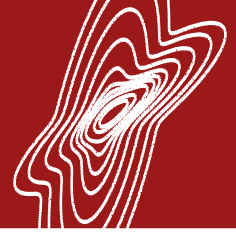


METODI STATISTICI PER LA BIOINGEGNERIA (B)

PARTE 8: REGRESSIONE LINEARE (PRIMA PARTE)

A.A. 2025-2026

Prof. Martina Vettoretti



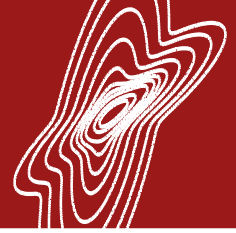
ANALISI DELLE RELAZIONI TRA VARIABILI (1 / 2)



- Nell'analisi di dati sperimentali, spesso siamo interessati a capire se sussiste una **relazione** tra una certa variabile di interesse, Y , e un'altra variabile, X , o un insieme di altre m variabili, X_i , $i = 1, \dots, m$.
 - Esempio: abbiamo un insieme di dati raccolti su una popolazione di individui e vogliamo studiare se esiste una relazione tra la pressione sistolica (Y) e due altre variabili, ovvero l'età (X_1) e il peso corporeo (X_2).
- In pratica, siamo interessati a capire se esiste una funzione f che consente di **predire** i valori di Y a partire dai valori delle variabili X_i :

$$Y = f(X_i)$$

- Esempio: siamo in grado di predire i valori di pressione sistolica, conoscendo l'età e il peso corporeo degli individui?

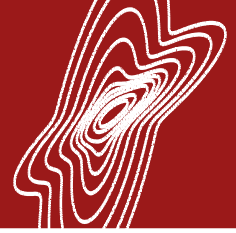


ANALISI DELLE RELAZIONI TRA VARIABILI (2/2)



$$Y = f(X_i)$$

- **Stima di f** a partire dall'analisi di un campione di valori di Y e dei campioni appaiati di valori delle X_i , $i = 1, \dots, m \rightarrow$ problema di **inferenza statistica**.
- Se Y è una variabile continua $\rightarrow$ la stima di f è un problema di **regressione**.
- Se assumiamo che f sia lineare $\rightarrow$ la stima di f è un problema di **regressione lineare**.



IL MODELLO DI REGRESSIONE LINEARE SEMPLICE



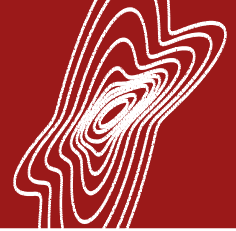
- Se vogliamo studiare la relazione tra due sole variabili, e ipotizziamo che questa sia di tipo lineare → **modello di regressione lineare semplice**

$$Y = \beta \cdot X + \beta_0$$

- Y: variabile dipendente, variabile di uscita, o outcome
 - X: variabile indipendente, variabile di ingresso, variabile esplicativa, regressore o predittore
 - β, β_0 : coefficienti di regressione, parametri del modello. β_0 è anche detto intercetta.
- In realtà la relazione tra le variabili in gioco non sarà mai perfettamente lineare → viene incluso nel modello un termine di errore casuale ε

$$Y = \beta \cdot X + \beta_0 + \varepsilon$$

- ε : errore di approssimazione del modello



ESEMPIO



L'emoglobina glicata (HbA1c) è un indicatore legato alla glicemia media negli ultimi 3 mesi, utilizzato nella diagnosi di diabete mellito. Tipicamente se $\text{HbA1c} > 6.5\%$ si sospetta la presenza di diabete.

La misura di HbA1c nel sangue viene prescritta nei soggetti a rischio di diabete, ma non fa parte degli esami del sangue di routine.

Negli esami di routine si misura tipicamente la glicemia a digiuno.

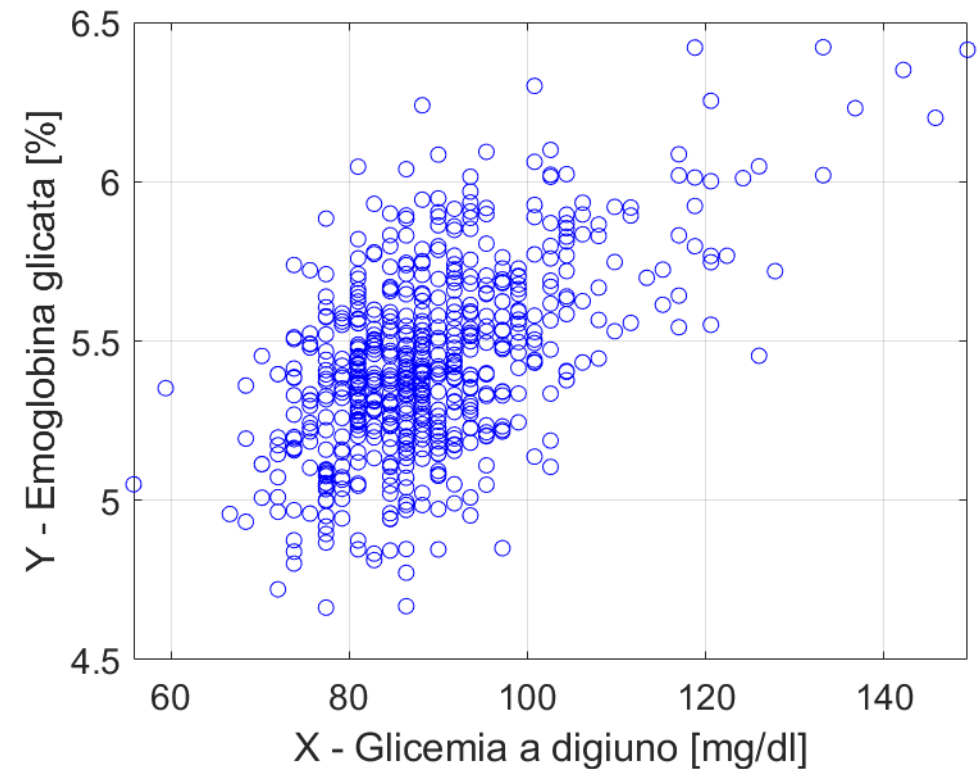
- Domanda: esiste una relazione tra la glicemia a digiuno (variabile X) e l'emoglobina glicata (variabile Y)? E' possibile predire i valori di emoglobina glicata a partire dalla glicemia a digiuno?
- Ipotizziamo che la relazione sia di tipo lineare...

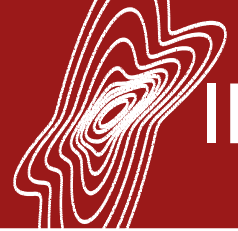


Domanda: esiste una relazione lineare tra le variabili X e Y?

1. Raccogliamo un campione bivariato contenente n osservazioni indipendenti di X e di Y appaiate: (x_i, y_i) , $i=1, \dots, n$.

Esempio: coppie di valori di glicemia a digiuno (x_i) ed emoglobina glicata (y_i) raccolte in 600 individui privi di diagnosi di diabete ($n=600$ osservazioni indipendenti).





Domanda: esiste una relazione lineare tra le variabili X e Y?

2. Consideriamo il modello di regressione lineare semplice per descrivere i dati del campione.

$$Y = \beta \cdot X + \beta_0 + \varepsilon$$

HbA1c Glicemia a digiuno

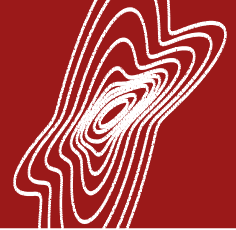
3. Con i dati a disposizione, stimiamo i parametri del modello di regressione.
 $\rightarrow \hat{\beta}, \hat{\beta}_0$



Domanda: esiste una relazione lineare tra le variabili X e Y?

4. Valutiamo la bontà del modello così identificato.

- Se il modello descrive bene i dati → possiamo concludere che sussiste una relazione tra X e Y e questa è approssimativamente lineare
- Se il modello non descrive bene i dati allora:
 - o non c'è alcuna relazione rilevante tra X e Y
 - oppure la relazione sussiste ma non è approssimativamente lineare.



CARATTERIZZAZIONE STATISTICA DEL MODELLO

Applicando l'equazione del modello di regressione semplice ai dati del campione si ottiene:

$$Y_i = \beta \cdot x_i + \beta_0 + \varepsilon_i, \quad i = 1, \dots, n$$

- $x_i \rightarrow$ quantità deterministica, nota, non casuale
- $\beta, \beta_0 \rightarrow$ parametri costanti incogniti
- $\varepsilon_i \rightarrow$ errore casuale $\rightarrow$ variabile aleatoria che assumiamo avere una distribuzione normale con media 0 e varianza σ_i^2 incognita:

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

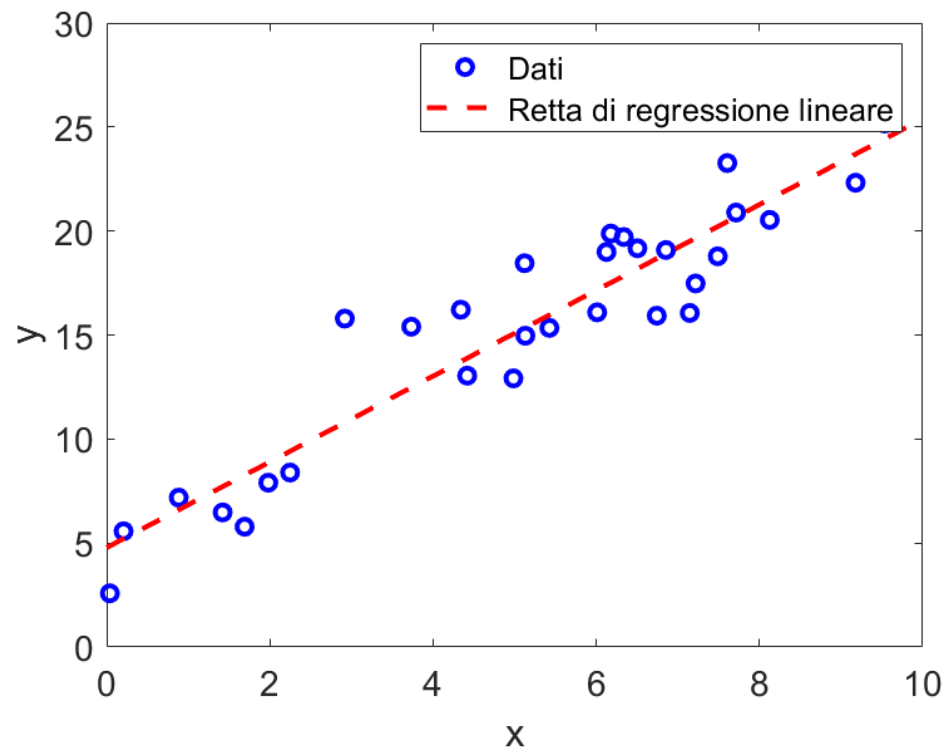
- Noto x_i , Y_i risulta anch'essa una variabile aleatoria normale:

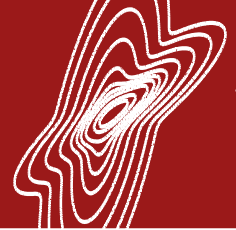
$$Y_i \sim N(\beta \cdot x_i + \beta_0, \sigma_i^2)$$

- Assumiamo che le diverse realizzazioni dell'errore siano indipendenti tra loro:

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$$

Avendo a disposizione **n osservazioni di X e Y**, (x_i, y_i) , $i=1, \dots, n$, il problema di identificazione del modello di regressione lineare consiste nello **stimare i valori dei parametri β e β_0** della retta che meglio approssima la relazione lineare tra X e Y.





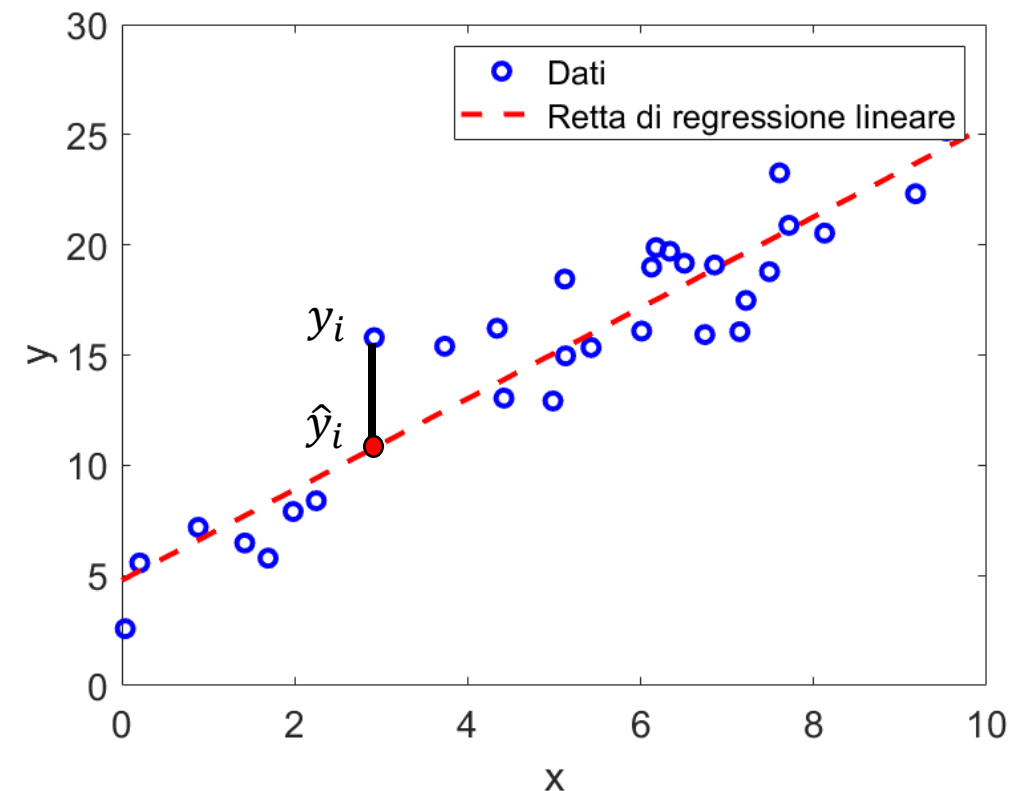
METODO DEI MINIMI QUADRATI LINEARI PER LA STIMA DEI COEFFICIENTI DI REGRESSIONE



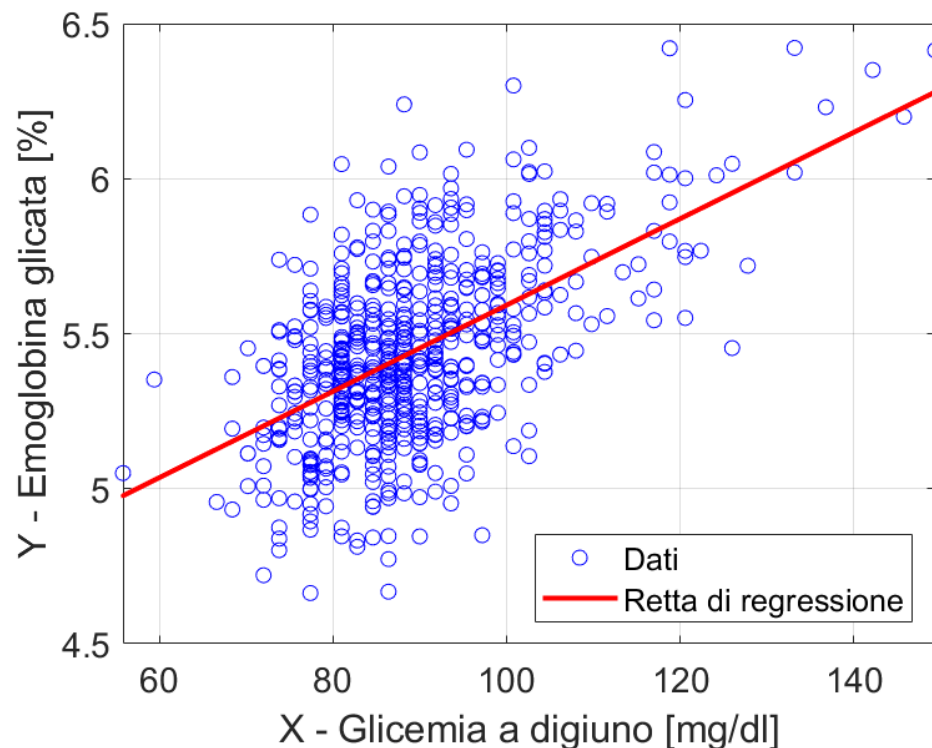
- **Assunzione:** la varianza dell'errore è costante $\rightarrow Var(\varepsilon_i) = \sigma^2 \forall i$.
- **Metodo dei minimi quadrati lineari:** si stimano i valori di β e β_0 che minimizzano la somma dei quadrati degli scarti tra i valori di uscita reali, y_i , e quelli predetti dal modello, $\hat{y}_i$.

$$\hat{y}_i = \beta \cdot x_i + \beta_0$$

$$\begin{aligned} \hat{\beta}, \hat{\beta}_0 &= \operatorname{argmin}_{\beta, \beta_0} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \\ &= \operatorname{argmin}_{\beta, \beta_0} \sum_{i=1}^n (y_i - (\beta \cdot x_i + \beta_0))^2 \end{aligned}$$



Domanda: esiste una relazione lineare tra la glicemia a digiuno (variabile X) e l'emoglobina glicata (variabile Y)? → stimiamo i parametri di un modello di regressione lineare semplice, con il metodo dei minimi quadrati lineari.



$$Y_i = \beta \cdot x_i + \beta_0 + \varepsilon_i$$



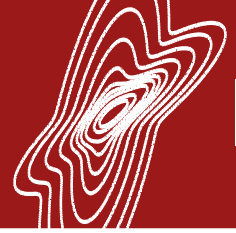
Stime dei parametri:
 $\hat{\beta} = 0.0139 [1/\text{mg/dl}]$

$$\hat{\beta}_0 = 4.2006$$



Retta di regressione:

$$\hat{y} = \hat{\beta} \cdot x + \hat{\beta}_0$$

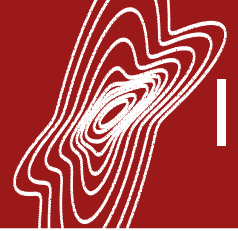


IL MODELLO DI REGRESSIONE LINEARE MULTIPLA



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m + \varepsilon$$

- Y : variabile dipendente, variabile di uscita, o outcome
- X_j : variabili indipendenti, di ingresso, esplicative, regressori o predittori
- β_j : coefficienti di regressione, parametri del modello. β_0 è anche detto intercetta.
- ε : errore di approssimazione del modello

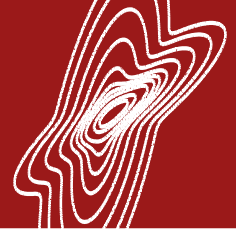


IL PROBLEMA DI REGRESSIONE LINEARE MULTIPLA



Domanda: esiste una relazione lineare tra delle variabili esplicative X_j e una variabile di outcome Y ? Ipotizzando che ci sia una dipendenza di tipo lineare tra Y e le variabili X_j , è possibile spiegare (o predire) i valori della variabile Y una volta noti i valori delle variabili X_j ?

1. **Raccogliamo un campione** contenente n osservazioni indipendenti di $X_j, j = 1, \dots, m$ e Y , appaiate tra loro: $(x_{i1}, x_{i2}, \dots, x_{im}, y_i), i=1, \dots, n$.
2. Consideriamo il **modello di regressione lineare multipla** per descrivere i dati.
3. Con i dati a disposizione, **stimiamo i parametri** del modello di regressione.
4. Valutiamo la **bontà del modello**.
 - Se il modello descrive bene i dati $\rightarrow$ sussiste una relazione tra Y e le variabili X_j e questa è approssimativamente lineare.
 - Se il modello non descrive bene i dati $\rightarrow$ non c'è alcuna relazione rilevante tra Y e le variabili X_j , oppure la relazione sussiste ma non è approssimativamente lineare.



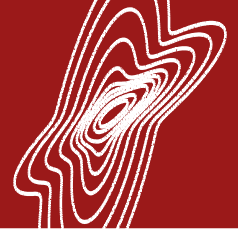
1. RACCOLTA DEL CAMPIONE

➤ Raccogliamo un campione $m+1$ -variato contenente n osservazioni appaiate per la variabile Y e le variabili X_j .

■ I dati del campione possono essere rappresentati in una tabella del tipo:

		Variabili			
Individui	Y	X_1	X_2	$\dots$	X_m
	y_1	x_{11}	x_{12}	$\dots$	x_{1m}
	y_2	x_{21}	x_{22}	$\dots$	x_{2m}
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	y_n	x_{n1}	x_{n2}	$\dots$	x_{nm}

■ Ogni colonna riporta una variabile e ogni riga riporta le osservazioni delle $m+1$ variabili appaiate (es. raccolte sullo stesso individuo/unità statistica).



2. IL MODELLO DI REGRESSIONE LINEARE MULTIPLA APPLICATO AI DATI RACCOLTI

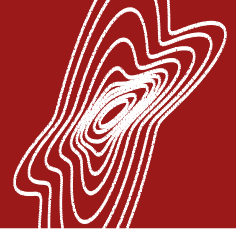


$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \varepsilon_i, \quad i = 1, \dots, n$$

Forma matrice-vettore:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

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



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

➤ $x_{ij} \rightarrow$ quantità note, deterministiche

➤ $\varepsilon_i \rightarrow$ errore del modello casuale e distribuito come una normale a media 0

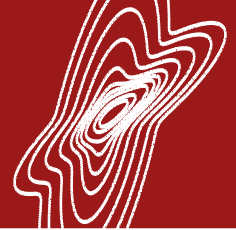
$$\varepsilon_i \sim N(0, \sigma_i^2)$$

➤ $Y_i \rightarrow$ variabili aleatorie normali:

$$Y_i \sim N\left(\beta_0 + \sum_{j=1}^m \beta_j \cdot x_{ij}, \sigma_i^2\right)$$

➤ Assunzione: le variabili aleatorie ε_i sono tra loro indipendenti

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$$



3. STIMA DEI PARAMETRI DEL MODELLO MEDIANTE I MINIMI QUADRATI LINEARI (1/2)



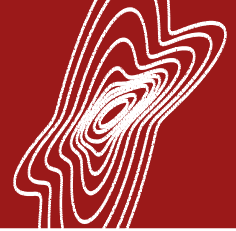
Assunzione: la varianza dell'errore è costante $\rightarrow Var(\varepsilon_i) = \sigma^2 \forall i$

Metodo dei minimi quadrati lineari: si seleziona la combinazione di parametri β_j tale per cui risulta minima la somma dei quadrati degli scarti tra i valori della variabile di uscita realmente osservati, y_i , e quelli predetti dal modello, $\hat{y}_i$.

$$\hat{y}_i = \beta_0 + \sum_{j=1}^m \beta_j x_{nj}$$

$$\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_m = \underset{\beta_0, \beta_1, \dots, \beta_m}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$y_i - \hat{y}_i$ sono
detti **residui**

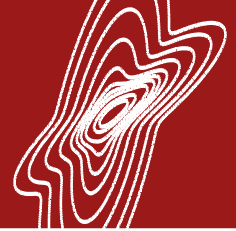


3. STIMA DEI PARAMETRI DEL MODELLO MEDIANTE I MINIMI QUADRATI LINEARI (2/2)



$$\begin{aligned}\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_m &= \operatorname{argmin}_{\beta_0, \beta_1, \dots, \beta_m} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \\ &= \operatorname{argmin}_{\beta_0, \beta_1, \dots, \beta_m} \sum_{i=1}^n (y_i - (\beta_0 + \sum_{j=1}^m \beta_j x_{ij}))^2 \\ \hat{\boldsymbol{\beta}} &= \operatorname{argmin}_{\boldsymbol{\beta}} \underbrace{(Y - X \cdot \boldsymbol{\beta})^T (Y - X \cdot \boldsymbol{\beta})}\end{aligned}$$

**Somma dei residui al quadrato o
*sum of squared error (SSE)***



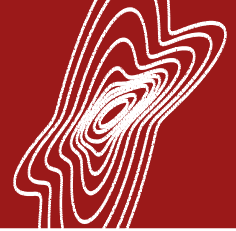
LO STIMATORE DEI PARAMETRI



- Se la matrice $X^T X$ è invertibile, la soluzione del problema di ottimizzazione precedente è:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

- $\hat{\beta}$ è il vettore contenente le stime dei parametri β_j del modello di regressione lineare multipla nelle variabili X_j che minimizza SSE, ovvero che meglio approssima i dati.
- **Nota:** la stima $\hat{\beta}$ è definita solo se $X^T X$ è invertibile, ovvero ha rango pieno o determinante $\neq 0$. Ciò si verifica solo se le colonne di X sono linearmente indipendenti $\rightarrow$ nessuna variabile è combinazione lineare delle altre.



DIMOSTRAZIONE (BONUS)



$$SSE = (Y - X \cdot \beta)^T (Y - X \cdot \beta) = Y^T Y - Y^T X \beta - \beta^T X^T Y + \beta^T X^T X \beta =$$

Scalars di uguale valore

$$= Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta$$

Per trovare il minimo, deriviamo SSE rispetto a β e poniamo il risultato a 0:

$$\frac{\partial SSE}{\partial \beta} = -2X^T Y + 2X^T X \beta = 0$$
$$\hat{\beta} = (X^T X)^{-1} X^T Y$$



Poiché Y è un vettore aleatorio di distribuzione normale, lo stimatore

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

è anch'esso un vettore aleatorio normale.

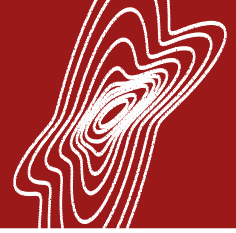
Proprietà:

- $E[\hat{\beta}_j] = \beta_j \rightarrow$ lo stimatore è corretto (non distorto)
- $Cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$



$$\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$$

- La varianze degli stimatori $\hat{\beta}_j$ sono gli elementi sulla diagonale di $\text{Cov}(\hat{\boldsymbol{\beta}})$.
- Per calcolarle abbiamo bisogno di $\mathbf{X}$ (nota) e σ^2 , tipicamente incognita $\rightarrow \sigma^2$ si può stimare a posteriori (= dopo la stima dei β_j)
- La deviazione standard dello stimatore $\hat{\beta}_j$ viene detta **standard error (SE) del parametro $\hat{\beta}_j$, SE_j** .
- **Coefficiente di variazione** delle stime dei parametri: $CV_j = \frac{SE_j}{|\hat{\beta}_j|} \cdot 100$
- SE_j e CV_j rappresentano l'incertezza sulla stima di β_j .



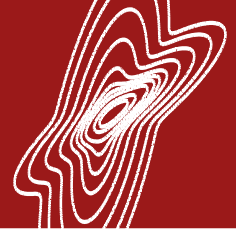
STIMA DELLA VARIANZA DELL'ERRORE



- La varianza dell'errore del modello, σ^2 , può essere stimata a posteriori usando lo stimatore:

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

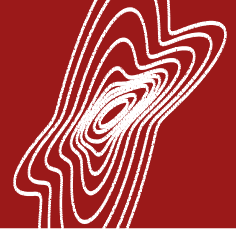
- Nota: la divisione per $n-m-1$ anziché per n garantisce che lo stimatore sia non distorto, ovvero che $E[\hat{\sigma}^2] = \sigma^2$.



ESEMPIO



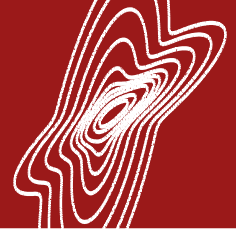
- Obesità, ipertensione e ipercolesterolemia sono condizioni spesso associate al diabete mellito.
- Domanda: vogliamo investigare se sussiste una relazione lineare tra l'emoglobina glicata (Y) e un insieme di altre 6 variabili:
 - X_1 : glicemia a digiuno [mg/dl]
 - X_2 : indice di massa corporea (IMC) [Kg/m^2]
 - X_3 : colesterolo totale [mg/dl]
 - X_4 : colesterolo HDL [mg/dl]
 - X_5 : pressione arteriosa sistolica [mmHg]
 - X_6 : pressione arteriosa diastolica [mmHg]



ESEMPIO: IL DATASET

- **Dataset:** misure delle variabili X_1 - X_6 e Y raccolte in 600 diversi individui privi di diagnosi di diabete ($n=600$ osservazioni indipendenti).

Individuo	Emoglobina glicata Y [%]	Glicemia a digiuno X_1 [mg/dl]	IMC X_2 [Kg/m ²]	Colesterolo totale X_3 [mg/dl]	Colesterolo HDL X_4 [mg/dl]	Pressione sistolica X_5 [mmHg]	Pressione diastolica X_6 [mmHg]
1	5.4	102.4	24.5	191.3	80.3	149.4	89.5
2	5.6	99.8	23.6	202.3	67.8	108.5	71.5
3	6.2	110.3	27.3	231.1	37.4	152.3	86.3
4	5.5	78.4	24.9	210.3	65.2	110.5	65.8
5	6.5	138.5	30.4	275.4	39.2	144.1	83.4
...	...	...	...	...	...	...	...



ESEMPIO: IL MODELLO DI REGRESSIONE



- Ipotizziamo il seguente modello di regressione lineare multipla:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon$$

- Stimiamo i parametri del modello con i dati a disposizione mediante il metodo dei minimi quadrati lineari.

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$



$$\hat{\beta}_0 = 3.8828$$

$$\hat{\beta}_1 = 0.0115 \text{ [1/mg/dl]}$$

$$\hat{\beta}_2 = 0.0145 \text{ [1/Kg/m}^2\text{]}$$

$$\hat{\beta}_3 = 0.0007 \text{ [1/mg/dl]}$$

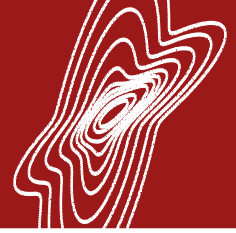
$$\hat{\beta}_4 = -0.0029 \text{ [1/mg/dl]}$$

$$\hat{\beta}_5 = 0.0029 \text{ [1/mmHg]}$$

$$\hat{\beta}_6 = -0.0032 \text{ [1/mmHg]}$$

- $SSE = 76.68$

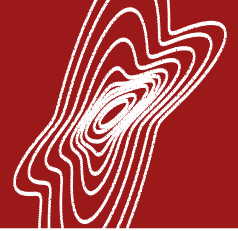
- $\hat{\sigma}^2 = \frac{SSE}{n-m-1} = \frac{76.68}{600-7} = 0.1293$



4. VALUTAZIONE DELLA BONTA' DEL MODELLO



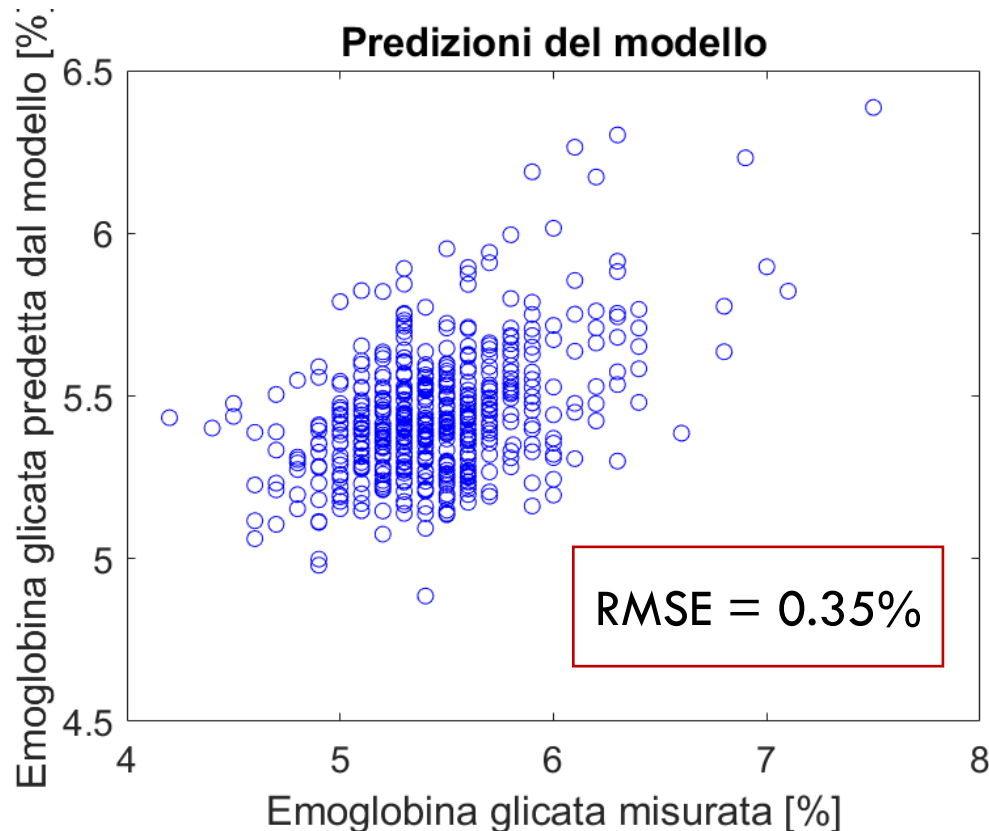
- Dopo aver stimato i parametri del modello è importante chiedersi se il modello risultante descrive i dati in maniera soddisfacente, ovvero approssima in maniera soddisfacente la relazione tra le variabili considerate.
- Criteri per valutare la bontà del modello:
 - A. Confronto tra i valori dell'outcome reali e quelli predetti
 - B. Coefficiente di determinazione
 - C. F test
 - D. Analisi dei residui



A. CONFRONTO TRA I VALORI DELL'OUTCOME REALI E QUELLI PREDETTI



- Possiamo confrontare in un grafico a dispersione i valori dell'outcome predetti dal modello (asse y) con quelli realmente misurati (asse x).



- Metriche per valutare lo scostamento tra outcome predetta e outcome reale:

- **Mean square error (MSE)**

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- **Root mean square error (RMSE)**

$$RMSE = \sqrt{MSE}$$

- Che unità di misura hanno MSE e RMSE?

- **Total sum of squares (SST):** devianza campionaria di $y_i \rightarrow$ rappresenta la variabilità della variabile di uscita

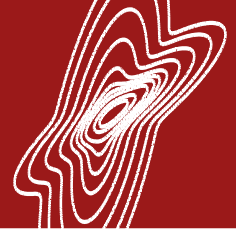
$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

- SST si può scrivere come somma di due componenti:

$$SST = \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{SSR}}$$

Sum of squared errors (SSE): devianza dei residui, componente di variabilità della variabile di uscita dovuta all'errore

Regression sum of squares (SSR): componente di variabilità della variabile di uscita spiegata dalle variabili di ingresso X_i



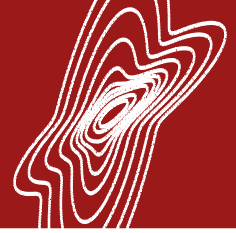
B. IL COEFFICIENTE DI DETERMINAZIONE R^2



- **Coefficiente di determinazione R^2 :** frazione della variabilità della variabile di uscita spiegata dalle variabili di ingresso (adimensionale).

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

- R^2 varia tra 0 e 1 ed è tanto maggiore quanto più il modello di regressione lineare è in grado di spiegare i valori della variabile di uscita.
- $R^2=1 \rightarrow$ la relazione tra le variabili di ingresso e di uscita è perfettamente lineare
 - $R^2=0 \rightarrow$ la variabile di uscita non è affatto spiegabile con una regressione lineare delle variabili di ingresso



ESEMPIO: R^2

➤ Per valutare la bontà del modello di regressione lineare multipla identificato nell'esempio precedente, calcoliamo il coefficiente R^2 .

- $SSE = 76.68$

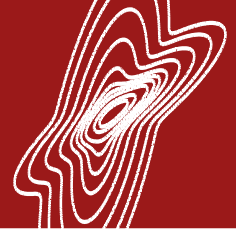
- $SST = 109.80$

- $R^2 = 1 - \frac{SSE}{SST} = 0.3016$



Le variabili X_1 - X_6 considerate sono in grado di spiegare circa il 30% della variabilità di Y .

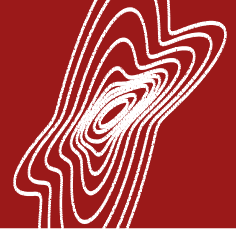
Si tratta di una frazione significativa?



C. F TEST (1 / 2)



- Quando R^2 è basso, viene spontaneo chiedersi se esso sia significativamente diverso da 0. Questo equivale a chiedersi se almeno uno dei coefficienti β_i associati alle variabili x_i sia significativamente diverso da 0.
- Rispondiamo a questa domanda con un test di verifica di ipotesi: **F test**.
- Assunzioni: i termini di errore ε_i hanno distribuzione normale.
- Sistema di ipotesi:
 - $H_0: \beta_1 = \beta_2 = \dots = \beta_m = 0$
 - $H_1: \text{almeno un coefficiente } \beta_i \neq 0, i \neq 0$



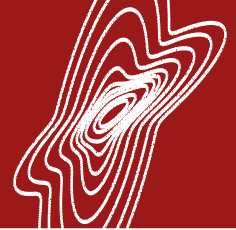
C. F TEST (2/2)



➤ Statistica del test:

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

- Quando vale H_0 , F ha una distribuzione F di Fisher con gradi di libertà m e $n-m-1$.
- Regola decisionale (test a una coda destro, ci interessa capire se il numeratore di F è significativamente maggiore del denominatore):
- Se $F > F_{\alpha, m, n-m-1} \rightarrow$ rifiutiamo H_0
 - Se $F \leq F_{\alpha, m, n-m-1} \rightarrow$ non possiamo rifiutare H_0



ESEMPIO: F TEST

➤ Il modello è in grado di spiegare una porzione significativamente diversa da 0 della variabilità di Y? → F test

➤ Sistema di ipotesi:

- $H_0: \beta_1 = \beta_2 = \dots = \beta_6 = 0$
- $H_1: \text{almeno un coefficiente } \beta_j \neq 0, j \neq 0$

➤ Valore osservato per la statistica F:

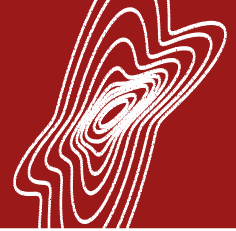
$$F_{oss} = \frac{(SST - SSE)/m}{SSE/(n - m - 1)} = \frac{(109.80 - 76.68)/6}{76.68/(600 - 7)} = 42.7$$

➤ Possiamo rifiutare l'ipotesi nulla con livello di significatività $\alpha=5\%$?

- $F_{\alpha,m,n-m-1} = F_{0.05,6,600-7} = 2.11$
- $p\text{-value} = 2.44 * 10^{-44}$



Rifiutiamo $H_0 \rightarrow$ almeno uno dei coefficienti β_j è significativamente $\neq 0$, e il modello predice una porzione significativa della variabilità di Y.



D. ANALISI DEI RESIDUI

- **Residui:** differenza tra i valori osservati di Y e le predizioni modello

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

- Se il modello di regressione lineare è una buona approssimazione per descrivere i dati, i residui devono presentare le proprietà statistiche dell'errore del modello ε_i .

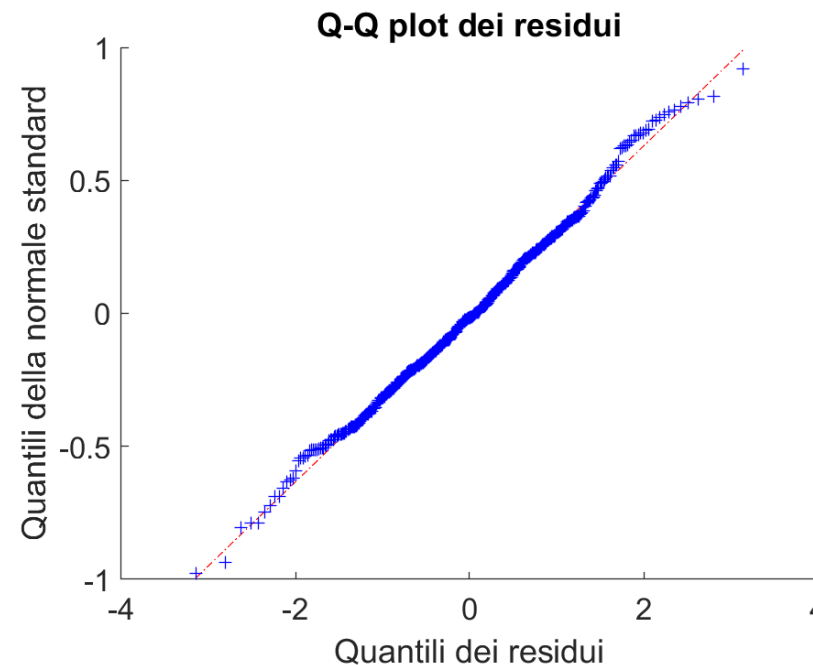
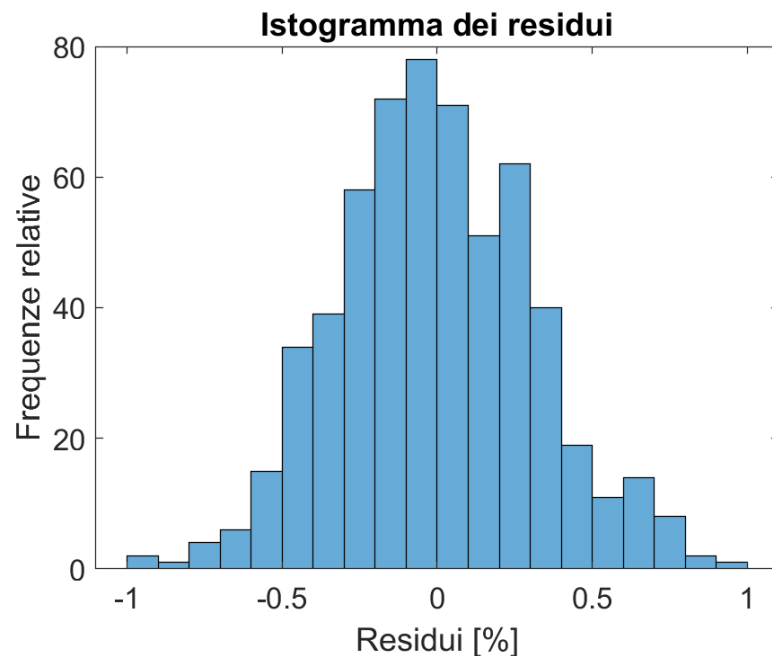
1. I residui devono avere **distribuzione approssimativamente normale**.
2. I residui devono avere **media nulla**.
3. I residui devono essere **scorrelati**.
4. I residui devono avere **varianza omogenea**.

D.1 DISTRIBUZIONE NORMALE



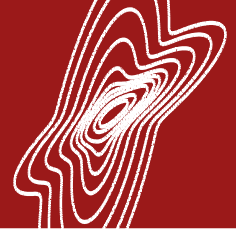
- Per verificare se la distribuzione dei residui è normale possiamo usare l'istogramma delle frequenze relative, un test di normalità, il q-q plot e gli indici di forma campionaria.

Risultati per il nostro esempio:



- Test di Lilliefors:
 - P-value = 0.26
- Indice di skewness campionaria: 0.14
- Indice di curtosi campionaria: 3.01

Cosa concludiamo?



D.2 MEDIA NULLA



- Calcoliamo la media campionaria dei residui:

$$\bar{r} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)$$

- Applichiamo un **t test** per verificare se la media dei residui è significativamente diversa da 0.

- Risultati per il nostro esempio:

- $\bar{r} = -0.0059$
- P-value del t test: 0.65

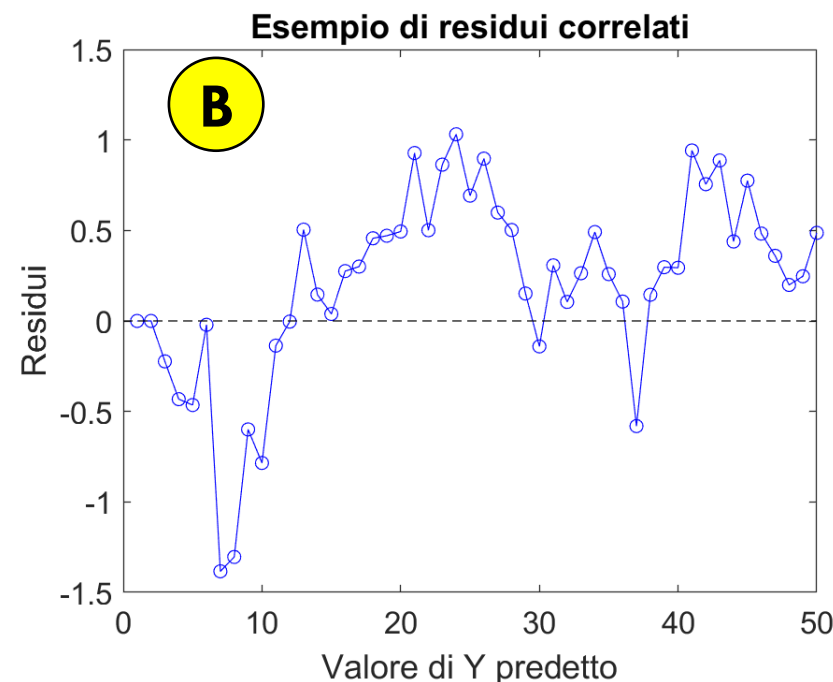
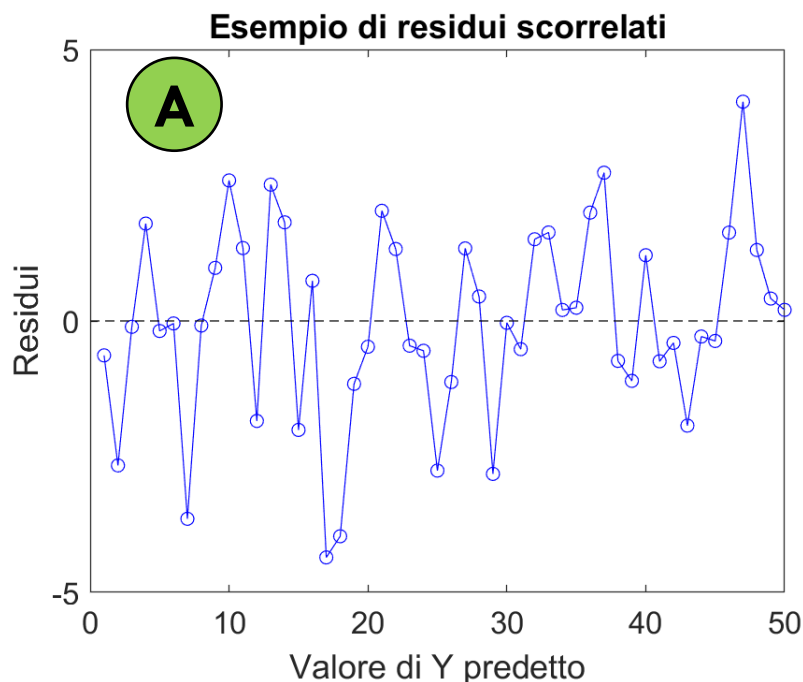


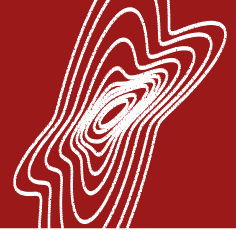
Il t test di per sé non mi consente di dire nulla, essendo il risultato negativo. Tuttavia poiché osserviamo una media campionaria molto vicina a 0 (effect size) possiamo pensare che i residui siano ragionevolmente a media nulla.

D.3 AUTOCORRELAZIONE DEI RESIDUI



- Poiché le osservazioni del dataset sono indipendenti, è ragionevole attendersi che i residui siano a **campioni scorrelati**, ovvero che il valore del residuo k -esimo non dipenda dai valori dei residui in posizioni precedenti a $k \rightarrow$ **bianchezza dei residui**
- Ispezione visiva: analisi del plot dei residui vs il valore predetto di Y ($\hat{y}_i$)

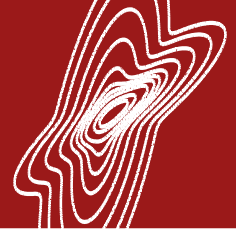




VALUTAZIONE QUANTITATIVA DELLA BIANCHEZZA DEI RESIDUI



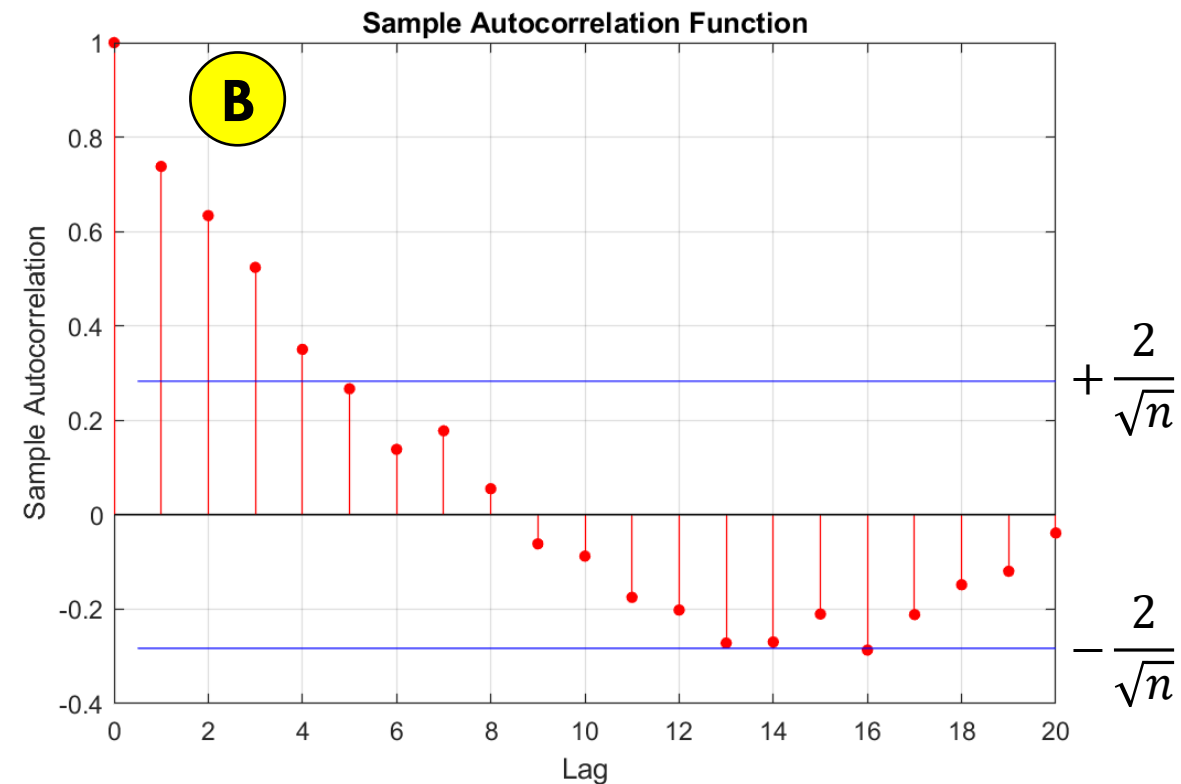
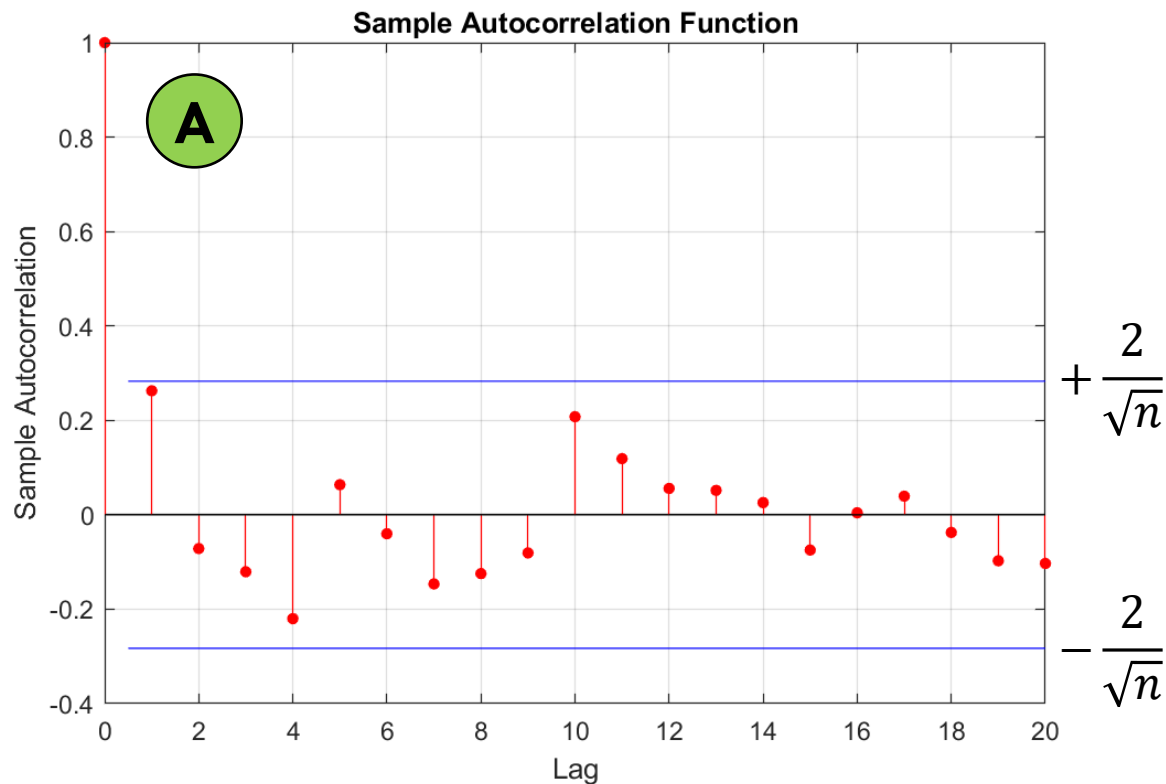
- Calcoliamo la **funzione di autocorrelazione** dei residui ordinati in base alle predizioni del modello ($\hat{y}_i$).
- I valori della funzione di autocorrelazione corrispondono alla correlazione tra il segnale dei residui e la sua versione ritardata di un certo numero di campioni (lag).
 - Lag 0 → correlazione del segnale $r_1, r_2, \dots, r_n$ correlato con se stesso (sempre pari a 1)
 - Lag 1 → correlazione del segnale $r_2, \dots, r_n$ con $r_1, r_2, \dots, r_{n-1}$
 - Lag 2 → correlazione del segnale $r_3, r_2, \dots, r_n$ con $r_1, r_2, \dots, r_{n-2}$
 - ...
 - Lag k → correlazione del segnale $r_k, r_{k+1}, \dots, r_n$ con $r_1, r_2, \dots, r_{n-k}$
- Se il segnale dei residui è scorrelato, ci aspettiamo che i valori della funzione di autocorrelazione stiano all'interno della banda di confidenza $\pm 2/\sqrt{n}$



ESEMPI DI FUNZIONE DI AUTOCORRELAZIONE



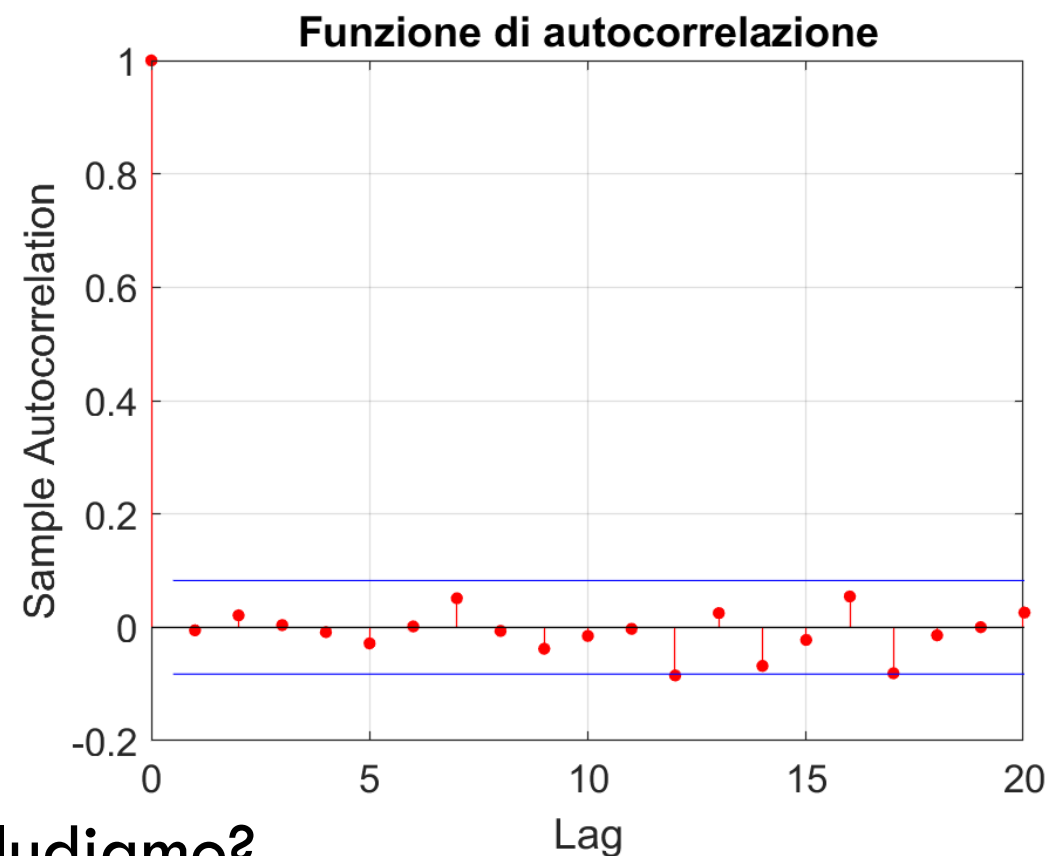
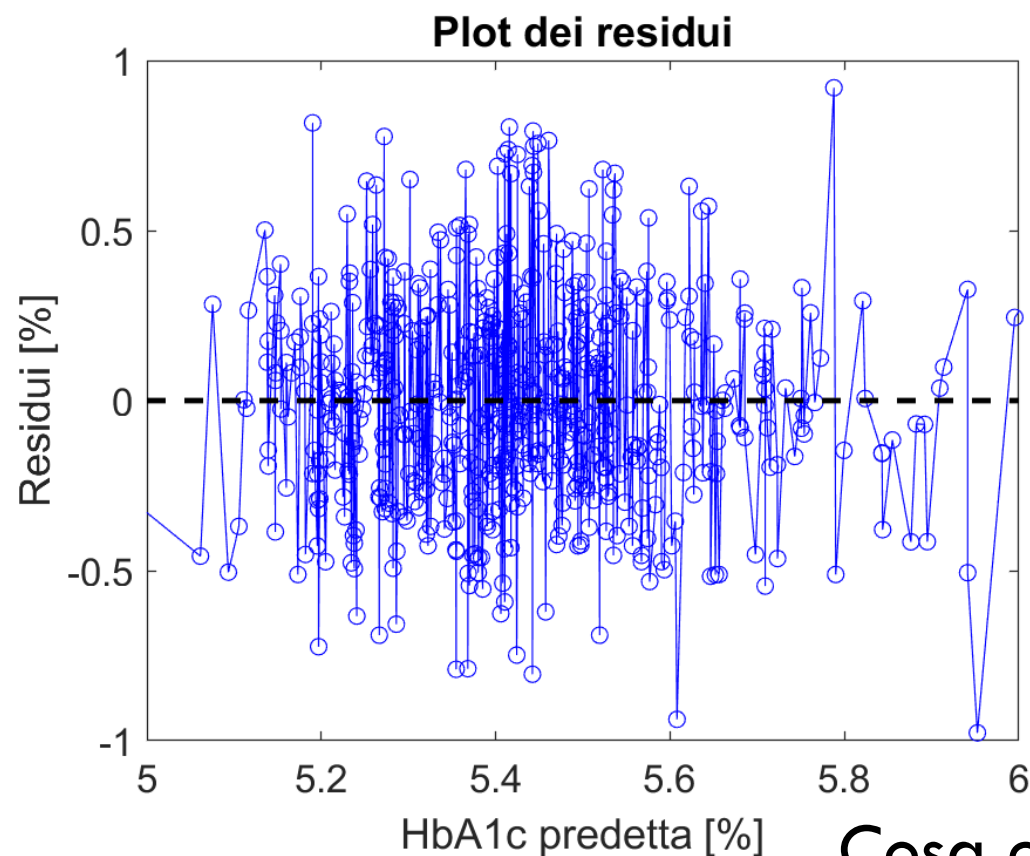
➤ Funzione di autocorrelazione stimata per le due serie di residui di slide 40.



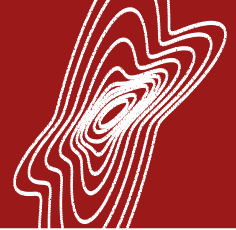


ESEMPIO: ANALISI DELLA BIANCHEZZA DEI RESIDUI

- Analizziamo la bianchezza dei residui del modello di regressione lineare multipla per la predizione dell'emoglobina glicata.

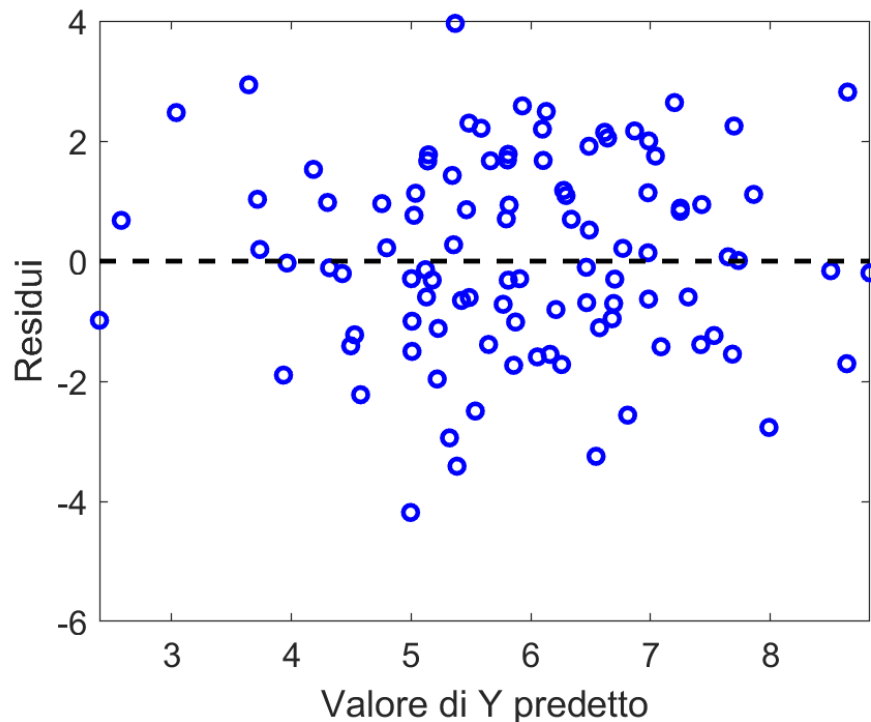


Cosa concludiamo?

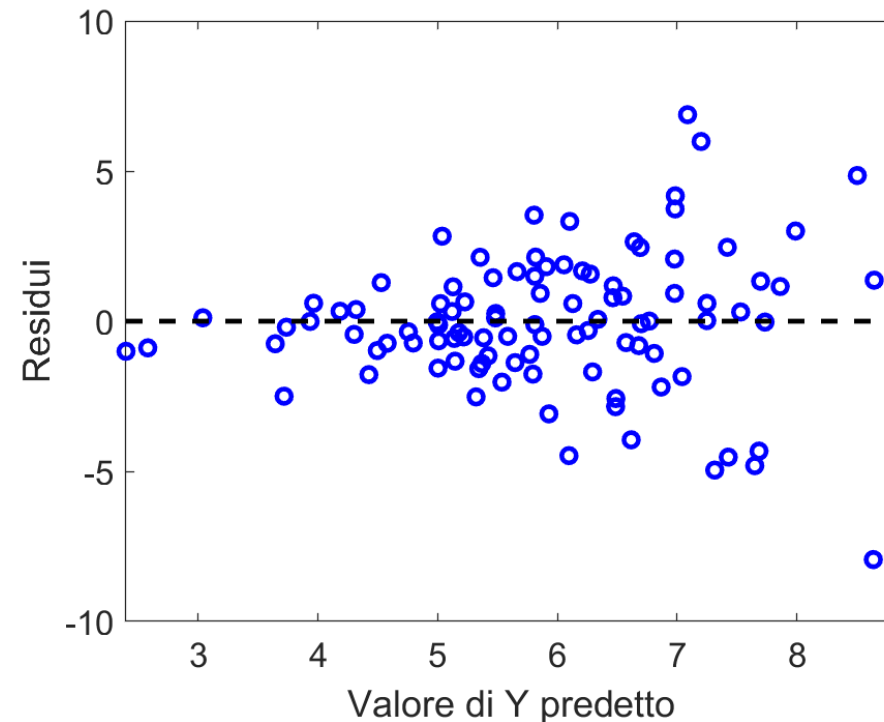


D.4 VARIANZA OMOGENEA

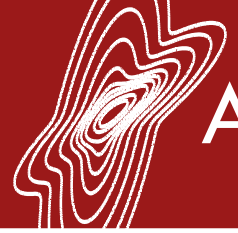
- La varianza dei residui deve essere omogenea al variare del valore predetto di Y.
- Valutazione tramite ispezione visiva con un grafico di dispersione avente i valori dei residui sull'asse y e i valori dell'outcome Y sull'asse x.



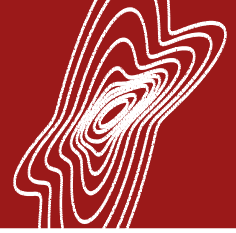
Varianza omogenea



Varianza non omogenea

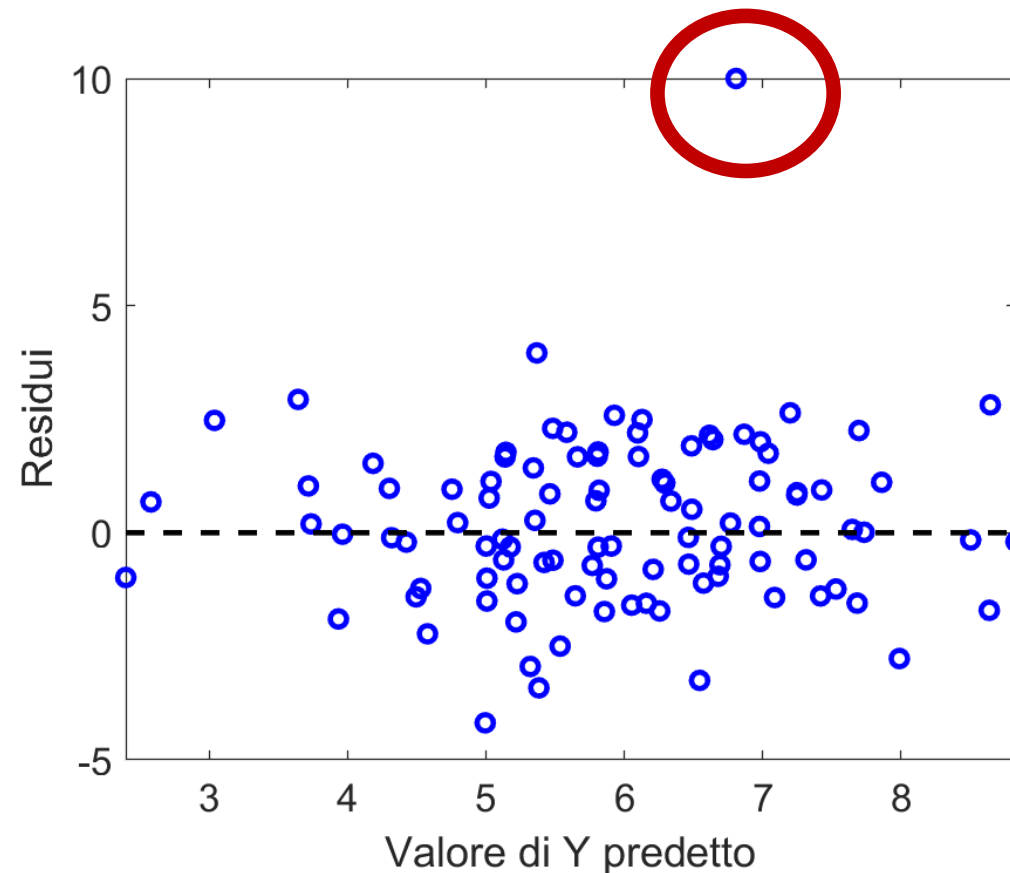


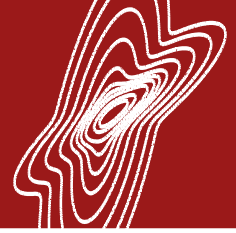
- Il plot dei residui al variare del valore predetto di Y ci consente anche di riscontrare eventuali altre anomalie nel comportamento dei residui, quali:
 - Outlier
 - Trend nell'andamento dei residui



OUTLIER NEI RESIDUI

- **Outlier:** osservazioni per le quali il modello commette un errore considerevolmente maggiore rispetto alle altre osservazioni.

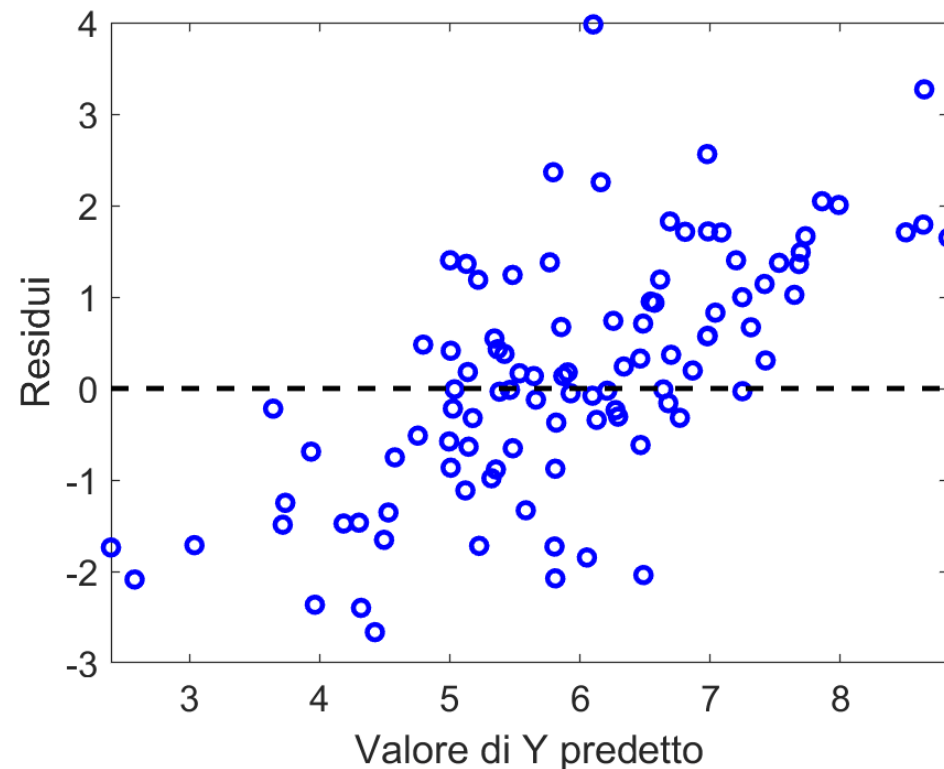


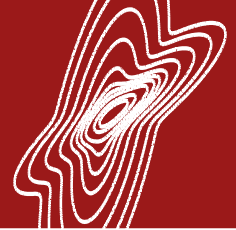


TREND NELL'ANDAMENTO DEI RESIDUI



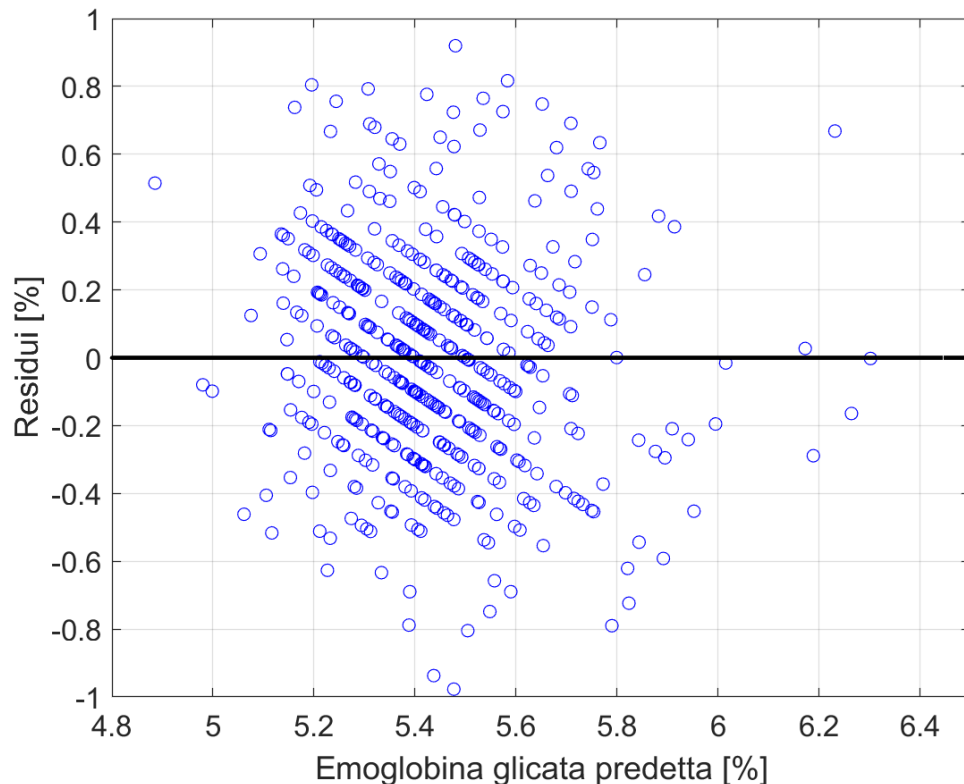
- **Trend nell'andamento dei residui:** l'errore del modello non è casuale, ma dipende dal valore di Y . Ciò significa che l'approssimazione lineare non è adeguata a descrivere i dati.





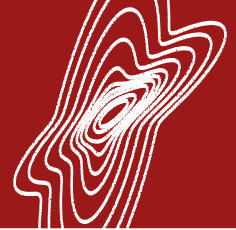
ESEMPIO: PLOT DEI RESIDUI VS. Y

- Riprendiamo l'esempio sulla regressione lineare multipla dell'emoglobina glicata e visualizziamo il plot dei residui al variare dell'emoglobina glicata.



- La varianza dei residui risulta omogenea?
- Sono visibili significativi outlier o trend nei residui?

- Una volta appurato che la bontà del modello è soddisfacente possiamo analizzarne i coefficienti per ricavare utili informazioni relativamente all'effetto delle variabili esplicative sull'outcome.
- **Intercetta β_0 :** valore medio di Y quando le variabili X_j sono tutte nulle → componente di Y **indipendente dai valori delle X_j**
- **Coefficiente β_j :** l'incremento medio di Y che si ottiene aumentando X_j di 1 unità e tenendo costanti tutte le altre variabili → **impatto di X_j su Y**
- **Valore assoluto di β_j :**
 - Se β_j è vicino a 0 → la variabile X_j ha un impatto trascurabile su Y
 - Se β_j è significativamente diverso da 0 → la variabile X_j ha un impatto significativo su Y
- **Segno di β_j :**
 - $\beta_j > 0$ → all'aumentare di X_j aumenta anche Y
 - $\beta_j < 0$ → all'aumentare di X_j diminuisce Y



SIGNIFICATIVITA' STATISTICA DEI COEFFICIENTI DI REGRESSIONE LINEARE



Domanda: il coefficiente β_j è significativamente diverso da 0? La variabile X_j ha un impatto significativo su Y ?

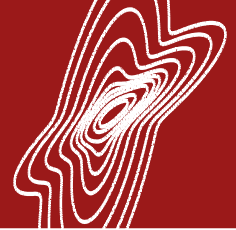
- Valutiamo il valore di $\hat{\beta}_j$ e il suo intervallo di confidenza.
 - Lo stimatore $\hat{\beta}_j$ ha distribuzione normale con deviazione standard $\hat{\sigma}\sqrt{v_j}$, dove v_j è l'elemento in posizione j della diagonale di $(X^T X)^{-1}$.
 - **Intervallo di confidenza 95%:** $\hat{\beta}_j \pm 1.96 \cdot \hat{\sigma}\sqrt{v_j} \rightarrow$ il valore vero di β_j è compreso in questo intervallo con probabilità circa pari al 95%
- Valutiamo l'ampiezza dell'intervallo di confidenza e se questo comprende lo 0.



- **Test statistico** per verificare l'ipotesi che il coefficiente β_j sia significativamente $\neq 0$.
- Assunzioni: gli errori ε_i hanno distribuzione normale con media 0 e varianza σ^2 .
- Sistema di ipotesi:
 - $H_0: \beta_j = 0$
 - $H_1: \beta_j \neq 0$
- Statistica del test:

Z-score del coefficiente β_j $\longrightarrow$
$$z_j = \frac{\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{v_j}}$$
 $\longleftarrow$ v_j è l'elemento in posizione j della diagonale di $(\mathbf{X}^T \mathbf{X})^{-1}$

- Quando vale H_0 , z_j ha distribuzione t di Student con $n-m-1$ gradi di libertà.
 - Se $|z_j| > t_{\frac{\alpha}{2}, n-m-1} \rightarrow$ rifiutiamo H_0
 - Se $|z_j| \leq t_{\frac{\alpha}{2}, n-m-1} \rightarrow$ non possiamo rifiutare H_0



ESEMPIO: INTERPRETAZIONE DEI COEFFICIENTI DI REGRESSIONE



- Analizziamo le stime dei coefficienti del modello di regressione lineare multipla dell'emoglobina glicata.

Variabile	Coefficiente stimato	Intervallo di confidenza al 95%	Z-score	P-value
Glicemia a digiuno	0.0115	[0.097 0.0134]	12.60	$1.99 \cdot 10^{-32}$
IMC	0.0145	[0.0075 0.0214]	4.16	$3.6 \cdot 10^{-5}$
Colesterolo totale	0.0007	[-0.0001 0.0014]	1.767	0.0777
Colesterolo HDL	-0.0029	[-0.0053 -0.0005]	-2.45	0.0145
Pressione sistolica	0.0029	[0.0008 0.0051]	2.72	0.0067
Pressione diastolica	-0.0032	[-0.0069 0.0005]	-1.71	0.0885

Che commenti possiamo fare?

$$t_{0.025, 600-7} = 1.96$$