78185 Yleistetyt lineaariset mallit

Harjoitus 4, syksy 2014

1. Tarkastellaan havaintoaineistoa `esoph.dat`. Kyseessä on käytännössä sama aineisto kuin harjoituksen 3 tehtävässä 3, ainoastaan muuttujien arvot on muutettu numeerisiksi.

   a) Muuta muuttujat `agegp`, `alcgp` ja `tobgp` kaksiluokkaisiksi 0/1-muuttujiksi ja tutki ruokatorvensyöpään sairastumisen vaaraa logistisen regressioanalyysin avulla, kun selittävinä muuttujina ovat 0/1-muuttujat ikä, tupakointi ja alkoholinkäyttö.

   b) Plottaa residuaalidevianssit (tai Pearsonin residuaalit) lineaarisen prediktorin estimaatteja ($\hat{\eta}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$) vastaan. Jos logistinen malli on kunnossa, niin suurin osa pisteistä pitäisi sijaita tasaisesti vaaka-akselin molemmilla puolilla rajojen -3 ja $+3$ välissä. Huom! Tämä plotti ei ole kovin hyödyllinen, jos n_i:t ovat liian pieniä.

   c) Testaa linkkifunktion hyvyyttä seuraavan yksinkertaisen testin avulla (ks. esim. McCullagh & Nelder, s. 401). Sovita logistinen regressiomalli ja tallenna estimoidut lineaariset prediktorit $\hat{\eta}_i$. Lisää muuttuja $\hat{\eta}_i^2$ malliin. Jos $\hat{\eta}_i^2$ on tilastollisesti merkitsevä, niin linkkifunktio on väärin määritelty.

   Avuksi:

```
> mod1 <- glm(cbind(n1,n0)~x1+x2*x3, data=tied, family=binomial)
> res1<-residuals(mod1)  #residuaalidevianssit
> res2<-residuals(mod1, type = "pearson") #Pearsonin residuaalit
> pred<-mod1$linear.predictors #lineaaristen prediktorien estimaatit
```

2. Tuholaismyrkyn tehoa on tutkittu suihkuttamalla myrkkyä lehtikirvojen päälle ja laskemalla jälkeenpäin kuolleiden kirvojen lukumäärät. Kirvat jaettiin kymmeneen 200 yksilön ryhmään, joita käsiteltiin eri vahvuisilla myrkkyseoksilla. Aineisto löytyy tiedostosta `myrkky.dat`.

   a) Tutki graafisesti logistisen regressiomallin sopivuutta aineistoon. Paraneeko sopivuus, jos käytetään muuttujan x sijasta muuttujaa $logx = \log(x)$?
   Vihje: Jos malli on kunnossa, niin pisteiden

   $$\left(x_i, \log\left(\frac{p_i}{1-p_i} \right) \right)$$

   pitäisi sijaita likimain suoralla. Todennäköisyydet p_i voi laskea suoraan havaintoaineistosta.

   b) Tutki aineistoa logistisen regressiomallin avulla (esim. summary, anova).

   c) Tutki b)-kohdassa saadun mallin sopivuutta graafisesti vertaamalla ennustettuja arvoja $\hat{p}_i$ todellisiin arvoihin $p_i = r_i/n_i$ (eri $logx$:n arvoilla).

Annos (mg/l)	Hyönteisten lkm	Kuolleiden lkm
x	n	r
1.0	200	2
2.0	200	6
3.0	200	33
4.0	200	68
5.0	200	108
6.0	200	137
7.0	200	152
8.0	200	162
9.0	200	178
10.0	200	186

3. 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:

Multinomimalli.

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

```
> 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)

> 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
```