78185 Yleistetyt lineaariset mallit

Harjoitus 5, syksy 2014

1. Tutkimuksessa selvitettiin ihmisten asenteita maahanmuuttoa kohtaan. Vastaajilta kysyttiin muun muassa seuraavia 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
```

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

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

Käytettäessä uudelleenparametrisointia

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

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.

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

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

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.