Logistična regresija

Predpostavljamo, da imamo m neodvisnih spremenljivk $X_1, \ldots, X_m$ in odvisno spremenljivko Y, ki ima lahko samo vrednosti 0 in 1. Zapišimo $\mathbf{X} = (X_1, \ldots, X_m)$. Predpostavljamo, da velja naslednji posplošen regresijski model, ki je znan pod imenov *logistični regresijski model*:

$$P(Y = 1|\mathbf{X}) = \frac{\exp(\beta_0 + \sum_{k=1}^{n} \beta_k X_k)}{1 + \exp(\beta_0 + \sum_{k=1}^{n} \beta_k X_k)},$$

kjer so $\beta_0, \beta_1, \ldots, \beta_m$ parametri modela.

V datoteki `logistic.dat` so podatki za izpis iz S-PLUS, ki je vključen.

```
*** Import Data ***

Import Successful
File name:   /valjhun/mihael/teaching/istat/verstat/data/logit.dat
Data name:   logit

Number of rows:   1000
Number of columns:   3

Columns:
  Name    Type
1 Col1 numeric
2 Col2 numeric
3 Col3 numeric

*** Generalized Linear Model ***

Call: glm(formula = Col1 ~ Col2 + Col3,
          + family = binomial(link = logit),
          + data = logit, na.action = na.exclude,
          + control = list(epsilon = 0.0001, maxit = 50, trace = F))

Deviance Residuals:
       Min        1Q    Median        3Q       Max
```

```
 -2.061219 -1.047069 0.5614318 1.049224 2.122089

Coefficients:
                 Value Std. Error   t value
(Intercept) 0.05832206 0.06777698 0.8604996
       Col2 0.41705971 0.08130599 5.1295077
       Col3 0.56273870 0.07989416 7.0435527

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 1385.81 on 999 degrees of freedom

Residual Deviance: 1252.243 on 997 degrees of freedom

Number of Fisher Scoring Iterations: 3

Correlation of Coefficients:
     (Intercept)       Col2
Col2  0.0623449
Col3 -0.0199886  -0.3277031

Analysis of Deviance Table

Binomial model

Response: Col1

Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev
NULL                   999   1385.810
Col2  1 79.92338       998   1305.887
Col3  1 53.64399       997   1252.243
```

Odgovorite na naslednja vprašanja:

a. Za dane podatke $y_1, y_2, \ldots, y_n$ in $\mathbf{x}_1, \ldots, \mathbf{x}_n$ zapišite (pogojno) funkcijo verjetja.

b. Prepričajte se, da so parametri v izpisu ocenjeni po metodi največjega verjetja.

c. Zapišite analitični izraz za Fisherjevo matriko verjetja.

d. Prepričajte se, da so standardne napake na izpisu ocenjene tako, da izračunamo inverz Fisherjeve matrike informacije.

e. Kako bi (pogojno) preizkusili domnevo $H_0\colon \beta_1 = \beta_2 = 0$? Na kaj bi se sklicali?

f. V načelu moramo upoštevati tudi, da so $\mathbf{x}_1, \ldots, \mathbf{x}_n$ tudi vzorec iz neke populacije. Privzemite, do so $\mathbf{X}_1, \ldots, \mathbf{X}_n$ neodvisne dvorazsežne normalne z $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = \sigma_2^2 = 1$ in nekim $\rho \in (-1, 1)$. Kako bi ocenili brezpogojno standardno napako cenilk za parametre β, če poznate ρ?

g. Kako bi ocenili brezpogojno standardno napako, če ne poznate parametra ρ? S simulacijo se prepričajte za konkreten primer, da so vaše ocene prave. Komentirajte, zakaj je bolj "pošteno" navajati brezpogojne standardne napake?

h. Kako bi brezpogojno preizkusili domnevo $H_0\colon \beta_1 = \beta_2 = 0$ v primeru, ko ρ poznate?