

METODI STATISTICI PER LA BIOINGEGNERIA

Laboratorio 6

A.A. 2025-2026

Enrico Longato



Dal lab 5: Sanity check sul p value

Sanity check = controllo di sensatezza dei risultati

t-test appaiato a due campioni (i dati sono due istanti di tempo dello stesso individuo)

- H_0 era che la media delle differenze fosse nulla
- Alfa era stato lasciato di default a 0.05.
- Il test era a due code.
- Il p-value è risultato $1.7667e-39$

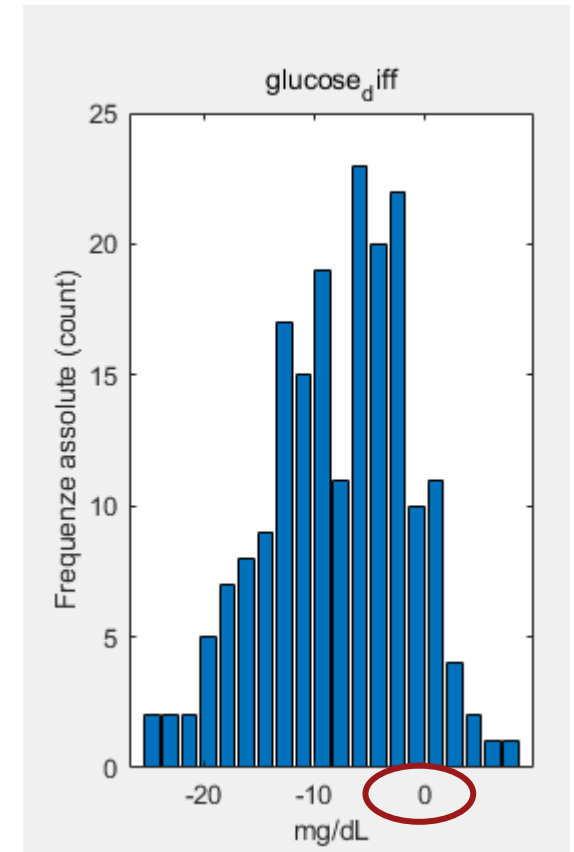
=> Rifiuto H_0 : con livello di significatività 0.05, posso affermare che le due medie sono diverse

Domanda: ha senso che il p value sia molto piccolo?

Risposta: Sì, perché la media della differenza è "distantissima" da 0!

Domanda bonus: cambia qualcosa fare la differenza "in un senso o nell'altro"? Ovvero "basal - ss" oppure "ss - basal"?

Risposta: No, se il test è a due code; se il test è a una coda basta stare attenti alla direzione.





Laboratorio 6: Contenuti e obiettivi

1. Esercitazione "alla lavagna"
 - Regressione lineare univariata
2. Esercizi da svolgere in autonomia (per superare la "paura del file bianco")
 - Regressione lineare multivariata (2 variabili)
 - Regressione lineare multivariata ("tante" variabili)
 - Esercizio motivazionale: la regressione lineare è lineare nei parametri β e non nei dati.
3. Ripasso di teoria (**parte integrante del programma d'esame di teoria!**)
 - Raccordo tra teoria e pratica.



Raccordo fondamentale tra teoria e pratica: la costruzione della forma matriciale del problema di regressione lineare a partire dai dati.

Attenzione a **non** copiare la **variabile dipendente** anche nella matrice X !

Vettore costante di 1 che ci permette (se necessario, altrimenti la colonna non serve) di stimare il β_0 dell'intercetta

La variabile dipendente è un vettore **colonna** (colonna!)

Variabile dipendente

$Y_{n \times 1}$

=

Variabile indipendente 1

Variabile indipendente 2

...

Variabile indipendente m

$X_{n \times (m+1)}$

Parametri da stimare

$\beta_{(m+1) \times 1}$

I parametri sono da stimare, quindi non vanno inseriti a priori, bensì ricavati dalla formula $\hat{\beta} = (X^T X)^{-1} X^T Y$

+ ε
Rumore

Del rumore non ci interessa "direttamente", ma solo per come **modifica le formule** (tipicamente, quella per stimare $\hat{\beta}$ e il suo standard error)



Prima di svolgere l'esercitazione (oppure al bisogno), utilizzare il comando **help** di MATLAB seguito dal nome delle seguenti function, utili allo svolgimento degli esercizi.

Esercizi 1-4

- `corrplot`, `stem`

Inoltre, ci serviranno le formule che trovate nelle **slide di teoria sulla regressione lineare**.

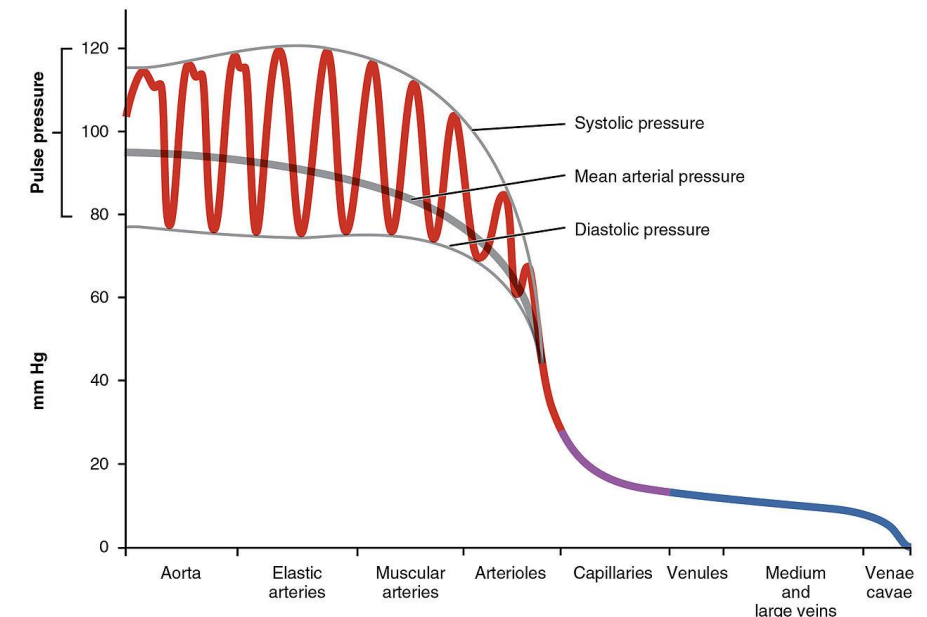
- Per l'esame, vanno imparate.

CONTESTO DELL'ESERCITAZIONE E DATI

Di nuovo il dataset ELSA (**ELSA.mat**).

Vedi workspace dopo il caricamento per il contenuto.

- La *pulse pressure* sarà la variabile dipendente per l'esercitazione.





ESERCIZIO 1 - PARTE 1: CARICAMENTO E PULIZIA DATI (svolto)

- Caricare il file **ELSA.mat**
 - **elsa** = matrice di dimensione 1000x16 ("mille soggetti per sedici variabili")
 - **elsa_labels** = cell array di etichette con i nomi delle 16 variabili
 - **elsa_units** = cell array con le unità di misura corrispondenti
- Individuare i dati
 - Mancanti (**NaN**)
 - Negativi
 - Infiniti
- Sostituire ai dati negativi e infiniti il valore **NaN** in modo da omogeneizzarli a quelli mancanti.
- Salvare nella matrice **elsa_reduced** le sole righe che non presentano dati mancanti.



ESERCIZIO 1 - PARTE 2: VARIABILI DI INTERESSE E IMPOSTAZIONE DEL PROBLEMA (svolto)

- Utilizzare l'istruzione **corrplot** per mostrare l'istogramma delle variabili pulse pressure e systolic blood pressure assieme ai loro scatter plot e coefficienti di correlazione (basta una riga).
 - Dire se sembrano correlate fra di loro.
- Con l'obiettivo di effettuare la regressione lineare
$$\text{pulse pressure} = \beta_{sbp} \times \text{systolic blood pressure} + \beta_0 + \varepsilon$$
 - Inserire nel vettore Y i dati corrispondenti alla variabile dipendente (pulse pressure)
 - Inserire nella matrice X i dati corrispondenti all'unica variabile indipendente (systolic blood pressure)
 - Si consideri di dover stimare anche l'intercetta β_0

ESERCIZIO 1 - PARTE 3: VERIFICARE LA GAUSSIANITA' DELLE VARIABILI IN GIOCO (proposto)

- Verificare la gaussianità delle variabili pulse pressure e systolic blood pressure.
 - NB: non confondetevi! È un esercizio per non perdere dimestichezza con i test statistici, ma "non c'entra nulla" con le assunzioni circostanti la regressione lineare.
 - Al più, quelle riguardano la gaussianità dei residui, non delle variabili!



ESERCIZIO 1 - PARTE 4: REGRESSIONE LINEARE parte 1 (svolto in parte)

- Stimare, utilizzando la formula chiusa, il valore dei parametri (intercetta inclusa).
- Calcolare la predizione del modello a partire dai parametri stimati e dai dati in ingresso (matrice X).
- Calcolare il vettore dei residui.
- Calcolare l'errore quadratico medio (RMSE).
- Rappresentare in una figura con tre pannelli (parte proposta)
 - Il valore vero della variabile pulse pressure contro il valore predetto (**scatter**)
 - NB: questo grafico è disegnabile sempre ed è sempre 2D.
 - Il residui di ciascun campione (**stem**). Suggerimento: visualizzare sull'asse y il valore dei residui (ovvio), sull'asse x l'indice progressivo del soggetto, cioè una sequenza da 1 a N campioni (meno ovvio).
 - Lo scatter plot di pulse pressure contro systolic blood pressure e, sovrapposta ad esso, la retta di regressione predetta corrispondente ai valori di systolic blood pressure da 70 a 220 con passo di campionamento 0.5 mmHg.
 - NB: questo grafico è disegnabile fino a un massimo di due variabili indipendenti (che, con quella dipendente, fanno le 3 variabili di un plot 3D).



ESERCIZIO 1 - PARTE 5: REGRESSIONE LINEARE parte 2 (svolto)

- Calcolare il coefficiente di determinazione R^2
 - Suggerimento: formula con ricordata la definizione di SST

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

- Calcolare lo standard error dei parametri.
 - Suggerimento: la matrice di varianza-covarianza dei parametri si calcola così $Cov(\hat{\beta}) = \sigma^2(X^T X)^{-1}$

e la varianza a posteriori si stima così
$$\hat{\sigma}^2 = \frac{SSE}{n - m - 1} = \frac{1}{n - m - 1} \sum_{i=1}^n (Y_i - \hat{y}_i)^2$$

- Calcolare il coefficiente di variazione (CV) percentuale dei parametri.
 - Suggerimento: formula $CV_j = \frac{SE_j}{|\hat{\beta}_j|} \cdot 100$

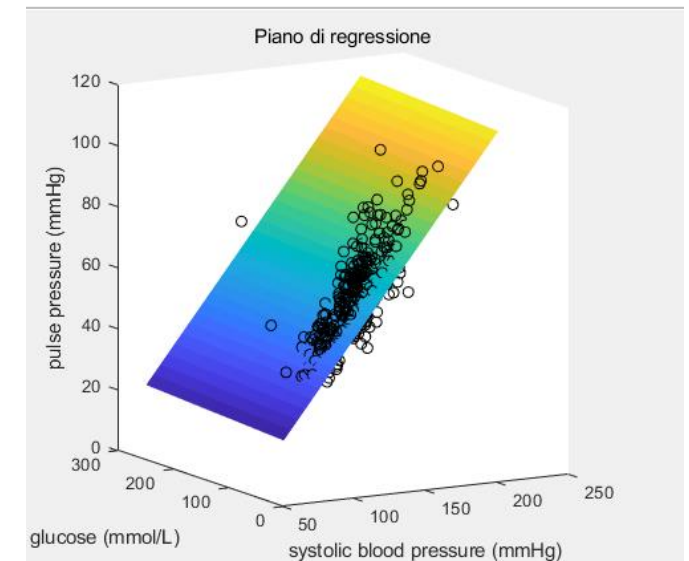


ESERCIZIO 2 – REGRESSIONE LINEARE CON DUE VARIABILI INDIPENDENTI (proposto)

- Ripetere l'esercizio 1 nella sua interezza con, soltanto, le seguenti modifiche
 - Le variabili indipendenti ora sono systolic blood pressure e glucose, quindi il modello di regressione lineare diventa
$$\text{pulse pressure} = \beta_{sbp} \times \text{systolic blood pressure} + \beta_{gluc} \times \text{glucose} + \beta_0 + \varepsilon$$
 - Il terzo pannello delle rappresentazioni grafiche della parte 4 può essere omesso
 - Nelle soluzioni troverete come, eventualmente, si sarebbe potuto rappresentare con un piano di regressione.

ESERCIZIO 3 – REGRESSIONE LINEARE MULTIVARIATA (proposto)

- Ripetere l'esercizio 1 nella sua interezza con, soltanto, le seguenti modifiche
 - Le variabili indipendenti ora sono tutte quelle del database elsa **tranne** systolic blood pressure e diastolic blood pressure.
 - Scrivere su carta l'equazione del modello di regressione così formato.
 - Il terzo pannello delle rappresentazioni grafiche della parte 4 deve essere omesso in quanto non rappresentabile.





ESERCIZIO 4 – LINEARITA' NEI PARAMETRI (proposto)

- Ripetere l'esercizio 2 nella sua interezza con, soltanto, le seguenti modifiche
 - Le variabili indipendenti ora sono il quadrato di systolic blood pressure e il cubo di glucose, quindi il modello di regressione lineare diventa
$$pulse\ pressure = \beta_{sbp^2} \times (systolic\ blood\ pressure)^2 + \beta_{gluc^3} \times (glucose)^3 + \beta_0 + \varepsilon$$
 - Il terzo pannello delle rappresentazioni grafiche della parte 4 può essere omissso

NB: questo è un esercizio motivazionale che dovrebbe farvi riflettere sul fatto che la regressione lineare si chiama lineare perché è lineare il sistema di equazioni in β .

- In altre parole, il punto è che si può usare una regressione lineare perché ad essere lineare è la relazione tra pulse pressure e le versioni trasformate della sistolica e del glucosio e che non importa che queste versioni trasformate siano il risultato di trasformazioni non lineari (quadrato e cubo).

NB2: anche se questo genere di trasformazione vi sembra strana e immotivata (perché, in un certo senso, lo è), "si può usare davvero" in un ambito adiacente a quello che affrontiamo nel corso, ovvero il machine learning (<https://scikit-learn.org/dev/modules/generated/sklearn.preprocessing.PolynomialFeatures.html>).

NB3: esiste, invece, un concetto analogo con indubbia "dignità statistica", ovvero quello dei termini di interazione.