TILM 3558 Harjoitustyö, Tehtävä 1
Lasse Rintakumpu, 63555

Kategoristen vastemuuttujien mallitus

Tarkastellaan vuoden 2011 eduskuntavaaleja käsittelevää havaintoaineistoa (`EK2011.sav`). Havaintoaineisto sisältää n_{kaikki} = 1318 tilastoyksikköä, joista tähän tarkasteluun on valittu satunnaisesti n_{filter} = 800 tilastoyksikköä. [1] Havaintoaineisto sisältää 25 muuttujaa, joista kiinnostuksen kohteena ovat sukupuoli (`d2`, kategorinen dikotominen, 1 = "mies", 2 = "nainen"), edellisissä eduskuntavaaleissa äänestäminen (`q23`, kategorinen, 1 = "kyllä", 2 = "en", 3 = "ei äänioikeutta edellisissä vaaleissa", 7 = "ei halua sanoa", 8 = "ei osaa sanoa") sekä omaa sukupuolta edustavan ehdokkaan äänestäminen vuoden 2011 eduskuntavaaleissa (`k23`, 1 = "kyllä", 2 = "en", 3 = "en osaa sanoa").

Tarkastellaan ensin havaintoaineistoa frekvenssijaukaumilla sekä ristiintaulukoinnilla, jonka jälkeen pyritään tarkastelemaan ja tulkitsemaan kyseisten kolmen muuttujan välisiä riippuvuuksia log-lineaaristen mallien avulla.

1. Frekvenssijakaumat ja ristiintaulukointi

Poistetaan satunnaisesti valituista 800 tilastoyksiköstä ne tilastoyksiköt, joista ei ole saatavilla kaikkia kiinnostuksen kohteena olevia tietoja. Nyt havaintoaineiston kooksi saadaan n = 685 tilastoyksikköä. Muuttujien yksiulotteisiksi frekvenssijakaumiksi saadaan:

```
d2 (=sukupuoli)
          f  Prosenttiosuus
Mies     341      0.4978102
Nainen   344      0.5021898
Yhteensä 685      1.0000000

q23 (=vuoden 2007 vaaleissa äänestäminen)
                 f     Prosenttiosuus
Kyllä          602    0.878832117
Ei              49    0.071532847
Ei äänioikeutta 27    0.039416058
Ei halua sanoa   1    0.001459854
Ei osaa sanoa    6    0.008759124
Yhteensä       685    1.000000000
```

[1] Tarkasteltavat tilastoyksiköt on poimittu SPSS-ohjelmalla. Itse analyysissa on käytetty R-ohjelmistoa. Poimintaan käytetty SPSS-syntaksi sekä analyysiin käytetty R-koodi tehtävän liitteenä.

```
k23 (=oman sukupuolen äänestäminen)
               f     Prosenttiosuus
Kyllä         402     0.58686131
Ei            278     0.40583942
Ei osaa sanoa   5     0.00729927
Yhteensä      685     1.00000000.
```

Frekvenssitauluista havaitaan log-lineaarisen mallintamisen kannalta hankalia pienifrekvenssisiä luokkia (n<10). Koska luokista ei yhdistämälläkään saada suurempia, pudotetaan kyseiset luokat (q23:n “ei osaa sanoa” ja “ei halua sanoa” sekä k23:n “ei osaa sanoa) tarkastelusta. Nyt saadaan kolmen muuttujan välinen ristiintaulukko:

```
Sukupuoli = Mies
                           Oman sukupuolen äänestäminen
Äänestäminen vuonna 2007  Kyllä Ei
                   Kyllä  197   101
                      Ei  15    8
         Ei äänioikeutta  8     6

Sukupuoli = Nainen
                           Oman sukupuolen äänestäminen
Äänestäminen vuonna 2007  Kyllä Ei
                   Kyllä  165   134
                      Ei  12    14
         Ei äänioikeutta  1     12
```

Yhdessäkään ristiintaulukon solussa ei esiinny nollafrekvenssiä. Siirrytään mallintamaan muuttujien välistä riippuvuutta log-lineaarisella mallilla.

2. Muuttujien välisen riippuvuuden log-lineaarinen mallinnus

Pyritään mallintamaan muuttujien välistä riippuvuutta etsimällä taaksepäin askeltaen mahdollisimman yksinkertainen, merkitsevä log-lineaarinen malli. Aloitetaan askeltaminen täydestä mallista, jossa generoivana luokkana on {d2*q23*k23}. Käytetään askeltamiseen R:n stats-kirjaston funktiota drop1. [2] Funktion avulla päädytään malliin:

[2] Raportissa esitetty tulostus lyhennetty niin, että siitä ilmenee ainoastaan viimeinen askellus. Täysi tulostus liitteenä olevassa R-koodissa.

```
Single term deletions
Model:
~q23 + k23 + d2 + k23:d2 + q23:k23

        Df AIC     LRT     Pr(>Chi)
<none>     23.11
q23     0  21.872  0.0000
k23     0  21.872  0.0000
d2      0  21.872  0.0000
k23:d2  1  31.696  11.8239 0.0005848
q23:k23 2  26.059  8.1868  0.0166825
```

Jossa merkitseviä ($p<0.05$) termejä ovat k23:d2 ja q23:k23. Näin päädytään ehdolliseen riippumattomuusmalliin, jonka generoiva luokka on {k23*d2, q23*k23}. Yhteys löydetään siis oman sukupuolen äänestämisen ja vuoden 2007 eduskuntavaaleissa äänestämisen sekä oman sukupuolen äänestämisen ja oman sukupuolen välille. Oma sukupuoli ja vuoden 2007 eduskuntavaaleissa ääneständminen ovat kuitenkin riippumattomia. Summary-funktiolla saadaan seuraavaa yhteenveto mallista:

```
Statistics:
                      X^2 df P(> X^2)
Likelihood Ratio 5.872300  4 0.2088902
Pearson          5.153074  4 0.2719487.
```

Yhteenvedon perusteella mallin yhteensopivuustestin p-arvoksi saadaan likimain 0.209. Mallin standardoiduiksi jäännöksiksi saadaan:

```
Sukupuoli = mies
                                    Oman sukupuolen äänestäminen
Vuoden 2007 vaaleissa äänestäminen  Kyllä      Ei
                             Kyllä -0.21975329  0.2738556
                                Ei  0.01949492 -0.4047331
                   Ei äänioikeutta  1.24515522 -0.5772916

Sukupuoli = nainen
                                    Oman sukupuolen äänestäminen
Vuoden 2007 vaaleissa äänestäminen  Kyllä      Ei
                             Kyllä  0.24290246 -0.2340208
                                Ei -0.02171404  0.3303641
                   Ei äänioikeutta -1.80696957  0.4611147.
```

Standardoitujen jäännösten vaihteluväli on [-1.81,1.25], mutta suuret jäännökset ovat keskittyneet yksin omaa sukupuolta äänestäneisiin, joilla ei ole ollut äänioikeutta vuoden

2007 vaaleissa. Jäännösten ja p-arvon (>0.05) perusteella malli soveltuu kohtalaisesti kuvaamaan kiinnostuksen kohteena olevien muuttujien välisiä yhteyksiä.

3. Mallin yhteyksien jatkotarkastelu ja tulkinta

Löydetyssä mallissa on kaksi yhteyttä, oman sukupuolen äänestämisen ja vuoden 2007 eduskuntavaaleissa äänestämisen sekä oman sukupuolen äänestämisen ja oman sukupuolen välillä. Oma sukupuoli ja vuoden 2007 eduskuntavaaleissa äänestäminen ovat kuitenkin riippumattomia. Tehdään mallin jatkotarkastelu toisistaan riippuvien muuttujien d2 ja k23 sekä q23 ja k23 ristiintaulukoilla:

```
             Oman sukupuolen äänestäminen
                   Kyllä          Ei
      Sukupuoli    %        f     %        f
           Mies    65.67   220   34.33   115
         Nainen    52.66   178   47.34   160

Pearson's Chi-squared test with Yates' continuity correction

data:  table(d2, k23)
X-squared = 11.2505, df = 1, p-value = 0.000796
```

```
                   Oman sukupuolen äänestäminen
                   Kyllä          Ei
Äänestäminen 2007  %      f       %       f
Kyllä              60.64  362     39.36   235
Ei                 55.10  27      44.90   22
Ei äänioikeutta    33.33  9       66.67   18

Pearson's Chi-squared test

data:  table(q23, k23)
X-squared = 8.3251, df = 2, p-value = 0.01557.
```

Tuloksista havaitaan, että aineiston perusteella miehet äänestivät vuoden 2011 eduskuntavaaleissa tilastollisesti merkitsevästi naisia useammin omaa sukupuolta edustavaa ehdokasta. Samoin vuoden 2007 eduskuntavaaleissa äänestäneet äänestivät vuoden 2011 eduskuntavaaleissa tilastollisesti merkitsevästi useammin omaa sukupuoltaan edustavaa ehdokasta kuin eduskuntavaaleissa vuonna 2007 äänestämättä jättäneet tai ne, joilla vuoden 2007 eduskuntavaaleissa ei ollut äänioikeutta.

Liitteet

Liitteenä tilastoyksikköjen valinnassa käytetty SPSS-syntaksi sekä analyysissa käytetty R-koodi.

Liite 1. SPSS-syntaksi

```
SET SEED=63555.
DATASET ACTIVATE DataSet1.
USE ALL.
do if $casenum=1.
compute #s_$_1=800.
compute #s_$_2=1318.
end if.
do if  #s_$_2 > 0.
compute filter_$=uniform(1)* #s_$_2 < #s_$_1.
compute #s_$_1=#s_$_1 - filter_$.
compute #s_$_2=#s_$_2 - 1.
else.
compute filter_$=0.
end if.
VARIABLE LABELS filter_$ '800 from the first 1318 cases (SAMPLE)'.
FORMATS filter_$ (f1.0).
FILTER  BY filter_$.
EXECUTE.
```

Liite 2. R-koodi

```
########################################
# TILM3558 Harjoitustyö, Osa 1, R-koodi #
# Lasse Rintakumpu, 63555               #
# 18.8.2015                             #
########################################

# Asetetaan työhakemisto
wd <- "D:/Dropbox/Edu/Statistics/TILM 3558 Harjoitustyö" #Hipatlaptop
setwd(wd)

# Funktio kirjastojen asentamiselle / lataamiselle
lataa_kirjasto <- function(kirjasto) {
  if(kirjasto %in% rownames(installed.packages()) == FALSE)
  {install.packages(kirjasto)}
  library(kirjasto, character.only = TRUE)
}
```

```
# Ladataan/asennetaan käytetyt kirjastot
lapply(c("MASS", "MissMech", "tables"), lataa_kirjasto)

# Hoidetaan puuttuvat havainnot poistamalla tilastoyksiköt joissa puuttuvia
havaintoja
ek2011 <-
na.omit(read.csv("https://raw.githubusercontent.com/rintakumpu/tilm3558/maste
r/EK2011_filtered.csv", header=TRUE, row.names=NULL, fileEncoding =
"UTF-8-BOM"));

# Tallennetaan käsiteltävä data omiin muuttujiinsa
sukupuoli <- ek2011$d2[ek2011$filter==1] # Sukupuoli
aanestys2007 <- ek2011$q23[ek2011$filter==1] #Äänestysaktiivisuus vuoden 2007
vaaleissa
omasupu <- ek2011$k23[ek2011$filter==1] # Oman sukupuolen äänestäminen

########################################
# 1. Frekvenssit ja ristiintaulukointi #
########################################

# Luodaan frekvenssitaulut
sukupuoli_t<-table(sukupuoli)
sukupuoli_t<-addmargins(sukupuoli_t)
row.names(sukupuoli_t) <- c("Mies", "Nainen", "Yhteensä")
sukupuoli_t<-cbind( f=sukupuoli_t, Prosenttiosuus=prop.table(sukupuoli_t)*2)

#           f     Prosenttiosuus
#Mies     341      0.4978102
#Nainen   344      0.5021898
#Yhteensä 685      1.0000000

aanestys2007_t<-table(aanestys2007)
aanestys2007_t<-addmargins(aanestys2007_t)
row.names(aanestys2007_t) <- c("Kyllä", "Ei", "Ei äänioikeutta", "Ei halua
sanoa", "Ei osaa sanoa","Yhteensä")
aanestys2007_t<-cbind( f=aanestys2007_t,
Prosenttiosuus=prop.table(aanestys2007_t)*2)

#                  f  Prosenttiosuus
#Kyllä           602    0.878832117
#Ei               49    0.071532847
#Ei äänioikeutta  27    0.039416058
#Ei halua sanoa    1    0.001459854
#Ei osaa sanoa     6    0.008759124
#Yhteensä        685    1.000000000

omasupu_t<-table(omasupu)
omasupu_t<-addmargins(omasupu_t)
row.names(omasupu_t) <- c("Kyllä", "Ei", "Ei osaa sanoa", "Yhteensä")
```

```
omasupu_t<-cbind( f=omasupu_t, Prosenttiosuus=prop.table(omasupu_t)*2)

#                f   Prosenttiosuus
#Kyllä         402     0.58686131
#Ei            278     0.40583942
#Ei osaa sanoa   5     0.00729927
#Yhteensä      685     1.00000000

# Poistetaan pienifrekvenssiset luokat, tallennetaan data
# uusiin _mod-muuttujiin
aanestys2007_mod<-aanestys2007[!aanestys2007 %in% c(7,8)]
sukupuoli_mod<-sukupuoli[!aanestys2007 %in% c(7,8)]
omasupu_mod<-omasupu[!aanestys2007 %in% c(7,8)]

sukupuoli_mod<-sukupuoli_mod[!omasupu_mod %in% c(3)]
aanestys2007_mod<-aanestys2007_mod[!omasupu_mod %in% c(3)]
omasupu_mod<-omasupu_mod[!omasupu_mod %in% c(3)] # Vektori ylikirjoitetaan

ct<-xtabs(~aanestys2007_mod+omasupu_mod+sukupuoli_mod) # Luodaan kolmen
muuttujan ristiintaulukko

#   Sukupuoli = Mies
#                               Oman sukupuolen äänestäminen
#   Äänestäminen vuonna 2007  Kyllä Ei
#                       Kyllä  197   101
#                          Ei  15    8
#            Ei äänioikeutta  8     6

#   Sukupuoli = Nainen
#                               Oman sukupuolen äänestäminen
#   Äänestäminen vuonna 2007  Kyllä Ei
#                       Kyllä  165   134
#                          Ei  12    14
#            Ei äänioikeutta  1     12

################################################################
# 2. Muuttujien välisen riippuvuuden loglineaarinen mallinnus #
################################################################

# Askelletaan taaksepäin täydestä mallista
malli1 <- loglm(~aanestys2007_mod*omasupu_mod*sukupuoli_mod, data=ct)
drop1(malli1, scope = ~aanestys2007_mod*omasupu_mod*sukupuoli_mod,
test="Chisq", trace=TRUE)

#Single term deletions
#Model:
#~aanestys2007_mod * omasupu_mod * sukupuoli_mod
#                                          Df    AIC    LRT Pr(>Chi)
#<none>                                       24.000
```

```
#aanestys2007_mod                              0 24.000 0.0000
#omasupu_mod                                   0 24.000 0.0000
#sukupuoli_mod                                 0 24.000 0.0000
#aanestys2007_mod:omasupu_mod                  0 24.000 0.0000
#aanestys2007_mod:sukupuoli_mod                0 24.000 0.0000
#omasupu_mod:sukupuoli_mod                     0 24.000 0.0000
#aanestys2007_mod:omasupu_mod:sukupuoli_mod  2 25.424 5.4243  0.06639 .

# Poistetaan aanestys2007_mod:omasupu_mod:sukupuoli_mod ei-merkitsevänä
malli2 <-
loglm(~aanestys2007_mod*omasupu_mod*sukupuoli_mod-aanestys2007_mod:omasupu_mo
d:sukupuoli_mod, data=ct)
drop1(malli2, scope =
~aanestys2007_mod*omasupu_mod*sukupuoli_mod-aanestys2007_mod:omasupu_mod:suku
puoli_mod, test="Chisq")

#Single term deletions
#Model:
#  ~aanestys2007_mod * omasupu_mod * sukupuoli_mod -
aanestys2007_mod:omasupu_mod:sukupuoli_mod
#                               Df    AIC     LRT Pr(>Chi)
#<none>                            25.564
#aanestys2007_mod                0 25.564  0.0000
#omasupu_mod                     0 25.564  0.0000
#sukupuoli_mod                   0 25.564  0.0000
#aanestys2007_mod:omasupu_mod    2 29.850  8.4257 0.0148043 *
#aanestys2007_mod:sukupuoli_mod  2 21.872  0.4480 0.7993029
#omasupu_mod:sukupuoli_mod       1 35.487 12.0628 0.0005144 ***

#Poistetaan aanestys2007_mod:sukupuoli_mod
#mallista ei-merkitsevänä.
#omasupu_mod:sukupuoli_mod jää malliin merkitsevänä
#p = 0.00148
#aanestys2007_mod:omasupu_mod jää malliin merkitsevänä
#p = 0.01480

malli3 <-
loglm(~aanestys2007_mod+omasupu_mod+sukupuoli_mod+omasupu_mod:sukupuoli_mod+a
anestys2007_mod:omasupu_mod, data=ct)
drop1(malli3, scope =
~aanestys2007_mod+omasupu_mod+sukupuoli_mod+omasupu_mod:sukupuoli_mod+aanesty
s2007_mod:omasupu_mod, test="Chisq")

#Single term deletions
#Model:
#  ~aanestys2007_mod + omasupu_mod + sukupuoli_mod +
omasupu_mod:sukupuoli_mod +
#aanestys2007_mod:omasupu_mod
#                           Df    AIC    LRT  Pr(>Chi)
```

```
#<none>                          23.11
#aanestys2007_mod               0 21.872  0.0000
#omasupu_mod                    0 21.872  0.0000
#sukupuoli_mod                  0 21.872  0.0000
#omasupu_mod:sukupuoli_mod      1 31.696 11.8239 0.0005848 ***
#aanestys2007_mod:omasupu_mod   2 26.059  8.1868 0.0166825 *

# Lopulliseen malliin jäävät omasupu_mod:sukupuoli_mod
# ja aanestys2007_mod:omasupu_mod

malli4 <- loglm(~aanestys2007_mod:omasupu_mod+omasupu_mod:sukupuoli_mod,
data=ct)

# Standardoidut jäännökset
residuals(malli4)

#, , sukupuoli_mod = 1

#                 omasupu_mod
#aanestys2007_mod           1          2
#               1 -0.21975329  0.2738556
#               2  0.01949492 -0.4047331
#               3  1.24515522 -0.5772916

#, , sukupuoli_mod = 2

#               omasupu_mod
#aanestys2007_mod           1          2
#               1  0.24290246 -0.2340208
#               2 -0.02171404  0.3303641
#               3 -1.80696957  0.4611147

# Yhteensopivuustesti
summary(malli4)

# Statistics:
#                       X^2 df P(> X^2)
# Likelihood Ratio 5.872300  4 0.2088902
# Pearson          5.153074  4 0.2719487

#######################################################
# 3. Mallin yhteyksien jatkotarkastelu ja tulkinta  #
#######################################################

# Mallin generoiva luokka on {aanestys2007_mod:omasupu_mod,
omasupu_mod:sukupuoli_mod}
# Mallissa on kaksi yhteyttä, Tulkitaan yhteyttä ristiintaulukoimalla:

omasupu_mod <- factor(omasupu_mod, levels=c(1,2), labels=c("Kyllä","Ei"))
```

```
sukupuoli_mod <- factor(sukupuoli_mod, levels=c(1,2), labels=c("Mies",
"Nainen"))
aanestys2007_mod <- factor(aanestys2007_mod, levels=c(1,2,3),
labels=c("Kyllä", "Ei", "Ei äänioikeutta"))

ct2 <-
tabular((Sukupuoli=sukupuoli_mod)~(Heading(Oman_sukupuolen_aanestaminen)*omas
upu_mod*(Percent("row")+ 1)))

#           Oman sukupuolen äänestäminen
#           Kyllä                            Ei
# Sukupuoli Percent                      All Percent All
# Mies      65.67                        220 34.33   115
# Nainen    52.66                        178 47.34   160

chisq.test(table(sukupuoli_mod, omasupu_mod))

# Pearson's Chi-squared test with Yates' continuity correction
# data:  table(sukupuoli_mod, omasupu_mod)
# X-squared = 11.2505, df = 1, p-value = 0.000796

ct2 <-
tabular((Aanestys_2007=aanestys2007_mod)~(Heading(Oman_sukupuolen_aanestamine
n)*omasupu_mod*(Percent("row")+ 1)))

#                Oman_sukupuolen_aanestaminen
#                Kyllä                           Ei
#Aanestys_2007   Percent                      All Percent All
#Kyllä           60.64                        362 39.36   235
#Ei              55.10                         27 44.90    22
#Ei äänioikeutta 33.33                          9 66.67    18

chisq.test(table(aanestys2007_mod, omasupu_mod))

# Pearson's Chi-squared test
# data:  table(aanestys2007_mod, omasupu_mod)
# X-squared = 8.3251, df = 2, p-value = 0.01557

# Taulukon perusteella miehet näyttäisivät äänestävän naisia
# useammin omaa sukupuolta edustavaa ehdokasta.
# Samoin tekivät vuoden 2007 eduskuntavaaleissa äänestäneet.
```