Esercizi di riepilogo

Analisi statistica dei dati biologici

Correlazione – Es 17 pag. 278

Se incontraste difficoltà nel risolvere questi problemi, sarebbe meglio dormirci sopra e ritentare il mattino dopo? Huber et al (2004) hanno chiesto a 10 persone di eseguire al computer un complesso esercizio di apprendimento spaziale subito prima di coricarsi. Hanno poi utilizzato l'encefalografia per misurare l'attività elettrica delle cellule durante il sonno. Il valore dell'aumento del sonno a «onde lente» dopo lo svolgimento dell'esercizio complesso, rispetto al livello di riferimento, è riportato nella tabella; è indicato anche l'aumento della prestazione registrato quando ai soggetti è stato proposto lo stesso esercizio al risveglio.

Correlazione – Es 17 pag. 278

Aumento dell'attività a onde lente durante il sonno (%)	Miglioramento della prestazione nell'esecuzione dell'esercizio dopo il risveglio (%)
8	8
14	3
13	0
15	0
17	8
18	15
31	14
32	10
44	27
54	26

Correlazione – Es 17 pag. 278

a. Rappresentate i dati in un diagramma a dispersione.
b. Calcolate il coefficiente di correlazione tra l'aumento del sonno a onde lente e il miglioramento della prestazione al risveglio.
c. Qual è l'errore standard della stima calcolata nella parte (b)?
d. Interpretate la grandezza che avete calcolato nella parte (c). Che cosa misura?
e. Verificate l'ipotesi che le due variabili siano correlate nella popolazione.

Correlazione – Es 17 pag. 278

```
#carico i dati presenti nel file «Esercizi_Correlazione_Es17Pag178.txt»
dati<-read.table(choose.files(),header=T)
#a.visualizzo i dati con un diagramma a dispersione
matplot(dati$AumentoAttivita,dati$MiglioramentoPrestazione,pch=1,col="red")
#b.calcolo il coefficiente di correlazione
r<-cor(dati$AumentoAttivita,dati$MiglioramentoPrestazione,method="pearson")
#c.calcolo l'Errore Standard
ESr<-sqrt((1-r^2)/(length(dati$AumentoAttivita)-2))
#d. L'Errore standard di r è la deviazione standard della distribuzione campionaria di r.
#e.test d'ipotesi:
#H0: nessuna relazione tra le due variabili (ρ=0)
#H1: le due variabili sono correlate (ρ≠0)
cor.test(dati$AumentoAttivita,dati$MiglioramentoPrestazione,
method="pearson", conf.level=0.95)
#Rifiuto H0, esiste una correlazione positiva tra l'aumento del sonno a onde lente e il miglioramento delle prestazioni al risveglio.
```

Modelli Lineari Generali – Es 7 pag. 326

Si è scoperto che il gene foraggiamento (*for*) è alla base della variazione del comportamento (relativo al foraggiamento) in parecchie specie di insetti. Ben-Shahar et al (2002) hanno esaminato se il gene è in grado di influenzare le differenze comportamentali nell'ape domestica (*Apis mellifera*). Nell'alveare le api operaie svolgono compiti quali la cura della covata (api nutrici) quando sono giovani, ma passano al foraggiamento di nettare e polline fuori dell'alveare (api bottinatrici) quando invecchiano. Gli autori hanno confrontato l'espressione del gene *for* nelle api nutrici e nelle api bottinatrici in tre colonie. I risultati sono riportati nella tabella che accompagna questo problema. L'espressione del gene è misurata in unità arbitrarie.

Modelli Lineari Generali – Es 7 pag. 326

Tipo di ape operaia	Colonia	Espressione gene for
Nutrice	1	0.99
Bottinatrice	1	1.93
Nutrice	2	1.00
Bottinatrice	2	2.36
Nutrice	3	0.24
Bottinatrice	3	1.96

Modelli Lineari Generali – Es 7 pag. 326

a. Costruite il diagramma di interazione per questi dati.
b. Formulate un modello lineare generale da adattare a questi dati. Cosa è rappresentato da ciascun termine del modello?
c. Valutare se i tipi di operaie differiscono nella loro espressione genica media.
d. Valutare se il termine COLONIA sia statisticamente significativo oppure no. Nel caso non lo sia, si dovrebbe eliminare dal modello lineare generale?

Modelli Lineari Generali – Es 7 pag. 326

#carico i dati presenti nel file «Esercizi_MLG_Es7Pag326.txt»

#a. Grafico di interazione

```
interaction.plot(dati$TIPODIOPERAIA,dati$COLONIA,dati$ESPRESSIONE)
```

#non sembra esserci interazione tra i fattori

#b.Formulo il modello lineare generale:

#ESPRESSIONE=COSTANTE+TIPODIOPERAIA+COLONIA+TIPODIOPERAIA*COLONIA

#ESPRESSIONE: livello di espressione del gene for

#COSTANTE: media generale dell'espressione del gene for

#TIPODIOPERAIA: tipo di ape operaia

#COLONIA: colonia di provenienza delle api

#TIPODIOPERAIA*COLONIA: interazione tra il tipo di operaia e la colonia di provenienza

Modelli Lineari Generali – Es 7 pag. 326

#c. Per prima cosa testiamo se è presente interazione

#1-Confrontiamo il modello completo con il modello senza interazione

```
anova(lm(ESPRESSIONE~TIPODIOPERAIA+COLONIA+TIPODIOPERAIA*COLONIA, data=dati))
```

#l'interazione non è significativa, il modello nullo (senza interazione) è preferito

#2) escludiamo dal modello completo l'interazione e verifichiamo l'effetto dei fattori

```
fit<-lm(ESPRESSIONE~TIPODIOPERAIA+COLONIA, data=dati)
anova(fit)
```

#il termine TIPODIOPERAIA risulta statisticamente significativo, in altre parole, il modello completo con il termine TIPODIOPERAIA viene favorito rispetto al modello nullo che non lo include. Questo significa che i due tipi di ape operaia differiscono nell'espressione media per il gene for.

d. Il termine COLONIA non è significativo, cioè viene favorito il modello nullo dove questo termine è assente. Non sarebbe corretto eliminare questo termine dal modello perché fa parte dell'impianto sperimentale e inoltre aumenta l'abilità nell'identificare l'effetto degli altri fattori (TIPODIOPERAIA)

Verosimiglianza – Es 10 pag. 358

Lo spazzaforno (*Thymelaea hirsuta*), un arbusto mediterraneo, ha 5 tipi sessuali, il più curioso dei quali è «genere labile»: gli individui di questo gruppo mutano di anno in anno il loro sesso predominante. Ramadan et al (1994) hanno trovato 13 individui genere-labili in un campione di 68 arbusti in un singolo habitat in Egitto.

Verosimiglianza – Es 10 pag. 358

a. Quale distribuzione di probabilità usereste per calcolare la probabilità che ci sia un particolare numero di individui genere-labili in un campione di dimensione uguale a 68?
b. Scrivete la formula per la verosimiglianza di p, in base ai dati. Che cosa misura questa verosimiglianza?
c. Scrivete la formula per la log-verosimiglianza di p.
d. Identificate la stima di massima verosimiglianza.
e. Generate un intervallo di confidenza al 95% di p basato sulla verosimiglianza.

Verosimiglianza – Es 10 pag. 358

#a. La distribuzione binomiale

#b. La formula della verosimiglianza è:

$L[p|13\ labile] = \binom{68}{13} p^{13}(1-p)^{55}$

#Questa quantità misura la probabilità di avere 13 individui «genere labile» in un campione di 68 se la proporzione di individui «genere labile» nella popolazione è p.

c. La formula della log-verosimiglianza è:

$$lnL[p|13\ labile] = ln\left[\binom{68}{13}\right] + 13ln[p] + (55)\ln[p]$$

Verosimiglianza – Es 10 pag. 358

```
#d. Stima di massima verosimiglianza:
n<-68
y<-13
#definiamo 100 punti in cui calcolare la L nell'intervallo di p 0.05-0.95
int_p<-seq(from=0.05, to=0.95, length.out=100)
#calcolo la logL per tutti i punti nell'intervallo
logL<-lchoose(n,y)+y*log(int_p)+(n-y)*log(1-int_p)
#trovo il massimo di logL nell'intervallo
logLmax<-max(logL)
stima<-int_p[logL==logLmax]
#creo la curva di massima verosimiglianza
matplot(int_p, logL, type="l")
#e. intervallo di confidenza di p al 95%
#calcolo la distanza di ogni punto dal valore massimo di logL
difL<-logL-max(logL)
#creo una matrice per il controllo visivo della distanza
cbind(int_p,difL,difL<=-1.92)
 #IC95%: 0.104 – 0.295
```