Linear Modeling of Analyzing Bike Sharing and Environmental Factors

Yuge Xue, Zihan Zhao, Linan Zhang, Yifan Zhang, James Bishop, Zhiying Yip

Abstract

In this project, linear regression model analysis has been conducted on Bike Sharing Dataset. The Dataset contains data from the Capital Bike share system in Washington D.C. The goal of this project is to determine whether various environmental factors have direct or in-direct correlation with bike user amount. The quadratic regression model, hypothesis tests, and predictions are performed with R programming. The analysis corporate response variables "cnt", "casual" and "registered" in order to distinguish the different correlational relationship with different users. Covariates are all environmental factor variables. The conclusion of this project would be that the 2011 data we trained our model with does not produce accurate prediction for 2012 data.

Introduction

Bike sharing is a newly developed market with the goal to make environmental improvement and solve urban traffic problems. This new transportation method allows its customers to access the service as a one-time user by paying a small fee or choose the registered option with a discount for each ride. This new service provides people with an easy access to easy transportation methods at designated locations throughout the city. Customers can simply scan and pick up their bike at these stations and return them at another station whenever they want. Since the bloom of this service, it keeps on a pace of steady profiting and research can be done to predict customer amounts in order to help businesses make efficient bike distribution plans. Because of the nature of bike riding, this service hugely depends on the environmental and seasonal factors when determining user amounts. Additionally, casual users and registered users might think differently when facing the choice of using the service or not. This research is determined to find the most significant environmental factor that causes people to choose if they want to ride a bike while evaluating the difference between casual users and registered users and predict potential user amount with this finding.

Background

Linear model is a fundamental approach when determine the relationship between one or more models. Creating a model and training the model with 2011 data, this research would validate the result with 2012 data. This research includes multiple linear regression model, standard residual comparison and etc., hoping to provide the most accurate result possible.

The datasets we used for this model was created by Hadi Fanaee-T of Laboratory 50 of Artificial Intelligence and Decision Support (LIAAD) of 51 the University of Porto (3). The core data set is from the 52 Capital Bikeshare system in Washington D.C., USA made 53 publicly available at http://capitalbikeshare.com/system- 54 data. Weather information was extracted from 55 http:/ /www.freemeteo.com. Holiday information was 56 extracted from http://dchr.dc.gov/page/holiday-schedule.

Variable	Dataset Attribute	Description
Count*	cnt	Number of rides counted per day
Record Index	instant	index of the record
Season	season	Coded 1 → 4 for each season starting in Spring
Month	mnth	Coded 1 → 12 each month starting in Jan.
Holiday	holiday	One-hot encoding of a federal holiday
Weekday	weekday	0 to 7 coding of days of the week starting on Sunday
Working Day	workingday	One-hot encoding of whether it is a work day (no holidays or weekends).
Weather Situation	weathersit	qualitative ratings of weather: 1 (clear, good weather), 2 (misty or cloudy), 3 (light rain or snow), 4 (heavy rain or snow or ice)
Temperature	temp	Normalized temperature in Celsius
Absolute Temperature	atemp	Normalized feeling temperature
Date	dteday	String of the date in the format "YYYY-MM-DD"
Humidity	hum	relative percent humidity
Wind Speed	windspeed	normalized wind speed

Figure 1

Modeling and Analysis

First of all, how data are being collected and what each variable represents in shown in Figure 1. This research focuses on environmental factors and their correlation with user amount (casual and registered), but many variables in the data collection could be interpreted as significant environmental variables. Thus, the initial covariates were chosen to be every variable in the dataset.

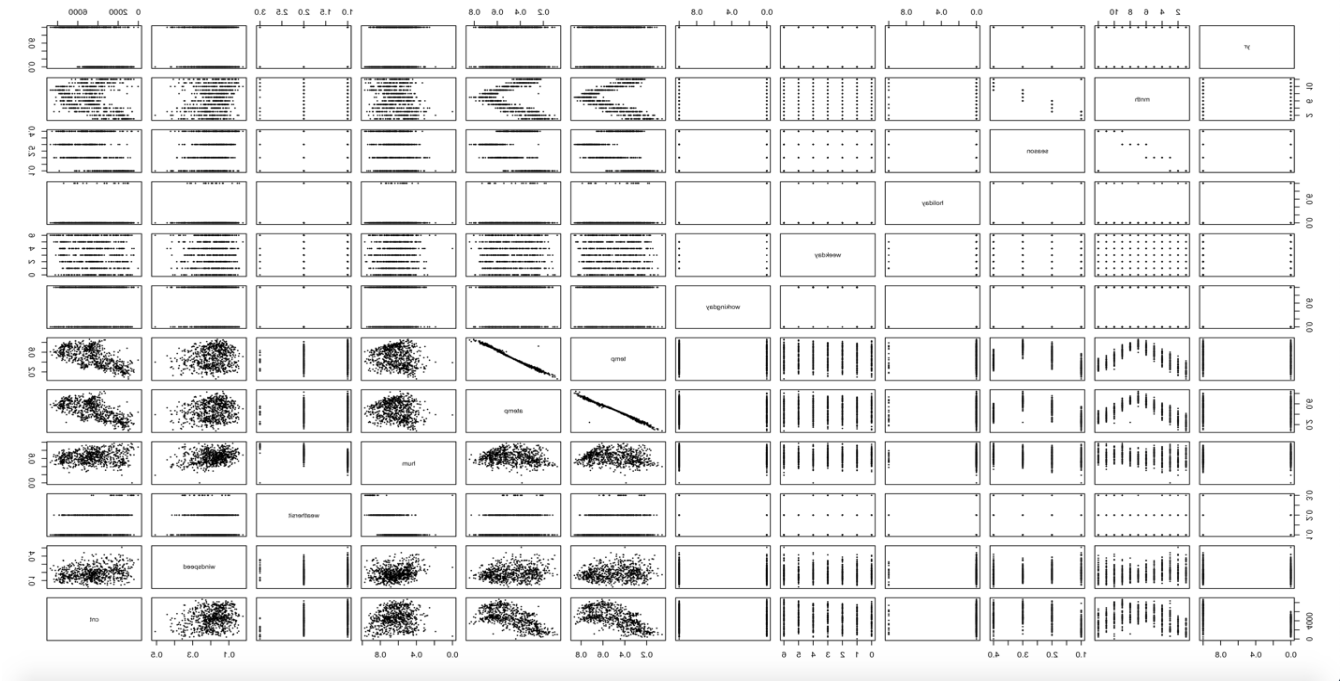


Figure 2

Then, response variables were chosen to be “cnt”. Initially, the linearity of each variable against each other is represented by the scatterplot matrix generated by R’s built in function pairs(). In figure 2, the relationship between each variable and the response variable are represented. Clearly, “temp” and “atemp” has a strong linear relationship and it is predictable by the description of the two datasets. Also, in this plot we can clearly see that there are linear or non-linear relationship between “mnth”, “season”, “weathersit”, “temp” and “atemp”. “mnth” and the response variables shows non-linear relationship. However, the humidity factor “hum” and response variables doesn’t show strong enough relationship in these matrices, further analysis is needed.

To further understand the finding and determine the best covariates that correlates with response variable not only “cnt”, but differentiated as “casual” and “resgistered”, correlation matrices are generated. In these graph, variable “season” has not been taken into consideration because the indication overlaps with “weathersit” and “temp”. The covariates are still being compared with three response variables in three different graphs, show in Figure 2, 3, and 4. This correlation matrix indicates the exact correlation coefficient between each candidate covariates. The correlation coefficient, also called R value is a value help determine how strong is the linear relationship between two variables. The graphs indicate both “temp” and “atemp” have relative strong correlation with all three response variables. However, since they also have strong positive correlation with each other, these two variables may cause multicollinearity. Therefore, only one can be used as a covariate. Looking at data show in the correlation matrix, there is no significant difference between “temp” and “atemp” and “temp” was kept because it is a more accurate temperature measurement.

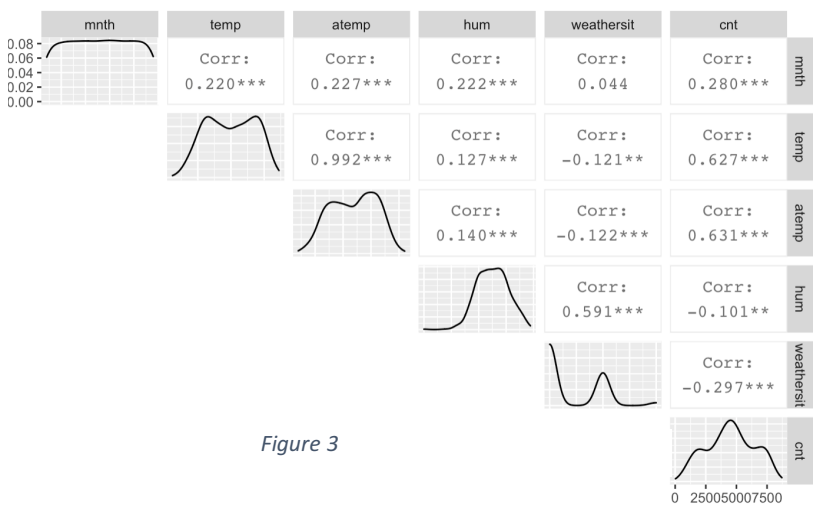


Figure 3

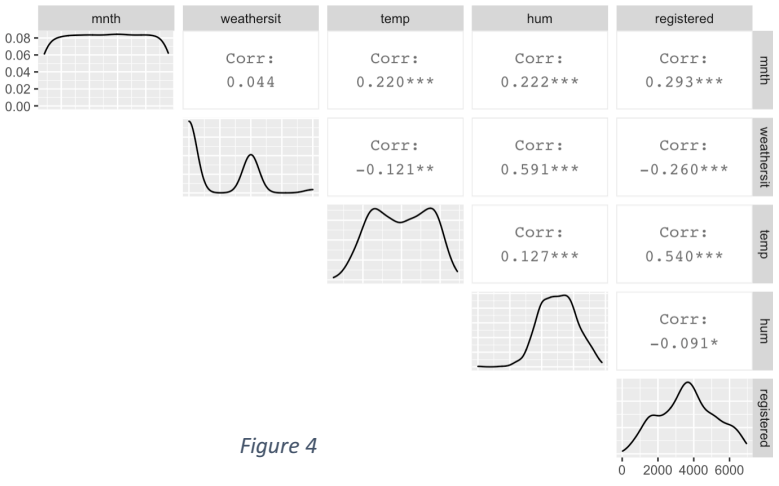


Figure 4

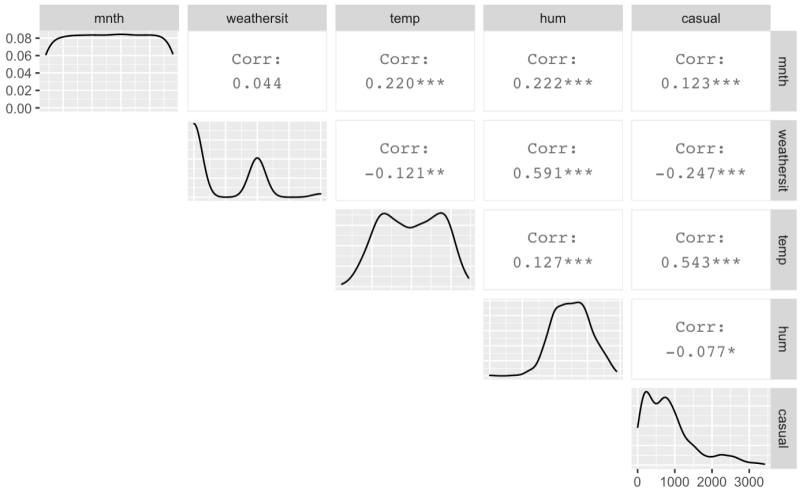


Figure 5

```
Call:
lm(formula = cnt ~ hum + mnth + temp + weathersit)

Residuals:
    Min      1Q  Median      3Q     Max
-3967.8 -1051.1  -172.4  1075.1  4143.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2604.57     263.10   9.899  < 2e-16 ***
hum         -1601.16     476.16  -3.363 0.000813 ***
mnth          103.97      15.78   6.588 8.58e-11 ***
temp         6155.62     299.82  20.531  < 2e-16 ***
weathersit   -589.20     122.15  -4.823 1.72e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1403 on 726 degrees of freedom
 ultiple R-squared:  0.478,     Adjusted R-squared:  0.4751
r-statistic: 166.2 on 4 and 726 DF,  p-value: < 2.2e-16
```

Figure 6

From figure 6, MLR model between response variable “cnt” and covariates “season”, “mnth”, “temp” and “weathersit” is shown. **R-squared is equal to 0.478, which means that approximately 48% variation of cnt can be explained by model with mnth, hum, temp and weathersit.** P-value shown on the last line is for the overall test of significance of MLR model. It tests the null hypothesis that all model coefficients are 0. Since **p-value in the figure is less than 2.2e-16, null hypothesis can be rejected at a = 0.01 or more.** The same result can be seen from the p-value of each covariates. Each p-value of covariates tests the null hypothesis that corresponding covariate’s coefficient is 0. Since all of those p-values are smaller than a = 0.01, so at this significance, each null hypothesis can be rejected. Residual standard error shows how far the observed values of “cnt” are from the prediction values Coefficient estimates are also shown in figure 5. From that multi-linear regression function of “cnt” and covariates can be written as

prediction(cnt) = 2604.57+103.97*mnth-1601.16*hum+6155.62*temp589.20*weathersit.

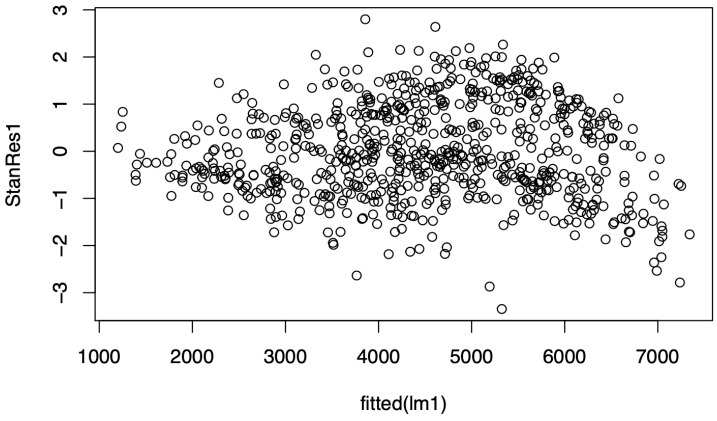


Figure 7

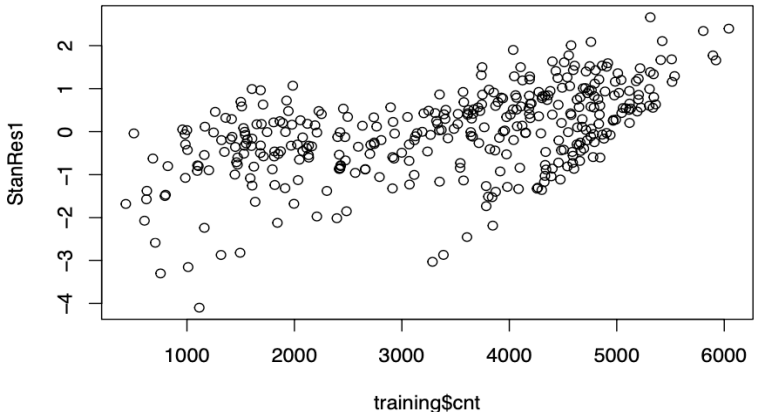


Figure 8

Each coefficient estimate is based on the adjusting or controlling of other covariates. Coefficient estimate of interception gives the mean value of “cnt” which is response variable(Y) when all covariates(Xs) are 0.

Furthermore, the t-values can also be used for testing the null hypothesis or confidence interval for model coefficients. From figure 7 below, the relationship between Y and Xs can be determined approximately linear, and the variation looks constant. From figure 8, data point is shown randomly separated at the first half of the diagram, but not at the end. Y, which is cnt, given Xs, which are “season” “mnth”, “temp” and “weathersit”, is approximately normal, except the end of the diagram, shown from figure 9.

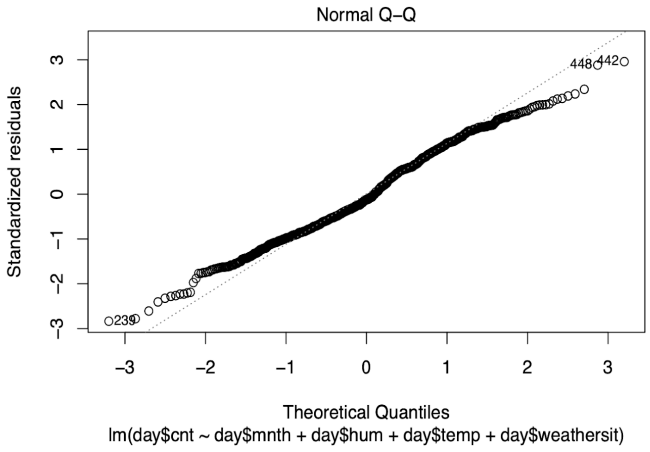


Figure 9

```
##
## Call:
## lm(formula = day$registered ~ day$mnth + day$hum + day$temp +
##     day$weathersit)

##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3208.8  -947.9  -157.5  1050.9  2762.6
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2322.15     231.67  10.024  < 2e-16 ***
## day$mnth          98.56      13.90   7.092 3.14e-12 ***
## day$hum        -1299.53     419.28  -3.099 0.002014 **
## day$temp        4177.38     264.00  15.823  < 2e-16 ***
## day$weathersit  -402.76     107.56  -3.745 0.000195 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1236 on 726 degrees of freedom
## Multiple R-squared:  0.3761, Adjusted R-squared:  0.3726
## F-statistic: 109.4 on 4 and 726 DF,  p-value: < 2.2e-16
```

Figure 10

For MLR model for Y as “registered”, and Xs, “mnth”, “hum”, “temp” and “weathersit”, shown in figure 10, R-squared is 0.3761, which means that 37.6% variation of registered can be explained by model with “mnth”, “hum”, “temp” and “weathersit”. The overall p-value shown on the last line is less than 2.2e-16, which is smaller than a = 0.01 or more. **Therefore, for the null hypothesis that all model coefficients are 0 can be rejected when a = 0.01.** The same result of the null hypothesis on model coefficients can be reached from p-values of each Xs, since each of them is smaller than a=0.01. According coefficients shown in figure 10, a MLR function can be written as

prediction(registered)=2322.15+98.56* mnth-1299.53*hum+4177.38*temp-402.76*weathersit

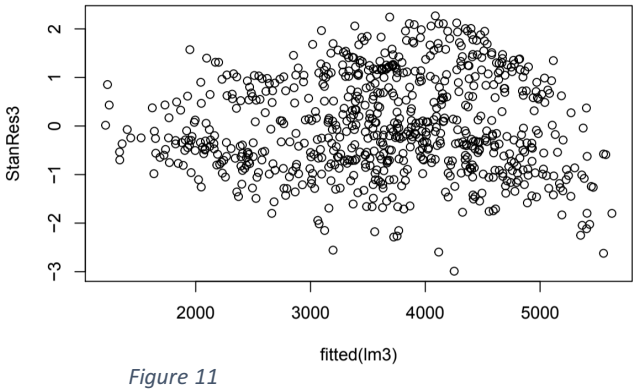


Figure 11

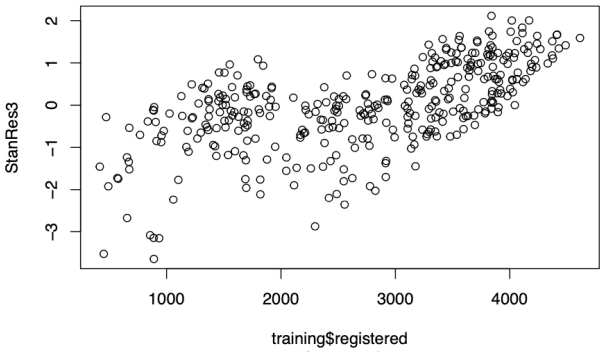


Figure 12

All of those coefficients are based on the adjusting or controlling of other covariates. Coefficient estimate of interception gives the mean value of registered when all covariates are 0. t-values can also be used for testing the null hypothesis or confidence interval for model coefficients. As Y being Registered customer and the relationship with “mnth”, “hum”, “temp” and “weathersit” is approximately normal, shown from figure 11. From figure 12, data point is shown randomly separated at the first half of the diagram, but not at the end. From figure 13 below, the relationship between Y and Xs can be determined approximately linear, and he variation looks constant, except at the end of the diagram.

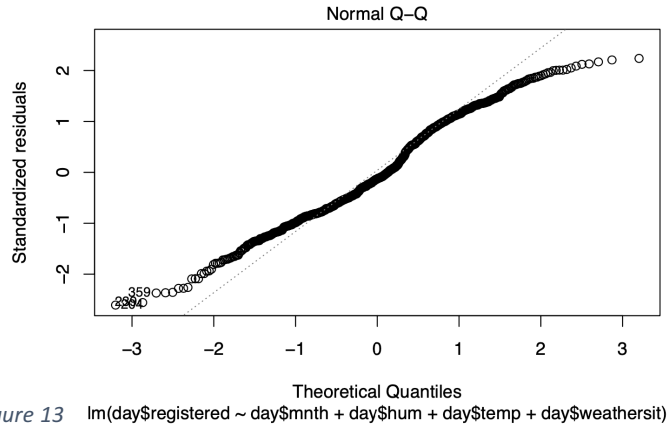


Figure 13

```
##
## Call:
## lm(formula = day$casual ~ day$mnth + day$hum + day$temp + day$weathersit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1075.0  -329.2  -156.2   139.0  2440.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)

## (Intercept)     282.423    105.547   2.676 0.007623 **
## day$mnth          5.413      6.331   0.855 0.392843
## day$hum        -301.626    191.020  -1.579 0.114765
## day$temp       1978.243    120.277  16.447  < 2e-16 ***
## day$weathersit -186.435     49.003  -3.805 0.000154 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 563 on 726 degrees of freedom
## Multiple R-squared:  0.3313, Adjusted R-squared:  0.3276
## F-statistic: 89.92 on 4 and 726 DF,  p-value: < 2.2e-16
```

Figure 14

The MLR model of Y as “casual”, and Xs, “mnth”, “hum” “temp” and “weathersit” is shown in figure 14. R-square is 0.3313, which means that 33.1% variation of casual can be explained by model with “mnth”, “hum”, “temp” and “weathersit”. The overall p-value is less than 2.2e-16, so the null hypothesis that all model coefficients are 0 can be rejected when a = 0.01 or more. P-values of each covariate can also show this result, except p-value of “mnth”. **P-value of “mnth” equals to 0.2717, and P-value of hum equals to 0.1148, which are bigger than a = 0.01, 0.05 or even 0.1.** Therefore, the null hypothesis that coefficient of “mnth” is 0 cannot be rejected, which means the **relationship between casual and “mnth” is undetermined.** the null hypothesis that coefficient of hum is 0 cannot be rejected as well, which also means the **relationship between casual and hum is undetermined**. According coefficients shown in figure 14, a MLR function can be written as

prediction(casual) = 282.423+5.413* mnth - 301.626*hum +1978.243*temp-186.435*weathersit

All of those coefficients are based on the adjusting or controlling of other covariates. Coefficient estimate of interception gives the mean value of registered when all covariates are 0. The t-values can also be used for testing the null hypothesis or confidence interval for model coefficients. Response variable “casual” given “mnth”, “hum”, “temp” and “weathsit” is not normal for this MLR, shown from figure 15. From figure 16, data point also shown it is a non-linear relationship. Figure 17 shows the relationship between Y and Xs can be concluded the same way, since the variation does not look constant, and data points are not randomly spreading, especially at the end of the diagram.

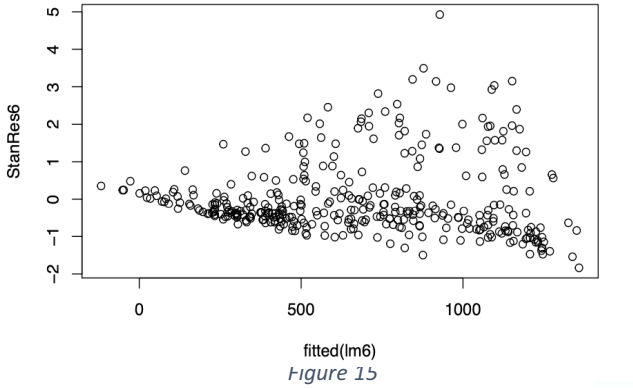


Figure 15

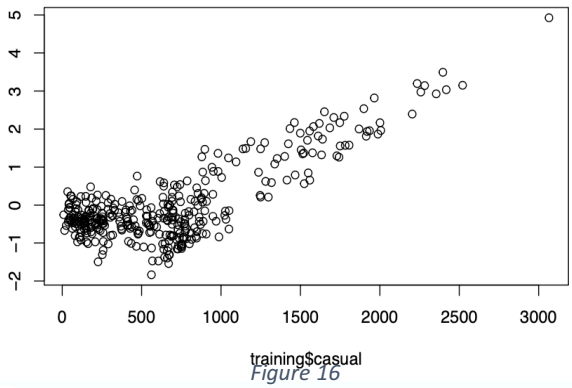


Figure 16

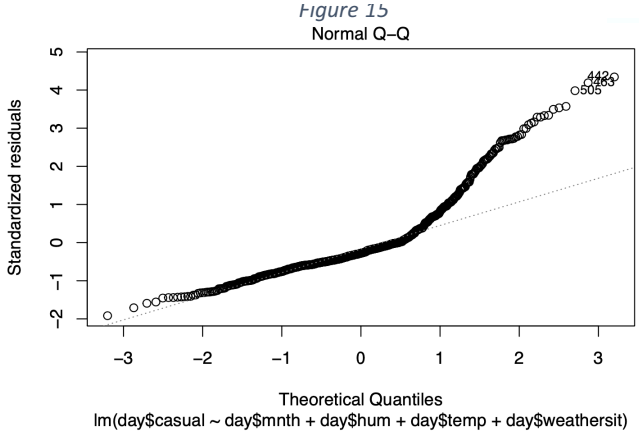


Figure 17

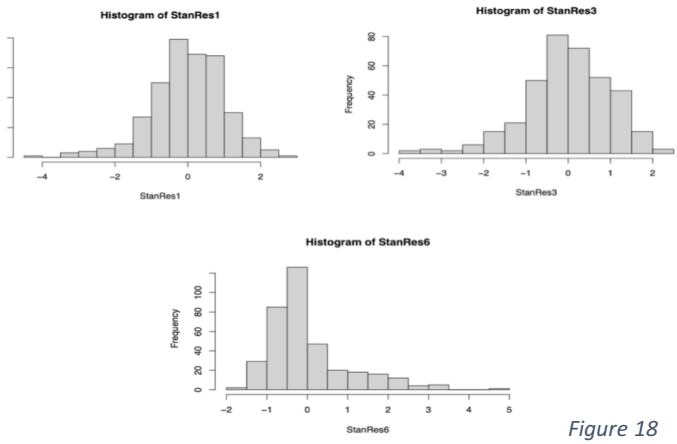


Figure 18

Also, the normality of three linear model can be seen from histogram. From figure 18, the diagram on the left corner shows the nearly normal distribution of linear model for “cnt”, and the diagram on the right corner shows the closely normal distribution of linear model for registered, except on the right side. However, the diagram at the bottom in figure 18 shows that linear model for casual is not normal. These conclusions correspond to results shown from qqplots previously. The

linear model for “cnt” and “registered” is acceptable, but they may not be optimal models, and the linear model for “casual” is not reasonable. Therefore, more models need to be applied.

Prediction:

From results of the Mean Square Error of validations, the models that give the lowest values among other models for each response variables “cnt”, “registered “and “casual” were used as prediction model. The Relative Mean Square Error for each response variable are 0.08163295, 0.09241478, and 0.3713214 respectively. Since the casual model has extremely high Relative Mean Square Error comparing to models for the other two response variables, this model is applied wls and log transformation. The prediction F statistics and graph are then generated. The quadratic model for predicting “cnt” is shown in figure 20 and the associated prediction graph is shown in figure 21

Predict(cnt) =-423.238-2219.062*hum+195.396*mnth+22437.868*temp-8.862*mnth^2-16890.683*temp^2-627.772*weathersit.

```
##
## Call:
## lm(formula = day$cnt ~ day$hum + day$mnth + day$temp + I(day$mnth^2) +
##     I(day$temp^2) + day$weathersit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3763.5 -1004.5   -75.3  1090.8  3689.1
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -423.238    372.391  -1.137    0.256
## day$hum         -2219.062    452.995  -4.899 1.19e-06 ***
## day$mnth          195.396    128.694   1.518    0.129
## day$temp        22437.868   1859.615  12.066  < 2e-16 ***
## I(day$mnth^2)      -8.862      9.364  -0.946    0.344
## I(day$temp^2)  -16890.683   1666.687 -10.134  < 2e-16 ***
## day$weathersit   -627.772    114.106  -5.502 5.22e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1305 on 724 degrees of freedom
## Multiple R-squared:  0.5502, Adjusted R-squared:  0.5465
## F-statistic: 147.6 on 6 and 724 DF,  p-value: < 2.2e-16
```

Figure 20

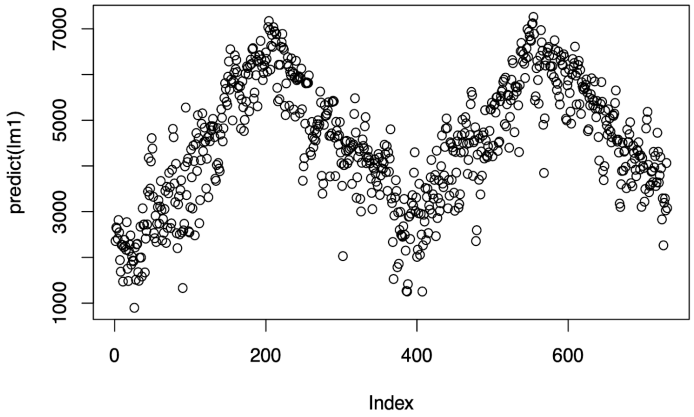


Figure 21

```
##
## Call:
## lm(formula = day$registered ~ day$hum + day$mnth + day$temp +
##     I(day$mnth^2) + I(day$temp^2) + day$weathersit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3152.2  -866.0  -112.0   920.4  2666.2
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        86.495    335.124   0.258    0.796
## day$hum         -1864.205    407.661  -4.573 5.66e-06 ***
## day$mnth          -13.726    115.815  -0.119    0.906
## day$temp        17655.086   1673.513  10.550  < 2e-16 ***
## I(day$mnth^2)       6.604      8.426   0.784    0.433
## I(day$temp^2)  -13250.431   1499.892  -8.834  < 2e-16 ***
## day$weathersit   -416.079    102.687  -4.052 5.63e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1174 on 724 degrees of freedom
## Multiple R-squared:  0.4385, Adjusted R-squared:  0.4338
## F-statistic: 94.22 on 6 and 724 DF,  p-value: < 2.2e-16
```

Figure 22

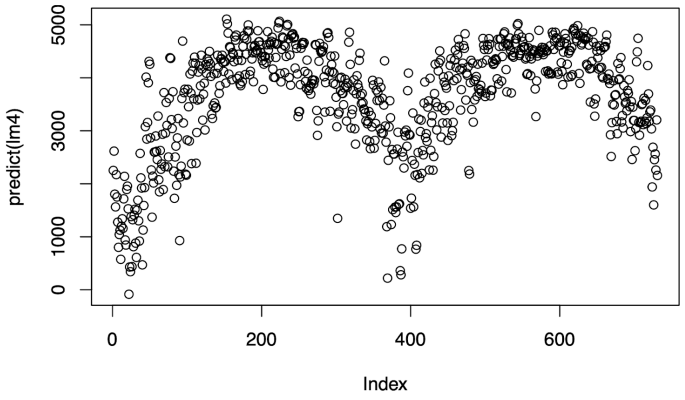


Figure 23

From figure 22, the quadratic model for predicting registered is shown as

Predict(registered)= 86.495-1864.205*hum-13.726*mnth+17655.086*temp+6.604*mnth^2-13250.431*temp^2-416.079*weathersit and the associated prediction graph is shown in figure 23

From figure 24, the exponential model for predicting casual is shown and the associated prediction graph is shown in figure 25. Because of the log transformation added to variable casual previously to achieve less MSE, in the forecast model, it has to be transformed back.

Predict(casual)=exp(3.26286-0.36371*hum+0.57768*mnth+9.04891*temp-0.04096*mnth^2-7.80091*temp^2-0.518*weathersit)

The normality of three models are shown from figure 25,26,27 in Appendix. Even they are not perfectly normal, with the lowest MSE value and similar normality with other models, those three models are the best among other models.

```
##
## Call:
## lm(formula = log(casual) ~ hum + mnth + temp + I(mnth^2) + I(temp^2) +
##     weathersit, data = training, weights = wts)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4359 -0.8925 -0.2289  0.8228  3.8353
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.262864   0.215138  15.166  < 2e-16 ***
## hum         -0.363709   0.256973  -1.415    0.158
## mnth         0.577679   0.088918   6.497 2.75e-10 ***
## temp         9.048911   1.195084   7.572 3.16e-13 ***
## I(mnth^2)   -0.040958   0.006304  -6.497 2.75e-10 ***
## I(temp^2)   -7.800913   1.021464  -7.637 2.05e-13 ***
## weathersit  -0.517814   0.065145  -7.949 2.48e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.218 on 358 degrees of freedom
## Multiple R-squared:  0.6876, Adjusted R-squared:  0.6824
## F-statistic: 131.3 on 6 and 358 DF,  p-value: < 2.2e-16
```

Figure 24

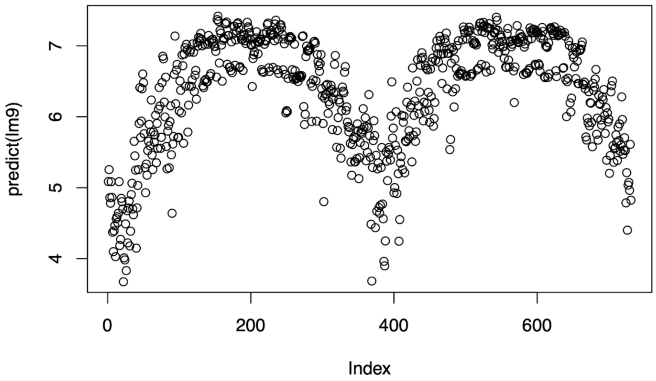


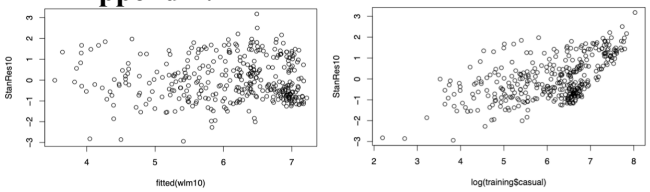
Figure 25

Discussion:

The goal of this project was to find relevant environmental variables that can accurately predict bike sharing's user amount. From the 2011 data, the conclusion is that environmental conditions have different influences on various kind of bike renters. For casual customers, temperature has the most influence on their decision of renting. For registered customers, temperature and humidity both have significant influence on their decisions.

However, the prediction of this model is not accurate since the prediction of 2012 user amount has been validated by comparing already known data, and the MSE is too large. According to the model for different kinds of customers, predictions for potential customers cannot be reached by this model. Other factors would need to be taken into consideration if further analysis needed to be done on the topic. The environmental factors are seemly important, but the condition of each year's economics, political environment, etc, would also play a huge role in bike sharing business.

Appendix:



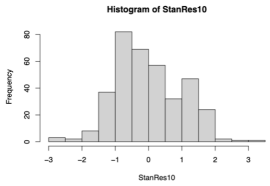
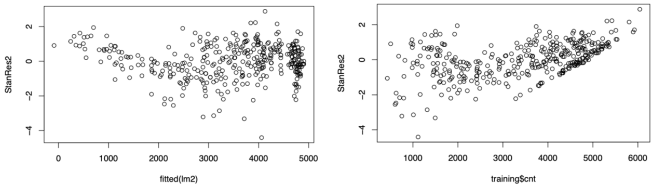


Figure 26



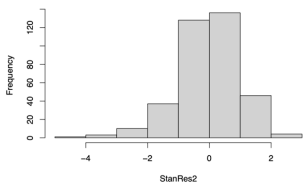
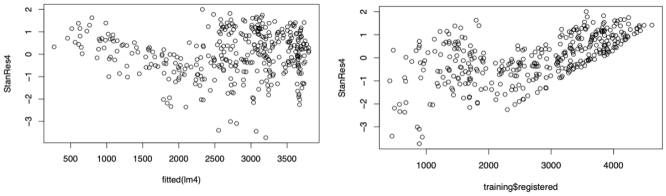


Figure 27



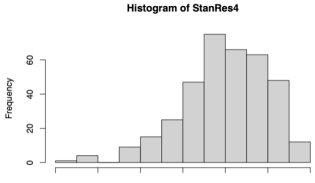


Figure 28