METODI STATISTICI PER LA BIOINGEGNERIA (B)

PARTE 8: REGRESSIONE LINEARE (PRIMA PARTE)

A.A. 2024-2025

Prof. Martina Vettoretti

ANALISI DELLE RELAZIONI TRA VARIABILI

- 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,…,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:

$$Y = f(X_i)$$

- Il problema di **stimare f** a partire dall'analisi di un campione di valori di Y e dei campioni appaiati di valori delle X_i, i =1,…,m è un problema di **inferenza statistica**.
- Quando Y è una variabile continua, si tratta di un problema di **regressione**.
- Se assumiamo che f sia lineare, il problema è di **regressione lineare**.

IL MODELLO DI REGRESSIONE LINEARE SEMPLICE

- Quando vogliamo studiare la relazione tra due sole variabili, e ipotizziamo che questa sia di tipo lineare, si utilizza il 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

IL PROBLEMA DI REGRESSIONE LINEARE SEMPLICE

Ci chiediamo se sussista una relazione lineare tra le variabili X e Y. Come possiamo rispondere a questa domanda?

- Raccogliamo un campione bivariato contenente n osservazioni indipendenti di X e di Y appaiate: (x_i, y_i), i=1,…,n.
- Consideriamo il modello di regressione lineare semplice per descrivere i dati del campione.
- Con i dati a disposizione, stimiamo i parametri del modello di regressione.
- 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:

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

- x_i viene considerata una quantità deterministica, nota, non casuale
- β, β_0 sono delle costanti incognite
- ε_i è considerato un errore casuale → 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:

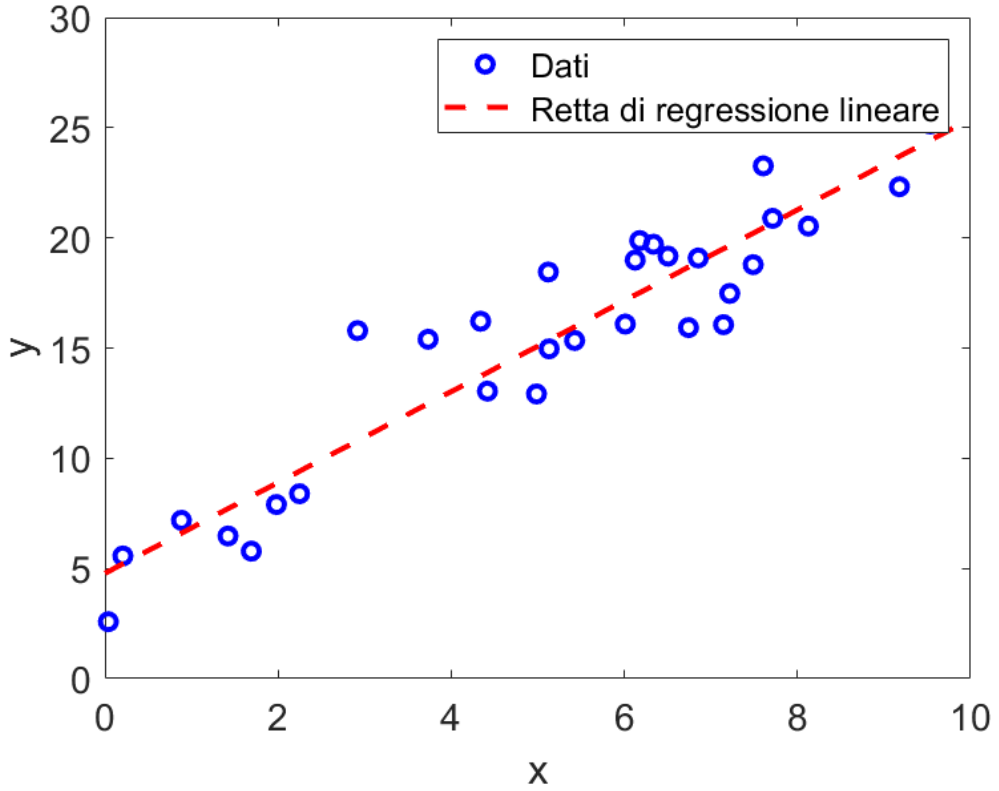
$$Y_i \sim N(\beta \cdot x_i + \beta_0, \sigma_i^2)$$

- Assumiamo inoltre che le diverse realizzazioni dell'errore casuale siano indipendenti tra loro, ovvero:

$$Cov(\varepsilon_i, \varepsilon_j) = 0, \qquad i \neq j$$

IDENTIFICAZIONE DEL MODELLO DI REGRESSIONE LINEARE

Avendo a disposizione n osservazioni di X e Y, (x_i, y_i), i=1,…,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.



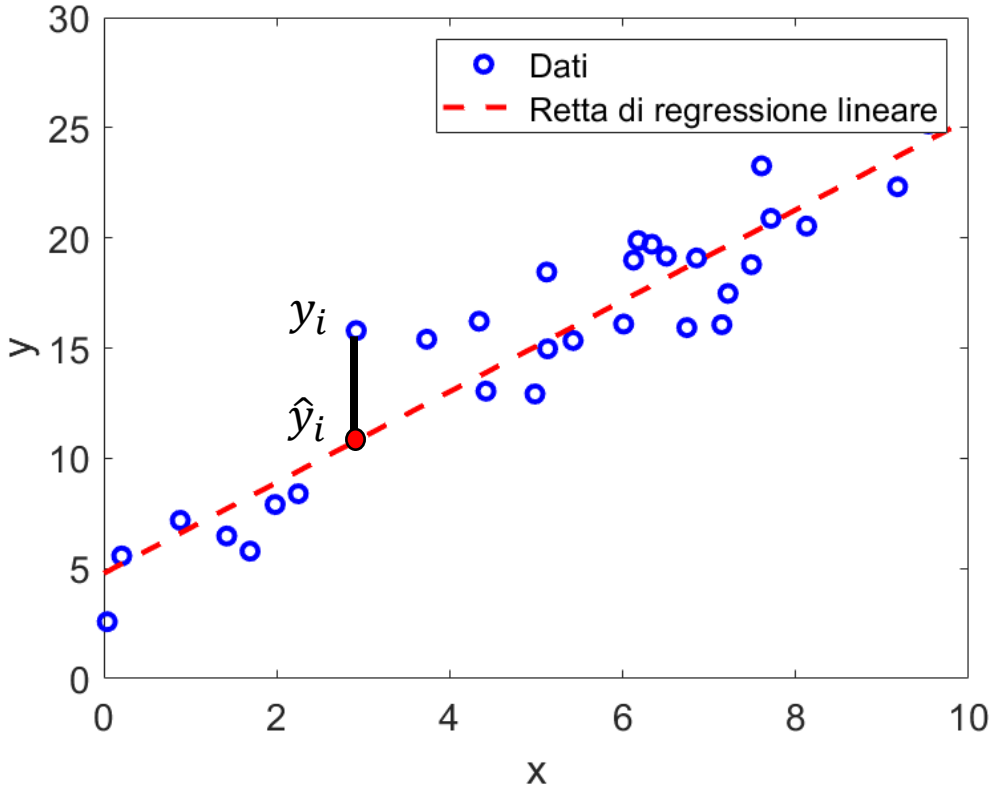
METODO DEI MINIMI QUADRATI LINEARI PER LA STIMA DEI COEFFICIENTI DI REGRESSIONE

Assumendo che la varianza dell'errore sia costante, ovvero $Var(\varepsilon_i) = \sigma^2 \; \forall i$, possiamo stimare i valori dei parametri β e β_0 mediante il **metodo dei minimi quadrati lineari**: si minimizza la somma dei quadrati degli scarti tra i valori di uscita reali, y_i, e quelli predetti dal modello, $\hat{y}_i$.

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

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

$$= \underset{\beta,\beta_0}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - (\beta \cdot x_i + \beta_0))^2$$



ESEMPIO: IL QUESITO

L'emoglobina glicata (HbA1c) rappresenta un importante parametro clinico utile alla diagnosi del diabete mellito. E' un indicatore legato alla glicemia media negli ultimi 3 mesi. Tipicamente se HbA1c>6.5% si sospetta la presenza di diabete.

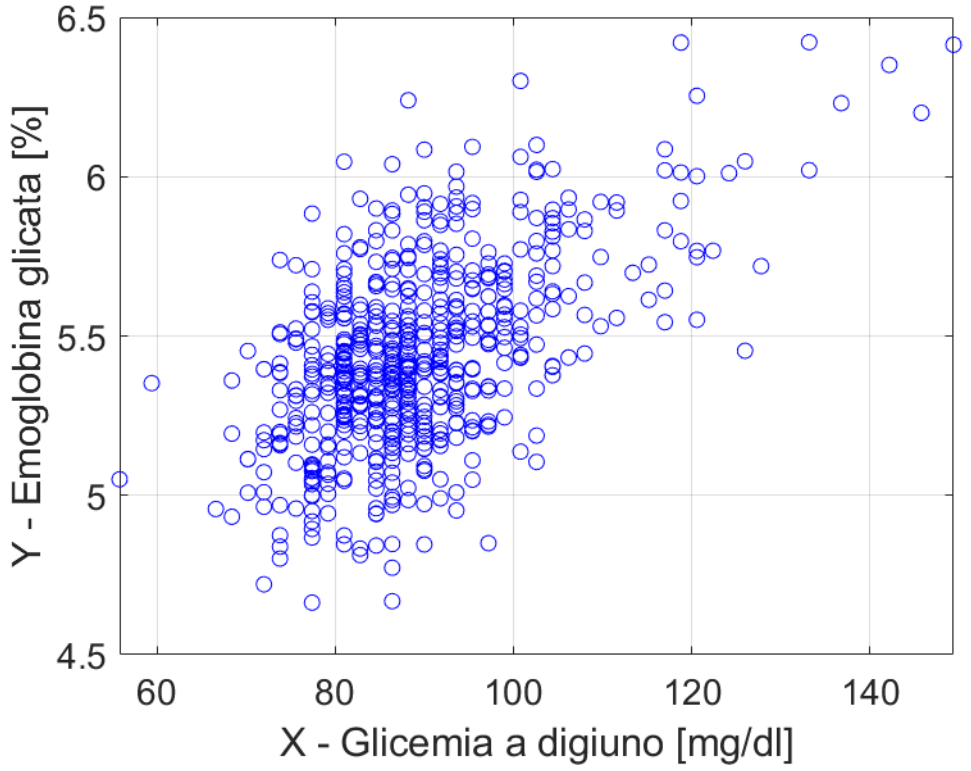
La misura di HbA1c nel sangue viene prescritta nei soggetti a rischio di diabete, ma non fa parte degli esami del sangue di routine.

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

Quesito: esiste una relazione lineare tra la glicemia a digiuno (variabile X) e l'emoglobina glicata (variabile Y)?

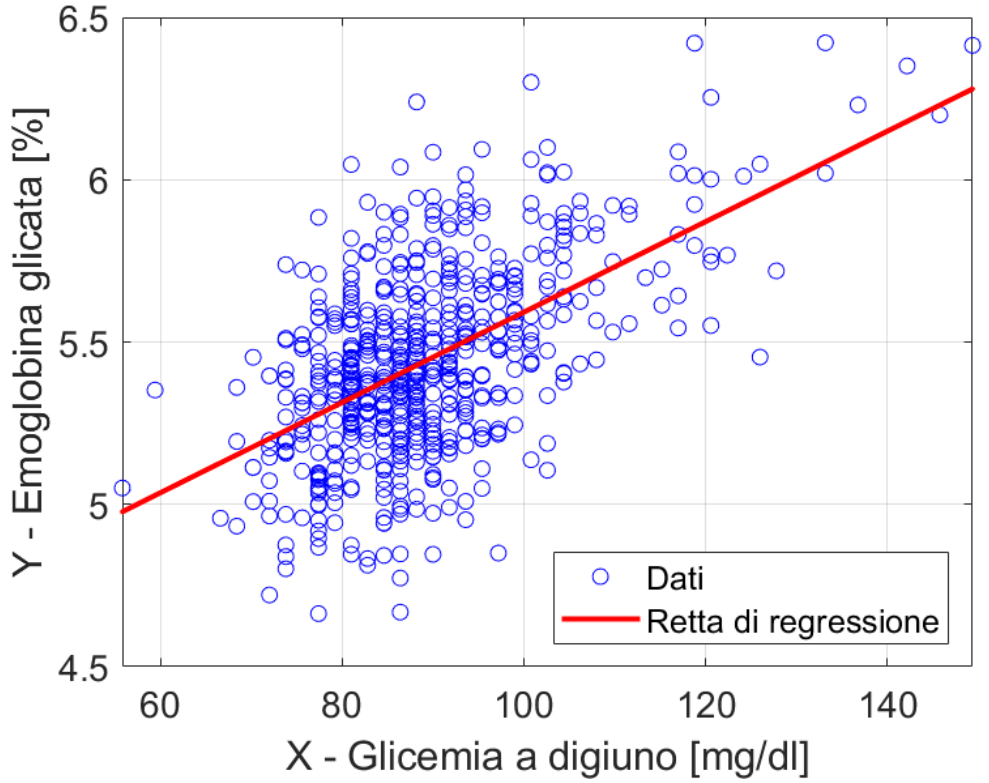
ESEMPIO: IL DATASET

- **Dataset**: 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).



ESEMPIO: IL MODELLO DI REGRESSIONE

- Per rispondere alla domanda, 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:

$$\hat{\beta} = 0.0139$$

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

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

Ci chiediamo se sussista una relazione lineare tra delle variabili esplicative X_j e una variabile di outcome Y. Ovvero, 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? Per rispondere a questa domanda:

- 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,…,n.
- Consideriamo il modello di regressione lineare multipla per descrivere i dati del campione.
- Con i dati a disposizione, stimiamo i parametri del modello di regressione.
  - Se il modello descrive bene i dati → possiamo concludere che sussiste una relazione tra Y e le variabili X_j e questa è approssimativamente lineare
  - Se il modello non descrive bene i dati allora:
    - o non c'è alcuna relazione rilevante tra Y e le variabili X_j
    - oppure la relazione sussiste ma non è approssimativamente lineare.

RACCOLTA DATI

- 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:

Y	**X_1**	**X_2**	**…**	**X_m**
y_1	x_{11}	x_{12}	…	x_{1m}
y_2	x_{21}	x_{22}	…	x_{2m}
…	…	…	…	…
y_n	x_{n1}	x_{n2}	…	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).

IL MODELLO DI REGRESSIONE LINEARE MULTIPLA APPLICATO AI DATI RACCOLTI

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

Forma matrice-vettore:

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

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

CARATTERIZZAZIONE STATISTICA DEL MODELLO

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

- Le quantità x_{ij} sono considerate deterministiche e note
- L'errore del modello è considerato casuale e distribuito come una normale a media nulla:

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

- I valori di uscita sono anch'essi rappresentati da variabili aleatorie normali:

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

- Le ε_i sono tra loro indipendenti $\rightarrow Cov(\varepsilon_i, \varepsilon_j) = 0, \;\; i \neq j$

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

Se assumiamo che la varianza dell'errore sia costante:

$$Var(\varepsilon_i) = \sigma^2 \; \forall i$$

i valori dei parametri β_j possono essere stimati a partire dai dati mediante il **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**

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

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

$$= \underset{\beta_0,\beta_1,\dots,\beta_m}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - (\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}))^2$$

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} (\boldsymbol{Y} - \boldsymbol{X} \cdot \boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X} \cdot \boldsymbol{\beta})$$

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

LO STIMATORE DEI PARAMETRI

- Se la matrice $\boldsymbol{X}^T\boldsymbol{X}$ è non singolare, la soluzione del problema di ottimizzazione precedente è:

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

- $\widehat{\boldsymbol{\beta}}$ rappresenta 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.

DIMOSTRAZIONE

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

Scalari di uguale valore

$$= \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$$

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

$$\frac{\partial SSE}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^T\boldsymbol{Y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = 0$$

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

- Nota: la stima $\widehat{\boldsymbol{\beta}}$ è definita solo se $\boldsymbol{X}^T\boldsymbol{X}$ è invertibile, ovvero ha rango pieno. Ciò si verifica solo se le colonne di X sono linearmente indipendenti, ovvero nessuna variabile è combinazione lineare delle altre.

PROPRIETA' DELLO STIMATORE AI MINIMI QUADRATI

Poiché Y è un vettore aleatorio di distribuzione normale, lo stimatore ai minimi quadrati di $\boldsymbol{\beta}$

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

è anch'esso un vettore aleatorio normale.

Si può dimostrare che:

- $E[\hat{\beta}_j] = \beta_j$ → lo stimatore è corretto (non distorto)
- $Cov(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$
  - La varianze degli stimatori $\hat{\beta}_j$ sono date dagli elementi sulla diagonale di $Cov(\widehat{\boldsymbol{\beta}})$
  - Per calcolarli abbiamo bisogno di **X** (nota) e σ^2, tipicamente incognita.
  - Possiamo però usare uno stimatore per stimare il valore di σ^2.
  - La deviazione standard dello stimatore $\hat{\beta}_j$ viene detto **standard error (SE)**.
  - Possiamo calcolare il coefficiente di variazione delle stime dei parametri come $CV_j = \frac{SE_j}{|\widehat{\beta}_j|} \cdot 100$

STIMA DELLA VARIANZA DELL'ERRORE

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

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

- 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: IL QUESITO

- Obesità, ipertensione e ipercolesterolemia sono condizioni spesso associate al diabete mellito.
- Quesito: 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.

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

$$\hat{\beta}_0 = 3.8828$$
$$\hat{\beta}_1 = 0.0115$$
$$\hat{\beta}_2 = 0.0145$$
$$\hat{\beta}_3 = 0.0007$$
$$\hat{\beta}_4 = -0.0029$$
$$\hat{\beta}_5 = 0.0029$$
$$\hat{\beta}_6 = -0.0032$$

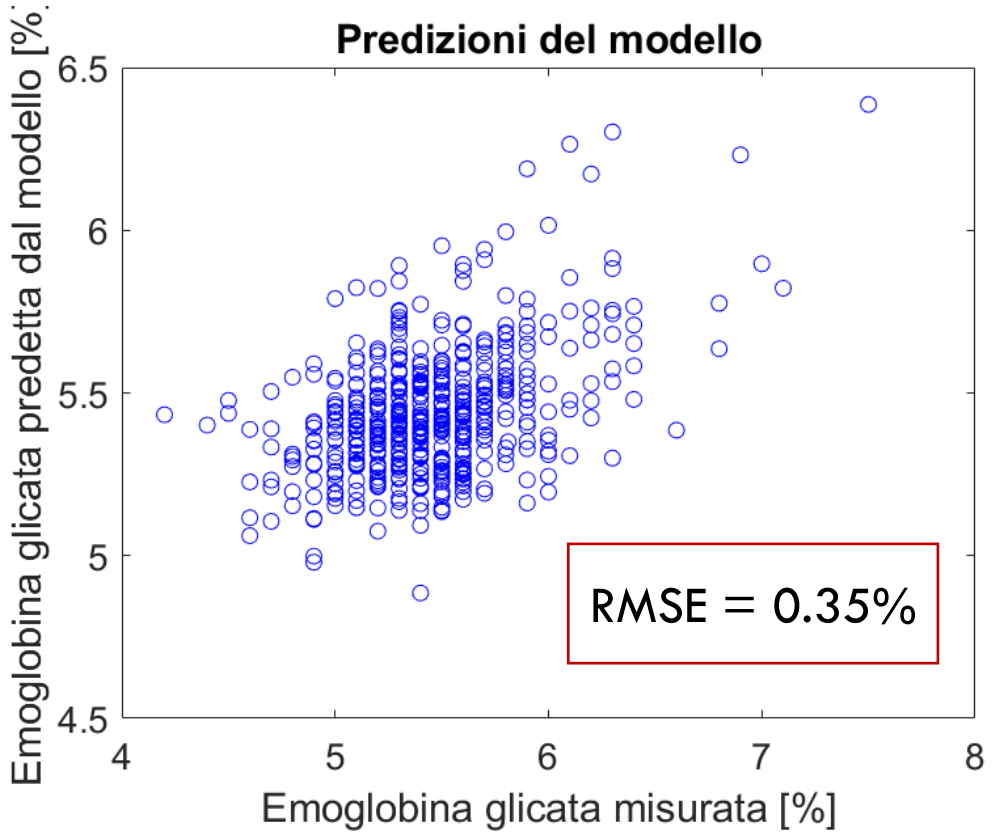
- $SSE = 76.68$
- $\hat{\sigma} = \frac{SSE}{n-m-1} = \frac{76.68}{600-7} = 0.1293$

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:
  - Confronto tra i valori dell'outcome reali e quelli predetti
  - Coefficiente di determinazione
  - F test
  - Analisi dei residui

CONFRONTO TRA I VALORI DELL'OUTCOME REALI E QUELLI PREDETTI

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



- Diverse metriche possono essere calcolate per valutare lo scostamento tra outcome predetta e outcome reale.
- Esempio: root mean square error (RMSE)

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

SCOMPOSIZIONE DELLA DEVIANZA DELL'OUTCOME

- **Total sum of squares (SST)**: devianza campionaria di y_i → 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:

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

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

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.
  - $R^2=1$ → la relazione tra le variabili di ingresso e di uscita è perfettamente lineare
  - $R^2=0$ → la variabile di uscita non è affatto spiegabile con una regressione lineare delle variabili di ingresso

F TEST

- Quando il valore di 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**.
- L'F test si basa sull'assunzione che i termini di errore ε_i abbiano distribuzione normale.
- Sistema di ipotesi:
  - H_0: $\beta_1 = \beta_2 = \cdots = \beta_m = 0$
  - H_1: almeno un coefficiente $\beta_i \neq 0,\ i \neq 0$
- Statistica del test:

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

- Quando vale H_0, F ha una distribuzione F di Fisher con gradi di libertà m e n-m-1.
  - Se $F > F_{\alpha,m,n-m-1}$ → rifiutiamo H_0
  - Se $F \leq F_{\alpha,m,n-m-1}$ → non possiamo rifiutare H_0

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.

ESEMPIO: F TEST

- Possiamo dire che il modello è in grado di spiegare una porzione significativamente diversa da 0 della variabilità di Y?
- Applichiamo l'F test per cercare di dare una risposta.
- Sistema di ipotesi:
  - H_0: $\beta_1 = \beta_2 = \cdots = \beta_6 = 0$
  - H_1: almeno un coefficiente $\beta_j \neq 0,\ j \neq 0$
- Calcoliamo la statistica F:

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

- Possiamo rifiutare l'ipotesi nulla con livello di significatività α=5%?
  - $F_{\alpha,m,n-m-1} = F_{0.05,6,600-7} = 2.11$
  - $p - value = 2.44 * 10^{-44}$

→ Rifiutiamo H0 → almeno uno dei coefficienti β_j è significativamente ≠0, e il modello predice una porzione significativa della variabilità di Y.

ANALISI DEI RESIDUI

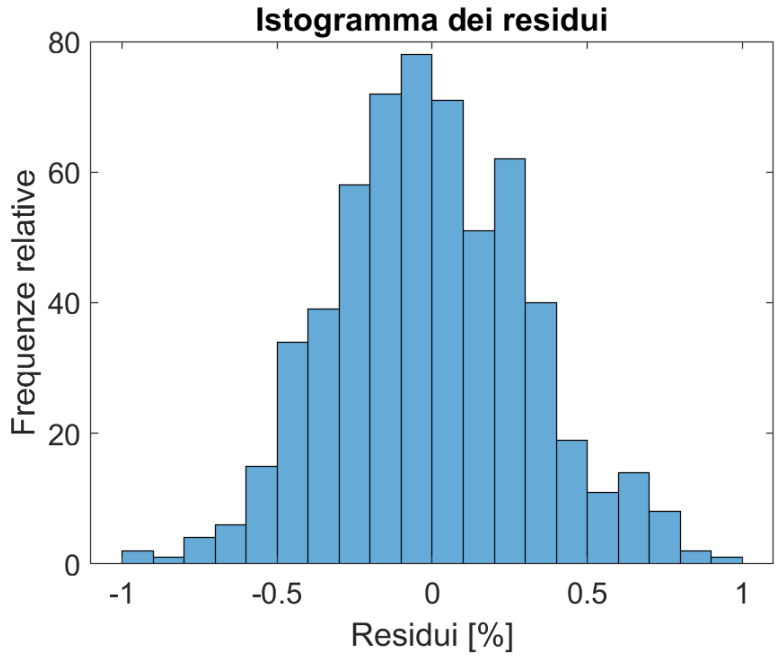
- **Residui:** differenza tra i valori osservati di Y e le predizioni modello

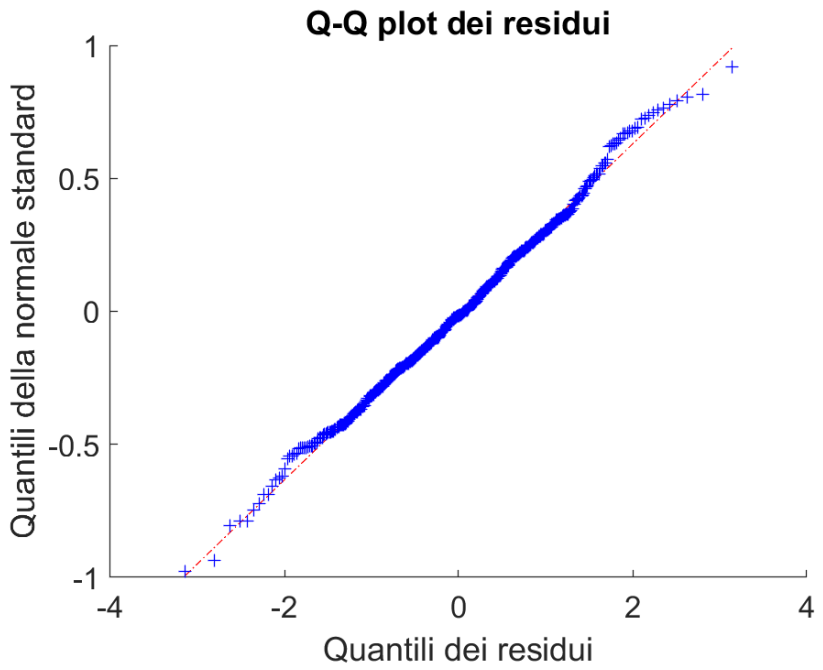
$$r_i = y_i - \hat{y}_i, \qquad i = 1, \dots, n$$

- 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.
  - I residui devono avere **distribuzione approssimativamente normale.**
  - I residui devono avere **media nulla.**
  - I residui devono essere **scorrelati.**
  - I residui devono avere **varianza omogenea.**

DISTRIBUZIONE NORMALE

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





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

Cosa concludiamo?

MEDIA NULLA

- Calcoliamo la media campionaria dei residui.

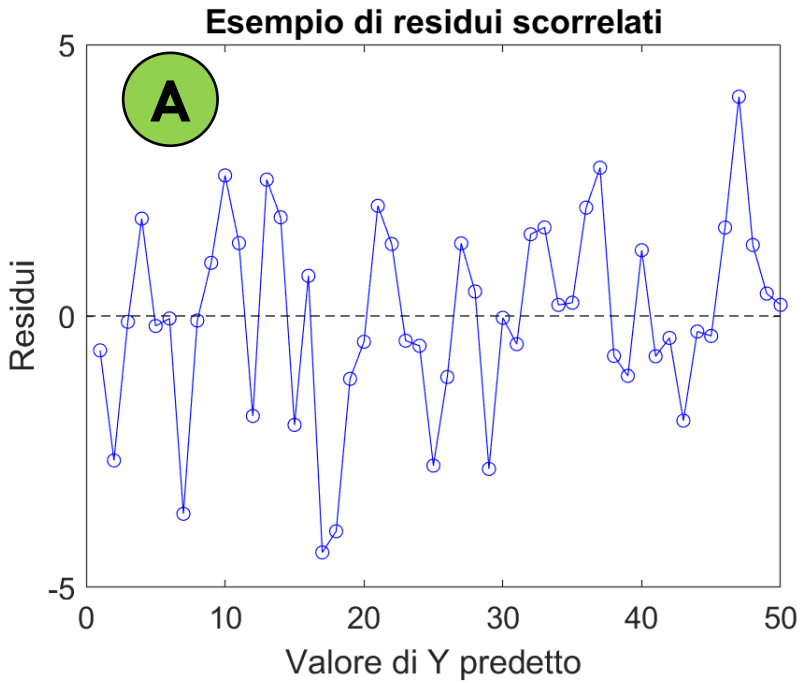
$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

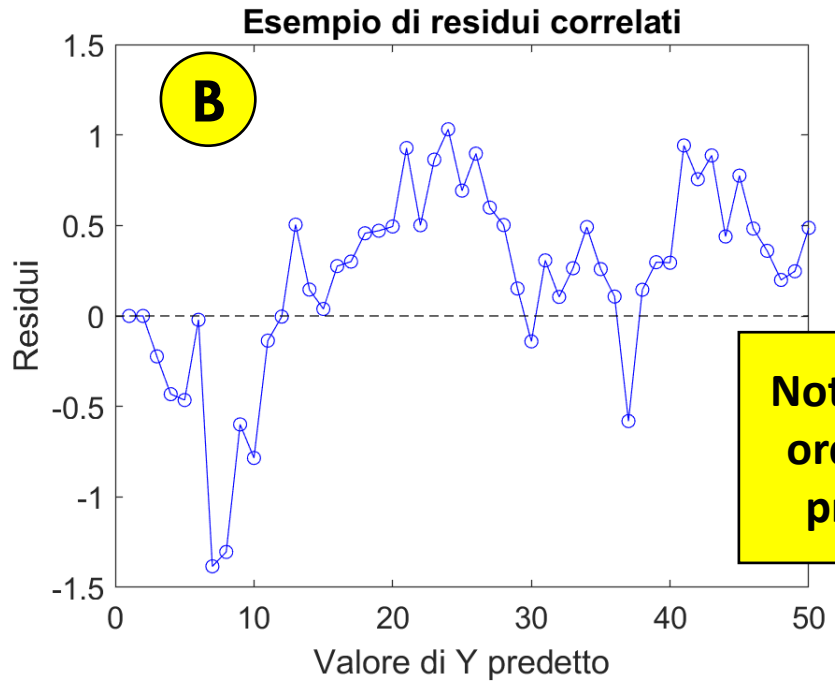
- Applichiamo un **t test** per verificare se la media dei residui è significativamente diversa da 0.
- Risultati per il nostro esempio:
  - $\bar{r} = -0.0059$
  - P-value del t test: 0.65

→ 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.

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.
  - Questa proprietà si chiama anche **bianchezza** dei residui.
- Ispezione visiva:





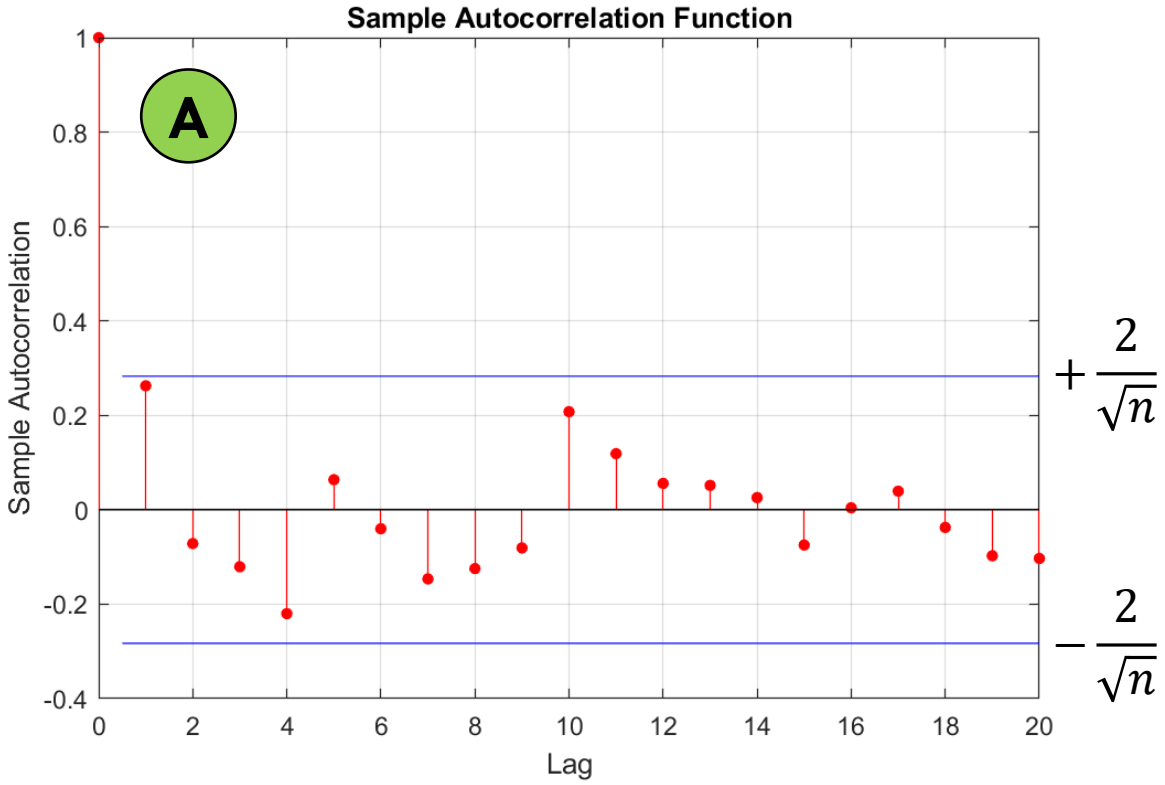
Nota: i residui vanno ordinati per il valor predetto di Y ($\hat{y}_i$)

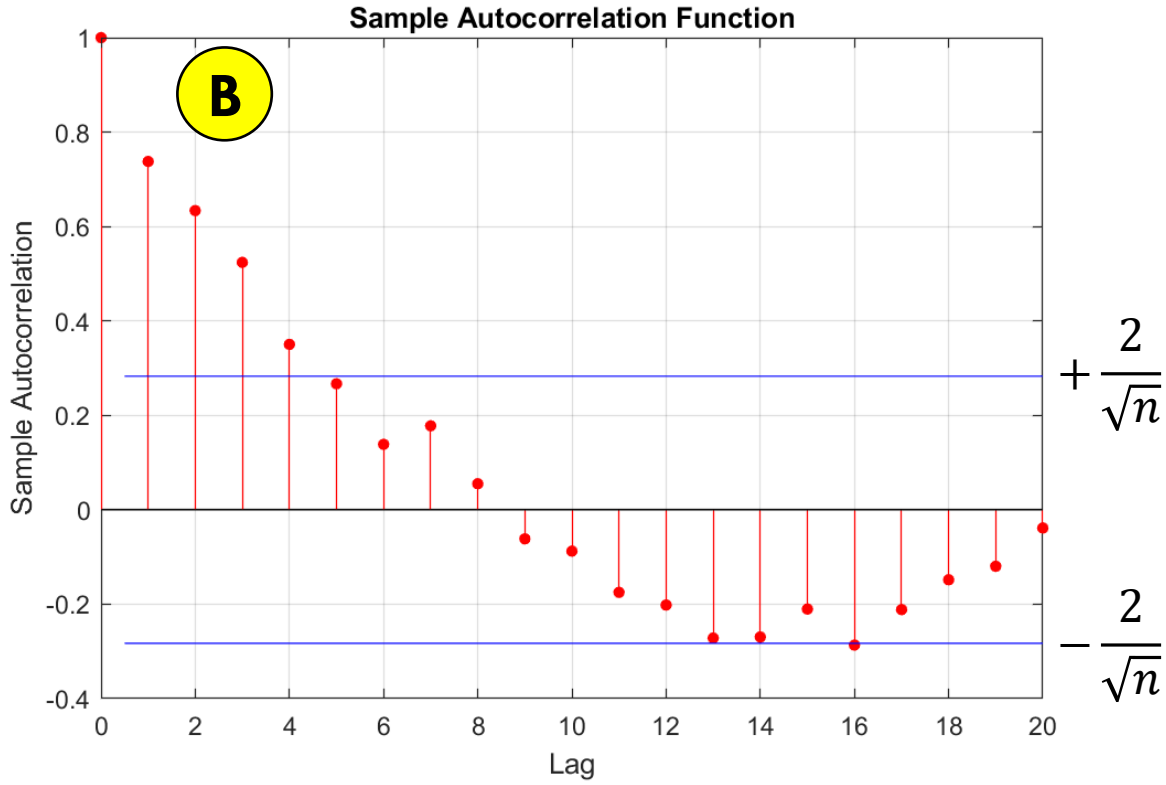
VALUTAZIONE QUANTITATIVA DELLA BIANCHEZZA DEI RESIDUI

- Possiamo valutare dal punto di vista quantitativo la bianchezza dei residui calcolando 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 $r_1, r_2, \ldots, r_n$ correlato con se stesso (sempre pari a 1)
  - Lag 1 → correlazione del segnale $r_2, \ldots, r_n$ con $r_1, r_2, \ldots, r_{n-1}$
  - Lag 2 → correlazione del segnale $r_3, r_2, \ldots, r_n$ con $r_1, r_2, \ldots, r_{n-2}$
  - …
  - Lag k → correlazione del segnale $r_k, r_{k+1}, \ldots, r_n$ con $r_1, r_2, \ldots, r_{n-k}$
- Se il segnale dei residui è scorrelato, ci aspettiamo che i valori della funzione di autocorrelazione stiano all'interno della banda di confidenza $\pm 2 \cdot 1/\sqrt{n}$

ESEMPI DI FUNZIONE DI AUTOCORRELAZIONE

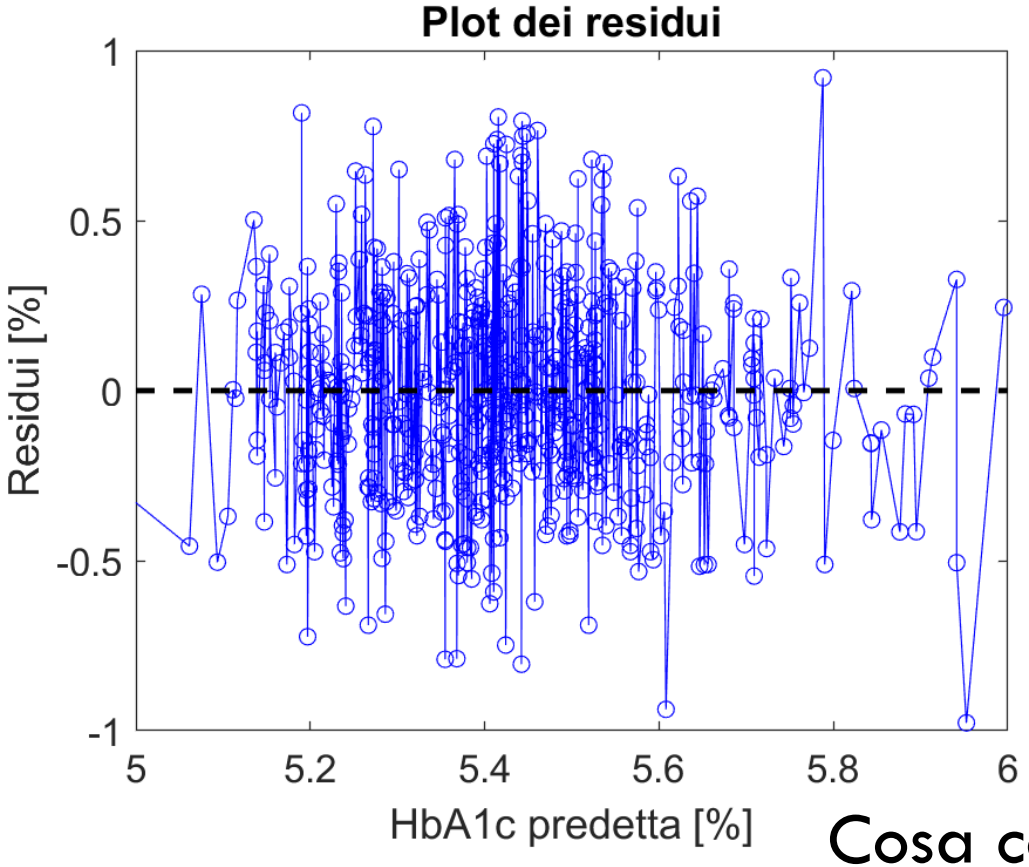
- Funzione di autocorrelazione stimata per le due serie di residui di slide 35.

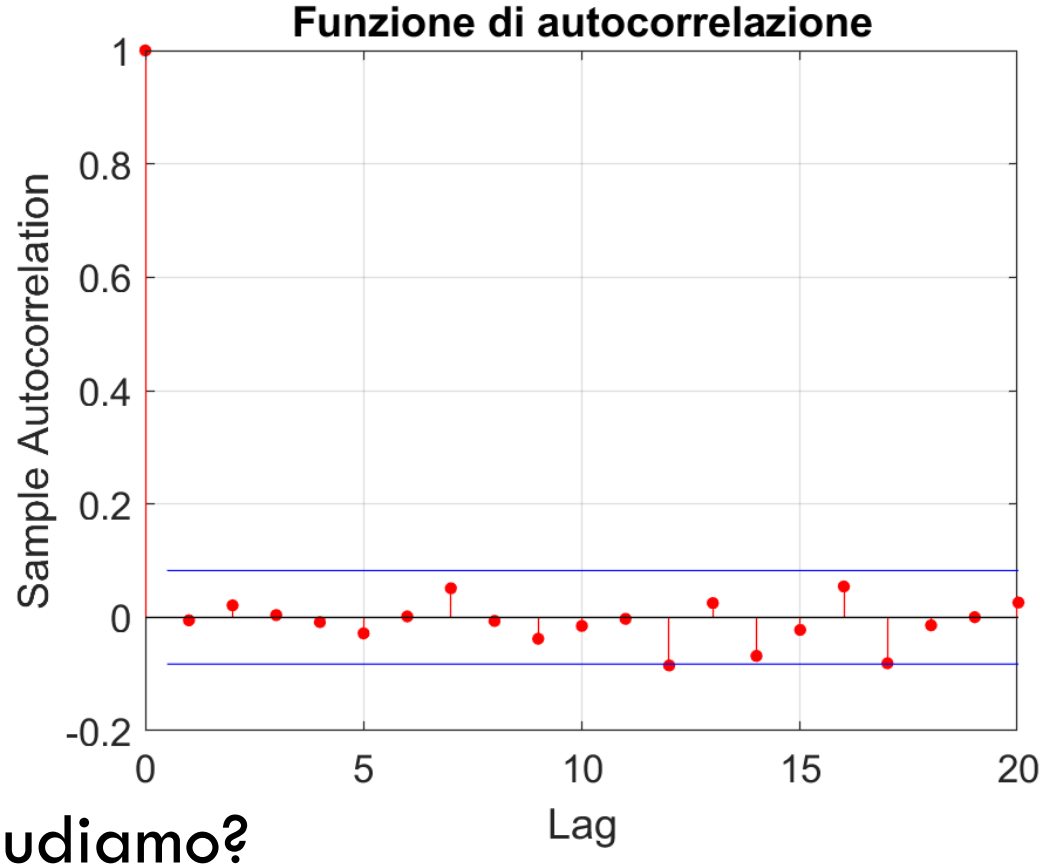




ESEMPIO: ANALISI DELLA BIANCHEZZA DEI RESIDUI

- Analizziamo la bianchezza dei residui del modello di regressione lineare multipla per la predizione dell'emoglobina glicata (slide 24).

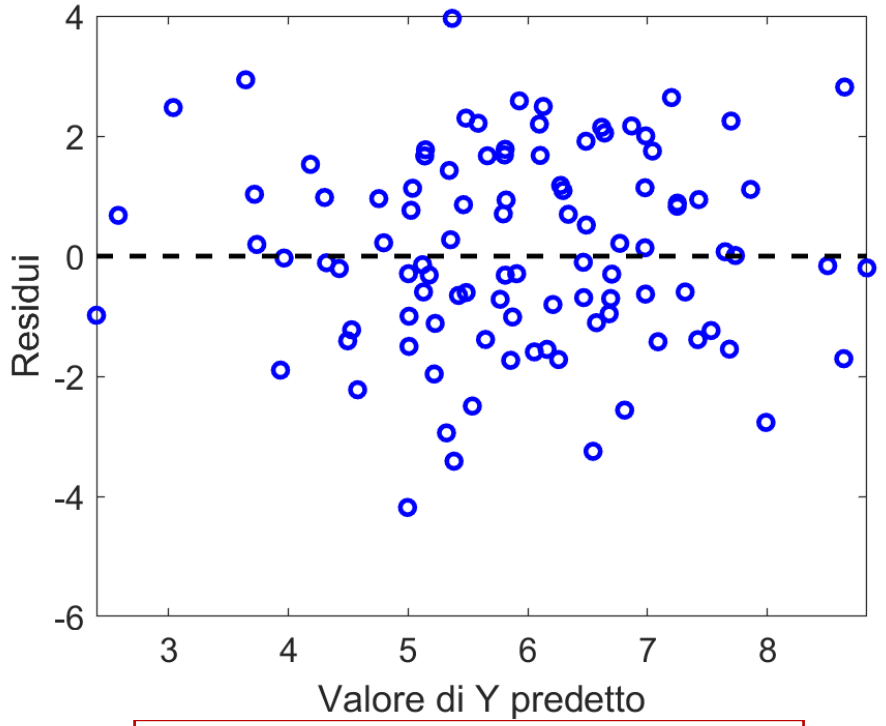




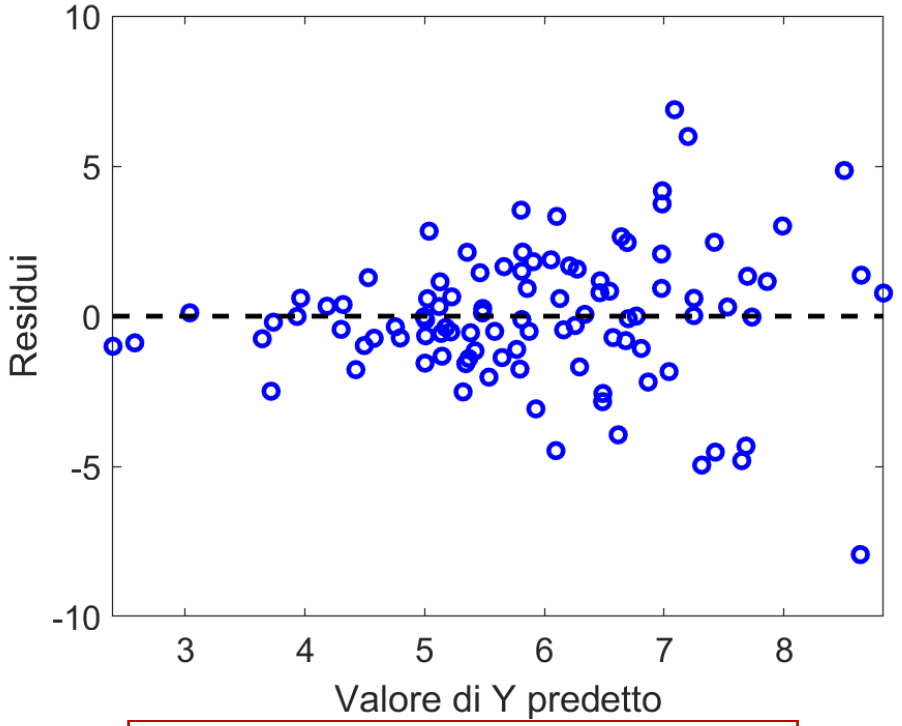
Cosa concludiamo?

VARIANZA OMOGENEA

- La varianza dei residui deve essere omogenea al variare del valore predetto di Y.
- Questo si può valutare 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



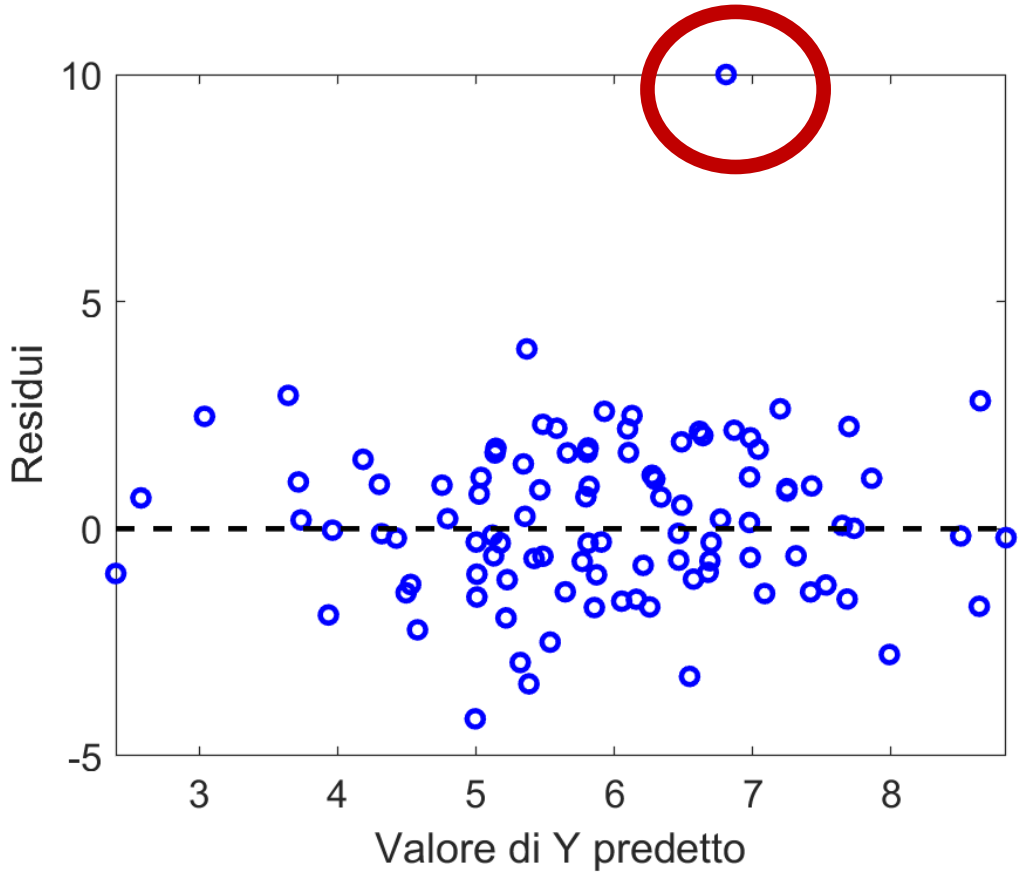
Varianza non omogenea

ALTRE ANOMALIE RISCONTRABILI DAL PLOT DEI RESIDUI

- 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

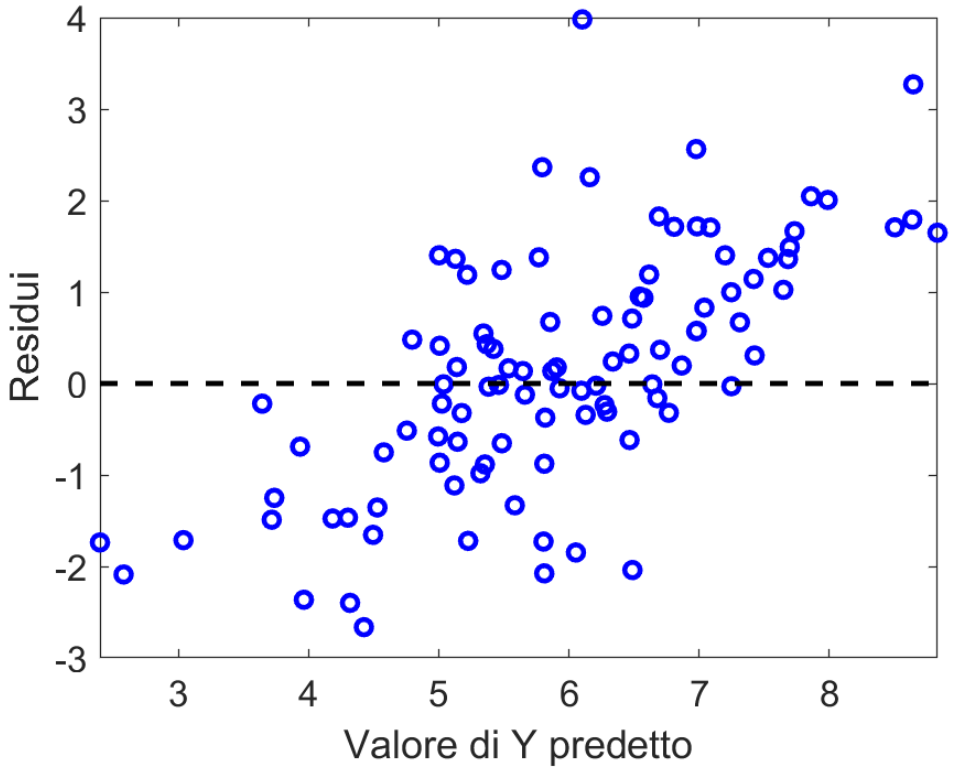
OUTLIER NEI RESIDUI

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



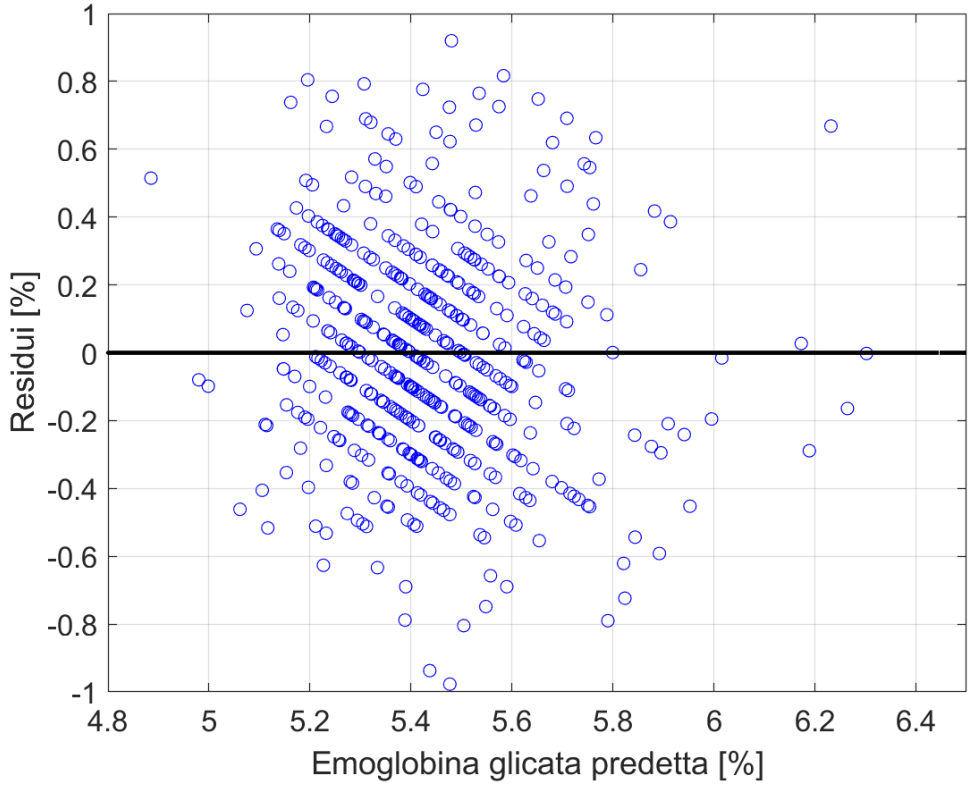
TREND NELL'ANDAMENTO DEI RESIDUI

- **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?

INTERPRETAZIONE DEI COEFFICIENTI DI REGRESSIONE

- 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.
- **L'intercetta** β_0 rappresenta il valore medio di Y quando le variabili X_j sono tutte nulle. Essa è la componente di Y **indipendente dai valori delle** X_j.
- Il **coefficiente** β_j quantifica **l'impatto di** X_j **su Y**. Esso è l'incremento medio di Y che si ottiene aumentando X_j di 1 unità e tenendo costanti tutte le altre variabili.
- **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
- **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

SIGNIFICATIVITA' STATISTICA DEI COEFFICIENTI DI REGRESSIONE LINEARE

- Come facciamo a determinare se il coefficiente β_j è significativamente diverso da 0?
- Innanzitutto valutiamo il valore di $\hat{\beta}_j$ e il suo intervallo di confidenza.
- Sappiamo che 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 $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.
- **Intervallo di confidenza 95%:** $\hat{\beta}_j \pm 2 \cdot \hat{\sigma}\sqrt{v_j}$ → il valore vero di β_j sarà compreso in questo intervallo con probabilità circa pari al 95%
- Valutiamo l'ampiezza dell'intervallo di confidenza e se questo comprende lo 0.

VERIFICA DI IPOTESI SUI COEFFICIENTI DI REGRESSIONE

- Possiamo anche applicare un **test statistico** per verificare l'ipotesi che il coefficiente β_j sia significativamente diverso da 0.
- Il test si basa sull'assunzione che i termini di errore ε_i abbiano 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 → $z_j = \dfrac{\hat{\beta}_j}{\sqrt{Var(\hat{\beta}_j)}} = \dfrac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$ ← v_j è l'elemento in posizione j della diagonale di $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$

- Quando vale l'ipotesi nulla, z_j ha distribuzione t di Student con n-m-1 gradi di libertà.
  - Se $|z_j| > t_{\frac{\alpha}{2},n-m-1}$ → rifiutiamo H_0
  - Se $|z_j| \leq t_{\frac{\alpha}{2},n-m-1}$ → 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	$1.99*10^{-32}$
IMC	0.0145	[0.0075 0.0214]	4.16	$3.6*10^{-5}$
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?

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