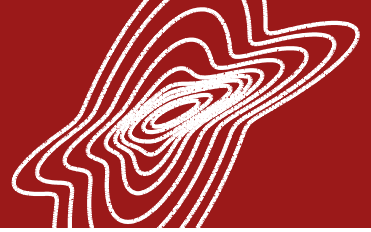


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

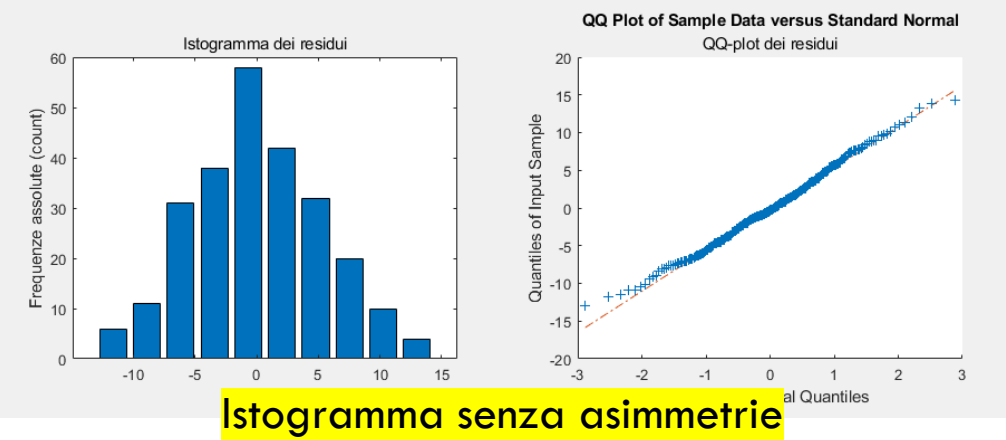
METODI STATISTICI PER LA BIOINGEGNERIA

Laboratorio 8

A.A. 2024-2025

Enrico Longato

Dal lab 7: Riassunto analisi dei residui



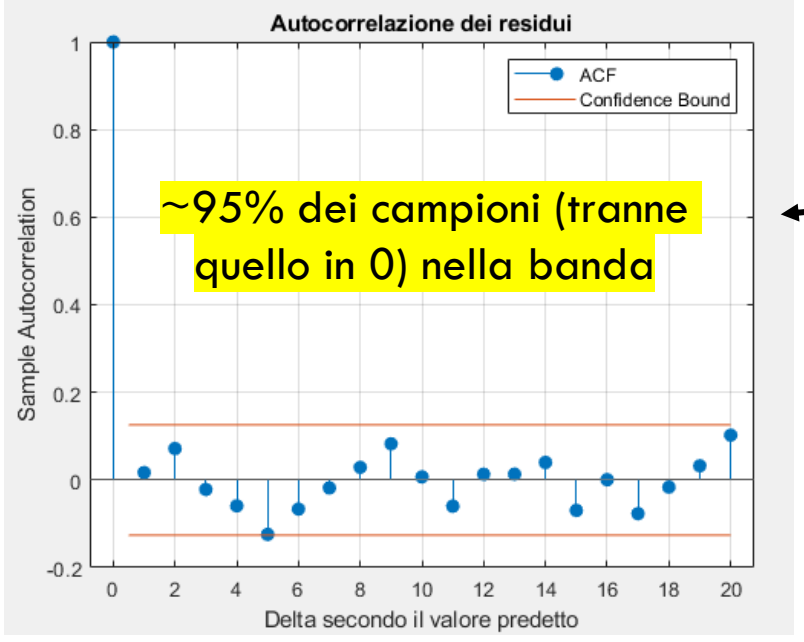
Istogramma senza asimmetrie
QQ-plot lungo una retta

```
% Skeweness e curtosi
residuals_skewness = skewness(residuals);   = 0
residuals_kurtosis = kurtosis(residuals);   = 3

% Test di gaussianità
[h_gauss, p_gauss, kstat_gauss, critval_gauss] = lillietest(residuals);
```

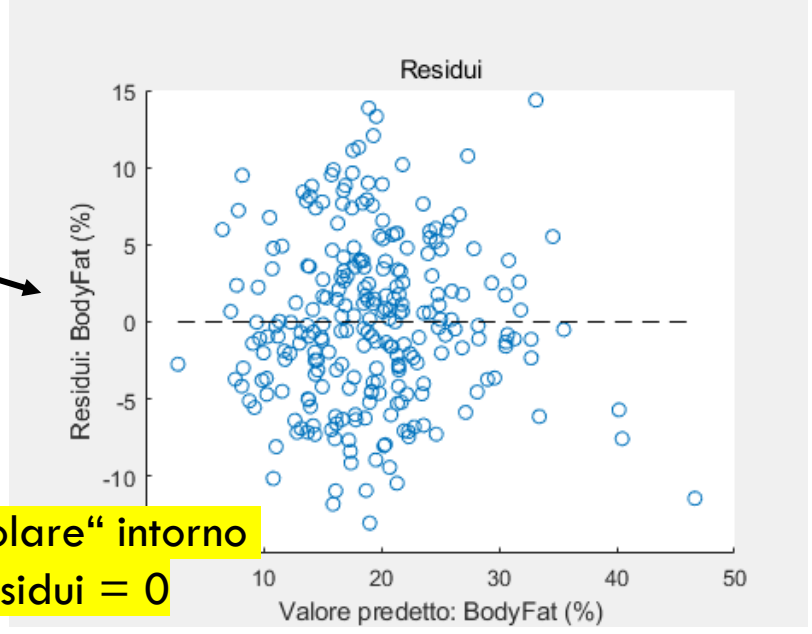
Commenti che sapete

```
% Verifica della media nulla
residuals_mean = mean(residuals);   = 0
```



~95% dei campioni (tranne quello in 0) nella banda

1. Normalità
2. Media nulla
3. Scorrelazione
4. Varianza omogenea



Banda "rettangolare“ intorno alla retta residui = 0

Dal lab 7: Correlazioni (e non) con i residui (1 di 2)

Residui e predizione sono scorrelati

Traccia di dimostrazione semplificata per Y a media nulla (altrimenti ci servono troppi lemmi)

$$\hat{Y} = X\hat{\beta}$$

$$\begin{aligned} res &= Y - \hat{Y} = \\ &= Y - X\hat{\beta} = \\ &= Y - X(X^T X)^{-1} X^T Y = \\ &= (I - X(X^T X)^{-1} X^T) Y \end{aligned}$$

$$\begin{aligned} \widehat{cov}(\hat{Y}, res) &= \frac{1}{N} \cdot \hat{Y}^T \cdot res = \\ &= \frac{1}{N} \cdot \left(X\hat{\beta}\right)^T \cdot (I - X(X^T X)^{-1} X^T) Y = \\ &= \frac{1}{N} \cdot \hat{\beta}^T X^T \cdot (I - X(X^T X)^{-1} X^T) Y = \\ &= \frac{1}{N} \cdot \hat{\beta}^T \cdot (X^T - X^T X(X^T X)^{-1} X^T) Y = \\ &= \frac{1}{N} \cdot \hat{\beta}^T \cdot (X^T - X^T) Y = 0 \end{aligned}$$

Dal lab 7: Correlazioni (e non) con i residui (2 di 2)

Residui e valore vero sono correlati

Traccia di dimostrazione semplificata per Y a media nulla (altrimenti ci servono troppi lemmi)

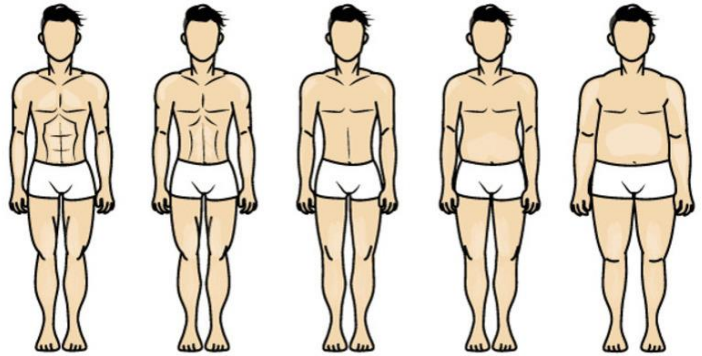
$$
\begin{aligned}
Y &= \hat{Y} + res \\
res &= Y - \hat{Y} = \\
&= Y - X\hat{\beta} = \\
&= Y - X(X^T X)^{-1} X^T Y = \\
&= (I - X(X^T X)^{-1} X^T) Y \\
\widehat{cov}(Y, res) &= \frac{1}{N} \cdot Y \cdot res = \\
&= \frac{1}{N} \cdot \left(\hat{Y} + res\right)^T \cdot res = \\
&= \frac{1}{N} \cdot (\hat{Y}^T \cdot res + res^T \cdot res) = \\
&= \frac{1}{N} \cdot (0 + \mathrm{SSE}) > 0
\end{aligned}
$$

CONTESTO DELL'ESERCITAZIONE E DATI (v. lab 7)

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.

Test statistici sul modello di regressione

- `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.

Collinearità e model selection

- Nulla: bastano le slide di teoria

ESERCIZIO 1 - PARTE 1: 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 2: 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).

ESERCIZIO 1 - PARTE 3: COLLINEARITA' (proposto)

- Possiamo continuare lo stesso esercizio di prima oppure ricaricare il file `bodyfat.mat` e identificare `BodyFat` come variabile dipendente e tutte le altre come variabili indipendenti.

- **Riscaldamento: calcolare il numero di condizionamento**
  1. Calcolare gli autovalori della matrice $X^T X$.
  2. Dividere l'autovalore massimo per l'autovalore minimo.
  3. Trarre le conclusioni del caso (v. slide di teoria).

- **Calcolare il VIF** di tutte le variabili nella configurazione iniziale in cui siano tutte possibili predittori
  1. Per ogni variabile
     1. Impostare una regressione con quella variabile "nel ruolo di Y" e tutte le altre "nel ruolo di X"
     2. Calcolare il valore di R^2 corrispondente
     3. Calcolare il VIF per quella variabile come $VIF = 1/(1 - R^2)$
  2. Una volta calcolati tutti i VIF, individuare il massimo e trarre le conclusioni del caso (v. slide di teoria)

- **Bonus** (esercizio di programmazione "difficile"; non in programma d'esame): portare a termine l'analisi basata sul VIF fino alla situazione in cui tutte le variabili hanno VIF < 5
  - Si tratta di ripetere il punto precedente sul calcolo del VIF togliendo, di volta in volta, la variabile con VIF massimo e >5.

ESERCIZIO 2: CONFRONTO TRA MODELLI (proposto)

- Caricare il file `bodyfat.mat` e mettere a punto due scenari
  1. Y = `BodyFat` e X = tutte le altre variabili (+ l'intercetta)
  2. Y = `BodyFat` e X1 = {`Age`, `Height`, `Weight`} (indicizzarle pure come `[2 3 4]`) (+ l'intercetta)

- Calcolare AIC e BIC per entrambi i modelli
- Calcolare R^2 e R^2 adjusted per entrambi i modelli
- Trarre le conclusioni del caso

Suggerimento: le quantità utili al calcolo delle quattro metriche di cui sopra sono

- Valori stimati dei parametri $\hat{\beta}$
- Predizione $\hat{Y}$
- Residui $Y - \hat{Y}$
- SSE e SST
- Numero di campioni
- Numero di parametri