Poisson and Negative Binomial Regressions

Manbir Singh Panesar

10/26/2021

2. For this question you need to use the warpbreaks dataset from the datasets package. That is you need to run the following code,

Run this code to get the veteran dataset library(datasets) data(warpbreaks)

You can find the details about the dataset by using `?warpbreaks' code. We are interested in the count of warp breaks per loom (i.e., variable =`breaks') by wool and tension level.

(a) Execute a Poisson regression to estimate the mean number of breaks by wool type and tension level.

(b) Execute a negative binomial regression to estimate the mean number of breaks by wool type and tension level.

(c) Compare the models using the AIC values. Interpret the dispersion parameter of the negative binomial regression. Which model performed better?

```
rm(list = ls())
set.seed(1000660251)

### Question 2 ###
library(datasets)
data(warpbreaks)
head(warpbreaks)
```

```
##   breaks wool tension
## 1     26    A       L
## 2     30    A       L
## 3     54    A       L
## 4     25    A       L
## 5     70    A       L
## 6     52    A       L
```

```
## 2a) Poisson Regression ##
summary(glm(breaks ~ wool + tension, family = poisson, data = warpbreaks))
```

```
##
## Call:
## glm(formula = breaks ~ wool + tension, family = poisson, data = warpbreaks)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6871  -1.6503  -0.4269   1.1902   4.2616
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.69196    0.04541  81.302  < 2e-16 ***
## woolB       -0.20599    0.05157  -3.994 6.49e-05 ***
```

```
## tensionM    -0.32132    0.06027  -5.332 9.73e-08 ***
## tensionH    -0.51849    0.06396  -8.107 5.21e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 297.37  on 53  degrees of freedom
## Residual deviance: 210.39  on 50  degrees of freedom
## AIC: 493.06
##
## Number of Fisher Scoring iterations: 4
```

Our mean value would be exp(intercept) = exp(3.69196)

For woolB this is the estimate that for one increase point while other variables remain constant. For an increase in woolB by one point, the difference in the logs of expected counts would be expected to decrease by 0.20599 units, while holding the other variables in the model constant. woolB's p value is much smaller than an alpha level of 0.05, therefore we reject the null hypothesis and woolB is statistically significant.

For tensionM its estimate for one point increase would be the difference in the logs of expected counts and it would decrease by 0.32132, while holding other variables constant. its p value < 0.05 therefore we reject the null hypothesis that tensionM has no effect on wool breaks and tensionM is statistically significant.

For tensionH its estimate for one point increase would be the difference in the logs of expected counts and it would decrease by 0.051849, when other variables are held constant. tensionH's pvalue < 0.05, therefore we reject the null hypothesis and tensionH is statistically significant.

```r
## 2b) Negative Binomial Regression ##
library(MASS)
summary(glm.nb(breaks ~ wool + tension, data = warpbreaks))
```

```
##
## Call:
## glm.nb(formula = breaks ~ wool + tension, data = warpbreaks,
##     init.theta = 9.944385436, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0144  -0.9319  -0.2240   0.5828   1.8220
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.6734     0.0979  37.520  < 2e-16 ***
## woolB        -0.1862     0.1010  -1.844   0.0651 .
## tensionM     -0.2992     0.1217  -2.458   0.0140 *
## tensionH     -0.5114     0.1237  -4.133 3.58e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(9.9444) family taken to be 1)
##
##     Null deviance: 75.464  on 53  degrees of freedom
## Residual deviance: 53.723  on 50  degrees of freedom
## AIC: 408.76
##
## Number of Fisher Scoring iterations: 1
```

```
##
##
##               Theta:  9.94
##           Std. Err.:  2.56
##
##  2 x log-likelihood:  -398.764
```

The mean for this model is the dispersion parameter of 9.94 for one increase in woolB there is a decrease in the difference in the logs of the expected counts by 0.1862. woolB's pvalue > alpha = 0.05 therefore we fail to reject the null hypothesis

For one increase in tensionM there is a decrease in the diff in the logs of the expected counts by 0.2992. tensionM's pvalue < alpha = 0.05 therefore we reject the null hypothesis and tensionM is statistically significant.

For one increase in tensionH there is a decrease in the diff in the logs of the expected counts by 0.5114. tensionH's pvalue «« alpha=0.05 therefore we reject the null hypothesis and tensionH is statistically significant.

2c) Model Comparison The AIC for the poisson regression is 493.06, whereas for the negative binomial regeression it is 408.76. Therefore the negative binomial regression is a better fit given the smaller AIC value. For neg. bin. the dispersion factor is: 1/k = 1/theta = 1/9.94. This value is not close to zero. Therefore we cannot use the poisson regression model since it would lead to overdispersion. Negative binomaial regression model is a better model to use.