78185 Yleistetyt lineaariset mallit

Harjoitus 2, syksy 2014

Esimerkkiratkaisut

1. Seuraavassa taulukossa on ristiintaulukoituna syntymäpaino (**matala** syntymäpaino, **normaali** syntymäpaino) ja vauvojen kuolleisuus (**kuollut** vuoden sisällä syntymästä, **elossa** vuoden kuluttua syntymästä). Estimoi riskisuhde ja ristitulosuhde (OR) kuolemiselle vuoden sisällä syntymästä. Laske myös likimääräiset 95% luottamusvälit.

	Kuollut	Elossa	Yhteensä
Matala	618	4597	5215
Normaali	422	67093	67515
Yhteensä	1040	71690	72730

Ratkaisu:

Vaaran estimaatti matalapainoisille:

$$\hat{p}_1 = \frac{618}{5215} \approx 0.12.$$

Vaaran estimaatti normaalipainoisille:

$$\hat{p}_2 = \frac{422}{67515} \approx 0.0063.$$

Riskisuhteen estimaatti on

$$\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{618/5215}{422/67515} \approx 19.0.$$

Ristitulosuhteen estimaatti on

$$\widehat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_1)} = \frac{(618/5215)/(4597/5215)}{(422/67515)/(67093/67515)} = \frac{618 \cdot 67093}{422 \cdot 4597} \approx 21.4.$$

Luentomonisteen perusteella likimääräiset 95% luottamusvälit lasketaan kaavoilla

$$\widehat{RR} \times \exp\left\{\pm 1.96\sqrt{\frac{(1-\hat{p}_1)}{\hat{p}_1 n_1} + \frac{(1-\hat{p}_2)}{\hat{p}_2 n_2}}\right\}$$

ja

$$\widehat{OR} \times \exp\left\{\pm 1.96\sqrt{\frac{1}{\hat{p}_1(1-\hat{p}_1)n_1} + \frac{1}{\hat{p}_2(1-\hat{p}_2)n_2}}\right\}.$$

Kun sijoitetaan estimaattien arvot kaavoihin, niin saadaan

- riskisuhteen likimääräinen 95% luottamusväli: $(16.8, 21.4)$
- ristitulosuhteen likimääräinen 95% luottamusväli: $(18.8, 24.2)$.

.

```
> n11<-618
> n12<-4597
> n21<-422
> n22<-67093
> p1<-n11/(n11+n12)
> p1
[1] 0.1185043
> p2<-n21/(n21+n22)
> p2
[1] 0.006250463
> RR<-p1/p2
> RR
[1] 18.95929
> OR<-(p1/(1-p1))/(p2/(1-p2))
> OR
[1] 21.37365
> RR.lo<-RR*exp(-1.96*sqrt((1-p1)/(p1*(n11+n12))+(1-p2)/(p2*(n21+n22))))
> RR.lo
[1] 16.80658
> RR.hi<-RR*exp(+1.96*sqrt((1-p1)/(p1*(n11+n12))+(1-p2)/(p2*(n21+n22))))
> RR.hi
[1] 21.38773
> OR.lo<-OR*exp(-1.96*sqrt(1/(p1*(1-p1)*(n11+n12))+1/(p2*(1-p2)*(n21+n22))))
> OR.lo
[1] 18.81832
> OR.hi<-OR*exp(+1.96*sqrt(1/(p1*(1-p1)*(n11+n12))+1/(p2*(1-p2)*(n21+n22))))
> OR.hi
[1] 24.27597

> n1<-c(618,422)
> n0<-c(4597,67093)
> bwt<-c(1,0)
> bwt<-factor(bwt,levels=c(0,1),labels=c("normal","low"))
> mod<-glm(cbind(n1,n0)~bwt,family=binomial)
> summary(mod)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) -5.06883    0.04883 -103.80   <2e-16 ***
bwtlow       3.06216    0.06496   47.14   <2e-16 ***
...
> coefficients(mod)[2]
     bwt
3.062159
> exp(coefficients(mod)[2])
     bwt
21.37365
> b<-coefficients(mod)
> b
(Intercept)      bwtlow
  -5.068830    3.062159
> exp(sum(b))/(1+exp(sum(b)))
[1] 0.1185043
> exp(b[1])/(1+exp(b[1]))
(Intercept)
0.006250463
> predict(mod,type="response")
          1           2
0.118504314 0.006250463
> confint(mod)
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) -5.166073 -4.974607
bwtlow       2.935259  3.189984
> exp(confint(mod))
Waiting for profiling to be done...
                   2.5 %       97.5 %
(Intercept)  0.005706935  0.006911237
bwtlow      18.826378323 24.288048210
```

2. Prostatasyöpäpotilaan hoitomenetelmä riippuu suuresti siitä onko syöpä levinnyt ympäröiviin imusolmukkeisiin. Havaintoaineisto koostuu 53 prostatasyöpäpotilaasta, joilta on mitattu muun muassa seuraavien muuttujien arvot (muuttuja *nodalinv* on saatu leikkauksen jälkeen, muiden muuttujien arvot saatu ennen leikkausta):

   nodalinv: Syövän levinneisyys imusolmukkeisiin. 1=”kyllä”, 0=”ei”.

   grad: Neulalla otetun solunäytteen perusteella tehty arvio syövän asteesta. 1=”vakava”, 0=”vähemmän vakava”.

   xray: Röntgenkuvan perusteella tehty arvio syövän asteesta. 1 = ”vakava”, 0 = ”vähemmän vakava”.

   Seuraavassa taulukossa on ristiintaulukoituna muuttujat *nodalinv*, *grad* ja *xray*:

```
                xray  0  1
nodalinv grad
0        0          21  3
         1           8  1
1        0           5  4
         1           4  7
```

(a) Laske vaaran $p = P(nodalinv = 1)$ estimaatti syövän levinneisyydelle ja vaaran likim. 95% luottamusväli, kun oletetaan, että populaatio on vaaran suhteen homogeeninen.

(b) Laske vaarasuhteen ja ristitulosuhteen estimaatti syövän levinneisyydelle ja vastaavat 95% luottamusvälit, kun verrataan ryhmiä $grad = 0$ ja $grad = 1$.

(c) Laske vaarasuhteen ja ristitulosuhteen estimaatti syövän levinneisyydelle ja vastaavat 95% luottamusvälit, kun verrataan ryhmiä $\{grad = 0, xray = 0\}$ ja $\{grad = 1, xray = 0\}$. Huom! Muuttuja $xray$ on vakio 0.

(d) Laske vaarasuhteen ja ristitulosuhteen estimaatti syövän levinneisyydelle ja vastaavat 95% luottamusvälit, kun verrataan ryhmiä $\{grad = 0, xray = 1\}$ ja $\{grad = 1, xray = 1\}$. Huom! Muuttuja $xray$ on vakio 1.

Ratkaisu:

(a) Vaaran estimaatti on

$$\hat{p} = \frac{5+4+4+7}{21+3+8+1+5+4+4+7} = \frac{20}{53} \approx 0.38.$$

ja vaaran likimääräinen 95% luottamusväli on

$$\left(\hat{p} \times \exp\left\{ -1.96 \times \sqrt{\frac{1-\hat{p}}{n\hat{p}}} \right\}, \;\; \hat{p} \times \exp\left\{ +1.96 \times \sqrt{\frac{1-\hat{p}}{n\hat{p}}} \right\} \right) = (0.27, 0.53).$$

```
> X<-data.frame(n1=20,n0=33)
> mod1 <- glm(cbind(n1, n0)~1,family=binomial(link="log"),data=X)
> exp(coef(mod1))
(Intercept)
  0.3773585
> exp(confint.default(mod1))
                2.5 %    97.5 %
(Intercept) 0.2670335 0.5332643
> p<-20/53
> p
[1] 0.3773585
```

```
> p*exp(-1.96*sqrt((1-p)/(53*p)))
[1] 0.2670318
> p*exp(+1.96*sqrt((1-p)/(53*p)))
[1] 0.5332677
```

(b)

	grad		Yht.
nodalinv	1	0	
1	11	9	20
0	9	24	33
Yht.	20	33	53

Vaaran estimaatit ovat $\hat{p}_1 = 11/20$ ($grad = 1$) ja $\hat{p}_0 = 9/33$ ($grad = 0$).

Vaarasuhteen estimaatti on

$$\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_0} = \frac{11/20}{9/33} \approx 2.02$$

ja käyttämällä monisteen kaavaa

$$\widehat{RR} \times \exp\left\{\mp 1.96\sqrt{\frac{(1-\hat{p}_1)}{\hat{p}_1 n_1} + \frac{(1-\hat{p}_2)}{\hat{p}_2 n_2}}\right\}$$

vaarasuhteen likimääräiseksi 95% luottamusväliksi saadaan $(1.02, 4.00)$.

Ristitulosuhteen estimaatti on

$$\widehat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_0/(1-\hat{p}_0)} = \frac{11 \cdot 24}{9 \cdot 9} \approx 3.26$$

ja käyttämällä monisteen kaavaa

$$\widehat{OR} \times \exp\left\{\pm 1.96\sqrt{\frac{1}{\hat{p}_1(1-\hat{p}_1)n_1} + \frac{1}{\hat{p}_2(1-\hat{p}_2)n_2}}\right\}$$

ristitulosuhteen likimääräiseksi 95% luottamusväliksi saadaan $(1.01, 10.47)$.

```
> X<-data.frame(grad=c(1,0),n1=c(11,9),n0=c(9,24))
> mod1 <- glm(cbind(n1, n0)~grad,family=binomial(link="log"),data=X)
> exp(coef(mod1))
(Intercept)        grad
  0.2727273   2.0166667
> exp(confint.default(mod1))
                2.5 %    97.5 %
```

```
(Intercept) 0.1562282 0.4760994
grad        1.0178107 3.9957767
> p0<-9/33
> p1<-11/20
> RR<-p1/p0
> RR
[1] 2.016667
> RR*exp(-1.96*sqrt((1-p0)/(33*p0)+(1-p1)/(20*p1)))
[1] 1.017798
> RR*exp(+1.96*sqrt((1-p0)/(33*p0)+(1-p1)/(20*p1)))
[1] 3.995827
> mod2 <- glm(cbind(n1, n0)~grad,family=binomial,data=X)
> exp(coef(mod2))
(Intercept)        grad
   0.375000    3.259259
> exp(confint.default(mod2))
                2.5 %     97.5 %
(Intercept) 0.1743106  0.8067495
grad        1.0141481 10.4745755
> OR<-(p1/(1-p1))/(p0/(1-p0))
> OR
[1] 3.259259
> OR*exp(-1.96*sqrt(1/(p0*(1-p0)*33)+1/(p1*(1-p1)*20)))
[1] 1.014126
> OR*exp(+1.96*sqrt(1/(p0*(1-p0)*33)+1/(p1*(1-p1)*20)))
[1] 10.4748
```

(c)

	grad		Yht.
nodalinv	1	0	
1	4	5	9
0	8	21	29
Yht.	12	26	38

Vaaran estimaatit ovat $\hat{p}_1 = 4/12$ ($grad = 1$, $xray = 0$) ja $\hat{p}_0 = 5/26$ ($grad = 0$, $xray = 0$).

Vaarasuhteen estimaatti on

$$\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_0} = \frac{4/12}{5/26} \approx 1.73.$$

ja käyttämällä monisteen kaavaa

$$\widehat{RR} \times \exp\left\{\mp 1.96\sqrt{\frac{(1-\hat{p}_1)}{\hat{p}_1 n_1} + \frac{(1-\hat{p}_2)}{\hat{p}_2 n_2}}\right\}$$

vaarasuhteen likimääräiseksi 95% luottamusväliksi saadaan $(0.56, 5.33)$. Huom! Ykkönen kuuluu luottamusvälille, joten testattaessa nollahypoteesia $H : RR = 1$ nollahypoteesi jää voimaan merkitsevyystasolla 0.05.

Ristitulosuhteen estimaatti on

$$\widehat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_0/(1-\hat{p}_0)} = \frac{4 \cdot 21}{5 \cdot 8} \approx 2.1.$$

ja käyttämällä monisteen kaavaa

$$\widehat{OR} \times \exp\left\{\pm 1.96 \sqrt{\frac{1}{\hat{p}_1(1-\hat{p}_1)n_1} + \frac{1}{\hat{p}_2(1-\hat{p}_2)n_2}}\right\}$$

ristitulosuhteen likimääräiseksi 95% luottamusväliksi saadaan $(0.45, 9.86)$. Huom! Ykkönen kuuluu luottamusvälille, joten testattaessa nollahypoteesia $H : OR = 1$ nollahypoteesi jää voimaan merkitsevyystasolla 0.05.

```
> X<-data.frame(grad=c(1,0),n1=c(4,5),n0=c(8,21))
> mod1 <- glm(cbind(n1, n0)~grad,family=binomial(link="log"),data=X)
> exp(coef(mod1))
(Intercept)        grad
  0.1923077   1.7333333
> exp(confint.default(mod1))
                 2.5 %    97.5 %
(Intercept) 0.08747487 0.4227757
grad        0.56394385 5.3275596
> p0<-5/26
> p1<-4/12
> RR<-p1/p0
> RR
[1] 1.733333
> RR*exp(-1.96*sqrt((1-p0)/(26*p0)+(1-p1)/(12*p1)))
[1] 0.5639319
> RR*exp(+1.96*sqrt((1-p0)/(26*p0)+(1-p1)/(12*p1)))
[1] 5.327672
> mod2 <- glm(cbind(n1, n0)~grad,family=binomial,data=X)
> exp(coef(mod2))
(Intercept)        grad
  0.2380952   2.1000000
> exp(confint.default(mod2))
                2.5 %    97.5 %
(Intercept) 0.0897804 0.6314222
grad        0.4472686 9.8598461
> OR<-(p1/(1-p1))/(p0/(1-p0))
> OR
```

```
[1] 2.1
> OR*exp(-1.96*sqrt(1/(p0*(1-p0)*26)+1/(p1*(1-p1)*12)))
[1] 0.4472559
> OR*exp(+1.96*sqrt(1/(p0*(1-p0)*26)+1/(p1*(1-p1)*12)))
[1] 9.860128
```

(d)

	grad		Yht.
nodalinv	1	0	
1	7	4	11
0	1	3	4
Yht.	8	7	15

Vaaran estimaatit ovat $\hat{p}_1 = 7/8$ ($grad = 1$, $xray = 1$) ja $\hat{p}_0 = 4/7$ ($grad = 0$, $xray = 1$).

Vaarasuhteen estimaatti on

$$\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_0} = \frac{7/8}{4/7} \approx 1.53.$$

ja käyttämällä monisteen kaavaa

$$\widehat{RR} \times \exp\left\{ \mp 1.96 \sqrt{\frac{(1-\hat{p}_1)}{\hat{p}_1 n_1} + \frac{(1-\hat{p}_2)}{\hat{p}_2 n_2}} \right\}$$

vaarasuhteen likimääräiseksi 95% luottamusväliksi saadaan $(0.77, 3.06)$.

Ristitulosuhteen estimaatti on

$$\widehat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_0/(1-\hat{p}_0)} = \frac{7 \cdot 3}{4 \cdot 1} \approx 5.25.$$

ja käyttämällä monisteen kaavaa

$$\widehat{OR} \times \exp\left\{ \pm 1.96 \sqrt{\frac{1}{\hat{p}_1(1-\hat{p}_1)n_1} + \frac{1}{\hat{p}_2(1-\hat{p}_2)n_2}} \right\}$$

ristitulosuhteen likimääräiseksi 95% luottamusväliksi saadaan $(0.40, 68.95)$.

```
> X<-data.frame(grad=c(1,0),n1=c(7,4),n0=c(1,3))
> mod1 <- glm(cbind(n1, n0)~grad,family=binomial(link="log"),data=X)
> exp(coef(mod1))
(Intercept)        grad
  0.5714286   1.5312500
> exp(confint.default(mod1))
                2.5 %   97.5 %
(Intercept) 0.3008436 1.085383
grad        0.7657747 3.061901
> p0<-4/7
> p1<-7/8
```

```
> RR<-p1/p0
> RR
[1] 1.53125
> RR*exp(-1.96*sqrt((1-p0)/(7*p0)+(1-p1)/(8*p1)))
[1] 0.7657648
> RR*exp(+1.96*sqrt((1-p0)/(7*p0)+(1-p1)/(8*p1)))
[1] 3.061941
> mod2 <- glm(cbind(n1, n0)~grad,family=binomial,data=X)
> exp(coef(mod2))
(Intercept)        grad
   1.333333    5.250000
> exp(confint.default(mod2))
                2.5 %    97.5 %
(Intercept) 0.2984165  5.957371
grad        0.3997715 68.945631
> OR<-(p1/(1-p1))/(p0/(1-p0))
> OR
[1] 5.25
> OR*exp(-1.96*sqrt(1/(p0*(1-p0)*7)+1/(p1*(1-p1)*8)))
[1] 0.3997526
> OR*exp(+1.96*sqrt(1/(p0*(1-p0)*7)+1/(p1*(1-p1)*8)))
[1] 68.9489
```

3. Olkoot $\{y_1, \ldots, y_{n_1}\}$, $\{y_{n_1+1}, \ldots, y_{n_1+n_2}\}$ ja $\{y_{n_1+n_2+1}, \ldots, y_{n_1+n_2+n_3}\}$ riippumattomat satunnaisotokset Bernoullijakaumista parametreilla p_1, p_2 ja p_3. Siis p_1, p_2 ja p_3 ovat populaatioihin liittyvät vaarat. Tilannetta vastaa logistinen regressiomalli

$$\text{logit}(p_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}, \quad i = 1, 2, 3,$$

jossa

$$x_{i2} = \begin{cases} 1 & \text{jos } i = 2, \\ 0 & \text{jos } i \neq 2. \end{cases} \quad \text{ja} \quad x_{i3} = \begin{cases} 1 & \text{jos } i = 3, \\ 0 & \text{jos } i \neq 3. \end{cases}$$

Johda parametrin $\boldsymbol{\beta}$ suurimman uskottavuuden estimaatti.

Vihje: 1) Satunnaisotokset ovat riippumattomia, 2) Laske parametrien p_1, p_2 ja p_3 su-estimaatit ja käytä su-estimaatin invarianssiominaisuutta (Tilastollisen päättelyn kurssilta tuttu asia).

Ratkaisu:

$$\boldsymbol{S}(\boldsymbol{\beta}) \;=\; \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\mu}) = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 - n_1 p_1 \\ y_2 - n_2 p_2 \\ y_3 - n_3 p_3 \end{pmatrix} = \begin{pmatrix} \sum y_i - \sum n_i p_i \\ y_2 - n_2 p_2 \\ y_3 - n_3 p_3 \end{pmatrix} \doteq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\Longrightarrow \hat{p}_i = y_i/n_i, i = 1, 2, 3.$$

Koska

$$\begin{aligned}
\log\left(\frac{p_1}{1-p_1}\right) &= \boldsymbol{\beta}^T\boldsymbol{x}_1 = \beta_1, \\
\log\left(\frac{p_2}{1-p_2}\right) &= \boldsymbol{\beta}^T\boldsymbol{x}_2 = \beta_1 + \beta_2, \\
\log\left(\frac{p_3}{1-p_3}\right) &= \boldsymbol{\beta}^T\boldsymbol{x}_3 = \beta_1 + \beta_3,
\end{aligned}$$

niin regressiokertoimiksi saadaan

$$\begin{aligned}
\beta_1 &= \log\left(\frac{p_1}{1-p_1}\right), \\
\beta_2 &= \log\left(\frac{p_2}{1-p_2}\right) - \log\left(\frac{p_1}{1-p_1}\right) = \log\left(\frac{p_2}{1-p_2}\Big/\frac{p_1}{1-p_1}\right), \\
\beta_3 &= \log\left(\frac{p_3}{1-p_3}\right) - \log\left(\frac{p_1}{1-p_1}\right) = \log\left(\frac{p_3}{1-p_3}\Big/\frac{p_1}{1-p_1}\right).
\end{aligned}$$

Parametrin $\boldsymbol{\beta}$ suurimman uskottavuuden estimaatti on tällöin

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \log\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) \\ \log\left(\frac{\hat{p}_2}{1-\hat{p}_2}\Big/\frac{\hat{p}_1}{1-\hat{p}_1}\right) \\ \log\left(\frac{\hat{p}_3}{1-\hat{p}_3}\Big/\frac{\hat{p}_1}{1-\hat{p}_1}\right) \end{pmatrix}$$

Huom! $\exp(\hat{\beta}_2)$ on estimaatti ristitulosuhteelle

$$\frac{p_2}{1-p_2}\Big/\frac{p_1}{1-p_1}$$

ja $\exp(\hat{\beta}_3)$ on estimaatti ristitulosuhteelle

$$\frac{p_3}{1-p_3}\Big/\frac{p_1}{1-p_1}.$$