W271 Group Lab

Analysis of Bike Share Demand in Korea

Violet Davis, Yuri Kinakin, W. Sean McFetridge and Emanuel Mejia

1 Introduction

The problem motivation is well stated in the Additional Information section of the dataset's webpage:

> *Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.*

We are being asked to generate the most accurate and generalizable model for the total number of bikes rented each hour (i.e. the *Rented Bike Count* variable).

Common wisdom would suggest that bikes are less likely to be rented when it is nighttime, precipitation is high, it's an off season (i.e. winter), it's cold outside, visibility is low or it is windy. Therefore our hypothesis is that rented bikes can be estimated using a function of these variables, and the coefficient is statistically different than zero for at least one.

Particularly, our null and alternative hypothesis can be defined as:

- $H_0 : \beta_{temperature} = \beta_{visibility} = \beta_{windspeed} = \beta_{precipitation} = \beta_{winter} = \beta_{time} = 0$
- $H_a : \beta_{temperature}$ *or* $\beta_{visibility}$ *or* $\beta_{windspeed}$ *or* $\beta_{precipitation}$ *or* β_{winter} *or* $\beta_{time} \neq 0$

In this report we will test that hypothesis, and build upon the original model to develop more accurate and refined solutions.

2 Data

2.1 Description

The data was taken from UC Irvine at Seoul Bike Sharing Demand The dataset under investigation consists of bike rental information collected in Seoul, South Korea in 2020. It consists of 8,760 observations of 14 features, with each row representing each hour over the course of a year.

Details of the features are:

Column Name	Type	Description/Comment
Date	Time	Day/Month/Year
Rental Count	Numeric	Number of bikes rented per hour.
Hour	Categorical	Hour of the day (from 0 to 23)
Temperature	Numeric	Temperature in °C
Humidity	Numeric	Relative humidity (percentage from 0 to 100)
Windspeed	Numeric	Wind speed (m/s)
Visibility	Numeric	Maximum visibility distance (10m increments)
Dew point	Numeric	Dew point in degrees Celsius
Solar radiation	Numeric	Solar radiation value (MJ/mˆ2)
Rainfall	Numeric	Measured rainfall during the hour (mm)
Snowfall	Numeric	Measured snowfall during the hour (cm)
Seasons	Categorical	Current season (Winter, Spring, Summer, Autumn)
Holiday	Categorical	Whether the day is a holiday or not
Functioning Day	Categorical	Whether or not bike rentals were available.

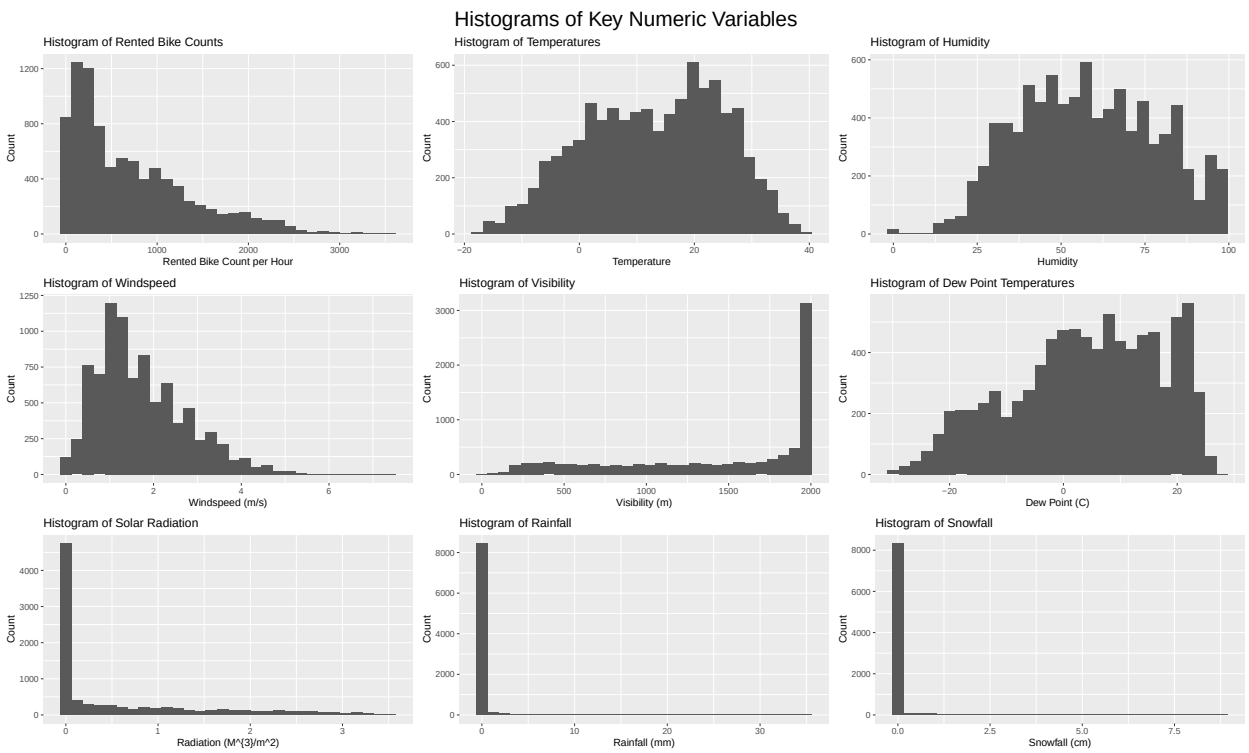
Limited information is provided about the data collection methodology. Particularly, the frequency at which the weather variables were collected is not mentioned, so it's unclear as to whether these are averages per hour or point measurements.

Reviewing the data, it's likely that many of these variables, particularly weather related variables (e.g. *Temperature, Humidity, Windspeed, Rainfall, etc.*) are correlated. In fact, assuming that the visibility will be reduced predominantly by fog/rain, the two variables of *Temperature* and *Dew Point Temperature* should be very good predictors for *Visibility*. For obvious physical reasons, we expect *Solar Radiation* is also very strongly correlated with time of day.

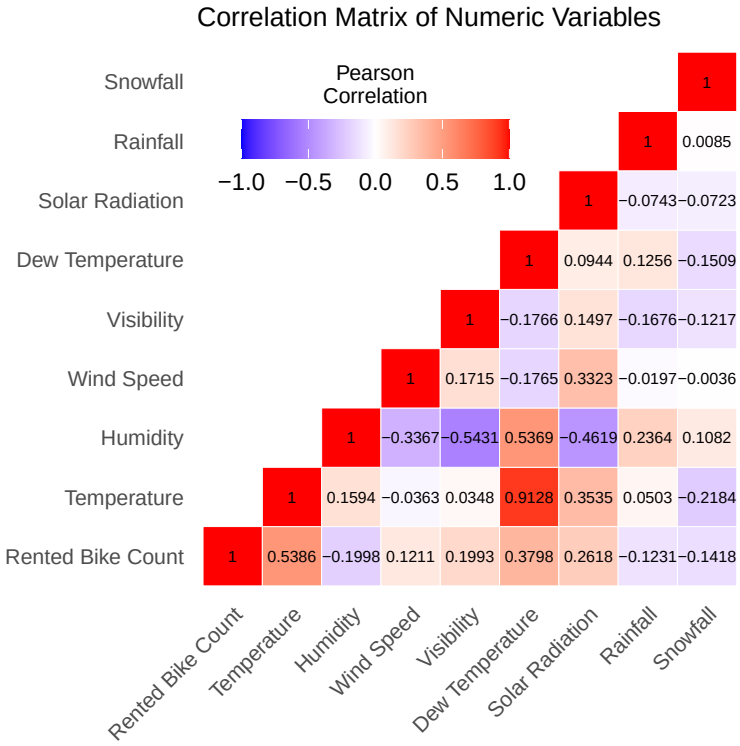
2.2 EDA

After loading in the data and fixing issues with non-unicode characters in the column names, we examined the data for any missing values. We then mutated the columns based on their format, in particular converting the *date* column to a proper date-time format in R and the *Seasons, Holiday and Functioning Day* columns to factor variables.

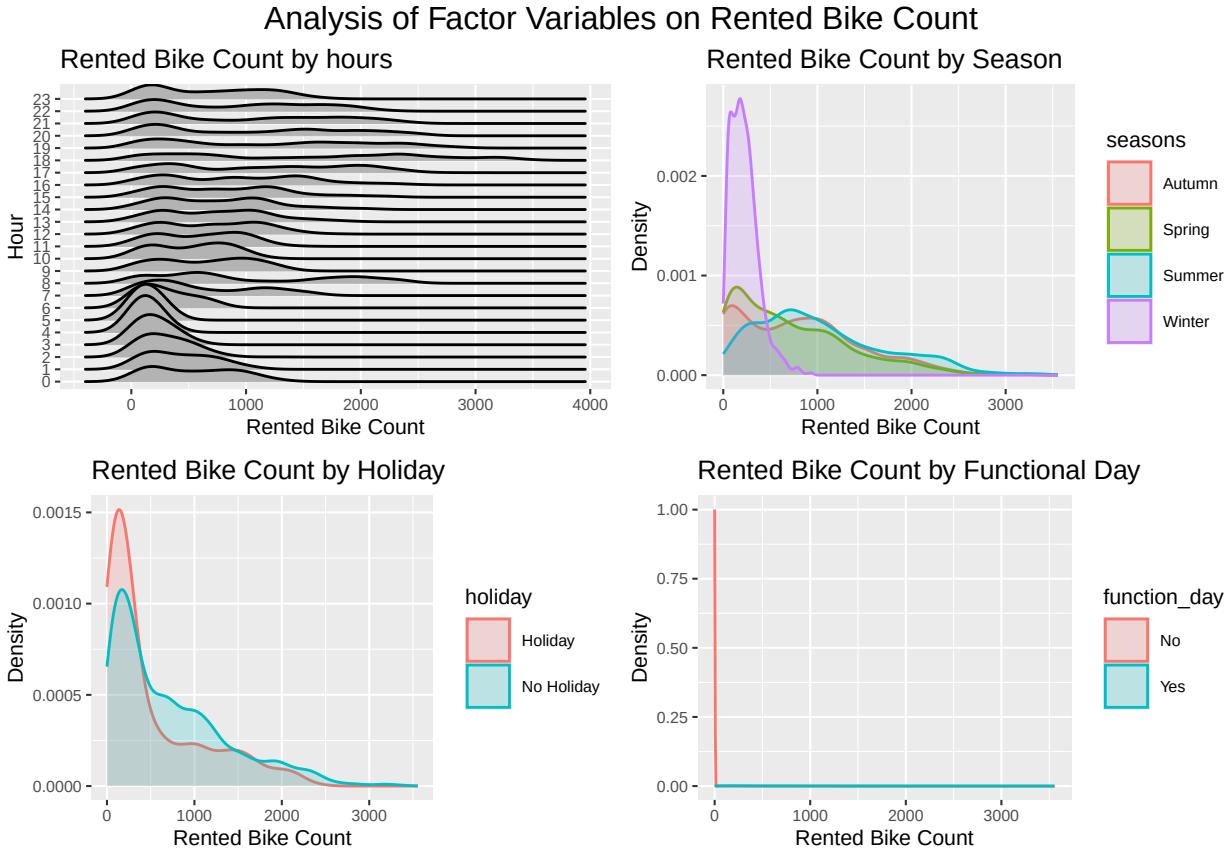
We first analyzed the histograms for the assumed key numeric variables to understand the distribution and frequency of the variables. These histograms are displayed below. From the histograms, we can see that rented bike counts are skewed to the right with many hours having very few bikes rented, whereas other variables, such as *Temperature*, *Dew Point Temperature* and *Wind Speed* are more normally distributed. We also noticed that on most days there is no rainfall or snowfall (i.e. there is a large "spike" at zero). We therefore combined rainfall and snowfall into a single binary variable for precipitation that is True when either rain or snowfall was recorded, and False when not.



To perform a more comprehensive check of all the numeric variables, a correlation matrix was created which is displayed below:



From this, we can see that the *Dew Point Temperature* and *Temperaure* have a high correlation. To avoid correlation between variables our analysis will focus on *Temperature* alone.



Lastly, we analyzed categorical variables which are summarized in the graphs above. The graphs suggest several other potential simplifications. First, *Hours* can be grouped into three sets of times, with "Morning" between the hours of 00:00 to 06:00, "Daytime" from 06:00 - 18:00, and "Evening" from 18:00 to 00:00. Second, the main differences in *Seasons* are for the "Winter" category, so we can reduce this categorical variable to a binary choice between "Winter" and "Not Winter." Third, as there are no bikes rented during non-functional hours, so we can use this to further filter our dataset as with NAs.

We will also create a *Weekday* factor variable that provides a name to each day. This will allow us to evaluate whether or not there is a difference in rental counts between days (e.g. Saturday/Sunday versus the rest of the week).

3 Model Development

3.1 Poisson regression

Our hypothesis for a base model is that the number of bike rentals depends on the following association:

$$Rented\ Bike\ Count = Temperature + Visibility + Windspeed + Precipitation + Winter + Time\ of\ Day$$

Where our null and alternative hypothesis are as follows:

- $H_0 : \beta_{temperature} = \beta_{visibility} = \beta_{windspeed} = \beta_{precipitation} = \beta_{winter} = \beta_{time} = 0$
- $H_a : \beta_{temperature}$ *or* $\beta_{visibility}$ *or* $\beta_{windspeed}$ *or* $\beta_{precipitation}$ *or* β_{winter} *or* $\beta_{time} \neq 0$

An overview of the variables and their significance is in Table 2.

Table 2: Base model summary

	Dependent variable:
	Rented Bike Count
Temperature	0.022^{***} (0.00005)
Visibility	0.0001^{***} (0.00000)
Wind Speed	0.016^{***} (0.0004)
Precipitation	-1.133^{***} (0.003)
Winter	-0.840^{***} (0.002)
Evening	0.344^{***} (0.001)
Morning	-0.836^{***} (0.001)
Constant	6.270^{***} (0.002)
Observations	8,465
Akaike Inf. Crit.	1,521,191.000
Note:	*p<0.1; **p<0.05; ***p<0.01

All of the variables are highly significant in this initial model.

Commentary on the numeric variables is as follows:

- *Temperature*: Increases in temperature are positively correlated with rented bikes per hour.
- *Visibility*: Increases in visibility are positively correlated with rented bikes per hour. However, the visibility variable appears to have little practical significance as small changes in visibility are unlikely to be noticed.
- *Windspeed*: Contrary to our natural understanding, increases in wind speed are positively correlated with rented bikes per hour.

Commentary on the factor variables are:

- *Precipitation*: Precipitation decreases the number of rented bikes per hour, relative to no precipitation.
- *Winter*: Winter decreases the number of rented bikes per hour, relative to non-winter.
- *Time of Day*: Evening increases the number of rented bikes per hour, whereas morning decreases the number of bikes per hour, relative to daytime.

Given the significance for each variable is less than 0.05, we reject the null hypothesis that none of the variables are significantly different than zero.

3.2 Model Comparison

To potentially improve the model accuracy we added the remaining explanatory variables, specifically, *Humidity, Dew Point Temperature, Solar Radiation, Holiday*, the *Weekday* variable that was

created after the EDA step, and utilizing the raw *Hour* variables rather than time of day. This last point significantly increases the number of variables in Model 2.

In Model 3, we then added:

- an interaction term to test whether or not a holiday falls a weekend, as the expectation is more people will rent a bike on weekend holidays;
- an interaction term between weekdays and precipitation as people are more likely to continue biking to work in the rain to work, but are less likely ride a bike when there is an option to stay home;
- an interaction term between temperature and weekend, as on warm weekends we expect more people will rent a bike to spend time outside.
- a quadratic term to the wind speed, as it may be more consequential when the winds speed is particularly high
- a quadratic terms for temperature, as small changes at the temperature extremes are likely to be more important than those near 0

All three models are summarized in Table 3. While certain variables are not shown (due to space), all variables have a high statistical significance, suggesting that Model 3 is likely to perform the best among the three models. This is validated at the bottom of the stargazer, we can see that Model 3 has the lowest AIC, AICc, and BIC. This is surprising as Model 3 is also the least parsimonious.

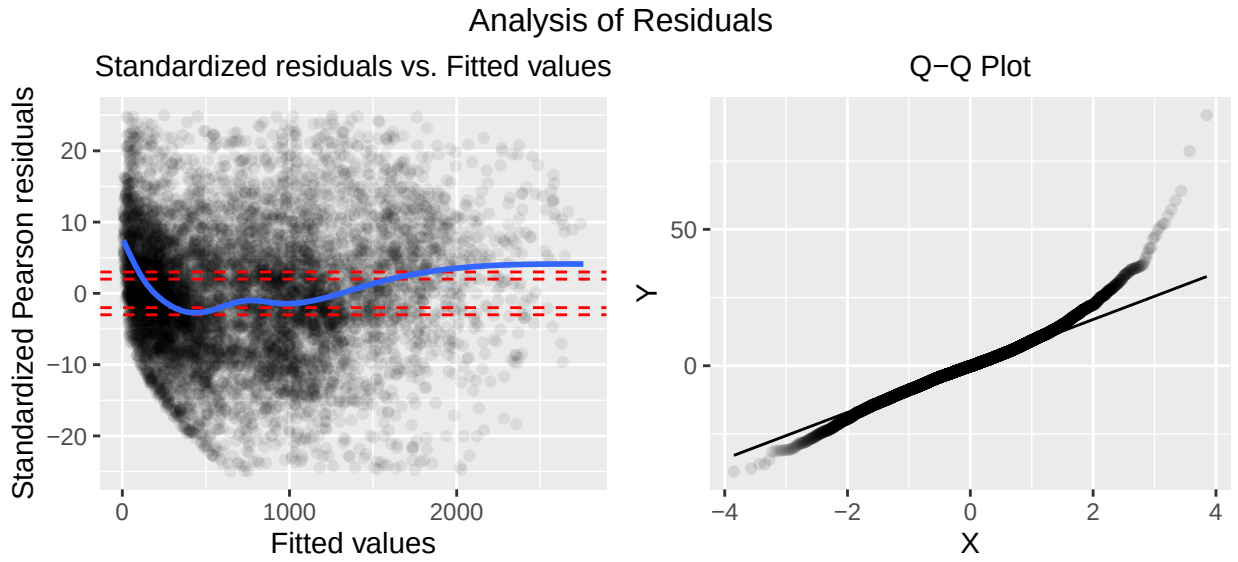
Table 3: Model comparisons

	Dependent variable:		
	Rented Bike Count		
	(1)	(2)	(3)
Temperature	0.022*** (0.00005)	0.003*** (0.0003)	0.048*** (0.0003)
Visibility	0.0001*** (0.00000)	0.00003*** (0.00000)	0.0001*** (0.00000)
Wind Speed	0.016*** (0.0004)	−0.034*** (0.0005)	0.001 (0.001)
Precipitation	−1.133*** (0.003)	−0.991*** (0.003)	−0.776*** (0.006)
Winter	−0.840*** (0.002)	−0.882*** (0.002)	−0.406*** (0.002)
Humidity		−0.011*** (0.0001)	−0.014*** (0.0001)
Dew Point Temperature		0.021*** (0.0003)	0.029*** (0.0003)
Solar Radiation		0.019*** (0.001)	0.055*** (0.001)
No Holiday		0.152*** (0.002)	0.772*** (0.012)
Temperature Squared			−0.001*** (0.00000)
Windspeed Squared			−0.006*** (0.0003)
Constant	6.270*** (0.002)	6.954*** (0.009)	6.045*** (0.014)
Time of Day	Yes	No	No
Hour	No	Yes	Yes
Weekdays	No	Yes	Yes
Interactions	No	No	Yes
Akaike Information Criterion	1521191	1057881	927718
Corrected Akaike Information Criterion	1521191	1057881	927719
Bayesian Information Criterion	1521248	1058155	928092
Observations	8,465	8,465	8,465
Note:			*p<0.1; **p<0.05; ***p<0.01

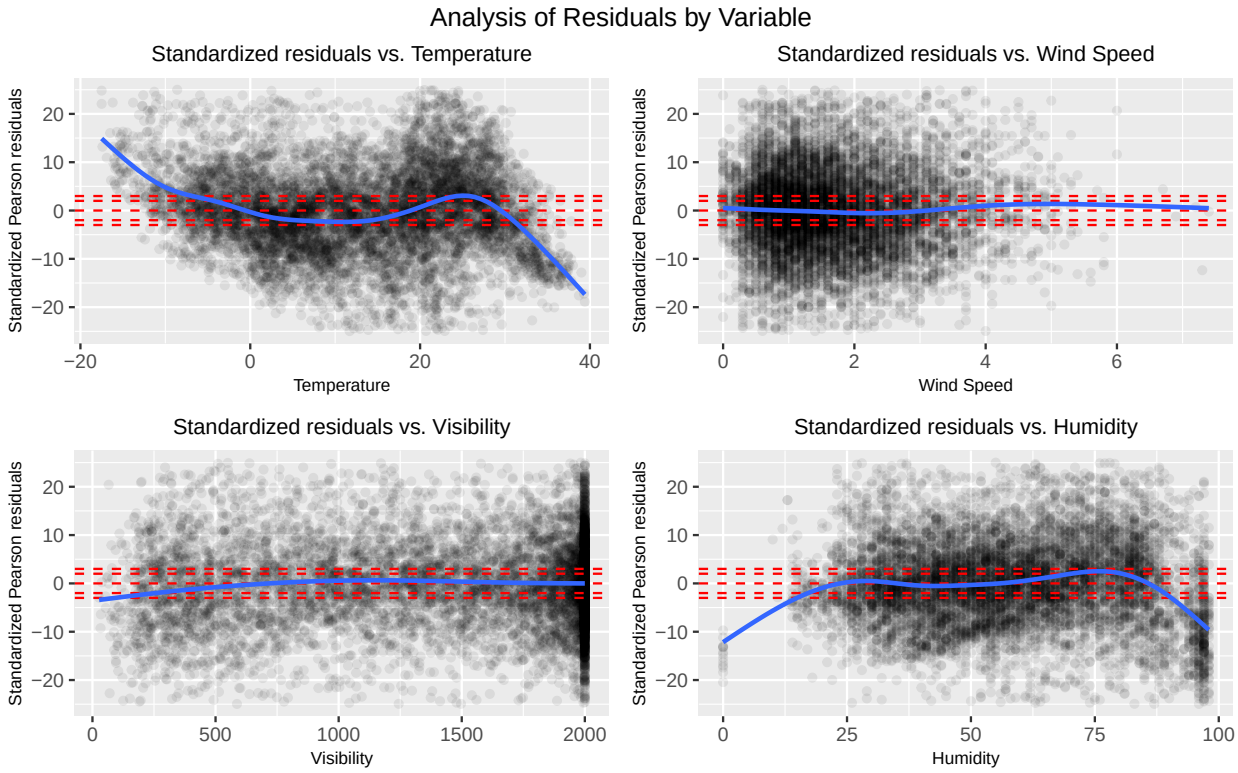
3.3 Model Assessment

Before accepting Model 3 as the best model, we spent some additional time analyzing the residuals to ensure the model model is a good approximation of the data. The first graph below shows the

fitted variables against the standardized pearson residuals.

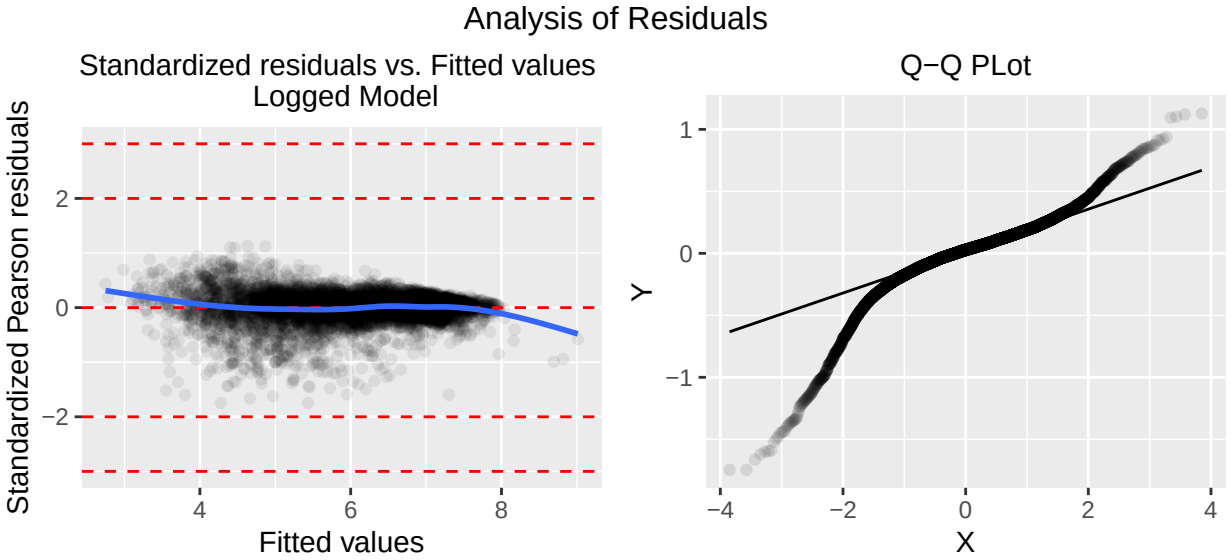


The blue line is a function of the residuals, and the red lines identify a standard deviation of $\pm$ 2 and $\pm$ 3. From the graph we can see that while the function of the residuals does track fairly close to zero, there is a fair bit of over dispersion, as many fitted values are more than three standard deviations away. To investigate the possible cause of this we analyzed a few different of the variables residuals directly.



Reviewing the matrix above, we can see there is overdispersion among all of the variables. This imples that the output variable should be transformed. As an alternative model, we've taken the

log of *Rented Bike Count* to help correct for the over dispersion from the original model. The revised graph is below:



Visually looking at the above model, we can see that by logging rented bikes per hour the standardized Pearson residuals are all within two standard deviations.

Table 4: Overdispersion test: `model_poisson_3_adj`

Test statistic	P value	Alternative hypothesis	alpha
-246.4	1	greater	-5.272

The dispersion test for the adjusted model shows a p value of 1, which is significantly larger than 0.05. Which means we fail to reject the null hypothesis, indicating we do not have over dispersion.

Table 5: Overdispersion test: `model_poisson_3`

Test statistic	P value	Alternative hypothesis	alpha
16.7	6.466e-63 * * *	greater	4523

If we ran the same test on our non-adjusted model, the p-value is less than 0.05 which indicates we have over dispersion. This means we were able to successfully correct for over dispersion by utilizing the log of *Rented Bike Count* as the response variable.

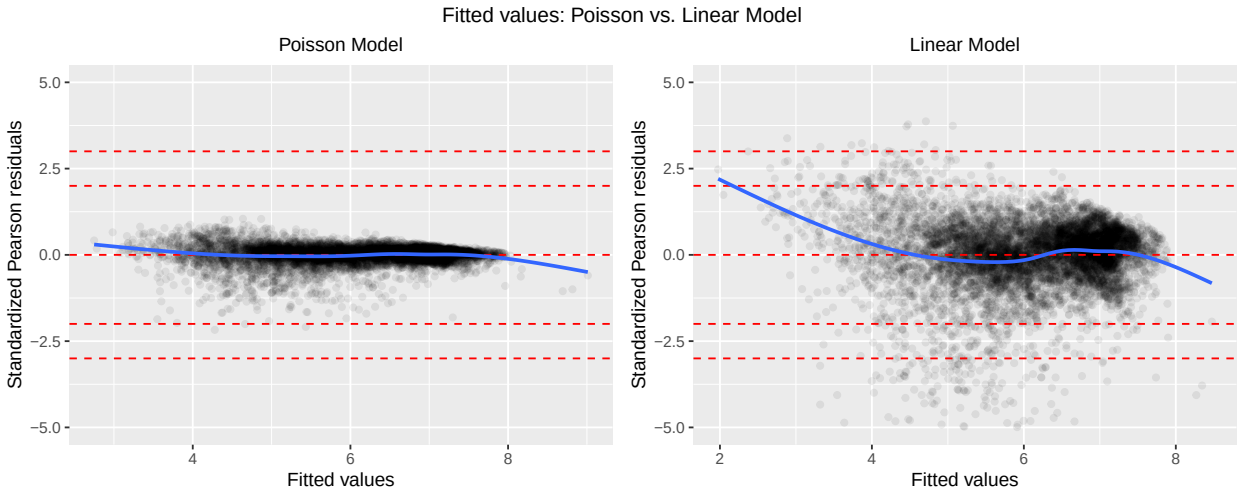
3.4 Alternative Specification

As an alternative specification, we prepared an OLS model, utilizing the same formula as that for adjusted Model 3.

Interpreting a few of the coefficients, we can see that for the OLS increasing the temperature by 1 unit decreases the log of rented bikes by 0.011, and when it is winter the log of rented bikes decreases by 0.422 (relative to non-winter seasons).

Table 6: Model Comparison: Logged Poisson and Linear Models

	Dependent variable:	
	Log of Rented Bike Count	
	Poisson	*OLS*
	(1)	(2)
Temperature	−0.001 (0.004)	−0.012** (0.005)
Visibility	0.00002* (0.00001)	0.0001*** (0.00001)
Wind Speed	−0.021 (0.015)	−0.101*** (0.022)
Precipitation	−0.127*** (0.039)	−0.636*** (0.053)
Winter	−0.071*** (0.019)	−0.426*** (0.028)
Humidity	−0.005*** (0.001)	−0.029*** (0.002)
Dew Point Temperature	0.012*** (0.004)	0.069*** (0.006)
Solar Radiation	0.014 (0.011)	0.123*** (0.017)
No Holiday	0.165** (0.068)	0.885*** (0.091)
Temperature Squared	−0.0002*** (0.00004)	−0.001*** (0.0001)
Windspeed Squared	0.003 (0.003)	0.015*** (0.005)
Constant	1.973*** (0.112)	7.242*** (0.167)
Hour	Yes	Yes
Weekdays	Yes	Yes
Interactions	Yes	Yes
Observations	8,465	8,465
Note:		*p<0.1; **p<0.05; ***p<0.01



Reviewing the plotted fitted values, we can see that the Poisson model performs better. The plotted residuals are flatter, and there is visually less over dispersion. For example, in the linear model, several observations fall outside two standard deviations.

4 Conclusion

The purpose of this report was to determine whether an accurate and generalizable model for total number of bikes per hour could be created. Our baseline was a Poisson model with a log link. Our natural understanding was that various weather-related variables and time would have an impact on our model, and therefore our null and alternative hypothesis were as follows:

- $H_0 : \beta_{temperature} = \beta_{visibility} = \beta_{windspeed} = \beta_{precipitation} = \beta_{winter} = \beta_{time} = 0$

- $H_a : \beta_{temperature}$ *or* $\beta_{visibility}$ *or* $\beta_{windspeed}$ *or* $\beta_{precipitation}$ *or* β_{winter} *or* $\beta_{time} \neq 0$

Upon running a review of the model coefficients, we were able to confirm that at least one of the variables were significantly different than zero and we therefore rejected the null hypothesis.

We also ran tests in an attempt to improve upon our original model, so we ran additional cases:

- Model 1: Poisson model with hypothesis parameters
- Model 2: Poisson model with all parameters
- Model 3: Poisson model with all parameters plus various interaction and quadratic terms

Comparing all of the models, we found that Model 3 performed the best as it had the lowest score for each of the three tested information criteria. However, a limitation of Model 3 was that, as originally developed, it was over-dispersed. We were able to correct for this by applying a logarithmic transfer to the response variable (*Rental Bike Count*).

Relative to our original model (Model 1), the adjusted Model 3 had significantly more variables and degrees of freedom. While the fit of this model was superior to the other three that were developed, a departure from normality in the residuals nevertheless remains. This implies that there may still be another omitted variable or missing interaction from the variables we reviewed, and predicting the number of rented bikes per hour therefore requires a more sophisticated model.

The key limitation of the model is its relatively limited extensibility. As the expected use case for this model is for a bike rental company to forecast the number of riders for a particular hour, the complexity and sensitivity of the selected model to variables like temperature and precipitation will make long term forecasting difficult. Also, it's difficult to understand how well the model might apply in a different market. So while we were able to develop a model that performs quite well in Seoul, there are a number of issues that have the potential to cause practical issues to implementation.