Weather Conditions and Their Impact on Seoul Bike Rentals

Rey Govea

11/29/2020

Introduction

The development of bike sharing options provides people with access to easily accessible transportation around cities. The interest in bike sharing locations provides easy accessibility of movement as well as tourist options to experience city life. With instant access to rental options, there is an interest in measuring the weather impacts on rental rates within the city. Considering the weather implications of rate of rentals allows for businesses consideration of what factors will impact the behavior of their customer base. We want to measure what weather conditions provide significant impacts on the bike rental counts for the city of Seoul.

The use of rental bikes in Seoul contributes to the commercial access to transport around the city. Since its inception in 1999, Asia has experienced a massive increase in their development of bike sharing programs. The first bike-sharing program was called TownBike in Signapore and now there are bike rental centers in South Korea, Taiwan, and China (Shaheen 163). This growth in the bike rental industry took about a decade to become accessible in South Korea. The first bike rental program in the country was started in 2008 under a program called Nubija in the city Chongwan (Shaheen 163). With bike sharing being a relatively new opportunity for business in the region, there is room for growth in a large city such as Seoul. If there became an opportunity to adjust access or management of the bike system, predictions of what weather conditions impact use can contribute to the management of needs such as maintenance of the property. With large populations having access to the bikes, scheduling times when the companies would maintain their bikes to maximize use could be one type of problem addressed to account for the use of their product.

Part of the success of rental bike use is the accessibility that technology provides and its alternative to climate change contributing transportation systems. When people want to be environmentally conscious about their transportation methods, biking is an easily accessible method in big cities to contribute to reducing carbon footprints. The bike sharing options within the city have provided a “healthy environment for citizens in Seoul to live” (Sathishkumar 166). On top of the conscientious customers that this model attracts, this increase in bike use has been made easier with cell phones and apps that allow for easy purchase of use. Bike sharing companies want their products to be easily accessible and their app systems allow for convenient features like nearby bike location information that allows for quick use (Sathishkumar 166). With the growth of the bike rental programs increasing and this business providing clean and easily accessible transportation, the benefit of these operations is clear.

If a bike rental program wants to maximize its supply in circulation in order to meet demand, an understanding of the weather impacts are important to understand. In order to assess the impacts of weather conditions on demand for bike renting, we want to consider three different models and assess which one best fits the true data. The null hypothesis is that no model is better at predicting the data than another while the alternative hypothesis is that some regression model is better than the others at predicting the relationship between bike rental use and weather conditions in the population.

Data Description

In order to measure the impact of weather conditions in Seoul, 8760 observations were collected from the "Seoul Bike Sharing Demand Data Set" from the UCI Machine Learning Repository. The data consists of a response variable called "Rental Bike Count" which measures the number of bikes rented each hour and 9 other integer and real variables. The real and integer variables include in order of appearance from $x_1 - x_9$ include: Hour, Temperature, Humidity, Windspeed, Visibility, Dew Point Temperature, Solar Radiation, Rainfall, and Snowfall. In order to measure what significant variables are in the data set we applied three different models to see which best fit the data.

Methedology

After collecting data from the UCI machine learning center, we applied different linear models to the data to assess whether the independent variables were influential on predicting the response. Overall, three models were used when attempting to apply regression analysis for predicting the Rented Bike Count and the best model was then applied for final analysis. The three models used where Multiple Linear Regression, Poisson Regression, and Negative Binomial Regression. In order to select the best model out of the nine independent variables, Stepwise-Selection was used. With our large data set (n=8760) relative to the amount of variables, this method to begin the analysis was deemed sufficient to select a significant model to fit the data. Before Stepwise-Selection our model was as follows…

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_9 x_9 + \epsilon$$

… with the application of Stepwise-Selection, we wanted to create a model that had the lowest AIC value to indicate a model that best fit the data. With a final AIC value of 107821, the best model for the collection of variables reduced the regression analysis to five variables.

$$\widehat{RentedBikeCount} = \beta_0 + \beta_1 Hour + \beta_2 Temperature + \beta_3 Humidity + \beta_7 SolarRadiation + \beta_8 Rainfall + \epsilon$$

Now that the model had been reduced to include significant independent variables, for every regression model used to assess the data there could not exist multi-colinearity between the variables in the model. In order to assess if there was a violation of this assumption in the new model, the Variance Inflation Factor (VIF) was applied to the data.

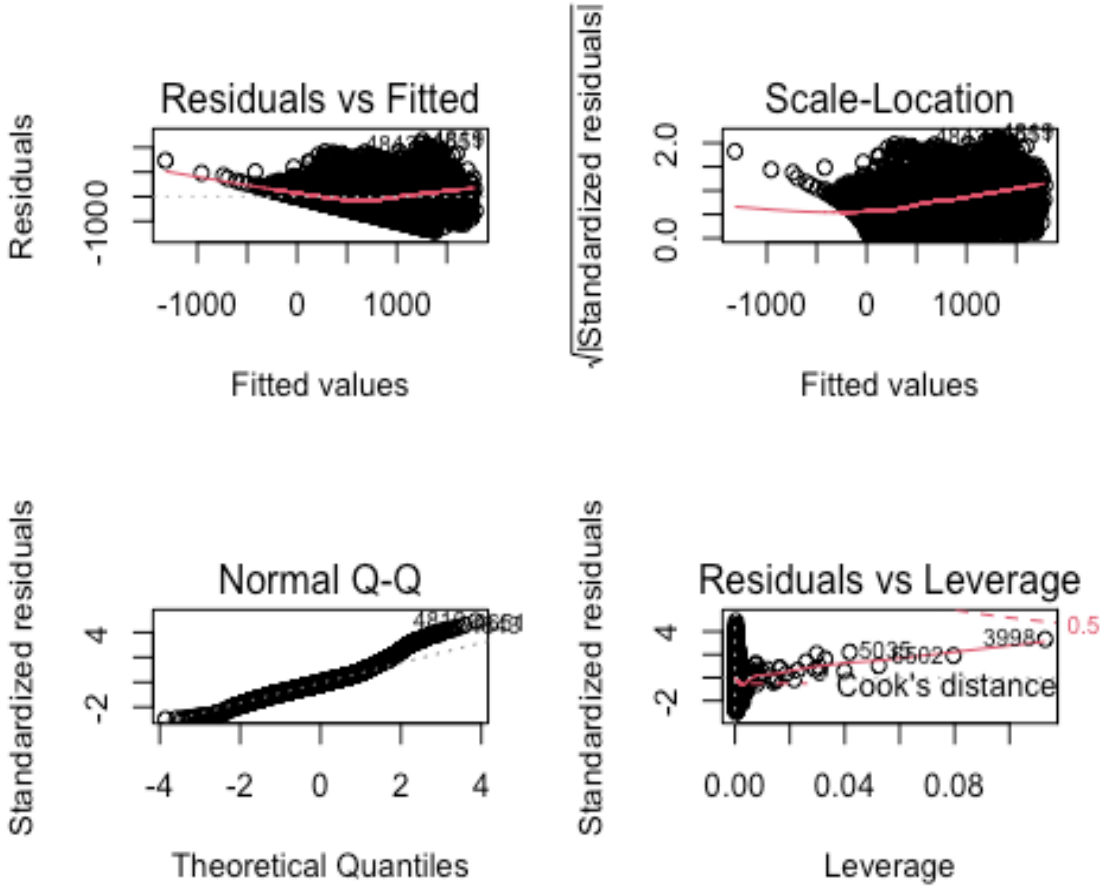
$$VIF_i = \frac{1}{1 - R_i^2}$$

```
##        Hour Temperature    Humidity Solar_Radiation    Rainfall
##    1.101892    1.388472    1.682460       1.675621    1.066328
```

This model measures the correlation that each independent variable has with all of the other independent variables. For assessing the impact that is present, we would consider a VIF of >10 to indicate high correlation with the other variables and a VIF >5 to be of moderate consideration and importance to the model. It is apparent that with the selected model, there is no significant VIF levels within the data. this indicates that for this model moving forward, we have validated the assumption in the linear models that there is no multi-colinearity present within the data.

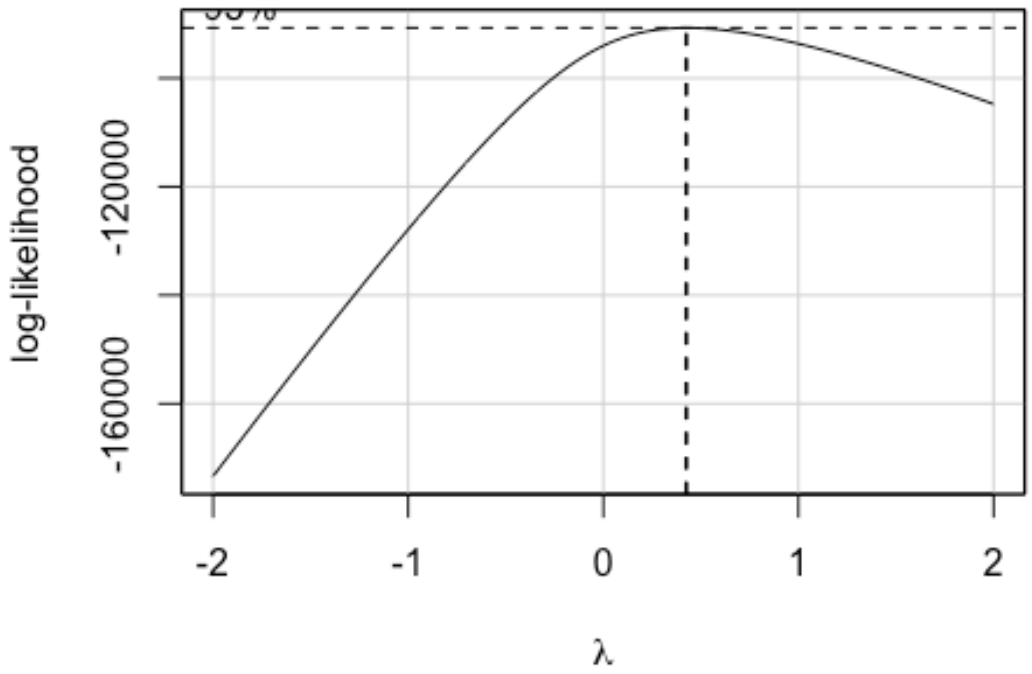
Multiple Linear Regression

The Multiple Linear regression method was first used as a means to apply the simplest model to the data and assess whether or not it's application sufficiently applied to the data. This model assumes that there is constant variance, the data is normally distributed, no multi-colinearity exists, and there is Homoscedsticity within the data. In order to address these assumptions, looking at a residuals plot would provide feedback on any disruptions within the data that would lead to potential adjustments in the model. When looking at the residuals and quantile plot we want to look for patterns in the data that indicate violations of the previous assumptions.

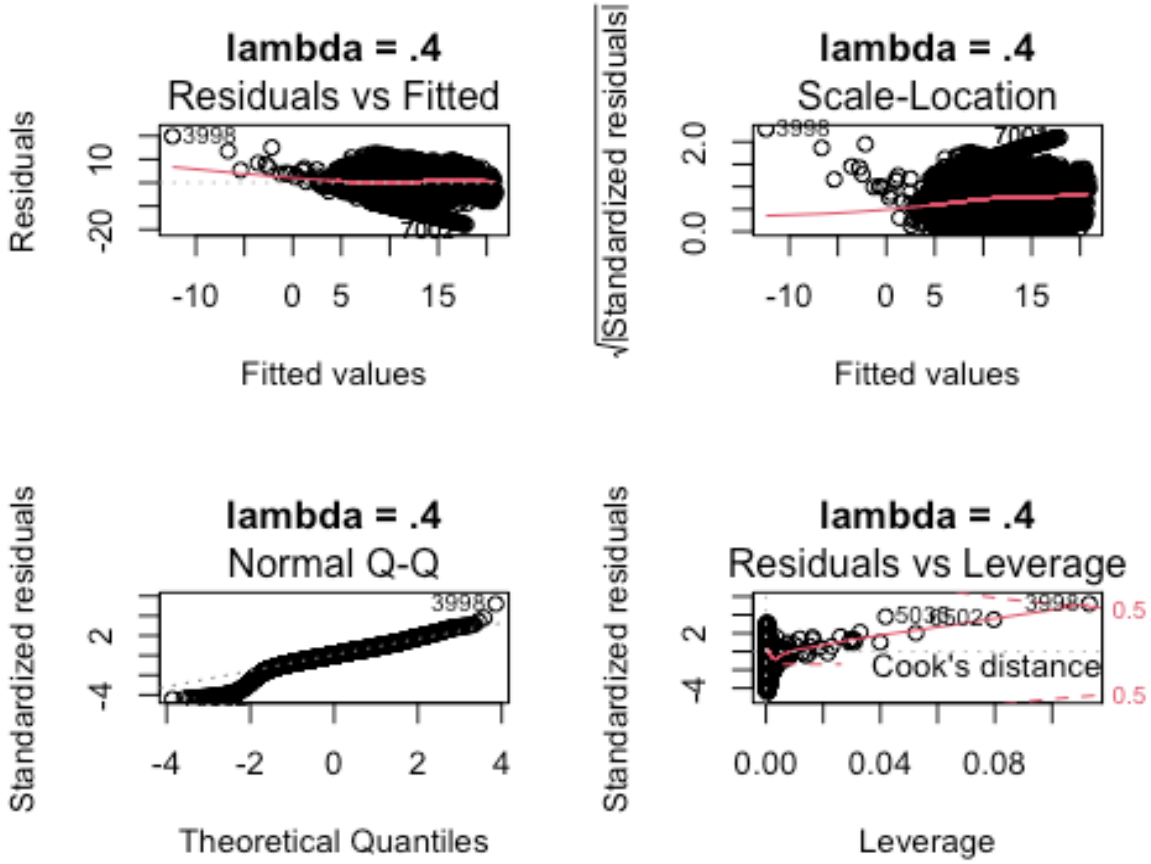


The multiple regression model shows clear problems with non-normality and patterns in the residuals. Addressing the residuals first, both the residuals v. fitted and standardized residuals v. fitted graph show patterns within the data that do not place it on a completely linear pattern and violate constant variance assumptions. The clear sloped line within the residuals also indicates that there could be a minimum value within the data. This observation makes sense because both our response variable and several of our independent variable can not be values below zero. Overall, the models non-constant variance with slight heteroscedsticity poses a potential violation to the multiple linear regression model that needs to be addressed. When assessing the Normal Q-Q plot, deviation from the theoretical quantiles at the right side indicates that our distribution is right heavy and not normal. The initial solution to this problem is to address the potential transformations to the data that would both normalize the model and correct some of the non-constant variance.

With The normal linear model not holding to the assumptions we wanted to apply an appropriate transformation to the data to see if our problems with the residuals would be fixed. In order to assign the correct model to the data, the Yeo-Johnson transformation estimator was applied to the data. This method was used instead of the Box-cox method because the bike rental data contains a lot of zero values in which the log transformations used in box cox produce undefined outcomes. The Yeo-Johnson method accounts for zero and negative values and outputs suggested λ values for power transformations to the data. The graphical outcome with the line estimating the suggested power transformation is as follows.



with a possible range of -2 to 2 we can see that with 95% confidence, the Yeo-Johnson method suggests that a power transformation of $\lambda \approx .4$ is the suggested rout to take to better the model. This new model was applied and the the residuals to the graph was analyzed for the same assumptions as the original model.



It appears as if there is still some pattern within the data that exists toward the left end of the residuals. However, the residuals appear relatively randomized. The two problems that continue to persist with this models assumptions lies in the non-normality of the data and the appearance of potentially influential outlines. Specifically, observation 3998 crosses the line of 0.5 and becomes flagged as influential and the Normal Q-Q plot inverses its graph to now indicate that the transformation suggests it is still non-normal and the tails are heavy on the left side now.

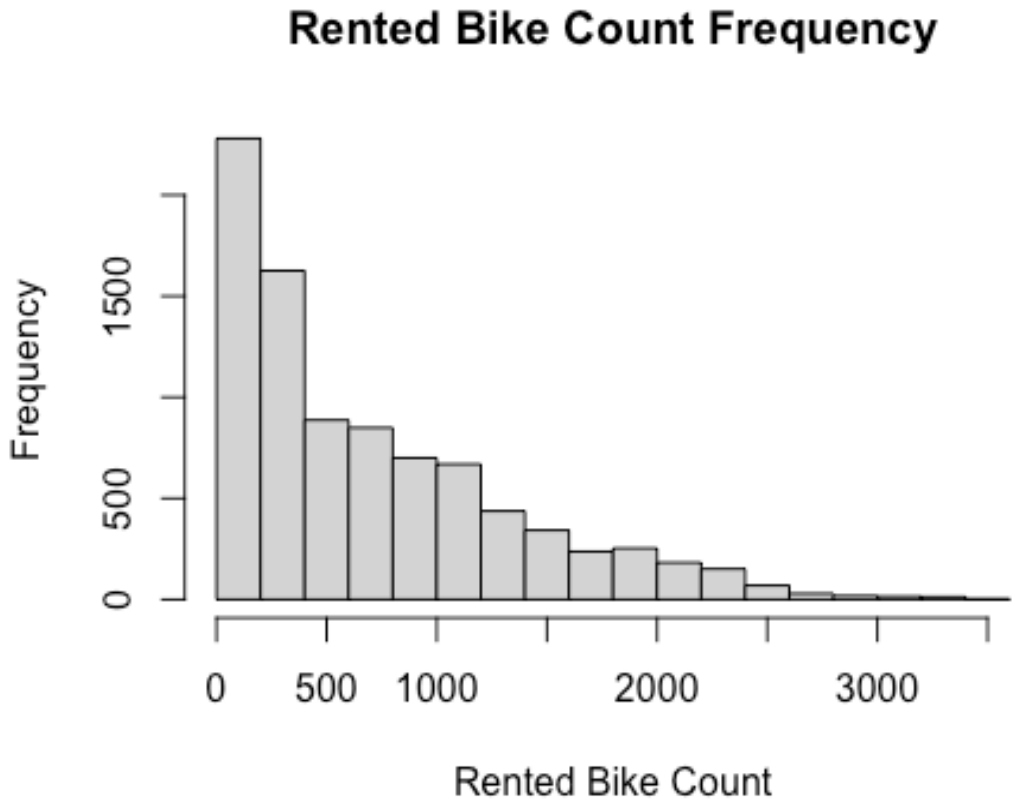
Poisson Model

The non-normality of the data suggests that multiple linear regression is not a completely appropriate model to represent the data. Another model that is good for representing non-negative count data is the Poisson regression model. In order for the Poisson regression model to adequately represent the data, there a few assumptions that the model requires. The assumptions that this model requires are that the response variable contains count data; all of the counts are positive integers; the distribution should have the mean be equal to the variance; the independent variables should be continuous, binary, or ordinal; and the observations should be independent.

Looking at the data, we can verify a few of the assumptions before any analysis. For example, the response variable is count data over time, all of the counts are positive integers, and we know that there is independence from the VIF of the model being used. This means that the important measures to verify the conditions of this model are to see if there is a Poisson distribution in the data's shape and if the mean and variance of the model are equal. The poisson model

```
## [1] 704.6021
```

```
## [1] 416021.7
```



The distribution seems to be skewed right like a Poisson distribution, the shape of the response variable then seems to satisfy the assumption of the model visually. However, there is a problem with the distribution of the data. The spread of the distribution violated the assumption of the Poisson model that the mean and variance are equal. The distribution of the data shows that $\overline{y} = 704.6021$ with $s^2 = 416{,}021.7$. The discrepancy in the variables indicate that there is overdispersion in the data and a model that conditions the two statistic being equal will not be a significant measure of the relationship that the independent variables will have on the response.

```
##
## Call:
## glm(formula = Rented_Bike_Count ~ Hour + Temperature + Humidity +
##     Solar_Radiation + Rainfall, family = poisson(link = "log"),
```

```
##     data = Bike)
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -55.426  -12.185   -3.408    7.859  108.966
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      5.841e+00  2.096e-03  2786.6   <2e-16 ***
## Hour             4.370e-02  6.662e-05   656.0   <2e-16 ***
## Temperature      4.507e-02  4.166e-05  1081.8   <2e-16 ***
## Humidity        -8.522e-03  2.760e-05  -308.7   <2e-16 ***
## Solar_Radiation -7.909e-02  5.416e-04  -146.0   <2e-16 ***
## Rainfall        -5.251e-01  2.161e-03  -243.0   <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: 4979261  on 8759  degrees of freedom
## Residual deviance: 2404219  on 8754  degrees of freedom
## AIC: 2471324
##
## Number of Fisher Scoring iterations: 6
```

We can also see that the residual deviance of the model shows that there is a large ratio when comparing the deviance to the degrees of freedom. This large deviance residual indicates that the model is not as good of a fit even if the variables are significant in predicting the response variable.

Zero Inflated Negative Binomial Model

With the Poisson model being too restrictive of a method to measure the bike rental data, we want to utilize a model that does not care about heteroscedsticity and places no restriction on the sample mean being greater than the variance. The negative binomial model provides a good measurement of the of the relationship between the independent variables and the dependent variable because of the condition of overdispersion does not apply to the analysis. Also, with 295 zero values within the data for rented bike counts, applying a zero inflation model controls for the large amounts of zero counts within the data.

```
##
## Call:
## zeroinfl(formula = Rented_Bike_Count ~ Hour + Temperature + Humidity +
##     Solar_Radiation + Rainfall, data = Bike, dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.3887 -0.6886 -0.1819  0.4458 20.1355
##
## Count model coefficients (negbin with log link):
##                   Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)      5.9020799  0.0372369 158.501  < 2e-16 ***
## Hour             0.0380900  0.0011903  32.002  < 2e-16 ***
## Temperature      0.0542963  0.0007908  68.657  < 2e-16 ***
## Humidity        -0.0108776  0.0005092 -21.362  < 2e-16 ***
## Solar_Radiation -0.0624389  0.0113823  -5.486 4.12e-08 ***
## Rainfall        -0.1230103  0.0043945 -27.992  < 2e-16 ***
## Log(theta)       0.7418173  0.0145034  51.148  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -3.815049   0.288422 -13.227  < 2e-16 ***
## Hour            -0.006804   0.009043  -0.752    0.452
## Temperature      0.025176   0.005993   4.201 2.66e-05 ***
## Humidity         0.003269   0.003880   0.843    0.400
## Solar_Radiation -0.028819   0.086206  -0.334    0.738
## Rainfall        -0.039452   0.070864  -0.557    0.578
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 2.0997
## Number of iterations in BFGS optimization: 1
## Log-likelihood: -6.24e+04 on 13 Df
```

When we run the Zero Inflated Negative Binomial model we can see that with a residual range of the data is from [-1.3887, 20.1355] and there is an IQR of 1.835. we can infer that there are some outliers present in the distribution, this may be of some concern and should be considered movin g forward. With the zero inflated model, we can see that Temperature is the only significant varia ble with a p-value of 2.66e-05. This indicates that there is a statistically significant relationship b etween the Temperature variable and the log expected Rental Bike Count. We can interpret the c oefficient to be that for every one unit change in temperature, the difference in the log of the exp ected Rental Bike Count is expected to change by 0.025176, given that the other predictors are h eld constant.

Model Comparison

After establishing the three different models to apply to the data, we want to apply statistics to understand which model best fits the data. The first tool of measurement to establish the model that is the best fit of the General Linear Models (GLM) is looking at the Akaike Information Criterion (AIC) values for each model. This statistic applies to GLM's by analyzing the the maximum likelihood function ($\hat{L}$). Using this comparison method, we can estimate the model that best fits the data by selecting the lowest AIC value of the group. Each AIC gives a value that corresponds to the degree in which the model best fits the true data. We apply this comparison method to the MLR, Poisson, and ZINB models to see which one best represents the data. The AIC values are provided in the order of Multiple Linear Regression, Poisson, and ZINB model.

$$AIC = 2k - 2ln(\hat{L})$$

```
## [1] 132651
```

```
## [1] 2471324
```

```
## [1] 124821.8
```

We also want to see if the ZINB model is a better fit than the Poisson regression model through a statistical test. In order to establish this model as a better tool for the regression analysis, we performed the Vuong test based on the likelihood ratio of both models. This model takes the likelihood estimates that were used in the AIC and applies a ratio between the two models and applies a function to estimate the better fitting model. The Poisson model and the ZINB model were used for comparison because of their lack of conditions for linearity and their previous AIC scores. Looking at both models it seems like they would be the two better fitting measurements for the regression so the test is applied to their comparison. By analyzing the two models, the test forms these hypotheses with $\alpha = .05$…

H_0:Both models have equal probability of generating the true data

H_1:The Zero Inflated Negative Binomial model is a closer fit to the true data

```
## NA or numerical zeros or ones encountered in fitted probabilities
## dropping these 292 cases, but proceed with caution
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##                Vuong z-statistic             H_A    p-value
## Raw                   -67.88394 model2 > model1 < 2.22e-16
## AIC-corrected         -67.88346 model2 > model1 < 2.22e-16
## BIC-corrected         -67.88175 model2 > model1 < 2.22e-16
```

The Vuong test showed that for all three measurements of raw scores, AIC corrected, and BIC corrected measurements, there is a statistically significant rejection of the null hypothesis. We can conclude that the Zero Inflated Negative Binomial model (ZINB) is a better fitting model than the Poisson model when measuring the relationship between the independent variables and Rental Bike Counts. There is a warning message in the model comparison test in that because our data set holds numerical zero's and 1's there has been enough dropped to consider that observations have been dropped from the model to account for the values unable to be accounted for in the log model. This limitation to the comparison method should be considered when comparing the models. However, realizing that the Poisson model has limitations with its over-dispersion problems, it should be reasonable to expect the ZINB model to be a better fit.

Discussion/Conclusion

We can see that when comparing the three models the Zero Inflated Negative Binomial Model (ZINB) best fits the data and best encompasses a significant representation of the impact that weather has on expected rental bike counts. While looking at the three models, the best model to predict the outcome of bike rentals was looking at temperature. We wanted to look at the best fit of the multiple regression, Poisson, and ZINB models and with the consideration of the model assumptions, the ZINB model was the better fit for the data. Using the AIC to compare all three

models, the ZINB model was greatly smaller than both the multiple linear regression model and the Poisson model. When considering the implications of the ZINB model, the other variables were not considered significant enough to be impactful on the log rental bike count. This allows us to draw the conclusion that the model that best represents the true population estimates that temperature is a significant predictor of bike rental counts.

As a means of model comparison, the three models that were used to address the problem of assessing rental bike counts highlighted the better way to solve predictions on Seoul's bike rental trends. Initially, it was surprising to see that power transformations on the multiple linear regression (MLR) could not completely control for the residuals in the data. When comparing the next two models, the adoption of the ZINB model was considered a better fit than the Poisson model because of the large amount of zero's and the overdispersion in the data. The count data representing the bike counts were left heavy in their distribution. This represents a general trend in count data in that with zero being the bottom count number, the distribution of bike rentals will generally decrease and the count increases.

The models that were used were the best fit for the situation at hand when considering GLM. However, there were some limitations to the data that should be addressed as well as potential better methods to address measuring the impacts of weather on rental bike counts. First, the ZINB model was considered statistically significant but when comparing to the poisson distribution there should be consideration with the log likelihood estimates. With the zero's and one's removed in the fitted probabilities, there is some error expected when considering the significant p-values for the AIC an BIC model comparison. It should still be considered that the ZINB model is better than the poisson model because of the apparent overdispersion in the data which is accounted for. Also, machine learning techniques would probably provide better estimates of which variables are significant predictors in the data. Authors like Sathishkumar have used methods like Classification and regression trees to assess the trends in the data and would probably come with closer estimates to the population trends.

The conclusions that can be drawn from this exploration of the Seoul bike data suggests that a bike rental company or program should consider the temperature of a day in making decisions for their company. When there is needs for bike maintenance for example, extreme weather conditions may be a time to remove the bikes from public use to minimize the loss in use. A method like this could potentially minimize losses from removing the supply from public use. Also, if there is variation in temperature across the city, there could be movement of bikes to more appropriate regions where the company would expect more use of their product. It should be considered that while this paper measured that temperature is important in considering the consumption of rental companies product, there is not specification on which range of temperatures are provide the most probable use of the rental bikes. Ultimately, the increased need for convenient transportation in the city has created an opportunity for bike rentals to be in wide use. Maximizing the ability to meet the demand for bike use involves considering how weather conditions will impact the consumption of the product. Considering the three GLM that could potentially measure such an impact results in the conclusion that ZINB models are the best in the group at predicting the behavior of customers as they contribute to the bike rental industry. The ability for technology to make transportation more accessible in the future is demonstrated by Seoul's efforts to make bike access universal. With the customer base requiring outside access to use the bike share company's product, weather conditions are especially important in assessing the demand trends of the bike rental industry.

Works Cited

Shaheen, Susan A., et al. "Bikesharing in Europe, the Americas, and Asia Past, Present, and Future." *TRANSPORTATION RESEARCH RECORD*, no. 2143, 2010, pp. 159–167. *EBSCOhost*, doi:10.3141/2143-20.

V E, Sathishkumar, and Yongyun Cho. "A Rule-Based Model for Seoul Bike Sharing Demand Prediction Using Weather Data." *European Journal of Remote Sensing*, vol. 53, Jan. 2020, pp. 166–183. *EBSCOhost*, doi:10.1080/22797254.2020.1725789.