78185 Yleistetyt lineaariset mallit

Harjoitus 3, syksy 2014

Esimerkkiratkaisut

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:

Suku-puoli	Perhe-suhde	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.

Ratkaisu:

Luetaan aineisto R-ohjelmistoon:

```
> alko<-read.table("alkoholi.dat",header=TRUE)
```

Tehdään muuttujista `sex`, `pesu` ja `ainoa` faktorimuuttujia:

```
> alko$sex<-factor(alko$sex)
> alko$pesu<-factor(alko$pesu)
> alko$ainoa<-factor(alko$ainoa)
```

Data frame -tyyppinen aineisto `alko`:

```
> 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
```

Antamalla `attach`-komento, aineiston muuttujiin voi viitata suoraan muuttujien nimillä. Muussa tapauksessa muuttujiin pitäisi viitata antamalla ennen muuttujan nimeä aineiston nimi ja dollari-merkki (esim. `alko$sex`).

```
> attach(alko)
```

Muodostetaan muuttuja `y1`, jossa on alkoholia kokeilleiden tai säännöllisesti käyttävien lasten lukumäärä:

```
> y1<-n2+n3
> y1
[1] 2475  100  594   62 2312  106  652   81
```

Muodostetaan muuttuja `y2`, jossa on alkoholia säännöllisesti käyttävien lasten lukumäärä:

```
> y2<-n3
> y2
[1] 922  37 279  32 941  36 337  38
```

Niiden lasten lukumäärät, jotka eivät ole lainkaan kokeilleet alkoholia (N1):

```
> n-y1
[1] 1816   83  306   29 1846   81  315   16
```

Sovitetaan logistinen regressiomalli, jossa vastemuuttujana on "lapsi on kokeillut tai käyttää säännöllisesti alkoholia". Selittäjinä ovat dikotomiset muuttujat `sex`, `pesu` ja `ainoa`. Argumentilla `family=binomial()` ilmoitetaan, että vaste on binomijakautunut ja linkkifunktiona on logit-funktio. Argumenttia `family=binomial(link = "log")` käyttämällä sovitettaisiin regressiomalli, jossa vaste on binomijakautunut ja linkkifunktiona luonnollinen logaritmi:

```
> model1<-glm(cbind(y1,n-y1)~sex+pesu+ainoa,family=binomial())
```

Tulostetaan sovitettuun logistiseen regressiomalliin liittyviä yhteenvetotietoja:

```
> 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
```

Logistisen regressiomallin regressiokerroinvektorin estimaatti on

$$\hat{\boldsymbol{\beta}} = (0.28333, -0.04323, 0.46507, 0.10132)'.$$

- $\exp(0.28333) = 1.3$ on "perusodds"eli vedonlyöntisuhde tilanteessa `sex`=1 (lapsi on poika), `pesu`=1 (normaali perhesuhde) ja `ainoa`=1 (perheessä useita lapsia).
- $\exp(-0.04323) = 0.96$ kertoo monikokertainen vedonlyöntisuhde (odds) alkoholinkäyttämiseen on tytöillä verrattuna poikiin. Regressiokerroin ei vaikuttaisi kuitenkaan olevan tilastollisesti merkitsevä (p-arvo=0.269).
- $\exp(0.46507) = 1.59$ kertoo monikokertainen vedonlyöntisuhde (odds) alkoholinkäyttämiseen on vajaaperheisillä verrattuna normaaliperheisiin. Regressiokerroin on tilastollisesti merkitsevä (p-arvo$< 2 \cdot 10^{-16}$).
- $\exp(0.10132) = 1.10$ kertoo monikokertainen vedonlyöntisuhde (odds) alkoholinkäyttämiseen on yksilapsisilla perheillä verrattuna monilapsisiin perheisiin. Regressiokerroin on ei ole tilastollisesti merkitsevä (p-arvo=0.263).

Devianssi jaettuna vapausasteilla on

$$\frac{D}{r-p} = \frac{15.22}{4} = 3.805.$$

Kyseinen osamäärä on selvästi yli ykkösen, joten aineistossa on ylihajontaa (ks. luentomonisteesta ylihajonnan mahdollisista syistä). Jos aineistossa on ylihajontaa, niin regressiokertoimien estimaattien keskivirheet on arvioitu liian pieniksi ja sen myötä p-arvoihin ei voi täysin luottaa. Yksi tapa tilanteen korjaamiseen on keskivirheiden kertominen luvulla $\sqrt{D/(r-p)}$. Tämän aineiston tapauksessa korjauskerroin on $\sqrt{3.805} = 1.95$.

Seuraavassa sovitetaan logistinen regressiomalli, jossa vasteena on "lapsi on käyttänyt säännöllisesti alkoholia". Vertailuluokkana on "lapsi on kokeillut alkoholia".

```
> 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
```

- $\exp(-0.51996) = 0.6$ on "perusodds"eli vedonlyöntisuhde tilanteessa `sex`=1 (lapsi on poika), `pesu`=1 (normaali perhesuhde) ja `ainoa`=1 (perheessä useita lapsia).
- $\exp(0.13781) = 1.2$ kertoo monikokertainen vedonlyöntisuhde (odds) säännölliseen alkoholinkäyttämiseen on tytöillä verrattuna poikiin. Regressiokerroin on tilastollisesti merkitsevä (p-arvo=0.007).
- $\exp(0.43309) = 1.5$ kertoo monikokertainen vedonlyöntisuhde (odds) säännölliseen alkoholinkäyttämiseen on vajaaperheisillä verrattuna normaaliperheisiin. Regressiokerroin on tilastollisesti merkitsevä (p-arvo$< 1.89 \cdot 10^{-12}$).
- $\exp(-0.10075) = 0.9$ kertoo monikokertainen vedonlyöntisuhde (odds) säännölliseen alkoholinkäyttämiseen on yksilapsisilla perheillä verrattuna monilapsisiin perheisiin. Regressiokerroin on ei ole tilastollisesti merkitsevä (p-arvo=0.37436).

Devianssi jaettuna vapausasteilla on

$$\frac{D}{r-p} = \frac{2.3192}{4} = 0.5798.$$

Kyseinen osamäärä on alle ykkösen, joten aineistossa ei ole ylihajontaa. Jos osamäärä on selvästi alle ykkösen, niin aineistossa on *alihajontaa*. Alihajontaa saattaa esiintyä esimerkiksi silloin, jos havainnot ovat korreloituneita.

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.

   Ratkaisu:

```
> attach(lapset85)
> y<-SYNTPAIN<250
> table(y)
y
FALSE  TRUE
  829    17
> aidtup<-factor(AIDINTUP)
> lapset85<-lapset85[(KAKSOSTU==1)&(PERINKUO==0),]
> parity<-(PARITEET>0)
```

   Sovitetaan logistinen regressiomalli, jossa vastemuuttujana on "lapsi alipainoinen". Selittäjänä muuttuja `aidtup`:

```
> mod1<-glm(y~aidtup,family=binomial)
> summary(mod1)

Call:
glm(formula = y ~ aidtup, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3094  -0.1836  -0.1836  -0.1836   2.9646

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.0745     0.3041 -13.400   <2e-16 ***
aidtup1      -0.3076     1.0512  -0.293   0.7698
aidtup2       1.0589     0.5498   1.926   0.0541 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 166.5  on 845  degrees of freedom
Residual deviance: 163.0  on 843  degrees of freedom
```

```
AIC: 169

Number of Fisher Scoring iterations: 7
> coefficients(mod1)
(Intercept)     aidtup1     aidtup2
 -4.0744510  -0.3075756   1.0589161
> exp(coefficients(mod1))
(Intercept)     aidtup1     aidtup2
 0.01700155  0.73522727  2.88324421
> exp(confint(mod1))
Waiting for profiling to be done...
                 2.5 %     97.5 %
(Intercept) 0.00879179 0.02929136
aidtup1     0.03995910 3.85452001
aidtup2     0.89329340 8.10350376
> exp(confint.default(mod1))  #Normaalijak.approksimaatio
                  2.5 %     97.5 %
(Intercept) 0.009368477 0.03085374
aidtup1     0.093684243 5.77001134
aidtup2     0.981538675 8.46945450
```

Paljon tupakoivilla äideillä vaikuttaisi olevan hieman kohonnut riski saada matalapainoinen lapsi verrattuna tupakoimattomiin äiteihin ($\widehat{OR} = 2.9$, p-arvo: 0.0541). Normaalijakauma-approksimaatioon perustuva likimääräinen 95% luottamusväli on $(1.0, 8.5)$.

Lisätään logistisen regressiomallin selittäjäksi dikotominen muuttuja `parity` (pariteetti):

```
> mod2<-glm(y~aidtup+parity,family=binomial)
> summary(mod2)

Call:
glm(formula = y ~ aidtup + parity, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3378  -0.2023  -0.1740  -0.1740   3.0080

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.8791     0.4306  -9.009   <2e-16 ***
aidtup1      -0.3301     1.0520  -0.314    0.754
aidtup2       1.0438     0.5506   1.896    0.058 .
```

```
parityTRUE   -0.3038     0.5004  -0.607    0.544
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 166.50  on 845  degrees of freedom
Residual deviance: 162.64  on 842  degrees of freedom
AIC: 170.64

Number of Fisher Scoring iterations: 7
```

Muuttujan `parity` lisääminen selittäjäksi ei selvästikään parantanut mallia tilastollisesti merkitsevästi.

Vaihdetaan muuttujan `parity` tilalle jatkuva muuttuja `AIDINIKA` (äidin ikä):

```
> mod3<-glm(y~aidtup+AIDINIKA,family=binomial)
> summary(mod3)

Call:
glm(formula = y ~ aidtup + AIDINIKA, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4218  -0.1996  -0.1805  -0.1666   2.9814

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.24255    1.31083  -3.999 6.35e-05 ***
aidtup1     -0.15309    1.06577  -0.144   0.8858
aidtup2      1.12629    0.55524   2.028   0.0425 *
AIDINIKA     0.04050    0.04327   0.936   0.3493
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 166.50  on 845  degrees of freedom
Residual deviance: 162.14  on 842  degrees of freedom
AIC: 170.14

Number of Fisher Scoring iterations: 7
```

```
> exp(coefficients(mod3))
(Intercept)     aidtup1     aidtup2    AIDINIKA
0.005286757 0.858054112 3.084196484 1.041334853
> exp(confint.default(mod3))
                  2.5 %    97.5 %
(Intercept) 0.0004049558 0.0690194
aidtup1     0.1062500509 6.9294730
aidtup2     1.0387748977 9.1571985
AIDINIKA    0.9566609179 1.1335033
```

Muuttujan `AIDINIKA` lisääminen malliin, pienensi hiukan muuttujan `aidtup2` regressiokertoimeen liittyvää p-arvoa. Paljon tupakoivilla äideillä vaikuttaisi olevan hieman kohonnut riski saada matalapainoinen lapsi verrattuna tupakoimattomiin äiteihin ($\widehat{OR} = 3.1$, p-arvo: 0.0425). Normaalijakauma-approksimaatioon perustuva likimääräinen 95% luottamusväli on $(1.0, 9.2)$.

Sovitetaan seuraavaksi malli, jossa on selittäjinä sekä `AIDINIKA` että `parity`:

```
> mod4<-glm(y~aidtup+AIDINIKA+parity,family=binomial)
> summary(mod4)

Call:
glm(formula = y ~ aidtup + AIDINIKA + parity, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5075  -0.2128  -0.1757  -0.1566   3.0549

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.53465    1.31677  -4.203 2.63e-05 ***
aidtup1     -0.09648    1.06904  -0.090   0.9281
aidtup2      1.13415    0.55601   2.040   0.0414 *
AIDINIKA     0.06529    0.04776   1.367   0.1716
parityTRUE  -0.65806    0.56884  -1.157   0.2473
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 166.50  on 845  degrees of freedom
Residual deviance: 160.84  on 841  degrees of freedom
AIC: 170.84
```

```
Number of Fisher Scoring iterations: 7
```

Malli ei muuttunut mitenkään merkittävästi. Testataan seuraavaksi onko muuttujilla `AIDINIKA` ja `parity` yhdysvaikutusta:

```
> mod5<-glm(y~aidtup+AIDINIKA*parity,family=binomial)
> summary(mod5)

Call:
glm(formula = y ~ aidtup + AIDINIKA * parity, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5248  -0.2016  -0.1913  -0.1512   3.1263

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -4.099288   2.130996  -1.924   0.0544 .
aidtup1             -0.130629   1.073351  -0.122   0.9031
aidtup2              1.133209   0.558223   2.030   0.0424 *
AIDINIKA             0.007366   0.085256   0.086   0.9311
parityTRUE          -3.014163   2.789055  -1.081   0.2798
AIDINIKA:parityTRUE  0.087219   0.102166   0.854   0.3933
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 166.50  on 845  degrees of freedom
Residual deviance: 160.08  on 840  degrees of freedom
AIC: 172.08

Number of Fisher Scoring iterations: 7

> anova(mod4,mod5,test="Chisq")
Analysis of Deviance Table

Model 1: y ~ aidtup + AIDINIKA + parity
Model 2: y ~ aidtup + AIDINIKA * parity
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       841     160.84
2       840     160.08  1  0.75418   0.3852
```

Muuttujilla `AIDINIKA` ja `parity` ei vaikuttaisi olevan uskottavuusosamäärän testin perusteella yhdysvaikutusta (p-arvo=0.3852).

Devianssi jaettuna vapausasteilla mallissa mod4,

$$\frac{D}{r-p} = \frac{160.84}{841} \approx 0.19,$$

on pienempi kuin yksi, joten mallissa vaikuttaisi olevan *alihajontaa*. Yksi mahdollinen ratkaisuna tähän voisi olla regressiokertoimien keskivirhe-estimaattinen kertominen luvulla $\sqrt{D/(r-p)}$. Toinen tapa on estimoida mallin parametrit ja keskivirheet quasi-uskottavuutta käyttäen (ks. esim McCullagh & Nelder). Tällöin malliin voi ottaa mukaan myös hajontaparametrin (Dispersion parameter.

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.

   Ratkaisu:

```
> data(esoph)
> summary(esoph)
```

   Muutetaan muuttujat tavallisiksi faktorimuuttujiksi:

```
> esoph$alcgp<-factor(esoph$alcgp,ordered=FALSE)
> esoph$tobgp<-factor(esoph$tobgp,ordered=FALSE)
> esoph$agegp<-factor(esoph$agegp,ordered=FALSE)
```

   Sovitetaan logistinen regressiomalli, jossa selittäjinä ovat faktorimuuttujat `alcgp` ja `tobgp`:

```
> mod1<-glm(cbind(ncases,ncontrols)~alcgp+tobgp,family=binomial,data=esoph)
> summary(mod1)

Call:
glm(formula = cbind(ncases, ncontrols) ~ alcgp + tobgp, family = binomial,
    data = esoph)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3326  -1.0436  -0.0775   0.6104   3.3347

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.8294     0.2078 -13.618  < 2e-16 ***
alcgp40-79    1.0688     0.2316   4.615 3.92e-06 ***
alcgp80-119   1.6086     0.2547   6.315 2.70e-10 ***
alcgp120+     2.1521     0.2755   7.810 5.70e-15 ***
tobgp10-19    0.3065     0.1983   1.545   0.1223
tobgp20-29    0.3424     0.2383   1.437   0.1508
tobgp30+      0.6515     0.2572   2.533   0.0113 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

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

```
    Null deviance: 227.24  on 87  degrees of freedom
Residual deviance: 131.48  on 81  degrees of freedom
AIC: 292.96

Number of Fisher Scoring iterations: 5
```

Lasketaan OR-estimaatit ja likimääräiset 95% luottamusvälit:

```
> exp(coefficients(mod1))
(Intercept)  alcgp40-79 alcgp80-119   alcgp120+  tobgp10-19  tobgp20-29
 0.05904545  2.91193723  4.99564088  8.60325338  1.35860539  1.40836342
   tobgp30+
 1.91834189
> exp(confint.default(mod1))
                 2.5 %     97.5 %
(Intercept) 0.03929411  0.0887249
alcgp40-79  1.84954133  4.5845845
alcgp80-119 3.03237483  8.2299944
alcgp120+   5.01323366 14.7641171
tobgp10-19  0.92104558  2.0040361
tobgp20-29  0.88276480  2.2469037
tobgp30+    1.15881334  3.1756932
```

Tarkistetaan onko mallissa ylihajontaa:

```
> 131.48/81
[1] 1.62321
> sqrt(1.62321)
[1] 1.274053
```

Mallissa on hyvin pientä ylihajontaa.

Testataan seuraavaksi onko alkoholinkäytöllä ja tupakoinnilla yhdysvaikutusta:

```
> mod2<-glm(cbind(ncases,ncontrols)~alcgp*tobgp,family=binomial,data=esoph)
> summary(mod2)
> anova(mod1,mod2,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(ncases, ncontrols) ~ alcgp + tobgp
Model 2: cbind(ncases, ncontrols) ~ alcgp * tobgp
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        81     131.48
2        72     123.95  9   7.5342   0.5817
```

Yhdysvaikutus ei ole tilastollisesti merkitsevä (p-arvo=0.5817).

Lisätään malliin vielä muuttuja agegp (ikäryhmä):

```
> mod3<-glm(cbind(ncases,ncontrols)~agegp+alcgp+tobgp,
+ family=binomial,data=esoph)
> anova(mod1,mod3,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(ncases, ncontrols) ~ alcgp + tobgp
Model 2: cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        81    131.484
2        76     53.973  5   77.511 2.782e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Muuttuja agegp on selvästi tilastollisesti merkitsevä (p-arvo=$2.782 \cdot 10^{-15}$).

Tulostetaan yhteenvetotuloksia sovitetusta mallista:

```
> summary(mod3)

Call:
glm(formula = cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp,
    family = binomial, data = esoph)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6891  -0.5618  -0.2168   0.2314   2.0642

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9108     1.0302  -5.737 9.61e-09 ***
agegp35-44    1.6095     1.0676   1.508 0.131652
agegp45-54    2.9752     1.0242   2.905 0.003675 **
agegp55-64    3.3584     1.0198   3.293 0.000991 ***
agegp65-74    3.7270     1.0253   3.635 0.000278 ***
agegp75+      3.6818     1.0645   3.459 0.000543 ***
alcgp40-79    1.1216     0.2384   4.704 2.55e-06 ***
alcgp80-119   1.4471     0.2628   5.506 3.68e-08 ***
alcgp120+     2.1154     0.2876   7.356 1.90e-13 ***
tobgp10-19    0.3407     0.2054   1.659 0.097159 .
tobgp20-29    0.3962     0.2456   1.613 0.106708
```

```
tobgp30+      0.8677     0.2765   3.138 0.001701 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 227.241  on 87  degrees of freedom
Residual deviance:  53.973  on 76  degrees of freedom
AIC: 225.45

Number of Fisher Scoring iterations: 6
```

Devianssi jaettuna vapausasteilla on

$$\frac{D}{r-p} = \frac{53.973}{76} \approx 0.71,$$

joten mallissa on hiukan alihajontaa. Lasketaan seuraavaksi ristitulosuhteiden estimaatit ja likimääräiset 95% luottamusvälit:

```
> exp(coefficients(mod3))
 (Intercept)   agegp35-44   agegp45-54   agegp55-64   agegp65-74     agegp75+
 0.002710046  5.000426461 19.592860766 28.741838956 41.554820823 39.716132031
  alcgp40-79  alcgp80-119    alcgp120+   tobgp10-19   tobgp20-29     tobgp30+
 3.069638047  4.250811157  8.292938857  1.405982889  1.486221090  2.381435327
> exp(confint.default(mod3))
                   2.5 %       97.5 %
(Intercept) 0.0003597915   0.02041279
agegp35-44  0.6169797158  40.52688306
agegp45-54  2.6318576780 145.85902430
agegp55-64  3.8942181518 212.13328950
agegp65-74  5.5708314582 309.97224501
agegp75+    4.9303452395 319.93117457
alcgp40-79  1.9236882114   4.89823542
alcgp80-119 2.5394302540   7.11553131
alcgp120+   4.7197039656  14.57142977
tobgp10-19  0.9400092121   2.10294523
tobgp20-29  0.9183446870   2.40525498
tobgp30+    1.3850613913   4.09457245
```

Johtopäätöksiä:

- Ruokatorvensyöpään sairastumisen odds on kolminkertainen alkoholinkäyttöryhmässä "40-79 g/vrk" verrattuna ryhmään "0-39 g/vrk". Likimääräinen 95% luottamusväli ristitulosuhteelle (OR) on $(1.9, 4.9)$.

- Ruokatorvensyöpään sairastumisen odds on 4.3-kertainen alkoholinkäyttöryhmässä "80-119 g/vrk"verrattuna ryhmään "0-39 g/vrk". Likimääräinen 95% luottamusväli ristitulosuhteelle (OR) on $(2.5, 7.1)$.
- Ruokatorvensyöpään sairastumisen odds on 8.3-kertainen alkoholinkäyttöryhmässä "120+ g/vrk"verrattuna ryhmään "0-39 g/vrk". Likimääräinen 95% luottamusväli ristitulosuhteelle (OR) on $(4.7, 14.6)$.
- Ruokatorvensyöpään sairastumisen odds on 2.4-kertainen tupakointiryhmässä "30+ g/vrk"verrattuna ryhmään "0-9 g/vrk". Likimääräinen 95% luottamusväli ristitulosuhteelle (OR) on $(1.4, 4.1)$.