Statistica Computazionale
Esempio di compito (3)

È richiesta una relazione contenente il codice R, debitamente commentato, necessario per risolvere i problemi posti. Le procedure usate devono essere giustificate a parole. Anche i grafici, se presenti, utilizzati nella discussione devono essere inseriti nell'elaborato.

Dove richiesto, devono essere usate 10000 replicazioni (Monte Carlo o bootstrap).

Non è possibile consultare alcun materiale, tranne l'"help" di R.

Il tempo a disposizione è di 135 minuti.

Esercizio 1

Il file `amianto.dat`[1] contiene 227 coppie di osservazioni (x_i, y_i). La prima colonna `Y` contiene le misure y_i, $i = 1, \ldots, 227$, del numero di fibre di amianto nei polmoni di lavoratori ex-esposti ad amianto (espresse in numero di fibre per grammo di polmone secco), mentre la seconda colonna `X` contiene una variabile dicotomica che indica il gruppo di esposizione ($x_i = 1$ se il lavoratore era esposto ad amianto *anfibolo*, mentre $x_i = 0$ se era esposto ad amianto *crisotilo*). Si vuole verificare se il tipo di esposizione influisce sul numero di fibre contenute nei polmoni.

a. Verificare, utilizzando un test t a due campioni, l'ipotesi che la media del numero di fibre sia uguale per i due gruppi (si verifichi l'assunzione di normalità alla base del test t).

b. Verificare, utilizzando un approccio bootstrap non parametrico, l'ipotesi che la mediana del numero di fibre sia la stessa per i due gruppi. Si fornisca il livello di significatività osservato del test.

c. Si forniscano degli intervalli di confidenza bootstrap al 95% per le mediane nei due gruppi.

Esercizio 2

Il file `cardio.dat` contiene l'informazione in merito alla presenza di sintomi da patologia cardiaca per 99 persone, divise in 43 classi di età x_i, $i = 1, \ldots, 43$ (colonna `age`). Per la i-ma classe di età x_i sono riportati il numero n_i di individui nella classe (colonna `n`) e il numero y_i di coloro che riportano dei sintomi connessi a problemi cardiaci (colonna `y`). L'obiettivo è studiare come la probabilità di sviluppare sintomi varia in funzione dell'età. In particolare, si suppone che la relazione tra tale probabilità e l'età sia definita da

$$\pi(\boldsymbol{\beta}, x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \ ,$$

dove $\boldsymbol{\beta} = (\beta_0, \beta_1)$. La variabile Y_i, che conta il numero di soggetti con problemi cardiaci nella i-ma classe di età, si distribuisce come una binomiale: $Y_i \sim Bin(n_i, \pi(\boldsymbol{\beta}, x_i))$ e la funzione di

[1] Per importare i dati, si usi `read.table("amianto.dat", header=T, sep=",")`.

probabilità associata è:

$$f_{y_i}(x_i, n_i, y_i; \boldsymbol{\beta}) = \binom{n_i}{y_i} \left[\pi(\boldsymbol{\beta}, x_i)\right]^{y_i} \left[1 - \pi(\boldsymbol{\beta}, x_i)\right]^{n_i - y_i} , \qquad i = 1, \ldots, 43.$$

Formalmente, questo è un problema di regressione logistica (in cui i dati sono stati raggruppati per classi di età), con funzione di legame `logit` (funzione di legame canonica).

a. Si scriva una funzione R che calcola la log-verosimiglianza in corrispondenza di due generici valori di β_0 e β_1.

b. Si disegni un grafico della funzione di log-verosimiglianza per $\beta_0 \in [-10, 0]$ e $\beta_1 \in [0, 0.2]$ e se ne commenti l'andamento: quali valori suggerisce il grafico per la stima di massima verosimiglianza?

c. Si ottenga la stima di massima verosimiglianza di $\boldsymbol{\beta} = (\beta_0, \beta_1)$ impiegando la funzione `nlm` di R e adottando una opportuna riparametrizzazione, se necessaria. Per la scelta dei valori di partenza, ci si aiuti con la rappresentazione grafica ottenuta al punto b. Si stimi poi lo stesso modello la funzione `glm`[2]. Ci sono differenze tra le stime di $\boldsymbol{\beta}$ ottenute con i due metodi?

d. Si ottengano gli intervalli di confidenza con grado di fiducia pari al 95% per i parametri β_0 e β_1, utilizzando la normalità asintotica dello stimatore di massima verosimiglianza[3].

Esercizio 3

Il file `esercizio_3.dat`[4] contiene 70 coppie di osservazioni (x_i, y_i).

a. Si stimi una regressione non parametrica basata sul metodo del nucleo, scegliendo un opportuno valore del parametro di lisciamento h[5].

b. Si stimi, usando la funzione `spline.regression`, una regressione non parametrica basata sulle spline, calcolando un opportuno valore del parametro di lisciamento λ tramite validazione incrociata[6].

c. Si sviluppi un test d'ipotesi bootstrap per verificare la linearità della relazione tra X e Y, sfruttando la regressione non parametrica basata sulle spline vista al punto b e fornendo il livello di significatività osservato per tale test.

---

[2] Il comando necessario è `glm(cbind(y, n - y) ~ x, family = binomial(logit))`, dove `n`, `y` e `x` sono i vettori contenenti i valori n_i, y_i e x_i, rispettivamente, per $i = 1, \ldots, 43$.

[3] Può essere utile la funzione fdHess nella biblioteca nlme.

[4] Per importare i dati, si usi `read.table("esercizio_3.dat", header=T, sep=",")`.

[5] Si utilizzi la biblioteca `sm`. Si ricorda che la stima della regressione richiesta è ottenuta tramite la funzione `sm.regression(x, y)`, mentre la funzione `h.select(x, y, method="cv")` restituisce il grado di lisciamento ottimale.

[6] La funzione `spline.regression` è contenuta nel file `spline.regression.R`, messo a disposizione. Inoltre, si ricorda che la stima della curva non parametrica è contenuta nell'oggetto `estimate` estratto con `spline.regression(x, y, display=none)$estimate`.