

Exercise 7.31

Per August Jarval Moen

26/10/2023

Exercise 7.31

We begin by loading the data:

```
url = "http://users.stat.ufl.edu/~aa/glm/data/Homicides.dat"
data = read.table(url, header=TRUE)
data = data[,c("race", "count")] #removing observation number
data$race=as.factor(data$race)
head(data)
```

```
##   race count
## 1    0     0
## 2    0     0
## 3    0     0
## 4    0     0
## 5    0     0
## 6    0     0
```

Note that the data is un-grouped, while table 7.5 in the book shows the grouped representation. You can check that the dataset is the same as the table. For instance we have that:

```
sum( data[data$race==1,2]==0) #119, agrees with table
```

```
## [1] 119
```

which agrees with table 7.5.

a)

We fit the Poisson GLM as specified in the exercise text:

```
poisson.model = glm(count~race, family=poisson(link="log"),data=data)
summary(poisson.model)
```

```
##
## Call:
## glm(formula = count ~ race, family = poisson(link = "log"), data = 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
```

Interpretation: The rate of blacks knowing someone who has been murdered is estimated to be $\exp(\hat{\beta}) = 5.66$ times larger than the same rate for whites.

b)

There are arguably other characteristics that determine the rate at which one knows someone who is murdered, for instance socio-economic background, postal code, etc. Since these are not included in the model, the response variable Y_i will show a larger variation/variance than the Poisson model suggests (remember that for the Poisson, the variance is equal to the mean/rate), making the Poisson model inadequate.

To fit a Negative Binomial model, we use the MASS package:

```
library(MASS)
negbin.model = glm.nb(count~race,link="log",data=data)
summary(negbin.model)

##
## Call:
## glm.nb(formula = count ~ race, data = data, link = "log", init.theta = 0.2023119205)
##
## 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
```

Notice that the estimated parameters are almost the same as in the Poisson GLM. Also, note that the interpretation of $\hat{\beta}$ is the same as in the Poisson model, only that we swap the term “rate” with “mean”.

The estimated dispersion parameter is $\hat{\gamma} = \frac{1}{0.2023} \approx 4.94$, which is quite large. Note that R parametrizes the overdispersion parameter as $\theta = \frac{1}{\gamma}$.

To formally compare the model fits, we use a likelihood-ratio test just as in exercise 5.30:

```
library(pscl)
odTest(negbin.model)

## Likelihood ratio test of H0: Poisson, as restricted NB model:
## n.b., the distribution of the test-statistic under H0 is non-standard
## e.g., see help(odTest) for details/references
##
## Critical value of test statistic at the alpha= 0.05 level: 2.7055
## Chi-Square Test Statistic = 122.1919 p-value = < 2.2e-16
```

The p-value is practically zero, so we conclude there is strong evidence that the Negative Binomial model is a better fit (so the Poisson is overdispersed).

c)

The Wald 95% confidence intervals for β are the usual confidence intervals based on asymptotic normality of the MLE: In both models,

$$\hat{\beta} \stackrel{d}{\approx} N(\beta, \text{se}(\hat{\beta})),$$

implying that a 95% CI for β using the Poisson GLM is

$$1.73314 \pm 1.96 \times 0.14657 = (1.44, 2.02),$$

and a 95% CI for β using the Negative Binomial model is

$$\hat{\beta} \pm 1.96 \times 0.14657 = (1.27, 2.2).$$

Exponentiating the above confidence intervals, we get that a 95% confidence interval for the ratio of means using the Poisson GLM is

$$(4.25, 7.55),$$

and a 95% confidence interval for the ratio of means using the Negative Binomial model is

$$(3.54, 9.03).$$

Which of these do we trust more? We have already seen strong evidence that the model fit of the Negative Binomial is better than that of the Poisson GLM, as the Poisson GLM is overdispersed. When GLMs are overdispersed, the model does not capture all of the variability of the data. The presence of overdispersion causes standard errors to be underestimated, and thus that confidence intervals are too narrow. A negative binomial model cannot be overdispersed (although it could of course have a poor model fit), so it is more robust.