIDUI: VARIANZA OMOGENEA, OUTLIER




- La varianza è omogenea?
- Sono presenti outlier?

# VALUTAZIONE STIME DEI PARAMETRI

| Variabili | Stime dei parametri $\widehat{\beta}_j$ | Standard error $SE_j$ | Coefficiente di variazione $CV_j$ | Intervallo di confidenza $[\widehat{\beta}_j - 1.96 * SE_j \; \widehat{\beta}_j + 1.96 * SE_j]$ | Z-score* $z_j$ |
|---|---|---|---|---|---|
| Intercetta | 21.0441 | 1.6262 | 7.73% | [17.86 24.23] | 12.94 |
| Altezza | 0.0884 | 0.0120 | 13.63% | [0.065 0.112] | 7.34 |
| Peso | 0.0735 | 0.0109 | 14.89% | [0.052 0.095] | 6.72 |
| Girovita | 0.0069 | 0.0115 | 166.37% | [-0.016 0.029] | 0.60 |
| Lunghezza piede | 0.5155 | 0.0578 | 11.22% | [0.402 0.629] | 8.91 |
| Età | -0.0111 | 0.0130 | 117.00% | [-0.037 0.014] | -0.85 |

*Con $\alpha$=0.05 la soglia critica è 1.96.

- Come valuteresti l'incertezza delle stime dei parametri?
- Quali variabili hanno un impatto statisticamente significativo sull'outcome?
- Quali variabili influiscono positivamente sul valore dell'outcome? Quali negativamente?