Predicting the Number of Expected Boating Trips to Lake Somerville using Generalized Linear Models

TJ Smith

Section 1 - Introduction

When people are not working, they are generally spending time with friends and family enjoying their hobbies. Recreational activities such as boating, skiing, fishing, and hiking are common hobbies. One important aspect of these activities is that space can be limited as there are not unlimited resources for recreational businesses. These businesses could benefit from having a way to measure the demand from these activities so they can properly allocate their resources.

The data set at hand provides some variables that relate to boating in the town of Somerville. Analyzing these variables will give us a great indication of which variables help predict the demand (number of trips) someone will take to Lake Somerville for boating. Since the response variable we are analyzing is measured in counts, a Poisson model or negative binomial model can be used to analyze the data.

The remainder of the report will proceed as follows. Section 2 will discuss the characteristics of the data, including the nature of the data and some summary information about its importance. The process of how the model is selected and its interpretation will be in section 3. The report finishes with section 4 which contains concluding remarks.

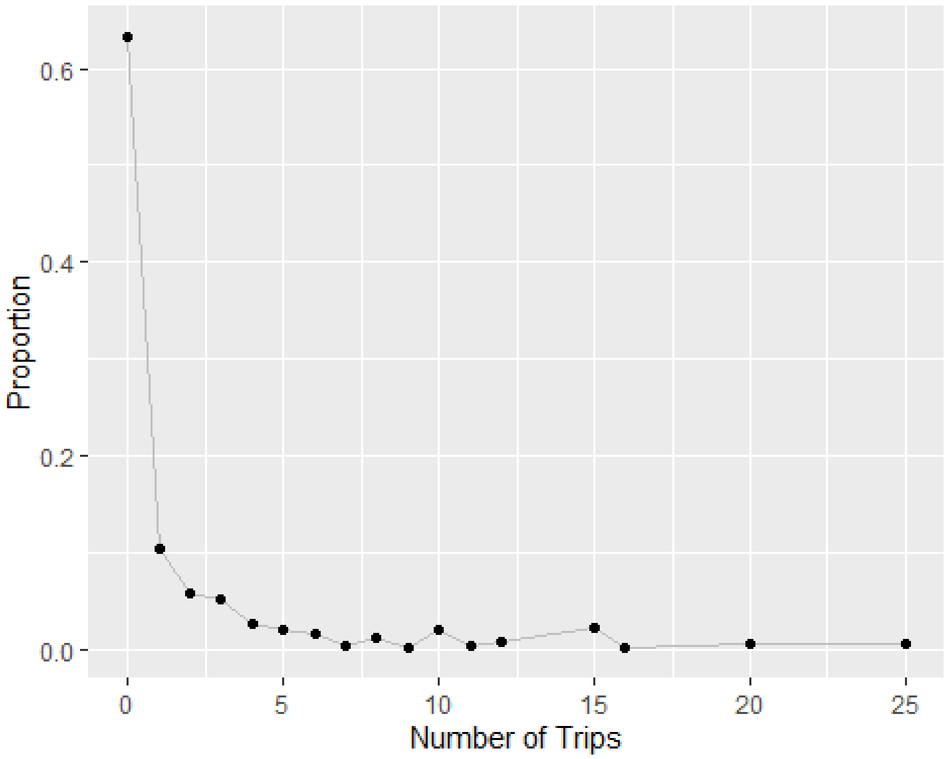
Section 2 - Data Characteristics

The data provided is observational and cross-sectional. It contains information from 659 surveys, which are filled out by potential users of boating trips. There are a total of 7 factors (independent variables) that can be used to predict how many boating trips a potential user might take. Some of these variables include expenditures when visiting certain lakes, the income of the head of the group, the facility's quality ranking, and whether or not the respondent engaged in water skiing while at the lake. The response variable is the actual number of boating trips that the person surveyed took over the last month. It is our job to analyze the data to determine which of the independent variables play a significant role in predicting the number of boating trips someone took at Lake Somerville.

The plot below provides some insight into the distribution of the response variable. It plots the total number of trips to Lake Somerville in the past month on the x-axis vs the proportion of trips on the y-axis. In context, the proportion shown below tells us that 63.26% of the people surveyed took 0 boating trips at Lake Somerville in the past month.

```r
pp <- pp[as.character(0)]
round(pp, 4)
```

```
##      0
## 0.6328

p <- ggplot(data=dftrips,mapping=aes(x=Trips,
                                y=freq.Trips))+xlim(0,25)
p <- p + geom_line(col = "gray")
p <- p + geom_point()
p <- p + labs(x = "Number of Trips",
              y = "Proportion")
p

## Warning: Removed 5 row(s) containing missing values (geom_path).

## Warning: Removed 5 rows containing missing values (geom_point).
```

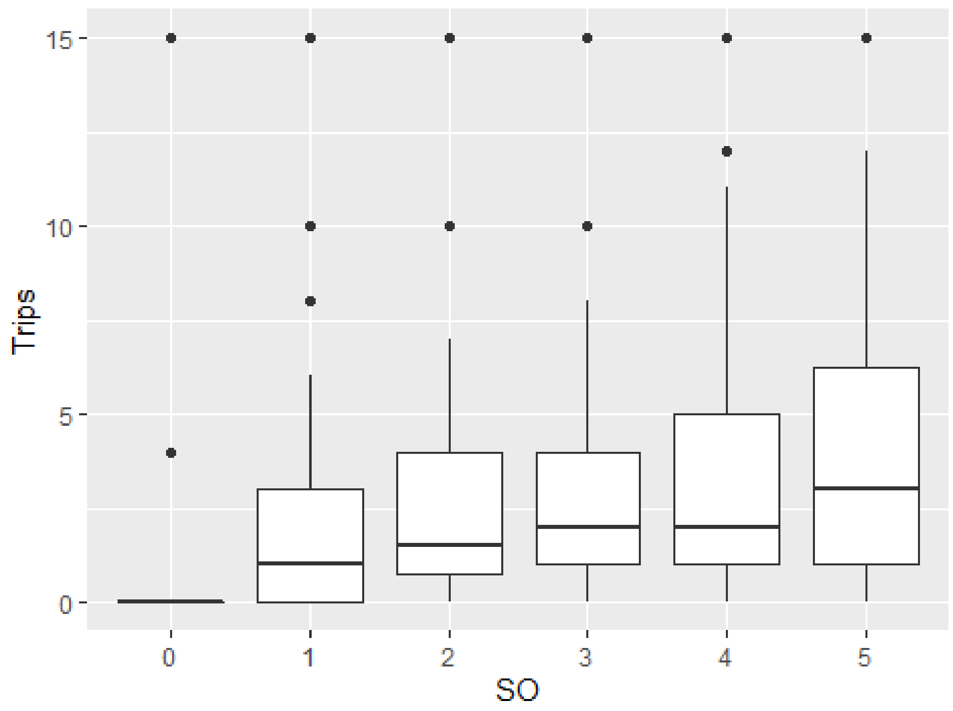


A vast majority of people take very few trips to Lake Somerville each month. There seems to be a constant decrease in the proportion value, but we observe slight increases at the values of 10 and 15. This pattern is most likely due to rounding as 10 and 15 might be common numbers to use if the person is unsure of the exact amount of trips they took to the lake.

An important note to mention, this plot only shows the number of trips up to the value of 25. In the data set, we have a few extreme data points going up to 50 and even 88 boating trips. The plot is limited to 25 for better visualization however We want to ensure that our fitted model will be able to account for these extreme observations.

Before fitting a model it is important to observe some exploratory plots that may give us an idea of which of the independent variables are significant in predicting the number of trips taken to Lake Somerville. Below shows the relationship between the variable SO (facility quality ranking) vs the number of trips taken to Lake Somerville. Since the quality ranking is categorical, not numerical, we must treat it as a factor.

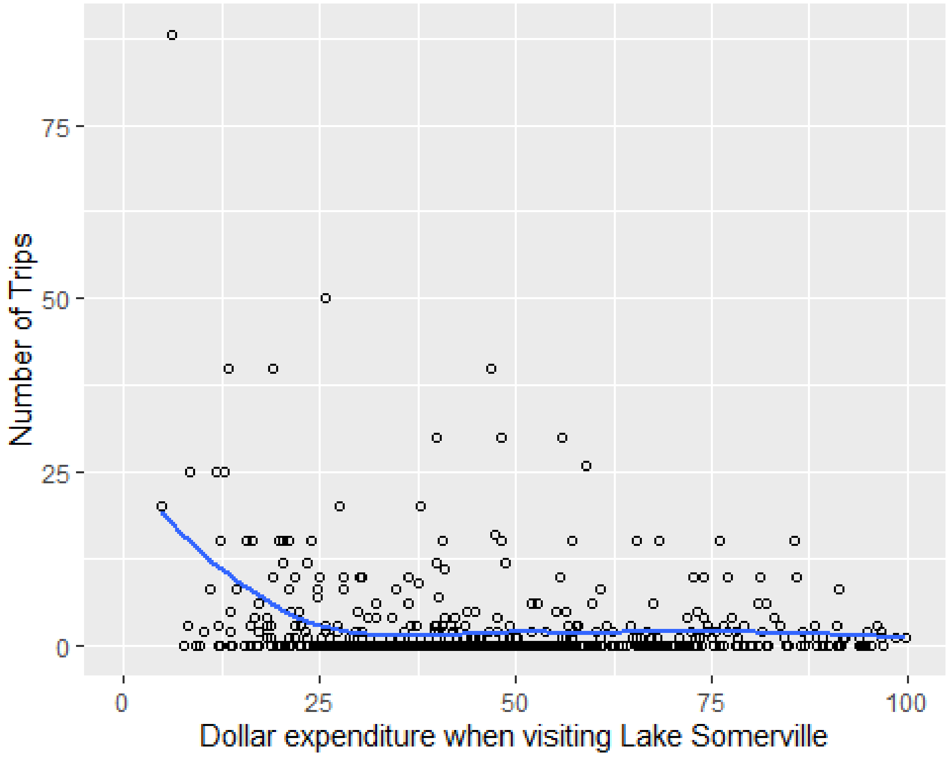
```
p <- ggplot(data = rec,
            mapping = aes(x = as.factor(SO),
                          y = TRIPS)) +ylim(0,15)
p <- p + geom_boxplot()
p <- p + labs(x = "SO",
              y = "Trips",
              title = "")
p

## Warning: Removed 16 rows containing non-finite values (stat_boxplot).
```



From the box plot, there seems to be a non-linear relationship between the different levels of SO. For example, the increase in the number of trips when SO increases from three to four is a lot smaller than when SO increases from four to five. As a result of this, it may be important to treat each level of SO as a different variable in the fitted model to capture these differences.

Similarly to the SO plot, when the variable C3 (dollar expenditures when visiting Lake Somerville) is plotted, there is a non-linear relationship. To ensure that the fitted model can account for this pattern, C3 should be included in the final model.

```
p <- ggplot(data = rec,
            mapping = aes(x = C3,
                          y = TRIPS)) +xlim(0,100)
p <- p + geom_point(pch = 1)
p <- p + geom_smooth(se=FALSE)
p <- p + labs(x = "Dollar expenditure when visiting Lake Somerville",
              y = "Number of Trips")
p

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 83 rows containing non-finite values (stat_smooth).

## Warning: Removed 83 rows containing missing values (geom_point).
```



We see from the model summary above that 83 rows have data have been removed due to not having finite values within the necessary. variables. While not an alarming amount, it is always important to be weary that models are not created with a high amount of data removal.

Section 3 - Model Selection

The previous section provided some preliminary plots and some indication of the distribution of the target variable (number of trips to Lake Somerville). This section will focus on narrowing down the data at hand to fit a generalized linear model that accurately predicts the number of trips.

As previously discussed, the response variable is given in counts (number of trips) so either a Poisson or a negative binomial model will be used to fit the data. This section will provide an interpretation of the fitted model as well as some graphics providing information on why the model fits well.

As discovered in the data characteristics section, the final model should include the quality ranking variable (SO) and expenditures when going to Lake Somerville(C3). To make sure that the non-linear relationship of SO is captured, each quality ranking (0-5) is included as its own individual variable. The following Poisson model was fit to see if these two variables fit the data on their own.

```
mi = glm(TRIPS ~ -1+SO0+SO1+SO2+SO3+SO4+SO5+C3,
         data=rec,family=poisson(link="log"))
summary(mi)
```

```
##
## Call:
## glm(formula = TRIPS ~ -1 + SO0 + SO1 + SO2 + SO3 + SO4 + SO5 +
##     C3, family = poisson(link = "log"), data = rec)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2969  -0.6219  -0.3405  -0.2561  14.1809
##
## Coefficients:
##       Estimate Std. Error z value Pr(>|z|)
## SO0 -2.2935951  0.2336534  -9.816   <2e-16 ***
## SO1  1.4659149  0.1247892  11.747   <2e-16 ***
## SO2  2.5515214  0.0768617  33.196   <2e-16 ***
## SO3  2.1794351  0.0608741  35.802   <2e-16 ***
## SO4  2.5018646  0.0667009  37.509   <2e-16 ***
## SO5  2.4047696  0.0757610  31.742   <2e-16 ***
## C3  -0.0132970  0.0009673 -13.747   <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: 5600.9  on 659  degrees of freedom
## Residual deviance: 2232.6  on 652  degrees of freedom
## AIC: 2999.7
##
## Number of Fisher Scoring iterations: 7
```

```
deviance(mi) / df.residual(mi)
```

```
## [1] 3.424263
```

Once the model was fit, it at first looks good as every variable is significant. However, if we observe the dispersion parameter (which is supposed to be a value of 1 for a Poisson model), we find a value of 3.42. If the dispersion parameter equals 1, that means that the mean and the variance of the data are equal. In our case, the variance is more than 3 times greater than the mean. If we use a Poisson model, we will see far too much variance within the fitted model so it might be better to fit a negative binomial model.

Since we found that the Poisson family would not fit the data, the following model was used instead which features a negative binomial function. The variables I, FC3, and SKI were added as they proved to improve the model. The variables used in the model are defined as follows:

SO0 -> Facility quality ranking of 0 SO1 -> Facility quality ranking of 1 SO2 -> Facility quality ranking of 2 SO3 -> Facility quality ranking of 3 SO4 -> Facility quality ranking of 4 SO5 -> Facility quality ranking of 5 C3 -> Dollar expenditure when visiting Lake Somerville I -> Household income of the head of the group ($1,000/year) FC3 -> A binary variable that equals 1 if the annual user's fee is paid at Lake Somerville. SKI -> Equals 1 if the person is engaged in water skiing at the lake.

A summary of the model is shown below.

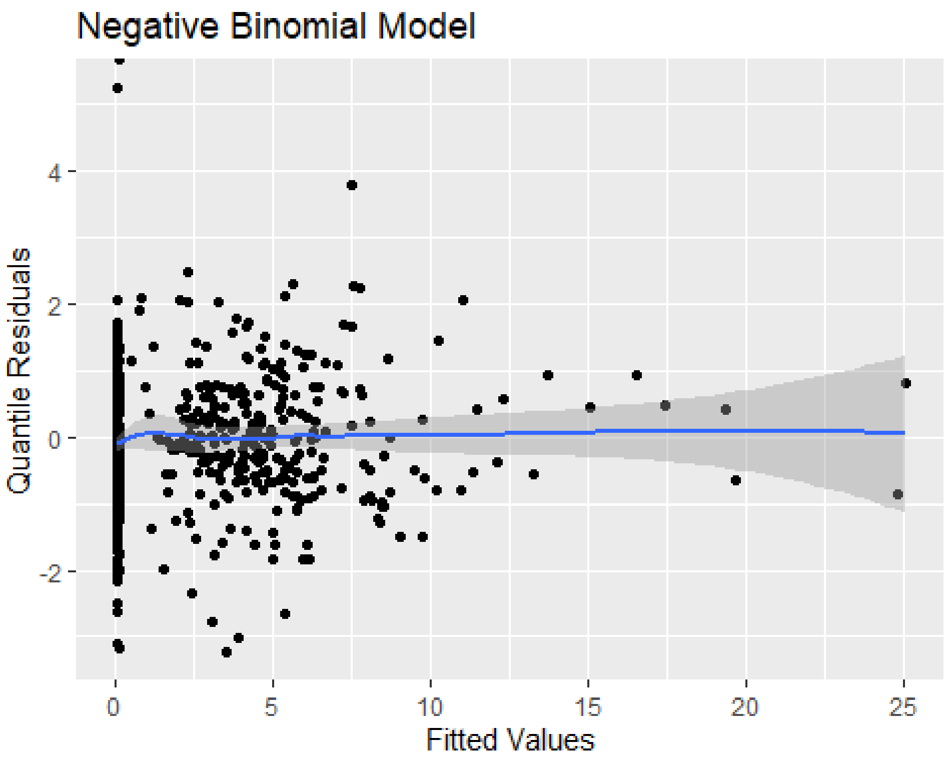
```
m0 = glm.nb(TRIPS ~ -1+SO0+SO1+SO2+SO3+SO4+SO5+C3+I+FC3+SKI,
         data=rec)
summary(m0)

## 
## Call:
## glm.nb(formula = TRIPS ~ -1 + SO0 + SO1 + SO2 + SO3 + SO4 + SO5 + 
##     C3 + I + FC3 + SKI, data = rec, init.theta = 0.7890820997, 
##     link = log)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.0224  -0.4014  -0.3129  -0.2355   8.1648  
## 
## Coefficients:
##      Estimate Std. Error z value Pr(>|z|)    
## SO0 -2.354980   0.287491  -8.191 2.58e-16 ***
## SO1  1.482257   0.298784   4.961 7.01e-07 ***
## SO2  2.305942   0.282872   8.152 3.58e-16 ***
## SO3  1.899590   0.214921   8.839  < 2e-16 ***
## SO4  2.412838   0.234231  10.301  < 2e-16 ***
## SO5  2.210585   0.249450   8.862  < 2e-16 ***
## C3  -0.008043   0.001981  -4.060 4.91e-05 ***
## I   -0.007532   0.002868  -2.626  0.00864 ** 
## FC3  1.039912   0.334810   3.106  0.00190 ** 
## SKI  0.606449   0.151227   4.010 6.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for Negative Binomial(0.7891) family taken to be 1)
## 
```

```
##     Null deviance: 1582.6  on 659  degrees of freedom
## Residual deviance:  404.2  on 649  degrees of freedom
## AIC: 1633.8
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.7891
##           Std. Err.:  0.0817
##
##  2 x log-likelihood:  -1611.7900
```

At a first glance, the negative binomial model fits well as every variable is significant and the dispersion of the data is close to one which is what is expected. However, just because the variables are significant does not mean that the fitted model actually analyzes the data well. An analysis of some diagnostic plots will give us an indication if the fitted model works well.

The first and most important plot to look at is a fitted values vs quantile residuals plot. The graph will tell us how well the model predicts the actual points in the data set. If the plot demonstrates a straight line across y=0, this will tell us that the model fits very well. It is also important to ensure that there are not too many data points that differ significantly from a y-value of zero, as that means the model does not predict those data points well.

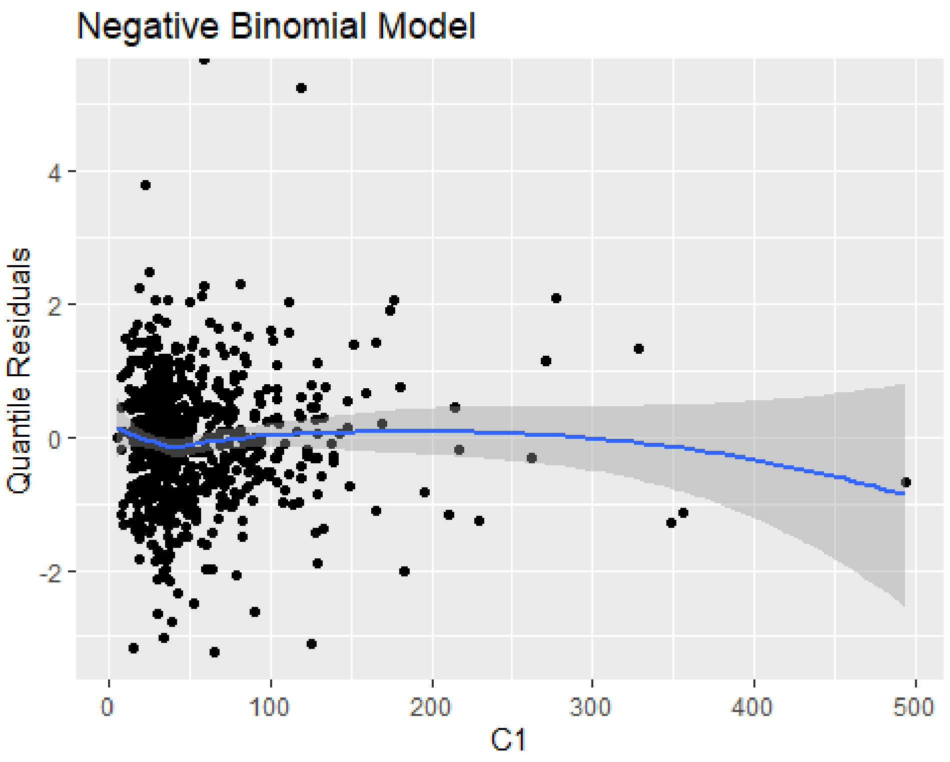
```
p <- ggplot(data = rec,
            mapping = aes(x = m0.mu,
                          y = m0.rQ))
p <- p + geom_point() + geom_smooth(se = TRUE)
p <- p + labs(x = "Fitted Values",
              y = "Quantile Residuals",
              title = "Negative Binomial Model")
p

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```



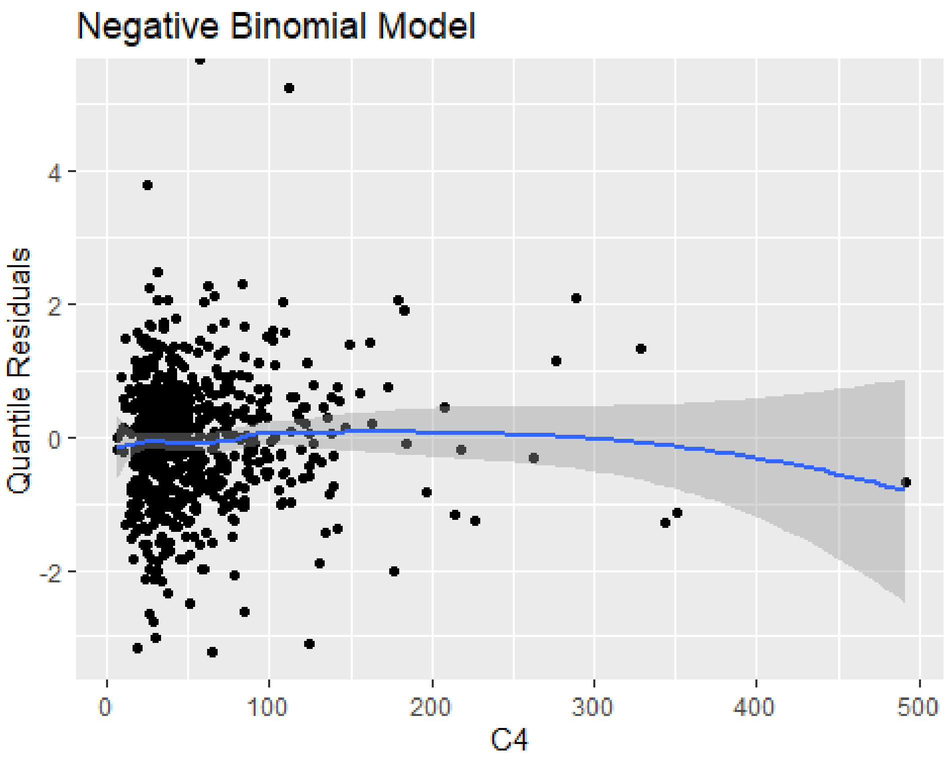
As we see from the plot above, there is a fairly straight line across y=0 which indicates that the fitted model works well. There are also not too many data points that sway too far from 0 which tells us that our model works well for most of the points in the data set. We do see a slight bump in the line initially but is not enough to indicate any issues with the model.

Now that we know our mode fits the data well, it is important to observe the variables excluded from the model that is in the data set to ensure that they wouldn't add any additional information that the model doesn't already capture. To do this we can plot the out-of-model variables against the model. If there is a pattern different from y=0, that may indicate that we should include the respective variables in the model.

```
p <- ggplot(data = rec,
            mapping = aes(x = C1,
                          y = m0.rQ))
p <- p + geom_point() + geom_smooth(se = TRUE)
p <- p + labs(x = "C1",
              y = "Quantile Residuals",
              title = "Negative Binomial Model")
p

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```



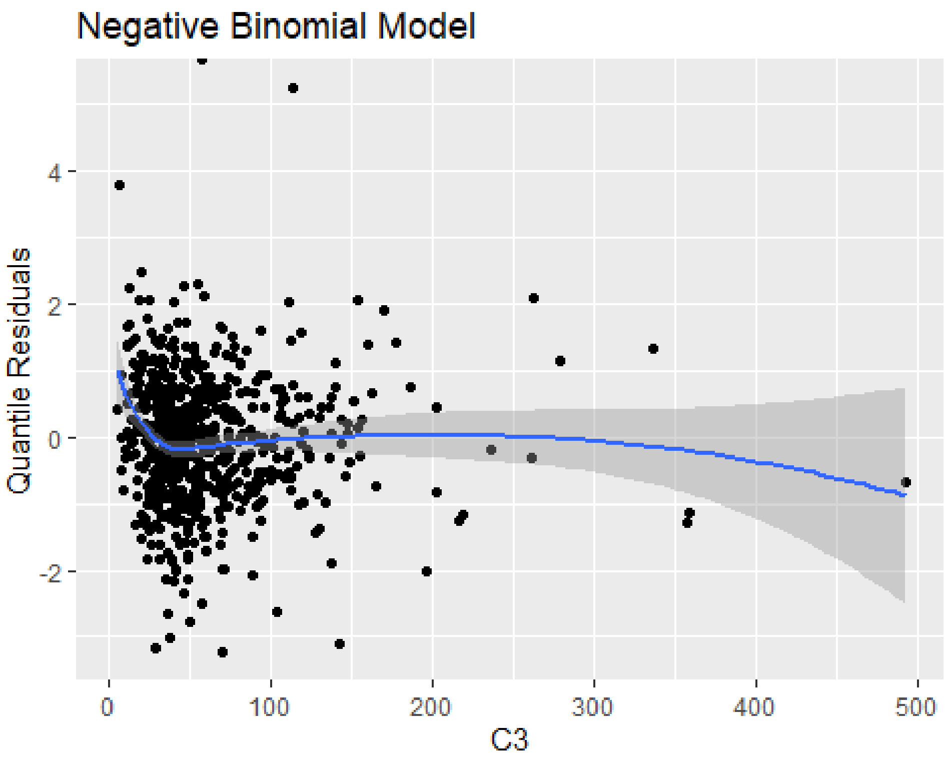
```
p <- ggplot(data = rec,
            mapping = aes(x = C4,
                          y = m0.rQ))
p <- p + geom_point() + geom_smooth(se = TRUE)
p <- p + labs(x = "C4",
              y = "Quantile Residuals",
              title = "Negative Binomial Model")
p

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```



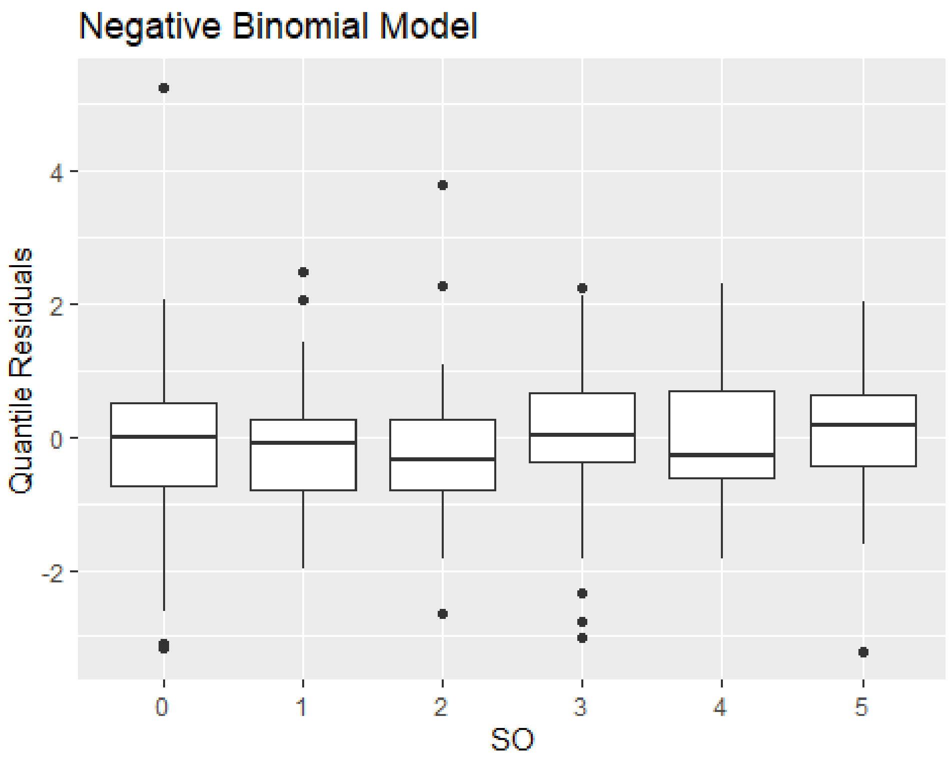
Based on the out-of-model plots for C1 and C4, there isn't any evidence that these variables will add anything to the model.

It is similarly important to look at the in-model variables just like the out-of-model variables. Since we know that the in-model variables are already significant, we want to ensure that they are on the correct scale within the model. Like the previous plots, if we observe a line across y=0, that indicates that the respective in-model variable is on the correct scale.

```
p <- ggplot(data = rec,
            mapping = aes(x = C3,
                          y = m0.rQ))
p <- p + geom_point() + geom_smooth(se = TRUE)
p <- p + labs(x = "C3",
              y = "Quantile Residuals",
              title = "Negative Binomial Model")
p

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```



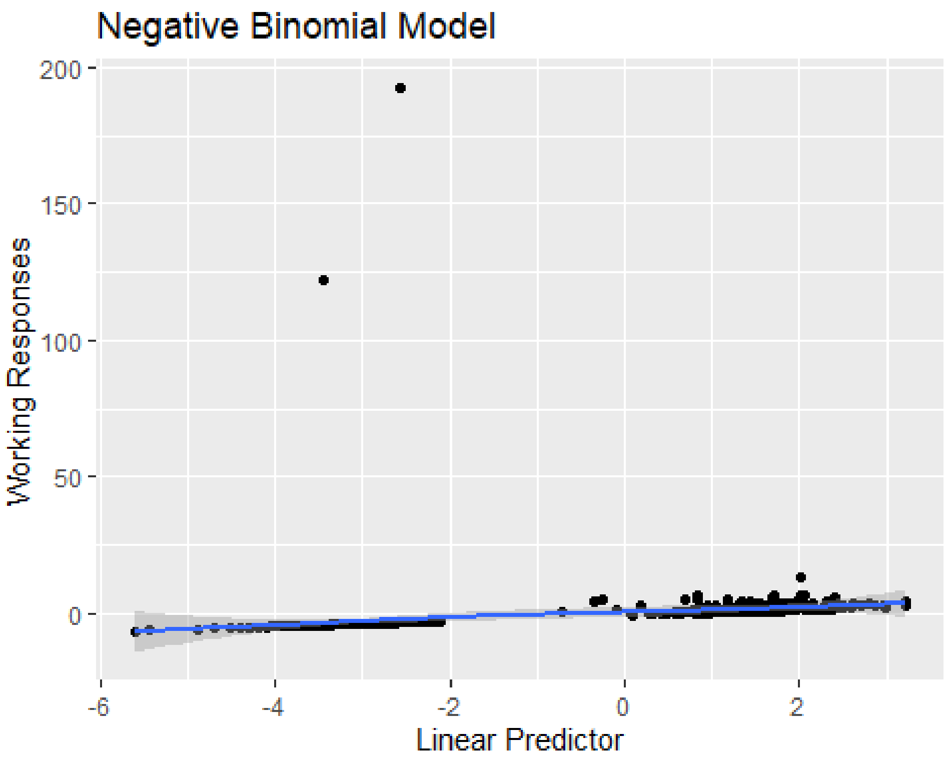
```
p <- ggplot(data = rec,
            mapping = aes(x = as.factor(SO),
                          y = m0.rQ))
p <- p + geom_boxplot()
p <- p + labs(x = "SO",
              y = "Quantile Residuals",
              title = "Negative Binomial Model")
p

## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



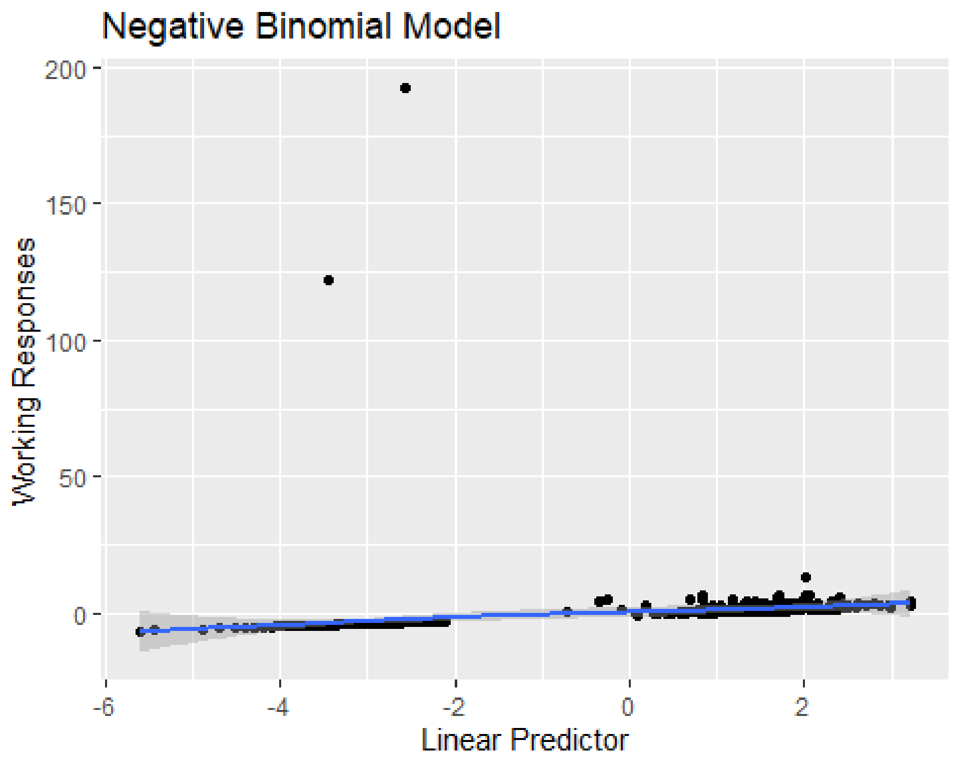
As seen from the plots above, the general trend for each variable does not sway far from y=0, telling us that the SO variables and C3 are on the right scale. Though not included, the same plot was fit for the other in-model variables, all proving to be on the correct scale as well.

The following linear predictor vs working responses plot indicates whether or not the link function used in the model fits well. In our case, we used a log link. We want to observe a linear relationship between the predictor and the responses.

```
p <- ggplot(data = rec,
            mapping = aes(x = m0.eta,
                          y = m0.wR))
p <- p + geom_point() + geom_smooth(se = TRUE)
p <- p + labs(x = "Linear Predictor",
              y = "Working Responses",
              title = "Negative Binomial Model")
p

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



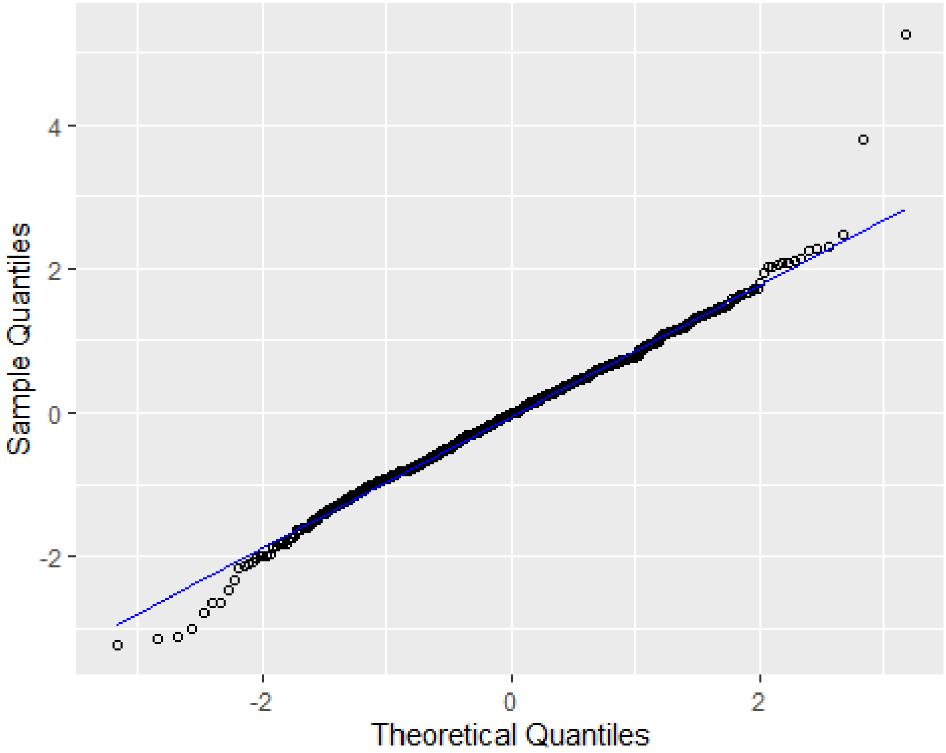
```
p <- ggplot(data = rec,
            mapping = aes(x = m0.eta,
                          y = m0.wR))
p <- p + geom_point() + geom_smooth(se = TRUE)
p <- p + labs(x = "Linear Predictor",
              y = "Working Responses",
              title = "Negative Binomial Model")
p

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The plot gives us a highly linear relationship indicating the log link fits the model well. We do see 2 data points swaying far from the trend line however that does not give us any alarming evidence that the log link does not work.

The final diagnostic plot to observe is a QQ plot which tells us if the negative binomial distribution is the correct distribution to use for this data set. We want to observe a linear pattern between the data points.

```
pt <- ggplot(data = rec,
            mapping = aes(sample = m0.rQ))
pt <- pt + geom_qq(pch = 1) + geom_qq_line(color = "blue")
pt <- pt + labs(x = "Theoretical Quantiles",
              y = "Sample Quantiles")
pt

## Warning: Removed 1 rows containing non-finite values (stat_qq).

## Warning: Removed 1 rows containing non-finite values (stat_qq_line).
```



As expected, there is a fantastic linear relationship in the QQ plot indicating that the negative binomial distribution is the correct one to use.

After reviewing each of the diagnostic plots, it seems that the fitted model fits well, the variables excluded will not add anything to the model, the variables in the model fit well, and the negative binomial distribution with a log link fits the data well. Considering this we can now interpret the model.

The equation of the final model:

$$\begin{aligned} log(\mu) \\ = -2.355(SO0) + 1.482(SO1) + 2.306(SO2) + 1.90(SO3) + 2.413(SO4) + 2.211(SO5) \\ - .008(C3) - .008(I) + 1.040(FC3) + .606(SKI) \end{aligned}$$

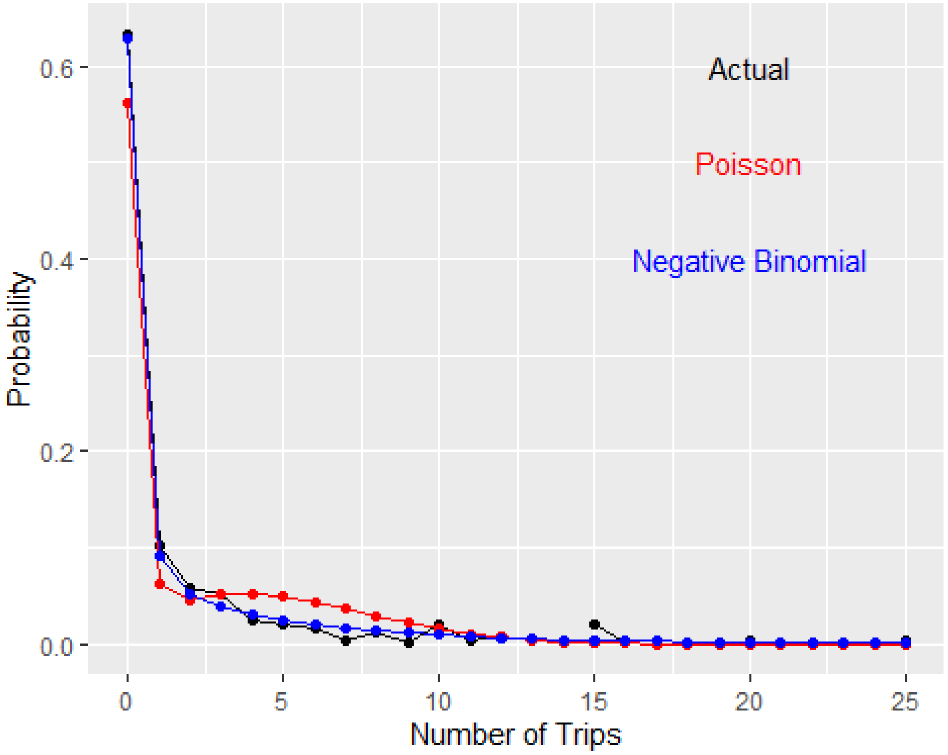
It is important to understand what each variable means in context. In our case, we have 2 different types of variables: binary (SO, SKI, and FC3) and continuous (C3 and I). The binary variables take a value of 0 if the condition is false, or a 1 if the condition is true. For example, if the quality ranking for a facility was chosen as 3, SO3 would be 1, and all other SO variables would be 0. Theoretically, the continuous variables can take any value, but since we are dealing with expenditures and income, the values can take any non-negative number. An interpretation for each type of variable is as follows:

If the quality ranking for a facility is 0, then on average, holding all other variables constant, we expect a person to have a decrease in trips to Lake Somerville by a multiplicative factor of e^-2.355. This interpretation can be used for the other quality rankings and their respective coefficients, as well as the SKI and FC3 variables when their respective condition is true.

If the head of the household’s income is $60,000, then on average, holding all other variables constant, we expect a person to have a decrease in trips to Lake Somerville by a multiplicative factor of e^-0.48. This interpretation can be applied to the C3 variable, which is expenditures when visiting Lake Somerville, with its respective coefficient.

The final plot that is good to observe is to compare the actual data points versus what the fitted model expects the values to be. The plot below shows the comparison between the actual data points in black, the Poisson models’ expected values in red, and the negative binomial models’ expected values in blue.

```
p <- ggplot(data = tbl,
            mapping = aes(x = TRIPS,
                          y = Actual))
p <- p + geom_line() + geom_point()
p <- p + geom_line(aes(y = Expected),
                   color = "red") +
  geom_point(aes(y = Expected),
             color = "red")
p <- p + geom_line(aes(y = Exp.NB),
                   color = "blue") +
  geom_point(aes(y = Exp.NB),
             color = "blue")
p <- p + annotate("text", x = c(20,20,20), y = c(0.6, 0.5,.4),
                  label = c("Actual", "Poisson","Negative Binomial"),
                  color = c("black", "red", "blue"))
p <- p + labs(x = "Number of Trips",
          y = "Probability")
p

## Warning: Removed 9 rows containing missing values (geom_point).
```



As expected, the negative binomial does a fantastic model of predicting the actual values given in the data set when compared to the initial Poisson model.

Section 4 - Conclusion

Many variables in someone's life may affect how often someone visits a lake to go boating. Despite how many factors there may be, it is possible to establish a model that determines the target variable well. As we have shown, a negative binomial model including the variables SO0, SO1, SO2, SO3, SO4, SO5, C3, I, FC3, and SKI proves to predict the number of trips someone takes to Lake Somerville well.

It is important to note that the analysis of this data set is based solely on trips to Lake Somerville as the target variable. This model may not translate well to predicting demand at other lakes. Different lakes across the world may have other variables that predict their demand better. Nonetheless, although the significant variables may differ, the techniques used in this report can be applied to other data sets that are attempting to predict a response variable measured in counts.