How the climate and Holidays affect Bike Rentals in Seoul

STAC67 Group 8 John Morales[*] Adil Karim[†] Naivil Patel[‡] Jalal Kassab[§]

06/04/2022

[*] Student Number 1000951703, Power-Point Slides, Forward Selection Process, Data Cleaning

[†] Student Number 1005399873, Model selection, validation, and diagnostics

[‡] Student Number 1005569369, Weighted Least Squares, Cooks Distance, some model diagnostics

[§] Student Number 1006097607, Case-Study Report, Plan Team Meetings, Team Leading, Presenting

Abstract

The motivation lies in predicting the amount of bikes sold on a given day in Seoul. In order to build an accurate and efficient model, model selection played an important part in the analysis. We considered the main effect model with an interaction between humidity and temperature as our model to predict the number of bike rentals from climate and calendar data. After transforming our variables and taking into account possible multi-collinearity within the data, we provide the reader with the most efficient linear model possible for this given data-set while avoiding over-fitting. We hope for this model to be used to help bike-share companies predict how many bikes to supply using climate data in other regions.

Background and Significance

Bike sharing holds a significant place in Seoul's city infrastructure, aiding in the reduction of cars and other forms of transportation that have a generally large individual carbon footprint.

An obvious trend of having higher sales during the warmer months of the year is to be expected, but this analysis can aid in the optimization of supply for the businesses so that they can have a relative idea for what they can expect for upcoming years based on this model. This in turn gives us our initial hypotheses that we'd like to test. We suspect that extreme temperatures in both directions should in theory negatively affect bike rentals, we also suppose that less bikes would be rented out on rainy days given that people do not want to get wet. Our hypotheses stem from looking at the scatterplot between bikes rented and explanatory variables in page 4 (part of the Explanatory Data Analysis section.)

Exploratory Data Analysis:

Our model was built using a data set containing 8760 observations that were collected over the span of a year. We also had access to 13 potential explanatory variables given that we already reserved for one variable to be the response.

As we see in the correlation plot in page 4, we expect a decrease in bike rentals when extreme temperatures are observed on both sides of the spectrum. We also expect a trend of more bike rentals to drop as more rainfall is present.

Date - This variable provides us with the date of rental of a given bike with the format being DD/MM/YY.

Rented bike count - this continuous variable is a measure for the number of bikes rented on a given period of a given day.

Hour- This variable provides us with the our of the day that the bike was rented.

Temperature - This is a continuous variable that is a measure for the temperature.

Humidity - A continuous variable representing the % of humidity in the air.

Wind speed - An independent variable providing us with the speed of the wind during that day.

Visibility - An independent variable representing the approximate maximum visible distance to the human eye on any given hour of any given day.

Dew point temperature - This describes the temperature the air needs to be cooled to (at constant pressure) in order to achieve a relative humidity (RH) of 100%.

Solar radiation - A variable that gives us a measure of the solar radiation on any given day/hour.

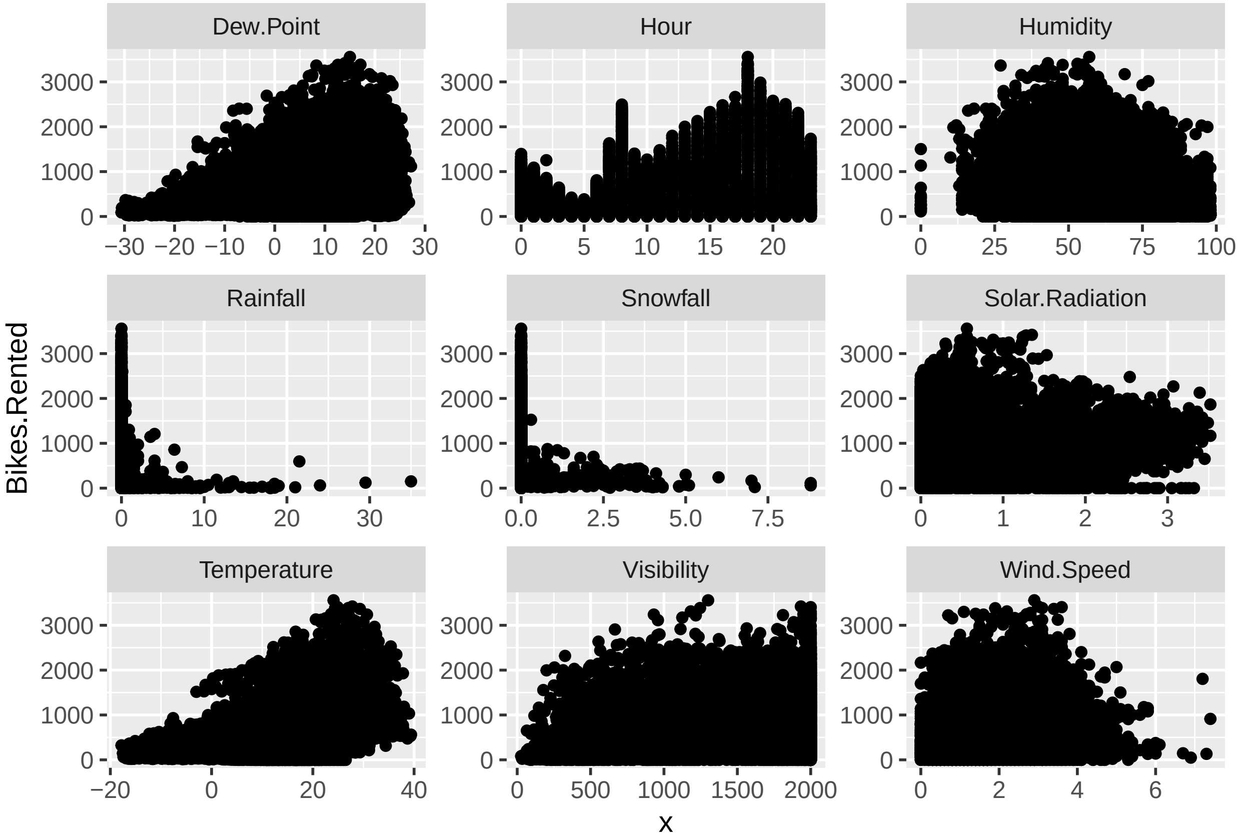
Rainfall - An independent variable that measures the total amount of rain.

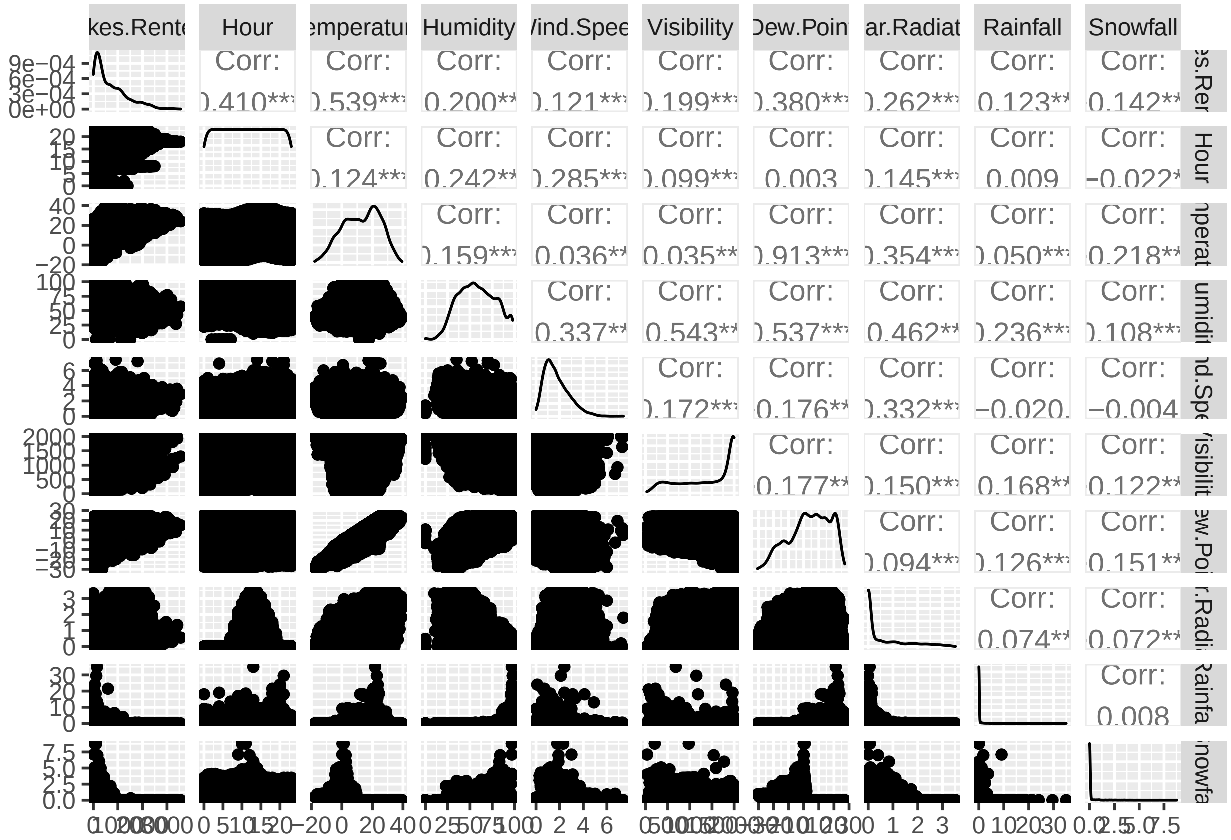
Snowfall - An independent variable that measures the total amount of snowfall.

Seasons - A variable that tells us the season of year on that given day.

Holiday - A Binary variable that tells us whether or not it is a holiday.

Functional day - A variable that tells us whether or not the bike rental shops are open on any given day/hour.





Kyoungok Kim found that high temperatures over 30 degrees celcius reduces the bicycle usage. We also found from Jangwoo Park that humidity causes high temperatures to feel more uncomfortable. From their research and the plot above showing no correlation between the two predictors, we decided to put the interaction between humidity and temperature into our model.

Data Preparation

Data Cleaning

We remove the Identifier column (date) and removed 0 bikes rented. Identifier column as it is unneeded in our model. Delete "0" from dataset because it skews data, making the prediction on number of bikes rented using the predictions less accurate.

Test-Train split

Split the modified data frame into a train and testing set for model validation. We chose to use 70% of the data to train the model to get better accuracy from the model. The rest of the data will then be used for testing. The model includes all the independent variables given except Functioning Day as most of the values were similar for that variable, meaning that it will not bring much to the model in regression. We add the interaction term between humidity and temperature as humidity is known to magnify the effects of extreme temperatures, which can have a prevalent effect on the number of bikes that are rented.

Model Building

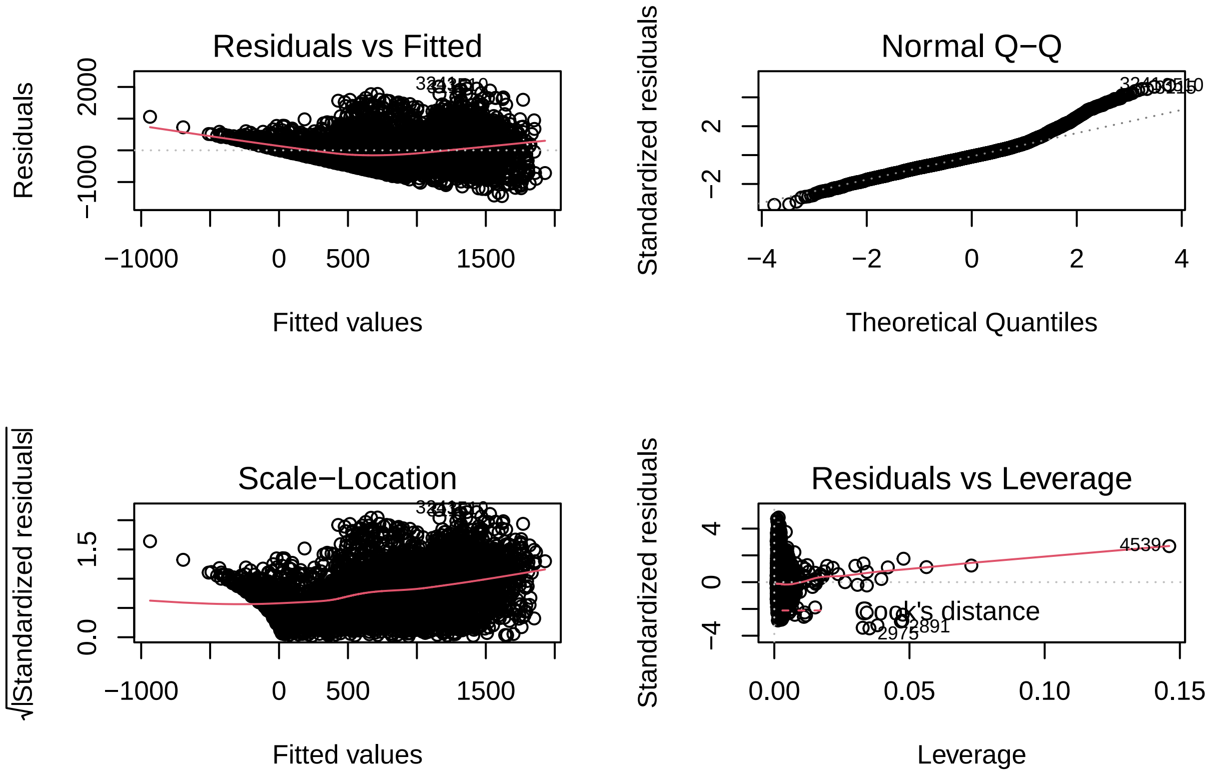
The process in building the model went as follows, initially we trained a main effects model with the interaction term between temperature and humidity. Then we used the stepAIC method to aid us in the reduction of the model by performing forward selection, This resulted in the elimination of the “Snowfall” predictor. Next, we found out that our response variable (The number of bikes rented) was in need of a necessary transformation due to the nature of the residual plots shown in the Residual Analysis subsection, to solve this problem we applied a box cox transformation on the variable and proceeded. Shortly afterward, a similar problem arose when the outlying observations led to a less accurate model, to fix that we used Cook’s distance to find the influential observations and get rid of them, more on this in the Residual Analysis subsection. Last but not least, heteroscedasticity was noticed so we simply applied the weighted least squares method to deal with that. Now we had a model void of any issues in regards to our regression assumptions and with a good R squared value.

```
## Step:  AIC=71780.03
## Bikes.Rented ~ Temperature + Hour + Humidity + Seasons + Rainfall +
##       Solar.Radiation + Dew.Point + Holiday + Wind.Speed + Visibility +
##       Temperature:Humidity
##
##              Df Sum of Sq        RSS   AIC
## <none>                    1072474063 71780
## + Snowfall    1  3468.5   1072470594 71782
```

Residual Analysis

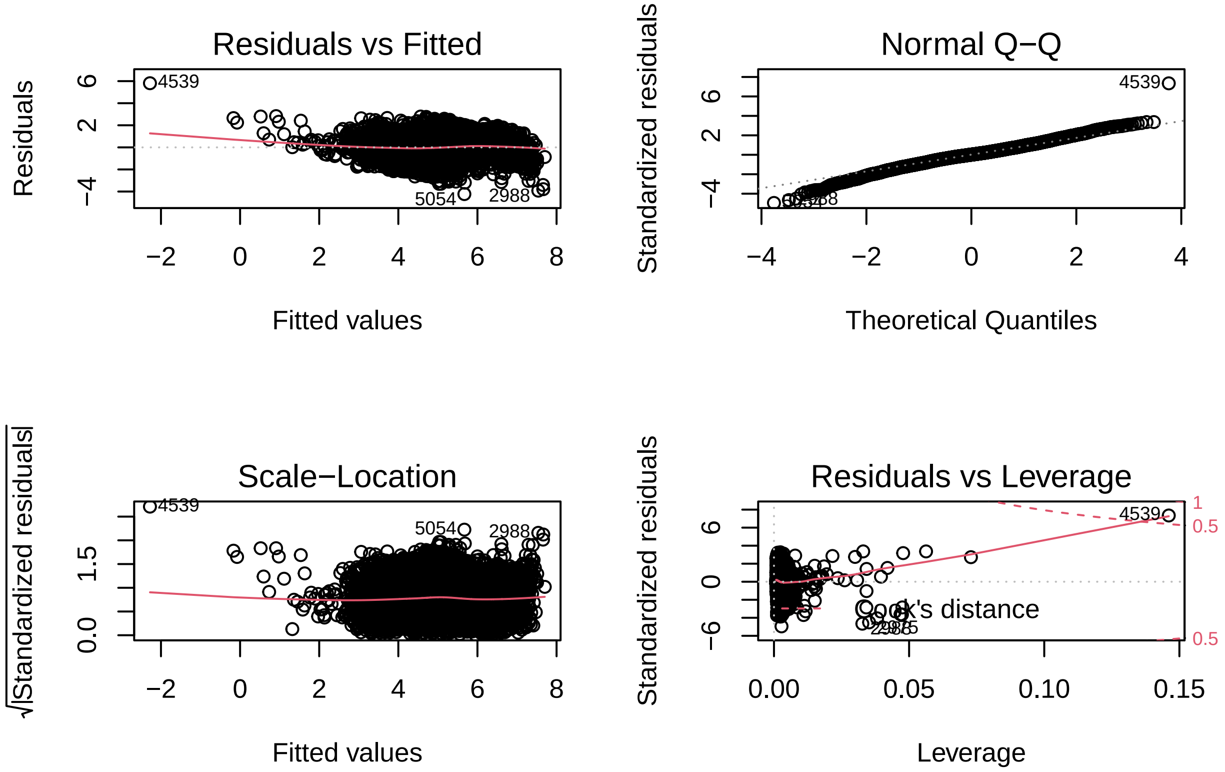
Testing Normality and Constant Variance Assumptions

The QQ plot graph shows that the majority of the residuals follow a normal distribution however it’s not perfect. Since the spread of the residuals is not equal at each level of the fitted values, we can claim that the constant variance assumption is violated — this can be displayed on the Residuals vs. Fitted values graph. In order to fix these violations, we consider a box cox transformation to fix the long tail causing a right skewed normal QQ plot.



Testing Normality and Constant Variance Assumptions after box cox

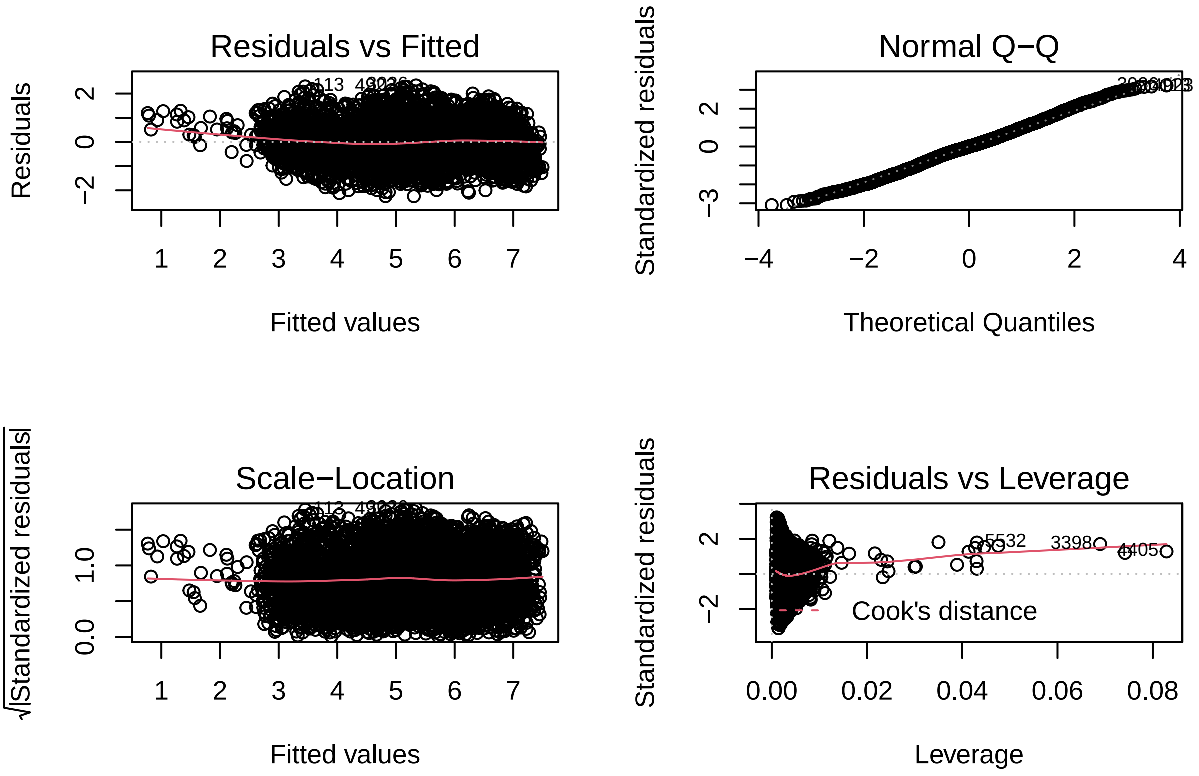
After a box cox transformation, the normal QQ plot appears to look more normally distributed with some noticeable outliers. Furthermore from the residual graph and excluding the visible outliers, the spread of the residuals seems more equal at each level compared to before the box cox transformation.



From the summary of the model, the box-cox adjusted model yields an R squared of 63.7% (or 0.6373). This means our model explains 63.7% of the variability of our dataset, adjusted for how many predictors we have. Thus the box-cox transformation improved our model's explanation of the data's variability.

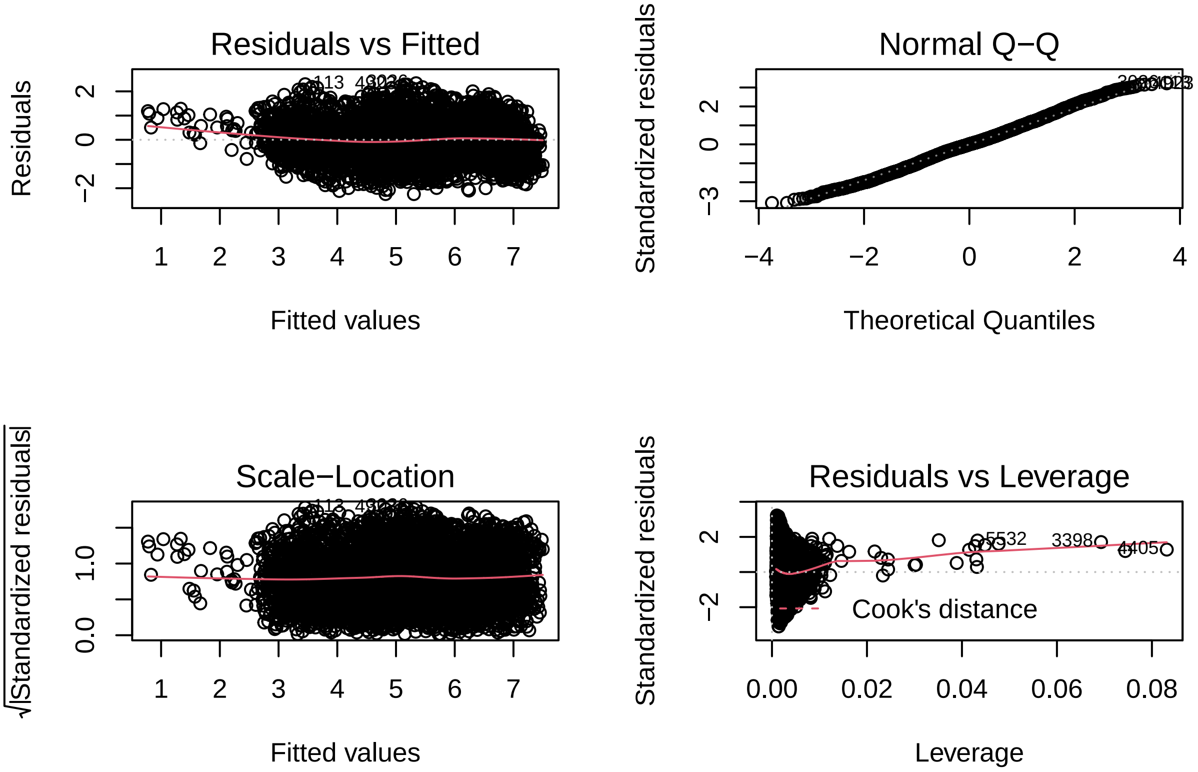
Testing Normality and Constant Variance Assumptions after Cook's Distance

With the removal of outlying observations, we get an adjusted Rˆ2 at 71% which is significant improvement over the previous model. The qq-plot is a nice straight line close to y=x, meaning the normality assumption is satisfied. This overall suggest the model is in good condition.



Testing Normality and Constant Variance Assumptions after Weighted Least Squares

WLS will add weights to the coefficients based on the variance of the individual prediction variables. This will include the variability in X to better fit the model. The qq-plot is a nice straight line close to y=x, meaning the normality assumption is satisfied. The residuals are a bit more scattered than the previous residual plot above.This overall suggest the model is in good condition. The Adjusted R^2 for the model is 0.72 which is better than the previous model, but its not as significant of an improvement as we saw when we removed outlying variables.



Model

For the validation of our model we brought back the "test" subset of our data and evaluated our model's performance on it. We found that the MSE (Mean squared error) and MSPR (Mean Squared Prediction Error) , 0.5 and 0.82 respectively, are extremely close. This is an indicator of a good model absent from overfitting and good at performing in most general cases. Seeing as we achieved our goal of putting forth a model that can aid bike share rental companies in seoul generally well, we get the idea that this is where we should stop.

Our final model to predict bikes rented

```
## # A tibble: 14 x 5
##    term                     estimate std.error statistic   p.value
##    <chr>                       <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)               7.29      0.248        29.4   2.45e-176
##  2 Temperature              -0.0259    0.00918      -2.82  4.83e-  3
##  3 Hour                      0.0653    0.00155      42.2   0
##  4 Humidity                 -0.0426    0.00268     -15.9   1.17e- 55
##  5 SeasonsSpring            -0.373     0.0294      -12.7   2.08e- 36
##  6 SeasonsSummer            -0.390     0.0373      -10.5   2.26e- 25
##  7 SeasonsWinter            -1.01      0.0406      -25.0   1.81e-130
##  8 Rainfall                 -0.339     0.0163      -20.7   3.82e- 92
##  9 Solar.Radiation          -0.0656    0.0159       -4.12  3.92e-  5
## 10 Dew.Point             0.109     0.00980      11.1   2.11e- 28
## 11 HolidayNo Holiday     0.399     0.0475        8.40  5.37e- 17
## 12 Wind.Speed            0.00455   0.0109        0.419 6.75e-  1
## 13 Visibility           -0.0000409 0.0000211    -1.94  5.30e-  2
## 14 Temperature:Humidity -0.000326  0.0000511    -6.38  1.91e- 10
```

Validating our model

```
anova(box_fit2)['Residuals', 'Mean Sq']
```

```
## [1] 0.5268118
```

```
preds <- as.vector(predict(box_fit2,Valid))
mspr = sum((Valid$Bikes.Rented^lambda - preds)^2)/(length(Valid$Bikes.Rented))
mspr
```

```
## [1] 0.8155791
```

Conclusion

Our goal was to make a model that predicts the number of bikes sold on a given day in Seoul so that the bike share companies can better manage their supply . To do this, we removed outlier cases because they tend to skew our data. We discovered that the cases where bike share services were not functioning were insignificant for predicting the number of bikes rented, due to the fact that there were very few non functioning days. In our findings, both temperature and humidity as well as rainfall had a negative impact on bikes rented which confirms our hypothesis. Moreover, when both humidity and temperature are at extremes we see a substantial impact on the amount of bikes rented, this was initially expected, by taking into account the interaction and presenting the model coefficient of the relationship, it led us to quantify our intuitions and solidify our position in regards to our hypotheses. A limitation of our model is that there are still some outlying variables in the training data set, which can be seen in the residual plots. These values are at the left side making a straight diagonal line angling downwards. Thus removing the randomness of the scattered residuals. This may be why our final model disagrees with Sathishkumar et. al's findings, where she deems that temperature positively affects the number of bikes rented. Getting rid of these values would create a model that is better at predicting the average cases seen in the data set.

Discussion

Several implications are realized after the finalization of our model, with the obvious yet most important one being that bike share rental companies along with most recreational commute-related companies are able to accurately model their sales on any given day based on a variety of predictable factors such as the weather or the visibility. In the broader sense, this opens up many doors to fellow statisticians and should serve as a reminder that, despite the unpredictability of this world, there always lies a trend underneath. A common belief is that humans are encoded in such a way where a variety of inputs are stored into the brain and a variety of neural networks take in that input and so an output or decision is generated, for example if I were to see the weather forecast and it predicted that we'll be seeing a snowstorm in an hour (input), I would decide to not ride a bike or go outdoors at all (output). This way of thinking can lead us to believe that most if not all things pertaining to human choice can be modeled to a certain extent.

References:

[1] Sathishkumar V E, Jangwoo Park, and Yongyun Cho. ‘Using data mining techniques for bike sharing demand prediction in metropolitan city.’ Computer Communications, Vol.153, pp.353-366, March, 2020
[2] Sathishkumar V E and Yongyun Cho. ‘A rule-based model for Seoul Bike sharing demand prediction using weather data’ European Journal of Remote Sensing, pp. 1-18, Feb, 2020
[3] Kim, Kyoungok. “Investigation on the Effects of Weather and Calendar Events on Bike-Sharing according to the Trip Patterns of Bike Rentals of Stations.” Journal of Transport Geography, vol. 66, no. 66, Jan. 2018, pp. 309–320, 14 May 2021.