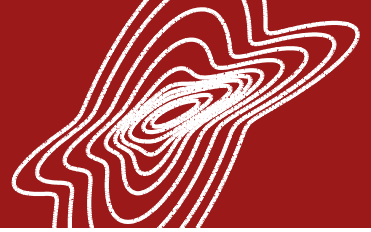


METODI STATISTICI PER LA BIOINGEGNERIA

Laboratorio 7

A.A. 2024-2025

Enrico Longato

Dal lab 6: Stimatori e denominatori

Regola mnemonica: in tantissimi casi, i denominatori degli stimatori non polarizzati correggono per il **numero di campioni** meno il **numero di parametri già stimati** prima di poter stimare la quantità di interesse

$$\sigma^2_{campionaria} = \frac{1}{N_{campioni} - 1} \sum_{i=1}^{N_{campioni}} (x_i - \mu_x)^2$$

Il parametro già stimato è 1: la media μ_x

$$\sigma^2_{a\ posteriori} = \frac{1}{N_{campioni} - N_{parametri}} \sum_{i=1}^{N_{campioni}} (y_i - \hat{y}_i)^2$$

I parametri già stimati sono tanti quanti sono i β_i, incluso, se presente nel modello, il β_0 dell'intercetta, perché sono quelli che servono per calcolare $\hat{y} = X\beta$

$$N_{parametri} = M_{variabili} + 1_{intercetta,se\ presente}$$

Dal lab 6: Lettura del codice altrui

Esercitazione di lettura del codice altrui (in questo caso, le soluzioni messe a disposizione dal docente) da svolgersi in aula.

Principi cardine

- Eseguire prima tutto il codice per sincerarsi che funzioni sul proprio dispositivo.
- Inizialmente, confrontare i soli risultati: se non tornano, verificare di star agendo correttamente (controllare dimensioni delle variabili nel workspace, denominatori, formule, …); non è esclusa una svista del docente.
- Solo in un secondo momento, confrontare il codice.
- Nel caso in cui un'istruzione risulti complicata
  - Eseguire, copiando nella command window, le sotto-istruzioni, dalla più interna alla più esterna.
  - Controllare, ad ogni passaggio, cosa sta facendo il codice (aiutandosi con la command window e il workspace).
  - Per avere un'idea più chiara, assegnare il risultato di ciascuna "sotto-istruzione interna" a una variabile, in modo da poter fare "doppio click" nel workspace.

Laboratorio 7: Contenuti e obiettivi

1. Esercitazione "alla lavagna"
   - Analisi dei residui
   - F-test
   - Test sui beta stimati
2. Esercizi da svolgere in autonomia (per superare la "paura del file bianco")
   - Stima dei parametri e della loro variabilità
   - Calcolo di R^2
   - Calcolo dell'RMSE

Analisi dei residui: i residui devono avere 4 proprietà, che cerchiamo di verificare come segue

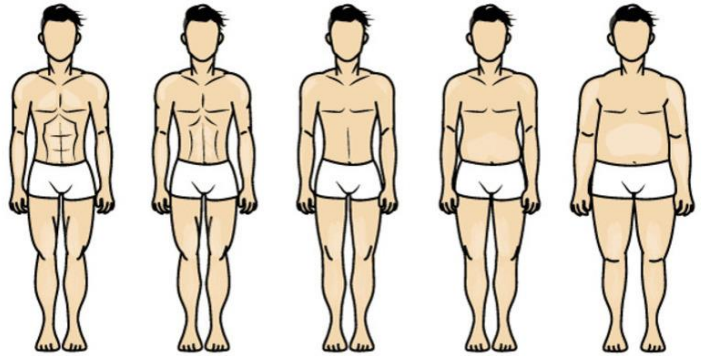
1. **Normalità**
   - Come quando lo facevamo per il test di ipotesi (stesse identiche deduzioni): istogramma (`hist`), QQ-plot (`qqplot`), test di gaussianità (`lillietest`), skewness (`skewness`), curtosi (`kurtosis`).
2. **Media nulla**
   - Semplicemente, guardiamo la media (`mean`), che deve essere vicina a 0.
3. **Scorrelazione**
   - Plot dell'autocorrelazione dopo aver ordinato secondo il valore predetto:
     1. `[Y_hat_sorted, i_sort] = sort(Y_hat)`
     2. `autocorr(residuals(i_sort))`
   - Escluso il valore in 0 (che fa sempre 1), circa il 95% dei campioni dell'autocorrelazione ottenuta dai residui così ordinati deve stare entro le bande di confidenza.
4. **Varianza omogenea**
   - Plot (`scatter`) dei residui (asse y) contro il valore predetto (asse x).
   - Deve vedersi una "banda" o "nuvola" omogenea intorno alla riga orizzontale che passa per $y = 0$.
   - **ATTENZIONE: NON fare il plot contro il valore vero** (che, per costruzione, è correlato ai residui).

CONTESTO DELL'ESERCITAZIONE E DATI

Dataset di misure antropometriche per la predizione della % di grasso corporeo (`bodyfat.mat`).

Dati di 252 uomini descritti da 8 variabili

1. BodyFat (in %) sarà la nostra variabile dipendente
2. Age (età in anni, years)
3. Weight (peso in libbre, lbs)
4. Height (altezza in pollici, inches)
5. Neck (circonferenza del collo in cm)
6. Chest (circonferenza del petto in cm)
7. Hip (circonferenza dei fianchi in cm)
8. Thigh (circonferenza dellla coscia in cm)



Adattato da https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset

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.

Analisi dei residui

- **v. slide precedente**

Test statistici sul modello di regressione (v. dopo)

- **`fitlm` per far fare la regressione lineare a MATLAB e verificare i nostri risultati.**
  - **A partire da `fitlm`: `coefTest` per il test di Fisher.**
  - **A partire da `fitlm`: accesso all'elemento `variabile_in_cui_è_salvato_il_modello.Coefficients.pValue` per i p value associati ai test sui beta.**

ESERCIZIO 1 - PARTE 1: CARICAMENTO DATI E REGRESSIONE LINEARE (proposto)

- Caricare il file `bodyfat.mat`

- **Considerando BodyFat come variabile dipendente e tutte le altre come variabili indipendenti**
  - Stimare i parametri di un modello di regressione lineare, intercetta inclusa
  - Calcolare lo standard error e il CV percentuale dei parametri
  - **[NEW ma semplice]** Calcolare l'intervallo di confidenza delle stime dei parametri come l'intervallo di estremi $\hat{\beta} \pm 1.96 \times SE(\hat{\beta})$
  - Plottare il valore vero contro il valore predetto
  - Calcolare l'RMSE
  - Calcolare R^2

ESERCIZIO 1 - PARTE 2: ANALISI DEI RESIDUI (svolto)

- Verificare le quattro proprietà fondamentali dei residui
  1. Normalità
  2. Media nulla
  3. Scorrelazione
  4. Varianza omogenea

- Seguire la traccia di soluzione nella slide introduttiva, che comprende:
  - Calcoli da fare
    - Stima della media
    - Skewness e curtosi
    - Test di gaussianità
  - Plot da disegnare
    - Istogramma dei residui
    - QQ-plot dei residui
    - Autocorrelazione dei residui (dopo aver ordinato per valore predetto)
    - Scatter plot dei residui contro il valore predetto

ESERCIZIO 1 - PARTE 3: VERIFICA DEI RISULTATI CON `fitlm` (svolto)

- Usare la funzione `mdl = fitlm(X, y, 'Intercept', false)` per effettuare la stima "automatica" dei parametri del modello di regressione lineare e di altri valori di interesse; poi confrontare i risultati con quelli ottenuti.
  - NB: `'Intercept', false` si usa solo se in X c'è già la colonna costante a 1; altrimenti, si deve usare `'Intercept', true` (che è il default)

- Come/dove si trovano i parametri di interesse da confrontare
  - **Valori dei parametri:** `beta_hat_fitlm = mdl.Coefficients.Estimate`
  - **Standard error:** `se_beta_hat_fitlm = sqrt(diag(mdl.CoefficientCovariance))`
  - **Varianza a posteriori** (attenzione al nome "strano"!): `sigma2_hat_fitlm = mdl.MSE`
  - **R^2:** `R2_fitlm = mdl.Rsquared.Ordinary`
  - **ATTENZIONE:** `rmse_fitlm_diverso_da_quello_che_usiamo_noi = mdl.RMSE` <-- questo stimatore ha al denominatore lo stesso $N_{campioni} - N_{parametri}$ della varianza a posteriori; **non** è quello che è tipicamente richiesto.

- Consiglio: se vi sembra troppo complicato (si tratta di struct innestate), prendete "per buoni" questi comandi "esplorativi" del modello come "indicazioni per trovare quello che ci interessa"; in ogni caso, la maggior parte di queste informazioni si vede da `disp(mdl)`
- *Maggiori informazioni alla pagina* https://it.mathworks.com/help/stats/linearmodel.html

ESERCIZIO 1 - PARTE 4: TEST SULLA REGRESSIONE LINEARE (svolto)

- Una volta ottenuto l'oggetto `mdl = fitlm(X, y, 'Intercept', false)`
  - Il p value associato all'F-test si trova con l'istruzione `p_F_test = coefTest(mdl)`
  - Il vettore dei p value associati a ciascun β_i si trova con l'istruzione `p_values = mdl.Coefficients.pValue`

- A fronte dei risultati ottenuti dai test di cui sopra, trarre le conclusioni del caso (v. slide di teoria).