78185 Yleistetyt lineaariset mallit

Harjoitus 3, syksy 2014

1. Vuosina 1980-81 toteutettiin postikyselynä Oulun ja Lapin lääneissä terveys- ja elämäntapatiedustelu koskien yhteensä 10,874 neljäntoista vuoden iässä olevaa lasta. Tutkimuksen yhteydessä kysyttiin mm. alkoholinkäyttöä. Tässä yhteydessä todettiin eri ryhmissä säännöllisiä alkoholin käyttäjiä seuraavasti:

Sukupuoli	Perhesuhde	Lapsien lkm	N	N1	N2	N3
1	1	1	4291	1816	1553	922
1	1	2	183	83	63	37
1	2	1	900	306	315	279
1	2	2	91	29	30	32
2	1	1	4158	1846	1371	941
2	1	2	187	81	70	36
2	2	1	967	315	315	337
2	2	2	97	16	43	38

N = Lapsien lukumäärä

N1 = Niiden lasten lukumäärät, jotka eivät ole lainkaan kokeilleet alkoholia

N2 = Niiden lasten lukumäärät, jotka ovat kokeilleet alkoholia

N3 = Niiden lasten lukumäärät, jotka käyttävät säännöllisesti alkoholia

Sukupuoli: 1=poika, 2=tyttö

Perhesuhde: 1=normaali, 2=vajaaperhe

Lapsien lkm: 1=perheessä useita lapsia, 2=lapsi ainoa

Aineistoon on sovitettu kahta mallia (Liite). Selitä, mistä mallituksissa on kysymys. Mitä johtopäätöksiä voit tehdä tulosten perusteella?

Vihje: Logistinen regressioanalyysi voidaan suorittaa R:llä esim. seuraavan esimerkin mukaisesti:

```
glm(cbind(ncases, ncontrols) ~ agegp + tobgp * alcgp,
    data = esoph, family = binomial())
```

jossa ncases on vektori, joka sisältää tapausten lukumäärät eri profiileilla ja ncontrols on vektori, joka sisältää verrokkien lukumäärät eri profiileilla. Selittävinä muuttujina ovat agegp, tobgp ja niiden yhdysvaikutus.

2. Tarkastellaan jälleen aineistoa `lapset85.dat`. Tutki logistisen regressioanalyysin avulla alipainoisen lapsen (< 2500 g) synnnyttämisen riskiä tupakoivilla äideillä. Huomaa, että muuttujan SYNTPAIN mittayksikkönä on 10 g. Mieti mitkä muuttujat ovat mahdollisia sekoittavia tekijöitä. Onko muuttujien välillä yhdysvaikutusta? Vertaa malleja devianssien avulla. Laske ristitulosuhteille likimääräiset 95% luottamusvälit.

   Vihjeitä:

```
#Havaintoaineiston muuttujiin voi viitata suoraan muuttujien nimilla
attach(lapset85)

#y<-(SYNTPAIN<250)

#parity=TRUE, kun aidilla aikaisempia raskauksia, muulloin FALSE.
parity<-(PARITEET>0)

#Tehdaan muuttujasta AIDINTUP faktorimuuttuja aidtup
aidtup<-factor(AIDINTUP)

# y on 0/1-muuttuja, kaytetaan logit-linkkifunktiota
# Selittavat muuttujat x1 ja x2
# Havaintoaineisto on "tied"
> mod1<-glm(y~x1+x2, family = binomial(), data = tied)

#Lisataan yhdysvaikutus
> mod2<-glm(y~x1*x2, family = binomial(), data = tied)

#Sama malli kuin mod2
> mod3<-glm(y~x1+x2+x1*x2, family = binomial(), data = tied)

> summary(mod1)
> anova(mod1,mod2,test="Chisq")  #Vertaillaan sis. malleja

> confint(mod1)      #Luott.valit log(OR):lle
> exp(confint(mod1)) #Luott.valit OR:lle
```

3. Tutki havaintoaineistoa ESOPH (saadaan käyttöön antamalla komento data(esoph) ja tietoa aineistosta saa komennolla help(esoph)) logistisen regressioanalyysin avulla. Havaintoaineisto koostuu ruokatorvensyöpään sairastuneista ja kontrollitapauksista. Tutki syöpään sairastumisen vaaraa, kun selittävinä tekijöinä ovat tupakointi ja alkoholinkäyttö.

Liite

```
> alko<-read.table("alkoholi.dat",header=TRUE)
> alko$sex<-factor(alko$sex)
> alko$pesu<-factor(alko$pesu)
> alko$ainoa<-factor(alko$ainoa)
> alko
  sex pesu ainoa    n   n1   n2  n3
1   1    1     1 4291 1816 1553 922
2   1    1     2  183   83   63  37
3   1    2     1  900  306  315 279
4   1    2     2   91   29   30  32
5   2    1     1 4158 1846 1371 941
6   2    1     2  187   81   70  36
7   2    2     1  967  315  315 337
8   2    2     2   97   16   43  38
> attach(alko)
> y1<-n2+n3
> y1
[1] 2475  100  594   62 2312  106  652   81
> y2<-n3
> y2
[1] 922  37 279  32 941  36 337  38
> n-y1
[1] 1816   83  306   29 1846   81  315   16
> model1<-glm(cbind(y1,n-y1)~sex+pesu+ainoa,family=binomial())
> summary(model1)

Call:
glm(formula = cbind(y1, n - y1) ~ sex + pesu + ainoa, family = binomial())

Deviance Residuals:
      1        2        3        4        5        6        7        8
 0.8510  -1.3304  -1.2039  -0.3973  -0.4805  -0.4899   0.3255   3.2468

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.28333    0.02913   9.726   <2e-16 ***
sex2        -0.04323    0.03911  -1.105    0.269
pesu2        0.46507    0.05199   8.945   <2e-16 ***
ainoa2       0.10132    0.09051   1.119    0.263
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 102.30  on 7  degrees of freedom
Residual deviance:  15.22  on 4  degrees of freedom
AIC: 75.719

Number of Fisher Scoring iterations: 4

> y1-y2
[1] 1553   63  315   30 1371   70  315   43
> model2<-glm(cbind(y2,y1-y2)~sex+pesu+ainoa,family=binomial())
> summary(model2)

Call:
glm(formula = cbind(y2, y1 - y2) ~ sex + pesu + ainoa, family = binomial())

Deviance Residuals:
      1        2        3        4        5        6        7        8
-0.0345   0.4255  -0.4197   0.9923   0.1371  -0.8959   0.2115  -0.3317

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.51996    0.03866 -13.448  < 2e-16 ***
sex2         0.13781    0.05115   2.694  0.00706 **
pesu2        0.43309    0.06150   7.043 1.89e-12 ***
ainoa2      -0.10075    0.11341  -0.888  0.37436
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 60.3320  on 7  degrees of freedom
Residual deviance:  2.3192  on 4  degrees of freedom
AIC: 59.885

Number of Fisher Scoring iterations: 3
```