Esercizi - Analisi Predittiva - II

aa 2023/2024

Esercizio 1

Si consideri il dataset `wbca` del pacchetto `faraway`. Il dataset contiene dati riguardo uno studio oncologico in cui si vuole poter individuare se un tumore è maligno o meno usando alcune caratteristiche delle cellule estratte usando un ago aspirato. Si veda anche `?faraway::wbca`.

1. Si esegua una prima stima in cui la variabile `Class`, cioè la variabile che indica la classificazione del tumore, dipende dalla variabile `Thick`: si produca un grafico che mostra come la probabilità che un tumore sia benigno dipende da `Thick`. Si crei inoltre un intervallo di confidenza al 96% per il parametro relativo alla variabile `Thick`.

2. Si stimi un modello in cui la la variabile `Class` dipende da tutte le altre variabili disponibili nel dataset. Si ottenga la devianza residua del modello e si verifichi se il modello ha un qualche valore predittivo. Si verifichi inoltre se il modello ha una miglior capacità predittiva del modello in cui solo la variabile `Thick` è usara come predittore.

3. Si usi la funzione `step` per costruire un modello in cui una sottoinsieme delle variabili viene usato. Si confrontino i modelli ottenuti quando vengono usati AIC o BIC come criteri per scegliere il sottoinsieme di variabili da usare come predittori nel modello. Se vi è qualche differenza tra i due modelli identificati usando i due criteri, si indichi il motivo alla base della differenza.

4. Usando il modello selezionato usando AIC si stimi la probabilità che due pazienti con le caratteristiche indicate nel dataset `nd` abbiano un tumore benigno (cioè un tumore con `Class = 1`):

   ```
   nd <- data.frame( Patient =  c("A","B"),
       Adhes = c(1,3), BNucl = c(1,3.5), Chrom = c(3,3.5), Epith = c(2,3.5),
       Mitos = c(1,1.6), NNucl = c(1,2.8), Thick = c(4,4.43), UShap = c(1,3.2),
       USize = c(1,3.14))
   ```

   Si fornisca una stima puntuale e una stima intervallare di questa probabilità usando un livello di confidenza del 96%. Si commenti l'ampiezza degli intervalli identificati.

Esercizio 2

Gli abitanti di sesso maschile dell'isola greca di Kalythos soffrono di una malattia congenita agli occhi, i cui effetti diventano più marcati in età avanzate. Su un campione di isolani di sesso maschile e di età diverse è stato contato il numero di individui ciechi. Il codice crea un dataset per i dati osservati creando una variabile per il numero di totale di uomini campionati per ogni età e una variabile per il numero di uomini ciechi individuati nel campione:

```
Kalythos <- data.frame(age = c(20,35,45,55,70),
                       total_sample = c(50,50,50,50,50),
                       n_blind = c(6,17,26,37,44))
```

1. Si stimi un modello che indaghi se la proporzione di persone con cecità nell'isola cambia in funzione dell'età degli individui. Si usi la funzione legame (link function) canonica.

2. Si crei un grafico che mostri la relazione stimata dal modello al punto precedente: si commenti come il modello stimato si adatta ai dati raccolti

3. Si calcoli un intervallo di confidenza della probabilità che ha un individuo nell'isola di essere cieco a 20, 50 e 70 anni. Si usi un livello di confidenza pari al 90%

4. Si proceda a fare un test per testare se il valore del coefficiente relativo al predittore `age` è uguale a 0.1. Si usi un livello di significatività del 10%.

5. Si delinei brevemente la base teorica usata per derivare il test svolto nel punto precedente, commentando la validità di tale base per l'applicazione al punto 4.

6. Si usino le due variabili specificate qui sotto come predittori in un modello che indaghi come la proporzione di persone con cecità nell'isola cambia in funzione dell'età degli individui. Si confrontino i valori stimati dei coefficienti: che interpretazione si può dare alla stima dell'intercetta nei diversi modelli?

   ```
   Kalythos$age_m20 <- Kalythos$age-20
   Kalythos$age_m45 <- Kalythos$age-45
   ```

Esercizio 3

In un esperimento sullo sviluppo infantile ad alcuni bambini è stato chiesto di costruire una torre più alta possibile usando dei blocchi: in una prima iterazione ai bambini venivano dati dei cubi, mentre in una seconda iterazione erano messi a loro disposizione dei cilindri. Ogni bambino è stato sottoposto al test due volte (quindi ogni bambino ha costruito quattro torri). Informazioni sulle variabili raccolte durante l'esperimento sono reperibili al dataset `blocks` nel pacchetto R `GLMsData` (`data(blocks, package = "GLMsData")`).

1. Si produca un grafico che indaghi se vi è una relazione tra il numero di blocchi usati (`Number`) e l'età dei bambini (`Age`)

2. Si produca un grafico che indaghi se la relazione tra il numero di blocchi usati (`Number`) e l'età dei bambini (`Age`) è diversa a seconda se i blocchi dati ai bambini sono cubi o cilindri (`Shape`)

3. Si usi un modello GLM con funzione di legame (link function) canonica per stimare un modello in cui il numero di blocchi usato dipende dall'età del bambino. Come si possono interpretare i valori dei coefficienti di regressione stimati. Si costruisca un intervallo di confidenza al livello di confidenza di 98% per il coefficiente relativo a `Age`.

4. Si crei un intervallo di confidenza per il numero di blocchi usati da un bambino di 4 e 9 anni: si commenti l'affidabilità di questi intervalli

5. Si usi un modello GLM con funzione di legame canonica per stimare un modello in cui il numero di blocchi usato dipende dall'età del bambino (`Age`) e dal tipo di blocco a disposizione (`Shape`): si verifichi se è necessario inserire un'interazione tra le due variabili nel modello

6. Si crei un grafico che mostri le stime ottenute per il valore atteso della variabile risposta usando il modello scelto al punto 5

7. Si esplicitino le assunzioni alla base del modello stimato al punto 5. Si commenti la validità delle assunzioni per il dataset in esame.

Esercizio 4

[Tratto da Salvan et al. *Modelli Lineari Generalizzati*, Springer]

Si prenda in esame il dataset contenuto nel file `bchem_phd.csv`. I dati contenuti nel data frame Biochemists (Long, 1990; Jackman, 2017) sono stati raccolti considerando dottori di ricerca in Biochimica che hanno conseguito il titolo nel periodo 1950-1967 in università degli Stati Uniti. Scopo dell'analisi era valutare le differenze di genere nella produttività scientifica. La variabile risposta è il numero di articoli scientifici, `art`, su riviste censite da Chemical Abstracts pubblicati nei 3 anni a cavallo del conseguimento del titolo. Le variabili concomitanti disponibili sono genere, `fem` (`Men`, `Women`), lo stato civile, `mar` (`Married`, `Single`), il numero di figli con non più di 5 anni, `kid5`, un indice di prestigio scientifico del dipartimento, `phd` (con valori tra 0 e 5), il numero di articoli scientifici pubblicati dal supervisore, `ment`, negli stessi 3 anni a cui è riferita la variabile `art`.

1. Si esamini la variabile risposta `art`.

2. Sarebbe possibile usare un modello di regressione multiplo per modellare la variabile risposta `art`? Quali vantaggi o svantaggi comporterebbe usare un modello di regressione multiplo per modellare la variabile risposta `art`?

3. Che modifiche si potrebbero apportare ad un modello di regressione multiplo per superare alcuni dei possibili svantaggi identificati al punto 2

4. Si usi un modello lineare generalizzato (GLM) in cui si assume che la variabile risposta `art` segua una distribuzione di Poisson per indagare se il numero di articoli pubblicati è influenzato dall'indice di prestigio scientifico del dipartimento, `phd`

5. Si trovi un sottoinsieme di predittori ottimali da usare in un modello lineare generalizzato simile a quello stimato al punto 3

6. Si produca una stima puntuale del valore atteso del numero di articoli per quattro persone con un dottorato in Biochimica con le seguenti caratteristiche:

```
nd <- data.frame(
    fem = c("Men","Women","Men","Women"), mar = c("Married","Married","Single","Single"),
    kid5=c(1,1,1,1), phh = c(3,3,3,3), ment = c(8,8,8,8))
rownames(nd) <- c("PHD A", "PHD B","PHD C","PHD D")
```

   Si commentino le stime trovate, esplicitando come i valori stimati dei coefficienti di regressione influiscono sui valori stimati

7. Si produca una stima intervallare usando un livello di confidenza del 90% per il valore atteso di articoli pubblicati dalle persone specificate al punto precedente

Esercizio 5

[Esercizio di Esame aa 2018/2019 - prof. Gaetan]

Si consideri una ricerca sul comportamento dei clienti di un negozio online e sul rapporto tra vendite e apprezzamento del sito Web. A un certo numero di visitatori del sito Web è stato chiesto di esprimere il proprio apprezzamento per il sito Web su una scala Likert a 5 punti, che va da 1 (pessimo) a 5 (ottimo). Per questi visitatori è stato anche registrato se hanno effettivamente acquistato qualcosa sul sito web. I dati sono contenuti nel *file* `online.txt`.

1. Di che tipo sono le variabili coinvolte?
2. Si consideri un visitatore del sito Web e si supponga di non avere informazioni su quanto questo cliente apprezza il sito Web. Si stimi la probabilità che questo cliente abbia effettivamente acquistato qualcosa.
3. Si stimi la probabilità di acquistare qualcosa separatamente per ogni livello di apprezzamento e si mostri in un grafico le probabilità stimate in funzione dell'apprezzamento.
4. Si espliciti perchè non è sensato specificare un modello lineare per stabilire la relazione tra le due variabili.
5. Si stimi un opportuno modello di regressione logistica.
6. E' vero che la variabile *apprezzamento del sito* ha un effetto sulla probabilità d'acquisto? Se si, qual è quest'effetto? Quanto è forte l'evidenza a supporto della vostra affermazione?

Esercizio 6

Il dataset `EdenRainfall` contiene informazioni sulle piogge registrate nel bacino del fioume Eden, nel nord dell'Inghilterra. Il dataset contiene le seguenti variabili:

- month: il mese a cui si riferisce l'osservazione
- year: l'anno a cui si riferisce l'osservazione
- ndays_prec: il numero di giorni con precipitazione $>$ 1mm nel mese
- high_prec: una variabile indicatore che ha valore 1 se nel mese è stata registrata una precipitazione estrema
- tot_prec: precipitazione totale accummulata nel mese
- mean_prec: precipitazione media accummulata nel mese
- tdays: il numero totale di giorni con records validi per il mese di riferimento
- nao: il valore del north atlantic oscillation (NAO) index per il mese.

- soi: il valore del southern oscillation index (SOI) per il mese.

Si indaghi se le variabili *nao* e *soi* influenzano la probabilità di osservare almeno un giorno con un elevata precipitazione (`high_precip`) e la proporzione di giorni piovosi in un mese per il mese di Gennaio. Si valutino in prima istanza modelli in cui i predittori vengono usati singolarmente: si creino grafici che mostrano l'impatto stimato dei singoli predittori sulla variabile di interesse. Si stimi un modello in cui i predittori sono entrambi inseriti nel modello. Si creino grafici che mostrano l'impatto dei predittori per diversi valori dell'altro predittore (ad esempio, 10o percentile, mediana e 90o percentile) sulla variabile di interesse.

Esercizio 7

Si prenda in esame la funzione `pois_gen_and_est` specificata nel seguente codice R:

```
pois_gen_and_est <- function(n, xrange = list(c(0,1)),
                             beta_true, out_est=FALSE){
   n_x <- length(xrange)
   X <- rep(1,n)
   for(j in 1:n_x) X <- cbind(X,
           runif(n, xrange[[j]][1], xrange[[j]][2]))
   y <- rpois(n, exp(X %*% beta_true))
   X <- X[,-1]
   if(!out_est) out <- data.frame(X,y)
   if(out_est) out <- as.numeric(coef(glm(y~X, family = poisson)))
   out
 }
```

1. Come vengono specificate le variabili risposta e i predittori? Che distribuzione ha la variabile risposta? Che distribuzione hanno i predittori? Si scriva in maniera estesa il modello sottostante la generazione dei dati nella funzione.

2. Si spieghi il contenuto dell'oggetto `pois_sim_n10` creato con il seguente codice:

   ```
   NSIM <- 1000
   set.seed(15496)
   pois_sim_n10 <- t(replicate(NSIM,
               pois_gen_and_est(n=10, xrange = list(c(0,1), c(5,6)),
                                beta_true = c(1.2,1,0.6), out_est = TRUE)))
   ```

3. Si usi `pois_sim_n10` per quantificare lo standard error degli stimatori dei coefficienti di regressione in un GLM

4. Si usi la funzione `pois_gen_and_est` per studiare come lo standard error degli stimatori dei coefficienti di regressione in un GLM varia in funzione della dimensione del campione

5. Si usi la funzione `pois_gen_and_est` per studiare come lo standard error degli stimatori dei coefficienti di regressione in un GLM varia in funzione del vero valore dei parametri di regressione (nota bene: si consiglia di usare un modello con un solo predittore)

6. Si verifichi che quanto osservato al punto 5 sia in accordo con i risultati teorici presentati nelle slides
7. Si crei una funzione `binom_gen_and_est` per indagare il comportamento degli stimatori in un modello GLM per dati di tipo binomiale o bernoulliano.
8. Si usi la funzione `binom_gen_and_est` per indagare l'effetto di numerosità campionaria e vero valore dei coefficienti di regressione sull'incertezza degli stimatori