University Of British Columbia

DATA 410: Regression and Generalized Linear Models

Project Report

Andrew Liawan

Ethan Nathanael

Kyle Nelwan

Dataset Introduction and Hypothesis

The dataset we are using examines how various environmental and seasonal settings affect the number of bike-sharing rental users at an hourly rate. The dataset is a combination of existing bike-sharing rental historical logs and aggregates the hourly data with corresponding weather and seasonal information. The bike-sharing dataset can be split up into three sets of data; one for casual users, one for registered users, and one for the total of both casual and registered users. For our project, we would create two unique models utilizing the casual users' set of data and registered users' set of data.

The purpose of this report is to examine how much these different independent variables contribute to the hourly amount of bike-sharing rental users. Also, we are interested if users pay attention more to the true temperature of the hour or the feeling temperature of the hour. We hypothesize that there are going to be several of the same variables that make up the causality of both amounts of casual and registered users. Furthermore, the true temperature will have a higher causality effect on the dependent variables than the feeling temperature. Throughout this report, we will be analyzing this data to see whether or not our hypotheses are correct.

Description of the Dataset

Variable Name	Variables Description	Type of Variables
Season	1: Spring 2: Summer 3: Fall 4: Winter	Categorical
Year	0: 2011 1: 2012	Categorical (binary)
Month	Numerical representation of months from 0	Discrete
Hour	Numerical representation of hour from 0	Discrete
Holiday	0: Not a national holiday 1: National Holiday	Categorical (binary)
Weekday	Numerical representation of the days in a week from 0	Discrete

Working days	0: Weekend or holidays 1: otherwise	Categorical
Weather situation	1: Clear or partly cloudy 2: Misty weather 3: light snow or rain, thunderstorm and scattered clouds 4: heavy snow or rain	Categorical
True temperature	Normalized temperature in Celsius.	Continuous (The values are divided by 41)
Feeling temperature	Normalized feeling temperature in Celsius	Continuous (The values are divided by 50)
Humidity	Normalized humidity	Continuous (The values are divided by100)
Windspeed	Normalized wind speed	Continuous (The values are divided by 67)
Casual	Count of Casual Users	Discrete
Registered	Count of Registered Users	Discrete

The two datasets examines the number of casuals and registered users at any hour and the hourly environmental situation from Washington D.C, United States of America over two years.

As seen from the chart above, there are twelve independent variables and two dependent variables. The variables are a mixture of categorical, discrete, and scaled numerical variables. The scaled numerical variables are treated as continuous variables. Overall, the dataset has 17,379 observations.

temp vs *atemp*

As previously mentioned, we would like to see the effect between hourly real temperature forecast and hourly feeling temperature forecast on the number of bike-sharing rental users. Therefore, we would create two different models for each dependent variable with a model using *temp* and the other using *atemp.* Hence, we would have a total of four regression models. Note, that we initially only consider a linear model for all four regressions. However, depending on the initial assessment of the models, we could create a polynomial model of order two as a potential model.

Variable Transformation and Selection

Initial Assessment

As a starting point for our analysis, we created two models with the numbers of casual users or registered users regressed on all other variables.

Using the summary function for the casual users model, we found that most of the independent variables are significant to the model. However, we notice that our adjusted R-squared is less than 0.5. Therefore, even though our model is significant linearly, it can only explain half of the available data. On further inspection, our model should not be a linear relationship. The residuals vs fitted diagnostic plot have a clear pattern. Our data set seems to have a quadratic relationship. The Normal Q-Q plot shows that our model is highly right-skewed. This is a huge problem for us. Notice that those rightly skewed data have a very large value of standardized residuals. Hence, this original linear model is flawed. The residuals vs leverage plot tell us that none of our data are influential points but have high residuals as stated before. Therefore, we realized that the casual users model should be non-linear, more specifically a quadratic model. Please see **Exhibit A.1** for the summary and diagnostic plots.

Again, we do the same analysis step for the registered users model. The summary of this model shows us that most of our variables are significant enough (5% significance level). However, this model has a low adjusted R-squared value at 0.335. Hence, this model is not great at explaining the data we have. On further inspection, our data may have a quadratic or a decaying relationship. This is shown in the residuals vs fitted plot. The Normal Q-Q plot shows that our data set has a light left tail and a heavy right tail. Hence, we can conclude that our data set is rightly skewed. Note that a lot of our data are beyond the 2 standardized residuals range. Similar to the previous model, this model does not have any influence points but has high residuals value. Therefore, we have a problem with our current linear model for the registered users model. It is more likely that the model is of quadratic or other non-linear relationship. Please see **Exhibit A.2** for the summary and diagnostic plots.

Polynomial and Box-Cox Transformations

Upon initial assessments, we decided to create polynomial models of order two for both of our existing models since the residuals vs fitted plots for both models have a quadratic pattern and a heavily right skewed sample residual distribution. This is done by squaring all the explanatory variables, and adding them to the model with all the linear variables. However, this did not solve the problem. We found out that the residuals vs fitted plot still have the same pattern and a still heavily skewed QQ-plot. Please see **Exhibit B.1** and **Exhibit B.2** for the summaries and diagnostic plots for the polynomial model.

Hence, we would use the box-cox transformation to try and fix the skewness. To maximize comparisons and ensure the best possible model is obtained, the transformation is applied to both linear and polynomial models of both *temp* and *atemp* models for casual and registered models, totaling up to eight different models and transformations.

After applying transformations to all eight models, the lambda values of each model are different, ranging between 0.101 - 0.263. Even though the values are reasonably close to 0 upon initial look, the zero value is not within 95% CI of the lambda values, assessing from the log-likelihood vs. lambda plots. Therefore, the final transformation applied on the response variable is the Box-Cox formula involving their respective lambdas for each model, instead of the simplified log(y). Details of Box-Cox transformation of each model can be found in the Rmd file.

As we see in **Exhibits B.1 - B.6**, the residual vs. fitted values plot of the transformed models show a less curved distribution (in varying degrees) of residual points, when plotted against their fitted values, than the initial non-transformed models. Also, the QQ-Plots show how the sample residuals are now a lot closer to their theoretical quantiles after the Box-Cox transformations.

Checking for Multicollinearity

To check for multicollinearity, we use the VIF function in R. This assesses the VIF (Variance Inflation Factor) of each independent variable in a model, which explains how much the variance of a variable is inflated due to the presence of multicollinearity. Large VIF values imply that the variables are likely to be collinear with other variables.

In our assessment (**Exhibit C.1**), the atemp and temp variables have the largest VIF values, with around 43. We expected these to be highly correlated due to their similar nature of explaining hourly temperatures, and is part of our research questions. It is supported by a quantitative evidence as well, as the VIF values of either variables shrink significantly into nearly a value of 1, when the other variable is removed (see **Exhibits C.2, C.3**)

Also, *season* and *month* have relatively high VIF values compared to the rest of the variables (around 3). This is logical since each month is identified by a season. This is also supported quantitatively, since the removal of either *season* or *month* from the model seems to reduce the VIF values of the other unremoved variable into almost 1 (see **Exhibits C.4, C.5**). For simplicity reasons, instead of creating another research question from this, which will further increase the amount of possible best model, we decided to remove *season* from the model, as the *month* variable gives us more insight into the data as there are more available inputs.

Variables Transformation and Variables Selection.

In our initial assessment, we found that the polynomial did not help in eliminating or reducing the problem in our diagnostic plots. Therefore, we decided to use Box-Cox transformation to change the format of our dependent variable in hopes of lowering the issues we have. Note that since Box-Cox transformation can only be used when the dependent variables are not equal to zero. Since some of our data is equal to 0, we add a value of one to each dependent variable data and do the Box-Cox transformation for all of our existing models.

As the Box-Cox transformation only changes the format of the output variables, we have not considered which set of input variables are significant for the models. Therefore, we use a shrinkage model selection method, specifically LASSO regression to find the best subset in finding a causality within our models. After running the algorithm, we found out that all of the regressors are significant for all our models. However, after creating the model and using the summary function, we found out that some variables are not contributing as much as we thought using LASSO regression. Hence, we tried to remove those regressors like a backward selection algorithm, but since it shows the same results, with all the regressors being significant, none of the explanatory variables are removed.

Model Selection

To have a fair comparison between all the registered models, the models should have the same type of regression. Therefore, if the registered *atemp* model is polynomial, we should compare it with a polynomial model of registered *temp*. The same treatment is given to casual models.

In selecting the casual model, we unanimously agree that it is best to use the linear causal model than the quadratic model. The linear model is more simple and is easier to understand. However, the linear model has a higher mean squared error and a lower adjusted R-squared. That being said, the linear model has a better diagnostic plot when compared to the quadratic model.

The casual linear model shows a more linear trend of residual points distribution, averaging around the zero mark, when plotted against their respective fitted values. On the other hand, the casual quadratic model seems to show a more curved and quadratic trend of residual point distribution when plotted against their fitted values. Hence, a better comparison can be made when using the linear model. For visual understanding, please see **Exhibit D.1** and **Exhibit D.5** for the summaries and diagnostic plots for the model.

In selecting the registered model, we used the linear model too since it has a better and more reasonable residual vs fitted plot than the quadratic model. Although the quadratic model has an adjusted R-squared value than the linear model, the diagnostic plots still show that the linear model is better.

The registered linear model shows a slight curving or quadratic trend on the distribution of residual points when plotted against the fitted values. However, this is comparatively better than the registered quadratic model, as it shows a much more curved distribution of residual points when plotted against their fitted values. For visual understanding, please see **Exhibit D.3** and **Exhibit D.6** for the summaries and diagnostic plots for the model.

Casual Model Comparison and Analysis

The temperature regressor is significant for each model. However, the *atemp* model has a marginally higher F-statistic value for its temperature regressor when compared to the *temp* model, implying that the *atemp* model is less likely to have zero values for all regression parameters (β), showing that the addition of predictors in the *atemp* model is slightly more significant than the *temp* model. Therefore, we conclude that *atemp* is the better temperature regressor for the Bike-Sharing Rental Model for casual users.

Not only that, the *atemp* model has a lower residual standard error at 1.229 while the *temp* model has 1.233. Also, the hourly temperature model has a higher adjusted R-squared with a value of 0.6013 while its comparison has a value of 0.5986. Hence, the *atemp* model can better explain the causality between the regressors and the number of casual users. It is also worth mentioning that the PRESS statistic for the *atemp* model is lower than the *temp* model with a difference of 180. Hence, the *atemp* model has a better prediction model than its counterpart.

Unfortunately, from the diagnostics perspective, the plots are very similar between the two models, that determining which model is better based on diagnostic visuals alone is very difficult. Please see **Exhibit D.1** and **Exhibit D.2** for the summaries and diagnostic plots for the model.

Registered Model Comparison and Analysis

Both the temperature regressor are significant for each model. However, the *atemp* regressor has a larger F-statistic value compared to the *temp* regressor, by a small difference. With the same implications and reasons mentioned in the casual model analysis above, we can say that the *atemp* model is the better fit for the data.

Furthermore, the *atemp* model has a lower residual standard error of 3.01 than the *temp* model with a residual standard error of 3.014. Also, the *atemp* model explains the data better with an adjusted R-squared value of 0.455, while the *temp* model has an adjusted R-squared value of 0.4537. Therefore, we can conclude that the *atemp* model will show better results for the Bike-Sharing Rental Model for registered users. Hence, the atemp model will be used in this case.

Like the casual case, the differences of diagnostic plots between the *atemp* and *temp* models is very miniscule, again implying that determining which is better purely based on diagnostics is very difficult. Please see **Exhibit D.3** and **Exhibit D.4** for the summaries and diagnostic plots for the model.

Other Findings and Adjustments

Why are the models so bad at explaining quality?

We can see that the adjusted R-squared for the casual model is around 0.6 while the adjusted R-squared for the is around 0.35. This is beyond what we expected and want. Therefore, we decided to think outside the data set, about other variables or factors that could affect our models. There could be many other omitted variables that we do not know about. However, we have some real situations that could explain our problem and the residuals between our model.

First, we do not know the number of available bikes at any given time. For example, there could be 100 interested people willing to try the service. However, there could only be 50 bikes available on one day and 100 bikes available the next day. Hence, only the data will only have 50 users on the first day and 100 on the next day.

Second, we do not know where the bikes are placed during their operational service. More bikes could be scattered in a residential area than in a business district on any given day. Hence, there will be more chances that people will be using bikes in a neighborhood than in a highrise complex. This would certainly affect the numbers as kids could enjoy the bikes during their free time near their house.

Third, we do not know if any advertisement for the bike-sharing rental service was displayed at any given time. We all know that advertisement is important in a business sense. This may affect the number of users at a given moment since it would create exposure of the service to a general audience. Thus, promoting the concept of the service could get people to use the service.

Fourth, we do not know if there is a promotion or a decrease in price happening on any of these days. The promotion could be an incentive for the public to try to use the service. The reduced paywall to use the bikes will certainly attract more users.

Note that the third and fourth points could increase in number when the weather condition is optimal for biking. For instance, it would be much more logical to advertise the bike-sharing rental service during the summer and spring than during the winter. Therefore, an interaction between *month* or *season* could have happened with these additional regressors if added.

Conclusion

Model Quality

While the models are successfully improved through transformations and proper variable selections, these final regression models that we deemed as optimal are still below expectations in terms of quality. As discussed previously, the relatively medium to low values of adjusted R-squared of the final models imply that the number of casual and registered users are not explained well enough by the existing explanatory variables in their respective models, caused by multiple factors such as omitted other factors, bike availability, bike placements, advertisements, and pricings.

Issues and Recommended Solutions

The main issue of the poor quality of models stems from the fact that the dataset has a low variety of explanatory variables, only containing variables regarding time and weather. This birthed the assumption that factors regarding time and weather are the only factors that matter, when in real life, a lot more factors take place, such as economic and supply factors. Furthermore, the data involved is taken throughout a period of time (2 years). This adds dynamicity to the number of rented bikes, or in other words, there might be certain weeks or days in the 2 years when the data was taken, where the number of bike rented suddenly shoots up, or falls down, due to certain events happening in the area that we might not know about, which adds complexity and unexplained parts of the dataset.

One solution is to expand the data, introducing more variables regarding other factors that might affect the number of rented bikes, including but not limited to economic-related variables, availability and placements of bikes, and whether an event or phenomenon is happening in the area at a certain day or week, treated as a categorical variable. The introduction of more possibly significant variables to the dataset might explain the unexplained part of the rented bike values, increasing the adjusted R-squared values.

It also should be noted that we split the models into casual and registered users. It is not clear how these two are differentiated, and upon first look at the dataset, the proportion of casual and registered users in a specific hour are not consistent throughout the dataset. Although the factors in the dataset are intuitively significant towards the total number of bikes rented, they might not necessarily affect whether a bike renter is categorized as casual or registered. This causes a decrease of adjusted R-squared values and overall quality of both casual and register models, when the other rent count variable is removed.

A possible solution to this is to fit a model that combines the two rent counts (casual and registered) into one combined and totaled number of bikes rented as the response variable. There is a possibility that the quality might improve when the response variable considers the whole total of bikes rented.

The final issue is the initial distribution of the number of bikes rented. We used a multiple linear regression method with the assumption that the number of bikes rented is normally distributed. From the initial diagnosis of the non-transformed data, we concluded that the data was very right-skewed. This opens the possibility that the data was not distributed normally in the first place, so Box-Cox transformation, while did improve the model fit, might not be the best transformation to apply. A solution is to figure out the actual distribution before transforming the data, and then fit a generalized linear model appropriate for the distribution, using the glm package in R, which might be a better fit than our current results.

Final Remarks

Throughout the whole process of obtaining the best possible model fit for the dataset, we had to decide between possible options in multiple steps along the way. In the model selection, we decided that the linear option is a better fit than the quadratic option in both casual and registered models. Both models have also undergone a Box-Cox transformation, specifically transforming the response variable using the Box-Cox formula with their respective optimum lambda values. The variable selection process suggests that no explanatory variables are removed, due to the high significance of all the variables after performing a LASSO selection in both casual and registered models. This answers our first research question, which asks the most appropriate set of explanatory variables to predict bike rental numbers (which is the set that includes all of them).

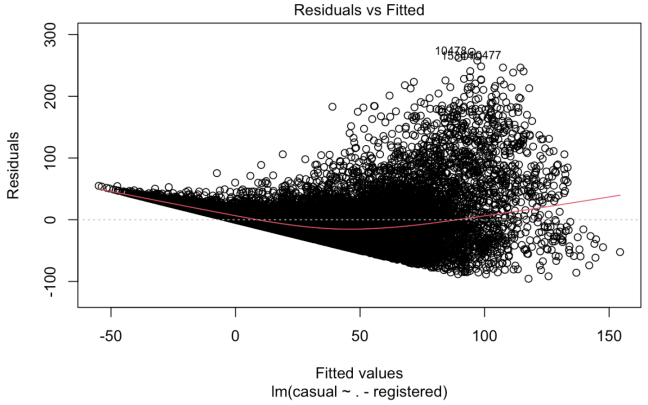
We also compared the models using exclusively either *atemp* or *temp* as the temperature variable. After examining and comparing their diagnostics and summaries, we concluded that the model using the *atemp* variable fits the data better than the *temp* variable for both casual and registered models. This implies that the ‘feels like’ temperature matters more than the actual temperature on determining how many people rent a bike in a period of time.

Although we did find the best possible model, the final quality of the models still seem to not be that good, based on the adjusted R-squared values. As discussed before, there are a lot of factors that cause this problem, and possible solutions do exist to solve, or at least reduce this problem, so that a better model can be obtained to better explain and predict the response variable of bike rental numbers.

Exhibits

Exhibit A.1 - Initial Casual Linear Model Assessment

```
Residual standard error: 36.38 on 17366 degrees of freedom
Multiple R-squared:  0.4559,     Adjusted R-squared:  0.4555
F-statistic:  1212 on 12 and 17366 DF,  p-value: < 2.2e-16
```



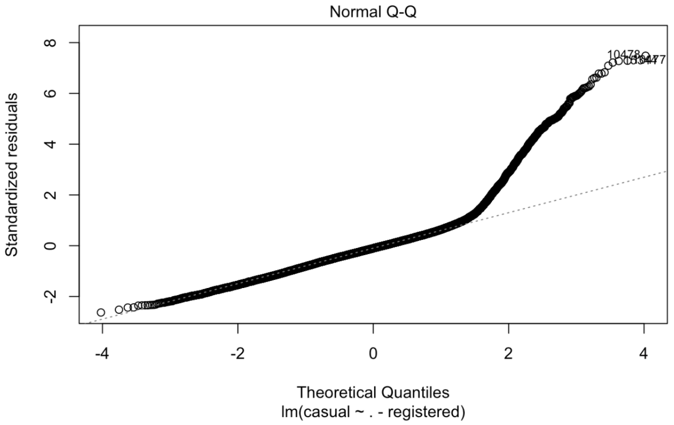
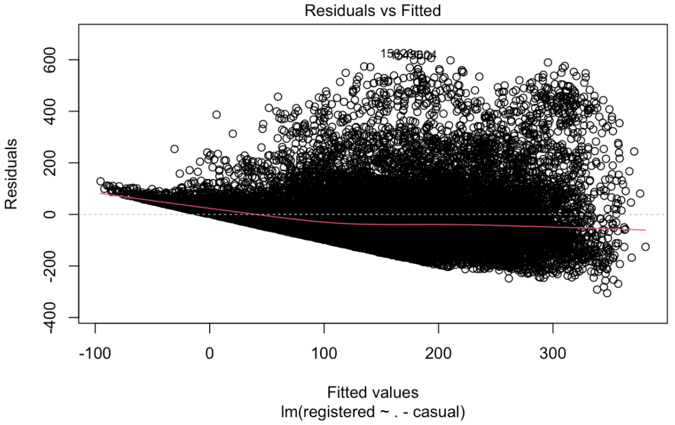


Exhibit A.2 - Initial Registered Linear Model Assesment

```
Residual standard error: 123.4 on 17366 degrees of freedom
Multiple R-squared:  0.3355,     Adjusted R-squared:  0.335
F-statistic: 730.6 on 12 and 17366 DF,  p-value: < 2.2e-16
```



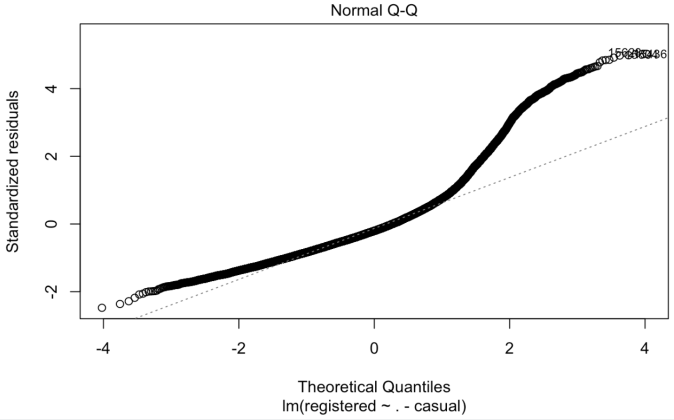


Exhibit B.1 - Initial Casual Quadratic Model Assesment

```
Residual standard error: 34.27 on 17361 degrees of freedom
Multiple R-squared:  0.5174,     Adjusted R-squared:  0.5169
F-statistic:  1095 on 17 and 17361 DF,  p-value: < 2.2e-16
```

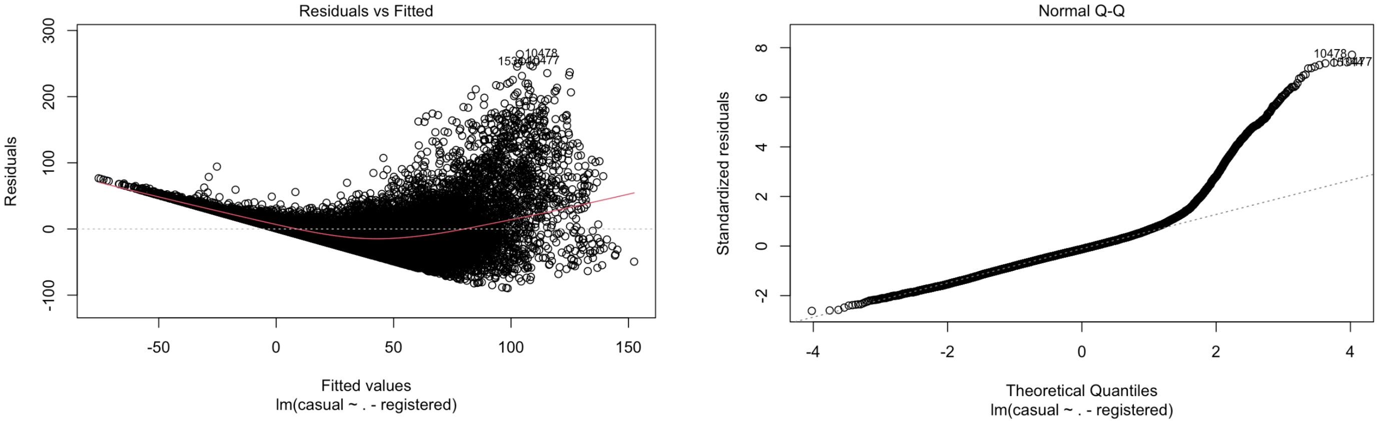


Exhibit B.2 - Initial Registered Quadratic Model Assesment

```
Residual standard error: 115 on 17361 degrees of freedom
Multiple R-squared:  0.4229,     Adjusted R-squared:  0.4224
F-statistic: 748.5 on 17 and 17361 DF,  p-value: < 2.2e-16
```

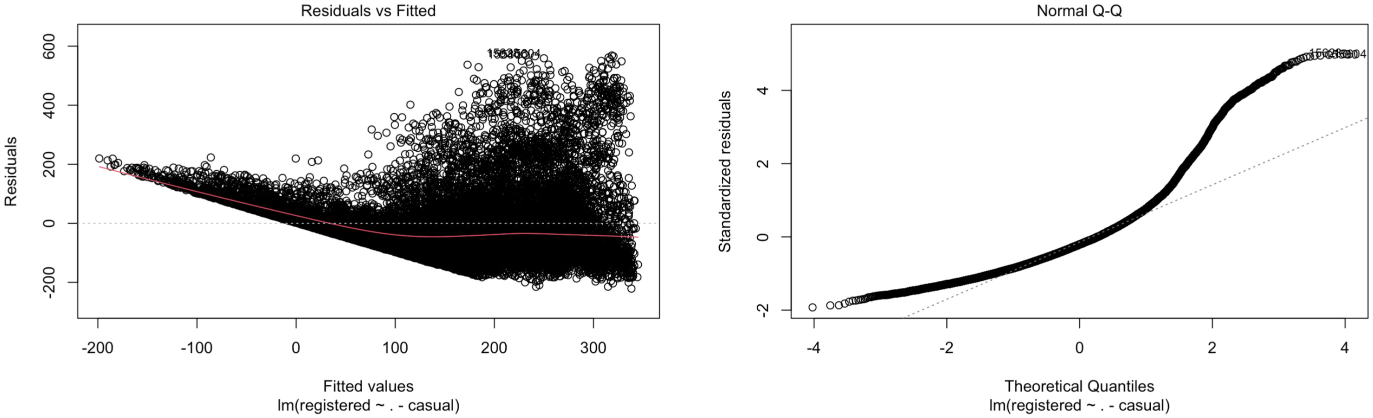


Exhibit C.1 - VIF Values of All Variables (Casual)

```
season         yr       mnth         hr    holiday    weekday
  3.500344   1.011587   3.285102   1.122967   1.081564   1.015767
workingday weathersit       temp      atemp        hum  windspeed
  1.073735   1.280636  43.737818  43.964035   1.525502   1.198389
```

Exhibit C.2 - VIF Values of All Variables Except *temp* (Casual)

```
    season         yr       mnth         hr    holiday    weekday
  3.499576   1.011470   3.285051   1.122679   1.080688   1.013937
workingday weathersit      atemp        hum  windspeed
  1.073458   1.279314   1.171256   1.520035   1.141924
```

Exhibit C.3 - VIF Values of All Variables Except a*temp* (Casual)

```
    season         yr       mnth         hr    holiday    weekday
  3.497427   1.011572   3.284558   1.121382   1.080560   1.013850
workingday weathersit       temp        hum  windspeed
  1.073672   1.278094   1.165229   1.520979   1.140001
```

Exhibit C.4 - VIF Values of All Variables Except *season* (Casual)

```
        yr       mnth         hr    holiday    weekday workingday
  1.011406   1.090735   1.122369   1.080214   1.015373   1.073696
weathersit       temp      atemp        hum  windspeed
  1.280136  43.728229  43.927403   1.523823   1.195957
```

Exhibit C.5 - VIF Values of All Variables Except *mnth* (Casual)

```
    season         yr         hr    holiday    weekday workingday
  1.162201   1.011524   1.121696   1.079505   1.014882   1.073598
weathersit       temp      atemp        hum  windspeed
  1.280550  43.737128  43.956742   1.519730   1.198358
```

Exhibit D.1 - Atemp Casual Linear Model Assesment (After Box-cox and LASSO)

```
Residual standard error: 1.229 on 17369 degrees of freedom
Multiple R-squared:  0.6015,     Adjusted R-squared:  0.6013
F-statistic:  2913 on 9 and 17369 DF,  p-value: < 2.2e-16
```

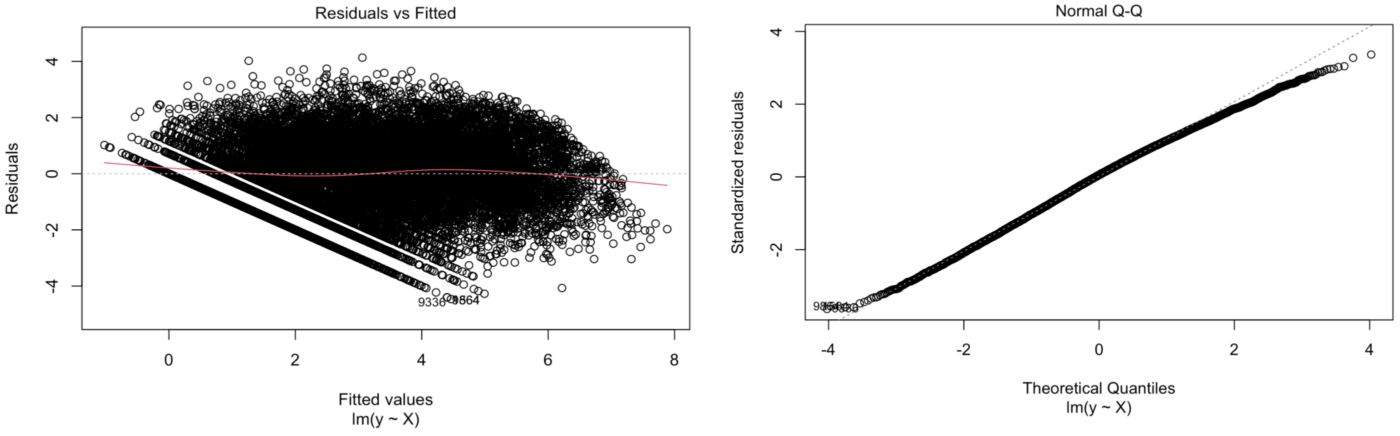


Exhibit D.2 - Temp Casual Linear Model Assesment (After Box-cox and LASSO)

```
Residual standard error: 1.233 on 17370 degrees of freedom
Multiple R-squared:  0.5987,     Adjusted R-squared:  0.5986
F-statistic:  3240 on 8 and 17370 DF,  p-value: < 2.2e-16
```

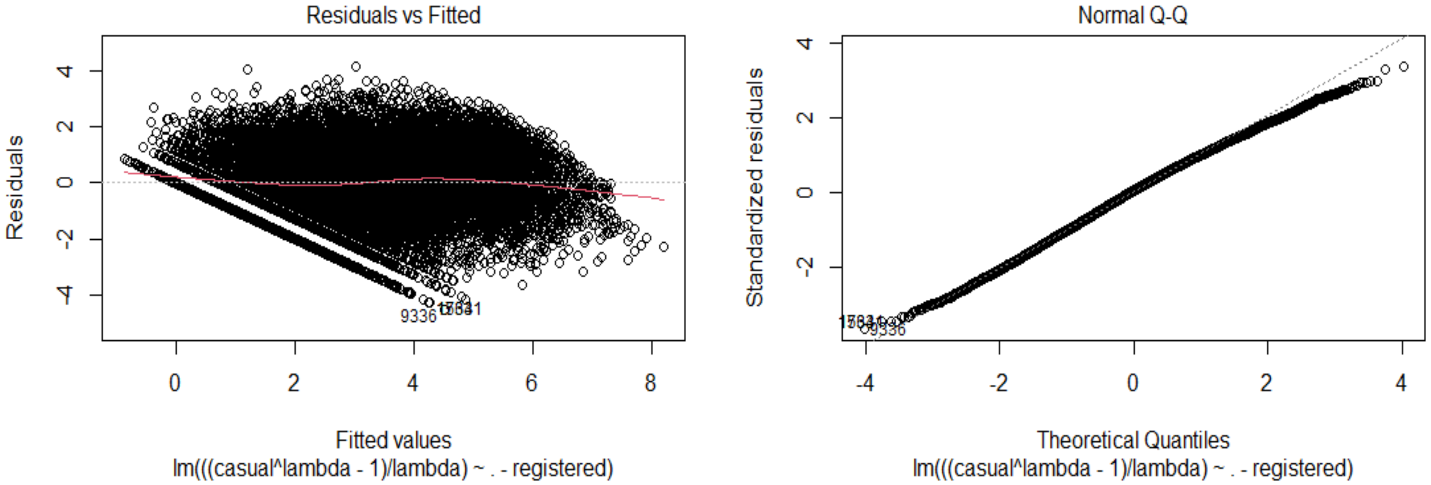


Exhibit D.3 - Atemp Registered Linear Model Assesment (After Box-cox and LASSO)

```
Residual standard error: 3.01 on 17368 degrees of freedom
Multiple R-squared:  0.4553,     Adjusted R-squared:  0.455
F-statistic:  1452 on 10 and 17368 DF,  p-value: < 2.2e-16
```

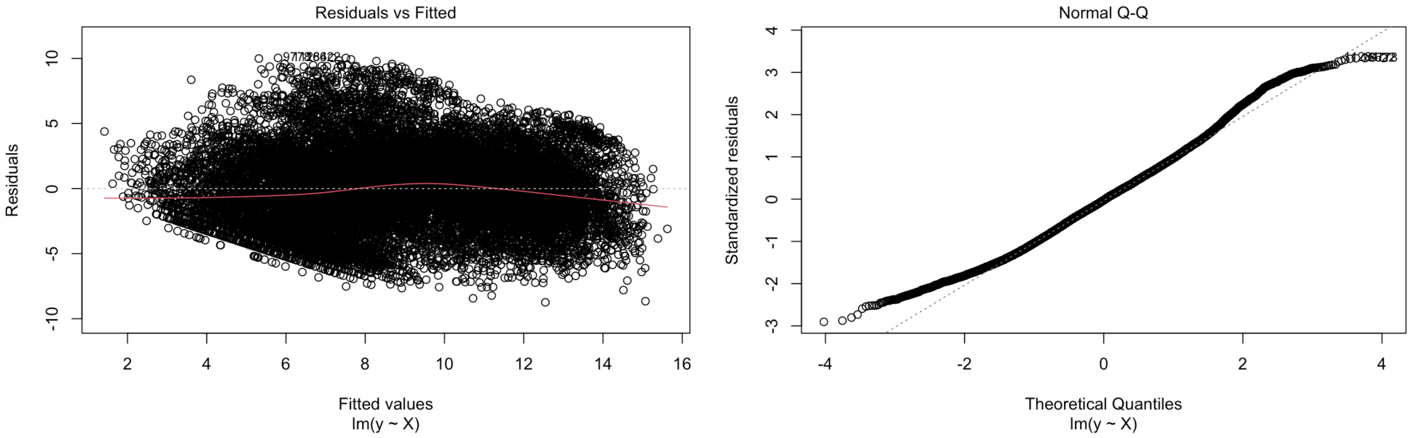


Exhibit D.4 - Temp Registered Linear Model Assesment (After Box-cox and LASSO)

```
Residual standard error: 3.014 on 17368 degrees of freedom
Multiple R-squared:  0.454,      Adjusted R-squared:  0.4537
F-statistic:  1444 on 10 and 17368 DF,  p-value: < 2.2e-16
```

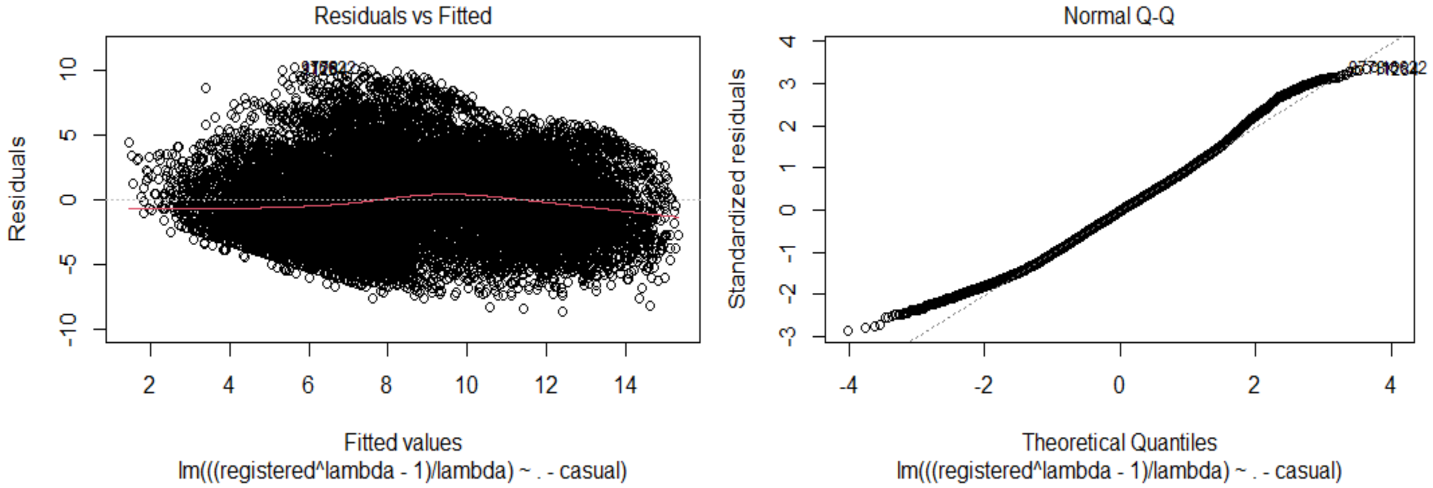


Exhibit D.5 - Atemp Casual Quadratic Model Assesment (After Box-cox and LASSO)

```
Residual standard error: 1.062 on 17361 degrees of freedom
Multiple R-squared:  0.7024,     Adjusted R-squared:  0.7021
F-statistic:  2411 on 17 and 17361 DF,  p-value: < 2.2e-16
```

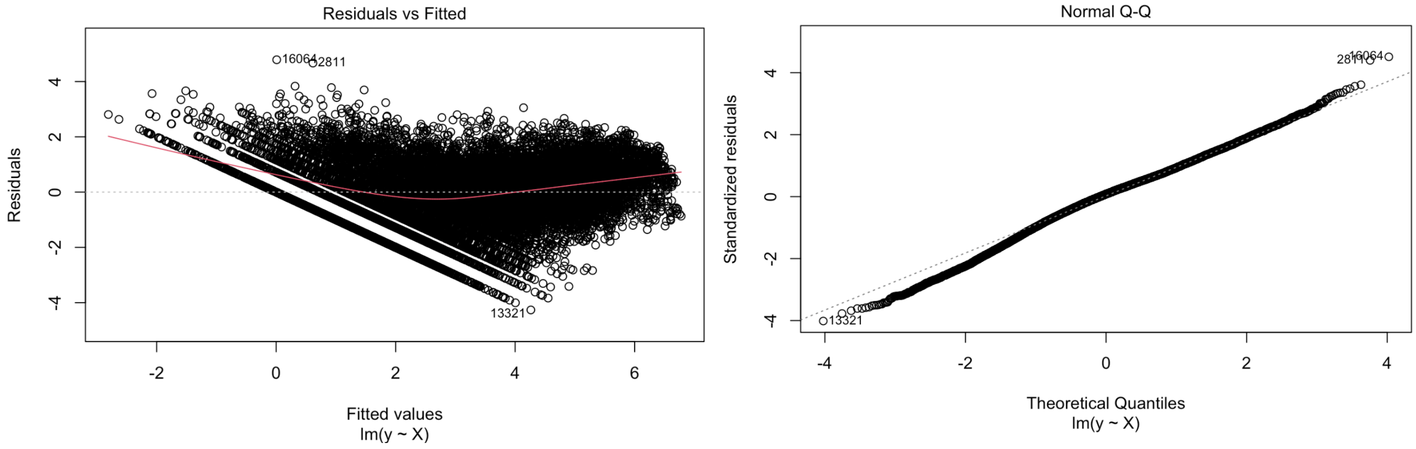
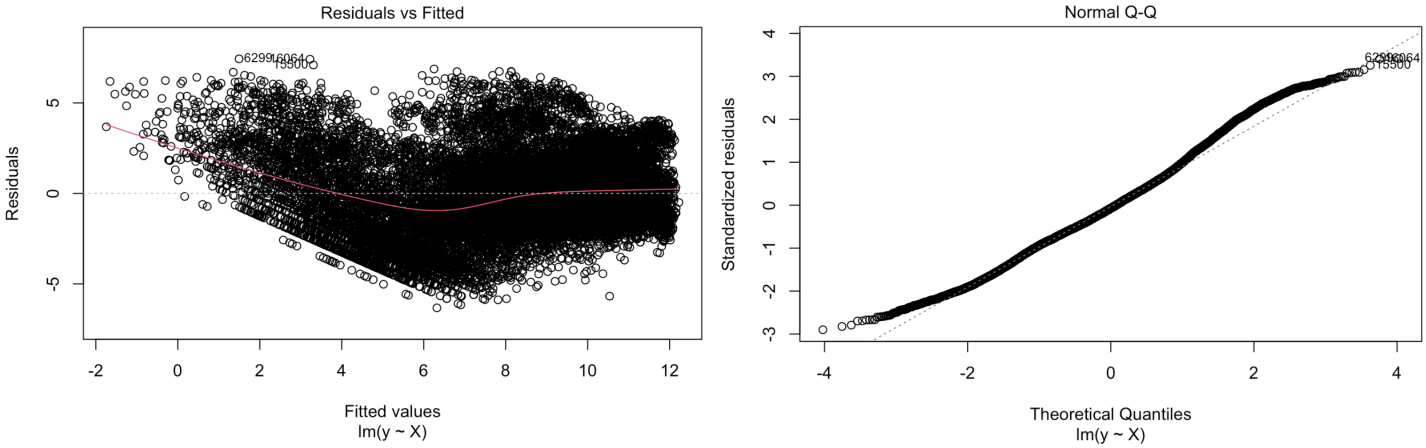


Exhibit D.6 - Atemp Registered Quadratic Model Assesment (After Box-cox and LASSO)

```
Residual standard error: 2.18 on 17361 degrees of freedom
Multiple R-squared:  0.5958,     Adjusted R-squared:  0.5954
F-statistic:  1505 on 17 and 17361 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ X)
Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(y ~ X)