78185 Yleistetyt lineaariset mallit

Harjoitus 5, syksy 2014

Esimerkkiratkaisut

1. Tutkimuksessa selvitettiin ihmisten asenteita maahanmuuttoa kohtaan. Vastaajilta kysyttiin muun muassa seuraavavia asioita:

   - Miten suhtaudut ulkomaalaisten työnhakijoiden muuttoon Suomeen. (*ulkom*)
   - Kuinka paljon sinulla on ammattikoulutusta (*ammattik*)
   - Pidätkö tärkeänä, että maahanmuuttajat voisivat säilyttää oman kulttuurinsa ja kielensä? (*kulttuuri*)

   Seuraavassa on ristiintaulukoitu muuttujat *ulkom* ja *kulttuuri* muuttujan *ammattik* kanssa.

	ulkom			
ammattik	1	2	3	4
1	14	69	26	5
2	18	77	31	5
3	16	72	39	11
4	7	102	70	19
5	2	22	32	15

	kulttuuri			
ammattik	1	2	3	4
1	21	54	32	7
2	28	55	35	13
3	31	54	40	13
4	37	101	48	12
5	26	33	9	3

$$
\begin{aligned}
ulkom &= \begin{cases} 1, & \text{En osaa sanoa,} \\ 2, & \text{Suomen tulisi ottaa vähemmän,} \\ 3, & \text{Nykyinen määrä on riittävä,} \\ 4, & \text{Suomen tulisi ottaa enemmän.} \end{cases} \\
ammattik &= \begin{cases} 1, & \text{Ei ollenkaan,} \\ 2, & \text{Ammattikurssi tai vastaava,} \\ 3, & \text{Ammattikoulu,} \\ 4, & \text{Opistoasteen koulutus,} \\ 5, & \text{Yliopistollinen loppututkinto.} \end{cases} \\
kulttuuri &= \begin{cases} 1, & \text{En lainkaan tärkeänä,} \\ 2, & \text{En kovin tärkeänä,} \\ 3, & \text{Kyllä, melko tärkeänä,} \\ 4, & \text{Kyllä, erittäin täärkeänä,} \end{cases}
\end{aligned}
$$

Aineisto on tiedostossa `asenne.dat`.

a) Sovita aineistoon multinomijakaumamalli käyttämällä luennolla esiteltyä R-funktiota `multinom` (R-pakkaus `nnet`) ja/tai R-funktiota `mlogit` (R-pakkaus `mlogit`). Käytä vastemuuttujana muuttujaa *ulkom* ja selittäjänä muuttujaa *ammattik*. Tulkitse tulokset.

b) Sovita aineistoon funktion `polr` (R-pakkaus `MASS`) avulla logistinen malli, jossa vastemuuttujana on ordinaaliasteikollinen *kulttuuri* ja selittäjänä *ammattik*. Tulkitse tulokset.

Aineiston lähde:

- Söderling, Ismo: Suomalaisten suhtautuminen ulkomaalaisiin 1996 [elektroninen aineisto]. FSD1111, versio 1.0 (2002-01-23). Turku: Turun yliopisto. Sosiaalipolitiikan laitos & Helsinki: Väestöliitto. Väestöntutkimuslaitos [tuottajat], 1996. Tampere: Yhteiskuntatieteellinen tietoarkisto [jakaja], 2002.

Vihjeitä tehtävään 1:

Luokat nominaaliasteikolla:

$$\begin{aligned} \boldsymbol{y} &= (y_1, \ldots, y_k) \sim Multin(n; (p_1, \ldots, p_k)) \\ p_j &= \frac{\exp(\boldsymbol{\beta}_j^T \boldsymbol{x})}{\sum_{i=1}^{k} \exp(\boldsymbol{\beta}_i^T \boldsymbol{x})}, \quad j = 1, \ldots, k. \end{aligned}$$

```
> library(nnet)
> y<-factor(y)
> mod1 <- multinom(y ~ 1)  #Mallissa ei selittajia (ainoastaan vakio)
> summary(mod1)
> head(fitted(mod1))
> coefficients(mod1)
> mod2 <- multinom(y ~ x)  #Yksi selittava muuttuja x
> summary(mod2)
> coefficients(mod2)
```

```
> library(mlogit)
> dat<-data.frame(y=y)
> mdat<-mlogit.data(dat,varying=NULL,choice="y",shape="wide")
> mod3<-mlogit(y~1,data=mdat) #Mallissa ei selittajia (ainoastaan vakio)
> summary(mod3)
> dat<-data.frame(y=y,x=x)
> mdat<-mlogit.data(dat,varying=NULL,choice="y",shape="wide")
> mod4<-mlogit(y~1|x,data=mdat)  #Yksi selittava muuttuja x
> summary(mod4)
```

Luokat ordinaaliasteikolla:

Merkitään

$$P_1 = p_1, \quad P_2 = p_1 + p_2, \quad \ldots \quad , P_j = p_1 + \ldots + p_j, \quad \ldots \quad , P_k = 1$$

ja rakennetaan "logistinen malli"olettaen, että

$$\log \frac{P_j}{1 - P_j} = \theta_j - \boldsymbol{\beta}^T \boldsymbol{x}.$$

```
> library(MASS)
> y<-factor(y,ordered=TRUE)
> mod<-polr(y~1) #Mallissa ei selittajia (ainoastaan vakio)
> summary(mod)
> mod$zeta
> mod2<-polr(y~x) #Yksi selittava muuttuja x
> summary(mod2)
> mod2$zeta
```

Ratkaisu:

a) Olkoon

$$x_i = \begin{cases} 1, & ammattik = i \\ 0, & ammattik \neq i, \end{cases}$$

jossa $i = 1, \ldots, 5$. Vertailtaessa luokkaa j luokkaan 1, todennäköisyyksien osamäärä on

$$\frac{p_j}{p_1} = \exp(\boldsymbol{\beta}_j^T \boldsymbol{x}) = \exp(\beta_{j1} + \beta_{j2}x_2 + \cdots + \beta_{j5}x_5), \quad j = 2, \ldots, 4.$$

Kun $x_1 = 1$ (eli $ammattik = 1$ eli "ei ollenkaan koulutusta"), niin

$$\frac{p_j}{p_1} = \exp(\beta_{j1}), \quad j = 2, \ldots, 4.$$

Kun $x_2 = 1$ (eli $ammattik = 2$ eli "Ammattikurssi tai vastaava"), niin

$$\frac{p_j}{p_1} = \exp(\beta_{j1} + \beta_{j2}) = \exp(\beta_{j2}) \exp(\beta_{j1}), \quad j = 2, \ldots, 4.$$

Luku $\exp(\beta_{j2})$ kertoo siis kuinka moninkertainen on osamäärä p_j/p_1 verrattaessa ryhmää $ammattik = 2$ ryhmään $ammattik = 1$. Kun $x_3 = 1$ (eli $ammattik = 3$ eli "Ammattikoulu"), niin

$$\frac{p_j}{p_1} = \exp(\beta_{j1} + \beta_{j3}) = \exp(\beta_{j3}) \exp(\beta_{j1}), \quad j = 2, \ldots, 4.$$

Luku $\exp(\beta_{j3})$ kertoo siis kuinka moninkertainen on osamäärä p_j/p_1 verrattaessa ryhmää $ammattik = 3$ ryhmään $ammattik = 1$. Kun $x_4 = 1$ (eli $ammattik = 4$ eli "Opistoasteen koulutus"), niin

$$\frac{p_j}{p_1} = \exp(\beta_{j1} + \beta_{j4}) = \exp(\beta_{j4}) \exp(\beta_{j1}), \quad j = 2, \ldots, 4.$$

Luku $\exp(\beta_{j4})$ kertoo siis kuinka moninkertainen on osamäärä p_j/p_1 verrattaessa ryhmää $ammattik = 4$ ryhmään $ammattik = 1$. Kun $x_5 = 1$ (eli $ammattik = 5$ eli "Yliopistollinen loppututkinto"), niin

$$\frac{p_j}{p_1} = \exp(\beta_{j1} + \beta_{j5}) = \exp(\beta_{j5}) \exp(\beta_{j1}), \quad j = 2, \ldots, 4.$$

Luku $\exp(\beta_{j5})$ kertoo siis kuinka moninkertainen on osamäärä p_j/p_1 verrattaessa ryhmää $ammattik = 5$ ryhmään $ammattik = 1$.

```
> asenne<-read.table("asenne.dat",header=TRUE)
> names(asenne)
[1] "ammattik"  "ulkom"     "kulttuuri"
> library(nnet)
> asenne$ulkom<-factor(asenne$ulkom)
```

```
> asenne$ammattik<-factor(asenne$ammattik)
> mod1<-multinom(ulkom~ammattik,data=asenne)
# weights:  24 (15 variable)
initial  value 903.863923
iter  10 value 709.526164
iter  20 value 704.249007
iter  20 value 704.249003
iter  20 value 704.249003
final  value 704.249003
converged
> summary(mod1)
Call:
multinom(formula = ulkom ~ ammattik, data = asenne)

Coefficients:
  (Intercept)   ammattik2   ammattik3 ammattik4 ammattik5
2   1.5950945 -0.14165747 -0.09103958  1.084012 0.8025778
3   0.6191115 -0.07550671  0.27183016  1.683511 2.1532574
4  -1.0294143 -0.25139840  0.65464729  2.028011 3.0441101

Std. Errors:
  (Intercept) ammattik2 ammattik3 ammattik4 ammattik5
2   0.2931300 0.3930242 0.4028801 0.4884591 0.7945277
3   0.3315014 0.4446431 0.4450086 0.5167612 0.8006481
4   0.5209625 0.7259024 0.6517773 0.6832974 0.9154028

Residual Deviance: 1408.498
AIC: 1438.498
> coefficients(mod1)
  (Intercept)   ammattik2   ammattik3 ammattik4 ammattik5
2   1.5950945 -0.14165747 -0.09103958  1.084012 0.8025778
3   0.6191115 -0.07550671  0.27183016  1.683511 2.1532574
4  -1.0294143 -0.25139840  0.65464729  2.028011 3.0441101
> B<-coefficients(mod1)
> B<-rbind(0,B)
> B
  (Intercept)   ammattik2   ammattik3 ammattik4 ammattik5
    0.0000000  0.00000000  0.00000000  0.000000 0.0000000
2   1.5950945 -0.14165747 -0.09103958  1.084012 0.8025778
3   0.6191115 -0.07550671  0.27183016  1.683511 2.1532574
4  -1.0294143 -0.25139840  0.65464729  2.028011 3.0441101
```

Lasketaan todennäköisyydet p_j, $j = 1, \ldots, 4$ eri profiileille $\boldsymbol{x}$:

```
> x<-c(1,0,0,0,0)    ## ammattik=1
```

```
> exp(B%*%x)/sum(exp(B%*%x))
        [,1]
  0.12280051
2 0.60525856
3 0.22807460
4 0.04386632
> x<-c(1,1,0,0,0)    ## ammattik=2
> exp(B%*%x)/sum(exp(B%*%x))
        [,1]
  0.13740401
2 0.58778584
3 0.23663774
4 0.03817241
> x<-c(1,0,1,0,0)    ## ammattik=3
> exp(B%*%x)/sum(exp(B%*%x))
        [,1]
  0.11594509
2 0.52174119
3 0.28260733
4 0.07970639
> x<-c(1,0,0,1,0)    ## ammattik=4
> exp(B%*%x)/sum(exp(B%*%x))
        [,1]
  0.03535204
2 0.51515221
3 0.35353371
4 0.09596204
> x<-c(1,0,0,0,1)    ## ammattik=5
> exp(B%*%x)/sum(exp(B%*%x))
        [,1]
  0.02817498
2 0.30985573
3 0.45070069
4 0.21126860
```

Edellä lasketut todennäköisyydet saadaan myös seuraavalla tavalla:

```
> newd<-data.frame(ammattik=factor(1:5))
> pr<-predict(mod1,type="probs",newdata=newd)
> pr
          1         2         3          4
1 0.12280051 0.6052586 0.2280746 0.04386632
2 0.13740401 0.5877858 0.2366377 0.03817241
3 0.11594509 0.5217412 0.2826073 0.07970639
4 0.03535204 0.5151522 0.3535337 0.09596204
```

```
5 0.02817498 0.3098557 0.4507007 0.21126860
```

Lasketaan seuraavaksi arvot $\exp(\beta_{ji})$, $j = 2, \ldots, 4$, $i = 1, \ldots, 5$:

```
> exp(coefficients(mod1))
  (Intercept) ammattik2 ammattik3 ammattik4 ammattik5
2   4.9287950 0.8679185 0.9129816  2.956516  2.231285
3   1.8572772 0.9272735 1.3123641  5.384428  8.612868
4   0.3572161 0.7777125 1.9244636  7.598955 20.991343
```

Todetaan vielä että edellä saatu taulukko koostuu todennäköisyyksien osamääristä ja osamäärien osamääristä:

```
> o2.1.1<-0.6052586/0.12280051;  o2.1.1  #exp(beta21)
[1] 4.928795
> o3.1.1<-0.2280746/0.12280051;  o3.1.1  #exp(beta31)
[1] 1.857277
> o4.1.1<-0.04386632/0.12280051; o4.1.1  #exp(beta41)
[1] 0.3572161
> o2.1.2<-0.5877858/0.13740401; o2.1.2/o2.1.1  #exp(beta22)
[1] 0.8679184
> o3.1.2<-0.2366377/0.13740401; o3.1.2/o3.1.1  #exp(beta32)
[1] 0.9272733
> o4.1.2<-0.03817241/0.13740401; o4.1.2/o4.1.1 #exp(beta42)
[1] 0.7777126
> o2.1.3<-0.5217412/0.11594509; o2.1.3/o2.1.1  #exp(beta23)
[1] 0.9129815
> o3.1.3<-0.2826073/0.11594509; o3.1.3/o3.1.1  #exp(beta33)
[1] 1.312364
> o4.1.3<-0.07970639/0.11594509; o4.1.3/o4.1.1 #exp(beta43)
[1] 1.924464
> o2.1.4<-0.5151522/0.03535204; o2.1.4/o2.1.1  #exp(beta24)
[1] 2.956516
> o3.1.4<-0.3535337/0.03535204; o3.1.4/o3.1.1  #exp(beta34)
[1] 5.384428
> o4.1.4<-0.09596204/0.03535204; o4.1.4/o4.1.1 #exp(beta44)
[1] 7.598957
> o2.1.5<-0.3098557/0.02817498; o2.1.5/o2.1.1  #exp(beta25)
[1] 2.231285
> o3.1.5<-0.4507007/0.02817498; o3.1.5/o3.1.1  #exp(beta35)
[1] 8.612869
> o4.1.5<-0.21126860/0.02817498; o4.1.5/o4.1.1 #exp(beta45)
[1] 20.99135
```

Edellä olevasta nähdään, että mitä korkeampi on koulutustaso niin sitä myönteisemmin suhtautuu työnhakijoiden muuttoon Suomeen. Esimerkik-

si, $\exp(\hat{\beta}_{45}) = 20.99135$ kertoo, että suhde p_4/p_1 on noin 21 kertainen yliopistollisen loppututkinnon suorittaneilla verrattuna sellaisiin, joilla ei ole ollenkaan koulutusta. Suhde p_4/p_1 on siis luokkien "Suomen tulisi ottaa enemmän" ja "En osaa sanoa" todennäköisyyksien suhde.

Tehdään vielä mallinnus käyttämällä funktiota `mlogit`:

```
\begin{verbatim}
> library(mlogit)
> mdat<-mlogit.data(asenne,varying=NULL,choice="ulkom",shape="wide")
> mod2<-mlogit(ulkom~1|ammattik,data=mdat)
> summary(mod2)

Call:
mlogit(formula = ulkom ~ 1 | ammattik, data = mdat, method = "nr",
    print.level = 0)

Frequencies of alternatives:
       1        2        3        4
0.087423 0.524540 0.303681 0.084356

nr method
6 iterations, 0h:0m:0s
g'(-H)^-1g = 1.58E-05
successive function values within tolerance limits

Coefficients :
               Estimate Std. Error t-value  Pr(>|t|)
2:(intercept)  1.595049   0.293123  5.4416 5.282e-08 ***
3:(intercept)  0.619039   0.331497  1.8674 0.0618449 .
4:(intercept) -1.029619   0.520988 -1.9763 0.0481228 *
2:ammattik2   -0.141616   0.393019 -0.3603 0.7186023
3:ammattik2   -0.075424   0.444639 -0.1696 0.8653017
4:ammattik2   -0.251314   0.725937 -0.3462 0.7291975
2:ammattik3   -0.090972   0.402877 -0.2258 0.8213530
3:ammattik3    0.271934   0.445007  0.6111 0.5411483
4:ammattik3    0.654926   0.651796  1.0048 0.3149920
2:ammattik4    1.084013   0.488449  2.2193 0.0264665 *
3:ammattik4    1.683546   0.516752  3.2579 0.0011223 **
4:ammattik4    2.028148   0.683313  2.9681 0.0029964 **
2:ammattik5    0.802842   0.794592  1.0104 0.3123115
3:ammattik5    2.153546   0.800712  2.6895 0.0071551 **
4:ammattik5    3.044519   0.915475  3.3256 0.0008822 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Log-Likelihood: -704.25
McFadden R^2:  0.037319
Likelihood ratio test : chisq = 54.601 (p.value = 2.1334e-07)
```

Lasketaan vielä luokkatodennäköisyydet p_j. Seuraavassa tehdään uusi mlogit.data -tyyppinen aineisto, jossa on viisi havaintoa. Todennäköisyydet lasketaan muuttujan `ammattik` arvoilla $1, \ldots, 5$. Muuttujien `ulkom` ja `kulttuuri` arvot voivat olla mitä tahansa kunhan vain muuttujan `ulkom` kaikki mahdolliset arvot esiintyvät aineistossa.

```
> newdat<-data.frame(ammattik=factor(1:5),ulkom=factor(c(1:4,1)),
+ kulttuuri=factor(rep(1,5)))
> newmdat<-mlogit.data(newdat,varying=NULL,choice="ulkom",shape="wide")
> pr<-predict(mod2,newdata=newmdat)
> pr
             1         2         3          4
[1,] 0.12280702 0.6052632 0.2280702 0.04385965
[2,] 0.13740458 0.5877863 0.2366412 0.03816794
[3,] 0.11594203 0.5217391 0.2826087 0.07971014
[4,] 0.03535354 0.5151515 0.3535354 0.09595960
[5,] 0.02816912 0.3098591 0.4507042 0.21126758
```

b) Olkoon Y = 'Havainto kuuluu luokkaan j', $j = 1, \ldots, k$. Tällöin $P_j = P(Y \leq j)$ ja

$$\frac{P_j}{1 - P_j} = \frac{P(Y \leq j)}{P(Y > j)}.$$

Monesti ordinaaliasteikon muuttuja on määritelty siten, että suuret arvot viittaavat "vakavampaan tilaan"(esimerkiksi sairauden eri asteita), jolloin on perusteltua käyttää edellisen osamäärän käänteislukua

$$\frac{1 - P_j}{P_j} = \frac{P(Y > j)}{P(Y \leq j)} = \frac{P(Y > j)}{1 - P(Y > j)}.$$

Olkoon (kuten a)-kohdassa)

$$x_i = \begin{cases} 1, & ammattik = i \\ 0, & ammattik \neq i, \end{cases}$$

jossa $i = 1, \ldots, 5$. Malliksi saadaan tällöin

$$\log \frac{1 - P_j}{P_j} = -\theta_j + \boldsymbol{\beta}^T \boldsymbol{x}.$$

eli

$$\frac{1 - P_j}{P_j} = \exp(-\theta_j) \exp(\boldsymbol{\beta}^T \boldsymbol{x}) = \exp(-\theta_j) \exp(\beta_1 + \beta_2 x_2 + \ldots + \beta_5 x_5).$$

Kun $x_1 = 1$, niin oddsiksi saadaan

$$o_{j1} = \frac{1 - P_j}{P_j} = \exp(-\theta_j)\exp(\beta_1)$$

Kun $x_i = 1$, $i = 2, \ldots 5$, niin oddsiksi saadaan

$$o_{ji} = \frac{1 - P_j}{P_j} = \exp(-\theta_j)\exp(\beta_1)\exp(\beta_i)$$

Ristitulosuhteeksi (OR) saadaan verrattaessa ryhmiä $ammattik = 1$ ja $ammattik = i$

$$OR_{ji} = \frac{o_{ji}}{o_{j1}} = \exp(\beta_i).$$

Huom! Edellä oleva ristitulosuhde ei riipu j:stä! Tämä mallioletus ei ole välttämättä realistinen kaikissa tapauksissa.

```
> library(MASS)
> asenne$kulttuuri<-factor(asenne$kulttuuri,ordered=TRUE)
> mod3<-polr(kulttuuri~ammattik,data=asenne)
> summary(mod3)

Re-fitting to get Hessian

Call:
polr(formula = kulttuuri ~ ammattik, data = asenne)

Coefficients:
             Value Std. Error t value
ammattik2  0.04013     0.2371  0.1693
ammattik3  0.05438     0.2347  0.2317
ammattik4 -0.09993     0.2144 -0.4662
ammattik5 -0.87817     0.2818 -3.1159

Intercepts:
    Value   Std. Error t value
1|2 -1.3993  0.1837    -7.6162
2|3  0.6366  0.1753     3.6326
3|4  2.4538  0.2149    11.4176

Residual Deviance: 1589.94
AIC: 1603.94
> coefficients(mod3)
  ammattik2   ammattik3   ammattik4   ammattik5
 0.04012650  0.05438387 -0.09993144 -0.87817135
> exp(coefficients(mod3))
```

```
ammattik2 ammattik3 ammattik4 ammattik5
1.0409424 1.0558898 0.9048995 0.4155421
> mod3$zeta
       1|2        2|3        3|4
-1.3993187  0.6366453  2.4538293
> exp(confint(mod3))
Waiting for profiling to be done...

Re-fitting to get Hessian

              2.5 %    97.5 %
ammattik2 0.6539929 1.6572688
ammattik3 0.6665165 1.6732747
ammattik4 0.5943578 1.3778785
ammattik5 0.2385509 0.7208242
```

Ainoastaan `ammattik5` (eli $ammattik = 5$) on tilastollisesti merkitsevä. Ristitulosuhteen estimaatiksi saatiin $\widehat{OR}_{j5} = 0.4155421 \approx 0.42$. Suuren vastemuuttuja-arvon $(Y > j)$ odds on ryhmässä $ammattik = 5$ noin 58% pienempi kuin ryhmässä $ammattik = 1$.

2. Oletetaan, että $y_1, \ldots, y_m$ ja $y_{m+1}, \ldots, y_{m+n}$ ovat riippumattomat satunnaisotokset jakaumista $Poi(\mu_1)$ ja $Poi(\mu_2)$. Merkitään

$$z_1 = y_1 + \ldots + y_m \quad \text{ja} \quad z_2 = y_{m+1} + \ldots + y_{m+n} \quad \text{sekä} \quad z = z_1 + z_2.$$

Käytettäessä uudelleenparametrisointia

$$\tau = m\mu_1 + n\mu_2 \quad \text{ja} \quad \Delta = \mu_2/\mu_1,$$

uskottavuusfunktio on muotoa (ks. luentomoniste)

$$L(\tau, \Delta) = vakio \cdot \tau^z e^{-\tau} \cdot \left(\frac{1}{1 + \frac{n}{m}\Delta}\right)^{z_1} \left(\frac{\frac{n}{m}\Delta}{1 + \frac{n}{m}\Delta}\right)^{z_2}.$$

a) Näytä, että edellä oleva uskottavuusfunktio on muotoa

$$L(\tau, \Delta) = vakio \cdot f(z; \tau) \cdot f(z_1 | z \ ; \ \theta),$$

jossa $f(z;\tau)$ on Poisson-jakauman pistetodennäköisyys, $f(z_1|z \ ; \ \Delta)$ on $Bin(z, \theta)$-jakauman pistetodennäköisyys ja $\theta = (1 + \frac{n}{m}\Delta)^{-1}$.

b) Laske parametrien τ ja Δ suurimman uskottavuuden estimaatit käyttämällä hyväksi a)-kohdan tulosta.

c) Johda parametrille Δ likimääräinen 95% luottamusväli.

Ratkaisu:

a) Uskottavuusfunktio on

$$\begin{aligned} L(\mu_1, \mu_2) \ = \ & \frac{z_1!}{\prod_{i=1}^m y_i!} \left(\frac{1}{m}\right)^{z_1} \cdot \frac{z_2!}{\prod_{i=m+1}^{m+n} y_j!} \left(\frac{1}{n}\right)^{z_2} \cdot \\ & \cdot \frac{z!}{z_1! z_2!} \left(\frac{m\mu_1}{m\mu_1 + n\mu_2}\right)^{z_1} \left(\frac{n\mu_2}{m\mu_1 + n\mu_2}\right)^{z_2} \cdot \\ & \cdot \frac{1}{z!} (m\mu_1 + n\mu_2)^z e^{-m\mu_1 - n\mu_2}. \end{aligned}$$

Tarkastellaa parametrisointia $\tau = m\mu_1 + n\mu_2$ ja $\Delta = \mu_2/\mu_1$. Ratkaisemalla parametrien μ_1 ja μ_2 suhteen, saadaan

$$\begin{cases} \tau = m\mu_1 + n\mu_2 \\ \Delta = \mu_2/\mu_1 \end{cases} \Leftrightarrow \begin{cases} \mu_1 = \frac{\tau}{m + n\Delta} \\ \mu_2 = \frac{\Delta\tau}{m + n\Delta} \end{cases}$$

Sijoittamalla $\mu_1(\tau, \Delta)$ ja $\mu_2(\tau, \Delta)$ uskottavuusfunktioon, saadaan

$$\begin{aligned}
L(\tau, \Delta) &= \frac{z_1!}{\prod_{i=1}^{m} y_i!}\left(\frac{1}{m}\right)^{z_1} \cdot \frac{z_2!}{\prod_{i=m+1}^{m+n} y_j!}\left(\frac{1}{n}\right)^{z_2} \cdot \\
&\quad \cdot \frac{z!}{z_1! z_2!}\left(\frac{m}{m+n\Delta}\right)^{z_1}\left(\frac{n\Delta}{m+n\Delta}\right)^{z_2} \cdot \\
&\quad \cdot \frac{1}{z!}\tau^z e^{-\tau} \\
&= \frac{z_1!}{\prod_{i=1}^{m} y_i!}\left(\frac{1}{m}\right)^{z_1} \cdot \frac{z_2!}{\prod_{i=m+1}^{m+n} y_j!}\left(\frac{1}{n}\right)^{z_2} \cdot \\
&\quad \cdot \frac{z!}{z_1! z_2!}\left(\frac{1}{1+\frac{n}{m}\Delta}\right)^{z_1}\left(\frac{\frac{n}{m}\Delta}{1+\frac{n}{m}\Delta}\right)^{z_2} \cdot \\
&\quad \cdot \frac{1}{z!}\tau^z e^{-\tau} \\
&= \frac{z_1!}{\prod_{i=1}^{m} y_i!}\left(\frac{1}{m}\right)^{z_1} \cdot \frac{z_2!}{\prod_{i=m+1}^{m+n} y_j!}\left(\frac{1}{n}\right)^{z_2} \cdot \\
&\quad \cdot \frac{1}{z!}\tau^z e^{-\tau} \cdot \binom{z}{z_1}\theta^{z_1}(1-\theta)^{z-z_1},
\end{aligned}$$

jossa $z = z_1 + z_2$ ja

$$\theta = \frac{1}{1+\frac{n}{m}\Delta} = \left(1 + \frac{n}{m}\Delta\right)^{-1}.$$

b) Parametrin τ suurimman uskottavuuden estimaatti saadaan maksimoimalla Poisson-jakautuneen havainnon uskottavuusfunktio

$$L(\tau) = \frac{\tau^z}{z!}e^{-\tau}.$$

Helposti nähdään, että estimaatiksi saadaan $\hat{\tau} = z$. Parametrin θ suurimman uskottavuuden estimaatti saadaan maksimoimalla binomijakauman pistetodennäköisyys (uskottavuusfunktio)

$$L(\theta) = \binom{z}{z_1}\theta^{z_1}(1-\theta)^{z-z_1}.$$

Tunnetusti $\hat{\theta} = z_1/z$. Koska

$$\theta = \frac{1}{1+\frac{n}{m}\Delta} \Leftrightarrow \Delta = \frac{m}{n} \cdot \frac{1-\theta}{\theta},$$

niin parametrin Λ suurimman uskottavuuden estimaatiksi saadaan

$$\hat{\Delta} = \frac{m}{n} \cdot \frac{1-\hat{\theta}}{\hat{\theta}} = \frac{m}{n} \cdot \frac{1-z_1/z}{z_1/z} = \frac{m}{n} \cdot \frac{z_2}{z_1}.$$

c) Olkoon

$$\hat{o} = \frac{z_2}{z_1} = \frac{z_2/z}{1 - z_2/z} = \frac{\hat{p}}{1 - \hat{p}}$$

kokeeseen $z_2 \sim Bin(z, p)$ liittyvä estimoitu vedonlyöntisuhde (odds). Luentomonisteen perusteella (kappale 4.1), vedonlyöntisuhteen o likimääräinen 95% luottamusväli on

$$\hat{o} \times \exp\left\{\pm 1.96 \times \sqrt{\frac{1}{n\hat{p}(1-\hat{p})}}\right\} = \frac{z_2}{z_1} \times \exp\left\{\pm 1.96 \times \sqrt{\frac{1}{z(z_2/z)(z_1/z)}}\right\}$$
$$= \frac{z_2}{z_1} \times \exp\left\{\pm 1.96 \times \sqrt{\frac{1}{z_1} + \frac{1}{z_2}}\right\}$$

Koska $\Delta = (m/n)o$, niin Δ:n likimääräinen 95% luottamusväli on

$$\frac{m}{n} \cdot \frac{z_2}{z_1} \times \exp\left\{\pm 1.96 \times \sqrt{\frac{1}{z_1} + \frac{1}{z_2}}\right\}$$

3. Olkoon $y_i \sim Poi(\mu_i)$, $i = 1, \ldots, n$, riippumaton satunnaisotos, jossa $g(\mu_i) = \log(\mu_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i$. Tällöin logaritminen uskottavuusfunktio on

$$l(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta})) = \text{const} + \sum_{i=1}^{n} \Big\{ y_i \log(\mu_i) - \mu_i \Big\}.$$

Johda luentomonisteen kappaleen 5.3 tulos

$$\boldsymbol{s}(\boldsymbol{\beta}) = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\mu}) \;\text{ ja }\; \boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X},$$

jossa

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T \;\text{ ja }\; \boldsymbol{W} = \text{diag}(\mu_1, \ldots, \mu_n).$$

Ratkaisu: Johdetaan pistemääräfunktio derivoimalla logaritminen uskottavuusfunktio vektorin $\boldsymbol{\beta}$ komponenttien suhteen:

$$\begin{aligned}
s_j(\boldsymbol{\beta}) &= \frac{\partial}{\partial \beta_j} l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big\{ y_i \frac{\partial}{\partial \beta_j} \log(\mu_i) - \frac{\partial}{\partial \beta_j} \mu_i \Big\} \\
&= \sum_{i=1}^{n} \Big\{ y_i \frac{\frac{\partial}{\partial \beta_j}\mu_i}{\mu_i} - \frac{\partial}{\partial \beta_j}\mu_i \Big\} = \sum_{i=1}^{n} \Big\{ y_i \frac{\frac{\partial}{\partial \beta_j} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)} - \frac{\partial}{\partial \beta_j} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i) \Big\} \\
&= \sum_{i=1}^{n} \Big\{ y_i \frac{x_{ij} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)} - x_{ij} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i) \Big\} = \sum_{i=1}^{n} \Big\{ y_i x_{ij} - x_{ij} \mu_i \Big\} \\
&= \sum_{i=1}^{n} x_{ij}(y_i - \mu_i).
\end{aligned}$$

Derivoidaan toiseen kertaan:

$$\begin{aligned} i_{jk}(\boldsymbol{\beta}) &= -\frac{\partial}{\partial \beta_k} s_j(\boldsymbol{\beta}) = -\frac{\partial}{\partial \beta_k} \sum_{i=1}^{n} x_{ij}(y_i - \mu_i) = \sum_{i=1}^{n} x_{ij} \frac{\partial}{\partial \beta_k} \mu_i \\ &= \sum_{i=1}^{n} x_{ij} \frac{\partial}{\partial \beta_k} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i) = \sum_{i=1}^{n} x_{ij} x_{ik} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i) \\ &= \sum_{i=1}^{n} x_{ij} x_{ik} \mu_i \end{aligned}$$

Kirjoitetaan matriisimuodossa:

$$\boldsymbol{s}(\boldsymbol{\beta}) = \begin{pmatrix} \sum_{i=1}^{n} x_{i1}(y_i - \mu_i) \\ \vdots \\ \sum_{i=1}^{n} x_{ip}(y_i - \mu_i) \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_{(1)}^T (\boldsymbol{y} - \boldsymbol{\mu}) \\ \vdots \\ \boldsymbol{x}_{(p)}^T (\boldsymbol{y} - \boldsymbol{\mu}) \end{pmatrix} = \boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{\mu}),$$

jossa $\boldsymbol{x}_{(i)}^T$ on matriisin $\boldsymbol{X}^T$ i:s rivi.

$$\begin{aligned} \boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) &= \begin{pmatrix} \sum_{i=1}^{n} x_{i1} x_{i1} \mu_i & \cdots & \sum_{i=1}^{n} x_{i1} x_{ip} \mu_i \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ip} x_{i1} \mu_i & \cdots & \sum_{i=1}^{n} x_{ip} x_{ip} \mu_i \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{x}_{(1)}^T \\ \vdots \\ \boldsymbol{x}_{(p)}^T \end{pmatrix} \begin{pmatrix} x_{11}\mu_1 & \cdots & x_{1p}\mu_1 \\ \vdots & \ddots & \vdots \\ x_{n1}\mu_n & \cdots & x_{np}\mu_n \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{x}_{(1)}^T \\ \vdots \\ \boldsymbol{x}_{(p)}^T \end{pmatrix} \begin{pmatrix} \mu_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mu_n \end{pmatrix} \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}, \end{aligned}$$

jossa $\boldsymbol{W} = \operatorname{diag}(\mu_1, \ldots, \mu_n)$.

4. Tarkastellaan koetta, jossa hiirille annettiin 18 kuukauden ajan hyvin pieniä annoksia tunnettua karsinogeeniä 2-asetamidofluoreenia (2-AAF). Seuraavassa taulukossa on annettu maksasyöpään sairastuneiden hiirien lukumäärät eri annostasoilla (parts per 10^4). Lähde: Zhang, H. & Zelterman, D. (1999). Binary Regression for Risks in Excess of Subject-Specific Thresholds, Biometrics, 55, pp. 1247–1251.

	Maksasyöpä	
Annos	Altistuneet	Tapaukset
0.00	555	6
0.30	2014	34
0.35	1102	20
0.45	550	15
0.60	441	13
0.75	382	17
1.00	213	19
1.50	211	24

Aineisto löytyy tiedostosta `liver.dat`.

a) Sovita aineistoon logistinen regressiomallio ja tulkitse tulokset. Paraneeko malli, jos selittäjäksi lisätään myös annoksen neliö?

b) Sovita aineistoon myös Poissonin regressiomalli log-linkkifunktiolla. Vertaa tuloksia a)-kohdan tuloksiin.

Vihje: Nyt oletetaan, että sairastuneiden lukumäärä on Poisson-jakautunut odotusarvolla μ_i ja

$$\log(\mu_i/n_i) = \boldsymbol{x}_i'\boldsymbol{\beta} \;\Leftrightarrow\; \log(\mu_i) = \log(n_i) + \boldsymbol{x}_i'\boldsymbol{\beta},$$

jossa n_i altistuneiden lukumäärä annoksella i. Mallissa oleva termi $\log(n_i)$ pakotetaan malliin regressiokertoimella yksi. R-ohjelmiston funktiossa glm tämä termi ilmoitetaan funktion argumentilla `offset`:

```
> mod<-glm(y~x1+x2,offset=log(n),family=poisson,data=aineisto)
```

Edellä oletetaan siis, että y on Poisson jakautunut ja $\log(\mu_i/n_i) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2$.

c) Tutki mallien sopivuutta esimerkiksi devianssien ja erilaisten graafisten kuvioiden avulla.

Ratkaisu:

a) Sovitetaan aineistoon logistinen regressiomalli:

```
> liver<-read.table("liver.dat",header=TRUE)
> liver
  dose    n  y
1 0.00  555  6
2 0.30 2014 34
3 0.35 1102 20
4 0.45  550 15
5 0.60  441 13
6 0.75  382 17
7 1.00  213 19
8 1.50  211 24
> mod1<-glm(cbind(y,n-y)~dose,family=binomial,data=liver)
> summary(mod1)

Call:
glm(formula = cbind(y, n - y) ~ dose, family = binomial, data = liver)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.89011  -0.46016  -0.08732   0.48732   1.65965

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.5083     0.1478 -30.501   <2e-16 ***
dose          1.7624     0.1864   9.457   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 77.6029  on 7  degrees of freedom
Residual deviance:  4.5217  on 6  degrees of freedom
AIC: 45.494

Number of Fisher Scoring iterations: 4

> exp(coefficients(mod1))
(Intercept)        dose
 0.01101739  5.82642544
```

Yhden yksikön lisäys annoksessa lisää kuolemisen oddsia 5.8-kertaiseksi. Lasketaan vielä likimääräiset 95% luottamusvälit:

```
> exp(confint.default(mod1))
                  2.5 %     97.5 %
(Intercept) 0.008246446 0.01471943
```

```
dose        4.043607499 8.39528402
```

Lisätään malliin annoksen neliö:

```
> mod2<-glm(cbind(y,n-y)~dose+I(dose^2),family=binomial,data=liver)
> summary(mod2)

Call:
glm(formula = cbind(y, n - y) ~ dose + I(dose^2), family = binomial,
    data = liver)

Deviance Residuals:
      1        2        3        4        5        6        7        8
 0.7754  -0.2081  -0.3916   0.3387  -0.5947  -0.1516   1.0935  -0.2846

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.8520     0.2958 -16.404  < 2e-16 ***
dose          2.9539     0.8833   3.344 0.000825 ***
I(dose^2)    -0.6980     0.5044  -1.384 0.166368
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 77.6029  on 7  degrees of freedom
Residual deviance:  2.5659  on 5  degrees of freedom
AIC: 45.539

Number of Fisher Scoring iterations: 4

> anova(mod1,mod2,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(y, n - y) ~ dose
Model 2: cbind(y, n - y) ~ dose + I(dose^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         6     4.5217
2         5     2.5659  1   1.9558    0.162
```

Annoksen toinen potenssi ei paranna mallia tilastollisesti merkitsevästi (p-arvo=0.162).

b) Sovitetaan aineistoon Poissonin regressiomalli log-linkkifunktiolla:

```
> mod3<-glm(y~dose,offset=log(n),family=poisson,data=liver)
> summary(mod3)
```

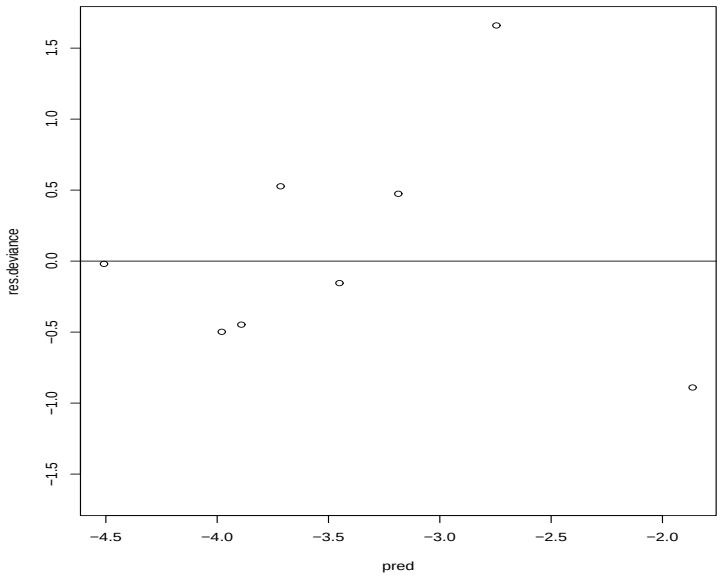
```
Call:
glm(formula = y ~ dose, family = poisson, data = liver, offset = log(n))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.89303  -0.47004  -0.09812   0.53392   1.67950

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.4919     0.1445 -31.096   <2e-16 ***
dose          1.6634     0.1767   9.414   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 74.4827  on 7  degrees of freedom
Residual deviance:  4.6882  on 6  degrees of freedom
AIC: 46.024

Number of Fisher Scoring iterations: 4

> exp(coefficients(mod3))
(Intercept)        dose
 0.01119887  5.27705573
```

Huomataan, että Poisson-malli antoi lähes samat regressiokerroinestimaatit kuin logistinen regressiomalli. Yhden yksikön lisäys annoksessa lisää kuolemisen riskiä 5.3-kertaiseksi. Lasketaan lopuksi likimääräiset 95% luottamusvälit:

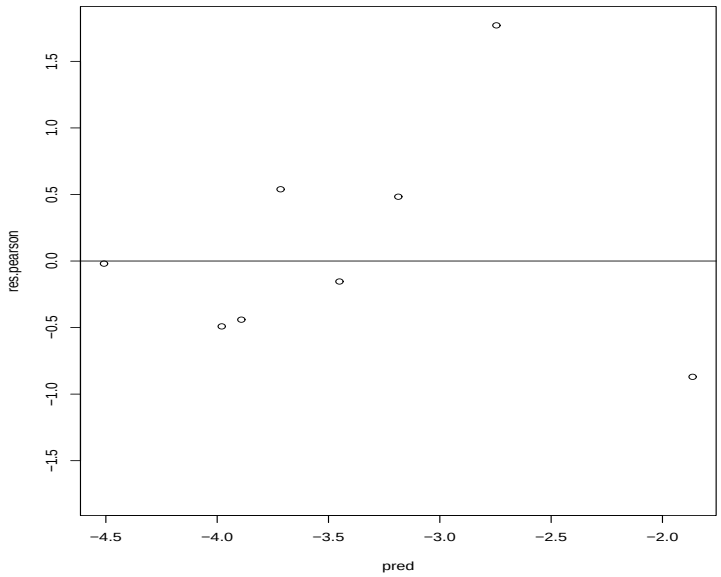
```
> exp(confint.default(mod3))
                  2.5 %     97.5 %
(Intercept) 0.008437521 0.01486394
dose        3.732459133 7.46084985
```

c) Tehdään residuaalikuviot:

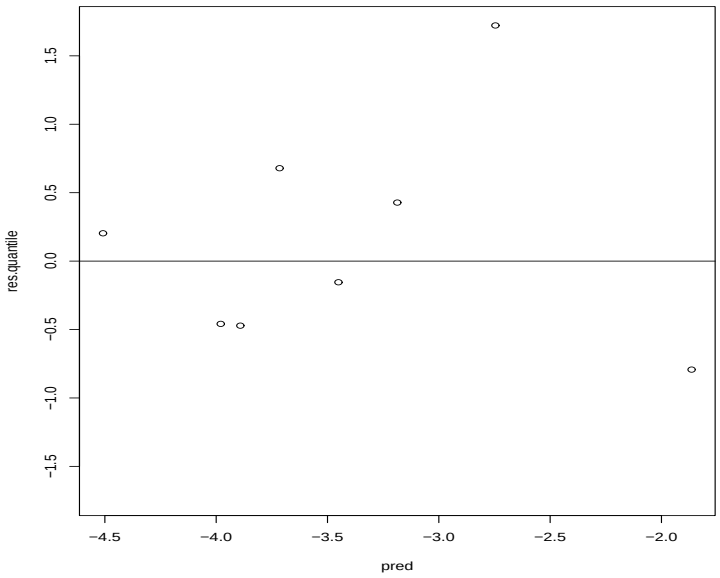
```
> res.deviance<-residuals(mod1)
> pred<-predict(mod1)
> plot(pred,res.deviance)
> res.pearson<-residuals(mod1,type ="pearson")
> plot(pred,res.pearson)
> library(statmod)
> res.quantile<-qresiduals(mod1)
> plot(pred,res.quantile)
```



Kuva 1: Residuaalidevianssit vs lineaarinen prediktori



Kuva 2: Pearsonin devianssit vs lineaarinen prediktori



Kuva 3: Satunnaistetut kvantiiliresiduaalit vs lineaarinen prediktori

Residuaaliploteissa erottuu kaksi selkeästi erottuvaa pistettä. Ne voivat johtua mallin sopimattomuudesta.