

Yleistetyt lineaariset mallit

10. harjoitukset, 13. vko 2002

10.1 # Aspiriinin käytön ja vatsahaavan yhteys

```
ulcer <- read.table("C:\\Kurssit\\Glim\\Glim02\\Datat\\ulcer.aspirin.txt", header =
attach(ulcer)
names(ulcer)
# [1] "ulcer"          "case.control" "aspirin"      "frequency"
```

```
ulcer.mhr<-glm(frequency ~ (aspirin+ulcer+case.control)^2,
data=ulcer, family=poisson)
```

```
#####
summary(ulcer.mhr)
```

Call:

```
glm(formula = frequency ~ (aspirin + ulcer + case.control)^2,
     family = poisson, data = ulcer)
```

Deviance Residuals:

1	2	3	4	5	6	7	8
0.4486	-1.2085	-0.5393	0.7281	-0.4661	1.4673	0.5073	-1.0829

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.81846	0.14515	26.307	< 2e-16 ***
aspirinuser	-1.37910	0.29514	-4.673	2.97e-06 ***
ulcergastric	-0.06977	0.20415	-0.342	0.73254
case.controlcontrol	0.21517	0.19172	1.122	0.26174
aspirinuser:ulcergastric	0.70005	0.34603	2.023	0.04306 *
aspirinuser:case.controlcontrol	-1.14288	0.35207	-3.246	0.00117 **
ulcergastric:case.controlcontrol	0.10574	0.26147	0.404	0.68590

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 127.749 on 7 degrees of freedom

Residual deviance: 6.283 on 1 degrees of freedom

AIC: 59.898

```
#####3
```

```
pchisq(6.283, 1, lower.tail = F)
```

```

# Koska devianssiin liittyvä p-arvo = 0.01219016 on verrattain pieni,
# mallin yhteensopivuus ei ole kovin hyvä. Huomattakoon kuitenkin, että
# aspiriinin päävaikutus on erittäin merkitsevä ja molemmat yhdysvaikutukset
# merkitseviä.

# Tässä tapauksessahan minimaalinen malli on

ulcer.m0<-glm(frequency ~ ulcer*case.control,data=ulcer, family=poisson)

# koska muuttujien "ulcer" ja "case.control" luokissa on reunafrekvenssit
# kiinnitetty.

anova(ulcer.m0,ulcer.mhr,test="Chisq")

anova(ulcer.m0,ulcer.mhr,test="Chisq")
Analysis of Deviance Table

Model 1: frequency ~ ulcer * case.control
Model 2: frequency ~ (aspirin + ulcer + case.control)^2
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1          4    126.708
2          1     6.283  3  120.425 6.251e-26

# Malli ulcer.mhr on erittäin merkitsevästi minimaalinen mallia parempi.

anova(ulcer.m0,ulcer.mhr,test="Chisq")

ulcer.mpv<-glm(frequency ~ aspirin+ulcer+case.control,data=ulcer, family=poisson)
anova(ulcer.mpv,ulcer.mhr,test="Chisq")
Analysis of Deviance Table

Model 1: frequency ~ aspirin + ulcer + case.control
Model 2: frequency ~ (aspirin + ulcer + case.control)^2
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1          4    21.7901
2          1     6.2830  3  15.5071  0.0014

# Malli ulcer.mhr on myös erittäin merkitsevästi päävaikutusten mallia ulcer.m0
# Aspiriinin käyttö näyttää siis selvästi olevan riskitekijä haavan synnylle
# Sen sijaan haavan sijaintiin vaikutus ei ole kovin merkitsevää (p-arvo = 0.0014)
# Tämä johtunee testin "heikkoudesta".

```

10.2

$$\begin{aligned} \log\left(\frac{\mu_{i2k}}{\mu_{i1k}}\right) &= \lambda + \lambda_i^X + \lambda_2^V + \lambda_k^Z + \lambda_{i2}^{XV} + \lambda_{ik}^{XZ} + \lambda_{2k}^{VZ} \\ &- (\lambda + \lambda_i^X + \lambda_1^V + \lambda_k^Z + \lambda_{i1}^{XV} + \lambda_{ik}^{XZ} + \lambda_{1k}^{VZ}) \\ &= \lambda_2^V + \lambda_{i2}^{XV} + \lambda_{2k}^{VZ}, \end{aligned}$$

joten

$$\log\left(\frac{\hat{\mu}_{i2k}}{\hat{\mu}_{i1k}}\right) = -1.37910 - 1.14288 + 0.70005.$$

10.3 Devianssi $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ voidaan esittää muodossa

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^n d_i(y_i; \hat{\mu}_i),$$

missä

$$d_i = -2 \int_{y_i}^{\hat{\mu}_i} \frac{y_i - t}{\varphi V(t)} dt.$$

Normaalijakauman tapauksessa $V(\mu_i) = 1$ ja $\varphi = \sigma^2$ (vakio). Silloin

$$\begin{aligned} d_i &= -2 \int_{y_i}^{\hat{\mu}_i} \frac{y_i - t}{\sigma^2} dt = -\frac{2}{\sigma^2} \Big|_{y_i}^{\hat{\mu}_i} (y_i t - t^2/2) \\ &= -\frac{2}{\sigma^2} (y_i \hat{\mu}_i - \hat{\mu}_i^2/2 - y_i^2 + y_i^2/2) \\ &= \frac{1}{\sigma^2} (y_i^2 - 2y_i \hat{\mu}_i - \hat{\mu}_i^2) = \frac{(y_i^2 - \hat{\mu}_i)^2}{\sigma^2}, \end{aligned}$$

joten

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^n \frac{(y_i^2 - \hat{\mu}_i)^2}{\sigma^2}.$$

10.4 `y <- scan()`

```
0 0 3 0 1 1 1 2 2
4 2 8 0 3 4 5 2 2
2 5 4 3 15 12 7 14 6
10 14 8 19 10 7 20 10 19
```

```
i <- 1:36
```

```
aids.reg <- glm(y ~ i, family=poisson)
```

```
aids.regq <- glm(y ~ i, quasipoisson(link = "log"))
```

```
# tai
aids.regq <- glm(y ~ i,family=quasi(link = "log", variance = "mu"))
```

```
summary(aids.reg)
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.039661	0.211933	0.187	0.852
i	0.079569	0.007708	10.323	<2e-16 ***

```
summary(aids.regq)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.03966	0.28619	0.139	0.89
i	0.07957	0.01041	7.644	6.94e-09 ***

Huomataan, että kvasiuskottavuusfunktion avulla estimoidussa mallissa ker
hajonnat ovat suurempia. Hajontojen suhde on

$0.01041/0.007708 = 1.350545,$

joten Y:n jakaumassa on ylihajontaa.

Jos ylihajontaa ei oteta huomioon, kertoimet "näyttävät" merkitsevämmiltä kuin ne todellisuudessa ovat.

Skaalausparametri ϕ =

$(0.01041/0.007708)^2 = 1.823971$

```
10.5 ldose <- rep(0:5, 2)
```

```
numdead <- c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
```

```
number <- rep(20,12) # painomuuttuja
```

```
y <- numdead/number # vastemuuttuja
```

```
budworm.lg <- glm(y ~ ldose,weight=number, family=binomial)
```

```
budworm.lgq <- glm(y ~ ldose,weight=number, family=quasibinomial(link = "log
```

```
summary(budworm.lg)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.7661	0.3701	-7.473	7.82e-14 ***
ldose	1.0068	0.1236	8.147	3.74e-16 ***

```
summary(budworm.lgq)
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-2.7661	0.4497	-6.151	0.000108	***
ldose	1.0068	0.1502	6.705	5.33e-05	***

```
# Kertoimien hajontojen suhde
```

```
0.1502/0.1236=1.215210,
```

```
joten hieman ylihajontaa.
```

```
Jos ylihajontaa ei oteta huomioon, kertoimet "näyttävät" merkitsevämmiltä kuin ne todellisuudessa ovat.
```

```
Skaalausparametri
```

```
phi = (0.1502/0.1236)^2= 1.476736
```