# Esame di Analisi dei dati 4 Luglio 2024

Avete 1h e 45m!

Risposte errate nelle domande a risposta multipla e vero/falso pesano negativamente sulla valutazione.

| Nome e cognome | Matricola | 1 |
|---|---|---|
| | | |

1. Una variabile aleatoria con distribuzione binomiale é sempre simmatrica per qualunque valore dei parametri n e p. (Vero/Falso)
2. L'odds puo essere negativo (Vero/Falso)
3. L'errore di campionamento dipende solo dalla varibilità nella popolazione (Vero/Falso)
4. Più $\sigma_R^2$ è piccola, più la retta di regressione spiega le variazioni della risposta (Vero/Falso)
5. L'analisi in componenti principali puo essere eseguita solo a partire dalla matrice di varianza-covarianza (Vero/Falso)
6. $R^2$ e $R^2 coretto$ indicano se le variabili esplicative sono la vera causa della variabilità della variabile dipendente (Vero/Falso)
7. La sovracopertura avviene quando alcuni elementi della lista sono inesistenti e/o non appartengono alla target population (Vero/Falso)
8. Il campionamento sistematico differisce dal casuale semplice solo per la tecnica di estrazione (Vero/Falso)

**Esercizio 1** Il dataset Titanic contiene i seguenti dati sul disastro del famoso transatlantico: **PassengerId**: un identificatore univoco per ciascun passeggero; **Sopravvissuto**: variabile che indica se il passeggero è sopravvissuto (0 = No, 1 = Sì); **Pclass**: la classe del passeggero ($1 = 1a, 2 = 2a, 3 = 3a$); **Nome**: il nome del passeggero; **Sesso**: il sesso del passeggero (maschio, femmina); **Età**: l'età del passeggero espressa in anni; **SibSp**: il numero di fratelli o coniugi che il passeggero aveva a bordo del Titanic; **Parch**: numero di genitori o figli che il passeggero aveva a bordo del Titanic; **Biglietto**: il numero del biglietto del passeggero; **Tariffa**: la tariffa pagata dal passeggero; **Cabina**: il numero di cabina assegnato al passeggero (se presente); **Imbarcato**: il porto d'imbarco (C = Cherbourg, Q = Queenstown, S = Southampton).
La tabella sottostante riporta i risultati del modello di regressione multipla che cerca di spiegare se il passeggero è sopravvissuto o meno sulla base delle sue caratteristiche.

```
Call:
lm(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +
    Fare + Embarked, data = titanic_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.06092 -0.21656 -0.08607  0.22393  1.00290

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.2638441  0.2753352   4.590 5.07e-06 ***
Pclass2     -0.1437791  0.0453742  -3.169  0.00158 **
Pclass3     -0.3346416  0.0425032  -7.873 1.01e-14 ***
Sexmale     -0.5021838  0.0283859 -17.691  < 2e-16 ***
Age         -0.0058328  0.0010819  -5.391 8.99e-08 ***
SibSp       -0.0409771  0.0130682  -3.136  0.00177 **
Parch       -0.0163464  0.0182360  -0.896  0.37029
Fare         0.0003474  0.0003401   1.021  0.30732
EmbarkedC   -0.1010612  0.2713371  -0.372  0.70964
EmbarkedQ   -0.1028007  0.2741059  -0.375  0.70772
EmbarkedS   -0.1702427  0.2710056  -0.628  0.53004
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3794 on 880 degrees of freedom
Multiple R-squared:  0.3988,    Adjusted R-squared:  0.3919
F-statistic: 58.36 on 10 and 880 DF,  p-value: < 2.2e-16
```

**a.** Qual è il nome preciso del modello di regressione lineare usato in questo esercizio?

**b.** Siete d'accordo nell'usare tale modello? Avreste scelto uno differente? Motivare la risposta.

**c.** In base alla vostra risposta al punto b quali variabili avreste inserito e perché? (spiegare la motivazione variabile per variabile scegliendone almeno 5)

**d.** Quali delle variabili nel dataset potrebbero essere correlate tra loro?

**e.** Provare a commentare comunque la bontà di adattamento del modello riportato in tabella.

**f.** Provare a commentare comunque i risultati del modello alla luce delle stime di tutti i coefficienti riportate in tabella. Cosa vi dice in particolare l'intercetta?

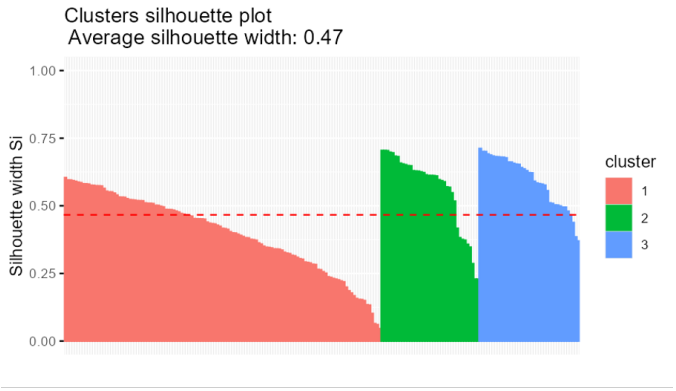
**2** Una grande catena di centri commerciali negli Stati Uniti ha deciso segmentare la clientela in gruppi. I dati raccolti consistono delle seguenti variabili: CustomerID (un identificatore univoco intero per ciascun cliente); Sesso (variabile categoriale che indica il sesso del cliente maschio/femmina); Etá (un numero intero che rappresenta l'etá del cliente in anni); Reddito annuo (una variabile numerica che rappresenta il reddito annuo del cliente in migliaia di dollari); Punteggio di spesa (1-100) (un punteggio numerico assegnato dal centro commerciale in base al comportamento dei clienti e alla natura della spesa, dove 1 é il piú basso e 100 é il piú alto)
Questo dataset include varie funzionalitá sui clienti che possono essere utilizzate per segmentarli in diversi cluster.

**a.** Sapendo che é stato utilizzato l'algoritmo delle k-medie quali delle varibaili disponibili nel dataset potrebbero essere state utilizzate? Motivare la scelta.

**b.** Spiegare brevemente la differenza tra k-medie e clustering gerarchica. Quando si usa l'una o l'altra?

**c.** Il seguente plot e tabella riporta i risultati della Silhouette, che fornisce informazioni su quanto bene ogni oggetto si trova all'interno del suo cluster. In particolare, cosa concludereste **dal plot e dalla tabella** sottostante?

| | cluster | size | ave.sil.width |
|---|---|---|---|
| 1 | 1 | 123 | 0.40 |
| 2 | 2 | 38 | 0.56 |
| 3 | 3 | 39 | 0.59 |




**3** E' stata condotta un'analisi in componenti principali di una serie di varibili legate alla criminalità negli Stati Uniti. Le variabili in oggetto sono le seguenti: MURDER (numero di arresti per omicidio su 100mila abitanti), ASSAULT (Numero di arresti per aggressioni su 100mila abitanti), URBANPOP (percentuale di popolazione che vive in aree urbane), RAPE (Numero di arresti per stupro su 100mila abitanti).

**a.** Si commenti la seguente matrice degli autovalori

| | eigenvalue | % of variance | cumulative % of variance |
|---|---|---|---|
| comp 1 | 2.48 | 62.01 | 62.01 |
| comp 2 | 0.99 | 24.74 | 86.75 |
| comp 3 | 0.36 | 8.91 | 95.66 |
| comp 4 | 0.17 | 4.34 | 100.00 |

**b.** La seguente tabella riporta la correlazione tra componenti principali e variabili. Cosa concludete?

| | Dim.1 | Dim.2 |
|---|---|---|
| Murder | 0.84 | -0.42 |
| Assault | 0.92 | -0.19 |
| UrbanPop | 0.44 | 0.87 |
| Rape | 0.86 | 0.17 |

**c.** Commentare i risultati osservando i grafici delle proiezioni di variabili e individui nello spazio fattoriale