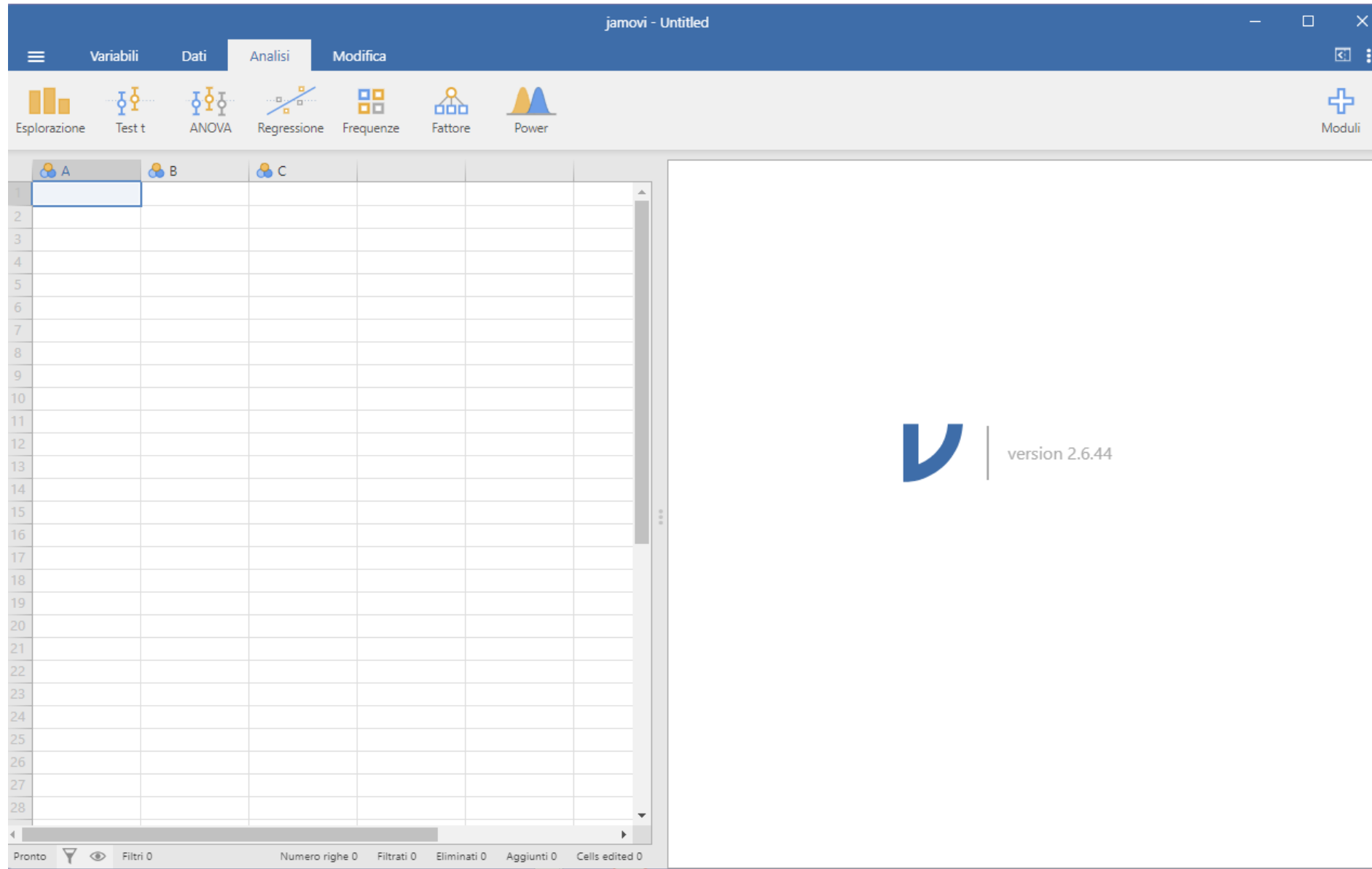


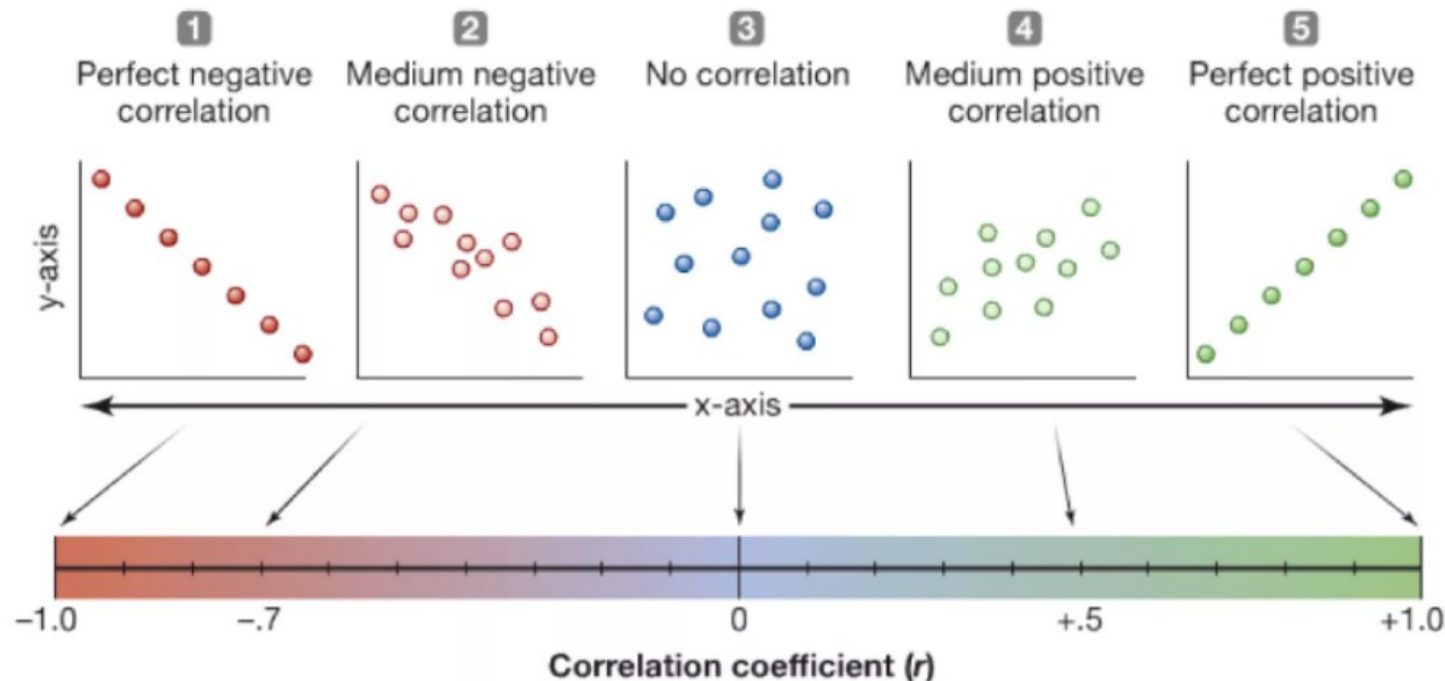
Correlazione & regressione



Indice di correlazione

r	Grado di associazione
Da 0.8 a 1.0 (da -1.0 a -0.8)	forte
Da 0.5 a 0.8 (da -0.8 a -0.5)	buona
Da 0.2 a 0.5 (da -0.5 a -0.2)	debole
Da 0.0 a 0.2 (da -0.2 a 0)	trascurabile

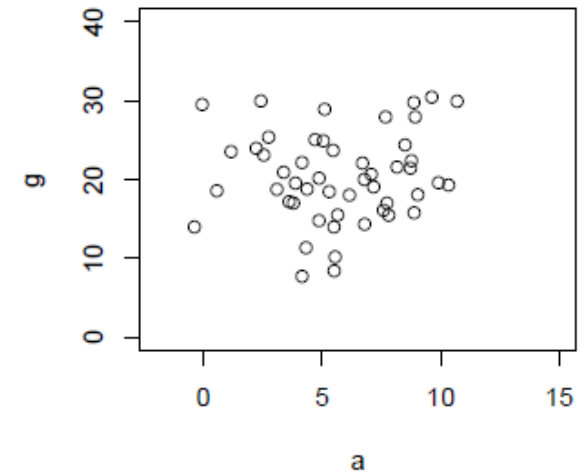
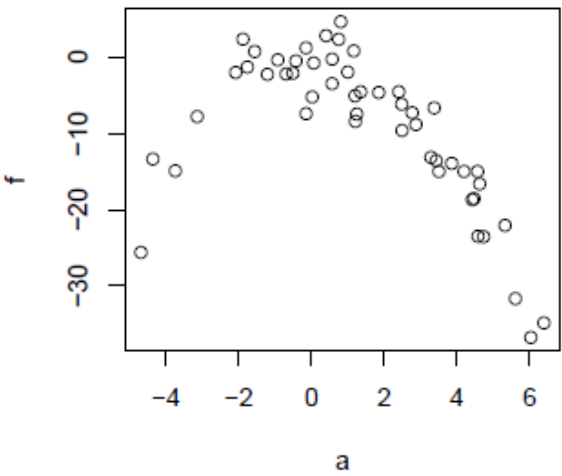
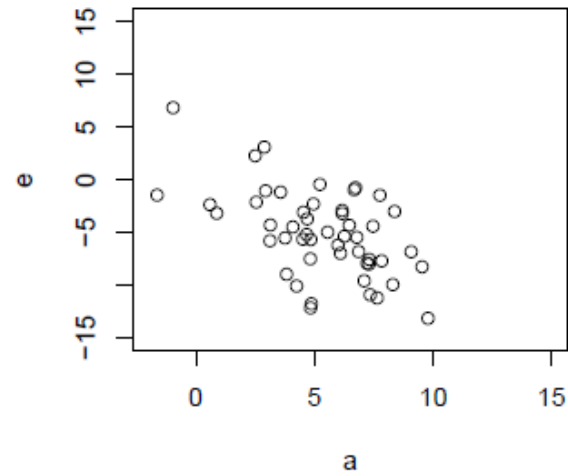
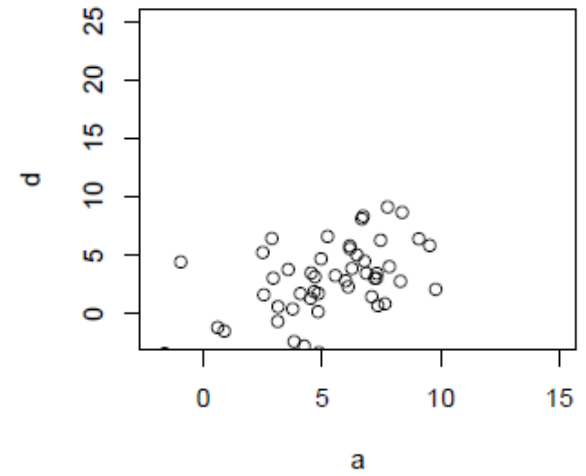
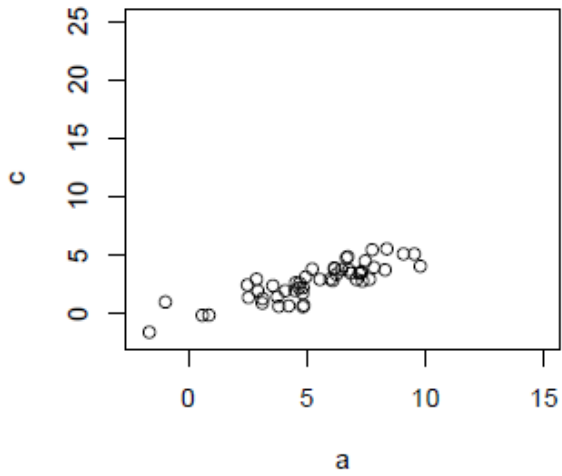
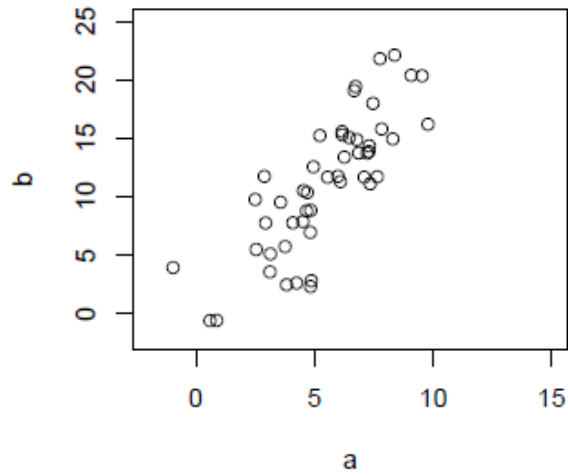
- Lo studio della correlazione permette di capire se esiste una **relazione** tra due variabili e la forza di tale associazione (Pearson→**relazione lineare**)



$$H_0: r = 0$$

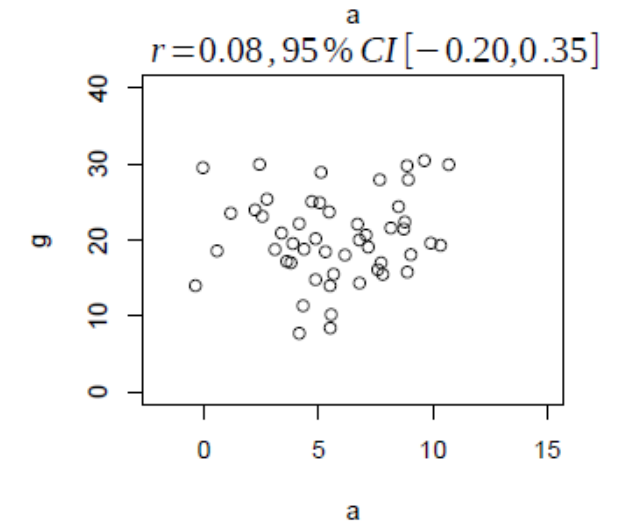
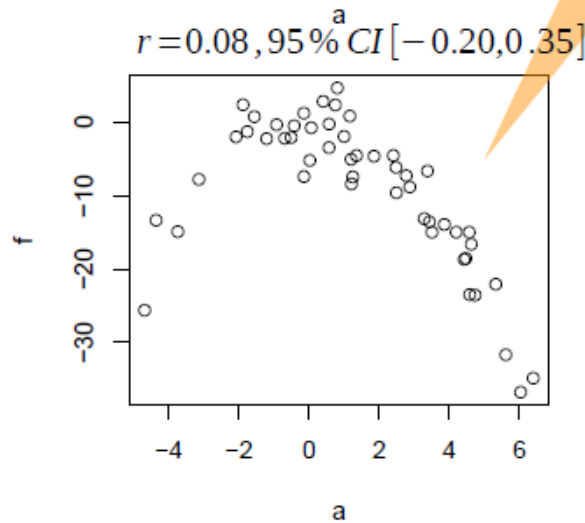
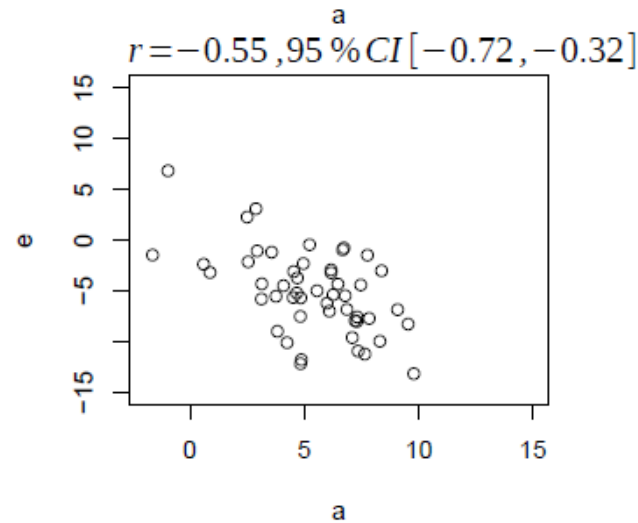
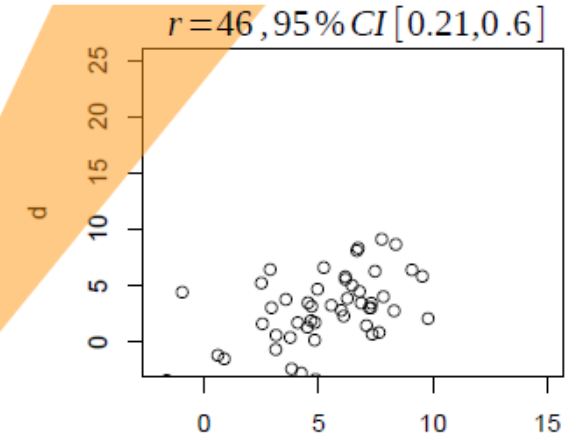
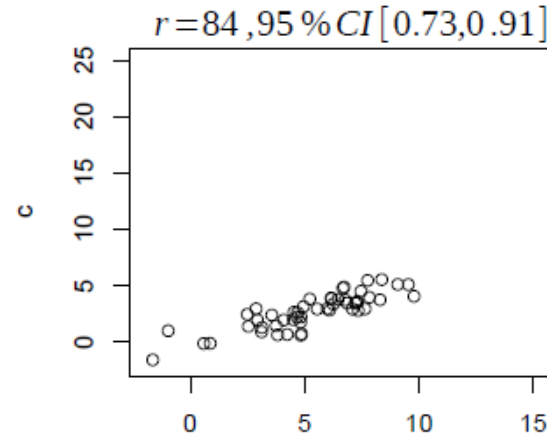
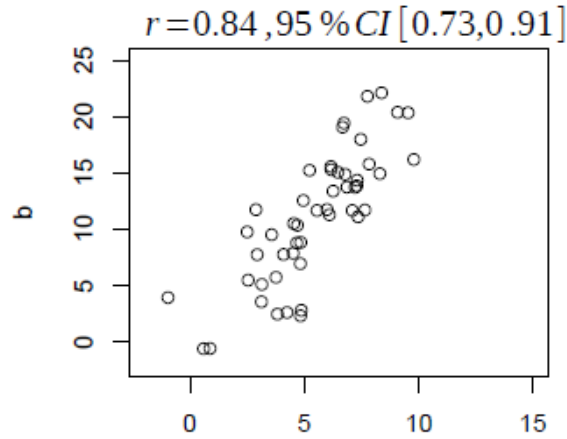
- Se il nostro studio è **inferenziale** (campione→popolazione), dobbiamo eseguire un **test di ipotesi** (e calcolare l'**intervallo di confidenza** quando possibile)

Dataset artificiale: *correlazione*

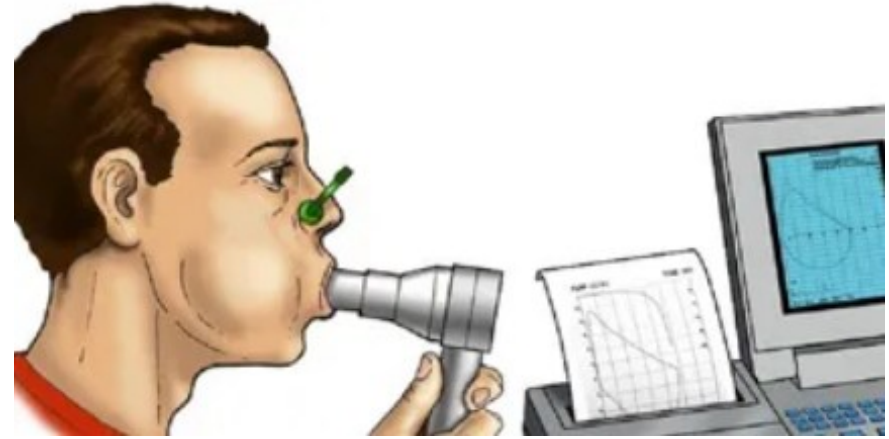


Dataset artificiale: *correlazione*

Gli indici di correlazione più utilizzati valutano la presenza di una **relazione lineare**



Dataset: *cystfibr*



- Il dataset contiene i risultati di uno studio su $n=25$ pazienti affetti da fibrosi cistica, una malattia molto grave causata da un difetto nel gene CFTR.
- Uno degli apparati più colpiti è quello respiratorio.
- Allo scopo di monitorare la funzionalità polmonare, vengono prese in considerazione due misurazioni: il **volume espiratorio massimo nel 1° secondo** (FEV1, l) e la **capacità funzionale residua** (FRC, ml).

Dataset: *cystfib*

- c'è una relazione tra FEV1 e FRC?

Stima della correlazione

- È possibile stimare il valore di FEV1 in base a FRC?

Dobbiamo quantificare la relazione lineare

$$\text{FEV1} = \alpha + \beta * \text{FRC} + \text{Errore}$$

- Come varia FEV1 in base a FRC?

Analisi del coefficiente di regressione β

Dataset: *cystfibr*

- Consideriamo l'individuo ID=6
 - Qual è il suo valore di FRC?
 - $x=155$
 - Qual è il suo valore di FEV1?
 - $y=4.4$
 - Qual è il valore di FEV1 predetto? $\hat{y} = 6.12 - 0.02 \times 155 = 3.5$
 - Qual è il suo residuo (errore)? $y - \hat{y} = 4.4 - 3.5 = 0.9$

Carichiamo adesso un set di dati chiamato «BP» che contiene delle informazioni su l'effetto del sale sulla pressione arteriosa di un campione di soggetti.

- C'è una **correlazione** significativa tra l'età dei soggetti e la SBP?
- Quale è l'impatto stimato sulla SBP di un anno di crescita dell'età?
- C'è un effetto circa le abitudini di utilizzo del sale sulla SBP «aggiustato» per l'età ?

Regressione Logistica

Carichiamo infine un set di dati chiamato «parti_regr_log» che contiene delle informazioni su l'impatto di alcune caratteristiche della donna sul rischio di parto cesareo.

- C'è una associazione significativa tra l'induzione al parto e avere un parto cesareo?
- Come si può quantificare l'effetto?
- C'è un effetto dell'induzione al parto «aggiustato» per il BMI della madre?