UNIVERSITÀ
DEGLI STUDI
DI PADOVA

METODI STATISTICI PER LA BIOINGEGNERIA

A.A. 2024-2025

Prof. Alessandra Bertoldo
Ing. Mattia De Francisci, Ing. Claudia Tarricone

ESERCIZIO 1

I dati a disposizione per il laboratorio sono stati raccolti per uno studio sull'insorgenza di stress in 20 soggetti affetti da ipertensione

L'obiettivo è investigare se esiste una relazione tra alcune variabili *pressione sanguigna, età, peso, estensione della superficie corporea, anni di ipertensione, battiti cardiaci, durata del sonno notte precedente* e il *livello di stress*

I dati sono contenuti nel file ***data_lab7.mat***, *sono presenti:*

- ***X:*** Matrice 2D di dimensioni 20x7 (20 soggetti, 7 variabili)
- ***Y:*** vettore 20x1 contenente le realizzazioni dell'indice di stress per i 20 soggetti
- ***labels:*** Array di celle di dimensioni 1x7 contenente i nomi delle variabili
  - ***BP:*** pressione sanguigna (mmHg)
  - ***Age:*** età (anni)
  - ***Weight:*** peso (kg)
  - ***BSA:*** body surface area (m^2)
  - ***Dur:*** durata dell'ipertensione (anni)
  - ***Pulse:*** battiti cardiaci (battiti al min)
  - ***Sleep hours:*** durata del sonno (ore)

- Caricare i dati (**`data_lab7.mat`**).

  NOTA: non viene chiesta l'analisi per la ricerca di valori non utilizzabili (tipo NaN) perchè sono già stati testati e normalizzati tramite **`zscore`**. *I dati sono quindi pronti all'uso.*

- Si richiede la stima di Y utilizzando tutte le variabili presenti in X. Calcolare:

  - residui (**`res`**) e somma dei residui al quadrato (SSE) (**`rss`**)
  - stima a posteriori dell'errore di misura (**`sigma_hat_2`**)
  - standard error (**`SE`**)
  - coefficiente di variazione (**`CV`**)
  - coefficiente di determinazione (**`R2`**)
  - R2 aggiustato (**`R2_adjusted`**)

  Mostrare **`scatterplot`** (dati vs predizione), plot dei residui e qq-plot per valutare il fit

$$R^2_{adj} = \bar{R}^2 = \left(R^2 - \frac{m}{n-1}\right)\frac{n-1}{n-m-1}$$

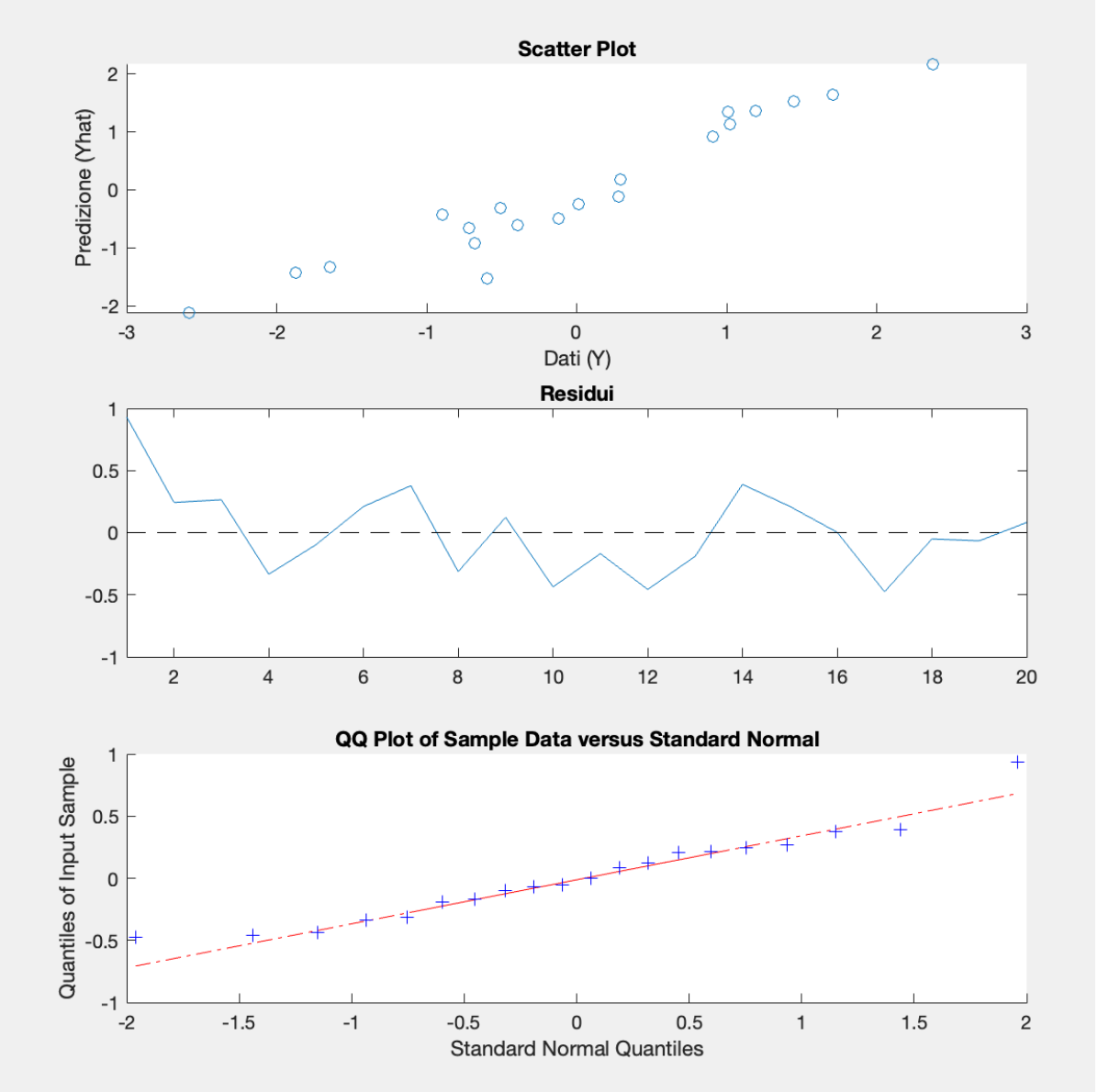
m=numero di colonne di X (esclusa intercetta) e n=numero di dati

- Commentare lo scatter-plot: com'è il fit?

- Commentare i residui:

  - Sono bianchi?
  - Ricalcano una distribuzione normale?

- Controllare i CV dei beta: sono accettabili?



- Valutare il Fattore di Inflazione della Varianza (**VIF**) per ogni variabile di X

$$VIF_j = (X^T X)^{-1}_{jj} = \frac{1}{1 - R_j^2}$$

- Per il suo calcolo, procedere con la formula $1/(1 - R_j^2)$, calcolando R_j^2 come segue:

$$\widehat{\beta_j} = \left(X(:, tutte\ tranne\ j)^T X(:, tutte\ tranne\ j)\right)^{-1} X(:, tutte\ tranne\ j)^T X(:, j)$$

Che equivale a stimare i parametri adottando $X(:, j)$ come variabile dipendente (cioè 'Y') e usando come predittori tutte le colonne (cioè le variabili) contenute in X eccetto la colonna j-esima.

$$R_j^2 = \left[corr(X(:, j), X(:, tutte\ tranne\ j) \cdot \widehat{\beta_j}\right]^2$$

- Selezionare le variabili da eliminare in base alla condizione VIF>10 e salvare le restanti in

`X_new`

- Si ripeta la stima di **Y** utilizzando le variabili in **X_new**. Calcolare gli stessi indici richiesti per il modello con **X**
- Stimare Y (**Y_hat_mod**) mediante **X_mod**, contenente solamente le **ultime due** colonne di (**X_mod=X_new(:,end-1:end)**)
- Confrontare il modello ottenuto da **X_new** con quello ottenuto da **X_mod** tramite R2 aggiustato e calcolando AIC e BIC per entrambi. Quale modello viene selezionato?

$$AIC = n \log\frac{SSE}{n} + 2m \qquad BIC = n \log\frac{SSE}{n} + m \log n$$

- Mostrare nella stessa figura **scatterplot** (dati vs predizione), plot dei residui e qq-plot per valutare il fit dei due modelli

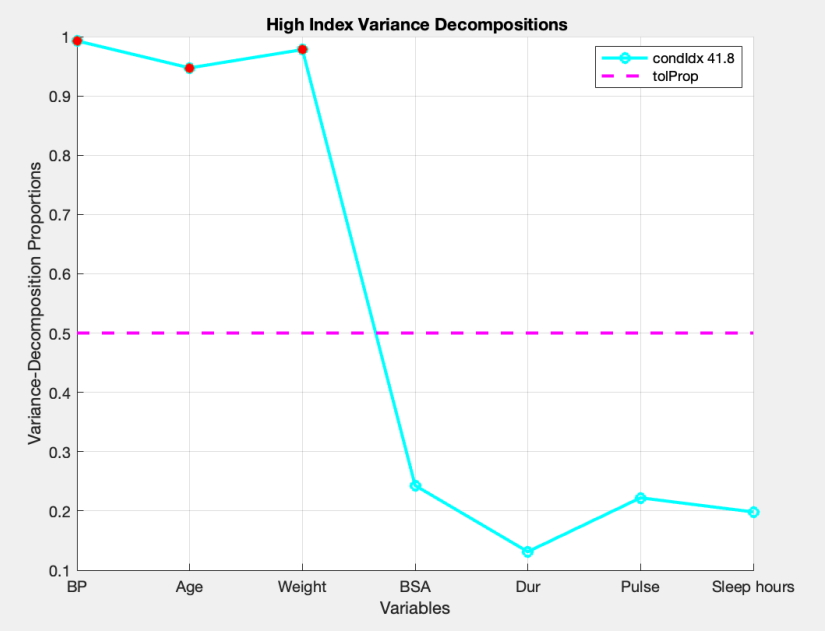
OPZIONALE

- Analizzare la presenza di multicollinearità tra variabili tramite **CN** (Numero di condizionamento) e stampare sulla command window il valore di CN

$$CN = \frac{\lambda_{max}}{\lambda_{min}}$$, con λ_i i=1…N, autovalori della matrice di covarianza $X^T X$

- Il valore di CN suggerisce la presenza di multicollinearità tra le variabili? (N.B. CN > 1000?)

- Esaminare la multicollinearità tra le variabili tramite l'Indice di Condizionamento (**CI**) mediante la funzione ***collintest*** (vedere l'help) e mostrare i risultati di CI in una figura (**`TolIdx=10`**). Sono state selezionate le stesse variabili di VIF?



BONUS: capire come viene calcolata la Variance Decomposition Proportions

(si vedano le prime 4 pagine il PDF allegato e il codice di collintest) N.B. non è materia d'esame