STK3100 Exercises, Week 11

Vinnie Ko, Jonas Moss

November 8, 2018

Exercise 7.30

```
> shark = c(33, 29, 29, 12, 17, 21, 31, 28, 19, 14, 11, 26, 23)
> summary(glm(shark ~ 1, family = poisson(link = "log")))
    .
    .
    .
    Null deviance: 31.392  on 12  degrees of freedom
    Residual deviance: 31.392  on 12  degrees of freedom
    AIC: 97.129
> summary(MASS::glm.nb(shark ~ 1))
    .
    .
    .
    Null deviance: 13.363  on 12  degrees of freedom
    Residual deviance: 13.363  on 12  degrees of freedom
    AIC: 92.608
```

The χ^2-test works out way better for the negative binomial. The AIC is also unusually much better. We conclude that the Poisson distribution can't account well for the data.

Exercise 7.31

a)

```
> # Read data
> homicide.data = read.table("http://www.stat.ufl.edu/~aa/glm/data/Homicides.dat",
    header = T)
> homicide.data[,"race"] = as.factor(homicide.data[,"race"])
> head(homicide.data)
Obs race count
1   1    0     0
2   2    0     0
3   3    0     0
4   4    0     0
5   5    0     0
6   6    0     0
> table(homicide.data[,"count"], homicide.data[,"race"])

0    1
```

```
0 1070  119
1   60   16
2   14   12
3    4    7
4    0    3
5    0    2
6    1    0
>
> # a)
>
> # Fit Poisson model
> Poisson.model = glm(count ~ race, family = poisson, data = homicide.data)
> summary(Poisson.model)

Call:
glm(formula = count ~ race, family = poisson, data = homicide.data)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-1.0218  -0.4295  -0.4295  -0.4295   6.1874

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.38321    0.09713  -24.54   <2e-16 ***
race1        1.73314    0.14657   11.82   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 962.80  on 1307  degrees of freedom
Residual deviance: 844.71  on 1306  degrees of freedom
AIC: 1122

Number of Fisher Scoring iterations: 6
```

$\widehat{\beta}_0$ can be interpreted as the log of the average number of known homicide victims for the reference group (white): $\mathrm{E}\left[Y|x_i = 0\right] = e^{\widehat{\beta}_0} = e^{-2.3832} = 0.0923$.

$\widehat{\beta}_1$ can be interpreted as the log rate ratio between the average number of known homicide victims for white and black: $\dfrac{\mathrm{E}\left[Y|x_i = 1\right]}{\mathrm{E}\left[Y|x_i = 0\right]} = e^{\widehat{\beta}_1} = e^{1.7331} = 5.6584$.

b)

Possible factors of heterogeneity might be socio-economic variables.

Fit negative binomial GLM

```
> # Fit negative binomial model
> negbin.model = glm.nb(count ~ race, data = homicide.data)
> summary(negbin.model)

Call:
```

```
glm.nb(formula = count ~ race, data = homicide.data, init.theta = 0.2023119205,
link = log)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-0.7184  -0.3899  -0.3899  -0.3899   3.5072

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3832     0.1172 -20.335  < 2e-16 ***
race1         1.7331     0.2385   7.268 3.66e-13 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for Negative Binomial(0.2023) family taken to be 1)

Null deviance: 471.57  on 1307  degrees of freedom
Residual deviance: 412.60  on 1306  degrees of freedom
AIC: 1001.8

Number of Fisher Scoring iterations: 1


Theta:  0.2023
Std. Err.:  0.0409

2 x log-likelihood:  -995.7980
>
> # Test overdispersion
> overdisp.test.statistic = -2*(logLik(Poisson.model) - logLik(negbin.model))
> 1 - pchisq(as.numeric(overdisp.test.statistic), df = 1)
[1] 0
> # We reject the null hypothesis with alpha = 0.05.
> # So, we conclude that there is overdispersion and choose for the negative binomial
   model.
```

The coefficient estimates are virtually the same as in the Poisson GLM, but the estimated dispersion parameter is $\widehat{\gamma} = \dfrac{1}{\theta} = \dfrac{1}{0.2023} = 4.94$. This suggests that there is overdispersion and that the Poisson GLM is inadequate.

c)

```
> # Wald 95% confidence interval
> exp(confint.default(Poisson.model))
2.5 %    97.5 %
(Intercept) 0.0762623 0.1115994
race1       4.2455738 7.5414329
> exp(confint.default(negbin.model))
2.5 %    97.5 %
(Intercept) 0.07332043 0.1160771
race1       3.54571025 9.0299848
```

As we saw in b), there is an evidence of overdispersion. Therefore, the confidence interval from negative binomial GLM is more reliable.

Extra: reconstruct table 7.5

```
> n.white = nrow(homicide.data[(homicide.data[,"race"] == 0),])
> n.black = nrow(homicide.data[(homicide.data[,"race"] == 1),])
> response.range = 0:6
>
> # Estimated number of reponses from Poisson model
> Pois.mu.hat.white = predict(Poisson.model, newdata = data.frame(race = as.factor(0))
    , type = "response")
> Pois.mu.hat.black = predict(Poisson.model, newdata = data.frame(race = as.factor(1))
    , type = "response")
> Poisson.estimation = data.frame(
+   black = n.black*dpois(x = response.range, lambda = Pois.mu.hat.black),
+   white = n.white*dpois(x = response.range, lambda = Pois.mu.hat.white)
+   )
> Poisson.estimation = round(Poisson.estimation, 1)
> Poisson.estimation
black  white
1  94.3 1047.7
2  49.2   96.7
3  12.9    4.5
4   2.2    0.1
5   0.3    0.0
6   0.0    0.0
7   0.0    0.0
>
> # Estimated number of reponses from negative binomial model
> negbin.mu.hat.white = predict(negbin.model, newdata = data.frame(race = as.factor(0)
    ), type = "response")
> negbin.mu.hat.black = predict(negbin.model, newdata = data.frame(race = as.factor(1)
    ), type = "response")
> negbin.estimation = data.frame(
+   black = n.black*dnbinom(x = response.range, size = negbin.model$theta, mu = negbin
    .mu.hat.black),
+   white = n.white*dnbinom(x = response.range, size = negbin.model$theta, mu = negbin
    .mu.hat.white)
+ )
> negbin.estimation = round(negbin.estimation, 1)
> negbin.estimation
black  white
1 122.8 1064.9
2  17.9   67.5
3   7.8   12.7
4   4.1    2.9
5   2.4    0.7
6   1.4    0.2
7   0.9    0.1
```

Exercise 8.6

In weighted least squares, β_j's are estimated by minimizing $Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{v_i}$. So,

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{v_i} = -2 \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_j} \frac{y_i - \mu_i}{v_i} \propto \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_j} \frac{y_i - \mu_i}{v_i} = 0.$$

Thus, when variance is known, equation (8.2) is equal to the weighted least square equation.

Exercise 8.8

Assume the null model $\mu_i = \beta$ and $v(\mu_i) = \sigma^2$, then

$$u(\beta) = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta} \frac{y_i - \mu_i}{v(\mu_i)} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i).$$

Thus, $u(\boldsymbol{\beta}) = 0$ gives $\widehat{\beta} = \overline{y}$ and by using formula (8.3) we obtain $V = \left[\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)^{\mathrm{T}} (v(\mu_i))^{-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right) \right]^{-1} = \left[\sum_{i=1}^{n} 1 \cdot \frac{1}{\sigma^2} \cdot 1 \right]^{-1} = \frac{\sigma^2}{n}$. A sensible model based estimate of V is $\widehat{V} = \frac{1}{n^2} \sum_{i=1}^{n} (y_i - \overline{y})^2$. The actual asymptotic variance of $\widehat{\beta}$ is by formula (8.4)

$$V \left(\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta} \frac{\mathrm{Var}(y_i)}{v(\mu_i)^2} \frac{\partial \mu_i}{\partial \beta} \right) V = \frac{\sigma^2}{n} \left(\sum_{i=1}^{n} \frac{\beta}{\sigma^4} \right) \frac{\sigma^2}{n} = \frac{\beta}{n}.$$

To find the robust estimate of the variance that adjusts for model misspecification, we replace $\mathrm{Var}(y_i)$ in the expression above with $(y_i - \overline{y})^2$. We find the robust estimate to be $\frac{1}{n^2} \sum_{i=1}^{n} (y_i - \overline{y})^2$.

Exercise 8.9

Assume the null model $\mu_i = \beta$ and $v(\mu_i) = \mu_i$, then

$$u(\beta) = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta} \frac{y_i - \mu_i}{v(\mu_i)} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mu_i} = \sum_{i=1}^{n} \frac{y_i - \beta}{\beta}.$$

Thus, $u(\boldsymbol{\beta}) = 0$ gives $\widehat{\beta} = \overline{y}$ and by using formula (8.3) we obtain $V = \left[\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)^{\mathrm{T}} (v(\mu_i))^{-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right) \right]^{-1} = \left[\sum_{i=1}^{n} 1 \cdot \frac{1}{\beta} \cdot 1 \right]^{-1} = \frac{\beta}{n}$. A sensible model based estimate of V is $\widehat{V} = \frac{\overline{y}}{n}$. The actual asymptotic variance of $\widehat{\beta}$ is by formula (8.4)

$$V \left(\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta} \frac{\mathrm{Var}(y_i)}{v(\mu_i)^2} \frac{\partial \mu_i}{\partial \beta} \right) V = \frac{\beta}{n} \left(\sum_{i=1}^{n} \frac{\sigma^2}{\beta^2} \right) \frac{\beta}{n} = \frac{\sigma^2}{n}.$$

To find the robust estimate of the variance that adjusts for model misspecification, we replace $\mathrm{Var}(y_i)$ in the expression above with $(y_i - \overline{y})^2$. We find the robust estimate to be $\frac{1}{n^2} \sum_{i=1}^{n} (y_i - \overline{y})^2$.

Exercise 8.14

```
> # Load packages
> library(MASS)
>
```

```
> # Read data
> homicide.data = read.table("http://www.stat.ufl.edu/~aa/glm/data/Homicides.dat",
    header = T)
> homicide.data[,"race"] = as.factor(homicide.data[,"race"])
> head(homicide.data)
Obs race count
1   1    0     0
2   2    0     0
3   3    0     0
4   4    0     0
5   5    0     0
6   6    0     0
> table(homicide.data[,"count"], homicide.data[,"race"])

0    1
0 1070  119
1   60   16
2   14   12
3    4    7
4    0    3
5    0    2
6    1    0
>
> # Fit Poisson model
> Poisson.model = glm(count ~ race, family = poisson, data = homicide.data)
> summary(Poisson.model)

Call:
glm(formula = count ~ race, family = poisson, data = homicide.data)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-1.0218  -0.4295  -0.4295  -0.4295   6.1874

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.38321    0.09713  -24.54   <2e-16 ***
race1        1.73314    0.14657   11.82   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 962.80  on 1307  degrees of freedom
Residual deviance: 844.71  on 1306  degrees of freedom
AIC: 1122

Number of Fisher Scoring iterations: 6

>
> # Fit negative binomial model
> negbin.model = glm.nb(count ~ race, data = homicide.data)
> summary(negbin.model)

Call:
glm.nb(formula = count ~ race, data = homicide.data, init.theta = 0.2023119205,
```

```
link = log)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-0.7184  -0.3899  -0.3899  -0.3899   3.5072

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3832     0.1172 -20.335  < 2e-16 ***
race1         1.7331     0.2385   7.268 3.66e-13 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for Negative Binomial(0.2023) family taken to be 1)

Null deviance: 471.57  on 1307  degrees of freedom
Residual deviance: 412.60  on 1306  degrees of freedom
AIC: 1001.8

Number of Fisher Scoring iterations: 1


Theta:  0.2023
Std. Err.:  0.0409

2 x log-likelihood:  -995.7980
>
> # Quasi likelihood approach.
> QL.model = glm(count ~ race, family = quasi(link = "log", variance = "mu"), data =
    homicide.data)
> summary(QL.model)

Call:
glm(formula = count ~ race, family = quasi(link = "log", variance = "mu"),
data = homicide.data)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-1.0218  -0.4295  -0.4295  -0.4295   6.1874

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.3832     0.1283  -18.57   <2e-16 ***
race1         1.7331     0.1937    8.95   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for quasi family taken to be 1.745693)

Null deviance: 962.80  on 1307  degrees of freedom
Residual deviance: 844.71  on 1306  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

For the QL-method, we obtain the same estimate as for Poisson model, but the standard errors are multiplied by $\sqrt{1.7457} = 1.3212$. Also, the negative binomial does here give the same estimates as Poisson model (which is not the case in general) and the standard errors are about the same as for the QL-method. (Check p.248 to see how variance is defined for negative binomial model.)

Exercise 8.17

Here's some R-code you can use to load the data into R.

```
games = c(1, 0, 4,
2, 7, 9,
3, 4, 11,
4, 3, 6,
5, 5, 6,
6, 2, 7,
7, 3, 7,
8 , 0, 1,
9, 1, 8,
10, 6, 9,
11, 0, 5,
12, 2, 5,
13, 0, 5,
14, 2, 4,
15, 5, 7,
16, 1, 3,
17, 3, 7,
18, 0, 2,
19, 8, 11,
20, 0, 8,
21, 0, 4,
22, 0, 4,
23, 2, 5,
24, 2, 7)

dim(games) = c(3, length(games)/3)
games = t(games)
colnames(games) = c("game", "yi", "ni")
games = as.data.frame(games)
```

The naive estimates are

```
> p_hat = sum(games$yi)/sum(games$ni)
> p_hat
[1] 0.3862069
> se_hat = 1/sqrt(sum(games$ni))*sqrt(p_hat*(1 - p_hat))
> se_hat
[1] 0.0404331
```

Overdispersion could be caused by different forms in different games.

The formula for robust variances is:

$$V \left[\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right) \frac{(y_i - \overline{y})^2}{[v(\mu_i)]} \left(\frac{\partial \mu_i}{\partial \beta} \right) \right] V$$

In our case $\mu_i = \beta$ for all i and the variance function is $v(\beta) = \beta(1 - \beta)$, while V equals $\beta(1 - \beta)/n$ as before. Thus

$$\begin{aligned}
\frac{\beta(1-\beta)}{n}\sum_{i=1}^{n}\frac{(y_i-\overline{y})^2}{\beta^2(1-\beta)^2}\frac{\beta(1-\beta)}{n} &= \frac{1}{n^2}\sum_{i=1}^{n}(y_i-\overline{y})^2 \\
&= \frac{1}{n}\widehat{\sigma^2}
\end{aligned}$$

And the confidence intervals are:

```
> lim = qnorm(.975)
> p_hat + c(-se_hat, se_hat)*lim
[1] 0.3069595 0.4654543
> p_hat + c(-1/sqrt(n)*sd(games$yi), 1/sqrt(n)*sd(games$yi))*lim
[1] -0.005687939  0.778101732
```

Additional Exercise 23

a) Law of Total Expectation

By definition $E(Y \mid X) = \int yp(y \mid X)\,dy$. This is a random variable since X is random. Moreover, the source of its randomness is captured entirely by X. Recall that the expecation $Ef(X) = \int f(x)\,p(x)\,dx$ or any suitable function f. Thus

$$E[E(Y \mid X)] = \int\left[\int yp(y \mid x)\,dy\right]p(x)\,dx$$

Now we assume the conditions for Fubini's theorem holds, so that we can change the order of integration. This is true provided expectation $E[Y]$ exist. In this case,

$$\begin{aligned}
E[E(Y \mid X)] &= \int\left[\int yp(y \mid x)\,dy\right]p(x)\,dx \\
&= \int\int yp(y \mid x)\,p(x)\,dxdy \\
&= \int\int yp(y,x)\,dxdy \\
&= \int yp(y)\,dy
\end{aligned}$$

b) Law of Total Variance

The definition of the conditional variance is

$$\operatorname{Var}(Y \mid X) = E\left(Y^2 \mid X\right) - [E(Y \mid X)]^2$$

Which implies

$$E[\operatorname{Var}(Y \mid X)] = E\left(Y^2\right) - E\left\{[E(Y \mid X)]^2\right\}$$

The variance of the conditional expectation is

$$\begin{aligned}\operatorname{Var}[E(Y \mid X)] &= E\left\{[E(Y \mid X)]^2\right\} - \{E[E(Y \mid X)]\}^2 \\ &= E\left\{[E(Y \mid X)]^2\right\} - [E(Y)]^2\end{aligned}$$

Combine the two identities to get

$$\begin{aligned}E[\operatorname{Var}(Y \mid X)] + \operatorname{Var}[E(Y \mid X)] &= E\left(Y^2\right) - [E(Y)]^2 \\ &= \operatorname{Var}(Y)\end{aligned}$$

Note: A similar result holds for covariances, see the law of total covariance.

Additional Exercise 24

a)

We already did this many times in earlies exercises.

b)

i)

$\exp[\beta_{\texttt{badh}}] = \dfrac{\mathrm{E}[Y|x_{\texttt{badh}} = 1]}{\mathrm{E}[Y|x_{\texttt{badh}} = 0]}$.

So, the estimate of rate ratio is

$$\exp\left[\widehat{\beta}_{\texttt{badh}}\right] = \exp[1.1409] = 3.1296.$$

95% confidence interval for this rate ratio is

$$\begin{aligned}&\left[\exp\left[\widehat{\beta}_{\texttt{badh}} - z_{0.975}\cdot \mathrm{SE}(\widehat{\beta}_{\texttt{badh}})\right],\ \exp\left[\widehat{\beta}_{\texttt{badh}} + z_{0.975}\cdot \mathrm{SE}(\widehat{\beta}_{\texttt{badh}})\right]\right] \\ &= [\exp[1.1409 - 1.96\cdot 0.0399], \exp[1.1409 + 1.96\cdot 0.0399]] \\ &= [\exp[1.0628], \exp[1.2190]] \\ &= [2.8944, 3.3839]\end{aligned}$$

ii)

$\exp\left[10\beta_{\texttt{age}}\right] = \dfrac{\mathrm{E}\left[Y|x_{\texttt{age}} = 50\right]}{\mathrm{E}\left[Y|x_{\texttt{age}} = 40\right]}$.

So, the estimate of rate ratio is

$$\exp\left[10\cdot\widehat{\beta}_{\texttt{age}}\right] = \exp[0.0556] = 1.0571.$$

95% confidence interval for this rate ratio is

$$\begin{aligned}&\left[\exp\left[10\cdot\widehat{\beta}_{\texttt{age}} - 10\cdot z_{0.975}\cdot \mathrm{SE}(\widehat{\beta}_{\texttt{age}})\right],\ \exp\left[10\cdot\widehat{\beta}_{\texttt{age}} + 10\cdot z_{0.975}\cdot \mathrm{SE}(\widehat{\beta}_{\texttt{age}})\right]\right] \\ &= [\exp[0.0556 - 1.96\cdot 0.0168], \exp[0.0556 + 1.96\cdot 0.0168]] \\ &= [\exp[0.0227], \exp[0.0884]] \\ &= [1.0230, 1.0924]\end{aligned}$$

iii)
$\{\widehat{\mu}|\texttt{age} = 40, \texttt{badh} = 0\} = \exp\left[\widehat{\beta}_0 + \widehat{\beta}_{\texttt{age}} \cdot 40 + \widehat{\beta}_{\texttt{badh}} \cdot 0\right] = \exp\left[0.5888 + 0.0056 * 40\right] = \exp\left[0.810971\right] =$ 2.2501

The confidence interval of this rate ratio is equal to the confidence interval of e^{η} where $\eta = \beta_0 + \beta_{\texttt{age}} \cdot 40$. We would then first find a confidence interval of η. This takes the form $\widehat{\eta} \pm z_{1-\frac{\alpha}{2}} \cdot \mathrm{SE}(\widehat{\eta})$, where $\mathrm{SE}(\widehat{\eta}) = \sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_0) + 40^2\widehat{\mathrm{Var}}(\widehat{\beta}_{\texttt{age}}) + 2 \cdot 40 \cdot \widehat{\mathrm{Cov}}\left(\widehat{\beta}_0, \widehat{\beta}_{\texttt{age}}\right)}$. So, we would need $\mathrm{Cov}\left(\widehat{\beta}_0, \widehat{\beta}_{\texttt{age}}\right)$.

c)

i)
We estimate parameters by solving quasi-likelihood equations instead of the likelihood equations. However, since ϕ will cancel, the estimates are the same as for the Poisson model.

ii)
We compute the Pearson statistic $X^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}$ for the Poisson model, and estimate ϕ by $\widehat{\phi} = \frac{X^2}{n-p}$ where $p = 3$.