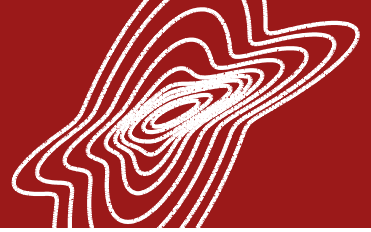


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

METODI STATISTICI PER LA BIOINGEGNERIA

Laboratorio 6

A.A. 2024-2025

Enrico Longato

Grazie alla Prof.ssa Alessandra Bertoldo e al Dott. Marco Castellaro per la versione iniziale delle slide.

Dal lab 5: Leggere l'help di MATLAB

```
>> help lillietest
 lillietest - Lilliefors test
    This MATLAB function returns a test decision for the null hypothesis
    that the data in vector x comes from a distribution in the normal
    family, against the alternative that it does not come from such a
    distribution, using a Lilliefors test.

    Syntax
      h = lillietest(x)
      h = lillietest(x,Name,Value)
      [h,p] = lillietest(___)
      [h,p,kstat,critval] = lillietest(___)

    Input Arguments
      x - Sample data
        vector

    Name-Value Arguments
      Alpha - Significance level
        0.05 (default) | scalar value in the range (0,1)
      Distribution - Distribution family
        'normal' (default) | 'exponential' | 'extreme value'
      MCTol - Maximum Monte Carlo standard error
        scalar value in the range (0,1)

    Output Arguments
      h - Hypothesis test result
        1 | 0
      p - p-value
        scalar value in the range (0,1)
      kstat - Test statistic
        nonnegative scalar value
      critval - Critical value
        nonnegative scalar value
```

```
>> help corrplot
 corrplot - Plot variable correlations
    This MATLAB function plots Pearson's correlation coefficients between
    all pairs of variables in the input matrix of time series data.

    Syntax
      [R,PValue] = corrplot(X)
      [R,PValue] = corrplot(Tbl)
      [___] = corrplot(___,Name=Value)

      corrplot(___)
      corrplot(ax,___)
      [___,H] = corrplot(___)

    Input Arguments
      X - Time series data
        numeric matrix
      Tbl - Time series data
        table | timetable
      ax - Axes on which to plot
        Axes object

    Name-Value Arguments
      Type - Correlation coefficient
        "Pearson" (default) | "Kendall" | "Spearman" | character vector
      Rows - Option for handling rows in input time series data that contain NaN values
        "all" | "complete" | "pairwise" (default) | character vector
      Tail - Alternative hypothesis
        "both" (default) | "right" | "left" | character vector
      VarNames - Unique variable names to use in plots
        string vector | character vector | cell vector of strings |
        cell vector of character vectors
      TestR - Flag for testing whether correlations are significant
        "off" (default) | "on" | character vector
      Alpha - Significance level
        0.05 (default) | scalar in [0,1]
      DataVariables - Variables in Tbl
        all variables (default) | string vector |
        cell vector of character vectors | vector of integers |
        logical vector
```

Dal lab 5: Indicazioni per l'esame

Elementi tipici che potrebbero essere in uno o più esercizi d'esame

1. Caricamento, selezione e pulizia dati.
2. Rappresentazione grafica e statistiche descrittive.
3. Verifica di ipotesi preliminari (gaussianità, ecc.).
4. Scelta del test statistico corretto.
5. Applicazione del test corretto e commento (preciso) del suo esito.

NB: questi elementi potrebbero ricorrere (e lo vedremo nei prossimi laboratori) anche in esercizi che, a un osservatore distratto, potrebbero apparire come "non sui test statistici" (per esempio, all'interno di un esercizio sulla regressione).

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 indipendente è un vettore **colonna** (colonna!)

$$\begin{pmatrix} \text{Variabile dipendente} \end{pmatrix} = \begin{pmatrix} \text{Variabile indipendente 1} & \text{Variabile indipendente 2} & \cdots & \text{Variabile indipendente m} & 1 \\ & & & & 1 \\ & & & & 1 \\ & & & & 1 \\ & & & & \cdots \\ & & & & 1 \\ & & & & 1 \\ & & & & 1 \\ & & & & 1 \end{pmatrix} \begin{pmatrix} \text{Parametri da stimare} \end{pmatrix} + \varepsilon$$

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

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

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 parte 8 sulla regressione lineare.

- Non le copio qui perché, per l'esame, vanno imparate.

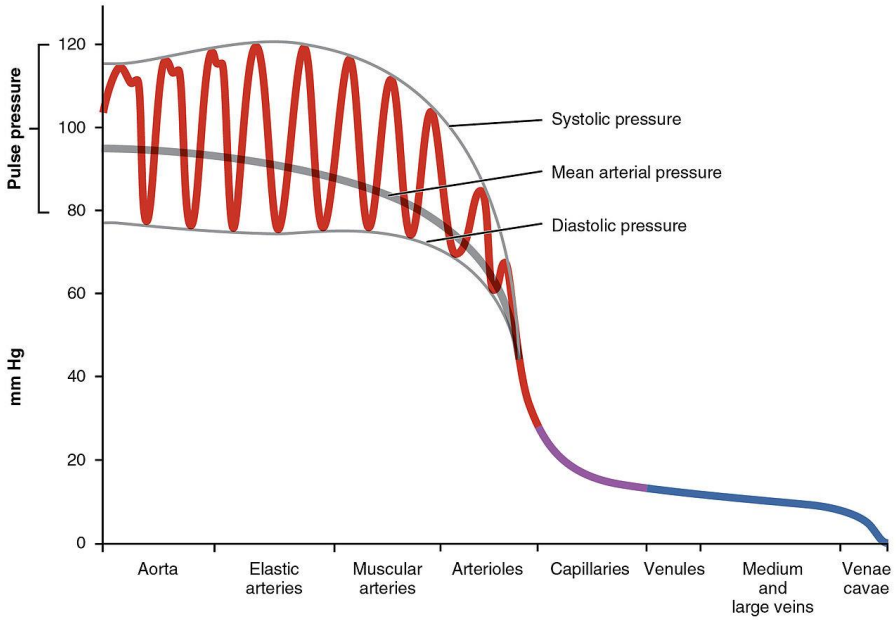
CONTESTO DELL'ESERCITAZIONE E DATI

Versione aggiornata del dataset ELSA (`ELSAv2.mat`).

Vedi workspace dopo il caricamento per il contenuto.

- **Unica modifica di rilievo** rispetto al laboratorio 3: la variabile che pensavamo essere "heart rate", cioè la frequenza cardiaca misurata in bpm (battiti al minuto) si è rivelata essere la ***pulse pressure*, ovvero la differenza tra pressione sistolica e diastolica misurata in mmHg.**

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



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

- Caricare il file `ELSAv2.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

$$pulse\ pressure = \beta_{sbp} \times systolic\ blood\ pressure + \beta_0 + \varepsilon$$

  - Inserire nel vettore Y i dati corrispondenti alla variabile dipendente (pulse presure)
  - 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)

- 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
  - 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)

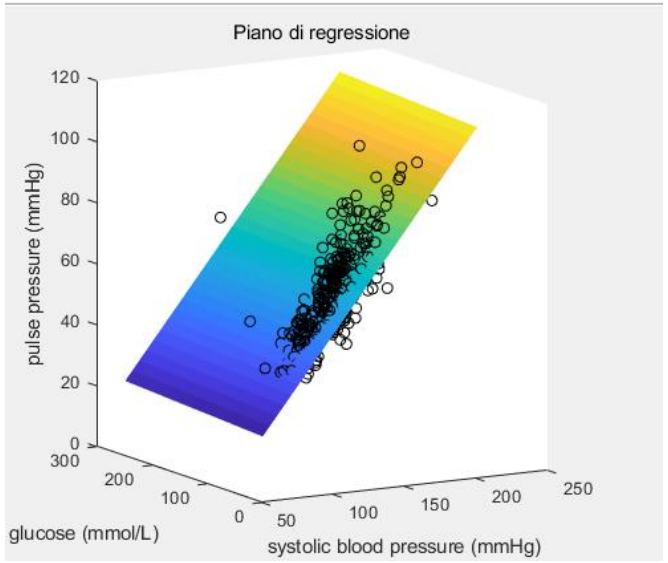
- Calcolare il coefficiente di determinazione R^2
- Calcolare lo standard error dei parametri.
- Calcolare il coefficiente di variazione (CV) percentuale dei parametri.

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

$$pulse\ pressure = \beta_{sbp} \times systolic\ blood\ pressure + \beta_{gluc} \times glucose + \beta_0 + \varepsilon$$

  - 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 omesso

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.