Data Analytics B
23 giugno 2025 - 1 ora e 30 minuti

Nome:	N. mat.:

Riportare nello script R `DataAnalytics20250623.R` le risposte della sezione **R** ed i comandi necessari allo svolgimento della sezione **Analisi dei dati**. Lo script va consegnato tramite moodle.

R

(10) Scrivere una funzione che prenda in input due scalari numerici interi diciamo $n > 1$ e $p > 0$. La funzione, dopo avere controllato che i due argomenti rispettino le condizioni date, creerà una matrice che abbia n righe e p colonne popolata da numeri generati casualmente da una variabile gaussiana di media 0 e varianza 1. La funzione restituirà una lista contenente la matrice suddetta, il vettore delle medie delle colonne, il vettore delle varianze delle colonne e la matrice di correlazione delle colonne. Chiamare la funzione `funzione`.

(5) Fornire un'istruzione `R` che su un data frame X che contiene un fattore $X\$y$ con $k > 4$ modalità selezioni solo le unità corrispondenti al secondo e quarto livello del fattore.

Analisi dei dati

Ogni risposta corretta vale 2.5 punti e ogni risposta errata dà luogo ad una penalizzazione di 0.5.

Si considerino i dati contenuti nel file `fev.cvs`. Si tratta di informazioni raccolte su 654 bambini e adolescenti. La variabile chiave è `FEV`, la capacità polmonare (in litri). Inoltre si dispone dell'età (`AGE`), della statura (`HEIGHT`, in pollici) e di due variabili dicotomiche, genere (`SEX`) e fumo (`SMOKE`).

1. Quanto vale la devianza della variabile `HEIGHT`

   a. circa 21274.66
   b. circa 21242.13
   c. circa 32.53
   d. circa 32.48

2. Si trasformino le variabili quantitative mediante standardizzazione e si calcoli la matrice di covarianza fra tali nuove variabili. Quale delle tre ha la covarianza più elevata?

   a. `AGE` e `HEIGHT`
   b. nessuna coppia di variabili mostra una relazione lineare
   c. `AGE` e `FEV`
   d. `FEV` e `HEIGHT`

3. Si calcoli la statistica X^2 per valutare l'associazione fra le variabili `SMOKE` e `SEX`. Quanto vale?

   a. circa 3.74
   b. 0.6
   c. circa -0.076
   d. circa -0.011

4. Si vuole usare come misura di asimmetria l'indice K definito come

$$K = \frac{x_{0.9} + x_{0.1} - 2x_{0.5}}{x_{0.9} - x_{0.1}}$$

per la variabile `FEV`. Quanto vale?

a. circa 0.004
b. circa 0.66
c. 0.33
d. circa 0.15

5. A partire dalle variabili `SMOKE` e `SEX` si consideri la varibile che è il prodotto logico delle due (ovvero una varibile che ha 4 modalità ("M e Fumatore", "F e Fumatore", "M e Non Fumatore", "F e Non Fumatore"). Per quale gruppo identificato da questa nuova variabile risulta la mediana di `FEV` più piccola?

a. "M e Fumatore"
b. "F e Fumatore"
c. "M e Non Fumatore"
d. "F e Non Fumatore"

6. Si determinino i parametri della funzione di regressione multipla che ha come variabile risposta `FEV` e come variabili esplicative `AGE` e `SMOKE`.

6.1 La devianza spiegata dalla regressione vale circa:

a. 443.3
b. circa 0.58
c. 0.43
d. circa 283.06

6.2 Quale tra le seguenti affermazioni è accettabile:

a. a parità di età la capacità polmonare dei fumatori è inferiore in media a quella dei non fumatori di circa 0.21 litri
b. la capacità polmonare dei fumatori è inferiore in media a quella dei non fumatori di circa 0.21 litri
c. a parità di età la capacità polmonare dei fumatori è superiore in media a quella dei non fumatori di circa 0.21 litri
d. la capacità polmonare dei fumatori è superiore in media a quella dei non fumatori di circa 0.21 litri

6.3 Quale delle seguenti affermazioni non è supportata dalle analisi. La variabile `AGE`

a. è molto rilevante per spiegare la capacità polmonare
b. all'aumentare dell'età la capacità polmonare aumenta in media del 23%
c. all'aumentare dell'età di 1 anno la capacità polmonare aumenta in media di 0.23 litri
d. per un soggetto di 10 anni e fumatore, la capacità polmonare prevista è di circa 2.46 litri

6.4 Qual è la percentuale di devianza dei residui della funzione di regressione

a. circa 0%
b. circa 44.3%
c. circa 56.5%
d. circa 42.3%