STK3100 Exercises, Week 10

Vinnie Ko, Jonas Moss

November 1, 2018

Exercise 6.15

a)

Consider proportional odds with a unit increase in $\boldsymbol{x}_i$ (i.e. adding 1 to every element of the vector $\boldsymbol{x}_i$):

$$\frac{\frac{\omega_{i',j}}{1-\omega_{i',j}}}{\frac{\omega_{i,j}}{1-\omega_{i,j}}} = \frac{\exp\left[\alpha_j + (\boldsymbol{x}_i + \mathbf{1})\boldsymbol{\beta}_j\right]}{\exp\left[\alpha_j + \boldsymbol{x}_i\boldsymbol{\beta}_j\right]} = e^{\boldsymbol{\beta}_j}.$$

Thus, $\boldsymbol{\beta}_j$ is the effect of unit increase in $\boldsymbol{x}_i$ on log odds ratio.
$\boldsymbol{\beta}_j$ is the effect for category j. If all $\boldsymbol{\beta}_j = \boldsymbol{\beta}$, then the effect is assumed to be the same for all j.

Such a way of sequential estimation can be useful for survival time analysis when we consider discrete time points as categories.

b)

By sequentially conditioning the joint probability from 1 to c, we obtain

$$P(y_{i,1}, \cdots, y_{i,c}) = P(y_{i,1})P(y_{i,2}|y_{i,1})P(y_{i,3}|y_{i,1}, y_{i,2}) \cdots P(y_{i,c-1}|y_{i,1}, \cdots, y_{i,c-2})P(y_{i,c}|y_{i,1}, \cdots, y_{i,c-1})$$

The log-likelihood for unit i becomes

$$\begin{aligned} L_i = &\log P(y_{i,1}) + \log P(y_{i,2}|y_{i,1}) + \log P(y_{i,3}|y_{i,1}, y_{i,2}) + \\ &\cdots + \log P(y_{i,c-1}|y_{i,1}, \cdots, y_{i,c-2}) + \log P(y_{i,c}|y_{i,1}, \cdots, y_{i,c-1}). \end{aligned}$$

Note that $P(y_{i,2} = 1|y_{i,1} = 1) = 0$ and $P(y_{i,2} = 1|y_{i,1} = 0) = \omega_{i,1}$. So, we can estimate α_2 and $\boldsymbol{\beta}_2$ by fitting a binary GLM to $y_{i,2}$ for all units i where $y_{i,1} = 0$. By using this technique sequentially, we end up fitting c binary GLMs and estimate remaining α_j's and $\boldsymbol{\beta}_j$'s.

Exercise 6.20

Recall the definition of the cumulative logit:

$$\begin{aligned} \text{logit}\left[P\left(y_i \leq j\right)\right] &= \log\left[\frac{P\left(y_i \leq j\right)}{P\left(y_i > j\right)}\right] \\ &= \alpha_j + x_i\beta \end{aligned}$$

The intercepts are increasing since the cumulative probabilities must increase. There are three categories, and since the last one is always 1, there are two intercepts, corresponding to staunch belief in heaven and agnosticism.

The race coefficient tells us that blacks are more likely to believe in heaven than whites, while the gender effect says that women are more likely to believe in heaven than men.

We use the Wald intervals: $CI\left(\beta\right) = \widehat{\beta} \pm z_{1-\frac{\alpha}{2}} n^{-1/2} \widehat{\sigma_\beta}$, where $\widehat{\sigma_\beta}$ is an estimate of the asymptotic standard deviation of $\sqrt{n}\left(\widehat{\beta} - \beta\right)$. These coincide with $n^{1/2}\mathrm{se}_\beta$ from the supplied table.

$$\begin{aligned} CI_{\alpha_0} &= 0.08 \pm 0.09 \\ CI_{\alpha_1} &= 2.32 \pm 0.14 \\ CI_{\text{gender}} &= 0.77 \pm 0.12 \\ CI_{\text{race}} &= 1.02 \pm 0.21 \end{aligned}$$

We use the residual deviance ($= 9.25$) for our goodness of fit test. Since $F_4^{-1}\left(0.95\right) = 9.49$, where F_ν^{-1} is the quantile function of the χ^2-distribution with ν degrees of freedom, we don't reject the model.

Exercise 7.2

The pmf of Poisson distribution is $f(y) = \dfrac{\mu^y}{y!} e^{-\mu}$. The corresponding log-likelihood function is $L(\boldsymbol{y}) = \log \mu \sum_{i=1}^{n} y_i - n\mu - \sum_{i}^{n} \log y_i!$. The MLE of μ is then $\widehat{\mu} = \overline{y}$. Thus, the likelihood-ratio statistic is

$$\begin{aligned} -2\left(L_0 - L_1\right) &= 2\left(\log \widehat{\mu}_1 \sum_{i=1}^{n} y_i - n\widehat{\mu}_1 - \sum_{i}^{n} \log y_i! - \log \mu_0 \sum_{i=1}^{n} y_i + n\mu_0 + \sum_{i}^{n} \log y_i!\right) \\ &= 2\left(\log \frac{\widehat{\mu}_1}{\mu_0} \sum_{i=1}^{n} y_i + n\left(\widehat{\mu}_1 - \mu_0\right)\right) \\ &= 2\left(n\overline{y} \log \frac{\widehat{\mu}_1}{\mu_0} + n\left(\widehat{\mu}_1 - \mu_0\right)\right) \\ &= 2\left(n\overline{y} \log \frac{\overline{y}}{\mu_0} + n\left(\overline{y} - \mu_0\right)\right). \end{aligned}$$

The likelihood-ratio statistics follows chi-squared distribution with degrees of freedom 1. So, a $100(1 - \alpha)\%$ confidence interval is given as the set of μ_0 that satisfy

$$2\left(n\overline{y} \log \frac{\overline{y}}{\mu_0} + n\left(\overline{y} - \mu_0\right)\right) < \chi_1^2(\alpha).$$

Exercise 7.3

The score and information of Poisson distribution are

$$\begin{aligned} S(\mu) &= \frac{\partial L(\boldsymbol{y})}{\partial \mu} = \frac{1}{\mu} \sum_{i=1}^{n} y_i - n = n\left(\frac{\overline{y}}{\mu} - 1\right) \\ H(\mu) &= \frac{\partial^2 L(\boldsymbol{y})}{\partial \mu^2} = -\frac{n\overline{y}}{\mu^2} \\ \mathcal{J}(\mu) &= \mathrm{E}\left[-\frac{\partial^2 L(\boldsymbol{Y})}{\partial \mu^2}\right] = \frac{n}{\mu^2} \mathrm{E}\left[\overline{Y}\right] = \frac{n}{\mu}. \end{aligned}$$

So, the score test statistic is

$$\frac{S(\mu_0)}{\sqrt{\mathcal{J}(\mu_0)}} = \frac{n\left(\frac{\overline{y}}{\mu_0} - 1\right)}{\frac{n}{\mu_0}} = \frac{\sqrt{n}(\overline{y} - \mu_0)}{\sqrt{\mu_0}} =\sim N(0,1).$$

Thus, a $100(1-\alpha)\%$ confidence interval is given as the set of μ_0 that satisfy

$$\left|\frac{\sqrt{n}(\overline{y} - \mu_0)}{\sqrt{\mu_0}}\right| < z_{1-\frac{\alpha}{2}}.$$

To find the confidence interval we solve a quadratic inequality:

$$n\left(\overline{y} - \mu\right)^2 \leq z^2_{\alpha/2}\mu$$

$$\mu^2 - \mu\left(2\overline{y} + n^{-1}z^2_{\alpha/2}\right) + \overline{y}^2 \leq 0$$

$$\frac{\left(2\overline{y} + n^{-1}z^2_{\alpha/2}\right) \pm \sqrt{\left(2\overline{y} + n^{-1}z^2_{\alpha/2}\right)^2 - 4\overline{y}^2}}{2}$$

$$\frac{\left(2\overline{y} + n^{-1}z^2_{\alpha/2}\right) \pm \sqrt{4\overline{y}^2 + 4n^{-1}z^2_{\alpha/2}\overline{y} + n^{-2}z^4_{\alpha/2} - 4\overline{y}^2}}{2}$$

This can be simplified to:

$$\left(\overline{y} + \frac{z^2_{\alpha/2}}{2n}\right) \pm \alpha n^{-1/2}\sqrt{\overline{y} + \frac{z^2_{\alpha/2}}{4n}}$$

Compare this to the Wald CI, which is:

$$\overline{y} \pm z_{\alpha/2}n^{-1/2}\sqrt{\overline{y}}$$

Additional Exercise 22

```
> # Enter the data
> lung.cancer.data = data.frame(
+   city = rep(1:4, each = 5),
+   age = rep(1:5, times = 4),
+   cases = c(11,11,11,10,11,13,6,15,10,12,4,8,7,11,9,5,7,10,14,8),
+   number = c(3059,800,710,581,509,2879,1083,923,834,634,3142,
+          1050,895,702,535,2520,878,839,631,539)
+   )
> lung.cancer.data[,"age"] = as.factor(lung.cancer.data[,"age"])
> lung.cancer.data[,"city"] = as.factor(lung.cancer.data[,"city"])
> head(lung.cancer.data)
city age cases number
1    1   1    11   3059
2    1   2    11    800
3    1   3    11    710
4    1   4    10    581
5    1   5    11    509
6    2   1    13   2879
```

a)

The expected value of the response variable (`cases`) is proportional to the total number of male inhabitants (`number`). We here model the rate, i.e. $\frac{y_{i,j}}{n_{i,j}}$, where $y_{i,j}$ and $n_{i,j}$ are the number of cases and the number of inhabitants in city i and age group j. The model $\mathrm{E}\left[\frac{Y_{i,j}}{n_{i,j}}\right] = e^{\eta_{i,j}}$ for a linear predictor $\eta_{i,j}$ hence corresponds to $\mathrm{E}[Y_{i,j}] = e^{\log n_{i,j} + \eta_{i,j}}$. So, we have to add $\log n_{i,j}$ as an offset to the linear predictor.

b)

```
> Poisson.model = glm(cases ~ offset(log(number)) + age + city, family = poisson, data
    = lung.cancer.data)
> summary(Poisson.model)

Call:
glm(formula = cases ~ offset(log(number)) + age + city, family = poisson,
data = lung.cancer.data)

Deviance Residuals:
Min        1Q    Median        3Q       Max
-1.20728  -0.59302  -0.09784   0.58493   1.46574

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.6455     0.2049 -27.555  < 2e-16 ***
age2          1.0961     0.2483   4.414 1.02e-05 ***
age3          1.5138     0.2317   6.534 6.39e-11 ***
age4          1.7584     0.2295   7.662 1.83e-14 ***
age5          1.8486     0.2354   7.855 4.01e-15 ***
city2        -0.1907     0.1910  -0.999   0.3180
city3        -0.4791     0.2103  -2.279   0.0227 *
city4        -0.2534     0.2033  -1.247   0.2125
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 115.434  on 19  degrees of freedom
Residual deviance:  11.598  on 12  degrees of freedom
AIC: 109.07

Number of Fisher Scoring iterations: 4
```

The interpretation of $\widehat{\beta}$ can be done in the same way as in exercise 7.31 a) from the book.

c)

```
> lung.cancer.data[,"Fredericia"] = as.factor(as.numeric(lung.cancer.data[,"city"] ==
    1))
> head(lung.cancer.data)
city age cases number Fredericia
1    1   1    11   3059          1
2    1   2    11    800          1
```

```
3    1   3    11    710          1
4    1   4    10    581          1
5    1   5    11    509          1
6    2   1    13   2879          0
> # Fit Poisson GLM
> Poisson.model.2 = glm(cases ~ offset(log(number)) + age + Fredericia, family =
    poisson, data = lung.cancer.data)
> summary(Poisson.model.2)

Call:
glm(formula = cases ~ offset(log(number)) + age + Fredericia,
family = poisson, data = lung.cancer.data)

Deviance Residuals:
Min        1Q    Median        3Q       Max
-1.62564  -0.59506  -0.03471   0.17297   1.81669

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9502     0.1818 -32.736  < 2e-16 ***
age2          1.0997     0.2483   4.429 9.47e-06 ***
age3          1.5187     0.2316   6.556 5.51e-11 ***
age4          1.7671     0.2294   7.704 1.32e-14 ***
age5          1.8582     0.2352   7.899 2.82e-15 ***
Fredericia1   0.2991     0.1606   1.863   0.0624 .
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 115.434  on 19  degrees of freedom
Residual deviance:  13.663  on 14  degrees of freedom
AIC: 107.14

Number of Fisher Scoring iterations: 4

>
> # Likelihood ratio test.
> anova(Poisson.model.2, Poisson.model.1)
Analysis of Deviance Table

Model 1: cases ~ offset(log(number)) + age + Fredericia
Model 2: cases ~ offset(log(number)) + age + city
Resid. Df Resid. Dev Df Deviance
1        14     13.663
2        12     11.598  2   2.0658
> 1 - pchisq(anova(Poisson.model.2, Poisson.model.1)$Deviance[2], df = 1)
[1] 0.1506362
```

The two models are nested. So, we use the likelihood ratio test.
$p = 0.1506 > 0.05$. So, we keep the null hypothesis and conclude that the model with simplified city effect is a better model.

d)

Rate ratio of new lung cancer cases in Fredericia compared to the three other cities: $\frac{\mathrm{E}[Y|x_{\text{Fredericia}}=1]}{\mathrm{E}[Y|x_{\text{Fredericia}}=0]} = \exp[\beta_{\text{Fredericia}}]$.

```
> # Estimated rate ratio
> exp(summary(Poisson.model.2)$coefficients["Fredericia1","Estimate"])
[1] 1.348685
> # 95% confidence interval of rate ratio
> exp(confint.default(Poisson.model.2))["Fredericia1",]
2.5 %    97.5 %
0.9845689 1.8474606
```

e)

```
> # Numeric version of variable age
> lung.cancer.data[,"age.numeric"] = rep(
+   c(mean(40,55), mean(55,60), mean(60,65), mean(65,70), mean(70,75)),
+   times = 4
+   )
> # Fit Poisson GLM
> Poisson.model.3 = glm(cases ~ offset(log(number)) + I(age.numeric-40) + Fredericia,
    family = poisson, data = lung.cancer.data)
> summary(Poisson.model.3)

Call:
glm(formula = cases ~ offset(log(number)) + I(age.numeric - 40) +
Fredericia, family = poisson, data = lung.cancer.data)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-1.8683  -0.6997  -0.1695   0.3869   1.6274

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)         -5.857611   0.158061 -37.059   <2e-16 ***
I(age.numeric - 40)  0.064311   0.006945   9.261   <2e-16 ***
Fredericia1          0.290371   0.160441   1.810   0.0703 .
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 115.434  on 19  degrees of freedom
Residual deviance:  16.141  on 17  degrees of freedom
AIC: 103.61

Number of Fisher Scoring iterations: 4

>
> # Likelihood ratio test.
> anova(Poisson.model.3, Poisson.model.2)
Analysis of Deviance Table

Model 1: cases ~ offset(log(number)) + I(age.numeric - 40) + Fredericia
```

```
Model 2: cases ~ offset(log(number)) + age + Fredericia
Resid. Df Resid. Dev Df Deviance
1        17     16.141
2        14     13.663  3   2.4776
> 1 - pchisq(anova(Poisson.model.3, Poisson.model.2)$Deviance[2], df = 1)
[1] 0.1154789
```

The two models are nested. So, we use the likelihood ratio test.
$p = 0.1155 > 0.05$. So, we keep the null hypothesis and conclude that the model with numerified age effect is a better model.

The interpretation of $\widehat{\beta}_{\text{age numeric}}$: log rate ratio of one unit increase in `age numeric`.

Exam 2016: Exercise 2

a)

The Poisson model is reasonable by the law of small numbers. Each accident has an extremely small chance of happening, but many occasions in which in can happen.

As for the model fit, the residual deviance looks pretty good, and the model can't be rejected with a χ^2-test. Still, one should not judge the suitability of a model from such a summary statistic. Try to look at residuals to check for patterns, and try another model as well, such as the negative binomial. These can easily be compared with the AIC.

b)

The offset is used because we are talking about relative instead of absolute magnitutes. Each combination of age group and gender have their own population, and the number of deaths only make sense with this number as a reference. See the textbook, p. 233.

c)

The exponentiated intercept tells us the expected rate of death for boys in the age group $0 - 17$. The rate is in number of deaths per 100000, as can be seen from the offset command.

d)

According to the table at the end, age group 5 contains people of age $45 - 54$. The fitted value for this one is

$$\begin{aligned} \exp\left[\beta_0 + \beta + \beta_5\right] &= \exp\left[0.15 - 1.02 + 1.54 + \log\left(338505/100000\right)\right] \\ &= 6.6. \end{aligned}$$

The table says that the observed number of deaths was 2, and the residual is -4.6.

e)

Recall the likelihood equation:

$$\sum_i \left(y_i - \mu_i\right) x_{ij} = 0, \, \forall j$$

When we look at the j corresponding to the gender, $j = 1$ for women, and the likelihood equation becomes

$$\sum_i y_i = \sum_i \mu_i.$$